Recall that Alg_1 has a runtime of $T^{1/2}$, and Alg_2 has a runtime of $100T^{1/3}$. Taking the limit of the quotient, we get

$$\lim_{T\to\infty} \frac{|T^{1/2}|}{|100T^{1/3}|} = \lim_{T\to\infty} \frac{T^{1/6}}{100} = \infty\ .$$

By Proposition 2, it follows that $100T^{1/3} \in \mathcal{O}(T^{1/2})$, and $T^{1/2} \notin \mathcal{O}(100T^{1/3})$. This verifies our previous assertion that Alg_2 is better, according to big-oh analysis.

EXAMPLE 1.9

Suppose one algorithm has runtime $n!$, and another algorithm has runtime n^n. Which algorithm is faster, according to big-oh analysis? To answer this question, we will determine the value of $\lim_{n\to\infty} \frac{n!}{n^n}$. To do so, observe that

$$\lim_{n\to\infty} \frac{n!}{n^n} \cdot \frac{(n-1)^{n-1}}{(n-1)!} = \lim_{n\to\infty} \left(1 - \frac{1}{n}\right)^{n-1} = \frac{1}{e}\ ,$$

and so by the ratio test, $\lim_{n\to\infty} \frac{n!}{n^n} = 0$. By Proposition 2, it follows that $n! \in \mathcal{O}(n^n)$ and $n^n \notin \mathcal{O}(n!)$. This tells us that the algorithm with runtime $n!$ is better, according to big-oh analysis.

EXERCISES

1.2.1 Suppose you have the option of running three algorithms, Alg_1, Alg_2, Alg_3, with running times $\mathcal{O}(n \ln n)$, $\mathcal{O}(n^3)$, $300n^2$, respectively. Which algorithm would you select (i.e., which one is the fastest according to big-oh analysis) to run on input instances with arbitrary sizes n?

1.2.2 Show that $\log_{b_1}(n) \in \mathcal{O}\big(\log_{b_2}(n)\big)$ for any $b_1, b_2 > 1$.

1.2.3 Show that if $f(n) \in \mathcal{O}(g(n))$, then $\lambda f(n) \in \mathcal{O}(g(n))$ as well, for any $\lambda \in \mathbb{R}$.

1.2.4 Let $f, g, h : \mathbb{N} \to \mathbb{R}$. Show that if $f \in \mathcal{O}(h)$ and $g \in \mathcal{O}(h)$, then $f + g \in \mathcal{O}(h)$.

1.3 Probability

Suppose we roll a fair, 6-sided die (i.e., a die for which each outcome $1, \ldots, 6$ is equally likely). Then the probability of rolling a "3" is $\frac{1}{6}$. To put this statement into standard mathematical notation, we could let X represent the outcome of rolling the die, and say:

$$\mathbb{P}(\{X = 3\}) = \frac{1}{6}\ .$$

In this example, we call $\Omega = \{1, \ldots, 6\}$ the *outcome space* and X a *random variable*. The event "a three is rolled" is denoted by $\{X = 3\}$, and can formally be thought of as the set of outcomes for which a three is rolled (i.e., just $\{3\}$). Now consider the event $E = \{X \leq 3\}$ of rolling no higher than a 3. Then E corresponds to the set of outcomes

$\{1,2,3\} \subseteq \Omega$. Suppose we would like to know $\mathbb{P}(E)$. Since we assumed the die was fair, each outcome has probability $\frac{1}{6}$, which allows us to use *additivity of probabilities*:

Proposition 3: additivity of probabilities

Let E_1 and E_2 be disjoint events (i.e., $E_1 \cap E_2 = \emptyset$). Then $\mathbb{P}(E_1 \cup E_2) = \mathbb{P}(E_1) + \mathbb{P}(E_2)$.

Using Proposition 3, we can calculate the probability of event $E = \{X \leq 3\}$:

$$\begin{aligned}
\mathbb{P}(\{X \leq 3\}) &= \mathbb{P}(\{X=1\} \cup \{X=2\} \cup \{X=3\}) \\
&= \mathbb{P}(\{X=1\}) + \mathbb{P}(\{X=2\}) + \mathbb{P}(\{X=3\}) && \text{by Prop. 3} \\
&= \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}
\end{aligned}$$

EXAMPLE 1.10

As above, let X denote the outcome of rolling a fair 6-sided die. Then the event $\{X \text{ is even}\}$ is the same as the set of outcomes $\{2,4,6\} \subseteq \Omega$. We can calculate the probability of this event using additivity again:

$$\begin{aligned}
\mathbb{P}(\{X \text{ is even}\}) &= \mathbb{P}(\{X=2\} \cup \{X=4\} \cup \{X=6\}) \\
&= \mathbb{P}(\{X=2\}) + \mathbb{P}(\{X=4\}) + \mathbb{P}(\{X=6\}) && \text{by Prop. 3} \\
&= \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}
\end{aligned}$$

When events intersect, then Proposition 3 cannot be applied as is. For example, consider the events $E_1 = \{X \leq 3\}$ and $E_2 = \{X \text{ is odd}\}$. Then

$$\frac{2}{3} = \mathbb{P}(E_1 \cup E_2) \neq \mathbb{P}(E_1) + \mathbb{P}(E_2) = 1.$$

However, we could still apply Proposition 3 if we decompose $E_1 \cup E_2$ into disjoint sets. For instance, we can write

$$E_1 \cup E_2 = (E_1 \setminus E_2) \cup (E_2 \setminus E_1) \cup (E_1 \cap E_2). \tag{1.1}$$

It follows that for any (possibly intersecting) events E_1 and E_2:

$$\begin{aligned}
\mathbb{P}(E_1 \cup E_2) &= \mathbb{P}\big((E_1 \setminus E_2) \cup (E_2 \setminus E_1) \cup (E_1 \cap E_2)\big) \\
&= \mathbb{P}(E_1 \setminus E_2) + \mathbb{P}(E_2 \setminus E_1) + \mathbb{P}(E_1 \cap E_2) && \text{by (1.1)} \\
&= \mathbb{P}(E_1) + \mathbb{P}(E_2) - \mathbb{P}(E_1 \cap E_2).
\end{aligned}$$

In summary, we have just shown the following "inclusion-exclusion" property of probabilities:

Proposition 4: inclusion-exclusion for probabilities

Let E_1 and E_2 be any two events. Then $\mathbb{P}(E_1 \cup E_2) = \mathbb{P}(E_1) + \mathbb{P}(E_2) - \mathbb{P}(E_1 \cap E_2)$.

EXAMPLE 1.11 inclusion-exclusion for probabilities

Consider the events $E_1 = \{X \leq 3\}$ and $E_2 = \{X \text{ is odd}\}$ from above. In this case, $E_1 \cap E_2 = \{1, 3\}$, and so Proposition 4,

$$\mathbb{P}(E_1 \cup E_2) = \mathbb{P}(E_1) + \mathbb{P}(E_2) - \mathbb{P}(E_1 \cap E_2) = \frac{1}{2} + \frac{1}{2} - \frac{1}{3} = \frac{2}{3}.$$

Conditional probabilities and independence. Let E_1 and E_2 be two events. Suppose we would like to know: what is the probability of event E_1 *given* that event E_2 occurs? We refer this as a *conditional probability*.

Definition 6: conditional probability

Let E_1 and E_2 be two events, with $\mathbb{P}(E_2) \neq 0$. The probability of E_1 given E_2, denoted $\mathbb{P}(E_1 \mid E_2)$, is $\mathbb{P}(E_1 \mid E_2) = \frac{\mathbb{P}(E_1 \cap E_2)}{\mathbb{P}(E_2)}$.

Notice that the conditional probability $\mathbb{P}(E_1 \mid E_2)$ is obtained by intersecting E_1 with E_2, and normalizing by $\mathbb{P}(E_2)$. The reason that we intersect with E_2 is that we are assuming E_2 occurs, which means that we should ignore all outcomes outside of E_2. The reason that we normalize by $\mathbb{P}(E_2)$ is to ensure that $\mathbb{P}(E_2 \mid E_2) = 1$, which is intuitively what we want.

Condition probabilities provide one way to measure how "related" two events are. If two events E_1 and E_2 are fairly unrelated, then one would expect $\mathbb{P}(E_1)$ to be quite similar to $\mathbb{P}(E_1 \mid E_2)$ (in other words, the occurrence of E_2 doesn't have a big impact on the probability of E_1). When two events are completely unrelated, they are call *independent*, which we define precisely below.

Definition 7: independence

Events E_1 and E_2 are *independent* if $\mathbb{P}(E_1 \cap E_2) = \mathbb{P}(E_1) \cdot \mathbb{P}(E_2)$.

Notice that if E_1 and E_2 are independent, then

$$\mathbb{P}(E_1 \mid E_2) = \frac{\mathbb{P}(E_1 \cap E_2)}{\mathbb{P}(E_2)} = \frac{\mathbb{P}(E_1)\mathbb{P}(E_2)}{\mathbb{P}(E_2)} = \mathbb{P}(E_1) \, ,$$

which matches our intuitive notion of independence.

Caution: independence is *not* the same thing as disjointness. For instance, let X_1 be the roll of a fair 6-sided die, and X_2 the roll of a different fair 6-sided die. Consider the events $E_1 = \{X_1 \leq 2\}$, $E_2 = \{X_1 \geq 3\}$, and $E_3 = \{X_2 = 5\}$. Then E_1 and E_2 are disjoint, but not independent; on the other hand, E_1 and E_3 are independent, but not disjoint.

Expected value. Again, let X denote the roll of a fair die. We would like to know the average of X, i.e., the average number rolled by a fair die. This average is referred to as the *expected value* of X.

Definition 8: expected value

Let X be a random variable with finite outcome space $\Omega = \{\omega_1, \dots, \omega_n\} \subseteq \mathbb{R}$. Then the *expected value of* X, denoted $\mathbb{E}[X]$, is

$$\mathbb{E}[X] = \sum_{\omega \in \Omega} \mathbb{P}(X = \omega) \cdot \omega = \mathbb{P}(X = \omega_1) \cdot \omega_1 + \cdots + \mathbb{P}(X = \omega_n) \cdot \omega_n.$$

EXAMPLE 1.12 expected die roll

Let X be the outcome of a die roll. In this case, the outcome space is $\{1, \dots, 6\}$, and the expected value is

$$\mathbb{E}[X] = \sum_{i=1}^{6} \mathbb{P}(\{X = i\}) \cdot i = \sum_{i=1}^{6} \frac{i}{6} = \frac{7}{2}.$$

EXAMPLE 1.13 Bernoulli random variables

Let X have outcome space $\Omega = \{0, 1\}$, with $\mathbb{P}(\{X = 1\}) = p$ and $\mathbb{P}(\{X = 0\}) = 1 - p$. In other words, X is a *Bernoulli random variable with parameter* p. The expected value of X is

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

Suppose we know that $\mathbb{E}[X_1] = m_1$ and $\mathbb{E}[X_2] = m_2$. Can we calculate $\mathbb{E}[X_1 + X_2]$? One useful quality of expected values is that they respect linear combinations. In particular, we can use *linearity of expectation* to simplify the expected value of a linear combination of random variables.

Proposition 5: linearity of expectation

Let $X_1, \dots, X_n$ be random variables with finite outcome spaces, and let $\alpha_1, \dots, \alpha_n \in \mathbb{R}$. Then

$$\mathbb{E}[\alpha_1 X_1 + \cdots + \alpha_n X_n] = \sum_{i=1}^{n} \alpha_i \mathbb{E}[X_i].$$

EXAMPLE 1.14 binomial random variables

Let $X_1, \ldots, X_n$ be independent Bernoulli random variables with parameter p, and let $X = \sum_{i=1}^n X_i$. Then X is called a *binomial random variable with parameters n and p*. By linearity of expectation, we have

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

Continuous random variables. All the examples and propositions above were for random variables with finite outcome spaces, but the theory extends easily to continuous random variables. Continuous random variables typically have $\mathbb{R}$ (or some subset of $\mathbb{R}$) as an outcome space, and so the outcome space cannot be "summed over" as we did above in defining expected value. Instead, we must use the continuous analog of a sum, which is to say, an integral.

Before we can discuss the expected value of a continuous random variable. For many continuous random variables, singleton outcomes have probability 0. For example, if X is normally distributed, then $\mathbb{P}(\{X = x\}) = 0$ for any $x \in \mathbb{R}$. In order to get around this issue and measure probabilities in small intervals, we use *probability density functions*.

Definition 9: density function

Let X be a continuous random variable with outcome space $\mathbb{R}$. A *density function* for X, denoted $f_X(x)$, is a function satisfying

$$\mathbb{P}(\{X \in [a, b]\}) = \int_a^b f_X(x)\,\mathrm{d}x$$

for all $a \leq b \in \mathbb{R}$.

EXAMPLE 1.15 uniform random variables

Let X be a uniform random variable on $[0, 1]$, i.e., $X \sim \mathrm{Unif}[0, 1]$. Convince yourself that

$$f_X(x) = \begin{cases} 0 & x < 0 \\ 1 & 0 \leq x \leq 1 \\ 0 & 1 < x \end{cases}.$$

We are now ready to define the expected value of a continuous random variable.

Definition 10: expected value of a continuous random variable

Let X be a random variable with outcome space $\mathbb{R}$ and density function $f_X(x)$. Then, provided that $x f_X(x)$ is integrable over $\mathbb{R}$, the *expected value* of X is

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x)\,\mathrm{d}x.$$

EXAMPLE 1.16 expected value of a uniform random variable

Let $X \sim \mathrm{Unif}[a,b]$ be uniformly distributed over the interval $[a,b]$. Then

$$\mathbb{E}[X] = \int_a^b \frac{x}{b-a}\,\mathrm{d}x = \frac{a+b}{2}\,.$$

Work out the details to convince yourself that this is true!

Concentration inequalities. One core idea in machine learning is that by observing many data points, one can estimate the "average" value of the underlying distribution. The probabilistic idea that bolsters this is that sample means converge to true means. Here we discuss several useful inequalities that relate a random variable to its mean. We begin with Chebyshev's inequality, which gives quadratic decay of random variables from their means.

Proposition 6: Chebyshev's inequality

Let X be a continuous random variable with finite mean μ and finite variance σ^2. Then for any $t > 0$,

$$\mathbb{P}[\{|X-\mu| \geq t\sigma\}] \leq \frac{1}{t^2}.$$

Hoeffding's inequality is useful in bounding the convergence rate of a sample mean. In particular, Hoeffding's inequality shows exponential convergence to the true mean as the sample size increases.

Proposition 7: Hoeffding's inequality

Let $X_1, \ldots, X_n$ be independent random variables with outcome space $\Omega \subseteq [0,1]$, and let $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$ be the sample mean. Then

$$\mathbb{P}(\{|\overline{X} - \mathbb{E}[\overline{X}]| \geq t\}) \leq 2\exp(-2nt^2).$$

EXERCISES

1.3.1 Consider k independent Bernoulli variables $X_1, X_2, \ldots, X_k$ with parameters $0 < p_i < 1$ for $i = 1, \ldots, k$, i.e., $\mathrm{Prob}(X_i = 1) = p_i$. What is the expectation of the sum $\sum_{i=1}^k X_i$? What if these variables are not independent?

1.4 Linear algebra

In this course, we will often work over the vector space $\mathbb{R}^d$ (d-dimensional Euclidean space). In addition to being a vector space, $\mathbb{R}^d$ is equipped with an *inner product*, which provides the geometric backbone of $\mathbb{R}^d$. The inner product we will use is the standard dot product.