Probability Theory

Jacob Denson

July 22, 2024

Table Of Contents

Chapter 1

Foundations

> So also the games in themselves merit to be studied and if some penetrating mathematician meditated upon them he would find many important results, for man has never shown more ingenuity than in his plays.
>
> Leibniz

These notes outline the basics of probability theory, the mathematical framework which allows us to interpret the statement that we are 8 times more likely to develop lung disease if you are smoker than if you are a non-smoker, or that there is a *50% chance* of rain on Saturday?

These statements seem intuitive, and we use them naturally in everyday conversation. But a closer analysis of these statements reveals a couple difficulties with understanding such a statement. On Saturday, it will either rain, or not rain, so it is nontrivial to see how one would calculate a reasonable universal 'chance' of such an event happening.

Mathematicians have a rigorous abstraction of these statements, interpreted in the language of measure theory. Soon, we will begin to work in this language. But probability theory is also firmly grounded in scientific applications. So at least for intuition, we should begin by exploring the various interpretations of probability theory in real life applications.

In this chapter, we will explore the two major interpretations of probability theory in real life, each of which use the same underlying mathematical theory to make judgements about the world. But regardless of which interpretation you have, the axiomatic system through which mathematicians study probability remains the same; the two interpretations differ only through which the system is applied to model real life events.

1.1 Frequentist Probability

Classical probability theory was developed according to the intuitions of what is now known as the frequentist school of probability theory, as developed prominantly by R.A. Fisher and R. von Mises, and is the simplest interpretation of probability to understand. It is most easily understood from the point of view of a scientist. Suppose you are repeatedly performing

some well-controlled experiment, in the sense that you expect a similar outcome to occur in each trial. Even under rigorously controlled conditions, the experiment will not always result in the same outcome. Slight experimental error results in slight changes in the outcome of the experiment. Nonetheless, some outcomes may occur more frequently than others.

Let us perform an experiment as often as desired, obtaining an infinite sequence of results. Let E be a certain statement about the outcome of the experiment. In probability theory, we call this an *event*. For instance, E may ask whether a flipped coin lands heads up when flipping a coin repeatedly. We define the *relative frequency* of E being true in n trials by the equation

$$P_n(E) := \frac{\#\{k \leqslant n : E \text{ is true for experiment } k\}}{n}$$

The key assumption of the frequentist school of probability is the existence of a *long term relative frequency* for these events; if our experiments are suitably controlled, then regardless of the specific sequence of measured outcomes, our relative frequencies will always converge to a well defined invariant ratio. We define this ratio to be the *probability* of a certain event, denoted $\mathbb{P}(E)$. That is,

$$\mathbb{P}(E) := \lim_{n\to\infty} P_n(E).$$

Let's explore some consequences of this doctrine:

- Since $0 \leqslant P_n(E) \leqslant 1$ for any n, taking limits shows $0 \leqslant \mathbb{P}(E) \leqslant 1$.
- If Ω is a tautological statement, i.e. an event which is true for any experiment, then $P_n(E) = 1$ for all n, so
$$\mathbb{P}(\Omega) = 1.$$
- If $E_1, \ldots, E_n$ are disjoint events, in the sense that no two of these propositions can be simultaneously true for any particular experiment, then
$$\begin{aligned} P_n\left(E_1 \vee \cdots \vee E_n\right) &= \frac{\#\{k \leqslant n : E_1 \vee \cdots \vee E_n \text{ is true for experiment } k\}}{n} \\ &= \sum_{i=1}^{n} \frac{\#\{k \leqslant n : E_i \text{ is true for experiment } k\}}{n} = \sum_{i=1}^{n} P_n(E_i). \end{aligned}$$
Thus
$$\mathbb{P}\left(\bigvee_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} \mathbb{P}(E_i).$$
Similarily, this result also holds for countable families of events $\{E_i\}$.

The properties described here turn out to be sufficient to describe all the mathematically important rules of probability theory. What's more, we can use these rules as axioms to *prove*, under certain mathematical assumptions, that the relative freuqncies of a sequence of controlled experiments eventually settles down to a well defined ratio, a fact known as the strong and weak laws of probability, which justifies the thought process of the frequentist school in the first place.

1.2 Bayesian Probability

The frequentist school is sufficient to use probability theory to model scientific experiments, but in everyday life we make a more expansive use of probabilistic language. If you turn on the news, it's common to hear that "there is an 80% chance of downpour this evening". It is difficult to interpret this result in the frequentist definition of probability. Even if we see each night's temperament as an experimental trial, it is hard to convince yourself that these experiments are controlled enough to converge to a probabilistic result. The Bayesian school of probability redefines probability theory to be attuned to a person's individual beliefs, so that we can interpret "there is an 80% chance of downpour this evening" as an individual's belief that they think it will rain this evening rather than not rain.

You might argue that, if probability is a personal belief in an unknown event, we can choose probabilities however we want, which would break down the logical structural required to make any mathematical progress. However, the probabilities that the Bayesian school studies are forced to be 'logically consistant', in a manner we will now describe. This forces the logical structure required for mathematical probability theory.

Consistancy can be formulated in various ways; here we discuss what is known as the Dutch book method, developed by the Italian probabilist Bruno de Finetti; if you assign to a certain unknown event E a probability $\mathbb{P}(E) \in [0,1]$, then you believe that a bet at $[\mathbb{P}(E) : 1 - \mathbb{P}(E)]$ odds is completely fair. In other words, for any $A > 0$, you would be willing to make a bet that if E occurs, you win $A \cdot (1 - \mathbb{P}(E))$ dollars, and if E does not occur, you have to pay up $A \cdot \mathbb{P}(E)$ dollars. Since you believe the bet is fair, you must also be willing to play a game where you lose $A \cdot (1 - \mathbb{P}(E))$ dollars if E occurs, and gain $A \cdot \mathbb{P}(E)$ dollars if E does not occur. For instance, you might be willing to bet a dollar against a dollar that a coin will turn up heads, which is $[1 : 1]$ odds, so we would assign the probability that a coin will turn up heads as $1/2$.

We saw a probability function $\mathbb{P}$ is *inconsistant* if it possible to make a series of bets that will guarantee a profit regardless of the outcome; this is known as a Dutch book. We view such functions as illogical, and so in probability theory we concentrate solely on *consistent* probability functions. It turns out that consistant probability functions satisfy the same properties as frequentist probabilities:

- Suppose an event Ω is tautological, and thus always occurs. We claim that any consistant probability function $\mathbb{P}$ must satisfy

  $$\mathbb{P}(\Omega) = 1.$$

  Suppose $\mathbb{P}(\Omega) < 1$. If we made a 1 dollar bet on Ω occuring, then we would make $1 - \mathbb{P}(\Omega) > 0$ dollars off the bet. This bet always makes money, so it is a Dutch book, and therefore $\mathbb{P}$ is inconsistant.

- Suppose $E_1, \ldots, E_n$ are disjoint propositions. We claim that if $\mathbb{P}$ is consistant, then

  $$\mathbb{P}\left(\bigvee_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} \mathbb{P}(E_i).$$

Suppose $\mathbb{P}$ is a probability function such that

$$\mathbb{P}\left(\bigvee_{i=1}^{n} E_i\right) < \sum_{i=1}^{n} \mathbb{P}(E_i).$$

Make a $[\mathbb{P}(E_i) : 1 - \mathbb{P}(E_i)]$ bet that each event E_i does *not* occur, for each index i, as well as a $[\mathbb{P}\left(\bigvee_{i=1}^{n} E_i\right) : 1 - \mathbb{P}\left(\bigvee_{i=1}^{n} E_i\right)]$ bet that $\bigvee_{i=1}^{n} E_i$ *does* occur. Suppose the event E_i occurs. Then we win every bet that E_j does not occur, for $j \neq i$, as well as the event that $\bigvee_{i=1}^{n} E_i$ occurs. Our profits are

$$\left(1 - \mathbb{P}\left(\bigvee_{i=1}^{n} E_i\right)\right) + \sum_{j \neq i} \mathbb{P}(E_j)$$

and we lose $1 - \mathbb{P}(E_i)$ dollars. Thus our total revenue is

$$\left(1 - \mathbb{P}\left(\bigvee_{i=1}^{n} E_i\right)\right) + \sum_{j \neq i} \mathbb{P}(E_j) - (1 - \mathbb{P}(E_i))$$
$$= \sum \mathbb{P}(E_i) - \mathbb{P}\left(\bigvee_{i=1}^{n} E_i\right) > 0.$$

On the other hand, if $\bigvee_{i=1}^{n} E_i$ does *not* occur, then we win every bet that E_i does *not* occur, but lose the bet that $\bigvee_{i=1}^{n} E_i$ occurs. Thus our profits are $\sum \mathbb{P}(E_i)$ dollars, but have to pay up $\mathbb{P}\left(\bigvee_{i=1}^{n} E_i\right)$ dollars. Our total revenue is

$$\sum \mathbb{P}(E_i) - \mathbb{P}\left(\bigvee_{i=1}^{n} E_i\right) > 0.$$

Thus we always make a positive revenue, so we have a Dutch book. On the other hand, if

$$\mathbb{P}\left(\bigvee_{i=1}^{n} E_i\right) > \sum_{i=1}^{n} \mathbb{P}(E_i),$$

then making the opposite series of bets also gives a Dutch book. Thus both inequalities gives a contradiction, so the required identity holds for any consistant probability function.

One technical difference in the Bayesian regime is that we are only able to prove that $\mathbb{P}(\bigvee E_i) = \sum \mathbb{P}(E_i)$ for *finite* disjoint unions of events. But allowing countable sums is so useful that it is difficult to solely use finite unions of events (though Definetti forced himself only to finite unions). But if you are willing to let this slight technical problem slip by, this means that Bayesian and Frequential probability theories both lead to the same fundamental laws of probability. We shall take the two laws we have derived, and use them to make a rigorous model so no more philosophical questions can enter the theory. This is where mathematical probability theory takes its form.

1.3 Axioms of Probability

The properties we have discovered are followed by both the Frequentist and Bayesian school formed the *axioms of probability*. The idea of the *event* and *sample space* was first introuced by R. Von Mises, and an axiomatic theory was eventually built up by Kolmogorov using the modern tools of measure theory. We consider a set Ω, which we view as a set of 'all possible outcomes' to some random phenomenon. This is called the *sample space*. We then fix a sigma-algebra Σ on Ω, whose elements we call *events*, and consider a positive measure $\mathbb{P} : \Sigma \to [0, \infty)$ with $\mathbb{P}(\Omega) = 1$, known as a *probability measure*. Together, $(\Omega, \Sigma, \mathbb{P})$ gives the structure of a *probability space*. We view Σ as the set of all 'acceptable' statements to ask the probability of. Measure theory tells us that we cannot assign probabilities to all subsets of Ω for sufficiently complicated probability distributions, which is the reason for this technicality.

From this summary, it seems probability theory should be viewed as a subset of measure theory, studying positive measures. But probability theory really only takes a probability space as the *model* on which to study certain phenomena. Once this is done, probability theory takes on a different mindset; as the theory progresses, the more technical aspects of measure theory tend to disappear, as probabilists concentrate more and more on properties of randomness independent of the particular measure space they are considering, to the point where mathematician Michael Talagrand called the more measure theoretic technical aspects of the theory as 'pre-1950s mathematics'.

More precisely, the probabilistic way of thinking eschews the particular sample space used to study properties of events that are 'independent of the sample spaces considered'. To consider a more precise approximation to this principle, we define an *extension* of a probability space Ω_0 to be a probability space Ω_1 together with a surjective, measure preserving map $T : \Omega_1 \to \Omega_0$. In the probabilistic way of thinking, we should identify an event $E_{\Omega_0} \subset \Omega_0$ with the event $E_{\Omega_1} = T^{-1}(E_{\Omega_0})$. In particular, we should only focus on studying properties of events which are preserved under this 'lifting'. Examples of properties preserved under lifting are the *probability* of an event, set operations like *union* and *intersection*, and properties like *independence*.

This way of thinking gets even easier with the tool of a *random variable*, which is a measurable function $X : \Omega \to S$, where S is some measurable space. In the probabilistic way of thinking, we think of X as a 'random element of S', i.e. with the probability of X taking a value in some $E \subset S$ as $\mathbb{P}(X^{-1}(E))$. In particular, we write this quantity as $\mathbb{P}(X \in E)$. Any of these probabilities is easily seen to preserved under lifting (i.e. identifying a random variable $X_{\Omega_0} : \Omega_0 \to E$ with $X_{\Omega_1} : \Omega_1 \to E$ given by $X_{\Omega_1} = X_{\Omega_0} \circ T$), and induces a probability measure μ_X on S given by $\mu_X(E) = \mathbb{P}(X \in E)$. We call this the *law* of the random variable. Conversely, if μ is a probability distribution on S, we write $X \sim \mu$ to mean that $\mu_X = \mu$. Using measure theory normally leads probability theory to introduce events first before random variables (this is not strictly necessary, since, e.g. using the Riesz-Markov-Kakutani extension theorem, we can define measures in terms of positive linear operators on families of functions which we consider as random variables). But random variables soon dominate the theory once we have enough measure theory to think probabilistically.

Classically, probability theory was the study of certain techniques used to calculate probabilities of events, and statistics of random variables, which motivated the development of the modern fields of combinatorics and integration theory. Though we are still interested in estimating the probabilities of certain events, probability thoery nowadays also focuses on more general principles underlying probability spaces, and integrates tools in more modern fields such as functional analysis to achieve this.

Example. *For discrete sets S, it is often easiest to define the probability distribution on S* atomically*, i.e. considering a function $f : S \to [0,1]$ such that $\sum f(s) = 1$, and then to define the distribution on S such that, for $E \subset S$,*

$$\mathbb{P}(E) = \sum_{s \in E} f(s).$$

we note that in this form, these measures are identified with a convex subset of l^∞, i.e. the set

$$\left\{ f : S \to [0,1] : \sum_s f(s) = 1 \right\}.$$

This is a convex subset of the unit ball in $l^\infty(S)$, which leads to some interesting functional analysis.

1.4 Discrete Probability

Let us consider some examples of discrete random variables, i.e. random variables $X : \Omega \to S$, where S is finite, or countable. We will begin by focusing mostly on *discrete random variables*, i.e. random variables $X : \Omega \to S$ taking values in a finite, or countable space S. In the finite case with $\#(S) = n$, we can identify S with $\{1, \ldots, n\}$. The most basic examples are those random variables which are *uniformly distributed* on S, but one could also consider any distribution on $\{1, \ldots, n\}$ induced by some probability vector $p \in [0,1]^n$.

Example. *If S is finite, we can put a* uniform distribution *on Ω atomically by defining*

$$\mathbb{P}(\omega) = \frac{1}{\#(\Omega)}.$$

If the law of a random variable X valued in S has the uniform distribution, we write $X \sim Uniform(S)$.

As an example, we could model the scenario of flipping a coin, and observing whether a head or tails is observed, by considering a random variable valued in the set $S = \{H, T\}$, and equipped with the uniform distribution.

Note that the space Ω does not really matter much from a probabilistic way of thinking; one choice could be to set $\Omega = \{H, T\}$, equipped with the uniform measure, and then to let X be the identity map. Another choice would be to set $\Omega = [0,1]$, equipped with the Lebesgue

measure, and then to set $X(t) = \mathbf{I}(t \leqslant 1/2)$. From the probabilistic perspective, both of the resulting random variables would be viewed as modelling the same probabilistic situation, i.e. any question we ask from a probabilistic perspective should not depend on which Ω we are using.

Example. *If $s \in S$ is fixed, we can put a* point mass distribution *on S by defining, for each $E \subset S$,*

$$\mathbb{P}(E) = \begin{cases} 1 & s \in E, \\ 0 & s \notin E. \end{cases}$$

The distribution represents an event where a outcome s is certain, and all other situations are impossible. If X is a random variable with the point mass distribution as it's law, we write $X \sim s$.

More generally, if $S = \{0, 1\}$, we can define a *Bernoulli* distribution Bernoulli(p) to be the distribution on S such that if $X \sim \text{Bernoulli}(p)$, then

$$\mathbb{P}(X = 1) = p.$$

More generally, if $S = \{1, \ldots, n\}$ and we consider $p \in [0, 1]^m$ with $\sum p_i = 1$, then we can consider random variables X valued in S such that for $1 \leqslant k \leqslant n$,

$$\mathbb{P}(X = k) = p_k.$$

Any random variable valued on S has such a probability vector.

We can obtain other useful distributions from a collection of random variables $X_1, \ldots, X_m : \Omega \to \{1, \ldots, n\}$. For instance, if we do not care about the particular ordering that these values take, we can consider the random variables $S_1, \ldots, S_n : \Omega \to \{1, \ldots, m\}$, where S_i gives the number of X_j with $X_j = i$. When we study $S_1, \ldots, S_n$, we often call the problem *indistinguishable*, since we do not care about distinguishing the particular values the variables $\{X_i\}$ take, but rather, their accumulated statistics. This is often much more simple, since the total number of possible different values the tuple $(S_1, \ldots, S_n)$ can take is equal to $\binom{m+n-1}{m}$, whereas the total number of configurations of the tuple $(X_1, \ldots, X_m)$ is equal to n^m, so the latter situation can be much more complicated. If $m \gg n$, then the former quantity roughly behaves like m^n, which is much smaller than n^m.

Let us consider some problems of indistinguishable type that occur in statistical mechanics. Imagine putting a certain number of indistinguishable molecules in a certain number of positions. If the positions are indistinguishable, it is reasonable to assume that each molecule independently occurs in any position with equal probability, an assumption leading to the *Maxwell-Boltzmann* theory of statistical mechanics. Given m independent random variables $X_1, \ldots, X_m$ with uniform distribution on $\{1, \ldots, n\}$, combinatorics then tells us that if $S_1, \ldots, S_n$ are as above, then for $k_1 + \cdots + k_n = m$,

$$\begin{aligned} \mathbb{P}_{\text{MB}}(S_1 = k_1, \ldots, S_n = k_n) &= \frac{1}{m^n} \binom{m}{k_1\ k_2\ \ldots\ k_n} \\ &= \frac{1}{n^m} \frac{m!}{k_1! \ldots k_n!}. \end{aligned}$$

Thus the random vector $S = (S_1, \ldots, S_n)$ have the *multinomial distribution*. In particular, we see that evenly spread out states are more likely to occur then highly concentrated states. More generally, if $X_1, \ldots, X_m$ are independent and have distribution given by a probability vector $p \in [0,1]^n$, then the resulting vector $S = (S_1, \ldots, S_n)$ will have law Multinomial(m, p), and will satisfy the formula

$$\mathbb{P}(S_1 = k_1, \ldots, S_n = k_n) = \binom{m}{k_1\ k_2\ \ldots\ k_n} p_1^{k_1} \cdots p_n^{k_n}.$$

Thus we have calculated the probability distribution of the $S_1, \ldots, S_n$. In particular, the distribution of S_1, given that $n = 2$, is called the *Binomial distribution*, denoted $\mathrm{Bin}(m, p)$, where $p \in [0,1]$ giving the probability that a value X takes on the value 1.

However, this turns out to be an *incorrect assumption* about the placement of certain types of physical particles in space. In the 20th century, Bose and Einstein found that in the study of certain particles, such as protons, nuclei, and atoms with an even number of elementary particles, any such configuration $(k_1, \ldots, k_n)$ has an *equal* chance of occuring, i.e. if $X_1, \ldots, X_m$ describes the position of these particles, then for $k_1 + \cdots + k_n = m$,

$$\mathbb{P}_{\mathrm{BE}}(S_1 = k_1, \ldots, S_n = k_n) = \frac{1}{\binom{n+m-1}{m}} = \frac{m!(n-1)!}{(n+m-1)!}.$$

This leads to the *Bose-Einstein* theory of statistical mechanics.

For other particles such as electrons, neutrons, and protons, a different set of assumptions hold as developed by Fermi and Dirac and called the *Fermi-Dirac* theory: no two particles can share the same box, and all permutations satisfying this constraint are equally likely. Thus for $(k_1, \ldots, k_n) \in \{0,1\}$ with $k_1 + \cdots + k_n = m$, in the Fermi-Dirac theory we therefore have

$$\mathbb{P}_{\mathrm{FD}}(S_1 = k_1, \ldots, S_n = k_n) = \frac{1}{\binom{n}{m}} = \frac{m!(n-m)!}{n!}.$$

Each theory of the distributions of the random variables $\{X_1, \ldots, X_m\}$ leads to a completely different distribution of the random variables $(S_1, \ldots, S_n)$, but all methods can be dealt with via the general probabilistic theory.

The Fermi-Dirac theory is larger than the corresponding probability for the Maxwell-Boltzmann probabilities, but becomes comparable for n significantly larger than m, and with k as above, we have

$$\mathbb{P}_{\mathrm{MB}}(S = k) = \left(1 - \frac{1}{n}\right) \cdots \left(1 - \frac{m-1}{n}\right) \mathbb{P}_{\mathrm{FD}}(S = k).$$

Thus

$$\mathbb{P}_{\mathrm{FD}}(S = k) \geqslant \mathbb{P}_{\mathrm{MB}}(S = k) \geqslant (1 - O(m^2/n))\, \mathbb{P}_{\mathrm{FD}}(S = k).$$

Thus the two quantities begin to become more and more equal as we have $m \lesssim n^{1/2}$, i.e. at the *birthday paradox* threshold.

Example. *An interesting application of combinatorial probability theory is in the so called Birthday paradox. Given a set of m points to place uniformly in n boxes, the probability that they all lie in distinct boxes is*

$$\frac{1}{n^m}\frac{n!}{(n-m)!} = (1-1/n)\cdots(1-(m-1)/n).$$

In particular, if $m = 365$, and $n = 23$, then we calculate the probability that two points lie in the same box exceeds one half, i.e. so that there is more than 50% of two children in an elementary school classroom sharing the same birthday. To determine, for each n, the number m required in order for this quantity to exceed 50%, we calculate using logarithms that

$$\begin{aligned}(1-1/n)\cdots(1-(m-1)/n) &= \exp\left(\sum_{k=1}^{m-1}\log(1-k/n)\right)\\ &\geqslant \exp\left(-\frac{(m-1)m}{2n}\right).\end{aligned}$$

This quantity is bigger than $1/2$ if $m \gtrsim n^{1/2}$.

Another distribution is obtained by changing our way of thinking of the discrete variables $\{X_1,\ldots,X_m\}$. Let us imagine a barrel of K balls with n different labels. There are k_i balls of type i, with $k_1+\cdots+k_n = K$. If we take balls one by one, let X_i denote the label of the ball we picked up, and then *replace* the ball back in the barrel, then the variables $\{X_i\}$ will be independent random variables, with $\mathbb{P}(X_i = j) = k_j/K$ for each i and j. As a result, for $l_1+\cdots+l_n = m$, the vector $(S_1,\ldots,S_n)$ will be Multinomial(n,p) distributed, where $p = k/K$, i.e.

$$\mathbb{P}(S_1 = l_1,\ldots,S_n = l_n) = \binom{m}{l_1\ \cdots\ l_n}\frac{k_1^{l_1}\ldots k_n^{l_n}}{K^m}.$$

On the other hand, if we do *not* replace the balls, then the vector $(X_1,\ldots,X_m)$ will not have independent coordinates. It is easiest to determine the resulting statistics of the totals $(S_1,\ldots,S_n)$, and this vector will satisfy the *hypergeometric distribution*, i.e. for $l_1+\cdots+l_n = m$, with $l_i \leqslant k_i$ for each i, we will have

$$\mathbb{P}(S_1 = l_1,\ldots,S_n = l_n) = \frac{\binom{k_1}{l_1}\cdots\binom{k_n}{l_n}}{\binom{K}{m}}.$$

Indeed, the denominator denotes the total number of subsets of K balls we can choose, whereas the numerator counts the number of subsets of K balls we can choose containing l_1 balls of type k_1, l_2 balls of type k_2, and so on and so forth. The name 'hypergeometric' comes from the fact that the generating function of this probability distribution is a hypergeometric function. In the simple case, but most important case where $n = 2$

On the other hand, we have

$$\mathbb{P}(X_1 = i_1,\cdots,X_n = i_n) =$$

If K is significantly larger than m, then sampling without replacement is essentially the same as sampling with replacement. For simplicity, let us deal with the case $n = 2$, and where we let k denote the number of the balls of type 1 (and thus there are $K - k$ balls of type 2). If $k/K = p$ is held fixed, but K is made much larger than m. We then calculate that

$$\mathbb{P}(S_1 = l, S_2 = m - l) = \binom{k}{l}\binom{K-k}{m-l}/\binom{K}{m}.$$

Applying Stirling's formula, this quantity is equal to

$$p^l(1-p)^{m-l}e^{O_p(m^2/K)},$$

where the implicit constant is uniform as we vary p as long as we stay away from $p = 0$ and $p = 1$. But this means that for large K (i.e. $K \gtrsim m^2$), the hypergeometric distribution behaves like the binomial distribution.

Example. *Consider a five digit number X selected uniformly at random. Then the sample space consists of the numbers* $[0, 99999]$*, and the probability that a particular number is selected is one in a 100,000. There are* $10!/5! = 30240$ *numbers all of whose digits are different, so the probability that a number is selected all of whose digits are different is* 0.3024*. If we look at the first 800 digits of the decimal expansion of the number e, and we take each 5 digit consecutive sequence in this expansion, we end up with approximately the same frequency of unique digits, leading us to believe the digits occuring in the number e are essentially random. A number with this property is called* normal*. Thus experiments lead us to believe that e is a normal number. But a proof of this fact remains an open problem.*

Example. *Which is more likely? Getting at least one six with four throws of a dice, or getting at least two sixes with twenty four rolls of the dice. As the story goes, the Chevalier de Mere, Antoine Gombaud had encountered a gambling rule which suggested both were equiprobable, and proposed the problem to the mathematician Blaise Pascal, since this seemed to disagree with the probabilistic calculations that had begun in the 17th century. Let's see the two different values. The probability of getting one six with four throws of the dice, is equal to one minus the probability of getting no six, i.e.*

$$1 - (5/6)^4 \approx 51.8\%.$$

On the other hand, the probability of getting at least two sixes with twenty four rolls of the dice is equal to 1 minus the probability that you get one or fewer sixes, i.e.

$$1 - (5/6)^{24} - (24 \cdot 5^{23})/6^{24} = 1 - 29 \cdot (5/6)^{23} \approx 56.2\%.$$

Together with this problem, the Chevalier proposed to Pascal the problem of determining the distribution of a prize pool at a competition (a shooting match, or ball game) that had to be cut short, given that each individual in the competition had a certain score. The answer of course is to determine the probabilities each individual has of winning. This problem has

a degree of complexity that requires a development of modern probability to solve, and it's solution by Pascal can be considered a decisive breakthrough in the initiation of probability thoery and established Pascal as one of the primordial founders of the modern theory. Later on, Pascal became more devout, joining a Jansenist group and formulated his famous wager, determining whether one should act as if god exists or god does not exist, even given the slight possiblity of a god, given that the reward for acting as a prospective god might want you to is incredibly high relative to the punishment.

Example. *In an ordered sequence, let us call a maximal subsequence of elements that are of the same type is called a* run. *Any sequence decomposes into a disjoint collection of runs, and the total number of runs is always equal to one plus the number of pairs of adjacent elements that differ from one another. If we have a sequence of n elements, each consisting of one of m different types, and each element of the sequence is selected uniformly at random, then the probability that any pair of adjacent elements differs from one another is equal to* $1 - 1/m$. *Thus (by linearity of expectation) we should expect the average number of runs to be equal to* $n(1 - 1/m)$. *In particular, if* $m = 2$, *then we should expect an average sequence to have about* $n/2$ *runs. If there are much fewer than this many runs, we should expect that the distribution of the sequence is not uniform, i.e. there is a tendency for like elements to cluster. If there are much more than this many runs, we should expect that the distribution has a tendency for like elements to separate from one another.*

Let us suppose that we have a sequence of n elements, containing a elements of one type α, *and b elements of another type* β. *Thus* $n = a + b$. *Let us assume we take a sequence taken from these elements uniformly at random. If l is the number of* β *runs, then either* $l = k - 1$, $l = k$, *or* $l = k + 1$. *Let us calculate the probability that there are k runs of* α *elements, and l runs of* β *elements. This is equivalent to counting the number of ways one can placing a elements into k boxes, such that no box is empty, and place b elements into l boxes, such that none of these boxes is empty. Using stars and bars, the total number of ways of doing this*

$$\binom{a-1}{k-1}\binom{b-1}{l-1}$$

The total number of sequences is equal to the total number of subsets of n elements of size a, i.e.

$$\binom{n}{a}.$$

Thus the probability of having runs of the sort above is equal to

$$\frac{\binom{a-1}{k-1}\binom{b-1}{l-1}}{\binom{n}{a}}.$$

Let's see some applications. For instance, suppose we look at a bar counter, and observe which seats are occupied, e.g. as indicated in a bar with 16 seats by a string of the form EOEEOEEEOEEEOEOE, where E indicates a seat is empty, and O that a seat is occupied. If the number of runs is significantly higher than the average, this indicates that

individuals intentional separate from one another. If the number is significantly lower, individuals are intentionally cluster (maybe we are watching the bar when coworkers meet after work to have a drink). One can never guarantee this is the case, since even with random seating any arrangement is possible; but repeated observation at a bar together with statistical techniques can be used to make stronger conclusions which are wrong with negligible probability.

Here's another application: Suppose that we have two populations, each associated with a numerical quantity (e.g. age, weight, and so on). To determine whether the two populations strongly differ with respect to this quantity, we might arrange them in order of this statistic. If the quantity clearly separates the groups, all of one quantity might precede all of the other quantity, so there are a small number of runs in the sequence we have ordered. On the other hand, if the runs seem random we might expect the groups are not so indistinguishable. This idea has been developed into the Wald-Wolfowitz runs test.

The theory of runs occurs in statistical physics, since in Ising's model of one dimenisonal lattices, the energy of the system is precisely the number of runs.

When studying countable probability spaces (also called discrete probability spaces), the sigma algebra plays almost no real role in the theory. This allows us to get away with discussing most of the basic principles of probability theory without running into too many technicalities. Nonetheless, even in the study of discrete phenomena understanding probability spaces with uncountably many points becomes necessary. For instance, in the study of the limiting average of a sequence of discrete coins flips, our sample space must consist of the space of infinite sequences of coin flips $\{0,1\}^\omega$, which is an uncountable sample space. And it is often necessary in applications to select a point in an interval uniformly at random, leading to continuous-valued probability spaces.

The first immediately obvious fact from the axioms is $\mathbb{P}(E^c) = 1 - \mathbb{P}(E)$, since E and E^c are disjoint events whose union is Ω. A similar discussion shows that $\mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F) - \mathbb{P}(E \cap F)$ because $E \cup F$ can be written as the union of the three disjoint events $E \cap F$, $E \cap F^c$, and $E^c \cap F$, and

$$\mathbb{P}(E) = \mathbb{P}(E \cap F) + \mathbb{P}(E \cap F^c) \quad \mathbb{P}(F) = \mathbb{P}(E \cap F) \cup \mathbb{P}(E^c \cap F)$$

This process can be generalized to unions of finitely many events. We have

$$\mathbb{P}(E \cup F \cup G) = \mathbb{P}(E) + \mathbb{P}(F) + \mathbb{P}(G) - \mathbb{P}(E \cap F) - \mathbb{P}(E \cap G) - \mathbb{P}(F \cap G) + \mathbb{P}(E \cap F \cap G)$$

which can be reasoned by looking at the number of times each element of $E \cup F \cup G$ is 'counted' on the right hand side. In general, we have the inclusion-exclusion principle

$$\mathbb{P}\left(\bigcup_{k=1}^{n} E_k\right) = \sum_{S \subset \{1,\ldots,k\}} (-1)^{|S|} \mathbb{P}\left(\bigcap_{k \in S} E_k\right)$$

This can be proven by a clumsy inductive calculation. More interestingly, but less useful, we often want to calculate the probability of an infinite union of sets E_k occuring. The

inclusion-exclusion principle can be taken 'in the limit' to conclude that

$$\mathbb{P}\left(\bigcup_{k=1}^{\infty} E_k\right) = \lim_{n\to\infty} \mathbb{P}\left(\bigcup_{k=1}^{n} E_k\right) = \sum_{\substack{S\subset\mathbf{N}\\ |S|<\infty}} (-1)^{|S|} \cdot \mathbb{P}\left(\bigcap_{k\in S} E_k\right)$$

where the sum on the right is taken as the limit of the partial sums where $S \subset \{1,\ldots,n\}$ – the sum need not convergence absolutely, so it is important to take the limit in the precise ordering given.

The inclusion-exclusion formula can be tricky to calculate in real examples, so we often rely on estimates to upper bound or lower the probability of a particular event occuring. The trivial *union bound*

$$\mathbb{P}\left(\bigcup E_i\right) \leqslant \sum \mathbb{P}(E_i)$$

can often be applied. This is a good inequality to apply if the E_i are 'nearly disjoint', or each have a neglible probability of occuring. On the other hand, the bound is shockingly bad if all the E_i are equal to one another.

Another useful fact to consider is that $\mathbb{P}(E_k) \to \mathbb{P}(E)$ if the sets E_k 'tend to' E in one form of another. If the E_k are an increasing sequence whose union is E, then we can certainly conclude $\mathbb{P}(E_k) \to \mathbb{P}(E)$. Similarily, if E_k is a decreasing sequence whose intersection is E, then $\mathbb{P}(E_k) \to \mathbb{P}(E)$. To obtain general results, we say that $E_k \to E$ if $\limsup E_k = \liminf E_k = E$, where

$$\limsup_{k\to\infty} E_k = \bigcap_{n=1}^{\infty} \bigcup_{k\geqslant n} E_k = \{\omega : \omega \in E_k \text{ for infinitely many } k\}$$

$$\liminf_{k\to\infty} E_k = \bigcup_{n=1}^{\infty} \bigcap_{k\geqslant n} E_k = \{\omega : \omega \in E_k \text{ for sufficiently large } k\}$$

We can then conclude that $\mathbb{P}(E_i) \to \mathbb{P}(E)$, since once can show

$$\limsup \mathbb{P}(E_k) \leqslant \mathbb{P}\left(\limsup E_k\right)$$

$$\liminf \mathbb{P}(E_k) \geqslant \mathbb{P}\left(\liminf E_k\right)$$

so we can apply the squeeze theorem. This already enables us to prove a very interesting theorem which can guarantee an event can 'never occur'.

Lemma 1.1 (Borel-Cantelli Lemma)**.** *If $E_1, E_2, \ldots$ is a sequence of events with $\sum \mathbb{P}(E_k) < \infty$, then $\mathbb{P}\left(\limsup E_k\right) = 0$. Thus none of the events E_k can happen infinitely often.*

Proof. Because

$$\mathbb{P}\left(\bigcup_{k\geqslant n} E_k\right) \leqslant \sum_{k\geqslant n} \mathbb{P}(E_k)$$

for any $\varepsilon > 0$ we can find an N such that for $n \geqslant N$, $\mathbb{P}\left(\bigcup_{k\geqslant n} E_k\right) < \varepsilon$. But for any n, $\limsup E_k \subset \bigcup_{k\geqslant n} E_k$, and so we conclude $\mathbb{P}(\limsup E_k) < \varepsilon$. We then let $\varepsilon \to 0$ to conclude $\mathbb{P}(\limsup E_k) = 0$. □

The next example shows that the hypothesis $\sum \mathbb{P}(E_k) < \infty$ cannot be relaxed without further analysis of the events E_k beyond their probabilities.

Example. *Take the Haar measure measure on* $\mathbf{T} = \mathbf{R}/\mathbf{Z}$*. Consider a sequence of positive numbers* $x_1, x_2, \ldots$*, define* $S_N = \sum_{n=1}^{N} x_n$*, and* $E_n = [S_{n-1}, S_n]$*, considered modulo* $\mathbf{Z}$ *of course. Then* $\mathbb{P}(E_n) = x_n$*, and* $\sum x_n = \infty$ *happens if and only if every point in* $\mathbf{T}$ *is contained in infinitely many of the* E_n*.*

Theorem 1.2. *If* $E_1, E_2, \ldots$ *are events with* $\inf \mathbb{P}(E_k) > 0$*, then infinitely many of the* E_i *occur at once with positive probability.*

Proof. The event that infinitely many of the $E_1, E_2, \ldots$ occur is the complement of the event that all but finitely many of the E_i *do not* occur, i.e. $\liminf E_i^c$, and it suffices to show $\mathbb{P}(\liminf E_i^c) < 1$. But by Fatou's lemma,

$$\mathbb{P}\left(\inf_{k \geqslant n} E_k^c\right) \leqslant \inf_{k \geqslant n} \mathbb{P}(E_k^c) = 1 - \sup_{k \geqslant n} \mathbb{P}(E_k) \leqslant 1 - \delta$$

and so, letting $n \to \infty$, we conclude $\mathbb{P}(\liminf E_k^c) \leqslant 1 - \delta$. Alternatively, if we consider the functions $S_n = \chi_{E_1} + \cdots + \chi_{E_n}$, then

$$S_n \leqslant m\mathbf{I}(S_n \leqslant m) + n\mathbf{I}(S_n > m) = m + (n-m)\mathbf{I}(S_n > m)$$

so if $\delta = \inf \mathbb{P}(E_i)$, then

$$\delta n \leqslant \mathbf{E}(S_n) \leqslant m + (n-m)\,\mathbb{P}(S_n > m)$$

which leads to the upper bound

$$\mathbb{P}(S_n > m) \geqslant \frac{\delta n - m}{n - m}$$

As $n \to \infty$, the events on the left hand side increasing to $\mathbb{P}(S_\infty > m)$, where we define S_∞ as the sum of all χ_{E_k}. Thus

$$\mathbb{P}(S_\infty > m) \geqslant \limsup_{n \to \infty} \frac{\delta n - m}{n - m} = \delta$$

But we can then let $m \to \infty$ to conclude that $\mathbb{P}(S_\infty = \infty) = \delta$. □

1.5 Conditional Probabilities

In the Bayesian interpretation of probability theory, it is natural for probabilities to change over time as more information is gained about the system in question. That is, given that we know some proposition F holds over the sample space, we obtain a new probability measure over Ω, denoted $\mathbb{P}(\cdot|F)$, which represents the ratio of winnings from the bet which is only played out if F occurs. That is

- You win $1 - \mathbb{P}(E|F)$ dollars if E occurs, and F occurs.
- You lose $\mathbb{P}(E|F)$ dollars if E does not occur, and F occurs.
- No money exchanges hands if F does not occur.

Suppose we had to assign values to $\mathbb{P}(E|F)$, subject to a consistancy principle which prevents a dutch book argument. It then follows that we must have $\mathbb{P}(F)\mathbb{P}(E|F) = \mathbb{P}(E \cap F)$ TODO: Fill in this argument.

In the empirical interpretation, $\mathbb{P}(E|F)$ is the ratio of times that E is true in experiments, where we only count experiments in which F also occurs. That is, we define $\mathbb{P}(E|F)$ as the limit of the ratios

$$P_n(E|F) = \frac{\#\{k \leqslant n : \omega_k \in E, \omega_k \in F\}}{\#\{k \leqslant n : \omega_k \in F\}}$$

But it is easy to calculate, by dividing the numerator and denominator by n, that $P_n(E|F) = P_n(E \cap F)/P_n(F)$, so by taking limits, we find

$$\mathbb{P}(E|F) = \lim_{n\to\infty} \frac{P_n(E \cap F)}{P_n(F)} = \frac{\mathbb{P}(E \cap F)}{\mathbb{P}(F)}$$

which gives us the formula $\mathbb{P}(F)\mathbb{P}(E|F) = \mathbb{P}(E \cap F)$. We must of course assume that $\mathbb{P}(F) \geqslant 0$, since overwise we are almost certain that F will never occur, and then we can almost guarantee that the limit of the values $P_n(E|F)$ does not exist.

Thus we have motivation to define conditional probabilities by the formula $\mathbb{P}(F)\mathbb{P}(E|F) = \mathbb{P}(E \cap F)$, provided that $\mathbb{P}(F) > 0$. It enables us to model the information gained by restricting our knowledge to a particular subset of sample space. In particular, we can use the definition to identify events which contain information 'useless' to learning about another event. We say two events E and F are independant if $\mathbb{P}(E \cap F) = \mathbb{P}(E)\mathbb{P}(F)$, or, provided $\mathbb{P}(F) > 0$, $\mathbb{P}(E|F) = \mathbb{P}(F)$; knowledge of F gives us no foothold over knowledge of the likelihood of E.

Example. *The Monty Hall problem is an incredible example of how paradoxical probability theory can seem. We are on a gameshow. Suppose there are three doors in front of you. A car (brand new!) is placed uniformly randomly behind one of the doors. After we pick a door (the first door, for instance), the gameshow host then opens the second door, which you didn't pick, revealing the car isn't behind the door. It is important to note that he picked randomly from the remaining doors which you didn't pick and don't have a car behind them. What is the chance that the door you picked has the brand new car? You likely would think the two doors have a 50-50 chance of containing the car given this info, but you'd be wrong. Let $X \in \{1,2,3\}$ denote the door chosen uniformly at random where the car lies, and let $Y \in \{1,2,3\}$ denote the door that the host randomly chose to open. We know $Y \neq 1$, because the gameshow host would never open the door we picked; that would give the game away! If $X = 1$, then Y is picked from $\{2,3\}$ with uniform possibility. However, if $X = 2$, something interesting occurs – the gameshow is forced to open door number 3, because that's the only door that (he thinks) won't give any information to the player, and similarily, if $X = 3$, then*

$Y = 2$. Now we know that since X is chosen uniformly at random $\mathbb{P}(X = k) = 1/3$ for each k. Similarily, we know that Y is then chosen uniformly at random from $\{2, 3\}$, given that $X = 1$, so assuming X and Y are independent, we conclude

$$\mathbb{P}(X = 1, Y = 2) = \mathbb{P}(X = 1)\,\mathbb{P}(Y = 2) = 1/6$$

$$\mathbb{P}(X = 1, Y = 3) = 1/6$$

But we also know that if $X = 2$, then $Y = 3$, so

$$\mathbb{P}(X = 2, Y = 3) = \mathbb{P}(X = 2) = 1/3$$

$$\mathbb{P}(X = 3, Y = 2) = \mathbb{P}(X = 3) = 1/3$$

It follows that

$$\begin{aligned}
&\mathbb{P}(\textit{door } 1 \textit{ has a car} | \textit{door } 2 \textit{ was opened}) \\
&\quad = \frac{\mathbb{P}(\textit{door } 1 \textit{ has a car}, \textit{door } 2 \textit{ was opened})}{\mathbb{P}(\textit{door 2 was opened})} \\
&\quad = \frac{\mathbb{P}(\{(X = 1, Y = 2)\})}{\mathbb{P}(\{(X = 1, Y = 2), (X = 3, Y = 2)\})} = \frac{1/6}{1/6 + 1/3} = 1/3
\end{aligned}$$

This means we should definitely change our minds about which door we were going to pick! The argument above causes a great media uproar when it was published in 1990 in a popular magazine, because of how convincing the fallacious argument below is. The total number of possibilities is

$$(X = 1, Y = 2), (X = 1, Y = 3), (X = 2, Y = 3), (X = 3, Y = 2)$$

and the car seems to be in door one half of the possibilities. However, these events do not have the same probability of occuring. However, if the host changes his strategy, the conditional probabilities fall more in line with intuition – if the host always picks door number 2 to open if door number 1 was picked and had the car behind it, then the two remaining doors have an equal chance of being picked.

We end this chapter with a final probability rule which is important in statistical analysis. If B is partitioned into a finite sequence of disjoint events $A_1, \ldots, A_n$, then we have the formula $\mathbb{P}(B) = \sum_i \mathbb{P}(B|A_i)\mathbb{P}(A_i)$. This easily gives us Bayes rule

$$\mathbb{P}(A_j|B) = \frac{\mathbb{P}(B|A_j)}{\sum_i \mathbb{P}(B|A_i)\mathbb{P}(A_i)}$$

If we view A_j as a particular hypothesis from the set of all hypotheses, and B as some obtained data, then Bayes rule enables us to compute the probability that A_j is the true hypothesis from the probability that B is the data generated given the hypothesis is true. This is incredibly important if you can interpret these probabilities correctly (if you are a

Bayesian), but not so useful if you are an empiricist (in which case we assume there is a 'true' result we are attempting to estimate from trials, so there is no probability distribution over the correctness hypothesis, other than perhaps a point mass, in which case Bayes rule gives us no information). We reiterate that Bayes rule is a theorem of probability theory, so is true in any interpretation, but can be used by Bayesians in a much more applicable way to their statistical analysis.

1.6 Kolmogorov's Zero-One Law

s

1.7 Beware of Intuition

Intuition is very useful in mathematics. It hints at which paths to a solution to a problem are most likely to bear fruit. In such a practical mathematical subject in probability, this is certainly true. But probability also has many pitfalls that often trip up our minds. And it is useful to list them in one palce so that we are able to better acknowledge, and avoid them.

- *The Law Of Averages*: If X is a random variable, then it is highly likely to be close to it's average. For instance, a British newspaper once made the claim that to pick a 'good' set of six numbers in the lottery, the average of the numbers should be close to 25. Of course, this is false; the numbers in the lottery have no actual bearing on the likelihood of being drawn. To those who are not easily convinced, ask if the game is the same if the balls were painted with various colours, rather than inscribed with numbers. If you accept this, ask if the colors can be swapped without changing the probabilities. And so if you accept this, then the numbers on the balls can be swapped, so all draws occur with equal probability.

  Of course, theoretically if $X = (X_1, \dots, X_6)$ are six independant random variabels uniformly distributed in $\{0, \dots, 9\}$, and we set $S(X) = X_1 + \cdots + X_6$, then

  $$\mathbb{E}(S(X)) = 6 \cdot \mathbb{E}(\text{Uni}(0, \dots, 9)) = 27.$$

  Certainly, $\text{argmax}_n \mathbb{P}(S = n) = 27$. But this does not mean that any particular choice of $(x_1, \dots, x_6) \in \{0, \dots, 9\}^6$ with $x_1 + \cdots + x_6 = 27$ is more likely to occur than $(x_1, \dots, x_6)$ with $x_1 + \cdots + x_6 \neq 27$; there are just more choices of $(x_1, \dots, x_6) \in \{0, \dots, 9\}^6$ with $x_1 + \cdots + x_6 = 27$.

  What further confuses matters is the weak law of large numbers, which says that if $X^1, X^2, \dots, X^n \in \{0, \dots, 9\}^6$ are n independant draws of 6 one digit numbers, for some large value n, then the random variable

  $$\frac{1}{n} \sum_{i=1}^{n} S(X^i)$$

is *very* likely to be close to 27. But just because the average is likely to be close to 27, does not at all mean that each $S(X^i)$ need not be close to 27.

- *Assuming Independence*: This is the opposite of the law of averages, assuming certain events are not related to one another, when in fact they are. Make sure to carefully think through whether certain events are independant, before you use this in an argument.
- *Assuming Equal Likelihood*: As David Williams' says, concentration on the 'equally likely' approach to probability is an invitation to disaster.
- *The Behaviour of Ratios*: TODO.

The remainder of this section is devoted to certain famous paradoxes in probability theory.

Example. *Suppose that a person has a 1% probability of contracting the disease. A test for the disease has a 90% accuracy of success. One person is chosen at random, tested for the disease, and the test comes back positive. It may seem that the person has reason to worry, but the person is actually still more likely to not have the disease, than to have the disease. If we let D the event that the person has the disease, and P the event that the test is positive, then using Bayes' rule, we conclude*

$$\mathbb{P}(D|P) = \frac{\mathbb{P}(P|D)\cdot\mathbb{P}(D)}{\mathbb{P}(P|D)\cdot\mathbb{P}(D)+\mathbb{P}(P|D^c)\cdot\mathbb{P}(D^c)} = \frac{0.9\cdot 0.01}{0.9\cdot 0.01+0.1\cdot 0.99} = 1/12.$$

So you are actually 11 times more likely to not have the disease than to have the disease.

Chapter 2

Random Variables

As mentioned before, our goal will be to try more and more to adopt the 'probabilistic way of thinking', where measure theory is eschewed more and more. A key tool in this process is the introduction of a *random variable*, which is a Borel measurable function $X : \Omega \to \mathbb{R}$. As a rough approximation, probabilistic thinking focuses on the distribution of the *outputs* of X rather than to consider the *inputs* Ω at all, i.e. thinking of X as a 'random number'. More categorically, we might say that probability theory studies properties of objects which are 'independent' of the sample space they are taken on. As a rough approximation, if $T : \Omega_1 \to \Omega_2$ is a surjective, measure preserving map between probability spaces (we view Ω_1 as an 'extension' of the space Ω_2, allowing more outcomes), then we should consider the random variable $X : \Omega_2 \to \mathbb{R}$ as the 'same' as the random variable $X \circ T : \Omega_1 \to \mathbb{R}$, and the concepts studied in probability theory (e.g. independence) should be preserved under this extension. As we reach further and further into statistical theory, samples spaces will soon become a distant memory, brought back only for the most technical of arguments.

In some formulations, one does not require a random variable X to be

Thus we emphasize the difference between how we formally define a random variable, and how we 'think' of a random variable. Formally, a random variable is a Borel measurable function $X : \Omega \to \mathbb{R}$. But we think of a random variable as a 'random number', which takes a particular value for each point in the sample space Ω. The measurability is needed so that the 'set of questions we ask about X' are still legal in the probability space Ω, i.e. the quantities

$$\mathbb{P}(X \in A) = \mathbb{P}(X^{-1}(A))$$

are well defined for each Borel set A.

2.1 Distributions

A random variable X induces a probability measure on the Borel sigma algebra of $\mathbb{R}$, which we call the *law*, or *distribution* of X. It is defined as $\mathbb{P}_X(A) = \mathbb{P}(X^{-1}(A))$. The distribution captures the size and shape of the random variable, but not it's relation to other random variables in the sample distribution.

If X is uniformly distributed on $[0,1]$, then μ_X is the Lebesgue measure on $[0,1]$, and if X is a *discrete random variable*, in the sense that the range of X takes on only countably many values, then μ_X is a discrete measure with $\mu_X(\{a\}) = \mathbb{P}(X = a)$.

We say two random variables X and Y are *identically distributed*, sometimes written

$$X \overset{(d)}{=} Y$$

if $\mu_X = \mu_Y$. Note that X and Y need not even be defined on the same sample space, let alone be actually equal to one another. For instance, if $\Omega = \{0,\dots,6\}^2$, with the uniform measure, and X and Y are random variables with $X(a,b) = a$ and $Y(a,b) = b$, then X and Y are identically distributed, but they are not equal to one another.

2.2 Expectation

Theorem 2.1. *For any $X \geqslant 0$,*

$$\mathbf{E}[X] = \int \mathbb{P}(X \geqslant x)dx$$

Proof. Applying Fubini's theorem,

$$\begin{aligned}\int_0^\infty \mathbb{P}(X \geqslant x)dx &= \int_0^\infty \int_x^\infty d\,\mathbb{P}_*(y)\,dx \\ &= \int_0^\infty \int_0^y dx\,d\,\mathbb{P}_*(y) \\ &= \int_0^\infty y d\,\mathbb{P}_*(y) = \mathbf{E}[X]\end{aligned}$$

□

Chapter 3

Useful Distributions

3.1 Discrete Distributions

The most basic discrete distribution is the *Bernoulli distribution*, which is $\{0, 1\}$ valued. The distribution, denoted $\mathrm{Ber}(p)$, is equal to 1 with probability p, and 0 with probability $1 - p$. It represents whether a single random quantity is true or false.

Suppoe we repeat a Bernoulli trial repeatedly and independently from one another. If we repeat the trial n times, and count the number of ones, we obtain the *Binomial distribution* $\mathrm{Bin}(n, p)$, which is valued in $\{0, \ldots, n\}$. The probability of getting k successes, for $0 \leqslant k \leqslant n$, is equal to

$$\binom{n}{k} p^k (1-p)^{n-k}.$$

As a practical example, suppose we have m balls in a bucket, and k of them are red. If $p = k/m$, and we repeatedly draw balls from the bucket, *replacing them every time*, then the number of red balls we draw will follow the $\mathrm{Bin}(n, p)$ distribution.

Suppose we draw balls from the bucket *without replacement*. Then we obtain the *hypergeometric distribution* $\mathrm{Hypergeometric}(n, m, k)$. This distribution is supported on $\max(0, n + k - m) \leqslant r \leqslant \min(k, n)$, and assigns to this point a probability

$$\frac{\binom{k}{r}\binom{m-k}{n-r}}{\binom{m}{n}}.$$

Indeed, a single outcome of the experiment can be seen by choosing k successes from the K possible successes, and drawing $n - k$ failures from the $N - K$ failures.

TODO: Give bounds here showing constraints ensuring hypergeometric and geometric distributions are roughly similar to one another provided that the sample size is large. TODO: Feller Vol 1. Chapter 2. 11. Problems and Complements of a Theoretic Character. TODO: Feller Vol 1. Chapter 3 onwards.

3.2 Normal Distribution

The Normal distribution

Consider a random point in the plane given by a distribution μ which is radial, and such that with respect to this distribution, the x coordinates and y coordinates of a point are independent of one another. Then we claim μ is a normal distribution. I believe this is a result due to Maxwell.

To prove this, we may assume without loss of generality that μ is a smooth distribution, since if $G(x) = C\exp(-x^2 - y^2)$ is the Gaussian distribution, then $\mu * G$ is smooth and satisfies the same properties as μ, and if we prove that $\mu * G$ is Gaussian, then it follows that μ itself must be Gaussian.

By radial symmetry, the x and y coordinates are identically distributed. Let $f : \mathbb{R} \to [0, \infty)$ denote their common density function. Then we claim that for any $x, y \in \mathbb{R}$, if $r = (x^2 + y^2)^{1/2}$, then

$$f(r) = f(x)f(y).$$

Indeed, by independence the probability that a random point lies in a δ by δ box with center (x_0, y_0) is, by independence, $\delta^2 f(x_0)f(y_0) + O(\delta^3)$. But by rotational invariance, this is the same as the probability that a random point lies in a rotated cube centered at $(r, 0)$ with sidelengths δ by δ, which is, again by independence, $\delta^2 f(r)f(0) + O(\delta^3)$. Taking $\delta \to 0$ gives that $f(x_0)f(y_0) = f(r)f(0)$. Since x_0 and y_0 are arbitrary here, by taking $y_0 = 0$, so that $x_0 = r$, we see that we must have $f(0) = 1$, which yields the identity.

The identity above implies that if $f(t_0) = 0$, then $f(t) = 0$ for $t \geqslant t_0$. let $I \subset [0, \infty)$ be the largest interval on which f has no zeroes. Since f is smooth, taking derivatives in x to the identity above gives that

$$f'(r)(x/r) = f'(x)f(y),$$

and also taking derivatives in y yields that

$$f'(r)(y/r) = f(x)f'(y).$$

Putting these two equations together, and using the non-vanishing of f on I, we conclude that there exists a constant C such that for all $t \in I$, $f'(t) = cxf(t)$. But this implies that

$$f(t) = ae^{cx^2/2}.$$

This function is non-vanishing, so that $I = [0, \infty)$. Since f gives a probability density, we must have $c < 0$. But then f is a Gaussian.

On the other hand, the function $g(x) = e^{-x^2}$ satisfies $G'(x) = -2xg(x)$. Thus

$$(f/g)'(r) = \frac{f'(r) + 2rf(r)}{g(r)}.$$

$$|\cos(\theta)x - \sin(\theta)y - x_0| \leqslant \delta \quad \text{and} \quad |\sin(\theta)x + \cos(\theta)y|$$

and so let ν be the distribution of x and y. Then $\mu = \nu \otimes \nu$.

Then we claim μ is a normal distribution. Indeed, let ν be the distribution of $\pi_*\mu$, where $\pi : \mathbb{R}^2 \to [0, \infty)$ is given by $\pi(x) = |x|$. Then $\nu =$

$$\mathbb{P}(|x| \leqslant R) = \int_0^R r d\nu$$

$$\mu(\sqrt{x^2 + y^2}) =$$

Chapter 4

The Law of Large Numbers

Chapter 5

Tail Bounds and Concentration

In most cases, it is very difficult, if not impossible, to find a closed form formula for the probability of certain events. But often, all that is important is to obtain some upper or lower bound on the probability of some *bad* event happening, or lower bound the probability of a *good* event happening. It will be helpful to utilize Tao's notation for the probabilities of certain events. Let E be an event depending on some parameter n:

- E holds *asymptotically almost surely* if $\mathbb{P}(E) = 1 - o(1)$ as $n \to \infty$.
- E holds with *high probability* if $\mathbb{P}(E) = 1 - O(n^{-c})$ for some $c > 0$.
- E holds with *overwhelming probability* if $\mathbb{P}(E) = 1 - O_c(n^{-c})$ for *every* $c > 0$.
- E holds *almost surely* if $\mathbb{P}(E) = 1$.
- E holds *surely* if $E^c = \varnothing$.

The union bound $\mathbb{P}(E \cup F) \leqslant \mathbb{P}(E) + \mathbb{P}(F)$ shows these classes are closed under various numbers of intersections:

- If $\{E_\alpha\}$ is an *arbitrary* family of events which occur surely, then $\bigcap E_\alpha$ occurs surely.
- If $\{E_k\}$ is a *countable* family of events which occur almost surely, then $\bigcap E_k$ occurs almost surely.
- If $E_1, \ldots, E_N$ are a *polynomial* number of events and these events hold with *uniform* overwhelming probability (i.e. implicit constants are independant of events), then $\bigcap E_k$ holds with overwhelming probability. In other words, if $N = O(n^{O(1)})$, and $\mathbb{P}(E_k) = 1 - O_c(n^{-c})$ independantly of k, then for all c,

  $$\mathbb{P}\left(\bigcap E_k\right) = 1 - \mathbb{P}\left(\bigcup E_k^c\right) \geqslant 1 - N \cdot O_c(n^{-c}) = 1 - O_c(n^{O(1)-c}).$$

  Since c can be arbitrarily large, this gives the result.

- If $E_1, \dots, E_N$ are a *sub-polynomial* number of events, holding with *uniform* high probabity, then $\bigcap E_k$ holds with high probability. In other words, if $N = O(n^{o(1)})$, and there is c such that $\mathbb{P}(E_k) = 1 - O(n^{-c})$ for all k, then for any $c_0 < c$,

$$\mathbb{P}\left(\bigcap E_k\right) = 1 - \mathbb{P}\left(\bigcup E_k^c\right) \geqslant 1 - O(n^{o(1)-c}) = 1 - O(n^{-c_0}).$$

- If $E_1, \dots, E_N$ are a *finite* number of events, holding asymptotically almost surely, then $\bigcap E_k$ holds asymptotically almost surely. In other words, if $N = O(1)$, then

$$\mathbb{P}\left(\bigcap E_k\right) = 1 - N \cdot o(1) = 1 - o(1).$$

Notice that as an event becomes more certain, we are more free to apply intersections over a larger family of intersections. Event that are likely remain likely under suitable types of condition, because

$$\mathbb{P}(F|E) = \frac{\mathbb{P}(F \cap E)}{\mathbb{P}(E)} \leqslant \frac{\mathbb{P}(F)}{\mathbb{P}(E)}$$

In particular, let F and E be events depending on a parameter n:

- If F occurs almost surely, it occur almost sure conditioned on E.

- If F occurs with overwhelming probability, it occurs with overwhelming probability conditioned on E provided that $\mathbb{P}(E) \gtrsim n^{-C}$ for some $C > 0$, since

$$\mathbb{P}(F^c|E) \leqslant \frac{\mathbb{P}(F^c)}{\mathbb{P}(E)} = \frac{O_c(n^{-c})}{O(n^{-C})} = O_c\left(n^{-(c-C)}\right)$$

- If F occurs with high probability, it occurs with high probability conditioned on E provided that $\mathbb{P}(E) \gtrsim n^{-c_0}$ for some sufficiently small c_0 depending on the decay of F. If $\mathbb{P}(F) = 1 + O(n^{-c})$, and $c_0 < c$, then

$$\mathbb{P}(F^c|E) \leqslant \frac{\mathbb{P}(F^c)}{\mathbb{P}(E)} \leqslant O\left(n^{-(c-c_0)}\right)$$

- If F occurs asymptotically almost surely, it occurs asymptotically almost surely conditioned on E, provided that $\mathbb{P}(E) \gtrsim 1$.

None of the consequences of this list involve any assumptions about the interactions of the evens. Given more interesting iteractions between the events, one can likely strengthen the bounds obtained here, e.g. via the method of *chaining*, which allows one to even obtain bounds on *uncountable intersections* given certain dependence information on the events.

Similarily, we can discuss probabilistic bounds on random variables. We say $X_n \lesssim_{\mathbb{P}} Y_n$ if for all $\varepsilon > 0$, there exists a constant C_ε such that

$$\mathbb{P}(X_n \leqslant C_\varepsilon Y_n) \geqslant 1 - \varepsilon.$$

Notations such as $\gtrsim_{\mathbb{P}}$, $\sim_{\mathbb{P}}$, $O_{\mathbb{P}}$, $\Omega_{\mathbb{P}}$, $\Theta_{\mathbb{P}}$ are defined similarily. We write $X_n = o(Y_n)$ if for all $\varepsilon > 0$, $X_n \leqslant \varepsilon Y_n$ asymptotically almost surely.

5.1 Tail Bounds

Tail bounds upper bound the probability that a random variable will take large values. Markov's inequality is a basic version, which connects the expectation of a random variable with a tail bound.

Theorem 5.1 (Markov's Inequality)**.** *If $X \geqslant 0$, then $\mathbb{P}(X \geqslant t) \leqslant \mathbf{E}(X)/t$.*

Proof. By Fubini's theorem,

$$\mathbf{E}(X) = \int \int_0^X dt \, d\mathbb{P} = \int_0^\infty \int \mathbf{I}(t \leqslant X) \, d\mathbb{P} \, dt = \int_0^\infty \mathbb{P}(X \geqslant t) \, dt$$

Markov's inequality is thus obtained as a weak-type inequality. □

Remark. The integral identity

$$\mathbf{E}X = \int_0^\infty \mathbb{P}(X > t) \, dt$$

also enables one to bound $\mathbb{E}(X)$ in terms of a tail bound.

One refers to applications of Markov's inequality as a *first moment* calculation, to compare with other methods for bounding tail probabilities. In general, if we bound $\mathbb{E}(X^p)$ for some $p > 0$, then Markov's inequality shows that

$$\mathbb{P}(X \geqslant t) = \mathbb{P}(X^p \geqslant t^p) \leqslant \mathbb{E}(X^p)/t^p$$

As p becomes larger, this inequality gives a sharper rate of decay on the random variable. This is known as a *pth moment* bound. We can be even more clever in this respect. If $\mathbb{E}(e^{\lambda X})$ is finite for some $\lambda > 0$, then

$$\mathbb{P}(X \geqslant t) = \mathbb{P}(e^{\lambda X} \geqslant e^{\lambda X}) \leqslant \mathbf{E}(e^{\lambda X}) e^{-\lambda t}$$

This is known as a *Chernoff bound*.

Remark. The linearity of expectation makes applying Markov's inequality quite simple. But higher order moments are non-linear, and so often one needs to analyze the distribution of X much more carefully. Chernoff bounds are even more difficult to obtain, but much more useful.

Remark. Technically, the dominated convergence theorem enables one to conclude the sharper results than given here. For instance, if $\mathbb{E}\, X^p < \infty$, then $\mathbb{P}(X > t) = o(1/t^p)$, and if $\mathbb{E}(e^{\lambda X}) < \infty$, then $\mathbb{P}(X > t) = o(e^{\lambda t})$. But the dominated convergence theorem gives no control on the rates of this little o term, which makes it relatively ineffective.

We can quantify tail bounds of a random variable as follows::

- If X has bounded expectation, then for any $\varepsilon > 0$, $|X| = O(n^\varepsilon)$ with uniform high probability.

- If X has bounded p'th moment, then $|X| = O(n^{\varepsilon})$ with probability $1 - O(n^{-\varepsilon p})$.
- If $\mathbf{E}(e^{\lambda X})$ is bounded, then $|X| = O(\log^{O(1)} n)$ with overwhelming probability.

Very similar methods establish upper bounds on $\mathbb{P}(|X - \mathbf{E}X| \geqslant t)$, known as a *deviation inequality*. The equivalent to moment bounds in the deviation regime is Chebyshev's inequality.

Theorem 5.2 (Chebyshev's Inequality)**.** *If X has finite variance σ^2, then*

$$\mathbb{P}(|X - \mathbb{E}(X)| \geqslant t) \leqslant \sigma^2/t^2$$

More generally,

$$\mathbb{P}(|X - \mu| \geqslant t) \leqslant \frac{\mathbb{E}\,|X - \mu|^p}{t^p} \quad \textit{and} \quad \mathbb{P}(|X - \mu| \geqslant t) \leqslant \mathbb{E}(e^{\lambda|X-\mu|})e^{-\lambda t}.$$

Proof. Applying Markov's inequality, we find

$$\mathbb{P}(|X - \mu| \geqslant t) = \mathbb{P}(|X - \mu|^2 \geqslant t^2) \leqslant \frac{\mathbf{E}|X - \mu|^2}{t^2} = \frac{\sigma^2}{t^2}.$$

The other equations was treated similarily. □

Example. *A* Median *for a random variable X is a quantity M such that*

$$\mathbb{P}(X \geqslant M), \mathbb{P}(X \leqslant M) \leqslant 1/2$$

For an absolutely continuous random variable, the median always exists and is unique. If $X \geqslant 0$ has a median M, and finite variance σ^2, Chebyshev's inequality implies

$$\mathbb{P}(|X - \mu| \geqslant \sqrt{2} \cdot \sigma) \leqslant 1/2$$

Thus $\mu - \sqrt{2} \cdot \sigma \leqslant M \leqslant \mu + \sqrt{2} \cdot \sigma$, so $M = \mu + O(\sigma)$. This explains why in most circumstances, the median rarely shows up in statistical estimates, because they are equivalent up to a small number of standard deviations.

When t is large, a bound on $\mathbb{P}(|X - \mathbf{E}X| \geqslant t)$ is known as a large deviation inequality. Chebyshev's inequality becomes useful only when t exceeds a single standard deviation, i.e. $t \geqslant \sigma$, and is therefore an example of a large deviation inequality. Second moment methods are often useful for bounding sums of independant random variables because if X and Y are independant, then $\mathbf{V}(X+Y) = \mathbf{V}(X)+\mathbf{V}(Y)$ is easy to calculate. We can continue establishing large deviation bounds, obtaining $\mathbb{P}(|X - \mathbf{E}X| > t) \lesssim 1/t^p$ by bounding $\mathbf{E}|X - \mathbf{E}X|^p$. We will refer to any of these inequalities as Chebyshev bounds.

5.2 Moments of Sums of Random Variables

We now study the tail inequalities for a sum $S_n = X_1 + \cdots + X_n$ of random variables, using the moment methods we have established. For simplicity, we assume all variables X_k have mean zero, so that deviation and tail inequalities agree with one another.

Using the first moment method, which requires no assumption on the interdependence of the X_k, we find $\mathbb{E}|S_n| \leqslant \mathbb{E}|X_1| + \cdots + \mathbb{E}|X_n|$. Combined with Markov's inequality, we conclude

$$\mathbb{P}(|S_n| \geqslant t) \leqslant \frac{\mathbb{E}|X_1| + \cdots + \mathbb{E}|X_n|}{t}$$

Nonetheless, this inequality can be sharp, especially given no independance between the random variables. which causes huge fluctuations in the magnitude of S_n. For instance, if all X_k are all equal to a single random variable $X \in \{-1, 1\}$, then $|S_n| = n$, so Markov's inequality is sharp when $t = n$.

Now we consider what the second moment method gives us. Note that

$$\mathbb{E}(S_n^2) = \sum_{i=1}^{n}\sum_{j=1}^{n} \mathbb{E}(X_i X_j)$$

If we assume the variables are pairwise independent, then we have $\mathbb{E}(X_i X_j) = 0$ for $i \neq j$, so $\mathbb{E}(S_n^2) = \sum \mathbb{E}(X_i^2)$. This is just the familiar statement that variance is additive. Chebyshev's inequality then yields that

$$\mathbb{P}(|S_n| \geqslant t) \leqslant \frac{\mathbb{V}(X_1) + \cdots + \mathbb{V}(X_n)}{t^2}$$

Informally, this means that with high probability,

$$S_n \lesssim \mathbb{V}(S_n)^{1/2} = (\mathbb{V}(X_1) + \cdots + \mathbb{V}(X_n))^{1/2}).$$

We cannot expect to obtain a much sharper region of concentration for S_n than that obtained by Chebyshev's inequality, because

$$\mathbb{V}(X_1) + \cdots + \mathbb{V}(X_n) = \mathbb{V}(S_n) = \int_0^\infty 2t \cdot \mathbb{P}(|S_n| \geqslant t)\, dt.$$

Thus we expect $\mathbb{P}(|S_n| \geqslant t) \gtrsim \mathbb{V}(S_n)/2t$ for significantly many values t.

Example. *Consider random variables $\{X_1, \ldots, X_n\}$, and let $S_n = X_1 + \cdots + X_n$. For simplicity, assume $|S_n| \lesssim n$ almost surely, and that $\mathbb{V}(S_n) \gtrsim n$. Suppose that $|S_n| \leqslant A_n$ with high probability for some $A_n > 0$, i.e. $\mathbb{P}(|S_n| \geqslant A) \lesssim 1/n^{1+\varepsilon}$ for some $\varepsilon > 0$. Then*

$$\begin{aligned}
\mathbb{V}(S_n) = \int_0^\infty 2t\mathbb{P}(|S_n| \geqslant t) &\lesssim \int_0^{A_n} 2t + \int_{A_n}^{O(n)} 2t \cdot n^{-1-\varepsilon} \\
&\lesssim A_n^2 + (1/n^{1+\varepsilon})(O(n^2) - A_n^2) \\
&\lesssim (1 - 1/n^{1+\varepsilon})A_n^2 + n^{1-\varepsilon}.
\end{aligned}$$

Since $\mathbb{V}(S_n) \gtrsim n$, we conclude that

$$A_n^2 \gtrsim \frac{\mathbb{V}(S_n) - n^{1-\varepsilon}}{1 - 1/n^{1+\varepsilon}} \gtrsim \mathbb{V}(S_n).$$

Thus the best concentration bound we can hope for concentrates in a region of length $O(\mathbb{V}(S_n)^{1/2})$.

The next example shows the tail bound is sharp in particular circumstances, especially when only pairwise independence is satisfied.

Example. *For each nonzero $k \in \{0,1\}^m$, let $X_k = (-1)^{k \cdot Y}$, where $Y = (Y_1, \ldots, Y_m)$ is drawn uniformly at random from the hypercube $\{0,1\}^m$. If $k \neq 0$, then there is j such that $k_i = 1$, and then $\mathbb{E}((-1)^{k_i Y_i}) = \mathbb{E}((-1)^{Y_i}) = 0$, so*

$$\mathbb{E}(X_k) = \prod_{j=1}^{m} \mathbb{E}((-1)^{k_j Y_j}) = 0$$

Also note $\mathbb{V}(X_k) = 1$ since $X_k^2 = 1$. If $k_i = 1$, then X_k is independant of the σ algebra generated by $\{Y_j : j \neq i\}$. This is because by looking at combinatorics on the hypercube, for any $y \in \{0,1\}^m$, where we view the hypercube as a vector space over $\mathbb{Z}_2$,

$$\begin{aligned}\mathbb{P}\left(\{X_k = 1\} \cap \{Y_j = y_j : j \neq i\}\right) &= \mathbb{P}\left(\{Y_i = k \cdot y - y_i\} \cap \{Y_j = y_j : j \neq i\}\right) \\ &= 1/2^m = \mathbb{P}(X_k = 1)\mathbb{P}\left(\bigcap_{j \neq i} Y_j = y_j\right)\end{aligned}$$

If $l \in \{0,1\}^m$ satisfies $l_i = 0$, then X_l is measurable with respect to $\sigma(Y_j : j \neq i)$, which implies it is independant to X_k. Thus the random variables are pairwise independant. On the other hand,

$$S = \sum_{k \in \{0,1\}^m} X_k = 2^m \cdot \mathbf{I}(Y = 0)$$

is the sum of pairwise independant random variables with mean one. By pairwise independance, it's variance is $2^m - 1$. Chebyshev's inequality gives

$$\frac{1}{2^m} = \mathbb{P}(S \geqslant 2^m) \leqslant \frac{1}{2^m - 1}$$

Thus the inequality is sharp here, up to constants which are insignificant for large m.

Regardless of the sharpness of the region of concentration, we can still hope to obtain sharper results for the tail bounds of random variables. To do this, we analyze higher moments of S_n. For simplicity, we now assume that each X_k has unit variance, still has mean zero, and furthermore, $|X_k| \leqslant C$ for some constant C. We note that if k is an even integer, then

$$\mathbb{E}\,|S_n|^k = \sum \{\mathbb{E}(X_{i_1} \ldots X_{i_k}) : 1 \leqslant i_1, \ldots, i_k \leqslant n\}$$

Suppose the random variables are k-wise independant, i.e. every collection of k random variables is independant from one another, this is fairly easy to calculate. Note that if some i_j appears only a single time in $(i_1, \ldots, i_k)$, then $\mathbb{E}(X_{i_1} \ldots X_{i_k}) = 0$. Thus at most $k/2$ distinct random variables appear in terms with nonzero expectation. If exactly $k/2$ distinct variables appear, then each random variable appears exactly twice, and since each X_k has unit variance, we conclude $\mathbb{E}(X_{i_1} \ldots X_{i_k}) = 1$. More generally, if $k/2 - r$ distinct variables appear, then we use the fact that $X_k \leqslant K$ together with the unit variance bound to conclude $\mathbb{E}(X_{i_1} \ldots X_{i_k}) \leqslant K^{2r}$. If N_r gives the total number of $(i_1, \ldots, i_k)$ such that $k/2 - r$ terms appear, then

$$\mathbb{E}\,|S_n|^k \leqslant \sum_{r=0}^{k/2} K^{2r} N_r$$

To estimate N_r, we note that there are $\binom{n}{k/2-r}$ ways to choose $k/2-r$ integers from $\{1, \ldots, n\}$. Once these numbers are fixed, each integer i_j must come from one of these $k/2 - r$ integers. Thus we obtain a crude bound,

$$N_r \leqslant \binom{n}{k/2-r} (k/2-r)^k \leqslant (en)^{k/2-r} (k/2)^{k/2+r},$$

where we have applied Stirling's formula. Thus

$$\mathbb{E}\,|S_n|^k \leqslant (enk/2)^{k/2} \sum_{r=0}^{k/2} (K^2 k/en)^r.$$

If $K^2 \leqslant n/k$, which holds for suitably large n, then we conclude

$$\mathbb{E}\,|S_n|^k \leqslant 2(enk/2)^{k/2}.$$

Thus

$$\mathbb{P}\left(|S_n| \geqslant t \cdot \sqrt{2nk/e}\right) \leqslant 2/t^k \quad \text{and} \quad \mathbb{P}\left(|S_n| \geqslant t\sqrt{n}\right) \leqslant \frac{2(enk/2)^{k/2}}{t^k}.$$

Note that as k increases, the rate of decay in t improves, but the range that S_n concentrates in grows from $O(\sqrt{n})$ to $O(\sqrt{nk})$.

Now we assume that the random variables $\{X_k\}$ are actually jointly independant. Then we can apply the last method for any k, and optimize for each value of t. In particular, if $\sqrt{nk}$ is a small multiple of t, then we obtain

$$\mathbb{P}\left(|S_n| \geqslant t \cdot \sqrt{n}\right) \leqslant C \exp(-ct^2)$$

for some constants $c, C > 0$. Thus we obtain square-exponential decay by controlling all the individual moments of S_n. A slicker way to do this is using exponential moments, as we do in the next section.

5.3 Concentration of Measure

We have seen that given a large number of random variables $X_1, \ldots, X_N$, with $X_k = O(1)$ almost surely for each k, then even though the random variable S_N ranges over an interval of length $O(N)$, it has a high probability of concentrating in a $\sqrt{N}$ region. The basic intuition is that for large N, it is difficult for independant random variables to 'work together' to push the function too far from it's mean. The tail bounds tend to be *sub-gaussian*, in the sense that S_N deviates t standard deviations from the mean with probability $O(e^{-ct^2})$. Concentration of measure is a primary tool for ensuring various events hold with overwhelming probability.

Lemma 5.3. *If X is centrally distributed and takes values in $[a,b]$, then*

$$\mathbb{E}(e^{\lambda X}) \leqslant e^{\lambda^2 (b-a)^2/8}.$$

Proof. By scale homogeneity, we may assume $b-a = 1$, so it suffices to prove that $\mathbb{E}(e^{\lambda X}) \leqslant e^{\lambda^2/8}$. Write $X = \Lambda a + (1-\Lambda)b$ for some random value $0 \leqslant \Lambda \leqslant 1$. Applying convexity,

$$e^{\lambda X} \leqslant \Lambda e^{\lambda a} + (1-\Lambda)e^{\lambda b}$$

Hence

$$\mathbb{E}(e^{\lambda X}) \leqslant be^{\lambda a} - ae^{\lambda b} = e^{\lambda a}(b - ae^{\lambda}) = e^{L(\lambda)}$$

where $L(x) = ax + \ln(b - ae^x) = ax + \ln(a + 1 - ae^x)$. Note $L(0) = L'(0) = 0$, and if $x \geqslant 0$,

$$L''(x) = \frac{-a(a+1)e^x}{(1+a-ae^x)^2} \leqslant 1/4$$

By Taylor's theorem, we conclude $L(x) \leqslant x^2/8$, and so $\mathbb{E}(e^{\lambda X}) \leqslant e^{\lambda^2/8}$, which completes the proof of the result. □

Theorem 5.4 (Hoeffding's Inequality)**.** *Let $X_1, \ldots, X_n$ be centrally distributed independant random variables, with $a_i \leqslant X_i \leqslant b_i$. Then*

$$\mathbb{P}\left(\sum X_i \geqslant t\right) \leqslant e^{-2t^2/\sum (b_i - a_i)^2}$$

Proof. For any $\lambda > 0$, a Chernoff bound gives that

$$\mathbb{P}\left(\sum X_i \geqslant t\right) \leqslant e^{-\lambda t} \prod \mathbf{E}(e^{\lambda X_i})$$

The last lemma thus allows us to conclude

$$\mathbb{P}\left(\sum X_i \geqslant t\right) \leqslant e^{-\lambda t} \prod_{i=1}^{n} e^{\lambda^2 (b_i - a_i)^2/8}$$

In particular, if

$$\lambda = \frac{t}{\sum (b_i - a_i)^2/4}$$

we obtain the required inequality. □

Suppose that $a_i \leqslant X_i \leqslant b_i$, but X_i is actually much more likely to concentrate in a much smaller region. Then Chernoff's bound gives tighter results, provided one can give better bounds on $\mathbb{E}(e^{tX_i})$ for each i. For instance, if the X_i are $\{0,1\}$ valued Bernoulli variables with parameters p_i, and $t > \mu$, where $\mu = \sum \mathbb{E}(X_i)$. Then one calculates that

$$\mathbb{E}(e^{\lambda X_i}) = p_i e^{\lambda} + 1 - p_i = 1 + (e^{\lambda} - 1)p_i \leqslant \exp((e^{\lambda} - 1)p_i)$$

If we sum up over the X_i, then $\mathbb{E}(e^{\lambda S_N}) \leqslant \exp((e^{\lambda} - 1)\mu)$. For any $t > 0$, picking $\lambda = \ln(t/\mu)$ gives

$$\mathbb{P}(S_N \geqslant t) \leqslant e^{-\mu}(e\mu/t)^t.$$

This tends to be tighter for smaller deviation values of t when μ is small.

Example. *Suppose we are performing a test of some property, which succeeds in obtaining the correct answer with probability $1/2 + \delta$, where δ is small. Then we obtain a stronger test by performing the test independantly N times, and taking the majority vote. If $X_1, \ldots, X_N$ denotes the $\{-1, 1\}$ valued outcome of the particular tests, our new test is just $sgn(S_N)$, where $S_N = X_1 + \cdots + X_N$ (We also assume here that N is odd, so that S_N never equals zero). Without loss of generality, if the property we are testing is true, then $\mathbb{P}(X_n = 1) = 1/2 + \delta$, and $\mathbb{P}(X_n = -1) = 1/2 - \delta$. Thus the random variable has expectation 2δ. If $\Delta_n = X_n - 2\delta$, then Δ_n is centrally distributed, and $S_N = \sum \Delta_n + 2N\delta$. By Hoeffding's inequality, we conclude that*

$$\mathbb{P}(\textit{Test fails}) = \mathbb{P}\left(S_N \leqslant 0\right) = \mathbb{P}\left(\sum \Delta_n \leqslant -2N\delta\right) \leqslant e^{-8N\delta^2}$$

In particular, if we want a bound like $\mathbb{P}(\textit{Test fails}) \leqslant \varepsilon$, then we need only choose $N = \Omega(\delta^{-2}\log(1/\varepsilon))$. This is why when performing a statistical test, a probability of success any amount higher than 50% is essentially comparable to a 99% probability of success.

The full assumption of independance is not necessary for a Chebyshev-type tail bound. In fact, it suffices that the differences $X_{i+1} - X_i$ can depend on the random variables $X_1, \ldots, X_i$, but have mean zero once conditioned on the X_i. This means that $\{S_n\}$ is a *martingale*. *Azuma's inequality* gives such a bound.

Theorem 5.5. *Let $\{S_k\}$ be a martingale (or super-martingale) with $|S_k - S_{k-1}| < a_k$ almost surely, and $S_0 = 0$. Then*

$$\mathbb{P}(S_n \geqslant t) \leqslant \exp\left(\frac{-2t^2}{\sum a_i^2}\right).$$

Proof. We again consider the exponential moment method. We apply Hoeffding's lemma since $|S_n - S_{n-1}| \leqslant a_n$ to conclude

$$\begin{aligned}
\mathbb{E}(\exp(\lambda S_n)) &= \mathbb{E}(\exp(\lambda(S_n - S_{n-1}))\exp(S_{n-1})) \\
&= \mathbb{E}(\exp(\lambda S_{n-1})\,\mathbb{E}(\exp(\lambda(S_n - S_{n-1}))|S_{n-1})) \\
&\leqslant e^{a_n^2\lambda^2/8}\,\mathbb{E}(\exp(\lambda S_{n-1}))
\end{aligned}$$

Iterating this argument gives

$$\mathbb{E}(e^{\lambda S_n}) \leqslant \exp\left(\sum a_i^2 \lambda^2/8\right)$$

And so performing a Chernoff bound gives

$$\mathbb{P}(S_n \geqslant t) \leqslant \exp\left(\sum a_i^2 \lambda^2/8\right) e^{-\lambda t}$$

Picking $\lambda = 4t/\sum a_i^2$, we conclude

$$\mathbb{P}(S_n \geqslant t) \leqslant e^{-2t^2/\sum a_i^2}. \qquad \square$$

For particular distributions, directly computing moments can sometimes give better tail bounds. For instance, here is a result using a Chernoff bound about $\mathrm{Ber}(p)$ random variables when p is small.

Lemma 5.6. *Let $X_1, \ldots, X_n$ be independent random variables, where $X_i \sim Ber(p_i)$ for $1 \leqslant i \leqslant n$. Let $S = X_1 + \cdots + X_n$, and let $\mu = \sum p_i$ be the mean of S. Then for $t > \mu$,*

$$\mathbb{P}(S \geqslant t) \leqslant e^{-\mu}\left(\frac{e\mu}{t}\right)^t,$$

and for $t < \mu$,

$$\mathbb{P}(S \leqslant t) \leqslant e^{-\mu}\left(\frac{e\mu}{t}\right)^t.$$

For $\delta \in (0,1)$, we have

$$\mathbb{P}(|S - \mu| \geqslant \delta\mu) \leqslant 2e^{-c\mu\delta^2}.$$

Proof. We have that for $\lambda > 0$,

$$\begin{aligned}
\mathbb{E}[e^{\lambda S}] &= \prod_{i=1}^n \mathbb{E}[e^{\lambda X_i}] \\
&= \prod_{i=1}^n [p_i e^{\lambda} + (1 - p_i)] \\
&\leqslant \prod_{i=1}^n (1 + p_i(e^{\lambda} - 1)) \\
&\leqslant \prod_{i=1}^n \exp\left(p_i(e^{\lambda} - 1)\right) \\
&= \exp\left(\mu(e^{\lambda} - 1)\right).
\end{aligned}$$

Thus we conclude that for any $\lambda > 0$,

$$\mathbb{P}(S \geqslant t) \leqslant \exp\left(\mu(e^{\lambda} - 1)\right) e^{-\lambda t}.$$

This bound is optimized for λ which satisfy

$$\frac{e^\lambda - 1}{\lambda} = t/\mu.$$

If t is significantly larger than μ, we we must choose λ to be large. Since for large λ, we expect that

$$\frac{e^\lambda - 1}{\lambda} \approx e^\lambda$$

this leads us to expect that we should choose $\lambda = \ln(t/\mu)$. Plugging this in for $t > \mu$ yields that

$$\mathbb{P}(S \geqslant t) \leqslant \exp(t-\mu)\,(\mu/t)^t = e^{-\mu}(e\mu/t)^t.$$

Conversely, for $t < \mu$,

$$\mathbb{P}(S \leqslant t) = \mathbb{P}(e^{-\lambda S} \geqslant e^{-\lambda t}) \leqslant \mathbb{E}[e^{-\lambda S}]e^{\lambda t}.$$

Similar to above, we calculate that

$$\mathbb{E}[e^{-\lambda S}] = \prod_{i=1}^{n}(1 - p_i(1-e^{-\lambda})) \leqslant \exp(-\mu(1-e^{-\lambda})).$$

Thus we find that

$$\mathbb{P}(S \leqslant t) \leqslant \exp(-\mu(1-e^{-\lambda}))e^{\lambda t}.$$

Setting $\lambda = \ln(\mu/t)$ yields that

$$\mathbb{P}(S \leqslant t) \leqslant e^{-\mu}(e\mu/t)^t.$$

Finally, we have

$$\mathbb{P}(|S-\mu| \geqslant \delta\mu) = \mathbb{P}(S \leqslant (1-\delta)\mu) + \mathbb{P}(S \geqslant (1+\delta)\mu).$$

Applying both deviation principles we have already proved yields the final result. □

Note that these tails behave like the *Poisson distribution*, i.e. if $X \sim \text{Poisson}(\mu)$, then for $t > \mu$,

$$\mathbb{P}(X \geqslant t) \leqslant e^{-\mu}\left(\frac{e\mu}{t}\right)^t.$$

Thus we begin to see how the Poisson distribution arises as a limit of a sum of Bernoulli distributions with very small mean.

5.4 Subgaussian Random Variables

Hoeffding's inequality only applies to bounded random variables. But we should still obtain fast tail decay in other circumstances. For instance, a Gaussian distribution has tail decay comparable to Hoeffding's inequality, and since the central limit theorem gives intuition saying that scaled sums of random variables begin to behave like a normal distribution, we should expect fast tail decay for much more general families of sums. If $g \sim N(0, \sigma^2)$, we calculate

$$\mathbb{P}(g - \mu \geqslant y) = \int_y^\infty \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} dx \leqslant \frac{1}{y\sqrt{2\pi\sigma^2}} \int_y^\infty x e^{-\frac{x^2}{2\sigma^2}} dx = \frac{\sigma e^{-\frac{y^2}{2\sigma^2}}}{y\sqrt{2\pi}}$$

This quantity is almost always better than Chebyshev's inequality, since the ratio $1/y$, which measures the inaccuracy of our inequality, is nullified by the exponential function. We can find similar equalities for random variables which are 'dominated' by normal distributions. We call these random variables *subgaussian*.

Theorem 5.7. *The following are equivalent, for comparable constants $K_i > 0$.*

(1) $\mathbb{P}(|X| \geqslant t) \leqslant 2\exp(-t^2/K_1^2)$.

(2) $\|X\|_p \leqslant K_2\sqrt{p}$, *for* $p \in [1, \infty)$.

(3) $M_{X^2}(\lambda^2) \leqslant \exp(K_3^2\lambda^2)$ *if* $|\lambda| \leqslant 1/K_3$.

(4) $M_{X^2}(1/K_4^2) \leqslant 2$.

(5) If, in addition, $\mathbf{E}X = 0$, $M_X(\lambda) \leqslant \exp(K_5^2\lambda^2)$.

If any of these equations hold, we say X is subgaussian.

Proof. We provide a proof that the inequalities implies one another, up to a constant change of coefficients.

- Suppose (1) holds. Then

$$\begin{aligned}\|X\|_p^p &= \int_0^\infty \mathbb{P}(|X| \geqslant \lambda^{1/p})\, d\lambda \leqslant 2\int_0^\infty \exp(-(\lambda^{1/p}/K_1)^2)\, d\lambda \\ &= 2K_1^p p \int_0^\infty \lambda^p \exp(-\lambda^2)\frac{d\lambda}{\lambda} = K_1^p p \cdot \Gamma(p/2) \leqslant K_1^p p (p/2)^{p/2}\end{aligned}$$

  Thus $\|X\|_p \leqslant (p^{1/p}2^{-1/2})\sqrt{p}K_1 \leqslant (e^{1/e}2^{-1/2})\sqrt{p}K_1$, so we can set $K_2 \leqslant 2^{-1/2}e^{1/e}K_1$.

- Suppose (2) holds. Using Stirling's approximation, we compute

$$\begin{aligned}M_{X^2}(\lambda^2) &= \sum_{k=0}^\infty \frac{\mathbf{E}(X^{2k})\lambda^{2k}}{k!} = \sum_{k=0}^\infty \frac{\|X\|_{2k}^{2k}\lambda^{2k}}{k!} \leqslant \sum_{k=0}^\infty k^k \frac{(2K_2^2\lambda^2)^k}{k!} \\ &\leqslant \sum_{k=0}^\infty k^k \frac{(2K_2^2\lambda^2)^k}{(2\pi)^{1/2}k^{k+1/2}e^{-k}} \leqslant \sum_{k=0}^\infty (2eK_2^2\lambda^2)^k\end{aligned}$$

For $|\lambda| \leqslant 1/2e^{1/2}K_2$, since $(1-x)^{-1} \leqslant \exp(2x)$ if $0 \leqslant x \leqslant 1/2$, we conclude

$$M_{X^2}(\lambda^2) = \frac{1}{1-2eK_2^2\lambda^2} \leqslant \exp(4eK_2^2\lambda^2)$$

Thus we can set $K_3 \leqslant 2e^{1/2}K_2$.

- The fact that (3) implies (4) is very simple, since $M_{X^2}(\log(2)/K_3^2) \leqslant \exp(\log(2)) = 2$. Thus we can set $K_4 \leqslant K_3/\log(2)^{1/2}$.

- The fact that (4) implies (1) is also pretty simple, since a Chernoff type bound gives

$$\mathbb{P}(|X| \geqslant t) = \mathbb{P}(e^{|X|^2/K_4^2} \geqslant e^{t^2/K_4^2}) \leqslant M_{X^2}(1/K_4^2)e^{-t^2/K_4^2} \leqslant 2e^{-t^2/K_4^2}$$

So we can set $K_1 \leqslant K_4$. This completes the argument that (1) through (4) are all equivalent.

- Now we show (3) implies (5), assuming X is centrally distributed. If $|\lambda| \leqslant 1/K_3$, we use the inequality $e^{\lambda x} \leqslant \lambda x + e^{\lambda^2 x^2}$ to conclude

$$M_X(\lambda) \leqslant \lambda \mathbf{E}X + M_{X^2}(\lambda^2) = M_{X^2}(\lambda^2) \leqslant \exp(K_3^2\lambda^2)$$

If $|\lambda| \geqslant 1/K_3$, we use the inequality $\lambda x \leqslant \lambda^2/2 + x^2/2$, so that

$$M_X(\lambda) \leqslant e^{K_3^2\lambda^2/2}M_{X^2}(1/2K_3^2) \leqslant e^{K_3^2\lambda^2/2}e^{1/2} \leqslant e^{K_3^2\lambda^2}$$

Thus we may set $K_5 \leqslant K_3$.

- Suppose (5) holds. Then a Chernoff bound gives

$$\mathbb{P}(X \geqslant t) \leqslant M_X(\lambda)e^{-\lambda t} \leqslant e^{K_5^2\lambda^2 - \lambda t}$$

If we set $\lambda = t/2K_5^2$, we find $\mathbb{P}(X \geqslant t) \leqslant e^{-t^2/4K_5^2}$. Thus $\mathbb{P}(|X| \geqslant t) \leqslant 2e^{-t^2/4K_5^2}$, and so we can set $K_1 \leqslant 2K_5$.

In particular, if we check all the constants, we see that $K_i \leqslant 10K_j$ for any minimal choice of K_i and K_j. □

The natural way to form a measure of being subgaussian is to come up with an Orlicz norm. Given a convex, increasing function $\psi : [0,\infty) \to [0,\infty)$, with $\psi(0) = 0$, and $\psi(x) \to \infty$ as $x \to \infty$. We define the Orlitz norm $\|X\|_\psi$ corresponding to ψ as the infinum of t such that $\mathbf{E}(\psi(t^{-1}|X|)) \leqslant 1$.

Theorem 5.8. *The Orlicz norm is actually a norm, and is also complete.*

Proof. It is obvious that $\|\lambda X\|_\psi = |\lambda|\|X\|_\psi$. To obtain the triangle inequality, we note that if $\mathbf{E}(\psi(t^{-1}|X|)) \leqslant 1$, and $\mathbf{E}(\psi(s^{-1}|Y|)) \leqslant 1$, then convexity implies

$$\mathbf{E}\left(\psi\left(\frac{|X+Y|}{s+t}\right)\right) \leqslant \mathbf{E}\left(\frac{|X|+|Y|}{s+t}\right) \leqslant \frac{t\mathbf{E}(t^{-1}|X|) + s\mathbf{E}(s^{-1}|Y|)}{s+t} \leqslant 1$$

This gives $\|X+Y\|_\psi \leqslant 1$. If $\|X\|_\psi = 0$, then the increasing nature of ψ makes it easy to see that $X = 0$ almost surely. If $X_1, X_2, \ldots$ is a Cauchy sequence with respect to the Orlitz norm, we may thin the sequence out so that $\|X_{n+1} - X_n\|_{\psi_2} \leqslant 1/2^n$. This means that $S = \sum |X_{n+1} - X_n|$ has finite Orlitz norm, because if $S_N = \sum_{n \leqslant N} |X_{n+1} - X_n|$, then $\|S_N\|_\psi \leqslant 1$, and by the monotone convergence theorem $\mathbf{E}(\psi(S)) = \lim \mathbf{E}(\psi(S_N)) \leqslant 1$. Thus $\|S\|_\psi \leqslant 1$. This means that S is finite almost everywhere, so $\sum X_{n+1} - X_n$ converges absolutely almost everywhere. Thus X_n converges almost everywhere to some X, and monotone convergence implies X_n converges to X in the Orlicz norm. But because the original sequence is Cauchy, this means the original sequence also converges to X. □

We can use the Orlicz norms to provide a norm measuring how subgaussian a random variable is. It is obtained by considering the convex, increasing function $\psi_2(x) = e^{x^2}/2$. Naturally, we let $\|X\|_{\psi_2}$ denote this norm. Because $\|X\|_{\psi_2} < \infty$ if and only if X is subgaussian, we think of the ψ_2 norm as being a measure of how 'subgaussian' a random variable is.

For the subspace of centered subgaussian random variables, we can use an alternate norm. For a subgaussian random variable X with $\mathbb{E}(X) = 0$, we let $\sigma(X)$ denote the smallest value σ such that $M_X(\lambda) \leqslant \exp((\sigma\lambda)^2/2)$. Of course, $\sigma(X)$ and $\|X\|_{\psi_2}$ are comparable to one another, in the sense that

$$\sigma(X) \leqslant \|X\|_{\psi_2} \leqslant 10\sigma(X)$$

But utilizing σ often makes certain calculations simpler. For instance, if X is centered, it's moment generating function is globally defined, so we can write

$$M_X(\lambda) = \mathbb{E}(\exp(\lambda X)) = \sum_{k=0}^{\infty} \mathbb{E}(X^k)\frac{\lambda^k}{k!} = 1 + \mathbb{V}(X)\lambda^2/2 + O(\lambda^3).$$

On the other hand,

$$\exp(\sigma(X)^2\lambda^2) = \sum_{k=0}^{\infty} \sigma(X)^{2k}\frac{\lambda^{2k}}{k!} = 1 + \sigma(X)^2\lambda^2 + O(\lambda^4).$$

Since $M_X(\lambda) \leqslant \exp(\sigma(X)^2\lambda^2)$ for all λ, if we take $\lambda \to 0$, we conclude that $\mathbb{V}(X) \leqslant \sigma(X)^2$.

It is obvious that σ is a homogenous quantity. To prove it is a norm, it therefore suffices to prove a triangle ienquality. If $\sigma(X+Y) \leqslant \sigma(X) + \sigma(Y)$, we apply Hölder's inequality to conclude that if $p^{-1} + q^{-1} = 1$,

$$\begin{aligned} M_{X+Y}(\lambda) &= \mathbf{E}(e^{\lambda X}e^{\lambda Y}) \leqslant \mathbf{E}(e^{p\lambda X})^{p^{-1}}\mathbf{E}(e^{q\lambda Y})^{q^{-1}} \\ &\leqslant \exp(p^{-1}(p\lambda\sigma(X))^2 + q^{-1}(q\lambda\sigma(Y))^2/2) \\ &\leqslant \exp((p\sigma(X)^2 + q\sigma(Y)^2)\lambda^2/2) \end{aligned}$$

Choosing $p = 1 + \sigma(X)/\sigma(Y)$ gives $\sigma(X+Y) \leqslant \sigma(X) + \sigma(Y)$. If X and Y are independant, we actually conclude that $X+Y$ is $\sqrt{\sigma^2 + \tau^2}$ subgaussian, because we needn't apply Hölder's inequality, instead computing

$$M_{X+Y}(\lambda) = M_X(\lambda) M_Y(\lambda) \leqslant \exp(\lambda^2(\sigma^2 + \tau^2)/2)$$

Thus we get *square root cancellation*.

Example. *Let $X_1, \ldots, X_n$ be independant subgaussian ,s and let $S = \sum X_i$. Then*

$$\mathbb{E}[S] = \sum_i \mathbb{E}[X_i] \quad \textit{and} \quad \sigma(S) \leqslant \left(\sum_i \sigma(X_i) \right)^{1/2}.$$

Thus we conclude that

$$\mathbb{P}(|S - \mathbb{E}(S)| \geqslant t) \leqslant 2 \exp\left(\frac{-t^2}{\sum \sigma(X_k)^2} \right)$$

Provided that we have a uniform bound $\sigma(X_k) \lesssim 1$, we conclude that

$$\mathbb{P}(|S - \mathbb{E}\, S| \geqslant t) \leqslant 2 \exp(-ct^2/n)$$

for some $c > 0$, which is a very useful deviation inequality for sums of random variables, i.e. it shows that $|S - \mathbb{E}\, S| \lesssim \log(n) \cdot n^{1/2}$ with high probability, and if $\{c_n\}$ is any sequence converging to infinity, regardless of how slowly, we have $|S - \mathbb{E}\, S| \lesssim c_n \log(n) n^{1/2}$ with overwhelming probability.

If $\sigma(X) = 0$, then $M_X(\lambda) \leqslant 1$ for all λ, which gives $\mathbb{P}(|X| \geqslant t) \leqslant e^{-\lambda t}$ for all $\lambda > 0$, and taking $\lambda \to \infty$ gives $\mathbb{P}(|X| \geqslant t) = 0$ for all $t > 0$. Thus $X = 0$ almost surely. If X is not centered, we still let $\sigma(X)$ denote $\sigma(X - \mathbb{E}\, X)$. Then σ is no longer a norm on the space of subgaussian random variables, but a seminorm, whose kernel is the space of random variables constant almost surely. The next lemma shows that $\sigma(X) \lesssim \|X\|_{\psi_2}$, but if the mean of X is unbounded, it is no longer true that $\|X\|_{\psi_2} \lesssim \sigma(X)$.

Lemma 5.9 (Centering)**.** *If X is subgaussian, $\|X - \mathbf{E}X\|_{\psi_2} \lesssim \|X\|_{\psi_2}$.*

Proof. Let δ be a point mass at zero. We use the triangle inequality to write

$$\|X - \mathbf{E}X\|_{\psi_2} \leqslant \|X\|_{\psi_2} + |\mathbf{E}X| \|\delta\|_{\psi_2} \lesssim \|X\|_{\psi_2} + |\mathbf{E}X|$$

Thus it suffices to prove $|\mathbf{E}X| \lesssim \|X\|_{\psi_2}$. But by Jensen's inequality,

$$|\mathbf{E}X| \leqslant \mathbf{E}|X| = \|X\|_1 \leqslant \|X\|_{\psi_2}.$$

This completes the argument. □

Theorem 5.10. *If $X_1, \dots, X_m$ are subgaussian random variables, and we have $\|X_i\|_{\psi_2} \leqslant K$ for all i, then*

$$\mathbb{E}(\max(|X_1|, \dots, |X_m|)) \lesssim K(\log m)^{1/2}.$$

Proof. Without loss of generality, assume $X_i \geqslant 0$ for all i. A union bound then gives

$$\mathbb{P}\left(\max(X_1, \dots, X_m) \geqslant t\right) \leqslant \sum_{i=1}^{m} \mathbb{P}\left(X_i \geqslant t\right) \leqslant m \cdot \exp(-t^2/K^2).$$

Thus

$$\begin{aligned}
\mathbb{E}(\max(X_1, \dots, X_m)) &= \int_0^\infty \mathbb{P}(\max(X_1, \dots, X_m) \geqslant t)\, dt \\
&= K \log m + m \int_{K \log m}^\infty \exp(-t^2/K^2) \\
&\lesssim K \log m + Km \log m \exp(-c(\log m)^2) \lesssim K \log m.
\end{aligned}$$

□

Example. *If X is a symmetric Bernoulli random variable with*

$$\mathbb{P}(X = -1) = \mathbb{P}(X = -1) = 1/2$$

Then $\mathbf{E}(e^{X^2/t^2}) = e^{1/t^2}$. This shows that $\|X\|_{\psi_2} = (\log 2)^{-1/2}$

Example. *If X is uniformly distributed on $[-1, 1]$, then*

$$\mathbf{E}[X^k] = \frac{1}{2} \int_{-1}^1 x^k \, dx = \frac{1 - (-1)^{k+1}}{2(k+1)} = \begin{cases} \frac{1}{k+1} & k \text{ even} \\ 0 & k \text{ odd} \end{cases}$$

So

$$\mathbf{E}[e^{\lambda X}] = \sum_{k=0}^\infty \frac{\lambda^{2k}}{(2k+1)(2k)!} \leqslant \sum_{k=0}^\infty \frac{\lambda^{2k}}{2^k k!} = e^{\lambda^2/2}$$

so $\sigma(X) \leqslant 1$. Since $\mathbf{V}(X) = 1$, we must have $\sigma(X) \geqslant 1$, so we actually have $\sigma(X) = 1$. More generally, by scaling, if X is uniformly distributed on $[-N, N]$, then $\sigma(X) = N$.

Example. *Suppose a centrally distributed random variable X satisfies $|X| \leqslant M$ almost surely, then $\sigma(X) \leqslant M$. Assume without loss of generality that $M = 1$. Set $Y = X + 1$, and $f(t) = (e^{2t} + 1)/2 - \mathbf{E}(e^{tY})$. Since $\mathbf{E}(Y) = 1$, $f'(t) = \mathbf{E}(Y[e^{2t} - e^{tY}])$. Since $Y \leqslant 2$ almost surely, $f'(t) \geqslant 0$, and so f is increasing. In particular, $f(0) = 1 - 1 = 0$, so that for $t \geqslant 0$,*

$$\mathbf{E}(e^{tX}) = e^{-t}\, \mathbf{E}(e^{tY}) \leqslant \frac{e^t + e^{-t}}{2} \leqslant e^{t^2/2}$$

Since we can perform the same argument for $-X$, X is 1 subgaussian.

If X is subgaussian, it 'looks' like a normal distribution if $\mathbb{V}(X)$ is approximately equal to $\|X\|_{\psi_2}$, i.e. so that with very large probability we have $|X| \lesssim \mathbb{V}(X)$. This fails for certain discrete distributions. For instance, a Bernoulli random variable $X \in \{0,1\}$ with $\mathbb{P}(X = 1) = p$ has $\mathbb{V}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = p(1-p)$, whereas $M_X(\lambda) = 1 + pe^{\lambda}$, which shows $\|X\|_{\psi_2} = 1/\log(1/p)$, which decreases at an arbitrarily fast rate as $p \to 0$. Thus for small values of p, $\|X\|_{\psi_2}$ is much larger than $\mathbb{V}(X)$. On the other hand, for sums of independant subgaussian random variables, these two quantities do agree with one another. This is Khintchine's ienquality.

Theorem 5.11 (Khintchine)**.** *Let $X_1, \dots, X_n$ be independant, mean zero subgaussian random variables, such that for all i,*

$$\|X_i\|_{\psi_2} \leqslant K.$$

Let $S = \sum_i X_i$. Then for any $p \in [2, \infty)$,

$$\mathbb{V}(S)^{1/2} \leqslant \|S\|_{L^p} \lesssim K p^{1/2} \mathbb{V}(S)^{1/2},$$

where $K = \max \|X_i\|_{\psi_2}$.

Proof. We have

$$\mathbb{V}(S)^{1/2} = \|S - \mathbb{E}\,S\|_{L^2} = \|S\|_{L^2} \leqslant \|S\|_{L^p},$$

which gives the first half of the inequality. To prove the second half of the inequality, we write

$$\|S\|_{L^p}^p = \int_0^\infty p t^{p-1}\, \mathbb{P}(|S| \geqslant t)\, dt.$$

We have $\|S\|_{\psi_2} \lesssim n^{1/2} K$, so plugging in the bound

$$\mathbb{P}(|S| \geqslant t) \leqslant 2\exp(-ct^2/nK)$$

Now fix $a > 0$, and calculate that

$$\begin{aligned}\int_{ap^{1/2}n^{1/2}K}^\infty p t^{p-1}\, \mathbb{P}(|S| \geqslant t) &\leqslant \int_{ap^{1/2}n^{1/2}K}^\infty 2pt^{p-1} \exp(-ct^2/nK^2) \\ &\lesssim p n^{p/2} K^p \int_{ap^{1/2}}^\infty t^{p-1} \exp(-ct^2)\, dt.\end{aligned}$$

If a is significantly large, then for $t \geqslant ap^{1/2}$,

$$\begin{aligned}t^{p-1}\exp(-ct^2) &\leqslant \exp((p-1)\log t - ct^2) \\ &\leqslant \exp\left(-\left(ct - (p-1)\frac{\log t}{t}\right)t\right) \\ &\leqslant \exp\left((-ct/2)\right).\end{aligned}$$

Thus we conclude that

$$\int_{ap^{1/2}}^\infty t^{p-1}\exp(-ct^2)\, dt \lesssim 1.$$

and so

$$\left(\int_{ap^{1/2}n^{1/2}K}^{\infty} pt^{p-1}\,\mathbb{P}(|S|\geqslant t)\right)^{1/p} \lesssim n^{1/2}K.$$

Conversely, we have

$$\int_0^{ap^{1/2}n^{1/2}K} pt^{p-1}\,\mathbb{P}(|S|\geqslant t) \leqslant p\int_0^{ap^{1/2}n^{1/2}K} t^{p-1}\,dt = p(ap^{1/2}n^{1/2}K)^p.$$

Thus

$$\left(\int_0^{ap^{1/2}n^{1/2}K} pt^{p-1}\,\mathbb{P}(|S|\geqslant t)\right)^{1/p} \lesssim p^{1/2}n^{1/2}K.$$

Putting these calculations together completes the proof. □

So if $K \lesssim 1$, $\|S\|_{\psi_2}^2$ is comparable to $\mathbb{V}(S)$, so the tails of S behave as if S was a Gaussian. This might be expected by virtue of the central limit theorem.

5.5 Subexponential Random Variables

The class of subgaussian random variables is very flexible, but some distributions just don't have that thin a tail. There are obviously random variables like the Cauchy distribution, whose averages do not settle down asymptotically whatsoever, but there are still distributions which do seem rather well behaved, possessing moments of all orders, while not lying in the category of subgaussian random variables. In this section we consider a slightly more general category of random variables for which we can get a tail bound, the subexponential random variables.

Theorem 5.12. *The following properties are equivalent, up to changes of constants:*

- $\mathbb{P}(|X|\geqslant t) \leqslant 2\exp(-t/K_1)$.
- $\|X\|_p \leqslant K_2\cdot p$.
- $\mathbf{E}(\exp(\lambda|X|)) \leqslant \exp(K_3\cdot\lambda)$ *for* $0\leqslant\lambda\leqslant 1/K_3$.
- *For some* K_4, $\mathbf{E}(\exp(|X|/K_4)) \leqslant 2$.
- *If* $\mathbf{E}X = 0$, $M_X(\lambda)\leqslant\exp(K_5^2\lambda^2)$ *for* $|\lambda|\leqslant 1/K_5$.

If the properties hold, we say that X is subexponential.

We leave the equivalence to the reader. A natural norm to place on this space is the Orlicz norm induced by $\psi_1(x) = e^{x/t}/2$. We have a variant of Hölder's inequality here.

Lemma 5.13. *A variable X is subgaussian if and only if X^2 is subexponential. The product of two subgaussian random variables is subexponential. More precisely,*

$$\|XY\|_{\psi_1} \leqslant \|X\|_{\psi_2}\|Y\|_{\psi_2}.$$

The analogy of Hoeffding's inequality for subgaussian functions is Bernstein's inequality, which describes the tail behaviour of a sum of subexponential random variables.

Theorem 5.14. *If $X_1, \ldots, X_N$ are independant, centrally distributed, subexponential random variables. Then*

$$\mathbb{P}\left(\left|\sum_{n=1}^{N} X_n\right| \geqslant t\right) \leqslant 2\exp\left(-c\min\left(\frac{t^2}{\sum \|X_m\|_{\psi_1}^2}, \frac{t}{\max \|X_m\|_{\psi_1}}\right)\right)$$

An important way to think of Bernstein's inequality is that it places the behaviour of deviation from the mean into two categories. For small deviations from the mean, the tail bound looks like that of a normal distribution. On the other hand, for large deviations, the tail becomes heavier, like an exponential distribution. The reason for the appearance of the exponential decay is a heuristic result of the central limit, which says the sum should look approximately like a normal distribution.

Chapter 6

Existence Theorems

In certain fields of probability theory, we wish to discuss collections of random variables defined over the same sample space. For instance, given a sequence $\mathbb{P}_1, \mathbb{P}_2, \ldots, \mathbb{P}_n$ of probability distributions defined over a space Y, we may want to talk about a sequence of independent random variables $X_i : \Omega \to Y$, such that $\mathbb{P}(X_i \in U) = \mathbb{P}_i(U)$. The construction here is simple; we take $\Omega = Y^n$, let $X_i = \pi_i$ be the projection on the i'th variable, and let $\mathbb{P}$ be the probability measure induced by

$$\mathbb{P}(U_1 \times U_2 \cdots \times U_n) = \mathbb{P}_1(U_1)\,\mathbb{P}_2(U_2) \ldots \mathbb{P}_n(U_n)$$

The construction here is simple because we have finitely many distributions, but the problem becomes much harder when we need to talk about an infinite family of distributions $\mathbb{P}_i$, or when we need to talk about non-independent random variables, with some specified relationships between the variables. The problem is to show there exists a sample space Ω 'big enough' for the random variables to all be defined on the space.

Chapter 7

Entropy

Let μ be a probability distribution. We would like to measure the expected 'amount of information' contained in the distribution – in essence, the average information entropy of μ. It was Claude Shannon who found the correct formula to measure this.

Shannon considered the problem of efficient information transfer. Suppose there was a channel of communication between two friends A and B. The friends have agreed on a standard dictionary X of possible messages, along with a probability distribution μ over the dictionary, and we would like to encode these messages into bits, in such a way that the average length of the message is smallest. We then define this to be the information entropy of μ. Shannon showed that if μ is discrete with probabilities $p_1, \ldots, p_n$, then the entropy can be calculated as

$$H(\mu) = \sum p_n \log_2 \left(\frac{1}{p_n} \right)$$

where the entropy is measured in bits, we can define the entropy in terms of the natural logarithm, in which case the entropy is said to be measured in nats. We assume that $p_i \log 1/p_i = 0$ for $p_i = 0$, which makes sense by the continuity of $x \log(1/x)$.

The entropy of a distribution also tells us

Now suppose that we were attempting to optimize a message with respect to a discrete distribution μ, and we instead encounter a distribution ν. Then the policy we have used for messages will be less optimal than if we had known that ν was the distribution in the first place. We define the relative difference in information between μ and ν as the difference between the encoding of ν with respect to μ, and the encoding of μ with respect to μ. This is not a linearly ordered relation, ν does not possess more information than μ, just different information. If μ takes probabilities p_i and ν takes relative probabilities q_i, the difference in information is calculated to be

$$D(\mu, \nu) = \sum p_i \log(1/q_i) - \sum p_i \log(1/p_i) = \sum p_i \log(p_i/q_i)$$

This is known as the *Kullback Leibler distance* between μ and ν.

Now suppose we are viewing independent samples $X_1, \ldots, X_n$, but we do not know where the samples are drawn from μ or ν. The larger $D(\mu, \nu)$ is, the less time we should take to

make an accurate decision that the distribution is μ or ν. Indeed, if $p_i > 0$ and $q_i = 0$, then $D(\mu, \nu) = \infty$, and we can conclude with certainty that the distribution is μ if we ever view the outcome corresponding to p_i.

It is necessary to define the 'entropy' of an arbitrary distribution, but it is then not clear how to interpret the entropy, since an encoding of uncountably many values will always have an infinite expected number of bits. However, we can defined the relative entropy by performing a discretization; Let μ and ν be distributions on some sample space X. Consider function $f : X \to \{1, \ldots, n\}$, and define

$$D(\mu, \nu) = \sup_f D(f_*\mu, f_*\nu)$$

where f_* pushes measures on X onto discrete measures on $\{1, \ldots, n\}$. For a fixed f, $D(\mu, \nu)$ upper bounds the difference in information we expect to see over a particular discretization. One can then calculate that

$$D(\mu, \nu) = \begin{cases} \infty & : \mu \not\ll \nu \\ \int \log(\frac{d\mu}{d\nu}) d\mu & : \mu \ll \nu \end{cases}$$

The relative entropies of well known distributions are easy to compute. Normal distributions, for instance, have

$$D(N(\mu_1, \sigma^2), N(\mu_2, \sigma^2)) = (\mu_1 - \mu_2)^2 / 2\sigma^2$$

For Bernoulli distributions, we have

$$D(B(p), B(q)) = p \log(p/q) + (1 - p) \log\left(\frac{1 - p}{1 - q}\right)$$

Which is true except perhaps at boundary conditions.

The Kullback Leibler distance gives us certain bounds which are essential to information theoretic lower bounds. The bound is useful, for it relates the probabilities of distributions by the difference in information contained within.

Theorem 7.1 (The High Probability Pinsker Bound)**.** *If μ and ν are probability measures on the same space X, and $U \subset X$ is measurable, then*

$$\mu(A) + \nu(A^c) \geqslant \frac{1}{2} e^{-D(\mu,\nu)}$$

Suppose we have a decision procedure which attempts to distinguish between events in probability distributions. If we choose an event A upon which the decision procedure fails to make the correct decision on the measure μ, and A^c measures the decision to fail under the measure ν, then the bound above shows the decision procedure cannot work reliably on both μ and ν.

Chapter 8

Appendix: Uniform Integrability

Uniform integrability provides stronger conditions on controlling convergence in the L^1 norm. For $p > 1$, inequalities often have 'smoothing' properties that are not apparent for the $p = 1$ case, so uniform integrability provides particular techniques to help us. We start with a basic result in measure theory, specialized to probabilistic language.

Lemma 8.1. *If $X \in L^1(\Omega)$ is a random variable, then for any $\varepsilon > 0$, there is $\delta > 0$ such that for any event E with $\mathbb{P}(E) \leqslant \delta$,*

$$\int_E |X| < \varepsilon$$

Proof. Suppose that there exists some ε, and events $E_1, E_2, \ldots$ with $\mathbb{P}(E_k) \leqslant 1/2^k$ but with

$$\int_{E_k} |X| \geqslant \varepsilon$$

By taking successive unions, we may assume the E_i are a decreasing family of sets, and then

$$\int_{\bigcap_{k=1}^{\infty} E_k} |X| = \lim_{k\to\infty} \int_{E_k} |X| \geqslant \varepsilon$$

and $\mathbb{P}(\bigcap E_k) = 0$, which is impossible. □

Corollary 8.2. *If $X \in L^1(\Omega)$, and $\varepsilon > 0$, then there is $K \in [0, \infty)$ with*

$$\int_{|X|>K} |X| < \varepsilon$$

A family of random variables $\{X_\alpha\}$ is called *uniformly integrable* if given $\varepsilon > 0$, there is $K \in [0, \infty)$ such that

$$\int_{|X_\alpha|>K} |X_\alpha| < \varepsilon$$

so that we can uniformly control the integral of X_α over large sets. We note that

$$\mathbf{E}|X_\alpha| = \int_{|X_\alpha|>K} |X_\alpha| + \int_{|X_\alpha|\leqslant K} |X_\alpha| \leqslant \varepsilon + K$$

so a family of uniformly integrable random variables is automatically in $L^1(\Omega)$, and *bounded* in $L^1(\Omega)$. However, a family of random variables bounded in $L^1(\Omega)$ is *not* necessarily uniformly integrable.

Example. *Let Ω be $[0,1]$ together with the Lebesgue measure. Let $E_n = (0, 1/n)$, and set $X_n = n\chi_{E_n}$. Then the X_n are bounded in $L^1(\Omega)$, but for $n \geqslant K$,*

$$\int_{X_n > K} X_n = 1$$

and so the random variables are not uniformly integrable.

These 'concentrating bumps' are essentially the only reason why we cannot always exchange expectations and limits, and require the application of the dominated convergence theorem. The condition of uniform integrability removes the ability for concentrating bumps to hide within the expectation of a family of random variables, and we find it also gives us conditions that guarantee we can exchange limits with integration. First, note that if we take a concentrated bump function $X = n\chi_{E_n}$, then $\mathbf{E}|X| = 1$ is bounded uniformly over n, but $\mathbf{E}|X|^{1+\varepsilon} = n^\varepsilon$ is unbounded, reflecting the fact that boundedness in $L^p(\Omega)$ for $p > 1$ removes concetrated bump functions by magnifying their effect.

Theorem 8.3. *Suppose that $\{X_\alpha\}$ is a class of random variables bounded in L^p for $p > 1$, then $\{X_\alpha\}$ is uniformly integrable.*

Proof. Consider some $A \in [0, \infty)$ which gives a uniform bound $\mathbf{E}|X_\alpha|^p < A$. Applying Hölder's inequality, we conclude

$$\int_{|X_\alpha| > K} |X_\alpha| \leqslant \int_{|X_\alpha| > K} \frac{|X_\alpha|^p}{K^{p-1}} \leqslant \frac{A}{K^{p-1}}$$

This is a uniform bound, and we may let $K \to \infty$ to let the bound go to zero. Thus the family $\{X_\alpha\}$ is uniformly integrable. □

Corollary 8.4. *If $|X_\alpha| \leqslant Y$ is a uniform bound over a family $\{X_\alpha\}$ of random variables, where $Y \in L^1(\Omega)$, then $\{X_\alpha\}$ is uniformly integrable.*

Proof. We find

$$\int_{|X_\alpha| > K} |X_\alpha| \leqslant \int_{|X_\alpha| > K} Y \leqslant \int_{Y > K} Y$$

and as $K \to \infty$, $\mathbb{P}(Y > K) \to 0$, and we can apply the continuity result to conclude that

$$\int_{Y > K} Y \to 0$$

and so we obtain a uniform bound. □

We recall that a sequence $X_1, X_2, \ldots$ of random variables *converges in probability* to a random variable X if, for every ε, $\mathbb{P}(|X_n - X| > \varepsilon) \to 0$. If $X_i \to X$ almost surely, then $X_i \to X$ in probability, because we can apply the reverse Fatou lemma to conclude

$$\begin{aligned} 0 &= \mathbb{P}\left(\liminf_{n\to\infty} |X_n - X| > \varepsilon\right) \\ &= \mathbb{P}\left(\limsup\{|X_n - X| > \varepsilon\}\right) \geqslant \limsup \mathbb{P}(|X_n - X| > \varepsilon) \end{aligned}$$

Hence $\mathbb{P}(|X_n - X| > \varepsilon) \to 0$. The bounded convergence theorem links L^1 convergence to convergence in probability using uniform integrability.

Theorem 8.5. *If X_n is a sequence of bounded random variables which tend to a random variable X in probability, then $X_n \to X$ in the L^1 norm.*

Proof. Let us begin by proving that if $|X_n| \leqslant K$, then $|X| \leqslant K$ almost surely. This follows because for any k,

$$\mathbb{P}(|X| > K + 1/k) \leqslant \mathbb{P}(|X - X_n| > 1/k) \to 0$$

so $\mathbb{P}(|X| > K + 1/k) = 0$, and letting $k \to \infty$ gives $\mathbb{P}(|X| > K) = 0$. Let $\varepsilon > 0$ be given. Then if we choose n large enough that

$$\mathbb{P}(|X_n - X| > \varepsilon) \leqslant \varepsilon$$

then

$$\begin{aligned} \mathbf{E}|X_n - X| &= \int_{|X_n - X| > \varepsilon} |X_n - X| + \int_{|X_n - X| \leqslant \varepsilon} |X_n - X| \\ &\leqslant 2K\varepsilon + \varepsilon \end{aligned}$$

we can then let $\varepsilon \to 0$ to obtain L^1 convergence. □

All this discussion concludes with a sufficient condition for L^1 convergence, showing that uniform integrability is really the right condition which removes the pathologies which prevent us from exchanging expectation with pointwise limits.

Theorem 8.6. *Let X_n be a sequence of integrable random variables, and X is another integrable random variable. Then $X_n \to X$ in the L^1 norm if and only if $X_n \to X$ in probability, and $\{X_n\}$ is uniformly integrable.*

Proof. Fix $K > 0$, and consider

$$f_K(x) = \begin{cases} K & : x > K \\ x & : |x| \leqslant K \\ -K & x < -K \end{cases}$$

Then for every $\varepsilon > 0$, we can choose K such that $\|f_K(X_n) - X_n\|_1 \leqslant \varepsilon$, $\|f_K(X) - X\|_1 \leqslant \varepsilon$ *uniformly across n* (adding a single variable to a uniformly integrable random variable keeps

it uniformly integrable). But it is easy to see that $f_K(X_n) \to f_K(X)$ in probability also, so by the bounded dominated convergence theorem, we conclude that $\|f_K(X_n) - f_K(X)\| \to 0$. A triangle inequality result gives the general result because the behaviour of X for large values is bounded by the uniform integrability.

To verify the reverse condition, note that if $\mathbf{E}|X_n - X| \to 0$, then Markov's inequality gives

$$\mathbb{P}(|X_n - X| \geqslant K) \leqslant \frac{\mathbf{E}|X_n - X|}{K} \to 0$$

to obtain uniform integrability, note that for each n, $\{X_1, \ldots, X_n, X\}$ is uniformly integrable, then for each $\varepsilon > 0$, there is δ such that if $\mathbb{P}(E < \delta)$,

$$\int_E |X_n| < \varepsilon \qquad \int_E |X| < \varepsilon$$

Since the entire set of X_n are bounded in $L^1(\Omega)$, we can choose K such that $\sup \mathbf{E}|X_k| < \delta K$, and then for $m > n$, $\mathbb{P}(|X_m - X| > K) < \delta$, and so

$$\int_{|X_m|>K} |X_m| \leqslant \int_{|X_m|>K} |X| + \mathbf{E}|X - X_m| \leqslant 2\varepsilon$$

where we assume we have chosen n large enough that $\mathbf{E}|X - X_m| \leqslant \varepsilon$. The fact that for $m \leqslant n$,

$$\int_{|X_m|>K} |X_m| \leqslant \varepsilon$$

follows from uniform integrability of the family $\{X, X_1, \ldots, X_n\}$, so we have shown the entire infinite sequence is uniformly integrable. □

Chapter 9

Percolation Theory

Let us consider the two dimensional theory of percolation. The two examples we have in mind are the lattice $\mathbf{Z}^2$, and the triangular lattice $\mathbf{T}$. For any $p \in [0,1]$, we define a graph structure on $\mathbf{Z}^2$, adding an edge between two adjacent elements of the lattice with independant probability p. On $\mathbf{T}$, we instead consider *site percolation*, where we keep a hexagon with probability p.

Theorem 9.1 (Russo-Seymour-Welsh)**.** *If $p = 1/2$, then for any $a, b > 0$, there exists c such that if A_n denotes the event that we can travel from the left edge to the right edge of the lattice $[0, a \cdot n] \times [0, b \cdot n] \cap \mathbf{Z}^2$, then $c < \mathbb{P}(A_n) < 1 - c$.*

One of the main problems in percolation theory is to determine how likely it is to find an infinite connected set of vertices, or cluster, in the randomly selected graph. As the probability of each edge becomes more likely, the graph becomes more and more connected. We find that for $p > 1/2$, there is almost surely an infinite cluster, and for $p < 1/2$, there is almost surely *not* a cluster. The value $p = 1/2$ is therefore called the *phase transition* point. A very related value to the phase transition problem is the percolation density function θ, which for each p, gives the probability $\theta(p)$ of the origin being in an infinite cluster of the graph. As an example, it is known that on $\mathbf{T}$, $\theta(p) = (p - 1/2)^{5/36+o(1)}$, as $p \downarrow 1/2$. Determining phase transition points is the main focus of this chapter's notes.

9.1 Duality

Note an important duality in these geometric scenarios. Given any graph on $\mathbf{Z}^2$, we can obtain another graph, the dual graph, by taking the vertices as unit squares with corners on $\mathbf{Z}^2$, and with an edge between adjacent squares if there is no edge separating the two squares. Then the probability that there is an edge between two squares is the same as the probability that we do *not* select the corresponding separating edge, i.e. with probability $1 - p$. This will be useful.

The book says the dual graph of $\mathbf{T}$ is $\mathbf{T}$, but I don't quite understand why?

9.2 Boolean Functions and Sharp Thresholds

If G is a graph, then the family of all graph structures on these vertices can be identified with $\{0,1\}^E = \{0,1\}^{O(V^2)}$. Thus a function f on the set of graphs can be identified with a boolean function, and we can apply boolean function techniques. In our case, the natural graphs will be the subgraphs of $\mathbf{Z}^2$ of the form $[0, a \cdot n] \times [0, b \cdot n]$. An example of a Boolean function on graphs is obtained by setting

$$f_n(G) = \mathbf{I}(\text{There is a path from left to right in } G)$$

Boolean analysis tells us the main features of G. Here we introduce some basics, which will help us get the job done.

If f is a boolean function, we say it is monotone if $x_i \leqslant y_i$ for each i implies $f(x) \leqslant f(y)$. An index i is pivotal for an input x if $f(x) \neq f(x^i)$, where i is x with the bit flipped at the i'th position. The influence $\mathbf{I}_i(f)$ of f in the variable i is then the probability that for a randomly chosen x, i is pivotal. If we instead choose an input to be equal to one with probability p, and zero with probability $1-p$, then the probability that i is pivotal is denoted $\mathbf{I}_i^p(f)$. The sum of the influence over all influences i is known as the *total influence*. Now if E is a monotone event, then it is obvious that as p increases, $\mathbb{P}(E)$ should increase. The degree to which it increases is quantified by the Margulis-Russo formula.

Theorem 9.2 (Margulis-Russo)**.** *Let E be monotone. Then $d\,\mathbb{P}(E)/dp = \mathbf{I}^p(E)$.*

Proof. Temporarily, let $\mathbf{I}_i^{p_1,\dots,p_n}(E)$ denote the chance that $X^i \neq X$, where the $X_j \in \{0,1\}$ are chosen uniformly at random with $\mathbb{P}(X_j = 1) = p_i$. Define $\mathbf{I}^{p_1,\dots,p_n}(E) = \sum \mathbf{I}_i^{p_1,\dots,p_n}(E)$. It suffices to show $\partial\,\mathbb{P}(E)/\partial p_i = \mathbf{I}_i^{p_1,\dots,p_n}(E)$, from which we can use the chain rule. We can write E as the union of two disjoint events E_0 and E_1, where $E_0 = E \cap \{\chi_E(X^i) \neq \chi_E(X)\}$, and $E_1 = E \cap \{\chi_E(X^i) = \chi_E(X)\}$. Now E_1 does not depend on the value of X_1 at all, so $\mathbb{P}(E_1)$ is independant of p_i, and so

$$\frac{\partial\,\mathbb{P}(E_1)}{\partial p_i} = 0$$

On the other hand, by monotonicity, E_0 then equals the probability that $\chi_E(X^i) \neq \chi_E(X)$, intersected with the event $X_i = 1$. These two events are independant, so $\mathbb{P}(E_0) = p_1\,\mathbb{P}(\chi_E(X^i) \neq \chi_E(X))$. The latter probability does not depend on the index 1, so

$$\frac{\partial\,\mathbb{P}(E_0)}{\partial p_i} = \mathbb{P}(\chi_E(X^i) \neq \chi_E(X)) = \mathbf{I}_1^{p_1,\dots,p_n}(E)$$

This completes the proof. □

To analyze the critical exponent, we rely on two results we will prove later on using Fourier analysis, which allow us to upper bound the influence by the variance of a function.

Theorem 9.3 (Bourgain, Kahn, Kalai)**.** *For any f and p, there exists i such that $\mathbf{I}_i^p(f) \gtrsim \mathbf{V}_p(f)[\log n/n]$, and $\mathbf{I}^p(f) \gtrsim \mathbf{V}_p(f)\log(1/\max \mathbf{I}_i^p(f))$.*

We also rely on a fact that, for exponentially many boolean inputs, the probability that a monotone event happens jumps from being unlikely to being likely in a very small small range of p values, i.e. of length approximately $1/\log n$. Think like the majority function. Once $p > 1/2$, the chance of a vote passing grows rapidly in p.

Theorem 9.4 (Friedgut, Kalai)**.** *If A is a monotone event, whose influences are the same for each index, and for $p = p_0$, if $\mathbb{P}(A) > \varepsilon$, then for $p \geqslant p_0 + c\log(1/\varepsilon)/\log n$, $\mathbb{P}(A) > 1 - \varepsilon$.*

Proof. If all the influences are the same, we find the total variance is

$$\mathbf{I}^p(E) \gtrsim \mathbf{V}_p(\chi_E)\log n = \min(\mathbb{P}(E), 1 - \mathbb{P}(E))\log n$$

Now the Margulis-Russo formula yields that if $\mathbb{P}(E) \leqslant 1/2$,

$$\frac{d(\log \mathbb{P}(E))}{dp} = \frac{\mathbf{I}^p(E)}{\mathbb{P}(E)} \gtrsim \log n$$

Thus if $p \geqslant p_0 + c/\log n$, $\mathbb{P}(E) \geqslant 1/2$. Now if $\mathbb{P}(E) \geqslant 1/2$, then

$$\frac{d(\log(1 - \mathbb{P}(E)))}{dp} = -\frac{\mathbf{I}^p(E)}{1 - \mathbb{P}(E)} \lesssim -\log n$$

In order to make $\mathbb{P}(E) \geqslant 1 - \varepsilon$, we need $\log(1 - \mathbb{P}(E)) \leqslant \log \varepsilon$. To move from $\log(1/2)$ to $\log \varepsilon$, we need $p \geqslant p_0 + c/\log n + c\log(1/\varepsilon)/\log n = p_0 + c\log(1/\varepsilon)/\log n$. □

Theorem 9.5 (FKG)**.** *Any two increasing events are positively correlated.*

We now try and prove the critical exponent for the square lattice is $1/2$. First, we show $\theta(1/2) = 0$. Consider the 'square annulus' between 4^n and $3 \cdot 4^n$, and let E_n be the event that there is a ring in this annulus. Because the edge sets involved in each event are disjoint, the events are independant. Furthermore, the probability that there is a cross on each side of the annulus is bounded below by some constant $c > 0$. Adding more edges doesn't hurt the probability that an event happens, so they are monotonic, and the FKG inequality says that $\mathbb{P}(E_n) = \mathbb{P}(A \cap B \cap C \cap D) \geqslant \mathbb{P}(A)\,\mathbb{P}(B)\,\mathbb{P}(C)\,\mathbb{P}(D) \geqslant c^4$. Thus Thus infinitely many occur almost surely, so $\theta(1/2) = 0$, and in fact, there is almost surely no infinite cluster anywhere.

To form a contradiction, we assume the critical point is $1/2 + \delta$ instead of $1/2$. Given n, we form a $2n \times n$ box, and we let J_n denote the event that there is a crossing in the box, i.e a path from left to right and an edge from bottom to top.

Lemma 9.6. *As $n \to \infty$, $\max \mathbf{I}_e^p(J_n) \to 0$, where the maximum is taken over $1/2 \leqslant p \leqslant 1/2 + \delta/2$ and all edges e.*

Proof. If e is pivotal on a given input for J_n, that means there is a path from a vertex adjacent to e to a vertex a distance $n/2$ away. Thus by translation invariance, $\mathbf{I}_e^p(J_n)$ is bounded by half the probability that there is a path of length $n/2$ from the origin, which is uniformly bounded for $p \leqslant 1/2 + \delta/2$. As $\theta(1/2 + \delta/2) = 0$, these values must tend to zero as $n \to \infty$. □

Lemma 9.7. *For some n, and with* $p = 1/2 + \delta/2$, $\mathbb{P}(J_n) \geqslant 0.98$.

Proof. We have already seen that $\inf \mathbb{P}(J_n) > 0$ when $p = 1/2$. If $\mathbb{P}(J_n) < 0.98$ for all n, then BKS shows that $d\,\mathbb{P}(J_n)/dp$ must tend to ∞ uniformly for all $1/2 \leqslant p \leqslant 1/2 + \delta/2$ as $n \to \infty$, by setting $\varepsilon < 0.02$. And this means that the probability of J_n increases massively over the range of p from $1/2$ to $1/2 + \delta/2$. □

Now we show this implies that we almost surely get an infinite cluster. If we can cross a $2n \times n$ box with probability $1 - \varepsilon$, crossing lines gives a probability of $1 - 5\varepsilon$ of cross a $4n \times n$ box. Thus the probability that we cross a $4n$ by $2n$ box is greater than the probability that we cross the top and bottom of the graph. Thus

$$\begin{aligned}\mathbb{P}(\text{Top} \cup \text{Bottom}) &= \mathbb{P}(\text{Top}) + \mathbb{P}(\text{Bottom}) - \mathbb{P}(\text{Top})\,\mathbb{P}(\text{Bottom}) \\ &\geqslant 2(1 - 5\varepsilon) - (1 - 5\varepsilon)^2 \geqslant 1 - 5\varepsilon^2\end{aligned}$$

If ε is small, this error is REALLY small. Thus almost surely all but finitely many of the J_n occur. Thus we just have to put these lines together in a way which guarantees

9.3 Conformal Invariance

Brownian motion is the limit of a random walk, and in the plane, is conformally invariant, in the sense that the image of a path of Brownian motion under an analytic map looks like the path of Brownian motion, up to a change in the measurement of time. Thus a random walk is conformally invariant 'in the limit'. In some sense, percolation should also look 'asymptotically' conformally invariant in the limit.

Let us describe what this principle should look like when discretized. We consider a conformal map ϕ from the unit disk to some other simply connected domain D fixing the origin, and with $\phi'(0) > 0$. Now for any δ, we can consider the lattice $\delta\mathbf{Z}^2$, restricted to the interior of the unit disk. If C_δ is the cluster around the origin in the interior. Similarily, we define C'_δ to be the cluster around the origin in $\delta\mathbf{Z}^2$ restricted to D. Now $\phi(C_\delta)$ and C'_δ don't even lie on the same lattice, but as $\delta \to 0$, they should still asymptotically describe the same law on space.

The simplest precise statement of conformal invariance was proved by Smirnov in 2001. Scale the hexagonal percolation problem by δ at the critical percolation value $p = 1/2$. If four points A, B, C, and D are chosen on the boundary of D, then the probability that there is a path from points on the boundary of D between A and B to points on the boundary of D between C and D converges as $\delta \to 0$. Furthermore, this convergent value is invariant under conformal mappings. In the case where D is a sidelength one equilateral triangle, A, B, and C are the three corner points, and D is on the line between A and C with distance x from C, the probability converges to x. By conformal invariance, this gives a general way to calculate the limiting probability. On $\mathbf{Z}^2$, even this statement is still open.

Chapter 10

High Dimensional Probability

Theorem 10.1. *Let f be L Lipschitz. Then*

$$\gamma(x \in \mathbf{R}^n : |f(x) - M| > t) \leqslant 2\exp(-t^2/2L^2)$$

Proof. Let $\mathbf{H}$ be a half space with Gaussian measure $1/2$, i.e. the standard upper half plane $\mathbf{H} = \{x : x_1 \leqslant 0\}$. Then $\gamma(H_\varepsilon^c) = \mathbb{P}(N(0,1) \geqslant \varepsilon) \leqslant e^{-\varepsilon^2/2}$. Thus $\gamma(H_\varepsilon) \geqslant 1 - \varepsilon^2/2$.

Next, if $A = \{x : f(x) \geqslant M\}$, we obtain

$$\gamma(x : f(x) \geqslant M - L\varepsilon) \geqslant \gamma(A_\varepsilon) \geqslant \gamma(\mathbf{H}_\varepsilon) \geqslant 1 - e^{-\varepsilon^2/2}$$

□

Theorem 10.2 (Isoperimetric Inequality in Gaussian Space)**.** *Among all measurable sets A in $\mathbf{R}^n$ with the same Gaussian measure, half spaces minimize the measure $\gamma_n(A_\varepsilon)$.*

Remark. We can replace M by $\mathbf{E}f$ using sub Gaussian centering.

Theorem 10.3. *Let X be sub Gaussian. Then the centered version satisfies $\|X - \mathbf{E}X\|_{\psi_2} \leqslant 2\|X\|_{\psi_2}$.*

Proof. We find

$$\|X - \mathbf{E}\|_p \leqslant \|X\|_p + \|\mathbf{E}X\|_p$$

Now $\mathbf{E}\|X\| \leqslant \|X\|_p$. □

The normal distribution concentrates on an annulus of radius $\sqrt{d}$ and width $O(1)$. Also, it concentrates on half spaces $\{X_1 \geqslant c\}$.

Bibliography

[1] Larry Wasserman, *All of Statistics*

[2] Walter Rudin, *Real and Complex Analysis*