Probability *

Zhiyuan Bai

Compiled on October 2, 2020

This document serves as a set of revision materials for the Cambridge Mathematical Tripos Part IA course *Probability* in Lent 2020. However, despite its primary focus, readers should note that it is NOT a verbatim recall of the lectures, since the author might have made further amendments in the content. Therefore, there should always be provisions for errors and typos while this material is being used.

Contents

*Based on the lectures under the same name taught by Dr. P. Sousi in Lent 2020.

0 Introduction

Probablity theory is the mathematical formulation of randomness. Examples include the modeling of random experiments like flipping a coin, throwing a die, shuffle a deck, and so on. What we want to do is to develop a mathematical framework to study randomness.

1 Definitions and Properties

Definition 1.1. Let Ω be a set, and $\mathscr{F}$ a set of subset of Ω. We call $\mathscr{F}$ a σ-algebra on Ω if
1. $\Omega \in \mathscr{F}$.
2. $A \in \mathscr{F} \implies A^c \in \mathscr{F}$.
3. For any countable sequence $(A_n)_{n\in\mathbb{N}} \in \mathscr{F}$, we have

$$\bigcup_{n\in\mathbb{N}} A_n \in \mathscr{F}$$

If $\mathscr{F}$ is a σ-algebra, then a function $\mathbb{P} : \mathscr{F} \to [0,1]$ is called a probablity measure if
1. $\mathbb{P}(\Omega) = 1$.
2. For every sequence of disjoint sets $(A_n)_{n\in\mathbb{N}} \in \mathscr{F}$, we have

$$\mathbb{P}\left(\bigcup_{n\in\mathbb{N}} A_n\right) = \sum_{n\in\mathbb{N}} \mathbb{P}(A_n)$$

If $\mathbb{P}$ is a probablity measure, then we call $(\Omega, \mathscr{F}, \mathbb{P})$ a probablity space.

Remark. When Ω is countable, we usually take $\mathscr{F} = 2^\Omega$.

Definition 1.2. The elements of Ω are called outcomes and the elements of $\mathbb{F}$ are called events. If $A \in \mathscr{F}$, we interpret $\mathbb{P}(A)$ as the probablity that A happens. Note that we only talk about probablity of events instead of outcomes.

We will see later that if we take a random point from $[0,1]$, the probablity of a certain point being taken is 0 (consequently any countable subset of $[0,1]$).

Proposition 1.1. *Let $A, B \in \mathscr{F}$. 1.* $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.
2. $\mathbb{P}(\varnothing) = 0$.
3. $A \subset B \implies \mathbb{P}(B) \geq \mathbb{P}(A)$.
4. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

Proof. All follow from definition. □

Example 1.1. 1. Rolling a fair die. So $\Omega = \{1,2,3,4,5,6\}$ and $\mathscr{F} = 2^\Omega$, and $\forall A \subset \Omega, \mathbb{P}(A) = |A|/6$.
2. Equality likely outcomes. Ω is just a finite set of size $n > 0$ and $\mathscr{F} = 2^\Omega$, and $\forall A \subset \Omega, \mathbb{P}(A) = |A|/n$. This is the model of a randomly chosen point of Ω. Taking $n = 6$ gives the first example.
3. Picking balls from a bag. Suppose we have a bag with n labelled balls $1, 2, \ldots n$ and indistinguishable by touch. We pick $k \leq n$ balls at random (i.e. all outcomes are equally likely) without looking. Then $\Omega = \{A \subset \{1,2,\ldots,n\} : |A| = k\}$ and $\mathscr{F} = 2^\Omega$, so $\forall A \subset \Omega, \mathbb{P}(A) = |A|/\binom{n}{k}$.
4. A deck of cards. Suppose we have a well-shuffled (i.e. all possible ordering of the cards are equally likely) deck of 52 cards (excluding jokers). So $\Omega = S_{52}, \mathscr{F} = 2^\Omega$ and $\forall A \subset \Omega, \mathbb{P}(A) = |A|/52!$. Hence $\mathbb{P}$(first two cards are aces) $= (4 \times 3 \times 50!)/52! = 1/221$.
5. Largest digit. Consider a string of random digits (all outcomes equally possible) $0, 1, \ldots, 9$ of length n. Take $\Omega = \{0,1,\ldots,9\}^n$ and $\mathscr{F}, \mathbb{P}$ as in before. Let $A_k = \{$no digit exceeds $k\}$ and $B_k = \{$largest digit is $k\}$. Since $|A_k| = (k+1)^n$ we have $|B_k| = |A_k \setminus A_{k-1}| = (k+1)^n - k^n$, so $\mathbb{P}(B_k) = ((k+1)^n - k^n)/10^n$.
6. Birthday problem. Suppose there are n people in the room. What is the probability of at least two people sharing the same birthday, given that nobody is born on 29 Feb and any other day in the year is equally probable. So $\Omega = \{1,2,\ldots,365\}^n$, and again $\mathscr{F}$ and $\mathbb{P}$ are taken as before (equally likely outcome). Let A be the event that all birthdays are different, then

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|} = \frac{365 \times 364 \times \ldots (365 - n + 1)}{365^n}$$

Then the probablity of two having the same birthday is $p = 1 - \mathbb{P}(A)$. If you take $n = 22$, then you get $p \approx .476$, and when $n = 23$, $p \approx .507$.

2 Some Counting and Stirling's Formula

2.1 Some Counting

If we have a finite set Ω, and let M be the number of ways of partition Ω into k subsets $S_1, S_2, \ldots, S_k$ with $|S_j| = n_j$ with $n_1 + n_2 + \cdots + n_k = |\Omega|$, then

$$M = \binom{|\Omega|}{n_1, n_2, \ldots, n_k} = \frac{|\Omega|}{n_1! n_2! \cdots n_k!}$$

We suddenly want to count the number of strictly increasing and nondecreasing functions from the set $\{1, 2, \ldots, k\}$ to $\{1, 2, \ldots, n\}$. Note that each strictly increasing function in this way are uniquely identified by their image, so the number of such functions equals $\binom{n}{k}$. [1] To count nondecreasing functions, however, we cannot use this trick. Nonetheless, we can consider the bijection

$$\{f \text{ nondecreasing} : \{1, 2, \ldots, k\} \to \{1, 2, \ldots, n\}\} \to$$

$$\{f \text{ strictly increasing} : \{1, 2, \ldots, k\} \to \{1, 2, \ldots, n + k - 1\}\}$$

by assigning a function f in the previous set to the function $g(i) = f(i) + i - 1$. So that number is $\binom{n+k-1}{k}$.

2.2 Stirling's Formula

Definition 2.1. Let $(a_n), (b_n)$ be two positive sequences, we say $a_n \sim b_n$ if $a_n/b_n \to 1$ as $n \to \infty$.

Theorem 2.1 (Stirling)**.**

$$n! \sim n^n \sqrt{2\pi n} e^{-n}$$

Proposition 2.2 (Weaker Statement of Stirling's Formula)**.** $\log(n!) \sim n \log n$

Proof. Let $\ell_n = \log(n!)$. So we have

$$\ell(n) = \log 2 + \log 3 + \cdots + \log n$$

Note that we have the trivial bound $\log\lfloor x \rfloor \le \log x \le \log\lfloor x + 1 \rfloor$, hence

$$\ell_{n-1} \le \int_1^n \log x \, \mathrm{d}x \le \ell_n$$

So

$$n \log(n) - n + 1 \le \ell_n \le (n + 1) \log(n + 1) - n$$

The proposition follows. □

Proof of Stirling's Formula. Note that

$$\int_a^b f(x) \, \mathrm{d}x = \frac{f(a) + f(b)}{2}(b - a) - \frac{1}{2} \int_a^b (x - a)(b - x) f''(x) \, \mathrm{d}x$$

[1] $\binom{n}{k} = 0$ for $k > n$

We take $f = \log$ we have

$$\int_k^{k+1} \log x \, \mathrm{d}x = \frac{\log(k) + \log(k+1)}{2} + \frac{1}{2} \int_0^1 \frac{x(1-x)}{(x+k)^2} \, \mathrm{d}x$$

Summing over $k = 1, 2, \ldots, n-1$ we have

$$n \log n - n + 1 = \log(n!) - \frac{\log n}{2} + \sum_{k=1}^{n-1} a_k$$

where

$$a_k = \frac{1}{2} \int_0^1 \frac{x(1-x)}{(x+k)^2} \, \mathrm{d}x$$

Note that it is easy to see the partial sum of a_k converges, so we define $A = \exp(1 - \sum_{k \in \mathbb{N}} a_k)$ to have

$$n! = n^n \sqrt{n} e^{-n} A \exp\left(\sum_{k=n}^{\infty} a_k \right)$$

Note that the last part goes to 1 as $n \to \infty$, so it remains to show $A = \sqrt{2\pi}$. We claim that

$$2^{-2n} \binom{2n}{n} \sim \frac{1}{\sqrt{\pi n}}$$

since $2^{-2n} \binom{2n}{n} \sim \sqrt{2}/(A\sqrt{n})$, it will prove what we want. Consider

$$I_n = \int_0^{\pi/2} \cos^n \theta \, \mathrm{d}\theta$$

It is trivial to see that it equals $I_{2n} = \binom{2n}{n} \pi / 2^{2n+1}$ and $I_{2n+1} = 2^{2n} \binom{2n}{n}^{-1} / (2n+1)$. We want to show that $I_{2n}/I_{2n+1} \to 1$ which will prove the result, but this is obvious since $I_{n+2}/I_n \to 1$ and I_n is decreasing. □

3 Some Properties of Probability Measure and More Counting

3.1 Elementary Results

Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space.

Proposition 3.1 (Countable Subadditivity)**.** *Let $(A_n)_{n \in \mathbb{N}} \in \mathscr{F}$, then*

$$\mathbb{P}\left(\bigcup_{n \in \mathbb{N}} A_n \right) \leq \sum_{n=1}^{\infty} \mathbb{P}(A_n)$$

Proof. Consider $B_1 = A_1$ and $B_n = A_n \setminus (A_1 \cup A_2 \cup \cdots \cup A_{n-1})$. The result follows. □

$\mathbb{P}$ is upward continuous.

Proposition 3.2. *Suppose $(A_n)_n \in \mathbb{N} \in \mathscr{F}$ is increasing, then*

$$\mathbb{P}(A_n) \to \mathbb{P}\left(\bigcup_{n\in\mathbb{N}} A_n\right)$$

as $n \to \infty$.

Proof. Consider $B_1 = A_1$ and $B_n = A_n \setminus (A_1 \cup A_2 \cup \cdots \cup A_{n-1})$. Note that $\mathbb{P}(A_n) = \mathbb{P}(B_1) + \cdots + \mathbb{P}(B_n)$, so

$$\mathbb{P}(A_n) = \sum_{k=1}^{n} \mathbb{P}(B_k) = \mathbb{P}\left(\bigcup_{k=1}^{n} B_n\right) = \mathbb{P}\left(\bigcup_{k=1}^{n} A_n\right) \to \mathbb{P}\left(\bigcup_{n\in\mathbb{N}} A_n\right)$$

As desired. □

By taking complement, we know that $\mathbb{P}$ is also downward continuous.

Proposition 3.3 (Inclusion-Exclusion Principle)**.** *Let $A_1, A_2, \ldots, A_n$ be events, then we have*

$$\mathbb{P}(A_1 \cup \cdots \cup A_n) = \sum_{k=1}^{n} (-1)^{k+1} \sum_{1\le i_1<\cdots<i_k\le n} \mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_k})$$

Proof. Induction. □

For the probability on finite sample space with equally likely outcomes, the principle reduced to the known inclusion-exclusion principle on finite sets.

Corollary 3.4. *Let $A_1, A_2, \ldots, A_n$ be sets, then we have*

$$|A_1 \cup \cdots \cup A_n| = \sum_{k=1}^{n} (-1)^{k+1} \sum_{1\le i_1<\cdots<i_k\le n} |A_{i_1} \cap \cdots \cap A_{i_k}|$$

Proof. Follows directly. □

Proposition 3.5 (Bonferroni Inequality)**.** *Truncating the inclusion-exclusion formula after the n^{th} term gives an overestimate if n is odd and underestimate if n is even.*

Proof. Trivial. □

3.2 Counting by Inclusion-Exclusion

Inclusion-Exclusion allows us to do some more counting.
We want to count the number of surjections $\{1, 2, \ldots, n\} \to \{1, 2, \ldots, m\}$ for $m \le n$. Take $\Omega = \{f : \{1, 2, \ldots, n\} \to \{1, 2, \ldots, m\}\}$ and A be the set of surjections. Let $A_i = \{f \in \Omega : i \notin \operatorname{Im} f\}$, so $A = A_1^c \cap \cdots \cap A_m^c = (A_1 \cup \cdots \cup A_m)^c$. Now since we have $|A_{i_1} \cap \cdots \cap A_{i_k}| = (m-k)^n$, hence by Inclusion-Exclusion, we have

$$|A| = \sum_{k=0}^{m} (-1)^k \binom{m}{k} (m-k)^n$$

(Note that a complement has been taken.)
We now want to count the number of derangements of $\{1, 2, \ldots, n\}$. Let $\Omega = S_n$ and A be the set of derangements. Let $A_i = \{\sigma \in \Omega : \sigma(i) = i\}$, so again $A = (A_1 \cup \cdots \cup A_n)^c$. But we have $|A_{i_1} \cap \cdots \cap A_{i_k}| = (n-k)!$, so assume an equally likely probability measure $\mathbb{P}$, we then have

$$\mathbb{P}(A^c) = \sum_{k=1}^{n} (-1)^{k+1} \frac{n}{k} \frac{(n-k)!}{n!} \implies \mathbb{P}(A) = \sum_{k=0}^{n} (-1)^k \frac{1}{k!} \to \frac{1}{e}$$

So asymptopically $|A| \sim n!/e$.

3.3 Independence

Definition 3.1. $A, B \in \mathscr{F}$ are said to be independent if $\mathbb{P}(A)\mathbb{P}(B) = (A \cap B)$. A countable sequence $(A_n)_{n\in\mathbb{N}}$ of events is independent if $\forall k \geq 2$ and for any set of indices $i_1, i_2, \ldots, i_k$ we have

$$\mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_k}) = \mathbb{P}(A_{i_1}) \cdots \mathbb{P}(A_{i_k})$$

Remark. 1. Pairwise independence does not imply independence. Say we flip a fair coin twice, then $\Omega = \{(0,0), (0,1), (1,0), (1,1)\}$ and $A = \{(0,0), (0,1)\}, B = \{(0,0), (1,0)\}, C = \{(0,1), (1,0)\}$, then they are pairwisely independent but not independent.
2. If A is independent of B, then A is also independence of B^c.

Definition 3.2. Let A, B be events such that $\mathbb{P}(B) > 0$. The conditional probability is defined as

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

In particular, $\mathbb{P}(A|B) = \mathbb{P}(A)$ iff A, B are independent.

Proposition 3.6. *Let $(A_n)_{n\in\mathbb{N}}$ be a sequence of disjoint events, then*

$$\mathbb{P}\left(\bigcup_{n\in\mathbb{N}} A_n \middle| B\right) = \sum_{n\in\mathbb{N}} \mathbb{P}(A_n|B)$$

Proof. Trivial. □

So conditional probability measure is also a probability measure.

Theorem 3.7 (Law of Total Probability)**.** *Suppose $(B_n)_{n\in\mathbb{N}}$ is a disjoint sequence of events such that $\mathbb{P}(B_n) > 0$ for all n and $\bigcup B_n = \Omega$. Then*

$$\mathbb{P}(A) = \sum_{n\in\mathbb{N}} \mathbb{P}(A|B_n)\mathbb{P}(B_n)$$

Proof. Obvious. □

Theorem 3.8 (Bayes' Formula)**.** *Let $(B_n)_{n\in\mathbb{N}}$ be as above and $\mathbb{P}(A) > 0$, then*

$$\mathbb{P}(B_n|A) = \frac{\mathbb{P}(A|B_n)\mathbb{P}(B_n)}{\sum_k \mathbb{P}(A|B_k)\mathbb{P}(B_k)}$$

Example 3.1 (False Positives for a Rare Condition)**.** Consider a rare medical condition A which affects 0.1% of the population and a medical test which is positive for 98% of the people affected and 1% of those unaffected by the condition. Take a random person, then we want to know the probablity that he has condition A given that he was tested positive. Let A be the event of individuals suffering from the condition and P be the individuals being tested positive. Hence

$$\mathbb{P}(A|P) = \frac{\mathbb{P}(P|A)\mathbb{P}(A)}{\mathbb{P}(P|A^c)\mathbb{P}(A^c) + \mathbb{P}(P|A)\mathbb{P}(A)} \approx 8.9\%$$

This may seems counterintuitive since this probability seems low, but the thing is that $\mathbb{P}(P|A^c) >> \mathbb{P}(P|A)$ since the medical condition is so rare.

Example 3.2 (Extra Knowledge Changes Probability)**.** Consider the following statements:
(a) I have 2 childrens, the elder of whom is a boy.
(b) I have 2 childrens, one of whom is a boy.
(c) I have 2 childrens, one of whom is a boy who is born on a Tuesday.
We assume equally likely distributions. Let GB denotes that the younger is a girl and the elder is a boy. Similar for BG, BB, GG.

$$\mathbb{P}(BB|a) = \mathbb{P}(BB|BG \cup BB) = \frac{1}{2}$$

$$\mathbb{P}(BB|b) = \mathbb{P}(BB|BG \cup GB \cup BB) = \frac{1}{3}$$

Write TN be that there are two boy, the younger is born on Tuesday but the elder is not, similar for (2-)combinations of T, N, G.

$$\begin{aligned}\mathbb{P}(BB|c) &= \mathbb{P}(TT \cup TN \cup NT | TT \cup TN \cup NT \cup TG \cup GT) \\ &= \frac{\mathbb{P}(TT \cup TN \cup NT)}{\mathbb{P}(TT \cup TN \cup NT \cup GT \cup TG)} \\ &= \frac{13}{27}\end{aligned}$$

Example 3.3 (Simpson's Paradox)**.** There are 50 men and 50 women applying to a college.

All applicants	Admitted	Rejected	Success Rate
State	25	25	50%
Indep	28	22	56%
Men Only	Admitted	Rejected	Success Rate
State	15	22	41%
Indep	5	8	38%
Women Only	Admitted	Rejected	Success Rate
State	10	3	77%
Indep	23	14	68%

So both men and women in state schools have higher acceptance rate, but the overall acceptance rate of state schools is still lower than that in independent schools. This is because the overall acceptance rate for women in this set of data is larger and there are much more of them in independent schools than state schools. Basically it is because $A/B > a/b$ and $C/D > c/d$ does not imply $(A + C)/(B + D) > (a + c)/(b + d)$.

4 Discrete Distributions

Definition 4.1. Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space where Ω is a countable set (we will most likely take it as finite), $\Omega = \{\omega_1, \omega_2, \ldots\}$ and $\mathscr{F} = 2^\Omega$. In order to determine $\mathbb{P}$, it sufficed to determine all $p_i = \mathbb{P}(\{\omega_i\})$. We call p_i as a discrete distribution.

Note that $p_i \geq 0$ and $\sum_i p_i = 1$.

4.1 Examples of Useful Distributions

Definition 4.2. $\Omega = \{0, 1\}$ and $p_1 = p, p_0 = 1 - p$ gives the Bernoulli distribution $\text{Bern}(p)$. This comes from flipped a p-coin.

Definition 4.3. Toss N p-coins and count the number of 1's (heads). So $\Omega = \{0, 1, \ldots, N\}$ and

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

this is called the Binomial distribution $\text{Bin}(N, p)$.

Note that

$$\sum_{i=0}^{N} p_k = \sum_{i=0}^{N} \binom{N}{k} p^k (1-p)^{N-k} = 1$$

by Binomial Theorem.

Definition 4.4. Consider k boxes and we throw N independent balls in them randomly, and p_i is the probability that one of the balls fall into box i. So $\Omega = \{(n_1, n_2, \ldots, n_k) \in \mathbb{N}_0^k : \sum_{r=1}^k n_r = N\}$ and we have

$$\mathbb{P}((n_1, n_2, \ldots, n_k)) = \binom{N}{n_1, n_2, \ldots, n_k} p_1^{n_1} \cdots p_k^{n_k}$$

This is the multinomial distribution.

Definition 4.5. Toss a fair coin until we reach a head. So $\Omega = \{1, 2, \ldots\}$ and

$$p_k = \mathbb{P}(\text{tossed } k \text{ coins until there is a head}) = (1-p)^{k-1} p$$

Sometimes we shift it by 1 and say $\Omega = \{0, 1, \ldots\}$ and

$$p_k = \mathbb{P}(\text{tossed } k \text{ coins before there is a head}) = (1-p)^k p$$

This is called the geometric distribution $\text{Geom}(p)$.

Definition 4.6. The Poisson distribution is used to model the number of occurences of events in a given period of time. For example, number of customer entering a shop in a day.
We take $\Omega = \{0, 1, 2 \ldots\}$ and

$$\mathbb{P}(\{k\}) = e^{-\lambda} \frac{\lambda^k}{k!}$$

We call this the Poisson distribution $\text{Pois}(\lambda)$ of parameter λ.

Consider the partition of $[0,1]$ in n intervals of length $1/n$ and in each interval customer arrives with probability p and at most n customers will arrive, then

$$\mathbb{P}(k \text{ customers arrived}) = \binom{n}{k} p^k (1-p)^{n-k}$$

which is $\text{Bin}(n,p)$. But if we take $p = \lambda/n$, we have

Proposition 4.1. $\text{Bin}(n, \lambda/n) \to \text{Pois}(\lambda)$ *as* $n \to \infty$.

Proof. Fix k, then

$$\begin{aligned}
\mathbb{P}(\{k\}) &= \binom{n}{k} \frac{\lambda^k}{n^k} \left(1 - \frac{\lambda}{n}\right)^{n-k} \\
&= \frac{\lambda^k}{k!} \frac{n!}{n^k (n-k)!} (1 + \frac{-\lambda}{n})^{n-k} \\
&\to \frac{\lambda_k}{k!} e^{-\lambda}
\end{aligned}$$

as $n \to \infty$, which is exactly the Poisson distribution. □

4.2 Random Variables

Definition 4.7. Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space, a random variable X is a function $\Omega \to \mathbb{R}$ such that
1. $\forall x \in \mathbb{R}, \{X \leq x\} = \{\omega \in \Omega : X(\omega) \leq x\} \in \mathscr{F}$. [2]

Example 4.1. Given $A \in \mathscr{F}$, the indicator of A is a function $1_A : \Omega \to \{0,1\}$ with

$$1_A(\omega) = \begin{cases} 1, \text{ if } \omega \in A \\ 0, \text{ otherwise} \end{cases}$$

This is a random variable

Definition 4.8. Let X be a random variable, then the probability distribution function of X to be a function $F_X : \mathbb{R} \to [0,1]$ given by $F_X(x) = \mathbb{P}(X \leq x) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \leq x\})$.

Proposition 4.2. *1.* $F_X(x) \to 1$ *as* $x \to \infty$ *and* $F_X(x) \to 0$ *as* $x \to -\infty$.
2. F_X *is increasing.*
3. F_X *is right continuous, that is,*

$$\lim_{u \to 0^+} F_X(x+u) = F_X(x)$$

Proof. Immediate. □

Definition 4.9. Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space and $(X_1, X_2, \ldots, X_n)$ is called a random variable in $\mathbb{R}^n$ if it is a function $\Omega \to \mathbb{R}^n$ and for any $(x_1, x_2 \ldots, x_n) \in \mathbb{R}^n$,

$$\{X_1 \leq x_1, \ldots, X_n \leq x_n\} = \{\omega \in \Omega : X_1(\omega) \leq x_1, \ldots, X_n(\omega) \leq x_n\} \in \mathscr{F}$$

Or equivalently, each X_i is a random variable in $\mathbb{R}$.

[2] Sometimes we also write $\{X \in A\} = \{\omega \in \Omega : X(\omega) \in A\}$

4.3 Discrete Random Variables

Definition 4.10. We call X a discrete random variable if it is a real random variable and the probablity space is discrete.
In this case, we define the function $p_k = \mathbb{P}(X = x) = \mathbb{P}(\{\omega \in \Omega : X(\omega) = x\})$ as the probability mass function of X.

Definition 4.11. The discrete random variables $X_1, X_2, \ldots, X_n$ are independent if

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \mathbb{P}(X_1 = x_1) \cdots \mathbb{P}(X_n = x_n)$$

Example 4.2. If we toss a p-coin n times independently, then $p_{\omega_1,\ldots,\omega_n} = \sum_{k=1}^{n} p^{\omega_k}(1-p)^{1-\omega_k}$ where $(\omega_1, \ldots, \omega_n) \in \{0,1\}^n$.
Define $X_k(\omega_1, \ldots, \omega_n) = \omega_k$, then $\mathbb{P}(X_k = 1) = p$ and $\mathbb{P}(X_k = 0) = p$, so X_k has distribution $\text{Bern}(p)$. Furthermore, any X_k are independent.
For $\omega = \omega_1, \ldots, \omega_n$, let $S_n(\omega) = \sum_k X_k(\omega)$, so S_n counts the number of heads and has distribution $\text{Bin}(n,p)$ as we have $\mathbb{P}(S_n = k) = \binom{n}{k} p^k (1-p)^{n-k}$.

4.4 Discrete Expectation

Consider a discrete probability space $(\Omega, \mathscr{F}, \mathbb{P})$ and a random variable $X : \Omega \to \mathbb{R}$, then

Definition 4.12. X is called nonnegative if $X(\Omega) \subset \mathbb{R}_{\geq 0}$.

Definition 4.13. For a non-negative random variable X, the expectation of X is the sum

$$\mathbb{E}[X] = \sum_{\omega \in \Omega} X(\omega)\mathbb{P}(\{\omega\})$$

Since X is nonnegative, this sum is either 0 or approaches $+\infty$.

Lemma 4.3.

$$\mathbb{E}[X] = \sum_{x \in \Omega_X} x\mathbb{P}(X = x)$$

Proof. Write $\Omega_X = \{X(\omega) : \omega \in \Omega\}$, then we immediately have

$$\Omega = \bigcup_{x \in \Omega_X} \{X = x\} \left(= \coprod_{x \in \Omega_X} \{X = x\} \right)$$

from definition. Use this to expand the formula yields

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{\omega \in \Omega} X(\omega)\mathbb{P}(\{\omega\}) \\
&= \sum_{x \in \Omega_X} \sum_{\omega \in \{X = x\}} x\mathbb{P}(\{\omega\}) \\
&= \sum_{x \in \Omega_X} x \sum_{\omega \in \{X = x\}} \mathbb{P}(\{\omega\}) \\
&= \sum_{x \in \Omega_X} x\mathbb{P}(X = x)
\end{aligned}$$

As desired. □

As one may expect, the expectation can be interpreted as a weighted average of the values taken by the random variable by the respective probabilities.

Example 4.3. 1. Suppose $X \sim \text{Bin}(n,p)$, then

$$\begin{aligned}\mathbb{E}[X] &= \sum_{k=0}^{n} k\binom{n}{k}p^k(1-p)^{n-k} \\ &= \sum_{k=1}^{n} n\binom{n-1}{k-1}p^k(1-p)^{n-k} \\ &= np\sum_{i=0}^{n-1}\binom{n-1}{i}p^i(1-p)^{n-i-1} \\ &= np(p+1-p)^{n-1} = np\end{aligned}$$

2. Suppose $X \sim \text{Pois}(\lambda)$, then

$$\begin{aligned}\mathbb{E}[X] &= \sum_{n=0}^{\infty} n\frac{\lambda^n}{n!}e^{-\lambda} \\ &= \lambda\sum_{n=1}^{\infty}\frac{\lambda^{n-1}}{(n-1)!}e^{-\lambda} \\ &= \lambda\end{aligned}$$

Let X be a general random variable, we can try to define its expectation by decomposing the positive and negative parts, that is

Definition 4.14. We decompose X into $X_+ = \max\{X,0\} = (X+|X|)/2$ and $X_- = \max\{-X,0\} = (|X|-X)/2$, then $X = X_+ - X_-$ and $|X| = X_+ + X_-$. If either $\mathbb{E}[X_+]$ or $\mathbb{E}[X_-]$ is finite, we define $\mathbb{E}[X] = \mathbb{E}[X_+] - \mathbb{E}[X_-]$.
Also, if $\mathbb{E}[|X|] < \infty$, we call X integrable..

Lemma 4.4. *We still have Lemma 4.3.*

Proof. Direct calculation. □

Proposition 4.5. *1. $X \geq 0 \implies \mathbb{E}[X] \geq 0$. In particular, if $X \geq 0$ and $\mathbb{E}[X] = 0$, then $\mathbb{P}(X=0) = 1$.*
2. Let $c \in \mathbb{R}$, then $\mathbb{E}[cX] = c\mathbb{E}[X]$ and $\mathbb{E}[c+X] = +E[X]$. In particular $\mathbb{E}[c] = c$.
3. Let X, Y be random variables, then $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$.

Proof. Trivial. □

2 and 3 are called *linearity of expectation*. And in fact, we can extend it and say for a sequence of random variables X_i we have $\mathbb{E}\left[\sum_n X_n\right] = \sum_n \mathbb{E}[X_n]$.

Proposition 4.6. *1. $\forall A \in \mathscr{F}$, we have $\mathbb{E}[1_A] = \mathbb{P}(A)$.*
2. (Law Of The Unconscious Statistician, LOTUS) Let $g : \mathbb{R} \to \mathbb{R}$ be a function and X a random variable, then

$$\mathbb{E}[g(X)] = \sum_{x\in\Omega_X} g(x)\mathbb{P}(X=x)$$

3. If a random variable X only takes integral value, then

$$\mathbb{E}(X) = \sum_{k=1}^{\infty} \mathbb{P}(X \geq k) = \sum_{k=0}^{\infty} \mathbb{P}(X > k)$$

Proof. 1 is from definition. For 2, we have

$$\begin{aligned}
\mathbb{E}[g(X)] &= \sum_{y \in \Omega_{g(X)}} y\mathbb{P}(Y = y) \\
&= \sum_{y \in \Omega_{g(X)}} y\mathbb{P}(\{\omega : g(X(\omega)) = y\}) \\
&= \sum_{y \in \Omega_{g(X)}} y\mathbb{P}(\{\omega : X(\omega) \in g^{-1}(\{y\})\}) \\
&= \sum_{y \in \Omega_{g(X)}} \sum_{x \in g^{-1}(\{y\})} y\mathbb{P}(X = x) \\
&= \sum_{x \in \Omega_X} g(x)\mathbb{P}(X = x)
\end{aligned}$$

3 is obvious since we can decompose $X = \sum_{k=1}^{\infty} 1_{\{X \geq k\}}$. □

One of the usage of 1 above is to give the following proof of Inclusion-Exclusion.

Another Proof of Inclusion-Exclusion. The indicator function satisfies $1_{A^c} = 1 - 1_A, 1_{A_1 \cap \cdots \cap A_n} = 1_{A_1} \cdots 1_{A_n}$, therefore

$$1_{A_1 \cup \cdots \cup A_n} = 1_{(A_1^c \cap \cdots \cap A_n^c)^c} = 1 - \prod_{i=1}^{n} (1 - 1_{A_i})$$

Expanding and taking expectation gives the result. □

Definition 4.15. Let X be a random variable and $n \in \mathbb{N}$, we call $\mathbb{E}[X^n]$ the n^{th} moment of X, if it is well-defined.

4.5 Variance

Definition 4.16. Let X be a random variable such that $\mathbb{E}[X]$ exists and is finite, then the variance of X is defined as $\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$.

Intuitively, the variance is the measure of how "spread out" the random variable is. So a smaller variance would mean that the random variable is largely concentrated in its expected value. Indubitably $\mathrm{Var}(X) \geq 0$, so we can define

Definition 4.17. The standard deviation is defined as $\mathrm{SD}(X) = \sqrt{\mathrm{Var}(X)}$.

Proposition 4.7. *1. If* $\mathrm{Var}(X) = 0$*, then* $\mathbb{P}(X = \mathbb{E}[X]) = 1$.
2. Let $c \in \mathbb{R}$*, then* $\mathrm{Var}(cX) = c^2 \mathrm{Var}(X)$.
3. $\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$.
4. $\forall x \in \mathbb{R}, \mathrm{Var}(X) \leq \mathbb{E}[(X - x)^2]$ *and equality hold iff* $x = \mathbb{E}[X]$.

Proof. Trivial. □

Example 4.4. 1. For $X \sim \text{Bin}(n,p)$, we have

$$\begin{aligned}\text{Var}(X) &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \\ &= \mathbb{E}[X(X-1)] + \mathbb{E}[X] - (\mathbb{E}[X])^2 \\ &= \sum_{k=0}^{n} k(k-1)\binom{n}{k}p^k(1-p)^{n-k} + np - n^2p^2 \\ &= p^2n(n-1)\sum_{k=0}^{n}\binom{n-2}{k-2}p^{k-2}(1-p)^{n-k} + np - n^2p^2 \\ &= np(1-p)\end{aligned}$$

2. For $X \sim \text{Pois}(\lambda)$, we can calculate

$$\begin{aligned}\text{Var}(X) &= \mathbb{E}[X(X-1)] + \mathbb{E}[X] - (\mathbb{E}[X])^2 \\ &= \sum_{k=0}^{\infty} k(k-1)\frac{\lambda^k}{k!}e^{-\lambda} + \lambda - \lambda^2 \\ &= \lambda^2 + \lambda - \lambda^2 \\ &= \lambda\end{aligned}$$

Definition 4.18. Let X, Y be two random variables, we define the covariance to be

$$\text{Cov}(X,Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

The covariance is a measure of the interdependence of X, Y.

Proposition 4.8. *1.* $\text{Cov}(X,Y) = \text{Cov}(Y,X)$.
2. $\text{Cov}(X,X) = \text{Var}(X)$.
3. $\text{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$.
4. Suppose c is a constant, then $\text{Cov}(cX,Y) = c\,\text{Cov}(X,Y), \text{Cov}(c+X,Y) = \text{Cov}(X,Y)$.
5. $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) + 2\,\text{Cov}(X,Y)$. *Easy to generalize it to finite sums*
6. For any constant c, $\text{Cov}(c,X) = 0$.
7. $\text{Cov}(X+Z,Y) = \text{Cov}(X,Y) + \text{Cov}(Z,Y)$.

Proof. Trivial. □

Definition 4.19. X, Y are called independent if $\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y)$.

Proposition 4.9. *Let X, Y be independent and f, g be nonnegative functions, then* $\mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)]\mathbb{E}[g(Y)]$

Proof.

$$\begin{aligned}\mathbb{E}[f(X)g(Y)] &= \sum_{x,y} f(x)g(y)\mathbb{P}(X=x,Y=y)\\ &= \sum_{x,y} f(x)g(y)\mathbb{P}(X=x)\mathbb{P}(Y=y)\\ &= \left(\sum_x f(x)\mathbb{P}(X=x)\right)\left(\sum_y f(y)\mathbb{P}(Y=y)\right)\\ &= \mathbb{E}[f(X)]\mathbb{E}[g(Y)]\end{aligned}$$

As desired. □

In particular, if X, Y are independent, then $\mathrm{Cov}(X,Y) = 0$. The converse, however, is not true.

Example 4.5 (Non-example)**.** Let $X_1, X_2, X_3 \sim \mathrm{Bern}(1/2)$. Define $Y_1 = 2X_1 - 1, Y_2 = 2X_2 - 1, Z_1 = Y_1X_3, Z_2 = Y_2X_3$. Now $\mathrm{Cov}(Z_1, Z_2) = 0$ but Z_1, Z_2 are not independent as $\mathbb{E}[Z_1 = 0, Z_2 = 0] = 1/2 \neq 1/4 = \mathbb{E}[Z_1 = 0]\mathbb{E}[Z_2 = 0]$.

5 Inequalities

Theorem 5.1 (Markov's Inequality)**.** *Suppose we have a nonnegative random variable X, then for any $a > 0$,*

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$$

Proof. Observe that $X \geq a1_{X\geq a}$ by checking each $\omega \in \Omega$, then we take expectation to get the inequality. □

Theorem 5.2 (Chebyshev's Inequality)**.** *Let X be a random variable with finite expectation, then for all $a > 0$,*

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq a) \leq \frac{\mathrm{Var}(X)}{a^2}$$

Proof.

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq a) = \mathbb{P}((X - \mathbb{E}[X])^2 \geq a^2) \leq \frac{\mathrm{Var}(X)}{a^2}$$

by Markov's Inequality. □

Theorem 5.3 (Cauchy-Schwarz Inequality)**.** *Let X, Y be random variables, then $\mathbb{E}[|XY|] \leq \sqrt{\mathbb{E}[X^2]\mathbb{E}[Y^2]}$*

Proof. It suffices to prove the case where X, Y are nonnegative, in which case we can discard the absolute value. Also it is trivial when one of $\mathbb{E}[X^2], \mathbb{E}[Y^2]$ is infinite, so we assume that they are finite afterwards. So $\mathbb{E}[XY] \leq (\mathbb{E}[X^2] + \mathbb{E}[Y^2])/2$ is also finite. Also if $\mathbb{E}[X^2]\mathbb{E}[Y^2] = 0$ then the (in)equality is trivial, so we assume henceforth that they are both positive. There we have $0 \leq (X - tY)^2 = X^2 - 2tXY + t^2Y^2$, from where taking expectation on both sides then yields $\mathbb{E}[X^2] - 2t\mathbb{E}[XY] + t^2\mathbb{E}[Y^2] \geq 0$ for any t. The inequality follows by taking determinant, also equality happens iff $X = tY$ for some t, i.e. X, Y are linearly dependent. □

Theorem 5.4 (Jensen's Inequality)**.** *Let f be a convex function defined on $\mathbb{R}$, that is for any $x, y \in \mathbb{R}, t \in [0,1]$,*

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)$$

Then if X is a random variable,

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$$

Proof. A convex function is equal to the supremum of all lines lying below it. That is, $\forall m \in \mathbb{R}, \exists a, b \in \mathbb{R}$ such that $f(m) = am + b$ and $\forall x \in \mathbb{R}, ax + b \leq f(x)$. This is obvious by choosing x, y with $x < m < y$ and apply the definition of convex functions, whence $\exists a \in \mathbb{R}$,

$$\frac{f(m) - f(x)}{m - x} \leq a \leq \frac{f(y) - f(m)}{y - m}$$

For any x it gives $f(x) \geq a(x - m) + f(m)$, so we just take $b = -am + f(m)$. Now we go back to our proof. Let $m = \mathbb{E}[X]$, then we can choose a, b such that $f(X) \geq aX + b$ and $f(m) = am + b$, taking expectation gives $f(\mathbb{E}[X]) = a\mathbb{E}[X] + b \leq \mathbb{E}[f(X)]$. □

Obviously we want to know when does the equality case hold in Jensen's Inequality. Let f be a convex function, assume that $\exists m \in \mathbb{R}$ such that $f(m) = am + b$ and $f(x) > ax + b$ for some a, b real. Suppose $m = \mathbb{E}[X]$, then consider $Y = f(X) - (aX + b)$ which is a nonnegative random variable. Assume that $\mathbb{E}[f(X)] = f(\mathbb{E}[X])$, then $\mathbb{E}[Y] = 0$, so $\mathbb{P}(Y = 0) = 1$, hence $\mathbb{P}(f(X) = aX + b) = 1$, therefore $\mathbb{P}(X = m) = 1$.

Corollary 5.5. *Let f be convex and let $x_1, \ldots, x_n$ be real, then*

$$\frac{1}{n}\sum_{k=1}^{n} f(x_k) \geq f\left(\frac{1}{n}\sum_{k=1}^{n} x_k\right)$$

Proof. Take X be a random variable taking values uniformly in $x_1, \ldots, x_n$ and apply Jensen's Inequality. □

Corollary 5.6. *For positive $x_1, \ldots, x_n$,*

$$\frac{1}{n}\sum_{k=1}^{n} x_k \geq \left(\prod_{k=1}^{n} x_k\right)^{1/n}$$

Proof. Take $f = -\log$. □

6 Conditional Expectation

Definition 6.1. Let B be an event such that $\mathbb{P}(B) > 0$ and let X be a random variable, then we define

$$\mathbb{E}[X|B] = \frac{\mathbb{E}[X \cdot 1_B]}{\mathbb{P}(B)}$$

Proposition 6.1 (Law of Total Expectation)**.** *Let Ω_n be a sequence of events that partitions Ω, then*

$$\mathbb{E}[X] = \sum_n \mathbb{E}[X|\Omega_n]\mathbb{P}(\Omega_n)$$

Proof. Write $X = \sum_n X \cdot 1_{\Omega_n}$. □

Definition 6.2. Let $X_1, \ldots, X_n$ be random variables, then the joint distribution is $\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n), x_i \in \omega_{X_i}$. Given such a joint distribution, the marginal distribution of X_i is defined as

$$\mathbb{P}(X_i = x_i) = \sum_{j \neq i, x_j \in \Omega_{X_j}} \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n)$$

Definition 6.3. Let X, Y be random variables, then the conditional probability given $Y = y, y \in \Omega_Y$ is defined by

$$\mathbb{P}(X = x|Y = y) = \mathbb{P}(X = x, Y = y)/\mathbb{P}(Y = y)$$

Immediately

$$\mathbb{P}(X = x) = \sum_{y \in \Omega_Y} \mathbb{P}(X = x, Y = y) = \sum_{y \in \Omega_Y} \mathbb{P}(X = x|Y = y)\mathbb{P}(Y = y)$$

Note that if X, Y are independent, then

$$\begin{aligned}\mathbb{P}(X + Y = z) &= \sum_{y \in \Omega_Y} \mathbb{P}(X + Y = z|Y = y)\mathbb{P}(Y = y) \\ &= \sum_{y \in \Omega_Y} \mathbb{P}(X = z - y|Y = y)\mathbb{P}(Y = y) \\ &= \sum_{y \in \Omega_Y} \mathbb{P}(X = z - y)\mathbb{P}(Y = y)\end{aligned}$$

Similarly

$$\mathbb{P}(X + Y = z) = \sum_{x \in \Omega_X} \mathbb{P}(X = x)\mathbb{P}(Y = z - x)$$

Example 6.1. Let $X \sim \text{Pois}(\lambda), Y \sim \text{Pois}(\mu)$, and X, Y are independent, then

$$\mathbb{P}(X + Y = n) = \sum_{r=0}^{\infty} \mathbb{P}(X = r)\mathbb{P}(X = n - r) = e^{-(\lambda+\mu)} \frac{(\lambda + \mu)^n}{n!}$$

Definition 6.4. Let X, Y be random variables, The conditional expectation of X given $Y = y$ is defined to be the expectation of the conditional distribution of X given $Y = y$, so

$$\mathbb{E}[X|Y = y] = \frac{\mathbb{E}[X \cdot 1_{Y=y}]}{\mathbb{P}(Y = y)} = \sum_{x \in \Omega_X} x\mathbb{P}(X = x|Y = y)$$

Note that $g(y) = \mathbb{E}[X|Y = y]$ is a function in $\Omega_Y \to \mathbb{R}$.

Definition 6.5. The conditional expectation of X given Y is defined to be $\mathbb{E}[X|Y] = g(Y)$ which is a random variable as a function of Y.

So obviously,

$$\begin{aligned}\mathbb{E}[X|Y] &= g(Y) \\ &= \sum_{y\in\Omega_Y} g(Y)\mathbb{P}(1_{Y=y}) \\ &= \sum_{y\in\Omega_Y} 1_{Y=y}g(y) \\ &= \sum_{y\in\Omega_Y} 1_{Y=y}\mathbb{E}[X|Y=y]\end{aligned}$$

Example 6.2. Toss a p-coin several times independently, so they produce a sequence of random variables $(X_i)_{i\in\mathbb{N}} \sim \text{Bern}(p)$ and the number of heads are distributed in $Y_m = X_1 + X_2 + \cdots + X_m \sim \text{Bin}(m,p)$ We want to calculate $\mathbb{E}[X_i|Y_m]$. Note that $\mathbb{E}[X_i|Y_m = r] = \mathbb{P}(X_i = 1|Y_m = r) = r/m$, which means

$$\mathbb{E}[X_i|Y_m] = \frac{Y_m}{m}$$

Proposition 6.2. *1.* $\forall c \in \mathbb{R}, \mathbb{E}[cX|Y] = c\mathbb{E}[X|Y], \mathbb{E}[c|Y] = c$.
2.

$$\mathbb{E}\left[\sum_{i=1}^n X_i \middle| Y\right] = \sum_{i=1}^n \mathbb{E}[X_i|Y]$$

3. $\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X]$.

Proof. 1 and 2 are obvious. For 3,

$$\begin{aligned}\mathbb{E}[\mathbb{E}[X|Y]] &= \sum_{y\in\Omega_Y} \mathbb{P}(Y=y)\mathbb{E}[X|Y=y] \\ &= \sum_{y\in\Omega_Y}\sum_{x\in\Omega_X} x\mathbb{P}(X=x|Y=y)\mathbb{P}(Y=y) \\ &= \sum_{x\in\Omega_X} x\mathbb{P}(X=x) \\ &= \mathbb{E}[X]\end{aligned}$$

As desired. □

Intuitively we also have the following:

Proposition 6.3. *1. If X, Y are independent, then $\mathbb{E}[X|Y] = \mathbb{E}[X]$ which is constant.*
2. Suppose Y, Z are independent, then $\mathbb{E}[\mathbb{E}[X|Y]|Z] = \mathbb{E}[X]$.
3. Let $h : \mathbb{R} \to \mathbb{R}$, then $\mathbb{E}[h(Y)X|Y] = h(Y)\mathbb{E}[X|Y]$.

Proof. 1 is trivial.
For 2, we have $\mathbb{E}[X|Y]$ is independent of Z, so $\mathbb{E}[\mathbb{E}[X|Y]|Z] = \mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X]$ by 1.
As for 3, we have $\mathbb{E}[h(Y)X|Y = y] = h(y)\mathbb{E}[X|Y = y]$, the identity follows. □

Corollary 6.4. $\mathbb{E}[\mathbb{E}[X|Y]|Y] = \mathbb{E}[X|Y]$

Proof. Take $h(y) = \mathbb{E}[X|Y = y]$ in Proposition 6.3.3. □

7 Random Walks

Definition 7.1. A random (or stochastic) process is a sequence of random variables.

Definition 7.2. A random walk is a stochastic process $(X_n)_{n\in\mathbb{N}}$ such that $X_0 = x, X_n = x + Y_1 + \cdots + Y_n$ where $x \in \mathbb{R}$ and Y_i are i.i.d. random variables.

Now we focus on the case where Y_i is 1 with probability p and -1 with probability $1-p$ and x is an integer. We think of X_n as the fortune of a gambler, and he gets it go up by 1 with probability p and lose with probability $q = 1-p$. His objective is to get to $a \in \mathbb{Z}$ before he get to 0. We say this is the event A. Let $h_x = \mathbb{P}(A|X_0 = x)$, then $h_x = ph_{x+1} + (1-p)h_{x-1}, p_a = 1, p_0 = 0$. For $p = 1/2 = 1-p$, we get $h_x = x/a$ which is perhaps not surprising.
For general p, by trying solutions of the form λ^x, we get $p\lambda^2 - \lambda + q = 0$, which has solutions $\lambda = 1, q/p$, and plugging in the boundary conditions yield

$$h_x = \frac{(q/p)^x - 1}{(q/p)^a - 1}$$

Now we want to estimate the expected time to absorption, that is the smallest T with $X_T \in \{0, a\}$. We write τ_x for such T with $X_0 = x$. we have $\tau_x = 1 + p\tau_{x+1} + q\tau_{x-1}, 0 < x < a, \tau_0 = \tau_a = 0$. For $p = 1/2 = 1-p$, we get the solution $\tau_x = x(a-x)$. For general p, we get the solution

$$\tau_x = \frac{1}{q-p}x - \frac{a}{q-p}\frac{(q/p)^x - 1}{(q/p)^a - 1}$$

8 Probability Generating Functions

8.1 Definitions and Examples

Definition 8.1. Let X random variable with $\Omega_X \subset \mathbb{N}$, the Probability distribution function as defined previously would be $p_r = \mathbb{P}(X = r), \forall r \in \mathbb{N}$. The probability generating function (PGF) is defined to be

$$p(z) = \sum_{r=0}^{\infty} p_r z^r = \mathbb{E}[z^X]$$

For $|z| \leq 1$, the series converges absolutely by comparison test, so this function is well defined in the interval $[-1, 1]$.

Theorem 8.1. *The PGF uniquely determines the distribution of the random variable.*

Proof. Suppose

$$\sum_{r=0}^{\infty} p_r z^r = \sum_{r=0}^{\infty} q_r z^r$$

Plugging in $z = 0$ gives $p_0 = q_0$. Suppose $p_r = q_r$ for any $r \in \{1, 2, \ldots, n\}$, we can get

$$\sum_{r=n+1}^{\infty} p_r z^r = \sum_{r=n+1}^{\infty} q_r z^r$$

dividing both sides by z^{n+1} and sending $z \to 0$ shows $p_{n+1} = q_{n+1}$. □

Theorem 8.2.

$$\lim_{z \to 1^-} p'(z) = p'(1-) = \mathbb{E}[X]$$

Proof. Assume first that $\mathbb{E}[X] < \infty$. For $0 < z < 1$ we have

$$p'(z) = \sum_{r=0}^{\infty} r p_r z^{r-1} \leq \sum_{r=0}^{\infty} r p_r = \mathbb{E}[X] \implies \lim_{z \to 1^-} p'(z) \leq \mathbb{E}[X]$$

The limit exists since p' is increasing. Also for any $\epsilon > 0$, for large enough N we have

$$\sum_{r=0}^{N} r p_r \geq \mathbb{E}[X] - \epsilon$$

Now since $p'(z) \geq \sum_{r=0}^{N} r p_r z^{r-1}$, we have

$$\lim_{z \to 1^-} p'(z) \geq \sum_{r=0}^{N} r p_r \geq \mathbb{E}[X] - \epsilon$$

Hence we must have $p'(1-) = \mathbb{E}[X]$.
Now if $\mathbb{E}[X] = \infty$, then $\forall M \in \mathbb{N}$, for large enough N we have

$$\sum_{r=0}^{N} r p_r > M$$

Thus by the same argument $p'(1-) \geq M$, hence $p'(1-) = \infty$. □

Similarly and by induction, we have

$$p^{(k)}(1-) = \mathbb{E}[X(X-1)\cdots(X-k+1)]$$

In particular

$$\mathrm{Var}(X) = p''(1-) + p'(1-) - (p'(1-))^2$$

Also by simply differentiating we get $p_n = \mathbb{P}(X = n) = P^{(n)}(0)/n!$.

Proposition 8.3. *Let $X_1, \ldots, X_n$ be independent random variables with PGFs q_i, then the PGF p of $X_1 + \cdots + X_n$ satisfies $p = \prod_i q_i$.*

Proof. Obvious. □

In particular, if X_i's are i.i.d., then $p = q^n$.

Example 8.1. 1. Let $X \sim \text{Bin}(n,p)$, then

$$p(z) = \mathbb{E}[z^X] = \sum_{k=0}^{n} \binom{n}{k} z^k p^k (1-p)^{n-k} = (1-p+zp)^n$$

Therefore if we have independent random binomial variables $X \sim \text{Bin}(n,p), Y \sim \text{Bin}(m,p)$, then $\mathbb{E}[z^{X+Y}] = (1-p+pz)^{n+m}$, hence $X+Y \sim \text{Bin}(n+m,p)$ as expected.
2. Let $X \sim \text{Geom}(p)$, then

$$p(z) = \mathbb{E}[z^X] = \sum_{r=0}^{\infty} z^r (1-p)^r p = \frac{p}{1-z(1-p)}$$

3. Suppose $X \sim \text{Pois}(\lambda)$, then

$$p(z) = \sum_{r=0}^{\infty} z^r e^{-\lambda} \frac{\lambda^r}{r!} = e^{(z-1)\lambda}$$

So for independent $X \sim \text{Pois}(\lambda), Y \sim \text{Pois}(\mu)$, we have $\mathbb{E}[z^{X+Y}] = e^{(z-1)(\lambda+\mu)}$, so $X+Y \sim \text{Pois}(\lambda+\mu)$.

8.2 Summing up a Random Number of Random Variables

Suppose $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables and N be a random variable, independent of all X_i, taking value in $\mathbb{N}$. Let $S_n = X_1 + \cdots + X_n$ and we consider the random variable S_N with $S_N(\omega) = X_1(\omega) + \cdots + X_{N(\omega)}(\omega)$. Legt q be the PGF of N and p the PGF of X_1 and r be the PGF of S_N, then we have

$$\begin{aligned}
r(z) &= \mathbb{E}[z^{S_N}] \\
&= \mathbb{E}[z^{X_1+\cdots+X_N}] \\
&= \mathbb{E}\left[\sum_{n=0}^{\infty} z^{X_1+\cdots+X_n} 1_{(N=n)}\right] \\
&= \sum_{n=0}^{\infty} \mathbb{E}[z^{X_1+\cdots+X_n} 1_{(N=n)}] \\
&= \sum_{n=0}^{\infty} \mathbb{E}[z^{X_1+\cdots+X_n}]\mathbb{P}(N=n) \\
&= \sum_{n=0}^{\infty} p(z)^n \mathbb{P}(N=n) \\
&= \mathbb{E}[p(z)^N] \\
&= q(p(z))
\end{aligned}$$

We can do it another way using conditional expectations. We have

$$r(z) = \mathbb{E}[z^{S_N}] = \mathbb{E}[\mathbb{E}[z^{X_1+\cdots+X_N}|N]]$$

We have $\mathbb{E}[z^{X_1+\cdots+X_N}|N=n] = \mathbb{E}[z^{X_1}]^n = p(z)^n$, therefore $\mathbb{E}[z^{X_1+\cdots+X_N}|N] = p(z)^N$, hence $r(z) = \mathbb{E}[p(z)^N] = q(p(z))$.
In particular, $[S_N] = \lim_{z\to 1^-} r'(z) = q'(p(1-))p'(1-) = \mathbb{E}[N]\mathbb{E}[X_1]$. Similarly $\text{Var}(S_N) = \mathbb{E}[N]\,\text{Var}(X_1) + \text{Var}(N)\mathbb{E}[X_1]^2$.

8.3 Branching Processes

The branching processes is a family of stochastic process that is developed by Bienaymé, Galton and Watson (therefore also called the Bienaymé-Galton-Watson processes).
Suppose $(X_n)_{n\geq 0}$ is a random process with $X_0 = 1$ and X_n equals the number of individials in generation n. The individual at time 0 produces a random number of offspring with probability distribution $\mathbb{P}(X_1 = k) = g_k, k = 0, 1, \ldots$, and every individual in the first generation produces an independent number of offspring (which is independent for different individials in generation n) with the same distribution. We call the distribution of X_1 the offspring distribution. Continuing this way, we consider a sequence $(Y_{k,n})_{k\geq 1, n\geq 0}$ be i.i.d. as X_1, then we can write

$$X_{n+1} = \begin{cases} Y_{1,n} + \cdots + Y_{X_n,n}, \text{ if } X_n \geq 1 \\ 0, \text{ otherwise} \end{cases}$$

So we can say $Y_{k,n}$ is the number of offspring of k^{th} individual in generation n.

Theorem 8.4. $\mathbb{E}[X_n] = (\mathbb{E}[X_1])^n, \forall n \geq 1$.

Proof. We proceed by induction. $n = 1$ is obvious. We set $\mu = \mathbb{E}[X_1]$. Suppose $\mathbb{E}[X_n] = \mu^n$, then for $n + 1$, $\mathbb{E}[X_{n+1}] = \mathbb{E}[\mathbb{E}[X_{n+1}|X_n]]$. But note that for any m, we have

$$\mathbb{E}[X_{n+1}|X_n = m] = \mathbb{E}[Y_{1,n} + \cdots + Y_{m,n}|X_n = m] = m\mu$$

SO $\mathbb{E}[X_{n+1}|X_n] = X_n\mu$, therefore $\mathbb{E}[X_{n+1}] = \mu^n\mu = \mu^{n+1}$. □

Theorem 8.5. *Set* $G(z) = \mathbb{E}[z^{X_1}]$ *and* $G_n(z) = \mathbb{E}[z^{X_n}]$, *then*

$$G_n(z) = G_{n-1}(G(z)) = G^n(z)$$

Proof. Induction again. $n = 1$ is from definition. Now assuming the previous cases, we have

$$G_{n+1}(z) = \mathbb{E}[z^{X_{n+1}}] = \mathbb{E}[\mathbb{E}[z^{X_{n+1}}|X_n]]$$

But now we have $\mathbb{E}[z^{X_{n+1}}|X_n = m] = \mathbb{E}[z^{X_1}]^m$ by independence, therefore $\mathbb{E}[z^{X_{n+1}}|X_n] = G(z)^{X_n}$, hence

$$\mathbb{E}[z^{X_{n+1}}] = \mathbb{E}[G(z)^{X_n}] = G_n(G(z))$$

As wanted. □

Certainly we would want to calculate the probability of extinction. Write q as the probability that $X_n = 0$ for some $n \geq 1$, and define q_n be the probability of $X_n = 0$. Let $A_n = \{X_n = 0\}$, then $A_n \subset A_{n+1}$ and by continuity of probability measure we get

$$q = \mathbb{P}\left(\bigcup_{n\in\mathbb{N}} A_n\right) = \lim_{n\to\infty} \mathbb{P}(A_n) = \lim_{n\to\infty} q_n$$

But we know that for any n, $q_n = \mathbb{P}(X_n = 0) = G_n(0)$, therefore $q_{n+1} = G(G_n(0)) = G(q_n)$. So since q_n does tends to a limit and G is continuous, we have $q = G(q)$.

Another way to get $q_{n+1} = G(q_n)$ is by observing that we can also condition on X_1. Conditioning on $X_1 = m$, we can write $X_{n+1} = X_n^{(1)} + \cdots + X_n^{(m)}$ where $X_n^{(i)}$ are i.i.d. branching processes corresponding to the i^{th} individual in the first generation. So

$$\begin{aligned}
q_{n+1} &= \mathbb{P}(X_{n+1} = 0) \\
&= \sum_m \mathbb{P}(X_{n+1} = 0 | X_1 = m)\mathbb{P}(X_1 = m) \\
&= \sum_m \mathbb{P}(X_n^{(1)} + \cdots + X_n^{(m)} | X_1 = m)\mathbb{P}(X_1 = m) \\
&= \sum_m \mathbb{P}(X_n^{(1)} = \cdots = X_n^{(m)} = 0 | X_1 = m)\mathbb{P}(X_1 = m) \\
&= \sum_m q_n^m \mathbb{P}(X_1 = m) \\
&= G(q_n)
\end{aligned}$$

Theorem 8.6. *The extinction probability q is the smallest non-negative solution to the equation $q = G(q)$. Also provided $\mathbb{P}(X_1 = 1) \leq 1$ we have $q \leq 1 \iff \mu \geq 1$.*

Proof. Let t be the smallest nonnegative solution to $q = G(q)$, we will show $q = t$. Note that $q_0 = 0 \leq t$, and if $q_n \leq t$, $q_{n+1} = G(q_n) \leq G(t) = t$ since G is increasing in $[0, 1]$. Hence $0 \leq q \leq t$ by Squeeze Theorem, so $q = t$ by minimality of t.
Now consider $H(z) = G(z) - z$, then $H''(z) = \sum_{r=2}^{\infty} r(r-1)g_r z^{r-2}$. If we assume $g_0 + g_1 < 1$ (trivial otherwise), then $H''(z) > 0$ for any $z \in (0, 1)$, therefore H' is strictly increasing in $[0, 1]$, which means (by Rolle's Theorem) that H can have at most one solution other than 1 in $[0, 1]$.
If H has no root in $[0, 1)$, then since $H(0) = G(0) \geq 0$, $\forall z \in [0, 1], H(z) \geq 0$. Now

$$H'(1-) = \lim_{z \to 1^-} \frac{H(1) - H(z)}{1 - z} = \lim_{z \to 1^-} -\frac{H(z)}{1 - z} \leq 0$$

So $\mu = G'(1-) = H'(1-) + 1 \leq 1$.
Otherwise, let r be a root of H in $[0, 1)$, then $G(r) = r$. But G' is strictly increasing, so H' must have a root z in $[r, 1)$ by Rolle's. Hence $G'(z) = 1$, but G' is increasing, so $\mu = G'(1-) > G'(z) = 1$. □

9 Continuous Random Variables

For a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ and $X : \Omega \to \mathbb{R}$ such that $\{X \leq x\} = \{\omega \in \Omega : X(\omega) \leq x\} \in \mathscr{F}$. Recall that $\forall x \in \mathbb{R}$ we defined the probability distribution function $F(x) = \mathbb{P}(X \leq x)$.

Proposition 9.1. *1. $x \leq y \implies F(x) \leq F(y)$.*
2. $\forall a < b, \mathbb{P}(a < X \leq b) = F(b) - F(a)$.
3. F is right continuous and the left limits always exists as well. Also,

$$F(x-) = \lim_{y \to x^-} F(y) \leq F(x)$$

4. We have

$$\lim_{x\to-\infty} F(x) = 0, \lim_{x\to\infty} F(x) = 1$$

Proof. Trivial. □

Definition 9.1. X is a continuous random variable if F is continuous.

So for a continuous random variable X and any $x \in \mathbb{R}$,

$$F(x-) = F(x) \implies \mathbb{P}(X < x) = \mathbb{P}(X \le x) \implies \mathbb{P}(X = x) = 0$$

In this course, we shall only consider the case where F is also differentiable. Set $f(x) = F'(x)$. Call f the probability density function of X.

Proposition 9.2. *1.* $f(x) \ge 0$.
2.

$$\int_{-\infty}^{x} f(y)\,\mathrm{d}y = F(x)$$

In particular,

$$\int_{-\infty}^{\infty} f(y)\,\mathrm{d}y = 1$$

Proof. Obvious. □

Intuitively, for Δx small, we have

$$\mathbb{P}(x < X \le x + \Delta x) = \int_{x}^{x+\Delta x} f(y)\,\mathrm{d}y \approx f(x)\Delta x$$

which is like approximating a continuous random variable by the probability density function of a discrete random variable.
Now define, as usual, $X_+ = \max\{X, 0\}$ and $X_- = \min\{X, 0\}$.

Definition 9.2. Let $Y \ge 0$ be a nonnegative continuous random variable, then define

$$\mathbb{E}[Y] = \int_{0}^{\infty} yf(y)\,\mathrm{d}y$$

And for $g \ge 0$ we define, for any Y (not necessarily positive),

$$\mathbb{E}[g(Y)] = \int_{-\infty}^{\infty} g(y)f(y)\,\mathrm{d}y$$

If at least one of $\mathbb{E}[X_+], \mathbb{E}[X_-]$ is finite, we define

$$\mathbb{E}[X] = \mathbb{E}[X_+] - \mathbb{E}[X_-] = \int_{-\infty}^{\infty} xf(x)\,\mathrm{d}x$$

Now the expectation is again a linear function from definition.

Proposition 9.3. *Let X be nonnegative, then*

$$\mathbb{E}[X] = \int_{0}^{\infty} \mathbb{P}(X \ge x)\,\mathrm{d}x$$

Note that due to lack of measure theory some of the steps are not fully justified.

First proof. Write

$$X = \int_0^\infty 1_{X \geq x} \, \mathrm{d}x$$

then taking expectation on both sides yields what we want. □

Second proof.

$$\begin{aligned}
\mathbb{E}[X] &= \int_0^\infty x f(x) \, \mathrm{d}x \\
&= \int_0^\infty \int_0^x f(x) \, \mathrm{d}y \, \mathrm{d}x \\
&= \int_0^\infty \int_y^\infty f(x) \, \mathrm{d}x \, \mathrm{d}y \\
&= \int_0^\infty \mathbb{P}(X \geq x) \, \mathrm{d}x
\end{aligned}$$

As we wish. □

Definition 9.3. The variance is also defined as $\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$.

Example 9.1. 1. The uniform distribution $\mathrm{Unif}(a, b)$ for $a < b$, where

$$f(x) = \begin{cases} 1/(b-a), \text{ if } x \in [a, b] \\ 0, \text{ otherwise} \end{cases}$$

Suppose X has density f, then

$$F(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^x f(x) \, \mathrm{d}x = \begin{cases} 0, \text{ if } x < a \\ (x-a)/(b-a), \text{ if } x \in [a, b] \\ 1, \text{ if } x > b \end{cases}$$

So we have

$$\mathbb{E}[X] = \int_{\mathbb{R}} x f(x) \, \mathrm{d}x = \int_a^b x f(x) \, \mathrm{d}x = \frac{a+b}{2}$$

2. The exponential distribution $\mathrm{Exp}(\lambda)$ has density function $f(x) = \lambda e^{-\lambda x}$ where $\lambda, x > 0$. Therefore for x positive,

$$F(x) = \int_0^x \lambda e^{-\lambda x} \, \mathrm{d}x = 1 - e^{-\lambda x}$$

and one can check that $\mathbb{E}[X] = \lambda^{-1}$.
The exponential distribution is kind of a limit of the geometric distribution. Let $T \sim \mathrm{Exp}(\lambda)$ and $T_n = \lfloor nT \rfloor$, so

$$\mathbb{P}(T_n \geq k) = \mathbb{P}(T \geq k/n) = e^{-k\lambda/n} = (e^{-\lambda/n})^k$$

So T_n is distributed geometrically with parameter $p_n = 1 - e^{-\lambda/n}$, so as $n \to \infty$, we have $p_n \sim \lambda/n$ and $T_n/n \to T$. So the exponential distribution arises as the

limit of a rescaled sequence of geometrics.
The exponential distribution also has the memoryless property, that is, for any $t, s \geq 0$,

$$\mathbb{P}(T > s + t | T > s) = \mathbb{P}(T > t)$$

Conversely, if T has the memoryless property, then let $g(t) = \mathbb{P}(T > t)$ (which is decreasing), then $g(t + s) = g(t)g(s)$. Now g must be exponential in $\mathbb{Q}$ by simple iteration. Also g is continuous, so g must be the exponential function (with appropriate scalings), so T is distributed exponentially.

Theorem 9.4. *Let X be a continuous random variable with density f. Let g be a continuous, strictly monotone function with g^{-1} differentiable. Then $g(X)$ is a continuous random variable with density $f(g^{-1}(x))|\mathrm{d}g^{-1}(x)/\mathrm{d}x|$.*

Proof. First assume that g is strictly increasing, then $\mathbb{P}(g(X) \leq x) = \mathbb{P}(X \leq g^{-1}(x)) = F(g^{-1}(x))$, so the density of $g(X)$ would be

$$\frac{\mathrm{d}}{\mathrm{d}x} F(g^{-1}(x)) = f(g^{-1}(x)) \frac{\mathrm{d}g^{-1}(x)}{\mathrm{d}x} = f(g^{-1}(x)) \left| \frac{\mathrm{d}g^{-1}(x)}{\mathrm{d}x} \right|$$

As for g strictly decreasing, $\mathbb{P}(g(X) \leq x) = \mathbb{P}(X \geq g^{-1}(x)) = 1 - F(g^{-1}(x))$, so then density is

$$\frac{\mathrm{d}}{\mathrm{d}x}(1 - F(g^{-1}(x))) = -f(g^{-1}(x)) \frac{\mathrm{d}g^{-1}(x)}{\mathrm{d}x} = f(g^{-1}(x)) \left| \frac{\mathrm{d}g^{-1}(x)}{\mathrm{d}x} \right|$$

as desired. □

Example 9.2. For $\mu \in \mathbb{R}, \sigma > 0$, we define the density of the normal distribution to be

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)}$$

We know how to integrate the Gaussian integral, so we can verify that this is indeed a density. Now for a random variable X having this density, we have

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} \frac{x}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)} \, \mathrm{d}x = \mu$$

So μ is actually the expected value. Also

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \int_{-\infty}^{\infty} \frac{(x-\mu)^2}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)} \, \mathrm{d}x = \sigma^2$$

This is called the normal distribution $\mathcal{N}(\mu, \sigma^2)$. We call $\mathcal{N}(0, 1)$ the standard normal. So the distribution and density of the standard normal would be

$$\Phi(x) = \int_{-\infty}^{x} \frac{e^{-t^2/2}}{\sqrt{2\pi}} \, \mathrm{d}t, \phi(x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}}$$

If we have $X \sim \mathcal{N}(\mu, \sigma^2)$, then for $a, b \in \mathbb{R}$ with $a \neq 0$, $Y = aX + b$ would have $\mathbb{E}[Y] = a\mu + b, \mathrm{Var}(Y) = a^2\sigma^2$ (note that X needs not be normal to get

these), we shall show that Y is also normal with $Y \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$. Let $g(x) = ax + b$, then $Y = g(X)$, we know that the density f_Y of Y has

$$f_Y(y) = f(g^{-1}(Y)) \left| \frac{\mathrm{d}g^{-1}(y)}{\mathrm{d}y} \right| = \frac{1}{\sqrt{2\pi a^2\sigma^2}} e^{-(y-(a\mu+b))^2/(2a^2\sigma^2)}$$

by some calculations. And this is what we wanted. So $(X - \mu)/\sigma \sim \mathcal{N}(0, 1)$. More than 95% of the normal distribution lies within two standard deviations of the mean, that is $\mathbb{P}(|(X - \mu)/\sigma| < 2) = \mathbb{P}(\mu - 2\sigma < X \leq \mu + 2\sigma) \geq 95\%$. By some computer stuff or other calculation methods, $\mathbb{P}(|(X - \mu)/\sigma| < 2) \geq \mathbb{P}(|(X - \mu)/\sigma| < 1.96) \approx 0.95$

Definition 9.4. Let X be a continuous random variable with density f, the median m of X is a number satisfying $\mathbb{P}(X > m) = \mathbb{P}(X \leq m) = 1/2$.

The median always exists due to intermediate value theorem. For symmetric distributions, i.e. $f(\mu + x) = f(\mu - x)$, then μ is its median.

10 Multivariate Distributions

10.1 Joint and Marginal Distributions

Let X be a continuous random variable has density f, then we know that

$$\mathbb{P}(X \leq x) = \int_{-\infty}^{x} f(y)\,\mathrm{d}y$$

It can be proved that this can generalize to an arbitrary (measurable) subsets $B \subset \mathbb{R}$, that is,

$$\mathbb{P}(X \in B) = \int_B f(x)\,\mathrm{d}x$$

Definition 10.1. For a tuple of random variables $X = (X_1, \ldots, X_n) \in \mathbb{R}^n$ where X_i are continuous, X is said to have density f if

$$\begin{aligned} F(x_1, \ldots, x_n) &= \mathbb{P}(X_1 \leq x_1, \ldots, X_n \leq x_n) \\ &= \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f(y_1, \ldots, y_n)\,\mathrm{d}y_n \cdots \mathrm{d}y_1 \end{aligned}$$

So $f(x_1, \ldots, x_n) = \partial^n F/(\partial x_1 \cdots \partial x_n)$. We can also generalize this to (measurable) subsets $B \subset \mathbb{R}^n$ by saying

$$\mathbb{P}(X \in B) = \int_B f(x_1, \ldots, x_n)\,\mathrm{d}x_1 \cdots \mathrm{d}x_n$$

Let $g : \mathbb{R}^n \to \mathbb{R}^+$, we define

$$\mathbb{E}[g(X)] = \int_{\mathbb{R}^n} g(x_1, \ldots, x_n) f(x_1, \ldots, x_n)\,\mathrm{d}x_1 \cdots \mathrm{d}x_n$$

Definition 10.2. We say $X_1, \ldots, X_n$ are independent if $\mathbb{P}(X_1 \leq x_1, \ldots, X_n \leq x_n) = \mathbb{P}(X_1 \leq x_1) \cdots \mathbb{P}(X_n \leq x_n)$.

Theorem 10.1. *Let $X = (X_1, \dots, X_n)$ have density f, then*
1. If $X_1, \dots, X_n$ are independent and X_i has density f_i, then

$$f(x_1, \dots, x_n) = \prod_{i=1}^{n} f_i(x_i)$$

2. Conversely, suppose we have the above formula for some nonnegative functions f_i, then $X_1 \dots, X_n$ are independent and have density functions proportional to f_i.

Proof. 1. We know

$$\begin{aligned}\mathbb{P}(X_1 \le x_1, \dots, X_n \le x_n) &= \prod_{i=1}^{n} \int_{-\infty}^{x} f_i(y_i)\,\mathrm{d}y_i \\ &= \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} \prod_{i=1}^{n} f(y_i)\,\mathrm{d}y_n \cdots \mathrm{d}y_1\end{aligned}$$

2. By multiplying constant, we may assume that

$$\int_{-\infty}^{\infty} f_i(x)\,\mathrm{d}x = 1$$

for all i. So for $B_i \subset \mathbb{R}$ (measurable),

$$\mathbb{P}(X \in B_1 \times \cdots \times B_n) = \int_{B_1} \cdots \int_{B_n} \prod_{i=1}^{n} f(y_i)\,\mathrm{d}y_n \cdots \mathrm{d}y_1$$

Now fix i and let $B_j = \mathbb{R}$ for any $j \neq i$, expanding the integral then gives

$$\mathbb{P}(X_i \in B_i) = \mathbb{P}(X_i \in B_i, X_j \in B_j, j \neq i) = \int_{B_i} f_i(y_i)\,\mathrm{d}y_i$$

So f_i is the density of X_i. Independence follows. □

Proposition 10.2. *Let $X = (X_1, \dots, X_n)$ have density f, we have*

$$\mathbb{P}(X_i \le x) = \int_{-\infty}^{x} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \dots, x_n)\,\mathrm{d}x_1 \cdots \mathrm{d}x_{i-1}\,\mathrm{d}x_{i+1} \cdots \mathrm{d}x_n\,\mathrm{d}x_i$$

So the density of X_i is

$$f_{X_i}(x) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \dots, x_n)\,\mathrm{d}x_1 \cdots \mathrm{d}x_{i-1}\,\mathrm{d}x_{i+1} \cdots \mathrm{d}x_n$$

Proof. Obvious. □

Definition 10.3. This is called the marginal density.

Definition 10.4. Let f, g be two densities on $\mathbb{R}$, their convolution is defined as

$$(f * g)(x) = \int_{-\infty}^{\infty} f(x - y) g(y)\,\mathrm{d}y$$

Let X, Y be two independent random variables with densities f_X, f_Y, then

$$\begin{aligned}
\mathbb{P}(X+Y \le z) &= \int_{\mathbb{R}^2} 1_{x+y\le z} f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \int_{x+y\le z} f_X(x) f_Y(y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{z-x} f_X(x) f_Y(y)\,\mathrm{d}y\,\mathrm{d}x \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{z} f_X(x) f_Y(y-x)\,\mathrm{d}y\,\mathrm{d}x \\
&= \int_{-\infty}^{z}\int_{-\infty}^{\infty} f_X(x) f_Y(y-x)\,\mathrm{d}x\,\mathrm{d}y \\
&= \int_{-\infty}^{z} (f*g)(y)\,\mathrm{d}y
\end{aligned}$$

So basically the convolution of f_X and f_Y is the density of $X+Y$. There is another way to obtain the same result, but it is not rigorous at all.

$$\begin{aligned}
\mathbb{P}(X+Y \le z) &= \int_{-\infty}^{\infty} \mathbb{P}(X+Y\le z, y\in \mathrm{d}y) \\
&= \int_{-\infty}^{\infty} \mathbb{P}(X+y\le z, y\in \mathrm{d}y) \\
&= \int_{-\infty}^{\infty} \mathbb{P}(X\le z-y)\mathbb{P}(y\in \mathrm{d}y) \\
&= \int_{-\infty}^{\infty} F_X(z-y) f_Y(y)\,\mathrm{d}y
\end{aligned}$$

Just differentiating and changing the order gives the result.

10.2 Conditionals

Definition 10.5. Let X, Y be continuous random variables with joint density $f_{X,Y}$, then the conditional density of X given $Y = y$ is

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

Given that f_Y is nonzero.

Definition 10.6. The conditional expectation of X given Y is $g(Y)$ where

$$g(y) = \int_{-\infty}^{\infty} x f_{X|Y}(x|y)\,\mathrm{d}x (= \mathbb{E}(X|Y=y))$$

We write $\mathbb{E}[X|Y] = g(Y)$.

10.3 Law of Total Probability

Proposition 10.3 (Law of Total Probability)**.**

$$f_X(x) = \int_{-\infty}^{\infty} f_{X|Y}(x|y) f_Y(y)\,\mathrm{d}y$$

Proof. Plug in definition. □

Example 10.1. Let $X \sim \text{Exp}(\lambda), Y \sim \text{Exp}(\mu)$ be independent. Set $Z = \min(X, Y)$. So

$$\begin{aligned}\mathbb{P}(Z \le z) &= 1 - \mathbb{P}(Z > z) = 1 - \mathbb{P}(X > z, Y > z) \\ &= 1 - \mathbb{P}(X > z)\mathbb{P}(Y > z) = 1 - e^{-\lambda z}e^{-\mu z} = 1 - e^{-(\lambda+\mu)z}\end{aligned}$$

Therefore $Z \sim \text{Exp}(\lambda + \mu)$. So by the same way, for $X_i \sim \text{Exp}(\lambda_i)$, then $\min(X_1, \dots, X_n) \sim \text{Exp}(\sum_i \lambda_i)$

10.4 Transformation of a Multidimensional Random Variable

Theorem 10.4. *Let X be a continuous random variable in $D \subset \mathbb{R}^d$ with density f_X. Let g be a bijection $D \to g(D)$ with continuous derivative on D and $\forall x \in D, \det g' \neq 0$. Set $y = g(x)$ and $Y = g(X)$, then the density of Y is given by*

$$f_Y(y) = f_X(x)|J| = f_X(x)\left\|\left(\frac{\partial x_i}{\partial y_j}\right)_{i,j=1}^{d}\right\|$$

Proof. Omitted. □

Example 10.2. Let $X, Y \sim \mathcal{N}(0, 1)$ be independent, and $R = \sqrt{X^2 + Y^2}, \Theta = \arg(X, Y)$, so $X = R\cos\Theta, Y = R\sin\Theta$. So we want the joint density of (R, Θ). We have

$$f_{R,\Theta}(r, \theta) = f_{X,Y}(x, y)|J| = rf_{X,Y}(x, y) = rf_X(x)f_Y(y) = \frac{r}{2\pi}e^{-r^2/2}$$

So R, Θ are independent with $\Theta \sim \text{Unif}(0, 2\pi)$ and R has density $f_R(r) = re^{-r^2/2}$.

10.5 Order Statistics of a Random Sample

Let $X_1, \dots, X_n$ be i.i.d. with distribution F and density f. If we put them in order from smallest to biggest, $X_{(1)} \le \cdots X_{(n)}$, then $Y_i = X_{(i)}$ is called the order statistics. So $\mathbb{P}(Y_1 \le x) = 1 - \mathbb{P}(Y_1 > x) = 1 - (1 - F(x))^n$ is the distribution of Y_1. The density then is $nf(x)(1-F(x))^{n-1}$. Also $\mathbb{P}(Y_n \le x) = (F(x))^n$, so it has density $nf(x)F(x)^{n-1}$. Now we want to find the joint density of $(Y_1, \dots, Y_n)$. Let $x_1 < \cdots < x_n$, we have

$$\begin{aligned}\mathbb{P}(Y_1 \le x_1, \dots, Y_n \le x_n) &= \sum_{\sigma \in S_n} \mathbb{P}(X_1 \le x_{\sigma(1)}, \dots X_n \le x_{\sigma(n)}) \\ &= n!\mathbb{P}(X_1 \le x_1, \dots X_n \le x_n) \\ &= n! \int_{-\infty}^{x_1} f(u_1) \int_{-\infty}^{x_2} f(u_2) \cdots \int_{-\infty}^{x_n} f(u_n)\,\mathrm{d}u_n \cdots \mathrm{d}u_1\end{aligned}$$

So by differentiating the density f_Y of Y_i's is $f_Y(y_1, \dots, y_n) = n!f(y_1) \cdots f(y_n)$ for $y_1 < y_2 < \cdots < y_n$ and 0 otherwise.

Example 10.3. Let $X_1, \dots, X_n$ be i.i.d. $\text{Exp}(\lambda)$ and Y_i be the order statistics and $Z_1 = Y_1, Z_i = Z_i - Z_{i-1}$ for $i = 2, \dots, n$. Then

$$Z = \begin{pmatrix} Z_1 \\ \vdots \\ Z_n \end{pmatrix} = A \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, A = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ -1 & 1 & 0 & \dots & 0 \\ 0 & -1 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix}$$

where we have $\det A = 1$ and $Y_j = \sum_{i=1}^{j} Z_i$, so for $y_j = \sum_{i=1}^{j} z_i$, we have

$$\begin{aligned} f_{Z_1,\dots,Z_n}(z_1, \dots, z_n) &= f_{Y_1,\dots,Y_n}(y_1, \dots, y_n)|J| \\ &= n! e^{-\lambda y_1} \cdots e^{-\lambda y_n} \\ &= n! \lambda^n e^{-\lambda(nz_1 + (n-1)z_2 + \cdots + 2z_{n-1} + z_n)} \\ &= \prod_{i=1}^{n} (n - i + 1) \lambda e^{-\lambda(n-i+1)z_1} \end{aligned}$$

So $(Z_1, \dots, Z_n)$ are independent exponentials with $Z_i \sim \text{Exp}(\lambda(n - i + 1))$.

11 Generation of Random Variables

Example 11.1. Suppose $U \sim \text{Unif}(0, 1)$ and set $Y = -\log U$, so $\mathbb{P}(Y \le x) = \mathbb{P}(U \ge e^{-x}) = 1 - e^{-x}$ which is $\text{Exp}(1)$.

Theorem 11.1. *Let X be a continuous random variable with distribution F and $U \sim \text{Unif}(0, 1)$, then $F^{-1}(U)$ has the same distribution as X.*

Proof. Obvious. □

Another way to generate is called the rejection sampling. Suppose $A \subset [0, 1]^d$ (measurable) and $f(x) = 1_A(x)/|A|$ where $|A|$ is the volume of A. We want a random variables X to have density f. Let U_n be an i.i.d. sequence of d-dimensional uniforms, i.e. $U_n = (U_{k,n} : k = 1, \dots, d)$ where $U_{k,n}$ are i.i.d. $\text{Unif}[0, 1]$. Let $N = \min\{n \ge 1 : U_n \in A\}$ and set $X = U_N$.

Proposition 11.2. *Let $B \in [0, 1]^d$, then $\mathbb{P}(X \in B) = |B \cap A|/|A|$.*

Proof.

$$\begin{aligned} \mathbb{P}(X \in B) &= \sum_{n=1}^{\infty} \mathbb{P}(X \in B, N = n) \\ &= \sum_{n=1}^{\infty} \mathbb{P}(U_n \in A \cap B, U_{n-1} \notin A, \dots, U_1 \notin A) \\ &= \sum_{n=1}^{\infty} |A \cap B|(1 - |A|)^{n-1} \\ &= \frac{|A \cap B|}{|A|} \\ &\left(= \int_B f(x) \, \mathrm{d}x \right) \end{aligned}$$

As desired. □

Let f be a bounded density on $[0,1]^{d-1}$, that is $\exists \lambda, \sup f \leq \lambda$ and we want to find $X \sim f$. Consider $A = \{(x_1, \ldots, x_d) \in [0,1]^d : x_d \in f(x_1, \ldots, x_{n-1})/\lambda\}$. Let $Y = (X_1, \ldots, X_d)$ be uniform on A generated as above and set $X = (X_1, \ldots, X_{d-1})$.

Proposition 11.3. *$X \sim f$.*

Proof. Let $B \subset [0,1]^{d-1}$, then

$$\begin{aligned}
\mathbb{P}(X \in B) &= \mathbb{P}((X_1, \ldots, X_d) \in (B \times [0,1]) \cap A) \\
&= \frac{(B \times [0,1] \cap A)}{|A|} \\
&= \frac{1}{|A|} \int \cdots \int_{[0,1]^d} 1_{(B \times [0,1]) \cap A}((x_1 \cdots x_d)) \, \mathrm{d}x_1 \cdots \mathrm{d}x_n \\
&= \frac{1}{|A|} \int \cdots \int_{[0,1]^d} 1_B((x_1, \ldots, x_{d-1})) \frac{f(x_1, \ldots, x_{d-1})}{\lambda} \, \mathrm{d}x_1 \cdots \mathrm{d}x_n \\
&= \int_B f(x) \, \mathrm{d}x
\end{aligned}$$

which is what we wanted. □

12 Moment Generating Functions

12.1 Moment Generating Function of One Random Variable

Definition 12.1. Let X have density f, then the moment generating function (MGF) is defined by

$$m(\theta) = \mathbb{E}[e^{\theta X}] = \int_{-\infty}^{\infty} e^{\theta x} f(x) \, \mathrm{d}x$$

wherever it is finite.

Note that $m(0) = 1$.

Theorem 12.1. *The MGF uniquely determines the distribution of random variable provided it is defined on some open interval around* 0.

Proof. Omitted. □

Theorem 12.2. *Suppose the MGF is defined for some open interval around* 0, *then* $m^{(r)}(0) = \mathbb{E}[X^r]$.

Proof. Omitted as well. □

Example 12.1. 1. Gamma distribution $X \sim \Gamma(n, \lambda)$ for $n \in \mathbb{N}, \lambda \geq 0$.

$$f(x) = e^{-\lambda x} \lambda^n \frac{x^{n-1}}{(n-1)!}$$

defined for $x \geq 0$. For $n = 1$, we get $\text{Exp}(\lambda)$. One can show by reduction formula that this is indeed a density.
In this case, we have

$$\begin{aligned} m(\theta) &= \mathbb{E}[e^{\theta X}] \\ &= \int_0^\infty e^{-(\lambda-\theta)x} \frac{\lambda^n}{(\lambda-\theta)^n} (\lambda-\theta)^n \frac{x^{n-1}}{(n-1)!} \, \mathrm{d}x \\ &= \left(\frac{\lambda}{\lambda-\theta}\right)^n \end{aligned}$$

taking $n = 1$ then gives the MGF of the exponential. Suppose now that $X_1, \cdots, X_n$ are independent, then the MGF of the sum

$$m(\theta) = \mathbb{E}[e^{\theta(X_1+\cdots+X_n)}] = \mathbb{E}[e^{\theta X_1}] \cdots \mathbb{E}[e^{\theta X_n}]$$

So if $X \sim \Gamma(n, \lambda), Y \sim \Gamma(m, \lambda)$ are independent, then

$$\mathbb{E}[e^{\theta(X+Y)}] = \left(\frac{\lambda}{\lambda-\theta}\right)^n \left(\frac{\lambda}{\lambda-\theta}\right)^m = \left(\frac{\lambda}{\lambda-\theta}\right)^{m+n} \sim \Gamma(n+m, \lambda)$$

In particular, if $X_1, \ldots, X_n$ are i.i.d. $\text{Exp}(\lambda)$, then $X_1 + \cdots + X_n \sim \Gamma(n, \lambda)$.
2. Suppose $X \sim \mathcal{N}(\mu, \sigma^2)$, then the MGF of X is

$$\begin{aligned} m(\theta) &= \mathbb{E}[e^{\theta X}] \\ &= \int_{-\infty}^\infty e^{\theta x} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)} \, \mathrm{d}x \\ &= e^{\mu\theta + \theta^2\sigma^2/2} \end{aligned}$$

After some completing square and calculation. Now if $X \sim \mathcal{N}(\mu, \sigma^2)$, then we know that $aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$. We can prove this again using MGF, indeed,

$$\mathbb{E}[e^{\theta(aX+b)}] = e^{\theta b} e^{a\theta\mu + (a\theta)^2\sigma^2/2} = e^{\theta(a\mu+b) + \theta^2(a^2\sigma^2)/2}$$

So $aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$. Now let $Y = \mathcal{N}(\nu, \tau^2)$ independent of X, then by using the same trick (MGF), we get $\mathbb{E}[e^{\theta(X+Y)}] = e^{\theta(\mu+\nu) + \theta^2(\sigma^2+\tau^2)/2}$ so $X + Y \sim \mathcal{N}(\mu + \nu, \sigma^2 + \tau^2)$.
3. (non-example) Cauchy's Distribition is obtained by $f(x) = 1/(\pi(1 + x^2))$ for $x \in \mathbb{R}$. Then we can get $m(\theta) = \infty$ for any $\theta \neq 0$, so if $X \sim f$ then $X, 2X, \ldots$ all have the same MGF but not the same distribution. So the assumption on $m(\theta)$ being finite on an open interval is essential.

12.2 Multivariate Moment generating Function

Definition 12.2. Let $X = (X_1, \ldots, X_n)$ be a random variable in $\mathbb{R}^n$, then the MGF of X is defined to be a function $m : \mathbb{R}^n \to \mathbb{R}$ with

$$m(\theta) = \mathbb{E}[e^{\theta^\top X}] = \mathbb{E}[e^{\theta_1 X_1 + \cdots + \theta_n X_n}]$$

where $\theta = (\theta_1, \ldots, \theta_n)^\top$.

Provided that $m(\theta)$ is finite for a certain range (which is out of the scope of this course) of values of θ, it uniquely characterizes the distribution of X. Also

$$\left.\frac{\partial^r m}{\partial \theta_i^r}\right|_{\theta=0} = \mathbb{E}[X_i^r], \left.\frac{\partial^{r+s} m}{\partial \theta_i^r \partial \theta_j^s}\right|_{\theta=0} = \mathbb{E}[X_i^r X_j^s]$$

Also $X_1, \ldots, X_n$ are independent iff

$$m(\theta) = \prod_{i=1}^{n} \mathbb{E}[e^{\theta_i X_i}]$$

Definition 12.3. Let $(X_n, n \in \mathbb{N})$ be a sequence of random variables, and let X be a random variable. We say $X_n \to X$ in distribution if

$$F_{X_n}(x) \to F_X(x), F_{X_n}(x) = \mathbb{P}(X_n \leq x), F_X(x) = \mathbb{P}(X \leq x)$$

For each $x \in \mathbb{R}$ such that F_X is continuous at x.

Theorem 12.3 (Continuity Theorem for MGFs)**.** *Let X be a random variable with MGF m and $m(\theta) < \infty$ for some $\theta \neq 0$. Suppose we have*

$$m_n(\theta) = \mathbb{E}[e^{\theta X_n}] \to m(\theta), \forall \theta \in \mathbb{R}$$

Then $X_n \to X$ in distribution.

Proof. Omitted. □

13 Limit Theorems

13.1 Law(s) of Large Numbers

Theorem 13.1 (Weak Law of Large Numbers)**.** *Let $(X_n : n \in \mathbb{N})$ be an iid sequence of random variables with finite expectation μ. Set $S_n = X_1 + \cdots + X_n$, then for any $\epsilon > 0$, we have*

$$\mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| > \epsilon\right) \to 0$$

as $n \to \infty$.

We shall prove this assuming $\mathrm{X}_1 = \sigma^2 < \infty$.

Proof. We have $\mathbb{E}[S_n/n] = \mu$ and $\mathrm{Var}(S_n/n) = \sigma^2/n$, then by Chebyshev's Inequality,

$$\mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| > \epsilon\right) \leq \frac{\mathrm{Var}(S_n/n)}{\epsilon^2} = \frac{\sigma^2}{\epsilon^2 n} \to 0$$

As $n \to \infty$. □

Definition 13.1. A sequence (X_n) converges to X in probability, written as

$$X_n \xrightarrow{\mathbb{P}} X, n \to \infty$$

if $\forall \epsilon > 0, \mathbb{P}(|X_n - X| > \epsilon) \to 0$ as $n \to \infty$.

So the weak law of large numbers says that $S_n/n \xrightarrow{\mathbb{P}} \mu$ as $n \to \infty$.

Definition 13.2. (X_n) converges to X with probability 1 (or "almost surely", a.s.) if

$$\mathbb{P}\left(\lim_{n\to\infty} X_n = X\right) = 1$$

Theorem 13.2 (Strong Law of Large Numbers)**.** *Let the setting be as before, then*

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{S_n}{n} \to \mu\right) = 1$$

Proof. Assume further that the 4^{th} moment is finite. BY considering $Y_i = X_i - \mu$, WLOG $\mu = 0$. We have

$$S_n^4 = \left(\sum_{i=1}^n X_i\right)^4 = \sum_{i=1}^n X_i^4 + 6 \sum_{1\le i<j\le n} X_i^2 X_j^2 + R$$

where R is a sum of terms of the form $X_i^3 X_j, X_i^2 X_j X_k, X_i X_j X_k X_l$ for i, j, k, l all distinct. Since X_i are independent with zero mean, $\mathbb{E}[R] = 0$, so

$$\begin{aligned}\mathbb{E}[S_n^4] &= n\mathbb{E}[X_1^4] + 3n(n-1)\mathbb{E}[X_1^2]^2 \\ &\le (n + 3n(n-1))\mathbb{E}[X_1^4] \\ &\le 3n^2\mathbb{E}[X_1^4]\end{aligned}$$

Hence

$$\mathbb{E}\left[\sum_{n=1}^\infty \left(\frac{S_n}{n}\right)^4\right] = \sum_{n=1}^\infty \mathbb{E}\left[\left(\frac{S_n}{n}\right)^4\right] \le 3\mathbb{E}[X_1^4]\sum_{n=1}^\infty \frac{1}{n^2} < \infty$$

So

$$\mathbb{P}\left(\sum_{n=1}^\infty \left(\frac{S_n}{n}\right)^4 < \infty\right) = 1 \implies \mathbb{P}\left(\lim_{n\to\infty} \frac{S_n}{n} = 0\right) = 1$$

as claimed. □

As the name suggests, he strong law of large numbers implies the weak law of large numbers.

Proposition 13.3. *Suppose $X_n \to \mu$ almost surely, then $X_n \xrightarrow{\mathbb{P}} \mu$*

By shifting, it suffices to consider the case where $\mu = 0$.

Proof. Assuming $X_n \to 0$ almost surely, then we have

$$\mathbb{P}(|X_n| \le \epsilon) \ge \mathbb{P}\left(\bigcap_{m=n}^\infty \{|X_m| \le \epsilon\}\right)$$

write the event in the right hand side as A_n, then $A_n \subset A_{n+1}$ and $\bigcup_n A_n$ is the event that $|X_m| \le \epsilon$ for any sufficiently large m. So

$$\lim_{n\to\infty} \mathbb{P}(|X_n| \le \epsilon) \ge \mathbb{P}\left(\bigcup_{n\in\mathbb{N}} A_n\right) \ge \mathbb{P}(X_n \to 0) = 1$$

So $\mathbb{P}(|X_n| < \epsilon) = 0$. □

13.2 Central Limit Theorem

We saw $S_n/n - \mu \to 0$ a.s. from the law of large numbers. We know also that $\mathrm{Var}(S_n/n) = \sigma^2/n$ where σ^2 is the variance of X_1. So if we want to normalize,

$$\frac{S_n/n - \mu}{\sqrt{\mathrm{Var}(S_n/n)}} = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

from which we will expect

Theorem 13.4 (Central Limit Theorem)**.** *Let $(X_n : n \in \mathbb{N})$ be a sequence of i.i.d. random variables with mean μ and variance σ^2, then set $S_n = X_1 + \cdots + X_n$ as before, we have*

$$\mathbb{P}\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \le x\right) \to \Phi(x) = \int_{-\infty}^{x} \frac{e^{-y^2/2}}{\sqrt{2\pi}}\,\mathrm{d}y$$

for all $x \in \mathbb{R}$.

In other words, $(S_n - n\mu)/(\sigma\sqrt{n}) \to \mathcal{N}(0,1)$ in distribution. What this means is that for n large enough, $S_n \approx n\mu + \sigma\sqrt{n}\mathcal{N}(0,1) = \mathcal{N}(\mu n, \sigma^2 n)$. In fact, not only does it converge, we can also estimate the rate of convergence, which is sadly beyond the course.

Proof. WLOG $\mu = 0, \sigma = 1$ by considering $(X_i - \mu)/\sigma$. Assume further that $\exists \delta > 0, \mathbb{E}[e^{\pm\delta X_1}] < \infty$. By continuity of MGFs, it suffices to show that, for $Z \sim \mathcal{N}(0,1)$,

$$\mathbb{E}[e^{\theta S_n/\sqrt{n}}] \to \mathbb{E}[e^{\theta Z}] = e^{\theta^2/2}$$

So if let m be the MGF of X_1,

$$\mathbb{E}[e^{\theta S_n/\sqrt{n}}] = \mathbb{E}[e^{\theta X_n/\sqrt{n}}]^n = m(\theta/\sqrt{n})^n$$

Note that when $|\theta| < \delta/2$,

$$m(\theta) = \mathbb{E}\left[\sum_{n=0}^{\infty} \frac{1}{n!}\theta^n X_1^n\right] = 1 + \frac{\theta^2}{2} + \mathbb{E}\left[\sum_{n=3}^{\infty} \frac{1}{n!}\theta^n X_1^n\right]$$

We will prove that the last term is $o(\theta^2)$ as $\theta \to 0$, which immediately implies the result.
We have

$$\begin{aligned}\sum_{n=3}^{\infty} \frac{1}{n!}|\theta|^n |X_1|^n &= |\theta X_1|^3 \sum_{k=0}^{\infty} \frac{|\theta X_1|^k}{(k+3)!} \\ &\le |\theta X_1|^3 e^{\delta|X_1|/2} \\ &\le 3! \left(\frac{2\theta}{\delta}\right)^3 e^{\delta|X_1|}\end{aligned}$$

Now by Jensen's Inequality

$$\begin{aligned}\left|\mathbb{E}\left[\sum_{n=3}^{\infty}\frac{1}{n!}\theta^n X_1^n\right]\right| &\le \mathbb{E}\left[\sum_{n=3}^{\infty}\frac{1}{n!}|\theta|^n|X_1|^n\right]\\ &\le 3!\left(\frac{2\theta}{\delta}\right)^3\mathbb{E}[e^{\delta|X_1|}]\\ &\le 3!\left(\frac{2\theta}{\delta}\right)^3(\mathbb{E}[e^{\delta|X_1|}]+\mathbb{E}[e^{-\delta X_1}])\\ &= o(|\theta|^2)\end{aligned}$$

As desired. □

There are few important application of central limit theorem.

Example 13.1. 1. Let (X_n) be i.i.d. Bern(p) and hence $S_n \sim \text{Bin}(n,p)$. Recall that $\mathbb{E}[S_n] = np$ and $\text{Var}(S_n) = np(1-p)$, so

$$\frac{S_n - np}{\sqrt{np(1-p)}} \to \mathcal{N}(0,1)$$

in distribution. So for n large, $S_n \approx \mathcal{N}(np, np(1-p))$ for n large. In the Poisson approximation of the binomial, we scaled p to λ/n, while in this approximation, we kept p constant.
2. We now want $S_n \sim \text{Pois}(n)$, and to accomplish this we can write $S_n = X_1 + \cdots + X_n$ where (X_n) are i.i.d. Pois(1). So $(S_n - n)/\sqrt{n} \approx \mathcal{N}(0,1)$.

13.3 Sampling Error by Central Limit Theorem

A proportion p of the population votes "yes" and $1-p$ votes "no" in a referendum. We want to estimate p with error at most $\pm 4\%$ in probability at least 0.99. We pick N individuals at random. Let S_N be the number of people who voted "yes", so we want to estimate p by $\hat{p}_N = S_N/N$, so what we want is

$$\mathbb{P}(|\hat{p}_N - p| \le 4\%) \ge 0.99$$

Now $S_N \sim \text{Bin}(N,p)$, so by previous,

$$\hat{p}_N = \frac{S_N}{N} \approx p + \sqrt{\frac{p(1-p)}{N}}Z, Z \sim \mathcal{N}(0,1)$$

for N large. So what we want is

$$\mathbb{P}\left(\sqrt{\frac{p(1-p)}{N}}|Z| \le 4\%\right) \ge 0.99$$

Now $\mathbb{P}(Z \ge z) = 2(1-\Phi(z))$, then $\mathbb{P}(|Z| \ge 2.58) = 0.01$. So in the worse case where $p = 1/2$ gives $N \ge 1040$.

14 Geometrical Probability

14.1 Buffon's Needle

Suppose we have parallel horizontal lines, each with distance L apart from each other and we have a needle with length $\ell \leq L$. Throw the needle at random, then we want to know the probbaility that it intersects at least one line. Suppose we have dropped the needle, then let Θ be its horizontal inclination and X the distance between the distance between the left end to the line above. So we take $X \sim \text{Unif}[0, L]$ and $\Theta \sim \text{Unif}[0, \pi]$ and they are independent. Hence

$$p = \mathbb{P}(\text{The needle intersects the lines}) = \mathbb{P}(X \leq \ell \sin \Theta)$$

So we can now calculate this by

$$\begin{aligned} p &= \mathbb{P}(\text{The needle intersects the lines}) \\ &= \mathbb{P}(X \leq \ell \sin \Theta) \\ &= \int_0^L \int_0^\pi 1_{x \leq l \sin\theta} f_{X,\Theta}(x, \theta) \, \mathrm{d}\theta \, \mathrm{d}x \\ &= \int_0^L \int_0^\pi 1_{x \leq l \sin\theta} \frac{1}{\pi L} \, \mathrm{d}\theta \, \mathrm{d}x \\ &= \frac{1}{\pi L} \int_0^\pi \ell \sin\theta \, \mathrm{d}\theta \\ &= \frac{2\ell}{\pi L} \end{aligned}$$

Hence $\pi = 2\ell/(pL)$. Now we want to approximate π by this experiment. Throw n needles independently and let $\hat{p}_n$ be the proportion of needles intersecting a line. We want to approximate p by $\hat{p}_n$ thus approximate π by $\hat{\pi}_n = 2\ell/(\hat{p}_n L)$. Suppose we want $\mathbb{P}(|\hat{\pi}_n - \pi| \leq 0.001) \geq 0.99$, we want to know how large n has to be.

Define $f(x) = 2\ell/(xL)$, then $f(p) = \pi$ and $f'(p) = -\pi/p$. Also $f(\hat{p}_n) = \hat{\pi}_n$. Let S_n be the number of needles intersecting a line, then $S_n \sim \text{Bin}(n, p)$, so $S_n \approx np + \sqrt{np(1-p)}Z$ where $Z \sim \mathcal{N}(0, 1)$. So $\hat{p}_n \approx p + \sqrt{p(1-p)/n}Z$. By Taylor's Theorem,

$$\hat{\pi}_n = f(\hat{p}_n) \approx f(p) + (\hat{p}_n - p) f'(p) = \pi - (\hat{p}_n - p)\pi/p$$

So when we substitute back, we obtain

$$\hat{\pi}_n - \pi \approx -\pi \sqrt{\frac{1-p}{pn}} Z$$

So

$$\mathbb{P}(|\hat{\pi}_n - \pi| \leq 0.001) = \mathbb{P}\left(\pi \sqrt{\frac{1-p}{pn}} |Z| < 0.001\right)$$

Now $\mathbb{P}(|z| \geq 2.58) < 0.01$. Also the variance of $\pi\sqrt{(1-p)/(pn)}Z$ is $\pi^2(1-p)/(pn)$ which is decreasing in p. We can minimize the variance by taking $\ell = L$, so $p = 2/\pi$ and the variance is $\pi^2(\pi/2 - 1)/n$, so in this case we need

$$\sqrt{\frac{\pi^2}{n}(\frac{\pi}{2} - 1)} 2.58 = 0.001 \implies n = 3.75 \times 10^7$$

which is quite large.

14.2 Bertrand's Paradox

We have a circle of radius r and draw a chord at random. We want to know the probability that it has length at most r. There are two ways to do this. The first approach is let $X \sim \text{Unif}(0, r)$ to be the perpendicular distance between the chord and the center of the circle. Let C be the length, then $C \leq r \iff 4X^2 \geq 3r^2$, so the probability is $1 - \sqrt{3}/2 \approx 0.134$.
There is a second approach. Fix one point of the chord and choose $\Theta \sim \text{Unif}(0, 2\pi)$ to be the angle between this point and the other point of the chord. So $C \leq r \iff \Theta \leq \pi/3 \vee \Theta \geq 2\pi - \pi/3$, hence the probability is $1/3$ which is far enough from 0.134.
But this is not a paradox since we are using essentially different sample spaces.

15 Multidimensional Gaussian

A random variable X in $\mathbb{R}$ is called Gaussian if $X \sim \mu + \sigma Z$ for $Z \sim \mathcal{N}(0, 1)$. It has density, as we have seen,

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)}$$

and we denote this by $X \sim \mathcal{N}(\mu, \sigma^2)$. We want to generalize this to higher dimensions.

Definition 15.1. A random variable $X = (X_1, \ldots, X_n)$ is Gaussian if for any $u \in \mathbb{R}^n$, $u^\top X$ is Gaussian in $\mathbb{R}$.
We call X a Gaussian vector.

Suppose now that we have an $n \times n$ matrix A and $b \in \mathbb{R}^n$, then $AX + b$ is also Gaussian. This is obvious from definition.
Set

$$\mu = \mathbb{E}[X] = \begin{pmatrix} \mathbb{E}[X_1] \\ \vdots \\ \mathbb{E}[X_n] \end{pmatrix}, V = \text{Var}(X) = \mathbb{E}[(X - \mu)(X^\top - \mu)] = (\text{Cov}(X_i, X_j))_{i,j}$$

Let $u \in \mathbb{R}^n$, then $\mathbb{E}[u^\top X] = u^\top \mu$ and the variance is $\text{Var}(u^\top X) = u^\top V u$, so $u^\top X \sim \mathcal{N}(u^\top \mu, u^\top V u)$. V is symmetric as Cov is, also by above $u^\top V u$ is always nonegative, so V is nonnegative definite. We want to know the MGF of X. Let $\lambda \in \mathbb{R}^n$, then $m(\lambda) = \mathbb{E}[e^{\lambda^\top X}]$, but we know the distribution of $\lambda^\top X$ which is $\mathcal{N}(u^\top \mu, u^\top V u)$, therefore $m(\lambda) = e^{\lambda^\top \mu + \lambda^\top V \lambda/2}$. So by uniqueness of MGFs, we see that the distribution of a Gaussian vector is uniquely characterized by its mean μ and variance V, so we write $X \sim \mathcal{N}(\mu, V)$.
As V is real symmetric and nonnegative definite, we can write $V = U^\top D U$ where U is real orthogonal and $D = \text{diag}(\lambda_1, \ldots, \lambda_n)$ where $\lambda_i \geq 0$ for all i. So we define the square root matrix of V to be

$$\sigma = U^\top \sqrt{D} U, \sqrt{D} = \text{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n})$$

So $\sigma^2 = V$, therefore we can write $X \sim \mathcal{N}(\mu, \sigma^2)$.
Let $Z_1, \ldots, Z_n$ be i.i.d. $\mathcal{N}(0, 1)$ and $Z = (Z_1, \ldots, Z_n)^\top$, then Z is Gaussian (by

e.g. looking at its MGF) and $Z \sim \mathcal{N}(0, I)$. Let $X = \mu + \sigma Z$ where $\mu \in \mathbb{R}^n$ and σ is real symmetric and nonnegative definite, then X is Gaussian since $x \mapsto \mu + \sigma x$ is linear. Also $\mathbb{E}[X] = \mu$ and

$$\operatorname{Var}(X) = \mathbb{E}[(X-\mu)(X-\mu)^\top] = \sigma \mathbb{E}[ZZ^\top]\sigma = \sigma I \sigma = \sigma^2$$

So $X \sim \mathcal{N}(\mu, \sigma^2)$.
We want to find the density function of a multivariate normal. Assuming that V is positive definite and $X \sim \mathcal{N}(\mu, V)$, then $\det V = \prod_i \lambda_i > 0$. Also since $X = \mu + \sigma Z$, we can invert to get $Z = \sigma^{-1}(X - \mu)$, hence

$$\begin{aligned} f_X(x) &= f_Z(z)|J| \\ &= \prod_{i=1}^n \frac{e^{-z_i^2/2}}{\sqrt{2\pi}} \det(\sigma^{-1}) = \frac{1}{\sqrt{(2\pi)^n \det V}} e^{-z^\top z/2} \\ &= \frac{1}{\sqrt{(2\pi)^n \det V}} e^{-(x-\mu)^\top V^{-1}(x-\mu)/2} \end{aligned}$$

If we only assume V is nonnegative definite, then we can have 0 eigenvalues hence 0 determinant. In this case, we can change the basis to get something of the form

$$\begin{pmatrix} U & 0 \\ 0 & 0 \end{pmatrix}$$

where U is positive definite $m \times m$ matrix and $\mu = (\lambda^\top, \nu^\top)^\top$ where $\lambda \in \mathbb{R}^m, \nu \in \mathbb{R}^n$, then we write $X = (Y^\top, \nu^\top)^\top$ where Y has

$$f_Y(y) = \frac{1}{\sqrt{(2\pi)^n \det U}} e^{-(y-\lambda)^\top U^{-1}(y-\lambda)/2}$$

Proposition 15.1. *Let $X = (X_1, \dots, X_n)$ be Gaussian, and suppose that $\operatorname{Cov}(X_i, X_j) = 0$ when $i \neq j$, then X_i are independent Gaussians.*

Note that the converse is obviously true.

Proof. The covariant matrix is diagonal, so the density factorizes. □

Another way to see it is by simply looking at the MGF.
Now we consider the bivariate Gaussians. Suppose $X = (X_1, X_2)$ is Gaussian and we set $\mu_k = \mathbb{E}[X_k], \sigma_k = \sqrt{\operatorname{Var}(X_k)}$ and suppose $\sigma_k > 0$, we define

$$\rho = \operatorname{corr}(X_1, X_2) = \frac{\operatorname{Cov}(X_1, X_2)}{\sqrt{\operatorname{Var}(X_1)\operatorname{Var}(X_2)}}$$

Hence we have

$$V = \operatorname{Var}(X) = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

We want to show that V is nonnegative definite for any ρ. Indeed, for any $x = (x_1, x_2)^\top$, we have $x^\top V x \geq 0$ by calculation. in particular, if $\rho = 0$ and $\sigma_k > 0$, then $f_{X_1,X_2}(x_1, x_2)$ can be found by multiplying the $\mathcal{N}(\mu_k, \sigma_k)$ since X_1, X_2 are independent as we have seen before.

Let $a \in \mathbb{R}$, then $\operatorname{Cov}(X_2 - aX_1, X_1) = \operatorname{Cov}(X_1, X_2) - a\operatorname{Var}(X_1) = \rho\sigma_1\sigma_2 - a\sigma_1^2$. Take $a = \rho\sigma_1/\sigma_2$ and $Y = X_2 - aX_1$, then $\operatorname{Cov}(X_1, Y) = 0$ and we can write

$$\begin{pmatrix} X_1 \\ Y \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -a & 1 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$$

So $(X_1, Y)^\top$ is also Gaussian. X_1, Y are independent and we can write $X_2 = Y + aX_1$ and $\mathbb{E}[X_2|X_1] = \mathbb{E}[Y] + aX_2$.

Theorem 15.2. *Let X be a Gaussian vector in $\mathbb{R}^k$ with finite variance. Suppose X has covariance matrix Σ. Let $X_1, \dots$ be i.i.d. copies of X, then*

$$S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mathbb{E}[X_i]) \to \mathcal{N}(0, \Sigma)$$

in the sense that for any (measurable) $B \in \mathbb{R}^k$, $\mathbb{P}(S_n \in B) \to \mathbb{P}(\mathcal{N}(0, \Sigma) \in B)$.

16 Bonus lectures

We have n bins $1, 2, \dots, n$ and n indistinguishable balls. For every ball, we pick a bin uniformly randomly and place it independently of other balls. Let X_i be the number of balls in bin i. Define the maximum load to be $M_n = \max_{i \le n} X_i$. Now for every i, $X_i \sim \operatorname{Bin}(n, 1/n)$. We first want to find heuristically the value of $\mathbb{P}(M_n \ge x)$. Note that we have $\mathbb{P}(M_n \ge x) \le n\mathbb{P}(X_1 \ge x)$. This is quite strict but works in this case. Now for large n, we approximate X_1 by $\operatorname{Pois}(1)$, so we can estimate (by something proved in example sheet) by $\mathbb{P}(\operatorname{Pois}(\lambda) \ge x) \le \exp(-x\log(x/\lambda) - \lambda + x)$, so

$$\mathbb{P}(M_n \ge x) \le n\mathbb{P}(X_1 \ge x) \approx n\mathbb{P}(\operatorname{Pois}(1) \ge x) \le \exp(-x\log(x) - 1 + x)$$

So for $\mathbb{P}(M_n \ge x) \to 0$, we need $x = (1 + \epsilon)\log n / \log\log n$.

Theorem 16.1. *We have*

$$\frac{M_n}{\log n / \log\log n} \xrightarrow{\mathbb{P}} 1$$

as $n \to \infty$.

Let $N \sim \operatorname{Pois}(\lambda)$ and let $X = \sum_{k=1}^{N} \xi_k$ where $\xi_k \sim \operatorname{Bern}(p)$, then $X \sim \operatorname{Pois}(\lambda p)$ and $N - X \sim \operatorname{Pois}(\lambda(1-p))$ and $X, N - X$ are independent, so

$$\mathbb{P}(X = x, N - X = y) = e^{-\lambda} \frac{\lambda^{x+y}}{(x+y)!} \binom{x+y}{x} p^x (1-p)^y$$

We cast a method called Poissonization.

Proof. Suppose we throw $\operatorname{Pois}(n(1+\epsilon))$ balls instead and let Y_i be the load of bin i, then $Y_i \sim \operatorname{Pois}(1+\epsilon)$ are i.i.d.. Set $\tilde{M}_n = \max_{i \le n} Y_i$, then

$$\begin{aligned}
\mathbb{P}(M_n \ge x) &\le \mathbb{P}(\tilde{M}_n \ge x, \operatorname{Pois}(n(1+\epsilon)) \ge n) + \mathbb{P}(\operatorname{Pois}(n(1+\epsilon)) < n) \\
&\le \mathbb{P}(\tilde{M}_n \ge x) + \mathbb{P}(\operatorname{Pois}(n(1+\epsilon)) < n) \\
&\le \mathbb{P}(\tilde{M}_n \ge x) + \exp(n\log(1+\epsilon) - \epsilon n) \\
&\le \mathbb{P}(\tilde{M}_n \ge x) + \exp\left(-\frac{n\epsilon^2}{10}\right)
\end{aligned}$$

for $\epsilon \in (0,1)$. Hence

$$\mathbb{P}\left(M_n \geq (1+\epsilon)\frac{\log n}{\log\log n}\right) \leq \mathbb{P}\left(\tilde{M}_n \geq (1+\epsilon)\frac{\log n}{\log\log n}\right) + \exp\left(-\frac{n\epsilon^2}{10}\right)$$

Note that $\exp(-n\epsilon^2/10) \to 0$ as $n \to \infty$. Now

$$P\left(\tilde{M}_n \geq (1+\epsilon)\frac{\log n}{\log\log n}\right) \leq n\mathbb{P}\left(Y_1 \geq (1+\epsilon)\frac{\log n}{\log\log n}\right)$$

which is bounded by

$$n\exp\left(-(1+\epsilon)\frac{\log n}{\log\log n}\log\left(\frac{\log n}{\log\log n}\right) - (1+\epsilon) + (1+\epsilon)\frac{\log n}{\log\log n}\right)$$

which is at most

$$n\exp\left(-(1+\epsilon)\log n + 10\frac{\log n \log\log\log n}{\log\log n}\right) \to 0$$

as $n \to \infty$, hence

$$\lim_{n\to\infty} \mathbb{P}\left(M_n \geq (1+\epsilon)\frac{\log n}{\log\log n}\right) \to 0$$

which establishes the upper bound.
For the lower bound, we need to show that for all $\epsilon \in (0,1)$, we have

$$\lim_{n\to\infty} \mathbb{P}\left(M_n \leq (1-\epsilon)\frac{\log n}{\log\log n}\right) \to 0$$

Note first that $\mathbb{P}(\text{Pois}(n(1-\epsilon)) > n) \leq e^{-n\epsilon^2/10}$, so

$$\mathbb{P}\left(M_n \leq (1-\epsilon)\frac{\log n}{\log\log n}\right) \leq e^{-n\epsilon^2/10} + \mathbb{P}\left(\tilde{M}_n \leq (1-\epsilon)\frac{\log n}{\log\log n}\right)$$

where $\tilde{M}_n = \max_{i\leq n} \tilde{Y}_i$ with $\tilde{Y}_i$ are i.i.d. Pois$(1-\epsilon)$. So

$$P\left(\tilde{M}_n \leq (1-\epsilon)\frac{\log n}{\log\log n}\right) = P\left(\tilde{Y}_1 \leq (1-\epsilon)\frac{\log n}{\log\log n}\right)^n$$

Now we have

$$P\left(\tilde{Y}_1 \geq (1-\epsilon)\frac{\log n}{\log\log n}\right) \geq e^{-(1-\epsilon)}\frac{(1-\epsilon)^M}{M!}, M = (1-\epsilon)\frac{\log n}{\log\log n}$$

Hence

$$\begin{aligned}\mathbb{P}(\tilde{M}_n < M) &\leq \left(1 - e^{-(1-\epsilon)}\frac{(1-\epsilon)^M}{M!}\right)^n \\ &\leq \exp\left(-ne^{-(1-\epsilon)}\frac{(1-\epsilon)^M}{M!}\right)\end{aligned}$$

Now $M! \leq M(M/e)^M$ for large enough M, so we (finally!) get $\mathbb{P}(\tilde{M}_n < M) = o(1)$, whcih establishes the result. $\square$

Throw n balls into n bins. Every time, we pick $d \geq 2$ bins at random and place the ball in the least loaded bin.

Theorem 16.2. *After all balls have been placed, the maximum load is*

$$\frac{\log \log n}{\log d} + O(1)$$

with probability $1 - o(1)$.

We will use the Chernoff inequality for binomials, i.e.

$$\mathbb{P}(\text{Bin}(n, p) \geq 2np) \leq e^{-np/3}$$

The height of the ball is the number of balls already placed in the bin it is placed +1. Let ν_i be the number of bins with load $\geq i$ and μ_i is the number of bins with height $\geq i$, then $\nu_i \leq \mu_i$. The idea is that we want to define a sequence β_i such that $\nu_i \leq \beta_i$ with high probability for all $i \leq i^\star$ where $i^\star$ is to be determined and we will show that $i^\star = \log \log n / \log d$. Then at this time, $\beta_{i^\star}$ will have order n and we can finish the proof easily from there. Suppose we condition on $\nu_i \leq \beta_i$, then the probability that a ball has height at least $i+1$ is bounded by $(\beta_i/n)^d$ since all of the d choices have to come from bins with load $\geq i$, so

$$\mathbb{P}(\nu_{i+1} > k) \leq \mathbb{P}(\text{Bin}(n, (\beta_i/n)^d))$$

Lemma 16.3. *Let $X_1, \ldots$ be a sequence of random variables and let Y_i be a function of $X_1, \ldots, X_i$ that takes values in $\{0, 1\}$. If $\mathbb{P}(Y_i = 1 | X_1, \ldots, X_{i-1}) \leq p$, then*

$$\mathbb{P}\left(\sum_{i=1}^{n} Y_i > k\right) \leq \mathbb{P}(\text{Bin}(n, p) > k)$$

Proof. Each Y_i is upper bounded by Bern(p). And for the sum we use induction. □

Proof of the Theorem. Let $\nu_i(t)$ be the number of bins with load $\geq i$ at time t (after the t^{th} ball is placed). Let $\mu_i(t)$ be the number of balls with height $\geq i$ at time t. Write $\nu_i(n) = \nu_i$ and $\mu_i(n) = \mu_i$, so $\nu_i(t) \leq \mu_i(t)$ for any i, t. Now we want to find a sequence β_i with $\nu_i \leq \beta_i$ for $i < i^\star$ with high probability. Let $\beta_4 = n/4$ and $\beta_{i+1} = 2n(\beta_i/n)^d$. Define $E_i = \{\nu_i \leq \beta_i\}$, so $\mathbb{P}(E_4) = 1$. Now we want to show that for all $4 \leq i < i^\star$ where $i^\star$ is to be determined, we have $\mathbb{P}(E_{i=1}^c) \leq \mathbb{P}(E_i^c) + 1/n^2$. Define a sequence of binomial variables $Y_t = 1_{h(t) \geq i+1, \nu_{i-1} \leq \beta_i}$. Let ω_j be bins selected by the j^{th} ball. Now

$$\mathbb{P}(Y_t = 1 | \omega_1, \ldots, \omega_{t-1}) \leq \left(\frac{\beta_i}{n}\right)^d = p_i$$

So by the preceding lemma, for any k,

$$\mathbb{P}\left(\sum_{t=1}^{n} Y_t > k\right) \leq \mathbb{P}(\text{Bin}(n, p) > k)$$

So

$$\mathbb{P}(E_{i+1}^c | E_i) = \mathbb{P}(\nu_{i+1} \geq \beta_{i+1} | E_i) \leq \mathbb{P}(\mu_{i+1} > \beta_{i+1} | E_i)$$

Conditioned on E_i, $Y_t = 1_{h(t)\geq i+1}$, so $\sum_{t=1}^n Y_t = \mu_{i+1}$, so

$$\begin{aligned}
\mathbb{P}(\mu_{i+1} > \beta_{i+1}|E_i) &= \mathbb{P}\left(\sum_{t=1}^n Y_t > \beta_{i+1}\middle| E_i\right) \\
&= \frac{1}{\mathbb{P}(E_i)}\mathbb{P}\left(\sum_{t=1}^n Y_t > \beta_{i+1}, E_i\right) \\
&\leq \frac{1}{\mathbb{P}(E_i)}\mathbb{P}\left(\sum_{t=1}^n Y_t > \beta_{i+1}\right) \\
&\leq \frac{1}{\mathbb{P}(E_i)}\mathbb{P}(\mathrm{Bin}(n,p_i) > \beta_{i+1}) \\
&\leq \frac{e^{-np_i/3}}{\mathbb{P}(E_i)}
\end{aligned}$$

Note that for any i with $np_i \geq 6\log n$ we have

$$\mathbb{P}(E_{i+1}^c|E_i) \leq \frac{1}{n^2\mathbb{P}(E_i)} \implies \mathbb{P}(E_{i+1}^c) \leq \frac{1}{n^2} + \mathbb{P}(E_i^c)$$

Let $i^\star$ be the first i such that $np_i < 6\log n$. Now we claim that

$$i^\star = \log\log n/\log d + O(1)$$

It suffices to show by induction that

$$\beta_{i+4} = \frac{n}{2^{2d^i} - \sum_{j=0}^{i-1} d^j}$$

Once we know that, we get $\beta_{i+4} \leq n/2^{d^i}$, so $\mathbb{P}(E_{i^\star}^c) \leq i^\star/n^2$, so $\beta_{i^\star+i} = 2np_{i^\star} \leq 2n(6\log n/n) = 12\log n$ Now

$$\begin{aligned}
\mathbb{P}(\nu_{i^\star+1} > 18\log n|E_{i^\star}) &\leq \mathbb{P}(\nu_{i^\star+1} > 18\log n|E_{i^\star}) \\
&\leq \frac{\mathbb{P}(\mathrm{Bin}(n, 6\log n/n^2) > 18\log n)}{\mathbb{P}(E_{i^\star})} \\
&\leq e^{-2\log n}/\mathbb{P}(E_{i^\star}) \\
&= \frac{1}{n^2\mathbb{P}(E_{i^\star})}
\end{aligned}$$

So

$$\mathbb{P}(\nu_{i^\star+1} \geq 18\log n) \leq \frac{1}{n^2} + \mathbb{P}(E_{i^\star}^c) \leq \frac{i^\star+1}{n^2}$$

Now $\{\nu_{i^\star+1} \geq 1\} \subset \{\mu_{i^\star+1} \geq 1\} \subset \{\mu_{i^\star+2} \geq 2\}$.

$$\begin{aligned}
\mathbb{P}(\mu_{i^\star+1} \geq 2|\nu_{i^\star+1} < 18\log n) &\leq \mathbb{P}(\mathrm{Bin}(n, (8\log n/n)^d) \geq 2)/\mathbb{P}(\nu_{i^\star+1} < 18\log n) \\
&\leq \binom{n}{2}\left(\frac{18\log n}{n}\right)^{2d} + \frac{i^\star+1}{n^2} \\
&\leq \frac{n^2}{n^{2d}}(18\log n)^{2d} + \frac{i^\star+1}{n^2} \\
&= o(1/n)
\end{aligned}$$

and this shows what we wanted. □