The Mathematics of Machine Learning

Kevin Atteson

December 13, 2020

Chapter 1

Why is Mathematics Important in Machine Learning?

1.1 What is Machine Learning?

In order to discuss the mathematics of machine learning, we should first investigate what the term "machine learning" means. For the purpose of this course, we borrow from Wikipedia:

Definition 1.1.1. Machine learning *(ML) is the study of computer algorithms that improve automatically through experience.*

A commonly used definition by Tom Mitchell[Mit97] is more precise:

Definition 1.1.2. *A computer program is said to* **learn** *from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.*

There are two closely related disciplines:

Definition 1.1.3. Statistics *is the discipline that concerns the collection, organization, analysis, interpretation and presentation of data.*

Definition 1.1.4. Econometrics *is the application of statistical methods to economic data in order to give empirical content to economic relationships.*

All of these disciplines have much in common: all of them involve "data" or "experience" and all of them involve "algorithms" or "methods" that apply to this data. Roughly speaking, if you hail from a statistics department, you call it statistics, if you hail from an economics department, you call it econometrics and if you hail from a computer science department, you call it machine learning. Of

course, to get funding for your startup company in the early 2020's, you should call it machine learning.

In this course, we will talk about theoretical limitations of all machine learning approaches are subject to, that is, what it is impossible for any machine learning method to achieve. In addition, we will talk about the capabilities that specific machine learning approaches can be proven to have, that is, what certain specific machine learning approaches can achieve. We will motivate these results largely with examples from portfolio selection.

1.2 The Curse of Dimensionality

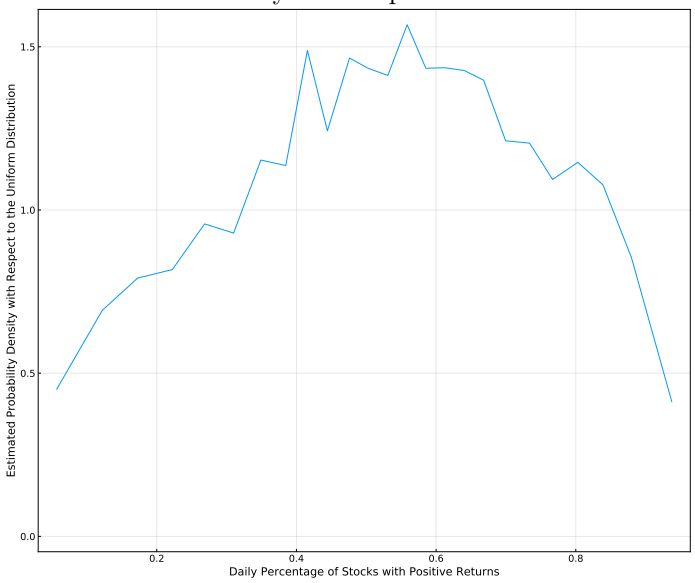
In order to motivate the discussion, we give a simple example that demonstrates a major limitation of all machine learning approaches, namely, the curse of dimensionality. Suppose we wish to decide whether to go long or short the largest 500 US stocks on a given day, which we denote by y_{t+1}. We are interested in using the prior day's returns to help inform our decision. In order to simplify the data, we use only information on whether each of the 500 stocks went up or down on the prior day. Since this is reducing the amount of information, it only makes the problem easier and we'll see later in the course that the same curse of dimensionality holds very generally. Let $x_{t,i}$ denote whether the *ith* stock went up or down today. We will assume that y_{t+1} a deterministic function of $x_{t,1}, x_{t,2}, \ldots, x_{t,500}$, that is $y_{t+1} = f(x_t)$ where x_t is the vector $(x_{t,1}, x_{t,2}, \ldots, x_{t,500})$ and f is some unknown function. Not only are we restricting our model to "up" or "down" returns, but we're eliminating all noise by making y_{t+1} a deterministic function. It turns out that that even this simplified model can easily be shown to be difficult.

How many days do we need to observe this process before we can know the function f for certain? There are $2^{500} \approx 10^{150}$ possible input strings and so, given that we need to know the output for each input string, we would need to observe the process for many many times the current scientific estimates of the age of the universe of about 14 billion years. What if we only need to be correct 90% of the time? We would still need to see 90% of the possible strings which would still be exorbitantly long. This issue is referred to as the curse of dimensionality. If we know nothing about the underlying problem and we have many predictor variables, there is little hope of learning a function of those variables.

We'll now change the problem a bit in order to demonstrate how we can avoid the curse of dimensionality so that useful knowledge can be learned. We'll now assume that whether each stock goes up or down is random and attempt learn the joint distribution, which, we'll see, will be useful for hedging. Let $X_{t,i}$ denote the random variable which indicates that the *ith* stock went up at time t (we adhere the convention, common in probability theory, of using capital letters for random variables, as discussed in more detail later). We'll introduce the following additional assumptions:

1. We assume that the sequence $X_t = (X_{t,1}, \ldots, X_{t,500})$ are independent

Figure 1.1: Estimated Density with respect to the Uniform Distribution



over time t. For the moment, we assume the reader knows the term "independent" which we will more carefully discuss later in the text.

2. We assume that the distribution of X_t for a fixed t is independent of the labeling of the the individual stocks, that is, for any permutation σ on the set $\{1, 2, \ldots, 500\}$, we have that:

$$\begin{aligned}&\mathbb{P}(X_{t,1} = x_{t,1}, X_{t,2} = x_{t,2}, \ldots, X_{t,500} = x_{t,500}) \\ &\quad = \quad \mathbb{P}\big(X_{t,\sigma(1)} = x_{t,\sigma(1)}, X_{t,\sigma(2)} = x_{t,\sigma(2)}, \ldots, X_{t,\sigma(500)} = x_{t,\sigma(500)}\big)\end{aligned}$$

where $\mathbb{P}$ is the probability distribution and $x_{t,1}, x_{t,2}, \ldots, x_{t,500}$ are constants. For example, this means that the chance of the *3rd* stock going up and all others going down is the same as the chance of the *17th* stock going up and all others going down. Similarly, it means that the chance of the *4th*, *7th* and *399th* stocks being the only ones to go up is the same as the chance of the *478th*, *485th* and *497th* stocks being the only ones to go up.

These assumptions imply that distribution only depends upon the number of stocks which go up or down, or, dividing by the total number of stocks, the percentage of stocks that go up or down. Figure 1.1 shows a histogram estimate of the distribution as a function of the percentage of stocks going up for the data from 2000 through 2019. There is nothing that particularly stands out in this chart other than the more central numbers being more frequent.

On further reflection, it can be seen that the distribution is highly skewed. Suppose only 1 stock goes up. This could be any one of the 500 stocks so there

are 500 ways in which this could happen. If no stocks go up, however, there is only a single way in which this could happen. For 2 stocks going up, there 500 possibilities for the first stock and 499 for the second. However, each of these possibilities can happen in two ways: the lower numbered stock could be chosen first or second. Hence, there are $\frac{500\times 499}{2} = 124\,750$ ways for 2 of the 500 stocks to go up. In general, the number of ways for m stocks to go up is given by the binomial coefficients:

$$\left(\begin{array}{c} 500 \\ m \end{array} \right) = \frac{500!}{m!(500-m)!}$$

Hence, if the patterns were uniformly distributed, the middle patterns would be much more likely. Figure 1.2 shows the ratio of the estimated density to the density derived from a uniform distribution of patterns. The y-axis of the chart is on a logarithmic scale since the values become large. For patterns with only a small number of stocks going up or only a small number going down, the number of occurences are 10^{80} times as likely as they would be from a uniform distribution. This is due to a well-known phenemon: stocks move in a highly correlated fashion. Indeed, risk can be reduced significantly by going long some names and short others. It does not seem that risk can completely eliminated by this method, that is, this method does not produce arbitrage. However, this reduction of risk is the basis of what is referred to as statistical arbitrage, one of the most common quantitative trading strategies. By making assumptions about the underlying process, we were able to derive some further knowledge, albeit in this case, knowledge that was already known.

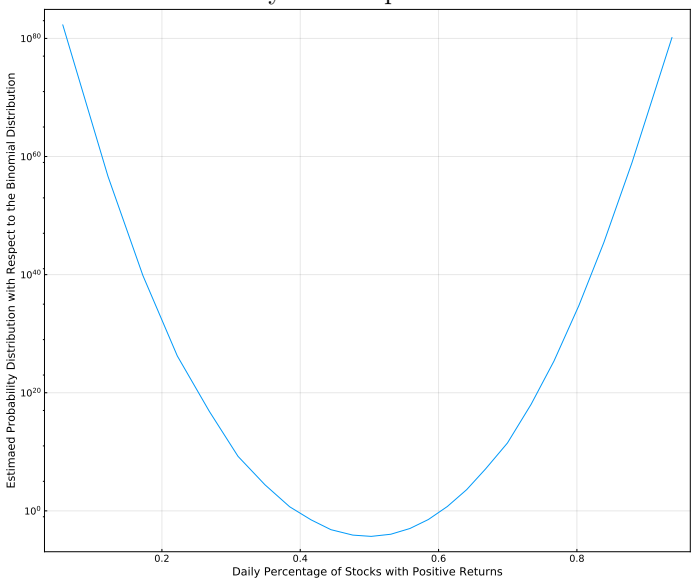
Let's now summarize the situation:

1. If one has no knowledge or makes no assumptions about the process, it requires an exponential number of data samples in the number of independent variables to learn a function, making it prohibitive for even moderately sized problems.

2. More efficient learning is possible by putting assumptions into the model. All practical methods of machine learning make assumptions. In this course, we'll examine some specific assumptions, demonstrate how many data samples are required for learning with those assumptions and show some methods which perform almost as well as possible.

These will be the themes of this course. While we motiviate these ideas from portfolio selection, the ideas are applicable to any area that machine learning can be applied to. In order to explore these areas, we will require some mathematics that not all readers will have seen, particularly from real analysis, probability theory and stochastic processes. We develop this mathematics in the process of demonstrating the results. Finally, we'd like to note that the example discussed in this section was for illustrative purposes only. We don't believe that all stocks are equivalent to each other and don't recommend investing using that viewpoint. We will present more realistic applications of machine learning to portfolio selection later in the text.

Figure 1.2: Estimated Density with respect to the Binomial Distribution



1.3 What This Course Isn't

This course won't teach you the latest techniques in deep learning. We hope to look at neural networks as one of several types of learning mechanisms which achieves near optimal performance in certain situations. Convolutional neural networks in particular are a way of incorporating particular types of knowledge about invariants or symmetries of certain machine learning problems such as image recognition. This course will teach you methods of analysis which are applicable to all machine learning techniques. This course will teach you how to apply certain machine learning techniques to portfolio selection and optimization. However, it will not teach the latest individual techniques.

Chapter 2

The St. Petersburg Paradox, Probability and Utility

2.1 The St. Petersburg Paradox

In order to discuss machine learning for portfolio selection, we will first explore portfolio optimization criteria given perfect information about the statistics of the underlying financial instruments. Machine learning will allow us to extend these results to cases when the statistics are unknown.

We begin the story in the year 1713, when Nicolas Bernoulli, a member of a family of influential mathematicians, proposed the problem which has come to be known as the St. Petersburg paradox. At the time, games of chance and gambling had become popular and the Bernoullis made fundamental contributions to the development of classical probability theory to analyze such games. In this problem, one is offered a wager where a coin is flipped repeatedly until a head comes up. If a head comes up on the first flip, the bettor is paid \$1. The payoff is doubled for every additional flip required before the first head appears so that the bettor is paid \$2 if the first head appears on the second flip, \$4 if it appears on the third flip, etc. The question is what is a fair value for the bettor to pay up front to enter into this wager.

Let's first consider some simpler games. Consider a game in which the bettor is paid \$1 if a head comes up on a single flip of a fair coin and nothing otherwise. If one were to play this game repeatedly, over many trials, the frequency with which a head came up would tend towards the probability of $\frac{1}{2}$. Hence, a fair value for the bettor to pay to enter this game would be \$0.50 since, playing it repeatedly many times, one would tend towards breaking even. Now consider extending this to the first 2 flips of the St. Petersburg paradox game. If a head comes up on the first flip, the bettor gets \$1 and, otherwise, if a head comes up

on the second flip. the bettor gets \$2. On the first toss, a head comes up with probability $\frac{1}{2}$ whence the bettor is paid \$1. Given that a head doesn't come up on the first toss, which occurs with the remaining probability of $\frac{1}{2}$, there is a $\frac{1}{2}$ probability that a head comes up on the second toss. Again, the frequency of occurrence of a first toss head or a first toss tail and a second toss head will tend towards their probabilities. Hence, if the game is played repeatedly many times, the total payoff to the bettor will tend towards:

$$\frac{1}{2}\$1 + \frac{1}{2}\frac{1}{2}\$2 = \$1$$

Generalizing this to n tosses, we see that the the expected value is:

$$\sum_{i=1}^{n} \left(\frac{1}{2}\right)^n \$2^{n-1} = \$\sum_{i=1}^{n} \frac{1}{2} = \$\frac{n}{2}$$

Note that this converges to ∞ as $n \to \infty$. Does this mean that the fair value for the original game, where we toss the die until we get the first head, is infinite? Would you be willing to pay an arbitrarily large amount of money to play this game? Intuitively, it seems not. This is called the St. Petersburg paradox.

Note that this idea that the frequency of occurence converges towards the probability is called the **law of large numbers** and was in fact first proven by Jacob Bernoulli, the uncle of Nicolas Bernoulli, who discovered the St. Petersburg paradox. We'll go over the law of large numbers in much more generality when we discuss modern probability theory. However, we first discuss the mathematics behind the classical probability theory of the Bernoullis and others.

2.2 Classical Probability Theory

We will now present the classical theory of probability. In order to make it easier to present the modern theory later, we focus on algebras which requires a review of basic set theory first. Algebras and modern probability theory will allow us to apply probability theory to many phenomena that classical probability wouldn't apply to.

2.2.1 Finite Set Theory

Defining Sets

Set theory is commonly used for the basis of all of mathematics and especially for probability theory. A set is a collection of items, called members of the set. As an example, we write a set with elements 1, 2 and 3 as $\{1, 2, 3\}$. Note that order doesn't matter for sets, that is, $\{1, 2, 3\} = \{3, 2, 1\}$. Also, an element can not occur in a set multiple times so that $\{1, 1, 2, 3\} = \{1, 2, 3\}$. Other examples of sets that will be useful later are the set of numbers of the faces of a die,

$\{1, 2, 3, 4, 5, 6\}$ and the set of possible outcomes of a coin toss, which we write as $\{H, T\}$, where H means heads and T means tails.

We now define 2 important relations (the mathematical term for a property) on sets. We write $s \in S$ when the element s is a member of the set S. Also, we say that a set S is a **subset** of a set T, written $S \subseteq T$ if every element of S is in T, that is, $s \in S$ implies that $s \in T$.

Our discussion of set theory will be greatly facilitated by using set comprehension notation, $\{x : P(x)\}$ which means the set of all x such that some property P holds. For example, the set:

$$\{x : x \text{ is even}\}$$

denotes the set of even numbers. This notation may be familiar to readers fluent in the Python programming language which has syntax for list comprehensions. The notation originated in mathematics and has migrated to modern programming languages in recent decades. In more formal set theory, care must be taken in using set comprehensions to avoid certain mathematical paradoxes but our setup will already prohibit these.

We now define two special sets:

Definition 2.2.1.

Empty set: *the empty set, $\varnothing$, is set with no elements.*

Universal set: *for our purposes, we assume there is some universal set, denoted by Ω, which contains all base elements we could be interested in. Note in addition to subsets of the universal set, we will also use sets of subsets of the universal set.*

Union, Intersection and Complement

Sets can be combined into other sets by the operations of union and intersection:

Definition 2.2.2.

Union: *the union of sets S_1 and S_2, written $S_1 \cup S_2$ is defined as:*

$$S_1 \cup S_2 = \{s : s \in S_1 \text{ or } s \in S_2\}$$

The union is the set of elements contained in either set.

Intersection: *the intersection of sets S_1 and S_2, written $S_1 \cap S_2$ is defined as:*

$$S_1 \cap S_2 = \{s : s \in S_1 \text{ and } s \in S_2\}$$

The intersection is the set of elements contained in both sets.

Union and intersection have the following mathematical properties:

- **Associativity:**

$$S \cup (T \cup U) = (S \cup T) \cup U$$
$$S \cap (T \cap U) = (S \cap T) \cap U$$

- **Identity:**

$$S \cup \varnothing = S$$
$$S \cap \Omega = S$$

- **Commutativity:**

$$S \cup T = T \cup S$$
$$S \cap T = T \cap S$$

- **Idempotence:**

$$S \cup S = S$$
$$S \cap S = S$$

- **Distributivity:**

$$S \cup (T \cap U) = (S \cup T) \cap (S \cup U)$$
$$S \cap (T \cup U) = (S \cap T) \cup (S \cap U)$$

We now define the unary operation of set complement:

Definition 2.2.3. *The* **complement** *of a set $S \subseteq \Omega$ is defined as:*

$$\tilde{S} = \{s \in \Omega : s \notin S\}$$

The complement is the set of element which are in the universal set Ω but not in S. Complement has the following properties:

- **Empty/universal set duality:**

$$\tilde{\varnothing} = \Omega$$

- **Involution:**

$$\tilde{\tilde{S}} = S$$

The following formulas govern the use of unions, complements and intersections together:

- **De Morgan's laws:**

$$\widetilde{S \cup T} = \tilde{S} \cap \tilde{T}$$
$$\widetilde{S \cap T} = \tilde{S} \cup \tilde{T}$$

We can also define the difference between two sets S and T as:

$$S - T = S \cap \tilde{T}$$

While many of the properties of addition hold for union, set difference is not the inverse operation of union. For addition and subtraction, we have that $x + y - y = x$. However, for union and set difference:

$$S \cup T - T = (S \cup T) \cap \tilde{T} = \left(S \cap \tilde{T}\right) \cup \left(T \cap \tilde{T}\right) = \left(S \cap \tilde{T}\right) \cup \varnothing = S \cap \tilde{T}$$

Note that $S \cap \tilde{T} = S$ only if $S \cap T = \varnothing$ in which case we say that S and T are **disjoint**. In fact, there is no inverse for union in general.

Cartesian Product

In many cases, we wish to determine probabilities of combined observations. For example, in the St. Petersburg paradox, we wish to know the probability that a head comes up on the first n tosses of a coin. Each toss is a separate observation. The Cartesian product is the mathematical operation which allows us to combine separate observations.

A **sequence** is a collection of objects in a fixed order. As an example, the sequence of Fibonacci numbers is $(1, 1, 2, 3, 5, 8, \ldots)$. While readers might

be familiar with infinite sequences from calculus, sequences can also be finite. We write the $(1, 1, 2, 3)$ for the sequence of the first 4 Fibonacci numbers. As mentioned, for sequences, unlike sets, order matters so that $(1, 2, 3) \neq (3, 2, 1)$. Also unlike sets, an element can occur multiple times in a sequence so that $(1, 1, 2, 3) \neq (1, 2, 3)$ A sequence of toss of 4 heads followed by 4 tails on a coin might be represented as (H, H, H, H, T, T, T, T).

We can now define the Cartesian product:

Definition 2.2.4. *The* **Cartesian product** *of two sets S and T, written $S \times T$, is the set of all sequences with the first element in S and the second element in T:*

$$S \times T = \{(s, t) : s \in S \text{ and } t \in T\}$$

As an example, if $\{H, T\}$ represents either a head or a tail occuring in a toss of a coin, then $\{H, T\} \times \{H, T\}$ represents the set of possible outcomes of two coin tosses:

$$\{H, T\} \times \{H, T\} = \{(H, H), (H, T), (T, H), (T, T)\}$$

The Cartesian product of any number of sets, $S_1, S_2, \ldots, S_n$ is similarly defined:

$$S_1 \times S_2 \times \ldots \times S_n = \{(s_1, s_2, \ldots, s_n) : \text{for all } i,\ s_i \in S_i\}$$

Note that the Cartesian product will make algebras, which we define below, more useful than densities or distributions of random variables, the basis of classical probability theory as it is often taught. Combining random variables in infinite Cartesian products will allow us to define probability measures, essentially the same as distributions but which can be defined on infinite sets of random variables which can be related to each other in arbitrary ways. However, we will first need to avoid certain paradoxes related to infinities which we discuss after defining algebras and finite probability theory.

Note that, if we strictly follow the definition, the sequences (a, b, c), $((a, b), c)$ and $(a, (b, c))$ are distinct. This can be seen, for example, in programming languages which have tuples which correspond to finite sequences in mathematics. However, for our purposes, we will equate these constructs. Sets of sequences are typically constructed by Cartesian products and so we are assuming that Cartesian products "flatten" any sequences in the sets so that, for example:

$$A \times B \times C = \{(a, b, c) : a \in A,\ b \in B \text{ and } c \in C\}$$

2.2.2 Algebras

An algebra of sets defines the events whose probability we might be interested in in finite probability theory. We define an algebra as follows:

Definition 2.2.5. *An* **algebra** *is a set of subsets, $\mathcal{F}$ of a universal set Ω with the following properties:*

1. $\Omega \in \mathcal{F}$
2. *If $F_1, F_2 \in \mathcal{F}$, then $F_1 \cup F_2 \in \mathcal{F}$.*
3. *If $F \in \mathcal{F}$, then $\tilde{F} \in \mathcal{F}$.*

As we will see, an algebra ultimately captures all the sets of observations that we wish to determine the probability of. Some straightforward implications of these definitions follow:

- The empty set is in $\mathcal{F}$, that is, $\varnothing \in \mathcal{F}$. This can be shown as follows:
  1. $\Omega \in \mathcal{F}$ by Item 1 of Definition 2.2.5.
  2. $\varnothing = \tilde{\Omega} \in \mathcal{F}$ by the empty/universal set duality and Item 3 of Definition 2.2.5.
- Intersections of sets in $\mathcal{F}$ are in $\mathcal{F}$:

$$S_1, S_2 \in \mathcal{F} \Rightarrow S_1 \cap S_2 \in \mathcal{F}$$

- Arbitrary finite unions are in $\mathcal{F}$:

$$S_1, S_2, \ldots, S_n \in \mathcal{F} \Rightarrow S_1 \cup S_2 \cup \ldots \cup S_n = \bigcup_{i=1}^{n} S_i \in \mathcal{F}$$

  The same holds for intersections.

We now provide some examples of algebras:

Example 2.2.1. *Given any set Ω, the set of all of subsets of that set is called the power set and denoted $\mathcal{P}(\Omega)$. $\mathcal{P}(\Omega)$ is always an algebra. For example, Ω could be $\{H, T\}$ in which case $\mathcal{P}(\Omega) = \{\varnothing, \{H\}, \{T\}, \{H, T\}\}$ which represents the events corresponding to a single flip of a coin.*

Example 2.2.2. *Similar to the last example, if $C = \{H, T\}$ then $\mathcal{P}(C \times C)$ is the algebra representing the events corresponding to two flips of a coin.*

Example 2.2.3. *Extending the last example, define C^n as the n-fold Cartesian product of C:*

$$C^n = \underbrace{C \times C \times \ldots \times C}_{n \text{ times}}$$

$\mathcal{P}(C^n)$ is the algebra corresponding with the observation of n tosses of a coin.

Example 2.2.4. *To see an example of an algebra which is not of the form $\mathcal{P}(\Omega)$, consider the set of natural numbers $\mathbb{N} = \{1, 2, 3, \ldots\}$. Let $\mathcal{F}$ be the set of all finite subsets of $\mathbb{N}$ or complements of such sets:*

$$\mathcal{F} = \{S : S \text{ is finite or } \mathbb{N} - S \text{ is finite}\}$$

So, for example $\{1, 2, 3\} \in \mathcal{F}$ but $\{x : x \text{ is even}\}$ is not. It can be demonstrated that $\mathcal{F}$ is an algebra.

Example 2.2.5. *Consider the St. Petersburg game. The possible outcomes are given by:*

$$\Omega = \{(H), (T, H), (T, T, H), \ldots\} \cup \{(T, T, T \ldots)\}$$

Since this can be parameterized by the flip on which the first head appears, we could also use $\Omega = \mathbb{N} \cup \{\infty\}$. Hence, this is similar to the previous example where we could use sets with a finite number of elements and their complements as the algebra or we could use the entire power set.

Many authors refer to the sets of an algebra as "events". For example, if $\Omega = C^n$, then the set $S_{i,x} = \{(c_1, c_2, \ldots, c_n) : c_i = x\}$ corresponds to the "event" that the *ith* toss of the coin is x. We will later place further restrictions on algebras when we start to talk about infinite numbers of events. There are basically two uses for such restricted algebras in probability theory. These can represent the events which we know the probability of. Alternatively, the can represent the events which we know the outcome of. Given that the power set is always an algebra, couldn't we always use that as the algebra? It is useful to restrict what we know, i.e. use a strict subset of the power set as the algebra, so that we will be able to represent what we know at different points in time. We know the price of a stock yesterday but not tomorrow. There are other, more technical reasons for restricting algebras which we will discuss further when we discuss infinite sequences of events.

2.2.3 Finite Probability Measures and Finite Probability Spaces

Now that we have defined an algebra, we can define a probability measure:

Definition 2.2.6. *A* **finite probability measure***, sometimes called a probability distribution, on algebra $\mathcal{F}$, is a set function with values in the interval $[0,1]$, that is, a function $\mathbb{P}:\mathcal{F}\to[0,1]$, such that:*

1. **Normalization:** $\mathbb{P}(\Omega)=1$
2. **Additivity:** *if $S_1,S_2\in\mathcal{F}$ are disjoint, then $\mathbb{P}(S_1\cup S_2)=\mathbb{P}(S_1)+\mathbb{P}(S_1)$*

Some immediate consequences of this definition are:

- $\mathbb{P}(\varnothing)=0$
- If $S_1,S_2,\ldots,S_n$ are pairwise disjoint, $\mathbb{P}(\bigcup_{i=1}^n S_i)=\sum_{i=1}^n\mathbb{P}(S_i)$ where we are using the following notation:

$$\bigcup_{i=1}^{n} S_i = S_1\cup S_2\cup\ldots\cup S_n$$

- If $S_1\subseteq S_2$, then $\mathbb{P}(S_1)\leq\mathbb{P}(S_2)$
- $\mathbb{P}\left(\tilde{S}\right)=1-\mathbb{P}(S)$

We now give some examples of methods of constructing probability measures:

Example 2.2.6. *Let $p:\Omega\to[0,1]$ be a function such that:*

$$\sum_{\omega\in\Omega} p(\omega)=1$$

Then, the set function, $\mathbb{P}:\mathcal{P}(\Omega)\to[0,1]$, given by:

$$\mathbb{P}(S)=\sum_{s\in S} p(s)$$

is a probability measure. We can see that Item 1 of Definition 2.2.6 holds because:

$$\mathbb{P}(\Omega)=\sum_{s\in\Omega} p(s)=1$$

Item 2 of Definition 2.2.6 (finite additivity) holds because, if S_1 and S_2 are disjoint:

$$\mathbb{P}(S_1\cup S_2) \quad = \quad \sum_{s\in S_1\cup S_2} p(s)=\sum_{s\in S_1} p(s)+\sum_{s\in S_2} p(s)=\mathbb{P}(S_1)+\mathbb{P}(S_2)$$

When such a function p exists, which need not always be the case, it is typically called a **point mass function***.*

Example 2.2.7. *If Ω is any finite set then the function $p : S \to [0,1]$ defined by:*

$$p(s) = \frac{1}{|\Omega|}$$

where $|\Omega|$ denotes the size of the set Ω is a point mass function since:

$$\sum_{s\in\Omega} p(s) = \sum_{s\in\Omega} \frac{1}{|\Omega|} = \frac{|\Omega|}{|\Omega|} = 1$$

The corresponding distribution is given by:

$$\mathbb{P}(S) = \frac{|S|}{|\Omega|}$$

and called the **uniform** *distribution.*

Example 2.2.8. *Let $\Omega = S^n$ for a finite set S. Suppose we have point mass functions $p_i : S \to [0,1]$ for $i = 1, 2, \ldots, n$. We can define a point mass function, q, on Ω as follows:*

$$q\left((s_1, s_2, \ldots, s_n)\right) = \prod_{i=1}^{n} p_i\left(s_i\right)$$

We can see that this is a point mass function since:

$$\begin{aligned}
\sum_{(s_1,s_2,\ldots,s_n)\in S^n} q\left(s\right) &= \sum_{(s_1,s_2,\ldots,s_n)\in S^n} \prod_{i=1}^{n} p_i\left(s_i\right) \\
&= \sum_{s_1\in S}\sum_{s_2\in S}\cdots\sum_{s_n\in S} p_1\left(s_1\right) * p_2\left(s_2\right) * \ldots * p_n\left(s_n\right) \\
&= \sum_{s_1\in S} p_1\left(s_1\right) * \sum_{s_2\in S} p_2\left(s_2\right) * \ldots * \sum_{s_n\in S} p_n\left(s_n\right) \\
&= \prod_{i=1}^{n}\sum_{s_i\in S} p_i\left(s_i\right) = \prod_{i=1}^{n} 1 = 1
\end{aligned}$$

Example 2.2.9. *Suppose $\mathbb{P}_1, \mathbb{P}_2, \ldots, \mathbb{P}_n$ are probability measures all defined on the same algebra $\mathcal{F}$. Let p be a point mass function on $\{1, 2, \ldots, n\}$. We define a probability measure $\mathbb{Q} : \mathcal{F} \to [0,1]$ as follows:*

$$\mathbb{Q}(S) = \sum_{i=1}^{n} p(i)\,\mathbb{P}_i(S)$$

To see that Item 1 of 2.2.6 holds:

$$\mathbb{Q}(\Omega) = \sum_{i=1}^{n} p(i)\,\mathbb{P}_i(\Omega) = \sum_{i=1}^{n} p(i) = 1$$

Finite additivity holds because, for disjoint sets $S_1, S_2 \in \mathcal{F}$:

$$\begin{aligned}
\mathbb{Q}(S_1 \cup S_2) &= \sum_{i=1}^{n} p(i)\,\mathbb{P}_i(S_1 \cup S_2) \\
&= \sum_{i=1}^{n} p(i)\,(\mathbb{P}_i(S_1) + \mathbb{P}_i(S_2)) \\
&= \sum_{i=1}^{n} p(i)\,\mathbb{P}_i(S_1) + \sum_{i=1}^{n} p(i)\,\mathbb{P}_i(S_2) \\
&= \mathbb{Q}(S_1) + \mathbb{Q}(S_2)
\end{aligned}$$

This contruction is called a **mixture** *distribution.*

The concept of a probability space encapsulates the universal set, the algebra and the probability measure:

Definition 2.2.7. *A* **finite probability space** *is a triple $(\Omega, \mathcal{F}, \mathbb{P})$ where Ω is the universal set, $\mathcal{F}$ is an algebra and $\mathbb{P}$ is a probability measure.*

2.2.4 Random Variables

We will now introduce random variables. We first define the inverse image of a function:

Definition 2.2.8. *Let A and B be sets and $S \subseteq B$. The* **inverse image** *of S under the function $f : A \to B$, is given by:*

$$f^{-1}(S) = \{a \in A : f(a) \in S\}$$

The inverse image is the set of values which map into S under the function f.

Example 2.2.10. *Let $f : \{-1, 0, 1\} \to \{0, 1\}$ be given by $f(x) = x^2$. Hence:*

$$\begin{aligned}
f^{-1}(\{0\}) &= \{0\} \\
f^{-1}(\{1\}) &= \{-1, 1\}
\end{aligned}$$

Note that inverse image respects unions:

$$\begin{aligned} f^{-1}(A \cup B) &= \{x : f(x) \in A \cup B\} = \{x : f(x) \in A \text{ or } f(x) \in B\} \\ &= \{x : f(x) \in A\} \cup \{x : f(x) \in B\} = f^{-1}(A) \cup f^{-1}(B) \end{aligned}$$

In fact, it also respects intersections, complements and infinite unions and intersections. Hence, the inverse images of the set of an algebra is an algebra as will be useful later.

We can now define random variables:

Definition 2.2.9. *A* **random variable** *on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is a functon $X : \Omega \to \mathbb{R}$ such that the inverse image of every interval $X^{-1}([a,b]) \in \mathcal{F}$ for all $a, b \in \mathbb{R}$.*

We denote random variables with capital roman letters, e.g. $X, X_1, Y, \ldots$. The intuition is that a random variable is a variable, that is, a quantity that could take on multiple possible values, and its value is chosen randomly. We insist that $X^{-1}([a,b]) \in \mathcal{F}$ so that we can measure the probability that the value of X is in $[a,b]$ since we can only determine the probability of sets in $\mathcal{F}$. We now give a common construction of random variables.

Example 2.2.11. *Consider the probability space $(C^n, \mathcal{P}(C^n), \mathbb{P})$. The random variable X_i is defined as $X_i((c_1, c_2, \ldots, c_n)) = \mathbb{1}_H(c_i)$ where the function $\mathbb{1}$ is defined as:*

$$\mathbb{1}_x(y) = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{otherwise} \end{cases}$$

X_i is the random variable which is 1 if the ith *flip of the coin comes up head and 0 otherwise. In this case, since the algebra is the power set, the condition on the inverse image automatically holds.*

To demonstrate a function which is not (always) a random variable, we give the following example:

Example 2.2.12. *An example of a function that might or might not be a random variable, depending upon the algebra is given by the function of Example 2.2.10. Since $f^{-1}(\{1\}) = \{-1, 1\}$, if the algebra is $\mathcal{F} = \{\varnothing, \{-1\}, \{0,1\}, \{-1,0,1\}\}$, then f is not a random variable. On the other hand, if the algebra is $\mathcal{P}(\{-1,0,1\})$, then f will be a random variable.*

The following intuitive notation will greatly simplify our discussion of random variables:

$$\mathbb{P}(X_1 \in S) = \mathbb{P}\big(X^{-1}(S)\big) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \in S\})$$

Similarly, we write:

$$\mathbb{P}(X_1 = x) = \mathbb{P}\big(X^{-1}(\{x\})\big) = \mathbb{P}(\{\omega \in \Omega : X(\omega) = x\})$$

and likewise for other relatiions such as inequalities.

2.2.5 Independence

One of the most important concepts in probability theory is the concept of independence. We say that two events are independent if the outcome of either of the events does not effect the probability of the outcome of the other. This corresponds to the following mathematical notion which we use as a definition:

Definition 2.2.10. *Given a probability space* $(\Omega, \mathcal{F}, \mathbb{P})$*, sets* $F_1, F_2 \in \mathcal{F}$ *are* **independent** *if:*

$$\mathbb{P}(F_1 \cap F_2) = \mathbb{P}(F_1) * \mathbb{P}(F_2)$$

We note some consequences of this definition:

1. **Symmetry:** If F_1 is independent of F_2 then F_2 is independent of F_1.
2. **Negation stability:** If F_1 and F_2 are independent, then so are any pair from $\left\{F_1, F_2, \tilde{F}_1, \tilde{F}_2\right\}$. For example:

$$\begin{aligned}\mathbb{P}\left(\tilde{F}_1 \cap F_2\right) &= \mathbb{P}((\Omega - F_1) \cap F_2) = \mathbb{P}(F_2 - F_1 \cap F_2) \\ &= \mathbb{P}(F_2) - \mathbb{P}(F_2) * \mathbb{P}(F_1) = (1 - \mathbb{P}(F_1)) * \mathbb{P}(F_2) \\ &= \mathbb{P}\left(\tilde{F}_1\right) * \mathbb{P}(F_2)\end{aligned}$$

3. **Pairwise does not imply n-wise independence:** Note that each pair from $\{F_1, F_2, F_3\}$ is independent, it does NOT imply that:

$$\mathbb{P}(F_1 \cap F_2 \cap F_3) = \mathbb{P}(F_1) * \mathbb{P}(F_2) * \mathbb{P}(F_3) \tag{2.1}$$

To see this, let $\Omega = C^2$ and consider a probability space $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$. Define:

$$\begin{aligned}F_1 &= \{(c_1, c_2) : c_1 = H\} \\ F_2 &= \{(c_1, c_2) : c_2 = H\} \\ F_3 &= \{(H, T), (T, H)\}\end{aligned}$$

Let the probability measure $\mathbb{P}$ be uniform. Then, it can be seen that:

$$\begin{aligned}&\mathbb{P}(F_1) = \mathbb{P}(F_2) = \mathbb{P}(F_3) = \frac{1}{2} \\ &\mathbb{P}(F_1 \cap F_2) = \mathbb{P}(F_1 \cap F_3) = \mathbb{P}(F_2 \cap F_3) = \frac{1}{4} \\ &\mathbb{P}(F_1 \cap F_2 \cap F_3) = 0\end{aligned}$$

which is incompatiable with Equation 2.1.

Due to Item 3 above, we use 2.1 to define 3-way independence and analogously for n-wise independence.

Random variables are independent if all relevant events corresponding to the are independent:

Definition 2.2.11. *Random variables $X_1, X_2, \ldots, X_n$ defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ are independent if:*

$$\begin{aligned} & \mathbb{P}(X_1 \in [a_1, b_1], X_2 \in [a_2, b_2], \ldots, X_n \in [a_n, b_n]) \\ = & \prod_{i=1}^{n} \mathbb{P}(X_i \in [a_i, b_i]) \end{aligned}$$

for all $a_1, a_2, \ldots, a_n \in \mathbb{R}$ and $b_1, b_2, \ldots, b_n \in \mathbb{R}$.

We now give examples of random variables which are independent and one of random variables which are not:

Example 2.2.13. *Consider the probability space of Example 2.2.8 with the random variables defined in Example 2.2.11. In this case, the random variables $X_1, X_2, \ldots, X_n$ are independent.*

Example 2.2.14. *Consider the case of the previous example where all the point mass functions are equal, $p_1 = p_2 = \ldots = p_n$. This is the most common situation in statistics and is referred to as* **independent and identically distributed**, *abbreviated as IID.*

Example 2.2.15. *Consider picking a coin from a hat containing one fair coin and one coin with heads on both sides. We then flip the coin we had picked n times. This is a mixture distribution, as mentioned in Example 2.2.9, of two IID distributions with point mass function $p_1, p_2 : C \to [0, 1]$:*

$$\begin{aligned} p_1(x) &= \frac{1}{2} \\ p_2(x) &= \begin{cases} 1 & \text{if } x = H \\ 0 & \text{if } x = T \end{cases} \end{aligned}$$

Each point mass function occurs with equal probability so that the mixture distribution has point mass function on C^n is given by:

$$p((c_1, c_2, \ldots, c_n)) = \frac{1}{2}\left(\frac{1}{2}\right)^n + \frac{1}{2}\prod_{i=1}^{n} \mathbb{1}_H(c_i)$$

Note that:

$$\mathbb{P}(X_i = H) = \frac{3}{4} \tag{2.2}$$

for each i but:

$$\mathbb{P}(X_1 = H, X_2 = H, \ldots, X_n = H) = \frac{1}{2}\left(\frac{1}{2}\right)^n + \frac{1}{2}$$

which is not the product of equation (2.2) so that X_i are not independent.

2.2.6 Expectation

The expected value of a random variable corresponds with the intuitive notion of the average value and is defined as follows:

Definition 2.2.12. *For a random variable X on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ having a point mass function $p : \Omega \to [0, 1]$, the* **expected value** *of X, denoted $E[X]$ is defined as:*

$$E[X] = \sum_{\omega \in \Omega} p(\omega)\, X(\omega)$$

The expected value has several important properties:

1. For any constant random variable a, we have that:

$$E[a] = a \tag{2.3}$$

   Note that we are using a to denote the random variable $X : \Omega \to \mathbb{R}$ which is a constant with value $a \in \mathbb{R}$, that is, $X(\omega) = a$ for all $\omega \in \Omega$. This notation, which is standard, makes no distinction between the real number $a \in \mathbb{R}$ and the random variable, $a : \mathcal{F} \to \mathbb{R}$. Indeed, the a on the left hand side of Equation (2.3) represents the random variable and that on the right hand side represents the numerical constant.

2. **Linearity:** For constants a and b and random variables X and Y:

$$E[aX + bY] = aE[X] + bE[Y]$$

3. **Monotonicity:** If X is a positive random variable, which we write as $X \geq 0$ then $E[X] \geq 0$.

2.3 Utility Theory

2.3.1 Bernoulli's Resolution of the St. Petersburg Paradox

Recall that the St. Petersburg paradox what the problem of determining of the fair value of a game in which a fair coin is tossed until a head occurs on the *nth* toss at which point the player is payed 2^{n-1}. The paradox occurs because the expected value of the game is ∞ which we show more fully below before we discuss the solution of Daniel Bernoulli, the cousin of Nicolas Bernoulli, who had first posed the paradox.

We now put this into the language of the probability theory that we have learned. Since we can potentially toss the coin an infinite number of times, we will use an infinite Cartesian product as the sample space, that is, the sample space is the space of all possible infinite sequences of coin tosses:

$$\Omega = C^{\infty} = \{(c_1, c_2, \ldots) : c_i \in C \text{ for all } i \in \mathbb{N}\}$$

We want to be able to determine the probability of each coin toss so let X_n be the random variable corresponding the *nth* coin toss as mentioned in Example 2.2.11 but now for the infinite coin toss case:

$$X_n((c_1, c_2, \ldots)) = \mathbb{1}_H(c_n)$$

For any $c^* \in C$, we will want to determine the probability of $\{X_n = c^*\}$, using the notation of Section 2.2.4, that is:

$$\begin{aligned}&\{X_n = \mathbb{1}_H(c^*)\} = \\ &\quad \{(c_1, \ldots, c_{n-1}, c^*, c_{n+1}, \ldots) : c_i \in C \text{ for all } i \in \{1, \ldots, n-1, n+1, \ldots\}\}\end{aligned}$$

Since the algebra must be closed under intersections, it must also contain the intersection of such sets. In particular, for any $c_1^*, \ldots, c_n^*$, the algebra must contain the set:

$$\begin{aligned}&\{X_1 = \mathbb{1}_H(c_1^*), \ldots, X_n = \mathbb{1}_H(c_n^*)\} \\ &\quad = \bigcap_{i=1}^{n} \{X_i = \mathbb{1}_H(c_i^*)\} = \{X_1 = \mathbb{1}_H(c_1^*)\} \cap \ldots \cap \{X_n = \mathbb{1}_H(c_n^*)\} \\ &\quad = \{(c_1^*, \ldots, c_n^*, c_{n+1}, \ldots) : c_{n+1}, \ldots \in C \text{ for all } i \in \{n+1, \ldots\}\}\end{aligned}$$

If we take finite unions of sets of this nature and add the empty set, the result is an algebra that we denote by $\mathcal{F}$. This algebra contains any possible combination of outcomes on any finite sequence of the coins that we intend to flip. However,

it does not contain possible outcomes of an infinite sequences of coin flips. We will discuss this in detail later.

We now derive the probability measure which we denote by $\mathbb{P}$. Note that, since each coin flip is independent, we have that, for any $i_1, i_2, \ldots, i_n \in \mathbb{N}$ and any $c_1^*, c_2^*, \ldots, c_n^* \in C$:

$$\begin{aligned}&\mathbb{P}(X_{i_1} = \mathbb{1}_H(c_1^*), X_{i_2} = \mathbb{1}_H(c_2^*), \ldots, X_{i_n} = \mathbb{1}_H(c_n^*)) = \\ &\quad \mathbb{P}(X_{i_1} = \mathbb{1}_H(c_1^*)) \times \mathbb{P}(X_{i_2} = \mathbb{1}_H(c_2^*)) \times \ldots \times \mathbb{P}(X_{i_n} = \mathbb{1}_H(c_n^*))\end{aligned}$$

From this we can derive the probability that the first head comes up on the *nth* toss:

$$\begin{aligned}&\mathbb{P}(X_1 = 0, X_2 = 0, \ldots, X_{n-1} = 0, X_n = 1) \\ &= \mathbb{P}(X_1 = 0) \times \mathbb{P}(X_2 = 0) \times \ldots \times \mathbb{P}(X_{n-1} = 0) \times \mathbb{P}(X_n = 1) \\ &= \underbrace{\frac{1}{2} \times \frac{1}{2} \times \ldots \times \frac{1}{2} \times \frac{1}{2}}_{n \text{ times}} = 2^{-n}\end{aligned}$$

The triple $(\Omega, \mathcal{F}, \mathbb{P})$ is the probability space corresponding to this game. Note that, as alluded to in Example 2.2.5, we can also place this game on the probablity space:

$$(\mathbb{N} \cup \{\infty\}, \mathcal{P}(\mathbb{N} \cup \{\infty\}), \mathbb{Q})$$

where $\mathbb{Q}$ is defined according to the point mass function:

$$q(x) = \begin{cases} 2^{-n} & \text{if } x \in \mathbb{N} \\ 0 & \text{if } x = \infty \end{cases}$$

Note that the set $\{\infty\}$ corresponds with always throwing tails, that is, $\{(T, T, \ldots)\}$ which is not in the algebra $\mathcal{F}$ discussed previously. We will discuss this subtlety later.

In thinking about this problem, Daniel Bernoulli posited that the "utility" (usefulness) of an incremental amount of wealth to a person is inversely proportional to the amount of wealth that person has already accrued. If we let $U(w)$ denote the utility of wealth w, this means that:

$$\frac{dU(w)}{dw} = \frac{c}{w}$$

This is be solved by $U(w) = c \log_2(w)$. Note the constant c will have no impact on the solution, assuming it's positive, will have no impact on the solution since this function, which will be called the utility function, will be used to decide

between participating and not participating in a game and multiplication by a positive constant will affect both decisions equally. We choose $c = 1$ for convenience. Assuming a person has initial wealth w and that the cost to enter the game is p, the expected value of the utility is:

$$\sum_{i=1}^{\infty} q(i) \log_2 \left(w - p + 2^{i-1}\right) = \sum_{i=1}^{\infty} 2^{-i} \log_2 \left(w - p + 2^{i-1}\right)$$

However, we have that $w - p \leq 2^{i-1}(w - p)$ since $w - p \geq 0$ and $i \geq 1$. Hence:

$$\begin{aligned}
&\sum_{i=1}^{\infty} 2^{-i} \log_2 \left(w - p + 2^{i-1}\right) && (2.4)\\
&\leq \sum_{i=1}^{\infty} 2^{-i} \log_2 \left(2^{i-1} * (w - p) + 2^{i-1}\right)\\
&= \sum_{i=1}^{\infty} 2^{-i} \log_2 \left(2^{i-1} * (w - p + 1)\right)\\
&= \sum_{i=1}^{\infty} 2^{-i} (i - 1) + \sum_{i=1}^{\infty} 2^{-i} \log_2 (w - p + 1)\\
&= \sum_{i=1}^{\infty} 2^{-i} (i - 1) + \log_2 (w - p + 1)\\
&= \sum_{i'=0}^{\infty} \left(\frac{1}{2}\right)^{i'+1} i' + \log_2 (w - p + 1) && (2.5)\\
& && (2.6)
\end{aligned}$$

where there is an implicit assumption that the player can not borrow money to play the game since that would require taking the logarithm of a negative number. Now note that:

$$\sum_{i=0}^{\infty} i\alpha^{i-1} = \sum_{i=0}^{\infty} \frac{d}{d\alpha}\alpha^i = \frac{d}{d\alpha} \sum_{i=0}^{\infty} \alpha^i = \frac{d}{d\alpha} \frac{1}{1-\alpha} = \frac{1}{(1-\alpha)^2}$$

Substituting $\alpha = \frac{1}{2}$ into Equation (2.5) yields:

$$\begin{aligned}
&\sum_{i=1}^{\infty} q(i) \log_2 \left(w + 2^{i-1} - p\right)\\
&\leq \frac{1}{4} \frac{1}{\left(1 - \frac{1}{2}\right)^2} + \log_2 (w - p + 1) = 1 + \log_2 (w - p + 1)
\end{aligned}$$

Hence, the expected value is finite and there will be some finite cost which is equivalent to it.

Bernoulli's solution brings up the important point that expected value is not the best way to value games or investments. It is often better to evaluate the expected value of a utility function as we discuss in the next section. However, it still doesn't fully resolve the problem. If, instead of paying 2^{n-1} for the first head occuring on the *nth* toss, the game pays 2^{2^n} then the expected value of the $\log_2$ utility function will still be infinite. Another way to resolve this is that, in practice, no one person has infinite credit to extend to pay the winner of the game.

2.3.2 Expected Utility Theory

Before Daniel Bernoulli's solution, mathematicians considered the expected value of a bet or investment to be an effective way of determining its value. Consider the following example:

Example 2.3.1. *Choose from the following* 2 *bets:*

1. *Double your money* 100% *of the time.*
2. *A* 99% *chance to lose all your money and a 1% chance to multiply your money by* 200.

Both of these bets have same expected value but most people would choose Bet 1 over Bet 2 even though they have the same expected value. Indeed Bet 1 has no uncertainty or is risk-free wherease Bet 2 is risky.

In expected utility theory, agents choose among bets or investments by maximizing expected value of a function, $u : \mathbb{R} \to \mathbb{R}$ called the **utility function**. If X is a random variable representing the value of some bet or investment, the expected utility of X is $E[u(X)]$. For example, Daniel Bernoulli chose $u(x) = \log_2(x)$ which we will call log utility.

The utility function typically has the following desired properties:

1. **Monotonicity:** If $X_1 \leq X_2$, that is, X_2 yields a higher result for every possible $\omega \in \Omega$, then the utility of X_1 is less than the utility of X_2, that is, $E[x(X_1)] \leq E[u(X_2)]$. This occurs if and only if u is a monotonically increasing function, that is, if $x_1 \leq x_2$ implies that $u(x_1) \leq u(x_2)$. Log utility function is monotonically increasing. Finally, note that a differentiable function is monotonically increasing if and only if the derviative is positive.

2. **Concavity:** Example 2.3.1 shows that most people believe that risky bets should have less utility than riskless bets. If c is a constant random variable and X is a risky random variable with $E[X] = c$, then the utility of c is higher than that of X:

$$E[u(X)] \leq E[u(c)] \tag{2.7}$$

To understand what this means mathematically, consider the simplest risky bet which has only two possible outcomes. Suppose X yields a with probability p and b with the remaining probability. Equation (2.7 becomes:

$$E[u(X)] = pu(a) + (1-p)\,u(b) \leq u(E[X]) = u(pa + (1-p)\,b) \tag{2.8}$$

Functions u with the property that $pu(a)+(1-p)\,u(b) \leq u(pa+(1-p)\,b)$ are called **concave** and concave utility functions are called **risk-averse**. In fact, an important result in probability theory tells us that if Equation (2.8) holds for all risky bets with two outcomes, then Equation (2.7) holds for all risk bets:

Theorem 2.3.1. Jensen's inequality: *For any random variable X and any concave function u:*

$$E[u(X)] \leq u(E[X])$$

Log utility is concave. Finally, note that a twice differentiable function is concave if and only if the second derivative is negative.

We now given an example of a commonly used class of utility functions:

Example 2.3.2. *The constant relative risk aversion (CRRA) utility function is given by:*

$$u_\gamma(x) = \begin{cases} \frac{x^\gamma - 1}{\gamma} & \text{if } \gamma \leq 1 \text{ and } \gamma \neq 0 \\ \ln(x) & \text{if } \gamma = 0 \end{cases}$$

The exception for $\gamma = 0$ makes the function continuous at γ for each x. The first and second derivatives of u_γ are given by:

$$\frac{d}{dx}u_\gamma(x) = \frac{d}{dx}\frac{x^\gamma - 1}{\gamma} = \frac{\gamma x^{\gamma-1}}{\gamma} = x^{\gamma-1}$$
$$\frac{d^2}{dx^2}u_\gamma(x) = \frac{d}{dx}x^{\gamma-1} = (\gamma-1)\,x^{\gamma-2}$$

Note that for $x \geq 0$ (expected utility theory always assumes non-negative wealth), $u_\gamma(x)$ is monotonic since it's derivative is positive. Similarly, the u_γ is concave for $\gamma \leq 1$ since the second deriviative is negative.

In recent decades, expected utility has been shown to have failings as a descriptive theory, that is, as a theory of how people make economic decisions. The pioneering work of Kahneman and Tversky and many behavioral economists after them showed that people are subject to a number of fallacies that can't be explained by expected utility theory. Kahneman and Tversky developed theory own theory, prospect theory, to explain these. The burgeonging academic field of behavioral finance blossomed from their work. However, these developments are beyond the scope of this text. However, that still leaves the possibility that expected utility theory could be appropriate as a normative theory, that is, a theory of how people should make economic decisions. We investigate this next.

Chapter 3

The Growth Optimal Portfolio and the Laws of Large Numbers

3.1 The Importance of Sums of Random Variables

We now suppose that we know the probability distribution corresponding to a vector of stock prices and investigate how we should choose a portfolio based on this information. We will assume that we are interested in the long-term performance of the portfolio. In the next chapter, we begin to look at how we can use machine learning to determine this probability distribution.

Let $X_{t,i}$ denote the value of the *ith* stock at time $t \in \mathbb{N}$. Let's suppose we invest in a portfolio consisting of $w_{t,i} \in [0,1]$ fraction of the value of our portfolio in the *ith* stock at time t. Let V_t denote the value of our portfolio at time t. The value that we have invested in the *ith* stock at time t is $V_t w_{t,i}$. Hence, the number of shares of the *ith* stock that we hold at time t is given by $\frac{V_t w_{t,i}}{X_{t,i}}$. The total value of our portfolio at time $t+1$ is then, ignoring transaction costs:

$$V_{t+1} = \sum_i \frac{V_t w_{t,i}}{X_{t,i}} X_{t+1,i} = V_t \sum_i w_{t,i} \frac{X_{t+1,i}}{X_{t,i}}$$

Over n time periods, our wealth becomes:

$$V_{n+1} = V_1 \prod_{t=1}^{n} \sum_i w_{t,i} \frac{X_{t+1,i}}{X_{t,i}} = V_1 \exp\left(\sum_{t=1}^{n} \ln\left(\sum_i w_{t,i} \frac{X_{t+1,i}}{X_{t,i}}\right)\right)$$

where we have used the fact that $\ln(ab) = \ln(a) + \ln(b)$.

Since we wish to maximize our wealth over the long-term, we will maximize V_n for large n. Since ln is a monotonic function, maximizing V_n is the same as maximizing $\ln(V_n)$ or:

$$\ln(V_{n+1}) = \ln(V_1) + \sum_{t=1}^{n} \ln\left(\sum_i w_{t,i} \frac{X_{t+1,i}}{X_{t,i}}\right) \tag{3.1}$$

Hence, we are interested in the behavior of sums of random variables which we will explore in what follows. Much work on modern probability theory, which began in the 1930's, has focused on this. First, we must learn more about modern probability theory.

3.2 The Limitations of Finite Additivity

In exploring the behavior of sums of random variables, we will find the finite addivity axiom, Axiom 2 of Definition 2.2.6 to be overly restrictive. In order to accomodate this, we will also need to further restrict our algebras so that they are closed under more than just finite unions, i.e. Axiom 2 of Definition 2.2.5. The following example demonstrates how these assumptions are overly restrictive.

Example 3.2.1. *Consider a sequence of random variables $X_1, X_2, \ldots$. We wish to determine the behavior of $S_n = \sum_{i=1}^{n} X_i$, which can be shown to be a random variable, as n gets large. This will often diverge so we consider $\lim_{n\to\infty} \frac{S_n}{n}$. For example, if X_i is the* ith *flip of a fair coin, we might want to know the probability that this limit is equal to $\frac{1}{2}$. Recalling the definition of a limit, this means that for every $\epsilon > 0$, there is some N such that for all $n \geq N$:*

$$\frac{S_n}{n} \in \left[\frac{1}{2} - \epsilon, \frac{1}{2} + \epsilon\right]$$

For a particular ϵ and N, this corresponds with the following subset of Ω:

$$\bigcap_{n=N}^{\infty} \left\{ \frac{S_n}{n} \in \left[\frac{1}{2} - \epsilon, \frac{1}{2} + \epsilon\right] \right\}$$

Similarly, over all $\epsilon > 0$ and N, this corresponds to the following set:

$$\bigcap_{\epsilon>0} \bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} \left\{ \frac{S_n}{n} \in \left[\frac{1}{2} - \epsilon, \frac{1}{2} + \epsilon\right] \right\}$$

However, this set is not a finite intersection so we cannot guarantee that it is in our algebra. In fact, if we are using the algebra $\mathcal{F}$ of Section 2.3.1, this set is not in the algebra and so it is not possible to determine its probability.

The next example demonstrates that finite additivity sometimes does not correspond to our intuition about probabilities at all.

Example 3.2.2. *Consider the algebra of Example 2.2.4, that is:*

$$\mathcal{F} = \Big\{ S \subseteq \mathbb{N} : S \text{ is finite or } \tilde{S} \text{ is finite} \Big\}$$

Now consider the probability measure, $\mathbb{P}$ defined as follows:

$$\mathbb{P}(S) = \begin{cases} 0 & \text{if } S \text{ is finite} \\ 1 & \text{if } S \text{ is infinite} \end{cases}$$

This can be shown to be a probability measure. However, the probability of every number $n \in \mathbb{N}$ is $\mathbb{P}(\{n\}) = 0$. Every individual number has 0 probability any infinite set of numbers in the algebra has probability 1.

The solution to the above issues is that we would like for our algebras to be closed under infinite unions or intersections and that we would like to be able to calculate probabilities across such infinite set operations. However, the next example demonstrates that we cannot do this arbitrarily:

Example 3.2.3. *Let our universal set be $[0,1]$ and our algebra be its power set $\mathcal{P}([0,1])$. We wish to define a uniform distribution on this set, that is, a distribution such that:*

$$\mathbb{P}([a,b]) = b - a$$

for any real numbers $a < b$. Note that for any individual number $x \in (a,b)$ and any $\epsilon > 0$:

$$\mathbb{P}(\{x\}) \leq \mathbb{P}\Big(\Big[x - \frac{\epsilon}{2}, x + \frac{\epsilon}{2}\Big]\Big) = x + \frac{\epsilon}{2} - \Big(x - \frac{\epsilon}{2}\Big) = \epsilon$$

Since $\epsilon > 0$ was arbitrary, it must be that $\mathbb{P}(\{x\}) = 0$. Suppose we extend the additivity axiom, Axiom 2 of Definition 2.2.6 to infinite unions:

2'. **Infinite Additivity:** *If some infinite set of sets S_α for $\alpha \in T$ are disjoint, then $\mathbb{P}(\bigcup_\alpha S_\alpha) = \sum_\alpha \mathbb{P}(S_\alpha)$*

In this case:

$$\mathbb{P}([0,1]) = \mathbb{P}\left(\bigcup_{x \in [0,1]} \{x\} \right) = \sum_{x \in [0,1]} \mathbb{P}(\{x\}) = \sum_{x \in [0,1]} 0 = 0$$

Hence, $\mathbb{P}$ is not a probability measure since it doens't satisfy Axiom 1 of Definition 2.2.6

Based on the last example, while we would like to be able to calculate probabilities of infinite unions, we must be careful in this endeavour. Cantor's distinction among infinite sets, which we present next, will help here.

3.3 Cantor Diagonalization

Cantor's diagonalization and cardinal numbers are the foundation of much of modern mathematics and particularly modern probability theory. Cantor demonstrated that there are different sizes, or, to use the mathematical terminology, cardinalities of infinite sets as we now present.

In order to determine if sets are of different cardinalities, we first consider how we first learned to count. Counting involves setting up a 1-1 correspondence between an initial segment of the natural numbers $\mathbb{N}$ and some set of objects. For example, consider the set $C \times C = \{HH, HT, TH, TT\}$. We could make a 1-1 correspondence between this set and the first 4 natural numbers as follows:

$$\begin{array}{cc} 1 & HH \\ 2 & HT \\ 3 & TH \\ 4 & TT \end{array}$$

Hence, the set has 4 elements. We could form different 1-1 correspondences with initial segments of this set but all of them will reveal the set to have 4 elements.

Now we consider counting infinite sets. Cantor called a (finite or infinite) set S **countable** if there is a 1-1 correspondence between the natural numbers, $\mathbb{N}$ and the set S. Hence, the natural numbers are countable because the identity function is a 1-1 correspondence. We now present further examples of countable sets.

1. **The even numbers:** the set of even numbers are countable as demonstrated by the following 1-1 correspondence:

$$\begin{array}{cc} 1 & 2 \\ 2 & 4 \\ 3 & 6 \\ \ldots & \end{array}$$

   that is, the 1-1 correspondence $f(n) = 2n$. A similar argument shows that the odd numbers are countable. In fact, every subset of a countable set is countable. The prime numbers are also countable.

2. **The integers:** Let $\mathbb{Z}$ be the set of integers, i.e., $\mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$. The integers are countable as demonstrated by the following 1-1 correspondence:

$$\begin{array}{rr} 1 & 0 \\ 2 & 1 \\ 3 & -1 \\ 4 & 2 \\ 5 & -2 \\ \dots & \dots \end{array}$$

that is, the 1-1 correspondence $f(n) = (-1)^n \left\lfloor \frac{n}{2} \right\rfloor$ where $\lfloor x \rfloor$ denotes the largest integer which is less than or equal to x.

3. **Correspondences with countable sets:** Suppose a set A, demonstrated by a 1-1 correspondence $f : \mathbb{N} \to A$. Suppose also that there is 1-1 correspondence $g : A \to B$ between A and another set B. Then B is also countable, since we have the 1-1 correspondence given by $h(n) = g(f(n))$.

4. **Cartesian products:** For any two countable sets A and B, the Cartesian product $A \times B$ is countable. To see this, let $A = \{a_1, a_2, a_3, \dots\}$ and $B = \{b_1, b_2, b_3, \dots\}$. In order to form a 1-1 correspondence, we first count all the pairs which have indices summing to 2 which consists of only (a_1, b_1). We then count all pairs with indices summing to 3 which consists $(a_1, b_2$ and (a_2, b_1). We then count all pairs with indices summing to 4, specifically (a_1, b_3), (a_2, b_2) and (a_3, b_1), and so on. This 1-1 correspondence is demonstrated by the following diagram:

$$\begin{array}{lllll} (a_1, b_1) & & (a_1, b_2) & & (a_1, b_3) \quad \dots \\ & \swarrow & & \swarrow & \\ (a_2, b_1) & & (a_2, b_2) & & \dots \\ & \swarrow & & & \\ (a_3, b_1) & & \dots & & \\ \dots & & & & \end{array}$$

In this case, explicit forms for the indices of a and b can be found but it does not provide much insight.

5. **The rational numbers:** The rational numbers, $\mathbb{Q}$, consist of fractions of integers:

$$Q = \left\{ \frac{n}{m} : n, m \in \mathbb{Z} \right\}$$

The rational numbers are countable because there is a 1-1 correspondence between them and $\mathbb{Z} \times \mathbb{Z}$.

While it might seem that all infinite sets are countable, we now demonstrate that the real numbers, $\mathbb{R}$ are not countable. Real numbers have infinite precision.

We show this by demonstrating that there is no 1-1 correspondence between the interval $[0,1]$ and the natural numbers. The proof is by contradiction. For this form of proof, we first assume that there is such a 1-1 correspondence and demonstrate that this implies a contradiction. This means the assumption must be false.

Suppose there is a 1-1 correspondence between $[0,1]$ and $\mathbb{N}$ Let the real number in correspondence with 1 be r_1, in correspondence with 2 be r_2, etc. Each of these numbers can be expanded in an infinite decimal. For the number r_i, we denote the *jth* decimal digit as $r_{i,j}$, that is, r_i corresponds with the number whose decimal expansion is $0.r_{i,1}r_{i,2}r_{i,3}\ldots$. We will construct a number q such that $q \neq r_i$ for all i. We do this by choosing the number to be different from r_i in the *ith* digit. This could be done, for example, by transforming the digits as follows:

$$f(n) = \begin{cases} n+1 & \text{if } n \in \{0,1,2,3,4,5,6,7,8\} \\ 0 & \text{if } n = 9 \end{cases}$$

We choose $q_i = f(r_{i,i})$. Note that $q_i \neq r_{i,i}$ since $f(n) \neq n$ for all n. Hence, $q \in r_i$ for all i. This argument can be visualized with the following diagram:

$\boldsymbol{q}$**:**	0	.	$f(r_{1,1})$	$f(r_{2,2})$	$f(r_{3,3})$	$\ldots$
$\boldsymbol{r_1}$**:**	0	.	$\boxed{r_{1,1}}$	$r_{1,2}$	$r_{1,3}$	$\ldots$
$\boldsymbol{r_2}$**:**	0	.	$r_{2,1}$	$\boxed{r_{2,2}}$	$r_{2,3}$	$\ldots$
$\boldsymbol{r_3}$**:**	0	.	$r_{3,1}$	$r_{3,2}$	$\boxed{r_{3,3}}$	$\ldots$
$\ldots$				$\ldots$		

The boxes in the above show which digit the number q differs from each of the numbers in the 1-1 correspondence. We started out with the assumption that there is a 1-1 correspondence and demonstrated that there is a number which is not in the correspondence. This is a contradiction so that it must be no such correspondence exists.

Note that if $\mathbb{R}$ were countable, then $[0,1]$ would also be countable since subsets of countable sets are countable. Since we have shown that $[0,1]$ is not countable, this cannot be the case.

Cantor's work was not well accepted by the mathematical community of his time. The well-known mathematician Kronecker called Cantor a "scientific charlatan" and a "corrupter of youth". Cantor was sensitive to these criticims, suffered through bouts of depression and spent time in hospitals in his later years. Today, Cantor's theory is widely accepted by mathematicians.

We take a brief digression to discuss the effect of Cantor's work on the foundations of mathematics. Based on Cantor's theory of sets, the logician Gottlob Frege developed a formal system in which all of mathematics could be developed. The logician Bertrand Russell wrote to Frege that he had discovered the following paradox in Frege's work.

> **Russell's paradox:** Consider the set of all sets which don't contain themselves. Does it contain itself?

Notice that this is the same style of argument, called diagonalization, that Cantor used in proving that the real numbers are uncountable. In this simple statement, Russell invalidated Frege's life's work. The most common resolution of Russell's paradox is to carefully define what a set is.

Russell went on to spend the better part of a decade, with Alfred North Whitehead, trying to fix the problem with Frege's work, ultimately publishing *Principia Mathematica*. This 3 volume work is so detailed that it requires hundreds of pages to prove that $1 + 1 = 2$. A simpler formal system for set theory, called Zermelo Fraenkel, is in current use today and most mathematicians consider it to be a basis for all of mathematics. However, in a brilliant PhD thesis in 1931, Kurt Gödel proved that, in any such system, one can create a sentence corresponding to the phrase "This sentence cannot be proven". After Russell invalidated Frege's work using a diagonalization argument, Gödel invalidated Russell's subsequent efforts and demonstrated that no formal system can prove everything also using diagonalization.

After Godël, Turing developed used a diagonalization argument to show that there are problems that can't be solved by computers, in 1937, before such machines existed in any meaningful way. To do this, he had to develop the notion of a universal machine, a machine that plausibly can do any mechanical computation that any other machine can do. While no one person can be said to have invented the computer, though this and other work, Turing made significant advances in the history of that development. While diagonalization arguments demonstrated that mathematics could not be put on a solid formal basis, they also contributed to the early development of computers.

3.4 Modern Probability Theory

As discussed in the Section 3.2, we will want to determine the probability of infinite unions of sets. While Example 3.2.3 shows that we can't do this for arbitrary infinite unions of sets in a meaningful way, it turns out that we can do this for countable unions. Indeed this is the approach used in modern probability theory.

In order to guarantee the existence of countable unions in an algebra, we define a restricted algebra which contains countable unions:

Definition 3.4.1. *A* **σ-algebra** *(or* **σ-field***) is an algebra such that if* $F_1, F_2, \ldots \in \mathcal{F}$*, then* $\bigcup_{i=1}^{\infty} F_i \in \mathcal{F}$*.*

Notice that, if we were defining σ-algebras without the definition of algebra, we could just replace Item 2 with

2". If $F_1, F_2, \ldots \in \mathcal{F}$, then $\bigcup_{i=1}^{\infty} F_i \in \mathcal{F}$.

Similarly, we define probability measures to be additive over disjoint countable unions:

Definition 3.4.2. *A* **probability measure** *is a finite probability measure, on a σ-algebra $\mathcal{F}$, is a function $\mathbb{P} : \mathcal{F} \to [0,1]$ such that:*

1. **Normalization:** $\mathbb{P}(\Omega) = 1$
2. **Countable additivity:** *if $F_1, F_2, \ldots \in \mathcal{F}$ are disjoint, then $\mathbb{P}(\bigcup_{i=1}^{\infty} F_i) = \sum_{i=1}^{\infty} \mathbb{P}(F_i)$*

Note that the sum of the probabilities of a countable number of sets in Item 2 above is just an infinite series, or limit of finite sums, as taught in calculus classes. Finally, we define a probability space:

Definition 3.4.3. *A* **probability space** *is a triple of a $(\Omega, \mathcal{F}, \mathbb{P})$ where Ω is the universal set, $\mathcal{F}$ is a σ-algebra and $\mathbb{P}$ is a probability measure.*

We now give some examples of probability measures.

Example 3.4.1. *As in Example 2.2.6, we define a point mass function to be a function $p : A \to [0,1]$ such that:*

$$\sum_s p(s) = 1$$

Just as in Example 2.2.6, we define the probability measure:

$$\mathbb{P}(S) = \sum_{s \in S} p(s)$$

This can be shown to be a probability measure. Note that in this case, countable additivity automatically holds as Equation (3.2) is interpreted as an infinite series when appropriate. As such, counterintuitive behavior as in Example 3.2.2 does not occur.

Example 3.4.2. *Let us start with a continuous function $f : (a,b) \to [0,\infty)$, for some $a, b \in \mathbb{R} \cup \{-\infty, \infty\}$, such that:*

$$\int_a^b f(x)\,dx = 1$$

Such a function is called a **probability density**. *Define a set function $\mathbb{P}$ such that:*

$$\mathbb{P}((c,d)) = \int_c^d f(x)\,dx$$

We can extend this set function to a probability measure by using the rules of Definition 3.4.3. It can be shown that this yields a consistent probability measure. For example, consider the set $[\tilde{c,d}]$. *We could define the probability of this set in two different ways:*

$$\begin{aligned}
\mathbb{P}\left(\widetilde{(c,d)}\right) &= \mathbb{P}((a,c] \cup [d,b)) = \mathbb{P}((a,c]) + \mathbb{P}([d,b)) \\
&= \int_a^c f(x)\,dx + \int_d^b f(x)\,dx \\
\mathbb{P}\left(\widetilde{(c,d)}\right) &= 1 - \mathbb{P}((c,d)) = 1 - \int_c^d f(x)\,dx = \int_a^b f(x)\,dx - \int_c^d f(x)\,dx \\
&= \int_a^c f(x)\,dx + \int_d^b f(x)\,dx
\end{aligned}$$

Both of these yield the same result. This probability measure is defined on a special σ*-algebra called the Borel* σ*-algebra. For any set of sets, there is always a smallest* σ*-algebra containing that set of sets. The Borel* σ*-algebra, which we denote by* $\mathcal{B}$*, is the smallest* σ*-algebra such that includes all intervals of the form* (c,d).

Examples of probability distributions defined via probability density functions include:

1. **The uniform probability distribution:**

$$f_{a,b}(x) = \begin{cases} \frac{1}{b-a} & \text{if } x \in (a,b) \\ 0 & \text{otherwise} \end{cases}$$

where $a < b \in \mathbb{R}$ *are parameters of the distribution. This example demonstrates how countable additivity allows us to avoid the issues of Example 3.2.3.*

2. **The normal distribution:**

$$f_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

where $\mu \in \mathbb{R}$ *and* $\sigma \in (0,\infty)$ *are parameters of the distribution.*

Example 3.4.3. *As in Example 2.2.9, the mixture of any two probability measures is a probability measure. However, we can now talk about mixtures of a countable number of probability distributions. Let* $p : \mathbb{N} \to [0,1]$ *be a point mass function and let* $\mathbb{P}_1, \mathbb{P}_2, \ldots$ *be probability distributions. The set function* $\mathbb{Q}$ *defined by the following:*

$$\mathbb{Q}(S) = \sum_{i \in \mathbb{N}} p(n)\,\mathbb{P}_i(S)$$

is a probability distribution.

Example 3.4.4. *Consider a function, $F : \mathbb{R} \to [0,1]$ with the following properties:*

1. **Monotonically increasing:** *if $x \leq y$ then $F(x) \leq F(y)$.*
2. **Right continuity:** $\lim_{x \downarrow a} F(x) = F(a)$
3. **Limits:**

$$\begin{aligned} \lim_{x \to -\infty} F(x) &= 0 \\ \lim_{x \to \infty} F(x) &= 1 \end{aligned}$$

If one then defines a set function by:

$$\mathbb{P}((c,d)) = F(d) - F(c)$$

and extends it by using the rules of Definition 3.4.3, the result is a probablity measure. This function is called a **cumulative distribution function**.

In fact, every probability measure on any subset of $\mathbb{R}$ can be defined in this way. For a point mass function $p : S \to [0,1]$ defined on a set $S \subseteq \mathbb{R}$, the cumulative distribution function is given by:

$$F(x) = \sum_{s \leq x} p(s)$$

For a probability density function $f : (a,b) \to [0,\infty)$, the cumulative distribution function is given by:

$$F(x) = \int_a^x f(x)\, dx$$

Let $\mathbb{Q}$ be a mixture distribution with point mass function p and such that the underlying distributions $\mathbb{P}_1, \mathbb{P}_2, \ldots$ have cumulative distribution functions $F_1, F_2, \ldots$. The cumulative distribution function of $\mathbb{Q}$ is given by:

$$F(x) = \sum_{i \in \mathbb{N}} p(i)\, F_i(x)$$

Example 3.4.5. *Classical probability theory discusses both point mass functions and densities. Mixtures of these two types of distributions can extend those types. However, there are some distributions on $\mathbb{R}$, referred to as* **singular** *which do not fit into any of these categories. As mentioned, all distributions on subsets $\mathbb{R}$ have cumulative distribution functions. Unlike classical probability theory, modern probability theory encapsulates all possible distributions which fit into Definition 3.4.3.*

Example 3.4.6. *We now discuss of infinite Cartesian products introduced in Section 2.3.1. In particular, suppose we have an infinite sequence of quantities which we wish to model as random such as stock prices. In particular, we consider an infinite Cartesian product of real number,* $\mathbb{R}^\infty$.

Let $\mathcal{B}$ *be the Borel* σ*-algebra. Consider the subsets of* $\mathbb{R}^\infty$ *of the form* $S_1 \times S_2 \times \ldots \times S_n \times \mathbb{R}^\infty$ *for some* $n \in \mathbb{N}$ *and* $S_1, S_2, \ldots, S_n \in \mathcal{B}$*. These are infinite Cartesian products where the first* n *elements are restricted for some* n *and the remaining elements are unrestricted and are called* **cylinder sets**. *The smallest* σ*-algebra containing all cylinder sets is called the cylinder* σ*-algebra. As an example, if we assume the elements of the infinite Cartesian product are IID with some distribution* $\mathbb{P}$*, we can define a probability measure* $\mathbb{Q}$ *on the cylinder sets as follows:*

$$\mathbb{Q}(S_1 \times S_2 \times \ldots \times S_n \times \mathbb{R}^\infty) = \prod_{i=1}^{n} \mathbb{P}(S_i)$$

The cylinder σ*-algebra allows us to define probability distributions on infinte sequences of random quantities. These quantities can all be dependent upon each other in a way which would be difficult to model with joint distributions presented in classical probability theory.*

3.5 Random Variables

We can define random variables on full probability spaces analogously to how we defined them for finite probability spaces in Definition 2.2.9:

Definition 3.5.1. *A* **random variable** *on a probability space* $(\Omega, \mathcal{F}, \mathbb{P})$ *is a functon* $X : \Omega \to \mathbb{R}$ *such that the inverse image of every interval* $X^{-1}([a, b]) \in \mathcal{F}$ *for all* $a, b \in \mathbb{R}$.

Some further properties of random variables include:

1. If $X_1, X_2, \ldots, X_n$ are random variables on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and $f : \mathbb{R}^n \to \mathbb{R}$ is a continuous function, then $f(X_1, X_2, \ldots, X_n)$ is a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$. For example, for random variables X_1 and X_2, both $X_1 + X_2$ and $X_1 X_2$ are random variables.
2. Let $X_1, X_2, \ldots$ be an infinite sequence of random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. If $\lim_{n\to\infty} X_n(\omega) = X^*(\omega)$ for some function $X^* : \Omega \to \mathbb{R}$, then X^* is a random variable.

We can now demonstrate how countable additivity allows us to avoid the problem of Example 3.2.1.

Example 3.5.1. *We detail the setup of Example 3.2.1. Let* $X_1, X_2, \ldots$ *be random variables on a probability space* $(\Omega, \mathcal{F}, \mathbb{P})$*. In Example 3.2.1, we demonstrated that the set of* $\omega \in \Omega$ *such that* $\lim_{n\to\infty} \frac{S_n}{n}$ *exists and equals* $\frac{1}{2}$*, where* $S_n = \sum_{i=1}^{n} X_i$*, is given by the following set:*

$$\bigcap_{\epsilon>0} \bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} \left\{ \frac{S_n}{n} \in \left[\frac{1}{2} - \epsilon, \frac{1}{2} + \epsilon \right] \right\} \tag{3.2}$$

To ease notation, we define $A_{n,\epsilon}$ as:

$$A_{n,\epsilon} = \left\{ \frac{S_n}{n} \in \left[\frac{1}{2} - \epsilon, \frac{1}{2} + \epsilon \right] \right\} \tag{3.3}$$

Given Item 1 above, we can see that $\frac{S_n}{n}$ is a random variable. Hence, $A_{n,\epsilon} \in \mathcal{F}$ for any n and ϵ. Furthermore, intersections of a countable number of sets in $\mathcal{F}$ are also in $\mathcal{F}$ since $\mathcal{F}$ is a σ-algebra. Therefore, $\bigcap_{n=N}^{\infty} A_{n,\epsilon} \in \mathcal{F}$. Unions of a countable number of these sets are also in $\mathcal{F}$ so that:

$$\bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} A_{n,\epsilon} \in \mathcal{F}$$

However, now we run into an issue. The outer most intersections in Equation (3.2) is over all $\epsilon \in (0, \infty)$. This is an uncountable set.

To resolve this, define:

$$B_\epsilon = \bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} A_{n,\epsilon}$$

Note that for $\epsilon_1 < \epsilon_2$, we have that $A_{n,\epsilon_1} \subseteq A_{n,\epsilon_2}$ since the interval in Equation (3.3) becomes larger. Also note that union and intersection are monotonic, that is, if $A_{n,\epsilon_1} \subseteq A_{n,\epsilon_2}$, then we have that:

$$\begin{aligned} \bigcap_{n=N}^{\infty} A_{n,\epsilon_1} &\subseteq \bigcap_{n=N}^{\infty} A_{n,\epsilon_2} \\ B_{\epsilon_1} = \bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} A_{n,\epsilon_1} &\subseteq \bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} A_{n,\epsilon_2} = B_{\epsilon_2} \end{aligned} \tag{3.4}$$

Note that for all $\epsilon \in (0, \infty)$, there is a $m(\epsilon) \in \mathbb{N}$ such that $\frac{1}{m(\epsilon)} < \epsilon$. Using this and Equation (3.4), we see that:

$$B_{\frac{1}{m(\epsilon)}} \subseteq B_\epsilon$$

Therefore, again using monotonicity of intersection:

$$\bigcap_{m\in\mathbb{N}} B_{\frac{1}{m}} \subseteq \bigcap_{\epsilon>0} B_\epsilon \tag{3.5}$$

However, since the left hand side of Equation (3.5) is an intersection over a smaller number of sets than the right hand side, we also have that:

$$\bigcap_{\epsilon>0} B_\epsilon \subseteq \bigcap_{m\in\mathbb{N}} B_{\frac{1}{m}}$$

so that:

$$\bigcap_{m\in\mathbb{N}} B_{\frac{1}{m}} = \bigcap_{\epsilon>0} B_\epsilon$$

The latter is a countable intersection of sets in $\mathcal{F}$ and so is in $\mathcal{F}$.

3.6 Lebesgue Integration

We wish to define the expected value of a random variable in a way that applies to as many random variables as possible. In his disseratation of 1902, Henri Lebesgue demonstrated that by subdividing the range (output) of a function rather than the domain (input), we arrive at a more general approach to integration. This is particularly suited to probability theory since we wish to take expected values of random variables whose range is $\mathbb{R}$ but whose domain is an arbitrary set Ω.

In order to define the expected value of a positive random variable, we must first define the supremum, which is a generalization of the concept of a maximum. The **supremum** of a set of real numbers $S \subseteq \mathbb{R}$, written $\sup S$ is the least upper bound of S. The supremum can be positive or negative infinity if the set is unbounded or empty respectively. Every set of real numbers has a supremum. Some examples of suprema are:

1. $\sup\left\{x : x < \sqrt{2}\right\} = \sqrt{2}$
2. $\sup\left\{x : x \leq \sqrt{2}\right\} = \sqrt{2}$
3. $\sup(-\infty, r) \cap \mathbb{Q} = r$
4. $\sup(-\infty, r) \cap \mathbb{N} = \lfloor r \rfloor$
5. $\sup \mathbb{R} = \infty$
6. $\sup \varnothing = -\infty$

If $\sup S \in S$, then we say that the supremum is **achieved** and we call it the **maximum** of the set. For example, the supremum is not achieved in Item 1 above but is achieved in Item 2.

A random variable which takes on only a finite number of values is called a **simple** random variable. We will first define the Lebesgue integral for simple random variables, then positive and then general random variables:

1. **Simple random variables:** let X be a simple random variable taking on values $x_1, x_2, \ldots, x_n$. As in Definition 2.2.12, we define the expected value of X as follows:

$$E[X] = \sum_{i=1}^{n} x_i \mathbb{P}(X = x_i)$$

2. **Positive random variables:** Now if X is a non-negative random variable, we define the expected value of X as:

$$E[X] = \sup \{E[Y] : Y \text{ is simple and } Y \leq X\}$$

Note that the expected value can be infinite, as we have seen with the St. Petersburg paradox.

3. **General random variables:** Finally, let X be a general random variable. We can divide X into positive and negative parts, written X^+ and X^-, as follows:

$$\begin{aligned} X^+ &= X \mathbb{1}_{\{X \geq 0\}} \\ X^- &= -X \mathbb{1}_{\{X \leq 0\}} \end{aligned}$$

Both of these are non-negative random variables. We define the expected value of X as:

$$E[X] = E\left[X^+\right] - E\left[X^-\right]$$

Note that, if both of the expectations above are infinite, the expected value of X is undefined.

Note that the properties of expectation given in Section 2.2.6 for finite random variables, as well as Jensen's inequality, Theorem 2.3.1, also hold for general random variables.

Before presenting some examples of random variables, we introduce the distribution of a random variable. The **distribution** of a random variable X on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is the probability measure $\mathbb{P}_X$ defined by:

$$\mathbb{P}_X(S) = \mathbb{P}(X \in S)$$

Note that $\mathbb{Q}$ is a probability measure with respect to the Borel σ-algebra $\mathcal{B}$ on the real nubmers $\mathbb{R}$.

We now introduce some examples of the calculation of expectations.

Example 3.6.1. *If $\mathbb{P}_X$ has a density function $f_X : \mathbb{R} \to [0, \infty)$, then $E[X]$ is given by:*

$$E[X] = \int_{-\infty}^{\infty} x f_X(x)\, dx \tag{3.6}$$

If we interpret the integral above as a Lebesgue integral described previously, it will be well defined for any function which is the limit of a continuous function, which can be highly discontinuous. Such functions can't be integrated using the Riemann integral presented in ordinary calculus.

We can break the integral in Equation (3.6) into integrals over the positive and negative values of X:

$$\int_{-\infty}^{\infty} x f_X(x)\, dx = \int_{-\infty}^{0} x f_X(x)\, dx + \int_{0}^{\infty} x f_X(x)\, dx \tag{3.7}$$

We can classify the possible non-finite values of the integral as follows:

1. *Both integrals on the right hand side of Equation (3.7) are finite. In this case, the expected value is the sum of these values. For example, this is the case for the normal distribution.*

2. *The integral over negative values is finite and the integral over positive values equals ∞. In this case, we say that the expected value is ∞. This is the case for the Pareto distribution with $\alpha \leq 1$:*

$$f_X(x) = \begin{cases} \frac{\alpha x_m^\alpha}{x^{\alpha+1}} & \text{if } x > x_m \\ 0 & \text{otherwise} \end{cases}$$

3. *The integral over negative values equals $-\infty$ and the integral over positive values is finite. In this case, we say that the expected value is $-\infty$.*

4. *The integral over negative values equals $-\infty$ and the integral over positive values equal ∞. In this case, the integral does not exist. This is the case for the Cauchy distribution:*

$$f_X(x) = \frac{1}{\pi\,(1+x^2)}$$

If $g : \mathbb{R} \to \mathbb{R}$ is a function such that $g(X)$ is a random variable then:

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)\, f_X(x)\, dx$$

Example 3.6.2. *Suppose a random variable X on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ takes on a countable number of values $\{x_1, x_2, \ldots\}$ and let $p_i = \mathbb{P}(X = x_i)$, then:*

$$E[X] = \sum_{i=1}^{\infty} p_i x_i$$

where the convergence of the series is as in ordinary calculus. Note that the expected value might be $\pm\infty$ if the series converges to one of those values. The series might also diverge in which case the expectation doesn't exist.

Similar to the last example, for any function $g : \{x_1, x_2, \ldots\} \to \mathbb{R}$, we have:

$$E[g(X)] = \sum_{i=1}^{\infty} p_i g(x_i)$$

3.7 The Law of Large Numbers

We now present the first result which will help us work out the limits of sums of random variables, the law of large numbers. There are many variants of this result and we first present the strong law of large numbers:

Theorem 3.7.1. Strong Law of Large Numbers: *Let $X_1, X_2, \ldots$ be IID random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. If $E[X_i]$ exists and is finite, then:*

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{S_n}{n} = E[X_1]\right) = 1 \tag{3.8}$$

where $S_n = \sum_{i=1}^{n} X_i$.

This result says that the long-term mean of IID random variables converges to their average, as long as their average exists. An early version of this result was demonstrated by Jacob Bernoulli, another one of the Bernoulli family. The strong law of large numbers implies, for example, that if there is a game you play repeatedly, your average winnings will converge to the expected value of the game. This is what lead the Bernoulli's to think about games where this might not be the best strategy, such as the St. Petersburg paradox.

We make some comments on the strong law of large numbers:

1. We used $E[X_1]$ in Equation (3.8) but, since the variables are IID, we have that $E[X_i] = E[X_j]$ for all i and j so we could have used $E[X_i]$ for any i.

2. The strong law of large numbers says that a limit exists and converges to a certain value "with probability 1". This limit might not exist or it might converge to some other value. For example, in tossing an infinite sequence of fair coins, it is possible that we obtain all heads. This happens with probability 0 but this is not surprising since any infinite sequence occurs with probability 0. What is more surprising is that there are an uncountable number of sequences for which $\frac{S_n}{n}$ doesn't converge to $\frac{1}{2}$. In fact, $\frac{S_n}{n}$ could converge to any number in $[0, 1]$ which is an uncountable set. From countable additivity, a countable union of disjoint sets of probability 0 has probability 0. However, we cannot deduce the strong law of large numbers from the fact that each individual sequence of tosses occurs with probability 0 since the set which converges to other values is uncountable. We will see other "with probability 1" results later.

3. Also, $E[X_1]$ need not exist in which case the strong law of large numbers does not hold. This is the case, for example. if the random variables have a Cauchy distribution.

We now show how the strong law of large numbers helps us choose a portfolio for long-term growth. Start with Equation (3.1):

$$\begin{aligned}
\ln(V_{n+1}) &= \ln(V_1) + \sum_{t=1}^{n} \ln\left(\sum_i w_{t,i} \frac{X_{t+1,i}}{X_{t,i}}\right) \\
&= \ln(V_1) + \sum_{t=1}^{n} \ln\left(\sum_i w_{t,i} \left(1 + R_{t+1,i}\right)\right) \\
&= \ln(V_1) + \sum_{t=1}^{n} \ln\left(1 + w_t^T R_{t+1}\right) \qquad (3.9)
\end{aligned}$$

where $R_{t+1,i}$ is the return of the *ith* stock from period t to period $t+1$. We've also used the w_t and R_t for the vector of all stock investment fractions and returns, respectively, at time t. The notation w_t^T denotes the transpose of the vector w_t so that $w_t^T R_t$ is the inner product of the vectors w_t and R_t. Finally, we used the fact that $\sum_i w_{t,i} = 1$ since we assume all wealth is invested.

Stock prices tend to grow over time as companies and the economy in general grows. Hence, stock prices themselves are not well-modeled as IID. However stock returns have a more stationary character and may be modeled as IID over time, that is, the vectors $R_1, R_2, \ldots$ of all stock returns are independent. Note that individual stock returns tend to be highly correlated for a particular time, so that $R_{t,i}$ and $R_{t,j}$ are not well modeled as independent and we do not assume that they are. For now, we make the assumption that stock returns are IID over time. We'll later investigate what happens when we relax this assumption.

At present, we only consider portfolios which are constant over time, that is, $w_{t,i} = w'_i$ for all t for some w'. Recall that $w_{t,i}$ is the fraction of our wealth invested in the *ith* stock at time t. Since each stock will move up or down by a different amount, a constant w_t will mean that we will need to trade small amounts to rebalance our portfolio each period. As an example, suppose w' has us placing $\frac{1}{2}$ of our wealth in stock 1 and $\frac{1}{2}$ of our wealth in stock 2. If stock 1 goes up 20% is period 1 and stock and stock 2 is flat then we will have $\frac{1}{2}(1+0.2) = 0.6$ fraction of our original wealth in stock 1 and 0.5 fraction of our original wealth in stock 2. We would have to sell 0.05 worth of stock 1 and buy 0.05 worth of stock 2 in order to achieve the same portfolio for the second period. Portfolios of this kind are often called constant rebalanced portfolios.

Since the random variables R_t are independent for different t, then the random variables $\sum_i w'_i (1 + R_{t+1,i})$ are independent for different t. We calculate the average log growth of our portfolio over time:

$$\begin{aligned}\lim_{n\to\infty} \frac{1}{n} \ln\left(\frac{V_n}{V_1}\right) &= \lim_{n\to\infty} \frac{\sum_{t=1}^{n} \ln\left(1 + (w')^T R_{t+1}\right)}{n} \\ &= E\left[\ln\left(1 + (w')^T R_1\right)\right]\end{aligned}$$

with probability 1 by the strong law of large numbers. Hence:

$$\lim_{n\to\infty} \frac{1}{n} \ln\left(\frac{V_n}{V_1}\right) = E\left[\ln\left(1 + (w')^T R_1\right)\right] \leq \sup_{w'} E\left[\ln\left(1 + (w')^T R_1\right)\right]$$

where $\sup_{w'}$ denotes the supremum over all constant rebalanced porfolios w'. Hence, any portfolio w^* such that:

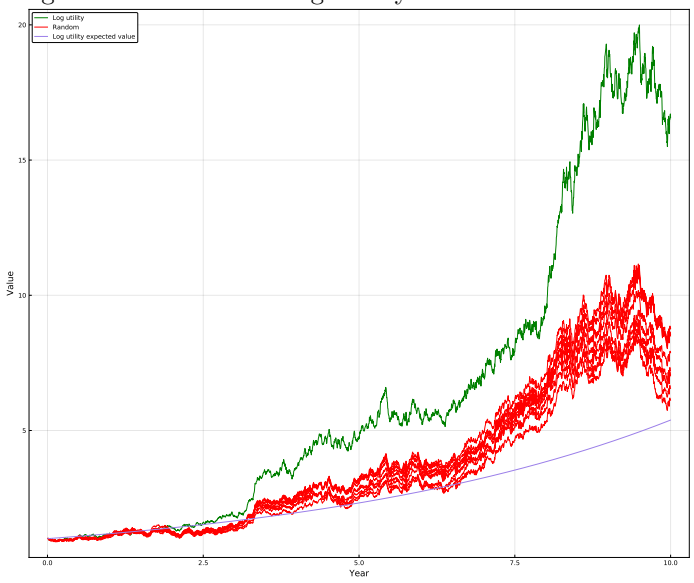
$$E\left[\ln\left(1 + (w^*)^T R_1\right)\right] = \sup_{w'} E\left[\ln\left(1 + (w')^T R_1\right)\right] \tag{3.10}$$

is called a **growth optimal portfolio** and will grow with at higher average rate than any other portfolio with probability 1. While it is not implied by the law of large numbers, it can also be shown that this portfolio grows faster than any time varying portfolio. Such portfolios will always exist because Equation (3.10) is concave in w' and the set of possible w' is a compact (closed under limits and bounded) set. Note that the reasoning here is very similar to the reasoning that expected value is the fair value of a gambling game in which the same amount is bet every time but incorporates compound growth when winnings are reinvested in the game.

We now explore how much faster the log optimal portfolio grows than a competing portfolio. The implication of the strong law of large numbers is that, with probability 1:

$$\lim_{n\to\infty} \frac{\ln(V_{n+1})}{n} = E\left[\ln\left(1 + (w')^T R_1\right)\right]$$

Figure 3.1: Growth of Log Utility and Random Portfolios



From the definition of the limit, this means that for any $\epsilon > 0$, there is an N such that for all $n \geq N$:

$$\left| \frac{\ln(V_{n+1})}{n} - E\Big[\ln\Big(1 + (w')^T R_1\Big)\Big] \right| < \epsilon$$

Hence:

$$V_{n+1} = \exp\Big(nE\Big[\ln\Big(1 + (w')^T R_1\Big)\Big] \pm n\epsilon\Big)$$

The growth optimal portfolio will grow exponentially faster than any portfolio with lower expected log.

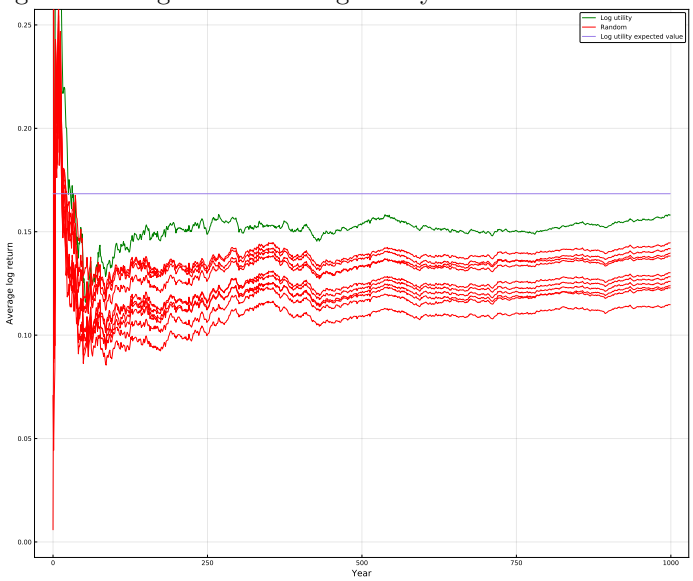
In order to develop an intuition for the log utility portfolio, we present some simulations. Figure 3.1 shows the growth of the maximum log utility portfolio in comparison with random portfolios. In this case, the market consists randomly generated log normal returns using the sample means and covariances from the top 10 stocks from 2000 through 2019. Also, shown in the figure is the growth in wealth expected from the strong law of large numbers.

We make some observations on this figure:

1. The growth optimal portfolio performs significantly better than the random portfolios.

2. The growth optimal portfolio is far from the expoentital of the expected log predicted by the law of large numbers.

Figure 3.2: Log Growth of Log Utility and Random Portfolios



3. Even the random portfolios perform better than the law of large numbers prediction.

Figure 3.2 shows the quantity that the strong law of large numbers is predicting, the average log growth, and over a longer time frame. The growth optimal portfolio does seem to be converging to the expected log growth rate but it is taking a very long time. Even after 1,000 years, the average growth rate is 15.8% compared with a predict 16.8%. In the next section, we explore how far off the law of large numbers estimate tends to be by discussing the next major limit theorem of probability theory: the central limit theorem.

3.8 The Central Limit Theorem

As can be seen from Figure 3.2, the deviation of the average from the expected value, which it is gradually decreasing to 0, seems to be random. In the early years, the average was higher and subsequently it was lower than the expected value, sometimes by bigger amounts and sometimes smaller. The central limit theorem tells how this randomness is distributed:

Theorem 3.8.1. Central Limit Theorem: *For IID random variables* $X_1, X_2, \ldots$ *such that* $\mu = E[X_1]$ *and* $\sigma = \sqrt{E\left[(X_1 - \mu)^2\right]}$ *exist and are finite:*

$$\lim_{n\to\infty} \mathbb{P}\left(\sqrt{n}\frac{\frac{S_n}{n} - \mu}{\sigma} \leq x\right) = \Phi(x) \tag{3.11}$$

where Φ *is the cumulative distribution function of the normal distribution:*

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{x^2}{2}\right) dx$$

We make some comments on the central limit theorem:

1. While the law of large numbers requires that the mean, $\mu = E[X_1]$, exists and is finite, the central limit theorem also requires that the variance, $\sigma^2 = E\left[(X_1 - \mu)^2\right]$, exists (which is always the case) and is finite. Hence, there are cases when the law of large numbers holds but the central limit theorem does not.

2. As mentioned previously, the deviation of the sample average from the mean goes to 0. Hence, in order to find the distribution, we need to scale this deviation as n becomes large. This is done in Equation (3.11) by multiplying by $\sqrt{n}$. To see that this is the correct scale, consider the variance of the resulting deviation:

$$\begin{aligned}
E\left[\left(\sqrt{n}\left(\frac{S_n}{n} - \mu\right)\right)^2\right] &= E\left[\left(\sqrt{n}\frac{\sum_{i=1}^n (X_i - \mu)}{n}\right)^2\right] \\
&= E\left[\left(\frac{\sum_{i=1}^n (X_i - \mu)}{\sqrt{n}}\right)^2\right] = E\left[\frac{\sum_{i=1}^n \sum_{j=1}^n (X_i - \mu)(X_j - \mu)}{n}\right] \\
&= \frac{\sum_{i=1}^n E\left[(X_i - \mu)^2\right] + 2\sum_{i=1}^n \sum_{j=i+1}^n E[(X_i - \mu)(X_j - \mu)]}{n} \\
&= \frac{\sum_{i=1}^n \sigma^2}{n} = \sigma^2
\end{aligned}$$

   where we have used the fact that the expected value of the product of independent random variables is the product of the expected values. Hence, the variance is constant. For faster growing scales, the variance would blow up and for slower growing ones, it would converge to 0.

3. While the law of large numbers discusses the probability of a limit, the central limit theorem discusses the limit of a probability. The central limit theorem does not imply that the scaled deviation converges to a random variable. In general, this does not hold.

We now investigate applying the central limit theorem to help us determine the deviation of the growth of our growth optimal portfolio. Applying Equation (3.11) to $\ln(V_{n+1})$ yields:

$$\begin{aligned}\lim_{n\to\infty}\mathbb{P}\left(\sqrt{n}\frac{\frac{\ln(V_{n+1})}{n}-\mu}{\sigma}\leq x\right) &= \lim_{n\to\infty}\mathbb{P}\left(\frac{\ln(V_{n+1})}{n}-\mu\leq\frac{\sigma x}{\sqrt{n}}\right)\\ &= \lim_{n\to\infty}\mathbb{P}\left(\frac{\ln(V_{n+1})}{n}\leq\mu+\frac{\sigma x}{\sqrt{n}}\right)\end{aligned}$$

where:

$$\mu = E\big[\ln\big(1+w^T R_1\big)\big] \tag{3.12}$$

$$\sigma = \sqrt{E\Big[\big(\ln(1+w^T R_1)-\mu\big)^2\Big]} \tag{3.13}$$

$$\tag{3.14}$$

Figure 3.3 shows a histogram of 10,000 simulations of the average growth of the growth optimal portfolio over 10 yearss superposed with the density function of the normal distribution predicted by the central limit theorem. There are two sources of noise here:

1. The noise from simulating only 10,000 paths. The number of paths need to go to ∞ for the histogram to converge to the distribution.

2. The noise from running only 10 years. The central limit theory holds in the limit as time goes to ∞.

However, these sources of noise seem to be within a reasonable margin for these simulation parameters as the fit is fairly close.

3.9 The Law of the Iterated Logarithm

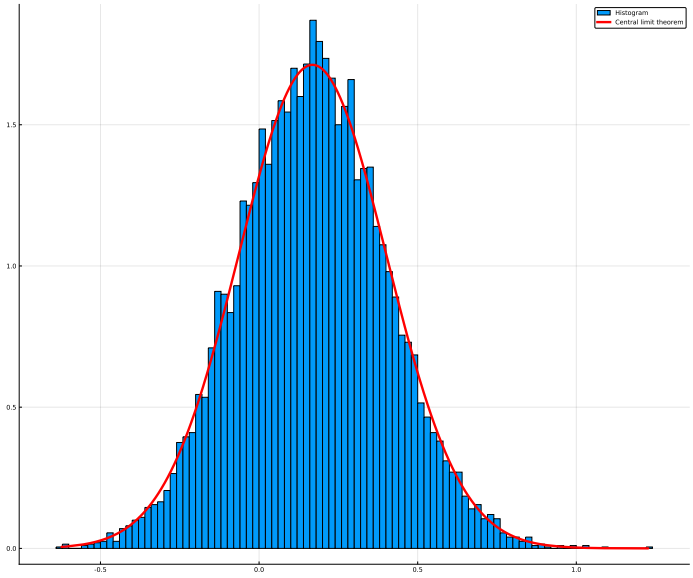
One important attribute of portfolio performance which has been missing from our analysis so far is the best case and worst case behavior of portfolios. The law of the iterated logarithm is a limit theorem that provides insight into the maximum and minimum of the portfolio value in the limit. In order to state the law of the iterated logarithm, we must first introduce the limit supremum of a sequence $a_1, a_2, \ldots \in \mathbb{R}$:

$$\limsup_{n\to\infty} a_n = \lim_{n\to\infty}\sup_{m>n} a_m \tag{3.15}$$

A few comments on limit suprema:

1. While limits don't always exist, the limit supremum always exists or is ∞.

Figure 3.3: Histogram of Growth Optimal Portfolio



2. A sequence comes arbitrarily close to its limit supremum infinitely often. More precisely, for any $\epsilon > 0$ and any n, there is an $m > n$ such that $|a_m - \limsup_n a_n| < \epsilon$. It might not stay near it and it might not ever hit it but always returns to near the limit supremum.

3. A sequence only goes above its limit supremum by some fixed $\epsilon > 0$ a finite number of times. More precisely, for any $\epsilon > 0$, there is an n such tha for all $m > n$, we have that $a_m < \limsup_{n\to\infty} a_n + \epsilon$.

4. The "dual" of the limit supremum is the limit infimum defined by $\liminf_{n\to\infty} a_n = -\limsup_{n\to\infty}(-a_n)$. Note that if $\liminf_{n\to\infty} a_n = \limsup_{n\to\infty} a_n = a^*$ then $\lim_{n\to\infty} a_n$ exists and $\lim_{n\to\infty} a_n = a^*$.

We can now present the law of the iterated logarithm:

Theorem 3.9.1. Law of the Iterated Logarithm: *For IID random variables* $X_1, X_2, \ldots$ *such that* $\mu = E[X_1]$ *and* $\sigma = \sqrt{E\left[(X_1 - \mu)^2\right]}$ *exist and are finite:*

$$\mathbb{P}\left(\limsup_{n\to\infty} \frac{S_n - n\mu}{\sigma\sqrt{2n\log(\log(n))}} = 1\right) = 1$$

and:

$$\mathbb{P}\left(\limsup_{n\to\infty} \frac{-(S_n - n\mu)}{\sigma\sqrt{2n\log(\log(n))}} = 1\right) = 1$$

We provide some comments on the law of the iterated logarithm.

1. As in the law of large numbers, the law of the iterated logarithm is a probability 1 result. In fact, it can be seen that the law of the iterated logarithm, as stated here, implies the law of large numbers. For any $\epsilon > 0$, we have, with probability 1.

$$\begin{aligned}\limsup_{n\to\infty} \frac{S_n - n\mu}{n^{\frac{1}{2}+\epsilon}} &= \limsup_{n\to\infty} \frac{S_n - n\mu}{\sigma\sqrt{2n\log(\log(n))}} \frac{\sigma\sqrt{2n\log(\log(n))}}{n^{\frac{1}{2}+\epsilon}} \\ &= \limsup_{n\to\infty} \frac{\sigma\sqrt{2\log(\log(n))}}{n^{\epsilon}} \leq \limsup_{n\to\infty} \frac{n^{\frac{\epsilon}{2}}}{n^{\epsilon}} \\ &= \limsup_{n\to\infty} \frac{1}{n^{\frac{\epsilon}{2}}} = 0\end{aligned}$$

   where we have used the law of the iterated logarithm and that $\log(n) < n^{\epsilon}$ for any $\epsilon > 0$ and sufficiently large n. Similarly, using the other direction of the law of the iterated logarithm yields:

$$\liminf_{n\to\infty} \frac{S_n - n\mu}{n^{\frac{1}{2}+\epsilon}} = 0$$

   with probability 1. Putting those together yields:

$$\lim_{n\to\infty} \frac{S_n - n\mu}{n^{\frac{1}{2}+\epsilon}} = 0 \tag{3.16}$$

   with probability 1. Using $\epsilon = \frac{1}{2}$ yields the strong law of large numbers Note that this proves more than the law of large numbers since we Equation (3.16) holds for any $\epsilon > 0$ and not just $\frac{1}{2}$ required for the law of large numbers.

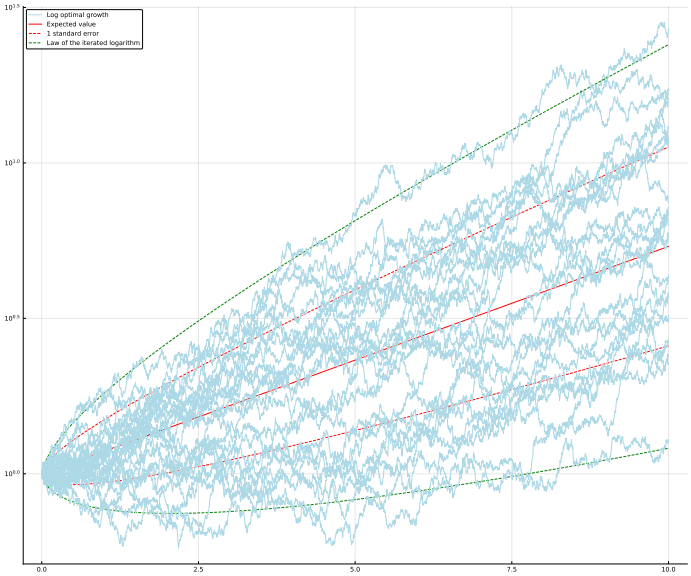
We now show how the law of the iterated logarithm can give us insight into the asymptotics of drawdowns. As mentioned in Item 3 of the comments on limit suprema, for any $\epsilon > 0$ we have that:

$$\frac{\log V_n - n\mu}{\sigma\sqrt{2n\log(\log(n))}} \leq 1 + \epsilon$$

except for a finite number of times with probability 1 where μ and σ are defined in Equations (3.12) and (3.13). Similarly:

$$\frac{\log V_n - n\mu}{\sigma\sqrt{2n\log(\log(n))}} \geq -1 - \epsilon$$

Figure 3.4: Growth Optimal Portfolio and the Law of the Iterated Logarithm Bounds



except for a finite number of times with probability 1. Also, the process comes arbitrarily close to these values infinitely often. Since the process is growing at roughly $n\mu$ and $\mu > 0$, we can ignore the finite number of exceptions to the above equations in calculating the maximum since the process will eventually exceed those. Hence, the maximum will grow approximately as:

$$\max_{m \in [1,n] \cap \mathbb{N}} \log(V_m) \approx n\mu + \sigma\sqrt{2n\log(\log(n))}$$

Similarly, minimum values are approximately:

$$\log(V_n) \approx n\mu - \sigma\sqrt{2n\log(\log(n))}$$

Figure 3.4 shows these bounds with 25 paths of the growth optimal portfolio value on a log scale. While it isn't guaranteed for any particular finite value, these bounds seem to largely contain the data.

We now wish to discuss the long-term behavior of sums of random variables when we relax the assumption that those random variables are IID. In order to do this, we first discuss conditional expectations.

3.10 The Radon-Nikodym Derivative and Conditional Expectations

In order to discuss conditional expectation, we will need to introduce several new concepts. We first discuss a generalization of a probability measure, called a signed measure. The main difference is that a signed measure can be positive or negative and need not sum to 1.

Definition 3.10.1. *Given a σ-algebra $\mathcal{F}$, a* signed measure*, μ, is a countably additive set function, that is, as set function $\mu : \mathcal{F} \to [-\infty, \infty]$ such that:*

Countable additivity: *For disjoint sets $F_1, F_2, \ldots \in \mathcal{F}$:*

$$\mu\left(\bigcup_{i=1}^{\infty} F_i\right) = \sum_{i=1}^{\infty} \mu(F_i)$$

A signed measure, μ, is called a **measure** *if $\mu(F) \geq 0$ for all $F \in \mathcal{F}$.*

Some properties of signed measures include:

1. For any sets $F_1 \subseteq F_2$ with $F_1, F_2 \in \mathcal{F}$:

   $$\mu(F_2 - F_1) = \mu(F_2) - \mu(F_1)$$

   since $F_2 - F_1$ and F_1 are disjoint so that:

   $$\mu(F_2) = \mu((F_2 - F_1) \cup F_1) = \mu(F_2 - F_1) + \mu(F_1)$$

2. $$\mu(\varnothing) = 0$$

   since $\varnothing \subseteq \varnothing$ so that:

   $$\mu(\varnothing) = \mu(\varnothing - \varnothing) = \mu(\varnothing) - \mu(\varnothing)$$

3. Analogous to a random variable for a probability measure, a function, $f : \Omega \to \mathbb{R}$, is called a **measurable function** if:

   $$f^{-1}([a, b]) \in \mathcal{F}$$

   for all $a, b \in \mathbb{R}$.

4. For any measurable function, the **integral**, written $\int f d\mu$ can be defined analogously to how we defined expected value. It has many of the same properties such as linearity:

   Linearity:

$$\int (\alpha f + \beta g)\, d\mu = \alpha \int f d\mu + \beta \int g d\mu$$

We now introduce a way to turn the expected value of a random variable under a probability measure into a signed measure. For a probability space, $(\Omega, \mathcal{F}, \mathbb{P})$, and a random variable X, define $\mu : \mathcal{F} \to [-\infty, \infty]$ by:

$$\mu(F) = E[X \mathbb{1}_F] \tag{3.17}$$

It can be shown that this defines a signed measure. In fact, under fairly broad circumstances, any signed measure μ on a σ-algebra $\mathcal{F}$ will have a density with respect to probability measure $\mathbb{P}$ on $\mathcal{F}$ called the Radon-Nikodym derivative:

Theorem 3.10.1. Radon-Nikodym theorem: *If, for a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a signed measure μ on $\mathcal{F}$, we have that, for all $F \in \mathcal{F}$:*

$$\mathbb{P}(F) = 0 \Rightarrow \mu(F) = 0$$

then there is a random variable X such that:

$$\mu(F) = E[X \mathbb{1}_F]$$

for all $F \in \mathcal{F}$.

The property needed for the theorem is called **absolute continuity** and written $\mu \ll \mathbb{P}$, that is, $\mu \ll \mathbb{P}$ if and only if for all $F \in \mathcal{F}$, we have that $\mathbb{P}(F) = 0 \Rightarrow \mu(F) = 0$.

We now introduce conditional expectation. Let us start with a probaiblity space $(\Omega, \mathcal{F}, \mathbb{P})$ and a random variable X on that probability space. By the conditional expectation of X, we mean the expectation conditioned on some event, say, a set $F \in \mathcal{F}$. If $\mathbb{P}(F) > 0$, the conditional expectation, written $E[X|F]$ can be defined as:

$$E[X|F] = \frac{E[X \mathbb{1}_F]}{\mathbb{P}(F)}$$

which matches the intuition of the expected value given F. However, if $\mathbb{P}(F) = 0$ such as, for example, when $F = \{Y = y\}$ for some continuous random variable

Y and some value $y \in \mathbb{R}$, the above definition doesn't work. In addition to the conditional expectation with respect to a single random variable, $E[X|\, Y = y]$, we will also be interested in the conditional expectation with respect to multiple random variables, $E[X|\, Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_N]$. We will do this by encapsulating the information corresponding to $Y_1, Y_2, \ldots, Y_n$ into a σ-algebra, $\mathcal{G}$, and taking the conditional expectation with respect to the σ-algebra, $E[X|\, \mathcal{G}]$.

Consider a σ-algebra $\mathcal{G} \subseteq \mathcal{F}$. Any measure on $\mathcal{F}$, restricted to $\mathcal{G}$, is a measure on $\mathcal{G}$. Let μ be the signed measure defined by Equation (3.17). Note that $\mu \ll \mathbb{P}$ since, if $\mathbb{P}(F) = 0$ then:

$$\mu(F) = E[X \mathbb{1}_F] = 0$$

Hence, there is a Radon-Nikoydym derivative of μ restricted to $\mathcal{G}$ with respect to $\mathbb{P}$ restricted to $\mathcal{G}$. This is the conditional expectation $E[X|\, \mathcal{G}]$. For the purpose of conditional expectation, a σ-algebra corresponds intuitively to a state of knowledge, that is, to everything that is known. $E[X|\, \mathcal{G}]$ then refers to the mean of X given that we know $\mathcal{G}$.

Properties of the conditional expectation include:

1. **Linearity:** $E[aX + bY|\, \mathcal{G}] = aE[X|\, \mathcal{G}] + bE[Y|\, \mathcal{G}]$
2. If X is $\mathcal{G}$ measurable then $E[X|\, \mathcal{G}] = X$.
3. **The law of iterated expectation:** If $\mathcal{G}_1$ and $\mathcal{G}_2$ are σ-algebras then $\mathcal{G}_1 \subseteq \mathcal{G}_2$ corresponds with knowing more under $\mathcal{G}_2$ than under $\mathcal{G}_1$. Let X be a random variable. One of the most important properties of conditional expectation is the law of iterated expectations:

   $$E[\, E[X|\, \mathcal{G}_2]|\, \mathcal{G}_1] = E[X|\, \mathcal{G}_1]$$

   We also note that by the definition, $E[X|\, \mathcal{G}_1]$ is $\mathcal{G}_1$ measurable so that:

   $$E[\, E[X|\, \mathcal{G}_1]|\, \mathcal{G}_2] = E[X|\, \mathcal{G}_1]$$

   Whenever iterated conditional expectations occur, whichever one is taken first, only the conditional expectation with respect to the smaller σ-algebra remains.
4. If X is a random variable, we let $\sigma(X)$ denote the smallest σ-algebra under which X is measurable. In this case, for another random variable Y, we write $E[Y|\, X]$ instead of $E[Y|\, \sigma(X)]$. Similarly, for a finite or infinite sequence of random variables $X_1, X_2, \ldots$, we let $\sigma(X_1, X_2, \ldots)$ denote the σ-algebra generated by $X_1, X_2, \ldots$ and we write $E[Y|\, X_1, X_2, \ldots]$ as a shorthand notation for $E[Y|\, \sigma(X_1, X_2, \ldots)]$. An example of a random variable that is measurable (a random variable) with respect to $\sigma(X_1, X_2, \ldots)$ is $\limsup_n \frac{1}{n} \sum_{i=1}^n X_i$ even though it is not measurable with respect to any finite set of these random variables.

5. **Jensen's inequality:** for a concave function $f : \mathbb{R} \to \mathbb{R}$ a random variable X, and a σ-algebra $\mathcal{G}$:

$$E[\, f(X) |\, \mathcal{G}] \leq f(E[X |\, \mathcal{G}])$$

We now demonstrate the calculation for random variables with densities or point mass functions.

Example 3.10.1.

3.11 Dependent Returns

Thus far, we have investigated the behavior of sums of random variables under the assumption that returns are independent and identically distributed. As all models, this is an approximation of reality. One well studied phenomenon which deviates from this assumption is stochastic volatility. Stocks tend to move in clusters of either large or small moves. In this section, we briefly discuss relaxing the assumptions of independence and identical distribution.

3.11.1 The Growth Optimal Portfolio for Dependent Returns

We will now introduce the growth optimal portfolio for arbitary, dependent or independent, returns. First, let $\mathcal{F}_t$ denote the σ-algebra of all information known at time t. In particular, we assume that R_t is a random variable with respect to $\mathcal{F}_t$. Also, since the portfolio could be chosen with information available at time t, we assume that w_t is a random variable with respect to $\mathcal{F}_t$. We will now show that, in the case of arbitrary returns, dependent or independent, the growth optimal portfolio, w_t^*, is chosen by:

$$w_t^* = \underset{w}{\operatorname{argmax}}\, E\big[\ln\big(1 + w^T R_{t+1}\big)\big|\, \mathcal{F}_t\big]$$

where argmax_w denotes the w which yields the maximum of the subsequent expression (note, in this case, the maximum can be shown to exist under broad conditions and be unique under somewhat less broad conditions). In order to show this, first consider the expected relative value of another portfolio w compared to w^* from Equation (3.9):

$$E\big[\ln\big(1 + w_t^T R_{t+1}\big)\big|\, \mathcal{F}_t\big] \leq E\Big[\ln\Big(1 + (w_t^*)^T R_{t+1}\Big)\Big|\, \mathcal{F}_t\Big]$$

so that:

$$\begin{aligned}\ln\left(E\left[\left.\frac{1+w_t^T R_{t+1}}{1+(w_t^*)^T R_{t+1}}\right|\mathcal{F}_t\right]\right) &\leq E\left[\left.\ln\left(\frac{1+w_t^T R_{t+1}}{1+(w_t^*)^T R_{t+1}}\right)\right|\mathcal{F}_t\right] \\ &= E\big[\ln\big(1+w^T R_{t+1}\big)\big|\,\mathcal{F}_t\big] - E\Big[\ln\Big(1+(w^*)^T R_{t+1}\Big)\Big|\,\mathcal{F}_t\Big] \leq 0\end{aligned}$$

where the first inequality follows from Jensen's inequality. Taking the exponential:

$$E\left[\left.\frac{1+w_t^T R_{t+1}}{1+(w_t^*)^T R_{t+1}}\right|\mathcal{F}_t\right] \leq 1 \tag{3.18}$$

Define $V_n(w)$ to be the value of a portfolio w taken from Equation (3.9):

$$V_n(w) = V_1 \prod_{t=1}^{n-1}\left(1+w_t^T R_{t+1}\right)$$

and let $V_n(w,w^*) = \frac{V_n(w)}{V_n(w^*)}$ be the relative value of the portfolio w with respect to the growth optimal portfolio w^*. Note that $V_n(w,w^*)$ can be seen to be a random variable as long as $1+(w_t^*)^T R_{t+1} > 0$ since division has a discontinuity when the denominator is 0. However, the growth optimal portfolio will have this property if any portfolio does, since $1+(w_t^*)^T \geq 0$ by the definition of arithmetic returns but the log utility will be $-\infty$ if $1+(w_t^*)^T = 0$. Furthermore, $V_n(w,w^*)$ is $\mathcal{F}_n$ measurable since it is only directly dependent upon $w_1, w_2, \ldots, w_{n-1}$ and $w_1^*, w_2^*, \ldots, w_{n-1}^*$ which are $\mathcal{F}_{n-1}$ measurable and on $R_2, \ldots, R_{n+1}$ which are $\mathcal{F}_n$ measurable. Consider the expected value of $V_{n+1}(w,w^*)$ conditional on $\mathcal{F}_n$:

$$\begin{aligned}E[V_{n+1}(w,w^*)|\,\mathcal{F}_n] &= E\left[\left.V_n(w,w^*)\,\frac{1+w_n^T R_{n+1}}{1+(w_n^*)^T R_{n+1}}\right|\mathcal{F}_n\right] \\ &= V_n(w,w^*)\,E\left[\left.\frac{1+w_n^T R_{n+1}}{1+(w_n^*)^T R_{n+1}}\right|\mathcal{F}_n\right] \leq V_n(w,w^*)\end{aligned}$$

where the first equality is because $V_n(w,w^*)$ is $\mathcal{F}_n$ measurable as mentioned above and the second equality is from Equation (3.18). A sequence of random variables with this property is called a supermartingale which we discuss in the next section.

3.11.2 Supermartingales

Definition 3.11.1. *A* **supermartingale** *with respect to a sequence of σ-algebras $\mathcal{F}_1, \mathcal{F}_2, \ldots$ is a sequence of random variables $X_1, X_2, \ldots$ such that X_i is $\mathcal{F}_i$ measurable and:*

$$E[\,X_{n+1}|\,\mathcal{F}_n] \leq X_n$$

Supermartinagles are a substantial and powerful generalization of sums of independent random variables since if $X_1, X_2, \ldots$ are independent:

$$E\left[\sum_{i=1}^{n+1} X_i \,\middle|\, X_1, X_2, \ldots, X_n\right] = \sum_{i=1}^{n} X_i + E[\,X_{n+1}|\,X_1, X_2, \ldots, X_n] = \sum_{i=1}^{n} X_i$$

Hence, $Y_1, Y_2, \ldots$ defined by $Y_n = \sum_{i=1}^{n} X_i$ are a supermartingale with respect to the sequence of σ-algebras $\mathcal{F}_1, \mathcal{F}_2, \ldots$ defined by $\mathcal{F}_n = \sigma(X_1, X_2, \ldots, X_n)$. An important property of supermartingales is that they converge if they are nonnegative:

Theorem 3.11.1. Supermartingale convergence theorem: *Let $S_1, S_2, \ldots$ be a nonnegative supermartingale with respect to the σ-algebras $\mathcal{F}_1, \mathcal{F}_2, \ldots$, then there is a random variable S^* such that:*

$$\mathbb{P}\Big(\lim_{n\to\infty} S_i = S^*\Big) = 1$$

Furthermore:

$$\mathbb{P}\left(\limsup_n \frac{1}{n}\ln(S_n) \leq 0\right) = 1$$

Applying the supermartingale convergence theorem to $V_n(w, w^*)$ yields:

Theorem 3.11.2. Competitive optimality of the growth optimal portfolio: *If w^* is the growth optimal portfolio and w is any other portfolio, then:*

$$\mathbb{P}\left(\limsup_n \frac{1}{n}\ln\left(\frac{V_n(w)}{V_n(w^*)}\right) \leq 0\right) = 1$$

3.11.3 Stationary and Ergodic Random Sequences

We first consider relaxation of the assumption of independence of returns. We will maintain the equivalent definition of "identical distribution" which is called stationarity or strict stationarity:

Definition 3.11.2. *A sequence of random variables $X_1, X_2, \ldots$ is called* **stationary** *or* **strictly stationary** *if:*

$$\begin{aligned}&\mathbb{P}(X_1 \in [a_1, b_1], X_2 \in [a_2, b_2], \ldots, X_n \in [a_n, b_n])\\ &\quad = \quad \mathbb{P}(X_{1+k} \in [a_1, b_1], X_{2+k} \in [a_2, b_2], \ldots, X_{n+k} \in [a_n, b_n])\end{aligned}$$

for any $n, k \in \mathbb{N}$ and any $a_1, a_2, \ldots, a_n, b_1, b_2, \ldots, b_n \in \mathbb{R}$.

Note that while stock prices are not stationary: they tend to move up over time, stock returns may well be modeled as a stationary process. Indeed, a stochastic volatility model is stationary as long as the model of the dynamics of volatility is stationary which practical models of stochastic volatility are.

The law of large numbers does not hold for every stationary process. For example, consider a process in which returns are chosen as a constant 5% per annum in perpetuity with 50% probability and 10% per annum in perpetuity with the other 50%. The expected value, at any time is 7.5% but in neither case will the sample average converge to this. The extra property required is called **ergodicity** and the law of large numbers for all stationary and ergodic stochastic processes and is referred to as the **ergodic theorem**. However, the **ergodic theorem** does not imply any rate of convergence such as the law of the iterated logarithm. Both the law of the iterated logarithm and the central limit theorem require additional assumptions such as the process being the a martingale difference (the difference between consecutive random variables in a martingale). While martingale differences have expected value 0, it is easy to accomodate a non-zero mean return by subtracting it out.

3.12 Limitations of the Growth Optimal Portfolio

We have demonstrated that the growth optimal portfolio maximizes long-term wealth. However, this rule has several limitations: some making it take on too much risk and some too little. For one, as seen in the simulations of the previous section, it might not perform well in reasonable amounts of time. Indeed, in industry, this rule is considered to produce quite risky portfolios. Some quantitative portfolio management firms halt trading for portfolio managers whose portfolios go down by 2.5% of its peak value. The simulated growth optimal portfolio discussed here hits that negative benchmark very frequently. This could be addressed, for example, by using the worst case long-term value given by the law of the iterated logarithm.

On the other hand, because the log becomes $-\infty$ if the portfolio value goes to 0, the growth optimal portfolio will never short nor utilize leverage. It is not clear how to address this limitation. While losing 100% of the portfolio value is a negative outcome, it might not be considered to be infinitely negative.

Ultimately, whether using the growth optimal portfolio, another utility function or some other approach entirely, there are several systematic ways to choose a portfolio once the distribution of returns is known. We now focus on the determination of the distribution of returns.

Chapter 4

Nearest Neighbor Learning and Friends

4.1 Introduction

In the last chapter, we saw how the log optimal portfolio yields the optimal long-term growth portfolio when the distribution of returns is known. In this chapter, we introduce a machine learning technique called nearest neighbor and show its application to portfolio selection. One advantage of this technique, as well as of a few other techniques, is that it is known to have a universal property by which it yields the optimal solution for many problems under arbitrary underlying distributions. For the problem of portfolio selection, we will later see that the nearest neighbor technique will yield portfolios with optimal long-term growth for almost any stationary ergodic stochastic process. To give a sense of how this works, we demonstrate these results in a more limited setting in this section before delving into more mathematical detail in the next section.

The setup for nearest neighbor learning, and indeed for many learning techniques, is a sequence of pairs of random variables or vectors $(X_1, Y_1), (X_2, Y_2), \ldots$ such that the pairs are typically assumed to be independent and identically distributed across time though X_i and Y_i typically have significant dependence. We think of X_i as an observation to help us predict Y_i. At each step, we have observed say $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ as well as X_{n+1} and wish to predict Y_{n+1}. If, for example, the Y_i's are discrete random variables corresponding to classes, then the problem is called classification. As example of a classification problem is image recognition where the X_i's are random vectors corresponding to the colors of the pixels in an image and the Y_i's are labels corresponding to the class of the object in the image. When Y_i's are continuous random variables, the problem is typically called regression. An example of regression, which will be important to us here, is when the X_i's are random vectors of stock characteristics, and the Y_i's are random vectors corresponding to the stock returns in a future period.

We illustrate nearest neighbor learning with the case where the X_i's are random variables and not random vectors. Consider the case where we have observed $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ and X_{n+1} and we wish to form a prediction of Y_{n+1}. First, for any $x \in \mathbb{R}$, define $X_{k,n}(x)$ to be the *kth* nearest neighbor of x, that is, $X_{1,n}, X_{2,n}, \ldots, X_{n,n}$ are chosen such that:

$$|X_{1,n} - x| \leq |X_{2,n} - x| \leq \ldots \leq |X_{n,n} - x|$$

and let $Y_{k,n}(x)$ be the Y_i corresponding to $X_{k,n}(x)$. The *kth* nearest neighbor (k-NN) estimate of Y_{n+1} is given by:

$$\hat{Y}_{n+1} = \frac{1}{k} \sum_{i=1}^{k} Y_{i,n}(X_{n+1})$$

that is, we take the sample average of the Y_i's corresponding to the k nearest neighbors of X_{n+1} among $X_1, X_2, \ldots, X_n$. In practice, k is typically chosen to be grow with n and we will denote it by k_n.

k_n-NN learning can be shown to converge to the optimal estimate under a broad range of conditions. We demonstrate its universality in the following simplified setting for now. We will present generalizations of this result subsequently.

Theorem 4.1.1. Simplified universality of nearest neighbor: *Let $X_1, X_2, \ldots$ be a sequence of IID random variables. and let $Y_i = f(X_i) + \epsilon_i$ for a continuous and bounded function $f : \mathbb{R} \to \mathbb{R}$ and for $\epsilon_1, \epsilon_2, \ldots$ IID, independent of $X_1, X_2, \ldots$ and such that $E[\epsilon_i] = 0$. If $\lim_{n\to\infty} k_n = \infty$ and $\lim_{n\to\infty} \frac{k_n}{n} = 0$ then:*

$$\mathbb{P}\Big(\lim_{n\to\infty} \Big|\hat{Y}_n - f(X_n)\Big| = 0\Big) = 1$$

Note that this is a nonlinear regression. However, rather than being a nonlinear regression with a particular parametric form, this technique works for all possible parametric forms of which there are infinitely many. Furthermore, the only requirement on the noise is that the expected value exists and is 0. Hence, the noise could have a density, a point mass function, be singular or be a mixture of any of these.

In order to demonstrate this result, we must first define the support of a probability distribution:

Definition 4.1.1. *Given a probability space $(\mathbb{R}, \mathcal{B}, \mathbb{P})$, a point $x \in \mathbb{R}$ is in the support of $\mathbb{P}$ if for all $\epsilon > 0$, we have that $\mathbb{P}((x - \epsilon, x + \epsilon)) > 0$. The set of all points in the support of $\mathbb{P}$ is written as* $\text{supp}(\mathbb{P})$.

Note that $\mathbb{P}(\text{supp}(\mathbb{P})) = 1$. We will now present a simple lemma that shows that as n becomes large, the k_n nearest neighbors, $X_{1,n}(x), X_{2,n}(x), \ldots, X_{k_n,n}(x)$, all converge to x. This lemma is useful for more than the above universality result:

Lemma 4.1.1. *Let $X_1, X_2, \ldots$ be a sequence of IID random variables with distribution $\mathbb{P}$. Suppose that $k_1, k_2, \ldots \in \mathbb{N}$ are such that $\frac{k_n}{n} \to 0$. For any $x \in \text{supp}(\mathbb{P})$, we have that:*

$$\lim_{n\to\infty} X_{k_n,n}(x) = x$$

Proof. First, let k'_n be a sequence such that $k'_n \geq k_n$ and that $k'_n \to \infty$ and that $\frac{k'_n}{n} \to 0$ (for example, $k'_n = \max(k_n, \sqrt{n})$ has these properties). Let $x \in \text{supp}(\mathbb{P})$ and define $S_{x,\epsilon} = \{y \in \mathbb{R} : |x - y| < \epsilon\}$. We will have that $X_{k'_n,n} \in S_{x,\epsilon}$ if:

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{X_i \in S_{x,\epsilon}} \geq \frac{k'_n}{n}$$

However, by the strong law of large numbers, the left hand side above converges, with probability 1, to $\mathbb{P}(S_{x,\epsilon}) > 0$ since $x \in \text{supp}(\mathbb{P})$. Furthermore, the right hand side converges to 0 by assumption. Hence, the result will be true for sufficiently large n.

To see that this is true for k_n as well as k'_n, note that:

$$|X_{k_n,n} - x| \leq \left|X_{k'_n,n} - x\right| \to 0$$

□

We now discuss why Theorem 4.1.1 works. First note that, from Lemma 4.1.1, we have that $\lim_{n\to\infty} X_{k_n,n}(x) = x$ for $x \in \text{supp}(\mathbb{P})$. This is called pointwise convergence of the function $X_{k_n,n}$:

Definition 4.1.2. *A sequence of functions $f_n : S \to \mathbb{R}$ for some set S* **converges pointwise** *to a function $f : S \to \mathbb{R}$ if for all $s \in S$:*

$$\lim_{n\to\infty} f_n(s) = f(s)$$

From pointwise convergence of $X_{k_n,n}$ to x (the identity function) on the support of the distribution, we would like to conclude that:

$$\lim_{n\to\infty} |X_{k_n,n}(X_{n+1}) - X_{n+1}| = 0$$

However, pointwise convergence alone does not imply this as the following example shows:

Example 4.1.1. *Consider the sequence of functions $f_n : (-1, 1) \to \mathbb{R}$ given by:*

$$f_n(x) = \frac{1}{n\,(x^2 - 1)}$$

For any point $x \in (-1,1)$, we have:

$$\lim_{n\to\infty} f_n(x) = 0$$

However, if we take $x_n = 1 - \frac{1}{n}$ then:

$$\begin{aligned}\lim_{n\to\infty} f_n(x_n) &= \frac{1}{n\left(\left(1-\frac{1}{n}\right)^2 - 1\right)} = \frac{1}{n\left(1-\frac{2}{n}+\frac{1}{n^2}-1\right)} \\ &= \frac{1}{n\left(\frac{2}{n}+\frac{1}{n^2}\right)} = \frac{1}{-2+\frac{1}{n}} = -\frac{1}{2}\end{aligned}$$

Hence, pointwise convergence, even to a continuous function doesn't imply that $\lim_{n\to\infty} |f_n(x_n) - x_n| = 0$ as we would like.

It turns out the result can be proven by more refined techniques which we will see later. To finish the result:

$$\begin{aligned}\lim_{n\to\infty} \left|\hat{Y}_{n+1} - f(X_{n+1})\right| &= \lim_{n\to\infty} \left|\frac{1}{k_n}\sum_{i=1}^{k_n} Y_{i,n}(X_{n+1}) - f(X_{n+1})\right| \\ &= \lim_{n\to\infty} \left|\frac{1}{k_n}\sum_{i=1}^{k_n} \left(f(X_{i,n}(X_{n+1})) + \epsilon_{i,n}(X_{n+1})\right) - f(X_{n+1})\right| \\ &= \lim_{n\to\infty} \left|\frac{1}{k_n}\sum_{i=1}^{k_n} \left(f(X_{n+1}) + \epsilon_{i,n}(X_{n+1})\right) - f(X_{n+1})\right| \\ &= \lim_{n\to\infty} \left|\frac{1}{k_n}\sum_{i=1}^{k_n} \epsilon_{i,n}(X_{n+1})\right| = 0\end{aligned}$$

with probability 1. The last equality above follows from the law of large numbers.

4.2 Universality of Nearest Neighbor

We now show a more general form of universality for the nearest neighbor method of machine learning. In order to measure how closely the method comes to the true values, we will use the mean squared error measure:

$$E\left[\left(\hat{Y}_n - Y_n\right)^2\right]$$

for some estimate $\hat{Y}_n$ which is measurable with respect to some σ-algebra $\mathcal{G}$. To begin, let's consider the best possible estimate, assuming that we know the probability measure $\mathbb{P}$:

$$\begin{aligned} E\left[\left(\hat{Y}_n - Y_n\right)^2\right] &= E\left[\left(\hat{Y}_n - E[Y_n|\,\mathcal{G}] + E[Y_n|\,\mathcal{G}] - Y_n\right)^2\right] \\ &= E\left[\left(\hat{Y}_n - E[Y_n|\,\mathcal{G}]\right)^2 + 2\left(\hat{Y}_n - E[Y_n|\,\mathcal{G}]\right)\left(E[Y_n|\,\mathcal{G}] - Y_n\right) + \left(E[Y_n|\,\mathcal{G}] - Y_n\right)^2\right] \\ &= E\left[\left(\hat{Y}_n - E[Y_n|\,\mathcal{G}]\right)^2\right] + 2E\left[\left(\hat{Y}_n - E[Y_n|\,\mathcal{G}]\right)\left(E[Y_n|\,\mathcal{G}] - Y_n\right)\right] \\ &\quad + E\left[\left(E[Y_n|\,\mathcal{G}] - Y_n\right)^2\right] \end{aligned} \tag{4.1}$$

where we have used linearity of expectation. Now consider the middle term from Equation (4.1):

$$\begin{aligned} &2E\left[\left(\hat{Y}_n - E[Y_n|\,\mathcal{G}]\right)\left(E[Y_n|\,\mathcal{G}] - Y_n\right)\right] \\ &= 2E\left[E\left[\left(\hat{Y}_n - E[Y_n|\,\mathcal{G}]\right)\left(E[Y_n|\,\mathcal{G}] - Y_n\right)\middle|\,\mathcal{G}\right]\right] \\ &= 2E\left[\left(\hat{Y}_n - E[Y_n|\,\mathcal{G}]\right)E[E[Y_n|\,\mathcal{G}] - Y_n|\,\mathcal{G}]\right] \\ &= 2E\left[\left(\hat{Y}_n - E[Y_n|\,\mathcal{G}]\right)\left(E[Y_n|\,\mathcal{G}] - E[Y_n|\,\mathcal{G}]\right)\right] \\ &= 2E\left[\left(\hat{Y}_n - E[Y_n|\,\mathcal{G}]\right)0\right] = 0 \end{aligned} \tag{4.2}$$

where we have used the law of iterated expectations and the fact that terms and factors which are measurable with respect to $\mathcal{G}$ can be pulled out of conditional expectations with respect to $\mathcal{G}$. Putting together Equations (4.1) and (4.2) yields:

$$\begin{aligned} E\left[\left(\hat{Y}_n - Y_n\right)^2\right] &= E\left[\left(\hat{Y}_n - E[Y_n|\,\mathcal{G}] + E[Y_n|\,\mathcal{G}] - Y_n\right)^2\right] \\ &= E\left[\left(\hat{Y}_n - E[Y_n|\,\mathcal{G}]\right)^2\right] + E\left[\left(E[Y_n|\,\mathcal{G}] - Y_n\right)^2\right] \end{aligned} \tag{4.3}$$

Note that the right hand side of Equation (4.3) is unaffected by choices of $\hat{Y}_n$ and the left hand side, which is non-negative, becomes 0 if we set $\hat{Y}_n = E[Y_n|\,\mathcal{G}]$.

Now consider IID pairs $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$. If we set $m(x) = E[Y_i|\,X_i = x]$ then, the optimal estimate $\hat{Y}_n$ has:

$$E\left[\left(\hat{Y}_n - m(X_n)\right)^2\right] = 0$$

If this is true asymptotically for a set probability measures, we call the estimates universally consistent on that set:

Definition 4.2.1. *A sequence of estimators is* **universally consistent** *on a set of probability measures $\mathcal{P}$ if, for all probability distributions $\mathbb{P} \in \mathcal{P}$, we have that:*

$$\lim_{n\to\infty} E^{\mathbb{P}}\left[\left(\hat{Y}_n - m(X_n)\right)^2\right] = 0$$

The nearest neighbor method can be shown to be universally consistent for a broad set of probability measures:

Theorem 4.2.1. *If $\lim_{n\to\infty} k_n = \infty$ and $\lim_{n\to\infty} \frac{k_n}{n} = 0$ then k_n-NN is universally consistent on the set of probability measures such that:*

1. $\mathbb{P}(X_i = X_j) = 0$ *for* $i \neq j$
2. $E^{\mathbb{P}}\left[Y_i^2\right] < \infty$

Note that Item 1 is about breaking ties. It is true if X_i has a continuous distribution. It can be eliminated if ties are broken in a reasonable way.

Note that, for the problem of Theorem 4.1.1, we have that:

$$E[Y_n|\, X_n] = E[f(X_n) + \epsilon_n|\, X_n] = f(X_n)$$

so that Theorem 4.2.1 implies that, under mild circumatances on $\mathbb{P}$:

$$\lim_{n\to\infty} E^{\mathbb{P}}\left[\left(\hat{Y}_n - f(X_n)\right)^2\right] = 0$$

When Y_i is bounded, this implies that:

$$\mathbb{P}\Big(\lim_{n\to\infty}\left|\hat{Y}_n - f(X_n)\right|\Big) = 1$$

4.3 Nearest Neighbor in Portfolio Selection

We now investigate how nearest neighbor performs in portfolio selection. The theory for this section is laid out in [LGW08] and the empirical results in Chapter 2 of [LGW12]. We first point out a few ways in which the results of the previous sections are not sufficient for this problem:

1. As we have discussed, our measure of performance in portfolio selection is the long-term growth of the value of the portfolio rather than the mean-squared error used in Theorem 4.2.1.
2. As also discussed, returns are can be better modeled by incorporating dependence such as stochastic volatility whereas 4.2.1 assumes that the pairs are IID.

We will address these issues in this section.

While the nearest neighbor method allows us to use arbitary information in the regression, we will only investigate using past returns to predict future ones. We could, for example, let X_i be a vector of past returns and let Y_i be the vector of returns from the current period. However, in this case, the pairs are not IID in this case. Furthermore, returns could be dependent upon returns arbitrarily far back. Hence, we have two "windows" to consider:

1. How many neighbors do we look at? As discussed, we would like to choose some k_n such that $\lim_{n\to\infty} k_n = \infty$ and $\lim_{n\to\infty} \frac{k_n}{n} = 0$.
2. How far back do we look in comparing neighbors? We could use $X_i = (R_{i-l}, R_{i-l+1}, \ldots, R_{i-1})$ for any l.

For each k, l, there is a growth optimal portfolio $w^*_{k,l}$ chosen based on the k nearest neighbors using a lookback length of l. Rather than calculate how fast either l or k_n should move, the authors of [LGW08] instead use a mixture of all possible values. They allow the the choice a point mass function $q : \mathbb{N}^+ \times \mathbb{N}^+ \to (0, 1]$ where $\mathbb{N}^+ = \mathbb{N} - \{0\}$. For a particular $k, l \in \mathbb{N}^+$, we write $q_{k,l}$ for the point mass of k and l. This mixture could be achieved by putting a $q_{k,l}$ portion of one's wealth into the strategy which looks back l and uses k neighbors. Under this scheme, the growth of the value of the portfolio is dominated by the supremum of the growth over all $k, l \in \mathbb{N}^+$.

$$\begin{aligned}
\liminf_{n\to\infty} \frac{1}{n} \log(V_{n+1}) &= \liminf_{n\to\infty} \frac{1}{n} \log\left(\sum_{k,l\in\mathbb{N}^+} q_{k,l} V_{n+1}\big(w^*_{k,l}\big) \right) \\
&\geq \liminf_{n\to\infty} \frac{1}{n} \log\left(\sup_{k,l\in\mathbb{N}^+} q_{k,l} V_{n+1}\big(w^*_{k,l}\big) \right) \\
&= \liminf_{n\to\infty} \sup_{k,l\in\mathbb{N}^+} \frac{1}{n} \log\big(q_{k,l} V_{n+1}\big(w^*_{k,l}\big)\big) \\
&\geq \sup_{k,l\in\mathbb{N}^+} \liminf_{n\to\infty} \frac{1}{n} \log\big(q_{k,l} V_{n+1}\big(w^*_{k,l}\big)\big) \\
&= \sup_{k,l\in\mathbb{N}^+} \liminf_{n\to\infty} \left(\frac{\log(q_{k,l})}{n} + \frac{\log\Big(V_{n+1}\Big(w^*_{k,l}\Big)\Big)}{n} \right) \\
&= \sup_{k,l\in\mathbb{N}^+} \liminf_{n\to\infty} \frac{\log\Big(V_{n+1}\Big(w^*_{k,l}\Big)\Big)}{n}
\end{aligned}$$

The growth of the mixture of strategies is dominated by the supremum of the strategies. It can be shown that this will achieve the optimal growth for all models where the geometric returns are stationary and ergodic. This is a very wide class of processes which includes all Markov processes. Note that, in cases

where there are dependencies going arbitrarily far back, this supremum would not be achieved since it would require a sequence of longer and longer l to capture all of the dependencies. Finally, we mention that this mixture idea works for any learning technique and is not restricted to nearest neighbor.

We now discuss the empirical performance of nearest neighbor portfolio selection. There is an unpublished early version of [LGW08] which uses data from the 1980's used in some earlier papers. On some of this data, the method makes a geometric return of 135% per year over 22 years. However, the data is outdated, ending in the 1985 and also highly volatile with some possible survivorship bias: the stocks were chosen to be volatile but were known to have survived through the period. Chapter 2 of [LGW12] shows results for data which goes through 2006. On this data set, the average annual return, over 44 years is 35%.

4.4 Four Nonparametric Paradigms

While there are many individual techniques for machine learning, [LGW10] discuss 4 different paradigms which encompass many machine learning techniques with a particular focus on nonparametric techniques, that is, techniques which have the capacity for universal consistency such as nearest neighbor.

1. **Local averaging techniques:** these techniques involve averaging over local data points, that is, estimating Y_{n+1} by averaging Y_i for all X_i which are near X_{n+1}. Examples of these technique include:
   (a) **Nearest neighbor:** as discussed above
   (b) **Histograms:** where the X values are partitioned into buckets and the Y's from the bucket of a particular X are used to estimate the next Y.
   (c) **Kernel smoothing:** where an average of the Y's weighted by a function decaying in the distance of the corresponding X is used to estimate the next Y.
2. **Local modeling:** these techniques involve fitting a finitely parameterized model, such as a polynomial, to local data points.
3. **Global modeling:** these techniques build parametric models which are fit globally, that is, to all the data. These techniques can be made nonparametric by allowing piecewise models. Example of these techniques include:
   (a) Piecewise linear regression
   (b) Piecewise polynomial regression
   (c) Neural networks

4. **Penalized modeling:** these techniques fit parametric models with differing number of parameters and minimize a function which includes a penalty term for the number of parameters. These techniques are often called **regularization**.

Techniques from all of these paradigms can be shown to exhibit universal consistency. We have seen this for k-NN, a local averaging result. We now present a universal consistency result for global modeling. In order to describe this result, we first introduce the concept of denseness:

Definition 4.4.1. *Let $\mathbb{P}$ be a probability measure. A set of functions $\mathfrak{F}$ is dense in $L_2(\mathbb{P})$ if for any g such that $E\left[g^2(x)\right]$ exists and is finite and for any $\epsilon > 0$, there is an $f \in \mathfrak{F}$ such that:*

$$E\left[(f-g)^2\right] < \epsilon$$

Denseness of a set of functions means that the set of functions comes arbitrarily close to any well-behaved function. We will discuss denseness in more detail later. The universal consistency of global modeling estimates is:

Theorem 4.4.1. *Let $\phi_1, \phi_2, \ldots$ be bounded functions such that:*

$$\bigcup_{k=1}^{\infty} \left\{ \sum_{i=1}^{k} a_i \phi_i : a_1, a_2, \ldots, a_k \in \mathbb{R} \right\}$$

is dense in $L_2(\mathbb{P})$ for a probability measure $\mathbb{P}$ on $\mathbb{R}^d$. For each n, if we choose $a_1, a_2, \ldots, a_{k_n}$ to minimize:

$$\frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{k_n} a_i \phi_i(X_i) - Y_i \right)^2$$

such that:

$$\sum_{i=1}^{k_n} |a_i| < \beta_n$$

where k_n and β_n satisfy:

$$\begin{aligned}
&\lim_{n\to\infty} k_n = \infty \\
&\lim_{n\to\infty} \beta_n = \infty \\
&\lim_{n\to\infty} \frac{k_n \beta_n^4 \log(\beta_n)}{n} = 0 \\
&\lim_{n\to\infty} \frac{\beta_n}{n^{1-\delta}} = 0
\end{aligned}$$

for some $\delta > 0$*, then the estimate* $\sum_{i=1}^{k_n} a_i \phi_i$ *is universally consistent.*

Chapter 5

Rates of Convergence

5.1 Naive Rates of Convergence Can be Arbitrarily Slow

As discussed, there are several different machine learning techniques which demonstrate universal consistency, meaning that their predictions converge to the best possible prediction. However, it turns out that this convergence can be arbitrarily slow.

Theorem 5.1.1. *Let* $(X_1, Y_1), (X_2, Y_2), \ldots$ *be random variables with* $X \in [0,1]$ *and* $Y \in \{0,1\}$ *and* X_i *IID with a uniform distribution. Let* $\hat{Y}_1, \hat{Y}_2, \ldots$ *be any sequence of estimates such that* $\hat{Y}_{n+1}$ *is measurable with respect to* $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n), X_{n+1}$*. For any sequence* a_n *such that* $\lim_{n\to\infty} a_n = 0$*, there is a function* $m : [0,1] \to \{0,1\}$ *such that, for* $Y_i = m_i(X)$*, we have:*

$$\limsup_{n\to\infty} \frac{E\left[\left(\hat{Y}_n - m(X_n)\right)^2\right]}{a_n} \geq 1$$

We make some observations about this result:

1. This result means that the variance of any estimator can be guaranteed to go down at any rate, even a slow rate such as $\frac{1}{\log(\log(n))}$ which would be around 0.3 after a trillion observations. It will be more than this rate an infinite number of times. The same is true of any rate, even rates that go to 0 more slowly.

2. For this result, there is no noise since $Y_i = m(X_i)$.

3. Note that, in contrast to Theorem 4.1.1, this result is for discontinuous functions since $m : \mathbb{R} \to \{0,1\}$.

5.2 Metric Spaces, Norms and Lipschitz Functions

In this section, we introduce some conditions on the function m which will allow us to guarantee certain rates of convergence. These conditions involve more than continuity. We must have a bound on the amount that the function, or its derivatives, move when the argument moves by a certain amount. In order to present this, we will first need to discuss a mathematical notion of distance, called a metric.

Definition 5.2.1. *A* **metric** *on a set S is a function $d : S \times S \to [0, \infty)$ such that:*

1. **Positive definite:** *for all $x, y \in S$, we have $d(x, y) = 0 \Rightarrow x = y$.*
2. **Symmetry:** *for all $x, y \in S$, we have $d(x, y) = d(y, x)$.*
3. **Triangle inequality:** *for all $x, y, z \in S$, we have $d(x, y) + d(y, z) \geq d(x, z)$.*

The pair (S, d) where S is a set and d is a metric on S is called a **metric space**. *We often write S in place of (S, d).*

The value $d(x, y)$ is thought of as the distance between x and y. The nearest neighbor method can be applied to an arbitrary metric space: the distance between points defined by the metric can be used to find which neighbors are nearest. Metrics will also later play an important part in determining how "large" a set of functions in order to be able to approximate them.

We now give some examples of metrics:

Example 5.2.1. *For any set S, the discrete metric is given by:*

$$d(x, y) = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{otherwise} \end{cases}$$

Example 5.2.2. *For the set $S = \mathbb{R}$, a commonly used metric is given by:*

$$d(x, y) = |x - y|$$

The conditions required for a metric are easily verified.

Example 5.2.3. *A generalization of the metric given above is, for the set $S = \mathbb{R}^n$ and for any $p \in [1, \infty)$:*

$$d(x, y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}}$$

This is called the Euclidean metric. We will always use this metric for $\mathbb{R}^n$.

The metrics given in Example 5.2.3 have the additional properties that they make sense with respect to vector space operations, namely, vector addition and scalar multiplication:

1. **Translation invariance:** $d(x+z, y+z) = d(x, y)$ for all $x, y, z \in \mathbb{R}^n$.

2. **Positive homogeneity:** $d(\alpha x, \alpha y) = |\alpha|\, d(x, y)$ for all $\alpha \in \mathbb{R}$ and $x, y \in \mathbb{R}^n$.

For metrics on vector spaces which have these properties, the function $\|x\| = d(x, 0)$ is called a norm. The definition of a norm, without reference to a metric space, is:

Definition 5.2.2. *A* **norm** *is a function from a vector space V into $[0, \infty)$, which, for $v \in V$, is written as $\|v\|$, such that:*

1. **Positive definiteness:** *for all $v \in V$, we have $\|v\| = 0 \Rightarrow v = 0$.*

2. **Subadditivity:** *for all $x, y \in V$, we have $\|x + y\| \leq \|x\| + \|y\|$.*

3. **Positive homogeneity:** *for all $\alpha \in \mathbb{R}$ and $v \in V$, we have $\|\alpha x\| = |\alpha|\, \|x\|$.*

A pair $(V, \|\cdot\|)$ consisting of a vector space V and a norm $\|\cdot\|$ is called a normed space. Often the norm will be supressed in the notation.

For any norm $\|x\|$, the function $d(x, y) = \|x - y\|$ defined a metric which is translation invariant and positive homogeneous.

We can now define Lipschitz functions:

Definition 5.2.3. *A function $f : X \to Y$, between two metric spaces X and Y, is called Lipschitz with constant $L \in [0, \infty)$ if for all $x_1, x_2 \in X$:*

$$d(f(x_1), f(x_2)) \leq L d(x_1, x_2)$$

Some properties of Lipschitz functions are:

1. **Function with bounded derivatives are Lipschitz:** if a function $f : \mathbb{R} \to \mathbb{R}$ is differentiable and has derivatives bounded by L:

$$\sup_x |f'(x)| \leq L$$

then f is Lipschitz with constant L. There is a generalization of this to multivariate functions.

2. **Lipschitz functions have bounded derivatives almost everywhere:** if $f : \mathbb{R} \to \mathbb{R}$ is Lipschitz with constant L, then the set on which f is not differentiable has "length" 0. More precisely:

$$\mu(f \text{ is not differentiable}) = 0$$

where μ is the Lebesgue measure, that is, the measure defined by the length function on intervals:

$$\mu([a, b]) = |b - a|$$

Some examples of functions which are or are not Lipschtiz follow:

Example 5.2.4. *Every affine function, $f : \mathbb{R}^m \to \mathbb{R}^n$, defined by:*

$$f(x) = Ax + b$$

for a matrix $A \in \mathbb{R}^{n \times m}$ and vector $b \in \mathbb{R}^n$, is Lipschitz with constant $L \leq \max_{i,j} |A_{i,j}|$.

Example 5.2.5. *Every polynomial:*

$$p(x) = \sum_{i=0}^{k} a_i x^i$$

is Lipschitz on any interval $[a, b]$ with $-\infty < a < b < \infty$.

Example 5.2.6. *The absolute value function $f(x) = |x|$ is Lipschitz with constant 1.*

Example 5.2.7. *The square root function $f(x) = \sqrt{x}$ is not Lipschitz on $[0, \infty)$. In particular, the derivative goes to ∞ at 0. Note that this function is Lipschitz on any interval $[a, \infty]$ for $a > 0$.*

Finally, we define Hölder functions:

Definition 5.2.4. *For normed spaces X and Y, a function $f : X \to Y$ is called (p, C)-Hölder if all its* pth *partial derivatives are Lipschitz continuous with constant C, that is, for $p_1, p_2, \ldots, p_d \in \mathbb{N}$ such that $\sum_i p_i = p$:*

$$\left\| \frac{\partial^p f}{\partial x_1^{p_1} \partial x_2^{p_2} \ldots \partial x_d^{p_d}}(x) - \frac{\partial^p f}{\partial x_1^{p_1} \partial x_2^{p_2} \ldots \partial x_d^{p_d}}(y) \right\| \leq C\|x - y\| \tag{5.1}$$

We will write $f \in \mathcal{F}^{(p,C)}$ for the set of (p, C)-Hölder functions. Note that $(1, C)$-Hölder functions are Lipschitz functions.

Note that in the mathematical literature, the definition of Hölder allows for any $C < \infty$ and they also place an exponent on the right hand side of Inequality (5.1).

5.3 Lower Bounds on Rates of Convergence for Lipschitz Functions and the Curse of Dimensionality

We will now show negative results that no learning method can perform better than certain lower bounds. For (p, C)-Hölder functions from $\mathbb{R}^n$ to $\mathbb{R}$, it can be shown that no learning method can perform better than a bound given by the following result from [LGW10]:

Theorem 5.3.1. *Let $(X_1, Y_1), (X_2, Y_2), \ldots$ be IID and let $m(x) = E[Y_i | X_i = x]$. For any fixed p and C:*

$$\liminf_{n\to\infty} \inf_{\hat{Y}_n} \sup_{m \in \mathcal{F}^{(p,C)}} \frac{E\left[\left(\hat{Y}_n - m(X_n)\right)^2\right]}{C^{\frac{2d}{2p+d}} n^{-\frac{2p}{2p+d}}} \geq C_1 > 0$$

for some fixed constant C_1.

We make the following comments on this result:

1. We can unpack the statement of the theorem into the following statement. For all $\epsilon > 0$ and for all $N \in \mathbb{N}$, there is an $n \geq N$ such that for all $\hat{Y}_n$, there is an $m \in \mathcal{F}^{(p,C)}$ such that:

$$E\left[\left(\hat{Y}_n - m(X_n)\right)^2\right] \geq (C_1 - \epsilon) C^{\frac{2d}{2p+d}} n^{-\frac{2p}{2p+d}}$$

2. Let's see what this theorem says about how many observations it takes to achieve some variance V:

$$\begin{aligned} V &\geq (C_1 - \epsilon) C^{\frac{2d}{2p+d}} n^{-\frac{2p}{2p+d}} \\ n^{\frac{2p}{2p+d}} &\geq \frac{(C_1 - \epsilon) C^{\frac{2d}{2p+d}}}{V} \\ n &\geq \left(\frac{(C_1 - \epsilon) C^{\frac{2d}{2p+d}}}{V}\right)^{\frac{2p+d}{2p}} \end{aligned}$$

3. Note that this is exponential in d, the dimensionality of the problem. While the constant, C_1, is not made explicit in [LGW10], let's suppose that we wish to achieve a variance $V = \frac{\beta}{2}$ and we have 100 feature dimensions. This requires more than 2^{100} observations which is approximately 10^{30}, an unrealistic number with any foreseeable technology. If we wish to achieve $\frac{1}{2}$ that variance, it would require this amount for 50 feature dimensions.

This is the curse of dimensionality. It is important to keep the number of feature dimensions to a minimum in learning problems when one knows nothing about the function.

4. Note that it is also exponential in C, the Lipschitz constant, which corresponds with the upper bound on the derivatives of the function. Hence, for the example above, if we wish to regress a function which could be moving around twice as fast (twice the derivative), it would require 2^{100} times as many observations for 100 feature dimensions.

5. Note that, if we can accept regressing functions which are highly differentiable, it can mitigate the curse of dimensionality. For example, if we assume that our function is 50-times differentiable, that might allow us to estimate regression functions with 100 feature dimensions without requiring an excessive number of observations. It is difficult to determine whether functions do or do not have this amount of differentiability. For example, an arbitarily differentiable function could be made to go through any set of pairs of x and y coordinates as long as all the x coordinates are separated by some minimum distance.

5.4 Upper Bounds for Nearest Neighbor

It turns out that several common methods have been shown to achieve the lower bounds within a constant factor. In this section, we present the result, from [LGW10], that provides an upper bound for the nearest neighbor mean squared error:

Theorem 5.4.1. *Suppose that $X_i \in \mathbb{R}^d$ for $d \geq 3$ and let $(X_1, Y_1), (X_2, Y_2), \ldots$ be IID. Define $m(x) = E[Y_n | X_n = x]$. If the following conditions hold:*

1. $m \in \mathcal{F}^{(1,C)}$
2. *X is bounded, that is, there is some $M \in [0, \infty)$ such that $|X_i| \leq M$*
3. $$E\Big[(Y_i - m(X_i))^2 \Big| X_i \Big] \leq \sigma^2$$
4. *$\hat{Y}_{n+1}$ be the nearest neighbor estimate of Y_{n+1} for $k_n = c' \left(\frac{\sigma^2}{C^2} \right)^{\frac{d}{2+d}} n^{\frac{2}{2+d}}$*

then there is a constant c'' such that for $d \geq 3$:

$$E\left[\left(\hat{Y}_{n+1} - m(X_{n+1}) \right)^2 \right] \leq c'' \sigma^{\frac{4}{d+2}} C^{\frac{2d}{2+d}} n^{-\frac{2}{d+2}}$$

Thus the nearest neighbor method achieves the lower bound given by Theorem 5.3.1 within a constant factor.

Note that nearest neighbor doesn't achieve the optimal rate of convergence given by Theorem 5.3.1 for class $\mathcal{F}^{(p,C)}$ with $p > 1$. This can be seen as coming from the fact that the nearest neighbor estimate has a discontinuity in the derivative as you move across a boundary where the neighbors change. There are ways of fixing this, such as using kernel smoothing (not to be confused with the kernel trick), which can be shown to achieve a similar guarantee for slightly more smooth functions and for local polynomial estimation, which can be shown to achieve a similar lower bound for arbitrarily smooth functions in 1 dimension. However, it is not known whether any methods could achieve similar results for arbitrarily smooth functions in n dimensions.

5.5 Projection Pursuit

In general, there is no way around the curse of dimensionality unless something is known about the function one is trying to predict. One source of information, as discussed above, is the smoothness of the function. However, another source of information is dimensionality reduction, that the function is dependent upon a smaller number of combinations of features. This is the assumption behind the projection pursuit method which we discuss here.

In projection pursuit, the model is given by:

$$Y_i = \sum_{j=1}^{K} m_j(\beta_j X_i) \tag{5.2}$$

where $X_i, \beta_j \in \mathbb{R}^d$ for all $i \in \{1, 2, \ldots, n\}$ and $j \in \{1, 2, \ldots, K\}$ and $m_j : \mathbb{R} \to \mathbb{R}$. This model can be estimated by using piecewise polynomial estimates for each of the m_j, see [LGW10] for details. In this case, the lower bound is given by the following result:

Theorem 5.5.1. *If $(X_1, Y_1), (X_2, Y_2), \ldots$ are IID and behave according to Equation (5.2) with the following conditions:*

1. *$X_i \in [0, 1]^d$*
2. *Y_i is bounded, that is, there is some $L \in [0, \infty)$ such that $|Y_i| \leq L$*
3. *$m_j \in \mathcal{F}^{p,C}$*

then there is a constant c, independent of n and C such that:

$$E\left[\left(\hat{Y}_{n+1} - m(X_{n+1})\right)^2\right] \leq cC^{\frac{2}{2p+1}} \left(\frac{\log(n)}{n}\right)^{\frac{2p}{2p+1}}$$

Note that, unlike the result presented about nearest neighbor, the result above does not come within a constant factor of the lower bound, Theorem 5.3.1. There is an extra factor of $\log(n)^{\frac{2p}{2p+1}}$. However, the result does avoid the curse of dimensionality, assuming one knows that the function behaves according to Equation (5.2) which is, albeit, difficult to determine.

Chapter 6

Empirical Risk Minimization

6.1 Statistical Decision Theory

We have seen that we often wish to optimize the expected value of some function with respect to a parameter. We now introduce statistical decision theory to generalize this.

A **statistical decision** problem consists of the following elements:

1. An unknown state of the world, θ, which may be a random variable, from a set Θ
2. An action that we can choose from a set A
3. A loss function $L : \Theta \times A \to \mathbb{R}$
4. The observations, a random vector $X \in \mathbb{R}^n$

If the state of the world turns out to be $\theta \in \Theta$ and we had chosen action a then we lose $L(\theta, a)$. We are allowed to choose the action after observing the observations so we allow it to be a function of the observations, $f^* : \mathbb{R}^n \to A$. We wish to minimize our expected loss:

$$\min_{f^*} E[L(Y, f^*(X))] \tag{6.1}$$

Statistical decision theory encompasses all of the problems that we have discussed in this course:

Example 6.1.1. *Let us consider choosing a portfolio of stocks w_t where the unknown state of the world is the returns R_{t+1} with loss function:*

$$L(R_{t+1}, w_t) = -\log\big(1 + w_t^T R_t\big) \tag{6.2}$$

In this case, we are optimizing Equation (3.10) and the optimum is the growth optimal portfolio. Note the negative sign in Equation (6.2) is because it represents a "value" to be maximized rather than a "loss" to be minimized. In order to fit this into the framework of statistical decision theory, a negative sign is required. More generally, we could choose a utility function $u : \mathbb{R} \to \mathbb{R}$ in which case the loss function would be:

$$L(R_{t+1}, w_t) = -u\big(1 + w_t^T R_{t+1}\big)$$

Example 6.1.2. *If we wish to choose an action $\hat{Y}_{n+1}$ to estimate a random variable Y_{n+1} based on another random variable X_{n+1} with squared error loss:*

$$L\Big(\hat{Y}_{n+1}, Y_{n+1}\Big) = \Big(\hat{Y}_{n+1} - Y_{n+1}\Big)^2$$

In this case, the expected loss is given by Equation (4.1) and, as shown there, the optimum is given by:

$$E[\, Y_{n+1} |\, X_{n+1}]$$

Example 6.1.3. *In the classification problem, we have feature vectors $X_1, X_2, \ldots$ which are in classes $Y_1, Y_2, \ldots$. For example, X_i might be an image and Y_i might be in the set $\{dog, cat, \ldots\}$. In this case, the unknown state of the world is the class, Y_{n+1} of some new observation X_{n+1}. The action is an estimate of the class $\hat{Y}_{n+1}$. A reasonable loss function would be:*

$$L\Big(Y_{n+1}, \hat{Y}_{n+1}\Big) = \mathbb{1}_{Y_{n+1} = \hat{Y}_{n+1}}$$

We minimize the expected loss which is the classification error:

$$E\Big[\mathbb{1}_{Y_{n+1} = \hat{Y}_{n+1}}\Big] = \mathbb{P}\Big(Y_{n+1} = \hat{Y}_{n+1}\Big)$$

We consider the case where we have IID observations $(X_1, Y_1), (X_2, Y_2), \ldots$ and wish to choose an estimate $\hat{Y}_{n+1} = \hat{f}(X_{n+1})$. If the distributions of the state of nature θ and the observations X are known, then we can optimize Equation (6.1). However, in the cases of interest here, we wish to choose rules which work for a variety of distributions. When the distribution is unknown, the expectation in Equation (6.1) can't be directly evaluated. A reasonable alternative is to optimize the sample average:

$$\min_{\hat{f}} \frac{1}{n} \sum_{i=1}^{n} L\Big(Y_i, \hat{f}(X_i)\Big) \tag{6.3}$$

This is called **empirical risk minimization** and is widely used, for example, in linear least squares, nonlinear least squares and neural network modeling.

We now investigate whether the optimum of Equation (6.3) is not far from the optimum of Equation (6.1). We first introduce $\hat{E}_n$ as a shorthand for the sample average:

$$\hat{E_n}[X] = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Assuming Equation (6.3) and Equation (6.1) have solutions f^* and $\hat{f}$, respectively:

$$E\Big[L\Big(Y_i, \hat{f}(X_i)\Big)\Big] - E[L(Y, f^*(X))] \tag{6.4}$$

$$= \Big(E\Big[L\Big(Y_i, \hat{f}(X_i)\Big)\Big] - \hat{E_n}\Big[L\Big(Y, \hat{f}(X)\Big)\Big]\Big) \tag{6.5}$$

$$+ \Big(\hat{E_n}\Big[L\Big(Y, \hat{f}(X)\Big)\Big] - \hat{E_n}[L(Y, f^*(X))]\Big) \tag{6.6}$$

$$+ \Big(\hat{E_n}[L(Y, f^*(X))] - E[L(Y, f^*(X))]\Big) \tag{6.7}$$

We examine the 3 terms from the above in reverse order:

- The Term (6.7) is the difference between an empirical expectation and the corresponding expectation. Hence, it is a sum of IID random variables which allows us to bring many tools to bear. For example, by the law of large numbers, converges to 0 with probability 1. In the next section, we'll discuss more explicit bounds on the probability that this difference exceeds a given level for a given number of observations. This will help us demonstrate results on sample-size complexity, that is, upper bounds on how many samples are required to reduce Equation (6.4 to a given level.

- The Term (6.6) is the difference between an empirical expectation evaluated at its minimum and the same empirical expectation evaluated at another point and so is negative

- Note that, in the Term (6.5), the law of large numbers does not apply because $\hat{f}$ is chosen based on the data according to Equation (6.3). We can, bound this term by taking the supremum of the difference over all functions:

$$\begin{aligned} & E\Big[L\Big(Y_i, \hat{f}(X_i)\Big)\Big] - \hat{E_n}\Big[L\Big(Y, \hat{f}(X)\Big)\Big] \\ \leq & \sup_{f \in \mathcal{F}} \Big| E[L(Y_i, f(X_i))] - \hat{E}_n[L(Y, f(X))] \Big| \end{aligned} \tag{6.8}$$

where $\mathcal{F}$ is the set of all functions we are using as estimators. In a subsequent section, we'll discuss how to bound this quantity using covering numbers which will help us demonstrate results on sample-size complexity.

6.2 Exponential Tail Inequalities

While the law of large numbers tells us that the empirical expectation converges to the expectation, it doesn't tell us how fast it converges. Under some sets of conditions, the probability that the empirical expectation of IID random variables diverges by more than a given amount from the expectation converges exponentially to 0 with the sample size. One such set of conditions is boundedness of the random variables in which case, Hoeffding's inequality gives us an upper bound:

Theorem 6.2.1. Hoeffding's inequality: *Let $X_1, X_2, \ldots, X_n$ be IID random variables with bounds $X_i \in [a, b]$. For any $\epsilon > 0$:*

$$\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^{n} X_i - E[X_1] \geq \epsilon \right) \leq \exp\left(-2n \left(\frac{\epsilon}{b-a} \right)^2 \right)$$

We make the following comments on Hoeffding's inequality:

1. Unlike the law of large numbers or central limit theorem, Hoeffding's inequality is not asymptotic: it holds for any finite sequence of random variables. This makes it extremely useful.

2. The assumption of boundedness is much stronger than the conditions for the law of large numbers, central limit theorem or law of the iterated logarithm to hold. These results hold for every sequence of bounded random variables but there are sequences of unbounded random variables for which the asymptotic results hold.

3. Hoeffding's inequality can be made 2 sided:

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - E[X_1]\right| \geq \epsilon\right) \\
&= \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - E[X_1] \geq \epsilon \text{ or } -\left(\frac{1}{n}\sum_{i=1}^{n} X_i - E[X_1]\right) \geq \epsilon\right) \\
&\leq \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - E[X_1] \geq \epsilon\right) + \mathbb{P}\left(\left(\frac{1}{n}\sum_{i=1}^{n} (-X_i) - E[-X_1]\right) \geq \epsilon\right) \\
&\leq 2\exp\left(-2n\left(\frac{\epsilon}{b-a}\right)^2\right)
\end{aligned}$$

We now show how to use Hoeffding's inequality to prove the law of large numbers for bounded random variables. In order to do this, we need to introduce the Borel-Cantelli lemma:

Lemma 6.2.1. Borel-Cantelli lemma: *If $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space and $A_1, A_2, \ldots \in \mathcal{F}$ are such that:*

$$\sum_{i=1}^{\infty} \mathbb{P}(A_i) < \infty$$

Then:

$$\mathbb{P}(A_i \text{ infinitely often}) = 0$$

Proof. We have that:

$$\begin{aligned}
\mathbb{P}(A_i \text{ infinitely often}) &= \mathbb{P}\left(\bigcap_{n=1}^{\infty}\bigcup_{m=n}^{\infty} A_m\right) \leq \mathbb{P}\left(\bigcup_{m=N}^{\infty} A_m\right) \\
&\leq \sum_{m=N}^{\infty} \mathbb{P}(A_m) \qquad (6.9)
\end{aligned}$$

for any N. However, the right hand side of Inequality (6.9) converges to 0. □

We can now show the law of large numbers for bounded IID random variables:

Theorem 6.2.2. *If $X_1, X_2, \ldots$ are IID random variables with $X_i \in [a, b]$, then:*

$$\mathbb{P}\Big(\lim_{n\to\infty} \hat{E}_n[X] = E[X]\Big) = 1$$

Proof. Fix any $\epsilon > 0$. Let $A_{n,\epsilon} = \left\{\left|\hat{E}_n[X] - E[X]\right| \geq \epsilon\right\}$. We have:

$$\begin{aligned}\sum_{n=1}^{\infty} \mathbb{P}(A_{n,\epsilon}) &= \sum_{n=1}^{\infty} \mathbb{P}\Big(\left|\hat{E}_n[X] - E[X]\right| \geq \epsilon\Big) \\ &\leq \sum_{n=1}^{\infty} 2\exp\left(-2n\left(\frac{\epsilon}{b-a}\right)^2\right) = \frac{2\exp\left(2\left(\frac{\epsilon}{b-a}\right)^2\right)}{1-\exp\left(2\left(\frac{\epsilon}{b-a}\right)^2\right)} < \infty\end{aligned}$$

where we have used Hoeffding's inequality for the first inequality. Hence, by the Borel-Cantelli lemma, $A_{n,\epsilon}$ only happens finitely often with probability 1. Since ϵ was arbitrary, the result holds. □

Hoeffding's inequality applies directly to Term (6.7) when the loss function L is bounded, such as is the case for the classification problem or mean squared error when Y is bounded. It will also help bound Term (6.5) in the next section in combination with the covering number bounds.

6.3 Covering Numbers and Bounds

Note that Hoeffding's inequality immediately allows us to bound Inequality (6.8) when the set of functions is finite:

$$\begin{aligned}&\mathbb{P}\left(\sup_{f\in\{f_1,f_2,\ldots,f_m\}} \left|E[L(Y,f(X))] - \hat{E}_n[L(Y,f(X))]\right| \geq \epsilon\right) \\ &\leq \sum_{i=1}^{m} \mathbb{P}\Big(\left|E[L(Y,f_i(X))] - \hat{E}_n[L(Y,f_i(X))]\right| \geq \epsilon\Big) \\ &\leq 2m\exp\left(-2n\left(\frac{\epsilon}{b-a}\right)^2\right) \qquad (6.10)\end{aligned}$$

Infinite sets of functions can often be approximated arbitrarily closely by finite sets of functions. In this section, we introduce covering numbers in metric spaces as a way to make such approximations.

Definition 6.3.1. *Let (X,d) be a metric space. The* **ball** *of radius ϵ centered at x is the set, $B_\epsilon(x)$ of points within ϵ of x:*

$$B_\epsilon(x) = \{x' : d(x,x') < \epsilon\}$$

A set $S \subseteq X$ is an **ϵ-cover** *of X if $X \subseteq \bigcup_{s\in S} B_\epsilon(s)$. The* **$\epsilon$-covering number** *of X with respect to d, written $\mathcal{N}(\epsilon, X, d)$, is the smallest number of elements in an ϵ-cover of X.*

We give some examples of these definitions:

Example 6.3.1. *Consider $X \subseteq \mathbb{R}^k$ and let $\|\cdot\|$ be one of the norms defined in Example 5.2.3. It can be shown that:*

$$\left(\frac{1}{\epsilon}\right)^k \frac{vol(X)}{vol(B_1(0))} \leq \mathcal{N}(\epsilon, X, \|\cdot\|) \leq \left(\frac{3}{\epsilon}\right)^k \frac{vol(X)}{B_1(0)}$$

where $vol(X)$ is the volume of the set X. Hence, the covering numbers are exponential in the dimension of the space. In fact, this is a defining characteristic of dimension.

The next result demonstrates how we can use the covering numbers to bound the error of empirical risk minimization. First, we introduce a norm on function spaces:

Definition 6.3.2. *The L_∞ norm, written $\|\cdot\|$ on a set of functions, $\mathcal{F}$, such that $f : S \to \mathbb{R}$ for some set S is:*

$$\|f\|_\infty = \sup_{s \in S} |f(s)|$$

The following result bounds the error in empirical risk minimization:

Theorem 6.3.1. *For each n, let $\mathcal{G}_n$ be a set of functions $g : \mathbb{R}^d \to [0, B]$.*

$$\mathbb{P}\left(\sup_{g \in \mathcal{G}_n} \left|\hat{E}_n[g(X)] - E[g(X)]\right| \geq \epsilon\right) \leq 2\mathcal{N}\left(\frac{\epsilon}{3}, \mathcal{G}_n, \|\cdot\|_\infty\right) \exp\left(-2n\left(\frac{\epsilon}{3B}\right)^2\right)$$

Furthermore, if:

$$\sum_{n=1}^{\infty} \mathcal{N}\left(\frac{\epsilon}{3}, \mathcal{G}_n, \|\cdot\|_\infty\right) \exp\left(-2n\left(\frac{\epsilon}{3B}\right)^2\right) < \infty \tag{6.11}$$

then:

$$\mathbb{P}\left(\lim_{n\to\infty} \sup_{g \in \mathcal{G}_n} \left|\hat{E}_n[g(X)] - E[g(X)]\right| = 0\right) = 1 \tag{6.12}$$

Proof. Let $\mathcal{G}_{n,\frac{\epsilon}{3}}$ be an $\frac{\epsilon}{3}$ covering of $\mathcal{G}_n$. For $g \in \mathcal{G}_n$, let $\tilde{g} \in \mathcal{G}_{n,\epsilon}$ be such that $\|g - \tilde{g}\| < \frac{\epsilon}{3}$:

$$\begin{aligned}
&\left|\hat{E}_n[g(X)] - E[g(X)]\right| \\
&\quad\leq \left|\hat{E}_n[g(X)] - \hat{E}_n[\tilde{g}(X)]\right| + \left|\hat{E}_n[\tilde{g}(X)] - E[\tilde{g}(X)]\right| + |E[\tilde{g}(X)] - E[g(X)]| \\
&\quad\leq \|g - \tilde{g}\|_\infty + \left|\hat{E}_n[\tilde{g}(X)] - E[\tilde{g}(X)]\right| + \|\tilde{g} - g\|_\infty \\
&\quad\leq \frac{2\epsilon}{3} + \left|\hat{E}_n[\tilde{g}(X)] - E[\tilde{g}(X)]\right|
\end{aligned}$$

Using Inequality (6.10) yields the first result. Applying the Borel-Cantelli lemma yields the second result. □

A few comments on this result:

1. Note that Equation (6.12) has a limit of a supremum. We would like to point out that this is not a limit supremum as defined in Equation (3.15) as the supremum is not over the natural numbers but over a set.

2. We can find a bound on the sample size required to achieve a probability p of exceeding some $\epsilon > 0$ as follows:

$$
\begin{aligned}
2\mathcal{N}\Big(\frac{\epsilon}{3}, \mathcal{G}_n, \|\cdot\|_\infty\Big) \exp\bigg(-n\left(\frac{\epsilon}{3B}\right)^2\bigg) &\leq p \\
\exp\bigg(-n\left(\frac{\epsilon}{3B}\right)^2\bigg) &\leq \frac{p}{2\mathcal{N}\big(\frac{\epsilon}{3}, \mathcal{G}_n, \|\cdot\|_\infty\big)} \\
-n\left(\frac{\epsilon}{3B}\right)^2 &\leq \log\left(\frac{p}{2\mathcal{N}\big(\frac{\epsilon}{3}, \mathcal{G}_n, \|\cdot\|_\infty\big)}\right) \\
n &\geq \frac{\log\big(2\mathcal{N}\big(\frac{\epsilon}{3}, \mathcal{G}_n, \|\cdot\|_\infty\big)\big) - \log(p)}{\big(\frac{\epsilon}{3B}\big)^2} \qquad (6.13) \\
& \qquad (6.14)
\end{aligned}
$$

3. Notice that the term containing the covering number $\mathcal{N}\big(\frac{\epsilon}{3}, \mathcal{G}_n, \|\cdot\|_\infty\big)$ in Inequality (6.13) is logarithmic in it. As mentioned in Example 6.3.1, the covering numbers tend to be exponential in the dimension of the set. Hence, Inequality (6.13) is affine in the dimensionality of the function space.

4. Many function spaces are infinite dimensional and so would have infinite covering numbers which would not produce a useful bound. Hence, in practice, we choose $\mathcal{G}_n$ to be finite dimensional and allow it to grow. If it grows slowly enough so that Inequality (6.11) holds, the error will still converge. In particular, we would like Inequality (6.11) to hold. This will happen if, for example:

$$
\mathcal{N}\Big(\frac{\epsilon}{3}, \mathcal{G}_n, \|\cdot\|_\infty\Big) \leq \left(\frac{c}{\epsilon}\right)^{d_n}
$$

such that $\lim_{n\to\infty} \frac{d_n}{n} = 0$. Here, d_n corresponds with the dimensionality of the set $\mathcal{G}_n$. This can grow to ∞ as long as it grows more slowly than n.

Theorem 6.3.2. *Let $f_1, f_2, \ldots, f_k$ be functions such that $\|f\|_\infty \leq M$ and let $\mathcal{F}$ be the set of functions defined by the following:*

$$\mathcal{F} = \left\{ \sum_{j=1}^{k} a_j f_j(x) : a_1, a_2, \ldots, a_k \in [0, R] \right\}$$

We have:

$$\mathcal{N}(\epsilon, \mathcal{F}, \|\cdot\|) \leq \left(\frac{R}{\delta}\right)^k = \left(\frac{RMk}{2\epsilon}\right)^k \tag{6.15}$$

Proof. We discretize each a_i by within an amount $\delta = \frac{2\epsilon}{Mk}$:

$$\begin{aligned}
&\sup_x \left| \sum_{j=1}^{k} a_i f_i(x) - \sum_{j=1}^{k} \delta \left\lfloor \frac{a_i}{\delta} + \frac{1}{2} \right\rfloor f_i(x) \right| \\
&= \sup_x \left| \sum_{j=1}^{k} \left(a_i - \delta \left\lfloor \frac{a_i}{\delta} + \frac{1}{2} \right\rfloor \right) f_i(x) \right| \\
&\leq \sup_x \sum_{j=1}^{k} \left| a_i - \delta \left\lfloor \frac{a_i}{\delta} + \frac{1}{2} \right\rfloor \right| |f_i(x)| \\
&\leq \sum_{j=1}^{k} \left| \delta \left(\frac{a_i}{\delta} - \left\lfloor \frac{a_i}{\delta} + \frac{1}{2} \right\rfloor \right) \right| \sup_x |f_i(x)| \\
&\leq \sum_{j=1}^{k} \left| \delta \frac{1}{2} \right| M = k \frac{\delta}{2} M = k \frac{2\epsilon}{2Mk} M = \epsilon
\end{aligned}$$

If we allow each a_i to span the entire real line, then there are an infinite number of such points. Hence, we choose each $a_i \in [0, R]$, there will be $\frac{R}{\delta}$ possible values for each a_i for a total of:

$$\mathcal{N}(\epsilon, \mathcal{F}, \|\cdot\|) \leq \left(\frac{R}{\delta}\right)^k = \left(\frac{RMk}{2\epsilon}\right)^k$$

□

Example 6.3.2. *As an example of the use of Inequality (6.13) in conjunction with Inequality (6.15), consider the following setup:*

1. *We start with a* basic spline function *from which we will generate other functions:*

$$\Lambda(x) = \begin{cases} 0 & \text{if } x < -1 \\ 1 + x & \text{if } x \in [-1, 0) \\ 1 - x & \text{if } x \in [0, 1) \\ 0 & \text{if } x \geq 1 \end{cases}$$

2. *For this exercise, we assume* $X_i \in [0, 1]$ *and* $Y_i \in [0, 1]$.

3. *We choose a parameters* γ *that determines the precision along the* x*-axis of our model*

4. *Define the set of functions to take linear combinations of as follows:*

$$\mathcal{F}_\gamma = \left\{ \Lambda\left(\frac{x - i\gamma}{\gamma} \right) : i \in \left\{ 0, 1, \ldots, \left\lfloor \frac{1}{\gamma} \right\rfloor \right\} \right\}$$

   Note that $k = \left\lfloor \frac{1}{\gamma} \right\rfloor$. *Also* $\sup_x \left| \Lambda\left(\frac{x - i\gamma}{\gamma} \right) \right| \leq 1$ *for any* i *so that* $M = 1$.

5. *Since* $Y_i \in [0, 1]$, *we choose* $a_i \in [0, 1]$ *so that* $R = 1$.

6. *Putting this all together into Inequality (6.13) in conjunction with Inequality (6.15), we get:*

$$n \geq \frac{\left\lfloor \frac{1}{\gamma} \right\rfloor \log\left(\frac{3\left\lfloor \frac{1}{\gamma} \right\rfloor}{\epsilon} \right) - \log(p)}{\left(\frac{\epsilon}{3} \right)^2}$$

7. *Hence, if we wish to model within* 0.1 *on the* x*-axis so that* $\gamma = 0.1$ *and have a generalization error, that is, the difference between the error on our sample and the true error, of less than* 0.1 *with* 99.99% *certainty, then we need at least about 60,000 samples.*

Note that tighter bounds are possible. Our goal in this section was to demonstrate the techniques rather than derive the tightest possible bounds.

6.4 VC Dimension

For classification problems, the Vapnik-Chervonenkis (VC) dimension, which we discuss in this section, helps determine the sample size bounds of empirical risk minimization similarly to covering numbers do for regression problems.

Consider a set of IID samples $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ where $Y_i \in \{0, 1\}$. We restrict ourselves to binary classification since n-ary classification can be achieved via multiple binary classifications. We choose a set of possible

classification rules $\mathcal{F}$ of the form $f : X \to \{0, 1\}$. We choose $\hat{f}_n$ to maximize the empirical risk using the classification error (sometimes called 0-1 loss) as the loss function:

$$\hat{f}_n = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \hat{E}_n[L(Y, f(X))] = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \hat{E}_n\left[\mathbb{1}_{Y_i = f(X_i)}\right]$$

The expected value of the loss is given by:

$$E[L(Y, f(X))] = \mathbb{P}(Y = f(X))$$

We can identify any classification rule $f \in \mathcal{F}$ with the set of points that are labeled as 1, which we denote by H_f:

$$H_f = \{x : f(x) = 1\}$$

We call H_f a **hypothesis** and let $\mathcal{H}$ denote the set of all hypotheses:

$$\mathcal{H} = \{H_f : f \in \mathcal{F}\}$$

Since the sample is finite, there are only a finite number of possible hypotheses, namely, the power set $\mathcal{P}(\{X_1, X_2, \ldots, X_n\})$, even if our hypothesis set is infinite. The growth function measures how fast the number of hypothesis represented by a hypothesis set $\mathcal{H}$ grows:

Definition 6.4.1. *The* **growth function** *of hypothesis set $\mathcal{H}$ is defined as:*

$$m_{\mathcal{H}}(n) = \sup_{\{C : |C| = n\}} |\{H \cap C : H \in \mathcal{H}\}| \tag{6.16}$$

Since the growth function gives us the number of hypotheses relevant to the sample, we can use it in Inequality (6.10) to determine the probability that the generalization error exceeds some $\epsilon > 0$:

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} \left|\hat{E}_n[L(Y, f)] - E[L(Y, f)]\right| \geq \epsilon\right) \leq 2 m_{\mathcal{H}}(n) \exp\left(-2n\epsilon^2\right) \tag{6.17}$$

The VC dimension will help us bound the rate of growth of the growth function. Since the expression under the supremum the right hand side of Equation (6.16) is the size of a set of subsets of C where $|C| = n$, the growth function can be bounded by:

$$m_{\mathcal{H}}(n) \leq |\mathcal{P}(C)| = 2^n$$

A set is called shattered if the growth function achieves this bound on it:

Definition 6.4.2. *A set C is* **shattered** *by the hypothesis set $\mathcal{H}$ if:*

$$\{H \cap C : H \in \mathcal{H}\} = \mathcal{P}(C)$$

The VC dimension is the supremum of the sizes of shattered sets, or, equivalently, the supremum of the numbers where $m_{\mathcal{H}}(n)$ achieves its upper bound of 2^n:

Definition 6.4.3. *The* **VC dimension** *of a set of subsets, $\mathcal{H}$, is the supremum of the numbers at which the growth function is maximal:*

$$d(\mathcal{H}) = \sup\{n : m_{\mathcal{H}}(n) = 2^n\}$$

Before turning to how the VC dimension bounds the growth function, we give some examples of the calculation of VC dimension.

Example 6.4.1. *Consider the hypothesis set consisting of threshold functions:*

$$\mathcal{H} = \{(-\infty, a] : a \in \mathbb{R}\}$$

This hypothesis set can shatter any singleton set $\{c\}$ since $(-\infty, c-1] \cap \{c\} = \varnothing$ and $(-\infty, c+1] \cap \{c\} = \{c\}$ which are the only two subsets of $\{c\}$. On the other hand, $\mathcal{H}$ can't shatter any set $\{c_1, c_2\}$ with $c_1 < c_2$ since $c_2 \in H$ impliers $c_1 \in H$ for any $H \in \mathcal{H}$. Hence, the $d(\mathcal{H}) = 1$.

Example 6.4.2. *Consider the hypothesis set consisting of intervals:*

$$\mathcal{H} = \{[a, b] : a, b \in \mathbb{R}\}$$

This set can shatter any 2-element set $C = \{c_1, c_2\}$ with $c_1 < c_2$ as follows:

$$\begin{aligned}
[c_1 - 1, c_1 - 2] \cap C &= \varnothing \\
\left[c_1 - 1, \frac{c_1 + c_2}{2}\right] \cap C &= \{c_1\} \\
\left[\frac{c_1 + c_2}{2}, c_2 + 1\right] \cap C &= \{c_2\} \\
[c_1 - 1, c_2 + 1] \cap C &= C
\end{aligned}$$

However, $\mathcal{H}$ can't shatter any set of 3 elements because no hypothesis in it can contain the outer points without containing the inner one. Hence $d(\mathcal{H}) = 2$.

Example 6.4.3. *Consider the hypothesis set of axis parallel rectangles in* $\mathbb{R}^2$*:*

$$\mathcal{H} = \{[a_1, b_1] \times [a_2, b_2] : a_1, b_1, a_2, b_2 \in \mathbb{R}\}$$

It can be seen that this hypothesis set can shatter the set $C = \{(-1, 0), (1, 0), (0, -1), (0, 1)\}$*. Furthermore, consider any set of* 5 *elements,* $C = \{(a_1, b_1), (a_2, b_2), \ldots, (a_5, b_5)\}$*. Define* l, r, t, b *be the indices of the left most, right most, top most and bottom most points respectively, that is:*

$$\begin{aligned} l &= \operatorname*{argmin}_i a_i \\ r &= \operatorname*{argmax}_i a_i \\ t &= \operatorname*{argmin}_i b_i \\ b &= \operatorname*{argmax}_i b_i \end{aligned}$$

Any element of $\mathcal{H}$ *which contains* $\{(a_l, b_l), (a_r, b_r), (a_t, b_t), (a_b, b_b)\}$ *must also contain the remaining points, of which there is at least one. Hence* $d(\mathcal{H}) = 4$*.*

We now turn to how the VC dimension bounds the growth function. This is given by Sauer's lemma:

Lemma 6.4.1. Sauer's Lemma: *For any set of sets* $\mathcal{H}$ *with* $d(\mathcal{H}) = d$*:*

$$m_{\mathcal{H}}(n) \leq \sum_{j=0}^{d} \binom{n}{j} \leq \left(\frac{en}{d}\right)^d \tag{6.18}$$

From Inequalities (6.17) and (6.18), we get:

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} \left|\hat{E}_n[L(Y, f)] - E[L(Y, f)]\right| \geq \epsilon\right) \leq 2\left(\frac{en}{d}\right)^d \exp\left(-2n\epsilon^2\right)$$

Using this, it can be shown that the number of samples required to guarantee that the generalization error is less than ϵ with probability $1 - \delta$ is at most:

$$n \leq \frac{128d \log\left(\frac{64d}{\epsilon^2}\right) + 64d \log\left(\frac{e}{d}\right) + 16 \log\left(\frac{4}{\delta}\right)}{\epsilon^2}$$

Using other techniques, a lower bound is also available: the number of samples needed is at least:

$$n \geq \frac{\max\left(\frac{1}{2} \log\left(\frac{1}{4\delta}\right), 8d\right)}{\epsilon^2}$$

We illustrate these bounds in the following table:

d	ϵ	δ	lower bound	upper bound
1	0.1	0.1	800	124,482
1	0.01	0.01	800	128,166
1	0.01	0.1	80,000	18,342,827
2	0.1	0.1	1,600	251,934
2	0.1	0.01	1,600	255,618
2	0.01	0.1	160,000	36,982,661

Chapter 7

Nearest Neighbor on General Metric Spaces

7.1 Definitions

In this chapter, we investigate the nearest neighbor method on general metric spaces. Much of this chapter is taken from [KP95]. We have introduced random variables as measurable real-valued functions form a probability space. For this chapter we need the generalization to functions whose values are values are in a metric space. Fix a metric space $(\mathfrak{X}, d)$.

Definition 7.1.1. *A* **random element** *or* X**-valued random variable** *on the probability space* $(\Omega, \mathcal{F}, \mathbb{P})$ *is a function* $X : \Omega \to \mathfrak{X}$ *such that:*

$$X^{-1}(B_r(x)) \in \mathcal{F}$$

for all $x \in \mathfrak{X}$ *and* $r \in [0, \infty)$*. Note that* $B_r(x)$ *has been defined in Definition 6.3.1.*

In addition to generalizing the space of the X's to a metric space, we will also generalize the X's from IID to more general sequences of random variables. We will still want to express that Y_i is dependent only upon X_i. For this, we need the concept of conditional independence.

Definition 7.1.2. *Let* X, Y, Z *be random elements (whose values need not necessarily be in the same space). The* **conditional probability** *that* $X \in S_x$ *given* $Z = z$ *is defined as:*

$$\mathbb{P}(X \in S_x | Z = z) = E[\mathbb{1}_{S_X}(X) | Z = z]$$

where:

$$\mathbb{1}_{S_X}(X) = \begin{cases} 1 & \text{if } X \in S_X \\ 0 & \text{otherwise} \end{cases}$$

X is **conditionally independent** *of Y given Z if:*

$$\begin{aligned} &\mathbb{P}(\, X \in S_X, Y \in S_Y |\, Z = z) \\ &\quad = \quad \mathbb{P}(\, X \in S_X |\, Z = z)\, \mathbb{P}(\, Y \in S_Y |\, Z = z) \end{aligned}$$

One important property of conditional independence is that factors commute with conditional expectation, that is, if X and Y are conditionally independent given Z:

$$E[\, f(X)\, g(Y) |\, Z] = E[\, f(X) |\, Z]\, E[\, g(Y) |\, Z]$$

Let $X_1, X_2, \ldots$ be $\mathfrak{X}$-valued random variables. Let $X_{n:m} = (X_n, X_{n+1}, \ldots, X_m)$ and $X_{n:\infty} = (X_n, X_{n+1}, \ldots)$. We will assume that Y_i are random variables conditionally independent of $X_{1:i-1}$, $X_{i+1:\infty}$, $Y_{1:i-1}$ and $Y_{i+1:\infty}$ given X_i. We also assume that the distribution of Y_i is stationary, that is:

$$\mathbb{P}(\, Y_i \in S |\, X_i = x_i) = \mathbb{P}(\, Y_j \in S |\, X_j = x_j)$$

for any i and j. We will consider the squared error loss:

$$L\Big(Y_{n+1}, \hat{Y}_{n+1}\Big) = \Big(Y_{n+1} - \hat{Y}_{n+1}\Big)^2$$

As shown in Equation (4.1), when the distributions are known, the optimum is given by the conditional expectation which we denote by $m(x)$:

$$m(x) = E[\, Y_i |\, X_i = x]$$

The mean squared error of this optimum will be denoted by $\sigma(x)$:

$$\sigma^2(x) = E\Big[\, (Y_i - m(X_i))^2 \Big|\, X_i = x\Big]$$

We will assume that m and σ^2 are Hölder continuous. We had previously defined Hölder continuous for norms instead of metrics. We now provide a definition appropriate to use here:

Definition 7.1.3. *We say that* $f : \mathfrak{X} \to \mathbb{R}$ *is* (α, C) **Hölder continuous** *if:*

$$|f(x) - f(y)| \leq Cd(x, y)^{\alpha}$$

In addition to being defined for functions on a metric space, this definition is also different in that it does involve derivatives, as these are not defined for metric spaces (though they are for normed spaces) and also in the use of the fractional power α. We will now proceed to provide bounds on the expected loss of k_N nearest neighbor in general metric spaces.

7.2 A Bound on the Mean Squared Error

We now present a bound on the k_n nearest neighbor expected loss conditioned on the independent variables:

Lemma 7.2.1. *If* m *is* (α_m, C_m) *Hölder and* σ^2 *is* $(\alpha_\sigma, C_\sigma)$ *Hölder, then:*

$$\begin{aligned} & E\left[\left(Y_{n+1} - \hat{Y}_{n+1}\right)^2 \middle| X_{1:n+1}\right] \\ \leq \; & \left(1 + \frac{1}{k_n}\right)\sigma^2(X_{n+1}) + C_m^2 d(X_{n+1}, X_{k_n,n}(X_{n+1}))^{2\alpha_m} \\ & + \frac{1}{k_n} C_\sigma d(X_{k_n,n}(X_{n+1}), X_{n+1})^{\alpha_\sigma} \end{aligned}$$

where $\hat{Y}_{n+1}$ *is the* k_n *nearest neighbor estimate of* Y_{n+1}*.*

Proof. We have:

$$\begin{aligned} & E\left[\left(Y_{n+1} - \hat{Y}_{n+1}\right)^2 \middle| X_{1:n+1}\right] \\ = \; & E\left[\left(Z_1 + Z_2 + Z_3\right)^2 \middle| X_{1:n+1}\right] \\ = \; & E\left[Z_1^2 + Z_2^2 + Z_3^3 + 2Z_1Z_2 + 2Z_2Z_3 + 2Z_1Z_3 \middle| X_{1:n+1}\right] \end{aligned} \tag{7.1}$$

where:

$$\begin{aligned} Z_1 &= Y_{n+1} - m(X_{n+1}) \\ Z_2 &= m(X_{n+1}) - \frac{1}{k_n}\sum_{i=1}^{k_n} m(X_{i,n}(X_{n+1})) \\ Z_3 &= \frac{1}{k_n}\sum_{i=1}^{k_n} m(X_{i,n}(X_{n+1})) - \hat{Y}_{n+1} \end{aligned}$$

where $X_{i,n}(X_{n+1})$ is the *ith* nearest neighbor of X_{n+1} in $X_{1:n}$. We now bound each of the 6 terms in the expectation of Equation (7.1):

1. We can write $E\big[\, Z_1^2 \big|\, X_{1:n+1}\big]$ as follows:

$$E\big[\, Z_1^2 \big|\, X_{1:n+1}\big] = E\Big[\left(Y_{n+1} - m(X_{n+1})\right)^2 \Big| \, X_{1:n+1}\Big] = \sigma^2(X_{n+1})$$

2. We can bound $E\big[\, Z_2^2 \big|\, X_{1:n+1}\big]$ as follows:

$$\begin{aligned}
E\big[\, Z_2^2 \big|\, X_{1:n+1}\big] &= E\left[\left(m(X_{n+1}) - \frac{1}{k_n}\sum_{i=1}^{k_n} m(X_{i,n}(X_{n+1}))\right)^2 \middle| \, X_{1:n+1}\right] \\
&= E\left[\left(\frac{1}{k_n}\sum_{i=1}^{k_n} \left(m(X_{n+1}) - m(X_{i,n}(X_{n+1}))\right)\right)^2 \middle| \, X_{1:n+1}\right] \\
&\leq E\left[\frac{1}{k_n^2}\left(\sum_{i=1}^{k_n} \left|m(X_{n+1}) - m(X_{i,n}(X_{n+1}))\right|\right)^2 \middle| \, X_{1:n+1}\right] \\
&\leq E\left[\frac{1}{k_n^2}\left(\sum_{i=1}^{k_n} C_m d(X_{n+1}, X_{i,n}(X_{n+1}))^{\alpha_m}\right)^2 \middle| \, X_{1:n+1}\right] \\
&\leq E\Big[\frac{1}{k_n^2}\left(k_n C_m d(X_{n+1}, X_{k_n,n}(X_{n+1}))^{\alpha_m}\right)^2 \Big| \, X_{1:n+1}\Big] \\
&= C_m^2 d(X_{n+1}, X_{k_n,n}(X_{n+1}))^{2\alpha_m}
\end{aligned}$$

3. We can bound $E\big[\, Z_3^2 \big|\, X_{1:n+1}\big]$ as follows:

$$\begin{aligned}
E\big[\, Z_3^2 \big|\, X_{1:n+1}\big] &= E\left[\left(\frac{1}{k_n}\sum_{i=1}^{k_n} m(X_{i,n}(X_{n+1})) - \hat{Y}_{n+1}\right)^2 \middle| \, X_{1:n+1}\right] \\
&= E\left[\left(\frac{1}{k_n}\sum_{i=1}^{k_n} m(X_{i,n}(X_{n+1})) - \frac{1}{k_n}\sum_{i=1}^{k_n} Y_{i,n}(X_{n+1})\right)^2 \middle| \, X_{1:n+1}\right] \\
&= E\left[\frac{1}{k_n^2}\left(\sum_{i=1}^{k_n} \left(m(X_{i,n}(X_{n+1})) - Y_{i,n}(X_{n+1})\right)\right)^2 \middle| \, X_{1:n+1}\right] \\
&= E\left[\frac{1}{k_n^2}\left(\sum_{i=1}^{k_n} Z_{3,i}\right)^2 \middle| \, X_{1:n+1}\right] \\
&= E\left[\frac{1}{k_n^2}\sum_{i=1}^{k_n}\sum_{j=1}^{k_n} Z_{3,i} Z_{3,j} \middle| \, X_{1:n+1}\right]
\end{aligned}$$

where $Z_{3,i} = m(X_{i,n}(X_{n+1})) - Y_{i,n}(X_{n+1})$. Now $Z_{3,i}$ and $Z_{3,j}$ are conditionally independent given $X_{1:n+1}$ for $i \neq j$ so that:

$$E[\, Z_{3,i} Z_{3,j} |\, X_{1:n+1}] \;=\; E[\, Z_{3,i} |\, X_{1:n+1}]\, E[\, Z_{3,j} |\, X_{1:n+1}]$$

but:

$$\begin{aligned} E[\, Z_{3,j} |\, X_{1:n+1}] &= E[m(X_{i,n}(X_{n+1})) - Y_{i,n}(X_{n+1}) |\, X_{1:n+1}] \\ &= m(X_{i,n}(X_{n+1})) - E[Y_{i,n}(X_{n+1}) |\, X_{1:n+1}] = 0 \end{aligned}$$

Hence:

$$\begin{aligned} E\big[\, Z_3^2 \big|\, X_{1:n+1}\big] &= E\left[\frac{1}{k_n^2} \sum_{i=1}^{k_n} \sum_{j=1}^{k_n} Z_{3,i} Z_{3,j} \,\middle|\, X_{1:n+1} \right] \\ &= E\left[\frac{1}{k_n^2} \sum_{i=1}^{k_n} Z_{3,i}^2 \,\middle|\, X_{1:n+1} \right] \\ &= E\left[\frac{1}{k_n^2} \sum_{i=1}^{k_n} \left(m(X_{i,n}(X_{n+1})) - Y_{i,n}(X_{n+1}) \right)^2 \,\middle|\, X_{1:n+1} \right] \\ &= \frac{1}{k_n^2} \sum_{i=1}^{k_n} \sigma^2(X_{i,n}(X_{n+1})) \end{aligned}$$

4. Note that Z_2 is measurable with respect to $X_{1:n+1}$ so that we can calculate $E[\, Z_1 Z_2 |\, X_{1:n+1}]$ as follows:

$$\begin{aligned} E[\, Z_1 Z_2 |\, X_{1:n+1}] &= Z_2 E[\, Z_1 |\, X_{1:n+1}] = Z_2 E[Y_{n+1} - m(X_{n+1}) |\, X_{1:n+1}] \\ &= Z_2 \left(E[Y_{n+1} |\, X_{1:n+1}] - m(X_{n+1}) \right) = 0 \end{aligned}$$

5. Again using that Z_2 is measurable with respect to $X_{1:n+1}$, we can calculate $E[\, Z_2 Z_3 |\, X_{1:n+1}]$ as follows:

$$E[\, Z_2 Z_3 |\, X_{1:n+1}] \;=\; Z_2 E[\, Z_3 |\, X_{1:n+1}] = Z_2 E\left[\frac{1}{k_N} \sum_{i=1}^{k_n} Z_{3,i} \,\middle|\, X_{1:n+1} \right] = 0$$

where we have used some results from Item 3 above.

6. Note that Z_1 and Z_3 are conditionally independent given $X_{1:n+1}$ since, of the Y's, Z_1 depends only on Y_{n+1} and Z_3 depends only on $Y_{1:n}$. Hence, we can evaluate $E[\, Z_1 Z_3 |\, X_{1:n+1}]$ as follows:

$$E[\, Z_1 Z_3 |\, X_{1:n+1}] = E[\, Z_1 |\, X_{1:n+1}]\, E[\, Z_3 |\, X_{1:n+1}] = 0$$

where the last equality is from Item 5 above.

Putting the above together in Inequality (7.1) yields:

$$\begin{aligned}
&E\left[\left(Y_{n+1} - \hat{Y}_{n+1} \right)^2 \middle| \, X_{1:n+1} \right] \\
&\quad = \quad E\big[\, Z_1^2 + Z_2^2 + Z_3^3 + 2Z_1Z_2 + 2Z_2Z_3 + 2Z_1Z_3 \big|\, X_{1:n+1}\big] \\
&\quad \leq \quad \sigma^2(X_{n+1}) + C_m^2 d(X_{n+1}, X_{k_n,n}(X_{n+1}))^{2\alpha_m} + \frac{1}{k_n^2} \sum_{i=1}^{k_n} \sigma^2(X_{i,n}(X_{n+1})) \\
&\quad = \quad \left(1 + \frac{1}{k_n} \right) \sigma^2(X_{n+1}) + C_m^2 d(X_{n+1}, X_{k_n,n}(X_{n+1}))^{2\alpha_m} \\
&\qquad + \frac{1}{k_n^2} \sum_{i=1}^{k_n} \left(\sigma^2(X_{i,n}(X_{n+1})) - \sigma^2(X_{n+1}) \right) \\
&\quad \leq \quad \left(1 + \frac{1}{k_n} \right) \sigma^2(X_{n+1}) + C_m^2 d(X_{n+1}, X_{k_n,n}(X_{n+1}))^{2\alpha_m} \\
&\qquad + \frac{1}{k_n^2} \sum_{i=1}^{k_n} C_\sigma d(X_{i,n}(X_{n+1}), X_{n+1})^{\alpha_\sigma} \\
&\quad \leq \quad \left(1 + \frac{1}{k_n} \right) \sigma^2(X_{n+1}) + C_m^2 d(X_{n+1}, X_{k_n,n}(X_{n+1}))^{2\alpha_m} \\
&\qquad + \frac{1}{k_n} C_\sigma d(X_{k_n,n}(X_{n+1}), X_{n+1})^{\alpha_\sigma}
\end{aligned}$$

which was to be shown. □

Bibliography

[KP95] Sanjeev R. Kulkarni and Steven E. Posner. Rates of convergence of nearest neighbor estimation under arbitrary sampling. *IEEE Transactions on Information Theory*, 41(4):1028–1039, July 1995.

[LGW08] Frederic Udina Lázló Györfi and Harro Walk. Nonparametric nearest neighbor based empirical portfolio selection strategies. *Statistics & Decisions*, 26:145–157, 2008.

[LGW10] Adam Krzyżak Lázló Györfi, Michael Kohler and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression.* Springer, 2010.

[LGW12] György Ottucsák Lázló Györfi and Harro Walk. *Machine Learning for Financial Engineering.* Imperial College Press, 2012.

[Mit97] Tom M. Mitchell. *Machine Learning.* McGraw-Hill, 1997.