Chapter 3
Basics on probability

3.1 Introduction

Let us motivate you by considering the following particular inverse problem, namely, the deconvolution problem. Given the observation signal $g(s)$, we would like to reconstruct the input signal $f(t) : [0,1] \to \mathbb{R}$, where the observation and the input obey the following relation

$$g(s_j) = \int_0^1 a(s_j,t) f(t)\, dt, \quad 0 \leq j \leq n. \tag{3.1}$$

Here, $a : [0,1] \times [0,1] \to \mathbb{R}$ is known as the blurring kernel. So, in fact we don't know the output signal completely, but at a finite number of observation points. A straightforward approach you may think of is to apply some numerical quadrature on the right side of (3.1), and then recover $f(t)$ at the quadrature points by inverting the resulting matrix. If you do this, you realize that the matrix is ill-conditioned, and it is not a good idea to invert it. There are techniques to go around this issue, but let us not pursue them here. Instead, we recast the deconvolution task into an optimization problem such as

$$\min_{f(t)} \sum_{j=0}^{n} \left(g(s_j) - \int_0^1 a(s_j,t) f(t)\, dt \right)^2, \tag{3.2}$$

that is, we minimize the *misfit* between the mathematical model $\int_0^1 a(s_j,t) f(t)\, dt$ and the actual observations. However, the ill-conditioning nature of our inverse problem does not go away. Indeed, (3.2) may have multiple solutions and multiple minima. In addition, a solution to (3.2) may not depend continuously on $g(s_j), 0 \leq j \leq n$. So what is the point of recast? Clearly, if the cost function (also known as the data misfit) is a parabola, then the optimal solution is unique. This immediately suggests that one should add a quadratic term to the cost function to make it more like a parabola, and hence making the optimization problem easier.

This is essentially the idea behind the *Tikhonov regularization*, which proposes to solve the nearby problem

$$\min_{f(t)} \sum_{j=0}^{n} \left(g(s_j) - \int_0^1 a(s_j,t) f(t)\, dt \right)^2 + \frac{\kappa}{2} \left\| \mathscr{R}^{1/2} f \right\|^2,$$

where κ is known as the regularization parameter, and $\|\cdot\|$ is some appropriate norm. Perhaps, two popular choices for $R^{1/2}$ are ∇ and Δ, the gradient and Laplace operator, respectively, and we discuss them in details in the following.

Now, in practice, we are typically not able to observe $g(s_j)$ directly but its noise-corrupted values

$$g^{obs}(s_j) = g(s_j) + e_j, \quad 0 \le j \le n,$$

where e_j, $j = 0, \ldots, n$, are some random noise. You can think of the noise as the inaccuracy in observation/measurement devices. The question you may ask is how to incorporate this kind of randomness in the above deterministic solution methods. There are works in this direction, but let us introduce a statistical framework based on the Bayesian paradigm to you in this note. This approach is appealing since it can incorporate most, if not all, kinds of randomness in a systematic manner.

Some portion of this chapter follows the presentation of the two excellent books by Somersalo *et. al.* [10, 17]. The pace is necessary slow since we develop this note for readers with minimal knowledge in probability theory. The only requirement is to either be familiar with or adopt the conditional probability formula concept. This is the corner stone on which we build the rest of the theory. Clearly, the theory we present here is by no means complete since the subject is vast, and still under development.

3.2 Some concepts from probability theory

We begin with the definition of randomness.

Definition 3.1. An even is *d*eterministic if its outcome is completely predictable.

Definition 3.2. A *random event* is the complement of a deterministic event, that is, its outcome is not fully predictable.

Example 3.1. If today is Wednesday, then "tomorrow is Thursday" is deterministic, but whether it rains tomorrow is not fully predictable. △

As a result, randomness means lack of information and it is the direct consequence of our ignorance. To express our belief[1] on random events, we use probability; probability of uncertain events is always less than 1, an event that surely happens

[1] Different person has different belief which leads to different solution of the Bayesian inference problem. Specifically, one's belief is based on his known information (expressed in terms of σ-algebra) and "weights" on each information (expressed in terms of probability measure). That is, people working with different probability spaces have different solutions.

has probability 1, and an event that never happens has 0 probability. In particular, to reflect the subjective nature, we call it *subjective probability* or *Bayesian probability* since it represents belief, and hence depending upon one's experience/knowledge to decide what is reasonable to believe.

Example 3.2. Let us consider the event of tossing a coin. Clearly, this is a random event since we don't know whether head or tail will appear. Nevertheless, we believe that out of n tossing times, $n/2$ times is head and $n/2$ times is tail.[2] We express this belief in terms of probability as: the (subjective) probability of getting a head is $\frac{1}{2}$ and the (subjective) probability of getting a tail is $\frac{1}{2}$. △

We define $(\Omega, \mathscr{F}, \mathbb{P})$ as a *probability space*. One typically call Ω the *sample space*, $\mathscr{F}$ a σ-algebra containing all events $A \subset \Omega$, and $\mathbb{P}$ a probability measure defined on $\mathscr{F}$. We can think of an event A as information and the probability that A happens, i.e. $\mathbb{P}[A]$, is the weight assigned to that information. We require that

$$0 \leq \mathbb{P}[A] \leq 1, \quad \mathbb{P}[\emptyset] = 0, \quad \mathbb{P}[\Omega] = 1.$$

Example 3.3. Back to the tossing coin example, we trivially have $\Omega = \{head, tail\}$, $\mathscr{F} = \{\emptyset, \{head\}, \{tail\}, \Omega\}$. The weights are $\mathbb{P}[\emptyset] = 0, \mathbb{P}[\{tail\}] = \mathbb{P}[\{head\}] = \frac{1}{2}$, and $\mathbb{P}[\{head, tail\}] = 1$. △

Two events A and B are independent[3] if

$$\mathbb{P}[A \cap B] = \mathbb{P}[A] \times \mathbb{P}[B].$$

One of the central ideas in Bayesian probability is the *conditional probability*[4]. The conditional probability of A on/given B is defined as[5]

$$\boxed{\mathbb{P}[A|B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}}, \tag{3.3}$$

This is the corner stone formula to build most of results in this note, make sure that you feel comfortable with it.

which can also be rephrased as the probability that A happens *provided* B has already happened. An intuitive way to understand this formula is to consider the probability measure as the standard area (or volume) measure. In this case, the conditional

[2] One can believe that out of n tossing times, $n/3$ times is head and $2n/3$ times is tail if he uses an *unfair* coin.

[3] Probability theory is often believed to be a part of measure theory, but independence is where it departs from the measure theory umbrella.

[4] A more general and rigorous tool is conditional expectation, a particular of which is conditional probability.

[5] This was initially introduced by Kolmogorov, a father of modern probability theory. An elegant derivation of this formula is based on the conditional expectation, but this would take us too deep into the probability theory [16].

measure is nothing more than the ratio of the area of the intersection $A \cap B$ and the area of B.

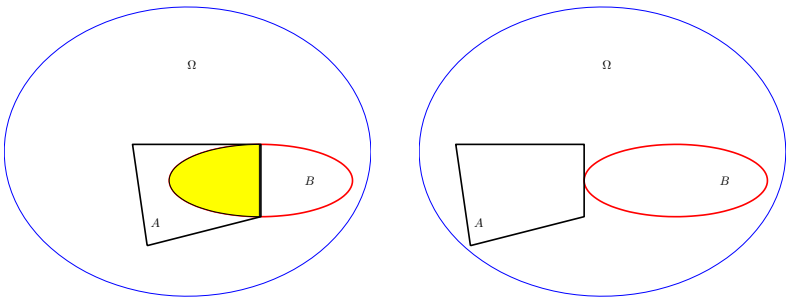
Example 3.4. Assume that we want to roll a dice. Denote B as the event of getting of face bigger than 4, and A the event of getting face 6. Using (3.3) we have

$$\mathbb{P}[A|B] = \frac{1/6}{1/3} = 1/2.$$

We can solve the problem using a more elementary argument. B happens when we either get face 5 or face 6. The probability of getting face 6 when B has already happened is clearly $\frac{1}{2}$. △

The conditional probability can also be understood as the probability when the sample space is restricted to B.

Exercise 3.1. Determine $\mathbb{P}[A|B]$ in Figure 3.1. •



(a) $\mathbb{P}[A|B] =$ (b) $\mathbb{P}[A|B] =$

Fig. 3.1 Demonstration of conditional probability.

Exercise 3.2. Show that the following Bayes formula for conditional probability holds

$$\boxed{\mathbb{P}[A|B] = \frac{\mathbb{P}[B|A]\,\mathbb{P}[A]}{\mathbb{P}[B]}.} \tag{3.4}$$

•

By inspection, if A and B are mutually independent, we have

$$\mathbb{P}[A|B] = \mathbb{P}[A], \quad \mathbb{P}[B|A] = \mathbb{P}[B].$$

The probability space $(\Omega, \mathscr{F}, \mathbb{P})$ is an abstract object which is useful for theoretical developments, but far from practical considerations. In practice, it is usually circumvented by probability densities over the *state space*, which are easier to handle and have certain physical meanings. We shall come back to this point in a moment.

Definition 3.3. The state space S is the set containing all the possible outcomes.

3.3 Appendix

Definition 3.4 (σ-algebra). A family $\mathscr{F}$ of subsets of a non-empty set Ω is called a σ-algebra if

- $\emptyset \in \mathscr{F}$,
- $A \in \mathscr{F}$ implies $A^c \in \mathscr{F}$ (here A^c is the complement of A), and
- $A_n \in \mathscr{F}$ for all $n \in \mathbb{N}$ implies $\cup_n A_n \in \mathscr{F}$.

Clearly the power set of Ω is a σ-algebra. Let $\mathscr{A}$ be a family of subsets and the intersection of all σ-algebras containing $\mathscr{A}$, i.e. the smallest σ-algebra containing $\mathscr{A}$, is called the σ-algebra generated by $\mathscr{A}$.The σ-algebra generated by all open sets in Ω is called the Borel algebra.