CHAPTER 1

Introduction To Mathematical Modeling

In his book "The Possible and the Actual", published by the University of Washington Press as a part of the Jessie and John Danz Lectures, Françis Jacob (1920-2013), a leading molecular biologist of the 20^{th} century, stated that a radically change in the Western art had occurred around the Renaissance time, from "symbolizing" to "representing" the real world. One can in fact view pure and applied mathematics as a change from the former to the latter. The task of ***applied mathematics*** is to quantitatively represent the real world in terms of mathematics. This task comes with two parts: ***studying representations*** and ***developing models***.

The idea that the real world phenomena can be represented in terms of mathematics originated in ancient Greek philosophy, and was convincingly demonstrated in the theory of mechanical motion, which you have already studied a little in high school: $\mathrm{d}^2x/\mathrm{d}t^2 = F/m$! There are other representations for other phenomena in the Nature; e.g., electromagnetism, electronic motion inside atoms, motions with the speed near that of light, etc. They are all part of theoretical physics, in which one usually studies highly idealized scenarios. Nevertheless, one can learn a great deal about the Nature by doing just that. In this class, we only deal with two presentations: the mechanical representation for motion and the chemical representation for counting.

A mathematical representation for the Reality always contains unknown terms that are specific to a real world, engineering problem. Specific models are required to obtain realistic results. There are fundamentally two types of mathematical modeling: (a) Numerical representation of scientific data in terms of mathematical formula or equations, and (b) mechanistic representation of a system's behavior (natural or engineered, physical or biological, electronic, chemical, economical, social, ...) based on existing, established mathematical formula and equations. For lack of better terminology, we shall call the former ***data-driven modeling*** and the latter ***mechanistic modeling***. Note, according to Karl Popper (1902–1994) and his philosophy of science, the only legitimate scientific activity is falsifying a hypothesis: That requires first to formulate a hypothesis, which sometime is just looking for patterns in the data (e.g., numerical hypothesis) and sometime is proposing a mechanism (e.g., the modeling we shall study in this class); and (b) to derive rigorous predictions from a hypothesis, which is a form of logical, or mathematical, deduction.

The first two chapters of the textbook present several well-known data-driven models. Some were from ancient times. This kind of endeavor has never stopped. Two of the most active current areas are perhaps *bioinformatics* and *financial engineering*.

Because the subject in biomedicine and in economics are very complex and full of uncertainties, the state-of-the-art data-driven modeling in these areas has to take chance, e.g., random probabilities, into account. That is why most of the modeling in these areas are statistical based.

1.1 Mechanistic model and applied mathematics

One of the most celebrated data-driven models in an earlier time was Kepler's Laws (Ch. 5). They are mathematical representations for the various aspects of the motion of planets around the Sun.

But Isaac Newton did something very novel, probably inspired by his believe in God: He described the mechanical motions represented by Kepler's laws in terms of a "cause", now widely called a "mechanism" or a "force". And furthermore, he generalized the same mechanism to many other systems beyond planetary motions, down to the earth, even for a falling apple! By doing so, he invented the philosophy of modern science and engineering. In a nutshell, Newton asked "why Kepler's laws?" — This is precisely the last word of the Ch. 1. The two key words here are "why" and "generalize". In the context of 383, knowing something about the "why" helps you to develop a *mechanistic* mathematical model; and mathematical computations can help you to "generalize" the predictions.

Therefore, when one finds a novel mathematical description for a set of interesting data, hopefully that will be one of you someday, the next question is "can it be explained by existing knowledge"? If it can, then you want to develop that. Here, the issue of *computability* and concept of *emergent behavior* come into play in modern mathematical modeling of complex phenomena. You should definitely check out these two buzzwords online.

The mechanistic approach to systems, processes, and scientific data leads to several distinct features:

(i) It is based on certain mechanistic hypotheses, which are themselves data-driven models from different systems in an earlier time;[*]

(ii) It believes "universality" more than diversity, even though it certainly does not oppose the latter;[†]

(iii) It is largely mathematical analysis in nature. In Professor Tung's preface, he called this type of modeling "deductive".

With predictions from a mechanistic mathematical model in hand, the true value of

[*] In classical physics, these are called *natural laws*. It is important to point out the notion of "intrinsic properties" which has led to reductionism, and the opposite notion of "systems perspective". There is an intimate relation between one's believe in the former and one's tendency to generalize. If one believes everything is "context dependent", then no data-driven model can and should be generalized.

[†] In current applied mathematics, diversity is often modelled in terms of a "distribution". This naturally leads to models in terms of partial differential euations and stochastic dynamics.

data lies in its disagreement with a prediction, i.e., a contradict to the hypothesis.[‡] Then, one needs to develop an improved model. This cycle turns out to be the most valuable:

mechanistic formulation → analysis/solution → interpretation/generalization/prediction → compare with data → improved mechanistic model.

We shall show this "paradigm" in several examples.

1.1.1 The certainty and confidence in exact sciences

You don't hear the term "exact science" very often anymore. It means a field of science, such as physics or chemistry, which has a mathematical foundation, with its theories expressed in terms of mathematics, and its measurements being quantitative and can be checked against predictions from the mathematical theory.

These days, we have a tremendous amount of confidence in *Science*. One of the easiest ways to shot-down an arguement is to show "it is non-scientific". But where is such a confidence from? If you exam carefully, you will see that the confidence really is about the absoluteness of a mathematical conclusion, or conclusions, based on an idealized mathematical model. Actually, a cautious person, correctly, should not have too much confidence in the idealized mathematical model representing a reality. The latter is always only approximately true.

1.2 Mechanical and chemical representations of the world

The successes of Newtonian theory of the Nature is so great that we now learn about $F = ma$, e.g., *mechanics*, even in high school. More significantly, the system of differential equations

$$m_i \frac{\mathrm{d}^2 x_i}{\mathrm{d}t^2} = F_i(x_1, \cdots, x_n), \tag{1.1}$$

is now considered to a *mathematical representation* of the world. It is no longer just a model; it has a certain *a priori* validity! One simply uses this mathematics to carry out engineering designs and beleives the outcomes from computer calculations!

But this mechanical approach to understanding the world runs into difficulties when the number of *point masses* are too large: One of the most telling example is the atomic motions in proteins, the biologically important molecules.

What is the difference between a point mass in mechanics and a molecule in chemistry? The former is a very abstract representation of a real object: It has no interior, it is completely specified with six numbers, three coordinates and three velocities, plus its mass. This can be done because of the notion *center of mass*. But in order

[‡] Here, one can see why it is generally claimed that a data-driven model does not lead to a "real understanding" while a mechanistic model does: The latter is built upon previous understandings. No wonder data-driven models are also called by some as "discoveries".

for the center-of-mass concept to work, there can not have frictional forces, which usually are dependent upon velocities. Indeed, the mechanics working well also because a series of assumptions: (i) $\vec{F}$ is independent of velocity; (ii) $\vec{F}$ has a potential function; (iii) the potential function is pairwise additive.

A molecule, however, has an interior: It is made of atoms which in turn are made of electons and nuclei, which in turn are made of elementary particules. So while physicists finding more and more fundamental constituents of subatomic world, it leaves behind an increasingly more and more complex molecules. It is fair to say that a molecule has an nearly infinite number of internal degrees of freedom. It is a complicated object. In fact, each and every molecule, at least for protein molecules in biological cells, is a different individual.

The scientific description of molecules, therefore, has to be based on *classifications* and *statistics*.

If the subjects with which chemistry deals are so complex, how does chemistry become a part of the *exact science*? The credit goes to the "Avogadro's number", 6.022×10^{23}! This is precisely the same reason why casinos are confident about their profitability. But such a confidence one can no longer be for cell biology: The total number of human cells in the Universe is 7.5 billion people (10^9) times 37 trillion cells per person (10^{12}), which is equal to only $= 2.78 \times 10^{23}$, less than one half of a mole!

It is all about counting the number of individuals in a particular "species", molecule species, ecological species, etc.

1.3 Statistical laws, or models

Kepler's laws and Fibonacci numbers give precise "predictions" for the values of next measurements. What happens if one can not give such precise, deterministic prediction? Does that mean things are hopeless? The answer is "no". *If one can not predict with certainty, maybe one can predict with certainty about the "uncertainty".* That is the probabilistic thinking. In modern economics, this is called "estimating the risk". Risk management is all about computing probabilities.

Let us consider the number of rain drops on a tin roof. Certainly, if you count the number of rain drops per unit time, it will not always be exactly the same. However, if you repeat the measurement again and again, a pattern emerges: Let the n be the number of rain drops. It is random with a distribution

$$\Pr\{n = k\} \approx \frac{\lambda^k}{k!} e^{-\lambda}, \tag{1.2}$$

where λ is a parameter. This is known as Poisson distribution.

But why? This leads to the concepts of statistical modeling and stochastic (mechanistic based, agent-based) modeling. Statistical modeling starts with data; and stochastic modeling starts with a mechanism. We now try to give a “mechanistic model” for the Poisson distribution about rain drops.

Let us assume that there are N number of total rain drops coming from the sky. Each one behaves in a similar, but independent fashion as another one. Let the probability of a single rain drop hitting the roof be p, and not hitting be $1-p$. Then the probability of having k rain drops hitting the roof is

$$p_k = \binom{N}{k} p^k (1-p)^{N-k} = \frac{N!}{k!(N-k)!} p^k (1-p)^{N-k}.$$

This certainly does not look like that in Eq. 1.2. We also note, however, that the N must be a very large number while the p a very small one. But Np is the average number of drops on the roof! So let us denote $Np = \lambda$, and now consider

$$p_k = \frac{N!}{k!(N-k)!} \left(\frac{\lambda}{N}\right)^k \left(\frac{N-\lambda}{N}\right)^{N-k} = \left\{ \frac{N!}{(N-k)!} \left(\frac{1}{N-\lambda}\right)^k \right\} \frac{\lambda^k}{k!} \left(1 - \frac{\lambda}{N}\right)^N .$$

We are interested in its limit when $N \to \infty$.

You can use the *Stirling formula*:[§] For a large number n:

$$\ln n! \approx n \ln n - n.$$

Therefore,

$$\begin{aligned}
& \ln \left\{ \frac{N!}{(N-k)!} \left(\frac{1}{N-\lambda}\right)^k \right\} \\
\approx\ & N \ln N - N - (N-k)\ln(N-k) + (N-k) - k\ln(N-\lambda) \\
=\ & N \ln \frac{N}{N-k} + k \ln \frac{N-k}{N-\lambda} - k \\
=\ & N \ln \left(1 + \frac{k}{N-k}\right) + k \ln \left(1 + \frac{\lambda - k}{N-\lambda}\right) - k \\
=\ & N \left(\frac{k}{N-k} - \frac{1}{2}\left(\frac{k}{N-k}\right)^2 + \cdots \right) + k \left(\frac{\lambda-k}{N-\lambda} - \frac{1}{2}\left(\frac{\lambda-k}{N-\lambda}\right)^2 + \cdots \right) - k \\
\to\ & 0.
\end{aligned}$$

And

$$\lim_{N\to\infty} \left(1 - \frac{\lambda}{N}\right)^N = e^{-\lambda}.$$

§

$$\ln n! = \sum_{k=1}^{n} \ln k \approx \int_1^n \ln x \, dx = x \ln x - x\big|_1^n = n \ln n - n + 1 \approx n \ln n - n.$$

Therefore, in the limit of $N \to \infty$,

$$\lim_{N\to\infty} \frac{N!}{k!(N-k)!}\left(\frac{\lambda}{N}\right)^k\left(\frac{N-\lambda}{N}\right)^{N-k} = \frac{\lambda^k}{k!}e^{-\lambda}.$$

From both examples, the Fibonacci sequence and the Poisson distribution, we see that mathematical modeling increases one's understanding of a natural phenomenon as well as provides some predictions. The ultimate goal of science is to understand, to cumulative knowledge; accurate predictions are valuable to engineering.

There is a wide range of issues in how to develop a data-driven mathematical model for a give set of data. This is much more so in statistical modeling than in deterministic modeling. Still there are several essential issues one should keep in mind:

(1) The definition of "goodness of fit";

(2) Over fitting and maximum parsimony;

(3) Choices of different "numerical models", with parameters.

One suggestion: Answers to the above questions are often not unique in mathematics. But if you ask yourself "what is the model for?", you would be able to come up with simple answers to these difficult questions. Each one of them becomes an entire area of research in statistics.