Mathematical Background on Probability Theory

Itay Hazan

December 24, 2017

Contents

1 Basic Concepts

In the general case, a *sample space* is a pair (Ω, p), where Ω is a non-empty set and $p\colon 2^{\Omega} \to [0,1]$ is a function called the *probability measure* which satisfies two properties: (1) $p(\Omega) = 1$, and (2) $p\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} p(A_i)$ for any countable collection $\{A_i\}_{i=1}^{\infty}$ of pairwise disjoint sets. Since we will mostly be interested in *finite* probability spaces, we may use the following alternative definition: a *finite* sample space

is a finite set Ω along with a function $p\colon \Omega \to [0,1]$ that satisfies $\sum_{\omega\in\Omega} p(\omega) = 1$. Intuitively, $p(\omega)$ is the probability that a random element sampled from Ω is ω.

An *event* is a subset $A \subseteq \Omega$, and the probability that event A occurs is $\Pr[A] = \sum_{\omega\in A} p(\omega)$, that is, $\Pr[A]$ is the probability that a random element sampled from Ω will be one of the elements in A.

1.1 Inclusion/Exclusion Principle

Let A and B be two disjoint events. From the definition of Pr, we have that

$$\Pr[A \cup B] = \sum_{\omega\in A\cup B} p(\omega) = \sum_{\omega\in A} p(\omega) + \sum_{\omega\in B} p(\omega) = \Pr[A] + \Pr[B],$$

i.e. for disjoint A and B, the probability that a random element from Ω is in the set $A \cup B$ is the probability that a random element from Ω is either in A or in B which is the sum of probabilities $\Pr[A] + \Pr[B]$. However, if A and B are not disjoint, then the sum $\Pr[A] + \Pr[B]$ counts the elements in $A \cap B$ twice, and so we have to subtract them once, i.e.

$$\Pr[A \cup B] = \sum_{\omega\in A\cup B} p(\omega) = \sum_{\omega\in A\setminus B} p(\omega) + \sum_{\omega\in B\setminus A} p(\omega) + \sum_{\omega\in A\cap B} p(\omega) = \sum_{\omega\in A} p(\omega) + \sum_{\omega\in B} p(\omega) - \sum_{\omega\in A\cap B} p(\omega) = \Pr[A] + \Pr[B] - \Pr[A \cap B].$$

This is a simple form of the inclusion/exclusion principle in probability.

Claim 1.1 (Inclusion/Exclusion Principle)**.** *For a set of events $A_1, \ldots, A_n$,*

$$\Pr\left[\bigcup_{i=1}^{n} A_i\right] = \sum_{i=1}^{n} \Pr[A_i] - \sum_{1\le i<j\le n} \Pr[A_i \cap A_j] + \cdots + (-1)^{n-1} \Pr[A_1 \cap \cdots \cap A_n].$$

It is relatively easy to see that every summand is at most as large as its preceding summand, i.e.,

$$\sum_{i=1}^{n} \Pr[A_i] \ge \sum_{1\le i<j\le n} \Pr[A_i \cap A_j] \ge \sum_{1\le i<j<k\le n} \Pr[A_i \cap A_j \cap A_k] \ge \cdots \ge \Pr[A_1 \cap \cdots \cap A_n].$$

Therefore, if we take only the first k summands of the right-hand side of the inclusion/exclusion principle, we achieve an upper bound on the right-hand side for odd k, and a lower bound for even k. Two important examples, presented hereafter, are for $k \in \{1, 2\}$.

Corollary 1.2 (Union bound)**.** *For any set of events $A_1, \ldots, A_n$,*

$$\Pr\left[\bigcup_{i=1}^{n} A_i\right] \le \sum_{i=1}^{n} \Pr[A_i].$$

Corollary 1.3 (Bonefforni inequality)**.** *For any set of events $A_1, \ldots, A_n$,*

$$\Pr\left[\bigcup_{i=1}^{n} A_i\right] \ge \sum_{i=1}^{n} \Pr[A_i] - \sum_{1\le i<j\le n} \Pr[A_i \cap A_j].$$

1.2 Conditional Probability and Independence

Suppose we have two events, A and B, and suppose we know that A occurred. What is the probability that B also occurred given that? We denote this probability by $\Pr[B \mid A]$, and read *the probability of B conditioned on (or given) A*. Let us restate what we are looking for: we are asking for the probability that a randomly sampled $\omega \in \Omega$ is in B given that $\omega \in A$. To understand this measure, we consider several examples (intuitively), figuratively presented in Figure 1:

1. If $A \cap B = \emptyset$, then clearly any ω that is known to be in A cannot be in B, and hence $\Pr[B \mid A] = 0$.

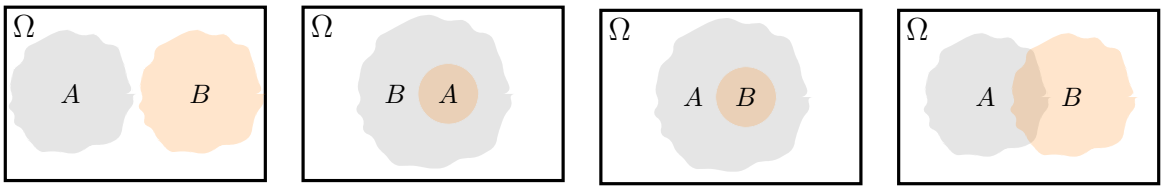
2. If $A \subseteq B$, then any ω that is known to also be in A is also in B, and hence $\Pr[B \mid A] = 1$.

3. If $B \subseteq A$, then the probability that a random $\omega \in A$ will be in B is the relative weight of B in A, i.e. $\Pr[B \mid A] = \frac{\Pr[B]}{\Pr[A]}$.

4. Finally, consider the case in which all of the previous cases do not hold, i.e. $B \nsubseteq A, A \nsubseteq B, B \cap A \neq \emptyset$. In that case, we may restate our question in following alternative form: what is the probability that a random $\omega \in A$ is also in $A \cap B$? This will be the relative weight of $A \cap B$ in A, i.e. $\Pr[B \mid A] = \frac{\Pr[A \cap B]}{\Pr[A]}$.

Figure 1: Four cases of conditional probability

(a) $\Pr[B \mid A] = 0$ (b) $\Pr[B \mid A] = 1$ (c) $\Pr[B \mid A] = \frac{\Pr[B]}{\Pr[A]}$ (d) $\Pr[B \mid A] = \frac{\Pr[A \cap B]}{\Pr[A]}$



Considering all four cases above, if we know that an event A already happened, this reduces the space from Ω to $\Omega \cap A$, where we need to scale the probabilities by $1/\Pr[A]$. We conclude the intuitive discussion above with the following definition:

Definition 1.4. Let A, B be two events such that $\Pr[A] > 0$. The probability of B *conditioned* on A is

$$\Pr[B \mid A] = \frac{\Pr[A \cap B]}{\Pr[A]}.$$

This definition gives us a way of computing the probabilities of events. Suppose we want to compute the probability that two events, A and B occur together, that is $\Pr[A \cap B]$. This may be a difficult task in the general case. However, if break that into steps, first computing $\Pr[A]$ and then $\Pr[B \mid A]$, then by multiplying these probabilities we get the desired probability.

A simple theorem worth mentioning in this context is *Bayes' theorem*.

Theorem 1.5 (Bayes' Theorem)**.** *Let A, B be two events such that $\Pr[A] > 0$. Then,*

$$\Pr[B \mid A] = \frac{\Pr[A \mid B] \Pr[B]}{\Pr[A]}.$$

Bayes' theorem allows us to replace the computation of $\Pr[B \mid A]$ by the computation of $\Pr[A \mid B]$, which may sometimes be easier to compute.

Example. A pregnancy test errs with probability $\frac{1}{100}$, that is, it produces false positives (telling a non-pregnant woman she is pregnant) with probability $\frac{1}{100}$ and false negatives (telling a pregnant woman she is not pregnant) with probability $\frac{1}{100}$. Suppose you know that a randomly selected woman in your city is pregnant with probability $\frac{1}{200}$. What is the probability that a randomly selected woman with a positive pregnancy test is pregnant?

Solution. Let us make the following notations: Preg is the event that the selected woman is pregnant and Pos is the event that the selected woman tested positive. From Bayes' theorem,

$$\begin{aligned}
\Pr[\text{Preg} \mid \text{Pos}] &= \frac{\Pr[\text{Pos} \mid \text{Preg}] \Pr[\text{Preg}]}{\Pr[\text{Pos}]} \\
&= \frac{\Pr[\text{Pos} \mid \text{Preg}] \Pr[\text{Preg}]}{\Pr[\text{Pos} \mid \text{Preg}] \Pr[\text{Preg}] + \Pr[\text{Pos} \mid \neg\text{Preg}] \Pr[\neg\text{Preg}]} \\
&= \frac{\frac{99}{100}\frac{1}{200}}{\frac{99}{100}\frac{1}{200} + \frac{1}{100}\frac{199}{200}}
\end{aligned}$$

$$= \frac{99}{298} \approx 0.33$$

This means that the probability that a random woman with a positive test is pregnant is 33%, which is rather disappointing...

Next, we wish to define the notion of *independence* of events. Intuitively, two events are independent if knowing that one event occurred does not affect the probability of the other one to occur, i.e. we would like our definition to satisfy $\Pr[A \mid B] = \Pr[A]$ and $\Pr[B \mid A] = \Pr[B]$. We thus define independence as follows:

Definition 1.6 (Independence)**.** Two events, A and B, are independent if $\Pr[A \cap B] = \Pr[A] \cdot \Pr[B]$.

It is easy to see that this definition implies both properties mentioned above.

Example. We pick a number $x \in [10]$ uniformly at random. Let A be the event that the chosen number is less than 7, and let B be the event the the chosen number if even. We therefore have that $A = \{1, \ldots, 6\}$, $B = \{2, 4, 6, 8, 10\}$, and $A \cap B = \{2, 4, 6\}$; namely, $\Pr[A] = 0.6$, $\Pr[B] = 0.5$, and $\Pr[A \cap B] = 0.3$, which means that A and B are independent (even if it goes against your intuition!).

A very common mistake is to confuse independence with disjointness. Recall that the notion of independence tells us that knowing that A occurred does not provide information on whether B occurred, and vice verse. However, if A and B are disjoint, then we know that they cannot occur together, and hence knowing that A occurred provides a lot of information on B — we can *determine* that it did not occur! Formally, it is easy to prove that if A and B are disjoint then they are not independent.

Independence and Set Theory Operations. We know that independence and intersection are closely related (by definition). Using De Morgan's laws, it is relatively easy to prove a connection between independence and *union*, by proving that if A and B are independent, then $\Pr[A \cup B] = 1 - (1 - \Pr[A])(1 - \Pr[B])$. Another simple exercise is showing that independence is closed under complement, i.e. if (A, B) is a pair of independent random variables, then so are $(A, \neg B)$, $(\neg A, B)$, and $(\neg A, \neg B)$.

We conclude this discussion by stating two useful definitions:

Definition 1.7. The events $A_1, \ldots, A_n$ are *mutually independent* if

$$\Pr\left[\bigcap_{i \in S} A_i\right] = \prod_{i \in S} \Pr[A_i]$$

for any subset $S \subseteq [n]$. We say that the events are k-wide independent (for $2 \leq k \leq n$) if the property above holds for any subset $S \subseteq [n]$ of size $|S| \leq k$.

2 Random Variables

In the general case, a *random variable* (in short, rv) is a mapping X from Ω to some measurable space. We will mostly be interested in the case where the range is the measurable space $\mathbb{R}$; namely, in our case, a random variable is a function $X \colon \Omega \to \mathbb{R}$. That is, X maps every event $\omega \in \Omega$ to some real value $\mathbb{R}$. Intuitively, if we conduct some experiment, the "raw" outcome is an event $\omega \in \Omega$ and $X(\omega)$ represents some measurable property of that event.

Example. Suppose we roll three fair dice. The raw outcome of this experiment is the actual three outcomes, e.g. $(2, 4, 1)$ or $(5, 4, 5)$. We can define infinitely many different random variables on this probability space, e.g. $X(\omega)$ is the number of fours in the outcome, or $Y(\omega)$ is the sum of outcomes.

The following example is a bit more interesting, and more related to our course:

Example *(Properties of $G(n,p)$).* Suppose we have an empty graph on n vertices, and we conduct the following experiment: for every pair of vertices, we toss a p-biased coin (Heads with probability p and Tails with probability $1-p$), and add the edge between them if and only if the result was Heads. The result is a random graph G whose distribution is usually denoted $G(n,p)$. We can now define many random variables on the probability space $G(n,p)$, e.g. $\chi(G)$ is the chromatic number of G, $\alpha(G)$ is the size of the largest independent set in G, $\omega(G)$ is the size of the largest clique is G, and so on. In fact, understanding different properties of $G(n,p)$ is what the study of random graphs tries to achieve!

Given a random variable X and a measurable set of reals $A \subseteq \mathbb{R}$, we define

$$\Pr[X \in A] \triangleq \Pr\left[\{\omega \in \Omega \colon X(\omega) \in A\}\right],$$

and given $r \in \mathbb{R}$, we define $\Pr[X = r] \triangleq \Pr[X \in \{r\}]$. Using this definition, we say that two random variables X and Y are independent if the events $X \in A$ and $Y \in B$ are independent for every measurable $A, B \subseteq \mathbb{R}$. In particular, if X, Y are independent then $\Pr[X = x, Y = y] = \Pr[X = x] \Pr[Y = y]$.

Definition 2.1 (CDF)**.** The *cumulative distribution function* (CDF) of X is a function, usually denoted $F_X \colon \mathbb{R} \to [0,1]$, which is given by

$$F_X(x) \triangleq \Pr(X \leq x).$$

From its definition, the following properties of the CDF are immediate:

- $0 \leq F_X(x) \leq 1$ for any $x \in \mathbb{R}$.
- $\lim_{x \to -\infty} F_X(x) = 0$.
- $\lim_{x \to +\infty} F_X(x) = 1$
- $F_X(a) \leq F_X(b)$ for any $a \leq b$.
- $\Pr[a < X \leq b] = F_X(b) - F_X(a)$ for any $a \leq b$.
- F_X is left-continuous: for every $a \in \mathbb{R}$, $\lim_{x \to a-} F_X(x) = F_X(a)$.

Note that F_X is not necessarily right-continuous. For example, if $X \equiv 0$ then $F_X(0) = 1$ but $F_X(x) = 0$ for all $x > 0$. A point which has positive probability is called an *atom*. A random variable without atoms is called *continuous*, and one composed only of atoms is called *discrete*.

2.1 Expectation, Variance, and Moments

The *expected value* of a random variable, sometimes also called its *mean value*, represents its average outcome.

Definition 2.2 (Expectation)**.** Let X be a random variable with range $R_X \subseteq \mathbb{R}$. The expected value of X, denoted by $\mathbb{E}[X]$ is

$$\mathbb{E}[X] \triangleq \sum_{x \in R_X} x \Pr[X = x].$$

If R_X is countably infinite, then the expectation is defined only if the sum converges. Note that if R_X is uncountably infinite, then the summation changes to a (Lebesgue) integral.

The intuition behind the notion of expectation is that if you repeat a random experiment a very large number of times and take the average of the observed data, the average becomes roughly $\mathbb{E}[X]$. Furthermore, the more you repeat your experiment, the closer the average is to the expectation. This intuition is formulated in the law of large numbers, explained in subsubsection 3.2.2.

One of the most important and useful properties of expectation is that it is linear, as stated in the following theorem:

Theorem 2.3 (Linearity of Expectation)**.** *Let X and Y be two random variables (not necessarily independent) and let $\alpha, \beta \in \mathbb{R}$. Then,*

$$\mathbb{E}[\alpha X + \beta Y] = \alpha \mathbb{E}[X] + \beta \mathbb{E}[Y].$$

Proof. We focus our proof on the discrete (or countably infinite) case. The proof for the uncountably infinite case is similar. We shall prove that $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$, and we leave if for the reader to prove that $\mathbb{E}[\alpha X] = \alpha \mathbb{E}[X]$.

$$\begin{aligned}
\mathbb{E}[X+Y] &= \sum_{t \in \mathbb{R}} t \Pr[X+Y=t] \\
&= \sum_{x \in \mathbb{R}} \sum_{y \in \mathbb{R}} (x+y) \Pr[X=x, Y=y] \\
&= \sum_{x \in \mathbb{R}} x \sum_{y \in \mathbb{R}} \Pr[X=x, Y=y] + \sum_{y \in \mathbb{R}} y \sum_{x \in \mathbb{R}} \Pr[X=x, Y=y] \\
&= \sum_{x \in \mathbb{R}} x \Pr[X=x] + \sum_{y \in \mathbb{R}} y \Pr[Y=y] \\
&= \mathbb{E}[X] + \mathbb{E}[Y].
\end{aligned}$$

□

The following corollary can be easily proven using a simple inductive argument:

Corollary 2.4. *Let $X_1, \ldots, X_n$ be* any *set of random variables (not necessarily independent), and let $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$. Then,*

$$\mathbb{E}\left[\sum\nolimits_{i=1}^{n} \alpha_i X_i\right] = \sum\nolimits_{i=1}^{n} \alpha_i \mathbb{E}[X_i].$$

Another type of useful arguments that involve expectation is *averaging* arguments.

Claim 2.5 (Averaging)**.** *Let X be some random variable with range R_X and with finite expected value $\mathbb{E}[X]$. Then there exists $a \in R_X$ such that $a \geq \mathbb{E}[X]$.*

Proof. Assume towards a contradiction that $a < \mathbb{E}[X]$ for all $a \in R_X$. Then,

$$\mathbb{E}[X] = \sum_{x \in R_X} x \Pr[X=x] < \sum_{x \in R_X} \mathbb{E}[X] \Pr[X=x] = \mathbb{E}[X] \cdot \sum_{x \in R_X} \Pr[X=x] = \mathbb{E}[X],$$

which is a contradiction. □

An example for a nice and simple averaging argument is given in subsubsection 2.2.3.

Claim 2.6. *Let $X_1, \ldots, X_n$ be two independent random variables. Then,*

$$\mathbb{E}[XY] = \mathbb{E}[X] \cdot \mathbb{E}[Y].$$

Proof.

$$\begin{aligned}
\mathbb{E}[XY] &= \sum_{x \in \mathbb{R}} \sum_{y \in \mathbb{R}} (xy) \Pr[X=x, Y=y] \\
&= \sum_{x \in \mathbb{R}} \sum_{y \in \mathbb{R}} (xy) \Pr[X=x] \Pr[Y=y] \\
&= \left(\sum_{x \in \mathbb{R}} x \Pr[X=x]\right) \cdot \left(\sum_{y \in \mathbb{R}} y \Pr[Y=y]\right) \\
&= \mathbb{E}[X] \cdot \mathbb{E}[Y].
\end{aligned}$$

□

In the following claim, the *law of total expectation* is stated. It is also commonly known as the *smoothing theorem*, and is very useful for computing the expectation of a random variable X. The proof is left for the reader.

Claim 2.7 (Law of Total Expectation)**.** *Let X be a random variable with finite expectation. Then for any random variable Y,*

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

Here the outer expectation is with respect to Y, and the inner expectation is with respect to X *given the value of* Y. For example, suppose that Y is supported on the non-negative integers, and let $X = X_1 + \cdots + X_Y$, where $X_1, X_2, \ldots$ are infinitely many i.i.d. copies of some random variable Z (i.i.d. stands for independent and identically distributed). Then

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X \mid Y]] = \mathbb{E}[Y\,\mathbb{E}[Z]] = \mathbb{E}[Y]\,\mathbb{E}[Z].$$

Some more basic properties of expectation (left for the reader to verify):

1. $\mathbb{E}[g(X)] = \sum_{x \in R_X} g(x) \Pr[X = x]$ for any function g.
2. If $X = 0$ then $\mathbb{E}[X] = 0$.
3. If $\mathbb{E}[X]$ exists then so does $\mathbb{E}[|X|]$.[1]
4. If $X \geq 0$ then $\mathbb{E}[X] > 0$.
5. If $X \geq Y$ then $\mathbb{E}[X] \geq \mathbb{E}[Y]$.

The next quantitative measure we discuss is the *variance* of a random variable, which tries to measure how spread out are the possible outcomes from the expectation. To understand that measure, we consider two random variables: X is a random variable that equals 49 with probability $1/2$ and 51 with probability $1/2$, and Y is a random variable that equals 0 with probability $1/2$ and 100 with probability $1/2$. Clearly, both random variables average at 50. However, a random sample from X will be far "closer" to the expectation than a random sample from Y. In that sense, a random sample from X *deviates* very little from the expectation, while a random sample from Y deviates a lot from the expectation. The notion of variance is formalized in the following definition:[2]

Definition 2.8 (Variance and Standard Deviation)**.** Let X be a random variable with finite expected value $\mathbb{E}[X]$. The *variance* of X, denoted $\mathbb{V}[X]$, is the average squared distance of X from its expectation, where distance is measured in L^2 norm; namely,

$$\mathbb{V}[X] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right].$$

The *standard deviation* of X is defined to be $\mathbb{SD}[X] = \sqrt{\mathbb{V}[X]}$.

The following claim provides a (sometimes) simpler way of calculating the variance of a random variable:

Claim 2.9 (Alternative Definition)**.** *For any random variable X with finite expected value $\mathbb{E}[X]$,*

$$\mathbb{V}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2.$$

Proof.

$$\mathbb{V}[X] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}\left[X^2 - 2X\,\mathbb{E}[X] + \mathbb{E}[X]^2\right] = \mathbb{E}[X^2] - 2\,\mathbb{E}[X]\,\mathbb{E}[X] + \mathbb{E}[X]^2 = \mathbb{E}[X^2] - \mathbb{E}[X]^2,$$

where the equalities hold since $\mathbb{E}[\mathbb{E}[X]] = \mathbb{E}[X]$. □

Some basic properties of variance are presented hereafter (left for the reader to verify):

1. $\mathbb{V}[X + a] = \mathbb{V}[X]$ for any constant $a \in \mathbb{R}$.
2. $\mathbb{V}[aX] = a^2\,\mathbb{V}[X]$ for any constant $a \in \mathbb{R}$.

[1] This technical property is a feature of the Lebesgue integral.

[2] Another, perhaps more intuitive, measure of concentration is $\mathbb{E}[|X - \mathbb{E}[X]|]$. However this measure is much harder to work with, and isn't additive for independent variables (i.e., doesn't satisfy $\mathbb{V}[X+Y] = \mathbb{V}[X] + \mathbb{V}[Y]$ for independent X, Y).

3. $\mathbb{V}[X] = 0$ if and only if X is a constant random variable (that is, $\Pr[X = a] = 1$ for some a).

Another two important measures are *covariance* and *correlation*, that reflect how two random variables relate to one another:

Definition 2.10 (Covariance and Correlation)**.** Let X and Y be random variables with finite expectancies and variances. The *covariance* of X and Y is defined to be

$$\mathbb{C}\text{ov}[X, Y] \triangleq \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])].$$

The correlation between X and Y is their normalized covariance, that is,

$$\rho[X, Y] \triangleq \frac{\mathbb{C}\text{ov}[X, Y]}{\sqrt{\mathbb{V}[X]\,\mathbb{V}[Y]}}.$$

X and Y are said to be *uncorrelated* if $\mathbb{C}\text{ov}[X, Y] = \rho[X, Y] = 0$.

To understand what the covariance between two random variables represents, consider the case of two *dependent* random variables, X and Y. Intuitively, if X and Y are dependent, then if X equals some value, it somehow "affects" the value of Y. Therefore, if we know that X deviates from its average by $(X - \mathbb{E}[X])$, then it must somehow "affect" how much Y deviates from its average, that is $(Y - \mathbb{E}[Y])$. The notion of covariance tries to quantify this intuition.

From the intuition above, one would expect that the covariance of two *independent* random variables would be 0, and that is, in fact, true; if X and Y are independent then they are uncorrelated. However, this does not work the other way: if X and Y are uncorrelated, it does not necessarily mean that they are independent, as suggested by the example below.

Example. For an example involving discrete random variables, let (X, Y) be a pair of random variables whose joint distribution is the following: $(X, Y) = (-1, 1)$ with probability $1/4$, $(X, Y) = (0, -1)$ with probability $1/2$, and $(X, Y) = (1, 1)$ with probability $1/4$. Clearly, X and Y are dependent, yet a simple calculation proves that $\mathbb{C}\text{ov}[X, Y] = 0$.
For an example involving continuous random variables, let X be a uniformly chosen number from the range $[-1, 1]$, and let $Y = X^2$. Again, it can be easily shown that X and Y are uncorrelated yet dependent.

The notion of correlation normalizes the covariance of two random variables: while the covariance is unbounded, correlation is always in the range $[-1, 1]$. Furthermore, it is fairly easy to prove that $\rho[X, Y] = \pm 1$ if and only if there exist $a \in \mathbb{R}^+$ and $b \in \mathbb{R}$ such that $X = \pm aY + b$. Observe that covariance and correlation, as opposed to variance, may have negative sign. That happens if $(X - \mathbb{E}[X])(Y - \mathbb{E}[Y]) < 0$ in the common case, that is, if "usually" (used informally) when one random variable is larger than its expectation then the other is smaller than its expectation.

Claim 2.11 (Variance of Sum)**.** *Let X and Y be random variables with finite expectancies and variances. Then,*

$$\mathbb{V}[X \pm Y] = \mathbb{V}[X] + \mathbb{V}[Y] \pm 2\,\mathbb{C}\text{ov}[X, Y].$$

Therefore, if X and Y are independent then $\mathbb{V}[X \pm Y] = \mathbb{V}[X] + \mathbb{V}[Y]$.

Proof.

$$\begin{aligned}
\mathbb{V}[X \pm Y] &= \mathbb{E}\left[((X \pm Y) - \mathbb{E}[X \pm Y])^2\right] \\
&= \mathbb{E}\left[X^2 \pm 2XY + Y^2 - 2(X + Y)\,\mathbb{E}[X + Y] + \mathbb{E}[X]^2 \pm 2\,\mathbb{E}[X]\,\mathbb{E}[Y] + \mathbb{E}[Y]^2\right] \\
&= \mathbb{E}[X^2] - \mathbb{E}[X]^2 + \mathbb{E}[Y^2] - \mathbb{E}[Y]^2 \pm 2\,\mathbb{E}[XY] \mp 2\,\mathbb{E}[X]\,\mathbb{E}[Y] \\
&= \mathbb{V}[X] + \mathbb{V}[Y] \pm 2\,\mathbb{C}\text{ov}[X, Y].
\end{aligned}$$

□

If X, Y are dependent random variables then we often think of (X, Y) as a distribution on pairs of real numbers called a *joint distribution*. Given a joint distribution, the distribution of the individual coordinates are known as *marginal distributions* or *marginals*. We discuss these notions further in subsection 2.3.

We conclude this section with a brief discussion on *moments*:

Definition 2.12 (Moments)**.** Let X be a random variable with finite expectation $\mathbb{E}[X]$. The *n-th moment* of X is defined to be $\mathbb{E}[X^n]$, and the *n-th normalized* or *central moment* is defined to be

$$\mu_n \triangleq \mathbb{E}\left[(X - \mathbb{E}[X])^n\right].$$

Generally speaking, moments give us information about the distribution *shape*. As we said earlier, the first moment (mean) tells us where is the center of the distribution. The second central moment (variance) tells us how concentrated the distribution is around its mean value. The third central moment, called *skewness*, tells us whether the distribution is symmetric around its mean value; if it is, then the skewness is zero. Otherwise, the distribution "leans to the right" (the tail is longer to the left) if the skewness is negative, and to the left if the skewness is positive. The fourth central moment, called *kurtosis*, is a measure of the heaviness of the tail; if the distribution has heavy tails, the kurtosis will be high, and otherwise, it would be low. (In practice the skewness and kurtosis are further normalized by appropriate powers of the standard deviation.)

Since moments provide information on the shape of the distribution, one would expect that two random variables that have the same distribution will have the same moments, and vice verse. This is, in fact, true. To formalize this intuition, we define *moment generating functions*.

Definition 2.13 (Moment Generating Function)**.** The moment generating function of a random variable X is defined to be

$$M_X(t) \triangleq \mathbb{E}\left[e^{tX}\right],$$

if the expectation exists.

The reason MGFs are useful is that they "encode" all of the moments of X in one function; it can be shown that

$$\mu_n = \frac{d^n}{dt^n} M_X(t)\Big|_{t=0}$$

Therefore, if two random variables have the same MGF, then they have the same moments, and hence the same distribution. Formally,

Claim 2.14. *Let X and Y be two random variables, and assume that $M_X(t)$ and $M_Y(y)$ exist. Then, $M_X(t) = M_Y(y)$ if and only if X and Y have the same distribution.*

2.2 Important Distributions

A note to the reader: To keep the review as concise as possible, the uniform, normal and exponential distributions are the only *continuous* distributions we present (since these will be the only ones you might see in your lectures or homework). Furthermore, in this entire review, we rarely discuss any notion of continuous distribution, and we even abstain from defining the most basic notions, such as *density functions*. We encourage you to read more on continuous random variables and other types of continuous distributions.

2.2.1 (Discrete) Uniform Distribution

Let $S \subseteq \mathbb{R}$. A random variable X is said to be *uniformly distributed* on S, and we denote $X \sim \text{Uni}(S)$ or $X \sim U(S)$, if $\Pr[X = s] = 1/|S|$ for every $s \in S$. In other words, the distribution $\text{Uni}(S)$ represents a "fair" random pick of an element in S. A simple example is the outcome of rolling a fair six-faced die, which is uniformly distributed on the set $[6]$; namely, if we denote the outcome by X then $X \sim \text{Uni}([6])$.

Expectation. The expectation of X is given by the (standard) average of the elements in S, namely, $\mathbb{E}[X] = \frac{1}{|S|}\sum_{s \in S} s$. An important special case is if S is a range of consecutive numbers, i.e. $S = \{a, a+1, \ldots, b\}$. In that case, $\mathbb{E}[X] = \frac{a+b}{2}$.

Variance. The variance of X can be easily calculated by definition:

$$\mathbb{V}[X] \;=\; \mathbb{E}[X^2] \;-\; \mathbb{E}[X]^2 \;=\; \frac{1}{|S|}\sum_{s\in S} s^2 \;-\; \left(\frac{1}{|S|}\sum_{s\in S} s\right)^2 \;=\; \frac{1}{|S|^2}\left((|S|-1)\sum_{s\in S} s^2 - 2\sum_{s<r} sr\right)$$

Giving a nicer-looking expression would require more information on the identity of S. For instance, if $S = \{a, a+1, \ldots, b\}$, then the expression above becomes $\mathbb{V}[X] = \frac{|S|^2-1}{12}$.

2.2.2 (Continuous) Uniform Distribution

The simplest and most common continuous uniform distribution is the uniform distribution over an interval $[a, b]$. A random variable X is said to be *uniformly distributed* over $[a, b]$, and we denote $X \sim \mathrm{Uni}([a, b])$ or $X \sim U([a, b])$ or $X \sim U(a, b)$, if $\Pr[X < a] = 0$, $\Pr[X < b] = 1$, and for $t \in [a, b]$, $\Pr[X < t] = \frac{t-a}{b-a}$. A simple example is an alternative way of defining the $G(n, p)$ random model: for each edge we choose a weight $w(e) \sim \mathrm{Uni}([0, 1])$, and the random graph contains all edges whose weights are smaller than p.

Expectation. The formulas above show that $\Pr[t < X < t + dt] = \frac{dt}{a-b}$ when $a < t < b$ (this is the *density* of the random variable), and so

$$\mathbb{E}[X] = \int_a^b \frac{t\,dt}{b-a} = \left.\frac{t^2}{2(b-a)}\right|_a^b = \frac{b^2-a^2}{2(b-a)} = \frac{a+b}{2},$$

an intuitive formula which also follows from symmetry considerations: if $X \sim \mathrm{Uni}([a, b])$ then $a+b-X \sim \mathrm{Uni}([a, b])$ as well, and so $a + b = \mathbb{E}[X + (a + b - X)] = 2\,\mathbb{E}[X]$.

Variance. We can similarly calculate the variance:

$$\mathbb{V}[X] = \int_a^b \frac{(t-\frac{a+b}{2})^2\,dt}{b-a} = \left.\frac{(t-\frac{a+b}{2})^3}{3(b-a)}\right|_a^b = \frac{2(\frac{b-a}{2})^3}{3(b-a)} = \frac{(b-a)^2}{12}.$$

Compare this to the variance of the uniform distribution over $\{a, \ldots, b\}$, which is $\frac{(b-a+1)^2-1}{12}$.

2.2.3 Bernoulli Distribution and Indicators

The Bernoulli distribution represents the distribution of a single yes/no experiment. Formally, given $p \in [0, 1]$, X is distributed as a Bernoulli random variable with parameter p, and denoted $X \sim \mathrm{Ber}(p)$, if $X = 1$ with probability p and 0 with probability $1 - p$. The simplest example is the outcome of tossing a fair coin, which is $\mathrm{Ber}(1/2)$.

Indicators are important special cases of Bernoulli random variables. As its name suggests, an indicator of an event $A \subseteq \Omega$ is a random variable that whose value is 1 if A occurs and 0 if A does not occur. We usually denote the indicator of the event A by $\mathbf{1}_A$, and hence

$$\mathbf{1}_A = \begin{cases} 1 & A \text{ occurs} \\ 0 & A \text{ does not occur} \end{cases} = \begin{cases} 1 & \text{w.p. } \Pr[A] \\ 0 & \text{w.p. } 1 - \Pr[A] \end{cases} \sim \mathrm{Ber}\left(\Pr[A]\right).$$

Note that for any two events A, B, $\mathbf{1}_A \cdot \mathbf{1}_B = \mathbf{1}_{A\cap B}$.

Expectation and Variance. The expectation of a Bernoulli random variable is simply p, since $\mathbb{E}[X] = 1 \cdot p + 0 \cdot (1 - p)$. The variance of X is given in a similar manner:

$$\mathbb{V}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[(X - p)^2] = (1 - p)^2 \cdot p + (0 - p)^2 \cdot (1 - p) = p(1 - p).$$

Example. A common use of indicators is for breaking "complex" random variables into their smaller parts. In this example we shall prove that any graph G has a cut of size at least $\frac{1}{2}|E|$.
Given a graph $G = (V, E)$, a *cut* is a partition of the vertices into two non-empty disjoint sets, $S, T \subseteq V$. An edge $e = \{u, v\}$ is said to *cross* the cut (S, T) if $u \in S$ and $v \in T$. The size of the cut is the number of edges crossing it.
Let $G = (V, E)$ be any graph. To prove that G has a cut of size at least $\frac{1}{2}|E|$, it suffices to show that the expected size of a randomly chosen cut of G is $\frac{1}{2}|E|$; by averaging, we get that there must *exist* such a cut.
We will choose the cut (S, T) uniformly at random, namely for every vertex $v \in V$, we put it in S with probability $1/2$ and in T with probability $1/2$. Let $E(S, T)$ denote the set of edges crossing the cut (S, T), and let $X \triangleq |E(S, T)|$ denote the size of the resulting cut. For every pair of vertices u, v, we define the indicator $\mathbf{1}_{\{\{u,v\}\in E(S,T)\}}$ that equals 1 if $\{u, v\}$ is a crossing edge and 0 otherwise. Therefore, $X = \sum_{\{u,v\}\in E} \mathbf{1}_{\{\{u,v\}\in E(S,T)\}}$. Since we chose u and v to be in either S or T with probability $1/2$, then the probability that $\{u, v\} \in E(S, T)$ is $\frac{1}{2}$, and hence $\mathbb{E}\left[\mathbf{1}_{\{\{u,v\}\in E(S,T)\}}\right] = \frac{1}{2}$. Applying the linearity of expectation, we get that

$$\mathbb{E}[X] = \sum_{\{u,v\}\in E} \mathbb{E}\left[\mathbf{1}_{\{\{u,v\}\in E(S,T)\}}\right] = \sum_{\{u,v\}\in E} \frac{1}{2} = \frac{1}{2}|E|.$$

The indicators presented above are usually referred to as $\{0, 1\}$-indicators, since their possible outputs are only 0 and 1. To conclude the discussion on the Bernoulli distribution, it may be worth mentioning that there is another type of indicators, called $\{-1, 1\}$-indicators. These may sometimes come in handy and simplify calculations. We shall not elaborate on this type of indicators, but only mention that there is a simple transformation between the two types of indicators. Given a $\{0, 1\}$-indicator X, the random variable $Y \triangleq 2X - 1$ is a $\{-1, 1\}$-indicator that satisfies $X = 1 \Leftrightarrow Y = 1$.

2.2.4 Geometric Distribution

Suppose you conduct a sequence of independent trials. Each trial either succeeds with probability p or fails with probability $1 - p$. When the first trial succeeds, you stop conducting more trials. Let X denote the number of trials you conduct (that is, until the first success). X is said to be distributed *geometrically* with success probability p, and this is denoted by $X \sim \text{Geo}(p)$ or $X \sim G(p)$. Given some $k \in \mathbb{N}$, the probability that $X = k$ is:

$$\Pr[X = k] = \Pr\left[\left(\begin{array}{l}\text{the } k\text{th trial}\\ \text{succeeded}\end{array}\right) \wedge \bigwedge_{i=1}^{k-1} \left(\begin{array}{l}\text{the } i\text{th trial}\\ \text{failed}\end{array}\right)\right] =$$

$$\Pr\left[\begin{array}{l}\text{the } k\text{th trial}\\ \text{succeeded}\end{array}\right] \cdot \prod_{i=1}^{k-1} \Pr\left[\begin{array}{l}\text{the } i\text{th trial}\\ \text{failed}\end{array}\right] = p \cdot (1-p)^{k-1}.$$

Memorylessness. A random variable is said to be *memoryless* if $\Pr[X \leq s + k \mid X > s] = \Pr[X \leq k]$. Intuitively, $\Pr[X \leq s + k \mid X > s]$ represents the following scenario: you have already conducted s experiments, all failed, and you ask for the probability that a success will occur in the next k trials. Since every experiment is independent of the other experiments, then the fact that the first s trials failed has no effect on the appearance of a success in the following k trials, and hence this simply equals $\Pr[X \leq k]$. It seems as if X "does not remember" previous failures, and is therefore said to be *memoryless*. In fact, every discrete random variable that satisfies the memorylessness property is geometric! The arguments above can be supported mathematically, but we leave this to the reader.

Expectation. Observe that if $X \sim \text{Geo}(p)$ then X has countably infinite many possible outcomes, and so the mean is a series!

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} k \Pr[X = k] = \sum_{k=1}^{\infty} kp(1-p)^{k-1} = p \sum_{k=1}^{\infty} k(1-p)^{k-1} =$$

$$-p\frac{d}{dp}\sum_{k=1}^{\infty}(1-p)^k = -p\frac{d}{dp}\frac{1}{p} = -p\cdot\frac{-1}{p^2} = \frac{1}{p}.$$

Intuitively, this result "makes sense": if the probability of success is $\frac{1}{3}$, then we would expect that the number of trials until the first success is 3.

Variance. The variance of X can be computed in a manner similar to its expectation. It is left for the reader to verify that $\mathbb{V}[X] = \frac{1-p}{p^2}$.

Example. Suppose you have a fair six-sided die. You toss the die repeatedly until a '4' turns up, and then you stop. In this case, the number of tosses is distributed geometrically with probability $\frac{1}{6}$, and the average number of tosses it would take before getting a '4' is 6.

2.2.5 Binomial Distribution

Suppose you conduct n independent experiments. Each experiment either succeeds with probability p or fails with probability $1-p$. The number of successful experiments is said to be *binomially* distributed on n trials with success probability p, and is denoted $X \sim \text{Bin}(n,p)$. Given some $k \in \{0,1\ldots,n\}$, the probability that $X = k$ is:

$$\Pr[X=k] = \binom{n}{k}p^k(1-p)^{n-k}.$$

Expectation. To compute the expected value of X, we will use indicators! For every $i \in \{1,\ldots,n\}$, let I_i be a $\{0,1\}$-indicator of the event "the ith experiment succeeded". Therefore, $X = \sum_{i=1}^{n} I_i$. Since every I_i is a Bernoulli random variable with expectation p, then

$$\mathbb{E}[X] = \sum_{i=1}^{n}\mathbb{E}[I_i] = np.$$

Variance. The variance of X can be computed in a manner similar to its expectation, using Claim 2.11. It is left for the reader to verify that $\mathbb{V}[X] = np(1-p)$.

Example. Suppose you have a fair coin, and you flip it 100 times, counting the number of Heads that turn up. In this case, the number of Heads is distributed $\text{Bin}(100, \frac{1}{2})$, and its expectation is 50.

2.2.6 Poisson Distribution

Suppose you know that some type of event occurs *independently* and with *constant rate*, e.g. there are 4.3 births per second around the world (according to [1]), and there are 23,607 deaths per year from the flu in the US (according to [2]). The Poisson distribution represents the number of such events in a fixed interval of time, and is used to answer questions like: given that there are 4.3 births per second around the world, what is the probability that 15 children were born in the past two seconds (while you were reading this sentence)?

A random variable $X \sim \text{Po}(\lambda)$ is said to be a Poisson random variable with rate $\lambda > 0$, and X satisfies:

$$\Pr[X=k] = e^{-\lambda}\frac{\lambda^k}{k!}.$$

We think of X as the number of events in a unit of time when the rate is λ events per one unit of time.

Expectation and Variance. From the intuition presented above, one can infer that the expectation of a Poisson random variable with rate λ is λ. This fact can be backed with the proper calculations. Furthermore, one can also prove that the variance of such a random variable is also λ.

Sum of Poissons. Suppose you have two Poisson random variables, $X \sim \text{Po}(\lambda_X)$ that represent the number of arrivals at restaurant X whose rate of arrival is λ_X, and $Y \sim \text{Po}(\lambda_Y)$ that represent the number of arrivals at restaurant Y whose rate of arrival is λ_Y. Consider the random variable $X+Y$, which is the number of arrivals at both restaurants combined. A simple calculation shows that $X+Y \sim \text{Po}(\lambda_X+\lambda_Y)$. This argument also goes the opposite way: if we know that the number of arrivals at both restaurants combined is distributed like $\text{Po}(\lambda_X + \lambda_Y)$, and we know that the number of arrivals a restaurant X is distributed like $\text{Po}(\lambda_X)$, then we can assume that the number of arrivals at restaurant Y is distributed like $\text{Po}(\lambda_Y)$.

Example. Suppose we know that the number of arrivals at a restaurant is a Poisson random variable with rate λ customers per hour. From the assumption that all arrivals are independent, and using summation of Poissons, we get that the number of arrivals at the restaurant in t hours is a $\text{Po}(t\lambda)$ random variable.

2.2.7 Exponential Distribution

The exponential distribution is a *continuous* distribution that represents the time between consecutive arrivals in a Poisson process. Consider a Poisson process in which people arrive at a restaurant with rate λ, and assume that the restaurant is empty. How long before the first customer arrives? Let $T \sim \text{Exp}(\lambda)$ denote the time of arrival of the first customer, and let $Y(t) \sim \text{Po}(\lambda t)$ be the number of people in the restaurant at time t (assume that no one leaves the restaurant).

$$\Pr[T > t] = \Pr[Y(t) = 0] = \frac{\lambda^0}{0!}e^{-\lambda} = e^{-\lambda}.$$

Therefore, the CDF of an exponential random variable with rate λ is $1 - e^{-\lambda}$, and its density function is

$$f_T(t) = \frac{d}{dt}\Pr[T \leq t] = \lambda e^{-\lambda}.$$

This intuition is somewhat analogous to the geometric distribution. As presented in the previous subsection, we can think of a geometric random variable as representing a discrete-time process, in which a random experiment occurs every minute, and we count the number of minutes until the first success (which is when the process terminates). The exponential distribution can be thought of as the continuous-time equivalent of this process, where we measure how long it takes before the first success occurs in a Poisson process with rate λ. Another intuition for the connection between the Poisson and the exponential distributions is presented in subsubsection 3.2.1.

Memorylessness. As presented above, the exponential distribution is, in some sense, the continuous equivalent of the (discrete) geometric distribution. This informal equivalence is further supported by the fact that the exponential distribution also satisfies the memorylessness property: if $T \sim \text{Exp}(\lambda)$ then

$$\Pr[T \leq s + t \mid T > s] = \Pr[T \leq t].$$

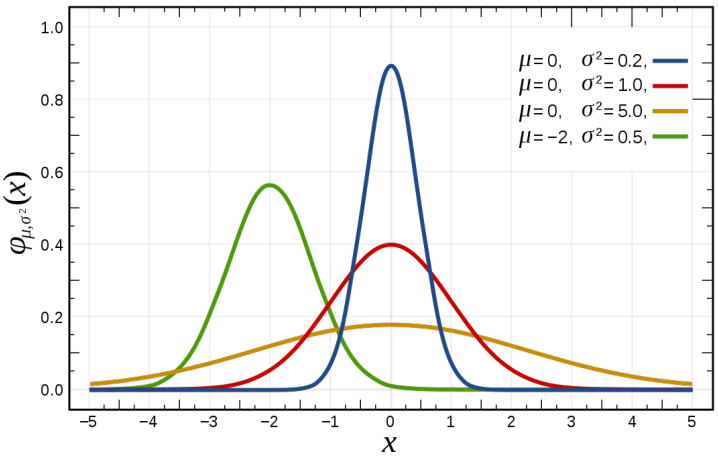
As for the geometric case, it can also be proven that any continuous distribution satisfying the memorylessness property is necessarily an exponential distribution.

MGF. The exponential distribution has a, perhaps surprisingly, simple moment generating function: if $X \sim \text{Exp}(\lambda)$ then $M_X(t) = \frac{\lambda}{\lambda - t}$ for $t < \lambda$. Therefore, a simple calculation shows that the nth moment of X is given by $\mathbb{E}[X^n] = \frac{n!}{\lambda^n}$.

Expectation and Variance. The expectation of an exponential random variable with rate λ can be easily derived from its MGF, giving $\mathbb{E}[X] = \frac{1}{\lambda}$. Similarly, the variance of X is

$$\mathbb{V}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.$$

Figure 2: The normal distribution (from Wikipedia [4])



Exponential Race. Suppose you have two exponential random variables, $X \sim \text{Exp}(\lambda_X)$ and $Y \sim \text{Exp}(\lambda_Y)$. Let $Z = \min(X, Y)$. Intuitively, X represents the time of the first arrival in a Poisson process with rate λ_X, Y represents the time of the first arrival in a Poisson process with rate λ_Y, and Z represents the time of the first arrival in both the processes. It is fairly easy to prove that $Z \sim \text{Exp}(\lambda_X + \lambda_Y)$.

2.2.8 Normal (Gaussian) Distribution

The normal distribution is one of the most common continuous distributions. Due to the *central limit theorem* (discussed in subsubsection 3.2.3), that states that average samples from a large population tend to be normally distributed, the normal distribution naturally appears in statistics and other sciences.

The normal distribution $\text{N}(\mu, \sigma^2)$ takes two parameters: its expectation μ and its variance σ^2 (σ is often used for standard deviation). In the general case, if $X \sim \text{N}(\mu, \sigma^2)$ then its density function is

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

Several examples of different normal distributions are presented in Figure 2. An important special case is $\text{N}(0, 1)$, called the standard normal distribution. In this case, the density function is usually denoted by the small Greek letter phi, $\varphi(x)$.

Some Properties of the Normal Distribution.

- *Symmetry*: The normal distribution is symmetric around its mean.
- *Highly Concentrated*: The normal distribution is highly concentrated around its mean value. It can be proven that roughly 99.7 percent of its weight lies within 3 standard deviations from the mean, i.e. $\Pr[|X - \mu| \leq 3\sigma] \approx 0.997$.
- *Standardization*: It is fairly easy to prove that if $X \sim \text{N}(\mu, \sigma^2)$, then $X + a \sim \text{N}(\mu + a, \sigma^2)$ and $bX \sim \text{N}(b\mu, b^2\sigma^2)$. Therefore, any general normal variable $X \sim \text{N}(\mu, \sigma^2)$ can be "turned into" a standard normal variable as follows: $\frac{X-\mu}{\sigma} \sim \text{N}(0, 1)$.
- *Summation*: If $X \sim \text{N}(\mu_X, \sigma_X^2)$ and $Y \sim \text{N}(\mu_Y, \sigma_Y^2)$ then $X + Y \sim \text{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$.

CDF. By definition, the cumulative distribution function of $X \sim \text{N}(\mu, \sigma^2)$ is

$$F_X(x) = \Pr[X \leq x] = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx.$$

However, this integral cannot be expressed in terms of elementary functions, and is hence unpleasant to work with. To be able to work with it, we turn to the standard normal distribution. The CDF of the

standard normal, usually denoted $\Phi(x)$, has many numerical approximations, and there exist tools such as *normal tables* that help compute Φ up to some precision. We shall not elaborate on such methods, but will mention that using standardization one can apply them on general normal random variables: if $X \sim \mathrm{N}(\mu, \sigma^2)$ then

$$\Pr[X \le x] = \Pr\left[\frac{X-\mu}{\sigma} \le \frac{x-\mu}{\sigma}\right] = \Phi\left(\frac{x-\mu}{\sigma}\right).$$

We conclude our discussion on the normal distribution by presenting several important properties of its CDF. Most of these properties are easily derived from the properties of the normal distribution presented above.

- $\Phi(x) = 1 - \Phi(-x)$.
- $\Pr[a \le X \le b] = \Phi(b) - \Phi(a)$ for a standard normal variable X.
- $\Pr[|X| \le x] = \Pr[-x \le X \le x] = \Phi(x) - \Phi(-x) = 2\Phi(x) - 1$ for a standard normal X.
- $\Pr[X > x] < \frac{1}{x} \cdot \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ for $x > 0$, and the bound is asymptotically tight.

2.3 Coupling

Coupling is a probabilistic technique with a wide variety of applications. The idea is to take two unrelated random variables (or distributions), X and Y, and construct a joint random variable $(\tilde{X}, \tilde{Y})$ such that the marginal distributions correspond to X and Y.

To better understand this technique, we present two applications of coupling.

Coupling of Bernoulli Variables. Let X and Y be independent Bernoulli random variables with parameters $0 \le q < r \le 1$, respectively. One way of coupling these two random variable is to construct the joint random variable $(\tilde{X}, \tilde{Y})$ to be (X, Y). In that case,

$$\begin{bmatrix} \Pr\left[(\tilde{X},\tilde{Y}) = (0,0)\right] & \Pr\left[(\tilde{X},\tilde{Y}) = (0,1)\right] \\ \Pr\left[(\tilde{X},\tilde{Y}) = (1,0)\right] & \Pr\left[(\tilde{X},\tilde{Y}) = (1,1)\right] \end{bmatrix} = \begin{bmatrix} (1-q)(1-r) & (1-q)r \\ q(1-r) & qr \end{bmatrix}.$$

Clearly, the marginal distributions remain the marginal distributions of X and Y. Another (and usually, more useful) way of coupling X and Y is called *monotone coupling*. Here, we define $(\tilde{X}, \tilde{Y})$ as follows: we first sample a uniform value from the range $[0, 1]$, $U \sim \mathrm{Uni}[0, 1]$. We then define $\tilde{X} = \mathbf{1}_{\{U \le q\}}$ and $\tilde{Y} = \mathbf{1}_{\{U \le r\}}$. Clearly, the marginal distributions of the new pair corresponds to X and Y:

$$\begin{bmatrix} \Pr\left[(\tilde{X},\tilde{Y}) = (0,0)\right] & \Pr\left[(\tilde{X},\tilde{Y}) = (0,1)\right] \\ \Pr\left[(\tilde{X},\tilde{Y}) = (1,0)\right] & \Pr\left[(\tilde{X},\tilde{Y}) = (1,1)\right] \end{bmatrix} = \begin{bmatrix} 1-r & r-q \\ 0 & q \end{bmatrix}.$$

Monotone coupling is useful since it ensures that $\tilde{X} \le \tilde{Y}$ (namely, $\tilde{X}$ and $\tilde{Y}$ are dependent).

Monotonicity of Binomial Distribution. This is an example for an application of monotone coupling of Bernoulli variables. Given n, we claim that $\mathrm{Bin}(n, p)$ is monotone in p, i.e.: if $0 \le q < r \le 1$, $X \sim \mathrm{Bin}(n, q)$, and $Y \sim \mathrm{Bin}(n, r)$, then $\Pr[X \ge k] \le \Pr[Y \ge k]$ for any natural $0 \le k \le n$. To prove the theorem, recall that the Binomial distribution can be represented by a sum of mutually independent Bernoulli random variables; let $X_1, \ldots, X_n \sim \mathrm{Ber}(q)$ and $Y_1, \ldots, Y_n \sim \mathrm{Ber}(r)$. Clearly, $X = \sum_{i=1}^{n} X_i$ and $Y = \sum_{i=1}^{n} Y_i$. Now, we use monotone coupling (as seen above) to couple each pair X_i and Y_i. We get two new sets of Bernoulli random variables, $\tilde{X}_1, \ldots, \tilde{X}_n$ and $\tilde{Y}_1, \ldots, \tilde{Y}_n$, that satisfy three properties:

1. $\tilde{X}_i$ has the same distribution as X_i, and similarly for $\tilde{Y}_i$ and Y_i.
2. $\tilde{X}_1, \ldots, \tilde{X}_n$ are mutually independent (since the coupling only "binds" X_i and Y_i), and so are $\tilde{Y}_1, \ldots, \tilde{Y}_n$.
3. $\tilde{X}_i \le \tilde{Y}_i$ for all $i \in [n]$.

From the first two properties, we have that $X = \sum_{i=1}^{n} X_i$ and $\sum_{i=1}^{n} \tilde{X}_i$ have the same distribution (and similarly for the Ys), and hence:

$$\Pr[X \geq k] = \Pr\left[\sum_{i=1}^{n} \tilde{X}_i \geq k\right] \leq \Pr\left[\sum_{i=1}^{n} \tilde{Y}_i \geq k\right] = \Pr[Y \geq k],$$

where the inequality follows from the third property.

3 Convergence of Random Variables

Suppose we are given a sequence of random variables $\{X_n\}_{n\in\mathbb{N}}$. Intuitively, we say that the sequence converges to some random variable X if X_n "behaves" more and more like X as $n \to \infty$. For an intuitive example, if $X_n \sim \mathrm{Po}(1 - 1/n)$, then clearly X_n " $\underset{n\to\infty}{\longrightarrow}$ " $\mathrm{Po}(1)$.

There are several notions of convergence of random variables. In this section, we define four such notions, and state some theorems that might help us better understand the relations between these different notions.

Definition 3.1 (Convergence in Distribution)**.** A sequence of random variables $\{X_n\}_{n\in\mathbb{N}}$ *converges in distribution* to a random variable X, and we denote $X_n \xrightarrow{D} X$, if for any $x \in \mathbb{R}$,

$$\lim_{n\to\infty} \Pr[X_n \leq x] = \Pr[X \leq x].$$

When X and X_n (for all n) is supported on natural numbers, then this definition can also be stated equivalently as follows:

Claim 3.2. *Assume X and X_n (for all n) is supported on natural numbers. Then $X_n \xrightarrow{D} X$ if and only if $\lim_{n\to\infty} \Pr[X_n = x] = \Pr[X = x]$ for all $x \in \mathbb{N}$.*

The proof of this claim is straightforward, and we leave it as an exercise.

Definition 3.3 (Convergence in Probability)**.** A sequence of random variables $\{X_n\}_{n\in\mathbb{N}}$ *converges in probability* to a random variable X, and we denote $X_n \xrightarrow{P} X$, if for any $\epsilon > 0$,

$$\lim_{n\to\infty} \Pr[|X_n - X| \geq \epsilon] = 0.$$

Definition 3.4 (Convergence in Mean)**.** A sequence of random variables $\{X_n\}_{n\in\mathbb{N}}$ *converges in the rth mean* (or, converges in the L^r-norm) to a random variable X, and we denote $X_n \xrightarrow{L^r} X$, if for any $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{E}[|X_n - X|^r] = 0.$$

When $r = 1$, we say that X_n converges in mean to X, and denote $X_n \xrightarrow{\mathbb{E}} X$.

Claim 3.5. *If $X_n \xrightarrow{L^r} X$, then $\lim_{n\to\infty} \mathbb{E}[|X_n|^r] = \mathbb{E}[|X|^r]$.*

Definition 3.6 (Almost Sure Convergence)**.** A sequence of random variables $\{X_n\}_{n\in\mathbb{N}}$ *converges almost surely* to a random variable X, and we denote $X_n \xrightarrow{\text{a.s.}} X$, if for any $\epsilon > 0$,

$$\Pr[\lim_{n\to\infty} X_n = X] = 1.$$

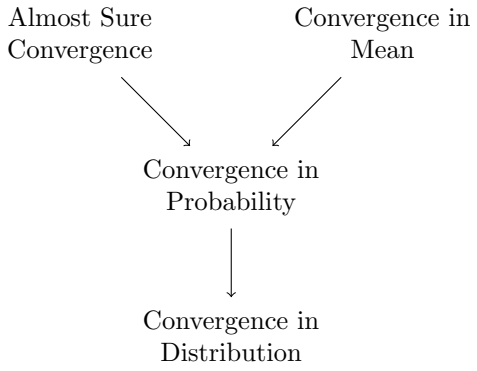
3.1 Relations between the Different Notions of Convergence

In this subsection, we state three theorems of the form "Type A convergence implies Type B convergence". The relations we state are visually presented in Figure 3, where a directed edge represents the relation "implies".

Theorem 3.7. *Provided that the probability space is complete,*[3] *the following properties hold:*

[3] A probability space is complete if every subset of a null set (an event whose probability is zero) is also an event. Every probability space relevant for us is complete.

Figure 3: Relations between the different notions of convergence



1. *If* $X_n \xrightarrow{\mathbb{E}} X$ *then* $X_n \xrightarrow{P} X$.
2. *If* $X_n \xrightarrow{a.s.} X$ *then* $X_n \xrightarrow{P} X$.
3. *If* $X_n \xrightarrow{P} X$ *then* $X_n \xrightarrow{D} X$.

Theorem 3.8. *Provided that the probability space is complete,*

1. *If* $X_n \xrightarrow{P} X$ *and* $Y_n \xrightarrow{P} Y$ *then* $X_nY_n \xrightarrow{P} XY$ *and* $aX_n + bY_n \xrightarrow{P} aX + bY$ *for any* $a, b \in \mathbb{R}$. *An analogous statement holds for convergence in mean and almost sure convergence.*
2. *If* $X_n \xrightarrow{P} X$ *and* $X_n \xrightarrow{P} Y$, *then* $X = Y$ *almost surely. An analogous statement holds if* X_n *converges to* X *and* Y *in mean or almost surely (note: in all three cases,* $X = Y$ almost surely*).*
3. ***Continuous mapping theorem**: If* $X_n \xrightarrow{P} X$ *then* $g(X_n) \xrightarrow{P} g(X)$ *for any continuous function* g. *A similar statement holds for convergence in distribution and almost sure convergence.*

3.2 Limit Theorems

In this subsection we present three limit theorems, which are theorems that characterize the behavior of sequences of random variables.

3.2.1 Poisson Limit Theorem

The Poisson limit theorem relates the Binomial and the Poisson distributions. Formally stated,

Theorem 3.9 (Poisson Limit Theorem)**.** *Let* $\lambda \in \mathbb{R}^+$ *be some constant, and let* $\{X_n\}_{n\in\mathbb{N}}$ *be a sequence of random variables such that* $X_n \sim \mathrm{Bin}(n, \frac{\lambda}{n})$. *Then,* $X_n \xrightarrow{D} \mathrm{Po}(\lambda)$.

There are several proofs of the Poisson limit theorem, e.g. using Stirling's approximation for factorials or generating functions. We chose to prove the theorem using more basic tools:

Proof. Let $X \sim \mathrm{Po}(\lambda)$. From Theorem 3.2, it suffices to prove that $\lim_{n\to\infty} \Pr[X_n = k] = \Pr[X = k]$.

$$\begin{aligned}
\lim_{n\to\infty} \Pr[X_n = k] &= \lim_{n\to\infty} \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \\
&= \lambda^k \lim_{n\to\infty} \frac{n!}{k!(n-k)!} \left(\frac{1}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \\
&= \frac{\lambda^k}{k!} \lim_{n\to\infty} \frac{n(n-1)\cdots(n-k+1)}{n^k} \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-k}
\end{aligned}$$

$$= \frac{\lambda^k}{k!} \cdot \left[\lim_{n\to\infty} \frac{n(n-1)\cdots(n-k+1)}{n^k}\right] \cdot \left[\lim_{n\to\infty} \left(1-\frac{\lambda}{n}\right)^n\right] \cdot \left[\lim_{n\to\infty} \left(1-\frac{\lambda}{n}\right)^{-k}\right]$$
$$= \frac{\lambda^k}{k!} \cdot 1 \cdot e^{-\lambda} \cdot 1$$
$$= \Pr[X = k]. \qquad \square$$

To better understand the Poisson limit theorem, we may think of the Poisson distribution as a sum of exponential clocks: consider an infinite sequence T_i of variables with standard exponential distribution Exp(1) (so $\Pr[T_i > t] = e^{-t}$), and the corresponding partial sums sequence $T_1, T_1 + T_2, \ldots$. We can think of the partial sums sequence as describing the following process: at time zero, we start an exponential clock, and when it "arrives", we mark this, and start another one, and so on. The partial sums sequence marks the arrival times. The number of arrivals until time λ has distribution $\mathrm{Po}(\lambda)$. We can see this by dividing the interval $[0, \lambda]$ into n parts, and using the alternative definition of the exponential distribution as a memoryless distribution which on an interval of infinitesimal length ϵ has a probability of ϵ to "buzz" (given that it hasn't buzzed so far). In some sense, this exponential process is the limit of $\mathrm{Bin}(n, \lambda/n)$ at $n = \infty$, as presented above.

3.2.2 Law of Large Numbers

The law of large numbers roughly states that re-sampling from the same probability distribution reduces the deviation of the average sample from the expectation. Formally stated,

Definition 3.10. Let $\{X_n\}_{n\in\mathbb{N}}$ be a sequence of independent identically distributed (i.i.d.) random variables, with finite expectation μ and finite variance σ^2. The *n-sample mean* is defined to be $\bar{X}_n \triangleq \frac{1}{n}\sum_{i=1}^n X_i$.

Two almost immediate observations, that follow from the linearity of expectation and variance[4], are that $\mathbb{E}[\bar{X}_n] = \mu$ and $\mathbb{V}[\bar{X}_n] = \frac{1}{n}\sigma^2$.

Theorem 3.11 (Law of Large Numbers)**.** $\bar{X}_n \xrightarrow{D} \mu$.

The proof is simple, and is hence omitted. It follows from applying Chebyshev's inequality, that is defined in Theorem 4.5, to $\Pr[|\bar{X}_n - \mu| \geq \epsilon]$.

This variant of the law of large numbers is called the *weak* law of large numbers. There is another variant of the law of large numbers, called the *strong* law of large numbers, in which the same statement is made for almost sure convergence.

3.2.3 Central Limit Theorem

The central limit theorem states that, under certain conditions, the sum of a large number of random variables behaves similarly to a normal random variable. Formally stated,

Theorem 3.12 (Central Limit Theorem)**.** *Let $\{X_n\}_{n\in\mathbb{N}}$ be a sequence of independent identically distributed (i.i.d.) random variables, with finite expectation μ and finite variance σ^2. We denote the nth sample mean by $\bar{X}_n$ and define $Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$ for all $n \in \mathbb{N}$. Then, $\bar{X}_n \xrightarrow{D} \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$ and $Z_n \xrightarrow{D} \mathrm{N}(0,1)$.*

For a theorem of such fundamental importance in probability and statistics, the proof is surprisingly simple. However, we will not provide it here since it uses *characteristic functions*, with which we do not deal for two main reasons: first, we will not need them in our course, and second, they require a bit of complex analysis. Nevertheless, we encourage curiosity!

4 Large Deviation Bounds and Concentration of Measure

In probability, the theory of large deviations concerns the asymptotic behavior of remote tails of sequences of probability distributions. It formalizes the heuristic ideas of concentration and convergence of measures, that is, that measurements do not tend to deviate from the expectation.

[4] Recall that linearity of variance can be stated for independent random variables.

4.1 Markov's inequality

Theorem 4.1 (Markov inequality)**.** *Let X be a nonnegative random variable over a probability space Ω, and let $\lambda \geq 0$. Then, $\Pr[X \geq \lambda] \leq \frac{\mathbb{E}[X]}{\lambda}$, or alternatively $\Pr[X \geq \lambda \mathbb{E}[X]] \leq \frac{1}{\lambda}$.*

Proof. For simplicity, assume X is a discrete random variable. Then,

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{\omega \in \Omega} X(\omega) \Pr[X = \omega] && \text{(definition of expectation)} \\
&\geq \sum_{X(\omega) \geq \lambda} X(\omega) \Pr[X = \omega] \\
&\geq \sum_{X(\omega) \geq \lambda} \lambda \Pr[X = \omega] \\
&= \lambda \Pr[X \geq \lambda]. && \square
\end{aligned}$$

4.1.1 Example: Las-Vegas and Monte-Carlo algorithms

When considering randomized algorithms, one might think of several possible definitions. Two of these, that are widely used in computer science theory, are the notions of *Las-Vegas* algorithms and *Monte-Carlo* algorithms.

Definition 4.2 (Monte-Carlo algorithms)**.** Monte-Carlo algorithms have a fixed running time (usually, polynomial in the input size), but they have a small probability of producing an incorrect result.

Definition 4.3 (Las-Vegas algorithms)**.** Las-Vegas algorithms have a finite (usually, polynomial in the input size) running time *in expectation*, but they always produce the correct answer.

Usually, when we think of randomized algorithms, we think of Monte-Carlo algorithms. However, a simple argument, that uses the Markov inequality, shows that any problem that can be solved with a Las-Vegas algorithm can also be using a Monte-Carlo algorithm as well.

Claim 4.4. *Any Las-Vegas algorithm that has expected time T can be "transformed" into a Monte-Carlo algorithm that runs in worst case time of $2T \log 1/\delta$ and produces the correct result with probability $\geq 1-\delta$.*

Proof. Let $\mathcal{A}$ be the Las-Vegas algorithm in hand. We define its Monte-Carlo "equivalent" as follows:

1. Repeat $\log 1/\delta$ times (independently):
   (a) Run $\mathcal{A}$ for $2T$ steps. If $\mathcal{A}$ stopped and returned a result, return it as well and terminate.
2. If the algorithm did not terminate earlier, return anything.

Clearly, the algorithm above runs in worst-case time of $2T \log 1/\delta$. As for its error probability, we observe that if one of the runs in step (1a) terminated, then the returned result is correct, since $\mathcal{A}$ is a Las-Vegas algorithm. Therefore,

$$\begin{aligned}
\Pr\left[\begin{array}{l}\text{our Monte-Carlo}\\ \text{algorithm errs}\end{array}\right] &\leq \Pr\left[\bigwedge_{i=1}^{\log 1/\delta} \begin{array}{l}\text{the } i\text{'th run in step (1a)}\\ \text{does not terminate in time}\end{array}\right] \\
&= \prod_{i=1}^{\log 1/\delta} \Pr\left[\begin{array}{l}\text{the } i\text{'th run in step (1a)}\\ \text{does not terminate in time}\end{array}\right] && \text{(independence of iterations)} \\
&= \prod_{i=1}^{\log 1/\delta} \Pr\left[\text{the time of } \mathcal{A} \text{ is } \geq 2T\right] \\
&\leq \prod_{i=1}^{\log 1/\delta} \frac{T}{2T} && \text{(Markov inequality)} \\
&= \left(\frac{1}{2}\right)^{\log 1/\delta} = \delta. && \square
\end{aligned}$$

4.2 Chebyshev's inequality

Theorem 4.5 (Chebyshev's inequality)**.** *Let X be a random variable with finite expectation $\mathbb{E}[X]$ and finite non-zero variance $\mathbb{V}[X]$. Then,* $\Pr\left[|X - \mathbb{E}[X]| \geq \lambda\right] \leq \frac{\mathbb{V}[X]}{\lambda^2}$.

Proof. The proof simply applies Markov's inequality to the random variable $(X - \mathbb{E}[X])^2$. Namely,

$$\begin{aligned}
\Pr\left[|X - \mathbb{E}[X]| \geq \lambda\right] &= \Pr\left[(X - \mathbb{E}[X])^2 \geq \lambda^2\right] \\
&\leq \frac{\mathbb{E}\left[(X - \mathbb{E}[X])^2\right]}{\lambda^2} && \text{(Markov inequality)} \\
&\leq \frac{\mathbb{V}[X]}{\lambda^2}. && \text{(definition of variance)}
\end{aligned}$$

□

4.3 Chernoff bound

Theorem 4.6 (Chernoff bound)**.** *Let $\{X_1, \ldots, X_n\}$ be mutually independent $\{0,1\}$-random variables (that is, X_i is 1 with probability p_i and 0 with probability $1 - p_i$), and let $X \triangleq \sum_{i=1}^{n} X_i$. $\mu \triangleq \mathbb{E}[X]$. Then for every $\delta > 0$,*

$$\Pr\left[X \geq (1+\delta)\cdot\mu\right] \leq \left[\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right]^{\mu} \tag{1}$$

$$\Pr\left[X \leq (1-\delta)\cdot\mu\right] \leq \left[\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right]^{\mu} \tag{2}$$

Proof. Since the proofs of the two inequalities are similar, we only tend to the first. The proof applies Markov's inequality to random variable $\exp(tX)$, where t is a positive dummy variable. To apply Markov's inequality on our newly defined random variable, we must first find its expectation.

$$\begin{aligned}
\mathbb{E}[\exp(tX)] &= \mathbb{E}\left[\exp\left(t\sum_{i=1}^{n} X_i\right)\right] \\
&= \mathbb{E}\left[\prod_{i=1}^{n} \exp(tX_i)\right] \\
&= \prod_{i=1}^{n} \mathbb{E}[\exp(tX_i)] && (X_i\text{s are independent}) \\
&= \prod_{i=1}^{n} ((1-p_i)\cdot\exp(t\cdot 0) + p_i\cdot\exp(t\cdot 1)) && \text{(definition of expectation)} \\
&= \prod_{i=1}^{n} (1 + p_i(e^t - 1)) \\
&\leq \prod_{i=1}^{n} \exp(p_i(e^t - 1)) && (1 + x \leq e^x \text{ for all } x \in \mathbb{R}) \\
&= \exp\left(\sum_{i=1}^{n} p_i(e^t - 1)\right) \\
&= \exp(\mu(e^t - 1)).
\end{aligned}$$

Now, applying Markov's inequality yields

$$\begin{aligned}
\Pr\left[X \geq (1+\delta)\mu\right] &= \Pr\left[t\cdot X \geq t\cdot(1+\delta)\mu\right] \\
&= \Pr\left[\exp(tX) \geq \exp(t(1+\delta)\mu)\right] \\
&\leq \frac{\mathbb{E}[\exp(tX)]}{\exp(t(1+\delta)\mu)} && \text{(Markov inequality)} \\
&= \frac{\exp(\mu(e^t - 1))}{\exp(t(1+\delta)\mu)}
\end{aligned}$$

$$= \exp(\mu(e^t - 1) - t(1+\delta)\mu).$$

Since t is a dummy variable, we have that $\Pr[X \geq (1+\delta)\mu] \leq \exp(\mu(e^t - 1) - t(1+\delta)\mu)$ for any choice of t. We thus seek for the value of t for which the right-hand side of the inequality is smallest, and that would produce the tightest upper bound. It suffices to derive what is inside the exponent to find its minimum:

$$\frac{d}{dt}((e^t - 1) - t(1+\delta)) = e^t - 1 - \delta = 0 \quad \Rightarrow \quad t = \ln(1+\delta)$$

and this value of t yields the bound from the theorem statement. □

For our purposes, the following corollary will mostly suffice:

Corollary 4.7. *For every $c > 0$,*

$$\Pr\left[\left|\sum_{i=1}^{n} X_i - \mu\right| \geq c\mu\right] \leq 2 \cdot e^{-\min\{c^2/4, c/2\}\cdot\mu}$$

The corollary above suggests that the probability that the sum of X_is diverges from its expectation is exponentially small, i.e. bounded by $2^{-\Omega(\mu)}$ (where the constant in the Ω notation depends on the choice of c).

Example *(Applying Chernoff's Bound to Coin Tosses).* Suppose we toss an unbiased coin (i.e. a coin whose probability to land on either side is 1/2) $n = 1000$ times, and let $X_i = \mathbf{1}_{\{\text{the } i\text{'th toss was Heads}\}}$. $X = \sum_{i=1}^{n} X_i$ is the total number of heads we saw, and clearly, the expected number of heads in these n tosses is $\mu = 500$. Suppose we ask what is the probability of seeing at least 600 heads. Using Chernoff's bound yields

$$\Pr[X \geq 600] = \Pr[X \geq (1+0.2)\cdot\mu] \leq \left[\frac{e^{0.2}}{1.2^{1.2}}\right]^{500} \approx 8.33 \cdot 10^{-5}.$$

And if we are interested in bounding the probability that we diverge from the expectation by 150 (i.e., that there are less than 350 Heads or more than 650), we can apply the corollary:

$$\Pr[|X - 500| \geq 150] = \Pr[|X - 500| \geq 0.3 \cdot 500] \leq 2 \cdot e^{-\min\{0.3^2/4, 0.3/2\}\cdot 500} = 2.601 \cdot 10^{-5}.$$

5 Useful Inequalities

5.1 Cauchy–Schwartz Inequality

Theorem 5.1. *Let $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$ be real numbers. Then*

$$\left(\sum_{i=1}^{n} a_i b_i\right)^2 \leq \sum_{i=1}^{n} a_i^2 \sum_{i=1}^{n} b_i^2.$$

To prove the theorem, we first state a simple lemma:

Lemma 5.2. *Let a and b be positive real numbers. Then $\sqrt{ab} \leq \frac{a+b}{2}$, and equality holds iff $a = b$.*

The lemma can be easily verified, since $(a+b)^2 - 4ab = (a-b)^2 \geq 0$.

Proof of Theorem 5.1. Assume $\sum_{i=1}^{n} a_i^2 > 0$ and $\sum_{i=1}^{n} b_i^2 > 0$ (otherwise, the inequality is trivial). Let $\alpha_j \triangleq \frac{a_j^2}{\sum_{i=1}^{n} a_i^2}$ and $\beta_j \triangleq \frac{b_j^2}{\sum_{i=1}^{n} b_i^2}$. Applying the lemma above, we get $\sqrt{\alpha_j \beta_j} \leq \frac{1}{2}(\alpha_j + \beta_j)$, i.e.

$$\frac{a_j}{\sqrt{\sum_{i=1}^{n} a_i^2}} \cdot \frac{b_j}{\sqrt{\sum_{i=1}^{n} b_i^2}} \leq \frac{1}{2}\left(\frac{a_j^2}{\sum_{i=1}^{n} a_i^2} + \frac{b_j^2}{\sum_{i=1}^{n} b_i^2}\right).$$

Summing the above inequality over all $j \in [n]$ yields

$$\frac{\sum_{j=1}^{n} a_j b_j}{\sqrt{\sum_{i=1}^{n} a_i^2}} \leq \frac{1}{2}\left(\frac{\sum_{j=1}^{n} a_j^2}{\sum_{i=1}^{n} a_i^2} + \frac{\sum_{j=1}^{n} b_j^2}{\sum_{i=1}^{n} b_i^2}\right) = \frac{1}{2}(1+1) = 1,$$

and that completes the proof. □

5.2 Jensen's Inequality

Before stating the inequality, we remind the reader the definitions of convex and concave functions.

Definition 5.3. A function $\varphi\colon A \to \mathbb{R}$ is *convex* if $\varphi(ta_1 + (1-t)a_2) \leq t\varphi(a_1) + (1-t)\varphi(a_2)$ for any $a_1, a_2 \in A$ and any $t \in [0,1]$. φ is said to be *concave* if $-\varphi$ is convex (that is, the inequality flips side).

Theorem 5.4 (Jensen's inequality)**.** *Let X be a real valued random variable, and let $\varphi\colon \mathbb{R} \to \mathbb{R}$ be a convex function. Then $\varphi(\mathbb{E}[X]) \leq \mathbb{E}[\varphi(X)]$. Similarly, if φ is concave, then the inequality flips side.*

5.3 FKG inequality

The *FKG theorem*, named after Fortuin, Kasteleyn, and Ginibre [3], is not an elementary, and its proof requires a bit of work. We state the theorem hereafter, and leave the details of the proof for curious readers.

Definition 5.5. Let Ω be a set and let $f\colon 2^{\Omega} \to \mathbb{R}^+$. μ is said to be *non-decreasing monotone* if $\mu(A) \leq \mu(B)$ for every $A \subseteq B \subseteq \Omega$.

Definition 5.6. Let Ω be a set and let $\mu\colon 2^{\Omega} \to \mathbb{R}^+$. μ is said to be *log super-modular* if for every $A, B \in 2^{\Omega}$,

$$\mu(A)\mu(B) \leq \mu(A \cup B)\mu(A \cap B).$$

Example. An example of a log super-modular function is the function $\mu_p\colon 2^{\binom{[n]}{2}} \to [0,1]$ that assigns every set $E \subseteq \binom{[n]}{2}$ the value $p^{|E|}(1-p)^{\binom{n}{2}-|E|}$, i.e. $\mu_p(E)$ is the probability that a the set of edges of a graph sampled from $G(n,p)$ is E. This function is log super-modular since

$$\mu_p(E)\mu_p(F) = p^{|E|+|F|}(1-p)^{2\binom{n}{2}-|E|-|F|} = p^{|E\cup F|+|E\cap F|}(1-p)^{2\binom{n}{2}-|E\cup F|-|E\cap F|} = \mu_p(E\cup F)\mu_p(E\cap F).$$

Theorem 5.7 (FKG Theorem, [3])**.** *For any log super-modular function $\mu\colon 2^{\Omega} \to \mathbb{R}^+$ and for any two non-decreasing monotone functions $f, g\colon 2^{\Omega} \to \mathbb{R}^+$,*

$$\left(\sum_{A\subseteq\Omega} \mu(A)f(A)\right)\left(\sum_{A\subseteq\Omega} \mu(A)g(A)\right) \leq \left(\sum_{A\subseteq\Omega} \mu(A)f(A)g(A)\right)\left(\sum_{A\subseteq\Omega} \mu(A)\right).$$

Example. An example for applications of the FKG theorem in random $G(n,p)$ graphs is for f, g that represent monotone properties of graphs. For instance, if we take μ to be μ_p (defined above), and we take $\chi(G)$ and $\kappa(G)$ to be the chromatic number and the size of the larges clique of G respectively, then the theorem proves that $\mathbb{E}[\kappa(G)] \cdot \mathbb{E}[\chi(G)] \leq \mathbb{E}[\kappa(G) \cdot \chi(G)]$.

In the μ_p case (or more generally, for any product measure), it is not too hard to prove the FKG inequality by induction. We will show by induction that if f, g are two monotone functions on $\{0,1\}^n$ then $\mathbb{E}[fg] \geq \mathbb{E}[f]\,\mathbb{E}[g]$, when the expectations are taken with respect to μ_p.

When $n = 1$, we have two monotone functions $f, g\colon \{0,1\} \to \mathbb{R}$. Notice that for any $x, y \in \{0,1\}$ we have $(f(x) - f(y))(g(x) - g(y)) \geq 0$. Indeed, if $x \geq y$ then both factors are non-negative, and if $x \leq y$ then both are non-positive. It follows that for $x, y \sim \{0,1\}$,

$$0 \leq \mathbb{E}[(f(x)-f(y))(g(x)-g(y))] = \mathbb{E}[f(x)g(x)]+\mathbb{E}[f(y)g(y)]-\mathbb{E}[f(x)g(y)]-\mathbb{E}[f(y)g(x)] = 2(\mathbb{E}[fg]-\mathbb{E}[f]\,\mathbb{E}[g]),$$

and so $\mathbb{E}[fg] \leq \mathbb{E}[f]\,\mathbb{E}[g]$.

Suppose now that the inequality holds for some n. We will prove it for $n+1$. Write

$$\mathop{\mathbb{E}}_{x_1,\ldots,x_{n+1}}[fg] = \mathop{\mathbb{E}}_{x_1,\ldots,x_n}\mathop{\mathbb{E}}_{x_{n+1}}[f(x_1,\ldots,x_n;x_{n+1})g(x_1,\ldots,x_n;x_{n+1})].$$

Applying the case $n = 1$ to estimate the inner expectation (with the functions $f_{x_1,\ldots,x_n}(x_{n+1}) = f(x_1,\ldots,x_n;x_{n+1})$ and $g_{x_1,\ldots,x_n}(x_{n+1}) = g(x_1,\ldots,x_n;x_{n+1})$), we obtain

$$\mathop{\mathbb{E}}_{x_1,\ldots,x_{n+1}}[fg] \geq \mathop{\mathbb{E}}_{x_1,\ldots,x_n}\mathop{\mathbb{E}}_{x_{n+1}}[f(x_1,\ldots,x_n;x_{n+1})]\mathop{\mathbb{E}}_{x_{n+1}}[g(x_1,\ldots,x_n;x_{n+1}].$$

Let $F(x_1,\ldots,x_n) = \mathbb{E}_{x_{n+1}}[f(x_1,\ldots,x_n;x_{n+1})]$ and $G(x_1,\ldots,x_n) = \mathbb{E}_{x_{n+1}}[g(x_1,\ldots,x_n;x_{n+1})]$. Applying the induction hypothesis, we deduce

$$\mathop{\mathbb{E}}_{x_1,\ldots,x_{n+1}}[fg] \geq \mathop{\mathbb{E}}_{x_1,\ldots,x_n}[F]\mathop{\mathbb{E}}_{x_1,\ldots,x_n}[G] = \mathop{\mathbb{E}}_{x_1,\ldots,x_{n+1}}[f]\mathop{\mathbb{E}}_{x_1,\ldots,x_{n+1}}[g].$$

References

[1] Birth Rate. Birth rate — Wikipedia, the free encyclopedia, 2017. [Online; accessed 17-November-2017].

[2] Centers of Disease Control and Prevention. Estimates of deaths associated with seasonal influenza — united states, 1976–2007, 2010. [Online; accessed 17-November-2017].

[3] Cees M Fortuin, Pieter W Kasteleyn, and Jean Ginibre. Correlation inequalities on some partially ordered sets. *Communications in Mathematical Physics*, 22(2):89–103, 1971.

[4] Normal Distribution. Normal distribution — Wikipedia, the free encyclopedia, 2017. [Online; accessed 17-November-2017].