Appendix B

Probability Theory

In this appendix we provide a brief overview of probability theory. In theoretical computer science and online algorithms in particular, we mostly deal with discrete probability spaces. Therefore, this appendix focuses on discrete probability theory.

B.1 Probability Space, Events, and Random Variables

Definition B.1.1. A *discrete probability space* is a pair (Ω, p), where

- Ω is called the *sample space* and is simply a finite or an infinitely countable set,
- p is called the *probability distribution* and is a function $p : \Omega \to \mathbb{R}_{\geq 0}$ satisfying $\sum_{\omega \in \Omega} p(\omega) = 1$.

Definition B.1.2. For a finite probability space (Ω, p), p is called *uniform* if for all $\omega \in \Omega$ we have $p(\omega) = \frac{1}{|\Omega|}$.

Definition B.1.3. *An event* E is a subset of Ω, i.e., $E \subseteq \Omega$.

Definition B.1.4. Given a set Ω, *the powerset* of Ω, denoted by 2^Ω, is defined as the set of all subsets of Ω, i.e.,

$$2^\Omega = \{E \mid E \subseteq \Omega\}.$$

Note: if Ω is a sample space, then its powerset is the set of all possible events.

Given a discrete probability space (Ω, p), we extend the probability distribution p defined on Ω to be defined on the powerset of Ω in the natural way: for any event $E \subseteq \Omega$ we define the probability of event E is

$$\Pr[E] = \sum_{\omega \in E} p(\omega).$$

Note that $\Pr[\{\omega\}] = p(\omega)$.

Definition B.1.5. Let E_1 and E_2 be two events such that $\Pr[E_2] \neq 0$. The probability of E_1 *conditioned on* E_2, denoted by $\Pr[E_1|E_2]$, is defined as

$$\Pr[E_1|E_2] = \frac{\Pr[E_1 \cap E_2]}{\Pr[E_2]}.$$

Theorem B.1.1. *Let $E_1, E_2, \ldots, E_n$ form a partition of the sample space Ω, that is for all $i \neq j$ we have $E_i \cap E_j = \emptyset$ and $\bigcup_{i\in[n]} E_i = \Omega$. Prove* the law of total probability*, which states for every event A we have:*

$$\Pr[A] = \sum_{i\in[n]} \Pr[A \cap E_i] = \sum_{i\in[n]} \Pr[A|E_i] \Pr[E_i],$$

where the second equality only holds when $\Pr[E_i] \neq 0$ for all i.

Definition B.1.6. Two events E_1 and E_2 are called *independent* if

$$\Pr[E_1 \cap E_2] = \Pr[E_1] \Pr[E_2].$$

Definition B.1.7. Events $E_1, \ldots, E_n$ are called *independent* if for every subset $I \subseteq [n]$ of events we have

$$\Pr\left[\bigcap_{i\in I} E_i\right] = \prod_{i\in I} \Pr[E_i].$$

Events $E_1, \ldots, E_n$ are called *pairwise* independent if for all $i \neq j$ we have

$$\Pr[E_i \cap E_j] = \Pr[E_i] \Pr[E_j].$$

Definition B.1.8. Let Γ be any set, and (Ω, p) be the discrete probability space. A Γ-valued *random variable* X is a function $X : \Omega \to \Gamma$. We are mostly interested in real-valued random variables, which means $\Gamma = \mathbb{R}$. From now on, when we write "random variable", we mean real-valued random variable, unless stated otherwise.

Observe that p itself is a random variable.

Definition B.1.9. Let X be a random variable and $x \in \mathbb{R}$ be a real number. Notation "$X = x$" is a short-hand for the event

$$\{\omega \mid X(\omega) = x\}.$$

The following exercise is both extremely easy and extremely important! It introduces the idea of the probability distribution *induced* on the image of a random variable X. Keeping in mind the distinction between the induced probability distribution and the true probability distribution can often avoid a lot of confusion in arguments based on probability theory.

Exercise B.1. *Let (Ω, p) be a discrete probability space and X be a random variable. Typically, X is not one-to-one. Let $\Im(X)$ denote the image of X. Observe that $\Im(X)$ is either finite or countable. For each $x \in \Im(X)$ we can define $\mu(x) := p(X = x)$.*

Prove that $(\Im(X), \mu)$ is a probability distribution. Note: μ is called the probability distribution *induced by p on $\Im(X)$.*

You may use the following exercise to check your understanding.

Exercise B.2. *Consider a randomized algorithm that takes n random unbiased independent bits and outputs their sum. What is the sample space and the probability distribution associated with this randomized experiment? Observe that the output of the algorithm can be viewed as a random variable. What's the definition of this random variable? What is the probability distribution induced on the output of the algorithm (i.e., the image of the associated random variable).*

Observe that going from (Ω, p) and X to $(\Im(X), \mu)$ loses information — in general, you cannot recover (Ω, p) from (X, μ) alone. In many scenarios, we are interested in the random variable itself and its distribution (i.e., $(\Im(X), \mu)$), and not the underlying probability space (i.e., (Ω, p)). In such applications of probability theory, you will often see $(\Im(X), \mu)$ being specified explicitly and the underlying probability space (Ω, p) omitted altogether! The unspoken agreement is that there is always an implicitly defined (Ω, p) that could be made explicit with some (often tedious) work. As can be seen, an explicitly defined (Ω, p) is usually not needed, since we can compute quantities of X (e.g., its moments) of interest from $(\Im(X), \mu)$ alone.

B.2 Expected Value, Independence, and Conditioning

Definition B.2.1. *The expected value* of a random variable X is defined as

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

Exercise B.3. *Prove that*

$$\mathbb{E}[X] = \sum_{x \in \Im(X)} x \Pr[X = x] = \sum_{x \in \Im(X)} x \mu(x).$$

The above exercise says that we don't need to know (Ω, p) to compute $\mathbb{E}[X]$ — it suffices to know $(\Im(X), \mu)$.

Exercise B.4. *Prove the* linearity *of expectation. That is for random variables X and Y and real numbers a and b we have*

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y].$$

Note: the above equality holds unconditionally regardless of what X and Y are. It generalizes to an any finite number of random variables via a straightforward induction.

Definition B.2.2. Random variables $X_1, \ldots, X_n$ are called independent if for all $x_1, \ldots, x_n \in \mathbb{R}$ the events $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$ are independent.

Exercise B.5. *Prove that if $X_1, \ldots, X_n$ are independent then $X_1, \ldots, X_{n-1}$ are independent.*

Exercise B.6. *Find two random variables X and Y such that*

$$\mathbb{E}[X \cdot Y] \neq \mathbb{E}[X] \cdot \mathbb{E}[Y],$$

where $X \cdot Y$ is the random variable that is the product of X and Y.

Exercise B.7. *Prove that if X and Y are* independent random variables *then we have*

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

Definition B.2.3. Random variables $X_1, \ldots, X_n$ are called *pairwise independent* if for all $i \neq j$ the variables X_i, X_j are independent.

Definition B.2.4. Let E be an event. *The indicator random variable* of event E, denoted by $\mathbb{I}_E$, is defined by

$$\mathbb{I}_E(\omega) = \begin{cases} 1 & \text{if } \omega \in E, \\ 0 & \text{otherwise} \end{cases}$$

Exercise B.8. *Prove that* the expected value *of an indicator random variable is equal to* the probability *of the event it is indicating. That is*

$$\mathbb{E}[\mathbb{I}_E] = \Pr[E].$$

B.3 Variance and Basic Inequalities in Probability Theory

Theorem B.3.1 (Markov's Inequality)**.** *Let X be a non-negative random variable, i.e., $X \geq 0$. For any $a > 0$ we have*

$$\Pr[X \geq a] \leq \frac{\mathbb{E}[X]}{a}.$$

Proof.

$$\begin{aligned}\mathbb{E}[X] &= \sum_{x \in \Im(X)} x \Pr[X = x] = \sum_{0 \leq x < a} x \Pr[X = x] + \sum_{x \geq a} x \Pr[X = x] \\ &\geq \sum_{x \geq a} x \Pr[X = x] \geq \sum_{x \geq a} a \Pr[X = x] = a \Pr[X \geq a].\end{aligned}$$

□

Definition B.3.1. *The variance* of a random variable X, denoted by $\mathrm{Var}(X)$, is defined as

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}(X))^2].$$

Exercise B.9. *Prove that the variance can alternatively be written as*

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2.$$

Exercise B.10. *Prove the following identity, where X is a random variable and a is a real number:*

$$\mathrm{Var}(aX) = a^2 \mathrm{Var}(X).$$

Exercise B.11. *Prove that if $X_1, \ldots, X_n$ are pairwise independent random variables, then the variance is additive:*

$$\mathrm{Var}(X_1 + \cdots + X_n) = \mathrm{Var}(X_1) + \cdots \mathrm{Var}(X_n).$$

Theorem B.3.2 (Chebyshev's Inequality)**.** *Let X be a random variable and a be a real number. Then we have*

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

Proof.

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

where the inequality step is justified by the Markov's Inequality applied to the nonnegative random variable $(X - \mathbb{E}[X])^2$. □

Theorem B.3.3 (Jensen's inequality)**.** *Let X be a random variable, and $f : \mathbb{R} \to \mathbb{R}$ be a* convex *function. Then we have*

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

Recall, that f is called convex if for all $x_1, x_2 \in \mathbb{R}$ and $t \in [0, 1]$ we have $f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2)$.

We state Jensen's inequality without a proof, although the following exercise asks you to prove it for finite probability spaces.

Exercise B.12. *Prove Jensen's inequality for finite probability spaces.* Hint: use the definition of convexity and induction.

Notation: we write i.i.d. to stand for independent identically distributed. This term is applied to a sequence of variables $X_1, \ldots, X_n$ and is self-explanatory.

Markov and Chebyshev inequalities are often used to bound the probability of tail events or large deviations of a random variable from the mean. The random variable of interest X often consists of a number of i.i.d. variables X_i, e.g., $X = \sum_{i=1}^{n} X_i$. In such cases, Markov and Chebyshev inequalities give bounds that decay inverse polynomially in n, i.e., $n^{-\Theta(1)}$. For Boolean X_i a much stronger exponential decay can be shown as well. This is known as the Chernoff bound. This is then often used in conjunction with a union bound over possibly exponentially many events to show that none of such events take place with high probability. We state two forms of the Chernoff bound without a proof.

Theorem B.3.4 (Chernoff bound (multiplicative form))**.** *Let $X_1, \ldots, X_n$ be i.i.d. random variables taking values in $\{0, 1\}$. Let $X = \sum_{i=1}^{n} X_i$. For any $\delta > 0$ we have*

$$\Pr[X \leq (1-\delta)\mathbb{E}[X]] \leq \exp\left(-\frac{\delta^2 \mathbb{E}[X]}{2}\right), \quad 0 \leq \delta \leq 1,$$

$$\Pr[X \geq (1+\delta)\mathbb{E}[X]] \leq \exp\left(-\frac{\delta^2 \mathbb{E}[X]}{2+\delta}\right), \quad 0 \leq \delta.$$

Theorem B.3.5 (Chernoff bound (additive form))**.** *Let $X_1, \ldots, X_n$ be i.i.d. random variables taking values in $\{0, 1\}$. Let $p = \mathbb{E}(X_i)$ and let $X = \sum_{i=1}^{n} X_i$. If $p \geq 1/2$ then for every $x > 0$ we have*

$$\Pr[X - np > x] \leq \exp\left(-\frac{x^2}{2np(1-p)}\right).$$

The above bounds can be generalized slightly to work with martingales resulting in the famous Azuma-Hoeffding inequality.

B.4 Martingales and Azuma-Hoeffding Inequality

A discrete-time martingale is a sequence of random variables arriving one at a time with some extra restrictions. The main restriction says that the expected value of the arriving random variable conditioned on all previous random variables is equal to the value of the most recent random variable. Formally, it is stated as follows.

Definition B.4.1. A sequence of random variables $X_0, X_1, \ldots$ is called *a discrete-time martingale* if for all i we have

- $\mathbb{E}[|X_i|] < \infty$
- $\mathbb{E}[X_{i+1} \mid X_0, X_1, \ldots, X_i] = X_i$

Let $Z_0, Z_1, \ldots$ be another sequence of random variables. Then $X_0, \ldots$ is called *a martingale with respect to the Z_i* if we have $\mathbb{E}[X_{i+1} \mid Z_0, Z_1, \ldots, Z_i] = X_i$ in addition to $\mathbb{E}[|X_i|] < \infty$.

In the above, we used the notion of the expectation of a random variable conditioned on another random variable. Before we define it formally, we define what it means for the expectation of a random variable to be conditioned on an event.

Definition B.4.2. Let X be a random variable, and E be an event such that $P(E) > 0$. *The expected value* of X conditioned on E is defined as

$$\mathbb{E}[X \mid E] = \sum_x x \Pr[X = x|E] = \sum_x \frac{x \Pr[X = x \cap E]}{\Pr[E]}.$$

Now, we can define what it means to condition on a random variable.

Definition B.4.3. Let X and Y be random variables. The expected value of X conditioned on Y, denoted by $\mathbb{E}[X \mid Y]$, is *a random variable* that is a function of values of y defined as follows:

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

Observe that on the right-hand side we have the expected value of X conditioned on *the event* "$Y = y$."

Exercise B.13. *Let X and Y be random variables. Prove the following identity:*

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

Exercise B.14. *Let X, Y and Z be random variables. Prove the following identity:*

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

Lastly, we state the Azuma-Hoeffding inequality without a proof.

Theorem B.4.1 (Azuma-Hoeffding inequality)**.** *Suppose X_i, $i \geq 0$, is a martingale (by itself or with respect to another sequence) such that there exists constants d_i such that $|X_i - X_{i-1}| \leq d_i$ almost surely (with probability 1). Then for every n and $x > 0$ we have*

$$\Pr[|X_n - X_0| \geq x] \leq 2 \exp\left(-\frac{x^2}{2\sum_{i=1}^n d_i^2}\right).$$

Let's go over an example of a problem that can be solved with the techniques just introduced. Suppose that you have n bins and you throw n balls into the bins. Each time you throw a ball, it has an equal chance of falling into one of the bins. How many empty bins are left after all n balls have been thrown?

Let's start by computing the expected value. Let Y denote the random variable that is equal to the number of empty bins at the end of the process. Let Z_i be the indicator random variable for the event that bin i is empty at the end of the process. Then we have $Y = \sum_{i=1}^n Z_i$. By linearity of expectation, we have $\mathbb{E}[Y] = \sum_{i=1}^n \mathbb{E}[Z_i] = n\mathbb{E}[Z_1]$, where the last equality follows since the Z_i are identically distributed. Now, using the fact that the expectation of the indicator random variable is equal to the probability of the event that the random variable indicates, we get that $\mathbb{E}[Z_1] = (1 - 1/n)^n \approx e^{-1}$. This is because the probability that ball j misses bin 1 is $1 - 1/n$, and all ball throws are independent. Therefore, we get that $\mathbb{E}[Y] = n(1 - 1/n)^n \approx n/e$.

Next, we show that the random variable Y is concentrated around its expected value. Let X_i denote the bin into which ball i lands. Define $Y_i = \mathbb{E}[Y \mid X_1, X_2, \ldots, X_i]$. Observe that $Y_0 = \mathbb{E}[Y] = n(1 - 1/n)^n \approx n/e$ and $Y_n = Y$. Also note that $|Y_i| \leq n$ for all i and

$$\mathbb{E}[Y_i \mid X_1, X_2, \ldots, X_{i-1}] = \mathbb{E}[\mathbb{E}(Y \mid X_1, \ldots, X_i) \mid X_1, \ldots, X_{i-1}] = \mathbb{E}[Y \mid X_1, \ldots, X_{i-1}] = Y_{i-1},$$

where the second equality follows from Exercise B.14. Thus, Y_i is a martingale with respect to X_i. Note that this construction is completely general — it is called *Doob's martingale*. Lastly, observe

that $|Y_{i+1} - Y_i| \leq 1$ since changing the bin into which ball $i+1$ lands can affect the number of non-empty bins by at most 1. Thus, applying Azuma-Hoeffding inequality to Y_i we get

$$\Pr[|Y_n - Y_0| \geq x] \leq \exp\left(-\frac{x^2}{2n}\right).$$

Thus, we get that $\Pr[|Y - n(1-1/n)^n| \geq c\sqrt{n}] \leq \exp(-c^2/2)$. Taking $c = \sqrt{2}\log n$ we get that the probability that the number of empty bins at the end of the process deviates from n/e by an additive term more than $\sqrt{2n}\log n$ is at most $1/n$.

B.5 Exercises

1. Let $E_1, E_2, \ldots, E_n$ be events in Ω. The following inequality is known as the *union bound*:

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

   Prove the union bound.

2. Let $E_1, E_2, \ldots, E_n$ be events in Ω. The following equation is known as the *inclusion-exclusion formula*:

$$\begin{aligned}\Pr\left[\bigcup_{i=1}^{n} E_i\right] &= \sum_{i=1}^{n} \Pr[E_i] - \sum_{1\leq i_1<i_2\leq n} \Pr[E_{i_1} \cap E_{i_2}] + \\ &\quad + \sum_{1\leq i_1<i_2<i_3\leq n} \Pr[E_{i_1} \cap E_{i_2} \cap E_{i_3}] - \cdots + (-1)^{n-1} \Pr[E_1 \cap E_2 \cap \cdots \cap E_n] \\ &= \sum_{k=1}^{n} (-1)^{k-1} \sum_{I\subseteq[n]:|I|=k} \Pr[E_I],\end{aligned}$$

   where $E_I := \cap_{i\in I} E_i$. Prove the inclusion-exclusion formula.

3. Let E_1 and E_2 be two events such that $\Pr[E_1], \Pr[E_2] \neq 0$. The following equation is known as the Bayes' rule:

$$\Pr[E_1|E_2] = \frac{\Pr[E_2|E_1]\Pr[E_1]}{\Pr[E_2]}.$$

   Prove it.

4. Is independence equivalent to $\Pr[E_1|E_2] = \Pr[E_1]$?

5. Does independence of n events imply their pairwise independence? If yes, prove it. If not, give a counter-example.

6. Does pairwise independence of n events imply their independence? If yes, prove it. If not, give a counter-example.

7. Consider an exact 3-SAT formula (meaning each clause contains exactly 3 variables) with n variables and m clauses, such that each variable is contained in at most k clauses. Generate an assignment by setting each variable to either 0 or 1 uniformly at random and independently of each other. What is the expected number of clauses satisfied by such random assignment? Use Doob's martingale together with Azuma-Hoeffding inequality to bound the probability that the number of satisfied clauses deviates from the mean by more than $\Theta(k\sqrt{n})$.