NOTES ON

Probability and Computing

Hao Su

Mainly based on *Probability and Computing* (2^{nd} ed.)

August 2, 2026

Contents

Chapter 1

Events and Probability

Example 1.1. Consider the following procedure that verifies if $F(x) \equiv G(x)$, where $F(x)$ is given as a product $F(x) = \prod_{i=1}^{d}(x - a_i)$ and $G(x)$ is given in its canonical form: assume that the maximum exponent of x in $F(x)$ and $G(x)$ is d, choose an integer r uniformly at random in the range $\{1, \dots, 100d\}$, and decide based on if $F(x) = G(x)$. Should an error occur then $F(x) \not\equiv G(x)$ and r is a root of $F(x) - G(x) = 0$, whose degree is no larger than d and thus has no more than d roots. Hence the chance of a wrong answer produced by this procedure is no more than $1/100$.

Definition 1.2. A *probability space* has three components:

1. a sample space Ω, which is the set of all possible outcomes of the random process modeled by the probability space;
2. a family of sets $\mathcal{F}$ representing the allowable events, where each set in $\mathcal{F}$ is a subset of the sample space; and
3. a probability function $\Pr : \mathcal{F} \to \mathbb{R}$ satisfying 1.3.

An element of Ω is called a *simple* or *basic* or *elementary* event.

In a discrete probability space $\mathcal{F} = 2^{\Omega}$, and Pr is uniquely defined by the probabilities of the simple events. The events need to be *measurable*, thus $\emptyset \in \mathcal{F}$ and $\mathcal{F}$ should be closed under complement and union and intersecton of countably many sets (a σ-algebra).

Definition 1.3. A *probability function* is any function $\Pr : \mathcal{F} \to \mathbb{R}$ that satisfies the following conditions:

1. for any event E, $0 \le \Pr(E) \le 1$;
2. $\Pr(\Omega) = 1$; and
3. for any finite or countably infinite sequence of pairwise mutually disjoint events

$E_1, E_2, E_3, \ldots,$

$$\Pr\left(\bigcup E_i\right) = \sum \Pr(E_i).$$

Lemma 1.4. *For any two events E_1 and E_2,*

$$\begin{aligned}
\Pr(E_1) &= \Pr(E_1 - (E_1 \cap E_2)) + \Pr(E_1 \cap E_2),\\
\Pr(E_2) &= \Pr(E_2 - (E_1 \cap E_2)) + \Pr(E_1 \cap E_2),\\
\Pr(E_1 \cup E_2) &= \Pr(E_1 - (E_1 \cap E_2)) + \Pr(E_2 - (E_1 \cap E_2)) + \Pr(E_1 \cap E_2).
\end{aligned}$$

by venn diagrams, which gives

$$\Pr(E_1 \cup E_2) = \Pr(E_1) + \Pr(E_2) - \Pr(E_1 \cap E_2).$$

Lemma 1.5. ***Union Bound*** *For any countable sequence of events $E_1, E_2, \ldots,$*

$$\Pr\left(\bigcup E_i\right) \leq \sum \Pr\left(E_i\right).$$

Lemma 1.6. ***Inclusion-Exclusion Principle*** *Let $E_1, \ldots, E_n$ be any n events. Then*

$$\begin{aligned}
\Pr\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i_1=1}^{n} \Pr\left(E_i\right) - \sum_{i_1<i_2} \Pr\left(E_i \cap E_j\right) + \sum_{i_1<i_2<i_3} \Pr\left(E_i \cap E_j \cap E_k\right)\\
- \cdots + (-1)^{l+1} \sum_{i_1<i_2<\cdots<i_l} \Pr\left(\bigcap_{r=1}^{l} E_{i_r}\right) + \cdots.
\end{aligned}$$

Definition 1.7. Tow events E and F are *independent* iff

$$\Pr\left(E \cap F\right) = \Pr\left(E\right) \cdot \Pr\left(F\right).$$

Events $E_1, E_2, \ldots, E_k$ are *mutually independent* iff for any $I \subseteq [1, k]$,

$$\Pr\left(\bigcap_{i \in I} E_i\right) = \prod_{i \in I} \Pr\left(E_i\right).$$

Definition 1.8. The *conditional probability* that E occurs given that F occurs is

$$\Pr(E \mid F) = \frac{\Pr(E \cap F)}{\Pr(F)},$$

which is well defined when $\Pr(F) > 0$.

This is very intuitive, looking for the probability of $E \cap F$ within the set of events defined by F. When E and F are independent and $\Pr(F) \neq 0$, we have

$$\Pr(E \mid F) = \frac{\Pr(E \cap F)}{\Pr(F)} = \frac{\Pr(E)\Pr(F)}{\Pr(F)} = \Pr(E).$$

We may sample with or without replacement in 1.1, and sampling without replacement slightly lowers the error rate. We may opt for sampling with replacement though, for easier analysis and implementation.

Example 1.9. Consider three $n \times n$ matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$, and assume we are working over integers modulo 2. We want to verify whether $\mathbf{AB} = \mathbf{C}$. Multiplying $\mathbf{A}$ and $\mathbf{B}$ takes roughly $\Theta(n^{2.37})$ operations. Instead we may choose a random vector $\overline{r} = (r_1, r_2, \dots, r_n) \in \{0,1\}^n$ and verify if $\mathbf{A}(\mathbf{B}\overline{r}) \neq \mathbf{C}\overline{r}$, which should take $\Theta(n^2)$ time in an obvious way. See 1.11 for analysis.

Theorem 1.10. ***Law of Total Probability*** *Let $E_1, E_2, \dots, E_n$ be mutually disjoint events in the sample space Ω, and let $\bigcup_{i=1}^n E_i = \Omega$. Then*

$$\Pr(B) = \sum_{i=1}^n \Pr(B \cap E_i) = \sum_{i=1}^n \Pr(B \mid E_i)\Pr(E_i).$$

Theorem 1.11. *In 1.9, if $\mathbf{AB} \neq \mathbf{C}$ and if $\overline{r}$ is chosen uniformly at random from $\{0,1\}^n$, then*

$$\Pr(\mathbf{AB}\overline{r} = \mathbf{C}\overline{r}) \leq \frac{1}{2}.$$

First Proof. Let $\mathbf{D} = \mathbf{AB} - \mathbf{C} \neq 0$. There exists $r' \in \{0,1\}^n$ where $\mathbf{D}r \neq 0$. For any $r \in \{0,1\}^n$, we have that if $\mathbf{D}r = 0$ then $\mathbf{D}(r + r') \neq 0$. Also, for any distinct $r_1, r_2 \in \{0,1\}^n$, we have that $r_1 + r' \neq r_2 + r'$. That is, $r \mapsto r + r'$ is injective and we hence have at least as many r's that make $\mathbf{D}r \neq 0$ as those do not. ⊣

Second Proof. Define $\mathbf{D}$ as in the first proof. W.L.O.G, let $d_{11} \neq 0$, where d_{11} is an entry of $\mathbf{D}$. For

$\mathbf{D}\overline{r} = 0$, it must be that case that $\sum_{j=1}^{n} d_{1j} r_j = 0$. Or equivalently,

$$r_1 = -\frac{\sum_{j=2}^{n} d_{1j} r_j}{d_{11}}.$$

Note that choosing $\overline{r} \in \{0,1\}^n$ uniformly at random is *equivalent* to choosing each r_i independently and uniformly from $\{0,1\}$ for that the latter process yields an $\overline{r}$ uniformly at random. Now, intead of reasoning about $\overline{r}$, suppose that we choose r_k independently and uniformly at random from $\{0,1\}$ in order, from r_n down to r_1, and consider the stage just before r_1 is chosen. That is, the sample space has collapsed to $\{0,1\}$. This idea, namely *principle of deferred decisions*, is formalized as follows, by 1.10. Note that r_1 making $\mathbf{D}\overline{r} = 0$ may not exist.

$$\begin{aligned}
&\Pr(\mathbf{AB}\overline{r} = \mathbf{C}\overline{r}) \\
&= \sum_{(x_2,\ldots,x_n)\in\{0,1\}^{n-1}} \Pr\left((\mathbf{AB}\overline{r} = \mathbf{C}\overline{r}) \cap ((r_2,\ldots,r_n) = (x_2,\ldots,x_n))\right) \\
&\leq \sum_{(x_2,\ldots,x_n)\in\{0,1\}^{n-1}} \Pr\left(\left(r_1 = -\frac{\sum_{j=2}^{n} d_{1j} r_j}{d_{11}}\right) \cap ((r_2,\ldots,r_n) = (x_2,\ldots,x_n))\right) \\
&= \sum_{(x_2,\ldots,x_n)\in\{0,1\}^{n-1}} \Pr\left(r_1 = -\frac{\sum_{j=2}^{n} d_{1j} r_j}{d_{11}}\right) \cdot \Pr((r_2,\ldots,r_n) = (x_2,\ldots,x_n)) \\
&\leq \sum_{(x_2,\ldots,x_n)\in\{0,1\}^{n-1}} \frac{1}{2} \Pr((r_2,\ldots,r_n) = (x_2,\ldots,x_n)) = \frac{1}{2}.
\end{aligned}$$

⊣

Theorem 1.12. ***Bayes' Law*** *Assume that $E_1, E_2, \ldots, E_n$ are muturally disjoint events in the sample space Ω such that $\bigcup_{i=1}^{n} E_i = \Omega$. Then*

$$\Pr\left(E_j \mid B\right) = \frac{\Pr\left(E_j \cap B\right)}{\Pr(B)} = \frac{\Pr\left(B \mid E_j\right)\Pr\left(E_j\right)}{\sum_{i=1}^{n} \Pr(B \mid E_i)\Pr(E_i)}.$$

Example 1.13. **Naïve Bayesian Classifier** Consider a training set D of size n, where each object x is represented as a features vector of length m, and a set C of possible classifications for objects. We want to model $\Pr(c \mid x)$. If each feature has two values we would need to estimate 2^m conditional probabilities per class, which is $|C|2^m$ probabilities in total. Now assume a 'naïve' model where the m features are independent, and we have

$$\begin{aligned}\Pr\left(c_j \mid x\right) &= \frac{\Pr\left(x \mid c_j\right) \cdot \Pr\left(c_j\right)}{\Pr(x)} \\ &= \frac{\prod_{k=1}^{m} \Pr\left(x_k \mid c_j\right) \cdot \Pr\left(c_j\right)}{\Pr(x)},\end{aligned}$$

indicating $O(m|C|)$ probabilities to estimate.

Definition 1.14. Min-Cut A *cut-set* in a graph is a set of edges whose removal breaks the graph into two or more connected components. Given a graph $G = (V, E)$ with n vertices, the minimum cut (or *min-cut*) problem is to find a minimum cardinality cut-set in G.

Algorithm 1.15. We randomly *contract* edges for $n - 2$ iterations. In contracting an edge (u, v) we merge the two vertices u and v into one vertex, eliminate all edges connecting u and v. The new graph may have parallel edges but no self-loops.

Theorem 1.16. *1.15 outputs a min-cut set with probability at least* $2/(n(n-1))$.

Proof. Let k be the size of the min-cut set of G. There may be several cut-sets with minimum size, each determining a partition of G. We consider one of such cut-sets C and let the partition be $\{S, V \setminus S\}$.

If, after $n-2$ iterations, no edges in C are contracted, the algorithm should return a correct answer. (Some intuition for randomness: if k is small, then it is not highly probable that we contract at random an edge in C.)

Let E_i be the event that the edge contracted in iteration i is not in C and let $F_i = \bigcap_{j=1}^{i} E_j$ the event that no edge of C was contracted in the first i iterations. Since the minimum cut-set has k edges, all vertices in the graph must have degree k or larger. If each vertex is adjacent to at least k edges, then the graph must have at least $nk/2$ edges. That is, the probability that we do not choose an edge of C in the first iteration is given by

$$\Pr(E_1) \geq 1 - \frac{k}{nk/2} = \frac{n-2}{n}.$$

Similarly we have

$$\Pr(E_i \mid F_{i-1}) \geq 1 - \frac{k}{k(n-i+1)/2} = \frac{n-i-1}{n-i+1}.$$

and

$$\begin{aligned}\Pr(F_{n-2}) &= \Pr(E_{n-2} \mid F_{n-3}) \cdot \Pr(E_{n-3} \mid F_{n-4}) \cdots \Pr(E_2 \mid F_1) \cdot \Pr(F_1) \\ &\geq \prod_{i=1}^{n-2} \left(\frac{n-i-1}{n-i+1}\right) \\ &= \left(\frac{n-2}{n}\right)\left(\frac{n-3}{n-1}\right)\left(\frac{n-4}{n-2}\right)\cdots\left(\frac{4}{6}\right)\left(\frac{3}{5}\right)\left(\frac{2}{4}\right)\left(\frac{1}{3}\right) \\ &= \frac{2}{n(n-1)}\end{aligned}$$

⊣

Comment. Since the algorithm has a one-sided error (can only outputs cut-sets with sizes greater than or equal to k), we can run it $n(n-1)\ln n$ times to bound the probability that the output is not a min-cut set by

$$\left(1 - \frac{2}{n(n-1)}\right)^{n(n-1)\ln n} \leq e^{-2\ln n} = \frac{1}{n^2},$$

for that $1 - x \leq e^{-x}$.

Exercises

Exercise 1.

to do

Chapter 2

Discrete Random Variables and Expectation

We are often interested in some value associated with the random event rather than in the event itself.

Definition 2.1. Let Ω be a sample space. A measurable function $X : \Omega \to \mathbb{R}$ is called a *random variable*. A discrete random variable takes on only countably many values. If X is discrete, denote $\{s \in \Omega \mid X(s) = a\}$ by "$X = a$".

Definition 2.2. Two discrete random variables X and Y are *independent* iff

$$\forall x \forall y,\ \Pr\left((X = x) \cap (Y = y)\right) = \Pr\left(X = x\right) \cdot \Pr\left(Y = y\right).$$

Similarly, discrete random variables $X_1, \ldots, X_k$ are *mutually independent* iff

$$\forall\, I \subseteq [k]\ \forall\, (x_i)_{i \in I},\ \Pr\left(\bigcap_{i \in I} \{X_i = x_i\}\right) = \prod_{i \in I} \Pr(X_i = x_i)\,.$$

Definition 2.3. The *expectation* of a discrete random variable X, denoted by $\mathbb{E}[X]$, is given by

$$\mathbb{E}[X] = \sum_{i \in \mathrm{ran}(X)} i \Pr(X = i)\,.$$

Say that the expectation is *finite* if it converges absolutely and *unbounded* if otherwise. Use $\mathbb{E}[X] = \infty$ to express that $\mathbb{E}[X]$ is unbounded.

Theorem 2.4. ***Linearity of Expectations*** *For any finite collection of discrete random* *variables $X_1, \dots, X_n$ with finite expectations,*

$$\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i]\,.$$

Proof. We prove the statement for two random variables X and Y; the general case follows by induction. The summations that follow are understood to be over the ranges of the corresponding random variables:

$$\begin{aligned}
\mathbb{E}[X+Y] &= \sum_i \sum_j (i+j)\Pr\big((X=i)\cap(Y=j)\big)\\
&= \sum_i \sum_j i\Pr\big((X=i)\cap(Y=j)\big) + \sum_i \sum_j j\Pr\big((X=i)\cap(Y=j)\big)\\
&= \sum_i i \sum_j \Pr\big((X=i)\cap(Y=j)\big) + \sum_j j \sum_i \Pr\big((X=i)\cap(Y=j)\big)\\
&= \sum_i i\Pr(X=i) + \sum_j j\Pr(Y=j)\\
&= \mathbb{E}[X] + \mathbb{E}[Y].
\end{aligned}$$

The first equality follows from 1.3, in that

$$(X+Y=t) = \bigcup_{\substack{i,j\\ i+j=t}} \{X=i\}\cap\{Y=j\}.$$

⊣

Comment. The variables *can* be dependent, that is, not (mutually) independent. C.f. 2.29.

Lemma 2.5. *For any constant c and discrete random variable X,*

$$\mathbb{E}[cX] = c\mathbb{E}[X]\,.$$

Proof. The lemma is obvious for $c = 0$. For $c \neq 0$,

$$\begin{aligned}
\mathbb{E}[cX] &= \sum_j j\Pr(cX=j)\\
&= \sum_{j/c} c(j/c)\Pr(X=j/c)\\
&= c\sum_{j/c} (j/c)\Pr(X=j/c)\\
&= c\mathbb{E}[X]\,.
\end{aligned}$$

⊣

Definition 2.6. A function $f : \mathbb{R} \to \mathbb{R}$ is *convex* if for any x_1, x_2 and $0 \le \lambda \le 1$,

$$f(\lambda x_1 + (1-\lambda)x_2) \le \lambda f(x_1) + (1-\lambda) f(x_2).$$

If f is twice differentiable, then f is convex iff $f''(x) \ge 0$.

Theorem 2.7. ***Jensen's Inequality*** *If f is a convex function, then*

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

Proof. We prove the theorem assuming that f has a Taylor expansion. (C.f. 2.10). Let $\mu = \mathbb{E}[X]$. By Taylor's theorem there is a value c such that

$$f(x) = f(\mu) + f'(\mu)(x-\mu) + \frac{f''(c)(x-\mu)^2}{2} \ge f(\mu) + f'(\mu)(x-\mu),$$

since $f''(c) > 0$ by convexity. Taking expectations of both sides yields the result:

$$\begin{aligned} \mathbb{E}[f(X)] &\ge \mathbb{E}[f(\mu) + f'(\mu)(X-\mu)] \\ &= \mathbb{E}[f(\mu)] + f'(\mu)(\mathbb{E}[X] - \mu) \\ &= f(\mu) = f(\mathbb{E}[X]). \end{aligned}$$ ⊣

Definition 2.8. **Bernoulli** Suppose that we run an experiment that succeeds with probability p and fails with probability $1-p$. Let Y be a random variable such that

$$Y = \begin{cases} 1 & \text{if the experiment succeeds,} \\ 0 & \text{otherwise.} \end{cases}$$

The variable Y is called a *Bernoulli* or an *indicator* random variable.

Definition 2.9. **Binomial** A *binomial* random variable X with parameters n and p, denoted by $B(n,p)$, is defined by the following probability distribution on $0 \le j \le n$:

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

The linearity of expectations yields a simple derivation for $\mathbb{E}[X] = np$.

Definition 2.10. The *conditional expectation* of Y given $\mathcal{E}$ is

$$\mathbb{E}[Y \mid \mathcal{E}] = \sum_{y \in \operatorname{ran}(Y)} y \Pr(Y = y \mid \mathcal{E}).$$

Lemma 2.11. *For any finite collection of discrete random variables $X_1, \ldots, X_n$ with finite expectations and for any event $\mathcal{E}$,*

$$\mathbb{E}\left[\sum_{i=1}^{n} X_i \mid \mathcal{E}\right] = \sum_{i=1}^{n} \mathbb{E}[X_i \mid \mathcal{E}].$$

The proof is almost the same as that of 2.4.

Note that $\mathbb{E}[X \mid Y]$ is a function of the random variable Y and hence also a random variable.

Theorem 2.12. *For any random variables X and Y,*

$$\begin{aligned}
\mathbb{E}[\mathbb{E}[X \mid Y]] &= \sum_{y \in \operatorname{ran}(Y)} \Pr(Y = y)\, \mathbb{E}[X \mid Y = y] \\
&= \sum_{y \in \operatorname{ran}(Y)} \left(\Pr(Y = y) \sum_{x \in \operatorname{ran}(X)} x \Pr(X = x \mid Y = y) \right) \\
&= \sum_{y \in \operatorname{ran}(Y)} \sum_{x \in \operatorname{ran}(X)} x \Pr(X = x \cap Y = y) \\
&= \sum_{x \in \operatorname{ran}(X)} \sum_{y \in \operatorname{ran}(Y)} x \Pr(X = x \cap Y = y) \\
&= \sum_{x \in \operatorname{ran}(X)} x \Pr(X = x) = \mathbb{E}[X]
\end{aligned}$$

if all of the expectations are finite.

Example 2.13. **Branching Process** Consider a program that includes one call to a process $\mathcal{S}$. Assume that each call to process $\mathcal{S}$ recursively spawns new copies of the process $\mathcal{S}$, where the number of new copies is a binomial random variable with parameters n and p. Assume that these random variables are independent for each call to $\mathcal{S}$ and ask: what is the expected number of copies of the process $\mathcal{S}$ generated by the program?
We would intuitively answer $\sum_{i \geq 0} (np)^i$. But as follows is how this can be derived.

Let Y_i denote the number of $\mathcal{S}$ in generation i. Suppose that $Y_{i-1} = y_{i-1}$. Let Z_k be the number of copies spawned by the kth process spawned in the $(i-1)$th generation for $1 \le k \le y_{i-1}$. Then

$$\begin{aligned}
\mathbb{E}[Y_i \mid Y_{i-1} = y_{i-1}] &= \mathbb{E}\left[\sum_{k=1}^{y_{i-1}} Z_k \mid Y_{i-1} = y_{i-1}\right] \\
&= \sum_{j \ge 0} j \Pr\left(\sum_{k=1}^{y_{i-1}} Z_k = j \mid Y_{i-1} = y_{i-1}\right) \\
&= \sum_{j \ge 0} j \Pr\left(\sum_{k=1}^{y_{i-1}} Z_k = j\right) \\
&= \mathbb{E}\left[\sum_{k=1}^{y_{i-1}} Z_k\right] \\
&= \sum_{k=1}^{y_{i-1}} \mathbb{E}[Z_k] = y_{i-1} np
\end{aligned}$$

The third line is by that the Z_k's and Y_{i-1} are all mutually independent. By 2.12

$$\begin{aligned}
\mathbb{E}[Y_i] = \mathbb{E}[Y_i \mid Y_{i-1}] &= \sum_{y_{i-1}} \Pr(Y_{i-1} = y_{i-1}) \, \mathbb{E}[Y_i \mid Y_{i-1} = y_{i-1}] \\
&= np \sum_{y_{i-1}} y_{i-1} \Pr(Y_{i-1} = y_{i-1}) = np\mathbb{E}[Y_{i-1}]
\end{aligned}$$

By induction we have $\mathbb{E}[Y_i] = (np)^i$ and conclude that

$$\mathbb{E}\left[\sum_{i \ge 0} Y_i\right] = \sum_{i \ge 0} \mathbb{E}[Y_i] = \sum_{i \ge 0} (np)^i.$$

Definition 2.14. Geometric A *geometric* random variable X with a parameter p is given by the following probability distribution on $n \in \mathbb{N}^+$:

$$\Pr(X = n) = (1-p)^{n-1} p.$$

Lemma 2.15. *For a geometric random variable X with parameter p and for $n > 0$,*

$$\begin{aligned}\Pr(X = n + k \mid X > k) &= \frac{\Pr(X = n + k \cap X > k)}{\Pr(X > k)} \\ &= \frac{(1-p)^{n+k-1}p}{(1-p)^k} \\ &= (1-p)^n p = \Pr(X = n)\,.\end{aligned}$$

Hence geometric random variables are called *memoryless*.

Lemma 2.16. *Let X be a discrete random variable that takes on only nonnegative integer values. Then*

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} \Pr(X \geq i)\,.$$

Proof. Note that the region of summation is $\{(i, j) \in \mathbb{N}^2 \,:\, 1 \leq i \leq j < \infty\}$. Also, the interchange of possibly infinite summations is justified due to the nonnegative summands.

$$\begin{aligned}\sum_{i=1}^{\infty} \Pr(X \geq i) &= \sum_{i=1}^{\infty}\sum_{j=i}^{\infty} \Pr(X = j) \\ &= \sum_{j=1}^{\infty}\sum_{i=1}^{j} \Pr(X = j) \\ &= \sum_{j=1}^{\infty} j \Pr(X = j) = \mathbb{E}[X]\end{aligned}$$

⊣

Remark 2.17. For a geometric random variable X with parameter p,

$$\Pr(X \geq i) = \sum_{n=i}^{\infty} (1-p)^{n-1} p = (1-p)^{i-1}.$$

By 2.16 we have that $\mathbb{E}[X] = 1/p$.

Remark 2.18. Consider a geometric random variable X with parameter p. By 2.12,

$$\begin{aligned}\mathbb{E}[X] &= \Pr(X > 1)\,\mathbb{E}[X \mid X > 1] + \Pr(X = 1)\,\mathbb{E}[X \mid X = 1] \\ &= (1-p)\mathbb{E}[X \mid X > 1] + p \cdot 1 \\ &= (1-p)\mathbb{E}[X - 1 + 1 \mid X > 1] + p = (1-p)\mathbb{E}[X - 1 \mid X > 1] + 1\end{aligned}$$

By 2.15 we have that $\mathbb{E}[X - 1 \mid X > 1] = \mathbb{E}[X]$, which gives $\mathbb{E}[X] = 1/p$.

Example 2.19. Coupon Collector's Problem Suppose that each box of cereal contains one of n different coupons. Once you obtain one of every type of coupon, you can send in for a prize. Assuming that the coupon in each box is chosen independently and uniformly at random from the n possibilities, we would like to determine the number of boxes bought until at least one of every type of coupon is obtained.

Let X denote that. If X_i is the number of boxes bought while one have exactly $i - 1$ different coupons, then $X = \sum_{i=1}^{n} X_i$. Note that each X_i is a geometric random variable with parameter $p_i = 1 - (i-1)/n$. Thus,

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i] = \sum_{i=1}^{n} \frac{n}{n-i+1} = n\sum_{i=1}^{n}\frac{1}{i}.$$

For harmonic number $H(n) = \sum_{i=1}^{n} 1/i$, we have that

$$H(n) - 1 = \sum_{k=1}^{n} \frac{1}{k+1} \leq \int_{i=1}^{n} \frac{1}{x}\,dx \leq \sum_{k=1}^{n} \frac{1}{k} = H(n). \tag{2.1}$$

Hence $\mathbb{E}[X] = n \ln n + \Theta(n)$.

Theorem 2.20. *Suppose that, whenever a pivot is chosen for Random Quicksort, it is chosen independently and uniformly at random from all possibilities. Then, for any input, the expected number of comparisons made by Random Quicksort is $2n \ln n + O(n)$.*

Proof. For convenience suppose that the inputs are all distinct. (A standard treatment is three-way partitioning, into `<pivot`, `=pivot`, and `>pivot`, which is not discussed here.) Let $y_1, \dots, y_n$ be the same values as the input values $x_1, \dots, x_n$ but sorted in increasing order. For $i < j$, let X_{ij} be a random variable that indicates whether y_i and y_j are compared at any time over the course of the algorithm. Note that two values can be *only* compared at most once over the course of the algorithm. Then

$$X = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} X_{ij}$$

and

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} X_{ij}\right] = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\mathbb{E}\left[X_{ij}\right] = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\Pr\left(X_{ij}=1\right).$$

Observe that y_i and y_j are compared iff either of them is the *first pivot* selected from the set $Y_{ij} = \{y_i, \dots, y_j\}$, which yields

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\frac{2}{j-i+1} \\
&= \sum_{i=1}^{n-1}\sum_{k=2}^{n-i+1}\frac{2}{k} \\
&= \sum_{k=2}^{n}\sum_{i=1}^{n+1-k}\frac{2}{k} \\
&= \sum_{k=2}^{n}(n+1-k)\frac{2}{k} \\
&= (2n+2)\sum_{k=1}^{n}\frac{1}{k} - 4n.
\end{aligned}$$

By 2.1 we are done. ⊣

Comment. Suppose instead that whenever a pivot is chosen for Quicksort, the first element of the sublist is chosen and that the input is chosen uniformly at random from all possible permutations of the origin input values, the expected number of comparisons made is still $2n\ln n + O(n)$. This is due to that the first pivot selected from Y_{ij} is the first element from Y_{ij} in the input, yielding the same possibility that y_i and y_j are compared over the course of the algorithm.

Exercises

Exercise 10. Jensen's Inequality

(a) Show by induction that if $f : \mathbb{R} \to \mathbb{R}$ is convex then, for any $x_1, x_2, \dots, x_n$ and $\lambda_1, \lambda_2, \dots, \lambda_n$ with $\sum_{i=1}^{n}\lambda_i = 1$,

$$f\left(\sum_{i=1}^{n}\lambda_i x_i\right) \le \sum_{i=1}^{n}\lambda_i f(x_i). \qquad (*)$$

(b) Use Eqn. (*) to prove that if $f : \mathbb{R} \to \mathbb{R}$ is convex then

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

for any random variable X that takes on only finitely many values.

Proof of (a). We use induction on n. The base case where $n = 1$ is trivial.

Assume true for $k - 1$. We have that

$$\begin{aligned}
\sum_{i=1}^{k} \lambda_i f(x_i) &= \sum_{i=1}^{k-2} \lambda_i f(x_i) + \lambda_{k-1} f(x_{k-1}) + \lambda_k f(x_k) \\
&= \sum_{i=1}^{k-2} \lambda_i f(x_i) + (\lambda_{k-1} + \lambda_k) \left(\frac{\lambda_{k-1}}{\lambda_{k-1} + \lambda_k} f(x_{k-1}) + \frac{\lambda_k}{\lambda_{k-1} + \lambda_k} f(x_k) \right) \\
&\geq \sum_{i=1}^{k-2} \lambda_i f(x_i) + (\lambda_{k-1} + \lambda_k) f \left(\frac{\lambda_{k-1}}{\lambda_{k-1} + \lambda_k} x_{k-1} + \frac{\lambda_k}{\lambda_{k-1} + \lambda_k} x_k \right) \\
&\geq f \left(\sum_{i=1}^{k-2} \lambda_i x_i + (\lambda_{k-1} + \lambda_k) \left(\frac{\lambda_{k-1}}{\lambda_{k-1} + \lambda_k} x_{k-1} + \frac{\lambda_k}{\lambda_{k-1} + \lambda_k} x_k \right) \right) \\
&= f \left(\sum_{i=1}^{k} \lambda_i x_i \right)
\end{aligned}$$

⊣

Proof of (b). The first equality in the following derivation is not immediate.

$$\mathbb{E}[f(X)] = \sum_{i \in \operatorname{ran}(X)} f(i) \Pr(X = i) \geq f \left(\sum_{i \in \operatorname{ran}(X)} i \Pr(X = i) \right) = f(\mathbb{E}[X])$$

⊣

Exercise 11. 2.11.

Exercise 18. **Reservoir Sampling** Suppose we have a sequence of items passing by one at a time. We want to maintain a sample of one item with the property that it is uniformly distributed over all the items that we have seen at each step. Moreover, we want to accomplish this without knowing the total number of items in advance or storing all of the items that we see.
Consider the following algorithm: we store at most one item in memory, and when the kth item appears, store it with probability $1/k$. Explain why this algorithm solves the problem.

The correctness of this algorithm can be proved by induction.

Exercise 19. Suppose that we modify the reservoir sampling algorithm of 2.18 so that, it at the beginning stores the first item, and when the kth item appears, it replaces the item in memory with probability $1/2$. Describe the distribution of the item in memory.

After the kth item appears, the ith item is stored with probability $1/2^{k+1-i}$, where $2 \le i \le k$. The first item is stored with probability $1/2^{k-1}$. This can also be proved by induction.

Exercise 27. Consider the following distribution on integers $x \ge 1 : \Pr(X = x) = (6/\pi^2)x^{-2}$, which is valid since $\sum_{k=1}^{\infty} k^{-2} = \pi^2/6$. What is $\mathbb{E}[X]$?

$$S_n = \sum_{i=1}^{n} i \cdot \frac{6}{\pi^2 i^2} = \frac{6}{\pi^2} \sum_{i=1}^{n} \frac{1}{i} = \frac{6}{\pi^2} \ln n + \Theta(1).$$

Therefore $\mathbb{E}[X] = \lim_{n\to\infty} S_n$ diverges.

Exercise 29. Show that

$$\mathbb{E}\left[\sum_{i=1}^{\infty} X_i\right] = \sum_{i=1}^{\infty} \mathbb{E}[X_i]$$

whenever $\sum_{i=1}^{\infty} \mathbb{E}[|X_i|]$ converges.

to do

Chapter 3

Moments and Deviations

We examine techniques for bounding the *tail distribution*, the probability that a random variable assumes values that are far from its expectation. In the context of analysis of algorithms, these bounds are the major tool for estimating the failure probability of algorithms and for establishing high probability bounds on their run-time.

Theorem 3.1. ***Markov's Inequality*** *Let X be a random variable that assumes only non-negative values. Then, for all $a > 0$,*

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

Proof. We have that

$$\begin{aligned}\mathbb{E}[X] &= \sum_x x \Pr(X = x) \geq \sum_{x \geq a} x \Pr(X = x) \geq a \sum_{x \geq a} \Pr(X = x) \\ &= a \Pr(X \geq a)\end{aligned}$$

⊣

Markov's inequality gives the best tail bound possible when all we know is the expectation of the random variable and that the variable is nonnegative. C.f. 3.16.

Definition 3.2. The kth *moment* of a random variable X is $\mathbb{E}\left[X^k\right]$.

Definition 3.3. The *variance* of a random variable X is defined as

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

The *standard deviation* of a random variable X is

$$\sigma[X] = \sqrt{\mathrm{Var}[X]}.$$

Definition 3.4. The *covariance* of two random variables X and Y is

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

Theorem 3.5. *For any two random variables X and Y,*

$$\mathrm{Var}[X + Y] = \mathrm{Var}[X] + \mathrm{Var}[Y] + 2\,\mathrm{Cov}(X, Y) \,.$$

Theorem 3.6. *If X and Y are two independent random variables, then*

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

Corollary 3.7. *If X and Y are independent random variables, then*

$$\mathrm{Cov}(X, Y) = 0$$

and hence

$$\mathrm{Var}[X + Y] = \mathrm{Var}[X] + \mathrm{Var}[Y] \,.$$

Theorem 3.8. *Let $X_1, \ldots, X_n$ be mutually independent random variables. Then*

$$\mathrm{Var}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathrm{Var}[X_i] \,.$$

This could be due to 3.7 and 3.14.

Example 3.9. To compute $\mathrm{Var}[X]$ where X is a binomial random variable with parameters n and p, we may compute by definition $\mathbb{E}\left[X^2\right]$, using a technique where we seek to apply the binomial theorem. Or we may think of X as the sum of n independent Bernoulli trials, each with success probability p. Such a Bernoulli trial Y has variance

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

By 3.8 we have that $\mathrm{Var}[X] = n(p - p^2)$.

Theorem 3.10. ***Chebyshev's Inequality*** *For any $a > 0$,*

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

Proof. By 3.1 we have that,

$$\begin{aligned}\Pr(|X - \mathbb{E}[X]| \geq a) &= \Pr\left((X - \mathbb{E}[X])^2 \geq a^2\right) \\ &\leq \frac{\mathbb{E}\left[(X - \mathbb{E}[X])^2\right]}{a^2} = \frac{\mathrm{Var}[X]}{a^2}\end{aligned}$$

⊣

Corollary 3.11. *For any $t > 1$,*

$$\begin{aligned}\Pr(|X - \mathbb{E}[X]| \geq t \cdot \sigma[X]) &\leq \frac{1}{t^2}, \\ \Pr(|X - \mathbb{E}[X]| \geq t \cdot \mathbb{E}[X]) &\leq \frac{\mathrm{Var}[X]}{t^2(\mathbb{E}[X])^2}.\end{aligned}$$

Exercises

Exercise 14. Prove that for any finite collection of random variables $X_1, \dots, X_n$,

$$\mathrm{Var}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathrm{Var}[X_i] + 2\sum_{i=1}^{n}\sum_{j>i} \mathrm{Cov}\big(X_i, X_j\big)\,.$$

Proof. Expand.

$$\begin{aligned}
\mathrm{Var}\left[\sum_{i=1}^{n} X_i\right] =& \mathbb{E}\left[\left(\sum_{i=1}^{n} X_i\right)^2\right] - \left(\mathbb{E}\left[\sum_{i=1}^{n} X_i\right]\right)^2 \\
=& \mathbb{E}\left[\sum_{i=1}^{n} X_i^2 + 2\sum_{i=1}^{n}\sum_{j>i} X_i X_j\right] \\
& - \left(\sum_{i=1}^{n} (\mathbb{E}[X_i])^2 + 2\sum_{i=1}^{n}\sum_{j>i} \mathbb{E}[X_i]\,\mathbb{E}\big[X_j\big]\right) \\
=& \sum_{i=1}^{n} \mathrm{Var}[X_i] + 2\sum_{i=1}^{n}\sum_{j>i} \mathrm{Cov}\big(X_i, X_j\big)\,.
\end{aligned}$$

⊣

Exercise 16. Markov's inequality is as tight as it could possibly be: given a positive integer k, describe a random variable X that assumes only nonnegative values such that

$$\Pr(X \geq k\mathbb{E}[X]) = \frac{1}{k}.$$

Given a *real number* $k \geq 1$ and its expectation $\mathbb{E}[X] \geq 0$, X takes on the value $k\mathbb{E}[X]$ with probability $1/k$ and the value 0 with probability $1 - 1/k$.

Exercise 17. C.f. 3.16. Can you give a similar example that shows that Chebyshev's inequality is tight? If not, explain why not.

to do