Elementary Probability Theory

An online chapter by D. Forsyth and J. Ponce

Justus H Piater

Computing with uncertainty

• If a flipped coin has come up heads 8 times and tails 6 times, what do I think will come up at the next flip?
• If 2/3 of my background pixels are black and the others are white, and if this relationship is reversed for foreground pixels, and if 5% of my pixels typically belong to the foreground, how certain can I be that a given, white pixel belongs to the foreground?
• Given a set of object categories with associated appearance vectors, which category has most likely produced a given, previously unseen vector?
• How likely will it rain tomorrow?
• How likely was there life on Mars a billion years ago?
What does probability mean?

Frequentist probability
limit of relative frequency
in a large number of trials
events only
hypothesis testing

Bayesian probability
degree of belief in a
proposition
arbitrary statements
degree of belief that can
be inferred/updated in the
light of new evidence

Let’s not worry about this debate. We’ll use probability theory because it is useful.
Experiments and outcomes

Example 1. A single coin flip
• The space of potential outcomes: $\Omega = \{h, t\}$
• A space of events: $\mathcal{F} = \{\emptyset, \Omega, \{h\}, \{t\}\}$

Example 2. Two coin flips
$\Omega = \{hh, ht, th, tt\}$
$\mathcal{F} = \mathcal{P}(\Omega)$

Example 3. Second coin flip unobservable
$\mathcal{F} = \{\emptyset, \Omega, \{hh, ht\}, \{th, tt\}\}$

Quick-Thinks

• Define the event space $\mathcal{F}$ representing the notion that the first coin flip is unobservable.
• In the case of Example 2, what is the event that either both heads or both tails occurred?
• Define the event space $\mathcal{F}$ that contains the events “both heads”, “both tails”, “coins the same”, “coins different”, “not both heads”, “not both tails”.
• Is $\mathcal{F} = \{\emptyset, \Omega, \{hh, ht\}\}$ a reasonable event space?
Events

Sample space $\Omega$, the set of possible outcomes. Event space $\mathcal{F}$, the set of possible events (subsets of $\Omega$):

$\emptyset \in \mathcal{F}$

$S_1 \in \mathcal{F} \Rightarrow \bar{S_1} = \Omega - S_1 \in \mathcal{F}$

countable $S_i \in \mathcal{F} \Rightarrow \bigcup_i S_i \in \mathcal{F}$

(These define a so-called $\sigma$-algebra.) It follows (from De Morgan’s laws):

$\Omega \in \mathcal{F}$

countable $S_i \in \mathcal{F} \Rightarrow \bigcap_i S_i \in \mathcal{F}$

Probability

A probability measure is a function $P : \mathcal{F} \rightarrow \mathbb{R}$ that satisfies the following three axioms:

$P(A) \geq 0$

$P(\Omega) = 1$

$P(A \cap B) = 0 \Rightarrow P(A \cup B) = P(A) + P(B)$

It follows:

$P(\bar{A}) = 1 - P(A)$, $P(\emptyset) = 0$

$0 \leq P(A) \leq 1$ for any $A$

if $A \subset B$, then $P(A) \leq P(B)$

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

We call $(\Omega, \mathcal{F}, P)$ a probability model. $P(A \cap B)$ is usually written $P(A, B)$, and is called the joint probability of the events $A$ and $B$. 
Quick-Thinks

- Give a probability model for Example 1.
- Give a probability model for Example 3.
- Give a probability model for Example 2.

If we already know something about the event?

Given the event space 
$$\mathcal{F} = \{\emptyset, \Omega, \{hh\}, \{tt\}, \{hh, tt\}, \{ht, th\}, \{ht, th, tt\}\},$$
say, we know that \(A = \{ht, th, tt\} \in \mathcal{F}\) has happened.

This defines \(\textit{by intersecting the events } B \in \mathcal{F} \textit{ with } A\) a new event space 
$$\mathcal{F}_A = \{\emptyset, A, \{tt\}, \{ht, th\}\}$$
that satisfies the event axioms [7].

The probabilities of the events \(B_A \in \mathcal{F}_A\) can be defined in terms of their counterparts \(B \in \mathcal{F}\) as 
$$P_A(B_A) = \frac{P(B \cap A)}{P(A)}$$
(verify!).
Conditional Probability

What is the probability that event $B$ occurs, given that event $A$ has occurred?

$$P(B \mid A) = \frac{P(B \cap A)}{P(A)}$$

$$P(B, A) = P(B \mid A)P(A) = P(A \mid B)P(B)$$

- We require $P(A) \neq 0$.
- Set-size analogy: $P(B \mid A)$ measures the size of $B \cap A$ relative to $A$.
- This is symmetric.

- For disjoint sets $A_i$ with $\Omega = \bigcup A_i$, we define likewise

$$P(B) = P(B \mid \Omega) = \sum_i P(B \mid A_i)P(A_i) = \sum_i P(B, A_i)$$

Conditional Probability (Continued)

The operation on the right is called *marginalization*. 
Quick-Think

Illustrate

\[ P(B) = P(B \mid \Omega) = \sum_i P(B \mid A_i)P(A_i) = \sum_i P(B, A_i) \]

on \( \mathcal{F} = \{ \emptyset, \Omega, \{hh\}, \{tt\}, \{hh, tt\}, \{ht, th\}, \{ht, th, tt\}, \{hh, ht, th\} \} \)

with

\[ A_1 = \{hh\} \]
\[ A_2 = \{tt\} \]
\[ A_3 = \{ht, th\} \]
\[ B = \{hh, tt\} \]

Independence

If events \( A \) and \( B \) are independent, then

\[ P(A \mid B) = P(A) \]
\[ P(B \mid A) = P(B) \]
\[ P(A, B) = P(A \mid B)P(B) = P(A)P(B) \]

We say that \( P(A, B) \) factors into \( P(A) \) and \( P(B) \).

Note

This can drastically reduce the number of degrees of freedom of a model.
For the two-coin-flip example
\[ \Omega = \{ hh, ht, th, tt \} \]
\[ F = 2^\Omega \]
give two joint probability models \( P(A, B) \) in terms of the outcomes of the first and the second coin flip:
- one where the probabilities of the outcomes of the second flip may depend on the outcome of the first flip,
- one where they do not.
How many degrees of freedom does each of the models have?

### Conditional Independence

A and \( B \) are conditionally independent given \( C \) iff
\[ P(A, B | C) = P(A | C)P(B | C) \]

Here, \( P(A, B | C) \) factors in the way indicated.

**Example 4.**
Assume \( A, B \) and \( C \) contain 2, 3, and 4 different possible outcomes, respectively.
- How many degrees of freedom does \( P(A, B, C) \) generally have?
- How many degrees of freedom does \( P(A | C)P(B | C)P(C) \) have?
Grounding probabilities in frequencies

Frequency data can be converted to $P$ functions that satisfy the axioms.
See some examples.
In general, if $P(\text{heads}) = p$:

$$P(k \text{ heads in } n \text{ flips}) = \binom{n}{k} p^k (1 - p)^{n-k}$$

This is, by the way, the **binomial distribution**.
Subjective probabilities

What is the probability that it will rain tomorrow?
What is the probability that an unknown coin will come up heads?
What is the probability that a given coin will land on its edge and stand there?

Note

• Often, these can (in principle) be reduced to frequentist probabilities.
• Even frequentist probability models usually require the modeler to pull some parameters out of the sleeve.
Continuous Spaces: Mostly the same thing.

- Events
- Probability
- Conditional probability and marginalization
- Sums become integrals.
- Independence
- Conditional independence

Example 5. Drawing real numbers from [0, 1]

- What is \( P(a) \)?
- What is \( P([0, \frac{1}{2}]) \)?
- What is \( P([0, 1]) \)?
Probability Density Functions

We can no longer enumerate the events and assign a probability to each of them. We will instead assign probabilities to subsets of $\Omega$ by integrating probability density functions $p$ over them:

$$P(\text{event}) = \int_{\text{event}} p(u) \, du$$

$$p(u_1) \, du = P(\{u \in [u_1, u_1 + du]\})$$

A PDF is nonnegative, but can exceed unity, and $P(\Omega) = \int_{\Omega} p(u) \, du = 1$.

It follows that $P(A) = 0$ for any finite set $A$.

**Note**

This works for sample spaces $\Omega$ of any dimensionality.

---

Quick-Think

Give a probability density function for Example 5.
Marginalization

Example 6. Drawing vectors of real numbers from $[0, 1] \times [0, 1]$

The probability density at the first element equal to $u$ is

$$p(u) = \int_0^1 p(u, v) \, dv$$
Random Variables

A **random variable** is a function from a sample space $\Omega$ to a **state space** (typically $\mathbb{R}^n$).

Thus, given a **probability model** $(\Omega, \mathcal{F}, P)$, a random variable $\xi$ induces a new probability model $(\Omega', \mathcal{F}', P)$, where, for each $A \in \mathcal{F}$ and corresponding $A' \in \mathcal{F}'$, $P(\{\xi \in A' \mid \xi \in A\}) = P(A)$.

The inherited probability measure $P$ (defined on a state space instead of the sample space) is called a **probability distribution**.

**Examples:**
- The identity function.
- The *income* from a coin-flip gamble.

---

Expectations

The **expected value** (or **expectation**) of a real-valued (discrete or continuous) random variable $X$ is:

$$E[X] = \sum_x x P(x)$$

$$E[X] = \int_{\Omega} x p(x) \, dx$$

The **variance** of a random variable is

$$\text{var}(X) = E[X^2] - E[X]^2$$, where $E[f(X)] = \sum_i f(x_i) P(x_i)$ or

$$E[f(X)] = \int_{\Omega} f(x) p(x) \, dx$$.

The **covariance** matrix of a vector of random variables is


Quick-Think

What are the expectation and variance of the income from gambling on a single coin flip [27], using a fair coin and a bet of 1 €?

Marginalization

Example 7. Two types of coins

Assume we have two types of coins, one balanced and one biased. What is the probability of observing heads?

\[
\Omega = \{ \text{heads}, \text{tails} \}
\]

\[
P(\text{heads}) = P(\{\text{heads}\}) + P(\{\text{tails}\})
\]

\[
P(\text{heads}) = P(\{\text{heads}\}) + P(\{\text{tails}\})
\]

Marginalization in general:

\[
p(x_1, ..., x_n) = \sum_{x_1} p(x_1, x_2, ..., x_n)
\]

\[
p(x_1, ..., x_n) = \int_{x_1} p(x_1, x_2, ..., x_n) dx_1
\]
Some Popular Distributions and Densities

**Uniform density/distribution:** Not meaningful in infinite domains.

**Binomial distribution:** \( n \) i.i.d. samples from a distribution with two values will contain \( k \) instances of the first value with probability \( P(k) = \binom{n}{k} p^k (1-p)^{n-k} \), where the first value is drawn with probability \( p \). The mean value of \( k \) is \( np \), its variance \( np(1-p) \) (example [17]).

**Poisson distribution:** Typically, uniform spatial or temporal models. A Poisson process produces \( k \) events per unit interval with probability \( P(k) = \frac{\lambda^k e^{-\lambda}}{k!} \). The mean and variance of the number of events per unit interval are equal to \( \lambda \).
The Normal (or Gaussian) Distribution

\[ p(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left( -\frac{(x-\mu)^2}{2\sigma^2} \right) \]

\[ p(x; \mu, \Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left( -\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu) \right) \]

This distribution is important in theory and in practice:

- **The Central Limit Theorem**: The sum of any i.i.d. random variables of finite variance tends toward a normal distribution (which may justify Gaussian noise models).
- Many (noise) effects observed in practice are reasonably well modeled by a normal distribution.
- Functions of the exponential family are easily manipulated.

---

Probabilistic Inference
Maximum-Likelihood Estimation

**Maximum likelihood (ML) estimation:**
\[ \arg\max P(\text{data} | \text{parameters}) \]

**Example 8. Inferring the type of a coin.**
Assume we have two types of coins with \( P(\text{H} | \lambda) = \lambda_1 \) and \( P(\text{H} | \lambda) = \lambda_2 \).
Given \( n \) independent flips \( r_i \) of an unknown coin, we conclude that it is of the type \( t \) that maximizes \( \prod_{i=1}^{n} P(r_i | t) \).

Priors, Posteriors and Bayes’ Rule

What if Type-II coins are really rare? Shouldn’t we multiply the above term by \( P(t) \)?
From \( P(A, B) = P(A | B)P(B) = P(B | A)P(A) \) follows **Bayes’ Rule**:
\[ P(B | A) = \frac{P(A | B)P(B)}{P(A)} \]
This is useful for
- reversing the order in a conditional probability
- adding evidence to a belief (\( P(B) \) is the \textit{prior}, \( P(B | A) \) is the \textit{posterior})
Bayesian Inference

**Maximum a-posteriori (MAP) estimation:**
\[ \text{argmax} P(\text{parameters} \mid \text{data}) \]
This is typically computed using Bayes’ Rule.

**Note**
If the prior is uniform, MAP estimation reduces to ML estimation.
Instead of just the MAP estimate, we may be interested in the whole posterior.

---

Quick-Think

**Example 9. Estimating the parameter of a coin from a sequence of flips**

Our coin comes up heads with probability \( \lambda \), i.e.,
\[ P(\text{heads} \mid \lambda) = \lambda. \]
Our prior is \( p(\lambda) = 1. \)

- What is \( p(\lambda \mid \text{hh}) \)?
- By an analogous argument,
  \[ p(\lambda \mid k \text{ heads and } n-k \text{ tails}) \propto \lambda^k(1-\lambda)^{1-k}. \]
50 Percent Heads

20 Percent Heads
No Heads

![Graph showing probability density for different scenarios of heads in coin flips.]

Model Selection

\[ P(\text{model} | \text{data}) = \frac{P(\text{data} | \text{model}) P(\text{model})}{P(\text{data})} \]

\[ = \frac{\int P(\text{data} | \text{model}, \text{params}) P(\text{params}) \, d\text{params}}{P(\text{data})} \]

\[ \propto \int P(\text{data} | \text{model}, \text{params}) P(\text{params}) \, d\text{params} \]