Background on Probability

Chandra Chekuri*
University of Illinois at Urbana-Champaign

August 18, 2020

1 Introduction

These notes are intended to provide a quick refresher on basics of probability before introduction to the use of randomization in algorithms. Probabilistic analysis tends to be non-trivial even in simple settings. Various tools have been developed over the years and it is helpful to see some simple templates. For the most part we will work with discrete (in fact finite) probability spaces but we will need also some tools from continuous spaces. There are several subtleties involved in formal aspects of probability theory when dealing with continuous spaces. These will not be of concern to us but it is useful to know why one needs more involved definitions.

1.1 Discrete Probability Space and Events

We will start with simple finite probability spaces. Consider the experiment of tossing an unbiased coin which turns up heads with probability $1/2$ and tails with probability $1/2$. Or tossing a 6-sided die where each side is equally likely (with probability $1/6$). To formally understand probability we need to define a probability space. In the finite setting it consists of two pieces of information. A set of possible *atomic* or *elementary* events Ω and for each $\omega \in \Omega$ the probability that ω is realized.

Definition 1.1. *A finite probability space is a pair* $(\Omega, \mathbf{Pr})$ *consists of finite set* Ω *of elementary events and function* $\mathbf{Pr} : \Omega \to [0,1]$ *which assigns a probability* $\mathbf{Pr}[\omega]$ *for each* $\omega \in \Omega$ *such that* $\sum_{\omega \in \Omega} \mathbf{Pr}[\omega] = 1$.

Example 1. *An unbiased coin.* $\Omega = \{H, T\}$ *and* $\mathbf{Pr}[H] = \mathbf{Pr}[T] = 1/2$.

Example 2. *A biased coin.* $\Omega = \{H, T\}$ *and* $\mathbf{Pr}[H] = 1/4$ *and* $\mathbf{Pr}[T] = 3/4$.

Example 3. *A* 6*-sided unbiased die.* $\Omega = \{1, 2, 3, 4, 5, 6\}$ *and* $\mathbf{Pr}[i] = 1/6$ *for* $1 \leq i \leq 6$.

From simple probability spaces one often creates more complex ones by the *product* construction.

Definition 1.2. *Suppose* $(\Omega_1, \mathbf{Pr}_1)$ *and* $(\Omega_2, \mathbf{Pr}_2)$ *are two finite probability spaces. Then the product probability space* $(\Omega, \mathbf{Pr})$ *is defined as follows.* $\Omega = \Omega_1 \times \Omega_2$ *and* $\mathbf{Pr}(\omega_1, \omega_2) = \mathbf{Pr}_1(\omega_1)\,\mathbf{Pr}_2(\omega_2)$ *for all* $\omega_1 \in \Omega_1, \omega_2 \in \Omega_2$.

The reader unfamiliar with product constructions should verify that the product space is indeed a probability space. That is, $\sum_{\omega \in \Omega} \mathbf{Pr}[\omega] = 1$.

One can easily extend the above product definition to the product of a finite number of spaces. Things get tricky if one wants infinite products but we will not need it.

*chekuri@illinois.edu. Comments and corrections are welcome.

Example 4. *Two unbiased coins. One can view this as a single probability space* $(\Omega, \mathbf{Pr})$ *where* $\Omega = \{HH, HT, TH, TT\}$ *where the probability of each event is* $1/4$ *or view this as the product of two spaces where each space is that of one unbiased coin. The product view is often useful when considering independence. On the other hand when correlations are present it is useful to view it as a single space.*

Most things from finite probability spaces extend to a probability space over a countably infinite set. Recall that a set A is countably infinite if there is a bijection from A to the natural numbers. Examples include the set of even numbers, the set of primes, set of all strings over a finite alphabet.

Example 5. *Let* $\lambda > 0$ *be a fixed real number. Consider the space* $(\Omega, \mathbf{Pr})$ *where* $\Omega = \{0, 1, 2, \ldots, \}$ *is the set of non-negative integers and* $\mathbf{Pr}(i) = e^{-\lambda}\frac{\lambda^i}{i!}$ *for each* $i \geq 0$*. This is the Poisson distribution with mean* λ*. We see that* $\sum_{i\geq 0} \mathbf{Pr}(i) = 1$ *since* $e^{\lambda} = \sum_{i\geq 0} \frac{\lambda^i}{i!}$*.*

Question 1. *Suppose* Ω *is the set of positive integers. We want to have a probability space where* $\mathbf{Pr}(i) = c/i^2$ *for each* $i \geq 1$ *where* c *is a fixed constant. Is it possible? What about if we wanted have a probability space where* $\mathbf{Pr}(i) = c/i$ *for each* i*?*

Events: In most probability spaces what we are really after are interesting events.

Definition 1.3. *Given a probability space* $(\Omega, \mathbf{Pr})$ *an event is a subset of* Ω*. In other words an event is a collection of elementary events. The probability of an event* A*, denoted by* $\mathbf{Pr}[A]$*, is* $\sum_{\omega \in A} \mathbf{Pr}[\omega]$*. The complement event of an event* $A \subseteq \Omega$ *is the event* $\Omega \setminus A$ *frequently denoted by* $\bar{A}$*.*

Example 6. *A pair of independent dice.* $\Omega = \{(i, j) \mid 1 \leq i \leq 6, 1 \leq j \leq 6\}$*. Let* A *be the event that the sum of the two numbers on the dice is even. Then* $A = \{(i, j) \in \Omega : (i + j)$ *is even*$\}$*.* $\mathbf{Pr}[A] = |A|/36 = 1/2$*.*

Question 2. *Consider the probability space corresponding to the Poisson distribution with mean* λ*. Let* A *be the event that the outcome is an even number. What is* $\mathbf{Pr}[A]$*?*

The following question illustrates that one can easily define events over relatively simple probability spaces but whose probability may not be easy to compute.

Question 3. *Let* $(\Omega, \mathbf{Pr})$ *be the probability space corresponding to the product of the probability spaces of 100 unbiased coins. Let* A *be the event that the number of heads is a prime number. What is* $\mathbf{Pr}[A]$*?*

Algebra over events: Given two events A, B we can define their union as the event $A \cup B$ and their intersection as $A \cap B$. We can easily prove the following in the setting of finite probability spaces via basic set theory.

$$\mathbf{Pr}[A] + \mathbf{Pr}[B] = \mathbf{Pr}[A \cup B] + \mathbf{Pr}[A \cap B]$$

which can be written alternatively as

$$\mathbf{Pr}[A \cup B] = \mathbf{Pr}[A] + \mathbf{Pr}[B] - \mathbf{Pr}[A \cap B].$$

What about the union of 3 events A, B, C? You can convince yourself via Venn diagram that

$$\mathbf{Pr}[A \cup B \cup C] = \mathbf{Pr}[A] + \mathbf{Pr}[B] + \mathbf{Pr}[C] - \mathbf{Pr}[A \cap B] - \mathbf{Pr}[B \cap C] - \mathbf{Pr}[A \cap C] + \mathbf{Pr}[A \cap B \cap C].$$

A powerful generalization is the inclusion-exclusion formula for the union of a *finite* collection of events $A_1, A_2, \ldots, A_n$.

$$\mathbf{Pr}\left[\cup_{i=1}^{n} A_i\right] = \sum_{i=1}^{n} \mathbf{Pr}[A_i] - \sum_{1\leq i<j\leq n} \mathbf{Pr}\left[A_i \cap A_j\right] + \sum_{1\leq i<j<k\leq n} \mathbf{Pr}\left[A_i \cap A_j \cap A_k\right] + \ldots + (-1)^{n-1} \mathbf{Pr}\left[\cap_{i=1}^{n} A_i\right].$$

Probability space vs distribution: We often talk about probability distributions. What is the distinction between a probability space and a distribution? Distributions should be thought of as *templates* for an actual concrete experiment. We say $X_1, X_2, \ldots, X_n$ are distributed according to say Poisson distribution with mean λ. What this means is that we instantiate the probability space for each X_i using the values that we copy from the distribution. This is to avoid the task of specifying in gory detail the probability space for standard cases.

1.2 Continuous Probability Spaces

Consider the experiment of picking a uniformly random number from the continuous interval $[0,1]$. It is natural to view the underlying probability space Ω as $[0,1]$ where the elementary events are the real numbers in the interval $[0,1]$. How should we define **Pr** over this space? Since Ω is uncountably infinite we cannot associate any non-zero value to $\mathbf{Pr}[x]$ for any $x \in [0,1]$; thus we need to set $\mathbf{Pr}[x] = 0$. On the other hand we intuitively see that the probability that a number picked uniformly from $[0,1]$ lies in the interval $[a,b]$ with $b \geq a$ should be its length $b-a$. How do we reconcile this with setting the probability of each atomic event in $[a,b]$ to 0? One can see that it is not feasible to define probability *measures* over continuous spaces by defining the values over elementary events and then assigning the values to events. Hence a more general attempt is to define the probability measure over all events simultaneously as long as they satisfy certain basic and intuitive axioms. For the interval $[0,1]$ we would like $\mathbf{Pr}[A]$ for any event/subset $A \subseteq [0,1]$ to correspond to its "length" which is intuitive for interval events.

So an attempt would be define $\mathbf{Pr}(A)$ for all $A \subseteq \Omega$ and the natural axioms would be the following: (i) $\mathbf{Pr}[\Omega] = 1$ (ii) for each event A, $\mathbf{Pr}\left[\bar{A}\right] = 1 - \mathbf{Pr}[A]$ (iii) for any countably infinite set of disjoint events $A_1, A_2, \ldots$, we have $\mathbf{Pr}\left[\cup_{i=1}^{\infty} A_i\right] = \sum_{i=1}^{\infty} \mathbf{Pr}[A_i]$. The last property may appear strong since we ask for countably infinite unions but this is needed in many applications of probability. It turns out that one cannot create such a measure due to various technical reasons and paradoxes. To avoid these issues standard measure theory takes the view that we do not need to assign a measure to all possible subsets of Ω but only some subsets $\mathcal{F}$. We want $\mathcal{F}$ to have some basic closure properties as well as include all interesting sets that we can intuitively think of. For example, in the setting of $[0,1]$ all intervals and their complements and finite unions and intersections should be included. A set not in $\mathcal{F}$ is called non-measurable. Formally we define the notion of a σ-algebra.

Definition 1.4. *Given a set Ω a collection of subsets $\mathcal{F} \subseteq \mathcal{P}(\Omega)$ is a σ-algbera if the following properties hold:*

- $\Omega \in \mathcal{F}$
- *$A \in \mathcal{F}$ implies that $\bar{A} \in \mathcal{F}$ (closure under complementation)*
- *If $A_1, A_2, \ldots,$ is a countably infinite sequence of sets such that each $A_i \in \mathcal{F}$ then $\cup_{i=1}^{\infty} A_i \in \mathcal{F}$ (closure under countable union)*

Note that closure under complementation and countable union also implies closure under finite intersection. That is, if $A_1, A_2, \ldots, A_n \in \mathcal{F}$ then $\cap_{i=1}^{n} A_i \in \mathcal{F}$.

Remark 1.5. *For a finite or a countably infinite probability space we let $\mathcal{F} = \mathcal{P}(\Omega)$ and hence it is often not explicitly mentioned.*

With the discussion in place the formal definition of a probability space is as follows.

Definition 1.6. *A probability space is a triple* $(\Omega, \mathscr{F}, \mathbf{Pr})$ *which consists of a set* Ω*, a* σ*-algebra* $\mathscr{F}$ *over* Ω *and a probability measure* $\mathbf{Pr} : \mathscr{F} \to [0,1]$ *such that*

- $\mathbf{Pr}[\Omega] = 1$
- *For* $A \in \mathscr{F}$, $\mathbf{Pr}[A] = 1 - \mathbf{Pr}\left[\bar{A}\right]$.
- *For any countably infinite sequence of disjoint sets* $A_1, A_2, \ldots$, *such that* $A_i \in \mathscr{F}$ *for all* $i \geq 1$, $\mathbf{Pr}[\cup_i A_i] = \sum_{i=1}^{\infty} \mathbf{Pr}[A_i]$.

Density functions: Almost all of the time we don't deal with the above subtleties since we typically define continuous probability measures via *density* functions and cumulative distribution functions. In other words we define probability measure indirectly via integration. Consider $\Omega = [0,1]$. We specify a density function $f : [0,1] \to [0,1]$ we let $\mathbf{Pr}[A] = \int_A f\,dx$. For instance if we want the uniform distribution over $[0,1]$ we let $f(x) = 1$ and we see that $\mathbf{Pr}[[a,b]] = \int_a^b 1\,dx = b - a$. In fact what this formula hides is that it doesn't tell us how to compute the standard Riemann integral for an arbitrary set $A \subset [0,1]$ but only over "nice" sets such as intervals and their finite unions and intersections. However, for all practical purposes in this course and for most applications that most of us will encounter the definition of probability measures via integration of "nice" functions over "nice" sets suffices.

Example 7. *The standard Normal distribution over the entire real line* $(-\infty, \infty)$ *is defined by the density function* $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$.

Question 4. *Consider the density function* $f(x) = cx^2$ *over the set* $\Omega = [0,1]$ *where* c *is a fixed constant. What should* c *be so that* $\mathbf{Pr}[\Omega] = 1$*? What is* $\mathbf{Pr}[A]$ *where* $A = [0.2, 0.4]$*. What about when* $A = [0.1, 0.2] \cup [0.3, 0.6]$.

Remark 1.7. *We will not mention the issue of the* σ*-algebra* $\mathscr{F}$ *from now on even when discussing continuous probability measures.*

1.3 Union Bound

Union bound is a very simple bound but extremely useful because it does not make any assumptions.

Lemma 1.8. *Let* A, B *be two events in a probability space* $(\Omega, \mathbf{Pr})$ *then* $\mathbf{Pr}[A \cup B] \leq \mathbf{Pr}[A] + \mathbf{Pr}[B]$*. More generally let* $A_1, A_2, \ldots, A_n$ *be a finite collection of events then*

$$\mathbf{Pr}\left[\cup_{i=1}^{n} A_i\right] \leq \sum_{i=1}^{n} \mathbf{Pr}[A_i].$$

Proof: The discrete case is easy to prove. We consider the continuous case. $A \cup B$ is the disjoint union of $A - B, A \cap B, B - A$ (if A, B are in the σ-algebra then $A \cup B$, $A \cap B$, $A - B$, $B - A$ are also in the σ-algebra, do you see why?). Hence,

$$\begin{aligned}
\mathbf{Pr}[A \cup B] &= \mathbf{Pr}[A - B] + \mathbf{Pr}[A \cap B] + \mathbf{Pr}[B - A] \\
&\leq \mathbf{Pr}[A - B] + \mathbf{Pr}[A \cap B] + \mathbf{Pr}[A \cap B] + \mathbf{Pr}[B - A] \\
&= \mathbf{Pr}[A] + \mathbf{Pr}[B].
\end{aligned}$$

We used the third axiom of probability on additivity over disjoint sets in the first and third equalities and non-negativity of probability for the inequality in the second line. □

1.4 Independent Events

Independence is a fundamental notion in probability.

Definition 1.9. *Let A, B be two events in a probability space. They are independent iff* $\mathbf{Pr}[A \cap B] = \mathbf{Pr}[A]\mathbf{Pr}[B]$*, otherwise they are dependent.*

Example 8. *Two coins.* $\Omega = \{HH, TT, HT, TH\}$ *and* $\mathbf{Pr}[HH] = \mathbf{Pr}[TT] = \mathbf{Pr}[HT] = \mathbf{Pr}[TH] = 1/4$.

- *A is the event that the first coin is heads and B is the event that the second coin is tails. A, B are independent.*
- *A is the event that both are not tails and B is the event that the second coin is heads. A, B* **are dependent**.

The formal definition of independence is often not very useful in understanding when two events are independent since calculating the probabilities of various quantities is not straightforward. A more intuitive definition is that A and B are independent if knowing whether one of them happened does not tell us anything about whether the other happened or not. In complex settings, often the reason two events A, B can be seen to be independent is because they are based on different parts of the probability space. In fact, in most cases, we *define* or *construct* probability spaces via *independence*! Consider the following example.

Example 9. *Toss* 100 *independent unbiased coins. Let A be the event that in the first* 5 *coins there are at least* 2 *heads and let B be the event that the number of tails in the last* 10 *coins is more than* 7*. Then A, B are independent events.*

Note that A, B are *obviously* independent because the coins are defined to be independent. What is the probability space here? It is the product of the probability spaces of the individual coins. We can formalize this in a lemma below.

Lemma 1.10. *Let $(\Omega, \mathbf{Pr})$ be the product of k probability spaces $(\Omega_1, \mathbf{Pr}_1), \ldots, (\Omega_k, \mathbf{Pr}_k)$. For $S \subseteq [k]$ we say that event A depends only on S if $A = B \times C$ where $B = \prod_{i \notin S} \Omega_i$ and $C \subseteq \prod_{j \in S} \Omega_j$. If A is an event that depends only on $S \subset [k]$ and A' is an event that depends only on $T \subset [k]$ and $S \cap T = \emptyset$ then A, A' are independent.*

Proof: Exercise. □

2 Random Variables

The most important concept that you need is the idea of a random variable. A random variable in a general context is simply a mapping/function from a probability space $(\Omega, \mathbf{Pr})$ to another probability/measure space Ω'. However, for the most part, we will be mainly interested in *real-valued* random variables and *indicator* random variables.

Definition 2.1. *Given a probability space $(\Omega, \mathbf{Pr})$ a random variable X over Ω is a function that maps each elementary event to a real number. In other words $X : \Omega \to \mathbb{R}$.*

An indicator (or binary) random variable is one whose range is $\{0, 1\}$, that is $X(\omega) = 0$ or $X(\omega) = 1$ for all $\omega \in \Omega$.

Discrete versus continuous random variables: If the underlying probability space is discrete then even when X is a real-valued random variable the range space is also effectively discrete although it is embedded in a continuous space of real numbers. However when the original space Ω is continuous then one has to be more careful in general. Not all functions mapping Ω to $\mathbb{R}$ work. There is a notion of measurable functions that we will not discuss here. However, for calculations, we will need a density function f_X for the probability distribution induced by X over the real numbers. Given a density function f defining the probability distribution $(\Omega, \mathbf{Pr})$ we can usually compute the density function f_X associated with the transformation X via calculus.

For real-valued random variables an important notion closely related to the density function is the *cumulative distribution function* (cdf). Typically density is referred to by f_X and cumulative distribution function as F_X. The function $F_X : \mathbb{R} \to [0,1]$ is defined as: $F_X(t) = \mathbf{Pr}[X \leq t]$. One can see that $\mathbf{Pr}[X \leq t] = \int_{\infty}^{t} f_X(y)dy$. This also shows that $f_X = \frac{d}{dy}F_X$.

Example 10. *Consider the probability space* $(\Omega, \mathbf{Pr})$ *where* $\Omega = \{1, 2, \ldots, 6\}$ *corresponding to the outcomes of a 6-sided die with uniform probability. Let* $X : \Omega \to \mathbb{R}$ *where* $X(i) = i^2$*. Then* X *is an integer-valued random variable.*

Example 11. *Consider the probability space* $(\Omega, \mathbf{Pr})$ *where* $\Omega = \{1, 2, \ldots, 6\}$ *corresponding to the outcomes of a 6-sided die with uniform probability. Let* $X : \Omega \to \{0, 1\}$ *where* $X(i) = i \mod 2$*. Then* X *is an indicator random variable.*

Example 12. *Consider the probability space* $(\Omega, \mathbf{Pr})$ *where* $\Omega = [0, 1]$ *and* **Pr** *is the uniform distribution. Let* $X : \Omega \to \mathbb{R}$ *where* $X(a) = \exp(a)$*. Then* X *is a real-valued random variable.*

In the preceding example how do we compute the density function of X? Consider the uniform probability distribution over $[0, 1]$. The density function f corresponding to this is $f(x) = 1$ for all $x \in [0, 1]$. The random variable X maps $[0, 1]$ to $[1, e]$. In this case we see that X is a one-to-one increasing function. It is easier to work with the cumulative distribution functions F and F_X. We see that $F(x) = x$ for $x \in [0, 1]$. Since X is one-to-one we see that for any $t \in [1, e]$, $F_X(t) = \mathbf{Pr}[X \leq t] = F(\ln t) = \ln t$. And hence $f_X(t) = \frac{d}{dt}F_X(t) = 1/t$.

Why random variables? Random variables allow us to take specific views of a complex probability space and analyze that particular view. Consider any ω' in the range of a random variable $X : \Omega \to \Omega'$. Thus X collapses all atomic events in $X^{-1}(\omega')$ into ω'. Thus, one can view X as partitioning the original space Ω and associating with each part of this partition a symbol; hence X creates a new probability space implicitly. In the real-valued case, by associating a number instead of a symbol, we gain the advantage of *calculating* various quantities of interest. Moreover, as we will see, the judicious use of random variables to calculate and estimate various quantities of interest is the key for analysis of probabilistic methods and randomized algorithms.

Events and Indicator Random Variables: Recall that an event in a probability space is simply a subset of atomic events in the discrete case (which will be our main focus) or belongs to the σ-algebra in the continuous case. With each event A we can associate an indicator random variable X_A where $X_A(\omega) = 1$ if $\omega \in A$ and 0 otherwise. Why is this useful? It is useful because we have associated a number with the event and now we can do manipulations with numbers when considering multiple events instead of working with set algebra which can become quite complicated quickly. We will see several examples soon. Conversely, each indicator random variable $X : \Omega \to \{0, 1\}$ corresponds to an event A_X where $A_X = \{\omega \mid X(\omega) = 1\}$. Given a random variable X, what is $\mathbf{Pr}[X = b]$ for some b in the range of X? We consider the event $X^{-1}(b) = \{\omega \in \Omega \mid X(\omega) = b\}$ and $\mathbf{Pr}[X = b] = \mathbf{Pr}\left[X^{-1}(b)\right]$.

2.1 Independence of Random Variables

We saw the definition of independence of events. A more general notion is that of independence of random variables.

Definition 2.2. *Two discrete random variables X, Y defined over the same probability space $(\Omega, \mathbf{Pr})$ are independent if for all a, b we have $\mathbf{Pr}[X = a, Y = b] = \mathbf{Pr}[X = a]\mathbf{Pr}[Y = b]$. More generally X, Y are independent if $\mathbf{Pr}[X \in A, Y \in B] = \mathbf{Pr}[X \in A]\mathbf{Pr}[Y \in B]$ for all measurable sets A, B.*

As in the discussion of independence of events, the formal definition is less useful in knowing when random variables are actually independent. Intuitively, two random variables are independent if knowing the value of one does not tell us any information about the value of the other.

2.2 Expectation

Recall that the advantage of a real-valued random variable is that we can calculate with it. The most basic information associated with a real-valued random variable is its expectation.

Definition 2.3. *For a real-valued discrete random variable X over a probability space $(\Omega, \mathbf{Pr})$ the expectation of X, denoted by $\mathbf{E}[X]$, is defined as $\sum_{\omega \in \Omega} \mathbf{Pr}[\omega] X(\omega)$. In other words, the expectation is the average value of X according to the probabilities given by $\mathbf{Pr}[\cdot]$.*

Definition 2.4. *For a real-valued random variable X over a continuous probability space $(\Omega, \mathbf{Pr})$ where f_X is the density function of X the expectation of X is defined as $\int_{-\infty}^{\infty} y f_X(y) dy$* [1].

Remark 2.5. *Note that for infinite probability spaces the expectation of a random variable may not be finite in which case we say it does not exist. Consider the random variable X where $\mathbf{Pr}[X = i] = c/i^2$ for all integers $i \geq 1$ where $c = 6/\pi^2$. $\mathbf{E}[X] = \sum_{i \geq 1} c/i$ which diverges to ∞.*

Remark 2.6. *Often one works with functions of random variables which can be viewed as random variables themselves but calculating the expectation of the resulting function can be done more directly than computing the density function of the new random variable. Suppose X is a real-valued random variable with density function X and we are interested in the function $g(X)$ (say g is given by $g(y) = y^2$). Then $\mathbf{E}[g(X)] = \int g(y) f_X(y) dy$.*

Expectation of an indicator random variable: A very useful fact is the following easy observation.

Lemma 2.7. *For an event A let X_A be the indicator random variable for A. Then $\mathbf{E}[X_A] = \mathbf{Pr}[A]$.*

2.3 Linearity of Expectation and Examples

One of the most useful facts about expectation is the following simple one about linearity of expectation.

Lemma 2.8. *Let X, Y be two real-valued random variables over the same probability space. Then $\mathbf{E}[X + Y] = \mathbf{E}[X] + \mathbf{E}[Y]$. More generally for any finite collection $X_1, X_2, \ldots, X_n$ of random variables over the same probability space and real numbers $a_1, a_2, \ldots, a_n$ we have $\mathbf{E}\left[\sum_{i=1}^{n} a_i X_i\right] = \sum_{i=1}^{n} a_i \mathbf{E}[X_i]$.*

[1] Alternatively it is also the same as $\int_{-\infty}^{\infty} (1 - F_X(y)) dy$ where F_X is the cumulative distribution function of X. This can be seen by integration by parts.

The preceding lemma easily follows from the definition of expectation.

We now give an example to illustrate the use of linearity of expectation. Let $G = (V, E)$ be a graph with n vertices and m edges. Let G' be the graph resulting from independently deleting every vertex of G with probability $1/2$. What is the expected number of edges in G'? How do we analyze this? First we define a random variable X to denote the number of edges in G'. What we want is $\mathbf{E}[X]$. The straightforward formula for expectation is $\sum_{k=1}^{|E|} k \, \mathbf{Pr}[X = k]$. How do we calculate the quantity $\mathbf{Pr}[X = k]$ and how do we do the sum? This seems quite hard. The key is to express X as a sum and then use linearity of expectation. For each edge $e \in E$ let X_e be an indicator random variable for whether $e \in G'$. We see that $X = \sum_{e \in E} X_e$ and by linearity of expectation $\mathbf{E}[X] = \sum_{e \in |E|} \mathbf{E}[X_e]$. Since X_e is an indicator random variable we know that $\mathbf{E}[X_e] = \mathbf{Pr}[X_e = 1]$.

- Event $A_e =$ edge $e \in E$ is present in G'.
- $\mathbf{Pr}\left[A_{e=(u,v)}\right] = \mathbf{Pr}[u \text{ and } v \text{ both are present}] = \mathbf{Pr}[u \text{ is present}] \cdot \mathbf{Pr}[v \text{ is present}] = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$. Here we use independence in the experiment itself.
- $\mathbf{E}[X_e] = \mathbf{Pr}[A_e]$.

$$\mathbf{E}[X] = \mathbf{E}\left[\sum_{e \in E} X_e\right] = \sum_{e \in E} \mathbf{Pr}[A_e] = \frac{m}{4}.$$

Question 5. *Let $G = (V, E)$ be a graph with n vertices and m edges. Assume G has t triangles (i.e., a triangle is a simple cycle with three vertices). Let G' be the graph resulting from deleting independently each vertex of G with probability $1/2$. What is the expected number of triangles in G'?*

Question 6. *Suppose n balls are independently thrown into n bins where for each ball the bin is chosen uniformly at random. (i) What is the probability that bin 1 is empty? (ii) What is the expected number of empty bins?*

2.4 Independence and product of expectations

Linearity of expectation works even when the variables are not independent. There are situations when we need to work with products and here independence is useful.

Lemma 2.9. *Let X, Y be independent random variables. Then $\mathbf{E}[XY] = \mathbf{E}[X]\mathbf{E}[Y]$. More generally for any finite collection $X_1, X_2, \ldots, X_n$ of independent random variables $\mathbf{E}\left[\prod_i X_i\right] = \prod_i X_i$.*

We give a proof for the discrete setting.

Proof:

$$\begin{aligned}
\mathbf{E}[X \cdot Y] &= \sum_{\omega \in \Omega} \mathbf{Pr}[\omega](X(\omega) \cdot Y(\omega)) \\
&= \sum_{x,y \in \mathbb{R}} \mathbf{Pr}[X = x \wedge Y = y](x \cdot y) \\
&= \sum_{x,y \in \mathbb{R}} \mathbf{Pr}[X = x] \cdot \mathbf{Pr}[Y = y] \cdot x \cdot y \\
&= (\sum_{x \in \mathbb{R}} \mathbf{Pr}[X = x]\, x)(\sum_{y \in \mathbb{R}} \mathbf{Pr}[Y = y]\, y) = \mathbf{E}[X]\mathbf{E}[Y]
\end{aligned}$$

□

2.5 Variance and Higher Moments

Expectation of a random variable gives basic information about its behavior. Additional information can be obtained via higher moments. For an integer $t \geq 1$ the t'th moment of X is defined as $\mathbf{E}\left[X^t\right]$. Expectation is the first moment. The second moment is closely related to the more standard measure, namely the variance.

Definition 2.10. *Let X be a real-valued random variable over a probability space such that $\mathbf{E}[X]$ is finite. The variance of X, denoted by $\mathbf{Var}[X]$, is $\mathbf{E}\left[(X - \mathbf{E}[X])^2\right]$.*

Prove the following claim which shows that $\mathbf{Var}[X]$ differs from the second moment by an additive term.

Claim 1. $\mathbf{Var}[X] = \mathbf{E}\left[(X - \mathbf{E}[X])^2\right] = \mathbf{E}\left[X^2\right] - (\mathbf{E}[X])^2$.

Note that $\mathbf{Var}[X]$ may not be finite even when $\mathbf{E}[X]$ is finite. Can you think of an example? Note that $\mathbf{Var}[X]$ is a non-negative random variable and measures how far X deviates from its expectation. Since $\mathbf{Var}[X]$ is the square of the deviation we also use $\sigma_X = \sqrt{\mathbf{Var}[X]}$, called the standard deviation, to measure the deviation.

Example 13. *Consider a random variable X where $X = -B$ with probability $1/2$ and $X = B$ with probability $1/2$. $\mathbf{E}[X] = 0$ for all values of B. On the other hand $\mathbf{Var}[X] = B^2$ and $\sigma_X = B$.*

Question 7. *What is $\mathbf{Var}[aX]$ for a scalar a as a function of a and $\mathbf{Var}[X]$?*

Prove the following lemma.

Lemma 2.11. *Let X, Y be independent random variables. Then $\mathbf{Var}[X+Y] = \mathbf{Var}[X] + \mathbf{Var}[Y]$. More generally if $X_1, X_2, \ldots, X_n$ are independent random variables then $\mathbf{Var}\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n \mathbf{Var}[X_i]$.*

$\mathbf{E}\left[X^t\right]$ is the t'th moment of X. They are typically split into odd and even moments based on the parity of t and the even moments are non-negative and hence tend to be more useful as measures of deviation of X from its expectation. If $\mathbf{E}\left[X^t\right]$ is small compared to $\mathbf{E}[X]$ for larger values of t then it indicates that X is very close to its expectation and well-behaved. This leads to concentration inequalities that find many applications.

3 Conditional Probability and Expectation

Conditional probability allows us to focus on subsets of the probability space. Suppose we have a probability space $(\Omega, \mathbf{Pr})$ and we have information that event A happened but nothing else. If $A = \{\omega\}$ then there is no uncertainty left. However if A is a non-singleton set then we don't know which elementary event in A may have happened. How do we capture this? Since the discrete case is easier to understand and essentially captures all the intuition we will stick to it. A natural way to do it is to define a new probability space over the set A since only elementary events in A are feasible now. How should we assign probabilities to each $\omega \in A$? Let $\mathbf{Pr}'$ be the probability function over A that we wish to compute. The most natural choice is to set $\mathbf{Pr}'[\omega] = \mathbf{Pr}[\omega] / \mathbf{Pr}[A]$ for each $\omega \in A$. This ensures that $\sum_{\omega \in A} \mathbf{Pr}'[\omega] = 1$, and hence the new measure is indeed a probability measure over A. One can also think of $\mathbf{Pr}'$ as a probability measure over Ω instead of over A; we set $\mathbf{Pr}'[\omega] = 0$ for each $\omega \notin A$. $\mathbf{Pr}'$ is the probability measure *conditioned* on event A. Typically we write $\mathbf{Pr}'$ as $\mathbf{Pr}[|A]$. For an event B we say $\mathbf{Pr}[B \mid A]$ is the probability of B conditioned on event A which is simply $\mathbf{Pr}'[B]$. From our definition we have the following easy consequence. Note that this is really coming from the definition.

Lemma 3.1. *For any event B,* $\mathbf{Pr}[B \mid A] = \mathbf{Pr}'[B \cap A] = \frac{\mathbf{Pr}[A \cap B]}{\mathbf{Pr}[A]}$.

The above is often called Bayes's theorem or rule; I prefer rule because it comes easily from a definition rather than coming from a chain of non-trivial deductions. We note that Bayes's rule is applied in different ways depending on the context. In some applications we may have access to $\mathbf{Pr}[A \cap B]$ and $\mathbf{Pr}[A]$ and want to compute $\mathbf{Pr}[B \mid A]$. In others we may have access to $\mathbf{Pr}[B \mid A]$ and $\mathbf{Pr}[A]$ and want to compute $\mathbf{Pr}[A \cap B]$ (for instance in observational studies).

Verify the claim below and see why it makes intuitive sense.

Corollary 3.2. *A, B are independent events iff* $\mathbf{Pr}[B \mid A] = \mathbf{Pr}[B]$.

One can extend conditional probability to random variables in the natural way. Recall that a real-valued random variable X over a probability space $(\Omega, \mathbf{Pr})$ is simply a function from Ω to $\mathbb{R}$. As such X does not explicitly depend on the probability measure. However $\mathbf{E}[X]$ does indeed depend on $\mathbf{Pr}$. We define $\mathbf{E}[X \mid A]$ as the conditional expectation of X given A. Its formal definition is simply the expectation of X under the modified probability measure $\mathbf{Pr}' = \mathbf{Pr}[\,\cdot \mid A]$. More formally, in the discrete case,

$$\mathbf{E}[X \mid A] = \sum_{\omega \in \Omega} \mathbf{Pr}[\omega \mid A] X(\omega) = \sum_{\omega \in A} \frac{\mathbf{Pr}[\omega]}{\mathbf{Pr}[A]} X(\omega).$$

4 Some probability distributions of interest

We will encounter a small number of common distributions often and it is helpful to catalog them and some of their known properties. Often these distributions are parametrized. You can find a lot of useful information as well as pictures and examples on these distributions and much more on Wikipedia.

Bernoulli and Binomial distributions: Coin tosses are perhaps the simplest probability distribution that are extremely well-studied for their simplicity and applicability. Given a number $p \in [0,1]$ consider the two variable space $\{H, T\}$ with $\mathbf{Pr}[H] = p$ and $\mathbf{Pr}[T] = 1 - p$. This is called a Bernoulli distribution. Often we want to toss n independent coins. Then the probability space is of size 2^n corresponding to $\{H, T\}^n$. It is often simpler to use $\{1, 0\}$ instead of $\{H, T\}$ in which case the probability space is the set of all binary strings of length n. The Binomial distribution with parameters n and p is obtained by considering a random variable X where $X(\omega)$ is equal to the number of heads (or 0's) in the string ω. We see that $\mathbf{Pr}[X = k] = \binom{n}{k} p^k (1-p)^{n-k}$. Thus the Binomial distribution can thus be alternatively thought of as a distribution over $\{0, 1, 2, \ldots, n\}$ where $\mathbf{Pr}[i] = \binom{n}{k} p^k (1-p)^{n-k}$.

Suppose X has the Binomial distribution with parameters n, p. Then one way to understand X is via a sum of n independent Bernoulli random variables $Y_1, Y_2, \ldots, Y_n$ with parameter p. We then see that $\mathbf{E}[X] = \mathbf{E}\left[\sum_{i=1}^{n} Y_i\right] = np$. We also have that $\mathbf{Var}[X] = \sum_i \mathbf{Var}[Y_i] = np(1-p)$. Note that a direct computation with binomial coefficients would be much more messy.

Poisson distribution: The Poisson distribution has a single non-negative real parameter $\lambda > 0$ and is a discrete distribution over the non-negative integers $\{0, 1, 2, \ldots, \}$. For a random variable X distributed according to this distribution, $\mathbf{Pr}[X = i] = e^{-\lambda} \frac{\lambda^i}{i!}$. One can prove that $\mathbf{E}[X] = \mathbf{Var}[X] = \lambda$. Do you see why? The Poisson distribution is very useful for a variety of reasons. It is connected to the Binomial distribution as follows. It can be seen as the limit distribution of the Binomial distribution with parameters n, p where we take $n \to \infty$ while reducing p such that $np = \lambda$. Thus the Poisson distribution approximates the Binomial distribution when p is very small compared to n; this allows one to avoid messy computations with binomial coefficients and instead use the simpler formulas provided by the Poisson distribution.

One very interesting property satisfied by the Poisson distribution is the following. If $X_1, X_2, \ldots, X_n$ are independent Poisson random variables with means $\lambda_1, \lambda_2, \ldots, \lambda_n$ then $X = \sum_{i=1}^{n} X_i$ has the Poisson distribution with mean $\sum_{i=1}^{n} \lambda_i$.

Geometric distribution: This distribution is closely related to the Bernoulli distribution and hence has a single parameter p. There are two variants and both use the same name so one has to be careful. The first variant is over the support $\{1, 2, \ldots\}$ and the probability of i is the number of tosses of a Bernoulli random variable to see the first head. If X is distributed according to this then we see that $\mathbf{Pr}[X = i] = (1-p)^{i-1}p$. $\mathbf{E}[X] = 1/p$ and $\mathbf{Var}[X] = (1-p)/p^2$.

The second variant is to measure the number of failures before the first head. This is the same as above shifted by -1 and hence has support $\{0, 1, 2, \ldots\}$. For this second variant the expectation is $(1/p - 1)$ and variance is the same $(1-p)/p^2$.

Uniform distribution: The uniform distribution over a finite state space is pretty simple. If the state space has size k then each elementary event has probability $1/k$. This can be seen as a generalization of the Bernoulli distribution. Suppose we have n such independent distributions. Their joint distribution corresponds to the experiment of throwing n balls into k bins. Many aspects of balls and bins distributions are studied and we will see examples and applications and connections to hashing.

Uniform distribution is very important in continuous probability spaces. The most familiar one is to consider a uniform distribution over a closed interval $[a, b]$ of the real line. The density function f corresponding to this is $f(x) = 1/(b-a)$ for $x \in [a, b]$ and 0 outside the interval. In general we can define a uniform distribution over more complicated "shapes", especially in higher dimensions. Consider the uniform distribution over the unit circle in the Euclidean plane. Or over a specified rectangle in the plane. In such cases we have the density function $f(x, y)$ is constant over the interior of the shape where the constant depends on the normalization factor which is the area of the shape. Random variables can be used to modify the uniform distribution appropriately to create more complex distributions.

Normal distribution: One of the most fundamental probability distributions is the Normal or Guassian distribution. It is a distribution over the entire Euclidean space in any fixed dimension d. In one dimension it is defined by two parameters, the mean μ and the variance σ^2 (or standard deviation σ). The density function of this distribution is: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$. The standard Normal distribution is obtained by setting $\mu = 0$ and $\sigma = 1$ in which case it becomes the simpler expression $\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$. It is a distribution that is symmetric about the mean μ. If X is a random variable distributed according to the Normal distribution, by construction $\mathbf{E}[X] = \mu$ and $\mathbf{Var}[X] = \sigma^2$. The cumulative probability distribution $F_X(t) = \mathbf{Pr}[X \leq t]$ is often needed but there is no closed-form expression for it.

Acknowledgments: We thank Manuel Torres for reading and corrections.