EE 126 Probability and Random Processes
January 16, 2018 Sample Space and Probability

1 Reading

Probability can be defined as **a measure of the likelihood that an event will occur**, presumably in a context of uncertainty. One might describe this measure in terms of **frequency of occurrence** (percentage of positive outcomes across a large number of representative situations) or in terms of **subjective belief** (an individual's own degree of confidence, based in reason or not, that the event will occur).

The hope is that readers of *Introduction to Probability* will be able to reason about uncertainty using probabilistic models. In this section, the structure and properties of such models will be described.

1.1. Sets

Probabilistic models assign probabilities to **sets** of possible outcomes, which is why we care about sets in the first place. In general, a set is a collection of elements. The empty set is denoted by $\emptyset$. The universal set, which contains all objects that could conceivably be of interest in a certain context, is denoted by Ω. All sets in this context will then be subsets of the universal set (for every set S, $S \subset \Omega, \Omega \supset S$).

A **countably infinite set** contains infinitely many elements but can still be enumerated in a list ending in .... The set of even integers $\{0, 2, -2, 4, -4, ...\}$ is countably infinite. An **uncountable set** takes on a continuous range of values and cannot be enumerated in a list. These sets must be written as "the set of all x that satisfy a property P", i.e. $\{x \mid 0 \leq x \leq 1\}$.

Two sets S and T are equal if $S \subset T$ and $T \subset S$.

1.1.1. Set Operations

- The **complement** of set S (w.r.t. the universe Ω) is the set $\{x \in \Omega \mid x \notin S\}$, or all elements of Ω that are not in S. It is denoted by S^c. The complement of the universal set is the empty set. Roughly equivalent to "not."
- The **union** of two sets S and T, $S \cup T$, is the set $\{x \mid x \in S \text{ or } x \in T\}$ of all elements in either S or T. Roughly equivalent to "or."
- The **intersection** of two sets S and T, $S \cap T$, is the set $\{x \mid x \in S \text{ and } x \in T\}$ of all elements in both S and T. Several sets are **disjoint** if no two contain a common element (i.e. each pairwise intersection is equal to the empty set). Roughly equivalent to "and."

A **partition** of set S is a collection of disjoint sets whose union is S.

1.1.2. Properties of Set Operations

- $S \cup T = T \cup S$
- $S \cup (T \cup U) = (S \cup T) \cup U$
- $S \cap (T \cup U) = (S \cap T) \cup (S \cap U)$
- $S \cup (T \cap U) = (S \cup T) \cap (S \cup U)$

De Morgan's laws yield two especially useful properties:

- $(\bigcup S_i)^c = \bigcap S_i^c$ (the **complement** of the union of sets is equal to the **intersection** of the complements of sets; both describe the elements that are outside of all sets)
- $(\bigcap S_i)^c = \bigcup S_i^c$ (the **complement** of the intersection of sets is equal to the **union** of the complements of sets; both describe the elements that are not shared by all sets)

1.2. Probabilistic Models

A **probabilistic model** is a mathematical description of an uncertain situation. It must include a **sample space** Ω, or the set of all possible outcomes, and a **probability law**, which assigns to a set A of possible outcomes a nonnegative probability $P(A)$. $P(A)$ encodes a belief about the collective likelihood of the elements in A.

Note that another word for A (as used above) is "event." An **event** is a set of possible outcomes (i.e. a subset of the sample space).

The "uncertain situation" mentioned earlier is an **experiment** which will produce exactly **one** of several possible outcomes. In our formulation, there is only one experiment per probabilistic model. Since there can only be one unique outcome, elements of the sample space should be distinct and mutually exclusive. Also, the sample space should be collectively exhaustive, such that any conceivable outcome will be contained in the sample space.

Often, experiments can be represented via a **tree-based sequential description**, where each outcome is associated with a leaf of the tree and is depicted by the unique path to that leaf from the root. Each node of the tree can also be identified with an event (all leaves branching out of that node). This description makes sense for experiments of an inherently sequential character, like receiving eight successive digits at a communication receiver.

1.2.1 Probability Axioms

- For every event A, $P(A) \geq 0$.
- If $A_1, A_2, ...$ is a sequence of **disjoint** events, then the probability of their union satisfies $P(A_1 \cup A_2 \cup ...) = P(A_1) + P(A_2) + ...$
- The probability of the entire sample space Ω is equal to 1.

1.2.2 Discrete Models

Discrete Probability Law
If the sample space is finite, then the probability law is specified by the probabilities of single-element events. Namely, $P(\{s_1, s_2, ..., s_n\}) = P(s_1) + P(s_2) + ... + P(s_n)$.

Discrete Uniform Probability Law
If the sample space consists of n equally likely outcomes (i.e. all single-element events have the same probability), then the probability of any event A is given by $P(A) =$ (number of elements in A) / n.

1.2.3 Properties of Probability Laws

Let A, B, and C be events.

- If $A \subset B$, then $P(A) \leq P(B)$.
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- **Union bound:** $P\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} P(A_i)$
- $P(A \cup B \cup C) = P(A) + P(A^c \cap B) + P(A^c \cap B^c \cap C)$

1.3. Conditional Probability

Conditional probability enables us to reason about uncertainty in the presence of partial information. We are assumed to possess an experiment, a sample space, and a probability law, and we know that the outcome is within some given event B. We want to quantify the likelihood that the outcome also belongs to some other given event A. We must construct a new probability law that factors in the available knowledge, and then for any event A specifies the conditional probability of A given B [denoted $P(A|B)$].

By the definition of conditional probability, $P(A|B) = P(A \cap B)/P(B)$ (accordingly, if $P(B) = 0$ then $P(A|B)$ is undefined). In other words, out of the total probability of the elements of B, $P(A|B)$ is the fraction assigned to the outcomes that **also** belong to A. This specifies a new [conditional] probability law on the same sample space Ω. All properties of probability laws remain valid.

In short, the conditional probability $P(A|B)$ captures the partial information that [the occurrence of] event B provides about [the occurrence of] event A, and is read "the probability of A given B."

Multiplication Rule (useful in tree-based sequential description)
$P(\cap_{i=1}^{n} A_i) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \cdots P(A_n| \cap_{i=1}^{n-1} A_i)$

1.4. Total Probability Theorem and Bayes' Rule

Total Probability Theorem
Let $A_1, ..., A_n$ be disjoint events that form a partition of the sample space (each possible outcome is included in exactly one of the events $A_1, ..., A_n$) and assume that $P(A_i) > 0$ for all i. Then, for any event B, we have
$P(B) = P(A_1 \cap B) + ... + P(A_n \cap B) = P(A_1)P(B|A_1) + ... + P(A_n)P(B|A_n)$.

This is because we can decompose B into the disjoint union of its intersections with the sets A_i, i.e. $B = (A_1 \cap B) \cup ... \cup (A_n \cap B)$. Such a theorem is useful, for example, when we can structure our desired probability as a conditional over multiple events (e.g. we want $P(B)$ and we have $P(B|A_1), P(B|A_2), P(B|A_3)$ where A_1, A_2, A_3 are disjoint events comprising a partition of the sample space).

1.4.1. Inference and Bayes' Rule

Bayes' Rule relates conditional probabilities of the form $P(A|B)$ with conditional probabilities of the form $P(B|A)$. (Notice that the order of the conditioning is reversed.)

Bayes' Rule
Let $A_1, ..., A_n$ be disjoint events that form a partition of the sample space (each possible outcome is included in exactly one of the events $A_1, ..., A_n$) and assume that $P(A_i) > 0$ for all i. Then, for any event B such that $P(B) > 0$, we have

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)} = \frac{P(B|A_i)P(A_i)}{P(A_1)P(B|A_1) + ... + P(A_n)P(B|A_n)}$$

Often Bayes' rule will be used for **inference**. In inference, a number of causes may result in a certain effect. We observe the effect and wish to infer the cause. If the inference context is such that events A_i are the causes and event B is the effect (where we might have $P(A_i)$ and $P(B|A_i)$ right off the bat), then we would call $P(A_i|B)$ the **posterior probability** for each cause A_i. We would call each $P(A_i)$ a **prior probability**.

In general, the posterior probability is the probability $P(\theta|x)$ of the parameters θ given the evidence x, whereas the **likelihood function** is the probability $P(x|\theta)$ of the evidence given the parameters. Again, the probability distribution function $P(\theta)$ represents the prior. And finally everything is brought together through Bayes' rule:

$$P(\theta|x) = \frac{P(x|\theta)P(\theta)}{P(x)}$$

1.5. Independence

If the occurrence of B provides no information about the occurrence of A, i.e. $P(A|B) = P(A)$, then A is said to be **independent** of B. By the definition of conditional probability, the above is equal to $P(A \cap B) = P(A)P(B)$. We adopt the latter relation as the definition of independence (A and B are independent if $P(A \cap B) = P(A)P(B)$).

Note that independence is symmetric. If A is independent of B, then B is also independent of A. Also, if A and B are independent, so are A and B^c (the non-occurrence of B provides no information on the probability of A occurring).

Independence does not translate well to sample space visualizations, because (1) independent events will often have different sample spaces, and (2) even if they have the same sample space, they will overlap in arbitrary ways. One note: two disjoint events A and B with probabilities greater than zero will never be independent, since their intersection $A \cap B$ is be empty and has probability zero. ($P(A \cap B) = 0$, but $P(A)P(B) > 0$!)

1.5.1. Conditional Independence

Here, we refer to the independence of various events with respect to a conditional probability law. Given an event C, the events A and B are **conditionally independent** if $P(A \cap B|C) = P(A|C)P(B|C)$. Alternatively,

$$\begin{aligned} P(A \cap B|C) &= \frac{P(A \cap B \cap C)}{P(C)} \\ &= \frac{P(A|B \cap C)P(B|C)P(C)}{P(C)} \\ &= P(A|B \cap C)P(B|C) \end{aligned}$$

Note: this implies that $P(A|C) = P(A|B \cap C)$ **iff** A and B are conditionally independent on C. This relation states that if C is known to have occurred, the additional knowledge that B has occurred does not change the probability of A.

Independence does **not** imply conditional independence (and vice-versa)!

1.5.2. Independence of a Collection of Events

The events $A_1, A_2, ..., A_n$ are independent if $P(\bigcap_{i \in S} A_i) = \prod_{i \in S} P(A_i)$ for every subset S of $\{1, 2, ..., n\}$. Note that pairwise independence alone does not imply independence across all events. Independence for a collection of events means that the occurrence or non-occurrence of **any number** of the events carries no information on the remaining events or their complements.

For example, if all A_i are independent, then $P(A_1 \cup A_2|A_3 \cap A_4) = P(A_1 \cup A_2)$.

1.5.3. Reliability

Say we want to assess the reliability of a system consisting of components that can fail independently. Often we can divide the system into subsystems, where each subsystem's components are connected either in series or in parallel.

Let a subsystem consist of components 1, ..., m and let p_i be the probability that component i succeeds.

- A series subsystem succeeds if all of its components succeed. $P(\text{series succeeds}) = p_1 p_2 \cdots p_m$
- A parallel subsystem succeeds if any of its components succeed. $P(\text{parallel succeeds}) = 1 - (1 - p_1) \cdots (1 - p_m)$

1.5.4. Independent Trials and the Binomial Probabilities

A sequence of independent but identical trials, where there are only two possible results at each stage, is known as a sequence of independent **Bernoulli trials**. Thinking of this in terms of coin tosses, where the probability of any toss being a head is p, the probability of any particular n-long sequence with k heads and $n - k$ tails is $p^k(1 - p)^{n-k}$.

Meanwhile, $p(k) = P(k \text{ heads come up in an } n\text{-toss sequence}) = \binom{n}{k} p^k (1-p)^{n-k}$. The numbers $\binom{n}{k}$ are known as **binomial coefficients**, while the probabilities $p(k)$ are known as **binomial probabilities**. Note that

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} = \text{number of } n\text{-toss sequences with } k \text{ heads}$$

Also, the binomial probabilities $p(k)$ must sum to 1. Hence the **binomial formula** states $\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = 1$.

1.6. Counting

When calculating probability, we often need to count the number of outcomes in an event.

The Counting Principle
Consider a process that consists of r stages. Suppose that there are n_i possible results at the ith stage. Then the total number of possible results of the r-stage process is $n_1 n_2 \cdots n_r$.

In the following subsections, we focus on counting arguments that involve the selection of k objects out of a collection of n objects.

1.6.1. k-permutations

If the order of a selection matters, the selection is called a **permutation**. Here, we want to count the number of ways we can arrange k of n objects. There are n choices for the first slot, $n-1$ choices for the second, and $n-(k-1)$ choices for the last one. By the Counting Principle, the number of possible sequences (called k-permutations) is $n(n-1)\cdots(n-k+1) = n!/(n-k)!$

If $k = n$, the number of possible sequences (called **permutations**) is $n!$.

1.6.2. Combinations

If the order of a selection does not matter, the selection is called a **combination**. In this case, we're interested in counting the number of k-element subsets of an n-element set. The difference between this and k-permutations is that now there is no ordering of the selected elements.

Each individual combination is associated with $k!$ duplicate k-permutations, so dividing the number of k-permutations by $k!$ gives the number of possible combinations:

$$\frac{n!}{k!(n-k)!}$$

Note: selecting an n-toss sequence of k heads is the same as selecting k heads out of the n-element set of tosses, i.e. it is simply a combination of k out of n objects. Hence the binomial coefficient is given by the same formula:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

The number of subsets [of all possible cardinalities] of an n-element set is 2^n.

1.6.3. Partitions

Here, we would like to divide a collection of n objects into multiple subsets. For example, a combination can be viewed as a partition of the set in two: one part with k elements and the other with the remaining $n-k$ elements. In general, the goal with partitions is to divide an n-element set into r disjoint subsets where the ith subset contains exactly n_i elements (for given n_i). *In how many ways can this be done?*

Forming the subsets one at a time, we have $\binom{n}{n_1}$ ways of forming the first subset, $\binom{n-n_1}{n_2}$ ways of forming the second subset, and so on. Overall, the total number of choices is

$$\binom{n}{n_1}\binom{n-n_1}{n_2}\binom{n-n_1-n_2}{n_3}\cdots\binom{n-n_1-...-n_{r-1}}{n_r}$$

which, after canceling, is equal to

$$\frac{n!}{n_1!n_2!\cdots n_r!}$$

This is called the **multinomial coefficient** and is usually denoted by $\binom{n}{n_1,n_2,...,n_r}$. In words, it is the number of ways to partition a set of n elements into a group of n_1 elements, a group of n_2 elements, and so on (up to a group of n_r elements).

1.7. Summary and Discussion

A probability problem can often be broken down into a few basic steps:

- Describing the sample space (the set of possible outcomes of the experiment)
- Specifying the probability law (the probability of each event), possibly indirectly
- Calculating explicitly the probabilities and/or conditional probabilities of events of interest

Several methods for calculating probabilities have been shown.

- In the **counting method**, which applies when the number of outcomes is finite and all outcomes are equally likely, P(event) = number of elements in the event / number of elements in the sample space.
- In the **sequential method**, which applies when the experiment has a sequential nature and conditional probabilities are known for each branch of the tree, the probability of an event is obtained by taking products of conditional probabilities along relevant paths (using the multiplication rule).
- In the **divide-and-conquer method**, which applies when we have $P(B|A_i)$ and $P(A_i)$ for events A_i which form a partition of the sample space, $P(B) = P(A_1)P(B|A_1) + ... + P(A_n)P(B|A_n)$.

2 Lecture

There are two interconnected skillsets:

- modeling the real world in the language of probability (understanding the problem as an "experiment")
- solving the problem mathematically (calculus, logic, combinatorics)

Probability Spaces

The **sample space** (Ω) of an experiment is the set of all possible outcomes. Elements of the sample space must be mutually exclusive and collectively exhaustive.

- If the experiment is "toss two coins," $\Omega = \{HH, HT, TH, TT\}$
- If the experiment is "toss a coin till you get heads," $\Omega = \{H, TH, TTH, ...\}$
- If the experiment is "waiting time for the next bus," Ω is uncountable.

Events are allowable subsets of Ω.

- If the experiment is "toss two coins," an event might be "get one head."
- If the experiment is "toss a coin till you get heads," an event might be "get an even number of tosses."
- If the experiment is "waiting time for the next bus," an event might be "waiting time for the bus is less than 5 minutes."

A **probability space** is a collection of three components: $(\Omega, \mathcal{F}, \mathcal{P})$

- Ω is the set of all outcomes. We must have an experiment with enumerable outcomes.
- $\mathcal{F}$ is the set of all events. We must be able to enumerate all of them.
- $\mathcal{P}$ is a probability measure, or the assignment of probabilities to events. Specifically, $\mathcal{P} : \mathcal{F} \mapsto [0, 1]$.

Axioms of probability (Kolmogorov)

- $P(\emptyset) = 0$; $P(\Omega) = 1$
- $P(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i)$ if the A_i's are disjoint $\leftarrow$ "countably additive" (n can be finite or infinite)

Fundamental probability facts

- $P(A^c) = 1 - P(A)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- **Union bound:** $P(A \cup B) \leq P(A) + P(B)$

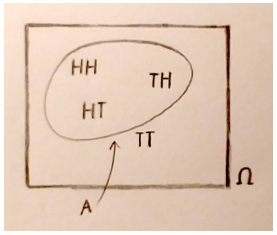
Inclusion-exclusion principle: $|A \cup B| = |A| + |B| - |A \cap B|$ (the number of elements in the union of two finite sets is equal to the sum of the sizes of the two sets minus the size of the intersection of the two sets). In probability,

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{k=1}^{n} \left((-1)^{k-1} \sum_{\substack{I \subseteq \{1,\ldots,n\} \\ |I|=k}} P\left(\bigcap_{i \in I} A_i\right) \right) = \sum_{i=1}^{n} P(A_i) - \sum_{\{i,j\}} P(A_i \cap A_j) + \sum_{\{i,j,k\}} P(A_i \cap A_j \cap A_k) - \ldots \pm P\left(\bigcap_{i=1}^{n} A_i\right)$$

Discrete probability: $P(A) = \sum_{\omega \in A} P(\omega)$

- In the experiment "toss two coins," P(at least one head) $= 3/4$ if fair coin.

P(**at least one head**)



Example: the birthday problem. We are interested in P(at least two people in a group of n people share the same birthday). Assume $n < 365$.

- $|\Omega| = 365^n$
- $A =$ "at least two people share the same birthday"
- $A^c =$ "no two people share the same birthday"

$$P(A^c) = \frac{|A^c|}{|\Omega|} = \frac{365 \cdot 364 \cdots (365 - n + 1)}{365^n}$$

If $n = 23$, $P(A) = 1 - P(A^c) \approx 50.4\%$. If $n = 50$, $P(A) \approx 97\%$. If $n = 100$, $P(A) \approx 99.9994\%$.

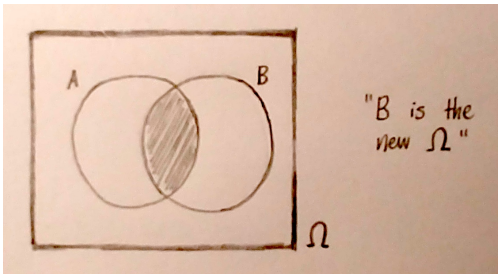
Conditional Probability

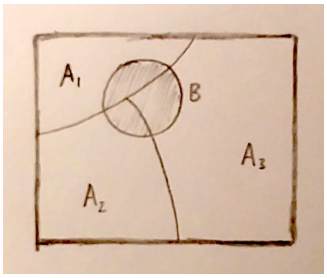
$P(A|B) = P(\text{event } A \text{ happens} \mid B \text{ happens}) = \frac{P(A \cap B)}{P(B)}$ for any event A (where $P(B) \neq 0$)

Example: rolling two six-sided dice. Event A is "the first die is a 6." Event B is "the sum of both dice is 7."

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(\{6,1\})}{P(\{6,1\} + \{5,2\} + ... + \{1,6\})} = \frac{1}{6}$$

Note: since $P(A|B) = P(A)$, we can say that A and B are independent.





Left: If probability is conditioned over B, B's region will become the new universe. Right: an illustration of total probability.

Multiplication Rule: $P(A_1 \cap A_2 \cap ... \cap An) = P(A_1)P(A_2|A_1) \cdots P(A_n|A_1, ..., A_{n-1})$

Total probability law: if $A_1, A_2, ...A_n$ form a partition of Ω (exhaustive and mutually exclusive), then

$$P(B) = P(A_1 \cap B) + P(A_2 \cap B) + ... + P(A_n \cap B) = \sum_{i=1}^{n} P(A_i)P(B|A_i)$$

Bayes' Rule

Inference setup:

- $A_1, ..., A_n$ form a partition of Ω
- We know the prior $P(A_i)$ for each A_i
- We know $P(B|A_i)$ for each A_i
- We want to determine $P(A_i|B)$

For example, we might be running a medical test for the diagnosis of a rare disease. Each A_i would represent whether or not a person has a given tumor, while B would represent the results of the test (positive or negative). While we might initially know B, what we really care about is whether a person has any of the tumors specified by the A_i's.

Independence

Two events are **independent** if the occurrence of one provides no information about the occurrence of the other.

$$P(A|B) = P(A) \iff P(A \cap B) = P(A)P(B) \iff A \text{ and } B \text{ are independent}$$

References

[1] D.P. Bertsekas and J.N. Tsitsiklis. *Introduction to Probability.* Athena Scientific books. Athena Scientific, 2002.