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

August 19, 2020

Contents

Projects

If you are interested to make a project presentation, please pick up of the subjects below. Please communicate your choice of the subject and discuss content with the instructor as soon as possible. First come first served. You may also suggest your own project for material which is relevant to the course but not covered in the class. You will need to prepare a Jupiter notebook presentation (in ipython or ijulia) for 10+5 minutes. We will have two presentation sessions, scheduled for Oct 22 and Dec 1 respectively during the regular class time. Oct 11 and November 22 are the last days to claim a project for the first and second sessions respectively.

List of suggested projects for the first session (Complex Analysis & Fourier Analysis):

1.1 Numerical Conformal mapping.

1.2 Complex numbers & analysis: AC electric circuit applications.

1.3 Laplace transform in systems engineering: linear-time-invariant and linear-time-varying systems.

1.4 Mellin transform and its applications.

1.5 Wavelets.

List of suggested projects for the second session (ODEs & PDEs):

2.1 Linear Stability/Instability in Fluid Mechanics: Kelvin-Helmholtz.

2.2 Susceptible-Infected-Susceptible (SIS) and Susceptible-Infected-Removed (SIR) of Epidemiology.

2.3 Sturm-Liouville Problem: Fokker-Planck equation (of statistical mechanics).

2.4 Wave equations and Eikonal (WKB) approximation of Classical Optics.

2.5 Nonlinear Schrödinger equation: solitons and integrability.

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 583), Theory (Math 527), and Algorithms (Math 575). 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). 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 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 sim-

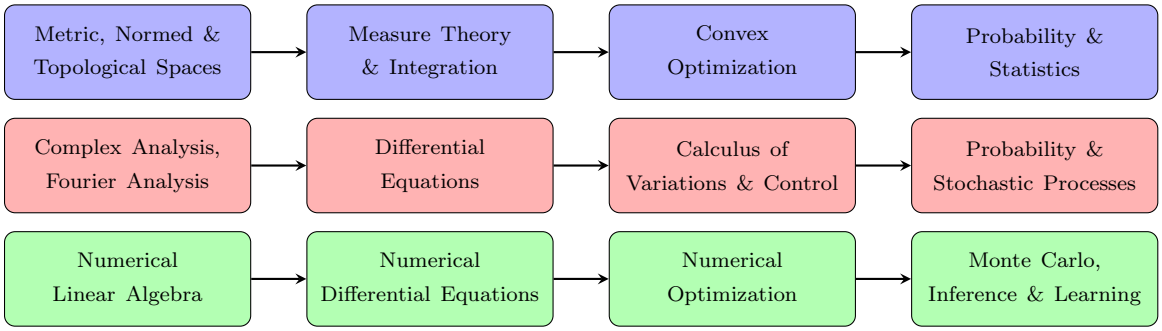


Figure 1: Topics covered in Theory (blue), Methods (red) and Algorithms (green) during the Fall semester (columns 1 & 2) and Spring semester (columns 3 & 4)

ulations 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. 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 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:

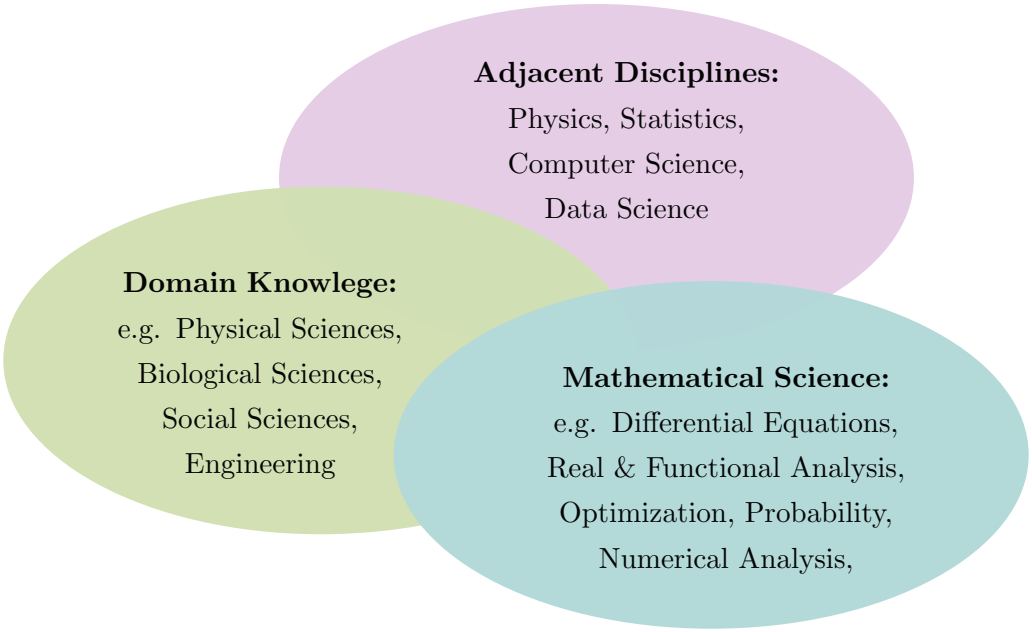


Figure 2: The key components studied under the umbrella of applied mathematics: (1) mathematical science, (2) domain-specific knowledge, and (3) a few 'math-adjacent' disciplines.

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 1

Complex Analysis

Complex analysis is the branch of mathematics that investigates functions of complex variables. A fundamental premise of complex analysis is that *most* of binary operations have natural extensions from real numbers to complex numbers. Furthermore, real-valued functions have natural extensions to complex-valued functions. Natural extensions of even the most elementary functions can lead to new and interesting behavior.

Complex-valued functions exhibit 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 in physics, (including the branches of hydrodynamics, thermodynamics, and particularly quantum mechanics), and engineering fields (such as aerospace, mechanical and electrical engineering).

1.1 Complex Variables and Complex-valued Functions

1.1.1 Complex Variables

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*, $i := \sqrt{-1}$. An *imaginary number* is any number that is a real multiple of the imaginary unit, for example $3i$, $i/2$ or $-\pi i$. A *complex number* is any number that has both a real and an imaginary component, and can therefore be represented by two real numbers, x and y, which we often write as $z = x + iy$.

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

Example 1.1.1. Let $z_1 = 4 + 3i$ and $z_2 = -2 + 5i$. Compute (a) $z_1 + z_2$ and (b) $z_1 - z_2$.

Solution.

(a) $z_1 + z_2 = (4 + 3i) + (-2 + 5i) = (4 + -2) + (3 + 5)i = 2 + 3i$

(b) $z_1 - z_2 = (4 + 3i) - (-2 + 5i) = (4 - -2) + (3 - 5)i = 6 - 7i$

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 1.1.1. 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 application of the definition $i^2 = -1$.

Example 1.1.2. Let $z_1 = 4 - 3i$ and $z_2 = -2 + 5i$. Compute (a) $z_1 z_2$, (b) $1/z_1$, (c) $1/z_2$, and (d) z_1/z_2.

Solution.

(a) $z_1 z_2 = (4 + 3i)(-2 + 5i) = -8 - 6i + 20i + 15i^2 = -23 + 14i$

(b) Note: $z_1^* = 4 + 3i$, and $z_1 z_1^* = (4 - 3i)(4 + 3i) = 16 + 12i - 12i - 9i^2 = 25$.
Therefore, $1/z_1 = (1/z_1)(z_1^*/z_1^*) = z_1^*/(z_1 z_1^*) = (4 + 3i)/25 = 4/25 + 3/25i$

(c) Note: $z_2^* = -2 + 5i$, and $z_2 z_2^* = (-2 + 5i)(-2 - 5i) = 4 - 10i + 10i - 25i^2 = 29$.
Therefore, $1/z_2 = (1/z_2)(z_2^*/z_2^*) = z_2^*/(z_2 z_2^*) = (-2 + 5i)/29 = -2/29 + 5/29i$

(d) $z_1/z_2 = (z_1/z_2)(z_2^*/z_2^*) = (z_1 z_2^*)/(z_2 z_2^*) = 7/29 - 26/29i$

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 addition of 2π.

$$x + iy \;\Leftrightarrow\; r\cos\theta + ir\sin\theta, \quad \text{where} \quad r = x^2 + y^2,\ \theta = \tan^{-1}(y/x)$$

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 $z_1 z_2 = r_1 r_2 \cos(\theta_1+\theta_2) + ir_1 r_2 \sin(\theta_1+\theta_2)$. 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

Definition 1.1.2. The *exponential function* is defined for imaginary arguments by

$$re^{i\theta} := r\cos(\theta) + ir\sin(\theta) = x + iy. \tag{1.1}$$

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

Example 1.1.3. Compute the polar representations of (a) $z_1 = 4-3i$ and (b) $z_2 = -2+5i$.

Solution.

(a) $r_1 = z_1 z_1^* = 5, \qquad \theta_1 = \tan^{-1}(3/4) \approx 0.64, \quad \Rightarrow \quad z_1 = 5e^{0.64i}$

(b) $r_2 = z_2 z_2^* = \sqrt{29}, \quad \theta_2 = \tan^{-1}(5/-2) \approx 1.95 \quad \Rightarrow \quad z_2 = \sqrt{29}e^{1.95i}$

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

Example 1.1.4. Find $\tilde{r}$ and $\tilde{\theta}$ such that the point $\omega = 1+5i$ can be written as $\omega = -1+\tilde{r}e^{i\tilde{\theta}}$

Solution. Given that $1+5i = -1+\tilde{r}e^{i\tilde{\theta}}$, solve 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 1.1.5. Express $z := (2+2i)e^{-i\pi/6}$ by its (a) cartesian and (b) polar representations.

Solution.

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

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

Definition 1.1.3. 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, that is the curve is simple if $z(t) \ne z(t')$ for $t \ne t'$. 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 1.1.6. Parameterize the following curves:

(a) The infinite horizontal line passing through $0 + i\pi$.

(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.

Solution.

(a) $x + \pi i$ for $-\infty < x < \infty$

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

(c) $\varepsilon e^{i\theta}$ for $0 \leq \theta \leq 2\pi$

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.

1.1.2 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, $f(z) =: u(x, y) + iv(x, y)$.

Example 1.1.7. Let $f(z) = \exp(iz)$ where $z = x + iy$. Express f as the sum $u + iv$ where u and v are real-valued functions of x and y.

Solution.

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

In equation (1.1) 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 incidently that $e^1 = 2.718\ldots$. This is not the only property we could have chosen to motivate the defintion e^z. We could have chosen to preserve any of the following properties:

- the function represented by the Taylor series $\sum z^n/n!$,
- the limiting expression $\lim_{n\to\infty}(1 + z/n)^n$,
- the solution to the ODE $z'(t) = z(t)$ subject to $z(0) = 1$.

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 yeilded the same results.

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

$$\cos(z) := \frac{e^{iz} + e^{-iz}}{2} \quad \text{and} \quad \sin(z) := \frac{e^{iz} - e^{-iz}}{2i} \tag{1.2}$$

Exercise 1.1.8. Find all values of $z \in \mathbb{C}$ satisfying the equation $\sin(z) = 3$.

Exercise 1.1.9. Investigate the asymptotic behavior of the complex-valued functions (a) $f(z) = \exp(z)$, (b) $f(z) = \sin(z)$, (c) $f(z) = \cos(z)$.

Example 1.1.10. Evaluate the functions (i) $f(z) = z^2$ and (ii) $g(z) = \exp(z+1)$ along the parameterized curves described in example 1.1.6.

Solution.

(a) For the infinite horizontal line passing through $0 + i\pi$.

(i) $f(x + i\pi) = (x + i\pi)^2 = x^2 - \pi^2 + 2\pi ix$ for $-\infty < x < \infty$.

(ii) $g(x + i\pi) = \exp(x + i\pi + 1) = -e^{x+1}$ for $-\infty < x < \infty$.

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

(i) $f(-1 + \rho e^{i\pi/3}) = (-1 + \rho e^{i\pi/3})^2 = 1 - 2\rho e^{i\pi/3} + \rho^2 e^{i2\pi/3}$ for $\rho < 0 < \infty$.

(ii) $g(-1 + \rho e^{i\pi/3}) = \exp(-1 + \rho e^{i\pi/3} + 1) = \exp(\rho\cos(i\pi/3) + i\rho\sin(i\pi/3)) = e^{\rho/2}\left(\cos(\rho\sqrt{3}/2) + i\sin(\rho\sqrt{3}/2)\right)$ for $\rho < 0 < \infty$.

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

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

(ii) $g(\varepsilon e^{i\theta}) = \exp(\varepsilon e^{i\theta} + 1) = \ldots$

Complex conjugates

Theorem 1.1.4. 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 1.1.11. Let us illustrate theorem 1.1.4 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 1.1.4 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 1.1.12. Use theorem 1.1.4 to show that the roots of a polynomial with real-valued coefficients of *arbitrary* order occur in complex conjugate pairs.

Exercise 1.1.13. Find all the roots of the polynomial, $z^4 - 6z^3 + 11z^2 - 2z - 10$, given that one of its roots is $2 - i$.

Exercise 1.1.14. Let $z_1 = x_1 + iy_1$ and $z_2 = x_2 + iy_2$. Show that if $\omega = z_1/z_2$, then $\omega^* = z_1^*/z_2^*$.

1.1.3 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 how we demonstrated how to parameterize curves in the complex plane in example 1.1.6 and how to evaluate a function along a parameterized curve in example 1.1.10. Consider example 1.1.10(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 1.1.15. Consider the example of $\omega = \sqrt{z}$. 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 $\sqrt{z}$, this translates into $\sqrt{r}\exp(i\theta/2+i\pi n)$, where therefore even and odd 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, so that $z = 0$ always stays on the left) direction 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. The point $z = 0$ is called a branch point of the second order of the two-valued $\sqrt{z}$ function.

Example 1.1.16. The generalization of example 1.1.15 to $\omega = z^{1/n}$ is straightforward. This function has n branches and thus $z = 0$ is an n^{th} order branch point.

Example 1.1.17. Another important example is $\omega = \log(z)$. 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.

To separate the branches one introduces cuts – lines which are forbidden to cross. After the introduction of appropriate branch cuts, each branch of a multi-valued, analytic function defines a single-valued function that is analytic everywhere except at the branch cut, where it is discontinuous. The choice of branch cuts need not be unique.

Remark. 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.

Definition 1.1.5. A multi-valued function $w(z)$ has a *branch point* at $z_0 \in \mathbb{C}$ if $w(z)$ is varies continuously along along a sufficiently small circuit surrounding z_0, but does not return to its starting values after one full circuit.

Definition 1.1.6. A *branch* of a multi-valued function $w(z)$ is a single-valued function that is obtained by restricting the image of the $w(z)$.

Definition 1.1.7. A *branch cut* is a curve in the complex plane along which a branch is discontinuous.

Example 1.1.18. Find the branch points of $\log(z-1)$, and sketch a set of possible branch cuts.

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 $\log(z-1)$. Similarly we observe that $z = \infty$ is also a branch point (thus infinite branch point) and there are no others. Therefore a valid branch cut for the function should connect the two branch points as illustrated in Fig. (1.1).

Example 1.1.19. Next 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. (1.2) show two examples of the $\log(z^2-1)$ branch cut.

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 $g(z)$ 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 $g(z)$.

Exercise 1.1.20. 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

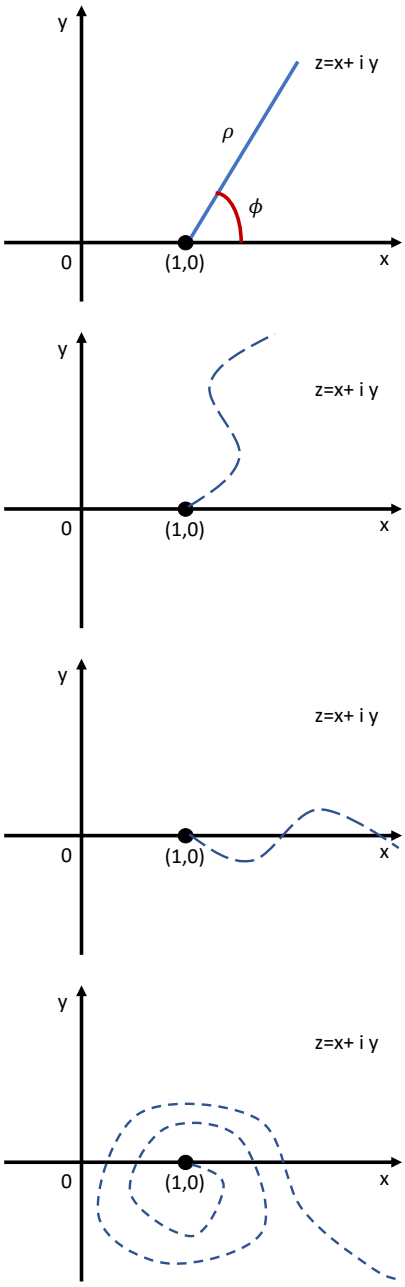


Figure 1.1: Polar parametrization of $\log(z-1)$ (left) and three examples of branch cut for the function connecting its two branch points, at $z = 1$ and at $z = \infty$.

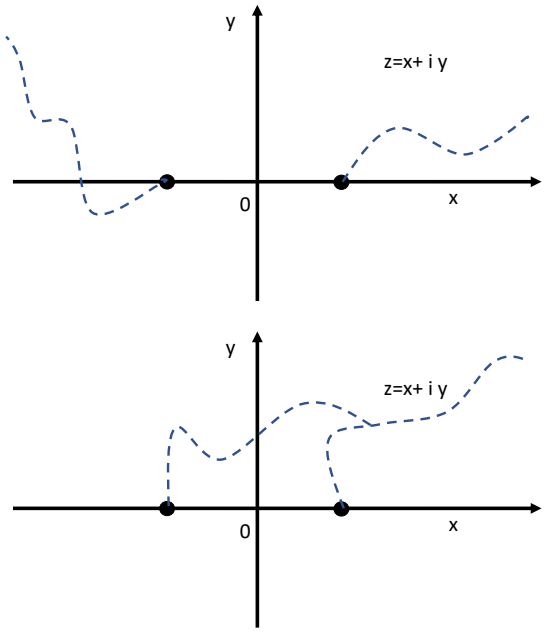


Figure 1.2

visualized in three-dimensional space with parallel projection and the image the 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.)

1.2 Analytic Functions and Integration along Contours

1.2.1 Analytic functions

The derivative of a real valued function is defined at a point x via a 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 1.2.1. The derivative of a complex function is defined via a limiting expression:

$$f'(z) = \lim_{\Delta z \to 0} \frac{f(z + \Delta z) - f(z)}{\Delta z}. \tag{1.3}$$

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. (1.3) 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,

$$\begin{aligned} u_x &= v_y, \\ u_y &= -v_x. \end{aligned} \tag{1.4}$$

and this gives a necessary condition for the following theorem.

Theorem 1.2.2 (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 (1.4) are satisfied in a neighborhood of z.

Notice that in the explanations which lead us to the Cauchy-Riemann theorem (1.2.2) 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 it one needs to show that Eq. (1.4) 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

$$\begin{aligned} \Delta f = f(z + \Delta z) - f(z) &= \frac{\partial f}{\partial x}\Delta x + \frac{\partial f}{\partial y}\Delta y + O\left((\Delta x)^2, (\Delta y)^2, (\Delta x)(\Delta y)\right) \\ &= \frac{1}{2}\left(\frac{\partial f}{\partial x} - i\frac{\partial f}{\partial y}\right)\Delta z + \frac{1}{2}\left(\frac{\partial f}{\partial x} + i\frac{\partial f}{\partial y}\right)\Delta z^* + O\left((\Delta x)^2, (\Delta y)^2, (\Delta x)(\Delta y)\right) \\ &= \frac{\partial f}{\partial z}\Delta z + \frac{\partial f}{\partial z^*}\Delta z^* + O\left((\Delta x)^2, (\Delta y)^2, (\Delta x)(\Delta y)\right) \\ &= \Delta z\left(\frac{\partial f}{\partial z} + \frac{\partial f}{\partial z^*}\frac{\Delta z^*}{\Delta z}\right) + O\left((\Delta x)^2, (\Delta y)^2, (\Delta x)(\Delta y)\right), \end{aligned} \tag{1.5}$$

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. (1.5) 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 the derivative, $f'(z)$, is well defined at any z, one needs to require that

$$\frac{\partial f}{\partial z^*} = 0, \tag{1.6}$$

i.e. that f does not depend on z^*. It is straightforward to check that the "independence of the complex conjugate" Eq. (1.6) is equivalent to Eq. (1.4). □

Definition 1.2.3 (Analyticity)**.** A function $f(z)$ is called (a) analytic (or holomorphic) at a point, z_0, if it is differentiable in a neigborhood of z_0; (b) analytic in a region of the complex plane (in the entire complex plane) if it is analytic at each point of the region (in the entire plane).

Exercise 1.2.1. Verify whether the functions (a) $\exp(z)$, (b) $\bar{z} := x - iy$, (c) $z\exp(\bar{z})$, and (d) $1/(1+z)$ are analytic.

Exercise 1.2.2. 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.

Exercise 1.2.3. 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 $v(x,y)$.

Exercise 1.2.4. Let $f(z) = u(x,y) + iv(x,y)$ be analytic. Given that $v(x,y) = -2xy$ and $f(0) = 1$, find $u(x,y)$.

The Cauchy-Riemann theorem 1.2.2 has a couple of other complementary interpretations discussed below.

Conformal Mappings

The Cauchy-Riemann condition (1.4) can be re-stated in the following compact form

$$i\frac{\partial f}{\partial x} = \frac{\partial f}{\partial y}. \tag{1.7}$$

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

$$J = \begin{pmatrix} \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} \\ \frac{\partial v}{\partial x} & \frac{\partial v}{\partial y} \end{pmatrix} = \begin{pmatrix} \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} \\ -\frac{\partial u}{\partial y} & \frac{\partial u}{\partial x} \end{pmatrix}. \tag{1.8}$$

Geometrically, the off-diagonal (skew-symmetric) part of the matrix represents rotation and the diagonal part of the matrix represents 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 the conditions for a function to be conformal.

Harmonic functions

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

$$(\partial_x^2 + \partial_y^2) f(x, y) = 0. \tag{1.9}$$

Eq. (1.9) 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. (1.9). This statement is a straightforward corollary of the Cauchy-Riemann theorem (1.2.2).

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.

1.2.2 Integration along Contours

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

Definition 1.2.4 (Complex Integration)**.** Let $f(z)$ be analytic in the neighborhood of a contour C. The integral of $f(z)$ along C is

$$\int_C f(z)\, dz := \lim_{n\to\infty} \sum_{k=0}^{n-1} f(\zeta_k)(\zeta_{k+1} - \zeta_k), \tag{1.10}$$

where for each n, $\{\zeta_k\}_{k=0}^n$ is an ordered sequence of points along the path breaking the path 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. Let $z(t)$ with $a \leq t \leq b$ be a parameterization of C, then definition 1.2.4 is equivalent to the Riemann integral of $f(z(t))z'(t)$ with respect to t. Therefore,

$$\int_C f(z)\,dz = \int_a^b f(z(t))\,z'(t)\,dt \tag{1.11}$$

Example 1.2.5. In example 1.1.6 we evaluated the functions (i) $f(z) = z^2$ and (ii) $g(z) = \exp(z+1)$ along the parameterized curves described in example 1.1.10. Now compute (i) $\int_C f(z)\,dz$ and (ii) $\int_C g(z)\,dz$ along the contours (a) C_a: the horizontal line segment from $-M + i\pi$ to $M + i\pi$, (b) C_b: the ray segment extending from the point $z = -1$ and to the point $\sqrt{3}i$, and (c) C_c: he circular arc of radius ε centered at 0.

Solution.

(a) Let $z = x + i\pi$ along C_a, then $dz = dx$ for $-\infty < x < \infty$.

(i) $\displaystyle\int_{C_a} z^2 dz = \int_{-M}^{M} (x+i\pi)^2 dx = \left|\tfrac{1}{3}x^3 - \pi^2 x + \pi i x^2\right|_{-M}^{M} = \left(\tfrac{2}{3}M^3 - 2\pi^2 M\right)$

(ii) $\displaystyle\int_{C_a} e^{z+1}\,dz = \int_{-M}^{M} e^{x+1+i\pi}\,dx = \left|e^{x+1}e^{i\pi}\right|_{-M}^{M} = -e^{M+1} + e^{-M+1}$

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

(i) $\displaystyle\int_{C_b} z^2 dz = \int_0^2 \left(-1 + \rho e^{i\pi/3}\right)^2 e^{i\pi/3} d\rho = \left|\rho e^{i\pi/3} - \rho^2 e^{i2\pi/3} + \tfrac{1}{3}\rho^3 e^{i3\pi/3}\right|_0^2 = \tfrac{1}{3} - i\sqrt{3}$

(ii) $\displaystyle\int_{C_b} e^{z+1} dz = \ldots$

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

(i) $\displaystyle\int_{C_c} z^2\,dz = \int_0^{2\pi} \left(\epsilon e^{i\theta}\right)^2 i\epsilon e^{i\theta} d\theta = \left|\tfrac{1}{3}\epsilon^3 e^{3i\theta}\right|_0^{2\pi} = 0$

(ii) $\displaystyle\int_{C_c} \exp(z+1)\,dz = \ldots$

Exercise 1.2.6. 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^2; (b) $1/z$; and (c) $\sqrt{z}$ along C_+ and C_-. For $\sqrt{z}$, use the branch where z is represented by $re^{i\theta}$ with $0 \leq \theta < 2\pi$. Suggest why the results are the same in (a) and different in (b) and (c). (You may look ahead to the next section for a hint.)

Exercise 1.2.7. Let C be the circular closed contour of radius R centered at the origin. Show that

$$\oint_C \frac{dz}{z^m} = 0, \quad \text{for } m = 2, 3, \ldots \tag{1.12}$$

by parameterizing the contour in polar coordinates.

Exercise 1.2.8. Use numerical integration to approximate the integrals in the exercises above and verify your results.

1.2.3 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 1.2.5 (Cauchy's Theorem, 1825)**.** If $f(z)$ is analytic in a single 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.

It is important to recognize that the use of Cauchy's theorem in what concerns integration of a multi-valued function. For Cauchy's theorem to hold one needs the integrand to be a single valued function. Cuts introduced in the preceding section are required for exactly this reason – 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 1.2.6 (Cauchy 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. (1.10), 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

$$\oint_C dz f(z) = \sum_k \oint_{C_k} f(z) dz, \tag{1.13}$$

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,

$$f(z) = f(z_k) + f'(z_k)(z - z_k) + O\left(\Delta^2\right) \tag{1.14}$$

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. (1.14) into Eq. (1.13) one derives

$$\oint_{C_k} dz f(z) = f(z_k) \oint_{C_k} dz + f'(z_k) \oint_{C_k} dz(z - z_k) + \oint_{C_k} dz O\left(\Delta^2\right) = 0 + 0 + \Delta^3. \tag{1.15}$$

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 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 1.2.7 (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

$$\left| \int_a^b dt g(t) \right| \leq \int_a^b dt |g(t)|,$$

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

$$\left| \int_\gamma f(z) dz \right| \leq \int_\gamma |f(z)||dz|,$$

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 as a Riemann sum

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

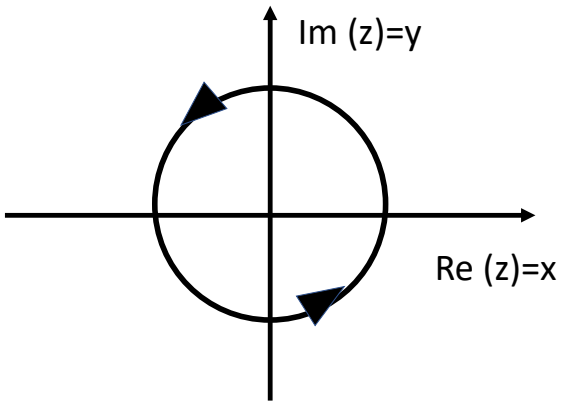


Figure 1.3

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

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

□

1.2.4 Cauchy's Formula

Recall from definition 1.1.3 that a curve is called *simple* if it does not intersect itself, and is called a *contour* if it consists of a finite number of connected smooth curves.

Theorem 1.2.8 (Cauchy's formula, 1831)**.** Let $f(z)$ be analytic on and interior to a simple closed contour C. Then,

$$f(z) = \frac{1}{2\pi i} \int_C \frac{f(\zeta)d\zeta}{\zeta - z}. \tag{1.16}$$

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 shown in Fig. (1.3a), we parameterize the contour explicitly in polar coordinates and derive

$$I = \oint \frac{dz}{z} = \int_0^{2\pi} \frac{r d\exp(i\theta)}{r\exp(i\theta)} = \int_0^{2\pi} \frac{r\exp(i\theta)id\theta}{r\exp(i\theta)} = i\int_0^{2\pi} d\theta = 2\pi i. \tag{1.17}$$

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. (1.17) and with the fact that $\log(z)$ is

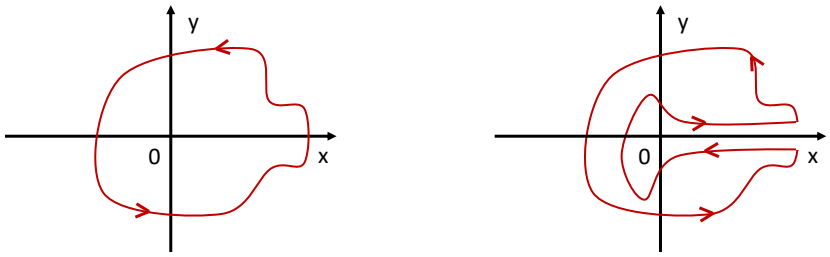


Figure 1.4

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. (1.17), i.e. $2\pi i$ multiplied by an integer, however it will not depend on the path.

Exercise 1.2.9. Compute, compare and discuss the difference (if any) between values of the integral $\oint dz/z$ over two distinct paths shown in Fig. (1.4).

The "small square" construction used above to prove the closed contour version of Cauchy's Theorem, i.e. Theorem 1.2.6, 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

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

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 1.2.9 (Cauchy's formula for derivatives, 1842)**.** Under the same conditions as in Theorem 1.2.8, higher derivatives are

$$f^{(n)}(z) = \frac{n!}{2\pi i} \int_C \frac{f(\zeta)d\zeta}{(\zeta - z)^{n+1}}. \tag{1.18}$$

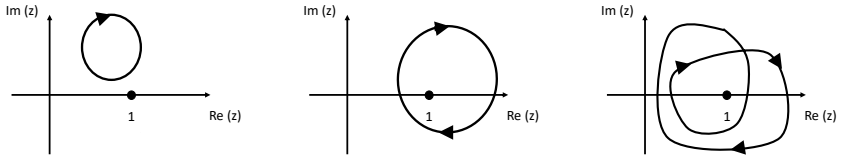


Figure 1.5

1.2.5 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 1.2.10. A function $f(z)$ that is analytic in the annulus $R_1 \leq |z-a| \leq R_2$ may be represented by the power series

$$f(z) = \sum_{n=-\infty}^{+\infty} c_n (z-a)^k. \tag{1.19}$$

in the (possible smaller) annulus $R_1 < \tilde{R}_1 \leq |z-a| \leq \tilde{R}_2 < R_2$ where

$$c_n = \frac{1}{2\pi i} \oint_C \frac{f(z)}{(z-a)^{n+1}} dz. \tag{1.20}$$

and C is any contour that is contained in the region of analyticity and circling a.

Suppose one needs to compute

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

where the contour surrounds $z = a$ in the positive (counter-clockwise) direction such that it contains no other singular points of $f(z)$. Then, we substitute $f(z)$ by its Laurent series, and observe that according to Cauchy's formula the only nonzero contribution will come from the $k = -1$ term

$$\oint f(z)dz = \oint \frac{c_{-1}dz}{z-a} = 2\pi i c_{-1}.$$

Due to this significance of the c_{-1} term, it has a special name, the *residue* of f at $z = a$, and is often denoted by $c_{-1} = \mathrm{Res}(f, a)$.

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

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

Theorem 1.2.11. 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 1.2.9 – that is we have an explicit formula for all the derivatives, so in particular the derivatives all exist. □

Theorem 1.2.12 (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 1.2.10. Prove Cauchy's Inequality Theorem utilizing Theorem 1.2.9. Illustrate the theorem on example of $\cos(z)$.

Theorem 1.2.13 (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 Cauchy's inequality (Theorem 1.2.12) states that $f'(z) \leq M/R$, but R can be arbitrarily large, thus $|f'(z_0)| = 0$ for every $z_0 \in C$. 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 1.2.14 (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 exercise 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. □

Exercise 1.2.11. Prove that $P(z) = \sum_{k=0}^{n} a_k z^k$ has at least one root. (Hint: Prove by contradiction and utilize the Liouville Theorem 1.2.13.)

Theorem 1.2.15 (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 1.2.16 (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 $z_0 + re^{i\theta}$, $\quad 0 \leq \theta \leq 2\pi$, $\quad \gamma'(\theta) = ire^{i\theta}$. 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 1.2.15. To sketch the proof we will use both the mean value property Theorem 1.2.16 and the triangle inequality Theorem 1.2.7. 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 1.2.13) it is plausible to expect that Theorem 1.2.15 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 1.2.15.

Theorem 1.2.17 (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 1.2.15 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. □

Exercise 1.2.12. Find the maximum modulus of $\sin(z)$ on the square, $0 \le x, y \le 2\pi$.

1.3 Residue Calculus

1.3.1 Singularities and Residues

Exercise 1.3.1. Use Cauchy's formula to compute

$$\oint \frac{\exp(z^2)dz}{z-1}, \tag{1.21}$$

for three contour examples shown in the Figure 1.5.

Exercise 1.3.2. Compute the integral $\oint dz/(e^z-1)$ over circle of the radius 4 centered around $3i$.

1.3.2 Evaluation of Real-valued Integrals by Contour Integration

Example 1.3.3. Evalute the integral

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

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

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

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

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

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. (1.7). 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

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

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

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

Exercise 1.3.4. Evaluate the following integrals:

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

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

(c) $\displaystyle\int_0^\infty \frac{\exp(ikx)dx}{x^4+a^4}$.

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

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

Cauchy Principal Value

Consider the integral

$$\int_0^\infty \frac{\sin(ax)dx}{x}, \tag{1.22}$$

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. (1.6). Then going through the following chain of evaluations we arrive at

$$\begin{aligned}\int_0^\infty \frac{\sin(ax)dx}{x} &= \frac{1}{2}\int_{[a\to b\to c\to d]} \frac{\sin(az)}{z}dz \\ &= \frac{1}{4i}\int_{[a\to b\to c\to d]} \left(\frac{\exp(iaz)}{z} - \frac{\exp(-iaz)}{z}\right) dz \\ &= \frac{1}{4i}\int_{[a\to b\to c\to d\to e\to a]} dz\frac{\exp(iaz)}{z} - \frac{1}{4i}\int_{[a\to b\to c\to d\to f\to a]} dz\frac{\exp(-iaz)}{z} = \frac{1}{4i}\left(2\pi i - 0\right) = \frac{\pi}{2}.\end{aligned} \tag{1.23}$$

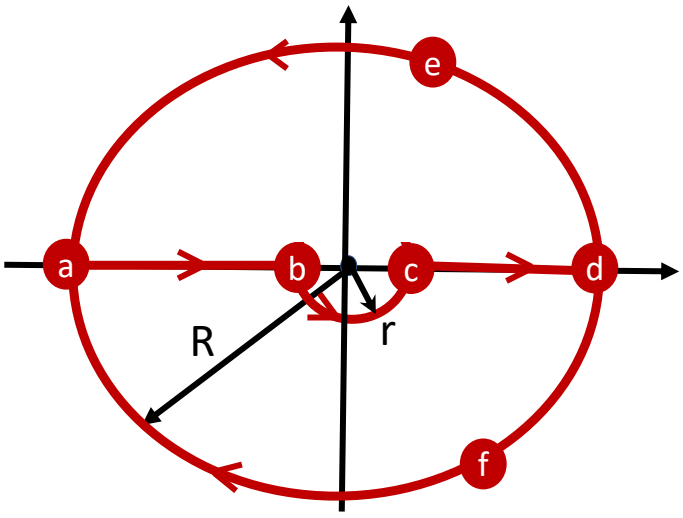


Figure 1.6

(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. (1.6) 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 Jordan's lemma.

The limiting process just explained is often refereed to as the (Cauchy) Principal Value of the integral

$$\mathrm{PV}\int_{-\infty}^{\infty}\frac{\exp(ix)dx}{x}=\lim_{R\to\infty}\int_{-R}^{R}\frac{\exp(ix)dx}{x}=i\pi. \tag{1.24}$$

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

$$\lim_{\varepsilon\to 0}\int_{-R}^{R}f(x)dx=\lim_{\varepsilon\to 0}\left(\int_{-R}^{c-\varepsilon}dxf(x)+\int_{c+\varepsilon}^{R}dxf(x)\right), \tag{1.25}$$

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

$$\int_a^b\frac{dx}{x}=\log\frac{b}{a}, \tag{1.26}$$

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

$$\mathrm{PV}\int_a^b\frac{dx}{x}\doteq\lim_{\varepsilon\to 0}\left(\int_a^{-\varepsilon}\frac{dx}{x}+\int_{\varepsilon}^{b}\frac{dx}{x}\right)=\lim_{\varepsilon\to 0}\left(\log\frac{\varepsilon}{-a}+\log\frac{b}{\varepsilon}\right)=\log\frac{b}{|a|}, \tag{1.27}$$

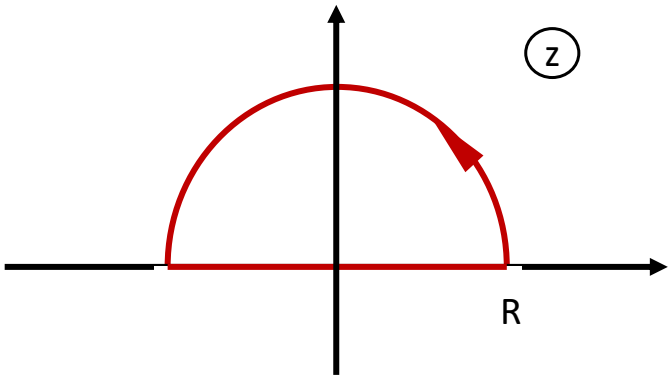


Figure 1.7

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.

1.3.3 Contour Integration with Multi-valued Functions

Proposed Addition: I would like to include a few more worked example for the students to reference.

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.

Consider the following standard integral and its contour version

$$\int_0^\infty \frac{dx}{\sqrt{x}(x^2+1)} \to \oint \frac{dz}{\sqrt{z}(z^2+1)} = \oint dz f(z). \tag{1.28}$$

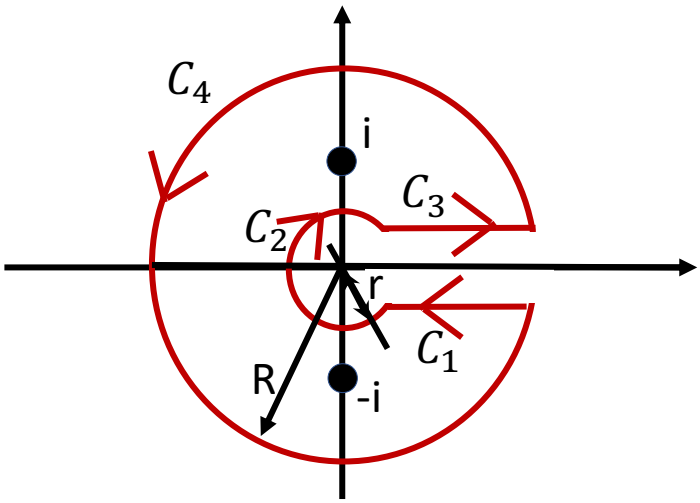


Figure 1.8

The square root in the integrand, $\sqrt{z} = \exp((\log z)/2$, is a multi-valued function, so it must be treated with a contour containing a branch cut. We consider contour shown in Fig. (1.8), then $\oint$ in Eq. (1.28) becomes $\int_{C_1} + \int_{C_2} + \int_{C_3} + \int_{C_4}$. The contour is chosen to guarantee that

$$r \to 0: \quad \int_{C_2} \frac{dx}{\sqrt{x}(x^2+1)} \to 0, \tag{1.29}$$

$$R \to \infty: \quad \int_{C_4} \frac{dx}{\sqrt{x}(x^2+1)} \to 0, \tag{1.30}$$

then resulting (under the $r \to 0$ and $R \to \infty$ limits) in

$$\oint \frac{dz}{\sqrt{z}(z^2+1)} = \int_{C_1} \frac{dz}{\sqrt{z}(z^2+1)} + \int_{C_3} \frac{dz}{\sqrt{z}(z^2+1)} = 2\int_0^\infty \frac{dx}{\sqrt{x}(x^2+1)}. \tag{1.31}$$

On the other hand the contour integral, with the (full) contour surrounding two poles of the integrand, at $z = \pm i$, thus

$$\oint \frac{dz}{\sqrt{z}(z^2+1)} = \pi i \left(\text{Res}\,(\text{at } z = i) + \text{Res}\,(\text{at } z = i)\right), \tag{1.32}$$

where

$$\text{Res}\,(\text{at } z = i) = \lim_{z\to i}\left(f(z)(z-i)\right) = \lim_{z\to i}\frac{1}{\sqrt{z}(z+i)} = \frac{\exp(3\pi i/4)}{2}, \tag{1.33}$$

$$\text{Res}\,(\text{at } z = -i) = \lim_{z\to i}\left(f(z)(z+i)\right) = \lim_{z\to i}\frac{1}{\sqrt{z}(z-i)} = \frac{\exp(-3\pi i/4)}{2}. \tag{1.34}$$

Summarizing one arrives at the following answer

$$\int_0^\infty \frac{dx}{\sqrt{x}(x^2+1)} = \pi i\left(\frac{\exp(3\pi i/4)}{2} - \frac{\exp(-3\pi i/4)}{2}\right) = \frac{\pi}{\sqrt{2}}. \tag{1.35}$$

Exercise 1.3.5. Evaluate the following integral

$$\int_1^\infty \frac{dx}{x\sqrt{x-1}}.$$

Aiming to compute the following integral along the real axis (notice asymptotics at $x \to 0$, $x \to 1$)

$$I = \int_0^1 \frac{dx}{x^{2/3}(1-x)^{1/3}}, \tag{1.36}$$

let us introduce and analyze contour integral with almost the same integrand

$$\oint \frac{dz}{z^{2/3}(z-1)^{1/3}} = \oint \frac{dz}{f(z)}, \tag{1.37}$$

where we introduce the contour, shown in Fig. (1.9a), 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 defined choice of the (originally multi-valued) function branches. Thus in the case under consideration $f(z) \doteq z^{2/3}(z-1)^{1/3}$ has the following parameterization as we go around the cut (in the negative direction):

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

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

$$\int_{C_1} \frac{dz}{f(z)} = \int_0^1 \frac{dx_1}{x_1^{2/3}(1-x_1)^{1/3}\exp(i\pi/3)} = \exp(-i\pi/3)I, \tag{1.38}$$

$$\int_{C_2} \frac{dz}{f(z)} = \int_\pi^{-\pi} \frac{ir\exp(i\theta_2)d\theta_2}{(1+r\exp(i\theta_2))^{2/3}(r\exp(i\theta_2)^{1/3}} \to_{r\to 0} 0 \tag{1.39}$$

$$\int_{C_3} \frac{dz}{f(z)} = \int_1^0 \frac{dx_3}{x_3^{2/3}\lvert 1-x_3\rvert^{1/3}\exp(-i\pi/3)} = -\exp(i\pi/3)I \tag{1.40}$$

$$\int_{C_4} \frac{dz}{f(z)} = \int_{2\pi}^0 \frac{ir\exp(i\theta_4)d\theta_4}{(r\exp(i\theta_4))^{2/3}(r\exp(i\theta_4)-1)^{1/3}} \to_{r\to 0} 0 \tag{1.41}$$

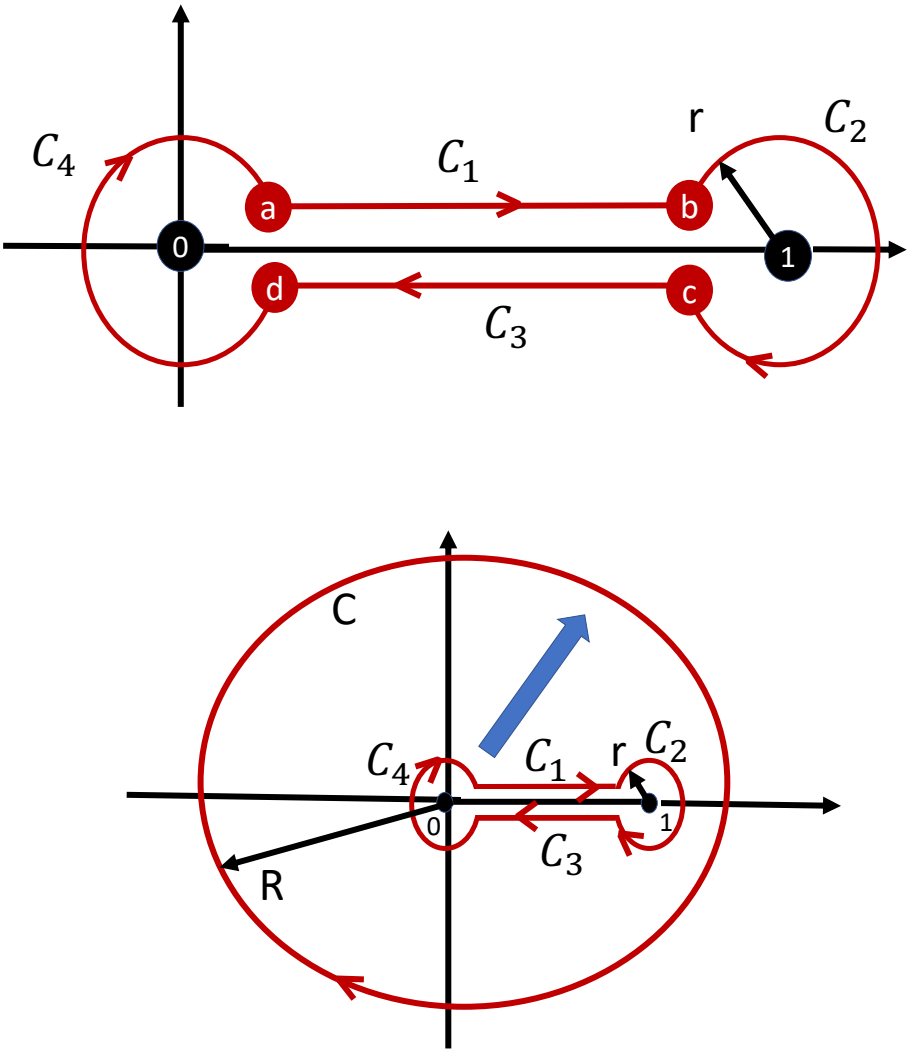


Figure 1.9

Finally, taking advantage of $f(z)$ analyticity everywhere outside the $[0,1]$ cut and using Cauchy's integral theorem one transform integral over, $C_1 \cup C_2 \cup C_3 \cup C_4$, into the same integral over the contour C shown in Fig. (1.9)

$$\int_{C_1} \frac{dz}{f(z)} + \int_{C_2} \frac{dz}{f(z)} + \int_{C_3} \frac{dz}{f(z)} + \int_{C_4} \frac{dz}{f(z)} = \int_{C} \frac{dz}{f(z)}. \tag{1.42}$$

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

$$\int_C \frac{dz}{f(z)} = \int_{2\pi}^{0} \frac{iR\exp(i\theta)d\theta}{R^{2/3}\exp(2i\theta/3)(R\exp(i\theta)-1)^{1/3}} \to_{R\to\infty} = -i\int_0^{2\pi} d\theta = -2\pi i. \tag{1.43}$$

Summarizing Eqs. (1.36, 1.37,1.38,1.39,1.40,1.41,1.42,1.43) one arrives at

$$I = \frac{-2\pi i}{-\exp(i\pi/3)+\exp(-i\pi/3)} = \frac{\pi}{\sin(\pi/3)} = \frac{2\pi}{\sqrt{3}}. \tag{1.44}$$

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

Exercise 1.3.6. Evaluate the integral

$$\int_{-1}^{1} \frac{dx}{(1+x^2)\sqrt{1-x^2}}, \tag{1.45}$$

by identifying and evaluating an equivalent contour integral.

1.4 Extreme-, Stationary- and Saddle-Point Methods

In this section we study family of 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, stationary-point or saddle-point method. We start discussing the extreme-point version, corresponding to estimating real-valued integrals over a real domain, then 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).

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

$$I_1 = \int_a^b dx \exp\left(f(x)\right), \tag{1.46}$$

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

$$f(x) = f(x_0) + \frac{(x-x_0)^2}{2} f''(x_0) + O\left((x-x_0)^3\right), \tag{1.47}$$

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. (1.47) in Eq. (1.46) 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

$$I_1 \to \sqrt{\frac{2\pi}{-f''(x_0)}} \exp\left(f(x_0)\right). \tag{1.48}$$

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

Example 1.4.1. Estimate the following integral

$$I = \int_{-\infty}^{+\infty} dx \exp(S(x)), \quad S(x) = \alpha x^2 - x^4/2, \tag{1.49}$$

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

Solution: Let us find all stationary points of $S(x)$ (saddle points of the integrand). Solving $S'(x_s) = 0$, one gets that either $x_s = 0$ or $x_s = \pm\sqrt{\alpha}$. Values of S at the saddle points are $S(0) = 0$ and $S(\pm\sqrt{\alpha}) = \alpha^2/2$, and we thus choose the dominating saddle points, $x_s = \pm\sqrt{\alpha}$, for further evaluations. In fact, and since the two (dominant) saddle points are fully equivalent, we pick one and then multiply estimation for the integral by two:

$$I \approx 2\exp(\alpha^2/2) \int_{-\infty}^{+\infty} dx \exp(S''(\sqrt{\alpha})x^2/2) = 2\exp(\alpha^2/2) \int_{-\infty}^{+\infty} dx \exp(-2\alpha x^2)$$
$$= \exp(\alpha^2/2)\sqrt{\frac{2}{\alpha\pi}}.$$

The same idea works for highly oscillatory integrals of the form

$$I_2 = \int_a^b dx \exp\left(if(x)\right), \tag{1.50}$$

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

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

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

Now in the most general case (of the saddle-point method) we consider the contour integral

$$I_3 = \int_C dz \exp\left(f(z)\right), \tag{1.51}$$

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

$$I_3 \to \sqrt{\frac{2\pi}{-f''(z_0)}} \exp\left(f(z_0)\right). \tag{1.52}$$

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 in the Example 1.4.1, one deforms the contour to pass through all the saddle-points then replacing rhs in Eq. (1.52) by sums of the saddle-point contributions.

Exercise 1.4.2. 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.

Chapter 2

Fourier Analysis

Fourier analysis is the study of the way functions may be represented or approximated by an integral, or a sum, of oscillatory basis functions. The process of decomposing a function into its oscillatory components, and the inverse process of recomposing the function from these components, are two themes of Fourier analysis. When the oscillatory components take a continuous range of wave-numbers (or frequencies), the decomposition and recomposition is achieved by integration, and is 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 is achieved by summation, and is 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 to apply to more general and abstract situations, and the field is now often called harmonic analysis.

2.1 The Fourier Transform and Inverse Fourier Transform

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

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} d\boldsymbol{k} \, \exp\left(i\boldsymbol{k}^T\boldsymbol{x}\right) \hat{f}(\boldsymbol{k}), \tag{2.1}$$

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

$$\hat{f}(\boldsymbol{k}) := \int_{\mathbb{R}^d} d\boldsymbol{x} \, \exp\left(-i\boldsymbol{k}^T\boldsymbol{x}\right) f(\boldsymbol{x}). \tag{2.2}$$

Eq. (2.1) and Eq.(2.2) are inverses of each other (meaning, for example, that substituting Eq. (2.2) into Eq. (2.1) 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

$$\delta(\boldsymbol{x}) := \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} d\boldsymbol{k}\, \exp(i\boldsymbol{k}^T\boldsymbol{x}). \tag{2.3}$$

We will discuss Dirac's δ-function in section 2.3, primarily for $d = 1$.

At first glance, it might appear that the appropriate class of functions for which Eq. (2.1) 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. (2.1) to be defined over a wider class of functions in section 2.4. 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 (Math527).

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 2.2), 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 2.4.

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.

2.2 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

$$\begin{aligned}\hat{h}(k) &= \int_{\mathbb{R}} dx\, h(x)e^{-ikx} = \int_{\mathbb{R}} dx\, \left(af(x) + bg(x)\right) e^{-ikx} = a\int_{\mathbb{R}} dx\, f(x)e^{-ikx} + b\int_{\mathbb{R}} dx\, g(x)e^{-ikx} \\ &= a\hat{f}(k) + b\hat{g}(k).\end{aligned} \tag{2.4}$$

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

$$\begin{aligned}\hat{h}(k) &= \int_{\mathbb{R}} dx\, h(x)e^{-ikx} = \int_{\mathbb{R}} dx\, f(x - x_0)e^{-ikx} = \int_{\mathbb{R}} dx'\, f(x')e^{-ikx'-ikx_0} \\ &= e^{-ikx_0}\hat{f}(k). \end{aligned} \tag{2.5}$$

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

$$\begin{aligned}\hat{h}(k) &= \int_{\mathbb{R}} dx\, h(x)e^{-ikx} = \int_{\mathbb{R}} dx\, f(x)e^{ik_0x}e^{-ikx} = \int_{\mathbb{R}} dx\, f(x)e^{-i(k-k_0)x} \\ &= \hat{f}(k - k_0). \end{aligned} \tag{2.6}$$

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

$$\begin{aligned}\hat{h}(k) &= \int_{\mathbb{R}} dx\, h(x)e^{-ikx} = \int_{\mathbb{R}} dx\, f(ax)e^{-ikx} = |a|^{-1}\int_{\mathbb{R}} dx'\, f(x')e^{-ikx'/a} \\ &= |a|^{-1}\hat{f}(k/a). \end{aligned} \tag{2.7}$$

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 $h(x) = \overline{(x)}$, then

$$\begin{aligned}\hat{h}(k) &= \int_{\mathbb{R}} dx\, h(x)e^{-ikx} = \int_{\mathbb{R}} dx\, (f(x))^* e^{-ikx} = \int_{\mathbb{R}} dx\, \overline{f(x)e^{ikx}} \\ &= \overline{\hat{f}(-k)}. \end{aligned} \tag{2.8}$$

Exercise 2.2.1. 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 $h(x) = \Re(f(x))$, then $\hat{h}(k) = \frac{1}{2}(\hat{f}(k) + (\hat{f}(-k))^*$.

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

Exercise 2.2.2. 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 $\rho^2 = k_1^2 + k_2^2$.

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

$$\begin{aligned}\hat{h}(k) &= \int_{\mathbb{R}} dx\, h(x)e^{-ikx} = \int_{\mathbb{R}} dx f'(x)e^{-ikx} = \Big[f(x)e^{-ikx}\Big]_{-\infty}^{\infty} - \int_{\mathbb{R}} dx(-ik)f(x)e^{-ikx} \\ &= (ik)\hat{f}(k). \end{aligned} \tag{2.9}$$

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 2.3), 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 $\int |f|^2 < \infty$,

$$\begin{aligned}\int_{-\infty}^{\infty} dx\,|f(x)|^2 &= \int_{-\infty}^{\infty} dx f(x)\overline{f(x)} = \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} \frac{dk_1}{2\pi} e^{ik_1 x}\hat{f}(k_1) \int_{-\infty}^{\infty} \frac{dk_2}{2\pi} e^{-ik_2 x}\overline{\hat{f}(k_2)} \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} dk\,|\hat{f}(k)|^2. \end{aligned} \tag{2.10}$$

Definition 2.2.1. The *integral convolution* of the function f with the function g, is defined as

$$(g * f)(x) := \int_{\mathbb{R}} dy\; g(x-y)f(y), \tag{2.11}$$

Proposed Addition: I think that we should (1) have an illustration of convolution here. It could be as simple as the schematic shown on the wikipedia page. (2) Write a computational snippet to demonstrate a moving average. **Convolution:** Suppose that h is the integral convolution of f with g, that is, $h(x) = (g * f)(x)$, then

$$\begin{aligned}\hat{h}(k) &= \int_{\mathbb{R}} dx\; h(x)e^{-ikx} = \int_{\mathbb{R}} dx \int_{\mathbb{R}} dy\; g(x-y)f(y)e^{-ikx} \\ &= \hat{g}(k)\hat{f}(k). \end{aligned} \tag{2.12}$$

The convolution of a function f with a kernel g is defined in Eq. (2.11). 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 $(f * g) = f$ 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 on the RHS of Eq. (2.26) could be nonzero.

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

2.3 Dirac's δ-function.

2.3.1 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

$$f_\epsilon(x) = \begin{cases} 1/\epsilon & |x| \leq \epsilon/2 \\ 0 & |x| > \epsilon/2 \end{cases} \tag{2.13}$$

The pointwise limit of f_ϵ is clearly zero for all $x \neq 0$, and therefore the integral of the limit of f_ϵ must also be zero:

$$\lim_{\epsilon\to 0} f_\epsilon(x) = 0 \quad \Rightarrow \quad \int_{-\infty}^{\infty} dx \lim_{\epsilon\to 0} f_\epsilon(x) = 0. \tag{2.14}$$

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

$$\int_{-\infty}^{\infty} dx\, f_\epsilon(x) = 1 \quad \Rightarrow \quad \lim_{\epsilon\to 0} \int_{-\infty}^{\infty} dx\, f_\epsilon(x) = 1. \tag{2.15}$$

Although Eq. (2.14) suggests that $f_\epsilon(x)$ may not be very interesting as a *function*, the behavior demonstrated by Eq. (2.15) motivates the use of $f_\epsilon(x)$ as a *functional*[1]. For any sufficiently nice function $\phi(x)$, define the functionals $f_\epsilon[\phi]$ and $f[\phi]$ by

$$f[\phi] := \lim_{\epsilon\to 0} f_\epsilon[\phi] := \lim_{\epsilon\to 0} \int_{-\infty}^{\infty} dx\, f_\epsilon(x)\phi(x) \tag{2.16}$$

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

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

[1]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

$$m_\epsilon \leq f_\epsilon[\phi] \leq M_\epsilon$$

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

$$f[\phi] = \lim_{\epsilon \to 0} f_\epsilon[\phi] = \phi(0)$$

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

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

$$g_\epsilon(x) = \frac{1}{\pi} \frac{\epsilon}{x^2 + \epsilon^2}$$

The pointwise limit $g_\epsilon(x)$ is also zero for every $x \neq 0$, so as before, the integral of the limit must be zero:

$$\lim_{\epsilon \to 0} g_\epsilon(x) = 0 \quad \Rightarrow \quad \int_{-\infty}^{\infty} dx \lim_{\epsilon \to 0} g_\epsilon(x) = 0$$

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

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

As with $f_\epsilon(x)$, we can use $g_\epsilon(x)$ to define the functionals $g_\epsilon[\phi(x)]$ and $g[\phi]$ by

$$g[\phi] := \lim_{\epsilon \to 0} g_\epsilon[\phi] := \lim_{\epsilon \to 0} \int_{-\infty}^{\infty} g_\epsilon(x)\phi(x)dx$$

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

$$g[\phi] = \lim_{\epsilon \to 0} g_\epsilon[\phi] = \phi(0)$$

That is, $g[\phi]$ also evaluates its argument at the point $x = 0$.

The sequences $f_\epsilon(x)$ and $g_\epsilon(x)$ 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 $\delta[\phi] = \phi(0)$.

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

Example 2.3.1. For $b, c \in \mathbb{R}$, show that $c\delta(x - b)f(x) = cf(b)$

$$\begin{aligned} c\delta(x-b)f(x) &= \int_{-\infty}^{\infty} dx\, c\delta(x-b)f(x) = c\int_{-\infty}^{\infty} dx'\, \delta(x')f(x'+b) \\ &= cf(b) \end{aligned} \tag{2.17}$$

Example 2.3.2. For $a \in \mathbb{R}$, show that $\delta(ax)f(x) = f(0)/|a|$

$$\begin{aligned}\delta(ax)f(x) &= \int_{-\infty}^{\infty} dx\, \delta(ax)f(x) = \int_{-\infty}^{\infty} \frac{dx'}{|a|}\, \delta(x')f(x'/a) \\ &= f(0)/|a| \end{aligned} \tag{2.18}$$

Corollary 2.3.1. Show that the Fourier transform of a δ-function is a constant.

Solution.

$$\begin{aligned}\hat{\delta}(k) &= \int_{-\infty}^{\infty} dx\, \delta(x)e^{-ikx} = e^{-ik\,0} \\ &= 1 \end{aligned} \tag{2.19}$$

Corollary 2.3.2. Show that

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

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

Solution. We identify the expression on the RHS as the inverse Fourier transform of the function $\hat{f}(k) = 1$

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} dk\, 1e^{ikx} \tag{2.20}$$

The constant function is not integrable in the traditional sense. The theory of distributions allows us to give meaning to this integral. We know that δ is defined so that for any suitable function $\phi(x)$, $\int dx\, \delta(x)\phi(x) = \phi(0)$. Even though we cannot integrate $f(x)$ directly, but if we can show that $\int dx\, f(x)\phi(x) = \phi(0)$, then we can assert that $f(x) = \delta(x)$.

$$\begin{aligned}f[\phi(x)] &= \int_{-\infty}^{\infty} dx\, \phi(x) \int_{-\infty}^{\infty} \frac{dk}{2\pi}\, 1e^{-ikx} = \int_{-\infty}^{\infty} \frac{dk}{2\pi} \int_{-\infty}^{\infty} dx\, \phi(x)e^{-ikx} \\ &= \int_{-\infty}^{\infty} \frac{dk}{2\pi}\, \hat{\phi}(k) = \int_{-\infty}^{\infty} \frac{dk}{2\pi}\, \hat{\phi}(k)e^{ik0} \\ &= \phi(0) \end{aligned} \tag{2.21}$$

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

Alternative Definitions of the δ-function

We have defined the δ-function in Eq. (2.3) as the limit of a particular δ-sequence, namely the 'top-hat' function given in Eq. (2.14). One has to wonder whether there may be other δ-sequences which give the same limit. For example, consider

$$\delta(t) = \lim_{\epsilon \to 0} \frac{2t^2\epsilon}{\pi(t^2 + \epsilon^2)^2}. \tag{2.22}$$

To validate the suitability of Eq. (2.22) 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 aroun $i\epsilon$ 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 2.3.3. Validate the following asymptotic representations for the δ-function

$$(a)\,\delta(t) = \lim_{\epsilon\to 0} \frac{1}{\sqrt{\pi\epsilon}} \exp\left(-\frac{t^2}{\epsilon}\right), \tag{2.23}$$

$$(b)\,\delta(t) = \lim_{n\to\infty} \frac{1-\cos(nt)}{\pi n t^2}. \tag{2.24}$$

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.

Exercise 2.3.4. Prove that for x on the interval $(-\pi, \pi)$

$$\lim_{r\to 1-0} \frac{1-r^2}{2\pi(1-2r\cos(x)+r^2)} = \delta(x).$$

2.3.2 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 2.3.3. The Fourier Transform of the convolution of the function f with the function g is the product $\hat{f}(k)\hat{g}(k)$

$$\begin{aligned}
\hat{f * g}(k) &= \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy\; g(x-y)f(y)e^{-ikx} && (2.25)\\
&= \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy \int_{-\infty}^{\infty} \frac{dk_1}{2\pi} \int_{-\infty}^{\infty} \frac{dk_2}{2\pi}\, \hat{f}(k_1)\hat{g}(k_2) \exp\left(-ikx + ik_1(x-y) + ik_2 y\right)\\
&= \int_{-\infty}^{\infty} dk_1 \int_{-\infty}^{\infty} dk_2\, \hat{f}(k_1)\hat{g}(k_2) \frac{1}{2\pi} \int_{-\infty}^{\infty} dx\, \exp\left(-ikx + ik_1 x\right) \frac{1}{2\pi} \int_{-\infty}^{\infty} dy\, \exp\left(-ik_1 y + ik_2 y\right)\\
&= \int_{-\infty}^{\infty} dk_1 \hat{f}(k_1)\delta(k-k_1) \int_{-\infty}^{\infty} dk_2\, \hat{g}(k_2)\delta(k_1-k_2)\\
&= \hat{f}(k)\hat{g}(k) && (2.26)
\end{aligned}$$

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 2.3.4. Unitarity [Parceval/Plancherel Theorem]:

$$\begin{aligned}\int_{-\infty}^{\infty} dx|f(x)|^2 &= \int_{-\infty}^{\infty} dx f(x)\overline{f(x)} = \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} \frac{dk_1}{2\pi} e^{ik_1 x}\hat{f}(k_1) \int_{-\infty}^{\infty} \frac{dk_2}{2\pi} e^{-ik_2 x}\overline{\hat{f}(k_2)} \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} dk_1 \int_{-\infty}^{\infty} dk_2\, \hat{f}(k_1)\overline{\hat{f}(k_2)}\, \frac{1}{2\pi}\int_{-\infty}^{\infty} dx \exp(ix(k_1-k_2)) \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} dk_1 \int_{-\infty}^{\infty} dk_2\, \hat{f}(k_1)\, \overline{\hat{f}(k_2)}\, \delta(x-y) \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} dk|\hat{f}(k)|^2. \end{aligned} \tag{2.27}$$

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

$$f(x) = \int dy \delta(x-y) f(y). \tag{2.28}$$

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

$$\delta\left(f(x)\right) = \sum_n \frac{1}{|f'(y_n)|}\delta(x-y_n). \tag{2.29}$$

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. □

2.3.3 The δ-function in Higher Dimensions

d-dimensional δ-function, which was instrumental for introducing d-dimensional Fourier transform in Section 2.1, is simply a product of one dimensional δ-functions, $\delta(\boldsymbol{x}) = \delta(x_1)\cdots\delta(x_n)$.

Example 2.3.5. Compute the δ-function in polar spherical coordinates.

2.3.4 The Heaviside Function and the Derivatives of the δ-function

One also derives from Eq. (2.28) that $\int_{-\infty}^{\infty} dx\delta(x) = 1$. This motivates introduction of a function associated with an incomplete integration of the $\delta(x)$

$$\theta(y) := \int_{-\infty}^{y} dx\delta(x) = \begin{cases} 0, & y < 0 \\ 1, & y > 0, \end{cases} \tag{2.30}$$

called Heaviside- or step-function.

Exercise 2.3.6. Prove the relation

$$\left(\frac{d^2}{dt^2} - \gamma^2\right) \exp\left(-\gamma|t|\right) = -2\gamma\delta(t). \tag{2.31}$$

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

One also gets, differentiating Eq. (2.30), that $\theta'(x) = \delta(x)$. We can also differentiate the δ-function. Indeed, integrating Eq. (2.28) by parts, and assuming that the respective anti-derivative is bounded, one arrives at

$$\int dy\delta'(y-x)f(y) = -f'(x) \tag{2.32}$$

Substituting in Eq. (2.32), $f(x) = xg(x)$ one derives

$$x\delta'(x) = -\delta(x). \tag{2.33}$$

Expanding $f(x)$ in the Taylor series around $x = y$, ignoring terms of the second order (and higher) in $(x - y)$, and utilizing Eq. (2.34) one arrives at

$$f(x)\delta'(x-y) = f(y)\delta'(x-y) - f'(y)\delta(x-y). \tag{2.34}$$

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

$$f'(x) = (f(y+0) - f(y-0))\delta(x-y) + g(x), \tag{2.35}$$

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 2.3.7. Express $t\delta''(t)$ via $\delta'(t)$.

2.4 Closed form representation for select Fourier Transforms

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

2.4.1 Elementary examples of closed form representations

Example 2.4.1. Show that the Fourier Transform of a δ-function is a constant.

Solution. See corollary 2.3.1 where we showed $\hat{\delta}(k) = 1$.

Example 2.4.2. Show that the Fourier Transform of a constant is a δ-function.

Solution. In corollary 2.3.2 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 2.4.3. 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.

$$\begin{aligned} \hat{f}(k) &= \int_{\mathbb{R}} dx f(x) e^{-ikx} = b\int_{-a}^{a} dx e^{-ikx} = \frac{b}{(-ik)} e^{-ikx}\Big|_{-a}^{a} = \frac{b}{-ik}\left(e^{-ika} - e^{ika}\right) \\ &= \frac{2b}{k}\sin(ka). \end{aligned} \tag{2.36}$$

Example 2.4.4. Show that the Fourier transform of a sinc function is a square pulse:

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

Example 2.4.5. Find the Fourier transform of a Gaussian function

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

Solution.

$$\begin{aligned} \hat{f}(k) &= \int_{\mathbb{R}} dx f(x) e^{-ikx} = a\int_{-\infty}^{\infty} dx e^{-bx^2} e^{-ikx} \\ &= a\exp\left(-\frac{k^2}{4b}\right)\int_{-\infty}^{\infty} dx \exp\left(-b\left(x + \frac{ik}{2b}\right)^2\right) \\ &= \frac{a}{\sqrt{b}}\exp\left(-\frac{k^2}{4b}\right)\int_{-\infty}^{\infty} dx' e^{-x'^2} \\ &= a\exp\left(-\frac{k^2}{4b}\right)\sqrt{\frac{\pi}{b}}. \end{aligned} \tag{2.37}$$

Exercise 2.4.6. Find the Fourier transform of $f(x) = \dfrac{1}{x^4 + a^4}$

Exercise 2.4.7. Find the Fourier transform of $f(x) = \operatorname{sech}(ax)$.

Exercise 2.4.8. Verify the following Fourier transform pair:

(a) Let $a > 0$. Show that

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

(b) Let $a > 0$. Show that

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

2.4.2 More complex examples of closed form representations

We can find closed form representations of other functions by combining the examples above with the properties in section 2.2.

Example 2.4.9. This problem is fantastically difficult. Let $f(t)$ be given by

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

where ω_0 and A are fixed, and $A > 0$.

(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. Coming Soon!

Exercise 2.4.10. Let

$$f_a(x) := \frac{2a}{a^2 + (4\pi x)^2}$$

for $a \in \mathbb{C}$ with $Re(a) > 0$. If also $b \in \mathbb{C}$ with $Re(b) > 0$, show that

$$f_a * f_b = f_{a+b}$$

Exercise 2.4.11. Show the following:

(a) Show that the Fourier transform of

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

(b) Show that the Fourier transform of

$$f(x) = \frac{\sin^2(x)}{x} \quad \text{is} \quad \hat{f}(k) = -\frac{i\pi}{2}\Big(\Pi(k-1) - \Pi(k+1)\Big)$$

where

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

2.4.3 Closed form representations in higher dimensions

Exercise 2.4.12. 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

(a) $g(\boldsymbol{x}) = \exp(-|\boldsymbol{x}|^2)$.

(b) (Bonus) $h(\boldsymbol{x}) = \exp(-|\boldsymbol{x}|)$ for $d = 3$ (i.e. in three dimensions).

2.5 Fourier Series

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, typically, 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), $\exp(i2\pi nx/L)$:

$$f(x) = \sum_{n=-\infty}^{\infty} f_n \exp\left(2\pi i n x/L\right). \tag{2.38}$$

This, so-called Fourier series representation of a periodic function immediately shows that the Fourier Series is a particular case of the Fourier integral. Indeed, a periodic function can be represented as a convolution of a function with a finite support in $[0, L]$ and of a sum of δ-functions

$$\sum_{n=-\infty}^{\infty} f_n \exp\left(2\pi i n x/L\right) \int_{-\infty}^{\infty} dk\, \delta(k-n) = \int_{-\infty}^{\infty} dk\, \exp\left(2\pi i k x/L\right) \sum_{n=-\infty}^{\infty} f_n \delta(k-n). \tag{2.39}$$

One can also consider a function with a finite support over $[0, L]$, i.e. one which is equal to zero outside of the interval. Fourier transform of this function and its (standard) Inverse

Fourier Transform are

$$\hat{f}(k) = \int_0^L \frac{dx}{2\pi} f(x) \exp\left(-ikx\right), \quad f(x) = \int_{-\infty}^{\infty} dk e^{ikx} \hat{f}(k). \tag{2.40}$$

Obviously if $x \in [0, L]$ assumption of periodicity and assumption of finite support are equivalent. Then comparing Eq. (2.39) and Eq. (2.40), one arrives at

$$f_n = \int_0^L \frac{dx}{L} f(x) \exp\left(-2\pi i \frac{nx}{L}\right), \tag{2.41}$$

which is the inverse Fourier Series relation for periodic and/or finite support functions.

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

$$\sum_{-\infty}^{\infty} \cdots \to \frac{L}{2\pi} \int_{-\infty}^{\infty} dk \cdots . \tag{2.42}$$

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

$$f_n = \int_0^{2\pi} \frac{dx}{2\pi} \exp(-inx + \alpha x) = \frac{1}{2\pi} \frac{1}{\alpha - in} \left(e^{2\pi\alpha} - 1\right). \tag{2.43}$$

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 2.5.1. Expand (a) $f(x) = x$, and (b) $g(x) = |x|$, both defined on the interval $-\pi < x < \pi$, in the Fourier series. Describe the difference between (a) and (b) in the dependence of the n-th Fourier coefficient on n.

Let us conclude this section reminding that constructing the Fourier Series (and also Fourier Integrals) we assume that the set of harmonic functions forms a complete set of basis functions for a properly integrable function. Proving the assumption requires an extra work which is not done in this course. Instead this proof (as well as many other proofs) is left for detailed discussion in the companion Math 525 course of the core AM series.

2.6 Riemann-Lebesgue Lemma

The Fourier series is infinite (contains infinite number of terms), thus computationally prohibitive, and one common approximation approach consists in truncating it.

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 2.6.1 (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 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 527) and also the algorithm course (Math 575), will be given at an inter-core recitation session.

2.7 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 a square wave

$$f(x) = \begin{cases} \pi/4, & \text{if} \quad 2n\pi \le x \le (2n+1)\pi, \quad n = 0, 1, 2, \dots \\ -\pi/4, & \text{if} \quad (2n+1)\pi \le x \le (2n+2)\pi, \quad n = 0, 1, 2, \dots \end{cases} \tag{2.44}$$

$$= \sum_{n=0}^{\infty} \frac{\sin((2n+1)x)}{2n+1}, \tag{2.45}$$

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:square-wave-Fourier and thus consider N-th partial Fourier Series

$$S_N(x) = \sum_{n=0}^{N} \frac{\sin((2n+1)x)}{2n+1}. \tag{2.46}$$

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

$$\frac{d}{d\epsilon} S_N(\epsilon) = \sum_{n=0}^{N} \cos((2n+1)\epsilon) = \frac{2(N+1)\epsilon}{2\sin\epsilon}, \tag{2.47}$$

where we have utilized formula for the sum of the geometric progression. Observe that $\frac{d}{d\epsilon} S_N(\epsilon) \to \frac{N+1}{2}$ 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 $S_N(\epsilon_*)$

$$S_N(\epsilon_*) = \sum_{n=0}^{N} \frac{\sin\left(\frac{(2n+1)\pi}{2(N+1)}\right)}{2n+1} = \sum_{n=0}^{N} \frac{\sin\left(\frac{n\pi}{N}\right)}{2n} + O(1/N)\Bigg|_{N\to 0}$$

$$\to \frac{1}{2}\int_0^{\pi} \frac{\sin t}{t} dt \approx \frac{\pi}{4} + 0.14, \tag{2.48}$$

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 2.7.1. Generalize the two functions from Exercise 2.5.1 beyond the $[-\pi, \pi)$ interval, so they are 2π-periodic function on $[-5\pi, 5\pi)$. Compute the respective partial Fourier

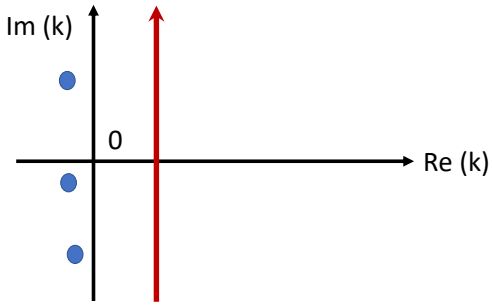


Figure 2.1: Integration contour, C, in Eq. (2.49) is shown red. C is often shown as straight line in the complex from $c - i\infty$ to $c + i\infty$, where c is an infinitesimally small positive number. Possible singularities of the LT $\tilde{\Phi}(k)$ may only be at the points with negative real number, which are shown schematically as blue dots.

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 discussion of the Fourier Series mentioning its, arguably most significant, application to the field of differential equations. Even though some differential equations can be analyzed or even solved analytically (this is the prime focus of the next two chapters of the course), most differential equations of interest can only be solved numerically. Looking for solution of an ODE or PDE in terms of a Fourier series and then truncating the series to a finite sum represents one of the most powerful numerical methods in the arsenal of Applied Mathematics. This, so-called spectral method, is to be discussed in the algorithm (Math 575) course of the core series.

2.8 Laplace Transform

The Laplace Transform (LT) may be considered as a Fourier transform applied to functions which are nonzero at $t \geq 0$. Then, the LT, defined at $t > 0$, is

$$\tilde{\Phi}(k) = \int_0^\infty dt \exp(-kt)\, \Phi(t). \tag{2.49}$$

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

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

$$\Phi(t) = \frac{1}{2\pi i}\int_C dk \exp(kt)\,\tilde{\Phi}(k). \tag{2.50}$$

over contour, C, shown in Fig. (2.1). C can be deformed arbitrarily within the domain, Re > 0, of the $\tilde{\Phi}(k)$ analyticity. Note that by construction, and consistently with the requirement imposed on $\Phi(t)$, the integral on the right hand side of Eq. (2.50) is equal to zero at $t < 0$. Indeed, given that $\tilde{\Phi}(k)$ 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.

It is instructive to illustrate similarities and differences between Laplace and Fourier transforms on examples.

Consider one sided exponential

$$f(t) = \theta(t)\exp(-\alpha t), \quad \alpha > 0, \tag{2.51}$$

$$\hat{f}(\omega) = \frac{\alpha - i\omega}{\alpha^2 + \omega^2}, \tag{2.52}$$

$$\tilde{f}(s) = \frac{1}{s+\alpha}, \quad \mathrm{Re}(s+\alpha) > 0, \tag{2.53}$$

which then turns into the step function, $\theta(t)$, at $\alpha \to 0^+$

$$f(t) = \theta(t), \tag{2.54}$$

$$\hat{f}(\omega) = \pi\delta(\omega) - \frac{i}{\omega}, \tag{2.55}$$

$$\tilde{f}(s) = \frac{1}{s}. \tag{2.56}$$

Shifting and rescaling the step-function we arrive at the following expressions for the signature function

$$f(t) = \mathrm{sign}(t), \tag{2.57}$$

$$\hat{f}(\omega) = -\frac{2i}{\omega}, \tag{2.58}$$

$$\tilde{f}(s) = \frac{1}{s}. \tag{2.59}$$

Exercise 2.8.1. Find the Laplace Transform of (a) $\Phi(t) = \exp(-\lambda t)$, (b) $\Phi(t) = t^n$, (c) $\Phi(t) = \cos(\nu t)$, (d) $\Phi(t) = \cosh(\lambda t)$, (e) $\Phi(t) = 1/\sqrt{t}$. Show details.

Exercise 2.8.2. Find the Inverse Laplace Transform of $1/(k^2 + a^2)$. Show details.

Part II

Differential Equations

Chapter 3

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 initial or boundary data.

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. $u = u(t)$ or $u = u(x)$) 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. $u = u(x, y)$, or $u = u(t, x, y, z)$) 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 $f(x)$ is a solution, so is $cf(x)$, 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 $\alpha_n(t)u^{(n)}(t) + \cdots + \alpha_1(t)u'(t) + \alpha_0(t)u(t) = f(t)$,

where $\alpha_i(t), i = 0, \ldots n$ and $f(t)$ 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 section 4.

3.1 ODEs: Simple cases

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

3.1.1 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.

$$\frac{dx}{dt} = \frac{f(t)}{g(x)} \quad \Rightarrow \quad g(x)dx = f(t)dt \quad \Rightarrow \quad \int g(x)dx = \int f(t)dt. \tag{3.1}$$

3.1.2 Method of Parameter Variation

To solve the following linear, inhomogeneous ODE

$$dy/dt - p(t)y(t) = g(t), \quad y(t_0) = y_0, \tag{3.2}$$

let us substitute,

$$y(t) = c(t) \exp\left(\int_{t_0}^{t} dt' p(t')\right), \tag{3.3}$$

where the second term on the right is selected based on solution of the homogeneous version of Eq. (3.2), i.e. $dy/dt = p(t)y(t)$, 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 $c(t)$

$$\frac{dc(t)}{dt}\exp\left(\int_{t_0}^{t} dt' p(t')\right) = g(t).$$

Applying the method of separable differential equations (see Eq. (3.1)) and then recalling the substitution (3.3), one arrives at

$$y(t) = \exp\left(\int_{t_0}^{t} dt' p(t')\right)\left(y_0 + \int_{t_0}^{t} dt' g(t')\exp(-\int_{t_0}^{t'} dt'' p(t''))\right).$$

Exercise 3.1.1. Solve $dx/dt - \lambda(t)x = f(t)/x^2$, where $\lambda(t)$ and $f(t)$ are known functions of t.

3.1.3 Integrals of Motion

Consider the conservative version of Eqs. (**??**) (conservative means there is no dissipation of energy)

$$\dot{x} = v, \quad \dot{v} = -\partial_x U(x), \tag{3.4}$$

describing the dynamics of a particle of unit mass in the potential, $U(x)$. The energy of the particle is

$$E = \frac{\dot{x}^2}{2} + U(x), \tag{3.5}$$

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,

$$\dot{x} = \pm 2\sqrt{E - U(x)}, \tag{3.6}$$

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. (3.7) is separable, and it can thus be integrated resulting in the following classic implicit expression for the particle coordinate as a function of time

$$\int_{x_0} \frac{dx}{\sqrt{E - U(x)}} = \pm t, \tag{3.7}$$

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 quantitiy H, called Hamiltonian, which is a twice-differentiable function of $2n$ variables, $p_1, \cdots, p_n$ (momenta) and $q_1, \cdots, q_n$ (coordinates), that satisfy the following system of equations, called Hamilton's canonical equations,

$$\forall i = 1, \cdots, N: \quad \dot{p}_i = -\frac{\partial H}{\partial q_i}, \quad \dot{q}_i = \frac{\partial H}{\partial p_i}. \tag{3.8}$$

Computing the rate of change of the Hamiltonian in time

$$\frac{dH}{dt} = \sum_{i=1}^{N} \left(\frac{\partial H}{\partial p_i}\dot{p}_i + \frac{\partial H}{\partial q_i}\dot{q}_i \right) = \sum_{i=1}^{N} \left(-\dot{q}_i\dot{p}_i + \dot{p}_i\dot{q}_i \right), \tag{3.9}$$

we observe that H is constant, that is, H is an integral of motion.

The one degree of freedom system (3.4) is an example of Hamilton's canonical system where the energy (3.5), considered as a function of x and v, is the Hamiltonian and x and v correspond to (scalar) q and p respectively. We will continue exploring the one degree of freedom system in section 3.2.

3.2 Phase Space Dynamics for Conservative and Perturbed Systems

3.2.1 Phase Portrait

Here we will follow material of [2] and Section 1.3 of [1]. Our starting point (and main example) will be the conservative (Hamiltonian) system with one degree of freedom (3.4).

We have established 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. (3.4).

Consider the quadratic potential, $U(x) = \frac{1}{2}kx^2$. The two cases of positive and negative k are illustrated in Fig. (3.1), 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. (3.1)).

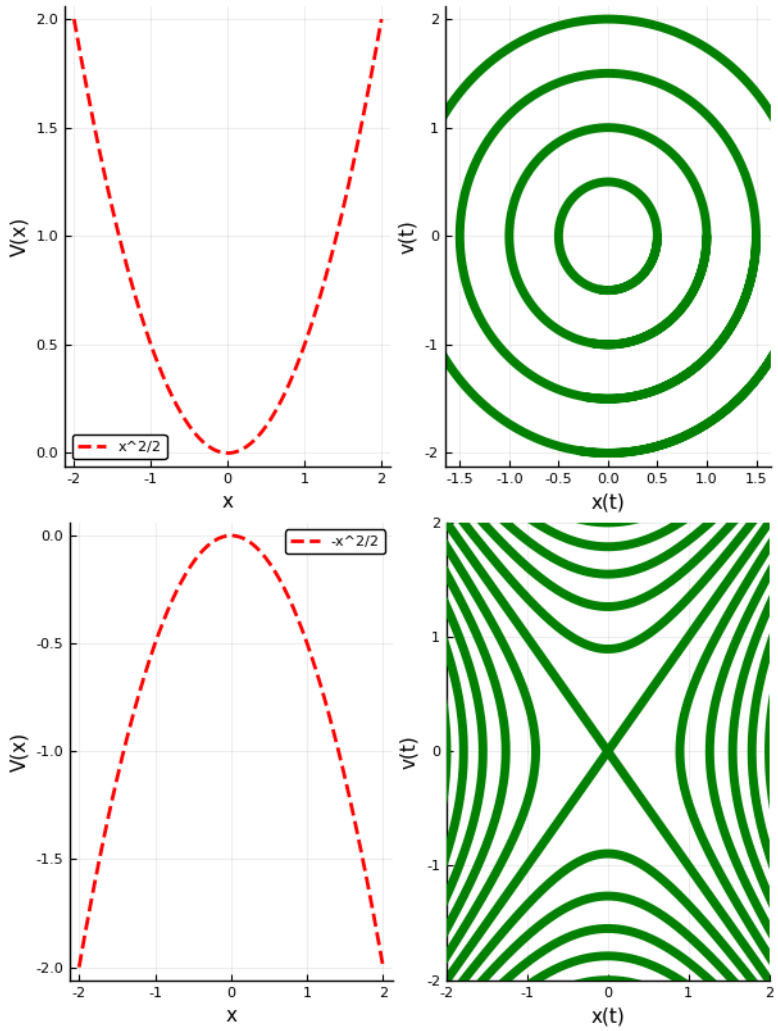


Figure 3.1: Phase portrait, i.e. (x, v) level-curves of the conservative system Eq. (3.4) with the potential, $U(x) = kx^2/2$ with $k > 0$ (top) and $k < 0$ (bottom).

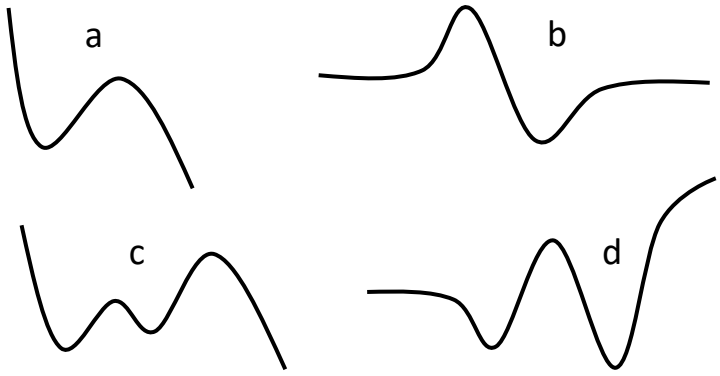


Figure 3.2: What is appearance of the level curves (phase portrait) of the energy for each of these potentials?

Consider the case where $k > 0$, and fix the value of the energy E. Due to Eq. (3.5), 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 motion is periodic, and its period, T, can be computed evaluating Eq. (3.7) between the turning points

$$T := \int_{x_-}^{x_+} \frac{dx}{\sqrt{E - U(x)}} = \int_{-\sqrt{2E/k}}^{\sqrt{2E/k}} \frac{dx}{\sqrt{E - kx^2/2}} = 2\pi. \tag{3.10}$$

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. (3.1)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.

Exercise 3.2.1. 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. (3.2).

3.2.2 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 $k = 1$:

$$\dot{x} = v + \varepsilon f(x, v), \quad \dot{v} = -x + \varepsilon g(x, v), \tag{3.11}$$

in the regime where $\varepsilon \ll 1$ and $x^2 + v^2 \leq R^2$.

For $\varepsilon = 0$, and assuming that $x^{(0)}(0) = x_0$, one derives

$$x^{(0)}(t) = x_0 \cos(t), \quad v^{(0)}(t) = -x_0 \sin(t).$$

We calculate the energy and find that $E = (x^{(0)})^2 + (v^{(0)})^2)/2$, which is obviously 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

$$\frac{d}{dt}E = x\dot{x} + v\dot{v} = \varepsilon\,(xf + vg) = \varepsilon\left(x^{(0)}f + v^{(0)}g\right) + O(\varepsilon^2). \tag{3.12}$$

Integrating over a period, one arrives at the following expression for the gain (or loss) of energy

$$\Delta E = \varepsilon \int_0^{2\pi} dt\left(x^{(0)}f + v^{(0)}g\right) + O(\varepsilon^2) = \varepsilon \oint (-f dv + g dx) + O(\varepsilon^2), \tag{3.13}$$

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

$$\ddot{x} = -x + \varepsilon\dot{x}(1 - x^2). \tag{3.14}$$

As in Eq. (3.13), we integrate $\frac{d}{dt}E$ over a period, which in this case gives

$$\begin{aligned}\Delta E &= \varepsilon \int_0^{2\pi} \dot{x}^2(1 - x^2)dt + O(\varepsilon^2) = \varepsilon x_0^2 \int_0^{2\pi} \sin^2 t\left(1 - x_0^2 \cos^2 t\right) dt + O(\varepsilon^2) \\ &= \pi\left(x_0^2 - \frac{x_0^4}{4}\right)\varepsilon + O(\varepsilon^2).\end{aligned} \tag{3.15}$$

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 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 3.2.2. Recall two properties of stable / unstable limit cycles:

$$\text{Stable Limit Cycle at } x = x_0 \text{ if } \Delta E(x_0) = 0 \quad \text{and} \quad \frac{d}{dx_0}\Delta E(x_0) < 0$$
$$\text{Unstable Limit Cycle at } x = x_0 \text{ if } \Delta E(x_0) = 0 \quad \text{and} \quad \frac{d}{dx_0}\Delta E(x_0) > 0$$

Suggest an example of perturbations, f and g, in Eq. (3.11) 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 $x_0 = 3$. Illustrate your suggested perturbations by building a computational snippet.

Consider another ODE example

$$\dot{I} = \varepsilon\,(a + b\cos\theta)\,, \quad \dot{\theta} = \omega, \tag{3.16}$$

where $\omega, \varepsilon, a, b$ are constants, and ε-term in the first Eq. (3.16) 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. (3.11).

Exercise 3.2.3. (a) Show that one can transform the unperturbed (i.e. $\varepsilon = 0$) version of the system described by Eq. (3.11) to the unperturbed version of the system described by Eq. (3.16) via the following transformation (change of variables)

$$v = \sqrt{I/2}\cos(\theta/\omega), \quad x = \sqrt{I/2}\sin(\theta/\omega). \tag{3.17}$$

(b) Restate Eq. (3.16) in the (x, v) variables.

The transformation discussed in the Exercise 3.2.3 is an example of 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. (3.16) as

$$\dot{\theta} = \frac{\partial H}{\partial I} = \omega, \quad \dot{I} = -\frac{\partial H}{\partial \theta} = 0, \tag{3.18}$$

therefore interpreting θ and I as the new coordinate and the new momentum respectively.

Averaging perturbed Eq. (3.16) over one ($2\pi\omega$) angle revolution, as done in Section 3.2.2, one arrives at

$$\Delta J = 2\pi\varepsilon a. \tag{3.19}$$

Taking many, $2\pi n\omega$, revolutions and replacing $2\pi n$ by t in the limit one arrives at the following equation for the averaged (over period) action

$$\dot{J} = \varepsilon a, \tag{3.20}$$

which has the solution, $J(t) = J_0 + \varepsilon a t$.

In fact Eqs. (3.16) can also be solved exactly

$$I(t) = \varepsilon a t + \frac{\varepsilon b \sin(\omega t)}{\omega}, \tag{3.21}$$

and one can check that indeed solution of the averaged Eq. (3.20) do not deviate (with time) from the exact solution of Eq. (3.16)

$$\omega \neq 0: \quad |J(t) - I(t)| \leq O(1)\varepsilon. \tag{3.22}$$

In a general n-dimensional case one considers the following system of bare (unperturbed) differential equations

$$\dot{\boldsymbol{I}} = 0, \quad \dot{\boldsymbol{\theta}} = \boldsymbol{\omega}(\boldsymbol{I}), \quad \boldsymbol{I} \doteq (I_1, \cdots, I_n), \quad \boldsymbol{\theta} \doteq (\phi_1, \cdots, \theta_n), \tag{3.23}$$

where thus each component of $\boldsymbol{I}$ is an integral of motion of the unperturbed system of equations. Perturbed version of Eq. (3.23) becomes

$$\dot{\boldsymbol{I}} = \varepsilon g(\boldsymbol{I}, \boldsymbol{\theta}, \varepsilon), \quad \dot{\boldsymbol{\theta}} = \boldsymbol{\omega}(\boldsymbol{I}) + \varepsilon \boldsymbol{f}(\boldsymbol{I}, \boldsymbol{\theta}, \varepsilon), \tag{3.24}$$

where f and g are 2π-periodic functions of each of the components of $\boldsymbol{\phi}$. 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, $\boldsymbol{J}(t) = \boldsymbol{I}(t) + \boldsymbol{O}(\varepsilon)$:

$$\dot{\boldsymbol{J}} = \varepsilon \boldsymbol{G}(\boldsymbol{J}), \quad \boldsymbol{G}(\boldsymbol{J}) \doteq \frac{\oint g(\boldsymbol{I}, \boldsymbol{\theta}, 0) d\boldsymbol{\theta}}{\oint d\boldsymbol{\theta}}, \tag{3.25}$$

where as in Section 3.2.2 $\oint$ 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. (3.24) become

$$\dot{\boldsymbol{I}} = -\frac{\partial H}{\partial \boldsymbol{\theta}}, \quad \dot{\boldsymbol{\theta}} = \frac{\partial H}{\partial \boldsymbol{I}}. \tag{3.26}$$

In this case averaging over $\boldsymbol{\theta}$ the rhs of the first equation in Eq. (3.26) 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.

3.3 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.

3.3.1 Homogeneous ODEs with Constant Coefficients

Consider the n-th order homogeneous ODE with constant coefficients

$$\mathcal{L}x(t) = 0, \quad \text{where} \quad \mathcal{L} \equiv \sum_{m=0}^{n} a_{n-m} \frac{d^{n-m}}{dt^{n-m}}. \tag{3.27}$$

(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. (3.27) in the form of a linear combination of exponentials

$$x(t) = \sum_{k=1}^{n} c_k \exp(\lambda_k t), \tag{3.28}$$

where c_k are constants. Substituting Eq.(3.28) into Eq.(3.27), one arrives at the condition that the λ_k are roots of the characteristic polynomial:

$$\left(\sum_{m=0}^{n} a_{n-m} \left(\lambda_k\right)^{n-m} \right) = 0. \tag{3.29}$$

Eq. (3.28) holds if the λ_k are not degenerate (that is, if there are n distinct solutions). In the case of degeneracy we generalize Eq. (3.28) 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.

$$x(t) = \sum_{k=1}^{m} \left(\sum_{l=0}^{d_k} c_k^{(l)} t^l \right) \exp(\lambda_k t), \tag{3.30}$$

where d_k is the degree of the k-th root degeneracy.

3.3.2 Inhomogeneous ODEs

Consider an inhomogeneous version of a generic linear ODE

$$\mathcal{L}x(t) = f(t). \tag{3.31}$$

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. (3.31) 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:

$$\ddot{x} + \omega_0^2 x = \cos(3t). \tag{3.32}$$

A generic solution of the homogeneous version of Eq. (3.32) is $x_0(t) = c\exp(i\omega_0 t)$, where c is a complex-valued constant, and a particular solution of Eq. (3.32) is $x_p(t) = \cos(3t)/(\omega^2 - 9)$. Therefore, a general solution of Eq. (3.32) is

$$x(t) = c\exp(i\omega_0 t) + \frac{\cos(3t)}{\omega^2 - 9}.$$

3.4 Linear Dynamics via the Green Function

Let us recall some of the empirical lessons of Section 3.2. If a 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 and the system is dissipative, so it relaxes back to the equilibrium. The relaxation may not be monotonical, and the system may show some oscillations. In the following we discuss the relaxation of a system back to its equilibrium state in response to a small perturbation. This type of relaxation is modeled by linear differential equations.

The method of Green function, or "response" functions, will be the working horse of our analysis for linear dynamics. It offers a powerful and intuitive approach which also extends to the case of PDEs. We will start exploring the method by revisiting the simple constant coefficient case of the linear scalar-valued first-order equation (3.2).

3.4.1 Evolution of a linear scalar

Consider the simplest example of scalar relaxation

$$\frac{d}{dt}x + \gamma x = \phi(t), \tag{3.33}$$

where γ is constant and $\phi(t)$ 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 $\phi(t)$ 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 is

$$x(t) = \int_0^\infty ds\, G(s)\phi(t-s), \tag{3.34}$$

where we have assumed that the evolution starts at $-\infty$ where $x(0) = 0$; and $G(t)$ is the so-called Green function which satisfies

$$\frac{d}{dt}G + \gamma G = \delta(t), \tag{3.35}$$

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. (3.33) would be defined unambiguously. Indeed, suppose $x_s(t)$ is a particular solution of Eq. (3.33), then $x_s(t) = C\exp(-\gamma t)$, where C is a constant, describes a family of solutions of Eq. (3.33). 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. (3.34) that defining the Green function, one also enforces that, $G(t) = 0$ at $t < 0$. This formal observation is, of course, consistent with the obvious—solutions of Eq. (3.33) at a particular moment in time t can only depend on external driving sources $\phi(t_0)$ that occured in the past, when $t \le t_0$, and cannot depend on external driving forces that will occur in the future, when $t > t_0$.

Now back to solving Eq. (3.35). Since $\delta(t) = 0$ at $t > 0$, one associates $G(t)$ with the zero mode of the aforementioned differential operator, $G(t) = A\exp(-\gamma t)$, where A is a constant. On the other hand due to the causality principle, $G(t) = 0$ at $t < 0$. Integrating Eq. (3.35) over time from $-\epsilon < 0$, where $0 < \epsilon \ll 1$, to τ, we observe that $G(t)$ should have a discontinuity (jump) at $t = 0$: $G(t) = A\exp(-\gamma t)\theta(t)$, where θ is the Heaviside function. Substituting the expression in Eq. (3.35) and integrating the result (left and right hand

sides of the resulting equality) over $-\epsilon < t < \epsilon$, one finds that $A = 1$. Substituting the expression into Eq. (3.34) one arrives at the solution

$$x(t) = \int_{-\infty}^{t} ds \exp(-\gamma(t-s))\phi(s). \tag{3.36}$$

We observe that the system "forgets" the past at the rate γ per unit time.

Exercise 3.4.1. Solve Eq. (3.33) at $t > 0$, where $x(0) = 0$ and $\phi(t) = A\exp(-\alpha t)$. Analyze the dependence on α and γ, including $\alpha \to \gamma$.

Notice that Eq. (3.35) assumes that the Green function depends on the difference between t and s, $t - s$, and not on the two variables separately. This assumption is justified for the case considered here, however it will not be correctfor 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 too, $G(t; x)$. In the case of the constant γ the Green function depends on the difference because of Eq. (3.35) symmetry with respect to the time translation (time homogeneity): the form of the equation does not change under the time shift, $t \to t + t_0$.

3.4.2 Evolution of a vector

Let us now generalize and consider

$$\frac{d}{dt}\boldsymbol{y} + \hat{\boldsymbol{\Gamma}}(t)\boldsymbol{y} = \boldsymbol{\chi}(t), \tag{3.37}$$

where $\boldsymbol{y}$ and $\boldsymbol{\phi}$ are n-dimensional vectors and $\hat{\boldsymbol{\Gamma}}$ is $n \times n$ time-independent matrix.

Note that this type of vector ODE appear in the result of "vectorization" of an n-the order ODE for a scalar variable x, where $y_1 = x,\ y_2 = dx/dt,\ \cdots,\ y_n = d^{n-1}x/dt^{n-1}$. Then $d\boldsymbol{y}/dt$ is expressed via the components of $\boldsymbol{y}$ and the original equation, thus resulting in Eq.(3.37).

Consider the following auxiliary linear algebra problem: find the eigen-set of the matrix Γ

$$\hat{\boldsymbol{\Gamma}}\boldsymbol{a}_i = \lambda_i \boldsymbol{a}_i, \tag{3.38}$$

where λ_i are eigen-values of Γ.

Let us assume, first, that the eigen-value problem is not degenerate. Then we expand $\boldsymbol{y}$ and $\boldsymbol{\chi}$ over the $\{\boldsymbol{a}_i|i\}$ basis,

$$\boldsymbol{y} = \sum_i x_i \boldsymbol{a}_i, \quad \boldsymbol{\chi} = \sum_i \phi_i \boldsymbol{a}_i. \tag{3.39}$$

Substituting the expansions into Eq. (3.37) one arrives at

$$\frac{dx_i}{dt} + \lambda_i x_i = \phi_i, \tag{3.40}$$

therefore reducing the vector equation to the set of scalar equations of the already considered type Eqs. (3.33).

To make this transformation invariant, and also extendable to the degenerate case (when at least two-eigenvalues of $\hat{\mathbf{\Gamma}}$ are equal) one introduces Green function $\hat{\boldsymbol{G}}$, which satisfies

$$\left(\frac{d}{dt} + \hat{\mathbf{\Gamma}}\right) \hat{\boldsymbol{G}}(t) = \delta(t)\hat{\mathbf{1}}. \tag{3.41}$$

The explicit solution of Eq. (3.41) is

$$\hat{\boldsymbol{G}}(t) = \theta(t) \exp\left(-\hat{\mathbf{\Gamma}}t\right), \tag{3.42}$$

which allows us to state the solution of Eq. (3.37) in the following invariant form

$$\boldsymbol{y}(t) = \int_{-\infty}^{t} ds \hat{\boldsymbol{G}}(t-s)\boldsymbol{\chi}(s) = \int_{-\infty}^{t} ds \theta(t-s) \exp\left(-\hat{\mathbf{\Gamma}}(t-s)\right) \boldsymbol{\chi}(s). \tag{3.43}$$

Notice that matrix exponential, introduced in Eq. (3.42) and utilized in Eq. (3.43), is the formal expression which may be interpreted in terms of the Taylor series

$$\exp\left(-t\hat{\mathbf{\Gamma}}\right) = \sum_{n=0}^{\infty} \frac{(-t)^n \hat{\mathbf{\Gamma}}^n}{n!}, \tag{3.44}$$

which is always convergent (for the matrix $\hat{\boldsymbol{G}}$ with finite elements).

To relate the invariant expression (3.43) to the eigen-value decomposition of Eqs. (3.39,3.40) one introduces the eigen-decomposition

$$\hat{\mathbf{\Gamma}} = \hat{\mathbf{A}}\hat{\mathbf{\Lambda}}\hat{\mathbf{A}}^{-1}, \tag{3.45}$$

where $\hat{\mathbf{\Lambda}}$ is the diagonal matrix formed from the eigenvalues of $\hat{\mathbf{\Gamma}}$ and the columns of $\hat{\mathbf{A}}$ are respective eigenvalues of $\hat{\mathbf{\Gamma}}$. Note that $\hat{\mathbf{\Gamma}}^n = \hat{\mathbf{A}}\hat{\mathbf{\Lambda}}^n\hat{\mathbf{A}}^{-1}$.

To illustrate the peculiarity of the degenerate case consider

$$\hat{\mathbf{\Gamma}} = \lambda\hat{\mathbf{1}} + \hat{\mathbf{N}}, \quad \hat{\mathbf{N}} \equiv \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix},$$

which is the canonical form of the Jordan (2×2) matrix/block, where $\hat{\mathbf{N}}$ is (2×2) nilpotent matrix, i.e. $\hat{\mathbf{N}}^2 = \hat{\mathbf{0}}$. Writing Eqs. (3.37) in components

$$\frac{dy_1}{dt} + \lambda y_1 + y_2 = \chi_1, \quad \frac{dy_2}{dt} + \lambda y_2 = \chi_2,$$

integrating the second equation, substituting result in the first equation, and then changing from y_1 to $y = y_1 + ty_2$, one arrives at

$$\frac{dy}{dt} + \lambda y = \chi_1 + t\chi_2.$$

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. Consistently, expression for the matrix exponential also show a secular term

$$\exp\left(-t\begin{pmatrix}\lambda & 1\\ 0 & \lambda\end{pmatrix}\right) = e^{-\lambda t}\left(1 - t\hat{\mathbf{N}}\right),$$

where we have accounted for the nilpotent property of $\hat{\mathbf{N}}$.

Exercise 3.4.2. Find the Green function of Eq. (3.37) for

$$\hat{\mathbf{\Gamma}} = \begin{pmatrix}\lambda & 1 & 0\\ 0 & \lambda & 1\\ 0 & 0 & \lambda\end{pmatrix}.$$

3.4.3 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

$$\frac{d^2}{dt^2}x + \omega^2 x = \phi(t). \tag{3.46}$$

To solve Eq. (3.46) note that its general solution can be expressed as a sum of its particular solution and solution of the homogeneous version of Eq. (3.46) with zero right hand side. Let us choose a particular solution of Eq. (3.46) in the form of convolution (3.34) of the source term, $\phi(t)$, with the Green function of Eq. (3.46)

$$\left(\frac{d^2}{dt^2} + \omega^2\right) G(t) = \delta(t). \tag{3.47}$$

As established above $G(t) = 0$ at $t < 0$. Integration of Eq. (3.47) 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. (3.47), defining $G(t)$ at $t > +0$. Finally, we arrive at the following result

$$G(t) = \theta(t)\frac{\sin(\omega t)}{\omega}, \tag{3.48}$$

where θ is the Heaviside function.

Furthermore, Eq. (3.34) gives the solution to Eq. (3.46) over the infinite time horizon, however one can also use the Green function to solve the respective Cauchy propblem (initial value problem). Since Eq. (3.46) 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. (3.46)

$$x(t) = \dot{x}(0)G(t) + x(0)\dot{G}(t) + \int_0^t dt_1 G(t - t_1)\phi(t_1). \tag{3.49}$$

Let us now generalize and consider

$$\mathcal{L}x = \phi(t), \quad \mathcal{L} \equiv \sum_{k=0}^{n} a_{n-k}\frac{d^{n-k}}{dt^{n-k}}, \tag{3.50}$$

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 3.3. We build a particular solution of Eq. (3.50) as the convolution (3.34) of the source term, $\phi(t)$, with the Green function, $G(t)$, of Eq. (3.50)

$$\mathcal{L}G = \delta(t), \tag{3.51}$$

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

$$x(t) = \sum_i b_i \exp(z_i t), \tag{3.52}$$

where b_i are arbitrary constants.

Let us now use the general representation (3.54) to construct the Green function solving Eq. (3.51). 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

$$\frac{d^{n-1}}{dt^{n-1}}G(0^+) = 1, \quad \forall 0 \le m < n-1: \quad \frac{d^m}{dt^n}G(0^+) = 0. \tag{3.53}$$

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

$$t > 0^+ : \quad G(t) = \sum_i b_i \exp(z_i t), \tag{3.54}$$

where b_i are arbitrary constants which are defined unambiguously from the system of algebraic equations for the coefficients one derives substituting Eq. (3.53) in Eq. (3.54).

Exercise 3.4.3. Find Green function of

(a) $\dfrac{d^2}{dt^2}x + 2\gamma\dfrac{d}{dt}x + \nu^2 x = \phi,$

(b) $\dfrac{d^4}{dt^4}x + 4\nu^2\dfrac{d^2}{dt^2}x + 3\nu^4 x = \phi.$

(c) $\left(\dfrac{d^2}{dt^2} + \nu^2\right)^2 x = \phi$

3.4.4 Laplace's Method for Dynamic Evolution

So far we have solved linear ODE 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 order of the differential equation). In this section we discuss an alternative way of solving the problem via application of the Laplace transform introduced in Section 2.8.

Laplace's method is natural for solving dynamic problems with causal structure. Let us see how it works for finding the Green function defined by Eq. (3.51). We apply the Laplace transform to Eq. (3.51), integrating it over time with the $\exp(-kt)$ Laplace weight from a small positive value, ϵ, to ∞. In this case 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

$$L(k)\tilde{G}(k) = 1, \quad L(k) \doteq \sum_{k=0}^{n} a_{n-k}(-k)^{n-1}. \tag{3.55}$$

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

$$G(t) = \int_{c-i\infty}^{c+\infty} \frac{dk}{2\pi i} \exp(kt)\tilde{G}(k). \tag{3.56}$$

The Laplace method also allows us to solve ODEs of the following type

$$\sum_{m=0}^{N} (a_m + b_m x) \frac{d^m Y}{dx^m} = 0, \tag{3.57}$$

where the coefficients are linear in x.

Let us look for solution of Eq. (3.57) in the form

$$Y(x) = \int_C dt Z(t) \exp(xt), \tag{3.58}$$

where C is a contour in the complex plane of t selected in a way that the integral has the value which is finite and nonzero. Substituting Eq. (3.57) with the weight Eq. (3.58), using the relation $xe^{xt} = de^{xt}/dt$, and assuming that the contour of integration in Eq. (3.57) is such that no "contact" term appears after the integration by parts (this is satisfied, e.g. when the contour is closed and the integrand is single-valued along the contour), one arrives at

$$\frac{d}{dt}(QZ) = PZ, \quad \text{where} \quad P(t) = \sum_{m=0}^{n} a_m t^m, \quad Q(t) = \sum_{m=0}^{n} b_m t^m, \tag{3.59}$$

which is solved by

$$Z(t) = \frac{1}{Q} \exp\left(\int \frac{P}{Q} dt \right), \tag{3.60}$$

where the integral is defined simply as the anti-derivative.

This is a generic recipe - let us now apply it to a particular case of the so-called Hermite equation

$$\frac{d^2Y}{dx^2} - 2x\frac{dY}{dx} + 2nY = 0. \tag{3.61}$$

In this case we derive

$$P = t^2 + 2n, \quad Q = -2t, \quad Z = -\frac{\exp(-t^2/4)}{2t^{n+1}}, \tag{3.62}$$

thus resulting in the following explicit solution of Eq. (3.61) (written in quadrature, and defined up to a multiplicative constant)

$$Y(x) = \int e^{xt - t^2/4} \frac{dt}{t^{n+1}} = e^{x^2} \int_C \frac{e^{-u^2} du}{(u-x)^{n+1}}, \tag{3.63}$$

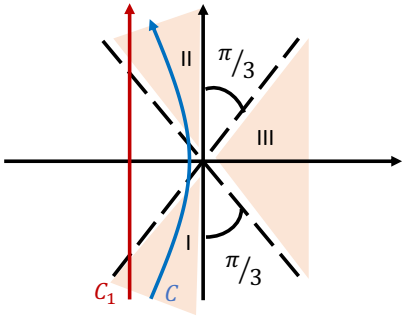


Figure 3.3: Layout of contours in the complex plane of t needed for saddle-point estimations of the Airy function described in Eq. (3.68).

where we also change variables $t \to u$ according to $t = 2(x - u)$.

When n is a nonegative integer, the integrand in Eq. (3.63) has a simple pole, and thus choosing the contour to go around the pole works (in the sense of satisfying the "no contact" term requirement). Applying the Cauchy formula to the resulting constour integral, one therefore arrives at the expression for the so-called Hermite polynomials

$$Y(x) = H_n(x) = (-1)^n e^{x^2} \frac{d^n}{dx^n} e^{-x^2}, \tag{3.64}$$

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.

Exercise 3.4.4. Prove that

$$\int_{-\infty}^{+\infty} dx\, e^{-x^2} H_n(x) H_m(x) = 2^n n! \sqrt{\pi} \delta_{nm}, \tag{3.65}$$

where δ_{nm} is unity when $n = m$ and it is zero otherwise (Kronecker symbol).

Hermite polynomials will come back later in the context of the Sturm-Liouville problem.

Consider another example of the equation which can be solved by the Laplace method

$$\frac{d^2}{dx^2} Y - xY = 0. \tag{3.66}$$

Following the general Laplace method we derive

$$P = t^2, \quad Q = -1, \quad Z = -\exp(-t^2/3). \tag{3.67}$$

According to Eq. (3.59) general solution of Eq. (3.67) can be represented as

$$Y(x) = \text{const} \int_C \exp\left(xt - t^3/3\right), \tag{3.68}$$

where we choose an infinite integration path shown in Fig. (3.3) 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 $\mathrm{Re}(t^2) > 0$ (shaded regions I, II, III in Fig. (3.3)). Moreover, by choosing that the contour starts in the region I and ends in the region II (blue contour C in Fig. (3.3)) we guarantee that the Airy function given by Eq. (3.66) remains finite at $x \to +\infty$. 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 t plane shown in Fig. (3.3), then Eq. (3.68) becomes (up to a constant) the so-called Airy function

$$Ai(x) = \frac{1}{\pi}\int\limits_0^\infty \cos\left(\frac{u^3}{3} + xu\right) = \frac{1}{2\pi}\mathrm{Re}\left(\int\limits_{-\infty}^{\infty} \exp\left(i\frac{u^3}{3} + ixu\right)\right). \tag{3.69}$$

Asymptotic expression for the Airy function at $x > 0,\ x \gg 1$, can be derived utilizing the saddle-point method described in Section 1.4. At $x = \pm\sqrt{x}$, the integrand in Eq. (3.68) 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. (3.3) which crosses the real axis at $t = -\sqrt{x}$.) The integral is dominated by the saddle-point at $t = -\sqrt{x}$, thus resulting (after substitution $t = \sqrt{x} + iu$, changing integration variable from t to u, making expansion over u, keeping quadratic term in u, ignoring higher order terms, and evaluating a Gaussian integral) in the following asymptotic estimation for the Airy function

$$x > 0,\ x \gg 1: \quad Ai(x) \approx \frac{1}{2\pi}\int\limits_{-\infty}^{+\infty} \exp\left(-\frac{2}{3}x^{3/2} - \sqrt{x}u^2\right) du = \frac{\exp(-2x^{3/2}/3)}{x^{1/4}\sqrt{4\pi}}. \tag{3.70}$$

(Notice that one can also provide an alternative argument and exclude contribution of the second, potentially dominating, saddle-point $t = \sqrt{x}$ 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. (3.69).)

3.5 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.

3.5.1 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 function $f(x)$ whose distribution over a finite spatial interval is described by the following set of equations

$$\frac{d^2}{dx^2}f = \psi(x), \quad \forall x \in (a,b) \quad \text{with} \quad f(a) = f(b) = 0. \tag{3.71}$$

We introduce the Green function which satisfies

$$\forall a < x, y < b: \quad \frac{d^2}{dx^2}G(x;y) = \delta(x-y), \quad G(a;y) = G(b;y) = 0. \tag{3.72}$$

Notice that the Green function now depends on both x and y.

According to Eq. (3.72), $\frac{d^2}{dx^2}G(x;y) = 0$ if $x = y$. Then enforcing the boundary conditions one derives

$$x > y: \quad G(x;y) = B(x-b), \tag{3.73}$$

$$y > x: \quad G(x;y) = A(x-a). \tag{3.74}$$

Furthermore, given that the differential equation in (3.72) 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

$$G(x;y) = \frac{1}{b-a}\begin{cases}(y-b)(x-a), & x < y\\ (y-a)(x-b), & x > y.\end{cases} \tag{3.75}$$

The solution the Eq. (3.71) is given by the convolution operator

$$f(x) = \int_a^b dy G(x;y)\psi(y). \tag{3.76}$$

Exercise 3.5.1. Find Green function of the operator $d^2/dx^2 + \kappa^2$ for periodic functions with the period 2π.

3.6 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.

3.6.1 Hilbert Space and its completeness

Let us first review some basic properties of a Hilbert space, in particular, and 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, $\|f\|^2 \doteq (f, f) > 0$, 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 turn into an ortho-normal basis with respect to a given scalar product, i.e. $x = \sum_{n=1}^{\infty}(x, f_n)f_n$, $\|x\|^2 = \sum_{n=1}^{\infty}|(x, f_n)|^2$. (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 $\int_\Omega dx|f(x)|^2 < \infty$ (one may say, casually, that the square modulus of the function is integrable). In this case the scalar product is defined as

$$(f, g) \doteq \int_\Omega dx f^*(x)g(x).$$

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

$$\sum_{n=1}^{\infty} f_n^*(x) f_m(y) = \delta(x-y). \tag{3.77}$$

As custom for the δ function (an other generalized functions), Eq. (3.77) should be understood as equality of integrals of the two sides of Eq. (3.77) integrated with a function from $L^2(\Omega)$.

3.6.2 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, $a \leq x \leq b$ 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:

$$(f, \hat{L}g) := \int_a^b dx\, f(x)\hat{L}g(x) = \int_a^b dx\, g(x)\hat{L}f(x) = (g, \hat{L}f). \tag{3.78}$$

It is clear from how the condition (3.78) 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

$$\hat{L} = \frac{d^2}{dx^2} + U(x), \tag{3.79}$$

where $U(x)$ is a function mapping from $\mathbb{R}$ to $\mathbb{R}$, is Hermitian.

Natural generalization of the Shrödinger operator 3.79 is the Sturm-Liouville operator

$$\hat{L} = \frac{d^2}{dx^2} + Q\frac{d}{dx} + U(x). \tag{3.80}$$

The Sturm-Liouville operator is not Hermitian, i.e. Eq. (3.78) 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. (3.78) holds

$$\int_a^b dx\rho(x) f(x)\hat{L}g(x) = \int_a^b dx\rho(x) g(x)\hat{L}f(x), \tag{3.81}$$

$$\text{where} \quad \frac{d}{dx}\rho = Q\rho \Rightarrow \rho = \exp\left(\int dx Q\right). \tag{3.82}$$

Consider now the eigen-functions f_n of the operator $\hat{L}$, which satisfy

$$\hat{L}f_n = \lambda_n f_n, \tag{3.83}$$

where λ_n is the spectral parameter (eigenvalue) of the eigen-function, f_n, of the Sturm-Liouville operator (3.80), 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. (3.84) 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-Liuville problem.

Observe that the conditions (3.81,3.82) translates into

$$\int dx\rho f_n \hat{L} f_m = \lambda_m \int dx\rho f_n f_m = \lambda_n \int dx\rho f_n f_m, \tag{3.84}$$

that becomes the following eigen-function orthogonality condition

$$\forall n \neq m: \quad \int dx\rho f_n f_m = 0. \tag{3.85}$$

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. (3.85) 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 $m \neq n$,

$$\int_0^{2\pi} dx \cos(nx)\cos(mx) = \int_0^{2\pi} dx \cos(nx)\sin(mx) = \int_0^{2\pi} dx \sin(nx)\sin(mx) = 0. \tag{3.86}$$

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

$$f = \sum_n c_n f_n. \tag{3.87}$$

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. (3.87) by ρf_n, integrating over the domain, and applying (3.85) to the right one derives

$$c_n = \frac{\int dx \rho f_n f}{\int dx \rho (f_n)^2}. \tag{3.88}$$

Note that for the example $\hat{L}_0$, Eq. (3.87) is a Fourier Series expansion of a periodic function.

Returning to the general case and substituting Eq. (3.88) back into (3.87), one arrives at

$$f(x) = \int dy \left(\rho(y) \sum_n \frac{f_n(x) f_n(y)}{\int dx \rho(x) (f_n(x))^2} \right) f(y). \tag{3.89}$$

If the set of functions $\{f_n(x)|n\}$ is complete relation (3.89) 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. (3.89) is just the $\delta(x)$, which is the special function which maps convolution of the function to itself, i.e.

$$\sum_n \frac{f_n(x) f_n(y)}{\int dx \rho(x) (f_n(x))^2} = \frac{1}{\rho(y)} \delta(x-y). \tag{3.90}$$

Therefore, one concludes that Eq. (3.90) is equivalent to the statement of the set of functions $\{f_n(x)|n\}$ completeness.

Exercise 3.6.1. Check validity of Eq. (3.90), 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.

3.6.3 Hermite Polynomials, Expansions

Let us now depart from our enabling example and consider the case of $Q(x) = -2x$ and $U(x) = 0$ over the class of functions mapping from $\mathbb{R}$ to $\mathbb{R}$ and decay sufficiently fast at $x \to \pm\infty$.

$$\hat{L}_2 = \frac{d^2}{dx^2} - 2x\frac{d}{dx}, \quad \rho(x) = \exp\left(-x^2\right). \tag{3.91}$$

That is we are discussing now

$$\hat{L}_2 f_n = \lambda_n f_n. \tag{3.92}$$

Changing from $f_n(x)$ to $\Psi_n(x) = f_n(x)\sqrt{\rho}$ one thus arrives at the following equation for Ψ_n:

$$e^{-x^2/2} \hat{L}_2 f_n(x) = e^{-x^2/2} \hat{L}_2 \left(e^{x^2/2} \Psi_n(x) \right) = \frac{d^2}{dx^2}\Psi_n + (1-x^2)\Psi_n = \lambda_n \Psi_n. \tag{3.93}$$

Observe that when $\lambda_n = -2n$, Eq. (3.92) coincides with the Hermite Eq. (3.61).

Let us look for solution of Eq. (3.92) in the form of the Taylor series around $x = 0$

$$f_n(x) = \sum_{k=0}^{\infty} a_k x^k. \tag{3.94}$$

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:

$$\forall k = 0, 1, \cdots : \quad a_{k+2} = \frac{2k + \lambda_n}{(k+2)(k+1)} a_k. \tag{3.95}$$

This results in the following two linearly independent solutions (even and odd, respectively, with respect to the $x \to -x$ transformation) of Eq. (3.92) represented in the form of a series

$$f_n^{(e)}(x) = a_0 \left(1 + \frac{\lambda_n}{2!}x^2 + \frac{\lambda_n(4+\lambda_n)}{4!}x^4 + \cdots\right), \tag{3.96}$$

$$f_n^{(o)}(x) = a_1 \left(x + \frac{(2+\lambda_n)}{3!}x^3 + \frac{(2+\lambda_n)(6+\lambda_n)}{5!}x^5 + \cdots\right), \tag{3.97}$$

where the two first coefficients in the series (3.94) are kept as the parameters. Observe that the series (3.96) and (3.97) 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. (3.93). Per statement of the Exercise 3.4.4, Hermite polynomials are normalized and orthogonal (weighted with ρ) to each other.

Exercise 3.6.2. Verify that the set of functions

$$\{\Psi_n(x) = \frac{1}{\pi^{1/4}\sqrt{2^n n!}} \exp(-x^2/2) H_n(x) | n = 0, 1, \cdots\}, \tag{3.98}$$

satisfy

$$\sum_{n=0}^{\infty} \Psi_n(x)\Psi_n(y) = \delta(x - y). \tag{3.99}$$

[Hint: use the following identity

$$\frac{d^n}{dx^n} \exp(-x^2) = \sqrt{\pi} \int_{-\infty}^{+\infty} \frac{dq}{2\pi} (iq)^n \exp\left(-q^2/4 + iqx\right).]$$

Statement of the Exercise 3.6.2 combined with the statement of Exercise 3.4.4 result in the statement of "completeness": the set of functions (3.99) 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.)

3.6.4 Schrödinger Equation in $1d$

Schrödinger equation

$$\frac{d^2\Psi(x)}{dx^2} + (E - U(x))\Psi(x) = 0 \tag{3.100}$$

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 or 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. (3.100) becomes

$$\underline{U_0 > E > 0:}$$
$$\Psi_E(x) = \begin{cases} c_L \exp(x\sqrt{U_0 - E}), & x < 0 \\ a_+ \exp(ix\sqrt{E}) + a_- \exp(-ix\sqrt{E}), & x \in [0, a] \\ c_R \exp(-x\sqrt{U_0 - E}), & x > a \end{cases}, \tag{3.101}$$
$$\underline{U_0 < E:}$$
$$\Psi_E(x) = \begin{cases} c_{L+} \exp(ix\sqrt{E - U_0}) + c_{L-} \exp(-ix\sqrt{E - U_0}), & x < 0 \\ a_+ \exp(ix\sqrt{E}) + a_- \exp(-ix\sqrt{E}), & x \in [0, a] \\ c_{R+} \exp(ix\sqrt{E - U_0}) + c_{R-} \exp(-ix\sqrt{E - U_0}), & x > a \end{cases}, \tag{3.102}$$

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. (3.101) or Eq. (3.102) result in an algebraic "consistency" conditions for E. We expect to get a continuous spectrum at $E > U_0$ and discrete at $U_0 > E > 0$.

Exercise 3.6.3. 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.

Example 3.6.4. Find eigen-functions and energy of stationary states of the Schrödinger equation for an oscillator:

$$\frac{d^2\Psi(x)}{dx^2} + (E - x^2)\Psi(x) = 0, \tag{3.103}$$

where $x \in \mathbb{R}$ and $\Psi : \mathbb{R} \to \mathbb{C}^2$.

As we saw already in the preceding section analysis of Eq. (3.103) is reduced to studying the Hermite equation, with its spectral version desribed by Eq. (3.92). However, we will follow another route here. Let us introduce the so-called "creation" and "annihilation" operators

$$\hat{a} = \frac{i}{\sqrt{2}}\left(\frac{d}{dx} + x\right), \quad \hat{a}^\dagger = \frac{i}{\sqrt{2}}\left(\frac{d}{dx} - x\right), \tag{3.104}$$

and then rewrite the Schrödinger Eq. (refeq:Schr-osc) as

$$\hat{H}\Psi(x) = \hat{a}^\dagger\hat{a}\Psi(x) = \left(2E - \frac{1}{2}\right)\Psi(x). \tag{3.105}$$

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

$$\hat{a}\hat{a}^\dagger\Psi(x) = \hat{a}^\dagger\hat{a}\Psi(x) + \Psi(x), \tag{3.106}$$

$$\begin{aligned}\hat{a}^\dagger\hat{a}\left(\hat{a}^\dagger\right)^n\Psi(x) &= \left(\hat{a}^\dagger\right)^2\hat{a}\left(\hat{a}^\dagger\right)^{n-1}\Psi(x) + \left(\hat{a}^\dagger\right)^n\Psi(x) \\ &= n\left(\hat{a}^\dagger\right)^n\Psi(x) + \left(\hat{a}^\dagger\right)^{n+1}\hat{a}\Psi(x).\end{aligned} \tag{3.107}$$

Introduce $\Psi_n(x) \doteq (\hat{a}^\dagger)^n\Psi_0(x)$. Since $\hat{a}\Psi_0(x) = 0$, the commutation relations (3.107) 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. (3.64),

$$\begin{aligned}\Psi_n(x) &= A_n\left(\frac{i}{\sqrt{2}}\left(\frac{d}{dx} - x\right)\right)^n\exp\left(-\frac{x^2}{2}\right) \\ &= A_n\frac{i^n}{2^{n/2}}\exp\left(\frac{x^2}{2}\right)\frac{d^n}{dx^n}\exp\left(-x^2\right),\end{aligned}$$

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 (3.65) one derives, $A_n = (n!\sqrt{\pi})^{-1/2}$.

Chapter 4

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 4.2). We will discuss how to generalize and solve elliptic PDE, normally associated with static problems, in section 4.3. Hyperbolic PDEs, discussed in section 4.4 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 4.5.

4.1 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 over the first-order derivatives.

Let $\theta(\boldsymbol{x}) : \mathbb{R}^d \to \mathbb{R}$ be a function of a d-dimensional coordinate, $\boldsymbol{x} := (x_1, \ldots, x_d)$. Introduce the gradient vector, $\boldsymbol{\nabla}_{\boldsymbol{x}}\theta := (\partial_{x_i}\theta; i = 1, \ldots, d)$, and consider the following linear in $\boldsymbol{\nabla}_{\boldsymbol{x}}\theta$ equation

$$(\boldsymbol{V} \cdot \boldsymbol{\nabla}_{\boldsymbol{x}}\theta) := \sum_{i=1}^{d} V_i \partial_{x_i}\theta = f, \tag{4.1}$$

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. (4.1)

$$(\boldsymbol{V} \cdot \boldsymbol{\nabla}_{\boldsymbol{x}} \theta) = 0. \tag{4.2}$$

Introduce an auxiliary parameter (or dimension) $t \in \boldsymbol{R}$, call it time, and then introduce the *characteristic equations*

$$\frac{d\boldsymbol{x}(t)}{dt} = \boldsymbol{V}(\boldsymbol{x}(t)), \tag{4.3}$$

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. (4.3) is a solution to Eq. (4.2), and that any function of the first integrals of Eqs. (4.3), $g(F_1, \ldots, F_k)$, is also a solution to Eq. (4.2).

Indeed, a direct substitution of $\theta = g$ in Eq. (4.2) leads to the following sequence of equalities

$$(\boldsymbol{V} \cdot \boldsymbol{\nabla}_{\boldsymbol{x}} g) = \sum_{i=1}^{k} \frac{\partial g}{\partial F_i} \sum_{j=1}^{d} \frac{\partial F_i}{\partial x_j} V_j = \sum_{i=1}^{k} \frac{\partial g}{\partial F_i} \frac{d}{dt} F_i = 0. \tag{4.4}$$

The system of equations (4.3) has $d-1$ first integrals independent of t (directly). Then a general solution to Eq. (4.4) is

$$\theta(\boldsymbol{x}(t)) = g\left(F_1(\boldsymbol{x}(t), \ldots, F_{d-1}(\boldsymbol{x}(t))\right), \tag{4.5}$$

where g is assumed to be sufficiently smooth (at least twice differential over the first integrals).

Eq. (4.2) 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. (4.2) means that derivative of θ 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 $\theta(\boldsymbol{x})$ 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. (4.2) becomes $d\theta/dt = 0$ along the characteristic.

Let us illustrate how to find a characteristic on the example of the following homogeneous PDE

$$\partial_x \theta(x, y) + y \partial_y \theta(x, y) = 0.$$

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 $\theta = g(F(x, y))$, 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.

Exercise 4.1.1. Find and visualize in the (x, y) plane characteristics of

(a) $\partial_x\theta - y^2\partial_y\theta = 0$,

(b) $x\partial_x\theta - y\partial_y\theta = 0$,

(c) $y\partial_x\theta - x\partial_y\theta = 0$.

Find general solutions to these PDEs, and verify your solutions by direct substitution.

Consider the following initial value (boundary) Cauchy problem: solve Eq. (4.2) subject to the boundary condition

$$\theta(\boldsymbol{x})|_{\boldsymbol{x}_0\in S} = \vartheta(\boldsymbol{x}_0), \tag{4.6}$$

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. (4.2). Consistently with what was described above solution to Eq. (4.3) with the initial/boundary condition Eq. (4.6) can be thought as the change of variables.

Let us illustrate the solution to the Cauchy problem on the example

$$\partial_x\theta = y\partial_y\theta, \qquad \theta(0, y) = \cos(y).$$

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

$$\theta(x, y) = g(y\exp(x)),$$

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:

$$\theta(0, y) = g(y) = \cos(y).$$

This results in the desired solution: $\theta(x, y) = \cos(y\exp(x))$.

Exercise 4.1.2. (a) Solve

$$y\partial_x\theta - x\partial_y\theta = 0,$$

for initial condition, $\theta(0, y) = y^2$. (b) Explain why the same problem with the initial condition $\theta(0, y) = y$ is ill-posed. (c) Discuss if the same problem with the initial condition, $\theta(1, y) = y^2$, is ill posed or not.

Exercise 4.1.3 (not graded)**.** Show that characteristic equations for the Liouville PDE

$$\partial_t f + \{H, f\} = 0, \quad \{H, f\} := \sum_{i=1}^{N} \left(\frac{\partial H}{\partial q_i} \frac{\partial f}{\partial p_i} - \frac{\partial H}{\partial p_i} \frac{\partial f}{\partial q_i} \right)$$

where $H(\boldsymbol{p}; \boldsymbol{q})$ is the Hamilton function and $\{H, f\}$ is the Poisson bracket of H and f, are the Hamilton Eqs. (3.8).

Now let us get back to the inhomogeneous Eq. (4.1). 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. (4.1) becomes

$$(\boldsymbol{V} \cdot \boldsymbol{\nabla}_{\boldsymbol{x}})\, \theta = (\dot{\boldsymbol{x}} \cdot \boldsymbol{\nabla}_{\boldsymbol{x}})\, \theta = \frac{d}{dt}\theta = f(\boldsymbol{x}(t)), \tag{4.7}$$

which can be integrated along the characteristic thus resulting in a desired particular solution to Eq. (4.1)

$$\theta_{inh} = \int_{t_0}^{t} f(\boldsymbol{x}(s)) ds (\boldsymbol{V} \cdot \boldsymbol{\nabla}_{\boldsymbol{x}}) \theta = (\dot{\boldsymbol{x}} \cdot \boldsymbol{\nabla}_{\boldsymbol{x}}) \theta. \tag{4.8}$$

Notice that this solution is not constant along characteristics.

Exercise 4.1.4. Solve the Cauchy problem for the following inhomogeneous equation

$$\partial_x \theta - y \partial_y \theta = y, \ \theta(0, y) = \sin(y).$$

The method of characteristics can also be generalized to quasi-liner first-order ODEs, (first-order ODEs (4.1) where $\boldsymbol{V}$ and f depend not only on the vector of coordinate, $\boldsymbol{x}$, but also on the function $\theta(\boldsymbol{x})$). In this case the characteristic equations become

$$\frac{d\boldsymbol{x}}{dt} = \boldsymbol{V}(\boldsymbol{x}, \theta), \quad \frac{d\theta}{dt} = f(\boldsymbol{x}, \theta). \tag{4.9}$$

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. (4.9).

Consider the example of the Hopf equation in $d = 1$

$$\partial_t u + u \partial_x u = 0, \tag{4.10}$$

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

$$u = u_0(x - ut). \tag{4.11}$$

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. (4.10) regularizes the non-physical breakdown, and explains creation of shock. The regularized second-order PDE is called Burger's equation.

4.2 Classification of linear second-order PDEs:

Consider the most general linear second-order PDE over two independent variables:

$$a_{11}\partial_x^2 u + 2a_{12}\partial_x\partial_y u + a_{22}\partial_y^2 u + b_1\partial_x u + b_2\partial_y u + cu + f = 0, \tag{4.12}$$

where all the coefficients may depend on the two independent variables x and y.

The methods of characteristics, (which applies to first-order PDEs, for example, when $a_{11} = a_{12} = a_{21} = c = 0$ in Eq. (4.12), can inform the analysis of second-order PDEs. Therefore, let us momentarity return to the first-order PDE,

$$b_1\partial_x u + b_2\partial_y u + f = 0, \tag{4.13}$$

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.

$$J = \det\begin{pmatrix} \partial_x\eta & \partial_y\eta \\ \partial_x\xi & \partial_y\xi \end{pmatrix} \neq 0, \infty. \tag{4.14}$$

Substituting $u = w(\eta(x, y), \xi(x, y))$ into the sum of the first derivative terms in Eq. (4.13) one derives

$$\begin{aligned} b_1\partial_x u + b_2\partial_y u &= b_1\left(\partial_x\eta\partial_\eta w + \partial_x\xi\partial_\xi w\right) + b_2\left(\partial_y\eta\partial_\eta w + \partial_y\xi\partial_\xi w\right) \\ &= \left(b_1\partial_x\eta + b_2\partial_y\eta\right)\partial_\eta w + \left(b_1\partial_x\xi + b_2\partial_y\xi\right)\partial_\xi w. \end{aligned} \tag{4.15}$$

Requiring that the second term in Eq. (4.15) 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. (4.12). We derive

$$a_{11}\partial_x^2 u + 2a_{12}\partial_x\partial_y u + a_{22}\partial_y^2 u = \left(A\partial_\xi^2 + 2B\partial_\xi\partial_\eta + C\partial_\eta^2\right) w, \tag{4.16}$$

where

$$\begin{aligned} A &:= a_{11}(\partial_x\xi)^2 + 2a_{12}(\partial_x\xi)(\partial_y\xi) + a_{22}(\partial_y\xi)^2 \\ B &:= a_{11}(\partial_x\xi)(\partial_x\eta) + a_{12}(\partial_x\xi\partial_y\eta + \partial_y\xi\partial_x\eta) + a_{22}(\partial_y\xi)(\partial_y\eta) \\ C &:= a_{11}(\partial_x\eta)^2 + 2a_{12}(\partial_x\eta)(\partial_y\eta) + a_{22}(\partial_y\eta)^2. \end{aligned}$$

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. (4.16) to zero, i.e. $A = C = 0$. This is achieved if we require that $\xi(y_+(x))$ and $\eta(y_-(x))$, where

$$\frac{dy_\pm}{dx} = \frac{a_{12} \pm \sqrt{D}}{a_{11}}, \quad \text{where} \quad D := a_{12}^2 - a_{11}a_{22}. \tag{4.17}$$

and D is called the *discriminant*. Eqs. (4.17) 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. (4.12) is called a *hyperbolic* PDE. In this case, the characteristics are real, and any real pair (ξ, η) is mapped to the real pair (η, ν). Eq. (4.12) gets the following canonical form

$$\partial_\xi\partial_\eta u + \tilde{b}_1\partial_\xi u + \tilde{b}_2\partial_\eta u + \tilde{c}u + \tilde{f} = 0. \tag{4.18}$$

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. (4.18) becomes

$$\partial_\alpha^2 u - \partial_\beta^2 u + \tilde{b}_1^{(2)}\partial_\alpha u + \tilde{b}_2^{(2)}\partial_\beta u + \tilde{c}^{(2)}u + \tilde{f}^{(2)} = 0. \tag{4.19}$$

If $D < 0$ Eq. (4.12) is called an *elliptic* PDE. In this case, Eqs. (4.18) 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:

$$\partial_\alpha^2 u + \partial_\beta^2 u + b_1^{(e)}\partial_\alpha u + b_2^{(e)}u_\beta + c^{(e)}u + f^{(e)} = 0. \tag{4.20}$$

$D = 0$ 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. (4.12) gets the following canonical parabolic form

$$\partial_\alpha^2 u + b_1^{(p)}\partial_\alpha u + b_2^{(p)}\partial_\beta u + c^{(p)}u + f^{(p)} = 0. \tag{4.21}$$

Exercise 4.2.1. 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) $\partial_x^2 u - 2\partial_x\partial_y u + \partial_y^2 u + \partial_x u + \partial_y u - u = 0$.

4.3 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. (4.20) to

$$\sum_{i,j=1}^{d} a_{ij}\partial_{x_i}\partial_{x_j}u(\boldsymbol{x}) + \text{lower order terms} = 0, \tag{4.22}$$

where it is assumed 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. (4.22) cannot be reduced to a canonical form (introduced, in the previous section, in the $d = 2$ case).

Our prime focus here will be on the $d \geq 2$ cases where, a_{ij}, in Eq. (4.22) is $\sim \delta_{ij}$, and also on solving inhomogeneous equations, where a nontrivial (nonzero) solution is driven by a an actual nonzero source. It is natural to approach solving these equations with the Green function method.

We have discussed in Section 3.5.1 how to solve static linear one dimensional case of the Poisson equation using Green functions. Here we generalize and consider Poisson equation in the space of higher dimension, and specifically in $d = 2$ and $d = 3$.

$$\boldsymbol{\nabla}_{\boldsymbol{r}}^2 f = \phi(\boldsymbol{r}), \tag{4.23}$$

where $\boldsymbol{\nabla}_{\boldsymbol{r}}^2 = \Delta_{\boldsymbol{r}}$ is the Laplacian operator, which is $\partial_x^2 + \partial_y^2$ in $d = 2$, $\boldsymbol{r} = (x,y) \in \mathbb{R}^2$, and $\partial_x^2 + \partial_y^2 + \partial_z^2$ in $d = 3$, $\boldsymbol{r} = (x,y,z) \in \mathbb{R}^3$.

The Poisson Eq. (4.23) has many applications, and in particular it describes the electrostatic potential of the charge distributed in $\boldsymbol{r}$ with the density $\rho(\boldsymbol{r})$, in which case, $\phi(\boldsymbol{r}) = -4\pi\rho(\boldsymbol{r})$. Note that the homogeneous case of $\rho = 0$ is still 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

$$\left(\boldsymbol{\nabla}_{\boldsymbol{r}}^2 - \kappa^2\right) f = \phi(\boldsymbol{r}), \tag{4.24}$$

which describes distribution of charge $\rho(\boldsymbol{\rho})$ in plasma for $\phi(r) = -4\pi\rho(\boldsymbol{r})$.

Functions which satisfy the homogeneous Laplace equation are called harmonic. Notice that there exists no nonzero harmonic function defined in the entire, $\mathbb{R}^2$, and approaching 0 as $|\boldsymbol{r}| \to \infty$. Indeed, applying Fourier transform to the homogeneous Laplace equation, one derives $q^2\hat{f}(\boldsymbol{q}) = 0$, which results in $\hat{f}(\boldsymbol{q}) \sim \delta(\boldsymbol{q})$, and then (applying Inverse Fourier transform), $f(\boldsymbol{r}) = \text{const}$. Finally, requiring that $f \to 0$ at $r \to \infty$ one observes that the constant is zero.

Let us stress that these arguments extends to any dimension, and also applies to the Debye equation: there exists no solution to the Debye equation defined in the entire space and decaying to zero at $r \to \infty$.

Therefore, nonzero harmonic function should be defined in a bounded domain, where the homogeneous Laplace equation should be supplemented with some kind of boundary conditions. For example, one can fix $f(x)$ at the boundary.

To solve the Laplace problem let us define the Green function, which is a solution to the inhomogenous equation with a point source on the right hand side,

$$\boldsymbol{\nabla}_{\boldsymbol{r}}^2 G = \delta(\boldsymbol{r}). \tag{4.25}$$

Then the solution to Eq. (4.23) becomes

$$f(\boldsymbol{r}) = \int d\boldsymbol{r}' G(\boldsymbol{r} - \boldsymbol{r}')\phi(\boldsymbol{r}'). \tag{4.26}$$

The solution to Eq. (4.25) can be found by applying the Fourier transform, resulting in the following algebraic equation, $q^2\hat{G}(\boldsymbol{q}) = -1$. Resolving it (trivially) and applying the

Inverse Fourier transform one derives

$$\begin{aligned} G(\boldsymbol{r}) &= -\int \frac{d^3\boldsymbol{q}}{(2\pi)^3}\frac{\exp(i(\boldsymbol{q}\boldsymbol{r}))}{q^2} \\ &= -\int \frac{d^2q_\perp}{(2\pi)^3}\int\limits_{-\infty}^{\infty} dq_\parallel \frac{\exp(iq_\parallel r)}{q_\parallel^2+q_\perp^2} \\ &= -\int \frac{d^2q_\perp}{(2\pi)^3}\frac{\pi}{q_\perp}\exp(-q_\perp r) \\ &= -\frac{1}{4\pi r}. \end{aligned} \tag{4.27}$$

Substituting Eq. (4.27) into Eq. (4.26) one derives

$$f(\boldsymbol{r}) = -\int d^3\boldsymbol{r}'\frac{\phi(\boldsymbol{r})}{4\pi|\boldsymbol{r}-\boldsymbol{r}'|} = \int d^3\boldsymbol{r}'\frac{\rho(\boldsymbol{r})}{|\boldsymbol{r}-\boldsymbol{r}'|}, \tag{4.28}$$

which is thus expression for the electrostatic potential for a given distribution of the charge density in the space.

Example 4.3.1. 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

$$\boldsymbol{\nabla}^2_{\boldsymbol{r}} G(\boldsymbol{r};\boldsymbol{r}') = \delta(\boldsymbol{r}-\boldsymbol{r}'), \tag{4.29}$$

for $\boldsymbol{r}$ such that $R \le r', r$, under condition that $G(\boldsymbol{r};\boldsymbol{r}') = 0,\ r = R \le r'$.

The Green function in this case is inhomogeneous, i.e. $G(\boldsymbol{r};\boldsymbol{r}') \neq G(\boldsymbol{r}-\boldsymbol{r}')$. It is direct to check that solution is given by

$$G = -\frac{1}{4\pi|\boldsymbol{r}-\boldsymbol{r}'|} + \frac{R}{4\pi r'|\boldsymbol{r}-\boldsymbol{r}''|},$$

where $\boldsymbol{r}'' := \boldsymbol{r}'R^2/(r')^2$. It is also clear that the solution is equivalent to placing two point sources (chargers), one at $\boldsymbol{r}'$ and another one at the image point, $\boldsymbol{r}''$, as if there is no zero boundary condition fixed at the surface, i.e. choosing $\rho(\boldsymbol{r}) = \delta(\boldsymbol{r}-\boldsymbol{r}') + \delta(\boldsymbol{r}-\boldsymbol{r}'')$ in Eq. (4.28). (This method of solution to the Laplace equation with zero condition at a surface is called, naturally, method of (mirror) images.)

Let us now turn to the Debye equation and find its Green function, defined by

$$\left(\boldsymbol{\nabla}^2_{\boldsymbol{r}} - \kappa^2\right) G = \delta(\boldsymbol{r}). \tag{4.30}$$

To solve Eq. (4.30) we act in the same way as above in the case of the Laplace equation. Equation for the Fourier transform of the Green function is, $(q^2+\kappa^2)\hat{G}(\boldsymbol{q}) = -1$, resolving

the algebraic equation and applying the Inverse Fourier transform to the result one finds

$$\begin{aligned} G(\boldsymbol{r}) &= -\int \frac{d^3\boldsymbol{q}}{(2\pi)^3}\frac{\exp(i(\boldsymbol{q}\boldsymbol{r}))}{q^2+\kappa^2} \\ &= -\int \frac{d^2q_\perp}{(2\pi)^3}\int\limits_{-\infty}^{\infty} dq_\parallel \frac{\exp(iq_\parallel r)}{q_\parallel^2+q_\perp^2+\kappa^2} \\ &= -\int \frac{d^2q_\perp}{(2\pi)^3}\frac{\pi}{\sqrt{q_\perp^2+\kappa^2}}\exp(-\sqrt{q_\perp^2+\kappa^2}r) \\ &= -\frac{\exp(-\kappa r)}{4\pi r}. \end{aligned} \tag{4.31}$$

Exercise 4.3.2. 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. $\rho(r)$.

4.4 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 $\boldsymbol{k}$

$$u(t;\boldsymbol{r}) = \int \frac{d\boldsymbol{k}}{(2\pi)^k}\exp\left(i(\boldsymbol{k}\cdot\boldsymbol{r})\right)\psi_{\boldsymbol{k}}(t)\hat{u}(\boldsymbol{k}), \quad \psi_{\boldsymbol{k}}(t)\equiv\exp\left(-i\omega(\boldsymbol{k})t\right), \tag{4.32}$$

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. (4.32) $\psi_{\boldsymbol{k}}(t)$ is a solution to the following first-order (in time) linear ODE

$$\left(\frac{d}{dt}+i\omega(\boldsymbol{k})\right)\psi_{\boldsymbol{k}}=0, \tag{4.33}$$

or alternatively of the following second-order linear ODE

$$\left(\frac{d^2}{dt^2}+\left(\omega(\boldsymbol{k})\right)^2\right)\psi_{\boldsymbol{k}}=0. \tag{4.34}$$

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 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. (4.34) becomes

$$\left(\frac{d^2}{dt^2} - c^2\boldsymbol{\nabla}_{\boldsymbol{r}}^2\right)\psi(t;\boldsymbol{r}) = 0. \tag{4.35}$$

Note that the two differential operators in Eq. (4.35), one in time and another in space, have opposite signs. Therefore, we naturally arrive at the case which generalizes the hyperbolic PDE (4.19). It is a generalization because $\boldsymbol{r}$ is not one-dimensional but d-dimensional, $d \geq 1$.

Eq. (4.23) 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. 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.

Another example is given by the electro-magnetic waves, described by the Maxwell equations on the electric, $\boldsymbol{E}$, and magnetic, $\boldsymbol{B}$, fields,

$$\partial_t\boldsymbol{E} = c\boldsymbol{\nabla}_{\boldsymbol{r}} \times \boldsymbol{B}, \quad \partial_t\boldsymbol{B} = -c\boldsymbol{\nabla}_{\boldsymbol{r}} \times \boldsymbol{E}, \tag{4.36}$$

supplemented by the divergence-free conditions,

$$(\boldsymbol{\nabla}_{\boldsymbol{r}} \cdot E) = (\boldsymbol{\nabla}_{\boldsymbol{r}} \cdot B) = 0, \tag{4.37}$$

where $\times$ is the vector product in $d = 3$[1], and c is the speed of light in the media. Differentiating first equation in the pair of Eqs. (4.36) 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

[1] $(\boldsymbol{\nabla}_{\boldsymbol{r}} \times \boldsymbol{B})_i = \varepsilon_{ijk}\nabla_j B_k$, where $i, j, = 1, \cdot 3$ and ε_{ijk} is the absolutely skew-symmetric tensor in $d = 3$

account that for the divergence-free, $\boldsymbol{E}$, $(\boldsymbol{\nabla}_{\boldsymbol{r}} \times (\boldsymbol{\nabla}_{\boldsymbol{r}} \times \boldsymbol{E})) = \boldsymbol{\nabla}^2_{\boldsymbol{r}} \boldsymbol{E}$, one arrives at Eq. (4.35) for all components of the electric field, i.e. with ψ replaced by $\boldsymbol{E}$.

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

$$\omega(\boldsymbol{k}) = \sqrt{gk + (\sigma/\rho)k^3}, \tag{4.38}$$

where g, σ and ρ are gravity coefficient, surface tension coefficient and density of the fluid, respectively. Eq. (4.38) is so complex because it accounts for both capillary and gravitational effects. Gravitational waves dominate at small q (large distances), where Eq. (4.38) 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. (4.33) or Eq. (4.34) 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. (4.38), 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 4.4.1. Solve the Cauchy (initial value) problem for amplitude of spin-waves which satisfy the following PDE

$$\partial_t^2 \psi = -(\Omega - b\boldsymbol{\nabla}^2_{\boldsymbol{r}})^2 \psi, \tag{4.39}$$

in $d = 3$, where $\psi(t = 0; \boldsymbol{r}) = \exp(-r^2)$ and $d\psi/dt(t = 0; \boldsymbol{r}) = 0$.

Note, first, that applying the Fourier transform over $\boldsymbol{r}$ to Eq. (4.39) one arrives at Eq. (4.33), where

$$\omega(k) = \Omega + bk^2, \tag{4.40}$$

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. (4.33,4.40) 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

$$\begin{aligned}
\psi(t;\boldsymbol{r}) &= \pi^{3/2}\int \frac{d^3k}{(2\pi)^3}e^{-q^2/4}\cos((\Omega+bk^2)t)\exp(i(\boldsymbol{k}\cdot\boldsymbol{r})) \\
&= \int_0^\infty \frac{kdk}{2\pi^{1/2}r}e^{-k^2/4}\cos((\Omega+bk^2)t)\sin(kr) \\
&= -\int_0^\infty \frac{dk}{2\pi^{1/2}r}e^{-k^2/4}\cos((\Omega+bk^2)t)\frac{d}{dr}\cos(kr) \\
&= -\mathrm{Re}\left(\frac{\exp(i\Omega t)}{4\pi^{1/2}r}\frac{d}{dr}\int_{-\infty}^\infty dk\exp\left(-\frac{1-4ibt}{4}k^2+ikr\right)\right) \\
&= \mathrm{Re}\left(\frac{\exp\left(i\Omega t-\frac{r^2}{1-4ibt}\right)}{(1-4ibt)^{3/2}}\right).
\end{aligned}$$

Exercise 4.4.2. Solve the Cauchy (initial value) problem for the wave Eq. (4.35) 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. (4.34) that arises from an source term $\chi(t;\boldsymbol{r})$ on the right hand side:

$$\left(\frac{d^2}{dt^2}+(\omega(-i\boldsymbol{\nabla}_{\boldsymbol{r}}))^2\right)\psi(t;\boldsymbol{r})=\chi(t;\boldsymbol{r}). \tag{4.41}$$

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

$$\left(\frac{d^2}{dt^2}+(\omega(-i\boldsymbol{\nabla}_{\boldsymbol{r}}))^2\right)G(t;\boldsymbol{r})=\delta(t)\delta(\boldsymbol{r}). \tag{4.42}$$

The solution to the inhomogeneous PDE, Eq. (4.41), can be expressed as the convolution of the source term $\chi(t_1;\boldsymbol{r}_1)$ with the Green function, $G(t;\boldsymbol{r})$

$$\psi(t;\boldsymbol{r})=\int dt_1 d\boldsymbol{r}_1 G(t-t_1;\boldsymbol{r}-\boldsymbol{r}_1)\chi(t_1;\boldsymbol{r}_1), \tag{4.43}$$

The solution to Eq. (4.41) is expressed as sum of the forced solution (4.43) and a zero mode of the respective free equation, i.e. Eq. (4.41) with zero right hand side.

To solve Eq. (4.42) for the Green function, or equivalently equation for its Fourier transform

$$\left(\frac{d^2}{dt^2} + (\omega(\boldsymbol{k}))^2\right)\hat{G}(t;\boldsymbol{k}) = \delta(t). \tag{4.44}$$

Recall that the inhomogeneous ODE. (4.44) was already discussed earlier in the course. Indeed Eq. (3.48) solves Eq. (4.44). Then recalling that ω depends on $\boldsymbol{k}$ and applying the inverse Fourier transform over $\boldsymbol{k}$ to Eq. (3.48) one arrives at

$$G(t;\boldsymbol{r}) = \theta(t)\int \frac{d^3k}{(2\pi)^3}\frac{\sin(\omega(k)t)}{\omega(k)}\exp\left(i(\boldsymbol{kr})\right). \tag{4.45}$$

Exercise 4.4.3 (not graded)**.** Show that the general expression (4.45) in the case of the linear dispersion law, $\omega(\boldsymbol{k}) = ck$, becomes

$$G(t;\boldsymbol{r}) = \frac{\theta(t)}{4\pi cr}\delta(r - ct), \tag{4.46}$$

where $r = |\boldsymbol{r}|$.

Substituting Eq. (4.46) into Eq. (4.43) one derives the following expression for linear dispersion (light or sound) radiation from a source

$$\psi(t;\boldsymbol{r}) = \frac{1}{4\pi c^2}\int \frac{d\boldsymbol{r}_1}{R}\chi(t - R/c;\boldsymbol{r}_1). \tag{4.47}$$

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.

Exercise 4.4.4 (not graded)**.** Solve the radiation Eqn. (4.41) in the case of the linear dispersion law for the case of a point harmonic source, $\chi(t;\boldsymbol{r}) = \cos(\omega t)\delta(\boldsymbol{r})$.

4.5 Diffusion Equation

The most common example of a multi-dimensional generalization of the parabolic equation Eq. (4.21) is the homogeneous diffusion equation

$$\partial_t u = \kappa\nabla^2_{\boldsymbol{r}} u, \tag{4.48}$$

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

$$\hat{u}(t;\boldsymbol{q})=\int dy_1\dots dy_d\exp\left(i\boldsymbol{q}\boldsymbol{x}\right)u(t;\boldsymbol{y}). \tag{4.49}$$

Integrating Eq. (4.48) with the Fourier weight one arrives at

$$\partial_t\hat{u}(t;\boldsymbol{q})=-q^2\hat{u}(t;\boldsymbol{q}) \tag{4.50}$$

Integrating the equation over time, $\hat{u}(t;\boldsymbol{q})=\exp(-q^2t)\hat{u}(0;\boldsymbol{k})$, and evaluating the inverse Fourier transform over $\boldsymbol{q}$ of the result one arrives at

$$u(t;\boldsymbol{x})=\int\frac{dy_1,\dots dy_d}{(4\pi t)^{d/2}}\exp\left(-\frac{(\boldsymbol{x}-\boldsymbol{y})^2}{4t}\right)u(0;\boldsymbol{y}). \tag{4.51}$$

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. (4.51) 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. (4.51), i.e.

$$u(t;\boldsymbol{x})\approx\frac{A}{(4\pi t)^{d/2}}\exp\left(-\frac{x^2}{4t}\right),\quad A=\int u(0;\boldsymbol{y})dy_1\dots dy_d. \tag{4.52}$$

Notice that the approximation (4.52) corresponds to the substitution of $u(0,\boldsymbol{y})\to A\delta(\boldsymbol{y})$ in Eq. (4.51). Another interpretation of Eq. (4.52) 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. (4.52) becomes

$$u(t;\boldsymbol{x})\approx\frac{(\boldsymbol{B}\cdot\boldsymbol{x})}{(4\pi t)^{d/2+1}}\exp\left(-\frac{x^2}{4t}\right),\quad B=2\pi\int\boldsymbol{y}u(0;\boldsymbol{y})dy_1\dots dy_d. \tag{4.53}$$

Exercise 4.5.1. Find asymptotic behavior of a one-dimensional diffusion equation at sufficiently long times for the following initial conditions

(a) $u(0;x)=x\exp\left(-\frac{x^2}{2l^2};\right)$

(b) $u(0;x)=\exp\left(-\frac{|x|}{l}\right)$

(c) $u(0;x)=x\exp\left(-\frac{|x|}{l}\right)$

(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.(4.51) in a series over y?

Our next step is to find the Green function of the heat equation, i.e. to solve

$$\partial_t G - \kappa \nabla^2_{\boldsymbol{r}} G = \delta(t)\delta(\boldsymbol{x}), \tag{4.54}$$

In fact, we have solved this problem already as Eq. (4.51) describes it with $u(0;\boldsymbol{y}) = G(+0;\boldsymbol{x}) = \delta(\boldsymbol{x})$ set as the initial condition. The result is

$$G(t;\boldsymbol{x}) = \frac{1}{(4\pi t)^{d/2}} \exp\left(-\frac{x^2}{4t}\right). \tag{4.55}$$

As always, the Green function can be used to solve the inhomogeneous diffusion equation

$$\partial_t u - \kappa \nabla^2_{\boldsymbol{x}} u = \phi(t;\boldsymbol{x}) \tag{4.56}$$

which solution is expressed via the Green function as follows

$$u(t;\boldsymbol{x}) = \int\limits_{-\infty}^{t} dt' \int d\boldsymbol{y} G(t';\boldsymbol{y})\phi(t-t';\boldsymbol{x}-\boldsymbol{y}), \tag{4.57}$$

where we assume that $u(\infty;\boldsymbol{x}) = 0$.

Exercise 4.5.2 (not graded)**.** Solve Eq. (4.56) for $\phi(t;\boldsymbol{x}) = \theta(t)\exp\left(-x^2/(2l^2)\right)$ in the $d = 4$-dimensional space.

4.6 Boundary Value Problems: Fourier Method

Consider the boundary value problem associated with sound waves:

$$\partial_t^2 u(t;x) - c^2 \partial_x^2 u(t;x) = 0, \tag{4.58}$$

$$0 \le x \le L, \quad u(t,0) = u(t,L) = 0, \quad u(0,x) = \varphi(x), \quad \partial_t u(0,x) = \psi(x). \tag{4.59}$$

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. (4.58) one arrives at

$$\frac{X''(x)}{X(x)} = \frac{T''(t)}{T(t)} = -\lambda, \tag{4.60}$$

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. (4.59). 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. (4.60) 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 (4.59):

$$\varphi(x) = \sum_{n=1}^{\infty} A_n X_n(x), \quad \psi(x) = \sum_{n=1}^{\infty} \lambda_n B_n X_n(x). \tag{4.61}$$

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. (4.61) on $X_m(x)$, integrating them from 0 to L, and accounting for the ortho-normality of the eigen-functions, one derives

$$A_m = \frac{2}{L}\int_0^{\vdots} dx\varphi(x)X_m(x), \quad B_m = \frac{2}{\lambda_m L}\int_0^{\vdots} dx\psi(x)X_m(x). \tag{4.62}$$

Exercise 4.6.1. The equation describing the 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 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 $x \in [0, L]$:

$$\partial_t u = a^2\partial_x^2 u, \quad u(t,0) = u(t,L) = 0, \quad u(0,x) = \begin{cases} x, & x < L/2 \\ L - x, & x > L/2. \end{cases} \tag{4.63}$$

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 4.6.2. Solve the following boundary value problem

$$\partial_t u = a^2\partial_x^2 u - \beta u, \quad u(t,0) = u(t,L) = 0, \quad u(0,x) = \sin\left(\frac{2\pi x}{L}\right).$$

4.7 Exemplary Nonlinear PDE: Burger's Equation

Burgers equation is a generalization of the Hopf equation, Eq. (4.10), discussed when illustrating the method of characteristics. Recall that the Hopf equation results in a wave breaking which leads to a non-physical multi-valued solution. Modification of the Hopf equation by adding dissipation/diffusion results in Burger's equation:

$$\partial_t u + u\partial_x u = \partial_x^2 u. \tag{4.64}$$

Like practically every other nonlinear PDE, Burger's equation seems rather hopeless to resolve at first glance. However, Burger's equation is in fact special. It allows the Cole-Hopf transformation, from $u(t;x)$ to $\Psi(t;x)$

$$u(t;x) = -2\frac{\partial_x \Psi(t;x)}{\Psi(t;x)}, \tag{4.65}$$

reducing Burger's equation to the diffusion equation

$$\partial_t \Psi = \partial_x^2 \Psi. \tag{4.66}$$

The solution to the Cauchy problem associated with Eq. (4.66) can be expressed as an integral convolving the initial profile $\Psi(0;x)$, with the Green function of the diffusion equation described in Eq. (4.55)

$$\Psi(t;x) = \int \frac{dy}{\sqrt{4\pi t}} \exp\left(-\frac{(x-y)^2}{4t}\right) \Psi(0;y). \tag{4.67}$$

This latter expression can be used to find *some* exact solutions to Burger's equation. Consider, for example, $\Psi(0;x) = \cosh(ax)$. Substitution into Eq. (4.67) and conducting integration over y, one arrives at $\Psi(t;x) = \cosh(ax)\exp(a^2 t)$, which results, according to Eq. (4.65), in stationary (time independent, i.e. standing) "shock" solution to Burger's equation, $u(t;x) = -2a\tanh(ax)$. Notice that the following more general solution to Burger's equation corresponds to a shock moving with the constant speed u_0

$$u(t;x) = u_0 - 2a\tanh(a(x - x_0 - u_0 t)).$$

Exercise 4.7.1 (not graded)**.** Solve the diffusion equation Eq. (4.66) with the initial conditions $\Psi(0,x) = \cosh(ax) + B\cosh(bx)$. Reconstruct respective $u(t;x)$ solving the Burgers Eq. (4.64). 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.

Part III

Optimization

Chapter 5

Calculus of Variations

The main theme of this section is the relation of equations to minimal principles. Oversimplifying 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 $S'(q_*) = Kq_* - g = 0$.

q in the example above 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 $I\{q(x)\}$ 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 $u(x)$ is called the variational derivative, and then by analogy with the finite dimensional example above, one finds that the Euler-Lagrange 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.

5.1 Examples

To have a better understanding of the calculus of variations we start describing four examples.

5.1.1 Fastest Path

Consider a robot navigating within the (x, y)-plane. We can describe the robot's path as $y = q(x)$. Assume that the plane constitutes a rugged terrain, so that robot's velocity (absolute value) when it passes a point on the plane is characterized by a scalar positive function, $g(x, y)$. Then the time it takes for the robot to move from x to $x + dx$ along the path $q(x)$, where dx is small, is

$$L(x, q(x), q'(x)) = g(x, q(x))\sqrt{1 + (q'(x))^2}dx.$$

The total length along the path which starts at $(x, y) = (0, 0)$ and ends at $(x, y) = (a, b)$, where $a > 0$ is

$$S\{q(x)\} = \int_0^a dx L(x, q(x), q'(x)).$$

We would like to find the path, $q(x)$, which minimizes the functional $S\{q(x)\}$, subject to $q(0) = 0$ and $q(a) = b$.

5.1.2 Minimal Surface

Consider making a three-dimensional bubble by dipping a wire loop into soapy water, and then asking the question what is the optimal shape of the bubble for the given loop. Physics suggests that the shape of the bubble minimizes area of the soap film.

We formalize this setting as follows. The surface of a bubble is described by the function, $q : \mathcal{D} \to \mathbb{R}$, where $\mathcal{D} \in \mathbb{R}^2$ is bounded (∞ is not contained in the set), and is the projection of the bubble on the $\mathbb{R}^2$ plane, and $u(\partial\mathcal{D}) = g(\partial\mathcal{D})$, where $\partial\mathcal{D}$ is the boundary of $\mathcal{D}$ (closed line in the $\mathbb{R}^2$ plane), and $g(\partial\mathcal{D})$ describes the coordinate of the wire loop along the third dimension. Then the optimal bubble results from minimizing the functional

$$S\{q(x)\} = \int_{\mathcal{D}} dx \sqrt{1 + |\nabla_x q(x)|^2},$$

over $q(x)$, subject to $q(\partial\mathcal{D}) = g(\partial\mathcal{D})$.

5.1.3 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, $[0, 1]^2 \in \mathbb{R}^2$, into a real number between white, 0, and black, 1. However, often only an image corrupted by a noise is observed. 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 signal, $f(x)$, by minimization of the following functional

$$S\{q(x)\} = \int\limits_{U=[0,1]^2} dx \left((q(x) - f(x))^2 + \lambda |\nabla_x q(x)| \right), \tag{5.1}$$

subject to the Neumann boundary condition, $n \cdot \nabla_x q = 0,\ x \in \delta U$, where n is the (unit) vector normal to δU (boundary of the domain U).

5.1.4 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 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,

$$S\{q\} \doteq \int_{t_1}^{t_2} dt L\left(t, q(t), \dot{q}(t)\right), \tag{5.2}$$

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

$$L\left(t, q(t), \dot{q}(t)\right) = \frac{\dot{q}^2}{2} - V(q). \tag{5.3}$$

5.2 Euler-Lagrange Equations

All the examples can be stated as the minimization of the functional

$$S\{q(x)\} = \int_{\mathcal{D} \in \mathbb{R}^n} dx L\left(x, q(x), \nabla_x q(x)\right),$$

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

$$L : \mathcal{D} \in \mathbb{R}^n \times \mathbb{R}^d \times \mathbb{R}^{d \times n} \to \mathbb{R},$$

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: $x \in \mathcal{D} \in \mathbb{R}^n$, $q \in \mathbb{R}^d$, and $p \in \mathbb{R}^{d \times n}$, and thus the

dimensionalities of L_x, L_q, and $L_{\nabla q}$ are n, d and $n \times d$ respectively.) We will assume in the following that both L and g are smooth.

Theorem 5.2.1 (Euler-Lagrange theorem (necessary condition for optimality))**.** Suppose that $u(x)$ is the minimizer of S, that is

$$\forall\ v(x) \in C^2(\bar{\mathcal{D}} = \mathcal{D} \cup \partial\mathcal{D}),\ \text{with } v(x) = g(x) \text{ on } \mathcal{D}: \quad S\{v\} \geq S\{u\},$$

then L satisfies

$$\nabla_x \left(L_{\nabla q}\left(x, q(x), \nabla_x q(x)\right)\right) - L_q\left(x, q(x), \nabla_x q(x)\right) = 0 \quad \text{in} \quad \mathcal{D}. \tag{5.4}$$

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. $\delta(x) = 0$ in $\mathcal{D}$. Then according to the assumption

$$S\{q\} \leq S\{\tilde{q}\} = S\{q + s\delta(x)\} \quad \forall s \in \mathbb{R}.$$

This means that

$$\frac{d}{ds} S\{q + s\delta(x)\}\bigg|_{s=0} = 0.$$

Notice that

$$S\{q + s\delta(x)\} = \int\limits_{\mathcal{D}} dx L\left(x, q(x) + s\delta(x), \nabla_x q(x) + s\nabla_x \delta(x)\right)$$

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

$$\begin{aligned}
\frac{d}{ds} S\{q + s\delta(x)\}\bigg|_{s=0} &= \int\limits_{\mathcal{D}} dx\ \frac{d}{ds} L\left(x, q(x) + s\delta(x), \nabla_x q(x) + s\nabla_x \delta(x)\right)\bigg|_{s=0} \qquad (5.5)\\
&= \int\limits_{\mathcal{D}} dx \left(L_q\left(x, q(x), \nabla_x q(x)\right) \cdot \delta(x) + L_p\left(x, q(x), \nabla_x q(x)\right) \cdot \nabla_x \delta(x)\right)\\
&= \int\limits_{\mathcal{D}} dx L_q\left(x, q(x), \nabla_x q(x)\right) \cdot \delta(x) + \int\limits_{\mathcal{D}} dx L_p\left(x, q(x), \nabla_x q(x)\right) \cdot \nabla_x \delta(x)\\
&= \int\limits_{\mathcal{D}} dx \left(L_q\left(x, q(x), \nabla_x q(x)\right) - \nabla_x \cdot L_{\nabla q}\left(x, q(x), \nabla_x q(x)\right)\right) \cdot \delta(x).
\end{aligned}$$

Since the resulting integral should be equal to zero for any $\delta(x)$ one arrives at the desired statement.

□

Exercise 5.2.1. Find the Euler-Lagrange equations (conditions) for

$$(a) \qquad \int dx \left((q'(x))^2 + \exp(q(x)) \right),$$

$$(b) \qquad \int dx\, q(x) q'(x),$$

$$(c) \qquad \int dx\, x^2 (q'(x))^2,$$

where $q : \mathbb{R} \to \mathbb{R}$.

Example 5.2.2. Consider the shortest path version of the fastest path problem set in Section 5.1.1, 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 5.2.3. (a) Write the Euler-Lagrange equation for the general case of the fastest path problem formulated in Section 5.1.1. (b) Find an example of metric, $g(x, y)$, resulting in the quadratic optimal path, i.e. $q(x) = \frac{b}{a^2} x^2$.

Example 5.2.4. Let us derive the Euler-Lagrange condition for the Minimal Surface problem introduced in Section 5.1.2:

$$\min_{\{u(x)\}} \int_{\mathcal{D}} dx \sqrt{1 + |\nabla_x q(x)|^2} \Bigg|_{q(\partial\mathcal{D})=g(\partial\mathcal{D})}.$$

In this case Eq. (5.4) becomes

$$\begin{aligned} 0 &= \nabla_x \left(L_{\nabla q} \left(x, q(x), \nabla_x q(x) \right) \right) - L_q \left(x, q(x), \nabla_x q(x) \right) \\ &= \nabla_x \cdot \left(\frac{\nabla_x q(x)}{\sqrt{1 + |\nabla_x q(x)|^2}} \right) \\ &\rightarrow -\nabla_x q(x) \cdot \nabla_x^2 q \nabla_x q + (1 + |\nabla_x q(x)|^2) \nabla_x^2 q = 0. \end{aligned} \tag{5.6}$$

Exercise 5.2.5. Show that, $q(x) = a \cdot x + b$, where a is a real n-dimensional vector and b is a real scalar solves a Minimal Surface Euler-Lagrange Eq. (5.6) on $\mathcal{D} = (-\pi/2, \pi/2)^2$. (Hint: Do not worry about the boundary conditions. We do not ask about them in the exercise.)

5.3 Phase-Space Intuition and Relation to Optimization (finite dimensional, not functional)

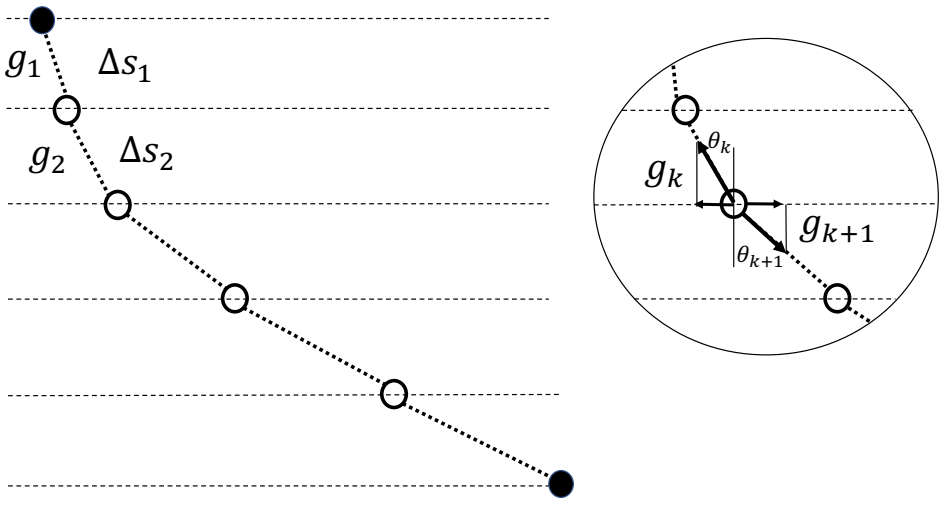


Figure 5.1: Variational Calculus via Discretization and Optimization.

Consider the special case of the fastest path problem of Section 5.1.1, which is still more general than the shortest path problem discussed in the Example 5.2.2, where the metric $g(x)$ depends only on x. In this case the action is

$$S\{q(x)\} = \int_0^a dx g(x) \sqrt{1 + (q'(x))^2} = \int_0^a ds g(x),$$

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, $L(x; q(x); q'(x)) = g(x)\sqrt{1 + (q'(x))^2}$, $L_q = 0$, $L_{q'} = g(x)q'/\sqrt{1 + (q')^2}$. Then the Euler-Lagrange equation becomes

$$\frac{d}{dx}\left(\frac{g(x)q'(x)}{\sqrt{1 + (q'(x))^2}}\right) = 0,$$

which results in

$$\frac{g(x)q'(x)}{\sqrt{1 + (q'(x))^2}} = g(x)\sin(\theta) = \text{constant}, \tag{5.7}$$

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. (5.7) 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, $S\{q(x)\}$:

$$\begin{aligned} S\{q(x)\} &\approx S_k(\cdots, q_k, \cdots) = \sum_k g_k \Delta s_k = \sum_k g_k \sqrt{1 + \left(\frac{q(x_k) - q(x_{k-1})}{\Delta}\right)^2}\Delta \\ &= \sum_k g_k \sqrt{1 + \left(\frac{q_k - q_{k-1}}{\Delta}\right)^2}\Delta \end{aligned}$$

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. (5.1). 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. (5.7):

$$\begin{aligned} \forall k: \qquad & \frac{g_{k+1}(q_{k+1} - q_k)}{\sqrt{1 + \left(\frac{q_{k+1} - q_k}{\Delta}\right)^2}} = \frac{g_k(q_k - q_{k-1})}{\sqrt{1 + \left(\frac{q_k - q_{k-1}}{\Delta}\right)^2}} \\ \rightarrow \quad & g_{k+1}\sin\theta_{k+1} = g_k \sin\theta_k. \end{aligned}$$

5.4 Towards Numerical Solutions of the Euler-Lagrange Equations

Here we discuss the image restoration problem set up in Section 5.1.3. 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 5.6.

5.4.1 Smoothing Lagrangian

The TV functional (5.1) is not differentiable at $\nabla_x u(x) = 0$, which creates difficulty for variations. One way to bypass the problem is to smooth the Lagrangian, considering

$$S_\varepsilon\{q\} = \int\limits_{[0,1]^2} dx \left(\frac{(q(x) - f(x))^2}{2} + \lambda \sqrt{\varepsilon^2 + (\nabla_x q(x))^2} \right), \tag{5.8}$$

where ε is small and positive. The Euler-Lagrange equations for the smoothed action (5.8) are

$$\forall x \in [0,1]^2 : \quad q - \lambda \nabla_x \cdot \frac{\nabla_x q}{\sqrt{\varepsilon^2 + (\nabla_x q(x))^2}} = f, \tag{5.9}$$

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. (5.9) for an arbitrary f is not possible. We will discuss ways to solve Eq. (5.9) numerically in the following.

5.4.2 Gradient Descent and Acceleration

We will start this part with a disclaimer. The discussion below of the numerical procedure for solving Eq. (5.9) is not fully comprehensive. We add it here for completeness, delegating details to Math 575, and also aiming to emphasize connections between numerical PDE analysis and forthcoming discussion (largely within 575) of the optimization algorithms.

A standard numerical scheme for solving Eq. (5.9) 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 continuous. Consider the following equation

$$\forall x \in [0,1]^2,\ t > 0 : \quad \partial_t \upsilon + \upsilon - \lambda \nabla_x \cdot \frac{\nabla_x \upsilon}{\sqrt{\varepsilon^2 + (\nabla_x \upsilon(x))^2}} = f, \tag{5.10}$$

for, $\upsilon(t;x)$, representing estimation at the computational time t for $q(x)$ solving Eq. (5.9), 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. (5.10) 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. (5.10) by the (damped) wave equation

$$\forall x \in [0,1]^2,\ t > 0: \quad \partial_t^2 \upsilon + a\partial_t \upsilon + \upsilon - \lambda \nabla_x \cdot \frac{\nabla_x \upsilon}{\sqrt{\varepsilon^2 + (\nabla_x \upsilon(x))^2}} = f, \tag{5.11}$$

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. (5.11) 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 5.6.2 where we discuss an alternative, primal-dual algorithm.

5.5 Variational Principle of Classical Mechanics

Here we apply the variational principle (also called Hamiltonian principle) to the classical mechanics highlighted in Section 5.1.4. See also [9], which logic we follow in this Section.

To streamline notations of this Section (and unless specificed otherwise) we will discuss dynamics of a particle in one dimension, i.e. $\forall t \in \mathbb{R}:\ u(t) \in \mathbb{R}^d$. Generalization of all the formulas discussed to higher dimensions is straightforward.

5.5.1 Noether's Theorem & time-invariance of space-time derivatives of action

In the case of the classical mechanics, introduced in Section 5.1.4, the Euler-Lagrange Eqs. (5.4) are

$$\frac{d}{dt}L_{\dot{q}} = L_q, \tag{5.12}$$

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,

$$\frac{d}{dt}\left(\dot{q} \cdot L_{\dot{q}} - L\right) \;=\; \ddot{q} \cdot L_{\dot{q}} + \dot{q} \cdot \frac{d}{dt}L_{\dot{q}} - L_q \cdot \dot{q} - L_{\dot{q}}\ddot{q} = \dot{q} \cdot \left(\frac{d}{dt}L_{\dot{q}} - L_q\right) = 0,$$

where the last equality is due to Eq. (5.12).

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 Noether's famous 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 a more general version of Noether's theorem we need the following definition.

Definition 5.5.1 (Invariance of 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 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 5.5.2 (Noether's theorem (1915))**.** If the Lagrangian L is invariant under the action of a one-parameter family of transformations, $h_s(u(t))$, then the quantity,

$$I(q(t), \dot{q}(t)) \equiv L_{\dot{q}} \cdot \frac{d}{ds} \left(h_s(q(t))\right)_{s=0}, \tag{5.13}$$

is constant along any solution of the Euler-Lagrange Eq. (5.12). Such a constant quantity is called an integral of motion.

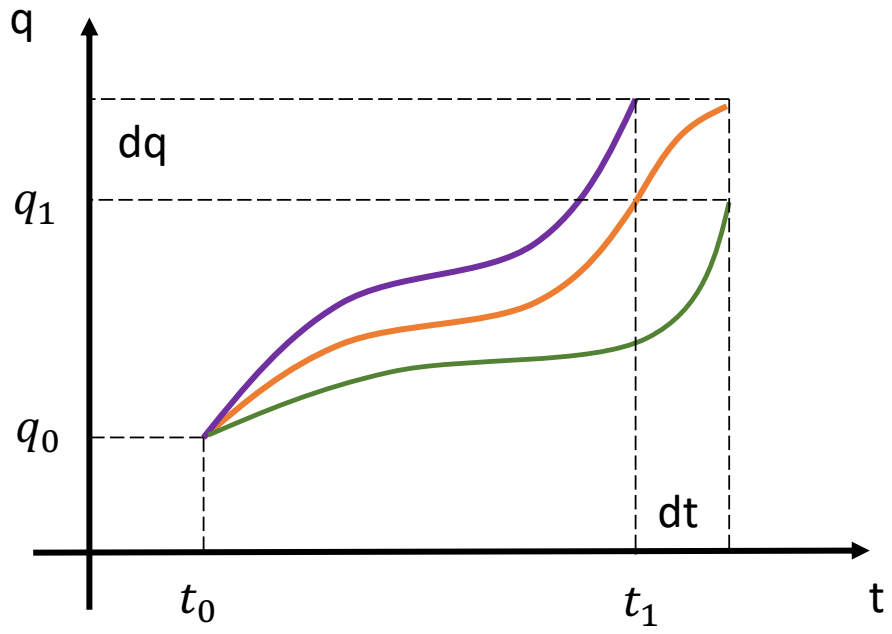


Figure 5.2: End point variation of a critical path.

Proof of Noether's theorem, which will be sketched below, is linked to analysis of the action viewed as a function (not functional!) of the end points of the critical path. Consider the critical/optimal path, $\{q(t)\}$, corresponding to the solution of the Euler-Lagrange Eq. (5.12), substitute it into the action-functional, $S\{q(t)\}$, and then consider the action as a function of the end points, $A_0 \doteq (t_0, q(t_0) = u_0)$ and $A_1 \doteq (t_1, q(t_1) = q_1)$. With a little abuse of notation we express this dependence of the action on A_0 and A_1 as, $S(A_0; A_1) = S(t_0, q_0; t_1, q_1)$. The following statement (sometimes presented as the main theorem of the Hamiltonian mechanics [9]) gives a very intuitive, geometrical interpretation for the derivatives of the action over the end-point parameters

Theorem 5.5.3 (End-point derivatives of the action)**.**

$$(a) \qquad \partial_{t_1} S(A_0; A_1) = (L - \dot{q} L_{\dot{q}})_{t=t_1} = -\partial_{t_0} S(A_0; A_1) = (L - \dot{q} L_{\dot{q}})_{t=t_0}, \tag{5.14}$$

$$(b) \qquad \partial_{q_1} S(A_0; A_1) = L_{\dot{q}}|_{t=t_1} = -\partial_{q_0} S(A_0; A_1) = -L_{\dot{q}}|_{t=t_0}\,. \tag{5.15}$$

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 t_1 and q_1, i.e. the final end point of the critical path.

Let us first keep the final time fixed at t_1 but move the final position by dq, as shown in Fig. (5.2). The trajectory $q(t)$ will vary by $\delta q(t)$, where $\delta q(t_0) = 0$ and $\delta q(t_1) = dq$. Variation of the action is

$$dS = \int_{t_0}^{t_1} dt \left(L_{\dot{q}} \delta \dot{q} + L_q \delta q \right). \tag{5.16}$$

One use the relation, $\delta \dot{q} = d\delta q/dt$, and also the Euler-Lagrange Eqs. (5.12) to rewrite Eq. (5.16)

$$dS = \int_{t_0}^{t_1} dt \left(L_{\dot{q}} \frac{d}{dt} \delta q + \delta q \frac{d}{dt} L_{\dot{q}} \right) = \int_{t_0}^{t_1} dt \left(L_{\dot{q}} \delta q \right) = \left(L_{\dot{q}} \delta q \right)_{t_0}^{t_1} = L_{\dot{q}}|_{t_1}\, dq. \tag{5.17}$$

Therefore, as we kept the final time fixed, $dS = \partial_{q_1} S dq$, and one arrives at the desired statement

$$\frac{\partial S}{\partial q_1} = L_{\dot{q}}|_{t_1}\,. \tag{5.18}$$

We compute variation of the action over the final time similarly. Consider variation of the action extended from $A_1 = (q_1, t_1)$ to $(q_1 + dq, t_1 + dt)$:

$$dS = L dt = \frac{\partial S}{\partial t_1} dt + \frac{\partial S}{\partial q_1} dq = \frac{\partial S}{\partial t_1} dt + L_{\dot{q}}|_{t_1}\, dq = \left(\frac{\partial S}{\partial t_1} + \dot{q} L_{\dot{q}} \right)_{t_1} dt,$$

where we utilize Eq. (5.18). Finally, we derive

$$\frac{\partial S}{\partial t_1} = \left(L - \dot{q} L_{\dot{q}} \right)_{t_1}.$$

□

We are now ready to sketch the proof of the Noether Theorem 5.5.2.

Proof. (of the Noether theorem) By the assumption of the theorem

$$S(t_0, h_s(q_0); t_1, h_s(q_1)) = S(t_0, q_0; t_1, q_1), \quad \forall s.$$

Differentiating both sides of the equality with respect to s at $s = 0$, and using Theorem 5.5.3 results in

$$\begin{aligned} 0 &= \partial_{u_0} S \cdot \frac{d}{ds} \left(h_s(q_0)\right)_{s=0} + \partial_{q_1} S \cdot \frac{d}{ds} \left(h_s(q_1)\right)_{s=0} \\ &= -L_p(q(t_0), \dot{q}(t_0)) \cdot \frac{d}{ds} \left(h_s(q_0)\right)_{s=0} + L_p(q(t_1), \dot{q}(t_1)) \cdot \frac{d}{ds} \left(h_s(q_1)\right)_{s=0} . \end{aligned}$$

Since t_1 can be chosen arbitrarily, it proves that Eq. (5.13) is constant along the solution of the Euler-Lagrange Eq. (5.12). □

Exercise 5.5.1. For $q(t) \in \mathbb{R}^3$ and each of the following families of transformations find the explicit form of the conserved quantity given by Noether's theorem (assuming that 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.

5.5.2 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. (5.3), 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. (5.3), in terms of an auxiliary optimization

$$\int dt \frac{\dot{q}^2}{2} = \max_{\{p(t)\}} \int dt \left(p\dot{q} - \frac{p^2}{2} \right) , \tag{5.19}$$

and substituting the result in Eqs. (5.2,5.3), one arrives at the following, alternative, variational formulation of the classical mechanics

$$\min_{\{q(t)\}} \max_{\{p(t)\}} \int dt \left(p\dot{q} - H(q;p)\right) \tag{5.20}$$

$$H(q;p) \doteq \frac{p^2}{2} + V(q), \tag{5.21}$$

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

$$\dot{q} = \frac{\partial H(q;p)}{\partial p}, \quad \dot{p} = -\frac{\partial H(q;p)}{\partial q}. \tag{5.22}$$

Exercise 5.5.2. (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 (5.22).
(b) [Conservation of Momentum] Show that if the Lagrangian does not depend explicitly on one of the coordinates, say q_1, 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.

The Hamiltonian system of equations becomes even more elegant in vector form

$$\dot{z} = -J\nabla_z H(z) = -\nabla_z J H(z), \quad z \doteq \begin{pmatrix} q \\ p \end{pmatrix}, \quad J \doteq \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \tag{5.23}$$

where the 2×2 matrix represents two-dimensional rotation (clock-wise in the (q,p)-space).

5.5.3 Hamilton-Jacobi equation

Let us work a bit more with the critical/optimal trajectory/path, $\{q(t'); t' \in [0,t]\}$, solving the Euler-Lagrange Eqs. (5.12), given the initial and final conditions for the position of the particle: $q(0)$ and $q(t)$. That is we continue the thread of Section 5.5.3, and specifically Theorem 5.5.2 and consider the action as a function of $A_1 = (t_1, q_1)$ – the final position of the critical path.

Let us re-derive in a bit different, but equivalent, form the main results of the Theorem 5.5.3. Assuming that the action is a sufficiently smooth function of the arguments, t and u, one would like to introduce (and interpret) derivatives of action over t and q, and then check if the derivatives are related to each other. Consider, first, derivative of the action over t:

$$\begin{aligned} \mathcal{S}_t \doteq \partial_t \mathcal{S}(t;q) = \partial_t \int_0^t dt' L\left(q(t'), \dot{q}(t')\right) = L + \int_0^t dt' \left(L_q \partial_t q(t') + L_{\dot{q}} \partial_t \dot{q}(t')\right) \\ = L + \int_0^t dt' \partial_t q(t') \left(L_q + \frac{d}{dt} L_{\dot{q}}\right) - L_{\dot{q}} \partial_t q(t')\big|_0^t = L - L_{\dot{q}} \dot{q}, \end{aligned} \tag{5.24}$$

where we have used that, $\partial_t q(t')|_{t'=0} = 0$, $\partial_t q(t')|_{t'=t} = \dot{q}(t)$, utilized the Euler-Lagrange equations Eq. (5.4), $t' \in [0,t]:\ (L_q - \frac{d}{dt} L_{\dot{q}})_{t'=t} = 0$.

Next, let us evaluate the derivative of the action over the coordinate, q:

$$\mathcal{S}_q \doteq \partial_q \mathcal{S}(t;q) = \partial_t \int_0^t dt' L\left(q(t'), \dot{q}(t')\right) = \int_0^t dt' \left(L_q \partial_{q(t)} q(t') + L_{\dot{q}} \partial_{q(t)} \dot{q}(t')\right)$$
$$= \int_0^t dt' \partial_{q(t)} q(t') \left(L_q + \frac{d}{dt} L_{\dot{q}}\right) + L_{\dot{q}} \partial_{q(t)} q(t')\Big|_0^t = L_{\dot{q}}. \tag{5.25}$$

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. (5.24) turns into the minus Hamiltonian, defined above in Eq. (5.21), and the right hand side of Eq. (5.25) 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. (5.24) and Eq. (5.25) as the definitions of the minus Hamiltonian of the system and of the system momentum, respectively,

$$p \equiv L_{\dot{q}}, \quad H(t;q;p) \doteq L_{\dot{q}} \dot{q} - L, \tag{5.26}$$

where the Hamiltonian is considered a function of time, t, coordinate, $q(t)$, and momentum, $p(t)$.

Combining Eqs. (5.24,5.25,5.26), that is (a) and (b) of the Theorem 5.5.3 and the definitions of the momentum and the Hamiltonian, one arrives at the Hamilton-Jacobi (HJ) equation

$$\mathcal{S}_t + H(q; \partial_q S) = 0, \tag{5.27}$$

which provides a nonlinear first order PDE representation of classical mechanics.

It is important to stress that, that if one knows the initial ($t = 0$) value of the action, the explicit expression of the Hamiltonian in terms of the time, coordinate and momentum, and the initial value of $\partial_q S$ at $t = 0$ and at all values of q and p Eq. (5.27 represents a Cauchy initial value problem, therefore resulting in solving the minimum action problem unambiguously. This is a rather remarkable and strong sentence with many important consequences and generalizations. The statement is remarkable because because one gets unique solution for 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 for the case of optimal control, called the Bellman-Hamilton-Jacobi (BHJ) equation. HJ equation, discussed here, and BHJ discussed in Section are linked ultimately to the concept of 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 time and over the coordinate. We have learned that, $\partial_t S$, is nothing but minus Hamiltonian, while $\partial_q S$, is simply momenta (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 $\{q(t'); t' \in [0,t]\}$, to representation as a function, of t and $q(t)$, $\mathcal{S}\{q(t')\} \to \mathcal{S}(t;q)$, one rewrites Eqs. (5.2,5.3)

$$\mathcal{S} = \int pdq - \int Hdt,$$

which then implies the following differential form

$$d\mathcal{S} = \frac{\partial \mathcal{S}}{\partial t}dt + \frac{\partial \mathcal{S}}{\partial q}dq,$$

so that

$$\partial_t \mathcal{S} = -H, \quad \partial_q \mathcal{S} = p,$$

resulting (in combination) in the HJ Eq. (5.27).

Example 5.5.3. 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 5.5.4. 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 original variables (u,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.

5.6 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. (5.19). 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 5.6.1 (Legendre-Fenchel (LF) transform)**.** Legendre-Fenchel transform of a function, $\Phi : \mathbb{R}^n \to \mathbb{R}$, is

$$\Phi^*(k) \doteq \sup_{x\in\mathbb{R}^n} \left(x \cdot k - \Phi(x)\right). \tag{5.28}$$

Often LF transform also refers to as "dual" transform. Then $\Phi^*(k)$ is dual to $\Phi(x)$.

Example 5.6.1. 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

$$\begin{aligned}
&\sup_x \quad \left(x \cdot k - \frac{1}{2} x \cdot A \cdot x + b \cdot x\right) \\
&= \quad \sup_x \left(-\frac{1}{2}(x - (k+b) \cdot A^{-1}) \cdot A \cdot (x - A^{-1}(k+b)) + \frac{1}{2}(b+k) \cdot A^{-1} \cdot (b+k)\right) \\
&= \quad \frac{1}{2}(b+k) \cdot A^{-1} \cdot (b+k),
\end{aligned} \tag{5.29}$$

where the maximum is achieved at $x_* = A^{-1}(k+b)$.

Definition 5.6.2 (Convex function over $\mathbb{R}^n$)**.** A function, $u : \mathbb{R}^n \to \mathbb{R}$ is convex if

$$\forall x, y \in \mathbb{R}^n,\ \lambda \in (0,1): \quad u(\lambda x + (1-\lambda)y) \le \lambda u(x) + (1-\lambda)u(y). \tag{5.30}$$

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 5.6.3 (Convexity and Involution of Legendre-Fenchel)**.** The Legendre-Fenchel transform of a convex function is convex, and it is also an involution, i.e. $(\Phi^*)^* = \Phi$.

5.6.1 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. [**?**]) stated in terms of the supporting lines.

Definition 5.6.4 (Supporting Lines)**.** $f : \mathbb{R} \to \mathbb{R}$ has a supporting line at $x \in \mathbb{R}$ if

$$\forall x' \in \mathbb{R} : \quad f(x') \geq f(x) + \alpha(x' - x).$$

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.

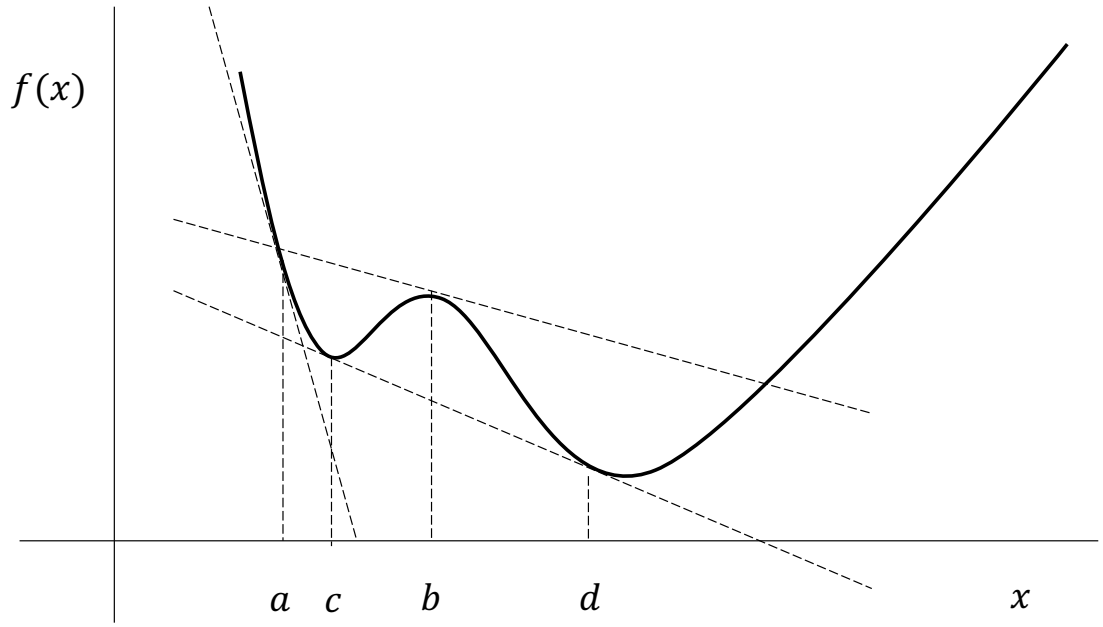


Figure 5.3: Geometric interpretation of supporting lines.

Example 5.6.2. 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 5.6.3. 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. (5.29), 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 5.6.5. 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 5.6.6. 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 5.6.4. Draw supporting lines for the example of a smooth non-convex function shown in Fig. (5.3).

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 5.6.7. $f^*(k)$ is always convex in k.

Corollary 5.6.8. $f^{**}(x)$ 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 5.6.9. $f^{**}(x) = f(x)$ iff $f(x)$ admits a supporting line at x.

The following two statements are immediate corollaries of the theorem.

Corollary 5.6.10. $f^{**} = f$ if f is convex.

Corollary 5.6.11. 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 5.6.12. A convex function can always be written as a LF transform of another function.

Theorem 5.6.13. $f^{**}(x)$ is the largest convex function satisfying $f^{**}(x) \leq f(x)$.

Because of the last statement we call $f^{**}(x)$ the convex envelope of $f(x)$.

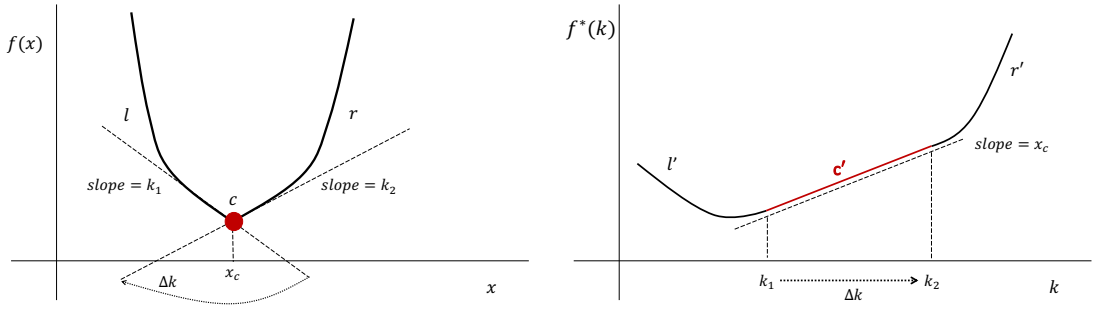


Figure 5.4: Function having a singularity cusp (left) and its LF transform (right).

Below we continue to illustrate the notion of supporting lines, as well as convexity and duality, on illustrative examples.

Example 5.6.5. Consider function containing a non-differentiable point (cusp), as shown in Fig. (5.4a). 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. (5.4a) 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. (5.4b). Overall left (l) and right (r) branches in Fig. (5.4a) transform into left (l') and right (l') branches in Fig. (5.4b)).

- (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. (5.4b), i.e. (c') branch is linear (affine).

The example is convex, therefore according to Corollary 5.6.10, $f^{**}(x) = f(x)$.

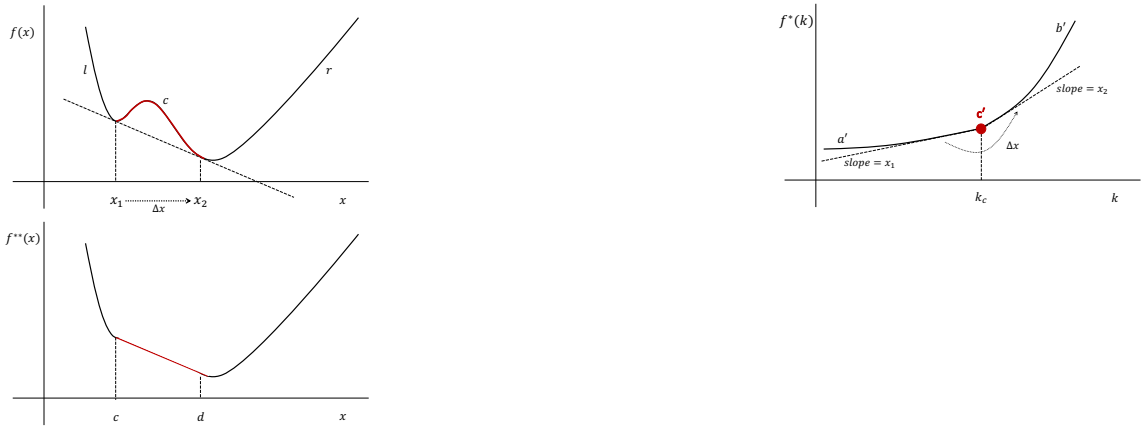


Figure 5.5: (a) An exemplary nonconvex function, $f(x)$; (b) its LT transform, $f^*(k)$; (c) its double LT transform $f^{**}(x)$.

Example 5.6.6. Show schematically $f^*(k)$ and $f^{**}(x)$ for $f(x)$ shown in Fig. (5.3).
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. (5.5) 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. (5.5b) 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. (5.5c). This straight line may be thought as a convex extrapolation, envelope, of $f(x)$ in its non-convex branch.

Exercise 5.6.7. (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.

5.6.2 Primal-Dual Algorithm and Dual Optimization

Now we are ready to return back to the image restoration problem set up in Section 5.1.3. Our task becomes to by-pass ε-smoothing discussed in Section 5.4.2 by using LF transform. This neat theoretical trick will then in developing computationally advantageous primal-dual algorithm. We will use Theorem 5.6.3 to accomplish this, transformation-to-dual, goal.

In fact, let us consider a more general set up than one discussed in Section 5.1.3. Assume that $\Phi : \mathbb{R}^n \to \mathbb{R}$ is convex and consider

$$\min_{\{u(x)\}} \int_U dx \left(\Psi(x, u(x)) + \Phi(\nabla_x u(x))\right), \tag{5.31}$$

where $u : U \to \mathbb{R}$. Let us now restate the formulation in terms of the Legendre-Fenchel transform of Φ, thus utilizing Theorem 5.6.3:

$$\min_{\{u(x)\}} \max_{\{p(x)\}} \int_U dx \left(\Psi(x, u(x)) + p(x) \cdot \nabla_x u(x) - \Phi^*(p(x))\right), \tag{5.32}$$

where $p : U \to \mathbb{R}^n$. (Notice the difference with $u(x) : \mathbb{U} \to \mathbb{R}$.) $u(x)$ is called the primal variable and $p(x)$ is called dual variable. We "dualized" only the second term in the integrand on the right hand side of Eq. (5.31) which is non-smooth, leaving the first (smooth) term unchanged. The optimization problem (5.32) is also called saddle-point formulation, due to its min-max structure.

Given the boundary condition $up \cdot n = 0$ on ∂U, we can apply integration by parts to the term in the middle in Eq. (5.32) then arriving at

$$\min_{\{u(x)\}} \max_{\{p(x)\}} \int\limits_U dx \left(\Psi(x, u(x)) - u(x) \nabla \cdot p(x) - \Phi^*(p(x)) \right). \tag{5.33}$$

We can attempt to solve Eq. (5.32) or Eq. (5.33) 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 5.1.3 – we can carry over the primal-dual min-max formulation further by exploring the structure of the argument (effective action), evaluating minimization over $\{u(x)\}$ 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

$$\Psi(x, u) = \frac{(u - f(x))^2}{2\lambda}, \quad \Phi(w = \nabla_x u(x)) = |w|,$$

in Eq. (5.31) thus arriving at the following optimization

$$\min_u \int\limits_U dx \left(\frac{(u-f)^2}{2\lambda} + |\nabla_x u| \right) \Bigg|_{x \in \partial U:\ n \cdot \nabla_x u = 0}. \tag{5.34}$$

Notice that $\Phi(w) = |w|$ is convex and thus, according to the high-dimensional generalization of what we have learned about LF transform, $\Phi^{**}(w) = \Phi(w)$. The LF dual of $\Phi(w)$ can be easily computed

$$\Phi^*(p) = \sup_{w \in \mathbb{R}^n} (p \cdot w - |w|) = \begin{cases} 0, & |w| \le 1 \\ \infty, & |w| > 1. \end{cases} \tag{5.35}$$

And then convexity of $\Phi(w) = |w|$ allows us, according to Theorem 5.6.3, to "invert" Eq. (5.35)

$$\Phi(w) = |w| = \sup_p \left(p \cdot w - \begin{cases} 0, & |w| \le 1 \\ \infty, & |w| > 1. \end{cases} \right) = \max_{|p| \le 1} p \cdot w. \tag{5.36}$$

Then min-max Eq. (5.33) becomes

$$\min_u \max_{|p| \le 1} \int\limits_U dx \left(\frac{(u-f)^2}{2\lambda} - u \nabla_x \cdot p) \right) \Bigg|_{x \in \partial U:\ n \cdot p = 0}. \tag{5.37}$$

Remarkably we can swap min and max in Eq. (5.37). This is guaranteed by the strong convexity theorem (yet to be discussed in the optimization part of the course/notes)

$$\max_{|p|\le 1}\min_{u}\int\limits_{U} dx\left(\frac{(u-f)^2}{2}-u\nabla_x\cdot p\right)\Bigg|_{x\in\partial U:\ n\cdot p=0}. \tag{5.38}$$

This trick is very useful because the optimization over u can be done explicitly. One finds that the minimum of the quadratic over u function in the integrand of the objective in Eq. (5.38) is achieved at

$$u(p)=f+\lambda\nabla\cdot p, \tag{5.39}$$

and then substituting the optimal value back in the objective we arrive at

$$\max_{|p|\le 1}\int\limits_{U} dx\left(f\nabla_x\cdot p-\frac{\lambda}{2}(\nabla_x\cdot p)^2\right)\Bigg|_{x\in\partial U:\ n\cdot p=0}. \tag{5.40}$$

which is thus the optimization dual to the primal optimization (5.34). If we are to ignore the constraint in Eq. (5.40), the objective is minimal at $\nabla\cdot p=f/\lambda$. To handle the constraint [10] has suggested to use the so-called projected gradient ascent algorithm

$$\forall x:\quad p^{k+1}(x)=\frac{p^k+\tau\nabla_x\cdot(\nabla_x\cdot p^k-f/\lambda)}{1+\tau|\nabla_x\cdot p^k-f/\lambda|}, \tag{5.41}$$

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. (5.41) 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. (5.39).

5.6.3 More on Geometric Interpretation of the LF transform

Here we inject some additional geometric meaning in the LF transform following [11]. 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).

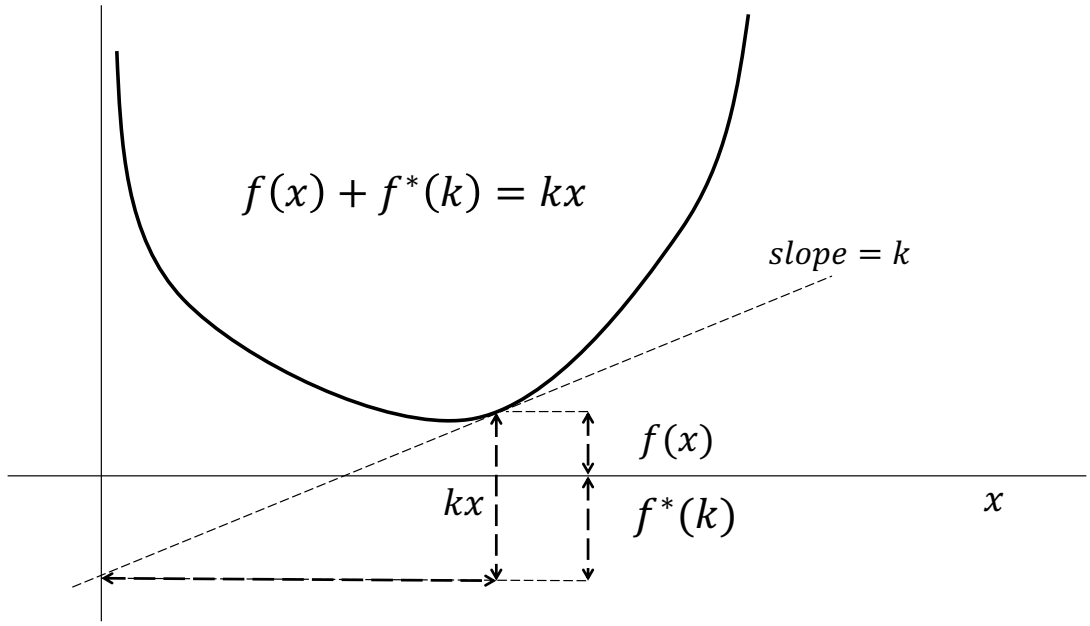


Figure 5.6: Graphic representation of the LF transform.

Second, consider relation, illustrated in Fig. (5.6), 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. $f^*(f'(x))$:

$$\forall x: \quad kx = f(x) + f^*(k), \quad \text{where } k = f'(x). \tag{5.42}$$

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. (5.42) 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 $x = (f^*)'(k)$

5.6.4 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 one 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. (5.42). 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 [11] 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).

5.6.5 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

$$\log F(k,n) = n \sup_{x\in\mathbb{R}} \left(kx - f(x)\right) + o(n).$$

5.7 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

$$\forall \epsilon, \eta \in \mathbb{R}^n_x : \quad \mathrm{Hess}_x(\epsilon,\eta) = \left.\frac{\partial^2 f(x + s\epsilon + w\eta)}{\partial s \partial w}\right|_{s=w=0}. \tag{5.43}$$

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. (5.43)

becomes

$$\begin{aligned}
\mathrm{Hess}_x\{\epsilon(t),\eta(t)\} &= \left.\frac{\partial^2\mathcal{S}\{u(t)+s\epsilon(t)+w\eta(t)\}}{\partial s\partial w}\right|_{s=w=0} \qquad (5.44)\\
&= \left(\frac{\partial}{\partial w}\left(\frac{\partial\mathcal{S}\{q(t)+s\epsilon(t)+w\eta(t)\}}{\partial s}\right)_{s=0}\right)_{w=0}\\
&= \left.\int dt\sum_{i=1}^{n}\left(\frac{\partial L(q+s\epsilon;\dot q+s\dot\epsilon)}{\partial q^i}-\frac{d}{dt}\frac{\partial L(q+s\epsilon;\dot q+s\dot\epsilon)}{\partial\dot q^i}\right)\eta^i\right|_{s=0}\\
&= \int dt\sum_{i,j=1}^{n}\left(\frac{\partial^2 L}{\partial q^j\partial q^i}\epsilon^j+\frac{\partial^2 L}{\partial\dot q^j\partial q^i}\dot\epsilon^j-\frac{d}{dt}\left(\frac{\partial^2 L}{\partial q^j\partial\dot q^i}\epsilon^j+\frac{\partial^2 L}{\partial\dot q^j\partial\dot q^i}\dot\epsilon^j\right)\right)\eta^i\\
&\doteq \int dt\sum_{i,j=1}^{n}J_{ij}\epsilon^j\eta^i,
\end{aligned}$$

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,

$$\mathcal{S}\{q(t)\}=\int_0^T dt\left(\dot q^2-q^2\right). \qquad (5.45)$$

with zero boundary conditions, $q(0)=q(T)=0$. To get some intuition about how the landscape of action (5.45) looks like, let us consider a subclass of functions, for example oscillatory functions consisting of only one harmonic,

$$\bar q(t)=a\sin\left(n\frac{\pi t}{T}\right), \qquad (5.46)$$

where $a\in\mathbb{R}$ (any real) and $n\in\mathbb{Z}\setminus 0$ (any nonzero integer). Substituting Eq. (5.46) into Eq. (5.45) 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 (5.45), 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 (5.45), 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 (5.45) is always positive (over the class of zero boundary condition, twice differentiable functions), and thus the respective quadratic form is always positive definite, if $T < \pi$.

Exercise 5.7.1. Prove that the action $\mathcal{S}\{q(t)\}$ given by Eq. (5.45) 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.)

5.8 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.

5.8.1 Functional Constraint(s)

Consider the shortest path problem discussed in Example 5.2.2, 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 5.8.1. 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, $\mathbb{E}_{\{P(x)\}}(x) \equiv \int_D dx x P(x) = \mu$.

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,

$$e^{-1-\lambda} \left(-\frac{\exp(-\lambda_1 x)}{\lambda_1}\right)\Big|_a^b = 1, \quad e^{-1-\lambda} \left(-\frac{x \exp(-\lambda_1 x)}{\lambda_1} - \frac{\exp(-\lambda_1 x)}{\lambda_1^2}\right)\Big|_a^b = \mu.$$

Exercise 5.8.2. Consider the setting of Example 5.8.1b with $a = -\infty,\ b = \infty$. Assuming additionally that the variance of the probability distribution is known, $\mathbb{E}_{\{P(x)\}}\left(x^2\right) = \sigma^2$, find $P(x)$ which maximizes the entropy.

5.8.2 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,

$$\mathcal{S}\{q(t)\} = \int dt L(t; q(t); \dot{q}(t)), \tag{5.47}$$

over $q : \mathbb{R} \to \mathbb{R}$, however constrained by the functional,

$$\forall t : \quad G(t; q(t); \dot{q}(t)) = 0. \tag{5.48}$$

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

$$\tilde{\mathcal{S}}\{q(t), \lambda(t)\} = \int dt \left(L(t; q(t); \dot{q}(t)) - \lambda(t) G(t; q(t); \dot{q}(t)) \right), \tag{5.49}$$

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 (5.49) over both $q(t)$ and $\lambda(t)$, will give a sufficient condition for the minimum of Eq. (5.47) constrained by Eq. (5.48).

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. (5.49) the modified action in this case becomes

$$\tilde{\mathcal{S}}\{q(t), \lambda(t)\} = \int dt \left(L(q(t); \dot{q}; \dot{v}) - \lambda(t) \left(v(t) - q(t) \right) \right). \tag{5.50}$$

Notice Then the modified Euler-Lagrange equations are

$$\frac{\partial L}{\partial q} = \frac{d}{dt} \left(\frac{\partial L}{\partial \dot{q}} + \lambda \right), \quad -\lambda = \frac{d}{dt} \frac{\partial L}{\partial \dot{v}}, \quad v = \dot{q}. \tag{5.51}$$

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:

$$\frac{\partial L}{\partial q} - \frac{d}{dt} \frac{\partial L}{\partial \dot{q}} + \frac{d^2}{dt^2} \frac{\partial L}{\partial \ddot{q}} = 0. \tag{5.52}$$

Exercise 5.8.3. Find extrema of $\mathcal{S}\{q(t)\} = \int_0^1 dt \|\dot{q}(t)\|$ for $q : [0, 1] \to \mathbb{R}^3$ subject to $\forall t : \quad \|q(t)\|^2 = 1$.

We will see more of the calculus of variations with (function) constraints later in the optimal control section of the course.

Chapter 6

Convex and Non-Convex Optimization

This Section was prepared by Dr. Yury Maximov from Los Alamos National Laboratory (and edited by MC). The material was presented in 6 lectures cross-cut between Math 583, Math 527 and Math 575. In the future the material will mainly be moved to Math 527 and only a brief (one-two lecture) summary will be kept within Math 583. The Section stays here for now, but may become an Appendix later on.

This Section is split into four Subsections. Sections 6.1 and 6.2 will be discussing basic convex and non-convex optimizations. (We focus primarily on finite dimensional case, noticing that generalizations of the basic methods to the infinite-dimensional case, e.g. corresponding to the variational calculus) is straightforward.) Then in Sections 6.3 and 6.4 we will turn to discussing iterative optimization methods for the optimization formulations, set in Sections 6.1 and 6.2, which are of constrained and unconstrained types.

The most general problem we will start our discussion from in Section 6.1 consists in minimization of a function, $f : S \subseteq \mathbb{R}^n \to \mathbb{R}$:

$$\begin{aligned} & f(x) \to \min \\ & \texttt{s.t.:} \quad x \in S \subseteq \mathbb{R}^n. \end{aligned} \tag{6.1}$$

Notice variability in notations – an absolutely equivalent alternative expression is

$$\min_{x \in S \subseteq \mathbb{R}^n} f(x).$$

Section 6.1 should be viewed as introductory (setting notations) leading us to discussion of the notion of (optimization) duality in Section 6.2.

Iterative algorithms, discussed in Sections 6.3 and 6.4, will be designed to solve Eq. (6.1. 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 6.3 and 6.4 primarily on the first-order and second order algorithms.

6.1 Convex Functions, Convex Sets and Convex Optimization Problems

Calculus of Convex Functions and Sets

An important class of functions one can efficiently minimize are convex functions, that were introduced earlier in Definition 5.6.2. We restate it here for convenience.

Definition 6.1.1 (Definition 5.6.2)**.** 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 6.1.2. 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 6.1.3. Let function $f : \mathbb{R}^n \to \mathbb{R}$ has smooth gradient. Then f is convex iff

$$\forall x:\ \nabla^2 f(x) \doteq (\partial_{x_i}\partial_{x_j} f(x); \forall i, j = 1, \cdots, n) \succeq 0,$$

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 6.1.4. Prove that the definitions above are equivalent for sufficiently smooth functions.

Proof. Assume that the function is convex according to the Definition 6.1.1. Then for any $h \in \mathbb{R}^n$, $\lambda \in [0,1]$, one has according to the Definition 6.1.1:

$$f(\lambda(x+h)+(1-\lambda)x) - f(x) = f(x+\lambda h) - f(x) \leq \lambda(f(x+h)-f(x)).$$

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. 6.1.2. 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 $\lambda \in [0,1]$:

$$f(y) \geq f(z) + \nabla f(z)^\top (y-z) = f(z) + \lambda \nabla f(z)^\top (y-x),$$
$$f(x) \geq f(z) + \nabla f(z)^\top (x-z) = f(z) + (1-\lambda)\nabla f(z)^\top (x-y)$$

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. 6.1.1 and Def. 6.1.2.

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 6.1.2 to the Definition 6.1.3 and vice versa. □

Definition 6.1.5. Function $f(x)$ is *concave* iff $-f(x)$ is convex.

Definition 6.1.2 is probably the most practical. To generalize it to non-smooth functions, we introduce the notion of *sub-gradient*.

Definition 6.1.6. 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 6.1.7. 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 6.1.8. 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 6.1.2). 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 6.1.1 (Math 575)**.** 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 6.1.2. 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:

$$\max(x_1, x_2, \ldots, x_n) \approx \frac{1}{\lambda} \log\left(\sum_{i=1}^n \exp(\lambda x_i)\right), \ \lambda \to 0, \ \lambda > 0. \tag{6.2}$$

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: $\|x\|_p \doteq (|x_i|^p)^{1/p}, x \in \mathbb{R}^n$, also called p-norm, or ℓ_p-norm when $p \geq 1$, is convex.

g) Dual norm $\|\cdot\|_*$ to $\|\cdot\|$ is $\|y\|_* \doteq \sup_{\|x\| \leq 1} x^\top y$. 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 6.1.3. 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 6.1.4. Check that all functions and all sets above are convex using Definition 6.1.1 of the convex function (or equivalent Definitions 6.1.2,6.1.3) and the Definition 6.1.7 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 6.1.9. 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 6.1.9 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 6.1.5 (Math 527)**.** 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 6.1.6 (Math 527)**.** 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 6.1.7 (Math 527)**.** 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 6.1.8 (Math 527)**.** 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?

6.2 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:

$$\begin{aligned} f(x) &\to \min \\ \texttt{s.t.}: g_i(x) &\le 0,\ 1 \le i \le m \\ h_j(x) &= 0,\ 1 \le j \le p \\ x &\in \mathbb{R}^n \end{aligned} \tag{6.3}$$

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 $\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$:

$$\mathcal{L}(x, \lambda, \mu) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) + \sum_{j=1}^{p} \mu_j h_j(x) = f(x) + \lambda^\top g(x) + \mu^\top h(x), \qquad \lambda \ge 0 \tag{6.4}$$

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 (6.4) allows us to formulate the constrained optimization, Eq. (6.3), as a min-max (also called saddle point) optimization problem:

$$p^* = \min_{x \in S \subseteq \mathbb{R}^n} \max_{\lambda \ge 0, \mu} \mathcal{L}(x, \lambda, \mu) \tag{6.5}$$

where the optimum of Eq. (6.3) is achieved at p_*.

Weak and Strong Duality

Let us consider the saddle point problem (6.5) 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

$$g(\lambda, \mu) = \min_{x \in S} \mathcal{L}(x, \lambda, \mu) \le \min_{x \in S} f(x) = p^* \Rightarrow \max_{\lambda \ge 0, \mu} \underbrace{\min_{x \in S} \mathcal{L}(x, \lambda, \mu)}_{g(\lambda, \mu)} \le p^* = \min_{x \in S} \max_{\lambda \ge 0, \mu} \mathcal{L}(x, \lambda, \mu),$$

where $g(\lambda,\mu) = \inf_{x\in\mathbb{R}^n} \mathcal{L}(x,\lambda,\mu) = \inf_{x\in\mathbb{R}^n} \left\{f(x) + \lambda^\top g(x) + \mu^\top h(x)\right\}$ 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 , $\max_{\lambda\geq 0,\mu} g(\lambda,\mu) = \max_{\lambda\geq 0,\mu}\min_{x\in S} \mathcal{L}(x,\lambda,\mu)$, 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 $g(\lambda,\mu)$ is always concave. Indeed for $(\bar{\lambda},\bar{\mu}) = \theta(\lambda_1,\mu_1) + (1-\theta)(\lambda_2,\mu_2)$ one has

$$\begin{aligned} g(\bar{\lambda},\bar{\mu}) &= \min_x \mathcal{L}(x,\bar{\lambda},\bar{\mu}) = \min_x\{\theta\mathcal{L}(x,\lambda_1,\mu_1) + (1-\theta)\mathcal{L}(x,\lambda_2,\mu_2)\} \\ &\geq \theta\min_x \mathcal{L}(x,\lambda_1,\mu_1) + (1-\theta)\min_x \mathcal{L}(x,\lambda_2,\mu_2) = \theta g(\lambda_1,\mu_1) + (1-\theta)g(\lambda_2,\mu_2) \end{aligned}$$

   The dual (maximization) problem $\max_{\lambda\geq 0,\mu} g(\lambda,\mu)$ is equivalent to the minimization of the convex function $-g(\lambda,\mu)$ over the convex set $\lambda\geq 0$.

2. *Lower bound property.* $g(\lambda,\mu)\leq p^*$ for any $\lambda\geq 0$.

3. *Weak duality:* For any optimization problem $d^*\leq p^*$. Indeed, for any feasible (x,λ,μ) we have $f(x)\geq \mathcal{L}(x,\lambda,\mu)\geq g(\lambda,\mu)$, thus $p^* = \min_{x\in\mathbb{R}^n} f(x)\geq \max_{\lambda\geq 0,\mu} g(\lambda,\mu) = d^*$.

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 6.2.1. 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

$$g(\lambda) = \inf_{y>0,x}(\exp(-x) + \lambda x^2/y) = 0.$$

That is the dual problem is $\max_{\lambda\geq 0} 0 = 0$, and the duality gap is $p^* - d^* = 1$.

Theorem 6.2.1 (Slater (sufficient) conditions)**.** Consider the optimization (6.3) 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 (6.5). 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. (6.5) 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. (6.4). 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

$$g(\lambda^*, \mu^*) = f(x^*) + g(x^*)^\top \lambda^* + h(x^*)^\top \mu^* = f(x^*) \tag{6.6}$$

where the first equality holds because of the problem stationarity, and the second conditions holds because of the complementary slackness.

Example 6.2.2. Find a duality gap and solve the dual problem for the following minimization

$$\begin{aligned} (x_1 - 3)^2 + (x_2 - 3)^2 &\to \min \\ \texttt{s.t.}: x_1 + 2x_2 &= 4 \\ x_1^2 + x_2^2 &\le 5 \end{aligned}$$

Note, that the problem is (strongly) convex and the Slater's condition is satisfied, therefore the minimum is unique. 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

$$g(\lambda, \mu) = \inf_{x \in \mathbb{R}^n} \mathcal{L}(x, \lambda, \mu).$$

The KKT conditions are

$$\nabla \mathcal{L} = \begin{pmatrix} 2(x_1 - 3) + \mu + \lambda x_1 \\ 2(x_2 - 2) + 2\mu + 2\lambda x_2 \end{pmatrix} = 0$$

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

$$g(\lambda) = \frac{9 + 16\lambda - 9\lambda^2}{5(1 + \lambda)^2} \to \max_{\lambda \ge 0}$$

Finally, the saddle point is $(x_1^*, x_2^*, \lambda_1^*, \lambda_2^*) = (2, 1, 2/3, 1/3)$.

Example 6.2.3. 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:

   $$g(\lambda,\mu) = \inf_{x,y} \mathcal{L}(x,y,\mu,\lambda) = \begin{cases} 2\mu+3\lambda, & \text{if } 3-\mu-\lambda=0 \text{ and } 7-5\mu-\lambda=0 \\ -\infty, & \text{otherwise} \end{cases}$$

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 6.2.4. (Math 583) 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 $4x+5y+7z$, `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 $\{(x_1-5/2)^2+7x_2^2-x_3^2\}$, `s.t.:` $x_1^2-x_2 \leq 0$, and $x_3^2+x_2 \leq 4$

Examples of Duality

Example 6.2.5 (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 the variational calculus Section (of the Math 586 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

$$f(x) \to \min$$
$$\texttt{s.t.}: x = b$$

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 6.2.6. [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

$$g(\lambda) = \inf_{x\in\mathbb{R}^n} \mathcal{L}(x, \lambda)$$
$$= \inf_{x\in\mathbb{R}^n} \left\{x^\top (c + A^\top \lambda) - b^\top \lambda\right\} = \begin{cases} -b^\top \lambda, & \text{if } c + A^\top \lambda = 0 \\ -\infty, & \text{otherwise} \end{cases}$$

The resulting dual optimization is

$$g(\lambda) = -b^\top \lambda \to \max_{c+A^\top\lambda=0,\ \lambda\ge 0}$$

Example 6.2.7 (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:

$$g(\lambda) = \inf_{x\in\mathbb{R}^n} \mathcal{L}(x, \lambda)$$
$$= \inf_{x\in\mathbb{R}^n} \left\{x^\top (A + \lambda I)x - 2b^\top x - \lambda\right\} = \begin{cases} -\infty, & A + \lambda I \not\succeq 0 \\ -\infty, & A + \lambda I \succeq 0, b \in Im(A + \lambda I) \\ -b^\top (A + \lambda I)^+ b - \lambda, & \text{otherwise} \end{cases}$$

The resulting dual optimization is

$$\begin{aligned} -b^\top (A+\lambda I)^+ b - \lambda \to \max \\ \mathtt{s.t.:} A+\lambda I \succeq 0 \\ b \in Im(A+\lambda I) \end{aligned}$$

Let us restate the optimization in a convex form by introducing an extra variable t

$$\begin{aligned} -t-\lambda \to \max \\ \mathtt{s.t.:} t \geq b^\top (A+\lambda I)^+ b \\ A+\lambda I \succeq 0 \\ b \in Im(A+\lambda I) \end{aligned}$$

Finally one arrives at

$$\begin{aligned} -t-\lambda \to \max \\ \mathtt{s.t.:} \begin{pmatrix} A+\lambda I & b \\ b^\top & t \end{pmatrix} \succeq 0 \end{aligned}$$

Example 6.2.8 (Dual to binary Quadratic Programming (QP))**.** Consider the following binary quadratic optimization

$$\begin{aligned} x^\top Ax \to \max \\ \mathtt{s.t.:} x_i^2 = 1,\ 1 \leq i \leq n \end{aligned}$$

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

$$\begin{aligned} \sum_{i=1}^n \mu_i \to \min \\ \mathtt{s.t.:} Diag(\mu) \succeq A \end{aligned} \tag{6.7}$$

Note that the optimization (6.7 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.).

Exercise 6.2.9. Find a dual problem, and estimate the duality gap in the following problem:

$$\begin{aligned} &\min\ -\frac{1}{2}x^\top Lx + b^\top x \\ &\text{s.t.: } \|x\|_\infty \le 1 \end{aligned}$$

if $bb^\top \preceq \varepsilon L$ for some small $\varepsilon > 0$. Consider the case $L \succeq 0$ and $L \preceq 0$. Is it true, that for sufficiently small $\varepsilon > 0$ one can solve the problem just stated exactly if $L \preceq 0$?

Conic Duality (additional material)

Standard formulation of the conic optimization is:

$$\begin{aligned} c^\top x &\to \min_x \\ \texttt{s.t.:} Ax &= b \\ x &\in \end{aligned} \tag{6.8}$$

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 $\alpha x + \beta y \in \mathcal{K}$, $\alpha, \beta \ge 0$;
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 6.2.8 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 $\mathcal{K}$ $\mathcal{K}^* = \{c :\ c^\top \langle c, x\rangle \ge 0, x \in \mathcal{K}\}$.

Exercise 6.2.10. 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, $Q^n = \{(x, t) \in \mathbb{R}^n_+ :\ t \ge \|x\|_2\}$

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 6.8 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

$$p^* = \min_{x \in \mathcal{K}} \max_{\lambda \in \mathcal{K}^*, \mu} \mathcal{L}(x, \lambda, \mu) \geq$$
$$d^* = \max_{\lambda \in \mathcal{K}^*, \mu} \min_{x \in \mathcal{K}} \mathcal{L}(x, \lambda, \mu)$$

And the dual problem is

$$g(\lambda, \mu) = \min_{x \in \mathcal{K}} \{c^\top x + \mu^\top (b - Ax) - \lambda^\top x\} = \begin{cases} \mu^\top b, & \text{if } c - A^\top \mu - \lambda = 0 \\ -\infty, & \text{otherwise} \end{cases}$$

And finally

$$d^* = \max \mu^\top b$$
$$\texttt{s.t.:} = c - A^\top \mu - \lambda = 0$$
$$\lambda \in \mathcal{K}^*$$

Finally, eliminating λ one has

$$\mu^\top b \to \max$$
$$\texttt{s.t.:} c - A^\top y \in \mathcal{K}^*.$$

Exercise 6.2.11. Find a dual problem (see Example 6.2.8) to

$$1^\top \mu = \sum_{i=1}^{n} \mu_i \to \min$$
$$\texttt{s.t.:} Diag(\mu) \succeq A.$$

Ensure, that your dual problem is equivalent to

$$\langle A, X \rangle \to \max$$
$$\texttt{s.t.:} X \in \mathbb{S}^n_+$$
$$X_{ii} = 1 \; \forall i$$

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.

6.3 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

$$\forall x, y: \ \|\nabla f(x) - \nabla f(y)\|_* \le \beta\|x - y\|, \tag{6.9}$$

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 6.3.1. 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 6.3.2. 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:

$$f(x_k) - f(x^*) \leq \frac{\|x_1 - x^*\|_2^2}{2\eta k}, \qquad \eta \leq 1/\beta, \tag{6.10}$$

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 6.3.3. 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 6.3.2: 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

$$\frac{d}{dt}V(t) = (X(t) - x^*)^\top \dot{X}(t) = -\nabla f(X(t))^\top (X(t) - x^*) \leq -(f(X(t)) - f^*), \tag{6.11}$$

where the last inequality is due to the convexity of f. Integrating Eq. (6.11 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 6.3.2: 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

$$f(y) \leq f(x) - \frac{\eta}{2}\|\nabla f(x)\|_2^2. \tag{6.12}$$

Note, that Eq. 6.12 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 6.3.1. 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 6.3.2 was $\eta \leq 1/\beta$.) Derive modified version of Eq. (6.10). Can one find a quadratic convex function for which the modified algorithm fails to converge?

Exercise 6.3.2 (not graded - difficult)**.** 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 6.3.3 (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 6.3.4. Restate and prove discrete time version of the Theorem 6.3.2 for GD in ℓ_p norm. (Hint: Consider the following Lyapunov function: $\|x - x^*\|_p^2$.)

Gradient Descent for Strongly Convex, Smooth Functions.

Theorem 6.3.4. 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 6.3.5. (not graded) Extend proof of the Theorem 6.3.2 to Theorem 6.3.4.

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:

$$x_{k+1} = x_k + \eta_k \nabla f(x_k) + \mu_k (x_k - x_{k-1}) \tag{6.13}$$

(1983) Nesterov Fast Gradient Method (FGM):

$$x_{k+1} = x_k + \eta_k \nabla f(x_k + \mu(x_k - x_{k-1})) + \mu_k (x_k - x_{k-1}). \tag{6.14}$$

The last term in Eqs. (6.13,6.14) 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 6.3.6. (not graded) Construct a convex function f with a piece-wise linear gradient such that the heavy ball algorithm (6.13) with some fixed μ and ηfails to converge.

Consider a slightly modified (less general, two-step recurrence) version of the FGM (6.14):

$$x_k = y_{k-1} - \eta \nabla f(y_{k-1}), \qquad y_k = x_k + \frac{k-1}{k+2}(x_k - x_{k-1}), \tag{6.15}$$

which can be re-stated in continuous time as follows

$$\ddot{X}(t) + \frac{3}{t}\dot{X}(t) + \nabla f(X) = 0. \tag{6.16}$$

Indeed, assuming $t \approx k\sqrt{\eta}$ and re-scaling one derives from Eq. (6.15)

$$\frac{x_{k+1} - x_k}{\sqrt{\eta}} = \frac{k-1}{k+2}\frac{x_k - x_{k-1}}{\sqrt{\eta}} - \sqrt{\eta}\nabla f(y_k). \tag{6.17}$$

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. (6.16).

To analyze convergence rate of the FGM (6.16) 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 6.3.5. 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 6.3.7. (not graded) 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 6.3.8. (not graded) Show that the FGM method, described by Eq. (6.15), transitions to Eq. (6.14) 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:

$$x_{k+1} = x_k - \eta_k g_k, \quad g_k \in \partial F(x_k), \tag{6.18}$$

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 (6.18) 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. (6.18) 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

$$\forall x: \ \|g(x)\| \le L, \ g(x) \in \partial f(x).$$

This condition follows, for example, from the Lipshitz condition, $|f(x) - f(y)| \le L\|x - y\|$, imposed on f. Let x^* be the optimal point of, $f(x) \to \min_{x\in\mathbb{R}^n}$, then

$$\begin{aligned}\|x_{k+1} - x^*\|_2^2 = \|x_k - \eta_k g_k - x^*\|_2^2 &= \|x_k - x^*\|_2^2 - 2\eta_k g_k^\top (x_k - x^*) + \eta_k^2 \|g_k\|_2^2 \\ &\le \|x_k - x^*\|_2^2 - 2\eta_k (f(x_k) - f(x^*)) + \eta_k^2 \|g_k\|_2^2, \end{aligned} \tag{6.19}$$

where the last inequality is due to convexity of f, i.e. $f(x^*) \ge f(x_k) + g_k^\top (x^* - x_k)$. Applying the inequality (6.19) recursively,

$$\|x_{k+1} - x^*\|_2^2 \le \|x_1 - x^*\|_2^2 - 2\sum_{j\le k} \eta_j (f(x_j) - f(x^*)) + \sum_{j\le k} \eta_j^2 \|g_j\|_2^2,$$

one derives,

$$\left(2\sum_{j\le k} \eta_j\right) (f(_{best}^{(k)}) - f(x^*)) \le 2\sum_{j\le k} \eta_j (f(x_j) - f(x^*)) \le \|x^{(1)} - x^*\|_2^2 + \sum_{j\le k} \eta_j^2 \|g_j\|_2^2,$$

which becomes

$$f(_{best}^{(k)}) - f(x^*) = \min_{j\le k} f(x_j) - f^* \le \frac{\|x_1 - x^*\|_2^2 + L_2^2 \sum_{j\le k} \eta_j^2}{2\sum_{j\le k} \eta_j},$$

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

$$\min_{j\leq k} f(x_j) - f^* \leq \min_{\eta} \frac{R_2^2 + L_2^2 \sum_{j\leq k} \eta_j^2}{2\sum_{j\leq k} \eta_j} = \frac{RL}{\sqrt{k}}, \tag{6.20}$$

where the step-size is $\eta_k = R/(L\sqrt{k})$. Note, that the $\sim 1/\sqrt{k}$ scaling in Eq. (6.20 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*

$$f(x) = g(x) + h(x) \to \min_{x\in\mathbb{R}^n}, \tag{6.21}$$

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:

$$f(x) = \|Ax - b\|_2^2 + \lambda\|x\|_1 \to \min_{x\in\mathbb{R}^n}. \tag{6.22}$$

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 (6.21).

Theorem 6.3.6. 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 (6.21) according to

$$f(x_{k+1}) - f^* \leq \frac{\|x_0 - x^*\|_2^2}{2\eta k}.$$

Proof of the Theorem (6.3.6) repeats the logic we use to prove Theorem 6.3.2 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 6.3.5:

Theorem 6.3.7. 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 (6.2) 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 $O(1/\varepsilon)$.

6.4 Constrained First-Order Convex Minimization

Projected Gradient Descent

The Projected Gradient Descent (PGD) is

$$\begin{aligned} x_{k+1} &= \Pi_C(x_k - \eta_k \nabla f(x_k)) \\ &= \arg\min_{y\in C}\left(f(x_k) - \nabla f(x_k)^\top (y-x) + \frac{1}{2\eta_k}\|x-y\|_2^2 + I_C(y)\right) \\ &= \mathrm{prox}_{I_C}(x_k - \eta_k \nabla f(x_k)), \end{aligned} \tag{6.23}$$

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 6.4.1. (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

$$f(x) \to \min, \quad \texttt{s.t.:} x \in S \tag{6.24}$$

In contrast to the PGD algorithm (6.23) making projection at each iteration, the Frank-Wolfe (FW) algorithm solves the following linear problem on C:

$$y_k = \arg\min_{y \in C} y^\top \nabla f(x_k), \quad x_{k+1} = (1-\gamma_k)x_k + \gamma_k y_k, \quad \gamma_k = 2/(k+1). \tag{6.25}$$

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. (6.24), then leading to

$$\max_{y \in C} \nabla f(x_k)^\top (x_k - y) \geq f(x_k) - f^*. \tag{6.26}$$

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 6.4.1. Given that $f(x)$ in Eq. (6.24) is a convex β-smooth function and C is a bounded, convex, compact set, Eq. 6.25 converges to the optimal solution, f^*, of Eq. (6.24) 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

$$\begin{aligned}&f(x)+\frac{\rho}{2}\sum_{i=1}^m(g_i(x))_+^2\to\min\\ &\qquad\qquad\texttt{s.t.:}g_i(x)\le 0,\ 1\le i\le m.\end{aligned}$$

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 6.4.2. 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 6.4.3. 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 6.4.2 (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 6.4.4. 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.

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 5, using optimization approaches, e.g. convex analysis and duality, described in Section 6.1, 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 $\{u(\tau) : [0,t] \to \mathbb{R}\}$. Solve the following constrained problem of the variational calculus type:

$$\min_{\{u(\tau),q(\tau)\}} \int_0^t d\tau (q(\tau))^2 \Bigg|_{\tau\in(0,1]:\ \dot{q}(\tau)=u(\tau),\ u(\tau)\le 1} \tag{7.1}$$

where $t > 0$ and the initial position, $q(0) = q_0$, are 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, $\forall \tau : \ \dot{q} \leq 1$. Then, following the logic of Section 6 we introduce the Lagrangian function,

$$L(q(\tau), \mu(\tau)) = q^2 + \mu(\dot{q} - 1),$$

and then write the KKT conditions, extended from the world of finite dimensional optimization discussed in the previous section to the world of infinite dimensional (variational calculus) optimization. Specifically the four KKT-conditions are:

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 7.0.2. Solve Example 7.0.1 with the condition $u \leq 1$ replaced by $|u| \leq 1$.

7.1 Linear Quadratic (LQ) Control via Calculus of Variations

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*:

$$\mathcal{S}\{q(\tau), u(\tau)\} \doteq \mathcal{S}_{eff}\{u(\tau)\} + \mathcal{S}_{des}\{q(\tau)\} + \mathcal{S}_{fin}(q_t), \tag{7.5}$$

$$\mathcal{S}_{eff}\{u(\tau)\} \doteq \frac{1}{2}\int_0^t d\tau u^T(\tau) R u(\tau), \tag{7.6}$$

$$\mathcal{S}_{des}\{q(\tau)\} \doteq \frac{1}{2}\int_0^t d\tau q^T(\tau) Q q(\tau), \tag{7.7}$$

$$\mathcal{S}_{fin}(q(t)) \doteq \frac{1}{2} q^T(t) Q_{fin} q(t), \tag{7.8}$$

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

$$\mathcal{S}\{q, u, \lambda\} \doteq \mathcal{S}\{q, u\} + \int_0^t d\tau \lambda^T(\tau) \left(-\dot{q} + Aq + Bu\right), \tag{7.9}$$

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 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 $q(t)$. Observe that Eqs. (7.11) are algebraic, thus allowing to express the control vector, u, via the adjoint vector, λ

$$u = -R^{-1}B^T\lambda. \tag{7.15}$$

Substituting it into Eqs. (7.10,7.12) one arrives at the following joint system of the original and adjoint equations

$$\begin{pmatrix} \dot{q} \\ \dot{\lambda} \end{pmatrix} = \begin{pmatrix} A & -BR^{-1}B^T \\ -Q & -A^T \end{pmatrix} \begin{pmatrix} q \\ \lambda \end{pmatrix}, \quad \begin{pmatrix} q(0) \\ \lambda(t) \end{pmatrix} = \begin{pmatrix} q_0 \\ Q_{fin}q(t) \end{pmatrix}. \tag{7.16}$$

The system of ODEs (7.16) 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 (7.16) one derives

$$\begin{pmatrix} q(\tau) \\ \lambda(\tau) \end{pmatrix} = W(\tau) \begin{pmatrix} q(0) \\ \lambda(0) \end{pmatrix}, \tag{7.17}$$

$$W(\tau) = \begin{pmatrix} W^{1,1}(\tau) & W^{1,2}(\tau) \\ W^{2,1}(\tau) & W^{2,2}(\tau) \end{pmatrix} \doteq \exp\left(\tau \begin{pmatrix} A & -BR^{-1}B^T \\ -Q & -A^T \end{pmatrix}\right), \tag{7.18}$$

which allows to express $\lambda(0)$ via $q(0) = q_0$

$$\lambda(0) = Mq_0, \quad M \doteq -\left(W^{2,2}(t) + Q_{fin}W^{1,2}(t)\right)^{-1}\left(W^{2,1}(t) + Q_{fin}W^{1,1}(t)\right). \tag{7.19}$$

Substituting Eqs. (7.17,7.19) into Eq. (7.15) one arrives at the following expression for the optimal control via q_0

$$u(\tau) = -R^{-1}B^T\left(W^{2,1}(\tau) + W^{2,2}(\tau)M\right)q_0. \tag{7.20}$$

A control of this type, dependent on the initial state, is called *open loop* control. This name suggests that at any moment of time, $\tau > 0$, we set the control based only on the

information about the initial state of the system at $t = 0$. The open loop control is normally juxtaposed with the so-called *feedback loop* control, which may also be called the *close loop* control. The feedback loop version of Eq. (7.20), is derived expressing $\lambda(\tau)$ and $q(\tau)$ via q_0 according to Eq. (7.17,7.19) and then substituting the result in Eq. (7.15):

$$\forall \tau \in (0,t]: \qquad u(\tau) = -R^{-1}B^T P(\tau) q(\tau), \tag{7.21}$$

$$P(\tau) \doteq \lambda(\tau) q^{-1}(\tau) \tag{7.22}$$

$$= \left(W^{2,1}(\tau) + W^{2,2}(\tau) M\right) \left(W^{1,1}(\tau) + W^{1,2}(\tau) M\right)^{-1}. \tag{7.23}$$

The feedback loop control, $\lambda(\tau)$, at any moment of time τ, i.e. as we go along, 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. 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

Exercise 7.1.1. Show, utilizing derivations and discussions above, that the matrix, $P(t)$, defined in Eq. (7.22), 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, $P(t) = Q_{fin}$.

Exercise 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 $R \to \infty$. [Hint: Analyze Riccatti Eq. (7.24) in the steady, $t \to \infty$, regime.]

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 $q(\tau)$ 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

$$\min_{\{q(\tau)\}} \int_0^t d\tau \left(\frac{(\dot{q}(\tau) - f(q(\tau)))^T (\dot{q}(\tau) - f(q(\tau)))}{2} + V(q(\tau)) \right) \Bigg|_{q(0)=q_0, \quad q(t)=q_t} . \tag{7.30}$$

Following Lagrangian and Hamiltonian approaches, described in details in the variational calculus portion of the course, see Section 5, one identifies action, Lagrangian, momentum and Hamiltonian for the functional optimization (7.30) as follows

$$\begin{aligned}
S\{q(\tau), \dot{q}(\tau)\} &= \int_0^t d\tau \frac{(\dot{q} - f(q))^T (\dot{q} - f(q))}{2} + V(q), && (7.31)\\
L &= \frac{(\dot{q} - f(q))^T (\dot{q} - f(q))}{2} + V(q), && (7.32)\\
p &\equiv \frac{\partial L}{\partial \dot{q}^T} = \dot{q} - f(q), && (7.33)\\
H &\equiv \dot{q}^T \frac{\partial L}{\partial \dot{q}^T} - L = \frac{\dot{q}^T \dot{q}}{2} - \frac{(f(q))^T f(q)}{2} - V(q)\\
&= \frac{p^T p}{2} + p^T f(q) - V(q). && (7.34)
\end{aligned}$$

Then the Euler-Lagrange equations are

$$\begin{aligned}
\forall i = 1, \cdots, d: \quad & \frac{d}{dt} \frac{\partial L}{\partial \dot{q}_i} = \frac{\partial L}{\partial q_i} && (7.35)\\
& \frac{d}{dt} (\dot{q}_i - f_i(q)) = - \sum_{j=1}^{d} (\dot{q} - f(q))_j \, \partial_{q_i} f_j(q) + \partial_{q_i} V(q),
\end{aligned}$$

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 equations 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 5.5.3 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 equations, in the control formulation, are called Bellman or Bellman-Hamilton-Jacobi (BHJ) equation, and sometimes just Bellman equations, to commemorate contribution of Bellman to the field, who has formulated the problem and resolved it deriving the BHJ equations.

In Section 7.4 we derive the BHJ equations in a more general setting.

7.3 Pontryagin Minimal Principle

Let us now consider the following (almost) most general optimal control problem formulated for a dynamical system in a state, $q(\tau) \in \mathbb{R}^d$, evolving in time, $\tau \in [0, t]$:

$$\min_{\{u(\tau), q(\tau)\}} \left(\phi(q(t)) + \int_0^t d\tau L\left(\tau, q(\tau), u(\tau)\right) \right) \Bigg| \begin{array}{ll} \forall \tau \in (0,t]: & \dot{q}(\tau) = f(\tau, q(\tau), u(\tau)),\ q(0) = q_0 \\ \forall \tau \in [0,t]: & u(\tau) \in U \subset \mathbb{R}^d \end{array} \quad (7.40)$$

where the control $u(\tau)$ is restricted to domain U of the d-dimensional space at all the times considered.

Analog of the standard variational calculus approach, consisting in the necessary Euler-Lagrange (EL) conditions over $\{u\}$ and $\{q\}$, is called Pontryagin Minimal Principle (PMP),

commemorating contribution of Lev Pontryagin to the subject [12] (see also [13] for extended discussion of the PMP bibliography, circa 1963). We present it here without much of elaborations (as it follows straightforwardly the same variational logic repeated by now many times in this Section). Introduce the effective action,

$$\tilde{\mathcal{S}} \doteq \mathcal{S} + \int_0^t d\tau \lambda(\tau) \left(f(\tau, q(\tau), u(\tau)) - \dot{q}(\tau) \right),$$

where $\{\lambda(\tau)\}$ is a Lagrangian multiplier (function) and then optimizing over $\{u\}$ and $\{q\}$, we arrive at the expression for the optimal control candidate, u^*, and at the adjoint (dual) equations, respectively

$$\forall \tau \in [0,t]: \qquad \min_{\{u\}} \tilde{S}: \quad u(\tau) = \arg\min_{\tilde{u}} \left(L\left(\tau, q(\tau), \tilde{u}(\tau)\right) + \lambda(\tau) f(\tau, q(\tau), \tilde{u}(\tau)) \right) \tag{7.41}$$

$$\frac{\delta \tilde{\mathcal{S}}}{\delta q(\tau)} = 0: \quad \dot{\lambda}(\tau) = -\frac{\partial}{\partial q} \left(L\left(\tau, q(\tau), u(\tau)\right) + \lambda(\tau) f(\tau, q(\tau), u(\tau)) \right), \tag{7.42}$$

$$\tau = t \qquad \frac{\partial \tilde{\mathcal{S}}}{\partial q(t)} = 0: \quad \lambda(t) = \partial \phi(q(t)) / \partial q(t). \tag{7.43}$$

Notice that Eq. (7.43) 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. (7.43) is equivalent to the derivation of the respective boundary condition (7.13) at $\tau = t$ in the case of the LQ control. Combination of Eqs. (7.41,7.42,7.43) with the (primal) dynamic equations supplemented by the initial condition on $q(0)$ (which are top conditions in Eq. (7.40) 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.

Exercise 7.3.1. Consider a rocket, modeled as a particle of constant (unit) mass moving in zero gravity (empty) two dimensional space. Assume that trust/force acting on the rocket, $f(\tau)$ is known (prescribed) function of time (dependent on, presumably pre-calculated, rate of the fuel burn), and that direction of the thrust can be controlled. Then equations of motion (of 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 $\forall \tau \in [0,t],\ u(\tau) > 0$. Show that $\min_{\{u\}} \phi(q(t))$, 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 is at rest initially, i.e. $q_1(0) = q_2(0) = 0$, and we aim to land the rocket at the furthest longitudinal position away from the origin, i.e. the optimization problem is

$$\max_{\{q\}} q_2(t)\Big|_{q_1(t)=0}.$$

Show that the optimal control in this case is of the following "linear tangent" type:

$$\tan\left(u(\tau)\right) = a + b\tau.$$

7.4 Dynamic Programming in Optimal Control

7.4.1 Discrete Time Optimal Control

Discretizing Eq. (7.40) in time one arrives at

$$\min_{u_{0:n-1}, q_{1:n}} \left(\phi(q_n) + \sum_{k=0}^{n-1} L(\tau_k, q_k, u_k) \right)\Bigg|_{k=0,\cdots,n-1:\ q_{k+1}=q_k+\Delta f(\tau_k,q_k,u_k)}, \tag{7.44}$$

where $k = 1, \cdots, n:\ \tau_k \doteq kt/n,\ q_k \doteq q(\tau_k),\ u_{k-1} \doteq u(\tau_k),\ \Delta \doteq t/n$, and q_0 is assumed fixed.

Main idea of the Dynamic Programming (DP) consists in making optimization in Eq. (7.44) not over all the variables at once, but sequentially, one after another, that is in a greedy fashion. Specifically, let us first optimize in Eq. (7.44) 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. (7.44) evaluated at $k = n-1$. One derives

$$S(n, q_n) \doteq \phi(q_n), \tag{7.45}$$

$$u^*_{n-1} \doteq \arg\min_{u_{n-1}\in U} S\left(n, q_{n-1} + \Delta f\left(\tau_{n-1}, q_{n-1}, u_{n-1}\right)\right) + L\left(\tau_{n-1}, q_{n-1}, u_{n-1}\right), \tag{7.46}$$

$$S(n-1, q_{n-1}) \doteq S\left(n, q_{n-1} + \Delta f\left(\tau_{n-1}, q_{n-1}, u^*_{n-1}\right)\right) + L\left(\tau_{n-1}, q_{n-1}, u^*_{n-1}\right), \tag{7.47}$$

where making optimization over u_{n-1} we took advantage of the Markovian, causal structure of the objective in Eq. (7.44), therefore taking into account only terms in the objective dependent on u_{n-1}. Repeating the same scheme and, first, excluding, q_{n-1}, 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 generalization of Eqs. (7.46,7.47)

$$k = n, \cdots, 1: \quad u^*_{k-1} \doteq \arg\min_{u_{k-1} \in U} S\left(k, q_{k-1} + \Delta f\left(\tau_{k-1}, q_{k-1}, u_{k-1}\right)\right) + L\left(\tau_{k-1}, q_{k-1}, u_{k-1}\right), \tag{7.48}$$

$$S(k-1, q_{k-1}) \doteq S\left(k, q_{k-1} + \Delta f\left(\tau_{k-1}, q_{k-1}, u^*_{k-1}\right)\right) + L\left(\tau_{k-1}, q_{k-1}, u^*_{k-1}\right), \tag{7.49}$$

where Eq. (7.45) sets initial condition for the backward in (discrete) time iterations. It is now clear that $S(0, q_0)$ is exactly solution of Eq. (7.44). $S(k, q_k)$, defined in Eq. (7.48), is called cost-to-go, or value function, evaluated at the (discrete) time τ_k. Eqs. (7.45,7.48,7.49) are summarized in the Algorithm 1.

Algorithm 1 Dynamic Programming [Backward in time Value Iteration]

Input: $L(\tau, q, u)$, $f(\tau, q, u)$ return the value of reward and the vector of incremental state corrections $\forall \tau, q, u$.

1: $\mathcal{S}(n, q) = \phi(q)$
2: **for** $k = n, \cdots, 0$ **do**
3: $\quad u^*_k(q) = \arg\min_u \left(L(\tau_k, q, u) + \mathcal{S}(\tau_k + 1, q_k + \Delta f(\tau_k, q_k, u))\right), \quad \forall q$
4: $\quad \mathcal{S}(k, q) = L(\tau_k, q, u^*_k(q)) + \mathcal{S}(k + 1, q_k + \Delta f(\tau_k, q, u^*_k(q))), \quad \forall q$
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 [14].

In accordance with the greedy nature of the DP algorithm construction – one step at a time, backward in time – it gives an example of what is called a greedy algorithm in Computer Science, that is an algorithm that makes locally optimal choice at each step. 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 prof of (through the sequence of transformations of Eqs. (7.45,7.48,7.49) resulted in the optimal solution of Eq. (7.44)), is that the greedy algorithm in this case is optimal/exact.

7.4.2 Continuous Time & Space Optimal Control

Taking a continuous limit of Eqs. (7.45,7.48,7.49) one arrives at the already familiar from Section 7.2 Bellman, or Bellman-Hamilton-Jacobi, equation

$$-\partial_\tau \mathcal{S}(\tau, q) = \min_{u \in U} \left(L(\tau, q, u) + f(\tau, q, u)\partial_q \mathcal{S}(\tau, q)\right). \tag{7.50}$$

Then expression for the optimal control, that is continuous time version of the line 3 in the Algorithm 1, is

$$\forall \tau \in (0,t]: \; u^*(\tau,q) = \arg\min_{u\in U} \left(L(\tau,q,u) + \partial_q \mathcal{S}(\tau,q) f(\tau,q,u) \right). \tag{7.51}$$

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:

$$\tau \in (0,t] : \ddot{x}(\tau) = -x(\tau) + u(\tau), \quad |u(\tau)| < 1, \tag{7.52}$$

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

$$\forall \tau \in (0,t]: \qquad q = \begin{pmatrix} q_1 \\ q_2 \end{pmatrix} \doteq \begin{pmatrix} x \\ \dot{x} \end{pmatrix}, \quad \dot{q} = Aq + Bu, \tag{7.53}$$

$$A \doteq \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \quad B \doteq \begin{pmatrix} 0 \\ 1 \end{pmatrix}. \tag{7.54}$$

We arrive at the optimal control problem (7.40) where, $\phi(q) = C^T q$, $C^T \doteq (-1,0)$, $L(t,q,u) = 0$, $f(t,q,u) = Aq + Bu$. Then Eq. (7.50) becomes

$$\forall \tau \in (0,t]: \; -\partial_\tau \mathcal{S} = (\partial_q S)^T Aq - \left| (\partial_q \mathcal{S})^T B \right|. \tag{7.55}$$

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. (7.55) one derives

$$\forall \tau \in (0,t]: \quad \dot{\psi} = -A^T \psi, \quad \dot{\alpha} = |\psi^T B|. \tag{7.56}$$

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. (7.51) one

derives

$$\forall \tau \in (0,t]: \ \psi(\tau) = \begin{pmatrix} -\cos(\tau - t) \\ \sin(\tau - t) \end{pmatrix}, \quad u(\tau, q) = -\text{sign}(\Psi_2(\tau)) = -\text{sign}\left(\sin(\tau - t)\right), \tag{7.57}$$

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

$$u(\tau) = \begin{cases} -1, & 0 < \tau < \pi \\ 1, & \pi < \tau < 2\pi \end{cases}, \tag{7.58}$$

and the optimal trajectory is

$$q^T = (q_1, q_2) = \begin{cases} (\cos(\tau) - 1, -\sin(\tau)) & 0 < \tau < \pi \\ (3\cos(\tau) + 1, -3\sin(\tau)) & \pi < \tau < 2\pi \end{cases} \tag{7.59}$$

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 7.4.2. Consider a soft version of the problem discussed in Example 7.4.1:

$$\min_{\{u(\tau)\},\{q(\tau)\}} \left(C^T q(t) + \frac{1}{2} \int_0^t d\tau (u(\tau))^2 \right) \Bigg|_{\forall \tau \in (0,t]: \ \dot{q}(\tau) = Aq(\tau) + Bu(\tau)}, \tag{7.60}$$

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.

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 raw is, $c(i, j)$. Then the total cost of placing n words in (presumably) nice looking text consisting of l rows is

$$c(1, j_1) + c(j_1 + 1, j_2) + \cdots + c(j_l + 1, n), \tag{7.61}$$

where $1 < j_1 < j_2 < \cdots < j_l < n$. We will seek for an optimal sequence minimizing the total cost. To make description of the problem complete one needs to introduce a plausible way of "pricing" the 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 margin, L, and define the cost to be a monotonically increasing function of the stretching factor, for example

$$c(i,j) = \begin{cases} +\infty, & L < (j-i) - \sum_{k=i}^{j} w_k \\ \left(\dfrac{L - (j-i) - \sum_{k=i}^{j} w_k}{j-i} \right)^3, & \text{otherwise} \end{cases} \tag{7.62}$$

(The cubic dependence in Eq. (7.62) is an empirical way to introduce preference for smaller stretching factors. Notice also that Eq. (7.62) 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.)

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 on if 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 answer to this question is affirmative, and in fact, as we will see below the solution is of the Dynamic Programming (DP) type. The key insight is relation between 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.

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

$$f(1) = \min_j (c(1,j) + f(j+1)). \tag{7.63}$$

while a partial cost satisfies the following recursive relation

$$\forall i: \quad f(i) = \min_{j:i\leq j} (c(i,j) + f(j+1)), \tag{7.64}$$

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. (7.64) is a full analog of the Bellman equation (7.49). Algorithm 2 is a recursive algorithm for $f(i)$ implementing Eq. (7.64).

Algorithm 2 Dynamic Programming for LaTeX Engine

Input: $c(i,j), \forall i,j = 1,\cdots,n$, e.g. according to Eq. (7.62). $f(n+1) = 0$.

```
for i = n, ··· , 1 do
    f_min = +∞
    for j = i, ··· , n do
        f_min = min (f_min, c(i, j) + f(j + 1))
    end for
end for
```

Output: $f(i),\ \forall i = 1,\cdots,n$

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 $O(n^2)$.

7.5.2 Shortest 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 $(i-1, j)$ or up $(i, j-1)$ 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 3 Dynamic Programming for Shortest Path over Grid

Input: Costs assigned: $a(i, j), \forall i = 1, \cdots, N;\ \forall j = 1, \cdots, M$. Boundary conditions fixed: $p(i, 0) = +\infty, \forall i = 1, \cdots, N$. $p(0, j) = +\infty, \forall j = 1, \cdots, M$. Initialization: $p(1, 1) = 0$.

```
for t = 2, ..., N + M do
    for i + j = t, i, j ≥ 0 do
        p(i, j) = min (p(i − 1, j), p(i, j − 1)) + a(i, j)
    end for
end for
```

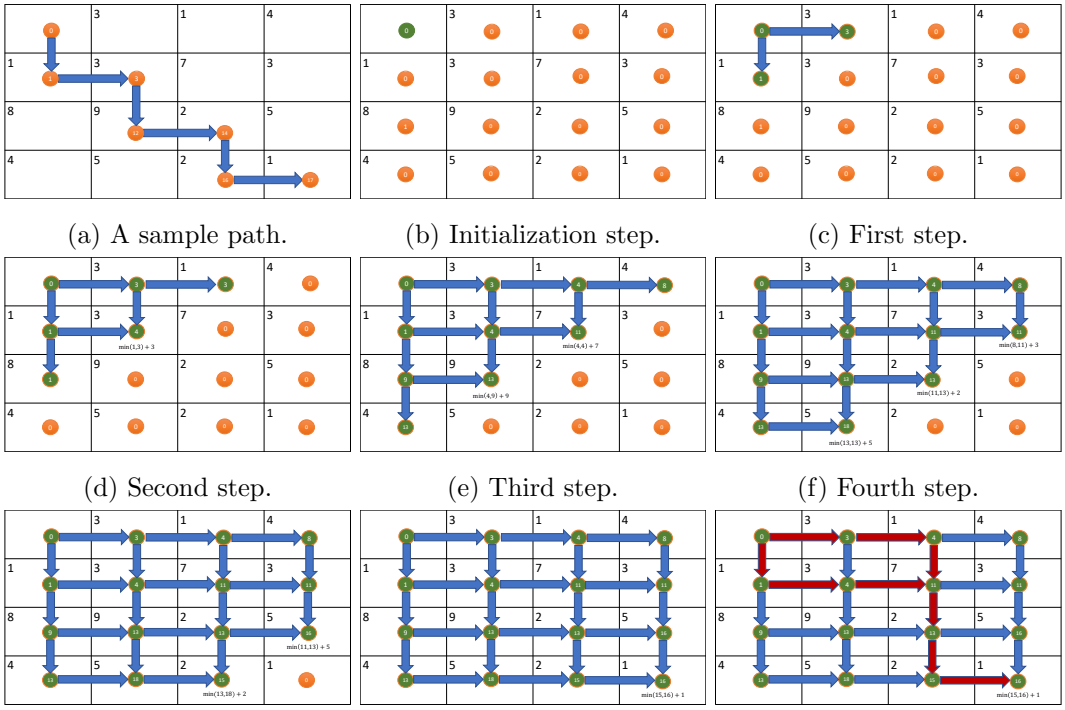
Output: $p(i, j),\ \forall i = 1, \cdots, N;\ j = 1, \cdots, M$.

Algorithm performance is illustrated in Fig. (7.1).

7.5.3 DP for Graphical Model Optimization

Number of optimization problems which can be solved with DP efficiently is remarkably broad. In particular, it appears that the following combinatorial optimization problem, over binary n-dimensional variable, x:

$$E \doteq \min_{x \in \{\pm 1\}^n} \sum_{i=1}^{n-1} E_i(x_i, x_{i+1}), \tag{7.65}$$



(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 Shortest-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).

which requires optimization over 2^n possible states, can be solved efficiently by DP in efforts linear in n. In the jargon of mathematical physics the problem just introduced is called "finding a ground state of the Ising model".

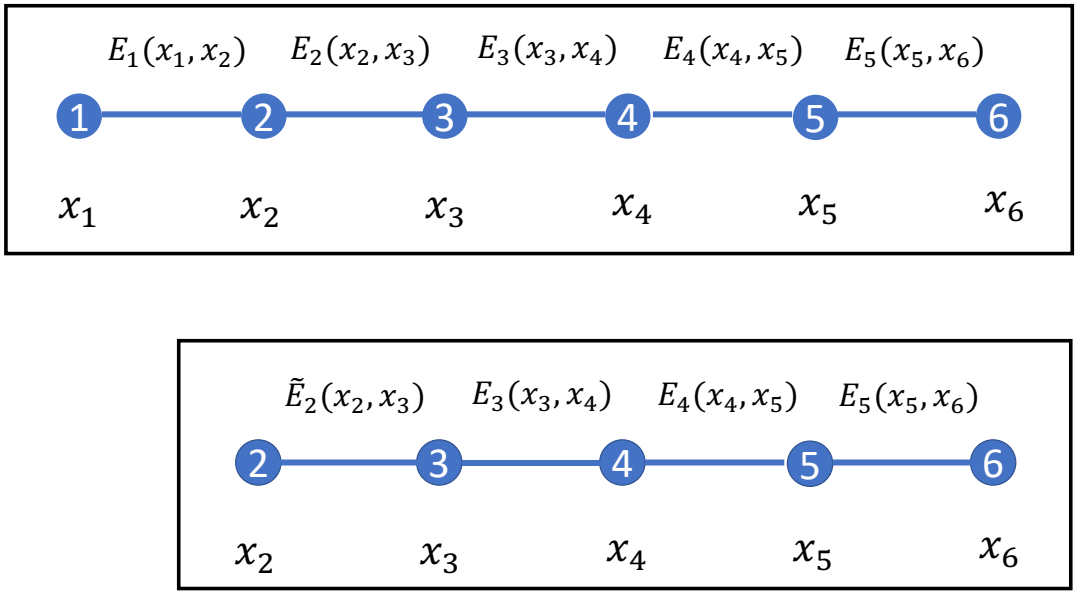


Figure 7.2: Top: Example of a linear Graphical Model (chain). Bottom: Modified GM (shorter chain) after one step of the DP algorithm.

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. (7.2). 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 (7.65) on the example in Fig. (7.2). 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

$$E = \min_{x_1} E_1(x_1, x_2) + \min_{x_2,\cdots,x_n} \sum_{i=2}^{n-1} E_i(x_i, x_{i+1})$$

$$= \min_{x_2,\cdots,x_n} \left(\tilde{E}_2(x_2, x_3) + \sum_{i=3}^{n-1} E_i(x_i, x_{i+1}) \right), \tag{7.66}$$

$$\tilde{E}_2(x_2, x_3) \doteq E_2(x_2, x_3) + \min_{x_1} E_1(x_1, x_2), \tag{7.67}$$

where we took advantage of the objective factorization (into sum of terms each involving

only a pair of neighboring components). Notice that in the result of minimization over x_1 we arrive at the problem with exactly the same structure we started from, i.e. a chain, which is however shorter by one node (and edge). The only change 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. (7.2) on 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.

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: $E = \sum_{x_{n-1}, x_n} E_{n-1}(x_{n-1}, x_n)$

Consider generalization of the combinatorial optimization problem (7.67) to the case of a single-connected tree, $\mathcal{T} = (\mathcal{V}, \mathcal{E})$, e.g. one shown in Fig. (7.3):

$$E \doteq \min_{x \in \Sigma^{|\mathcal{V}|}} \sum_{\{i,j\} \in \mathcal{E}} E_{i,j}(x_i, x_j), \tag{7.68}$$

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.

Exercise 7.5.1. Generalize Algorithm 4 to the case of the GM optimization problem (7.68) over a tree, that is compute E defined in Eq. (7.68). (Hint: one can start from any leaf node of the tree, and use induction as in any other DP scheme.)

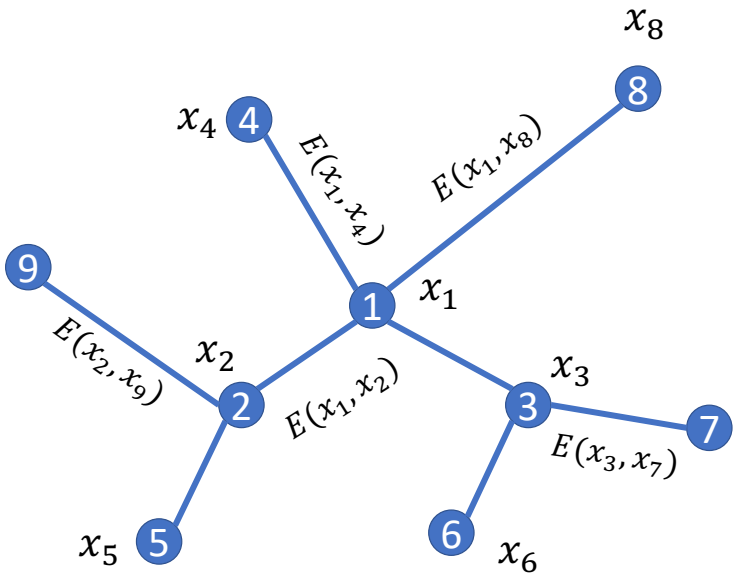


Figure 7.3: Example of a tree-like Graphical Model.

Part IV

Mathematics of Uncertainty

Chapter 8

Basic Concepts from Statistics

8.1 Random Variables: Characterization & Description.

8.1.1 Probability of an event

Consider events drawn from a sample space, Σ. In general Σ may be continuous, e.g. embedded into $\mathbb{R}^n$, however let us start discussing the simple case of a discrete binary space: $\Sigma = \{0, 1\}$. Let us draw a sequence of random variables from the space, for example by tossing a coin. Given that any new toss of a coin does not depend on any previous tossings and also assuming that the law/rule of tossing does not change as we progress, we arrive at the so-called Bernoulli i.i.d. (independent and identically distributed) process described by the probability of being in the state ς:

$$\forall \varsigma \in \Sigma : \quad \text{Prob}(\varsigma) = P(\varsigma) \tag{8.1}$$

$$0 \leq P(\varsigma) \leq 1 \tag{8.2}$$

$$\sum_{\varsigma \in \Sigma} P(\varsigma) = 1, \tag{8.3}$$

where thus, $P(1) = \beta$, $P(0) = 1 - \beta$. If $\beta \neq 1/2$ the coin is biased.

Another important i.i.d discrete event distribution is the Poisson distribution. An event can occur $k = 0, 1, 2, \cdots$ times in an interval. The average number of events in an interval is $\tilde{\lambda}$ - called event rate. The probability of observing k events within the interval is

$$\forall k \in \mathbb{Z}^* = \{0\} \bigcup \mathbb{Z} : \quad P(k) = \frac{\tilde{\lambda}^k e^{-\tilde{\lambda}}}{k!}. \tag{8.4}$$

(Check that the probability is properly normalized, in the sense of Eq. (8.1). Notice also that $\tilde{\lambda}$ is dimensionless. Later we will also be discussing the Poisson process where a related, but dimensional, object λ will be introduced. λ stands for rate of arrival per unit time.)

The distribution is also called exponential distribution (for obvious reason - look at the expression).

Standard notations for Bernoulli and Poisson distributions are Bernoulli(β) or and Poisson($\tilde{\lambda}$), respectively.

Example 8.1.1. Are Bernoulli and Poisson distributions related? Can you "design" Poisson from Bernoulli? Can you give an example of the Poisson process from life/science?

Solution:
Consider repeating Bernoulli, each time independently, thus drawing a Bernoulli process. You get sequence of zeros and ones. Then check only for ones and record times/slots associated with arrivals of ones. Study probability distribution of t arrivals in n step, and then analyze $n \to \infty$, to get the Poisson distribution. (The statement, also called in the literature Poisson Limit Theorem, will be discussed in details in one of the following lectures.) Some examples of processes associated with the Poisson distribution (what we call Possion processes) are: probability distribution of the number of phone calls received by a call center per hour, probability distribution of customers arrival at the shop/bank, probability distribution of the number of meteors greater than 1 meter in diameter that strike earth in a year, probability distribution of the number of typing errors per page page, and many other.

The domain, Σ, can also be continuous, bounded or unbounded. Example of an i.i.d. distribution which is bounded - is the uniform distribution from the $[0, 1]$ interval:

$$\forall x \in [0,1]: \quad p(x) = 1, \tag{8.5}$$

$$\int_0^1 dx p(x) = 1, \tag{8.6}$$

where $p(x)$ is the probability density distribution. (It is custom to use low-key p for the probability density and the upper-case, P, to denote actual probabilities.) Gaussian distribution is the most important (also most frequently used) continuous distribution:

$$\forall x \in \mathbb{Z}: \quad p(x|\sigma,\mu) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{8.7}$$

$p_{\sigma,\mu}(x)$ another possible notation. It is also called "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 so-called law of large numbers, also called central limit theorem, shortly.) The distribution is parameterized by the mean, μ, and by the variance σ^2. Standard notation in math for the Gaussian/normal distribution is $\mathcal{N}(\mu, \sigma^2) = N(\mu, \sigma^2)$.

There are many more 'standard' distributions (i.i.d. or not) beyond the golden three – Bernoulli, Poisson and Gaussian. In fact one can generate practically any other distribution from the 'golden set' (possibly extended with the uniform distribution).

Let us make a brief remark about 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, e.g. X. A random variable takes on values in some domain, and if we want to consider a particular instantiation, thus instance/sample, of the random variable (that is, it has been sampled and observed to have a particular value in the domain) then that non-random value is denoted by lower case e.g. x. $\mathbb{E}\left(f(x)\right) = \mathbb{E}_X\left(f(x)\cdots\right) = \mathbb{E}_{P_X}\left(f(x)\cdots\right) = \langle f(x)\rangle$ are all the different notations used for averaging of a function $f(x)$ of the variable x over the probability distribution, $P(x)$, that is $\sum_{x\in\Sigma} f(x)P(x)$. $x \sim P(x)$ denotes the fact that the random variable x is drawn from the distribution, $P(x)$.

8.1.2 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).

8.1.3 Moments. Generating Function.

Expectations.

$$\mathbb{E}\left[A(\varsigma)\right]_p = \langle A(\varsigma)\rangle_p = \sum_{\varsigma\in\Sigma} A(\varsigma)p(\varsigma).$$

Examples: mean,

$$\mathbb{E}\left[\varsigma\right],$$

variance,

$$\mathrm{Var}[\varsigma] = \mathbb{E}\left[(\varsigma - \mathbb{E}[\varsigma])^2\right].$$

We have already discussed these for the Gaussian process.

Example 8.1.2. What is the average number of the events in the Poisson process, Pois(λ), described by the probability distribution function (8.4)? What is the second moment (variance) of the Poisson distribution?

Solution:

The average number of events in the interval

$$\mu_1 = \sum_{k=0}^{\infty} \frac{k\lambda^k}{k!} e^{-\lambda} = \sum_{k=1}^{\infty} \frac{\lambda^k}{(k-1)!} e^{-\lambda} = \lambda \sum_{n=0}^{\infty} \frac{\lambda^n}{n!} e^{-\lambda} = \lambda.$$

The second moment is

$$\mu_2 = \sum_{k=0}^{\infty} \frac{k^2\lambda^k}{k!} e^{-\lambda} = \sum_{k=1}^{\infty} \frac{k\lambda^k}{(k-1)!} e^{-\lambda} = \lambda \sum_{n=0}^{\infty} \frac{(n+1)\lambda^n}{n!} e^{-\lambda} = \lambda(\lambda+1),$$

and then the variance is $\sigma^2 = \mu_2 - \mu_1^2 = \lambda$. Note, that the expectation value and variance of the Poisson distribution are both equal to the same value, λ.

Example 8.1.3. Consider the Cauchy distribution. (It 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 the distribution is

$$p(x) = \frac{1}{\pi} \frac{\gamma}{(x-a)^2 + \gamma^2}, \quad -\infty < x < +\infty. \tag{8.8}$$

Show that the probability distribution is properly normalized and find its first moment. What can you say about the second moment?

Solution:

The first moment is

$$\mu_1 = \frac{\gamma}{\pi} \int_{-\infty}^{+\infty} \frac{x dx}{(x-a)^2 + \gamma^2} = a. \tag{8.9}$$

(Recall that this integral is an example of the "principal value integral" we have studied in the fall.) The second moment μ_2 is not defined (infinite).

Moments of a probability distribution $P(\varsigma)$ are defined as follows

$$k = 0, \cdots, \quad m_k(\Sigma) \doteq \mathbb{E}_P\left[\varsigma^k\right] = \langle \varsigma^k \rangle_P = \sum_{\varsigma \in \Sigma} \varsigma^k P(\varsigma). \tag{8.10}$$

We can also extend the definition to the probability density $p(x) = p_X(x)$, over continuous valued X:

$$k = 0, \cdots, \quad \mu_k \doteq \mathbb{E}_p\left[x^k\right] = \langle x^k \rangle_p = \int dx x^k p(x). \tag{8.11}$$

Example 8.1.4. Find variance and moments of the Bernoulli distribution, Bernoulli(β), with the probability density function

$$p(x) = \beta\delta(1-x) + (1-\beta)\delta(x). \tag{8.12}$$

Solution:

$$k = 1, \cdots : \quad \mu_k = \langle X^n \rangle = \int_{-\infty}^{\infty} x^n p(x) dx = \beta, \quad n = 1, 2, \ldots . \tag{8.13}$$

In this case the variance is $\sigma^2 = \mu_2 - \mu_1^2 = \beta - \beta^2 = \beta(1-\beta)$.

Moment Generating and Characteristic Function

Moment generating function is defined by

$$M_X(t) = \mathbb{E}\left[\exp(tx)\right] = \int_{-\infty}^{\infty} dx p(x) \exp(tx) = \int_{-\infty}^{\infty} dx p(x) \sum_{k=0}^{\infty} \frac{x^k}{k!} = \sum_{k=0}^{\infty} \frac{\mu_k}{k!}. \tag{8.14}$$

where $t \in \mathbb{R}$ and all integrals are assumed well defined

Example 8.1.5. Consider standard example of Boltzmann distribution from statistical mechanics, where the probability density, $p(s)$, of a state, s is

$$p(s) = \frac{1}{Z} e^{-\beta E(s)}, \quad Z(\beta) = \sum_s e^{-\beta E(s)}, \tag{8.15}$$

where $\beta = 1/T$ is the inverse temperature and $E(s)$ is the known function of s, called energy of the state s. 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. (Notice that up to sign inversion of the argument the partition function is equivalent to the moment generating function (8.14), $Z(\beta) = M_X(-\beta)$.) Compute the expected mean value and the variance of the energy.

Solution:

The mean value (average) of the energy is

$$\langle E \rangle = \sum_s p(s) E(s) = \frac{1}{Z} \sum_s E(s) e^{-\beta E(s)} = -\frac{1}{Z} \frac{\partial Z}{\partial \beta} = -\frac{\partial \ln Z}{\partial \beta}. \tag{8.16}$$

The variance of the energy (energy fluctuations) is

$$\Delta E^2 = \langle (E - \langle E \rangle)^2 \rangle = \frac{\partial^2 \ln Z}{\partial \beta^2}, \tag{8.17}$$

Characteristic function is a related object, defined as a Fourier transform of the probability density:

$$G(k) \doteq \mathbb{E}_p\left[\exp(ikx)\right] = \int_{-\infty}^{+\infty} dx p(x) \exp(ikx), \tag{8.18}$$

where $i^2 = -1$. The characteristic function exists for any real k and it obeys the following relations

$$G(0) = 1, \quad |G(k)| \leq 1. \tag{8.19}$$

The characteristic function contains information about all the moments μ_m. Moreover it allows the Taylor series representation in terms of the moments:

$$G(k) = \sum_{m=0}^{\infty} \frac{(ik)^m}{m!} \langle x^m \rangle, \tag{8.20}$$

and thus

$$\langle x^m \rangle = \frac{1}{i^m} \frac{\partial^m}{\partial k^m} G(k)\Big|_{k=0}. \tag{8.21}$$

This implies that derivatives of $G(k)$ at $k = 0$ exist up to the same m as the moments μ_m.

Example 8.1.6. Find characteristic function of the Bernoulli distribution, Bernoulli(β).

Solution:

Substituting Eq. (8.12) into the Eq. (8.18) one derives

$$G(k) = 1 - \beta + \beta e^{ik}, \tag{8.22}$$

and thus

$$\mu_m = \frac{\partial^m}{\partial (ik)^m} [1 - \beta + \beta e^{ik}]\Big|_{k=0} = \beta. \tag{8.23}$$

The result is naturally consistent with Eq. (8.13).

Exercise 8.1.7. The probability density function of the so-called exponential distribution is

$$p(x) = \begin{cases} Ae^{-\lambda x}, & x \geq 0, \\ 0, & x < 0, \end{cases} \tag{8.24}$$

where the parameter $\lambda > 0$. Calculate

(1) The normalization constant A of the distribution.

(2) The *mean value* and the *variance* of the probability distribution.

(3) The characteristic function $G(k)$ of the exponential distribution.

(4) The $m-$th moment of the distribution (utilizing $G(k)$.

Cumulants

The cumulants are defined by the characteristic function as follows

$$\ln G(k) = \sum_{m=1}^{\infty} \frac{(ik)^m}{m!} \kappa_m. \tag{8.25}$$

According to Eq. (8.19) the Taylor series in Eq. (8.25) start from unity. Utilizing Eqs. (8.20) and (8.25), one derives the following relations between the cumulants and the moments

$$\kappa_1 = \mu_1, \tag{8.26}$$

$$\kappa_2 = \mu_2 - \mu_1^2 = \sigma^2. \tag{8.27}$$

The procedure naturally extends to higher order moments and cumulants.

Notice that moments determine the cumulants in the sense that any two probability distributions whose moments are identical will have identical cumulants 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 8.1.8. Find characteristic function and cumulants of the Poisson distribution (8.4).

Solution:

The respective characteristic function is

$$G(p) = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} e^{ipk} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{(\lambda e^{ip})^k}{k!} = exp\left[\lambda(e^{ip} - 1)\right], \tag{8.28}$$

and then

$$\ln G(p) = \lambda(e^{ip} - 1). \tag{8.29}$$

Next, using the definition (8.25), one finds that $\kappa_m = \lambda$, $m = 1, 2, \ldots$.

Example 8.1.9. Birthday's 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

$$p_m = \prod_{i=1}^{m} \left(1 - \frac{i-1}{366}\right). \tag{8.30}$$

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 8.1.10. (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 $(1, 0)$.

8.1.4 Probabilistic Inequalities.

Here are some useful probabilistic inequalities, which we present here mainly for a reference. (Proofs of the inequalities will be discussed in Math 527. See also `http://jeremykun.com/2013/04/15/probabilistic-bounds-a-primer/`.)

- (Markov Inequality)

$$P(x \geq c) \leq \frac{\mathbb{E}[x]}{c} \tag{8.31}$$

- (Chebyshev's inequality)

$$P(|x - \mu| \geq k) \leq \frac{\sigma^2}{k^2} \tag{8.32}$$

- (Chernoff bound)

$$P(x \geq a) = P(e^{tx} \geq e^{ta}) \leq \frac{\mathbb{E}[e^{tx}]}{e^{ta}} \tag{8.33}$$

where μ and σ^2 are mean and variance of x.

We will get back to discussion of these and some other useful probabilistic inequalities in the lecture devoted to entropy and to how compare probabilities.

Exercise 8.1.11 (not graded)**.** Play, e.g. in IJulia (notebook linked to the lecture), checking the three inequalities for the distributions mentioned through out the lecture. Provide examples of the distributions for which the tree inequalities are saturated (becomes equalities)?

8.2 Random Variables: from one to many.

8.2.1 Law of Large Numbers

Take n 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$(y_n)$? $\sqrt{n}(y_n - \mu)$, converges in distribution to Gaussian with mean, μ, and variance, σ^2:

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} x_i - \mu\right) \to \mathcal{N}(0,\, \sigma^2). \tag{8.34}$$

This is so-called Weak Version of the Central Limit Theorem. (Notice, that the "Large Deviation Theorem" is an alternative name.)

Let us sketch the prove of the weak-CLT (8.34) in a simple case $\mu = 0$, $\sigma = 1$. Obviously, $m_1(Y_n\sqrt{n}) = 0$. Compute

$$m_2(Y_n\sqrt{n}) = \mathbb{E}\left[\left(\frac{x_1 + \cdots + x_n}{\sqrt{n}}\right)^2\right] = \frac{\sum_i \mathbb{E}\left[x_i^2\right]}{n} + \frac{\sum_{i\neq j}\mathbb{E}\left[x_i x_j\right]}{n} = 1.$$

Now the third moment:

$$m_3(Y_n\sqrt{n}) = \mathbb{E}\left[\left(\frac{x_1 + \cdots + x_n}{\sqrt{n}}\right)^3\right] = \frac{\sum_i \mathbb{E}\left[x_i^3\right]}{n^{3/2}} \to 0,$$

at $n \to \infty$, assuming $\mathbb{E}\left[x_i^3\right] = O(1)$. Can you guess what will happen with the fourth moment? $m_4(Y_n\sqrt{n}) = 3 = 3m_2(Y_n)$. This is related to the so-called Wick's theorem (see discussion in the next lecture).

Example 8.2.1 (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 zero mean normal distribution

$$p(x) = \mathcal{N}(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{x^2}{2\sigma^2}\right),$$

exactly.

Solution:

Remind that moments, $\langle x^n\rangle$, over $p(x)$, can be calculated via the characteristic function

$$G(k) \doteq \int e^{ikx}p(x)dx = \exp\left(i\mu k - \frac{\sigma^2 k^2}{2}\right),$$

then resulting in

$$\langle x^{2n}\rangle = \frac{1}{i^m}\frac{\partial^m}{\partial k^m}G(k)\Big|_{k=0}\frac{(2n)!}{2^n n!}\sigma^{2n}, \quad \langle x^{2n+1}\rangle = 0.$$

Then the characteristic function of the distribution $p_n(y_n)$ is

$$G_n(k) = (G(k/n))^n = \exp\left(i\mu k - \frac{\sigma^2 k^2}{2n}\right). \tag{8.35}$$

Inverse Fourier transform of $G_n(k)$ results in

$$p_n(y_n) = \int_{-\infty}^{+\infty}\frac{dk}{2\pi}G_n(k)e^{-iky_n} = \int_{-\infty}^{+\infty}\frac{dk}{2\pi}\exp\left(-ik(y_n - \mu) - n\frac{\sigma^2 k^2}{2}\right) = \tag{8.36}$$

$$= \frac{\sqrt{n}}{\sqrt{2\pi}\sigma}\exp\left(-\frac{n(y_n - \mu)^2}{2\sigma^2}\right). \tag{8.37}$$

Example 8.2.2 (Violation 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

$$p(x) = \frac{\gamma}{\pi}\frac{1}{x^2+\gamma^2}, \tag{8.38}$$

and show that the CLT does not hold in this case. Explain why.

Solution:

The characteristic function of the Cauchy distribution is

$$G(k) = \frac{\gamma}{\pi}\int_{-\infty}^{+\infty}\frac{dx}{x^2+\gamma^2}e^{ikx} = e^{-\gamma k}. \tag{8.39}$$

The resulting characteristic functional expression is

$$G_n(k) = (G(k/n))^n = G(k). \tag{8.40}$$

This expression shows that for any n the variable y_n is Cauchy-distributed with exactly the same width parameter as the individual samples. The CLT is "violated" in this case because we have ignored an important requirement/condition for the CLT to hold – existence of the variance. (See Example 8.1.3.)

Exercise 8.2.3. 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) What is the probability that you win at least 200\$?

Exercise 8.2.4 (not graded)**.** Check Julia notebok for the lecture and experiment with the law of large numbers 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.)

If one is interested in not only the asymptotic, $n \to \infty$, by itself but also in how the asymptotic is approached, the so-called strong version of CLT (also known under the name of the Cramér theorem) states for the normalized sum, $y_n = \sum_{i=1}^{n} x_i/n$, of the i.i.d. variables,

$x_i \sim p_X(x)$,

$$\forall z > \mu : \quad \lim_{n\to\infty} \frac{1}{n} \log \text{Prob}\left(y_n \geq z\right) = -\Phi^*(z) \tag{8.41}$$

$$\Phi^*(x) \doteq \sup_{\lambda\in\mathbb{R}} \left(\lambda z - \Phi(\lambda)\right) \tag{8.42}$$

$$\Phi(\lambda) \doteq \log\left(\mathbb{E}\exp\left(\lambda x\right)\right), \tag{8.43}$$

Here, $\Phi(\lambda)$, is the characteristic function of $p_X(x)$ and $Phi^*(x)$ is the Legendre-Fenchel transform of the characteristic function, also called the Cramér function. This was a formal (mathematical) statement. A less formal ("physical") version of Eq. (8.41) is

$$n \to \infty : \quad \text{Prob}\left(y_n\right) \propto \exp\left(-n\Phi^*(x)\right). \tag{8.44}$$

Note, that the weak version of the CLT (8.34) is equivalent to approximating the Cramer function (asymptotically exact) by a Gaussian around its minimum.

Exercise 8.2.5 (not graded)**.** Prove the strong-CLT (8.41,8.42). [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 8.2.6. Compute Cramer function for the Bernoulli process, i.e. (generally unfair) coin toss

$$x = \begin{cases} 0 & \text{with probability } 1-\beta \\ 1 & \text{with probability } \beta \end{cases} \tag{8.45}$$

Solution:

$$\Phi(\lambda) = \log(\beta e^{\lambda} + 1 - \beta) \tag{8.46}$$

$$0 < x < 1 : \quad \Phi^*(z) = z \log \frac{z}{\beta} + (1-z) \log \frac{1-z}{1-\beta}. \tag{8.47}$$

Eqs. (8.46,8.47) 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

$$n! = \sqrt{2\pi n} n^n e^{-n} \left(1 + O(1/n)\right).$$

(Do you see how?) Second, the $z \log z$ structure is an "entropy" which will appear a number of times in following lectures - stay tuned.

Exercise 8.2.7. Consider n independent Poisson processes, $i = 1, \cdots, n:\ X_i \sim \text{Pois}(\lambda_i)$ each distributed according to its own rate, $\lambda_i > 0$. (That is each of the random numbers in the sequence is independent on others, but they are not identically distributed.) Show that the sum, $Y = \sum_{i=1}^{n} X_i$, is distributed according to the Poisson distribution with the rate $\lambda = \sum_{i=1}^{n} \lambda_i$, i.e. $Y \sim \text{Pois}(\lambda)$.

8.2.2 Multivariate Distribution. Marginalization. Conditional Probability.

Consider an n-component vector build of components each taking a value from a set, Σ, $\varsigma = (\varsigma_i \in \Sigma | i = 1, \cdots, n)$. Σ may be discrete, e.g. $\Sigma = \{0, 1\}$, or continuous, e.g. $\Sigma = \mathbb{R}$. Assume that any state, ς, occur with the probability, $P(\varsigma)$, where $\sum_{\varsigma} P(\varsigma) = 1$.

Consider statistical version the Ising model discussed in Section 7.5 in the discrete optimization setting. (We have used it back then to illustrate application of the Dynamic Programming in combinatorial optimization.) We introduce the following probability distribution over the 2^n-dimensional space Σ (space of cardinality 2^n):

$$\varsigma = (\varsigma_i = \pm 1 | i = 1, \cdots, n) : P(\varsigma) = Z^{-1} \prod_{i=1}^{n-1} \exp(J \varsigma_i \varsigma_{i+1}) \tag{8.48}$$

$$Z = \sum_{\varsigma} \prod_{i=1}^{n-1} \exp(J \varsigma_i \varsigma_{i+1}) \tag{8.49}$$

where Z is the normalization constant, also called the partition function introduced to guarantee that the sum over all the states is unity. For $n = 2$ one gets the example of a bi-variate probability distribution

$$P(\varsigma) = P(\varsigma_1, \varsigma_2) = \frac{\exp(J \varsigma_1 \varsigma_2)}{4 \cosh(J)}. \tag{8.50}$$

$P(\varsigma)$ is also called a joint probability distribution function of the ς vector components, $\varsigma_1, \cdots, \varsigma_n$. It is also useful to consider conditional distribution, say for the example above with $n = 2$,

$$P(\varsigma_1 | \varsigma_2) = \frac{P(\varsigma_1, \varsigma_2)}{\sum_{\varsigma_1} P(\varsigma_1, \varsigma_2)} = \frac{\exp(J \varsigma_1 \varsigma_2)}{2 \cosh(J \varsigma_2)} \tag{8.51}$$

is the probability to observe ς_1 under condition that ς_2 is known. Notice that, $\sum_{\varsigma_1} P(\varsigma_1 | \varsigma_2) = 1,\ \forall \varsigma_2$.

We can also marginalize the multivariate (joint) distribution over a subset of variables. For example,

$$P(\varsigma_1) = \sum_{\varsigma \setminus \varsigma_1} P(\varsigma) = \sum_{\varsigma_2, \cdots, \varsigma_n} P(\varsigma_1, \cdots, \varsigma_n). \tag{8.52}$$

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

$$p(x_1, \ldots, x_n) = \frac{1}{Z} \exp\left(-\frac{1}{2} \sum_{i,j=1,\cdots,n} x_i A_{ij} x_j \right), \tag{8.53}$$

where A is the symmetric, $A = A^T$, positive definite, $A \succ 0$, matrix. If the matrix is diagonal then the probability distribution (8.53) 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. (8.53) is the normalization factor, called the partition function, which is

$$Z = \frac{(2\pi)^{n/2}}{\sqrt{\det A}}. \tag{8.54}$$

Moments of the Gaussian distribution are

$$\forall i: \quad \mathbb{E}[x_i] = \mu_i; \quad \forall i,j: \quad \mathbb{E}[(x_i - \mu_i)(x_j - \mu_j)] = (A^{-1})_{ij} \doteq \Sigma_{ij}, \tag{8.55}$$

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 $\mathcal{N}_n(\mu, \Sigma)$.

Gaussian distribution is remarkable because of its "invariance" properties.

Theorem 8.2.1 (Invariance of Normal/Gaussian distribution under conditioning and marginalization)**.** Consider $x \sim \mathcal{N}_n(\mu, \Sigma)$ 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 of x, $p + q = n$. Assume also that the mean vector, μ, and the covariance matrix, Σ, are split into components as follows

$$\mu = (\mu_1, \mu_2); \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}, \tag{8.56}$$

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: $p(x_1) \doteq \int dx_2 p(x_1, x_2)$ is the following Normal/Gaussian distribution, $\mathcal{N}(\mu_1, \Sigma_{11})$.

- Conditioning: $p(x_1|x_2) \doteq \frac{p(x_1,x_2)}{p(x_2)}$ is the Normal/Gaussian distribution, $\mathcal{N}\left(\mu_{1|2}, \Sigma_{1|2}\right)$, where

$$\mu_{1|2} \doteq \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \quad \Xi_{1|2} \doteq \Sigma_{22} - \Sigma_{12}^T\Sigma_{11}^{-1}\Sigma_{12}. \tag{8.57}$$

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. (8.56).)

8.2.3 Bayes Theorem

We already saw how to get conditional probability distribution and marginal probability distribution from the joint probability distribution

$$P(x|y) = \frac{P(x,y)}{P(y)}, \quad P(y|x) = \frac{P(x,y)}{P(x)}. \tag{8.58}$$

Combining the two formulas to exclude the joint probability distribution we arrive at the famous Bayes formula

$$P(x|y)P(y) = P(y|x)P(x). \tag{8.59}$$

Here, in Eqs. (8.58,8.59) both x and y may be multivariate. Rewriting Eq. (8.59) as

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)}, \tag{8.60}$$

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/`

Exercise 8.2.8. The joint probability density of two real random variables X_1 and X_2 is

$$\forall x_1, x_2 \in \mathbb{R}: \quad p(x_1, x_2) = \frac{1}{Z}\exp(-x_1^2 - x_1x_2 - x_2^2). \tag{8.61}$$

(1) Calculate the normalization constant Z.

(2) Calculate the marginal probability $p(x_1)$.

(3) Calculate the conditional probability $p(x_1|x_2)$.

(4) Calculate the moments $\mathbf{E}[X_1^2X_2^2]$, $\mathbf{E}[X_1X_2^3]$, $\mathbf{E}[X_1^4X_2^2]$ and $\mathbf{E}[X_1^4X_2^4]$.

8.3 Information-Theoretic View on Randomness

8.3.1 Entropy.

Entropy is defined as an expectation of $-log$-probability

$$H = -\mathbf{E}_{P(X)}\left[\log(P(X))\right] = -\sum_{x\in\mathcal{X}} P(x)\log(P(x)), \tag{8.62}$$

where x is drawn from the space $\mathcal{X}$. Intuitively, entropy is a measure of uncertainty. Entropy of a deterministic process, that is a process when a state takes a value, say x_0, with the probability 1, is zero. Indeed, according to Eq. (8.62), $0\log 0 = 0$.

One remark on notations, before we proceed further. The notation used in Eq. (8.62) should be considered as a shortcut. A more accurate notation would be, $H(X)$, on the left hand side of Eq. (8.62), where thus X is the random variable which can take a value $x \in \mathcal{X}$. Following tradition of the information theory, we use H for entropy. Beware of an alternative notation, S, custom in Statistical Physics.

Somehow importantly, the logarithm of the probability distribution is chosen as a measure of the information in the definition of entropy (logarithm and not some other function) because it is **additive** for independent sources.

Let us familiarize ourselves with the concept of entropy on example of the Bernoulli $\{0,1\}$ process (8.45). In this case, there are only two states, $P(X=1)=\beta$ and $P(X=0)=1-\beta$, and therefore

$$H = -\beta\log\beta - (1-\beta)\log(1-\beta). \tag{8.63}$$

Notice that H, considered as a function of β, has the bell like shape with the maximum at $\beta = 1/2$. 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 p. 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 (8.62), has the following properties (some of these can be interpreted as alternative definitions):

- $H \geq 0$
- $H = 0$ 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 $\mathcal{X}$.
- Choice of the logarithm base is custom - just a re-scaling. (Base 2 is custom in the information theory, when dealing with binary variables.)

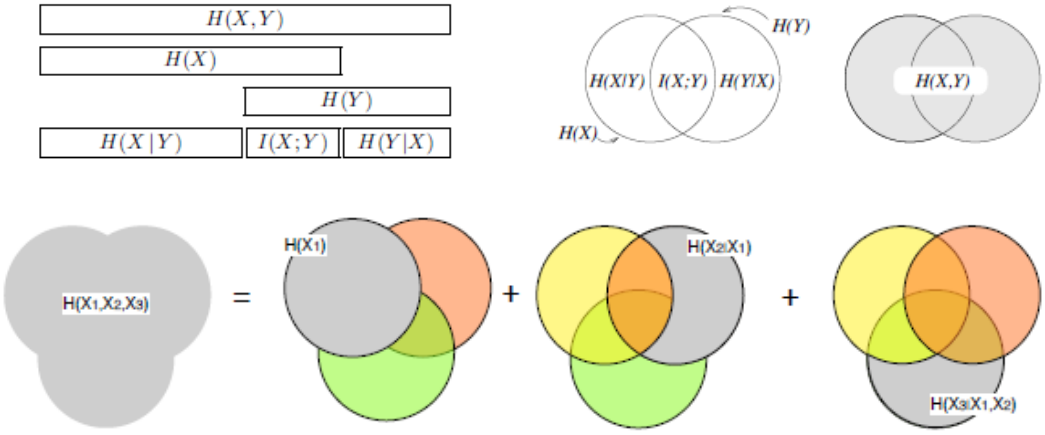


Figure 8.1: Venn diagram(s) explaining the chain rule for computing multivariate entropy.

- 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). (*)
- Entropy is the lower bound on the average length of the shortest description of a random variable

(*) requires a clarification. 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.

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

$$H(X,Y) \doteq -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} P(x,y)\log(P(x,y)), \tag{8.64}$$

and the conditional entropy is

$$H(Y|X) \doteq -\mathbb{E}_{P(X,Y)}\left[\log(P(Y|X)\right] = -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} P(x,y)\log(P(y|x)). \tag{8.65}$$

Note, that $H(Y|X) \neq H(X|Y)$.

Definitions of the joint and conditional entropies naturally lead to the following relation between the two

$$H(X,Y) = H(X) + H(Y|X), \tag{8.66}$$

derived from the Bayes theorem. (Checking it is a good exercise.) Eq. (8.66) is also called the chain rule.

One can naturally extend the chain rule from bi-variate to the multi-variate case $(X_1, \cdots, X_n) \sim P(x_1, \cdots, x_n)$ as follows

$$H(X_n, \cdots, X_1) = \sum_{i=1}^{n} H(X_i|X_{i-1}, \cdots, X_1). \tag{8.67}$$

Notice, that the choice of the order in the chain is arbitrary. The name "chain-rule" should become clear from (8.67). See also Fig. (8.1) for the Venn diagram illustration of the chain rule.

8.3.2 Independence, Dependence, and Mutual Information.

The essence of our next theme is in comparing random numbers, or more accurately their probabilities. Kullback-Leibler (KL) divergence offers a convenient way of doing the comparison

$$D(P_1\|P_2) \doteq \sum_{x\in\mathcal{X}} P_1(x) \log \frac{P_1(x)}{P_2(x)}. \tag{8.68}$$

Note that the KL difference 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. Any proper metric, d_{ab}, for the elements a and b from a space, should be a) positive (for all elements of the space), b) zero when comparing identical states, i.e. $d_{aa} = 0$; c) symmetric, i.e. $d_{ab} = d_{ba}$, and d) satisfy the triangle inequality, $d_{ab} \leq d_{ac} + d_{bc}$. The last two conditions do not hold in the case of the KL divergence. However, an infinitesimal version of the KL divergence - Hessian of the KL distance around its minimum, also called Fisher information, constitutes a proper metric.

Exercise 8.3.1. 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 KL-divergence, $D(P_1\|P_2)$, as a function of a vector $(P_1(x)|x \in \mathcal{X})$, with all the $|\mathcal{X}|$ components non-negative and related to each other via the probability normalization condition, $\sum_{x\in\mathcal{X}} P_1(x) = 1$. Show

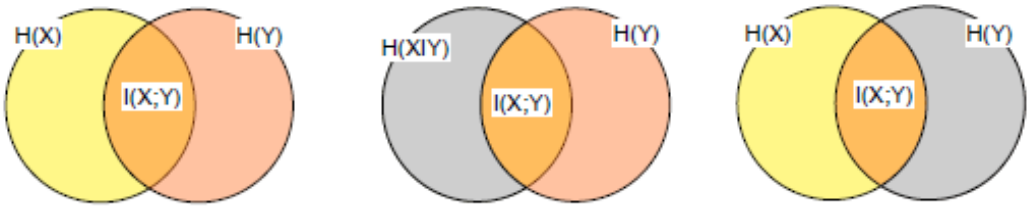


Figure 8.2: Venn diagram explaining relations between the mutual information and respective entropies.

that $D(P_1\|P_2)$ is non-negative and it achieves its minimum at $\forall x \in \mathcal{X}: \; P_1(x) = P_2(x)$, i.e.

$$\arg \min_{(P_1(x)|x\in\mathcal{X})} D(P_1\|P_2)\Big| \begin{array}{l} \sum_{x\in\mathcal{X}} P_1(x) = 1 \\ \forall x \in \mathcal{X}: \quad P_1(x) \geq 0 \end{array} = (P_2(x)|x \in \mathcal{X}). \tag{8.69}$$

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

$$I(X;Y) = \mathbb{E}_{P(x,y)}\left[\log \frac{P(x,y)}{P(x)P(y)}\right] = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P(x,y)\log\frac{P(x,y)}{P(x)P(y)}. \tag{8.70}$$

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,

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y) \tag{8.71}$$

The relation is illustrated in Fig. (8.2). Mutual Information also possesses the following properties

$$I(X;Y) = I(Y;X) \text{ (symmetry)} \tag{8.72}$$

$$I(X;X) = S(X) \text{ (self-information)} \tag{8.73}$$

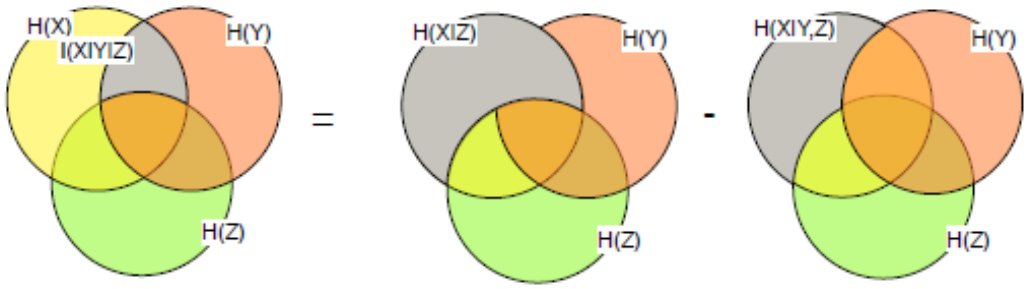


Figure 8.3: Venn diagram explaining the chain rules for mutual information.

The conditional mutual information between X and Y given Z is

$$I(X;Y|Z) \doteq H(X|Z) - H(X|Y,Z) = \mathbb{E}_{P(x,y,z)}\left[\log \frac{P(x,y|z)}{P(x|z)P(y|z)}\right] \tag{8.74}$$

The entropy chain rule (8.66) when applied to the mutual information of $(X_1, \cdots, X_n) \sim P(x_1, \cdots, x_n)$ results in

$$I(X_n, \cdots, X_1; Y) = \sum_{i=1}^{n} I(X_i; Y|X_{i-1}, \cdots, X_1) \tag{8.75}$$

See Fig. (8.3) for the Venn diagram illustration of Eq. (8.75).

See [15] for extra discussions on entropy, mutual information and related.

8.3.3 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

$$\mathbb{E}\left[f(X)\right] \geq f\left(\mathbb{E}\left[X\right]\right). \tag{8.76}$$

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

$$f\left(\mathbb{E}[X]\right) = f\left(x_1\beta + x_0(1-\beta)\right) \leq \beta f(x_1) + (1-\beta)f(x_0) = \mathbb{E}\left[f(X)\right], \tag{8.77}$$

where the critical inequality in the middle is simply expression of the function $f(x)$ convexity (taken verbatim from the definition).

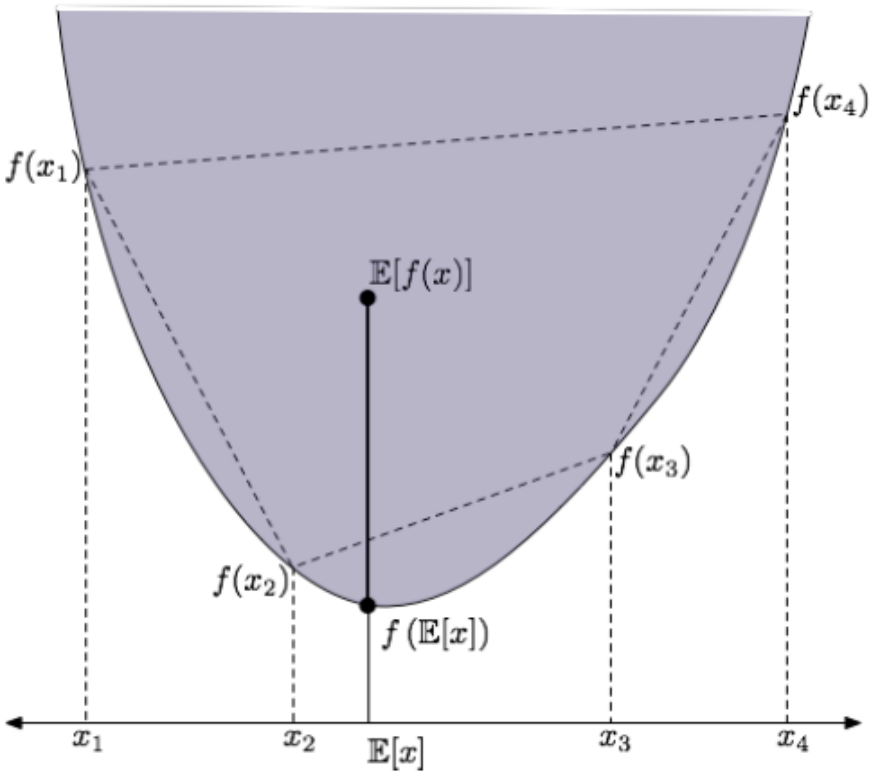


Figure 8.4

See also Fig. (8.4) with another (graphical) hint on the proof of the Jensen inequality.

In fact, the Jensen inequality holds over any spaces. Mathematically accurate proof of the Jensen inequality will be discussed in Math 527.

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]

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^{n} a_i \right) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}, \tag{8.78}$$

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 $H(P) \doteq H_P(X)$ (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 8.3.2. 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

$$\mathbf{E}[f(u)] \geq f[\mathbf{E}(u)]. \tag{8.79}$$

Let us define

$$f(u) = -\log_2 u, \quad u = 1/P(x)$$

Obviously, $f(u)$ is convex. In accordance with (8.79) 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.

Example 8.3.3. 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

$$p_n = \begin{cases} \frac{0.1}{n}, & \text{for } n \in 1, \ldots, 12367 \\ 0, & \text{for } n > 12367 \end{cases} \tag{8.80}$$

Under an assumption that English documents are generated by picking words at random according to Eq. (8.80) compute the entropy of the made-up English per word and also per character. Interpret the results

Solution. Substituting the distribution (8.80) into the definition of entropy one derives

$$H = -\sum_{i=1}^{12367} \frac{0.1}{n} \log_2 \frac{0.1}{n} \approx \frac{0.1}{\ln 2} \int_{10}^{123670} \frac{\ln x}{x} dx = = \frac{1}{20 \ln 2} (\ln^2 123670 - \ln^2 10) \approx 9.9 \text{ bits}.$$

Let us now calculate the entropy of the English per character. The resulting entropy is fairly low ~ 1 bit. Thus, 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.

Exercise 8.3.4. The joint probability distribution $P(x, y)$ of two random variables X and Y is described in Table 8.1. Calculate the marginal probabilities $P(x)$ and $P(y)$, conditional probabilities $P(x|y)$ and $P(y|x)$, marginal entropies $H(X)$ and $H(Y)$, as well as the mutual information $I(X;Y)$.

$P(x,y)$		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.

Chapter 9

Stochastic Processes

9.1 Markov Chains [discrete space, discrete time]

9.1.1 Transition Probabilities

So far we have studied random variables and events often assuming that these are i.i.d. = independent identically distributed. However, in real world we "jump" from one random state to another so that the consecutive states are dependent. The memory may last for more than one jump, however there is also a big family of interesting random processes which do not have long memory - only current state influences where we jump to. This is the class of random processes described by Markov Chains (MCs).

Before we define a Markov chain, it is a good idea to watch the introductory video, which explains the origin of Markov chains and briefly describes what they are.

A Markov chain p is a stochastic process with no memory other than its current state. We can think of a Markov chain as a random walk on a directed graph, where vertices correspond to states and edges correspond to transitions between states. Each edge $i \to j$ is associated with the probability $p(j \leftarrow i)$ of going from the state i to the state j. A useful interactive playground can be found here.

MCs can be explained in terms of directed graphs, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where the set of vertices, $\mathcal{V} = (i)$, is associated with the set of states, and the set of directed edges, $\mathcal{E} = (j \leftarrow i)$, 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 make description complete we need to associate to each vertex a transition probability, $p_{j \leftarrow i} = p_{ji}$ from the state i to the state j. Since p_{ji} is the probability, $\forall (j \leftarrow i) \in \mathcal{E}: \quad p_{ji} \geq 0$, and

$$\forall i: \quad \sum_{j:(j \leftarrow i) \in \mathcal{E}} p_{ji} = 1. \tag{9.1}$$

Then, the combination of $\mathcal{G}$ and $p \doteq (p_{ji}|(j \leftarrow i) \in \mathcal{E})$ 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 in the following stationary Markov chains, i.e. these with p_{ji} constant - not changing in time. However, for many of the following statements/considerations generalization to the time-dependent processes is straightforward.

MC generates a random (stochastic) dynamic process. Time flows continuously, however as a matter of convenient abstraction we consider discrete times (and sometimes, actually quite often, events do happen discreetly). 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.

How does one relates the directed graph with weights (associated to the transition probabilities) to samples? The relation, actually, has two sides. The direct one - is about how one generates samples. The samples are generated by advancing the trajectory from the current-time state flipping coin according to the transition probability p_{ij}. The inverse side is about reconstructing characteristics of Markov chain from samples or verifying if the samples where indeed generated according to (rather restrictive) MC rules.

Now let us get back to the direct problem where a MC is described in terms of $(\mathcal{V}, \mathcal{E}, p)$. However, instead of characterizing the system in terms of the trajectories/paths/samples, we can pose the question following evolution of the "state probability vector", or simply the "state vector":

$$\forall i \in \mathcal{V}, \quad \forall t = 0, \cdots: \quad \pi_i(t+1) = \sum_{j:(i \leftarrow j) \in \mathcal{E}} p_{ij} \pi_j(t). \tag{9.2}$$

Here, $\pi(t) \doteq (\pi_i(t) \geq 0 | i \in \mathcal{V})$ is the vector built of components each representing probability for the system to be in the state i at the moment of time t. Thus, $\sum_{i \in \mathcal{V}} \pi_i = 1$. We can also rewrite Eq. (9.2) in the vector/matrix form

$$\pi(t+1) = p\pi(t), \tag{9.3}$$

where $\pi(t)$ the column/state and $p(t)$ is the transition-probability matrix, which satisfies the so-called "stochasticity" property (9.1), to preserve the total probability.

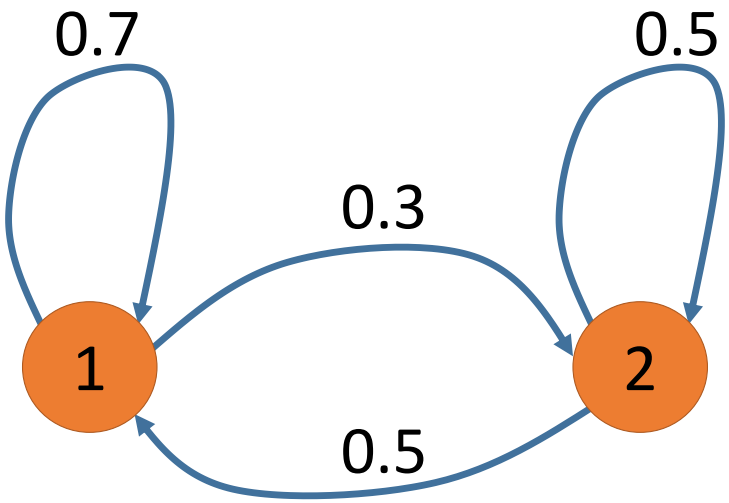


Figure 9.1: An exemplary Markov Chain (MC).

Definition 9.1.1. A matrix is called stochastic if all of its components are nonnegative and each column sums to 1.

Sequential application of Eq. (9.3) results in

$$\pi(t+k) = p^k \pi(t), \tag{9.4}$$

and we are interested to analyze properties of p^k, characterizing the Markov chain acting for k sequential periods.

Let us first study it on example of the simple MC illustrated in Fig. (9.1). In this case, p^k is 2×2 matrix which dependence on k as follows

$$p^1 = \begin{pmatrix} 0.7 & 0.5 \\ 0.3 & 0.5 \end{pmatrix}, \quad p^2 = \begin{pmatrix} 0.64 & 0.6 \\ 0.36 & 0.4 \end{pmatrix}, \quad p^{10} \approx p^{100} \approx \begin{pmatrix} 0.625 & 0.625 \\ 0.375 & 0.375 \end{pmatrix}. \tag{9.5}$$

9.1.2 Properties of Markov Chains

Definition 9.1.2 (Irreducability of MC)**.** MC is **irreducible** if one can access any state from any state, formally

$$\forall i, j \in \mathcal{V}: \quad \exists k > 1, \quad \text{s.t.} \quad (p^k)_{ij} > 0. \tag{9.6}$$

Example of Eq. (9.5) is obviously irreducible. However, if we replace $0.3 \to 0$ and $0.7 \to 1$ the MC becomes reducible – state 1 is not accessible from 2.

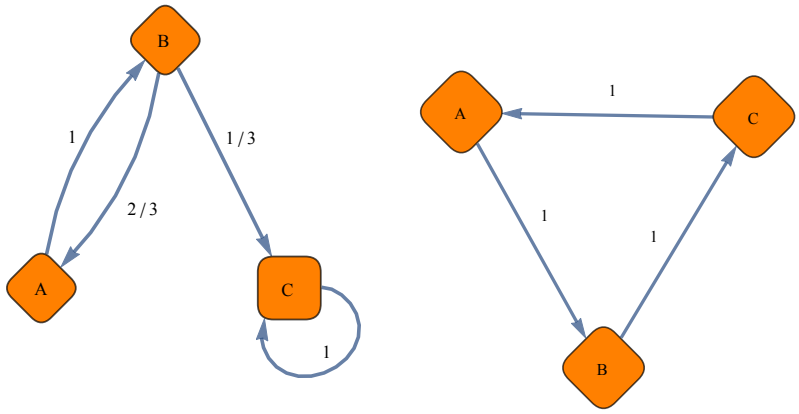


Figure 9.2: Some examples of Markov chains.

Definition 9.1.3 (Aperiodicity of MC)**.** A state i has period k if any return to the state must occur in multiples of k. Formally the period of state k is

$$k = \text{greatest common divisor}\left\{n > 0 : \text{Prob}(x_n = i | x_0 = i) > 0\right\},$$

provided that the set is not empty (otherwise the period is not defined). If $k = 1$ than the state is **aperiodic**. MC is **aperiodic** if all states are 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. Example #1 is obviously aperiodic. However, it becomes periodic with period two if the two self-loops are removed.

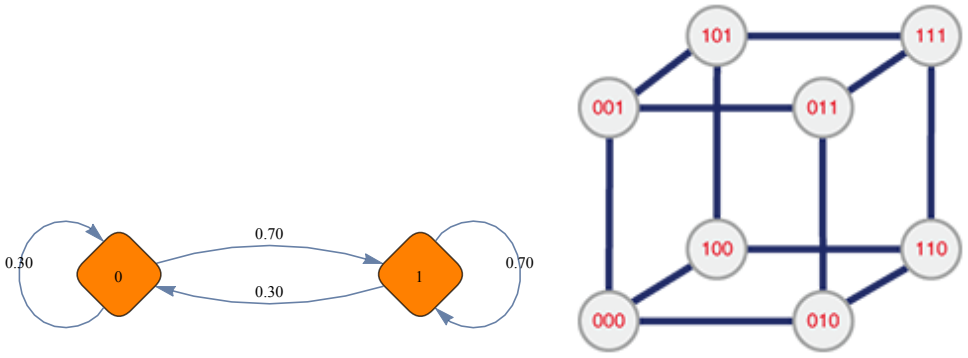
Exercise 9.1.1 (Not graded.)**.** Consider two MC examples shown in Fig. 9.2. Are these MC reducible or irreducble? Periodic or aperiodic?

Definition 9.1.4. A state i is said to be transient if, given that we start in state i, there is a non-zero probability that we will never return to i. State i is recurrent if it is not transient. State i is **positive-recurrent** if the expected return time (to the state) is positive.

Notice that the positivity is an important feature for analysis of MC over infinite graphs.

Exercise 9.1.2 (not graded)**.** 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.

Definition 9.1.5 (Ergodicity of MC)**.** A state is **ergodic** if the state is aperiodic and positive-recurrent. If all states in an irreducible MC are ergodic then the MC is ergodic. A



(a) Bernoulli Markov chain (b) Walking on Hypercube

Figure 9.3: Illustration of sampling.

MC is ergodic 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. (9.5) $k_* = 2$.

Note, that there are other (alternative) descriptions of ergodicity. A particularly intuitive one is: **the MC is ergodic if it is aperiodic and irreducible**. Notice that if we replace positive-recurrence by irreducibility in the definition of ergodicity, the ergodicity still holds. However, the combination of irreducibility and positive-recurrence does not guarantee ergodicity. In this course we will not go into related mathematical formalities and details, largely considering generic, i.e. ergodic, MC.

Practical consequences of the ergodicity are that the steady state is unique and it is universal. Universality means that the steady state does not depend on the initial condition.

9.1.3 Sampling

As already mentioned above MC are widely used to generate samples of some distribution. One can imagine a particle which travels over a graph according to edges' weights. After some time (for ergodic chain) the probability distribution of a particle becomes stationary (one say that the chain is mixed) and then the trajectory of the particle will represent the sample of a distribution. Analyzing the trajectory you can say a lot about distribution, e.g. calculate moments and expectation values of functions.

In the Figure 9.3a we see a Markov chain which corresponds to the Bernoulli distribution with probability of success equal to 0.7. More complicated example is shown in the Figure 9.3b. Imagine that you need to generate a random string of n bits. There is 2^n possible configurations. You can organize these configurations in a hypercube graph. The

hypercube has 2^n vertices and each vertex has n neighbors, corresponding to the strings that differ from it at a single bit. 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.

9.1.4 Steady State Analysis

Theorem 9.1.6 (Existence of Stationary Distribution)**.** Component-wise positive, normalized, π^*, is called stationary distribution (invariant measure) if

$$\pi^* = p\pi^* \tag{9.7}$$

An irreducible MC has a stationary distribution iff all of its states are positive recurrent.

Proof of this (and other statements used in this Section) will be discussed in Math 527. Solving Eq. (9.7) for the example of Eq. (9.5) one finds

$$\pi^* = \begin{pmatrix} 0.625 \\ 0.375 \end{pmatrix}, \tag{9.8}$$

which is naturally consistent with Eq. (9.5). In general,

$$\pi^* = \frac{e}{\sum_i e_i}, \tag{9.9}$$

where e is the eigenvector with the eigenvalue 1. And how about other eigenvalues of the transition matrix?

9.1.5 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

$$p = U^{-1}\Sigma U \tag{9.10}$$

where $\Sigma = \mathrm{diag}(\lambda_1, \cdots, \lambda_n)$, $1 = |\lambda_1| > \lambda_2 \geq |\lambda_3| \geq \cdots |\lambda_n|$ and U is the matrix of eigenvectors (each normalized to having an l_2 norm equal to 1) where each raw is a right eigenvector of p. Then

$$\pi^{(k)} = p^k\pi = (U^{-1}\Sigma U)^k\pi_0 = U^{-1}\Sigma^k U\pi_0. \tag{9.11}$$

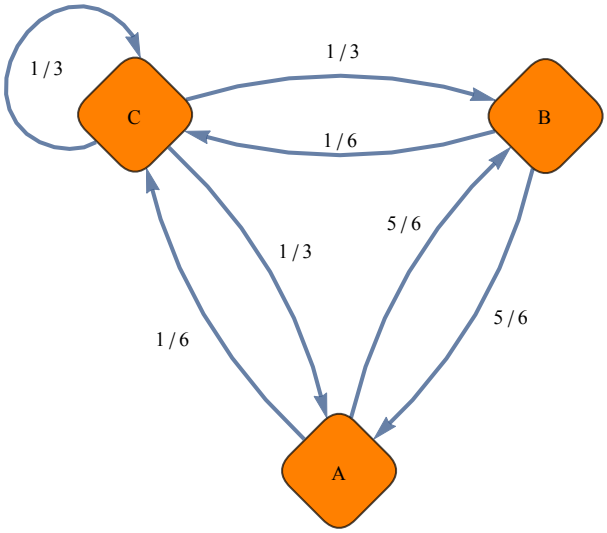


Figure 9.4: Illustration of the Detailed Balance (DB).

Let us represent p_0 as an expansion over the normalized eigenvectors, $u_i, \cdots i = 1, \cdots, n$:

$$\pi = \sum_{i=1}^{n} a_i u_i. \tag{9.12}$$

Taking into account orthonormality of the eigenvectors one derives

$$\pi^{(k)} = \lambda_1 \left(a_1 u_1 + a_2 \left(\frac{\lambda_2}{\lambda_1} \right)^k u_2 + \cdots + a_n \left(\frac{\lambda_n}{\lambda_1} \right)^k u_n \right) \tag{9.13}$$

Since $\pi^{(k)}_{k\to\infty} \to \pi^* = u_1$, the second term on the rhs of Eq. (9.13) describes the rate of convergence of $\pi^{(k)}$ to the steady state. The convergence is exponential in $\log(\lambda_1/\lambda_2)$.

Example 9.1.3. Find eigen-values for the MC shown in Fig. (9.4) with the transition matrix

$$p = \begin{pmatrix} 0 & 5/6 & 1/3 \\ 5/6 & 0 & 1/3 \\ 1/6 & 1/6 & 1/3 \end{pmatrix}. \tag{9.14}$$

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

$$\pi(t) = p^t \pi(0). \tag{9.15}$$

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

$$p\pi^* = \pi^*. \tag{9.16}$$

Thus, π^* is an eigenvector of p with eigenvalue 1 with all components positive and normalized. The matrix (9.14) has three eigenvalues $\lambda_1 = 1, \lambda_2 = 1/6, \lambda_3 = -5/6$ and corresponding eigenvectors are

$$\pi^* = \left(\frac{2}{5}, \frac{2}{5}, \frac{1}{5}\right)^T,\ u_2 = \left(-\frac{1}{2}, -\frac{1}{2}, 1\right)^T,\ u_3 = (-1, 1, 0)^T. \tag{9.17}$$

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

$$\pi(0) = \pi^* - \frac{u_2}{5} - \frac{u_3}{2}, \tag{9.18}$$

and then

$$\pi(t) = p^t\pi(0) = \pi^* - \frac{\lambda_2^t}{5}u_2 - \frac{\lambda_3^t}{2}u_3. \tag{9.19}$$

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 9.1.7 (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.

9.1.6 Reversible & Irreversible Markov Chains.

MC is called **reversible** if there exists π s.t.

$$\forall\{i, j\} \in \mathcal{E}: \quad p_{ji}\pi_i^* = p_{ij}\pi_j^*, \tag{9.20}$$

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

$$Q \doteq (Q_{ji} = p_{ji}\pi_i^* | (j \leftarrow i) \in \mathcal{E}), \tag{9.21}$$

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. (9.1)

$$Q = \begin{pmatrix} 0.7 * 0.625 & 0.5 * 0.375 \\ 0.3 * 0.625 & 0.5 * 0.375 \end{pmatrix} = \begin{pmatrix} 0.4375 & 0.1875 \\ 0.1875 & 0.1875 \end{pmatrix} \tag{9.22}$$

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.

9.1.7 Detailed Balance vs Global Balance. Adding cycles to accelerate mixing.

Note that if a steady distribution, π^*, satisfy the DB condition (9.20) 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

$$\sum_{j:(j \leftarrow i) \in \mathcal{E}} \tilde{p}_{ji} \pi_i^* = \sum_{j:(i \leftarrow j) \in \mathcal{E}} \tilde{p}_{ij} \pi_j^*. \tag{9.23}$$

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 (9.23).

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 antisymmetric 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,

$$Q_{ij} - Q_{ji} = \sum_{\alpha} J_\alpha \left(C_{ij}^\alpha - C_{ji}^\alpha \right) \tag{9.24}$$

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 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.

Exercise 9.1.4 (not graded)**.** Construct a Markov chain, which mixes in the shortest time regardless of the initial state, and which obeys the following properties. The state space contains N states, and the desired stationary distribution is such that the probability to be in a state i equals to known, p_i. What can you say about eigenvalues of the corresponding transition matrix? Construct the transition matrix explicitly.

Exercise 9.1.5 (Hardy-Weinberg Law)**.** Consider an experiment of mating rabbits. We watch evolution 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.

1) Write down the transition matrix P of the Markov chain thus defined. Is the Markov chain irreducible and aperiodic?

2) Assume that we start with a hybrid rabbit. Let μ_n be the probability distribution of the character of the rabbit of the n-th generation. In other words, $\mu_n(GG)$, $\mu_n(Gg)$, $\mu_n(gg)$ are the probabilities that the n-th generation rabbit is GG, Gg, or gg, respectively. Compute μ_1, μ_2, μ_3. Is there some kind of law?

3) Calculate P^n for general n. What can you say about μ_n for general n?

4) Calculate the stationary distribution of the MC. Does the Detailed Balance hold in this case?

9.2 Bernoulli and Poisson Processes [discrete space, discrete & continuous time]

The two processes discussed in the following are some of the simplest dynamic random processes, which are also building blocks for others. Simplicity here is related to the fact that the processes are defined with the least number of characteristics. We will focus on important features of the processes, such as memorylessness (also called Markov property), and will work out interesting (and rather general) questions one may ask (and answer).

9.2.1 Bernoulli Process: Definition

Bernoulli process is defined as a sequence of independent Bernoulli trials, i.e. at each trial $P(success) = P(x = 1) = \beta$ and $P(failure) = P(x = 0) = 1 - \beta$. The Bernoulli process can be represented as a simple MC (two nodes + two self-loops, please draw one). The sequence looks like $00101010001 = **S*S*S***S$. S here stands for "success".

Examples:

- Sequence of discrete updates – ups and downs (stock market).
- Sequence of lottery wins.
- Arrivals of buses at a station, checked every 1/5/? minutes.

9.2.2 Bernoulli: Number of Successes

As we discussed already earlier in the course, number of k successes in n steps follows the binomial distribution

$$\forall k = 0, \cdots, n: \quad P(S = k|n) = \left(\begin{array}{c} n \\ k \end{array} \right) \beta^k (1-\beta)^{n-k} \tag{9.25}$$

$$\text{mean}: \quad \mathbb{E}[S] = n\beta \tag{9.26}$$

$$\text{variance}: \quad var(S) = \mathbb{E}[(S - \mathbb{E}[S])^2] = n\beta(1-\beta) \tag{9.27}$$

Let us now discuss dynamic characteristics of the Bernoulli process.

9.2.3 Bernoulli: Distribution of Arrivals

Call T_1 the number of trials till the first success (including the success event too). The Probability Mass Function (PMF) for the time of the first success is

$$t = 1, 2, \cdots: \quad P(T_1 = t) = \beta(1-\beta)^{t-1} [\text{Geometric PMF}] \tag{9.28}$$

The answer is the product of the probabilities of $(t-1)$ failures and one success (thus memoryless). It is called geometric because checking that the probability distribution is normalized involves summing up the geometric sequence (progression). Naturally, $\sum_{t=1}^{\infty}(1-\beta)^{t-1} = 1/\beta$. Mean and variance of the geometric distribution are

$$\text{mean}: \quad \mathbb{E}[T_1] = \frac{1}{\beta} \tag{9.29}$$

$$\text{variance}: \quad var(T_1) = \mathbb{E}[(T_1 - \mathbb{E}[T_1])^2] = \frac{1-\beta}{\beta^2} \tag{9.30}$$

More on the memoryless property. Given n, the future sequence $x_{n+1}, x_{n+2}, \cdots$ is also a Bernoulli process and it is independent of the past. 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

$$P(T - n = k | T > n) = \beta(1-\beta)^{k-1} \tag{9.31}$$

And how about the k^{th} arrival? Let y_k be the number of trials until k^{th} success (inclusive), then we write

$$t = k, k+1, \cdots : \quad P(y_k = t) = \begin{pmatrix} t-1 \\ k-1 \end{pmatrix} \beta^k (1-\beta)^{t-k} [\text{Pascal PMF}] \tag{9.32}$$

$$\text{mean} : \quad \mathbb{E}[y_k] = \frac{k(1-\beta)}{\beta^2} \tag{9.33}$$

$$\text{variance} : \quad var(y_k) = \mathbb{E}[(y_k - \mathbb{E}[y_k])^2] = \frac{k(1-\beta)}{\beta^2} \tag{9.34}$$

The combinatorial factor accounts for the number of configurations of the "k arrivals in y_k trials" type.

Exercise 9.2.1 (Not graded.)**.** Define $T_k = y_k - y_{k-1}, \quad k = 2, 3, \cdots$, where thus T_k is the inter-arrival time between $k-1$-th and k-th arrivals. Write down the probability density distribution function for the k-th inter-arrival time, T_k.

9.2.4 Poisson Process: Definition

Examples:

- All examples from the Bernoulli case considered in continuous time.
- E-mail arrivals with infrequent check.
- High-energy beams collide at a high frequency (10 MHz) with a small chance of a good event (actual collision).
- Radioactive decay of a nucleus with the trial being to observe a decay within a small time interval.
- Spin flip in a magnetic field.

COVID-19 challenge: suggest an example of a Poisson process event inspired by our daily "infected" life.

Let us first recall the definition of the Poisson distribution we had in Section 8.1.1, and specifically relation between the Bernoulli distribution and the Poisson distribution.

The Poisson distribution, describing arrival of k customers in an interval (of unspecified duration) was defined as

$$\forall k \in \mathbb{Z}^* = \{0\} \bigcup \mathbb{Z}: \quad \text{Pois}(k|\tilde{\lambda}) = \frac{\tilde{\lambda}^k e^{-\tilde{\lambda}}}{k!}. \tag{9.35}$$

Then we notice that if we take the binomial distribution (9.25), describing probability of k arrivals in n intervals, with each arrival being independent with probability β, and then consider it in the limit, $n \to \infty$, $\beta \to \tilde{\lambda}/n$ we arrive at Eq. (9.35).

Now we would like to inject into consideration the notion of the time interval duration, and thus transition to continuous time. Then, the continuous time version of Eq. (9.35) describing probability density (per unit time) of getting one arrival in time t becomes

$$p_{T_1}(t) = \lambda \exp(-\lambda t), \tag{9.36}$$

where the normalization is chosen proper, i.e. $\int_0^\infty dt p_{T_1}(t) = 1$. Notice that with some minor abuse of notations we change from a dimensionless parameter $\tilde{\lambda}$ in Eq. (9.35) to λ in Eq. (9.36) where the latter has the dimension of the inverse time $[\lambda] = [1/t]$.

9.2.5 Poisson: Arrival Time

Then the probability density of the first arrival in time t is

$$P(T_1 \leq t) = \int_0^t dt' p_{T_1}(t') = 1 - \exp(-\lambda t).$$

By extension (generalizing), for the probability density of time of the k^{th} arrival one derives

$$p_{T_k}(t) = \frac{\lambda^k t^{k-1} \exp(-\lambda t)}{(k-1)!}, \quad t > 0 \quad \text{(Erlang "of order" k)}$$

Like Bernoulli, the Poisson process show the following two key properties

- **Fresh Start Property**: the time of the next arrival is independent of the past
- **Memoryless property:** suppose we observe the process for t seconds and no success occurred. Then the density of the remaining time of arrival is exponential.

Summary of the relations between the Bernoulli process and the Poisson process is summarized in the table

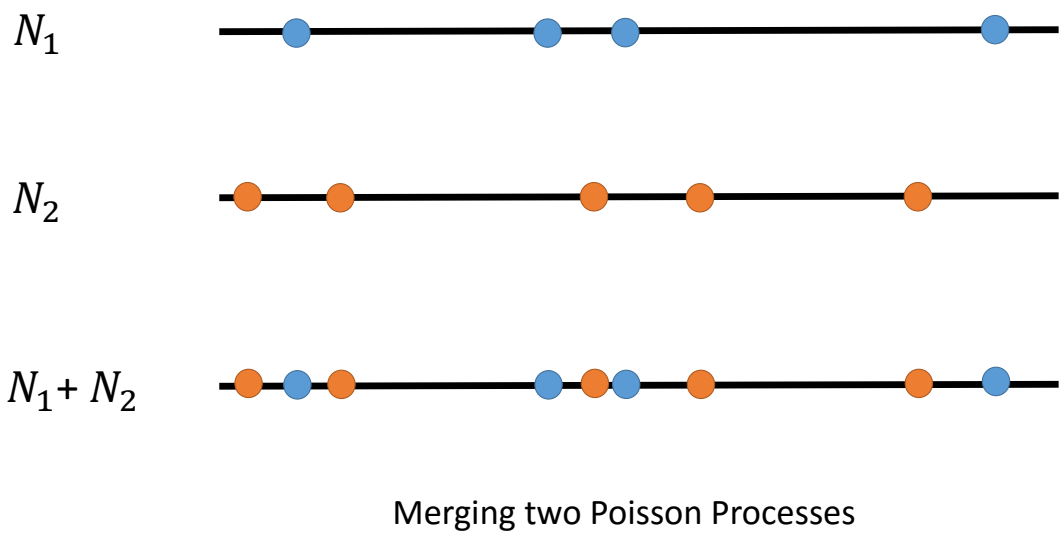


Figure 9.5: Merging two Poisson processes.

	Bernoulli	Poisson
Times of Arrival	Discrete	Continuous
Arrival Rate	p/per trail	λ/unit time
PMF of Number of arrivals	Binomial	Poisson
PMF of Interarrival Time	Geometric	Exponential
PMF of k^{th} Arrival Time	Pascal	Erlang

9.2.6 Merging and Splitting Processes

Most important feature shared by 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.

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. (9.5). 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 have an independent increment. 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 9.2.2. 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.

(1) What is the probability to observe at least one large meteor next year?

(2) What is the probability of observing no meteor hits within the next 1000 years?

(3) Calculate the probability distribution $P(t_n)$, where the random variable t_n represents the appearance time of the n-th meteor.

Solution:
The probability of observing n meteors in a time interval t is given by

$$P(n,t) = \frac{(\lambda t)^n}{n!} e^{-\lambda t}, \tag{9.37}$$

where $\lambda = 0.001$ (events per year) is the average hitting rate.

(1) $P(n > 0 \text{ meteors next year}) = 1 - P(0,1) = 1 - e^{-0.001} \approx 0.001$.

(2) $P(n = 0 \text{ meteors next 1000 years}) = P(0,1000) = e^{-1} \approx 0.37$.

(3) It is intuitively clear that

$$(\text{probability that } t_n > t) = (\text{probability to get } n-1 \text{ arrivals in interval } [0,t]),$$

Therefore

$$\int_t^\infty p(t_n)dt_n = P(n-1,t),$$
$$p(t_n) = \frac{\lambda^n t_n^{n-1}}{(n-1)!}e^{-\lambda t_n}.$$

Exercise 9.2.3. Customers arrive at a store with the Poisson rate of 10 per hour. 40%/60% of arrivals are males/females.

(1) Compute probability that at least 20 customers have entered between 10 and 11 am.

(2) Compute probability that exactly 10 woman entered between 10 and 11 am.

(3) Compute the expected inter-arrival time of men.

(4) Compute probability that there are no male customers between 2 and 4 pm.

9.3 Space-time Continuous Stochastic Processes

In this lecture we discuss stochastic dynamics of continuous variables governed by the Langevin equation. We discuss how to derive the so-called Fokker-Planck equations, describing temporal evolution of the probability of a state. We then go into some additional details for a basic example of stochastic dynamics in a free space (no potential) describing the Brownian motion where Fokker-Planck equations becomes the diffusion equation.

9.3.1 Langevin equation in continuous time and discrete time

Stochastic process in 1d is described in the continuous-time and discrete-time forms as follows

$$\dot{x} = -F(x) + \sqrt{D}\xi(t), \quad \langle\xi(t)\rangle = 0, \quad \langle\xi(t_1)\xi(t_2)\rangle = \delta(t_1 - t_2) \tag{9.38}$$
$$x_{n+1} - x_n = -\Delta F(x_n) + \sqrt{D\Delta}\xi(t_n), \quad \langle\xi(t_n)\rangle = 0, \quad \langle\xi(t_n)\xi(t_k)\rangle = \delta_{kn} \tag{9.39}$$

The first and second terms on the rhs of Eq. (9.38) stand for the force and the "noise" respectively. The noise is considered independent at each time step. These equations, also called Langevin equations, describe evolution of a "particle" positioned at $x \in \mathbb{R}$. The two terms on the rhs of Eq. (9.38)) correspond to deterministic 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. (9.39) may seem mysterious, let us clarify its origin on basic (no force/potential) example of $F(x) = 0$. (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 this case in

$$\forall t \geq 0 : \quad x(t) = \int_0^t dt' \xi(t'), \tag{9.40}$$

$$\forall t \geq 0 \quad \langle x^2(t) \rangle = \int_0^t dt_1 \int_0^t dt_2 D\delta(t_1 - t_2) = D \int_0^t dt_1 = Dt, \tag{9.41}$$

where we also set $x(0) = 0$. Infinitesimal version of Eq. (9.41) is

$$\delta x = \sqrt{D\Delta}, \tag{9.42}$$

which is thus the Brownian (no force) version of Eq. (9.39).

9.3.2 From the Langevin Equation to the Path Integral

The Langevin equation can also be viewed as relating the change in $x(t)$, i.e. dynamics of interest, to stochastic dynamics of the δ-correlated source $\xi(t_n) = \xi_n$ characterized by the Probability Density Function (PDF)

$$p(\xi_1, \cdots, \xi_N) = (2\pi)^{-N/2} \exp\left(-\sum_{n=1}^{N} \frac{\xi_n^2}{2} \right) \tag{9.43}$$

Eqs. (9.38,9.39,9.43) are 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, e.g. through the telegraph process.

Let us express ξ_n via x_n from Eq. (9.39) and substitute it into Eq. (9.43)

$$p(\xi_1, \cdots, \xi_{N-1}) \to p(x_1, \cdots, x_N) = (2\pi D)^{-(N-1)/2} \exp\left(-\frac{1}{2D\Delta} \sum_{n=1}^{N-1} (x_{n+1} - x_n + \Delta F(x))^2 \right) \tag{9.44}$$

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)

$$p\{x(t)\} \propto \exp\left(-\frac{1}{2D} \int_0^T dt \, (\dot{x} + F(x))^2 \right) \tag{9.45}$$

This object is called (in physics and math) "path integral" and/or Feynmann/Kac integral.

9.3.3 From the Path Integral to the Fokker-Plank (through sequential Gaussian integrations)

Probability Density Function 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 probability (density) distribution (of the path) conditioned to being at the initial position, x_1, at the moment of time t_1, $p(x_2, \cdots, x_N|x_1)$, and also from the prior/initial (distribution) $p_1(x_1)$ – both assumed known:

$$p_N(x_N) = \int dx_1 \cdots dx_N p(x_2, \cdots, x_N|x_1) p_1(x_1). \tag{9.46}$$

It is convenient to derive relation between $p_N(\cdot)$ and $p_1(\cdot)$ in steps, i.e. through a recurrence, integrating over $dx_1, \cdots, dx_N$ sequentially. Let us proceed analyzing the case of the Brownian motion where, $F = 0$. Then the first step of the induction becomes

$$p_2(x_2) = (2\pi D)^{-1/2} \int dx_1 \exp\left(-\frac{1}{2D\Delta}(x_2 - x_1)^2\right) P_1(x_1) \tag{9.47}$$

$$= (2\pi D)^{-1/2} \int d\epsilon \exp\left(-\frac{\epsilon^2}{2D\Delta}\right) P_1(x_2 - \epsilon) \tag{9.48}$$

$$\approx (2\pi D)^{-1/2} \int d\epsilon \exp\left(-\frac{\epsilon^2}{2D\Delta}\right) \left(p_1(x_2) - \epsilon \partial_x p_1(x_2) + \frac{\epsilon^2}{2}\partial_x^2 p_1(x_2)\right) \tag{9.49}$$

$$= p_1(x_2) + \Delta \frac{D}{2} \partial_x^2 p_1(x_2), \tag{9.50}$$

where transitioning from Eq. (9.48) to Eq. (9.49) 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

$$\partial_t p_t(x) = \frac{D}{2} \partial_x^2 p_t(x). \tag{9.51}$$

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 $U(x)$ term (derivation is straightforward) one arrives at the Fokker-Planck equation, generalizing the zero-force diffusion equation

$$\partial_t p_t(x) - \partial_x (U(x) p_t(x)) = \frac{D}{2} \partial_x^2 p_t(x). \tag{9.52}$$

9.3.4 Analysis of the 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. [17].

The Fokker-Planck equation (9.52) 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 - "diffusion" and "advecton" - the terminology is from fluid mechanics - indeed not only fluids but also probabilities can flow. The flow of probability is in the phase space. The diffusion originates from the stochastic source, while 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. (9.52) can be represented in the 'flux' form:

$$\partial_t p_t + \partial_x J_t(x) = 0 \tag{9.53}$$

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. (9.53) has a gradient form, corresponds to the global conservation of probability. Indeed, integrating Eq. (9.53) 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, $\partial_t \int dx p_t(x) = 0$. In the steady state, when $\partial_t p_t = 0$ 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, $p_{t=0}(x)$ is known, $p_t(x)$ for any consecutive t is well defined, in the sense that the Fokker-Planck is the Cauchi (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 $F = 0$. Another example, principal for the so-called 'equilibrium statistical physics', is of the potential force $F = \partial_x U(x)$, where $U(x)$ is a potential. Think, for example about x representing a particle connected to the origin by a spring. $U(x)$ is the potential/energy stored within the spring. In this case of the gradient force the stationary (i.e. time-independent) solution of the Fokker-Planck Eq. (9.52) can be found explicitly,

$$p_{st}(x) = Z^{-1} \exp\left(-\frac{U(x)}{D}\right). \tag{9.54}$$

This solution is called Gibbs distribution, or equilibrium distribution.

Brownian Motion

Example 9.3.1. Consider 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 (9.38) in this case becomes

$$\frac{dx}{dt} + \gamma x = \sqrt{D}\xi(t), \quad \langle \xi(t) \rangle = 0, \quad \langle \xi(t_1)\xi(t_2) \rangle = \delta(t_1 - t_2) \tag{9.55}$$

Write a formal solution of Eq. (9.55) 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 Fokker-Planck (FP) equation for $n(t, x)$, and solve it for the initial condition, $n(t, 0) = \delta(x)$.

Solution:

Eq. (9.55) has the formal solution

$$x(t) = x(0)e^{-\gamma t} + \int_0^t \xi(t')e^{-\gamma(t-t')}dt'. \tag{9.56}$$

For simplicity, we assume $x(0) = 0$. Then $\mathbb{E}[x(t)] = 0$ and

$$\begin{aligned} \langle x^2(t) \rangle &= \int_0^t \int_0^t dt'dt'' \langle \xi(t')\xi(t'') \rangle e^{-\gamma(t-t')}e^{-\gamma(t-t'')} \\ &= 2De^{-2\gamma t} \int_0^t \int_0^t dt'dt'' \delta(t' - t'')e^{\gamma(t'+t'')} = \frac{D}{\gamma}(1 - e^{-2\gamma t}). \end{aligned} \tag{9.57}$$

At small time scale $t \ll 1/\gamma$ we deal with usual diffusion $\langle x^2(t) \rangle \simeq 2Dt$, since the particle does not feel the potential, while at larger time scale $t \gg 1/\gamma$ the dispersion saturates, $\langle x^2(t) \rangle \simeq D/\gamma$.

The Fokker-Planck equation, $\partial_t n = (\gamma \partial_x x + D\partial_x^2)n$, should be supplemented by the initial condition $n(0,x) = \delta(x)$. Then, the solution (the Green function) is

$$n(t,x) = \frac{1}{\sqrt{2\pi\langle x^2(t)\rangle}} \exp\left[-\frac{x^2}{2\langle x^2(t)\rangle}\right]. \tag{9.58}$$

The meaning of the expression is clear: the probability function $n(t,x)$ is Gaussian, but the dispersion is time-dependent.

Exercise 9.3.2 (High order moments, not graded)**.** Prove that the moments $\mathbb{E}\left[x^{2k}(t)\right]$ for the Brownian motion in $\mathbb{R}^1$ obey the following recurrent equation

$$\partial_t \langle x^{2k}\rangle = 2k(2k-1)D\langle x^{2(k-1)}\rangle. \tag{9.59}$$

Solve this equation for a particle starting from $x = 0$ at $t = 0$.

Exercise 9.3.3 (Brownian motion in parabolic potential, not graded)**.** The concentration field, $n(t,x) : \mathbb{R}_+ \times \mathbb{R} \to \mathbb{R}_+$, for a Brownian particle in the potential, $U(x) = \alpha x^2/2$, is described by the advection-diffusion equation

$$D\partial_x^2 n + \alpha \partial_x (xn) = \partial_t n. \tag{9.60}$$

Write down stochastic ODE for the underlying stochastic process, $x(t) : \mathbb{R} \to \mathbb{R}_+$, and, given the initial condition for the concentration field, $n(0,x) = \delta(x)$, compute respective statistical moments $\langle x^k(t)\rangle$. [Hint: Reconstruct stochastic ODE correspondent to the PDE (9.60) and then follow the logic/strategy of Example 9.3.1.]

Exercise 9.3.4 (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 plane xy with fixed speed v_0. The Cartesian components of the particle velocity v_x, v_y in the polar coordinates are

$$v_x = v_0 \cos\varphi, \qquad v_y = v_0 \sin\varphi, \tag{9.61}$$

where the polar angle φ defines the direction of motion. Assume that φ evolves according to the stochastic equation

$$\frac{d\varphi}{dt} = \xi, \tag{9.62}$$

where $\xi(t)$ is the Gaussian white noise with zero mean and pair correlator $\langle \xi(t_1)\xi(t_2)\rangle = 2D\delta(t_1 - t_2)$. The initial condition are chosen to be $\varphi(0) = 0$, $x(0) = 0$ and $y(0) = 0$.

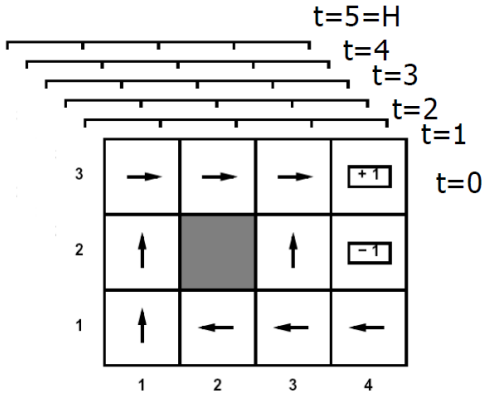


Figure 9.6: Optimal solution set of actions (arrows) for each state, for each time.

- Calculate $\langle x(t)\rangle$, $\langle y(t)\rangle$.
- 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.)

Solving MDP means finding optimal a, i.e. set of actions for each state at each moment of time, as illustrated on the GridWorld example (to be discussed next) in Fig. 9.6.

Our description here is intentionally terse/introductory. For a more colloquial, detailed and mathematical exposition of MDP check the lecture notes of Pieter Abbeel (UC Berkeley) `http://www.cs.berkeley.edu/~pabbeel/cs287-fa12/slides/mdps-exact-methods.pdf` from the Berkley AI course. In fact, the Berkeley course on AI also contains a very good repository of materials at `http://ai.berkeley.edu/lecture_videos.html`. Our running 'Grid World' example/illustration of MDP (comes next) is used intensively in the lecture series, see `http://aima.cs.berkeley.edu/demos.html` and also `http://www2.hawaii.edu/~chenx/ics699rl/grid/`.

9.3.5 MDP: Grid World Example

MDP can be considered as an interactive probabilistic game one plays against computer (random number generator). The game consists in defining transition rates between the states to achieve certain objectives. Once optimal (or suboptimal) rates are fixed the implementation becomes just a Markov Process we have studied already.

Let us play this 'Grid World' game a bit. The rules are introduced in Fig. (9.7). An agent lives on the grid (3×4). Walls block the agent's path. The agent actions do not always go as planned: 80% of time the action 'North' take the agent 'North' (if there is no

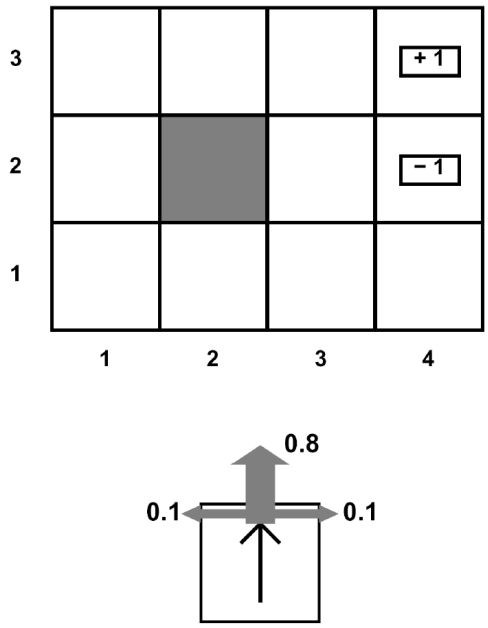


Figure 9.7: Canonical example of MDP from 'Grid World' game.

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. We will come to this example many times during this lecture.

We will consider the following *Value Iteration* algorithm [1]:

Algorithm 5 MDP – Value Iteration

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 $i = 0, \cdots, H-1$ **do**

$\forall s: \quad V_{i+1}^*(s) \leftarrow \max_a \sum_{s'} P(s'|s,a)\left[R(s,a,s') + \gamma V_i^*(s')\right]$ – [Bellman update/back-up] – the expected sum of rewards accumulated when starting from state s and acting optimally for a horizon of $i+1$ steps

end for

The Grid World implementation of the algorithm is illustrated in Fig. (9.8).

[1] The algorithm is justified through a standard Dynamic Programming arguments, of the type discussed above.

Figure 9.8: Value Iteration in Grid World.

9.3.6 Recitation. Dynamic Programming.

Chapter 10

Elements of Inference and Learning

??

10.1 Exact and Approximate Inference and Learning

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, $y_i \sim \rho(y) = \exp(-y)$ – 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:

$$x \sim \rho_0(x) = \begin{cases} e^{-x} & x > 0, \\ 0 & \text{otherwise} \end{cases}$$

- Second, aiming to get a sample from the positive half of Gaussian,

$$x \sim \rho_0(x) = \begin{cases} \sqrt{2/\pi}\exp(-x^2/2) & x > 0, \\ 0 & \text{otherwise} \end{cases}$$

, 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, $\max(\rho(x)/\rho_0(x)) = \sqrt{2/\pi}\, e^{1/2} \approx 1.32$, 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, $\rho(x)$, i.e. $\int dx\rho(x)f(x)$, in the regime where $f(x)$ and $\rho(x)$ are concentrated around very different x. In this case the overlap of $f(x)$ and $\rho(x)$ is small and as a result a lot of MC samples drawn from $\rho(x)$ 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 $\rho(x)$ to $\rho_a(x)$ and then utilizing the following obvious formula

$$\mathbb{E}_\rho[f(x)] = \int dx\rho(x)f(x) = \int dx\rho_a(x)\frac{f(x)\rho(x)}{\rho_a(x)} = \mathbb{E}_{\rho_a}\left[\frac{f(x)\rho(x)}{\rho_a(x)}\right]$$

See the IJulia notebook associated with this lecture contrasting DS example, $\rho(x) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)$ and $f(x) = \exp\left(-\frac{(x-4)^2}{2}\right)$, with IS where the choice of the proposal distribution is, $\rho_a(x) = \frac{1}{\sqrt{\pi}}\exp\left(-(x-2)^2\right)$. 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, $\rho_a(x)$, right. One way of fixing this problem is to search for good $\rho_a(x)$ 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 Brut-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 $x \doteq (x_i | i = 1, \cdots, N)$:

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 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 the Bias 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

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 $P(x_i|x^{(d)}) \doteq \sum_{x \setminus x_i; x^{(I)} = x^{(d)}} P(x)$ 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 $P(x)$.

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.

In general efforts of the partition function oracle are exponential. However in some special cases the partition function can be computed efficiently (polynomially in the number of steps). For example this is the case for (glassy) Ising model without magnetic field over planar graph. See the report and references there in for details.

Ising Model

Let us digress and consider the Ising model which is, in fact, an example of a larger class of important/interesting multi-variate statistics often referred to (in theoretical engineering) as Graphical Models (GM). We will study GM later in the course. Consider a system of spins or pixels (binary variables) on a graph, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is a set of nodes/vertices and $\mathcal{E}$ is the set of edges. The graph may be 1d chain, tree, 2d lattice ... or any other graph. (The cases of regular lattices are prevalent in physics, while graphs of a relevance to various engineering disciplines are, generally, richer.) Consider binary variables, residing at every node of the graph, $\forall i \in \mathcal{V}: \ \sigma_i = \pm 1$, we call them "spins". If there are N spins in the system, 2^N is the number of possible configuration of spins — notice exponential scaling with N, meaning, in particular that just counting the number of configurations is "difficult". If we are able to do the counting in an algebraic/polynomial number of steps, we

would call it "easy", or rather "theoretically easy", while the practically easy case - which is the goal – would correspond to the case when "complexity" of, say counting, would be $O(N)$ - linear in N at $N \to \infty$. (Btw $o(N)$ is the notation used to state that the behavior is actually slower than $O(N)$, say $\sim \sqrt{N}$ at $N \to \infty$, i.e. asymptotically $o(N) \ll O(N)$.) In magnetism (field of physics where magnetic materials are studied) probability of a spin configuration (vector) is

$$p(\sigma) = \frac{\exp\left(-\beta E(\sigma)\right)}{Z}, \quad E(\sigma) = -\frac{1}{2}\sum_{\{i,j\}\in\mathcal{E}} \sigma_i J_{ij} \sigma_j + \sum_{i\in\mathcal{V}} h_i \sigma_i, \tag{10.2}$$

$$Z = \sum_{\sigma} \exp\left(-\beta E(\sigma)\right). \tag{10.3}$$

$E(\sigma)$ is the energy of a given spin configuration, σ. The first term in $E(\sigma)$ is pair-wise (wrt nodal spins), spin exchange/interaction term. The last term in $E(\sigma)$ stands from (potentially node dependent) contribution of the magnetic field, $h = (h_i | i \in \mathcal{V})$ on individual spins. Z is the partition function, which is the weighted sum of the spin configurations. Formally the partition function is just the normalization condition introduced to enforce, $\sum_\sigma p(\sigma) = 1$. For a general graph with arbitrary values of J and h, Z is the difficult object to compute, i.e. complexity of computing Z is $O(2^N)$. (Notice that for some special cases, such as the case of a tree, or when the graph is planar and $h = 0$, computing the partition function becomes easy.) Moreover, computing other important characteristics, such as the most probable configuration of spins

$$\sigma_{ML} = \arg\max_{\sigma} p(\sigma), \tag{10.4}$$

also called Maximum Likelihood and Ground State in information sciences and physics respectively, or the (so-called marginal) probability of observing a particular node in the state σ_i (can be $+$ or $-$)

$$p_i(\sigma_i) = \sum_{\sigma\setminus\sigma_i} p(\sigma), \tag{10.5}$$

are also difficult problems. (Wrt notations – $\arg\max$ - pronounced argmax - stands for particular σ at which the maximum in Eq. (10.4) is reached. $\sigma \setminus \sigma_i$ in the argument of the sum in Eq. (10.5) means that we sum over all σ consistent with the fixed value of σ_i at the node i.)

Exercise 10.1.1. Consider Ising model on a square, 4×4 lattice, construct (write down on paper and code) and compare performance of two direct sampling algorithms, one by rejection and also the decimation algorithm (6).

10.1.2 Markov-Chain Monte-Carlo

Markov Chain Monte Carlo (MCMC) methods belong to the class of algorithms for sampling from a probability distribution 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 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 & Metropolis-Hastings.

Gibbs Sampling

Assume that the direct sampling is not feasible (because there are too many variables and computations are of "exponential" complexity — more on this latter). The main point of the Gibbs sampling is that given a multivariate distribution it is simpler to sample from a conditional distribution than to marginalize by integrating over a joint distribution. Then we create a chain: start from a current sample of the vector x, pick a component at random, compute probability for this component/variable conditioned to the rest, and sample from this conditional distribution. (The conditional distribution is for a simple component and thus it is easy.) We continue the process till convergence, which can be identified (empirically) by checking if estimation of the histogram or observable(s) stopped changing.

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

Example 10.1.2. Describe Gibbs sampling for example of a general Ising model. Build respective Markov chain. Show that the algorithm obeys Detailed Balance.

Solution: Starting from a state we pick a random node i and compare two candidate states, ($s_i = 1$ and $s_i = -1$). Then we calculate the corresponding conditional (all spins except i are fixed) probabilities p_+ and p_- , following $p_+ + p_- = 1, \quad p_+/p_- = e^{-\beta\Delta E}$, where ΔE is the energy difference between the two configurations. Next, one accepts the configuration $s_i = 1$ with the probability p_+ or the configuration $s_i = -1$ with the probability p_-.

Markov chain corresponding to the algorithm is defined on the hupercube. To check the DB condition compute the probability flux from the state with $s_i = +1$ to the state with $s_i = -1$. It is $Q_{-+} = \frac{1}{Z}e^{-\beta E(s_i=-1)}p_+$, and then the reversed probability flux is $Q_{+-} = \frac{1}{Z}e^{-\beta E(s_i=+1)}p_-$. One finds that, indeed, the DB is satisfied since $Q_{-+} = Q_{+-}$.

Metropolis-Hastings Sampling

Metropolis-Hastings sampling is an MCMC method which explores efficiently the DB condition, i.e. reversibility of the underlying Markov Chain. The algorithm also uses sampling from the conditional probabilities and smart use of the rejection strategy. Assume that the probability of any state x from which one wants to sample (call it the target distribution) is explicitly known up to the normalization constant, Z, i.e. $p(x) = p(x)/Z$, where $Z = \sum_x p(x)$. 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.

Algorithm 8 Metropolis-Hastings Sampling

Input: Given $\pi(x)$ and $p(x'|x)$. Start with a sample x_t.

loop Till convergence
  Draw a random $x' \sim p(x'|x_t)$.
  Compute $\alpha = \frac{p(x_t|x')\pi(x')}{p(x'|x_t)\pi(x_t)}$.
  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 recordered as a new sample
end loop

Note that the Gibbs sampling previously introduced can be considered as the Metropolis-

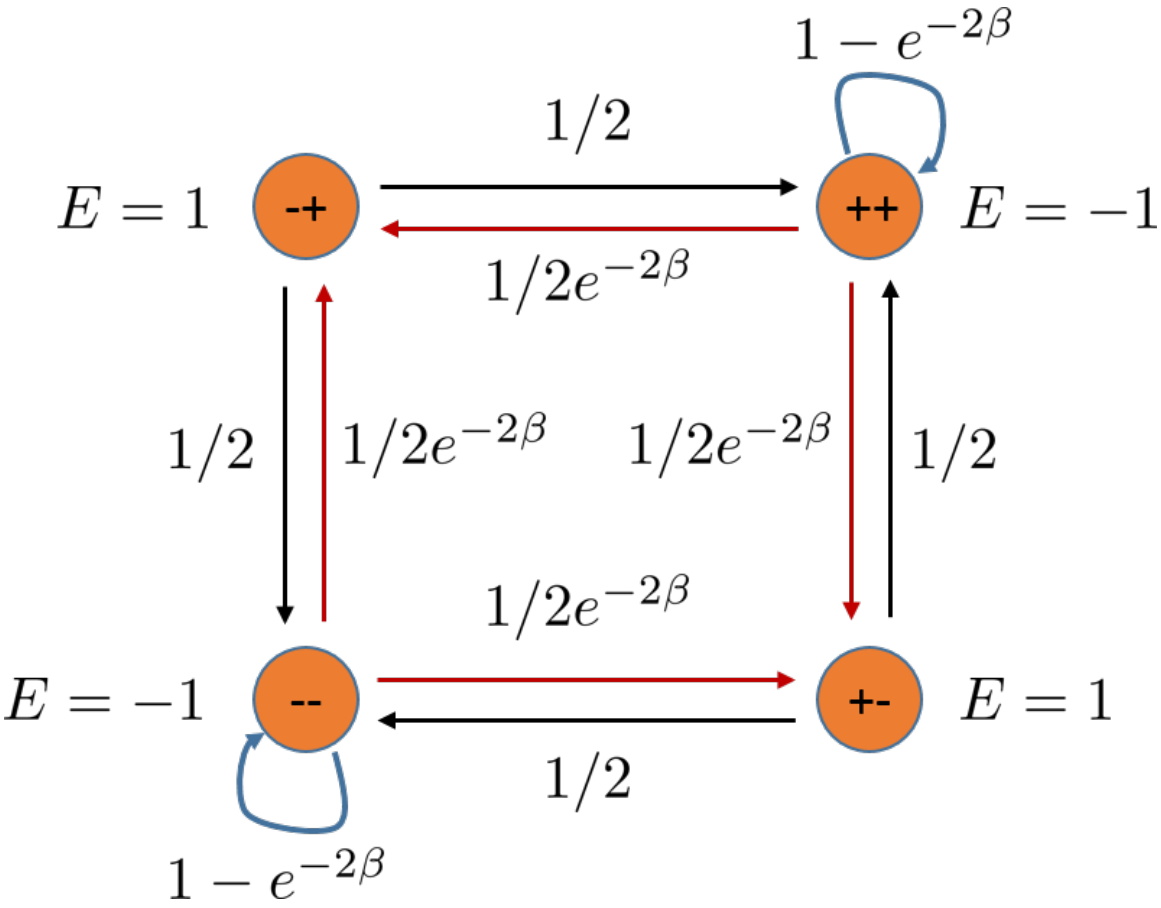


Figure 10.1: Metropolis-Hastings Markov chain example for two spins

Hastings without rejection (thus it is a particular case).

Example 10.1.3. Consider Markov chain example representing MH algorithm for two spins. Show that the Markov chain corresponds to an ergodic process. Describe the algorithm. Show that the algorithm obeys the DB condition. What is the resulting stationary distribution? MH algorithm contains the rejection step. What is the resulting steady distribution if the rejected step is removed from the consideration, as in the case of direct sampling by rejection?

Solution: We start from an arbitrary initial state and then perform a random walk in a state space flipping one spin at a time. Think about the algorithm as of a Markov chain defined over 2^N vertices of the hypercube. The algorithm works as follows: at each step one, first, chooses the random site i, then compute probabilities of keeping the spin value or flipping it (while other spins are kept instant), then flip the spin with the following

probability

$$p = \begin{cases} 1, & \text{if } \Delta E < 0 \\ e^{-\beta \Delta E}, & \text{if } \Delta E \geq 0. \end{cases} \tag{10.6}$$

Since our Markov chain is irreducible and aperiodic (contains self-loops), it is ergodic and thus has a unique stationary distribution. DB is checked directly. The algorithm converges to the Boltzmann/Gibbs distribution, $P_{eq}(s) = \frac{1}{Z} e^{-\beta E(s)}, \quad Z = \sum_s e^{-\beta E(s)}$.

If the spin flip is rejected one accepts the current state as a new configuration. This is an important difference with direct sampling by rejection. If the reject state is removed the resulting distribution is uniform.

The proposals (conditional probabilities) may vary. Details are critical (change mixing time), especially for large system. There is a (heuristic) rule of thumb: **lower bound on number of iterations of MH**. If the largest distance between the states is L, the MH will mix in time

$$T \approx (L/\varepsilon)^2 \tag{10.7}$$

where ε is the typical step size of the random walk.

Mixing may be extremely slow if the proposal distribution is not selected carefully. Let us illustrate how slow MCMC can be on a simple example. (See Section 29 of [15] for details.) Consider the following target distribution over N states

$$\pi(x) = \begin{cases} 1/N & x \in \{0, \cdots, N-1\} \\ 0 & \text{otherwise} \end{cases} \tag{10.8}$$

and proposal distribution over $N+2$ states (extended by -1 and N)

$$p(x'|x) = \begin{cases} 1/2 & x' = x \pm 1 \\ 0 & \text{otherwise} \end{cases} \tag{10.9}$$

Notice that the rejection can only occur when the proposed state is $x' = -1$ or $x' = N$.

A more sophisticated example of the Glauber algorithm (version of MH) on the example of the Ising Model is to be discussed next.

Glauber Sampling of Ising Model

Let us return to the special version of the Gibbs algorithm (and thus also a special case of the MH algorithm) developed specifically for the Ising model – the Glauber dynamics/algorithms:

Algorithm 9 Glauber Sampling

Input: Ising model on a graph. Start with a sample σ

loop Till convergence
  Pick a node i at random.
  $-\sigma_i \leftarrow \sigma_i$
  Compute $\alpha = \exp\left(\sigma_i \left(\sum_{j\in\mathcal{V}:\{i,j\}\in\mathcal{E}} J_{ij}\sigma_j - 2h_i\right)\right)$.
  Draw random $\beta \in U([0,1])$, uniform i.i.d. from $[0,1]$.
  if $\alpha < \beta < 1$ **then**
    $-\sigma_i \leftarrow \sigma_i$ [reject]
  end if
  Output: σ as a sample
end loop

Exercise 10.1.4 (not graded)**.** (a) What is the proposal distribution turning the MH sampling into the Glauber sampling (for the Ising model)? (b) Consider running parallel dynamics, based on the Glauber algorithm, i.e. at every moment of time update all variables in parallel according to the Glauber Sampling rule applied to the previous state. What is the resulting stationary distribution? Is it different from the Ising model? Does the algorithm satisfy the DB conditions?

Exercise 10.1.5 (Spanning Trees (not graded).)**.** Let G be an undirected complete graph. A simple MCMC algorithm to sample uniformly from the set of spanning trees of G is as follows: Start with some spanning tree; add uniformly-at-random some edge from G (so that a cycle forms); remove uniformly-at-random sample 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 probability proportional to the overall weight of the spanning for the following cases:
(i) the weight of any sub-graph of G is the sum of costs of its edges;
(ii) the weight of any sub-graph 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 some small (but non-trivial) weighted graph of your choice. 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 $\to \infty$:

$$\lim_{N\to+\infty} \frac{1}{N} \sum_{n=0}^{N} f(x_n) g(x_{n-1}) \to \mathbb{E}\left[f(x)\right] \mathbb{E}\left[g(x)\right], \tag{10.10}$$

where $f(x)$ and $g(x)$ are arbitrary functions (however such that respective expectations on the rhs of Eq. (10.10) 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):

$$\lim_{N\to+\infty} \frac{1}{N} \sum_{n=0}^{N} f(x_n) \to \mathbb{E}\left[f(x)\right], \tag{10.11}$$

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:

$$\lim_{s\to s_*} \lim_{N\to+\infty} \frac{1}{N} \sum_{n=0}^{N} f_s(x_n) \to \mathbb{E}\left[f_{s_*}(x)\right]. \tag{10.12}$$

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 [15]. An extensive set of modern references, discussions and codes are also available at the website 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.

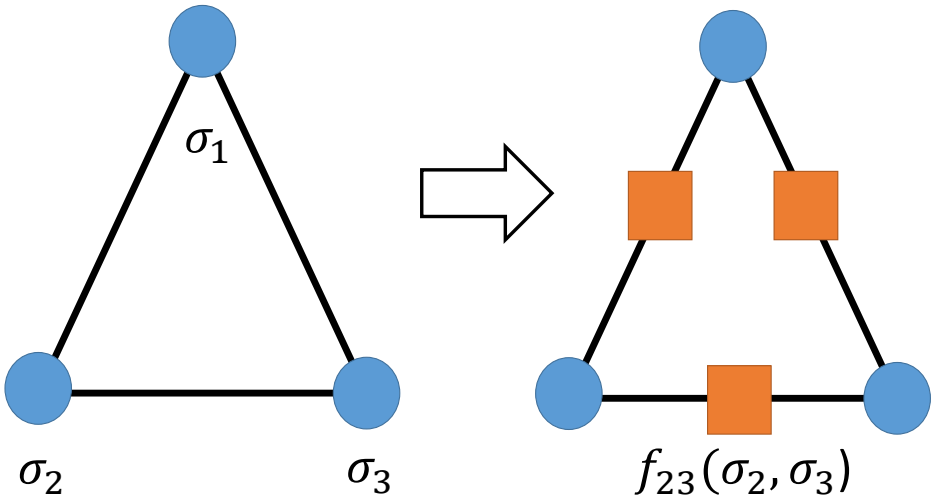


Figure 10.2: Factor Graph Representation for the (simple case) with pair-wise factors only. In the case of the Ising model: $f_{12}(\sigma_1, \sigma_2) = \exp\left(-J_{12}\sigma_1\sigma_2 + h_1\sigma_1 + h_2\sigma_2\right)$.

10.2 Graphical Models

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.

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).

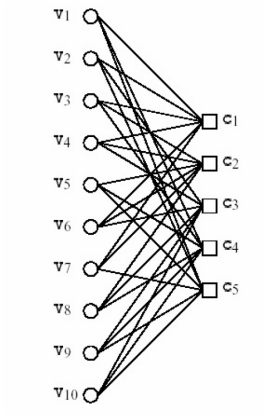


Figure 10.3: Tanner graph of a linear code, represented with $N = 10$ bits, $M = 5$ checks, and $L = N - M = 5$ information bits. This code selects 2^5 codewords from 2^10 possible patterns. This adjacency, parity-check matrix of the code is given by Eq. (10.14).

Ising Model, as well as other models discussed later in the lectures, can thus be stated in terms of the general factor-graph framework/model

$$P(\sigma) = Z^{-1} \prod_{a \in \mathcal{V}_f} f_a(\sigma_a), \quad \sigma_a \doteq (\sigma_i | i \in \mathcal{V}_n, \quad (i,a) \in \mathcal{E}), \tag{10.13}$$

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 interesting problem from Information Theory - decoding of error-correction codes.

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 [**?**] 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 zeros: $H_{i\alpha} = 1$ if $i \sim \alpha$ and $H_{i\alpha} = 0$ otherwise. The set of the codewords is thus defined as $\Xi^{(cw)} = (\varsigma | \boldsymbol{H}\boldsymbol{\varsigma} = \mathbf{0} \ (\mod 2))$. 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

$$\boldsymbol{H} = \begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 \end{pmatrix}. \tag{10.14}$$

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) 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 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:

$$\boldsymbol{\sigma}^{(MAP)} = \arg \min_{\boldsymbol{\sigma} \in \Xi^{(cw)}} \prod_{i=1}^{N} p(x_i|\sigma_i). \tag{10.15}$$

More generally, the probability of a codeword $\boldsymbol{\varsigma} \in \Xi^{(cw)}$ to be a pre-image for $\boldsymbol{x}$ is

$$\mathcal{P}(\boldsymbol{\sigma}|\boldsymbol{x}) = (Z(\boldsymbol{x}))^{-1} \prod_{i \in \mathcal{G}_{0;v}} g^{(ch)}(x_i|\varsigma_i), \quad Z(\boldsymbol{x}) = \sum_{\boldsymbol{\varsigma} \in \Xi^{(cw)}} \prod_{i \in \mathcal{G}_{0;v}} g^{(ch)}(x_i|\varsigma_i), \tag{10.16}$$

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

$$\forall\ i: \quad \varsigma_i^{(s-MAP)} = \arg \max_{\varsigma_i} \sum_{\boldsymbol{\varsigma} \setminus \varsigma_i}^{\boldsymbol{\varsigma} \in \Xi^{(cw)}} \mathcal{P}(\boldsymbol{\varsigma}|\boldsymbol{x}). \tag{10.17}$$

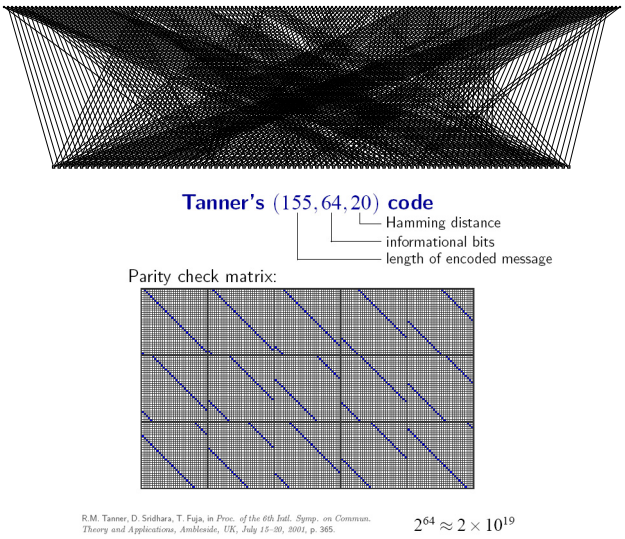


Figure 10.4: Tanner graph and parity check matrix of the $(155, 64, 20)$ Tanner code, where $N = 155$ is the length of the code (size of the code word), $L = 64$ and the Hamming distance of the code, $d = 20$.

Partition Function. Marginal Probabilities. Maximum Likelihood.

The partition function in Eq. (10.13) is the normalization factor

$$Z = \sum_{\sigma} \prod_{a \in \mathcal{V}_f} f_a(\sigma_a), \quad \sigma_a \doteq (\sigma_i | i \in \mathcal{V}_n, \quad (i,a) \in \mathcal{E}), \tag{10.18}$$

where $\sigma = (\sigma_i \in \{0,1\} \in \mathcal{V}_n)$. 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. (10.13) over a subset of variables, for example over all the elementary/nodal variables but one

$$P(\sigma_i) \doteq \sum_{\sigma \setminus \sigma_i} P(\sigma). \tag{10.19}$$

Expectation of σ_i computed with the probability Eq. (10.19) is also called (in physics) 'magnetization' of the variable.

Exercise 10.2.1 (not graded)**.** Does a partition function oracle sufficient for computing $P(\sigma_i)$? What is the relation in the case of the Ising model between $P(\sigma_i)$ and $Z(h)$?

Another object of interest is the so-called Maximum Likelihood. Stated formally, is the most probable state of all represented in Eq. (10.13):

$$\sigma_* = \arg\max_{\sigma} P(\sigma). \tag{10.20}$$

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 (10.20) 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 $P(\sigma_i)$ is a more difficult problem than computing σ_*.

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.

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. $b(\sigma)$ will be a belief - which is our probabilistic guess - for the probability of state σ. Consider it on the example of the triangle system shown in Fig. (10.2). There are 2^3 states in this case: $(\sigma_1 = \pm 1, \sigma_2 = \pm 1, \sigma_3 = \pm 1)$, which can occur with the probabilities, $b(\sigma_1, \sigma_2, \sigma_3)$. All the beliefs are positive and together should some to unity. We would like to compare a particular assignment of the beliefs with $P(\sigma)$, generally described by Eq. (10.13). Let us recall a tool which we already used to compare probabilities - the Kullback-Leibler (KL) divergence (of probabilities) discussed in Lecture #2:

$$D(b\|P) = \sum_{\sigma} b(\sigma) \log\left(\frac{b(\sigma)}{P(\sigma)}\right) \tag{10.21}$$

Note that the KL divergence (10.21) 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:

$$\forall \sigma : \quad b(\sigma) \geq 0, \tag{10.22}$$

$$\sum_{\sigma} b(\sigma) = 1. \tag{10.23}$$

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:

$$P = \arg\min_{b} D(b\|P), \quad \min_{b} D(b\|P) = 0. \tag{10.24}$$

Substituting Eq. (10.13) into Eq. (10.24) one derives

$$\log Z = -\min_{b} \mathcal{F}(b), \quad \mathcal{F}(b) \doteq \sum_{\sigma} b(\sigma) \log \left(\frac{\prod_a f_a(\sigma_a)}{b(\sigma)} \right), \tag{10.25}$$

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.

Variational Approximations. 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:

$$b(\sigma) \to b_{MF}(\sigma) = \prod_i b_i(\sigma_i) \tag{10.26}$$

$$\forall i \in \mathcal{V}_i, \quad \forall \sigma_i : \quad b_i(\sigma_i) \geq 0 \tag{10.27}$$

$$\forall i \in \mathcal{V}_i : \quad \sum_{\sigma_i} b_i(\sigma_i) = 1. \tag{10.28}$$

Clearly $b_i(\sigma_i)$ 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. (10.25) one arrives at the MF estimation for the partition function

$$\log Z_{mf} = -\min_{b_{mf}} \mathcal{F}(b_{mf}),$$
$$\mathcal{F}(b_{mf}) \doteq \sum_a \sum_{\sigma_a} \left(\prod_{i \sim a} b_i(\sigma_i) \right) \log f_a(\sigma_a) - \sum_i \sum_{\sigma_i} b_i(\sigma_i) \log(b_i(\sigma_i)). \quad (10.29)$$

To solve the variational problem (10.29) constrained by Eqs. (10.26,10.27,10.28) is equivalent to searching for the (unique) stationary point of the following MF Lagrangian

$$\mathcal{L}(b_{mf}) \doteq \mathcal{F}(b_{mf}) + \sum_i \lambda_i \sum_{\sigma_i} b_i(\sigma_i) \quad (10.30)$$

Exercise 10.2.2 (Not graded.)**.** Show that $Z_{mf} \geq Z$, and that $\mathcal{F}(b_{mf})$ is a strictly convex function of its (vector) argument. Write down equations defining the stationary point of $\mathcal{L}(b_{mf})$. Suggest an iterative algorithm converging to the stationary point of $\mathcal{L}(b_{mf})$.

The fact that Z_{mf} (see the exercise above) gives an upper 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.

Dynamic Programming for Inference over Trees

Consider Ising model over a linear chain of n spins shown in Fig. 10.5a, the partition function is

$$Z = \sum_{\sigma_n} Z(\sigma_n), \quad (10.31)$$

where $Z(\sigma_n)$ 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

$$Z(\sigma_n) = \sum_{\sigma_{n-1}} \exp(J_{n,n-1}\sigma_n\sigma_{n-1} + h_n\sigma_n) Z_{(n-1)\to(n)}(\sigma_{n-1}), \quad (10.32)$$

where $Z_{(n-1)\to(n)}(\sigma_i)$ 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

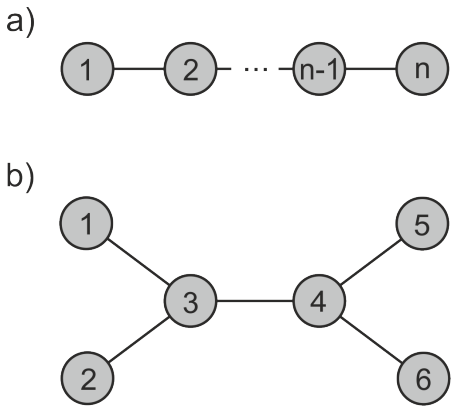


Figure 10.5: Exemplary interaction/factor graphs which are tree.

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

$$Z_{(i-1)\to(i)}(\sigma_{i-1}) = \sum_{\sigma_{i-2}} \exp(J_{i-1,i-2}\sigma_{i-1}\sigma_{i-2} + h_{i-1}\sigma_{i-1}) Z_{(i-2)\to(i-1)}(\sigma_{i-2}) \tag{10.33}$$

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 $Z(\sigma_i)$ is the partition function of the entire tree with a value of the spin at the site/node i fixed. We derive

$$Z(\sigma_i) = e^{h_i\sigma_i} \prod_{j\in\partial i} \left(\sum_{\sigma_j} e^{J_{ij}\sigma_i\sigma_j} Z_{j\to i}(\sigma_j) \right), \tag{10.34}$$

where ∂i denotes the set of neighbors of the i-th spin and

$$Z_{j\to i}(\sigma_j) = e^{h_j\sigma_j} \prod_{k\in\partial j\setminus i} \left(\sum_{\sigma_k} e^{J_{kj}\sigma_k\sigma_j} Z_{k\to j}(\sigma_j) \right) \tag{10.35}$$

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

$$Z = \sum_{\sigma_4} Z(\sigma_4), \tag{10.36}$$

The partition function, partially summed and conditioned to the spin value at the spin, σ_4, is

$$Z(\sigma_4) = e^{h_4\sigma_4} \sum_{\sigma_5} e^{J_{45}\sigma_4\sigma_5} Z_{5\to 4}(\sigma_5) \sum_{\sigma_6} e^{J_{46}\sigma_4\sigma_6} Z_{6\to 4}(\sigma_6) \sum_{\sigma_3} e^{J_{34}\sigma_3\sigma_4} Z_{3\to 4}(\sigma_3) \tag{10.37}$$

where

$$Z_{3\to 4}(\sigma_3) = e^{h_3\sigma_3} \sum_{\sigma_1} e^{J_{13}\sigma_1\sigma_3} Z_{1\to 3}(\sigma_1) \sum_{\sigma_2} e^{J_{23}\sigma_2\sigma_3} Z_{2\to 3}(\sigma_2). \tag{10.38}$$

Exercise 10.2.3 (not graded)**.** Demonstrate that the i-th spin is conditionally independent of all other spins, given values of spins of the i-th spin neighbors fixed, i.e.

$$p(\sigma_i|\sigma/\sigma_i) = p(\sigma_i|\sigma_j \sim \sigma_i), \tag{10.39}$$

where, $p(\sigma_i|\sigma/\sigma_i)$, is the probability distribution of the ith spin conditioned to the values of all other spins, and, $p(\sigma_i|\sigma_j \sim \sigma_i)$, is the probability distribution of ith spin conditioned to the spin values of its neighbors.

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

$$P(x_1, x_2) = P(x_1)P(x_2|x_1), \tag{10.40}$$

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

$$\begin{aligned} P(x_1, x_2, x_3) &= P(x_1, x_2)P(x_3|x_1, x_3) = P(x_1, x_2)P(x_3|x_2) = \\ &= P(x_1)P(x_2|x_1)P(x_3|x_2) = \frac{P(x_1, x_2)P(x_2, x_3)}{P(x_2)}, \end{aligned} \tag{10.41}$$

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

$$\begin{aligned} P(x_1, x_2, x_3, x_4) &= P(x_1, x_2, x_3)P(x_4|x_1, x_3, x_4) = P(x_1, x_2, x_3)P(x_4|x_2) = \\ &= P(x_1, x_2)P(x_3|x_1, x_2)P(x_4|x_2) = P(x_1, x_2)P(x_3|x_2)P(x_4|x_2) = \\ &= P(x_1)P(x_2|x_1)P(x_3|x_2)P(x_4|x_2) = \frac{P(x_1, x_2)P(x_2, x_3)P(x_2, x_4)}{P^2(x_2)}. \end{aligned} \tag{10.42}$$

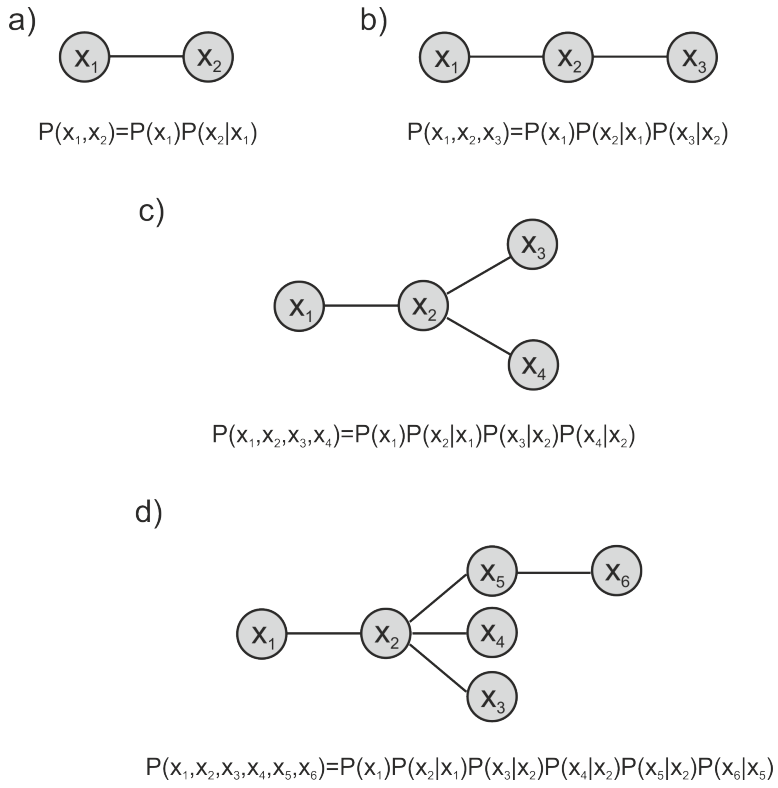


Figure 10.6: Examples of undirected tree-structured graphical models.

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

$$P(x_1, x_2, x_3, x_4, x_5, x_6) = P(x_1)P(x_2|x_1)P(x_3|x_2)P(x_4|x_2)P(x_5|x_2)P(x_6|x_5) = \\ = \frac{P(x_1, x_2)P(x_2, x_3)P(x_2, x_4)P(x_2, x_5)P(x_5, x_6)}{P^3(x_2)P(x_5)}. \quad (10.43)$$

In general, the joint probability distribution of a tree-like graphical model can be written as follows

$$P(x_1, x_2, \ldots, x_n) = \frac{\prod_{(i,j)\in\mathcal{E}} P(x_i, x_j)}{\prod_{i\in\mathcal{V}} P^{q_i-1}(x_i)}, \quad (10.44)$$

where q_i is the degree of the ith node. Eq. (10.44) can be proven by induction.

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 [**?**], a comprehensive review [**?**], and respective lecture notes, for an advanced/additional reading.

Instead of Eq. (10.26) one uses the following BP substitution

$$b(\sigma) \to b_{bp}(\sigma) = \frac{\prod_a b_a(\sigma_a)}{\prod_i (b_i(\sigma_i))^{q_i-1}} \tag{10.45}$$

$$\forall a \in \mathcal{V}_f, \quad \forall \sigma_a : \quad b_a(\sigma_a) \geq 0 \tag{10.46}$$

$$\forall i \in \mathcal{V}_n, \quad \forall a \sim i : \quad b_i(\sigma_i) = \sum_{\sigma_a \setminus \sigma_i} b_a(\sigma_a) \tag{10.47}$$

$$\forall i \in \mathcal{V}_n : \quad \sum_{\sigma_i} b_i(\sigma_i) = 1. \tag{10.48}$$

where q_i stands for degree of node i. The physical meaning of the factor $q_i - 1$ on the rhs of Eq. (10.45) 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. (10.45) comes as a correction for this over-counting.

Substitution of Eqs. (10.45) into Eq. (10.25) results in what is called Bethe Free Energy (BFE)

$$\mathcal{F}_{bp} \doteq E_{bp} - \mathcal{H}_{bp}, \tag{10.49}$$

$$E_{bp} \doteq -\sum_a \sum_{\sigma_a} b_a(\sigma_a) \log f_a(\sigma_a) \tag{10.50}$$

$$\mathcal{H}_{bp} = \sum_a \sum_{\sigma_a} b_a(\sigma_a) \log b_a(\sigma_a) - \sum_i \sum_{\sigma_i} (q_i - 1) b_i(\sigma_i) \log b_i(\sigma_i), \tag{10.51}$$

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

$$\arg\min_{b_a, b_i} \mathcal{F}_{bp}\Big|_{\text{Eqs. (10.46,10.47,10.48)}}, \tag{10.52}$$

$$\min_{b_a, b_i} \mathcal{F}_{bp}\Big|_{\text{Eqs. (10.46,10.47,10.48)}} \tag{10.53}$$

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. (10.52) results from the following optimization

$$\min_{b_a, b_i} E_{bp}\Big|_{\text{Eqs. (10.46,10.47,10.48)}} \tag{10.54}$$

Note the optimization is a Linear Programming (LP) — minimizing linear objective over set of linear constraints.

Belief Propagation & Message Passing

Let us restate Eq. (10.52) as an unconditional optimization. We use the standard method of Lagrangian multipliers to achieve it. The resulting Lagrangian is

$$\begin{aligned}\mathcal{L}_{bp}(b,\eta,\lambda) \doteq & \sum_a \sum_{\sigma_a} b_a(\sigma_a)\log f_a(\sigma_a) - \sum_a \sum_{\sigma_a} b_a(\sigma_a)\log b_a(\sigma_a) \\ & + \sum_i \sum_{\sigma_i} (q_i-1) b_i(\sigma_i)\log b_i(\sigma_i) \\ & - \sum_i \sum_{a\sim i} \sum_{\sigma_i} \eta_{ia}(\sigma_i)\left(b_i(\sigma_i) - \sum_{\sigma_a\setminus\sigma_i} b_a(\sigma_a)\right) + \sum_i \lambda_i \left(\sum_{\sigma_i} b_i(\sigma_i) - 1\right), \quad (10.55)\end{aligned}$$

where η and λ are the dual (Lagrangian) variables associated with the conditions Eqs. (10.47,10.48) respectively. Then Eq. (10.52) become the following min-max problem

$$\min_b \max_{\eta,\lambda} \mathcal{L}_{bp}(b,\eta,\lambda). \quad (10.56)$$

Changing the order of optimizations in Eq. (10.56) and then minimizing over η one arrives at the following expressions for the beliefs via messages (check the derivation details)

$$\begin{aligned}\forall a,\ \forall \sigma_a: \qquad b_a(\sigma_a) &\sim f_a(\sigma_a)\exp\left(\sum_{i\sim a}\eta_{ia}(\sigma_i)\right) \doteq f_a(\sigma_a)\prod_{i\sim a} n_{i\to a}(\sigma_i) \\ &\doteq f_a(\sigma_a)\prod_{i\sim a}\prod_{b\sim i}^{b\neq a} m_{b\to i}(\sigma_i) \quad (10.57)\\ \forall i,\ \forall \sigma_i: \qquad b_i(\sigma_i) &\sim \exp\left(\frac{\sum_{a\sim i}\eta_{ia}(\sigma_i)}{q_i-1}\right) \doteq \prod_{a\sim i} m_{a\to i}(\sigma_i), \quad (10.58)\end{aligned}$$

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

$$\begin{aligned}\forall i,\ \forall a\sim i: \quad & n_{i\to a}(\sigma_i) \doteq \exp\left(\eta_{ia}(\sigma_i)\right) \quad (10.59)\\ \forall a,\ \forall i\sim a: \quad & m_{a\to i}(\sigma_i) \doteq \exp\left(\frac{\eta_{ia}(\sigma_i)}{q_i-1}\right). \quad (10.60)\end{aligned}$$

Combining Eqs. (10.57,10.58,10.59,10.60) with Eq. (10.47) results in the following BP-equations stated in terms of the message variables

$$\forall i,\ \forall a \sim i,\ \forall \sigma_i : \quad n_{i\to a}(\sigma_i) = \prod_{b\sim i}^{b\neq a} m_{a\to i}(\sigma_i) \tag{10.61}$$

$$\forall a,\ \forall i \sim a,\ \forall \sigma_i : \quad m_{a\to i}(\sigma_i) = \sum_{\sigma_a\setminus\sigma_i} f_a(\sigma_a) \prod_{j\sim a}^{j\neq i} n_{j\to a}(\sigma_j). \tag{10.62}$$

Note that if the Bethe Free Energy (10.49) is non-convex there may be multiple fixed points of the Eqs. (10.61,10.62). The following iterative, so called Message Passing (MP), algorithm (10) is used to find a fixed point solution of the BP Eqs. (10.45,10.46)

Algorithm 10 Message Passing, Sum-Product Algorithm [factor graph representation]

Input: The graph. The factors.

1: $\forall i,\ \forall a \sim i, \forall \sigma_i : \quad m_{a\to i} = 1$ [initialize variable-to-factor messages]
2: $\forall a,\ \forall i \sim a, \forall \sigma_i : \quad n_{i\to 1} = 1$ [initialize factor-to-variable messages]
3: **loop**Till convergence within an error [or proceed with a fixed number of iterations]
4: $\quad \forall i,\ \forall a \sim i,\ \forall \sigma_i : \quad n_{i\to a}(\sigma_i) \leftarrow \prod_{b\sim i}^{b\neq a} m_{a\to i}(\sigma_i)$
5: $\quad \forall a,\ \forall i \sim a,\ \forall \sigma_i : \quad m_{a\to i}(\sigma_i) \leftarrow \sum_{\sigma_a\setminus\sigma_i} f_a(\sigma_a) \prod_{j\sim a}^{j\neq i} n_{j\to a}(\sigma_j)$
6: **end loop**

Exercise 10.2.4 (not graded)**.** Derive the $T = 0$ version of the aforementioned (see previous exercise) message-passing equations. [A hint: the iterative equations should contain alternating min- and sum- steps — thus the name min-sum algorithm.] Study performance of the message-passing algorithm on example of a small code decoding, for example check this student midterm paper for discussion of decoding of a binary $(3, 6)$ code over the Binary-Erasure Channel (BEC). Show how BP decodes and contrast the BP decoding against the MAP decoding. What is the (best) complexity of the MAP decoder for a code over the BEC channel? [Hint: Use Gaussian Elimination over GL(2).]

Sufficient Statistics

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) using a little bit different notations then before

$$P(\sigma) = \frac{1}{Z(\theta)} \exp\left\{\sum_{i\in V} \theta_i\sigma_i + \sum_{\{i,j\}\in E} \theta_{ij}\sigma_i\sigma_j\right\} = \exp\{\theta^T\phi(\sigma) - \log Z(\theta)\}, \tag{10.63}$$

where $\sigma_i \in \{-1, 1\}$ and the *partition function* $Z(\theta)$ serves to normalize the probability distribution. In fact, Eq. (10.63) describes what is called the *exponential family* - emphasizing 'exponential' dependence on the factors θ.

Exercise 10.2.5 (not graded)**.** Show 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) of the spin variables, $\mu^{(1)} \doteq (\mu_i = \mathbb{E}[\sigma_i], i \in V)$ and $\mu^{(2)} \doteq (\mu_{ij} = \mathbb{E}[\sigma_i\sigma_j], \{i, j\} \in E)$. The *sufficient statistics* statement is that to reconstruct θ, fully defining the GM, it is *sufficient* to know $\mu^{(1)}$ and $\mu^{(2)}$.

Maximum-Likelihood Estimation/Learning of GM

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)

$$\forall i: \quad \partial_{\theta_i} \log Z(\theta) = -\mu_i, \quad \forall i, j: \quad \partial_{\theta_{ij}} \log Z(\theta) = -\mu_{ij}. \tag{10.64}$$

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):

$$\theta^* = \arg\max_\theta \{\mu^T\theta - \log Z(\theta)\} \tag{10.65}$$

If P represents the empirical distribution of a set of independent identically-distributed (i.i.d.) samples $\{\sigma^{(s)}, s = 1, \dots, S\}$ then μ are the corresponding empirical moments, e.g. $\mu_{ij} = \frac{1}{S}\sum_s \sigma_i^{(s)}\sigma_j^{(s)}$.

General Remarks about GM Learning. The ML parameter Estimation (10.65) 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 needs to rely on 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 this recent paper for details.)

- Relay 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. But check this article and references therein, if interested.

Learning Spanning Tree

Eq. (10.44) 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 (10.44) 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.

Consider the candidate probability distribution, $P_T(x_1, \ldots, x_n)$ 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. (10.44) via marginal (pair-wise and single-variable) probabilities as follows

$$P_T(x_1, x_2, \ldots, x_n) = \frac{\prod_{(i,j)\in\mathcal{E}^F} P(x_i, x_j)}{\prod_{i\in\mathcal{V}^F} P(x_i)^{q_i-1}(x_i)}. \tag{10.66}$$

"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

$$D(P \parallel P_T) = -\sum_{\vec{x}} P(\vec{x}) \log \frac{P(\vec{x})}{P_T(\vec{x})}. \tag{10.67}$$

As discussed in Section 8.3, 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 (10.66) into Eq. (10.67) one arrives at the following chain of explicit transformations

$$\begin{aligned}
&\sum_{\vec{x}} P(\vec{x}) \left(\log P(\vec{x}) - \sum_{(i,j)\in\mathcal{E}} \log P(x_i, x_j) + \sum_{i\in\mathcal{V}} (q_i - 1) \log P(x_i) \right) = \\
&= \sum_{\vec{x}} P(\vec{x}) \log P(\vec{x}) - \sum_{(i,j)\in\mathcal{E}^{\mathcal{F}}} \sum_{x_i,x_j} P(x_i, x_j) \log P(x_i, x_j) + \\
&+ \sum_{i\in\mathcal{V}} (q_i - 1) \sum_{x_i} P(x_i) \log P(x_i) = - \sum_{(i,j)\in\mathcal{E}^{\mathcal{F}}} \sum_{x_i,x_j} P(x_i, x_j) \log \frac{P(x_i, x_j)}{P(x_i)P(x_j)} + \\
&+ \sum_{\vec{x}} P(\vec{x}) \log P(\vec{x}) - \sum_{i\in\mathcal{V}^{\mathcal{F}}} \sum_{x_i} P(x_i) \log P(x_i),
\end{aligned} \tag{10.68}$$

where the following nodal and edge marginalization relations were used, $\forall i \in \mathcal{V}^{\mathcal{F}}: \quad P(x_i) = \sum_{\vec{x}\setminus x_i} P(\vec{x})$, and, $\forall (i,j) \in \mathcal{E}^{\mathcal{F}}: \quad P(x_i, x_j) = \sum_{\vec{x}\setminus x_i, x_j} P(\vec{x})$, respectively. One observes that the Kullback-Leibler divergence becomes

$$D(P \parallel P_F) = -\sum_{(i,j)\in\mathcal{E}^{\mathcal{F}}} I(X_i, X_j) + \sum_{i\in\mathcal{V}^{\mathcal{F}}} S(x_i) - S(\vec{x}), \tag{10.69}$$

where

$$I(X_i, X_j) \doteq \sum_{x_i,x_j} P(x_i, x_j) \log \frac{P(x_i, x_j)}{P(x_i)P(x_j)} \tag{10.70}$$

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

$$\sum_{(i,j)\in\mathcal{E}^{\mathcal{F}}} I(X_i, X_j). \tag{10.71}$$

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):

- (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. (10.44) 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 (10.44) and the empirical joint distribution. This reconstruction of the optimal tree is based on the Chow-Liu algorithm.

Exercise 10.2.6. Find Chou-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 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.]

Table 10.1: Information available about an exemplary probability distribution of four binary variables discussed in the Exercise 10.2.6.

$x_1x_2x_3x_4$	$P(x_1, x_2, x_3, x_4)$	$P(x_1)P(x_2\|x_1)P(x_3\|x_2)P(x_4\|x_1)$	$P(x_1)P(x_2)P(x_3)P(x_4)$
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

10.3 Neural Networks

This Section is work in progress. If time permits, we plan to follow here material from Chapter V of the "Information Theory Inference and Learning Algorithms" book by David MacKay [15] devoted to Neural Networks. Some useful material can also be found in the recent book "Linear Algebra and Learning from Data" by Gilbert Strang [18], specifically in the Chapter VII "Learning from Data"; and also in the lecture on "Deep Learning and Graphical Models" by Eric Xing.

10.3.1 Single Neuron and Supervised Learning

Exercise 10.3.1. Consider a Neural Network (NN) with two layers, each with only one node. Assume that each node is assigned the activation function

$$\hat{y} = \tanh\Big(w_2 \tanh\big(w_1 x + b_1\big) + b_2\Big),$$

and assume that the weights are currently set at $(w_1, b_1) = (1.0, 0.5)$ and $(w_2, b_2) = (-0.5, 0.3)$. 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?

10.3.2 Hopfield Networks and Boltzmann Machines

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] A. Chambolle, "An algorithm for total variation minimization and applications," *Journal of Mathematical Imaging and Vision*, vol. 20, pp. 89–97, 2004.

[11] 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

[12] L. Pontryagin, V. Boltayanskii, R. Gamkrelidze, and E. Mishchenko, *The mathematical theory of optimal processes (translated from Russian in 1962).* Wiley, 1956.

[13] A. T. FULLER, "Bibliography of pontryagm's maximum principle," *Journal of Electronics and Control*, vol. 15, no. 5, pp. 513–517, 1963.

[14] R. Bellman, "On the theory of dynamic programming," *PNAS*, vol. 38, no. 8, p. 716, 1952.

[15] D. J. C. Mackay, *Information theory, inference, and learning algorithms.* Cambridge University Press, 2003.

[16] C. Moore and S. Mertens, *The Nature of Computation.* New York, NY, USA: Oxford University Press, 2011.

[17] N. V. Kampen, *Stochastic processes in physics and chemistry.* North Holland, 2007.

[18] G. Strang, *Linear Algebra and Learning from Data.* Wellesley-Cambridge Press, 2019.