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

June 25, 2020

Contents

Chapter 1

Applied Math Core Courses

Every student in the Program for Applied Mathematics at the University of Arizona takes the same three core courses during their first year of study. These three courses are called Methods (Math 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. 1.1). 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 simulations and computation. Understanding the theoretical aspects of a problem motivates better design and implementation of these methods and algorithms, and allows us to make precise statements about when and how they will work.

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

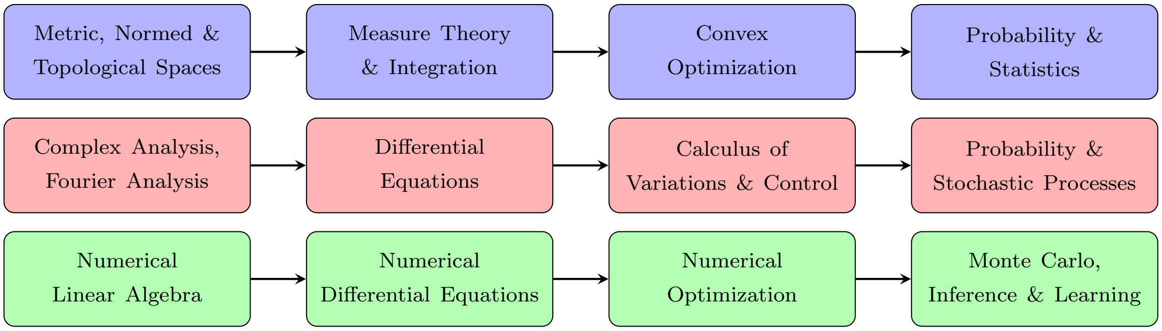
Metric, Normed & Topological Spaces
Measure Theory & Integration
Convex Optimization
Probability & Statistics
Complex Analysis, Fourier Analysis
Differential Equations
Calculus of Variations & Control
Probability & Stochastic Processes
Numerical Linear Algebra
Numerical Differential Equations
Numerical Optimization
Monte Carlo, Inference & Learning

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

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.

1.1 What is Applied Mathematics?

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

Applying mathematics to real-world problems requires mathematical approaches that have evolved to stand up to the many demands and complications of real-world problems. In some applications, a relatively simple set of governing mathematical expressions are able to describe the relevant phenomena. In these situations, problems often require very accurate solutions, and the mathematical challenge is to develop methods that are efficient (and sometimes also adaptable to variable data) without losing accuracy. In other applications, there is no set of governing mathematical expressions (either because we do no know them,

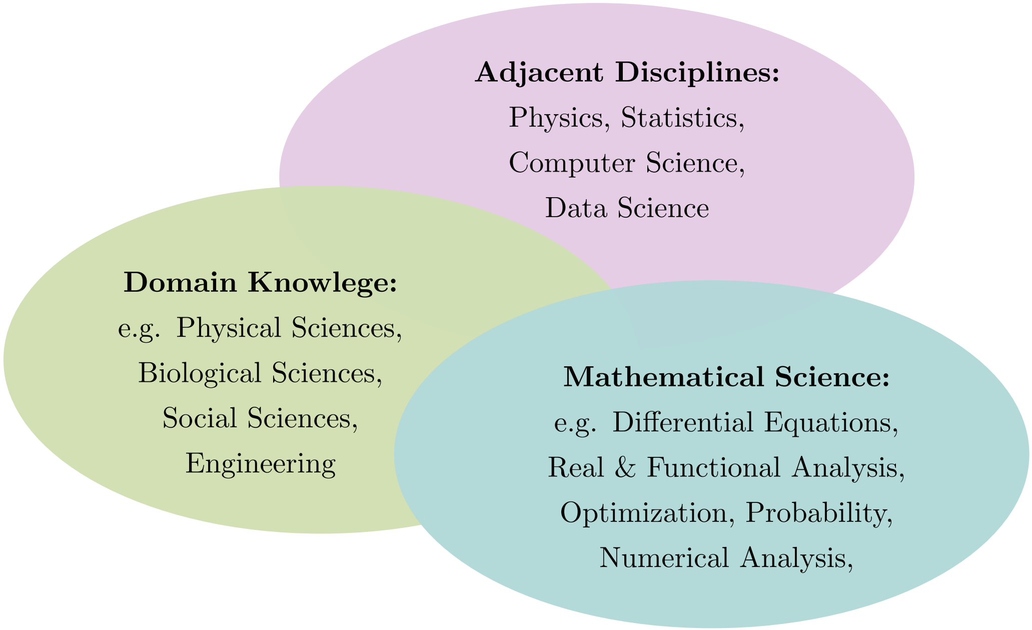

Figure 1.2: The key components studied under the umbrella of applied mathematics: (1) mathematical science, (2) domain-specific knowledge, and (3) a few 'math-adjacent' disciplines.

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:

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.

Chapter 2

Preview of Topics for Math 583

This introductory chapter provides a glimpse of the mathematical material we will discuss during the methods course, and more specifically, it demonstrates how this material may be applied to real-world problems. This chapter does *not* try to mimic the flavor of the course—during the course, we must necessarily spend most of the time learning the mathematics (and when and why it works), whereas in this chapter, we will present only some highlights of the material without working through any details. We have chosen to present these highlights by illustrating how they may be applied to real-world problems.

We recommend reading this preview chapter during the summer before taking the course. The chapter is not mandatory, nor is there any course credit to be earned; it is only intended to 'whet your appetite". We hope the exercises are thought provoking, but we do acknowledge how difficult they are without proper discussion of the yet-to-be-discussed material, and have thus provided many of the solutions.

2.1 Applied Analysis

The first two chapters of the methods course cover complex analysis and Fourier analysis. Neither of these topics lends itself quite as directly to real-world problems, and so we only include a brief description of what will be discussed and then quickly move to differential equations, optimization, and the mathematics of uncertainty.

The methods course begins with approximately four weeks of complex analysis. During this time, we discuss complex-valued functions with an emphasis on their representation by Laurent series, on singularities and on multi-valued functions. We then move to calculus of complex-valued functions, beginning with the Cauchy-Riemann conditions for differentiability, and then to contour integration. A considerable portion of this chapter will discuss integration along parameterized curves, Cauchy's theorem and residue calculus, and applications of Cauchy's theorem to integrals involving multi-valued functions (often with singularities). The chapter ends with a brief preview of the asymptotic approximation of integrals by the methods of stationary phase and of steepest descent.

The methods course then moves to about four weeks of Fourier analysis. During this time, we discuss properties and closed form solutions of both Fourier series and Fourier transforms. We use this opportunity to introduce generalized functions, which will be revisited in the chapter on differential equations. Some theoretical details of Fourier analysis are discussed in this course, e.g. convergence (briefly), Gibbs phenomenon and the Riemann-Lebesgue lemma, but the finer details of integrability and of convergence are discussed in the theory course.

2.2 Equations: From Algebraic to Differential

In this section we walk (fast and through examples) from algebraic equations to differential equations, first ordinary and then partial.

2.2.1 Algebraic Equations

An **equation** is the statement of equality between two expressions, for example $y = ax^2 + bx + c$. Here, a, b and c are parameters (quantities whose values are fixed or can be selected) and x and y are variables. This simple algebraic equation can be solved **analytically** and is guaranteed to have two (not-necessarily unique) solutions over complex space, $x_* \in \mathbb{C}$, however it may have no real solutions.

Not every problem is so simple. The Kirchoff equations

$$\forall a \in 1, \cdots, N: \quad Q_a = V_a \sum_{b=0,\cdots,N} \left(\frac{V_a - V_b}{z_{ab}}\right)^*, \quad V_0 = 1, \tag{2.1}$$

is a **system** of algebraic equations that appear in the context of power systems operating under the alternating current (AC) paradigm, and will be one of our exemplary engineered systems. Here, N is a finite positive integer determining the number of nodes of the system, V_a and Q_a represent the complex-valued voltage and power at node a, and z_{ab} is a complex-valued parameter denoting the impedance along the line joining nodes a and b. Here J^* is standard notation for the complex conjugation of J.

The system of (quadratic) algebraic equations (2.1) over a physically sensible subspace of N dimensional complex space, $\boldsymbol{V} = (V_a \mid a \in 1, \cdots, N) \in \mathbb{C}^N$, say $\forall a = 1, \cdots, N,\ |1 - V_a| \leq 0.2$, may have multiple solutions or no solutions. In the general case, understanding these equations analytically, even for questions like the existence of solutions (**feasibility**), can be hopeless. We may, however, try solving this system of equations numerically, which may produce answers that are just as valuable as analytic solutions.

Exercise 2.2.1. Some power systems operate with direct current (DC). DC systems operate according to the Kirchoff's equations (2.1) with the additional restriction that all the parameters and variables in the equations are real. Consider the simplest case of such a DC power systems consisting of two nodes (i.e. $N = 1$). Assume that $V_0 = 1$ at the slack bus (the generating node). Study the dependence of V_1 on Q_1 and identify the regime where the resulting equation has no solutions.

Solution. Take $N = 1$. Since all variables and parameters are real, we write

$$Q_1 = V_1 \left(\frac{V_1 - V_0}{z_{10}} \right), \quad V_0 = 1,$$

or

$$V_1^2 - V_1 - Q_1 z_{10} = 0$$

Hence, we have $V_1 = \frac{1 \pm \sqrt{1+4Q_1 z_{10}}}{2}$, and we need $Q_1 z_{10} \leq -1/4$ to have real solutions. □

Later on (and more in the spring) we will also consider inequalities and systems of inequalities. These either have infinitely many solutions or have no solution, thus the feasibility question may be posed and answered too, but in general only numerically.

An **analytic solution** generally refers to a formula, whereas a numerical solution generally refers to an **algorithm**. An algorithm may find the **exact** solution in a finite number of steps, but this is rarely the case. The algorithm may also only find an exact solution asymptotically, that is, only in the limit of infinitely many steps. In this case, the user usually terminates the algorithm prematurely and gives an **approximate** solution. Addressing the **approximation quality** of an algorithm involves (i) introducing a suitable measure of (approximation) accuracy, and (ii) analyzing how the measure of accuracy improves (or not) with number of the (algorithm) steps.

2.2.2 Ordinary Differential Equations

As discussed in Exercise 2.2.1, Kirchoff's equations (2.1) may have no solution. In pure mathematics this is the end of the story. However, we are doing applied mathematics where we would also like to understand what this "no solution" statement means for the application. Indeed, Kirchoff's equations describe the relation between the power injected (or consumed) at a particular node, $a = 1$, and the voltage at the same node (in the exercise, the node a is connected by a power line to another node, $b = 0$, where voltage is maintained constant). Kirchoff's equations assume a steady (i.e. time independent) situation. However, situations that do not admit any steady-state solutions may in fact admit transient (time dependent) solutions. Temporal equations should be introduced if we intend to fully explain power engineering. Therefore, assuming for simplicity that the power line connecting nodes a and b is purely inductive (i.e. $z_{ab} = i/\beta_{ab}$, where $\beta_{ab} > 0$), and all voltages are set to the same constant and then re-scaled to unity, one generalizes Kirchoff's equations (2.1) by the inhomogeneous Kuramoto-Sivashinsky system of ODEs, or swing-equations

$$\forall a \in \mathcal{V}: \quad \tau_a \frac{d^2}{dt^2}\theta_a + \frac{d}{dt}\theta_a = p_a - \sum_{b:\{a,b\}\in\mathcal{E}} \beta_{ab} \sin(\theta_a - \theta_b), \tag{2.2}$$

over the undirected graph, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, representing the power system, where $\mathcal{V}$ and $\mathcal{E}$ are the sets of nodes and edges of the graph. Here $i^2 = -1,\ p_a = \text{Re}[Q_a],\ V_a = \exp(i\theta_a)$, i.e. the voltages of the complex potentials, $|V_a|$, at all nodes of the system are set to unity, and the phases, θ_a are counted from a particular node of the system (the so-called slack bus, $a = 0$, which is typically chosen to be the biggest generator in the system).

The system of ODEs is difficult because the underlying equations are nonlinear and have too many degrees of freedom. (Even though the number of degrees of freedom is finite in the example of finite graph power systems.) It is reasonable to first analyze a simpler problem involving a single ODE. For example, in the AC power system example described by equation (2.2), we may consider a simple two-node network, with a generator connected to a load by an inductive line. Then, the phase at the load, $\theta(t)$, counted from the phase of the generator (the reference bus), evolves in time according to the following ODE

$$\tau\ddot{\theta} + \dot{\theta} = p - \beta\sin(\theta) = -U'(\theta), \quad U(\theta) = -p\theta + \beta\left(1 - \cos(\theta)\right), \tag{2.3}$$

where $\dot{U} := dU(\theta)/d\theta$, is the derivative of the potential with respect to its argument, the phase.

Equation (2.3) and similar ODEs appear in many other areas where engineered or natural systems need to be modeled. For example, it is very common to encounter similar problems in classical mechanics that describes the evolution in time t of a unit mass particle positioned at $x(t) \in \mathbb{R}$, moving with speed $v(t)$, and subject to a frictional force, $-v/\tau$, and a potential force, $-\partial_x U(x)$. The resulting second order ODE can be stated in terms of the following two first order ODEs

$$\dot{x} = v, \quad \dot{v} = -\frac{v}{\tau} - \partial_x U(x). \tag{2.4}$$

Dynamics, governed by this equation in an exemplary (and rather common) case of a double well potential is described in a supplementary simulation snippet [RelaxInertStoch.ipynb] for

$$U(x) = -\frac{a}{2}x^2 + \frac{b}{4}x^4. \tag{2.5}$$

Exercise 2.2.2. Build a simulation snippet for a frictionless version of the swing equation in the case where there are two nodes joined by a single edge (2.3),

$$\tau\frac{d^2}{dt^2}\theta = p - \beta\sin\theta. \tag{2.6}$$

Consider two regimes (a) $p = 0.1\beta$ and (b) $p = 1.2\beta$. Experimenting with the initial conditions for θ and $\omega := \dot{\theta}$, and also with the parameter τ study (1) the evolution in time of θ and ω; (2) the phase diagram (plotting ω vs θ as they evolve with time). Do you see a difference between the two regimes? Describe your findings.

Solution. Coming Soon

2.2.3 Partial Differential Equations

Mathematicians began studying partial differential equations (PDEs) in the 18th century when trying to explain various physical phenomena occuring in two or more dimensions, (either two or more spatial dimesions, or both a temporal dimension and one or more spatial dimensions). PDEs, like the phenomena they attempt to model, can exhibit rich spatial or spatio-temporal behavior, and although their solutions cannot usually be expressed analytically, various tools have been developed to extract quite a lot of information about a problem without actually solving it.

We will use the dynamics of fluids to illustrate what one of these methods, dimensional analysis, can say about PDEs. Fluid flow is often modeled by the Navier-Stokes equations

$$\partial_t \boldsymbol{u} + (\boldsymbol{u}\cdot\boldsymbol{\nabla})\,\boldsymbol{u} = -\boldsymbol{\nabla}p + \nu\boldsymbol{\nabla}^2\boldsymbol{u}, \tag{2.7}$$

$$\partial_t \rho + \boldsymbol{\nabla}\cdot(\rho\boldsymbol{u}) = \kappa\boldsymbol{\nabla}^2\rho, \tag{2.8}$$

$$p = c_s^2\rho \tag{2.9}$$

The independent variables in (2.7-2.9) are time and d-dimensional space, which are represented by $t \in \mathbb{R}$ and $\boldsymbol{r} \in \mathbb{R}^d$ respectively. The dependent variables are the fluid's velocity, pressure and density, which are represented by $\boldsymbol{u}(t;\boldsymbol{r}) \in \mathbb{R}^d$, $p(t;\boldsymbol{r}) \in \mathbb{R}$ and $\rho(t;\boldsymbol{r}) \in \mathbb{R}$ respectively. The nabla operator, $\boldsymbol{\nabla}$, is standard notation for the d-dimensional vector of partial derivatives and $\cdot$ indicates the scalar product operator (so the fluid's inertia vector is given by $(\boldsymbol{u}\cdot\boldsymbol{\nabla})\,\boldsymbol{u}$ and has components $\sum_{j=1}^{d} u_j(t;\boldsymbol{r})\,\partial_{r_j}u_i(t;\boldsymbol{r})$). The linear relation between p and ρ is due to the so-called "ideal gas" thermodynamic (equation of state) relation, where c_s is a constant characterizing speed of sound in the ideal gas. The constants ν and κ characterize the fluid's viscosity and diffusion coefficient resepctively.

If a fluid is incompressible (like water under normal temperature conditions), then the density field homogenizes over the entire system very fast (with the speed of sound) and we are in the regime when ρ is a constant (does not depend on t and $\boldsymbol{r}$). According to Eq. (2.8) this is possible if the following, so-called "incompressibility" condition is maintained

$$\forall t, \quad \forall \boldsymbol{r}: \quad (\boldsymbol{\nabla} \cdot \boldsymbol{u}) = 0. \tag{2.10}$$

On the other hand, the incompressibility condition dictates that the pressure and velocity fields are related to each other according to

$$\forall t, \quad \forall \boldsymbol{r}: \quad \boldsymbol{\nabla}^2 p = \sum_{u,\beta=1,\cdots,d} \left(\nabla^\alpha u^\beta\right) \left(\nabla^\beta u^\alpha\right), \tag{2.11}$$

which is derived by applying $\boldsymbol{\nabla}$ to Eq. (2.7) and then accounting for Eq. (2.10). Eq. (2.11) may be considered as implicitly expressing the pressure via the velocity. Under this consideration, i.e. finding the pressure field corresponding to a given velocity field, Eq. (2.11) is a Poisson equation.

For both compressible and incompressible fluids, the Navier-Stokes equations (2.7-2.9) present one of the most fundamental problems in science and engineering, that of turbulence. Turbulence occurs when the the fluid velocity becomes sufficiently large, and the flow transitions from a regime where flow patterns are quiet and simple, (called the laminar regime) to a regime when flow patterns are violent and complex, which can be observed when looking at patterns in the sky or in swirls of running water in a river. The complexity of turbulent flow is both spatial and temporal; eddies of many spatial scales emerge and evolve over many temporal scales.

Turbulence is the most mysterious phenomenon associated with the Navier-Stokes equations, and it is still very far from being fully understood. Many great scientists, starting from Leonardo de Vinci and including giants of last century such as Richardson, Kolmogorov, Landau, Heisenberg, Fermi, and Batchelor, have attempted to solve it, but with only limited success. We will demonstrate how dimensional analysis of the Navier-Stokes equations in the incompressible regime[1] can extract some insight on the laminar-to-turbulence transition, and make some sense of the many spatio-temporal scales.

First we establish the dimensionality of the fields entering the formulation. The basic dimensional characteristics are obviously time and length. If time is measured in seconds, $[t] = s$, and the spatial length is measured in meters, $[x] = m$, then the dimensionality of the velocity, $[\boldsymbol{u}]$, should be measured in m/s. A major principle of the dimensional analysis is *"only physical quantities having the same dimension, called commensurable quantities, may be compared, equated, added, or subtracted"*. Therefore each term in Eq. (2.7) must

[1] Note that the concept of the dimensional analysis is attributed to Joseph Fourier (1822), see e.g. `https://en.wikipedia.org/wiki/Dimensional_analysis`.

have the same dimensions, and we can equate $[\nu \boldsymbol{\nabla}^2 \boldsymbol{u}] = [\nu][\boldsymbol{u}]/m^2 = [\nu]/(ms)$ with $[\partial_t \boldsymbol{u}] = [\boldsymbol{u}]/s = m/s^2$, which gives $[\nu] = m^2/s$.

Furthermore objects which are measured with the same units can be compared. For example, the spatial extent of the system is one object of this type, call it L. The largest velocity (by absolute value) in a turbulent flow would naturally be associated with L, call it u_L. The dimensionality of L and the dimensionality of u_L are not commensurate, so we cannot compare them, but we can form an object, L/u_L, which describes the time scale associated with the largest spatial scale of the system. Since the only other dimensional object at our disposal is the viscosity coefficient ν, we should try to find the spatial and temporal scales associated with ν.

Given that the dimensions of ν and of $u_L L$ are equal, their quotient must be dimensionless. Informally, this dimensionless number can be said to measure the strength of the turbulence. Formally, it is defined as

$$\mathrm{Re} := u_L L/\nu \tag{2.12}$$

and is called the Reynolds number after Osborne Reynolds. When Re is sufficiently small, the viscous term, $\nu\nabla^2 \boldsymbol{u}$ in Eq. (2.7) is balanced with the other dominant terms (often the pressure-gradient, or internal stress term, $-\nabla p$). The flow is laminar in this viscosity-controlled regime, and all the prominent eddies are roughly comparable, $\sim L$.

The situation changes dramatically when Re exceeds a certain threshold, $\mathrm{Re} \geq \mathrm{Re}_c$, and we observe the emergence of a range of statial scales from the largest, so-called energy containing, scale which is typically $\sim L$, to the smallest, so-called viscous or Kolmogorov scale, η. This qualitative fact follows straightforwardly from a qualitative comparative analysis of different terms in the Navier-Stokes equations at different scales. The value of this threshold depends on many factors, but is typically found to be $\mathrm{Re}_c \sim 10^4 - 10^5$. We can also conclude from this rough comparative analysis that L/η grows with $\mathrm{Re}/\mathrm{Re}_c$, however to derive a more quantitative relation between the two dimensionless quantities, one needs to go significantly beyond our introductory discussion. (See book of U. Frisch [1] on “Turbulence: The Legacy of A. N. Kolmogorov” for extensive discussion of turbulence phenomenology, some mathematically exact results and many remaining challenges.)

We conclude that dimensional analysis of the key quantities entering a PDE, and subsequent comparative analysis of the different terms in the equation, can give useful insight into the underlying spatio-temporal phenomena. Intuition gained through these initial considerations can not replace mathematical analysis of the PDE, but proves to be extremely valuable in guiding it.

Exercise 2.2.3. Burgers' equation

$$\partial_t u + u\partial_x u = \nu \partial_x^2 u, \tag{2.13}$$

is a PDE appearing in various areas of applied mathematics, such as fluid mechanics, nonlinear acoustics, gas dynamics, traffic flow, and others. Here in Eq. (2.13), $u(t;x) : \mathbb{R}^2 \to \mathbb{R}$, and ν is the viscosity coefficient, which is a dimensional constant. Burgers equation describes the nonlinear dynamics of a traveling wave which steepens into a shock. Assume that a nonlinear traveling wave, representing a shock, has the form, $u(t;x) = cf((x-ct)/\delta)$, where, c and δ are dimensional constants that describe the velocity and the width of the shock respectively, and $f(y)$ is a dimensionless function of a dimensionless parameter, y. Establish (i) the dimensionality of the constants c, δ and ν, and (ii) find a relation between c, δ and ν.

Solution. To find the dimensionality of c, notice the expression $x-ct$ implies the dimensions of x must be the same as the dimensions of ct. That is, $[x] = [ct]$, and therefore $[c] = [x]/[t]$. Next, the expression $(x-ct)/\delta$ implies that $[\delta] = [x-ct]$ because we are given that $y := \frac{x-ct}{\delta}$ is dimensionless. Therefore, $[\delta] = [x]$. Finally, given that $\frac{\nu}{c\delta}$ is dimensionless, we must have $[\nu] = [x]^2/[t]$.

The nonlinear traveling wave form $u(t;x) = cf(\frac{x-ct}{\delta})$ is substituted into the PDE to give

$$cf \cdot (-\frac{c}{\delta}) + cfcf'\frac{1}{\delta} = \nu\frac{c}{\delta^2}f'' \tag{2.14}$$

Hence,

$$-f + ff' = \frac{\nu}{c\delta}f'' \tag{2.15}$$

Since f and its independent variable y are both dimensionless, so are f' and f''. Hence, $\frac{\nu}{c\delta}$ is dimensionless, ν, c and δ can be related by $\nu \sim c\delta$. □

2.3 Optimization and Control

So far we have discussed equations and we now turn to optimization problems. Formulating optimization problems are typically more involved than formulating systems of equations. Furthermore, their solutions are typically more complex, and require more careful interpretation and analysis.

Optimization problems are often discussed during a first semester calculus course where students learn that the optimizers of an objective function, say $f(x)$, over some set, say $A \subseteq \mathbb{R}$, can only occur where the derivative $f'(x)$ is zero (or undefined), or on the set boundary ∂A. In practice, optimization problems involve high (or infinte) dimensional variables with complex constraints. Nonetheless, the ‘first-order’ condition for optimality generalizes to these more challenging problems. We intend to keep our comments very general, and to wait until the spring semester to address important details like, “what, why and how”. By introducing this subject here, even though briefly, we establish the connections between the differential equations (to be studied during the fall semester) and optimization problems (to be studied in the spring semester).

2.3.1 What is an optimization problem?

We will introduce several examples illustrating applications for optimization problems: minimal cost, minimal error, optimal design, variational principles and optimal control. Here is a framework to use when thinking about these examples.

Problem formulation

- What are the basic features involved? (variables, cost, constraints)
- Are there any tricks available to express features differently?
- How can the problem be classified or categorized?

Solution techniques

- What is meant by a “solution” (analytic = formula, numerical = algorithm
- What to look for in setting up an optimization? , something in between)?
- How can the solution be determined? What are tools available?
- What is the complexity of the algorithm? Is it parallizable?
- What is the accuracy of the algorithm?
- What is quality measure of the algorithm? How to compare algorithms?

Analysis

- What are theoretical tools available for analysis?
- When do solution exist? How many are there? Counting? (Remember similar discussion for equations.)
- Can we certify that our solution is sufficiently close to the global optimum?

- How can solutions be categorized?
- What happens to solutions under perturbation? Where does it work and where it does not?

2.3.2 Minimal Cost (Operations Research and Engineering)

Consider the transport of a commodity (for example oil) over a network of pipes that connect a set of terminals (e.g. refinineries or industial centers, etc). We model the network as a graph, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of nodes (terminals) and $\mathcal{E}$ is the set of edges (pipes). The commodity is injected at some nodes, $\mathcal{V}_{in} \subset \mathcal{V}$, and extracted at other nodes, $\mathcal{V}_{out} \subset \mathcal{V}$.

Let $\boldsymbol{x} = (x_i \mid i \in \mathcal{V})$ be the vector representing the quantity injected or extracted at the nodes, so that $\boldsymbol{x}_{in} = (x_i \in \mathbb{R}_+ \mid i \in \mathcal{V}_{in})$, and $\boldsymbol{x}_{out} = (x_j \in \mathbb{R}_- \mid j \in \mathcal{V}_{out})$. We assume that $\boldsymbol{x}$ is known. If there are no losses in the system, so if there is no *net* injection or extraction in the system, then

$$\sum_{i \in \mathcal{V}_{in}} x_i + \sum_{j \in \mathcal{V}_{out}} x_j = 0. \tag{2.16}$$

and we say the the system is closed.

Let $\boldsymbol{\phi} = (\phi_{ij} \mid \{i,j\} \in \mathcal{E})$ be the quantity flowing from node i to j. The quantity flowing through the pipes can be prescribed in an arbitrary way, provided that the total balance at the nodes is maintained, that is,

$$\forall v_i \in \mathcal{V}: \quad x_i = \sum_{\{i,j\} \in \mathcal{E}} \phi_{ij}, \quad \forall \{i,j\} \in \mathcal{E}: \quad \phi_{ij} = -\phi_{ji} \in \mathbb{R}. \tag{2.17}$$

In most practical applications, there is a capacity constraint, $\bar{\phi}_{ij}$, for the maximum possible flow through each pipe,

$$\forall \{i,j\} \in \mathcal{E}: \quad |\phi_{ij}| \leq \bar{\phi}_{ij}. \tag{2.18}$$

The system of equations (2.16, 2.17) are called the *network flow equations.* Note that a system of network flow equations may have multiple (continuity of solutions) $\boldsymbol{\phi}$ for a given $\boldsymbol{x}$.

If you were an engineer who needed to pick one of the many possibilities, how would you act? One strategy is to price flows through each pipe by introducing a cost, $c_{ij} \geq 0$, per unit flow over the pipe $\{i,j\}$ (assuming for simplicity that the relation betweeen flow and cost is linear). From this perspective, the optimal solution can be found by minimizing the cost (or objective) function, which we write as

$$\boldsymbol{\phi}_* = \arg\min_{\boldsymbol{\phi}} \sum_{\{i,j\} \in \mathcal{E}} c_{ij} |\phi_{ij}|, \quad \text{s.t. Eqs. (2.17, 2.18)}, \tag{2.19}$$

where $\arg\min_{\boldsymbol{\phi}}$ reads as "argument of minimum". This is our first example of an optimization problem.

Optimization problems can be minimization problems (as above) or maximization problems. Every maximization problem can be converted to a minimization problem by switching the sign of the cost function, and vice versa. In general an optimization can be formulated as

$$\begin{cases} \text{Given an objective function } f : A \to \mathbb{R}, \text{ where } A \subseteq \mathbb{R}^n, \\ \text{Find } \boldsymbol{x}_* \in A \text{ such that } f(\boldsymbol{x}_*) \le f(\boldsymbol{x}) \ \forall \boldsymbol{x} \in A. \end{cases} \tag{2.20}$$

In the spring semester, we will discuss a number of optimization problems with varying levels of complexity, and the different ways to solve them. *Linear Programming (LP)* is a type of optimization problem that is often considered the "base" problem in the hierarchy of complexity. An LP is a special case where $f(\boldsymbol{x})$ is linear in $\boldsymbol{x}$ and A is a polytope, meaning that A can be defined through a number linear equalities or inequalities. Formally, an LP is an optimization that can be expressed in the following canonical form

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \boldsymbol{c}^\top \boldsymbol{x}, \qquad \text{subject to } \hat{\boldsymbol{A}}\boldsymbol{x} \preceq \boldsymbol{b} \tag{2.21}$$

where the vectors $\boldsymbol{c}$ and $\boldsymbol{b}$, and the matrix $\hat{\boldsymbol{A}}$ are determined when the problem is formulated, and the vector $\boldsymbol{x}$ represents the variables to be determined, and The symbol $^\top$ is standard notation for transposition and $\preceq$ represents an element-wise inequality between two vectors.

Exercise 2.3.1. Show that solving the minimum cost network flow optimization (2.19) is equivalent to solving an LP.

Solution. Equations (2.18) and (2.19) do not constitute an LP (2.21) as stated because an LP requires that both the objective function and the constraint functions are linear, and (2.19) contains absolute values which are non-linear. Hence we need a trick to replace (2.19) by an equivalent expression without absolute values.

The trick is to rewrite the un-directed graph as a directed graph with twice as many edges (and hence, twice as many variables for optimization). Let's rewrite ϕ_{ij} as $\phi_{ij} = \epsilon^+_{ij} - \epsilon^-_{ij}$, where both $\epsilon^\pm_{ij}$ are non-negative. Therefore, to minimize ϕ_{ij} we can look into the minimum $\epsilon^+_{ij} + \epsilon^-_{ij}$ instead. Therefore, the minimization

$$\arg\min_{\boldsymbol{\phi}} \sum_{i,j} c_{ij}|\phi_{ij}| \qquad \text{is equivalent to} \qquad \arg\min_{\boldsymbol{\epsilon}} \sum_{i,j} c_{ij}(\epsilon^+_{ij} + \epsilon^-_{ij}).$$

which is minimized when one of ϵ^+_{ij} and ϵ^-_{ij} is zero, and the other is equal to the optimal ϕ_{ij} (which may also be zero). The restrictions $\phi_{ij} = -\phi_{ji}$ and $|\phi_{ij}| < \bar{\phi}_{ij}$ can be re-written as

$$\epsilon^+_{ij} = \epsilon^-_{ji}, \quad \epsilon^-_{ij} = \epsilon^+_{ji}, \quad \text{and} \quad \epsilon^+_{ij} + \epsilon^-_{ij} \le \bar{\phi}_{ij}$$

which can be encoded into a matrix equation. □

Optimizing infrastructure flows for power-grids, natural gas pipe systems, district heating/cooling systems, water systems, transportation systems and communication systems can often be built from the minimum cost network flow optimization with further generalizaitons to the objective function and with additional network constraints.

Our discussion in the spring semester will also cover questions related to the computational complexity of different optimization algorithms, i.e. the computational effort needed to get a numerical approximation of a required quality. We will see that algorithms for LPs have nice properties, which can thus be used as a base for solving many more complex optimization problems.

2.3.3 Minimal Error (Data Science and Statistics)

A sizable portion of this course, especially in the fall, will be focused on analytic computations, where the results can be derived analytically with pen and paper or with symbolic software. In a majority of practical applications however, analytic evaluation is infeasible, and we must rely on some kind of approximation. It becomes critically important to estimate the approximation error and to then minimize this error by careful selection of the approximation parameters.

Consider the exemplary least-squares optimization problem. This "minimal error" type problem appears in the context of "fitting" (regression) in data science. Assume that one has input and output data vectors of the same dimensionality, N, $\boldsymbol{x} = (x_i \mid i = 1 \dots, N)$ and $\boldsymbol{y} = (y_i \mid i = 1 \dots, N)$, which are related to each other through a scalar function, $f(\,\cdot\,; \boldsymbol{\beta})$, where $\boldsymbol{\beta} = (\beta_j \mid j = 1 \dots M)$ is a "fitting" parameter. Then a natural minimal error formulation for the best fit becomes

$$\arg\min_{\boldsymbol{\beta}} \sum_{i=1,\dots,N} \Big(y_i - f(x_i; \boldsymbol{\beta})\Big)^2. \tag{2.22}$$

It is very popular, and often sufficent, to choose a function that is linear in the parameter. In this case, $f(x; \boldsymbol{\beta}) \to \sum_{j=1,\cdots,M} \beta_j \psi_j(x)$, where $\psi_j(x)$ are known (chosen basis functions). Then the linear least square optimization becomes

$$\arg\min_{\boldsymbol{\beta}} \sum_{i=1,\cdots,N} \left(y_i - \sum_{j=1,\cdots,M} \beta_j \psi_j(x_i)\right)^2. \tag{2.23}$$

One naturally poses the following question: given the number of parameters M, how many samples N are needed to reconstruct? Here a common key word "over-fitting" emerges. The term is used to emphasize that if N is too large, e.g. more samples than number of parameters, there may be a problem (or not). We will not attempt to answer this question here, leaving it for an extensive discussion later in the course.

The method of least squares is attributed to Gauss (1795) and Legendre (1805) who both introduced it as an algebraic procedure for fitting linear equations to data in the context of astronomy and geodesy during the "age of exploration". There is a profound link between optimization and statistics, which will be discussed in the spring semester. In fact, Gauss made connections between the method of least squares with the principles of probability and the normal distribution.

The general formula (2.23) and the linear least square formulations (2.22) are both examples of unconstrained optimization.

Exercise 2.3.2. Show that the linear least square optimization (2.23) admits the following analytic solution

$$\arg\min_{\boldsymbol{\beta}} \sum_{i=1,\cdots,N} \left(y_i - \sum_{j=1,\cdots,M} \beta_j \psi_j(x_i) \right)^2 = (\hat{\boldsymbol{\Psi}}^T \hat{X})^{-1} \hat{\boldsymbol{\Psi}}^T \boldsymbol{y}, \tag{2.24}$$

where $\hat{\boldsymbol{\Psi}} := (\psi_j(x_i) | i = 1, \cdots, N; j = 1, \cdots, M)$ is $N \times M$ matrix constructed from the input vector, $\boldsymbol{x}$.

Solution. This classical theorem can be found in many books on linear algebra. Here we present the proof from the perspective of optimization. Since $\{y_i\}$ and $\{\psi_j(x_i)\}$ are fixed,

$$\sum_{i=1,\cdots,N} \left(y_i - \sum_{j=1,\cdots,M} \beta_j \psi_j(x_i) \right)^2 \tag{2.25}$$

is a bilinear form for $\{\beta_j\}$. To get its minimum, it is necessary to have all first-order partial derivatives to be 0. That is:

$$\forall j, \quad 2 \sum_{i=1,\cdots,N} \left(y_i - \sum_{j=1,\cdots,M} \beta_j \psi_j(x_i) \right) \psi_j(x_i) = 0 \tag{2.26}$$

Taking y_i terms to right-hand side we get

$$\forall j, \quad \sum_{i=1,\cdots,N} \psi_j(x_i) \left(\sum_{j=1,\cdots,M} \psi_j(x_i) \beta_j \right) = \sum_{i=1,\cdots,N} \psi_j(x_i) y_i \tag{2.27}$$

Equations (2.27) exactly involve all the entries of $\hat{\boldsymbol{\Psi}}$, $\{\beta_j\}$ and y. Furthermore, the right-hand side of (2.27) is the dot product between the j^{th} row of $\hat{\boldsymbol{\Psi}}^\top$ and y. Likewise, the left-hand side is the dot product between the j^{th} row of $\hat{\boldsymbol{\Psi}}^T \hat{\boldsymbol{\Psi}}$ and $\{\beta_j\}$. Collecting them together gives $\boldsymbol{\beta} = (\hat{\boldsymbol{\Psi}}^\top \hat{\boldsymbol{\Psi}})^{-1} \hat{\boldsymbol{\Psi}}^\top \boldsymbol{y}$.

Since the bilinear form is positive definite, we conclude that the stationary point for $\boldsymbol{\beta}$ is its global minimum. □

	Infinite Dimensional	Finite Dimensional
"Variable"	$\{x(t)\}$	$\{x(t) \mid N\} := \{x_n = x(t_n) \mid n = 0, \ldots, N;\ t_n = Tn/N = n\Delta\}$
Objective Function	$\mathcal{S}\{x(t)\} := \int_0^T dt \mathcal{L}\left(x(t); \dot{x}(t)\right)$	$\mathcal{S}(x_0, \cdots, x_N) := \sum_{n=0}^{N-1} \mathcal{L}\left(x_{n+1}, x_n\right)$
Lagrangian	$\mathcal{L}\left(x(t); \dot{x}(t)\right) := \frac{\dot{x}^2}{2} - U(x)$	$\mathcal{L}\left(x_{n+1}; x_n\right) := \frac{(x_{n+1} - x_n)^2}{2\Delta^2} - U(x_n)$
Optimality Condition	$\forall t \in [0, T]: \quad \delta S\{x(t)\} = 0$	$\forall n = 0, \cdots, N-1: \quad \frac{\partial}{\partial x_n} \mathcal{S}\left(x_0, \cdots, x_N\right) = 0$
"Dynamics"	$\forall t \in [0, T]: \quad (2.28)$	$\forall n = 1, \cdots, N: \quad \frac{2x_n - x_{n+1} - x_{n-1}}{\Delta} - \Delta \partial_{\theta_n} U(\theta_n) = 0$

Table 2.1: Comparison between infitite dimesnional and finite dimensional optimization. Note that we do not optimize over x_0 and x_N to reflect on the fact that a second order ODE should be supplemented by two conditions.

2.3.4 Classical Mechanics & the Variational Principle

Variational calculus, or the *calculus of variations*, establishes a relationship between differential equations and optimization problems. For illustration, return to the swing dynamics governed by equation (2.3), switching from notation standard in energy systems, $\theta(t)$, to notation from classical mechanics, $x(t)$. Consider the regime where the inertia of the system dominates the damping, and the latter can be ignored, so

$$\ddot{x} = -\partial_x U(x). \tag{2.28}$$

Equation (2.28) can be restated in its equivalent *variational* form

$$x_*(t) = \arg\min_{\{x(t)\}} \int dt \left(\frac{\dot{x}^2}{2} - U(x) \right). \tag{2.29}$$

To understand the relation between equations (2.28) and (2.29), it is useful to pause and discuss the transition from finite dimensional optimization to infinite dimensional optimization (table 2.1).

The most famous differential equations of physics have an equivalent variational interpretation, including Newton's laws of motion. The principle of least action asserts that the solution to the differential equation coincides with the minimizers of the objective function, or action, $\mathcal{S}\{\boldsymbol{x}(t)\}$. This is not in itself surprising, as such a function could always be constructed. However, the action is often expressed as the integral of a function called a Lagrangian, $\mathcal{L}(\boldsymbol{x}(t); \dot{\boldsymbol{x}}(t))$, which, can usually be derived in a consistent, physically meaningful manner (in classical mechanics, it is always kinetic energy minus potential energy, for example).

Some problems are easier to pose as a differential equation, but easier to solve in their equivalent variational form. Others are easier to to pose as a variational statement, but easier to solve as a differential equation. Our next example will be of the latter.

Exercise 2.3.3. A cable of length ℓ and uniform density per unit length ρ is suspended from two anchors placed sufficiently high off the ground and separated by $2a < \ell$. Assume the cable is flexible (able to bend) and inelastic (unable to stretch). Let $y(x), -a < x < a$ represent the height of the cable at coordinate x. Use a variational formulation to develop a model for y.

Solution. The infinitessimal arc-length of $y(x)$ is given by $ds = \sqrt{1+|y'(x)|^2}dx$. Since we require that the total length of the cable is ℓ,

$$\text{Constraint:} \quad \ell = \int_{-a}^{a} ds = \int_{-a}^{a} \sqrt{1+|y'(x)|^2}\, dx$$

Our model will assume that the cable is suspended in such a configuration so as to minimize its gravitational potential energy, which is given by

$$\text{Objective:} \quad PE = \int_{-a}^{a} \rho g y(s)\, ds = \int_{-a}^{a} \rho g y(x) \sqrt{1+|y'(x)|^2}\, dx \tag{2.30}$$

Using a lagrange multiplier, we incorporate the constraint into the minimization function

$$y_*(x) = \underset{\{y(x)\}}{\arg\min} \int_{-a}^{a} \rho g y(x) \sqrt{1+|y'(x)|^2} + \lambda(\sqrt{1+|y'(x)|^2} - \ell)dx \tag{2.31}$$

We will learn how to derive the Euler-Lagrange equations, which for this problem are

$$\mathcal{L}_y - \frac{d}{dx}\mathcal{L}_{y'} = 0,$$

The solution to this is differential equation is.

$$y(x) = \lambda + a\cosh(x/a) \tag{2.32}$$

where the value of λ can be determined from ℓ and from a. □

Note that many equations of applied mathematics are associated with results of an unconstrained optimization. For an equation of the form $F(x) = 0$, involving a mapping $F : \mathbb{R}^n \to \mathbb{R}^n$, a variational principle is an expression of F as the gradient mapping ∇f associated with some function $f : \mathbb{R}^n \to \mathbb{R}^n$. Such an expression leads to the interpretation that the desired x satisfies a first-order optimality condition with respect to f. Under certain additional conditions on F, it may be concluded that x minimizes f, at least “locally.” Such variational reformulation is useful as it opens up the problem to many analytic and numerical benefits of optimization.

2.3.5 Control Theory

Our last exemplary problem in this section will be one from the field of **Control Theory**, which deals with systems whose evolution can be influenced by some external agent. Consider control systems that can be defined as a system of differential equations depending on some parameters $u(t)$ changing in time

$$\dot{x} = f(x(t), u). \tag{2.33}$$

For each initial point $x(0) = x_0$ there are many trajectories depending on the choice of the control parameters u assuming that they depend on time, $u(t)$. A classic question in control is the so-called "controllability" problem: given equation (2.33), describe which x can be reached starting from x_0. (Notice the relation between the controllability of a control problem and the feasibility of equations.) If controllability to a final point $x(T) = x_1$ is granted, one can try to reach x_T minimizing some cost, thus defining an **Optimal Control Problem**:

$$\min_{\{u(t)\}} \int\limits_0^T dt \mathcal{L}(x(t), u(t)) \text{ s.t. Eq. (2.33)}, \ x(0) = x_0, \ x(T) = x_T, \tag{2.34}$$

where $\mathcal{L}$ is the Lagrangian or running cost. Notice the similarity with the principle of least-action from classical mechanics, $\delta\mathcal{S} = 0$. The similarity and relations between the two subjects are deep and optimal control theory can be thought of as a generalization of the calculus of variations. We will discuss the details of this relation in the spring semester.

Exercise 2.3.4. A factory produces $x(t) \geq 0$ products per unit of time. The owner can reinvest some fraction, $0 \leq u(t) \leq 1$, of the factory output, which boosts the production rate according to

$$\begin{cases} \dot{x}(t) = k\, u(t)\, x(t), \\ x(0) = x_0, \end{cases} \tag{2.35}$$

where $x_0 > 0$ is the initial value of $x(t)$ and the constant, $k > 0$, defines efficiency of the reinvestment. The products are sold making a net profit

$$P = \int\limits_0^T (1 - u(t))\, x(t) dt. \tag{2.36}$$

Find the optimal control, $u_*(t)$, $0 \leq t \leq T$ to maximize the profit (2.36) accumulated over time T.

Solution. Start by constructing the solution intuitively. If T is very small, there may not be sufficient time to both re-invest the product and then sell whatever product is accumulated by the re-investment, so it may be best to sell the initial product for profit. However, if T is sufficiently large, it may be advantageous to re-invest the early production and recooperate profits later.

The optimal control problem is expressed as a constrained optimization problem (constrained by the dynamics of the governing ODE, (2.35)),

$$\max_{\{u(t),x(t)\}} \int_0^T (1-u(t))x(t)dt \Bigg|_{\forall t\in[0,T]:\ \dot{x}(t)=u(t)kx(t), 0\le u(t)\le 1}.$$

The constrained optimization can be re-fomulated as min-max problem by using a Lagrange multiplier, λ to include the constraint in the objective function. The constraints on the optimal control are similarly included in the objective function with associated lagrange multipliers μ_1 and μ_2

$$\max_{\{u(t),x(t)\}} \min_{\{\lambda(t),\mu_1(t),\mu_2(t)\}} \int_0^T dt L\left(t, x(t), u(t), \lambda(t), \mu_1(t), \mu_2(t)\right)$$

where

$$L := \left((1-u(t))x(t) + \lambda(\dot{x} - kux) + \mu_1(u-1) - \mu_2 u\right),$$

and

$$\mu_1(t), \mu_2(t) \ge 0 \quad \forall t \in [0,T].$$

The Euler-Lagrange (KKT) conditions are

$$\begin{aligned} \forall t \in [0,T]: \qquad & 1 - u - \dot{\lambda} - \lambda k u = 0 \\ & -x - \lambda k x + \mu_1 - \mu_2 = 0. \end{aligned}$$

The complementary slackness conditions are

$$\forall t \in [0,T]:\ \mu_1(u-1) = 0,\ \mu_2 u = 0.$$

We observe that μ_1 and μ_2 cannot be zero simultaneously (as KKT are in a contradiction) and thus for each $t \in [0,T]$, either $u(t) = 0$ or $u(t) = 1$. One derives

$$\begin{aligned} u - 1 = \mu_2 = 0: \quad & \dot{\lambda} + k\lambda = 0, \\ u = \mu_1 = 0: \quad & \dot{\lambda} = 1, \end{aligned}$$

where (according to our intuitive description above) the two solutions are realized short, $t \in [0, \tau]$, and long, $t \in [\tau, T]$, times respectively. Gluing the two solutions together and substituting the result in optimization formulation above we arrive at

$$\alpha_*(t) = \begin{cases} 1 & t < \tau^* \\ 0, & \text{otherwise} \end{cases}$$

$$x^*(t) = \begin{cases} x_0 \exp(tk), & t < \tau^* \\ x_0 \exp(\tau^* k), & \text{otherwise} \end{cases}$$

$$P^*(t = T) = \begin{cases} x_0 T & T \leq 1/k \\ x_0 \exp(kT - 1)/k & T \geq 1/k \end{cases}$$

where $\quad \tau^* = \max\{0, T - 1/k\}$. □

2.4 Mathematics of Uncertainty

The objective of this section is to leapfrog to the last portion of the core courses, where we discuss probability, stochastic processes, statistics and data science, all of which are branches of mathematics dealing with uncertainty.

2.4.1 Why Data Science (and Related)?

We devote a significant portion of our applied mathematics core curriculum to probability, statistics and data science because we find it hard to imagine a modern applied mathematics researcher not touching upon the subjects of data driven uncertainty modeling and prediction.

We will use a few examples to illustrate how to work with uncertainty. When discussing differential equations and optimizations in sections 2.2 and 2.3, we assumed that all the parameters in the problem formulation were known, and that the model assumptions did not need to be questioned. In practice, this is rarely the case, and in most real-world applications, many elements of the model formulation are uncertain.

2.4.2 Maximum Likelihood and Marginal Probability

We begin by revisiting the example on minimal error optimization formulation, discussed in the previous section on optimization, and we now look at it through probabilistic lens.

Consider a dependent, or output, variable $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^\top$ and a set of p explanatory, or input, variables $\boldsymbol{x}_j = (x_{1j}, x_{2j}, \ldots, x_{nj})^\top$, for $j = 1, \ldots, p$. A linear regression relation between $\boldsymbol{y}$ and $\hat{\boldsymbol{X}}$ can be written as

$$\boldsymbol{y} = \hat{\boldsymbol{X}}\boldsymbol{\beta} + \boldsymbol{\xi}, \tag{2.37}$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_p)^\top$ are the (as-of-yet unknown) weights in the linear model, and $\boldsymbol{\xi} = (\xi_1, \xi_2, \dots, \xi_n)^\top$ is a random vector that models the uncertainty, or the error, of the linear relation. The components of $\boldsymbol{\xi}$ are independent white-noise Gaussian random variables (zero mean and variance σ^2) described by the following probability density function

$$P(\boldsymbol{\Xi} = \boldsymbol{\xi}) = \prod_{i=1,\cdots,n} P(\Xi_i = \xi_i), \quad P(\Xi = \xi) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\xi^2}{2\sigma^2}\right). \tag{2.38}$$

The random variable $\boldsymbol{\Xi}$ represents errors which are added on the way from input to output[2]. Expressing $\boldsymbol{\xi}$ via $\hat{\boldsymbol{X}}, \boldsymbol{y}$ and $\boldsymbol{\beta}$, according to Eq. (2.37), substituting the result into Eq. (2.38), and then taking logarithm of the result one arrives at an expression for $\boldsymbol{\beta}$ called the log-likelihood

$$\mathcal{L}(\boldsymbol{\beta}) = -\frac{\|\boldsymbol{y} - \hat{\boldsymbol{X}}\boldsymbol{\beta}\|_2^2}{2\sigma^2} - \frac{N}{2}\log(2\pi\sigma^2), \tag{2.39}$$

which depends only on $\boldsymbol{\beta}$ and measures likelihood of $\boldsymbol{\beta}$. Taking the maximum of the log-likelihood over $\boldsymbol{\beta}$ results in the least-square formula (2.23).

Exercise 2.4.1. Generalize the approach just explained, leading from Eq. (2.37) to Eq. (2.39), and then resulting in Eq. (2.23), to the case when the random error vector, $\boldsymbol{\xi}$ in Eq. (2.37), is governed by the Gaussian probability density function of a general position

$$P(\boldsymbol{\Xi} = \boldsymbol{\xi}) = \frac{\exp\left(-\frac{1}{2}\sum\limits_{i,j=1,\cdots,N} (\xi_i - \mu_i)(\Sigma^{-1})_{ij}(\xi_j - \mu_j)\right)}{(2\pi)^{N/2}\sqrt{\det(\hat{\boldsymbol{\Sigma}})}} \tag{2.40}$$

characterized by the vector of means, $\boldsymbol{\mu}$, and by the symmetric, positive definite matrix of co-variance, $\hat{\boldsymbol{\Sigma}}$.

Solution. From (1.27) we have $\xi = \boldsymbol{y} - \hat{X}\boldsymbol{\beta}$. Getting this to (1.30) we have

$$P(\boldsymbol{\beta}) = \frac{\exp\left(-\frac{1}{2}\left(\boldsymbol{y} - \hat{X}\boldsymbol{\beta} - \boldsymbol{\mu}\right)\Sigma^{-1}\left(\boldsymbol{y} - \hat{X}\boldsymbol{\beta} - \boldsymbol{\mu}\right)\right)}{\left(2\pi\det(\hat{\Sigma})\right)^{N/2}} \tag{2.41}$$

[2] By convention, upper case variables denote random variables, e.g. Ξ. A random variable takes on values in some domain, and if we want to consider a particular instantiation 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. ξ.

Hence, the log-likelihood is

$$\mathcal{L}(\boldsymbol{\beta}) = -\frac{1}{2}\left(\boldsymbol{y} - \hat{X}\boldsymbol{\beta} - \boldsymbol{\mu}\right)\Sigma^{-1}\left(\boldsymbol{y} - \hat{X}\boldsymbol{\beta} - \boldsymbol{\mu}\right) - \frac{N}{2}\log(2\pi\det(\hat{\Sigma})) \tag{2.42}$$

□

More generally than for the case of the minimal error, the maximum likelihood projects from the world of probability to the world of optimization. Let us see how it works on an example of a joint probability distribution over two variables $P(X_1 = x_1, X_2 = x_2)$, say

$$P(X_1 = x_1, X_2 = x_2) = Z^{-1}\exp\left(-\frac{x_1^2 + x_2^2 + x_1 x_2}{2}\right), \tag{2.43}$$

where $x_1, x_2 \in \mathbb{R}$ and Z is a normalization constant called the "partition function" which guarantees that when we integrate over all possible x_1 and x_2 the total probability for all events is unity, that is, $\int dx_1 dx_2 P(X_1 = x_1, X_2 = x_2) = 1$. We did not write down Z in this example explicitly because knowing this coefficient will not be needed for estimating the maximum likelihood

$$\arg\max_{x_1, x_2 \in \mathbb{R}} P(X_1 = x_1, X_2 = x_2), \tag{2.44}$$

where arg max denotes the values of x_1 and x_2 at which the objective, $P(X_1 = x_1, X_2 = x_2)$, is maximized.

Marginalization of a multi-variate probability distribution is another very useful operation. In probability, statistics and related topics, the marginal distribution of a subset of a collection of random variables is the probability distribution of the variables contained in the subset. It gives the probabilities of various values of the variables in the subset without reference to the values of the other variables. (Notice the contrast with a conditional distribution, which gives the probabilities contingent upon the values of the other variables.) In the bi-variate case, $P_{X_1,X_2}(x_1, x_2)$ marginalization over one of the variables, say x_1, results in the marginal probability over another,

$$P(X_1 = x_1) = \sum_{x_2 \in \Sigma_2} P(X_1 = x_1, X_2 = x_2). \tag{2.45}$$

Here Σ_2 is the domain of x_2 and we assume that Σ_2 is countable (or even finite. In the case of a continuous variable, say with $\Sigma_2 \subset \mathbb{R}$, summation in the marginalization formula should be replaced by an integration.)

2.4.3 Stochastic ODEs

Probabilistic modeling of uncertainty can be extended from algebraic equations to differential equations. Imagine that we monitor a process which is "almost" described by an ODE.

Consider for example,

$$\dot{x} = v(t;x) + \xi(t), \tag{2.46}$$

where $x(t) \in \mathbb{R}$ is, say, the position of a particle at time t, and $v(t;x(t))$ is a deterministic function of t and $x(t)$. The uncertainty (the "almost" component of our formulation) is modeled by a random source, $\xi(t)$.

A popular choice of $\xi(t)$ is a continuous time white-noise Gaussian signal, which can be derived from the random white Gaussian vector by carefully taking a limit of the components of the vector defined for discrete time in Eq. (2.40) to make the transition to continuous time. Jumping over the details, the probability measure of the time-discretized white-noise Gaussian signal

$$\{\xi(t)\,|\,N\} \equiv \{\xi(t_n)\,|\,n = 1, \dots N; t_n = Tn/N = n\Delta\} \underset{N\to\infty}{\longrightarrow} \{\xi(t)\,|\,0 \le t \le T\} \tag{2.47}$$

is

$$\begin{aligned} P\Big(\{X(t);N\} = \{x(t);N\}\Big) &= \left(\frac{\Delta}{2\pi\sigma^2}\right)^{N/2} \exp\left(-\frac{\Delta}{2\sigma^2}\sum_{n=1}^{N}\xi_n^2\right) \\ \underset{N\to\infty}{\longrightarrow} P\Big(\{X(t)\} = \{x(t)\}\Big) &\sim \exp\left(-\frac{1}{2\sigma^2}\int\limits_0^T (\xi(t))^2\, dt\right), \end{aligned} \tag{2.48}$$

where $\sim$ indicates that the continuous-time probability density is defined up to a normalization factor called the partition function.

Direct substitution of $\xi(t)$, expressed via $x(t)$ according to Eq. (2.46), into Eq. (2.48) results in

$$P\Big(\{X(t)\} = \{x(t)\}\Big) \sim \exp\left(-\frac{1}{2\sigma^2}\int\limits_0^T (\dot{x} - v(t;x))^2\, dt\right). \tag{2.49}$$

We have just arrived at the Feynman-Kac "path-integral" for the realization (or trajectory) $\{x(t) \mid 0 \le t \le T\}$. The stochastic ODE defined by Eqs. (2.46,2.48), and the path-integral (2.49), provide complementary representations for the probability density of the path, $\{x(t)\}$.

One asks, naturally, to compute the marginal probability density of observing the particle at a particular moment of time, say T, and at a particular position, say x_T provided that the particle has started at the position x_0 at $t = 0$, (that is, to find $P(X(T) = x_T)$ given that $X(0) = x_0$). The respective marginal probability density, also called transition probability density from x_0 to $x(T)$, becomes

$$\begin{aligned} &P\Big(X(T) = x_T \mid X(0) = x_0\Big) \\ &\qquad \underset{N\to\infty}{\longleftarrow} \int dx_1 \cdots dx_{N-1} P\Big(\{X(t);N\} = \{x(t);N\} \mid X(0) = x_0\Big). \end{aligned} \tag{2.50}$$

The transition probability (density), considered as a function of $x = x_T$, can be shown to satisfy the PDE

$$\left(\partial_t + \partial_x v(x) - \frac{\sigma^2}{2}\partial_x^2\right) P\Big(X(T) = x \mid X(0) = x_0\Big) = 0, \tag{2.51}$$

called the Fokker-Planck, the forward Kolmogorov, or the advection-diffusion Smoluchowski equation, where $\partial_x vP$ is the advection term and $(\sigma^2/2)\partial_x^2 P$ is the diffusion term.

The coefficient $\sigma^2/2$ is a diffusion constant, and when the advective term is dropped (by setting v to zero), the stochastic ODE (2.46) describes the evolution of a Brownian particle, and the Fokker-Planck equation (2.51) describes the diffusive spread of the probability distribution of the particle's position.

Generally speaking, transitioning from a deterministic description to a stochastic description adds a dimension. This is because the deterministic version of a stochastic ODE (2.46) describes the evolution of a particle position in time, whereas the Fokker-Planck Eq. (2.51) describes evolution of the probability distribution of the particle's position in time and space.

The Fokker-Planck equation in "real" time is akin to the Schrödinger equation of Quantum Mechanics in "imaginary" time. The stochastic evolution of a particle is akin to many non-classical trajectories accounted for by quantum mechanics beyond classical mechanics.

Exercise 2.4.2. Use dimensional analysis in the regime where only diffusion is present, (i.e. $v = 0$), to estimate how the position $|x|$ of a typical particle grows with time, t, as $t \to \infty$. (b) In the regime where advection stabilizes the diffusive spread, the velocity depends on x, but not on t, (i.e. $v = v(x)$), find a relationship between $v(x_*)$ and x_*, where $|x_*|$ is the typical position of a stabilized particle.

Solution. Dimensional analysis can give a rough idea about the variables. (a) Balancing the two terms on the left-hand-side of Eq. (2.51) gives $1/t_* \sim \sigma^2/x_*^2$, concluding that $x_*^2 \sim \sigma^2 t_*$. (b) Stationarity means that $\partial_T P = 0$, thus the two terms we want to balance in this case give us, $v(x_*)/x_* \sim \sigma^2/x_*^2$, thus resulting in $v(x_*) \sim \sigma^2/x_*$ for large $|x_*|$. □

Markov Processes

A stochastic process is called "memoryless", or Markov (after Andrey Markov, 1856-1922), if the predictions for the future of the process that are based solely on its present state are just as good as the predictions based on the process's full history. Rephrasing, conditioned on the present state of the system, the past and future states of a Markov process are independent:

$$P\Big(X_{n+1} = x \mid X_1 = x_1; \ldots; X_n = x_n\Big) = P\Big(X_{n+1} = x \mid X_n = x_n\Big). \tag{2.52}$$

The stochastic ODE satisfying equation (2.46) and driven by Gaussian white noise described by equation (2.48) generates continuous time, continuous space Markovian paths $\{x(t)\}$.

How would one deal with a discrete-time discrete-space Markov Process? This is where our brief discussion of the linear Fokker-Planck equation for the evolution of the probability of being in a particular state, $X_n = x$, at the moment of time labeled by n, becomes useful. One expects the probability to depend linearly on the probability one step earlier:

$$P\Big(X_{n+1} = x_{n+1}\Big) = \sum_{x_n \in \Sigma} P\Big(X_{n+1} = x \mid X_n = x_n\Big) P\Big(X_n = x_n\Big), \tag{2.53}$$

where $\Sigma = \{s_1, \cdots, s_K\}$ is a finite set of allowed states for x_n, and $P\left(X_{n+1} = x \mid X_n = x_n\right)$ is the transition probability describing transition from a state $x_n \in \Sigma$ to a state $x_{n+1} \in \Sigma$. Assuming that cardinality of Σ is K, $|\Sigma| = K$, and that it does not change in time, one builds the so-called transition matrix

$$\hat{\Pi}_{n+1} \equiv \begin{pmatrix} P\Big(X_{n+1} = s_1 \mid X_n = s_1\Big) & \cdots & P\Big(X_{n+1} = s_K \mid X_n = s_1\Big) \\ \vdots & \ddots & \vdots \\ P\Big(X_{n+1} = s_K \mid X_n = s_1\Big) & \cdots & P\Big(X_{n+1} = s_K \mid X_n = s_K\Big) \end{pmatrix} \tag{2.54}$$

In general, the matrix may change in time, however we will limit our discussion here to the case of a stationary Markov process, where $\hat{\Pi}_n$ does not depend on n. Obviously, all elements of $\hat{\mathbf{\Pi}}$ are positive real and bounded between 0 and 1, as each corresponds to a probability. Moreover, due to normalization condition for $P(X_{n+1} = x_{n+1})$ the transition probability matrix satisfies the following normalization constraint

$$\forall k = 1, \cdots, K: \quad \sum_{k'=1}^{K} \Pi_{k'k} P\Big(X_{n+1} = k' \mid X_n = k\Big) = 1. \tag{2.55}$$

(A positive matrix satisfying Eq. (2.55) is called stochastic.)

It is also be useful to explain the transition probability matrix in terms of a directed graph (thus borrowing notations, and also results, from the graph theory). Let us use the example from `https://en.wikipedia.org/wiki/Markov_chain` to illustrate the construc-

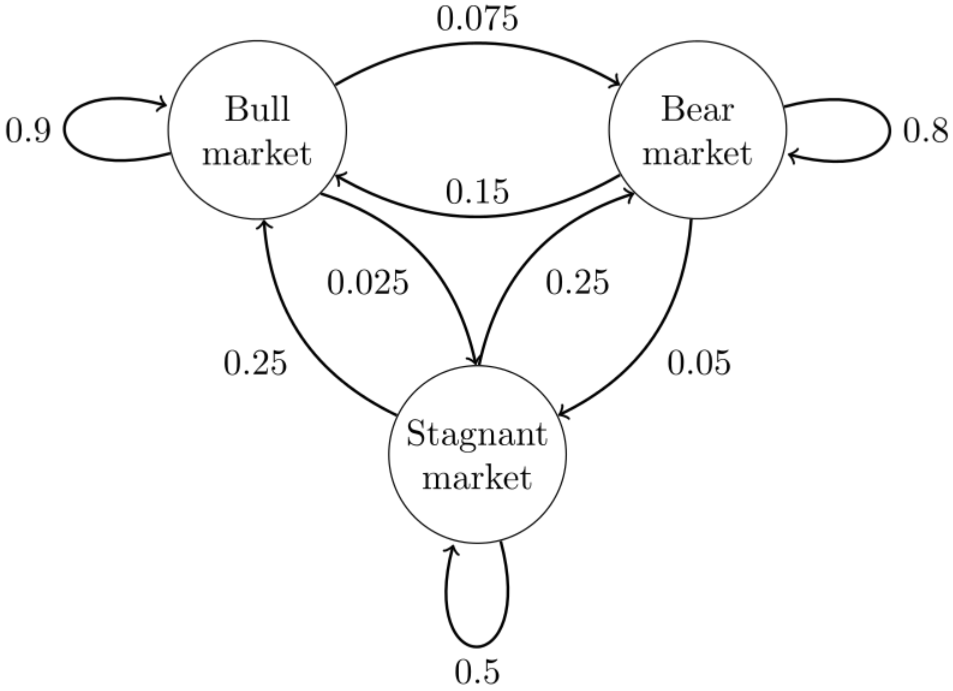

Figure 2.1: Hypothetical example of the stock market exhibiting a bull market, bear market, or stagnant market trend during a given week.

tion of a transition probability matrix

$$\begin{aligned}\hat{\mathbf{\Pi}} &\equiv \big(P\left(X_{n+1}=k' \mid X_n=k\right) \mid k',k=\{\text{bull market}, \text{bear market}, \text{stagnant market}\}\big)\\ &= \begin{pmatrix} 0.90 & 0.075 & 0.025 \\ 0.15 & 0.80 & 0.05 \\ 0.25 & 0.25 & 0.50 \end{pmatrix},\end{aligned} \tag{2.56}$$

visualized with a directed graph shown in the Fig. (2.1). The states represent whether a hypothetical stock market is exhibiting a bull market, bear market, or stagnant market trend during a given week. According to the figure, a bull week is followed by another bull week 90% of the time, a bear week 7.5% of the time, and a stagnant week the other 2.5% of the time.

Studying discrete-time, discrete space Markov Processes, one is usually interested in answering the following questions:

- Given an initial state (or distribution over states) how does the probability distribution evolve with time?
- Does it converge to a steady distribution?
- Will the steady distribution depend on the initial state?
- What is the time of convergence (so-called) mixing time?

Exercise 2.4.3. For the example of the hypothetical stock market described by Eq. (2.56) and Fig. (2.1), write a compute program and study the evolution of the probability vector, $\boldsymbol{P}_{n+1}=\hat{\mathbf{\Pi}}^n\boldsymbol{P}_1$, with n and its dependence on the initial probability vector, $\boldsymbol{P}_1$, where

$$\forall n: \quad \boldsymbol{P}_n = \begin{pmatrix} P(X_n=\text{bull market}) \\ P(X_n=\text{bear market}) \\ P(X_n=\text{stagnant market}) \end{pmatrix} \tag{2.57}$$

(Remember that components of P_1 should be non-negative and normalized.) Analyze the convergence of the process, and explain the results.

Solution. In progress

2.4.4 Markov Decision Processes

There are situations where the transition probability matrix can be manipulated. In these situations, it can also be of interest to study the dependence of the aforementioned characteristics on some principle properties of the transition matrix, such as its size, spectrum, etc. This engineering design approach to Markov Processes naturally lead us to an optimization formulation, called a **Markov Decision Process**. Consider the exemplary setting:

$$\min_{\hat{\mathbf{\Pi}}_2,\cdots,\hat{\mathbf{\Pi}}_N} \sum_{n=1}^{N-1} \left(\mathcal{M}(\hat{\mathbf{\Pi}}_{n+1};\hat{\mathbf{\Pi}}_t) + \boldsymbol{P}_n^T \boldsymbol{c}_n\right)_{\forall n=1,\cdots,N-1:\ \boldsymbol{P}_{n+1}=\hat{\mathbf{\Pi}}_{n+1}\cdots\hat{\mathbf{\Pi}}_2\boldsymbol{P}_1} \tag{2.58}$$

where $\forall n = 1,\cdots,N:\quad \hat{\mathbf{\Pi}}_n$ is a stochastic matrix now dependent on the time index n; $\mathcal{M}(\hat{\mathbf{\Pi}}_n;\hat{\mathbf{\Pi}}_t)$ stands for a (scalar) function penalizing deviation of the optimization (matrix) variable, $\hat{\mathbf{\Pi}}_n$, from a pre-defined target transition probability matrix, $\hat{\mathbf{\Pi}}_t$; $\boldsymbol{c}_n$ is the cost vector constructed from components representing cost of being in the respective state at the n-th moment. We expect the two terms in the objective of Eq. (2.58) to compete producing an optimal consensus solution.

2.4.5 Data Science: Modeling, Learning and Inference

This section of the course is introduced to reflect on very recent, and exciting, research topics with terminology, methodology and approaches which are not yet fully grounded in the rest of mathematics. The increased availability of data and the development of many new **Machine Learning** methods and algorithms are important driving forces for this new progress. Our modest task here will be to give a primer data science from the perspective of applied mathematics, thus emphasizing its relation to building, analyzing and solving the models we have discussed so far in this course.

Data science research starts by building a parameterized model to fit available data. The values of the parameters are optimized by comparing the model to the available data during the learning or training step. The optimal values of the parameters allow the model make inferences, or computational predictions, about the relevant phenomena. The three steps of are:

$$\text{Modeling} \quad\Rightarrow\quad \text{Learning/Training} \quad\Rightarrow\quad \text{Inference.}$$

The model may be agnostic to the known mechanism governing the specific science domain providing the data, in which case we may use a generic model coming from statistics, such as a regression model. Alternatively we may choose to use a model built primarily on considerations of, say, fluid mechanics, for reducing uncertainty in measurements of velocity of acceleration of particles in the flow.

For a basic neural network, the model is represented by a repeated combination of high-dimensional matrix-vector multiplication/addition and element-wise application of a non-linear 'activation' function.

$$\hat{\boldsymbol{y}} := \boldsymbol{F}(\boldsymbol{x}) = \sigma\left(\hat{\boldsymbol{W}}^{[n]}\sigma\left(\dots\left(\hat{\boldsymbol{W}}^{[2]}\sigma\left(\hat{\boldsymbol{W}}^{[1]}\boldsymbol{x}+\boldsymbol{b}^{[1]}\right)+\boldsymbol{b}^{[2]}\right)\dots\right)+\boldsymbol{b}^{[n]}\right). \tag{2.59}$$

Here $\hat{\boldsymbol{W}}^{[j]}$ and $\boldsymbol{b}^{[j]}$ are matrices and vectors of parameters, and σ is a non-linear function that acts on each element of its vector-valued argument.

A typical formulation of learning is in terms of an optimization as illustrated by the minimal error optimization discussed earlier, e.g. $\arg\min_{\hat{\boldsymbol{W}}^j,\boldsymbol{b}^j}\sum_i ||\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i||_2$. For toy-examples, the optimal parameters can be found with black-box solvers, but most real-world applications require careful model formulation and sophisticated algorithms that are efficient and scalable.

Exercise 2.4.4. (a) Read sections 1-5 in [2]. (Note: the details of algorithms discussed in sections 4 & 5 will be studied in the Spring. For now, it is sufficient to get the gist of what they do.) (b) Download and run the computational snippet `Excercise_1_5_5.ipynb` from D2L. The computational snippet is Julia code expanded on the Matlab code presented in section 6 of [2]. (c) Suggest a suitable means to describe the performance of the network (d) Study experimentally the dependence of the performance of the network on the number of samples in the training set, and report your findings. (e) Suggest (but do not implement) efficient ways to improve your ability to learn the function f.

We conclude this brief discussion of Neural Networks summarizing important features we have started to discuss

- construction of the model includes repetitive application of a chain of parameterized nonlinear functions;
- over-parametrization (more parameters than data points) can be ok (or not);
- training is an optimization;
- tests need to be developed to access quality of the Neural Network reconstruction.

There are many other important features of Neural Networks, like, convolution, randomization, iterative algorithms, other flavors of learning (not only supervised learning as in the example, but also unsupervised learning, clustering, reinforcement learning, etc), and many students are exploring these new avenues of research.

2.4.6 Graphical Models

A graphical model is a probabilistic model where the dependence structure between the random variables is expressed by a graph. In this section, we will limit our discussion to graphical on undirected graphs, and specifically to the Ising model. The Ising model consists of discrete variables that represent magnetic dipole moments of atomic spins at the location i that can be in one of two states, $\sigma_i = +1$ or $\sigma_i = -1$. The spins are arranged on a graph, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, often a lattice, allowing each spin to interact with its neighbors.

The Ising model was originally introduced as a mathematical model for ferromagnetism in statistical mechanics, and became popular in the late 1960s as the simplest model for identifying phase transitions. It is used as a model of statistical activity of neurons in the brain. Each neuron is either active (+) or inactive (-) at any given time. The active neurons are those that send an action potential down the axon in any given time window, and the inactive ones are those that do not. Because the neural activity at a given time is modelled by independent bits, Hopfield suggested that a dynamical Ising model would provide an approximation to a neural network that is capable of learning.

In modern neuroscience the Ising model is considered useful for any model of neural function, as based on the basic information-theoretic principle of maximum entropy. Given a collection of neurons represented by the state vector, $\boldsymbol{\sigma} = \{\sigma_i\}$, its probability distribution is given by

$$P(\boldsymbol{\Sigma} = \boldsymbol{\sigma}) = Z^{-1} \exp\left(-\frac{1}{2}\sum_{ij} J_{ij}\sigma_i\sigma_j - \sum_i h_i\sigma_i\right), \tag{2.60}$$

where $-h_i\sigma_i$ corresponds to the firing rate for a neuron, i, and $-(1/2)J_{ij}\sigma_i\sigma_j$, represents the pair-wise correlation in the firing of neurons i and j. J_{ij} is not necessarily restricted to neighbors. (Note that this beyond nearest neighbor interaction Ising model is sometimes called the quadratic exponential binary distribution in statistics.) This probability distribution introduces biases for the values taken by a spin, and by a pair of spins.

An activity pattern sampled from the distribution (2.60) requires the more bits than any other distribution with the same average activity and pairwise correlations when stored on a computer using the most efficient encoding scheme. This means that Ising models are relevant to any system which is described by bits which are as random as possible, with constraints on the pairwise correlations and the average number of 1s, which frequently occurs in both the physical and social sciences.

We will use the example of the Ising model to illustrate the main principles of statistical inference, learning and also (maximum likelihood) optimization. Inference may mean one of the following three related tasks:

- Computing the partition function (normalization factor):

$$Z = \sum_{\boldsymbol{\sigma}} \exp\left(-\frac{1}{2}\sum_{ij} J_{ij}\sigma_i\sigma_j - \sum_i h_i\sigma_i\right).$$

  In computer science, this is often referred to as "weighted counting".

- Computing the marginal probability for a particular spin at a particular site:

$$P_i(\Sigma_i = \sigma_i) = \sum_{\boldsymbol{\sigma}\setminus\sigma_i} P(\boldsymbol{\sigma}),$$

  or at a particular pair of sites:

$$P_{ij}(\Sigma_i = \sigma_i, \Sigma_j = \sigma_j) = \sum_{\boldsymbol{\sigma}\setminus\{\sigma_i,\sigma_j\}} P(\boldsymbol{\sigma}).$$

- Sampling an i.i.d. configuration, σ_i, from $P(\boldsymbol{\sigma})$.

Finding the most likely spin configuration $\arg\min_{\boldsymbol{\sigma}} P\left(\boldsymbol{\Sigma} = \boldsymbol{\sigma}\right)$ is generally easier to do computationally.

Investigating the computational complexity of inference and optimization generally involves understanding how the computational effort scales with the system size. In general, the Ising Model is known to be a $\#P$ (pronounced sharp P) hard—the complexity of inference is likely exponential in the system size—as there is no known algorithm to solve it in polynomial time. Indeed, consider the naive approach for a system of size N. The number of available states is 2^N, and therefore simply summing them up will require 2^N steps[3], which is computationally infeasible for large N.

Therefore, one needs to rely on an approximation. There exist three approximation methodologies in the literature for inference:

- Variational Methods, e.g. the so-called Mean Field and Belief Propagation, when the problem of weighted counting is replaced by an optimization formulation, or even better sequence of optimization formulations.

- Elimination Methods, like tensor networks, where we sum over variables one by one and then introduce approximations as complexity of intermediate object grows beyond a preset threshold.

[3] Some special cases may still be poly-tractable through some smart tricks – for example there exists an algorithm to compute partition function or sample in $N^{3/2}$ steps in the case of zero magnetic field, $\forall i: \quad h_i = 0$, and pair-wise interaction map forming a planar graph.

- Stochastic methods consisting in designing a Markov Chain with favorable mixing characteristics to achieve faster convergence of the Markov Chain to the pre-defined (e.g. by the Graphical Model) probability distribution.

We experiment with Markov Chain Monte Carlo (MCMC) algorithm, implementing the last of the three strategies, in the supplementary Jupyter notebook [Intro-MCMC-snippet.ipynb].

Learning a graphical model is similar to learning Bayesian neural networks. For an Ising model, learning might entail using a set of i.i.d. samples from a multivariate distribution to reconstruct the originating graph, magnetic fields, and pair-wise interactions. If all the nodes (spins) are observable, then graphical model learning can be done efficiently, that is with relatively few samples, $O(\log N)$ and through a convex optimization algorithm.

Consider training and inference with both **Neural Networks** and **Graphical Models**. For neural networks, learning is computationally intensive and the recent progress in **deep learning** comes from making the optimization behind training efficient and scalable. By construction, inference is straightforward as it is just explicit forward evaluation. On the contrary, graphical models can be learned efficiently (this observation is rather recent and not trivial), while inference requires computing marginal probabilities of a complicated multi-variate probability distribution expressed via graph. This bottleneck is addressed by using approximations. We conclude this brief preliminary tour of the course material by mentioning that even though on the practical side of Machine Learning, Graphical Models and Neural Networks are often viewed as distinctly different, the two approaches are ultimately related, and moreover the Graphical Model framework is broader, including in particular Neural Networks.

Chapter 3

Guide for Incoming Students by Colin Clark (former student)

Students join our program with diverse academic backgrounds: some with undergraduate degrees specifically in applied math, and many with degrees in pure math, physics, statistics, engineering or computer science. Sudents come with varying amounts of work experience: many come straight from their undergradute work, and others can have five or more years of industry work experience. Despite all these differences, most students need to review some material during the summer prior to joining the program. This chapter was written by a former student and it serves as an unofficial and very subjective to incoming students and how they might want to prepare for their first year of graduate school.

Topics for Review

- *What?* Undergraduate linear algebra, multi-variable calculus and ODEs.
  Who? Everyone. These three topics are the building blocks of applied math, and absolute mastery of these topics is crucial.
  How? To review this fundamental undergraduate topics, I would recommend working through homework exercises and exam problems. If you still have your old mid-terms and final exams, re-do the exercises (timed, closed book, etc) and then thoroughly review the topics that you found uncomfortable. If you have lost these resources, or if you think that they may have been lacking in any way, the MIT OCW site has plenty of excellent homework sets and exams (often with solutions), and excellent lecture videos.

- *What?* Undergraduate Analysis.
  Who? Students with backgrounds in physics, statistics, engineering or computer

science sometimes find the theory course particularly unfamiliar.
How? For this one, I would recommend learning from a book. Reading mathematics takes some practice, so working through the first few chapters of just about any book in undergraduate analysis would be very helpful. Read with a pencil in hand and work through any statement that is not immediately obvious. Be prepare for your progress to be very slow at first.

- *What?* A little bit of programming.
  Who? Anyone with no programming experience
  How? There are plenty of short courses on the web that teach introductory Matlab (or Octave), Python or Julia. (Some of your homework assignments and exam problems will require implementing and testing various algorithms in a language of your choice. Higher level languages will be sufficient. Although lower level languages like C or Fortran can be useful for some research problems, they tend to be over-kill for these ‘bite-size’ exercises.)

Not-so-much

It may be tempting to think that a successful start to graduate school requires knowledge of all the material from undergraduate applied math, and that any gap, real or perceived, amounts to a severe liability. I disagree. Applied math undergraduate courses like stochastic processes, complex variables, PDEs, dynamical systems, variational calculus, etc, are certainly valuable (and fun!), but learning or reviewing this material is *not* the top priority. Much of this material will be visited and revisited at deeper and deeper levels during graduate school. Deep understanding of these topics rely on a deep understanding of linear algebra (always), multi-variable calculus (often) and ODEs (sometimes). Go back and review linear algebra instead!

Chapter 4

Sample Problems

4.1 Problem Set #1

1. Diagonalize the matrix

$$A = \begin{bmatrix} 2 & 1 \\ 0 & 2 \end{bmatrix}$$

   i.e. Find some invertible T and diagonal D such that $A = TDT^{-1}$.

2. Use the Fredholm alternative to analyze whether the following linear system for the variables (x, y, z) is solvable for different values of the constant c. Specifically, for what values of c is the system (a) solvable, (b) not solvable, (c) uniquely solvable?

$$\begin{aligned} x + \ \ y + cz &= \ \ 4c \\ x + cy + \ \ z &= -2 \\ 2x + \ \ y + \ \ z &= -2 \end{aligned}$$

3. Using the change of variables

$$u = x + y, \quad v = \frac{y}{x+y}$$

   evaluate the integral

$$\iint_A \frac{1}{x+y}\, dx\, dy$$

   where A is the region bounded by the lines $x+y = 1$ and $x+y = 4$ and the coordinate axes.

4. Consider the paraboloid $z = 2 - (x^2 + y^2)$. Show that its surface area above the xy-plane is $13\pi/3$.

5. Find the particular solution of

$$y'' - 2y' + y = 0, \quad \text{with} \quad y_c(0) = 1, \quad y_c'(0) = 0.$$

6. Consider $y' = -y \ln(|y|)$. Solve it. Is the solution to an initial value problem unique?

7. Show that, if a function f is differentiable at a point x, then f is continuous at x. Give an example to show that the converse is not true.

8. Show that $e^x \geq 1 + x$ for all $x \in \mathbb{R}$.

9. Is the function $f(x) = 1/(x^2 - 2)$ defined for all rational numbers x? Is this a continuous function of the set of all rationals? Justify your answer.

10. Let $z = (2 - i)/(3 + 4i)$.

    (a) Express z in the form $a + bi$.

    (b) Find z^{10}.

4.2 Problem Set #2

1. (Modified from Linear Algebra and its Applications, Gilbert Strang). Suppose A is the sum of two matrices of rank one: $A = u_1 v_1^\top + u_2 v_2^\top$

   (a) Which vectors span the column space of A?

   (b) Which vectors span the row space of A?

   (c) The rank of A is less than 2 if ________________ or if ________________.

2. Find the eigenvalues and eigenvectors of the matrix

$$A = \frac{1}{4}\begin{bmatrix} 2 & 1 & 0 & 1 \\ 1 & 2 & 1 & 0 \\ 0 & 1 & 2 & 1 \\ 1 & 0 & 1 & 2 \end{bmatrix}$$

   Hint: You may use the fact that A is a symmetric, circulant matrix.

3. For the linear system of equations in the unknown x, y, z

$$\begin{aligned} 6x - 2y + 3kz &= 3 \\ 6x + y &= 2k \\ 2kx + 3z &= 3 \end{aligned}$$

   Determine the values of k for which the system has:

   (a) no solution

   (b) a unique solution

   (c) many solutions

   Note: We are not asking you to explicitly compute the solutions.

4. Convince yourself of the following equivalence:

$$\left(\int_{-\infty}^{\infty} e^{-t^2}\,dt\right)^2 = \left(\int_{-\infty}^{\infty} e^{-x^2}\,dx\right)\left(\int_{-\infty}^{\infty} e^{-y^2}\,dy\right).$$

Re-write the right-hand side as an iterated integral and use a suitable change in variables to show that

$$\int_{-\infty}^{\infty} e^{-t^2}\,dt = \sqrt{\pi}$$

5. (a) Evaluate by direct integration, the integral

$$\iint_S (\nabla \times \boldsymbol{V}) \cdot \mathbf{n}\,dS$$

where

$$\boldsymbol{V} = (2x - y)\mathbf{i} - yz^2\mathbf{j} - y^2z\mathbf{k},$$

S is the upper half surface of the sphere $x^2 + y^2 + z^2 = 1$, and $\mathbf{n}$ is the unit normal to S.

(b) Verify your result by using Stokes' theorem to evaluate the integral.

6. Find the general solution to the system of first-order linear ODE's:

$$\begin{pmatrix} x' \\ \\ y' \end{pmatrix} = \begin{pmatrix} 4 & 1 \\ \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x \\ \\ y \end{pmatrix}$$

7. (a) Find the general solution to the homogeneous ODE

$$x^2y'' + xy' - y = 0$$

(b) Use the variation of parameters to find the general solution to the inhomogeneous ODE

$$x^2y'' + xy' - y = x, \quad \text{for}\, x \neq 0.$$

8. Let $x \in \mathbb{R}$. Find $\lim_{n\to\infty}(1 + x/n)^n$. What about if x is replaced by $z \in \mathbb{C}$.

9. Use a computer to estimate

$$1 - 1/2 + 1/3 - 1/4 + 1/5 - 1/6 + \cdots$$

in two ways. One way is as written. The other way is

$$1 + 1/3 - 1/2 + 1/5 + 1/7 - 1/4 + 1/9 + 1/11 - 1/6 + \cdots,$$

which is a rearrangement. What do you observe? Is this what one should expect? Explain.

10. Use contour integration to evaluate the integral of $1/z^2$. Assume the contour encloses the origin and is oriented in a counter-clockwise direction.

4.3 Problem Set #3

1. Let

$$\mathbf{A} = \begin{pmatrix} 0 & \pi/2 \\ \pi/2 & 0 \end{pmatrix}$$

Express $\sin(\mathbf{A})$ in its simplest possible form. Justify each step.

Hint: $\sin : \mathbb{R}^n \to \mathbb{R}^n$ is defined via its Taylor series.

2. Consider the matrix

$$B = \begin{bmatrix} 2 & 4 & 6 \\ 1 & 2 & 1 \\ 1 & 2 & 5 \end{bmatrix}$$

(a) Find a vector $\boldsymbol{x}$ that is orthogonal to the row space of B.

(b) Find a vector $\boldsymbol{y}$ that is orthogonal to the column space of B.

3. Let $t_1, t_2, t_3, \ldots, t_n$ be real numbers. Define the $n \times n$ matrix by

$$T = \begin{pmatrix} t_1 & t_2 & t_3 & t_4 & \ldots & t_{n-1} & t_n \\ t_n & t_1 & t_2 & t_3 & \ldots & t_{n-2} & t_{n-1} \\ t_{n-1} & t_n & t_1 & t_2 & \ldots & t_{n-3} & t_{n-2} \\ t_{n-2} & t_{n-1} & t_n & t_1 & \ldots & t_{n-4} & t_{n-3} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ t_3 & t_4 & t_5 & t_6 & \ldots & t_1 & t_2 \\ t_2 & t_3 & t_4 & t_5 & \ldots & t_n & t_1 \end{pmatrix}$$

(a) Show that the vectors

$$v_j = (1, e^{i\alpha_j}, e^{i2\alpha_j}, e^{i3\alpha_j}, \ldots, e^{i(n-1)\alpha_j})$$

with $\alpha_j = \frac{2\pi}{n} j$ and $j = 0, 1, 2, \ldots, n-1$ are eigenvectors of T and find the corresonding eigenvalues.

(b) Suppose that n is even. Let $x = (1, 0, 1, 0, 1, 0, \ldots, 1, 0)$. Compute $T^{10}x$.

4. Use Lagrange multipliers to show that the maximum value of the function $f(x, y) := x^3y$, constrained to the ellipse $3x^2 + y^2 = 6$ is $9/4$. *i.e.* Show that

$$\max_{(x,y)} x^3 y \Big|_{3x^2+y^2=6} = \frac{9}{4}$$

5. Consider the helix, expressed in cylindrical coordinates, $(r, \theta, z) = (R_0, 2\pi t, \alpha t)$ where $0 \leq t \leq 1$ and R_0 and α are fixed. Let $\mathcal{S}$ represent the arclength of this helix. Compute $\mathcal{S}$ and $\frac{\partial \mathcal{S}}{\partial \alpha}|_{\alpha=0}$.

6. Let $\mathcal{S}$ be the disc $x^2 + y^2 \leq 4$ at $z = -3$, and let C denote the path around that disc. For the vector field

$$\mathbf{v} = y\mathbf{i} + xz^3\mathbf{j} - zy^3\mathbf{k}$$

where $(\mathbf{i}, \mathbf{j}, \mathbf{k})$ are the standard unit vectors in $\mathbb{R}^3$, evaluate

$$I = \oint_C \mathbf{v} \cdot d\mathbf{x}$$

where the path of integration is taken in the anti-clockwise sense.

7. Solve

$$\frac{dx}{dt} = \sin(t)\cos^2(x), \quad x(0) = 0$$

8. Solve

$$\frac{dx}{dt} - \lambda(t)x(t) = \frac{f(t)}{x(t)}$$

where $\lambda(t)$ and $f(t)$ are known functions.

9. Prove there is no largest prime number.

10. Let $f : [0, 1] \to \mathbb{R}$ be a continuous function.

(a) Suppose that

$$\inf_{x \in [0,1]} f(x) \leq 0 \leq \sup_{x \in [0,1]} f(x).$$

Show that there is a point c in $[0, 1]$ such that $f(c) = 0$.

(b) Is this conclusion true if f is defined on the open interval $(0, 1)$?

Bibliography

[1] U. Frisch, *Turbulence: The Legacy of A. N. Kolmogorov.* Cambridge University Press, 1995.

[2] C. F. Higham and D. J. Higham, “Deep Learning: An Introduction for Applied Mathematicians,” *arxiv:1801.05894*, 2018. [Online]. Available: https://arxiv.org/pdf/1801.05894.pdf