MATH 170B, PROBABILITY THEORY, WINTER 2017

STEVEN HEILMAN

ABSTRACT. These notes closely follow the book of Bertsekas and Tsitsiklis, available here.

CONTENTS

1. INTRODUCTION

These notes assume familiarity with the subject matter and notation from the Math 170A Probability class. For example, we will assume familiarity with sets, probability laws, independence, conditional expectation, and various commonly encountered random variables, both discrete and continuous.

In this course, we will generally focus on averages of independent random variables. In particular, we will prove two of the most significant theorems in probability: the Law of Large Numbers and the Central Limit Theorem.

Date: March 14, 2017

Suppose I flip a fair coin 10^9 times. Then I should expect to get roughly $\frac{1}{2}10^9$ heads and $\frac{1}{2}10^9$ tails. This is formalized in the Law of Large Numbers. Or, suppose I have a casino game where the casino wins 51% of the time. Then over a long period of time, the casino will make money; the Law of Large Numbers guarantees that! However, if I do flip 10^9 fair coins, it is unlikely that I will get *exactly* $\frac{1}{2}10^9$ heads. (What is the exact probability?) There will typically be some small fluctuations around $\frac{1}{2}10^9$. But about how close to $\frac{1}{2}10^9$ will the number of heads be? This question is answered precisely by the Central Limit Theorem. In your previous probability class, you may have mentioned the Central Limit Theorem applied to coin flips, which is known as the De Moivre-Laplace Theorem:

Theorem 1.1 (De Moivre-Laplace Theorem). *Let $X_1, \ldots, X_n$ be independent Bernoulli random variables with parameter $1/2$, so that $\mathbf{P}(X_1 = 1) = \mathbf{P}(X_1 = 0) = 1/2$. Recall that X_1 has mean $1/2$ and variance $1/4$. Let $a \in \mathbb{R}$. Then*

$$\lim_{n\to\infty} \mathbf{P}\left(\frac{X_1 + \cdots + X_n - (1/2)n}{\sqrt{n}\sqrt{1/4}} \leq a\right) = \int_{-\infty}^{a} e^{-t^2/2} \frac{dt}{\sqrt{2\pi}}.$$

That is, when n is large, the CDF of $\frac{X_1+\cdots+X_n-(1/2)n}{\sqrt{n}\sqrt{1/4}}$ is roughly the same as that of a standard normal. In particular, if you flip n fair coins, then the number of heads you get should typically be in the interval $(n/2 - \sqrt{n}/2, n/2 + \sqrt{n}/2)$, when n is large.

Remark 1.2. The random variable $\frac{X_1+\cdots+X_n-(1/2)n}{\sqrt{n}\sqrt{1/4}}$ has mean zero and variance 1, just like the standard Gaussian. So, the normalizations of $X_1 + \cdots + X_n$ we have chosen are sensible. Also, to explain the interval $(n/2 - \sqrt{n}/2, n/2 + \sqrt{n}/2)$, note that

$$\begin{aligned}
&\lim_{n\to\infty} \mathbf{P}\left(\frac{n}{2} - \frac{\sqrt{n}}{2} \leq X_1 + \cdots + X_n \leq \frac{n}{2} + \frac{\sqrt{n}}{2}\right) \\
&\quad= \lim_{n\to\infty} \mathbf{P}\left(-\frac{\sqrt{n}}{2} \leq X_1 + \cdots + X_n - \frac{n}{2} \leq \frac{\sqrt{n}}{2}\right) \\
&\quad= \lim_{n\to\infty} \mathbf{P}\left(-1 \leq \frac{X_1 + \cdots + X_n - \frac{n}{2}}{\sqrt{n}/2} \leq 1\right) = \int_{-1}^{1} e^{-t^2/2} \frac{dt}{\sqrt{2\pi}} \approx .6827.
\end{aligned}$$

Exercise 1.3. Using the De Moivre-Laplace Theorem, estimate the probability that 1000000 coin flips of fair coins will result in more than 501,000 heads. (Some of the following integrals may be relevant: $\int_{-\infty}^{0} e^{-t^2/2}dt/\sqrt{2\pi} = 1/2$, $\int_{-\infty}^{1} e^{-t^2/2}dt/\sqrt{2\pi} \approx .8413$, $\int_{-\infty}^{2} e^{-t^2/2}dt/\sqrt{2\pi} \approx .9772$, $\int_{-\infty}^{3} e^{-t^2/2}dt/\sqrt{2\pi} \approx .9987$.)

Casinos do these kinds of calculations to make sure they make money and that they do not go bankrupt. Financial institutions and insurance companies do similar calculations for similar reasons.

Exercise 1.4. Let X and Y be nonnegative random variables. Recall that we can define

$$\mathbf{E}X := \int_{0}^{\infty} \mathbf{P}(X > t)dt.$$

Assume that $X \leq Y$. Conclude that $\mathbf{E}X \leq \mathbf{E}Y$.

More generally, if X satisfies $\mathbf{E}\,|X| < \infty$, we define $\mathbf{E}X := \mathbf{E}\max(X,0) - \mathbf{E}\max(-X,0)$. If X, Y are any random variables with $X \leq Y$, $\mathbf{E}\,|X| < \infty$ and $\mathbf{E}\,|Y| < \infty$, show that $\mathbf{E}X \leq \mathbf{E}Y$.

2. Random Variables and Expectations

2.1. **Properties of Probability Laws.** Recall that a probability law $\mathbf{P}$ on a sample space Ω satisfies the following three axioms.

(i) For any $A \subseteq \Omega$, we have $\mathbf{P}(A) \geq 0$. (**Nonnegativity**)

(ii) For any $A, B \subseteq \Omega$ such that $A \cap B = \emptyset$, we have

$$\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B).$$

If $A_1, A_2, \ldots \subseteq \Omega$ and $A_i \cap A_j = \emptyset$ whenever i, j are positive integers with $i \neq j$, then

$$\mathbf{P}\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} \mathbf{P}(A_k). \qquad (\textbf{Additivity})$$

(iii) We have $\mathbf{P}(\Omega) = 1$. (**Normalization**)

In this course, we will make several limiting statements about probabilities. For this reason, the following property of probability laws will be quite useful.

Proposition 2.1 (**Continuity of a Probability Law**). *Let $\mathbf{P}$ be a probability law on a sample space Ω. Let $A_1, A_2, \ldots$ be sets in Ω which are increasing, so that $A_1 \subseteq A_2 \subseteq \cdots$. Then*

$$\lim_{n\to\infty} \mathbf{P}(A_n) = \mathbf{P}(\cup_{n=1}^{\infty} A_n).$$

In particular, the limit on the left exists.

Proof. First, recall that $A \smallsetminus B := A \cap B^c$ where $A, B \subseteq \Omega$. Now, let $B_1 := A_1$, let $B_2 := A_2 \smallsetminus A_1$, and for any $n \geq 1$, inductively define $B_n := A_n \smallsetminus A_{n-1}$. We claim that $B_1, B_2, \ldots$ are disjoint, and $\cup_{n=1}^{k} A_n = \cup_{n=1}^{k} B_n$ for any $1 \leq k \leq \infty$.

To see the first statement, let $i, j \geq 1$ with $i > j$. Since $i - 1 \geq j$, $A_j \subseteq A_{i-1}$, so $A_{i-1}^c \cap A_j = \emptyset$. So

$$B_i \cap B_j = (A_i \smallsetminus A_{i-1}) \cap (A_j \smallsetminus A_{j-1}) = A_i \cap A_{i-1}^c \cap A_j \cap A_{j-1}^c = \emptyset.$$

To see the second statement, let $x \in \cup_{n=1}^{k} A_n$. Let $m \geq 1$ such that $m = \min\{1 \leq n \leq k \colon x \in A_n\}$. If $m = 1$, then $x \in B_1 = A_1$. If $m > 1$, then $x \notin A_{m-1}$ so $x \in B_m = A_m \smallsetminus A_{m-1}$. So, in any case, $x \in \cup_{n=1}^{k} B_n$. For the reverse inclusion, let $x \in \cup_{n=1}^{k} B_n$. Then $x \in B_n$ for some $n \geq 1$. So $x \in A_n$ since $B_n \subseteq A_n$. So, $x \in \cup_{n=1}^{k} A_n$. The claim is proven.

Now, using our claim, we have by the second axiom for probability laws,

$$\begin{aligned}\mathbf{P}(\cup_{n=1}^{\infty} A_n) &= \mathbf{P}(\cup_{n=1}^{\infty} B_n) = \sum_{n=1}^{\infty} \mathbf{P}(B_n) = \lim_{k\to\infty} \sum_{n=1}^{k} \mathbf{P}(B_n) \\ &= \lim_{k\to\infty} \mathbf{P}(\cup_{n=1}^{k} B_n) = \lim_{k\to\infty} \mathbf{P}(\cup_{n=1}^{k} A_n) = \lim_{k\to\infty} \mathbf{P}(A_k).\end{aligned}$$

The last line used $A_k \supseteq A_{k-1} \supseteq \cdots \supseteq A_1$. □

A similar statement can be made for a decreasing sequence of sets.

Proposition 2.2 (**Continuity of a Probability Law**)**.** *Let* $\mathbf{P}$ *be a probability law on a sample space* Ω*. Let* $A_1, A_2, \ldots$ *be sets in* Ω *which are decreasing, so that* $A_1 \supseteq A_2 \supseteq \cdots$*. Then*

$$\lim_{n\to\infty} \mathbf{P}(A_n) = \mathbf{P}(\cap_{n=1}^{\infty} A_n).$$

Proof. Apply Proposition 2.1 to A_n^c for any $n \geq 1$, and then apply De Morgan's law:

$$\lim_{n\to\infty} \mathbf{P}(A_n) = 1 - \lim_{n\to\infty} \mathbf{P}(A_n^c) = 1 - \mathbf{P}(\cup_{n=1}^{\infty} A_n^c) = \mathbf{P}(\cap_{n=1}^{\infty} A_n).$$

□

Definition 2.3 (**Convergence of Real Numbers**)**.** Let $x_1, x_2, \ldots$ be a sequence of real numbers. Let $x \in \mathbb{R}$. We say that $x_1, x_2, \ldots$ **converges** to x if: $\forall\ \varepsilon > 0$, $\exists\ m = m(\varepsilon)$ such that, for all $n \geq m$, we have $|x_n - x| < \varepsilon$. If $x_1, x_2, \ldots$ converges to x, we denote this by writing

$$x = \lim_{n\to\infty} x_n.$$

Exercise 2.4. Using the definition of convergence, show that the sequence of numbers $1, 1/2, 1/3, 1/4, \ldots$ converges to 0.

Exercise 2.5 (**Uniqueness of limits**)**.** Let $x_1, x_2, \ldots$ be a sequence of real numbers. Let $x, y \in \mathbb{R}$. Assume that $x_1, x_2, \ldots$ converges to x. Assume also that $x_1, x_2, \ldots$ converges to y. Prove that $x = y$. That is, a sequence of real numbers cannot converge to two different real numbers.

2.2. **Derived Distributions.**

Proposition 2.6. *Let* X *be a continuous random variable with density function* $f_X \colon \mathbb{R} \to [0, \infty)$*. Let* $g \colon \mathbb{R} \to \mathbb{R}$ *be continuous. Let* $Y := g(X)$*. Assume that* F_Y *is differentiable, where* $F_Y(y) = \mathbf{P}(Y \leq y)$ *for all* $y \in \mathbb{R}$*. Then for any* $y \in \mathbb{R}$*,*

$$f_Y(y) = \frac{d}{dy} \int_{\{x \in \mathbb{R} \colon g(x) \leq y\}} f_X(x) dx.$$

Proof. Let $A \subseteq \mathbb{R}$. Recall that f_X is defined so that

$$\mathbf{P}(X \in A) = \int_A f_X(x) dx.$$

So, if we let $y \in \mathbb{R}$ and if we define $A := \{x \in \mathbb{R} \colon g(x) \leq y\}$, we have

$$F_Y(y) = \mathbf{P}(Y \leq y) = \mathbf{P}(g(X) \leq y) = \mathbf{P}(X \in A) = \int_A f_X(x) dx = \int_{\{x \in \mathbb{R} \colon g(x) \leq y\}} f_X(x) dx.$$

So, if F_Y is differentiable, $\frac{d}{dy} F_Y(y) = f_Y(y)$ for all $y \in \mathbb{R}$, completing the proof. □

Example 2.7. Let X be a uniformly distributed random variable on $[-1, 1]$, and let $g \colon \mathbb{R} \to \mathbb{R}$ so that $g(x) = x^3$ for any $x \in \mathbb{R}$. Let $Y := g(X)$. Then for any $y \in \mathbb{R}$,

$$f_Y(y) = \frac{d}{dy} \int_{\{x \in \mathbb{R} \colon g(x) \leq y\}} f_X(x) dx = \frac{d}{dy} \int_{\{x \in [-1,1] \colon x^3 \leq y\}} \frac{1}{2} dx.$$

If $y < -1$ the integral is zero. If $y > 1$, the integral is 1. And if $y \in [-1,1]$, we have

$$f_Y(y) = \frac{d}{dy}\frac{1}{2}\int_{x=-1}^{x=y^{1/3}} dx = \frac{1}{2}\frac{d}{dy}[y^{1/3}+1] = \frac{1}{6}y^{-2/3}.$$

And if $y \notin [-1,1]$, we have $f_Y(y) = 0$.

Exercise 2.8. Let X be a uniformly distributed random variable on $[-1,1]$. Let $Y := X^2$. Find f_Y.

Exercise 2.9. Let X be a uniformly distributed random variable on $[0,1]$. Let $Y := 4X(1-X)$. Find f_Y.

Example 2.10. Let X be a continuous random variable such that F_X is differentiable. Let $a, b \in \mathbb{R}$ with $a \neq 0$. Let $g(x) := ax + b$ for any $x \in \mathbb{R}$, and let $Y := g(X) = aX + b$. Then for any $y \in \mathbb{R}$, we will show that

$$f_Y(y) = \frac{1}{|a|}f_X\left(\frac{y-b}{a}\right).$$

Suppose $a > 0$. Then the function $\mathbf{P}(Y \leq y) = \mathbf{P}(aX + b \leq y) = \mathbf{P}(X \leq (y-b)/a) = F_X((y-b)/a)$ is differentiable with respect to y. So, for any $y \in \mathbb{R}$, the Chain Rule implies

$$f_Y(y) = \frac{d}{dy}\int_{\{x\in\mathbb{R}\colon\, g(x)\leq y\}} f_X(x)dx = \frac{d}{dy}F_X((y-b)/a) = \frac{1}{a}f_X((y-b)/a).$$

The case $a < 0$ is demonstrated similarly.

Example 2.11. Let X be a normal random variable with mean μ and variance $\sigma^2 > 0$ where $\sigma > 0$. That is,

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \qquad \forall\, x \in \mathbb{R}.$$

Let $a, b \in \mathbb{R}$ with $a > 0$. Let $Y := aX + b$. Then Y is a Gaussian random variable with variance $a^2\sigma^2$ and mean $b + a\mu$:

$$f_Y(y) = \frac{1}{a\sigma\sqrt{2\pi}}e^{-\frac{(((y-b)/a)-\mu)^2}{2a^2\sigma^2}} = \frac{1}{a\sigma\sqrt{2\pi}}e^{-\frac{(y-b-a\mu)^2}{2a^2\sigma^2}}$$

Definition 2.12 (**Monotonic Function**)**.** Let $I, J \subseteq \mathbb{R}$ be open intervals. Let $g\colon I \to J$. We say that g is **strictly increasing** if, for any $x, y \in I$ with $x > y$, we have $g(x) > g(y)$. We say that g is **strictly decreasing** if, for any $x, y \in I$ with $x > y$, we have $g(x) < g(y)$. We say that g is **strictly monotonic** if g is either strictly increasing or strictly decreasing.

Remark 2.13 (**Monotonic Functions are Invertible**)**.** Let $I, J \subseteq \mathbb{R}$ be open intervals. Let $g\colon I \to J$ be a monotonic function with range J. As we recall from calculus, g has an inverse. That is, there exists a monotonic function $h\colon J \to I$ such that $g(h(x)) = x$ for every $x \in J$ and $h(g(x)) = x$ for every $x \in I$. Also, as we recall from calculus, if g is differentiable with $g'(x) \neq 0$ for all $x \in I$, then h is differentiable, and by differentiating the identity $h(g(x)) = x$ and applying the chain rule, we get

$$\frac{d}{dx}h(g(x)) = \frac{1}{g'(x)}, \qquad \forall\, x \in I.$$

Or, written another way (defining $y := g(x)$, so that $x = h(y)$),

$$h'(y) = \frac{1}{g'(h(y))}, \qquad \forall\, y \in J.$$

If we graph g and h, then h is obtained by reflecting g across the line $\{(x, y) \in \mathbb{R}^2 \colon x = y\}$. Similarly, g is obtained by reflecting h across the line $\{(x, y) \in \mathbb{R}^2 \colon x = y\}$.

Proposition 2.14. *Let X be a continuous random variable such that F_X is differentiable. Let $I, J \subseteq \mathbb{R}$ be open intervals. Let $g \colon I \to J$ be a monotonic, differentiable function with range J. Assume that $g'(x) \neq 0$ for every $x \in I$. Let $Y := g(X)$. Let $h \colon J \to I$ be the inverse of g. Then for any $y \in J$,*

$$f_Y(y) = f_X(h(y)) \cdot \left|\frac{d}{dy}h(y)\right| = f_X(h(y)) \cdot \frac{1}{|g'(h(y))|}.$$

Proof. Let $y \in J$. First, assume g is strictly increasing. Then

$$F_Y(y) = \mathbf{P}(Y \leq y) = \mathbf{P}(g(X) \leq y) = \mathbf{P}(X \leq h(y)) = F_X(h(y)).$$

Since F_X and h are differentiable, the Chain Rule then proves the first equality. The second equality follows from Remark 2.13, where we noted that

$$\frac{d}{dy}h(y) = \frac{1}{g'(h(y))}, \qquad \forall y \in J.$$

□

Exercise 2.15. Let X be a uniformly distributed random variable on $[0, 1]$. Find the PDF of $-\log(X)$.

Exercise 2.16. Let X be a standard normal random variable. Find the PDF of e^X.

We can perform similar manipulations to find the joint PDF of functions of several random variables.

Example 2.17. Let X, Y be independent exponential random variables with parameter $\lambda = 1$. So, $f_X(x) = e^{-x}$ for any $x \geq 0$ and $f_X(x) = 0$ otherwise. Let $Z := \max(X, Y)$. Then for any $t \in \mathbb{R}$, $\{Z \leq t\} = \{X \leq t, Y \leq t\}$. So, using independence of X, Y,

$$\mathbf{P}(Z \leq t) = \mathbf{P}(X \leq t, Y \leq t) = \mathbf{P}(X \leq t)\mathbf{P}(Y \leq t) = (1 - e^{-t})^2, \qquad \forall\, t \geq 0.$$

So, using the chain rule,

$$f_Z(z) = \frac{d}{dz}\mathbf{P}(Z \leq z) = \begin{cases} 2(1 - e^{-z})e^{-z} & \text{, if } z \geq 0 \\ 0 & \text{, otherwise.} \end{cases}$$

Exercise 2.18. Let X, Y, Z be independent standard Gaussian random variables. Find the PDF of $\max(X, Y, Z)$.

Example 2.19. Let X, Y be independent standard Gaussian random variables. Let $Z := X/|Y|$. For any $t \in \mathbb{R}$, let $A_t := \{(x, y) \in \mathbb{R}^2 \colon x \leq t\,|y|\}$. Then, using polar coordinates, if

$t \geq 0$ we have

$$\begin{aligned}
\mathbf{P}(Z \leq t) = \mathbf{P}(X \leq t\,|Y|) = \mathbf{P}((X,Y) \in A_t) &= \frac{1}{2\pi}\int_{A_t} e^{-(x^2+y^2)/2}dxdy \\
&= \int_{y=-\infty}^{y=\infty}\int_{x=-\infty}^{x=t|y|} e^{-(x^2+y^2)/2}dxdy \\
&= \frac{1}{2\pi}\int_{\theta=\tan^{-1}(1/t)}^{\theta=2\pi-\tan^{-1}(1/t)}\int_{r=0}^{r=\infty} re^{-r^2/2}drd\theta \\
&= \frac{1}{2\pi}\int_{\theta=\tan^{-1}(1/t)}^{\theta=2\pi-\tan^{-1}(1/t)} d\theta = 1 - \frac{1}{\pi}\tan^{-1}(1/t).
\end{aligned}$$

Similarly, if $t < 0$, then $\mathbf{P}(Z \leq t) = \frac{1}{\pi}\tan^{-1}(1/|t|)$. So, from the Chain rule,

$$f_Z(z) = \frac{1}{\pi(z^2+1)}, \qquad \forall\, z \in \mathbb{R}.$$

Exercise 2.20. Let X be a random variable uniformly distributed in $[0,1]$ and let Y be a random variable uniformly distributed in $[0,2]$. Suppose X and Y are independent. Find the PDF of X/Y^2.

2.3. **Covariance.** Recall that the covariance of two random variables X and Y, denoted $\mathrm{cov}(X,Y)$, is

$$\mathrm{cov}(X,Y) = \mathbf{E}((X - \mathbf{E}(X))(Y - \mathbf{E}(Y))).$$

In particular, $\mathrm{cov}(X,X) = \mathbf{E}(X - \mathbf{E}(X))^2 = \mathrm{var}(X)$.

Definition 2.21. Let X, Y be random variables. We say that X, Y are **uncorrelated** if $\mathrm{cov}(X,Y) = 0$.

Exercise 2.22. Let X, Y be random variables with $\mathbf{E}X^2 < \infty$ and $\mathbf{E}Y^2 < \infty$. Prove the **Cauchy-Schwarz inequality**:

$$\mathbf{E}(XY) \leq (\mathbf{E}X^2)^{1/2}(\mathbf{E}Y^2)^{1/2}.$$

Then, deduce the following when X, Y both have finite variance:

$$|\mathrm{cov}(X,Y)| \leq (\mathrm{var}(X))^{1/2}(\mathrm{var}(Y))^{1/2}.$$

(Hint: in the case that $\mathbf{E}Y^2 > 0$, expand out the product $\mathbf{E}(X - Y\mathbf{E}(XY)/\mathbf{E}Y^2)^2$.)

Lemma 2.23. *Let $X_1, \ldots, X_n$ be random variables with $\mathrm{var}(X_i) < \infty$ for all $1 \leq i \leq n$. Then*

$$\mathrm{var}(\sum_{i=1}^n X_i) = \sum_{i=1}^n \mathrm{var}(X_i) + 2\sum_{1 \leq i < j \leq n} \mathrm{cov}(X_i, X_j).$$

Proof.

$$\begin{aligned}\mathrm{var}(\sum_{i=1}^{n}X_i) &= \mathbf{E}(\sum_{i=1}^{n}X_i-\mathbf{E}(\sum_{i=1}^{n}X_i))^2=\mathbf{E}(\sum_{i=1}^{n}(X_i-\mathbf{E}(X_i)))^2\\ &= \mathbf{E}\left(\sum_{i=1}^{n}(X_i-\mathbf{E}(X_i))^2\right)+2\mathbf{E}\left(\sum_{1\leq i<j\leq n}(X_i-\mathbf{E}(X_i))(X_j-\mathbf{E}(X_j))\right)\\ &= \sum_{i=1}^{n}\mathrm{var}(X_i)+2\sum_{1\leq i<j\leq n}\mathrm{cov}(X_i,X_j).\end{aligned}$$

The assumption $\mathrm{var}(X_i)<\infty$ for all $1\leq i\leq n$ and Exercise 2.22 ensure that all of the above quantities are finite. □

Lemma 2.23 immediately implies:

Corollary 2.24. *Let $X_1,\ldots,X_n$ be random variables which are pairwise uncorrelated. That is, $\mathrm{cov}(X_i,X_j)=0$ for any $i,j\in\{1,\ldots,n\}$ with $i\neq j$. Then*

$$\mathrm{var}(\sum_{i=1}^{n}X_i)=\sum_{i=1}^{n}\mathrm{var}(X_i).$$

Corollary 2.25. *Let $X_1,\ldots,X_n$ be independent random variables. Then*

$$\mathrm{var}(\sum_{i=1}^{n}X_i)=\sum_{i=1}^{n}\mathrm{var}(X_i).$$

Proof. Let $i,j\in\{1,\ldots,n\}$ with $i\neq j$. Then, using independence,

$$\mathrm{cov}(X_i,X_j)=\mathbf{E}((X_i-\mathbf{E}(X_i))(X_j-\mathbf{E}(X_j)))=\mathbf{E}(X_iX_j)-2\mathbf{E}(X_i)\mathbf{E}(X_j)+\mathbf{E}(X_i)\mathbf{E}(X_j)=0.$$

So, Corollary 2.24 concludes the proof. □

Exercise 2.26. Let X be a binomial random variable with parameters $n=2$ and $p=1/2$. So, $\mathbf{P}(X=0)=1/4$, $\mathbf{P}(X=1)=1/2$ and $\mathbf{P}(X=2)=1/4$. And X satisfies $\mathbf{E}X=1$ and $\mathbf{E}X^2=3/2$.

Let Y be a geometric random variable with parameter $1/2$. So, for any positive integer k, $\mathbf{P}(Y=k)=2^{-k}$. And Y satisfies $\mathbf{E}Y=2$ and $\mathbf{E}Y^2=6$.

Let Z be a Poisson random variable with parameter 1. So, for any nonnegative integer k, $\mathbf{P}(Z=k)=\frac{1}{e}\frac{1}{k!}$. And Z satisfies $\mathbf{E}Z=1$ and $\mathbf{E}Z^2=2$.

Let W be a discrete random variable such that $\mathbf{P}(W=0)=1/2$ and $\mathbf{P}(W=4)=1/2$, so that $\mathbf{E}W=2$ and $\mathbf{E}W^2=8$.

Assume that X,Y,Z and W are all independent. Compute

$$\mathrm{var}(X+Y+Z+W).$$

Exercise 2.27. Let $X_1,\ldots,X_n$ be random variables with finite variance. Define an $n\times n$ matrix A such that $A_{ij}=\mathrm{cov}(X_i,X_j)$ for any $1\leq i,j\leq n$. Show that the matrix A is positive semidefinite. That is, show that for any $y=(y_1,\ldots,y_n)\in\mathbb{R}^n$, we have

$$y^TAy=\sum_{i,j=1}^{n}y_iy_jA_{ij}\geq 0.$$

2.4. **Conditional Expectation as a Random Variable.** In your previous probability class, we defined conditional expectation to be a number. For example, if X, Y are discrete random variables, then for any fixed $y \in \mathbb{R}$ where $\mathbf{P}(Y = y) > 0$, we defined

$$\mathbf{E}(X|Y=y) = \sum_{x\in\mathbb{R}} x p_{X|Y=y}(x) = \sum_{x\in\mathbb{R}} x\mathbf{P}(X=x|Y=y) = \sum_{x\in\mathbb{R}} x\frac{\mathbf{P}(X=x, Y=y)}{\mathbf{P}(Y=y)}.$$

Or, if X, Y are continuous random variables, then for any fixed $y \in \mathbb{R}$ where $f_Y(y) > 0$, we defined

$$\mathbf{E}(X|Y=y) = \int_{-\infty}^{\infty} x f_{X|Y}(x|y)dx = \int_{-\infty}^{\infty} x\frac{f_{X,Y}(x,y)}{f_Y(y)}dx.$$

We can create a random variable from these definitions in the following natural way.

Definition 2.28 (**Conditional Expectation**)**.** Let X, Y be random variables. Let A be the range of Y. Define $g\colon A \to \mathbb{R}$ by $g(y) := \mathbf{E}(X|Y=y)$, for any $y \in A$. We then define the **conditional expectation** of X given Y, denoted $\mathbf{E}(X|Y)$, to be the random variable $g(Y)$.

Example 2.29. Let X, Y be random variables such that (X, Y) is uniformly distributed on the triangle $\{(x, y) \in \mathbb{R}^2 \colon x \geq 0, y \geq 0, x + y \leq 1\}$. If (x, y) is in this triangle, then $f_{X,Y}(x, y) = 2$, while $f_{X,Y}(x, y) = 0$ otherwise. If $y \in [0, 1]$,

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx = \int_{x=0}^{x=1-y} 2dx = 2(1-y).$$

Otherwise, $f_Y(y) = 0$. So, if $y \in [0, 1]$

$$\mathbf{E}(X|Y=y) = \int_{-\infty}^{\infty} x f_{X|Y}(x,y)dx = \int_{-\infty}^{\infty} x\frac{f_{X,Y}(x,y)}{f_Y(y)}dx = \int_{x=0}^{x=1-y} \frac{1}{1-y}xdx = \frac{1}{2}(1-y).$$

And $\mathbf{E}(X|Y=y)$ is undefined when $y \notin [0, 1]$, since $f_Y(y) = 0$ when $y \notin [0, 1]$.

Then, by definition of $\mathbf{E}(X|Y)$, we have

$$\mathbf{E}(X|Y) = \frac{1}{2}(1-Y).$$

(Strictly speaking, we have $\mathbf{E}(X|Y) = g(Y)$ where $g(y) = (1/2)(1 - y)$ for any $y \in [0, 1]$ and $g(y)$ is undefined for any $y \notin [0, 1]$. But since Y only takes values in $[0, 1]$, $g(Y) = \frac{1}{2}(1-Y)$.)

Exercise 2.30 (**Another Total Expectation Theorem**)**.** Using the definition of $\mathbf{E}(X|Y)$, prove the following theorem, which can be considered as a version of a Total Expectation Theorem:

$$\mathbf{E}(\mathbf{E}(X|Y)) = \mathbf{E}(X).$$

Exercise 2.31. If X is a random variable, and if $f(t) := \mathbf{E}(X-t)^2$, $t \in \mathbb{R}$, then the function $f\colon \mathbb{R} \to \mathbb{R}$ is uniquely minimized when $t = \mathbf{E}X$. This follows e.g. by writing

$$\begin{aligned}&\mathbf{E}(X-t)^2 = \mathbf{E}(X - \mathbf{E}(X) + \mathbf{E}(X) - t)^2\\ &\quad= \mathbf{E}(X-\mathbf{E}(X))^2 + (\mathbf{E}X - t)^2 + 2\mathbf{E}[(X-\mathbf{E}X)(\mathbf{E}X - t)] = \mathbf{E}(X-\mathbf{E}(X))^2 + (\mathbf{E}X-t)^2.\end{aligned}$$

So, the choice $t = \mathbf{E}X$ is the smallest, and it recovers the definition of variance, since $\mathrm{var}(X) = \mathbf{E}(X - \mathbf{E}X)^2$.

A similar minimizing property holds for conditional expectation. Let $h\colon \mathbb{R} \to \mathbb{R}$. Show that the quantity $\mathbf{E}(X - h(Y))^2$ is minimized among all functions h when $h(Y) = \mathbf{E}(X|Y)$. (Hint: Exercise 2.30 might be helpful.)

Lemma 2.32. *Let X, Y be random variables. Let $h\colon \mathbb{R} \to \mathbb{R}$. Then*

$$\mathbf{E}(Xh(Y)|Y) = h(Y)\mathbf{E}(X|Y).$$

Proof. Let $y \in R$. Then $\mathbf{E}(Xh(Y)|Y = y) = \mathbf{E}(Xh(y)|Y = y) = h(y)\mathbf{E}(X|Y = y)$. So, if $f(y) := \mathbf{E}(Xh(Y)|Y = y)$, we have $f(Y) = \mathbf{E}(Xh(Y)|Y) = h(Y)\mathbf{E}(X|Y)$ as desired. □

Definition 2.33 (Conditional Variance). Let X, Y be random variables. Let A be the range of Y and let $y \in A$. We then define the **conditional variance** of X given $Y = y$ to be

$$\mathrm{var}(X|Y = y) := \mathbf{E}\left[(X - \mathbf{E}(X|Y))^2 \,|Y = y\right].$$

Then, define $g\colon A \to \mathbb{R}$ so that $g(y) := \mathrm{var}(X|Y = y)$, for any $y \in A$, and define the **conditional variance** of X given Y, denoted $\mathrm{var}(X|Y)$, to be the random variable $g(Y)$.

Proposition 2.34. *Let X, Y be random variables. Then*

$$\mathrm{var}(X) = \mathbf{E}(\mathrm{var}(X|Y)) + \mathrm{var}(\mathbf{E}(X|Y)).$$

Proof. We square both sides of $X - \mathbf{E}(X) = (X - \mathbf{E}(X|Y)) + (\mathbf{E}(X|Y) - \mathbf{E}(X))$ to get

$$\begin{aligned}\mathrm{var}(X) &= \mathbf{E}(X - \mathbf{E}(X))^2 \\ &= \mathbf{E}(X - \mathbf{E}(X|Y))^2 + \mathbf{E}(\mathbf{E}(X|Y) - \mathbf{E}(X))^2 + 2\mathbf{E}\left[(X - \mathbf{E}(X|Y))(\mathbf{E}(X|Y) - \mathbf{E}(X))\right].\end{aligned}$$

By Exercise 2.30, the first term is $\mathbf{E}(X - \mathbf{E}(X|Y))^2 = \mathbf{E}(\mathbf{E}[(X - \mathbf{E}(X|Y))^2|Y]) = \mathbf{E}\mathrm{var}(X|Y)$. By Exercise 2.30, the second term is $\mathbf{E}[\mathbf{E}(X|Y) - \mathbf{E}(\mathbf{E}(X|Y))]^2 = \mathrm{var}(\mathbf{E}(X|Y))$. It remains to show that the last term is zero. Let $h(Y) := \mathbf{E}(X|Y) - \mathbf{E}(X)$. Using Lemma 2.32 and Exercise 2.30, the last term is

$$\begin{aligned}\mathbf{E}\left[(X - \mathbf{E}(X|Y))h(Y))\right] &= \mathbf{E}[Xh(Y)] - \mathbf{E}[\mathbf{E}(X|Y)h(Y)] \\ &= \mathbf{E}[Xh(Y)] - \mathbf{E}[\mathbf{E}(Xh(Y)|Y)] = \mathbf{E}[Xh(Y)] - \mathbf{E}(Xh(Y)] = 0.\end{aligned}$$

□

Exercise 2.35. Toys are stored in small boxes, small boxes are stored in large crates, and large crates comprise a shipment. Let X_i be the number of toys in small box $i \in \{1, 2, \ldots\}$. Assume that $X_1, X_2, \ldots$ all have the same CDF. Let Y_i be the number of small boxes in large crate $i \in \{1, 2, \ldots\}$. Assume that $Y_1, Y_2, \ldots$ all have the same CDF. Let Z be the number of large crates in the shipment. Assume that $X_1, X_2, \ldots, Y_1, Y_2, \ldots, Z$ are all independent, nonnegative integer-valued random variables, each with expected value 10 and variance 16.

Compute the expected value and variance of the number of toys in the shipment.

To demonstrate what to do in this exercise, we compute the number of toys in the first large crate.

Let T_1 be the number of toys in the first large crate. Let y be a nonnegative integer. Given that $Y_1 = y$, we know that $T_1 = X_1 + \cdots + X_y$. So, using independence,

$$\mathbf{E}(T_1|Y_1 = y) = \mathbf{E}(X_1 + \cdots + X_y|Y_1 = y) = y\mathbf{E}(X_1) = 10y.$$

So, by the definition of conditional expectation, $\mathbf{E}(T_1|Y_1) = 10Y_1$. Finally, by Exercise 2.30,

$$\mathbf{E}(T_1) = \mathbf{E}(\mathbf{E}(T_1|Y_1)) = \mathbf{E}(10Y_1) = 10\mathbf{E}Y_1 = (10)(10) = 100.$$

From the definition of conditional variance, and Corollary 2.25,

$$\begin{aligned}\mathrm{var}(T_1|Y_1 = y) &= \mathbf{E}\left((T_1 - 10y)^2|Y_1 = y\right) = \mathbf{E}\left((X_1 + \cdots + X_y - 10y)^2|Y_1 = y\right)\\ &= \mathrm{var}(X_1 + \cdots + X_y) = y\mathrm{Var}(X_1) = 16y.\end{aligned}$$

So, $\mathrm{var}(T_1|Y_1) = 16Y_1$, and by Proposition 2.34,

$$\mathrm{var}(T_1) = \mathbf{E}(16Y_1) + \mathrm{var}(10Y_1) = (16)(10) + (100)(16) = 160 + 1600 = 1760.$$

Exercise 2.36. Let $0 < p < 1$. Suppose you have a biased coin which has a probability p of landing heads, and probability $1-p$ of landing tails, each time it is flipped. Also, suppose you have a fair six-sided die (so each face of the cube has a distinct label from the set $\{1, 2, 3, 4, 5, 6\}$, and each time you roll the die, any face of the cube is rolled with equal probability.)

Let N be the number of coin flips you need to do until the first head appears. Now, roll the fair die N times. Let S be the sum of the results of the N rolls of the die. Compute $\mathbf{E}S$ and $\mathrm{var}(S)$.

Exercise 2.37. Let $f\colon \mathbb{R} \to \mathbb{R}$ be twice differentiable function. Assume that f is convex. That is, $f''(x) \geq 0$, or equivalently, the graph of f lies above any of its tangent lines. That is, for any $x, y \in \mathbb{R}$,

$$f(x) \geq f(y) + f'(y)(x - y).$$

(In Calculus class, you may have referred to these functions as "concave up.") Let X be a discrete random variable. By setting $y = \mathbf{E}(X)$, prove **Jensen's inequality**:

$$\mathbf{E}f(X) \geq f(\mathbf{E}(X)).$$

In particular, choosing $f(x) = x^2$, we have $\mathbf{E}(X^2) \geq (\mathbf{E}(X))^2$.

2.5. **Transforms.** Generally speaking, a transform is a way of creating one function from another function. For example, the moment generating function associates a real-valued function to a random variable. And the characteristic function (or Fourier transform) associates a complex-valued function to a random variable.

Definition 2.38 (**Moment Generating Function**)**.** Let X be a random variable. The **moment generating function** of X is a function $M_X\colon \mathbb{R} \to \mathbb{R}$ defined by

$$M_X(t) := \mathbf{E}(e^{tX}), \quad \forall\, t \in \mathbb{R}.$$

Remark 2.39. For certain random variables X, the moment generating function may not exist. For example, if X is a continuous random variable with density function $f_X(x) = x^{-2}$ for any $x > 1$, and $f_X(x) = 0$ otherwise. Then $M_X(t) = \int_1^\infty e^{tx} f_X(x)dx$ does not exist when $t > 0$.

Assume that $M_X(t)$ exists for all $t \in \mathbb{R}$, and assume we can differentiate under the expected value. Then

$$\frac{d}{dt}|_{t=0} M_X(t) = \mathbf{E}\left(\frac{d}{dt}_{t=0} e^{tX}\right) = \mathbf{E}(X).$$

That is, the first derivative of the moment generating function at $t = 0$ is equal to the first moment of X. More generally, the n^{th} derivative of the moment generating function at $t = 0$ is equal to the n^{th} moment of X:

Exercise 2.40. Let X be a random variable. Assume that $M_X(t)$ exists for all $t \in \mathbb{R}$, and assume we can differentiate under the expected value any number of times. For any positive integer n, show that

$$\frac{d^n}{dt^n}|_{t=0} M_X(t) = \mathbf{E}(X^n).$$

So, in principle, all moments of X can be computed just by taking derivatives of the moment generating function.

Example 2.41. Let X be an exponential random variable with parameter $\lambda > 0$. That is, $f_X(x) = \lambda e^{-\lambda x}$ for any $x \geq 0$, and $f_X(x) = 0$ otherwise. Then for any $t < \lambda$,

$$\begin{aligned} M_X(t) &= \lambda \int_0^\infty e^{tx} e^{-\lambda x} dx = \lambda \int_0^\infty e^{(t-\lambda)x} dx \\ &= \lambda \lim_{N\to\infty} \frac{1}{t-\lambda} [e^{(t-\lambda)x}]_{x=0}^{x=N} = \frac{\lambda}{\lambda - t}. \end{aligned}$$

From Exercise 2.40, $\mathbf{E}X = \frac{d}{dt}|_{t=0} M_X(t) = \frac{\lambda}{\lambda^2} = \lambda^{-1}$. More generally, it follows by induction that

$$\frac{d^n}{dt^n} M_X(t) = \lambda n! (\lambda - t)^{-n-1}, \qquad \forall\, n > 0. \qquad (*)$$

The case $n = 1$ is known. To complete the inductive step, note that

$$\frac{d^n}{dt^{n+1}} M_X(t) = \frac{d}{dt}\frac{d^n}{dt^n} M_X(t) = \frac{d}{dt} \lambda n! (\lambda - t)^{-n-1} = \lambda (n+1)! (\lambda - t)^{-(n+1)-1}.$$

So, from Exercise 2.40, for any $n > 0$,

$$\mathbf{E}X^n = \frac{d^n}{dt^n}|_{t=0} M_X(t) = n! \lambda^{-n}.$$

Exercise 2.42. Let X be a standard Gaussian random variable. Compute an explicit formula for the moment generating function of X. (Hint: completing the square might be helpful.) From this explicit formula, compute an explicit formula for all moments of the Gaussian random variable. (The $2n^{th}$ moment of X should be something resembling a factorial.)

Proposition 2.43. *Let $X_1, \ldots, X_n$ be independent random variables. Then*

$$M_{X_1+\cdots+X_n}(t) = \prod_{j=1}^n M_{X_j}(t), \qquad \forall\, t \in \mathbb{R}.$$

Proof. Since $X_1, \ldots, X_n$ are independent, $e^{tX_1}, \ldots, e^{tX_n}$ are independent, for any $t \in \mathbb{R}$. So,

$$M_{X_1+\cdots+X_n}(t) = \mathbf{E}e^{t(X_1+\cdots+X_n)} = \mathbf{E}\prod_{j=1}^n e^{tX_j} = \prod_{j=1}^n \mathbf{E}e^{tX_j} = \prod_{j=1}^n M_{X_j}(t)$$

□

Example 2.44. Let X be a binomial distributed random variable with parameters n and $0 < p < 1$. That is, X can be written as the sum of n independent Bernoulli random variables $X_1, \ldots, X_n$ with parameter p. Then by Proposition 2.43, for any $t \in \mathbb{R}$,

$$M_X(t) = \prod_{j=1}^{n} M_{X_j}(t) = (M_{X_1}(t))^n = ((1-p)e^{0\cdot t} + pe^t)^n = (1-p+pe^t)^n.$$

In some cases, the moment generating function uniquely determines the random variable.

Theorem 2.45 (**Lévy Continuity Theorem, Weak Form**). *Let X, Y be random variables. Assume that $M_X(t), M_Y(t)$ exist for all $t \in \mathbb{R}$, and that $M_X(t) = M_Y(t)$ for all $t \in \mathbb{R}$. Then X and Y have the same CDF.*

Exercise 2.46. Construct two random variables $X, Y\colon \Omega \to \mathbb{R}$ such that $X \neq Y$ but $M_X(t), M_Y(t)$ exist for all $t \in \mathbb{R}$, and such that $M_X(t) = M_Y(t)$ for all $t \in \mathbb{R}$.

Exercise 2.47. Unfortunately, there exist random variables X, Y such that $\mathbf{E}X^n = \mathbf{E}Y^n$ for all $n = 1, 2, 3, \ldots$, but such that X, Y do not have the same CDF. First, explain why this does not contradict the Lévy Continuity Theorem, Weak Form. Now, let $-1 < a < 1$, and define a density

$$f_a(x) := \begin{cases} \frac{1}{x\sqrt{2\pi}} e^{-\frac{(\log x)^2}{2}} (1 + a\sin(2\pi \log x)) & \text{, if } x > 0 \\ 0 & \text{, otherwise.} \end{cases}$$

Suppose X_a has density f_a. If $-1 < a, b < 1$, show that $\mathbf{E}X_a^n = \mathbf{E}X_b^n$ for all $n = 1, 2, 3, \ldots$. (Hint: write out the integrals, and make a change of variables $s = \log(x) - n$.)

From Exercise 2.40, the moment generating function of a random variable X contains all information about the moments of X. However, as mentioned in Remark 2.39, $M_X(t)$ may not exist for many values of t. So, studying the moment generating function may not be so helpful for certain random variables. Fortunately, the closely related characteristic function will always exist, and it also contains all information about the moments of X

Definition 2.48 (**Characteristic Function/ Fourier Transform**). Let $i := \sqrt{-1}$. Let X be a random variable. The **characteristic function** (or **Fourier transform**) of X is the function $\phi_X\colon \mathbb{R} \to \mathbb{C}$ defined by

$$\phi_X(t) := \mathbf{E}(e^{itX}), \quad \forall\, t \in \mathbb{R}.$$

Or equivalently,

$$\phi_X(t) = M_X(it), \quad \forall\, t \in \mathbb{R}.$$

Remark 2.49 (**Expectation of Complex-Valued Random Variables**). Any complex number $z \in \mathbb{C}$ can be written as $z = a + bi$ where $a, b \in \mathbb{R}$. We also define $|z| := \sqrt{a^2 + b^2}$. We call a the real part of z, and we call b the imaginary part of z. Similarly, if Z is a complex-valued random variable, we can write $Z = X + iY$ where X, Y are real-valued random variables. Then, we can define

$$\mathbf{E}Z := \mathbf{E}X + i(\mathbf{E}Y).$$

That is, taking the expected value of a complex-valued random variable is barely different from taking the expected value of a real-valued random variable.

Exercise 2.50. Compute the characteristic function of a uniformly distributed random variable on $[-1,1]$. (Some of the following formulas might help to simplify your answer: $e^{it} = \cos(t) + i\sin(t)$, $\cos(t) = [e^{it} + e^{-it}]/2$, $\sin(t) = [e^{it} - e^{-it}]/[2i]$, $t \in \mathbb{R}$.)

Remark 2.51. If $t \in \mathbb{R}$, then $|e^{it}| = |\cos(t) + i\sin(t)| = \sqrt{\cos^2(t) + \sin^2(t)} = 1$. The characteristic function is often more appealing to work with than the moment generating function, since the characteristic function always exists. For example, for any $t \in \mathbb{R}$,

$$|\phi_X(t)| = \left|\mathbf{E}e^{itX}\right| \leq \mathbf{E}\left|e^{itX}\right| = 1.$$

However, as mentioned in Remark 2.39, $M_X(t)$ may or may not exist for some $t \in \mathbb{R}$.

Exercise 2.52. Let X be a random variable. Assume we can differentiate under the expected value of $\mathbf{E}e^{itX}$ any number of times. For any positive integer n, show that

$$\frac{d^n}{dt^n}|_{t=0}\phi_X(t) = i^n\mathbf{E}(X^n).$$

So, in principle, all moments of X can be computed just by taking derivatives of the characteristic function.

Exercise 2.53. Let X be a random variable such that $\mathbf{E}|X|^3 < \infty$. Prove that for any $t \in \mathbb{R}$,

$$\mathbf{E}e^{itX} = 1 + it\mathbf{E}X - t^2\mathbf{E}X^2/2 + o(t^2).$$

That is,

$$\lim_{t\to 0} t^{-2}\left|\mathbf{E}e^{itX} - [1 + it\mathbf{E}X - t^2\mathbf{E}X^2/2]\right| = 0$$

(Hint: it may be helpful to use Jensen's inequality, Exercise 2.37, to first justify that $\mathbf{E}|X| < \infty$ and $\mathbf{E}X^2 < \infty$. Then, use the Taylor expansion with error bound: $e^{iy} = 1 + iy - y^2/2 - (i/2)\int_0^y (y-s)^2 e^{is} ds$, which is valid for any $y \in \mathbb{R}$.)

Actually, this same bound holds only assuming $\mathbf{E}X^2 < \infty$, but the proof of that bound requires things we have not discussed.

Since $\phi_X(t) = M_X(it)$, the proof of Proposition 2.34 immediately implies:

Proposition 2.54. *Let $X_1, \ldots, X_n$ be independent random variables. Then*

$$\phi_{X_1+\cdots+X_n}(t) = \prod_{j=1}^{n} \phi_{X_j}(t), \qquad \forall\, t \in \mathbb{R}.$$

The Gaussian density has the rather remarkable property that it is essentially its own Fourier transform.

Proposition 2.55. *Let X be a standard Gaussian random variable. Then*

$$\mathbf{E}e^{itX} = e^{-t^2/2}, \qquad \forall\, t \in \mathbb{R}.$$

Proof. Using $e^{itx} = \cos(tx) + i\sin(tx)$ for any $t, x \in \mathbb{R}$, with $t \neq 0$

$$\begin{aligned}
\phi_X(t) = \mathbf{E}e^{itX} &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{itx}e^{-x^2/2}dx \\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} (\cos(tx) + i\sin(tx))e^{-x^2/2}dx \\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \cos(tx)e^{-x^2/2}dx, \qquad \text{since } e^{-x^2/2}\sin(tx) \text{ is odd.}
\end{aligned}$$

Now, differentiating under the integral sign (which is valid, but we will not justify it), and integrating by parts,

$$\begin{aligned}
\frac{d}{dt}\phi_X(t) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} (-x)\sin(tx)e^{-x^2/2}dx = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \sin(tx)\frac{d}{dx}e^{-x^2/2}dx \\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} (-t)\cos(tx)e^{-x^2/2}dx = -t\phi_X(t).
\end{aligned}$$

Therefore,

$$\frac{d}{dt}[\phi_X(t)e^{t^2/2}] = [t\phi_X(t) - t\phi_X(t)]e^{t^2/2} = 0, \qquad \forall\, t \in \mathbb{R}.$$

That is, there exists a constant $c \in \mathbb{R}$ such that $\phi_X(t)e^{t^2/2} = c$, i.e. $\phi_X(t) = ce^{-t^2/2}$. Since $\phi_X(0) = 1 = c$, the proof is complete. □

2.6. **Sums of Independent Random Variables and Convolution.** Let X, Y be independent random variables. From Proposition 2.43, the moment generating function of $X+Y$ can be easily expressed as $M_{X+Y}(t) = M_X(t)M_Y(t)$, for any t such that both quantities on the right exist. On the other hand, the CDF of $X+Y$ has a more complicated dependence on X and Y.

Example 2.56. Let X, Y be independent integer-valued random variables. Then, repeatedly using properties of probability laws, and using that X, Y are independent,

$$\begin{aligned}
\mathbf{P}(X+Y=t) &= \sum_{j,k\in\mathbb{Z}\colon\, j+k=t} \mathbf{P}(X=j, Y=k) = \sum_{j\in\mathbb{Z}} \mathbf{P}(X=j, Y=t-j) \\
&= \sum_{j\in\mathbb{Z}} \mathbf{P}(X=j)\mathbf{P}(Y=t-j) = \sum_{j\in\mathbb{Z}} p_X(j)p_Y(t-j).
\end{aligned}$$

Definition 2.57 (Convolution on the integers). Let $g, h\colon \mathbb{Z} \to \mathbb{R}$ be functions. The **convolution** of g and h, denoted $g * h$, is a function $g * h\colon \mathbb{Z} \to \mathbb{R}$ defined by

$$(g*h)(t) := \sum_{j\in\mathbb{Z}} g(j)h(t-j), \qquad \forall\, t \in \mathbb{Z}.$$

Example 2.58. Let $g(k) := e^{-k}$ and let $h(k) := e^{-k}$ for any nonnegative integer $k \geq 0$, and let $g(k) = h(k) = 0$ for any other integer $k < 0$. Then if $t \geq 0$ is an integer,

$$(g*h)(t) = \sum_{k\in\mathbb{Z}} g(k)h(t-k) = \sum_{k=0}^{t} e^{-k}e^{-(t-k)} = \sum_{k=0}^{t} e^{-t} = (t+1)e^{-t}.$$

And $(g*h)(t) = 0$ for any negative integer t.

A similar formula holds for continuous random variables. That is, if X, Y are two continuous random variables, then the density of $X+Y$ is the convolution of f_X and f_Y.

Definition 2.59 (Convolution on the real line). Let $g, h\colon \mathbb{R} \to \mathbb{R}$ be functions. The **convolution** of g and h, denoted $g * h$, is a function $g * h\colon \mathbb{R} \to \mathbb{R}$ defined by

$$(g * h)(t) := \int_{-\infty}^{\infty} g(x)h(t-x)dx, \qquad \forall\, t \in \mathbb{R}.$$

Proposition 2.60. *Let X, Y be two continuous independent random variables such that $\mathbf{P}(X+Y \leq t)$ is differentiable with respect to $t \in \mathbb{R}$. Then*

$$f_{X+Y}(t) = (f_X * f_Y)(t), \qquad \forall\, t \in \mathbb{R}.$$

Proof. Let X, Y be independent continuous random variables. Then, changing variables,

$$\mathbf{P}(X+Y \leq t) = \int_{\{(x,y)\in\mathbb{R}^2 \colon x+y\leq t\}} f_{X,Y}(x,y)dxdy = \int_{x=-\infty}^{x=\infty}\int_{y=-\infty}^{y=t-x} f_X(x)f_Y(y)dydx.$$

Then, since $\mathbf{P}(X+Y \leq t)$ is differentiable with respect to t, we have by the Fundamental Theorem of Calculus,

$$f_{X+Y}(t) = \frac{d}{dt}\mathbf{P}(X+Y \leq t) = \int_{x=-\infty}^{x=\infty} f_X(x)\frac{d}{dt}\int_{y=-\infty}^{y=t-x} f_Y(y)dydx = \int_{x=-\infty}^{x=\infty} f_X(x)f_Y(t-x)dx.$$

□

Example 2.61. Let $g(x) = h(x) := \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$ for any $x \in \mathbb{R}$. Then if $t \in \mathbb{R}$, we complete the square and change variables twice to get

$$\begin{aligned}(g * h)(t) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-x^2/2}e^{-(t-x)^2/2}dx = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-x^2+xt-t^2/2}dx\\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-(x-t/2)^2+t^2/4-t^2/2}dx = e^{-t^2/4}\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-(x-t/2)^2}dx\\ &= e^{-t^2/4}\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-x^2}dx = e^{-t^2/4}\frac{1}{2\sqrt{\pi}}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-x^2/2}dx = e^{-t^2/4}\frac{1}{2\sqrt{\pi}}.\end{aligned}$$

And $(g * h)(t) = e^{-t^2/4}\frac{1}{2\sqrt{\pi}}$ for any $t \in \mathbb{R}$.

Alternatively, we know that if X, Y are independent standard Gaussian random variables, then $X+Y$ is a Gaussian random variable with mean zero and variance $\sigma^2 = 2$. That is, $X+Y$ has density $e^{-t^2/4}\frac{1}{2\sqrt{\pi}}$, $t \in \mathbb{R}$.

Exercise 2.62 (Convolution is Associative). Let $g, h, d\colon \mathbb{R} \to \mathbb{R}$. Then for any $t \in \mathbb{R}$,

$$((g * h) * d)(t) = (g * (h * d))(t)$$

Exercise 2.63. Let X, Y, Z be independent and uniformly distributed on $[0,1]$. Note that f_X is not a continuous function.

Using convolution, compute f_{X+Y}. Draw f_{X+Y}. Note that f_{X+Y} is a continuous function, but it is not differentiable at some points.

Using convolution, compute f_{X+Y+Z}. Draw f_{X+Y+Z}. Note that f_{X+Y+Z} is a differentiable function, but it does not have a second derivative at some points.

Make a conjecture about how many derivatives $f_{X_1+\cdots+X_n}$ has, where $X_1,\ldots,X_n$ are independent and uniformly distributed on $[0,1]$. You do not have to prove this conjecture. The idea of this exercise is that convolution is a kind of average of functions. And the more averaging you do, the more derivatives $f_{X_1+\cdots+X_n}$ has.

Exercise 2.64. Construct two random variables X,Y such that X and Y are each uniformly distributed on $[0,1]$, and such that $\mathbf{P}(X+Y=1)=1$.

Then construct two random variables W,Z such that W and Z are each uniformly distributed on $[0,1]$, and such that $W+Z$ is uniformly distributed on $[0,2]$.

(Hint: there is a way to do each of the above problems with about one line of work. That is, there is a way to solve each problem without working very hard.)

2.7. **Sums of a Random Number of Independent Random Variables.** We give a general solution for problems similar to Exercise 2.36

Proposition 2.65. *Let $\mu\in\mathbb{R}$ and let $\sigma>0$. Let $X_1,X_2,\ldots$ be random variables each with mean μ and variance σ^2. Let N be a random variable taking nonnegative integer values. Assume that $N,X_1,X_2,\ldots$ are all independent. Let $S:=\sum_{i=1}^N X_i$. Then*

- $\mathbf{E}S=\mu\mathbf{E}N$. *(**Wald's Equation**)*
- $\mathrm{var}(S)=\sigma^2\mathbf{E}N+\mu^2\mathrm{var}(N)$.
- *If additionally $X_1,X_2,\ldots$ all have the same CDF, then*

$$\mathbf{E}e^{tS}=\sum_{n=0}^{\infty}(M_{X_1}(t))^n\mathbf{P}(N=n),\qquad\forall\,t\in\mathbb{R}.$$

Proof. Let $n\geq 0$ be an integer. Conditioned on $N=n$, we know that $S=X_1+\cdots+X_n$. So, $\mathbf{E}(S|N=n)=\mathbf{E}(X_1+\cdots+X_n)=n\mu$. So, $\mathbf{E}(S|N)=N\mu$, and by Exercise 2.30,

$$\mathbf{E}S=\mathbf{E}(\mathbf{E}(S|N))=\mathbf{E}(N\mu)=\mu\mathbf{E}N.$$

The definition of conditional variance, independence of S and N, and Corollary 2.25 give

$$\mathrm{var}(S|N=n)=\mathbf{E}[(S-\mathbf{E}(S|N))^2|N=n]=\mathbf{E}[(S-N\mu)^2|N=n]=\mathbf{E}[(S-n\mu)^2|N=n]$$
$$=\mathbf{E}(X_1+\cdots+X_n-n\mu)^2=\mathrm{var}(X_1+\cdots+X_n)=\sum_{i=1}^n\mathrm{var}(X_i)=n\sigma^2.$$

So, $\mathrm{var}(S|N)=N\sigma^2$. And by Proposition 2.34 and Exercise 2.30

$$\mathrm{var}(S)=\mathbf{E}(\mathrm{var}(S|N))+\mathrm{var}(\mathbf{E}(S|N))=\sigma^2\mathbf{E}N+\mathrm{var}(N\mu)=\sigma^2\mathbf{E}N+\mu^2\mathrm{var}(N).$$

We now prove the final assertion. Let $t\in\mathbb{R}$. Using independence and Proposition 2.43,

$$\mathbf{E}(e^{tS}|N=n)=\mathbf{E}(e^{t(X_1+\cdots+X_n)}|N=n)=\mathbf{E}(e^{t(X_1+\cdots+X_n)})$$
$$=\prod_{i=1}^n\mathbf{E}e^{tX_i}=(M_{X_1}(t))^n.$$

So, $\mathbf{E}(e^{tS}|N)=(M_{X_1}(t))^N$, and by Exercise 2.30,

$$\mathbf{E}e^{tS}=\mathbf{E}(\mathbf{E}(e^{tS}|N))=\mathbf{E}(M_{X_1}(t))^N=\sum_{n=0}^{\infty}(M_{X_1}(t))^n\mathbf{P}(N=n).$$

□

3. Limit Theorems

We now start to build up the machinery that is used to prove the two "big theorems" of probability: the Law of Large Numbers, and the Central Limit Theorem. We begin with some useful inequalities.

3.1. Markov and Chebyshev Inequalities.

Markov's inequality says that a random variable with finite expected value cannot be too large very often.

Proposition 3.1 (The Markov Inequality). *Let X be a nonnegative random variable. Then*

$$\mathbf{P}(X \geq t) \leq \frac{\mathbf{E}X}{t}, \qquad \forall\, t > 0.$$

Proof. Let $t > 0$. Let Y be a random variable such that

$$Y = \begin{cases} t & , \text{ if } X \geq t \\ 0 & , \text{ if } X < t. \end{cases}$$

By definition of Y, we have $Y \leq X$. Therefore, $\mathbf{E}Y \leq \mathbf{E}X$ by Exercise 1.4. By the definition of Y, $\mathbf{E}Y = t\mathbf{P}(X \geq t)$. That is,

$$t\mathbf{P}(X \geq t) \leq \mathbf{E}(X).$$

□

Remark 3.2. A nearly identical proof shows that $\mathbf{P}(X > t) \leq \frac{\mathbf{E}X}{t}$, for all $t > 0$.

Markov's inequality is commonly applied in the following ways.

Corollary 3.3. *Let X be a random variable. Then*

$$\mathbf{P}(|X| \geq t) \leq \frac{\mathbf{E}\,|X|}{t}, \qquad \forall\, t > 0.$$

More generally, if n is a positive integer, then

$$\mathbf{P}(|X| \geq t) \leq \frac{\mathbf{E}\,|X|^n}{t^n}, \qquad \forall\, t > 0.$$

Proof. The first assertion follows immediately by applying Proposition 3.1 to $|X|$. For the second assertion, we use the first assertion to get

$$\mathbf{P}(|X| \geq t) = \mathbf{P}(|X|^n \geq t^n) \leq \frac{\mathbf{E}\,|X|^n}{t^n}, \qquad \forall\, t > 0.$$

□

The second inequality of Corollary 3.3 is fairly useful , since if many moments of $|X|$ are bounded, then $\mathbf{P}(|X| \geq t)$ decays very rapidly.

Replacing X by $X - \mu$ and taking $n = 2$ in Corollary 3.3 gives:

Corollary 3.4 (Chebyshev Inequality). *Let X be a random variable with mean μ. Then*

$$\mathbf{P}(|X - \mu| \geq t) \leq \frac{\mathrm{var}(X)}{t^2}, \qquad \forall\, t > 0.$$

Or, replacing t by $t\sqrt{\mathrm{var}(X)}$,

$$\mathbf{P}(|X-\mu| \geq t\sqrt{\mathrm{var}(X)}) \leq \frac{1}{t^2}, \qquad \forall\, t>0.$$

Exercise 3.5. Let X be a standard Gaussian random variable. Let $t>0$ and let n be a positive even integer. Show that

$$\mathbf{P}(X>t) \leq \frac{(n-1)(n-3)\cdots(3)(1)}{t^n}.$$

That is, the function $t \mapsto \mathbf{P}(X>t)$ decays faster than any monomial.

Exercise 3.6. Let X be a random variable. Let $t>0$. Show that

$$\mathbf{P}(|X|>t) \leq \frac{\mathbf{E}X^4}{t^4}.$$

Exercise 3.7 (The Chernoff Bound). Let X be a random variable and let $r>0$. Show that, for any $t>0$,

$$\mathbf{P}(X>r) \leq e^{-tr}M_X(t).$$

Consequently, if $X_1,\ldots,X_n$ are independent random variables with the same CDF, and if $r,t>0$,

$$\mathbf{P}\left(\frac{1}{n}\sum_{i=1}^n X_i > r\right) \leq e^{-trn}(M_{X_1}(t))^n.$$

For example, if $X_1,\ldots,X_n$ are independent Bernoulli random variables with parameter $0<p<1$, and if $r,t>0$,

$$\mathbf{P}\left(\frac{X_1+\cdots+X_n}{n} - p > r\right) \leq e^{-trn}(e^{-tp}[pe^t+(1-p)])^n.$$

And if we choose t appropriately, then the quantity $\mathbf{P}\left(\frac{1}{n}\left|\sum_{i=1}^n (X_i-p)\right| > r\right)$ becomes exponentially small as either n or r become large. That is, $\frac{1}{n}\sum_{i=1}^n X_i$ becomes very close to its mean. Importantly, the Chernoff bound is much stronger than either Markov's or Cheyshev's inequality, since they only respectively imply that

$$\mathbf{P}\left(\left|\frac{X_1+\cdots+X_n}{n} - p\right| > r\right) \leq \frac{2p(1-p)}{r}, \quad \mathbf{P}\left(\left|\frac{X_1+\cdots+X_n}{n} - p\right| > r\right) \leq \frac{p(1-p)}{nr^2}.$$

Proposition 3.8 (Borel-Cantelli Lemma). *Let $A_1,A_2,\ldots$ be events with $\sum_{n=1}^\infty \mathbf{P}(A_n) < \infty$. Let B be the event that only finitely many of the events $A_1,A_2,\ldots$ occur. Then $\mathbf{P}(B)=1$.*

Proof. For any $n\geq 1$, let 1_{A_n} be a random variable which is 1 if A_n occurs, and 0 otherwise. That is, $1_{A_n}(\omega)=1$ if $\omega\in A_n$, and $1_{A_n}(\omega)=0$ if $\omega\notin A_n$. Then $\mathbf{E}(\sum_{n=1}^\infty 1_{A_n}) = \sum_{n=1}^\infty \mathbf{P}(A_n) < \infty$. So, by Markov's inequality,

$$\mathbf{P}\left(\sum_{n=1}^\infty 1_{A_n} \geq t\right) \leq \frac{\sum_{n=1}^\infty \mathbf{P}(A_n)}{t}, \qquad \forall\, t>0.$$

Letting $t\to\infty$ and using Continuity of the Probability Law, Proposition 2.2,

$$\mathbf{P}(B) = \mathbf{P}\left(\sum_{n=1}^\infty 1_{A_n} = \infty\right) = \lim_{t\to\infty}\mathbf{P}\left(\sum_{n=1}^\infty 1_{A_n} \geq t\right) = 0.$$

□

3.2. **Weak Law of Large Numbers.**

Definition 3.9. Let $X_1, X_2, \ldots$ be random variables. We say that $X_1, X_2, \ldots$ are **identically distributed** if $X_1, X_2, \ldots$ all have the same CDF. That is, $\mathbf{P}(X_i \leq t) = \mathbf{P}(X_j \leq t)$ for all $t \in \mathbb{R}$ and for all positive integers i, j.

Remark 3.10. If $X_1, X_2, \ldots$ are identically distributed random variables, then $\mathbf{E}X_i = \mathbf{E}X_j$ for all positive integers i, j.

We know intuitively that, if the results of independent experiments are averaged, then the average will become close to the expected value of a single experiment. Indeed, one way to intuitively think about expected value is as the average of many repeated experiments. The Law of Large Numbers makes the previous statement rigorous. For now, we only prove a weak version of this statement, though a stronger version will be proven later.

Theorem 3.11 (**Weak Law of Large Numbers**)**.** *Let $X_1, X_2, \ldots$ be independent identically distributed random variables. Assume that $\mu \in \mathbb{R}$ and $\mathbf{E}X_1 = \mu$. Then, for any $\varepsilon > 0$,*

$$\lim_{n\to\infty} \mathbf{P}\left(\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| \geq \varepsilon\right) = 0.$$

Proof. We make the additional assumption that $\mathrm{var}(X_1) < \infty$. Removing this assumption relies on things outside of this class. From Corollary 2.24,

$$\mathrm{var}\left(\frac{X_1 + \cdots + X_n}{n}\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{var}(X_i) = \frac{1}{n}\mathrm{var}(X_1).$$

So, Chebyshev's inequality implies that

$$\mathbf{P}\left(\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| \geq \varepsilon\right) \leq \frac{1}{n}\varepsilon^{-2}\mathrm{var}(X_1).$$

Letting $n \to \infty$ concludes the proof. □

Remark 3.12. We saw in Exercise 3.7 that the Chernoff bound implies the Weak Law of Large Numbers. However, the Chernoff bound requires the moment generating function to exist and be close to 1 for small $t > 0$, which is a much stronger assumption than what we assumed in Theorem 3.11.

Example 3.13. Let $X_1, X_2, \ldots$ be independent Bernoulli random variables with parameter $1/2$. Let $n := 10^4$, $\varepsilon := 10^{-2}$. Then

$$\mathbf{P}\left(\left|\frac{X_1 + \cdots + X_n}{n} - \frac{1}{2}\right| \geq \frac{1}{100}\right) \leq 10^{-4}10^4(1/4) = \frac{1}{4}.$$

3.3. **Convergence in Probability.**

Definition 3.14. We say that a sequence of random variables $Y_1, Y_2, \ldots$ **converges in probability** to a random variable Y if: for all $\varepsilon > 0$

$$\lim_{n\to\infty} \mathbf{P}(|Y_n - Y| > \varepsilon) = 0.$$

More formally, if Ω is the sample space, then $\forall\ \varepsilon > 0$, $\lim_{n\to\infty} \mathbf{P}(\omega \in \Omega\colon\ |Y_n(\omega) - Y(\omega)| > \varepsilon) = 0$.

Remark 3.15. So, the Weak Law of Large numbers says: if X_1, X_2 are independent identically distributed random variables with $\mu := \mathbf{E}X_1 \in \mathbb{R}$, then the random variables $\frac{X_1+\cdots+X_n}{n}$ converge in probability to the constant μ.

Example 3.16. For any $n \geq 1$, let Y_n be a random variable such that $\mathbf{P}(Y_n = n^2) = 1/n$, and $\mathbf{P}(Y_n = 0) = 1 - 1/n$. Then $Y_1, Y_2, \ldots$ converges in probability to 0. For any $\varepsilon > 0$,

$$\mathbf{P}(|Y_n - 0| > \varepsilon) = \mathbf{P}(|Y_n| > \varepsilon) = \mathbf{P}(Y_n = n^2) = 1/n.$$

Therefore, $\lim_{n\to\infty} \mathbf{P}(|Y_n - 0| > \varepsilon) = 0$.

However, note that convergence in probability does not imply convergence in expected value, since $\lim_{n\to\infty} \mathbf{E}Y_n = \lim_{n\to\infty} n = \infty$, whereas the expected value of 0 is just 0.

Proposition 3.17 (Uniqueness of the Limit). *Suppose $Y_1, Y_2, \ldots$ converges in probability to Y. Also, suppose $Y_1, Y_2, \ldots$ converges in probability to Z. Then $\mathbf{P}(Z \neq Y) = 0$.*

Proof. From the triangle inequality, for any $n \geq 1$,

$$|Z - Y| = |Z - Y_n + Y_n - Y| \leq |Z - Y_n| + |Y - Y_n|.$$

So, for any $\varepsilon > 0$, if $|Z - Y| \geq \varepsilon$, then either $|Z - Y_n| \geq \varepsilon/2$ or $|Y - Y_n| \geq \varepsilon/2$. That is, for any $\varepsilon > 0$ and for any $n \geq 1$,

$$\{\omega \in \Omega\colon\ |Z(\omega) - Y(\omega)| \geq \varepsilon\}$$
$$\subseteq \{\omega \in \Omega\colon\ |Z(\omega) - Y_n(\omega)| \geq \varepsilon/2\} \cup \{\omega \in \Omega\colon\ |Y(\omega) - Y_n(\omega)| \geq \varepsilon/2\}.$$

Therefore, for any $\varepsilon > 0$ and for any $n \geq 1$,

$$\mathbf{P}(|Z - Y| \geq \varepsilon) \leq \mathbf{P}(|Z - Y_n| \geq \varepsilon/2) + \mathbf{P}(|Y - Y_n| \geq \varepsilon/2).$$

The left side does not depend on n. So, letting $n \to \infty$, we get $\mathbf{P}(|Z - Y| \geq \varepsilon) = 0$, for all $\varepsilon > 0$. Now,

$$\{Z \neq Y\} \subseteq \cup_{t=1}^{\infty}\{|Z - Y| \geq 1/t\}.$$

Therefore, $\mathbf{P}(Z \neq Y) \leq \sum_{t=1}^{\infty} \mathbf{P}(|Z - Y| \geq 1/t) = 0$. So, $\mathbf{P}(Z \neq Y) = 0$. □

Exercise 3.18. Let $X_1, X_2, \ldots$ be independent random variables, each with exponential distribution with parameter $\lambda = 1$. For any $n \geq 1$, let $Y_n := \max(X_1, \ldots, X_n)$. Let $0 < a < 1 < b$. Show that $\mathbf{P}(Y_n \leq a \log n) \to 0$ as $n \to \infty$, and $\mathbf{P}(Y_n \leq b \log n) \to 1$ as $n \to \infty$. Conclude that $Y_n / \log n$ converges to 1 in probability as $n \to \infty$.

Exercise 3.19. We say that random variables $X_1, X_2, \ldots$ converge to a random variable X in L_2 if

$$\lim_{n\to\infty} \mathbf{E}\,|X_n - X|^2 = 0.$$

Show that, if $X_1, X_2, \ldots$ converge to X in L_2, then $X_1, X_2, \ldots$ converges to X in probability.

Is the converse true? Prove your assertion.

Exercise 3.20. Let $X_1, X_2, \ldots$ be independent, identically distributed random variables such that $\mathbf{E}\,|X_1| < \infty$ and $\mathrm{var}(X_1) < \infty$. For any $n \geq 1$, define

$$Y_n := \frac{1}{n}\sum_{i=1}^{n} X_i^2.$$

Show that $Y_1, Y_2, \ldots$ converges in probability. Express the limit in terms of $\mathbf{E}X_1$ and $\mathrm{var}(X_1)$.

3.4. **Central Limit Theorem.** The following is a stronger version of Theorem 2.45.

Theorem 3.21 (**Lévy Continuity Theorem**)**.** *Let $X_1, X_2, \ldots$ be random variables and let X be a random variable. For any fixed $t \in \mathbb{R}$, assume that $\lim_{n\to\infty} \phi_{X_n}(t) = \phi_X(t)$. Assume also that $\phi_X(t)$ is continuous at $t = 0$. Then for any fixed $t \in \mathbb{R}$ such that $\mathbf{P}(X \leq t)$ is continuous, we have $\lim_{n\to\infty} \mathbf{P}(X_n \leq t) = \mathbf{P}(X \leq t)$.*

In particular, if X, Y are random variables with $\phi_X(t) = \phi_Y(t)$ for all $t \in \mathbb{R}$, and if $\phi_X(t)$ is continuous at $t = 0$, then X, Y are identically distributed.

We are finally able to prove the generalization of the De Moivre Laplace Theorem, Theorem 1.1, to arbitrary random variables.

Theorem 3.22 (**Central Limit Theorem**)**.** *Let $X_1, X_2, \ldots$ be independent, identically distributed random variables. Let Z be a standard Gaussian random variable. Let $\mu, \sigma \in \mathbb{R}$ with $\sigma > 0$. Assume that $\mathbf{E}X_1 = \mu$ and $\mathrm{var}(X_1) = \sigma^2$. Then for any $t \in \mathbb{R}$,*

$$\lim_{n\to\infty} \mathbf{P}\left(\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \leq t\right) = \int_{-\infty}^{t} e^{-x^2/2} \frac{dx}{\sqrt{2\pi}} = \mathbf{P}(Z \leq t).$$

Remark 3.23. The random variable $\frac{X_1+\cdots+X_n-\mu n}{\sigma\sqrt{n}}$ has mean zero and variance 1, just like the standard Gaussian X. So, the normalizations of $X_1 + \cdots + X_n$ we have chosen are sensible.

Remark 3.24. Let $f, g \colon \mathbb{R} \to \mathbb{R}$. Below we use the notation $f(t) = o(g(t))$ $\forall$ $t \in \mathbb{R}$ to denote $\lim_{t\to 0} \left|\frac{f(t)}{g(t)}\right| = 0$. For example, if $f(t) = o(t^2)$, then $\lim_{t\to 0} \left|\frac{f(t)}{t^2}\right| = 0$. Below we will use that $o(t^2 + o(t^2)) = o(t^2)$, and that $\lim_{n\to\infty} n \cdot o(1/n) = 0$.

Proof. For every $j \geq 1$, let $Y_j := (X_j - \mu)/\sigma$. Note that $Y_1, Y_2, \ldots$ are independent and identically distributed, $\mathbf{E}Y_j = 0$ and $\mathbf{E}Y_j^2 = 1$ for all $j \geq 1$. From Theorem 3.21 and Proposition 2.55, it suffices to show that, for any $t \in \mathbb{R}$,

$$\lim_{n\to\infty} \mathbf{E}e^{it\frac{Y_1+\cdots+Y_n}{\sqrt{n}}} = \mathbf{E}e^{itZ} = e^{-t^2/2}.$$

From Proposition 2.54,

$$\mathbf{E}e^{it\frac{Y_1+\cdots+Y_n}{\sqrt{n}}} = \prod_{j=1}^{n} \mathbf{E}e^{itY_j/\sqrt{n}} = (\mathbf{E}e^{itY_1/\sqrt{n}})^n.$$

We make the additional assumption that $\mathbf{E}\left|X_1\right|^3 < \infty$, so that $\mathbf{E}\left|Y_1\right|^3 < \infty$ and we can apply Exercise 2.53. (As remarked in Exercise 2.53, this assumption is not needed for the conclusion of Exercise 2.53 to hold.) By Exercise 2.53, and using $\mathbf{E}Y_1 = 0$ and $\mathbf{E}Y_1^2 = 1$,

$$\mathbf{E}e^{itY_1/\sqrt{n}} = 1 + \frac{it}{\sqrt{n}}\mathbf{E}Y_1 - \frac{t^2}{2n}\mathbf{E}Y_1^2 + o(t^2/n) = 1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right).$$

Therefore,

$$\mathbf{E}e^{it\frac{Y_1+\cdots+Y_n}{\sqrt{n}}} = \left(1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right)^n.$$

Taking logarithms, using $\log(1 + x) = x + o(x)$ for $-1 < x < 1$, and using Remark 3.24,

$$\log \mathbf{E}e^{it\frac{Y_1+\cdots+Y_n}{\sqrt{n}}} = n \log\left(1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right) = -\frac{t^2}{2} + n \cdot o\left(\frac{t^2}{n}\right).$$

Letting $n \to \infty$ and using Remark 3.24 completes the proof. □

Definition 3.25 (Convergence in Distribution). Let $X, X_1, X_2, \dots$ be random variables. We say that $X_1, X_2, \dots$ **converge in distribution** to X if, for any $t \in \mathbb{R}$ such that the CDF of X is continuous at t,

$$\lim_{n\to\infty} \mathbf{P}(X_n \leq t) = \mathbf{P}(X \leq t).$$

So, the Central Limit Theorem, Theorem 3.22, says: if $X_1, X_2, \dots$ are independent, identically distributed random variables with $\mu := \mathbf{E}X_1$ and $\sigma^2 := \mathrm{Var}(X_1)$ with $\sigma > 0$, then the random variables $\frac{X_1+\cdots+X_n-n\mu}{\sigma\sqrt{n}}$ converge in distribution to the standard Gaussian random variable. This fact is rather remarkable, since it holds no matter what distribution X_1 has! In this sense, the Gaussian random variable is "universal."

Exercise 3.26. This exercise demonstrates that geometry in high dimensions is different than geometry in low dimensions.

Let $x = (x_1, \dots, x_n) \in \mathbb{R}^n$. Let $\|x\| := \sqrt{x_1^2 + \cdots + x_n^2}$. Let $\varepsilon > 0$. Show that for all sufficiently large n, "most" of the cube $[-1,1]^n$ is contained in the annulus

$$A := \{x \in \mathbb{R}^n \colon (1-\varepsilon)\sqrt{n/3} \leq \|x\| \leq (1+\varepsilon)\sqrt{n/3}\}.$$

That is, if $X_1, \dots, X_n$ are each independent and identically distributed in $[-1,1]$, then for n sufficiently large

$$\mathbf{P}((X_1, \dots, X_n) \in A) \geq 1 - \varepsilon.$$

(Hint: apply the weak law of large numbers to $X_1^2, \dots, X_n^2$.)

Exercise 3.27 (Confidence Intervals). Among 625 members of a bank chosen uniformly at random among all bank members, it was found that 25 had a savings account. Give an interval of the form $[a, b]$ where $0 \leq a, b \leq 625$ are integers, such that with about 95% certainty, the number of any set of 625 bank members with savings accounts chosen uniformly at random lies in the interval $[1, 5]$. (Hint: if Y is a standard Gaussian random variable, then $\mathbf{P}(-2 \leq Y \leq 2) \approx .95$.)

Exercise 3.28 (Hypothesis Testing). Suppose we run a casino, and we want to test whether or not a particular roulette wheel is biased. Let p be the probability that red results from one spin of the roulette wheel. Using statistical terminology, "$p = 18/38$" is the null hypothesis, and "$p \neq 18/38$" is the alternative hypothesis. (On a standard roulette wheel, 18 of the 38 spaces are red.) For any $i \geq 1$, let $X_i = 1$ if the i^{th} spin is red, and let $X_i = 0$ otherwise.

Let $\mu := \mathbf{E}X_1$ and let $\sigma := \sqrt{\mathrm{var}(X_1)}$. If the null hypothesis is true, and if Y is a standard Gaussian random variable

$$\lim_{n\to\infty} \mathbf{P}\left(\left|\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}\right| \geq 2\right) = \mathbf{P}(|Y| \geq 2) \approx .05.$$

To test the null hypothesis, we spin the wheel n times. In our test, we reject the null hypothesis if $|X_1 + \cdots + X_n - n\mu| > 2\sigma\sqrt{n}$. Rejecting the null hypothesis when it is true is called a type I error. In this test, we set the type I error percentage to be 5%. (The type I error percentage is closely related to the p-value.)

Suppose we spin the wheel $n = 3800$ times and we get red 1868 times. Is the wheel biased? That is, can we reject the null hypothesis with around 95% certainty?

3.5. **Strong Law of Large Numbers.**

Exercise 3.29. Suppose random variables $X_1, X_2, \ldots$ converge in probability to a random variable X. Prove that $X_1, X_2, \ldots$ converge in distribution to X.

Then, show that the converse is false.

By Exercise 3.29, we see that the convergence guaranteed by the Central Limit Theorem is weaker than convergence in probability. We might hope to upgrade the Central Limit Theorem to get the stronger convergence in probability, but unfortunately this is impossible.

Exercise 3.30. Let $X_1, X_2, \ldots$ be independent identically distributed random variables with $\mathbf{P}(X_1 = 1) = \mathbf{P}(X_1 = -1) = 1/2$. For any $n \geq 1$, define

$$S_n := \frac{X_1 + \cdots + X_n}{\sqrt{n}}.$$

The Central Limit Theorem says that S_n converges in distribution to a standard Gaussian random variable. We show that S_n does not converge in probability to any random variable. The intuition here is that if S_n did converge in probability to a random variable Z, then when n is large, S_n is close to Z, $Y_n := \frac{\sqrt{2}S_{2n} - S_n}{\sqrt{2}-1}$ is close to Z, but S_n and Y_n are independent. And this cannot happen.

Proceed as follows. Assume that S_n converges in probability to Z.

- Let $\varepsilon > 0$. For n very large (depending on ε), we have $\mathbf{P}(|S_n - Z| > \varepsilon) < \varepsilon$ and $\mathbf{P}(|Y_n - Z| > \varepsilon) < \varepsilon$.
- Show that $\mathbf{P}(S_n > 0, Y_n > 0)$ is around $1/4$, using independence and the Central Limit Theorem.
- From the first item, show $\mathbf{P}(S_n > 0 | Z > \varepsilon) > 1 - \varepsilon$, $\mathbf{P}(Y_n > 0 | Z > \varepsilon) > 1 - \varepsilon$, so $\mathbf{P}(S_n > 0, Y_n > 0 | Z > \varepsilon) > 1 - 2\varepsilon$.
- Without loss of generality, for ε small, we have $\mathbf{P}(Z > \varepsilon) > 4/9$.
- By conditioning on $Z > \varepsilon$, show that $\mathbf{P}(S_n > 0, Y_n > 0)$ is at least $3/8$, when n is large.

The Weak Law of Large Numbers, Theorem 3.11, showed that the average $\frac{X_1 + \cdots + X_n}{n}$ of independent identically distributed random variables with finite mean converges to the mean in probability. We can upgrade this convergence in probability to a stronger notion of convergence, which we now define.

Definition 3.31 (**Almost Sure Convergence**). We say that random variables $X_1, X_2, \ldots$ converge **almost surely** (or **with probability one**) to a random variable X if

$$\mathbf{P}(\lim_{n\to\infty} X_n = X) = 1.$$

More rigorously, if Ω is the sample space, then $\mathbf{P}(\{\omega \in \Omega \colon \lim_{n\to\infty} X_n(\omega) = X(\omega)\}) = 1$

Exercise 3.32. Let $X_1, X_2, \ldots$ be random variables that converge almost surely to a random variable X. That is,

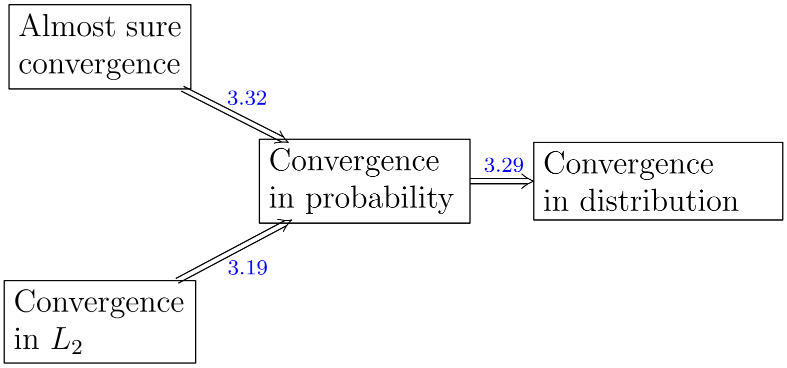
$$\mathbf{P}(\lim_{n\to\infty} X_n = X) = 1.$$

Show that $X_1, X_2, \ldots$ converges in probability to X in the following way.

- For any $\varepsilon > 0$ and for any positive integer n, let
$$A_{n,\varepsilon} := \bigcup_{m=n}^{\infty} \{\omega \in \Omega \colon\, |X_m(\omega) - X(\omega)| > \varepsilon\}.$$
Show that $A_{n,\varepsilon} \supseteq A_{n+1,\varepsilon} \supseteq A_{n+2,\varepsilon} \supseteq \cdots$.
- Show that $\mathbf{P}(\cap_{n=1}^{\infty} A_{n,\varepsilon}) = 0$.
- Using Continuity of the Probability Law, deduce that $\lim_{n\to\infty} \mathbf{P}(A_{n,\varepsilon}) = 0$.

Now, show that the converse is false. That is, find random variables $X_1, X_2, \ldots$ that converge in probability to X, but where $X_1, X_2, \ldots$ do not converge to X almost surely.

Remark 3.33. The following table summarizes our different notions of convergence of random variables. That is, the following table summarizes the implications of Exercises 3.19, 3.29 and 3.32.



Remark 3.34. Almost sure convergence does not imply convergence in L_2, and convergence in L_2 does not imply almost sure convergence.

To see the first, assertion, recall the random variables $Y_1, Y_2, \ldots$ constructed in Example 3.16. Then $Y_1, Y_2, \ldots$ converges almost surely to 0, since $\lim_{n\to\infty} Y_n(t) = 0$ for all $t \in (0, 1]$, so $\mathbf{P}(\lim_{n\to\infty} Y_n = 0) = \mathbf{P}((0, 1]) = 1$. On the other hand, $Y_1, Y_2, \ldots$ does not converge in L_2 to 0, since $\mathbf{E}\,|Y_n - 0|^2 = \mathbf{E}Y_n^2 = n^4/n = n^3$, so $\lim_{n\to\infty} \mathbf{E}\,|Y_n - 0|^2 \neq 0$.

We now show that convergence in L_2 does not imply almost sure convergence. Let $\mathbf{P}$ be the uniform probability law on $[1, 2]$. For any positive integer n, define $X_n \colon [1, 2] \to \mathbb{R}$ as follows. Let $j = j(n)$ be the nonnegative integer such that $2^j \leq n < 2^{j+1}$. Let $X_n(t) := 1$ if $t \in [n2^{-j}, (n+1)2^{-j}]$, and let $X_n(t) := 0$ otherwise. We claim that $X_1, X_2, \ldots$ converges to 0 in L_2, but $X_1, X_2, \ldots$ does not converge almost surely to 0. Note that $\mathbf{E}\,|X_n - 0|^2 = \mathbf{E}X_n^2 = 2^{-j}$, and as $n \to \infty$, $j \to \infty$, so that $\lim_{n\to\infty} \mathbf{E}\,|X_n - 0|^2 = 0$. However, for any $t \in [0, 1]$, there exist infinitely many values of n such that $X_n(t) = 1$ and infinitely many values of n such that $X_n(t) = 0$. Therefore, $\lim_{n\to\infty} X_n(t)$ does not exist, for every $t \in [0, 1]$. That is, $X_1, X_2, \ldots$ does not converge almost surely to any random variable.

From Corollary 3.4 and Corollary 2.24, if $X_1, \ldots, X_n$ are independent random variables with mean zero, then for any $t > 0$,
$$\mathbf{P}(|X_1 + \cdots + X_n| > t) \leq t^{-2}\mathrm{var}(X_1 + \cdots + X_n) = t^{-2}(\mathrm{var}(X_1) + \cdots + \mathrm{var}(X_n))$$
We used this inequality in our proof of the Weak Law of Large Numbers, Theorem 3.11. To prove the Strong Law of Large Numbers, we use the following stronger version of this inequality, where a maximum appears on the left side.

Theorem 3.35 (Kolmogorov Maximal Inequality). *Let $X_1, X_2, \ldots$ be independent random variables with mean zero and finite variance. Then for any $t > 0$, and for any $k > 0$,*

$$\mathbf{P}\left(\max_{1\leq n\leq k} |X_1 + \cdots + X_n| \geq t\right) \leq \frac{\mathrm{var}(X_1) + \cdots + \mathrm{var}(X_k)}{t^2}.$$

Proof. For an event A, we use the notation $1_A\colon \Omega \to \mathbb{R}$ where $1_A(\omega) := 1$ if $\omega \in A$, and $1_A(\omega) := 0$ if $\omega \notin A$.

Let $t > 0$. For any $n \geq 1$, define $S_n := X_1 + \cdots + X_n$. For any $n \geq 1$, let A_n be the event that $|S_n| \geq t$ and $|S_j| < t$ for all $1 \leq j < n$. Then $A_1, \ldots, A_k$ are disjoint, and $\cup_{n=1}^k A_n = \{\max_{1\leq n\leq k} |S_n| \geq t\}$. So, using $\mathbf{P}(\cup_{n=1}^k A_n) \leq \sum_{n=1}^k \mathbf{P}(A_n)$ and $\sum_{n=1}^k \mathrm{var}(X_n) = \mathbf{E}S_k^2$, it suffices to show that

$$\sum_{n=1}^k \mathbf{P}(A_n) \leq \frac{\mathbf{E}S_k^2}{t^2}. \qquad (*)$$

When A_n occurs, we have $1 \leq \frac{1}{t^2}S_n^2$. Therefore,

$$\mathbf{P}(A_n) = \mathbf{E}1_{A_n} \leq \mathbf{E}1_{A_n}\frac{1}{t^2}S_n^2, \qquad \forall\, 1 \leq n \leq k.$$

Below, we will show that

$$\mathbf{E}1_{A_n}S_n^2 \leq \mathbf{E}1_{A_n}S_k^2, \qquad \forall\, 1 \leq n \leq k. \qquad (**)$$

Then $(**)$ implies $(*)$, since the disjointness of the sets $A_1, \ldots, A_k$ implies $\sum_{n=1}^k 1_{A_n} \leq 1$, so

$$\sum_{n=1}^k \mathbf{P}(A_n) \leq \sum_{n=1}^k \mathbf{E}1_{A_n}\frac{1}{t^2}S_n^2 \overset{(**)}{\leq} \frac{1}{t^2}\mathbf{E}\sum_{n=1}^k 1_{A_n}S_k^2 \leq \frac{1}{t^2}\mathbf{E}S_k^2.$$

We now prove $(**)$. Let $1 \leq n \leq k$. Then, squaring both sides of $S_k = S_n + (S_k - S_n)$,

$$\begin{aligned} S_k^2 &= S_n^2 + (X_{n+1} + \cdots + X_k)^2 + 2S_n(X_{n+1} + \cdots + X_k) \\ &\geq S_n^2 + 2S_n(X_{n+1} + \cdots + X_k). \end{aligned}$$

Multiplying by 1_{A_n} and taking expected values,

$$\mathbf{E}S_k^2 1_{A_n} \geq \mathbf{E}S_n^2 1_{A_n} + 2\mathbf{E}[1_{A_n}S_n(X_{n+1} + \cdots + X_k)].$$

So, $(**)$ follows by showing the last term is zero. Note that $X_{n+1}, \ldots, X_k$ are independent of S_n, and $X_{n+1}, \ldots, X_k$ are independent of 1_{A_n}, since 1_{A_n} only depends on $X_1, \ldots, X_n$. Therefore,

$$\mathbf{E}[1_{A_n}S_n(X_{n+1} + \cdots + X_k)] = \mathbf{E}(1_{A_n}S_n) \cdot \mathbf{E}(X_{n+1} + \cdots + X_k) = 0.$$

The proof of $(**)$ is therefore complete. The Theorem follows. □

Exercise 3.36. Using the Central Limit Theorem, prove the Weak Law of Large Numbers.

Exercise 3.37. Let $m \geq 1$. Show by integral comparison of infinite series that

$$\sum_{j=m}^\infty \frac{1}{j^2} \leq \frac{10}{m}.$$

Theorem 3.38 (**Strong Law of Large Numbers**). *Let $X_1, X_2, \ldots$ be a sequence of independent identically distributed random variables. Let $\mu \in \mathbb{R}$. Assume that $\mu = \mathbf{E}X_1$. Then*

$$\mathbf{P}\left(\lim_{n\to\infty} \frac{X_1 + \cdots + X_n}{n} = \mu\right) = 1.$$

Proof. We make the additional assumption that $\mathrm{var}(X_1) < \infty$. Removing this extra assumption is beyond our course material.

For any $j \geq 1$, let $Y_j := X_j - \mu$. Note that $Y_1, Y_2, \ldots$ are independent identically distributed random variables with $\mathbf{E}Y_1 = 0$ and $\mathrm{var}(Y_1) = \mathrm{var}(X_1) < \infty$. We are required to show that

$$\mathbf{P}\left(\lim_{n\to\infty} \frac{Y_1 + \cdots + Y_n}{n} = 0\right) = 1.$$

Let $\varepsilon > 0$, $m \geq 1$. For any $n \geq 1$, let

$$A_n := \left\{\max_{m\leq k\leq n} \left|\sum_{j=m}^{k} \frac{Y_j}{j}\right| \geq \varepsilon\right\}.$$

From the Kolmogorov maximal inequality, Theorem 3.35,

$$\mathbf{P}(A_n) \leq \frac{1}{\varepsilon^2}\sum_{j=m}^{n} \frac{\mathrm{var}(Y_j)}{j^2} = \frac{\mathrm{var}(Y_1)}{\varepsilon^2}\sum_{j=m}^{n}\frac{1}{j^2}, \qquad \forall n \geq 1.$$

By their definition, $A_1 \subseteq A_2 \subseteq A_3 \subseteq \cdots$. So, by continuity of $\mathbf{P}$, Proposition 2.1, and Exercise 3.37,

$$\mathbf{P}\left(\max_{k\geq m}\left|\sum_{j=m}^{k}\frac{Y_j}{j}\right| \geq \varepsilon\right) = \mathbf{P}(\cup_{n=m}^{\infty} A_n) = \lim_{n\to\infty}\mathbf{P}(A_n) \leq \frac{10}{m}\frac{\mathrm{var}(Y_1)}{\varepsilon^2}.$$

Let A be the event that $\max_{k\geq m}\left|\sum_{j=m}^{k}\frac{Y_j}{j}\right| \geq \varepsilon$ for all $m \geq 1$. Then, A can be written as the decreasing intersection $A = \cap_{m=1}^{\infty}\{\max_{k\geq m}\left|\sum_{j=m}^{k}\frac{Y_j}{j}\right| \geq \varepsilon\}$. So, by continuity of $\mathbf{P}$, Proposition 2.2,

$$\mathbf{P}(A) = \lim_{m\to\infty}\mathbf{P}\left(\max_{k\geq m}\left|\sum_{j=m}^{k}\frac{Y_j}{j}\right| \geq \varepsilon\right) \leq \lim_{m\to\infty}\frac{10}{m}\frac{\mathrm{var}(Y_1)}{\varepsilon^2} = 0.$$

Since $\mathbf{P}(A) = 0$, $\mathbf{P}(A^c) = 1$. That is, with probability 1, for any $\varepsilon > 0$, there exists $m \geq 1$ such that $\max_{k\geq m}\left|\sum_{j=m}^{k}\frac{Y_j}{j}\right| < \varepsilon$.

The Proof is concluded by the following Propositions which we will not prove, since they are better suited for Math 131A. In particular, we apply the first Proposition to $S_k := \sum_{n=1}^{k}\frac{Y_n}{n}$, and we apply the second Proposition to $y_n := Y_n/n$ and $b_n := n$ for all $n \geq 1$. □

Proposition 3.39. *Let $s_1, s_2, \ldots$ be a sequence of real numbers such that: for all $\varepsilon > 0$ there exists $m \geq 1$ such that $\max_{k\geq m}|s_k - s_m| \leq \varepsilon$. Then $\lim_{k\to\infty} s_k$ exists.*

Proposition 3.40 (**Kronecker's Lemma**). *Let $y_1, y_2, \ldots$ be a sequence of real numbers. Let $0 < b_1 \leq b_2 \leq \cdots$ be a sequence of real numbers that goes to infinity. Assume that $\lim_{k\to\infty}\sum_{n=1}^{k} y_n$ exists. Then $\lim_{k\to\infty}\frac{1}{b_k}\sum_{n=1}^{k} b_n y_n = 0$.*

Remark 3.41. The Strong Law of Large Numbers Implies the Weak Law of Large Numbers by Exercise 3.32.

Exercise 3.42 (Renewal Theory). Let $t_1, t_2, \ldots$ be positive, independent identically distributed random variables. Let $\mu \in \mathbb{R}$. Assume $\mathbf{E}t_1 = \mu$. For any positive integer j, we interpret t_j as the lifetime of the j^{th} lightbulb (before burning out, at which point it is replaced by the $(j+1)^{st}$ lightbulb). For any $n \geq 1$, let $T_n := t_1 + \cdots + t_n$ be the total lifetime of the first n lightbulbs. For any positive integer t, let $N_t := \min\{n \geq 1\colon T_n \geq t\}$ be the number of lightbulbs that have been used up until time t. Show that N_t/t converges almost surely to $1/\mu$ as $t \to \infty$. (Hint: if c, t are positive integers, then $\{N_t \leq ct\} = \{T_{ct} \geq t\}$. Apply the Strong Law to T_{ct}.)

Exercise 3.43 (Playing Monopoly Forever). Let $t_1, t_2, \ldots$ be independent random variables, all of which are uniform on $\{1, 2, 3, 4, 5, 6\}$. For any positive integer j, we think of t_j as the result of rolling a single fair six-sided die. For any $n \geq 1$, let $T_n = t_1 + \cdots + t_n$ be the total number of spaces that have been moved after the n^{th} roll. (We think of each roll as the amount of moves forward of a game piece on a very large Monopoly game board.) For any positive integer t, let $N_t := \min\{n \geq 1\colon T_n \geq t\}$ be the number of rolls needed to get t spaces away from the start. Using Exercise 3.42, show that N_t/t converges almost surely to $2/7$ as $t \to \infty$.

Exercise 3.44 (Random Numbers are Normal). Let X be a uniformly distributed random variable on $(0, 1)$. Let X_1 be the first digit in the decimal expansion of X. Let X_2 be the second digit in the decimal expansion of X. And so on.

- Show that the random variables $X_1, X_2, \ldots$ are uniform on $\{0, 1, 2, \ldots, 9\}$ and independent.
- Fix $m \in \{0, 1, 2, \ldots, 9\}$. Using the Strong Law of Large Numbers, show that with probability one, the fraction of appearances of the number m in the first n digits of X converges to $1/10$ as $n \to \infty$.

(Optional): Show that for any ordered finite set of digits of length k, the fraction of appearances of this set of digits in the first n digits of X converges to 10^{-k} as $n \to \infty$. (You already proved the case $k = 1$ above.) That is, a randomly chosen number in $(0, 1)$ is normal. On the other hand, if we just pick some number such that $\sqrt{2} - 1$, then it may not be easy to say whether or not that number is normal.

(As an optional exercise, try to explicitly write down a normal number. This may not be so easy to do, even though a random number in $(0, 1)$ satisfies this property!)

Exercise 3.45. Let $X_1, X_2, \ldots$ be random variables with mean zero and variance one. The Strong Law of Large Numbers says that $\frac{1}{n}(X_1 + \cdots + X_n)$ converges almost surely to zero. The Central Limit Theorem says that $\frac{1}{\sqrt{n}}(X_1 + \cdots + X_n)$ converges in distribution to a standard Gaussian random variable. But what happens if we divide by some other power of n? This Exercise gives a partial answer to this question.

Let $\varepsilon > 0$. Show that

$$\frac{X_1 + \cdots + X_n}{n^{1/2}(\log n)^{(1/2)+\varepsilon}}$$

converges to zero almost surely as $n \to \infty$. (Hint: Re-do the proof of the Strong Law of Large Numbers, but divide by $n^{1/2}(\log n)^{(1/2)+\varepsilon}$ instead of n.)

4. Stochastic Processes

A **stochastic process** is a collection of random variables. These random variables are often indexed by time, and the random variables are often related to each other by the evolution of some physical procedure. Stochastic processes can then model random phenomena that depend on time.

Proposition 4.1 (**A Very Important Proposition**). *Let B be a fixed subset of some sample space Ω. Let $\mathbf{P}$ be a probability law on Ω. Assume that $\mathbf{P}(B) > 0$. Given any subset A in Ω, define $\mathbf{P}(A|B) = \mathbf{P}(A \cap B)/\mathbf{P}(B)$ as above. Then $\mathbf{P}(A|B)$ is itself a probability law on Ω, when viewed as a function of subsets A in Ω. Also, $\mathbf{P}(A|B)$ is a probability law on $\Omega \cap B$, when viewed as a function of subset A in $\Omega \cap B$*

Exercise 4.2. Let A, B be events in a sample space. Let $C_1, \ldots, C_n$ be events such that $C_i \cap C_j = \emptyset$ for any $i, j \in \{1, \ldots, n\}$, and such that $\cup_{i=1}^n C_i = B$. Show:

$$\mathbf{P}(A|B) = \sum_{i=1}^{n} \mathbf{P}(A|B,\, C_i)\mathbf{P}(C_i|B).$$

(Hint: consider using the Total Probability Theorem and Proposition 2.54.)

4.1. Bernoulli Process.

Our first example of a stochastic process will be the Bernoulli Process. Recall that a Bernoulli random variable X with parameter $0 < p < 1$ is a discrete random variable such that $\mathbf{P}(X = 1) = p$ and $\mathbf{P}(X = 0) = 1 - p$.

Remark 4.3. A set of random variables $X_1, X_2, \ldots$ is said to be independent if, for any integer $n \geq 1$, the set of random variables $X_1, \ldots, X_n$ is independent.

Definition 4.4 (**Bernoulli Process**). A **Bernoulli Process** with parameter $0 < p < 1$ is a sequence $X_1, X_2, \ldots$ of independent Bernoulli random variables, each with parameter p.

Remark 4.5. If $X_1, X_2, \ldots$ is a Bernoulli process with parameter $0 < p < 1$, and if $T := \min\{n \geq 1 \colon X_n = 1\}$, then T is the time of the first "successful" coin flip, and T has a geometric PMF: $p_T(t) = (1-p)^{t-1}p$ for any integer $t \geq 1$.

Proposition 4.6. *Let T be a geometric random variable with parameter $0 < p < 1$. Then T has the following **memoryless property**: for any integers $n, t \geq 1$*

$$\mathbf{P}(T - n = t \,|\, T > n) = \mathbf{P}(T = t).$$

Proof.

$$\mathbf{P}(T - n = t \,|\, T > n) = \frac{\mathbf{P}(T - n = t, T > n)}{\mathbf{P}(T > n)} = \frac{\mathbf{P}(T = t + n)}{\mathbf{P}(T > n)} = \frac{(1-p)^{t+n-1}p}{\sum_{k=n+1}^{\infty}(1-p)^{k-1}p}$$
$$= \frac{(1-p)^{t+n-1}p}{(1-p)^n} = (1-p)^{t-1}p = \mathbf{P}(T = t).$$

□

Remark 4.7. A Bernoulli Process $X_1, X_2, \ldots$ has the following **Markov property** or **fresh-start property**. For any integer $n \geq 1$, the sequence $X_{n+1}, X_{n+2}, \ldots$ is itself a Bernoulli process which is independent of $X_1, \ldots, X_n$.

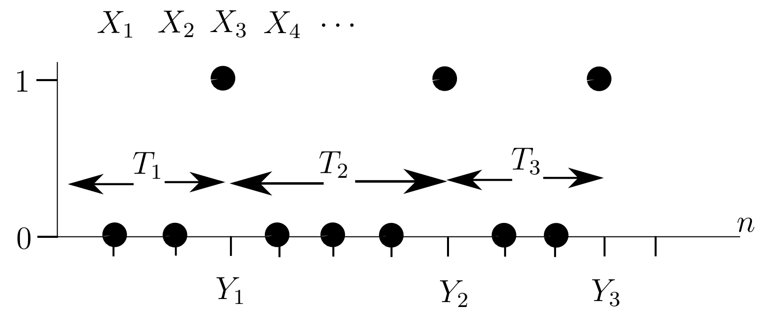


FIGURE 1. One Sample from a Bernoulli Process.

Proposition 4.8. *Let $X_1, X_2, \ldots$ be a Bernoulli Process with parameter $0 < p < 1$. Define $Y_1 := \min\{n \geq 1 \colon X_n = 1\}$. For any integer $k \geq 1$, inductively define $Y_k := \min\{n > Y_{k-1} \colon X_n = 1\}$ to be the time of the k^{th} "successful trial." Then, define $T_1 := Y_1$, and $T_k := Y_k - Y_{k-1}$ for any $k \geq 2$. Then the inter-arrival times $T_1, T_2, \ldots$ are independent geometric random variables with parameter p.*

Remark 4.9. When $A_1, \ldots, A_n, B_1, \ldots, B_m$ are events, we use the notation

$$\mathbf{P}(A_1, \ldots, A_n \,|\, B_1, \ldots, B_m) := \mathbf{P}(A_1 \cap \cdots \cap A_n \,|\, B_1 \cap \cdots \cap B_m).$$

Proof. The case $k = 1$ follows by definition of $Y_1 = T_1$ as noted in Remark 4.5. We now consider any $k \geq 2$. By its definition, $T_k = Y_k - Y_{k-1} = \min\{n \geq 1 \colon X_{n+Y_{k-1}} = 1\}$. Let $s, t \geq 1$ be integers, and let $x_1, \ldots, x_s \in \{0, 1\}$ so that $\{X_1 = x_1, \ldots, X_s = x_s\} \subseteq \{Y_{k-1} = s\}$. Then

$$\begin{aligned}
&\mathbf{P}(T_k = t \,|\, Y_{k-1} = s,\, X_1 = x_1, \ldots, X_n = x_n) \\
&= \mathbf{P}(X_{1+s} = 0, \ldots, X_{t-1+s} = 0, X_{t+s} = 1 \,|\, Y_{k-1} = s,\, X_1 = x_1, \ldots, X_s = x_s) \\
&= \mathbf{P}(X_{1+s} = 0, \ldots, X_{t-1+s} = 0, X_{t+s} = 1 \,|\, X_1 = x_1, \ldots, X_s = x_s) \\
&= \mathbf{P}(X_{1+s} = 0, \ldots, X_{t-1+s} = 0, X_{t+s} = 1) \qquad \text{, by Remark 4.7} \\
&= \mathbf{P}(X_1 = 0, \ldots, X_{t-1} = 0, X_t = 1) \qquad \text{, by Remark 4.7} \\
&= (1-p)^{t-1} p.
\end{aligned}$$

Summing over all $x_1, \ldots, x_s \in \{0, 1\}$ such that $\{X_1 = x_1, \ldots, X_s = x_s\} \subseteq \{Y_{k-1} = s\}$ and using the Total Probability Theorem as in Exercise 4.2, we get

$$\mathbf{P}(T_k = t \,|\, Y_{k-1} = s) = (1-p)^{t-1} p, \qquad \forall\, s, t \geq 1.$$

Summing over all $s \geq 1$, we get by the Total Probability Theorem,

$$\mathbf{P}(T_k = t) = (1-p)^{t-1} p, \qquad \forall\, t \geq 1. \qquad (*)$$

Finally, to prove the independence property, let $k \geq 1$, and let $t_1, \ldots, t_k \geq 1$, so that

$$\begin{aligned}
&\mathbf{P}(T_1 = t_1, \ldots, T_k = t_k) \\
&= \mathbf{P}(X_1 = 0, \ldots, X_{t_1 - 1} = 0, X_{t_1} = 1, X_{t_1+1} = 0, \ldots, X_{t_1+t_2-1} = 0, X_{t_1+t_2} = 1, \\
&\qquad\qquad \ldots, X_{t_1+\cdots+t_k-1} = 0, X_{t_1+\cdots+t_k} = 1) \\
&= \prod_{i=1}^{k} (1-p)^{t_i - 1} p \overset{(*)}{=} \prod_{i=1}^{k} \mathbf{P}(T_i = t_i).
\end{aligned}$$

□

Exercise 4.10. Let $T_1, T_2, \ldots$ be independent geometric random variables with parameter p. For any integer $k \geq 1$, let $Y_k := T_1 + \cdots + T_k$. Show that the PMF of Y_k is given by

$$p_{Y_k}(t) = \begin{cases} \binom{t-1}{k-1} p^k (1-p)^{t-k} & \text{, if } t \geq k,\ t \in \mathbb{Z} \\ 0 & \text{, otherwise.} \end{cases}$$

Proposition 4.8 shows that the inter-arrival times of the Bernoulli process are independent geometric random variables. In fact, we can reverse this implication. That is, if we assume that the inter-arrival times of some sequence of random variables are independent and geometric, then the sequence of random variables is a Bernoulli process.

Proposition 4.11 (**An Equivalent Definition of Bernoulli Process**). *Let $0 < p < 1$. Let $T_1, T_2, \ldots$ be independent geometric random variables with parameter p. Define a sequence of random variables $X_1, X_2, \ldots$ such that, for any integer $n \geq 1$*

$$X_n = \begin{cases} 1 & \text{, if } n = T_1, T_1 + T_2, T_1 + T_2 + T_3, \ldots \\ 0 & \text{, otherwise.} \end{cases}$$

Then $X_1, X_2, \ldots$ is a Bernoulli process with parameter p.

Proof. By its definition, $X_1 = 1$ only if $T_1 = 1$, and $X_1 = 0$ otherwise. So $\mathbf{P}(X_1 = 1) = \mathbf{P}(T_1 = 1) = p$, since T_1 is a geometric random variable with parameter p. Similarly, $\mathbf{P}(X_1 = 0) = 1 - \mathbf{P}(X_1 = 1) = 1 - p$. So, X_1 is a Bernoulli random variable.

We now consider the case of general $k \geq 1$. By the Total Probability Theorem,

$$\begin{aligned}
&\mathbf{P}(X_{k+1} = 1) \\
&= \sum_{x_1, \ldots, x_k \in \{0,1\}} \mathbf{P}(X_{k+1} = 1 \,|\, X_k = x_k, \ldots, X_1 = x_1) \mathbf{P}(X_k = x_k, \ldots, X_1 = x_1). \qquad (*)
\end{aligned}$$

Given any $x_1, \ldots, x_k \in \{0,1\}$, let $i_1 < \cdots < i_m$ such that $x_{i_1} = \cdots = x_{i_m} = 1$ and so that $x_j = 0$ if $j \notin \{i_1, \ldots, i_m\}$. Then

$$\begin{aligned}
&\mathbf{P}(X_{k+1} = 1 \,|\, X_k = x_k, \ldots, X_1 = x_1) \\
&= \mathbf{P}(X_{k+1} = 1 \,|\, T_{m+1} > k - i_m, T_m = i_m - i_{m-1}, \ldots, T_2 = i_2 - i_1, T_1 = i_1) \\
&= \mathbf{P}(T_{m+1} = k + 1 - i_m \,|\, T_{m+1} > k - i_m, T_m = i_m - i_{m-1}, \ldots, T_2 = i_2 - i_1, T_1 = i_1).
\end{aligned}$$

Using that the random variables $T_1, \ldots, T_{m+1}$ are independent and Proposition 4.6,

$$\begin{aligned}
\mathbf{P}(X_{k+1} = 1 \,|\, X_k = x_k, \ldots, X_1 = x_1) &= \mathbf{P}(T_{m+1} = k + 1 - i_m \,|\, T_{m+1} > k - i_m) \\
&= \mathbf{P}(T_{m+1} = 1) = p.
\end{aligned}$$

This shows the independence property. Substituting back into $(*)$,

$$\mathbf{P}(X_{k+1}=1)=\sum_{x_1,\ldots,x_k\in\{0,1\}} p\cdot\mathbf{P}(X_k=x_k,\ldots,X_1=x_1)=p.$$

$\square$

Exercise 4.12. Give an alternate proof that $\mathbf{P}(X_{k+1}=1)=p$ in Proposition 4.11 by using the following conditioning argument:

$$\mathbf{P}(X_{k+1}=1)=\sum_{n=1}^{k+1}\mathbf{P}(X_{k+1}=1\,|\,T_1=n)\mathbf{P}(T_1=n)$$

$$=\mathbf{P}(X_{k+1}=1\,|\,T_1=k+1)\mathbf{P}(T_1=k+1)+\sum_{n=1}^{k}\mathbf{P}(X_{k+1}=1\,|\,T_1=n)\mathbf{P}(T_1=n)$$

$$=\mathbf{P}(T_1=k+1)+\sum_{n=1}^{k}\mathbf{P}(T_1+\cdots+T_j=k+1\text{ for some }j\geq 2\,|\,T_1=n)\mathbf{P}(T_1=n)=\cdots$$

Remark 4.13 (**Splitting**)**.** Let $X_1,X_2,\ldots$ be a Bernoulli process with parameter $0<p<1$. Let $0<q<1$. For any integer $n\geq 1$, define a random variable Z_n so that $Z_n:=X_n$ with probability q, and $Z_n:=0$ with probability $1-q$. Since $X_1,X_2,\ldots$ are independent, $Z_1,Z_2,\ldots$ are independent. Also, for any integer $n\geq 1$, $\mathbf{P}(Z_n=1)=q\mathbf{P}(X_n=1)=pq$, and $\mathbf{P}(Z_n=0)=1-pq$. So, $Z_1,Z_2,\ldots$ is a Bernoulli process with parameter pq. We can think of $Z_1,Z_2,\ldots$ so that whenever the Bernoulli process $X_1,X_2,\ldots$ succeeds, we flip a coin, and record success for $Z_1,Z_2,\ldots$ with (conditional) probability q. In this way, $Z_1,Z_2,\ldots$ is "split" away from $X_1,X_2,\ldots$.

Remark 4.14 (**Merging**)**.** Let $X_1,X_2,\ldots$ be a Bernoulli process with parameter $0<p<1$. Let $Y_1,Y_2,\ldots$ be another Bernoulli process with parameter $0<q<1$ which is independent of $X_1,X_2,\ldots$. For any integer $n\geq 1$, define a random variable Z_n so that $Z_n:=\max(X_n,Y_n)$ Since $X_1,X_2,\ldots,Y_1,Y_2,\ldots$ are independent, $Z_1,Z_2,\ldots$ are independent. Also, for any integer $n\geq 1$, $\mathbf{P}(Z_n=0)=\mathbf{P}(X_n=0)\mathbf{P}(Y_n=0)=(1-p)(1-q)$, and $\mathbf{P}(Z_n=1)=1-\mathbf{P}(Z_n=0)$. So, $Z_1,Z_2,\ldots$ is a Bernoulli process with parameter $1-(1-p)(1-q)=p+q-pq$. We can think of $Z_1,Z_2,\ldots$ so that whenever either Bernoulli process $X_1,X_2,\ldots$ or $Y_1,Y_2\ldots$ succeeds, we record success for $Z_1,Z_2,\ldots$. In this way, $Z_1,Z_2,\ldots$ "merges" $X_1,X_2,\ldots$ and $Y_1,Y_2,\ldots$.

Exercise 4.15. Let $X_1,X_2,\ldots$ be a Bernoulli process with parameter $p=1/2$. What is the expected number of trials that have to occur before we see two consecutive "successes"?

Exercise 4.16. Let $X_1,X_2,\ldots$ be a Bernoulli process with parameter $p=1/2$. Define $N:=\min\{n\geq 1\colon X_n\neq X_1\}$. For any $n\geq 1$, define $Y_n:=X_{N+n-2}$. Show that $\mathbf{P}(Y_n=1)=1/2$ for all $n\geq 1$, but $Y_1,Y_2,\ldots$ is not a Bernoulli process.

4.2. **Poisson Process.** Let $\lambda>0$. For any $n\geq 1$, let $p_n:=\lambda/n$. Let $X_1,X_2,\ldots$ be a Bernoulli Process with parameter p_n. For any integer $i\geq 1$, let $N_{i,n}:=X_1+\cdots+X_i$. Intuitively, as $n\to\infty$, we hope to get a new "continuous-time" stochastic process from the Bernoulli Process. Note that $N_{i,n}$ is the number of "successes" of the Bernoulli process among the first i "trials." So, $N_{i,n}$ has a binomial distribution with parameters i and $p_n=\lambda/n$. So, if $i = sn$ for some rational constant s, then $N_{i,n} = N_{sn,n}$ has a binomial distribution with parameters sn and $p_n = \lambda/n$.

We now recall a Proposition from your previous probability class.

Proposition 4.17 (**Poisson Approximation to the Binomial**). *Let $\lambda > 0$. For each positive integer n, let $0 < p_n < 1$. Assume that $\lim_{n\to\infty} p_n = 0$ and $\lim_{n\to\infty} np_n = \lambda$. Let B_n be a binomial distributed random variable with parameters n and p_n. (So, $\mathbf{P}(B_n = t) = \binom{n}{t}p_n^t(1-p_n)^{n-t}$ for any integer $0 \leq t \leq n$.) Then, for any nonnegative integer t, we have*

$$\lim_{n\to\infty} \mathbf{P}(B_n = t) = e^{-\lambda}\frac{\lambda^t}{t!}.$$

That is, as $n \to \infty$ the binomial random variable B_n converges in distribution to a Poisson random variable with parameter λ.

We think of $N_{1,n}, N_{2,n}, \ldots$ as a sequence of random variables such that the value $N_{i,n}$ is plotted on the x-axis at time i/n. Then the random variables $N_{1,n}, N_{2,n}, \ldots$ "converge" as $n \to \infty$ to a set of random variables $\{N(s)\}_{s\geq 0}$, where $s \geq 0$ denotes any nonnegative real number. From Proposition 4.17, we anticipate that $N(s)$ has a Poisson distribution with parameter λs. This observation leads to our first informal definition of the Poisson Process $\{N(s)\}_{s\geq 0}$.

Definition 4.18 (**Poisson Process, Informal Definition**). Let $\lambda > 0$. For any $n \geq 1$, let $p_n := \lambda/n$ and let $X_1, X_2, \ldots$ be a Bernoulli Process with parameter p_n. For any integer $i \geq 1$, let $N_{i,n} := X_1 + \cdots + X_i$. Then the sequence of random variables $N_{1/n}, N_{2/n}, N_{3/n}, \ldots$ "converges" as $n \to \infty$ to a set of random variables $\{N(s)\}_{s\geq 0}$, which is defined to be the Poisson Process with parameter $\lambda > 0$. (Here $s \geq 0$ denotes any nonnegative real number.)

Remark 4.19. From this informal definition, we see that for any $s \geq 0$, $N(s)$ has nonnegative integer values. Also, we expect that $N(s+r) - N(s)$ is independent of $N(s)$ for any $r, s > 0$, since this independence property applies to the sequence $N_{1,n}, N_{2,n}, N_{3,n}, \ldots$, for any $n \geq 1$.

Proposition 4.20. *A Poisson Process with parameter $\lambda > 0$ has independent, exponential inter-arrival times with parameter λ.*

Proof Sketch. For any $n \geq 1$, let $p_n := \lambda/n$ and let $X_1, X_2, \ldots$ be a Bernoulli Process with parameter p_n. For any integer $i \geq 1$, let $N_{i,n} := X_1 + \cdots + X_i$.

Define $Y_1 := \min\{k \geq 1 \colon X_k = 1\}$. For any integer $i \geq 1$, inductively define $Y_i := \min\{k > Y_{i-1} \colon X_k = 1\}$. Then, define $T_1 := Y_1$, and $T_i := Y_i - Y_{i-1}$ for any $i \geq 2$. Then the inter-arrival times $T_1, T_2, \ldots$ are independent geometric random variables with parameter $p_n = \lambda/n$, by Proposition 4.8. That is, for any integers $i, t \geq 1$,

$$\mathbf{P}(T_i = t) = (\lambda/n)(1-\lambda/n)^{t-1}. \qquad (*)$$

Define a new sequences of random variables so that $Z_{i,n} := Y_i/n$ for any integer $i \geq 1$, and define $S_{i,n} := Z_{i,n} - Z_{(i-1),n} = T_i/n$. Then $S_{1,n}, S_{2,n}, \ldots$ are the inter-arrival times for the sequence $N_{1,n}, N_{2,n}, \ldots$, so

$$\mathbf{P}(S_{i.n} = t/n) = \mathbf{P}(T_i/n = t/n) = \mathbf{P}(T_i = t) \overset{(*)}{=} (\lambda/n)(1-\lambda/n)^{t-1}, \qquad \forall\, i, t \geq 1.$$

So,

$$\mathbf{P}(S_{i,n}\leq 1)=\sum_{t=1}^{n}\mathbf{P}(S_{i,n}=t/n)=\sum_{t=1}^{n}(\lambda/n)(1-\lambda/n)^{t-1}=\sum_{t=1}^{n}(\lambda/n)(1-\lambda/n)^{n(t/n)-1}.$$

Letting $n\to\infty$, and using $\lim_{n\to\infty}(1-\lambda/n)^n=e^{-\lambda}$, we see that the Riemann sum converges as follows

$$\lim_{n\to\infty}\mathbf{P}(S_{i,n}\leq 1)=\int_0^1\lambda e^{-\lambda x}dx.$$

More generally, for any $s\geq 0$,

$$\lim_{n\to\infty}\mathbf{P}(S_{i,n}\leq s)=\int_0^s\lambda e^{-\lambda x}dx.$$

That is, as $n\to\infty$, the inter-arrival times (divided by n) of the Bernoulli Process with parameter $p_n=\lambda/n$ converge to exponential random variables with parameter λ. □

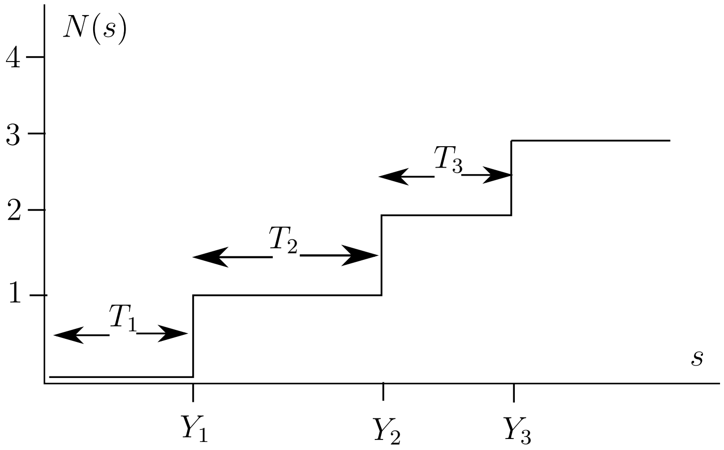


FIGURE 2. One Sample Path of a Poisson Process.

We have informally argued that the Poisson Process with parameter λ should have independent exponential inter-arrival times with parameter λ. We can actually use independent exponential random variables to give a formal definition of a Poisson Process.

Definition 4.21 (Poisson Process, Formal Definition). Let $\lambda>0$. Let $T_1,T_2,\ldots$ be independent exponential random variables with parameter λ. Let $Y_0=0$, and for any $n\geq 1$, let $Y_n:=T_1+\cdots+T_n$. A **Poisson Process** with parameter $\lambda>0$ is a set of integer-valued random variables $\{N(s)\}_{s\geq 0}$ defined by $N(s):=\max\{n\geq 0\colon Y_n\leq s\}$, $\forall\ s\geq 0$.

Here the random variables $T_1,T_2,\ldots$ are the inter-arrival times of the process. The following properties follow from the formal definition 4.21, though we anticipated these properties e.g. in Proposition 4.17.

Proposition 4.22 (Properties of the Poisson Process). *Let $\{N(s)\}_{s\geq 0}$ be a Poisson process with parameter $\lambda>0$. Then*

(i) $N(0)=0$.
(ii) $N(t+s)-N(s)$ *is a Poisson random variable with parameter λt for all $s,t>0$.*

(iii) $\{N(s)\}_{s\geq 0}$ *has* ***independent increments****. That is, for any* $0 < u_0 < \cdots < u_k$*, the following random variables are independent:*

$$N(u_1) - N(u_0), \ldots, N(u_k) - N(u_{k-1}).$$

We omit the proof of this Proposition, since it is covered in Math 171.

Exercise 4.23. Suppose the number of students going to a restaurant in Ackerman in a single day has a Poisson distribution with mean 500. Suppose each student spends an average of \$10 with a standard deviation of \$5. What is the average revenue of the restaurant in one day? What is the standard deviation of the revenue in one day? (The amounts spent by the students are independent identically distributed random variables.)

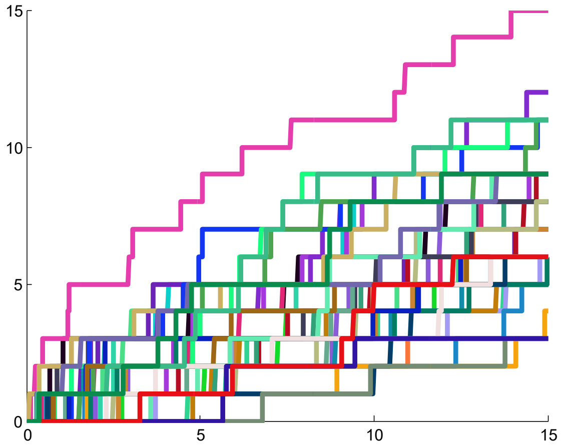


FIGURE 3. Several Sample Paths of a Poisson Process. The horizontal axis is the s-axis.

4.3. **Random Walks.**

Definition 4.24 (**Random Walk**)**.** Let $X_1, X_2, \ldots$ be independent identically distributed random variables with $\mathbf{E}X_1 = 0$ and $\mathbf{E}\,|X_1| < \infty$. Let $X_0 := 0$ and for any integer $n \geq 0$, define $S_n := X_0 + \cdots + X_n$. We call the sequence of random variables $S_0, S_1, \ldots$ a **random walk** started at 0. More generally, if $c \in \mathbb{R}$ is a constant and if $X_0 = c$, we call the sequence of random variables $S_0, S_1, \ldots$ a **random walk** started at c.

Definition 4.25 (**Stopping Time**)**.** A **stopping time** for a random walk $S_0, S_1, \ldots$ is a random variable T taking values in $0, 1, 2, \ldots, \cup\{\infty\}$ such that, for any integer $n \geq 0$, the event $\{T = n\}$ is determined by $S_0, \ldots, S_n$. More formally, for any integer $n \geq 1$, there is a set $B_n \subseteq \mathbb{R}^{n+1}$ such that $\{T = n\} = \{(S_0, \ldots, S_n) \in B_n\}$. Put another way, the indicator function $1_{\{T=n\}}$ is a function of the random variables $S_0, \ldots, S_n$.

For a real-world example of a stopping time, suppose $S_0, S_1, \ldots$ is a random walk which describes the price of a stock. Suppose the stock is currently priced at $S_0 = 100$ and you instruct your stock broker to sell the stock when its price reaches either \$110 or \$90. That is, define the stopping time $T = \min\{n \geq 1 \colon S_n \geq 110 \text{ or } S_n \leq 90\}$. Then T is a stopping

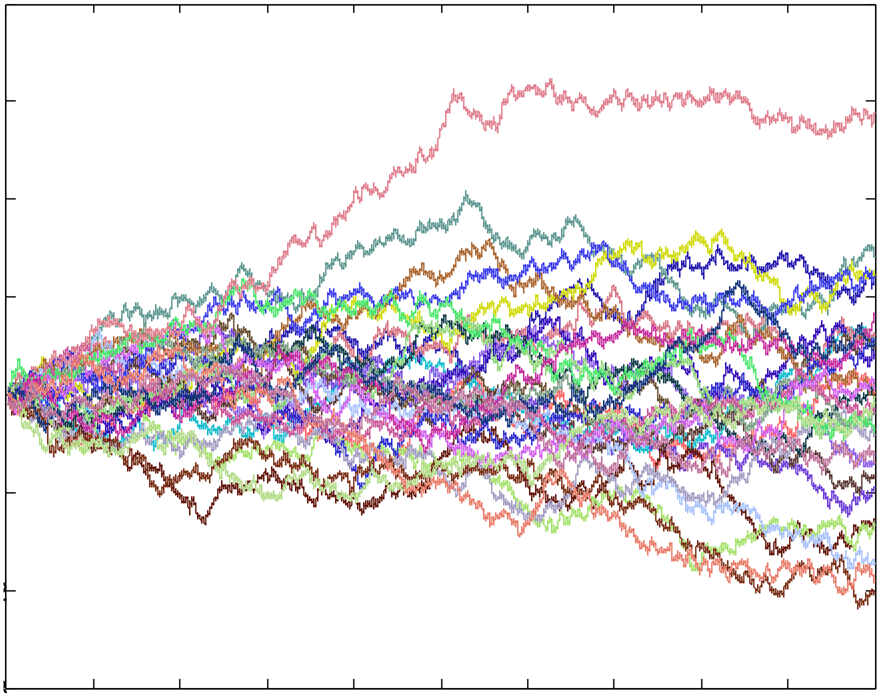


FIGURE 4. Several Sample Paths of a Random Walk.

time. From the Optional Stopping Theorem below, $\mathbf{E}S_T = \mathbf{E}S_0$. That is, you cannot make money off of this stock (if it is a random walk).

Exercise 4.26. Let $X_0 := 0$. Let $X_0, X_1, \ldots$ be independent random variables such that $\mathbf{P}(X_n = 1) = \mathbf{P}(X_n = -1) = 1/2$ for all $n \geq 1$. Let $S_0, S_1, \ldots$ be the corresponding random walk started at 0. Let $T := \min\{n \geq 1 \colon S_n = 1\}$. Show that T is a stopping time.

Exercise 4.27. Let $X_0 := x_0 \in \mathbb{Z}$. Let $X_0, X_1, \ldots$ be independent random variables such that $\mathbf{P}(X_n = 1) = \mathbf{P}(X_n = -1) = 1/2$ for all $n \geq 1$. Let $S_0, S_1, \ldots$ be the corresponding random walk started at x_0. Let $a, b \in \mathbb{Z}$ such that $a < x_0 < b$. Let $T := \min\{n \geq 1 \colon S_n \in \{a, b\}\}$. Show that T is a stopping time.

Remark 4.28. Let $a, b \in \mathbb{R}$. We use the notation $a \wedge b := \min(a, b)$. Note that if T is a stopping time, then $a \wedge T$ is a stopping time, for any fixed $a \in \mathbb{R}$.

Theorem 4.29 (Optional Stopping Theorem, Version 1). *Let $(S_0, S_1, \ldots)$ be a random walk, and let T be a stopping time such that $\mathbf{P}(T < \infty) = 1$. Then $\mathbf{E}S_{n\wedge T} = \mathbf{E}S_0$ for all $n \geq 0$.*

Proof. The proof is identical to that of Wald's equation in Proposition 2.65. Let $n \geq 0$ be an integer. Conditioned on $T = m$, we know that $S_{n\wedge T} = X_0 + \cdots + X_{n\wedge m}$. So, $\mathbf{E}(S_{n\wedge T}|T = m) = \mathbf{E}(X_0 + \cdots + X_{n\wedge m}) = \mathbf{E}S_0 + (n \wedge m)\mathbf{E}X_1 = \mathbf{E}S_0$, since $\mathbf{E}X_1 = 0$. So, $\mathbf{E}(S_{n\wedge T}|T) = \mathbf{E}S_0$, and by Exercise 2.30,

$$\mathbf{E}S_{n\wedge T} = \mathbf{E}(\mathbf{E}(S_{n\wedge T}|T)) = \mathbf{E}S_0.$$

□

Theorem 4.30 (**Bounded Convergence Theorem**). *Let $c \in \mathbb{R}$. Let $Y_0, Y_1, \ldots$ be a sequence of random variables such that $|Y_n| \leq c$ for all $n \geq 0$. Assume that $Y_0, Y_1, \ldots$ converges in probability to a random variable Z. Then $\lim_{n\to\infty} \mathbf{E}Y_n = \mathbf{E}Z$.*

Proof. Since $Y_0, Y_1, \ldots$ converges in probability to Z, it follows from Exercise 3.29 that $Y_0, Y_1, \ldots$ converges in distribution to Z. So, since $|Y_n| \leq c$ for all $n \geq 0$, we conclude that $\mathbf{P}(|Z| \leq c) = 1$.

Fix $\varepsilon > 0$. Let $A := \{|Y_n - Z| > \varepsilon\}$. Then

$$\mathbf{E}Y_n - \mathbf{E}Z = \mathbf{E}(Y_n - Z)(1_A + 1_{A^c}) = \mathbf{E}(Y_n - Z)1_A + \mathbf{E}(Y_n - Z)1_{A^c}.$$

We bound each term separately. We have

$$|\mathbf{E}(Y_n - Z)1_A| \leq \mathbf{E}\,|Y_n - Z|\,1_A \leq \mathbf{E}(|Y_n| + |Z|)1_A \leq 2c \cdot \mathbf{E}1_A = 2c \cdot \mathbf{P}(A). \qquad (*)$$

Also, since $A^c = \{|Y_n - Z| \leq \varepsilon\}$, we have

$$|\mathbf{E}(Y_n - Z)1_{A^c}| \leq \mathbf{E}\,|Y_n - Z|\,1_{A^c} \leq \varepsilon\mathbf{E}1_{A^c} \leq \varepsilon.$$

So, by the triangle inequality, for any $n \geq 1$,

$$|\mathbf{E}Y_n - \mathbf{E}Z| \leq 2c \cdot \mathbf{P}(A) + \varepsilon.$$

Letting $n \to \infty$ and using the definition of convergence in probability, we then get

$$\lim_{n\to\infty} |\mathbf{E}Y_n - \mathbf{E}Z| \leq \varepsilon, \qquad \forall\, \varepsilon > 0.$$

Since $\varepsilon > 0$ is arbitrary, we conclude that $\lim_{n\to\infty} |\mathbf{E}Y_n - \mathbf{E}Z| = 0$, as desired. □

Remark 4.31. Let $\mathbf{P}$ be the uniform probability law on $[0, 1]$. For any $n \geq 1$, consider the function $f_n \colon [0, 1] \to \mathbb{R}$ such that $f_n = n1_{[0,1/n]}$. Then $\mathbf{E}f_n = 1$, but f_n converges in probability to 0, so $1 = \lim_{n\to\infty} \mathbf{E}f_n \neq \mathbf{E}0 = 0$. So, the boundedness assumption is important in Theorem 4.30

Theorem 4.32 (**Optional Stopping Theorem, Version 2**). *Let $(S_0, S_1, \ldots)$ be a random walk, and let T be a stopping time such that $\mathbf{P}(T < \infty) = 1$. Let $c \in \mathbb{R}$. Assume that $|S_{n\wedge T}| \leq c$ for all $n \geq 0$. Then $\mathbf{E}S_T = \mathbf{E}S_0$.*

Proof. From Theorem 4.29, $\mathbf{E}S_{n\wedge T} = \mathbf{E}S_0$ for all $n \geq 0$. Also, since $\mathbf{P}(T < \infty) = 1$, we have

$$\mathbf{P}(\lim_{n\to\infty} S_{n\wedge T} = S_T) = 1.$$

That is, $S_{0\wedge T}, S_{1\wedge T}, \ldots$ converges almost surely to S_T. By Exercise 3.32, $S_{0\wedge T}, S_{1\wedge T}, \ldots$ converges in probability to S_T. So, by the Bounded Convergence Theorem 4.30,

$$\mathbf{E}S_0 = \lim_{n\to\infty} \mathbf{E}S_{n\wedge T} = \mathbf{E}S_T.$$

□

Example 4.33. Let $X_0 := x_0 \in \mathbb{Z}$. Let $X_0, X_1, \ldots$ be independent random variables such that $\mathbf{P}(X_n = 1) = \mathbf{P}(X_n = -1) = 1/2$ for all $n \geq 1$. Let $S_0, S_1, \ldots$ be the corresponding random walk started at x_0. Let $a, b \in \mathbb{Z}$ such that $a < x_0 < b$. Let $T := \min\{n \geq 1 \colon S_n \in \{a, b\}\}$. Then T is a stopping time by Exercise 4.27. Also, $|S_{n\wedge T}| \leq \max(|a|, |b|)$, so Theorem 4.32 applies. Let $c := \mathbf{P}(S_T = a)$. Then

$$x_0 = \mathbf{E}S_0 = \mathbf{E}S_T = ac + (1 - c)b.$$

Solving for c, we get

$$c = \frac{x_0 - b}{a - b}.$$

(It can be shown that $\mathbf{P}(T < \infty) = 1$, but we will not do so here.)

5. Appendix: Notation

Let n, m be a positive integers. Let A, B be sets contained in a universal set Ω.

$\mathbb{R}$ denotes the set of real numbers

$\in$ means "is an element of." For example, $2 \in \mathbb{R}$ is read as "2 is an element of $\mathbb{R}$."

$\forall$ means "for all"

$\exists$ means "there exists"

$\mathbb{R}^n = \{(x_1, x_2, \ldots, x_n) \colon x_i \in \mathbb{R}\, \forall\, 1 \leq i \leq n\}$

$f \colon A \to B$ means f is a function with domain A and range B. For example,

$f \colon \mathbb{R}^2 \to \mathbb{R}$ means that f is a function with domain $\mathbb{R}^2$ and range $\mathbb{R}$

$\emptyset$ denotes the empty set

$A \subseteq B$ means $\forall\, a \in A$, we have $a \in B$, so A is contained in B

$A \setminus B := \{a \in A \colon a \notin B\}$

$A^c := \Omega \setminus A$, the complement of A in Ω

$A \cap B$ denotes the intersection of A and B

$A \cup B$ denotes the union of A and B

$\mathbf{P}$ denotes a probability law on Ω

$\mathbf{P}(A|B)$ denotes the conditional probability of A, given B.

Let $a_1, \ldots, a_n$ be real numbers. Let n be a positive integer.

$$\sum_{i=1}^{n} a_i = a_1 + a_2 + \cdots + a_{n-1} + a_n.$$

$$\prod_{i=1}^{n} a_i = a_1 \cdot a_2 \cdots a_{n-1} \cdot a_n.$$

$\min(a_1, a_2)$ denotes the minimum of a_1 and a_2.

$\max(a_1, a_2)$ denotes the maximum of a_1 and a_2.

Let X be a discrete random variable on a sample space Ω, so that $X \colon \Omega \to \mathbb{R}$. Let $\mathbf{P}$ be a probability law on Ω. Let $x \in \mathbb{R}$. Let $A \subseteq \Omega$. Let Y be another discrete random variable

$p_X(x) = \mathbf{P}(X = x) = \mathbf{P}(\{\omega \in \Omega \colon X(\omega) = x\}), \forall\, x \in \mathbb{R}$

the Probability Mass Function (PMF) of X

$\mathbf{E}(X)$ denotes the expected value of X

$\mathrm{var}(X) = \mathbf{E}(X - \mathbf{E}(X))^2$, the variance of X

$\sigma_X = \sqrt{\mathrm{var}(X)}$, the standard deviation of X

$X|A$ denotes the random variable X conditioned on the event A.

$\mathbf{E}(X|A)$ denotes the expected value of X conditioned on the event A.

$1_A \colon \Omega \to \{0,1\}$, denotes the indicator function of A, so that

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

Let X, Y be a continuous random variables on a sample space Ω, so that $X, Y \colon \Omega \to \mathbb{R}$. Let $-\infty \leq a \leq b \leq \infty$, $-\infty \leq c \leq d \leq \infty$. Let $\mathbf{P}$ be a probability law on Ω. Let $A \subseteq \Omega$.

$f_X \colon \mathbb{R} \to [0,\infty)$ denotes the Probability Density Function (PDF) of X, so

$$\mathbf{P}(a \leq X \leq b) = \int_a^b f_X(x)dx$$

$f_{X,Y} \colon \mathbb{R} \to [0,\infty)$ denotes the joint PDF of X and Y, so

$$\mathbf{P}(a \leq X \leq b, c \leq Y \leq d) = \int_c^d \int_a^b f_{X,Y}(x,y)dxdy$$

$f_{X|A}$ denotes the Conditional PDF of X given A

$\mathbf{E}(X|A)$ denotes the expected value of X conditioned on the event A.

Let X be a random variable on a sample space Ω, so that $X \colon \Omega \to \mathbb{R}$. Let $\mathbf{P}$ be a probability law on Ω. Let $x \in \mathbb{R}$.

$$F_X(x) = \mathbf{P}(X \leq x) = \mathbf{P}(\{\omega \in \Omega \colon X(\omega) \leq x\})$$

the Cumulative Distibution Function (CDF) of X.

UCLA Department of Mathematics, Los Angeles, CA 90095-1555
E-mail address: heilman@math.ucla.edu