Intruduction to Discrete Probability

Mads Anker Nielsen
madsn20@student.sdu.dk

This document contains notes for an introductory lecture on discrete probability theory. The lecture is intended to be a somewhat self-contained introduction to the subject aimed at undergraduate computer science students. Most of the material covered is from Chapter 7 of "Discrete Mathematics and Its Applications" (8th edition) by Kenneth H. Rosen. These notes cover only a small subset of the important concepts in discrete probability theory, and the reader is encouraged to consult e.g. the aforementioned book for a more comprehensive treatment of the subject.

1 Fundamentals

A **sample space** S is a non-empty (possibly infinite) set. The elements of $s \in S$ are called **outcomes**. An **event** E in a sample space S is a set of outcomes or, equivalently, a subset of S. A **probability distribution** on a sample space S is a function $p : S \to \mathbb{R}$ satisfying

$$0 \leq p(s) \leq 1 \quad \text{for all } s \in S$$

and

$$\sum_{s \in S} p(s) = 1.$$

Given a probability distribution p on a sample space S and an event $E \subseteq S$, we define the **probability $p(E)$ of the event** E as

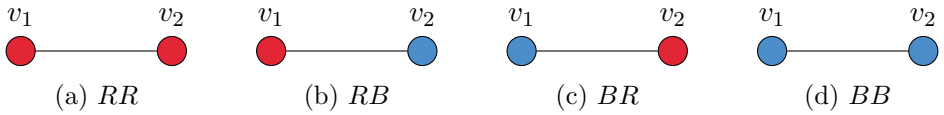
$$p(E) = \sum_{s \in E} p(s).$$

We can define a probability distribution however we like as long as we satisfy these requirements. However, we should define the probability distribution such that $p(s)$ is the ratio of the number of times s occurs to the number of times the experiment is performed when the number of times the experiment is performed tends to infinity.

Example 1

A coloring of a graph $G = (V, E)$ is a function c mapping each vertex $v \in V$ a color $c(v)$. In this example, we consider picking a 2-coloring $c : V(P_2) \to \{R, B\}$ of the path on 2 vertices P_2. We use R for red and B for blue (see Figure 1). Suppose we chose a coloring uniformly at random among all possible 2-colorings. That is, we pick each possible coloring with equal probability.

The sample space $S = \{RR, RB, BR, BB\}$ consists of all possible colorings. The probability distribution p on S is defined by $p(RR) = p(RB) = p(BR) = p(BB) = 1/4$. The event $E = \{RR, RB\}$ is the event that v_1 is red. The event $F = \{BR, RB\}$ is the event that the vertices receive different colors. The probability of E is $p(E) = p(RR) + p(RB) = 1/2$ and the probability of F is $p(F) = p(BR) + p(RB) = 1/2$.



(a) RR (b) RB (c) BR (d) BB

Figure 1: All possible 2-colorings of the path on 2 vertices.

Note that it is perfectly reasonable to let the outcomes in the sample space be any mathematical object we like. In Example 1, we represent a coloring (which we define as a function) by a string of two characters. In a more general setting, we might write $S = \{c | c \text{ is a 2-coloring of } P_2\}$ and treat $c \in S$ as a function.

In Example 1, the sample space is small enough that we can easily enumerate all outcomes explicitly. This is rarely, if ever, the case in real applications. Instead, we define events using standard set-builder notation. For example, we could define the event E that v_1 is red from Example 1 as $E = \{c \in S | c(v_1) = R\}$.

Perhaps slightly confusingly, we sometimes take probabilities of expressions which are strictly speaking not sets. For example, we might write $p(c(v_1) = R)$ or $p(c \text{ is a proper coloring})$. This is purely a notational convention. Formally, the probabilities referred to by these expressions

are $p(\{c \in S | c(v_1) = R\})$ and $p(\{c \in S | c \text{ is a proper coloring}\})$ respectively. Given that the formally correct notation is quite heavy and perhaps less readable, we often use the former. For the same reason, it is also common to define events by writing e.g. "Let E be the event that v_1 is red" instead of the more formal "Let $E = \{c \in S | c(v_1) = R\}$".

1.1 Intersections, unions, and complements of events

Events are subsets of the sample space. Therefore, we can apply standard set-theoretic operations to events to obtain new events. For example if $E, F \subseteq S$ are events, then $E \cup F$ is also an event. The intuitive interpretation of $E \cup F$ is the event that either E or F occurs since $E \cup F = \{s \in S | s \in E \text{ or } s \in F\}$. Likewise, $E \cap F$ is the event that both E and F occur since $E \cap F = \{s \in S | s \in E \text{ and } s \in F\}$.

Lemma 1

Let S be a sample space, p a probability distribution on S, and $E, F \subseteq S$ events. Then

$$p(E \cup F) = p(E) + p(F) - p(E \cap F).$$

Proof. By the definition of the probability of an event, the claim is equivalent to

$$\sum_{s \in E \cup F} p(s) = \sum_{s \in E} p(s) + \sum_{s \in F} p(s) - \sum_{s \in E \cap F} p(s).$$

Notice that both sides contain only terms of the form $p(s)$ for some $s \in E \cup F$. For each $s \in E \cup F$, we show that $p(s)$ occurs once on both sides of the equation. On the left-hand side, $p(s)$ occurs exactly once for each $s \in E \cup F$ by the definition of $p(E \cup F)$. We consider 3 cases for the right-hand side.

- If $s \in E$ and $s \notin F$, then $p(s)$ occurs only in the first sum.
- If $s \notin E$ and $s \in F$, then $p(s)$ occurs only in the second sum.
- If $s \in E$ and $s \in F$, then $p(s)$ occurs in all tree sums on the right-hand side and once with a negative sign. Thus, $p(s)$ occurs exactly once.

□

Example 2

Consider again the experiment from Example 1 of coloring P_2 the colors red and blue. What is the probability that v_1 is red or that the vertices receive different colors?

Recall the events $E = \{RR, RB\}$ and $F = \{BR, RB\}$, representing the events that v_1 is red and that the vertices receive different colors, respectively. The event $E \cup F$ is the event that v_1 is red or that the vertices receive different colors. $p(E) = 1/2$, $p(F) = 1/2$, and $p(E \cap F) = p(\{RB\}) = 1/4$. By Lemma 1,

$$p(E \cup F) = p(E) + p(F) - p(E \cap F) = 1/2 + 1/2 - 1/4 = 3/4.$$

Using Lemma 1, we can derive a simple upper bound on the probability that at least one of several events occurs. When working with discrete probability, it is very often the case that we are satisfied with bounds on the probability of an event rather than the exact answer. In some cases, we only care about the probability of an event being strictly larger than 0. In other cases, we might use probability theory to show that the asymptotic running time of an algorithm is polynomial in expectation, and we do not care about the exact constant factors.

Lemma 2: Union bound

Let S be a sample space, p a probability distribution on S, and $E_1, E_2, \ldots, E_n$ be any events in S. Then

$$p(E_1 \cup E_2 \cup \cdots \cup E_n) \leq p(E_1) + p(E_2) + \cdots + p(E_n).$$

Proof. By induction on n. For $n = 1$, the statement is $p(E_1) \leq p(E_1)$ which holds. Let $n \geq 2$ and let $E = E_1 \cup E_2 \cup \cdots \cup E_{n-1}$. By the induction hypothesis,

$$p(E) \leq p(E_1) + p(E_2) + \cdots + p(E_{n-1}).$$

By Lemma 1,

$$p(E \cup E_n) = p(E) + p(E_n) - p(E \cap E_n) \leq p(E) + p(E_n)$$

since $p(E \cap E_n) \geq 0$. Combining the two observations we get

$$p(E_1 \cup E_2 \cup \cdots \cup E_n) = p(E \cup E_n) \leq p(E) + p(E_n) \leq p(E_1) + p(E_2) + \cdots + p(E_n).$$

□

Despite its simplicity, the union bound has non-trivial consequences. Example 6 demonstrates this.

Lemma 2 holds with equality if the events $E_1, E_2, \ldots, E_n$ are pairwise disjoint (meaning $E_i \cap E_j = \emptyset$ for all distinct $i, j \in [n]$) as stated by the following lemma.

Lemma 3

Let S be a sample space, p a probability distribution on S, and $E_1, E_2, \ldots, E_n$ be parwise disjoint events in S. Then

$$p(E_1 \cup E_2 \cup \cdots \cup E_n) = p(E_1) + p(E_2) + \cdots + p(E_n).$$

The proof of Lemma 3 is very similar to the proof of Lemma 2 and is left as an exercise.

The complement of an event is the event that the original event does not occur. The following lemma gives a simple formula for the probability of the complement of an event.

Lemma 4

Let S be a sample space, p a probability distribution on S, and $E \subseteq S$ an event. Then

$$p(\overline{E}) = 1 - p(E).$$

Proof. Since p is a probability distribution, $p(S) = 1$. Furthermore, $E \cup \overline{E} = S$ and $E \cap \overline{E} = \emptyset$. By Lemma 1,

$$1 = p(S) = p(E \cup \overline{E}) = p(E) + p(\overline{E}) - p(E \cap \overline{E}) = p(E) + p(\overline{E})$$

and thus $p(E) = 1 - p(\overline{E})$. □

Lemma 4 nothing but a formalization of the intuitively simple observation that every event must either occur or not occur.

1.2 Conditional probability

Given information about the outcome of an experiment might change the probability that some event occurs. For example, suppose we are given that at least one vertex is red in the experiment from Example 1. Intuitively, the probability that v_1 is red should be larger given this information. Indeed, the only possible colorings that could have been chosen are RR, RB, and BR, and they were chosen with equal probability. Thus, the probability that v_2 is red is $2/3$ instead of the original $1/2$ when no information is given.

Definition 1

Let S be a sample space, p a probability distribution on S, and $E, F \subseteq S$ events with $p(F) > 0$. The **conditional probability of E given F** is denoted $p(E|F)$ and defined as

$$p(E|F) = \frac{p(E \cap F)}{p(F)}.$$

One can think of the conditional probability $p(E|F)$ as the probability that E occurs when the sample space is restricted to F and the probability $p(s)$ of each event $s \in S$ has been scaled by a factor of $1/p(F)$.

Example 3

Consider again the experiment of picking a random 2-coloring of a path on 2 vertices. Given that at least one vertex is red, what is the probability that v_1 is red?

The event that at least one vertex is red is $F = \{RR, RB, BR\}$. The event that v_1 is red $E = \{RR, RB\}$. We have $E \cap F = \{RR, RB\}$, thus $p(E \cap F) = 1/2$, $p(E) = 2/4$, $p(F) = 3/4$, and

$$p(E|F) = \frac{p(E \cap F)}{p(F)} = \frac{1/2}{3/4} = 2/3.$$

1.3 Independence

In the previous section, we saw that the information that F occurs can affect the probability that some other event E occurs. However, this is not always the case. It might be that F occurring does not change the probity of E occurring. In such cases, we say that the events E and F are *independent*.

Definition 2

Let S be a sample space, p a probability distribution on S, and E, F events from S. E and F are **independent** if and only if

$$p(E \cap F) = p(E)p(F).$$

Perhaps the statement $p(E|F) = p(E)$ or $p(F|E) = p(F)$ more closely matches the intuitive interpretation of independence, and the statements are in fact equivalent. Indeed,

$$p(E|F) = p(E) \Leftrightarrow p(E) = \frac{p(E \cap F)}{p(F)} \Leftrightarrow p(E)p(F) = p(E \cap F).$$

Example 4

Consider again the experiment of picking a random 2-coloring of P_2. The event $E = \{RB, RR\}$ that v_1 is red and the event $F = \{BR, RB\}$ that the coloring is proper are independent. Indeed,

$$p(E \cap F) = p(\{RB\}) = 1/4 = 1/2 \cdot 1/2 = p(E)p(F).$$

Independence can be extended to more than two events.

Definition 3

Let S be a sample space, p a probability distribution on S, and $E_1, E_2, \ldots, E_n \subseteq S$ events. If it for any distinct $i, j \in [n]$ it holds that $p(E_i \cap E_j) = p(E_i)p(E_j)$, then the events $E_1, E_2, \ldots, E_n$ are **pairwise independent**. If for any subset $I \subseteq [n]$ it holds that $p(\cap_{i \in I} E_i) = \prod_{i \in I} p(E_i)$, then $E_1, E_2, \ldots, E_n$ are **mutually independent**.

Pairwise and mutual independence of a set of events are not equivalent definitions. Intuitively, events are pairwise independent if knowing that any given event occurs does not change the probability of other events occurring. Mutual independence is a stronger condition that requires that knowing that any subset of the events occurs does not change the probability of any other event in the set occurring.

Example 5

Consider the experiment of picking a random 2-coloring from Example 1. The event E_1 of v_1 begin red, E_2 of v_2 being blue, and E_3 of the vertices receiving different colors are pairwise independent and all occur with probability 1/2. However, they are not mutually independent since given that E_1 and E_2 occur, E_3 occurs with probability 1. Formally, $p(E_1 \cap E_2 \cap E_3) = p(E_1)p(E_2) \neq p(E_1)p(E_2)p(E_3)$.

Example 6

A k-CNF formula $C = C_1 \wedge C_2 \wedge \ldots C_m$ over a set of variables $x_1, x_2, \ldots, x_n$ is a conjunction of clauses where each clause is a disjunction of exactly k literals. For example,

$$(x_1 \vee x_2 \vee \neg x_3) \wedge (\neg x_1 \vee x_2 \vee x_3)$$

is a 3-CNF formula with 2 clauses and 3 variables. An assignment is a function $\phi : \{x_i\}_{i \in [n]} \to \{T, F\}$. An assignment ϕ satisfies x_i if $\phi(x_i) = T$ and satisfies $\neg x_i$ if $\phi(x_i) = F$. An assignment satisfies a clause if it satisfies at least one literal in the clause and satisfies a formula if it satisfies all clauses in the formula. A formula is satisfiable if there exists an assignment that satisfies it.

In this example, we use the union bound to show that any k-CNF formula with less than 2^k clauses is satisfiable.

Let C be a k-CNF formula with $m < 2^k$ clauses. Denote by $C_{j,l}$ the l-th literal in the j-th clause of C. Consider picking a random assignment ϕ where

$$\phi(x_i) = \begin{cases} T & \text{with probability } 1/2 \\ F & \text{with probability } 1/2 \end{cases}$$

independently for each $i \in [n]$. For each $j \in [m]$ and $l \in [k]$, let $E_{j,l}$ be the event that $C_{j,l}$ is not satisfied by ϕ. Formally,

$$E_{i,j} = \{\phi |\ \phi \text{ does not satisfy } C_{j,l}\}.$$

The event

$$E_i = E_{i,1} \cap E_{i,2} \cap \cdots \cap E_{i,k}$$

is the event that C_i is not satisfied by ϕ. The events $E_{j,1}, E_{j,2}, \ldots, E_{j,k}$ are mutually independent by the way in which ϕ is chosen. Thus,

$$p(E_i) = p(E_{i,1})p(E_{i,2}) \ldots p(E_{i,k}) = \frac{1}{2^k}.$$

The event $E = E_1 \cup E_2 \cup \cdots \cup E_m$ is the event that some clause is not satisfied by ϕ and hence the event that the formula is not satisfied. By the union bound,

$$p(E) \leq p(E_1) + p(E_2) + \cdots + p(E_m) = m \cdot \frac{1}{2^k} < 1$$

since $m < 2^k$. The event $\overline{E}$ is the event that the formula is satisfied. By Lemma 4, $p(\overline{E}) = 1 - p(E) > 0$, so there exists an assignment that satisfies C.

2 Random variables

Random variables are a tremendously useful tool for modelling problems concerned with numerical values associated with outcomes. For example, we might be concerned with the number of successes in a sequence of experiments or the sum of outcomes when repeating an experiment multiple times.

Definition 4

A random variable $X : S \to \mathbb{R}$ on a sample space S is a function assigning real numbers to the outcomes of S.

The name "random variable" is somewhat misleading. A random variable is a function from the sample space to the real numbers. It is neither a variable nor random in any precise sense.

Example 7

On the same space $S = \{RR, RB, BR, BB\}$ from Example 1, the function

$$X(c) = |\{v \in V | c(v) = R\}|$$

is a random variable which maps a coloring c to the number of vertices colored red by c. On the sample space $S = \{T, F\}^n$ of all assignments from Example 6, the function

$$Y(\phi) = |\{j \in [m] | \phi \text{ satisfies } C_j\}|$$

is a random variable which maps an assignment ϕ to the number of clauses satisfied by ϕ.

On the sample space $S = \{1, 2, 3, 4, 5, 6\}$ which could be used to model the outcome of a die roll, the function

$$Z(s) = s$$

is a random variable which maps an outcome s to the value of the roll assuming that $s \in S$ actually represents the outcomes of rolling an s (anything else would certainly be a strange choice).

2.1 Expected value

Definition 5

Let X be a random variable on a sample space S with probability distribution p. The **expected value** of X is defined as

$$\mathrm{E}[X] = \sum_{s \in S} X(s)p(s).$$

The following lemma establishes a more convenient formula for the expected value of a random variable.

Lemma 5

Let S be a sample space, p a probability distribution on S, and X a random variable on S.

Denote by $X(S)$ the range of X (the set of values that X may take). Then

$$\mathrm{E}[X] = \sum_{x \in X(S)} xp(X = x).^a$$

[a] Here, we use the slightly abusive $p(X = x)$ to mean $p(\{x \in X(S) | X = x\})$ as remarked in the beginning.

Proof. The sum from the definition of expectation sums over all outcomes in the sample space. We split this sum into an outer part summing over all values x in the range $X(S)$ of X and an inner part summing over all the outcomes for which X takes the value x. Formally,

$$\begin{aligned}
\mathrm{E}[X] &= \sum_{s \in S} X(s)p(s) \\
&= \sum_{x \in X(S)} \sum_{s \in \{s | X(s) = x\}} X(s)p(s) \\
&= \sum_{x \in X(S)} x \sum_{s \in \{s | X(s) = x\}} p(s) \\
&= \sum_{x \in X(S)} xp(X = x).
\end{aligned}$$

□

One particularly useful kind of random variable is the indicator random variable. An indicator random variable is a random variable that takes only the values 0 and 1.

Lemma 6

Let S be a sample space, p a probability distribution on S, and X an indicator random variable. Then

$$\mathrm{E}[X] = p(X = 1).$$

Proof. Using the more convenient formula from Lemma 5 we have

$$\mathrm{E}[X] = \sum_{x \in X(S)} X(s)p(s) = \sum_{x \in \{0,1\}} xp(X = x) = p(X = 1).$$

□

2.2 Linearity of expectation

We stress that random variables are functions from the sample space to the real numbers and can be manipulated as such. The sum of two functions from a sample space S to the real numbers is a function from S to the real numbers. Thus, the sum of two random variables is a random variable. For the same reason, we can multiply random variables by each other or by any real number to obtain a new random variable.

The following lemma is *very* useful and used extensively.

Lemma 7: Linearity of expectation

Let S be a sample space, p a probability distribution on S, and $X_1, X_2, \ldots, X_n$ be random

variables on S. Then

$$\mathrm{E}[X_1 + X_2 + \cdots + X_n] = \mathrm{E}[X_1] + \mathrm{E}[X_2] + \cdots + \mathrm{E}[X_n]$$

Proof. By induction on n. For $n = 1$ the statement is $\mathrm{E}[X_1] = \mathrm{E}[X_1]$ which holds. Let $n \geq 2$ and let $X = X_1 + X_2 + \cdots + X_{n-1}$. By the induction hypothesis, $\mathrm{E}[X] = \mathrm{E}[X_1] + \mathrm{E}[X_2] + \cdots + \mathrm{E}[X_{n-1}]$. Now, using the definition of expectation and the induction hypothesis,

$$\begin{aligned}
\mathrm{E}[X_1 + X_2 + \cdots + X_n] &= \mathrm{E}[X + X_n] \\
&= \sum_{s \in S} (X(s) + X_n(s)) p(s) \\
&= \sum_{s \in S} X(s) p(s) + \sum_{s \in S} X_n(s) p(s) \\
&= \mathrm{E}[X] + \mathrm{E}[X_n] \\
&= \mathrm{E}[X_1] + \mathrm{E}[X_2] + \cdots + \mathrm{E}[X_n].
\end{aligned}$$

□