Chapter 0

Review of Probability

In this chapter, we will review some concepts from probability theory and linear algebra that will be useful in the rest of the course.

This review is not comprehensive. You can refer to the course webpage for more resources.

0.1 What is probability?

Intuitively, probability is a way of systematically studying events whose outcomes are uncertain. It enables us to quantify information and uncertainty (e.g., the probability of rolling a 6 is $1/6$ or the probability of rain on grounds at 10 am tomorrow is 20%). It can be used to describe relationships and provides ways to transfer our knowledge about one random quantity to another.

From a mathematical point of view, probability deals with sets, and functions that assign real values to those sets, in a way that certain axioms are satisfied. In this sense, probability is similar to geometry, number theory, etc. It can be used to model the real world, but it can also be studied as an abstract subject.

0.1.1 Definitions:

Assuming an experiment with different possible outcomes, consider the following definitions.[1]

- Ω: the sample space, the set of all possibilities (*outcomes*)
- $E \subseteq \Omega$: an event, i.e., a set of outcomes
- $\Pr$: A function from subsets of Ω to $\mathbb{R}$. $\Pr(E)$ is the probability of the event E.

0.1.2 Axioms:

- $\Pr(E) \geq 0$ for all $E \subseteq \Omega$.
- $\Pr(\Omega) = 1$
- $\Pr(E_1 \cup E_2) = \Pr(E_1) + \Pr(E_2)$ if $E_1 \cap E_2 = \varnothing$.

Based on these axioms, many theorems and other results can be proven. For $A, B \subseteq \Omega$:

- If $A \subseteq B$, then $\Pr(A) \leq \Pr(B)$.
- $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$.

More definitions for basic concepts:

[1] These definitions and the following axioms are simplified. We cannot always assign probability to all subsets of Ω. Also, for the third axiom, for any **countable** sequence of mutually exclusive events $E_1, E_2, \ldots$, we require that $\Pr(\bigcup_{i=1}^{\infty}) = \sum_{i=1}^{\infty} \Pr(E_i)$.

- Two events A and B are *independent*, denoted $A \perp\!\!\!\perp B$, if $\Pr(A \cap B) = \Pr(A)\Pr(B)$.
- If $\Pr(B) \neq 0$, the *conditional probability* of A given B is defined as

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}.$$

- Random variables, distributions, expected value, ...

What these theorems and definitions 'mean' depends on what we think probability means.

0.1.3 Interpretations of probability

> *Probability is the most important concept in modern science, especially as nobody has the slightest notion what it means.*
>
> Bertrand Russell

How do we assign probability to events? What does it mean, for example, to say that $\Pr(E) = 1/3$?

- Classical interpretation: If there are K possible outcomes, and we have no reason for some outcomes to be more likely than others, the probability of each outcome is $1/K$.
  - Probability of rolling a 3 is $1/6$.
  - Probability of heads is $1/2$ when tossing a fair coin.
- Frequentist interpretation: Assume that there is a "random" experiment that can be repeated many times. If we repeat it N times and N is very large, then the number of times that the event E occurs is approximately $N\Pr(E)$. In other words, the "frequency" of E occurring is $\Pr(E)$.
  - Probability of heads for a given coin is $\Pr(H) = 1/3$. So if we toss it 3000 times, we should see heads around 1000 times.
  - Probability **distribution** of the number N of children (≤ 18) of a randomly chosen American household:

	$\Pr(N=0)$	$\Pr(N=1)$	$\Pr(N=2)$	$\Pr(N \geq 3)$
1970	0.442	0.182	0.174	0.203
2008	0.541	0.195	0.169	0.095

- Bayesian interpretation: probability indicates the degree of belief in a way that is consistent with the axioms. This allows us to consider events that are, strictly-speaking, not random.
  - Pr(Heads) = 1/2 (both Bayesian and frequentist)
  - Pr(Stock market will hit a certain threshold this year)
  - Pr(Nuclear war this century)
  - Pr(A certain person is guilty of a given crime)

The classical interpretation is sometimes criticized as being circular. We call a coin fair if $\Pr(H) = \Pr(T) = 1/2$ and we say $\Pr(H) = 1/2$ if the coin is fair. Nevertheless, the definition is relied upon in practice, e.g., in games of chance. The frequentist definition can be criticized for being vague. What do "large" and "approximately" mean? How large is large enough? And how close should two values be for us to call them approximately equal? The Bayesian interpretation is criticized for being subjective and for assigning probabilities to experiments that happen only once (so any given event either happens or does not happen).

Criticism of interpretations of probability does not create any mathematical problems. Mathematically, we only need to assign probabilities in a way that the axioms are satisfied. Different interpretations however lead to different approaches to problems, potentially leading to different real-world decisions.

0.2 Sets and their sizes

Finding the probability of an event is easiest when all outcomes are equally likely. In such cases, if we can measure the size of the set A of desirable outcomes, dividing that by the size of the sample space, will yield the probability,

$$\Pr(A) = \frac{|A|}{|\Omega|},$$

where $|A|$ denotes the size of the set A.

Definition 0.1. A set A is **finite** if there is a natural number n such that the number of elements in A is less than n. Otherwise, it is **infinite**. If the elements of A can be counted, i.e., there is a one-to-one function from A to natural numbers, then A is **countable**. Otherwise, it is **uncountable**. A countable set may be finite (e.g., $\{1, 5, 6\}$) or infinite (e.g., integers, prime numbers, rational numbers).

If A is finite, we define its size (aka, cardinality) as the number of elements. This requires us to be able to count:

- **Sum rule:** If an action can be performed in m ways and another action can be performed in n ways, and further if we can choose which action to perform, in total we have $m + n$ options.
- **Product rule:** If the first action can be performed in m ways and the second action can be performed in n ways, and further if we must perform both actions in order, in total we have $m \times n$ options.
- **Permutations:** The number of ways we can arrange n objects is $n! = 1 \times 2 \times \cdots \times n$.
- **Combinations:** The number of ways we can choose k objects from a set of n objects is

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

Exercise 0.2. †[2] Prove that $\binom{n}{x}x = n\binom{n-1}{x-1}$. △

Exercise 0.3. How many 8-bit bytes are there? How many of these have exactly 3 ones? If we pick a random byte, what is the probability that it has exactly 3 ones (binomial distribution)? What is the probability that it has 6 or more consecutive ones? △

Exercise 0.4. How many binary sequences of length n that end with one are there with exactly k ones? △

If the sample space has an infinite, even uncountable, number of outcomes, we may still be able to think of the outcomes as equally likely. For example, if we pick a random number between 0 and 1, we may assume all outcomes are equally likely. In such cases, the size of the set can be measured via length, area, volume, etc.

Exercise 0.5. A random number in the interval $[0, 1]$ is chosen. What is the probability that it is more than $1/2$ but less than $2/3$? What is the probability that it is equal to $1/2$? What is the probability that it is rational (optional)? △

Exercise 0.6. A random point is chosen in a square of unit side. What is the probability that it is inside the circle of diameter one inscribed in the square? What is the probability that it is on the circle? △

0.3 Random variables and distributions

A **random variable** (**RV**) is a function that assigns real values to outcomes in Ω. In most cases, there is a very natural mapping. For example, let X denote the number showing on a dice. Now X is a random variable, mapping each outcome of the form "the dice shows i" to the real number i. For this reason, the

[2] This symbol indicates that the exercise, section, etc., is optional.

fact that random variables are really functions is often overlooked. Information about the probabilities of different outcomes is given by the **distribution** of the random variable.

A random variable is **discrete** if there are a countable number of possibilities (could be infinite but countable, like natural numbers). They can also be **continuous** (uncountable number of outcomes, defined over the real line or some subset of some Euclidean space).

For example, a random variable that is 1 if heads shows when a given coin is filliped and is 0 otherwise is discrete and finite; the number of phone calls made in a given hour is discrete and infinite; the arrival time of a plane from midnight is continuous.

0.3.1 Discrete distributions

The distribution of a discrete random variable X is given by its **probability mass function** (pmf) denoted by $p_X(x)$, where

$$p_X(x) = \Pr(X = x).$$

Clearly, $p_X(x) \geq 0$ for all x and

$$\sum_x p_X(x) = 1. \tag{0.1}$$

If clear from the context, we drop the X in the subscript.

Example 0.7 (**Poisson Distribution**). An RV X has the Poisson distribution with parameter λ if

$$p(x) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x \in \{0, 1, \dots\}.$$

The number of times an event, e.g., phone calls or car accidents, occurs in a given interval of time is often assumed to have a Poisson distribution (with good reason). △

Exercise 0.8. A red die and a blue die are rolled. Let X denote the number showing on the red die and Y denote the sum of the two dice. Find the pmf of X and the pmf of Y. △

Exercise 0.9. Two cards are drawn at random from a standard deck of 52 cards and let Z denote the number of Aces drawn. Find the pmf of Z. △

0.3.2 Continuous distributions

The distribution of a continuous random variable X is given by its **probability distribution function** (pdf) $p_X(x)$, also sometimes denoted $f_X(x)$. Roughly speaking,

$$\Pr\left(x - \frac{dt}{2} \leq X \leq x + \frac{dt}{2}\right) = p_X(x)dt.$$

For two real numbers a, b,

$$\Pr(a \leq X \leq b) = \int_a^b p_X(x)dx.$$

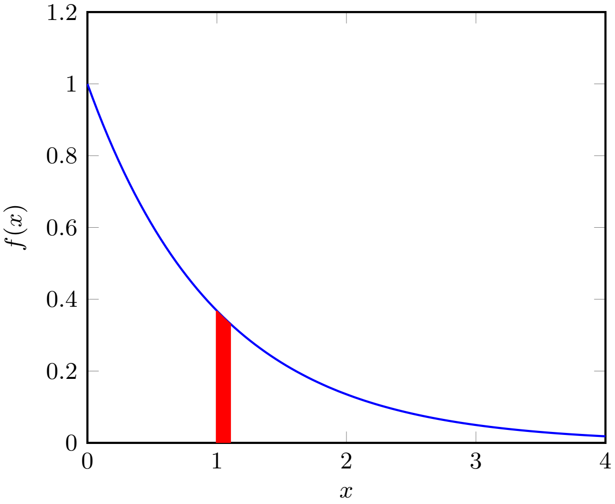
For any pdf, we have $p_X(x) \geq 0$ and

$$\int_{-\infty}^{\infty} p_X(x)dx = 1.$$

Exercise 0.10 (**Exponential distribution**). An exponential random variable X with parameter λ has distribution

$$f(x) = \lambda e^{-\lambda x}, \quad x \geq 0.$$

For $\lambda = 1$, the probability that X is between 1 and 1.1 is around $e^{-1} \times 0.1 = 0.37 \times 0.1 = 0.037$. In the figure below, the area colored red represents this probability.



△

0.3.3 Cumulative distribution functions

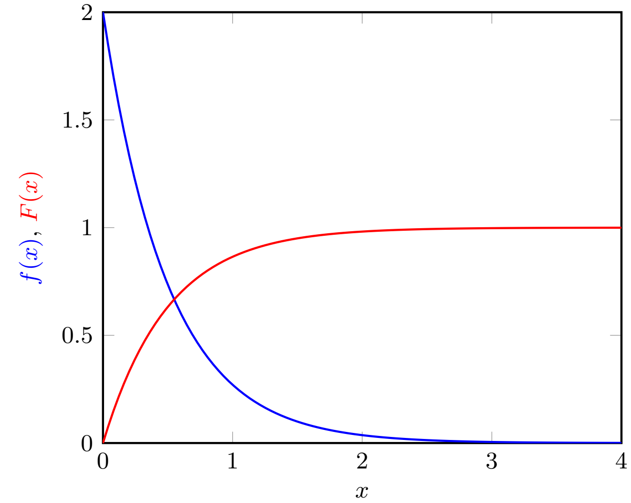
Cumulative distribution functions (CDFs) are defined for both discrete and continuous RVs as $F_X(x) = p_X(X \le x)$ and can be found via summation or integration:

$$F_X(x) = \sum_{k \le x} p_X(k)$$

$$F_X(x) = \int_{-\infty}^{x} p_X(t) dt$$

Example 0.11. The CDF of the exponential RV in Example 0.10 with $\lambda = 2$ is given by

$$F_X(x) = \int_{-\infty}^{x} \lambda e^{-\lambda t} dt = 1 - e^{-\lambda x}$$



△

0.3.4 Expected value

The **expected value** or the **mean** $\mathbb{E}[X]$ of a random variable X with distribution $p(x)$ is given by

$$\mathbb{E}[X] = \sum_x xp(x),$$
$$\mathbb{E}[X] = \int_{-\infty}^{\infty} xp(x)dx.$$

One way to think about the expected value is as the average of a large number of experiments. For example, if a game pays out $\$X$ each time you play with probability distribution $p(x)$, if you play the game many times, on average you will win $\$\,\mathbb{E}[X]$ per game. That is if you play n times, each time winning $\$x_n$, and n is large, then

$$\frac{1}{n}(x_1 + x_2 + ... + x_n) \simeq \mathbb{E}[X].$$

Exercise 0.12. Find the expected value of the discrete and continuous RVs in the examples above. △

Exercise 0.13. Find $\mathbb{E}[1]$. △

0.3.4.1 Expectation of functions of random variables

For an RV X and a function $f(x)$ it follows from the definition that

$$\begin{aligned}\mathbb{E}[f(X)] &= \sum_x f(x)p(x),\\ \mathbb{E}[f(X)] &= \int_{-\infty}^{\infty} f(x)p(x)dx.\end{aligned} \tag{0.2}$$

Exercise 0.14. A random variable X has distribution

$$p_X(-1) = 0.1,\ p_X(0) = 0.2,\ p_X(1) = 0.3,\ p_X(2) = 0.4.$$

Find $\mathbb{E}\,X$. Let $Y = X^2$. Find $\mathbb{E}\,Y$, both by finding the distribution of Y and by using (0.2). △

0.3.4.2 Linearity of expectation

For a RV X, functions $f(x)$ and $g(x)$, and real numbers a and b,

$$\mathbb{E}[af(X) + bg(X)] = a\,\mathbb{E}[f(X)] + b\,\mathbb{E}[g(X)],$$

which can be proven easily from the definition of expectation.

Example 0.15. $\mathbb{E}[(X-a)^2] = E[X^2 - 2aX + a^2] = \mathbb{E}[X^2] - 2a\,\mathbb{E}\,X + a^2$. △

Consider a collection of random variables $X_1, X_2, \ldots, X_n$. By the linearity of expectation

$$\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}\,X_i. \tag{0.3}$$

If all variables are identically distributed, then

$$\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = n\,\mathbb{E}\,X_1. \tag{0.4}$$

Example 0.16. In a class of n students, what is the expected number of pairs of students who have the same birthday? To find this, for two students i and j, let X_{ij} be equal to 1 if they share a birthday and 0 otherwise and let $X = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} X_{ij}$. Now,

$$\mathbb{E}\, X = \binom{n}{2} \mathbb{E}\, X_{12} = \binom{n}{2} \Pr(X_{12} = 1) = \binom{n}{2} \frac{1}{365} \simeq \frac{n^2}{730}. \tag{0.5}$$

In particular, having $n = \sqrt{730} \simeq 27$ students in a class is enough to have on average one pair with the same birthday. With $n = 60$ and $n = 85$ students, there should be around 5 and 10 such pairs, respectively. △

0.3.4.3 Variance

Suppose someone offers you a game in which your expected winning is \$100. Will you accept? Which game would you play?

- You always win exactly \$100.
- You win \$0 with probability $1/2$ and \$200 with probability $1/2$.
- You win \$1200 with probability $1/2$ and lose \$1000 with probability $1/2$.

All three have the same mean. So what's different between them?

The mean helps us represent a distribution with one value, which describes the average behavior of the RV. But as this example shows, the behavior around the mean is also important. Denoting the mean of X by μ_X, the variability around the mean is captured to a degree by the variance $\mathrm{Var}[X]$,

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

The variance gives a sense of *how far X is from its mean μ_X, on average.* The **standard deviation**, σ_X, is defined as

$$\sigma_X = \sqrt{\mathrm{Var}[X]},$$

and the variance is usually denoted as σ_X^2.

Exercise 0.17. Prove that

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

△

Exercise 0.18. Find the mean and variance of each of the following RVs [1]:

- $X + c$
- aX
- $aX + c$
- $\frac{X - \mu_X}{\sigma_X}$ (called the **standardized version** of X)

△

0.3.5 Common distributions

We denote X having distribution 'Dist' by $X \sim \mathrm{Dist}(a, b, \dots)$, where $a, b, \dots,$ are the parameters of the distribution.

0.3.5.1 Discrete distributions

- $X \sim \text{Ber}(p): \quad \Pr(X=1)=p,\ \Pr(X=0)=1-p, \quad \mathbb{E}[X]=p, \quad \text{Var}[X]=p(1-p).$
- $X \sim \text{Bin}(n,p):$ [3] $\quad p(x)=\binom{n}{x}p^x(1-p)^{n-x},\ 0\le x\le n, \quad \mathbb{E}[X]=np, \quad \text{Var}[X]=np(1-p).$
- $X \sim \text{Geo}(p): \quad p(x)=(1-p)^{x-1}p,\ x\ge 1, \quad \mathbb{E}[X]=1/p, \quad \text{Var}[X]=(1/p)^2-(1/p).$
- $X \sim \text{NegBin}(k,p): \quad p(x)=\binom{x-1}{k-1}(1-p)^{x-k}p^k,\ x\ge k, \quad \mathbb{E}[X]=k/p, \quad \text{Var}[X]=k[(1/p)^2-(1/p)].$
- $X \sim \text{Poi}(\lambda): \quad p(x)=\frac{\lambda^x e^{-\lambda}}{x!},\ x\ge 0, \quad \mathbb{E}[X]=\lambda, \quad \text{Var}[X]=\lambda.$
- $X \sim \text{Uni}[a,b]: \quad p(x)=\frac{1}{b-a+1},\ x\in\mathbb{Z}, a\le x\le b, \quad \mathbb{E}[X]=\frac{a+b}{2}, \quad \text{Var}[X]=\frac{(b-a+1)^2-1}{12}.$

Exercise 0.19. Prove that the mean of Bin(n,p) is as given using Exercise 0.2. △

0.3.5.2 Continuous distributions

- $X \sim \text{Uni}(a,b): \quad p(x)=\frac{1}{b-a},\ x\in(a,b), \quad \mathbb{E}[X]=\frac{a+b}{2}, \quad \text{Var}[X]=\frac{(b-a)^2}{12}.$
- $X \sim \mathcal{N}(\mu,\sigma^2): \quad p(x)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{(x-\mu)^2}{2\sigma^2}),\ x\in\mathbb{R}, \quad \mathbb{E}[X]=\mu, \quad \text{Var}[X]=\sigma^2.$
- $X \sim \text{Exp}(\lambda): \quad p(x)=\lambda e^{-\lambda x},\ x\ge 0, \quad \mathbb{E}[X]=1/\lambda, \quad \text{Var}[X]=1/\lambda^2.$

Sometimes, we drop the normalization constant, that is, the constant by which we divide to ensure that the distribution integrates to 1. This could be because the constant is not important (e.g., in Bayesian inference) or because it is hard to determine. In such cases, we use $\propto$ to show proportionality rather than equality. We should be careful which of the entities appearing is the *variable*. For example, viewed as a function of x, we have $f(x)=\frac{\lambda^x e^{-\lambda}}{x!}\propto\frac{\lambda^x}{x!}$ and as a function of λ, we have $g(\lambda)=\frac{\lambda^x e^{-\lambda}}{x!}\propto\lambda^x e^{-\lambda}$.

- $X \sim \text{Beta}(\alpha,\beta): \quad p(x)\propto x^{\alpha-1}(1-x)^{\beta-1},\ 0\le x\le 1, \quad \mathbb{E}[X]=\frac{\alpha}{\alpha+\beta}, \quad \text{Var}[X]=\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$
- $X \sim \text{Gamma}(\alpha,\beta): \quad p(x)\propto x^{\alpha-1}e^{-\beta x},\ x>0, \quad \mathbb{E}[X]=\frac{\alpha}{\beta}, \quad \text{Var}[X]=\frac{\alpha}{\beta^2}.$

Example 0.20. For the distributions given in this section, try changing what the variable is and what the parameters are and check whether another distribution from the list can be obtained with appropriate normalization. For example, Bin(n,p) viewed as a distribution in p turns into Beta$(x+1,n-x+1)$. △

0.4 Joint probability distributions

Joint probability distributions allow us to encode information about relationships between quantities, from independence to strong correlation.

For random variables X and Y, the CDF and the pmf/pdf give their joint distribution, depending on their type,

$$\begin{aligned}
F_{X,Y}(x,y) &= \Pr(X\le x, Y\le y), && \text{CDF for continuous and discrete}\\
p_{X,Y}(x,y) &= \Pr(X=x, Y=y), && \text{pmf for discrete}\\
p_{X,Y}(x,y)dxdy &\simeq \Pr\left(x-\frac{dx}{2}\le X\le x+\frac{dx}{2}, y-\frac{dy}{2}\le Y\le y+\frac{dy}{2}\right), && \text{pdf for continuous}
\end{aligned}$$

[3] Note that sometimes p is used both as a parameter and as the distribution. The meaning should be clear from the context.

We can find the distribution for each random variable (in this context these are called the **marginals**) by integration/summation,

$$p_X(x) = \sum_y p_{X,Y}(x,y), \qquad\qquad p_X(x) = \int_{-\infty}^{\infty} p_{X,Y}(x,y)dy.$$

0.4.1 Expectation, correlation, and covariance

Given two or more RVs, we may be interested in finding the expected value of a function of these RVs, e.g., $\mathbb{E}[XY]$. In such case, similar to (0.2), we have

$$\mathbb{E}[f(X,Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)p(x,y)dxdy, \tag{0.6}$$

and similarly for discrete variables.

The **correlation** between X and Y is $\mathbb{E}[XY] = \int\int xyp(x,y)dxdy$. The **covariance** $\text{Cov}(X,Y)$ and the **correlation coefficient** $\rho_{X,Y}$ are defined as

$$\text{Cov}(X,Y) = \mathbb{E}[(X-\mu_X)(Y-\mu_Y)]$$
$$\rho_{X,Y} = \frac{\text{Cov}(X,Y)}{\sigma_X\sigma_Y}.$$

It can be shown that $-1 \le \rho_{X,Y} \le 1$. If $\rho = 0$, then the random variables are **uncorrelated**.

What does the correlation coefficient mean? Let X and Y be random variables, for example, weight and height of a person chosen at random. Suppose that we want to predict the value of Y given X but we are restricted to linear functions of X. Then, in a certain sense,[4] the best predictor $\hat{Y}$ of Y is

$$\hat{Y} = \mathbb{E}\,Y + \rho\frac{\sigma_Y}{\sigma_X}(X - \mathbb{E}\,X),$$

with the “error” being

$$\sigma_Y^2\left(1-\rho^2\right).$$

In particular, if X and Y are standardized, $\hat{Y} = \rho X$ with error $1-\rho^2$.

Exercise 0.21. If $|\rho|$ is close to 1, the RVs are said to be **strongly correlated**. Why? △

Exercise 0.22. Show that $\text{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}\,X\,\mathbb{E}\,Y$. △

Example 0.23. The bivariate jointly Gaussian distribution for X,Y with means μ_X and μ_Y, variances σ_X and σ_Y, and correlation coefficient ρ is given as

$$p(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}e^{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2}+\frac{(y-\mu_Y)^2}{\sigma_Y^2}-\frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y}\right]}.$$

Examples of this pdf are given in Figure 1. △

Exercise 0.24. For random variables X,Y,Z and constants a,b,c,d,e, prove that

- $\text{Var}(X) = \text{Cov}(X,X)$
- $\text{Cov}(X+Y,Z) = \text{Cov}(X,Z) + \text{Cov}(Y,Z)$
- $\text{Cov}(aX,Y) = a\,\text{Cov}(X,Y)$
- $\text{Cov}(X,b) = 0$

[4] Minimizing the Mean Square Error

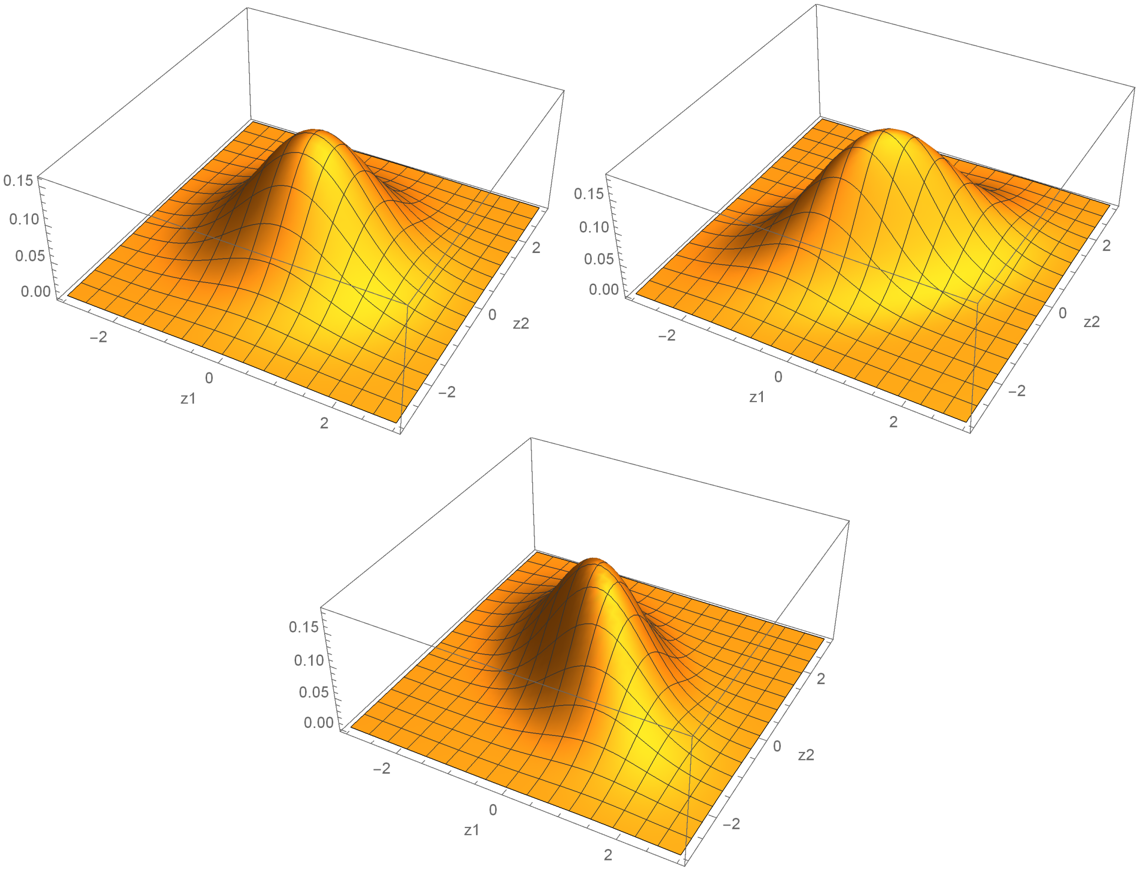


Figure 1: Bivariate Normal pdfs with $\mu_X = \mu_Y = 0$, $\sigma_X = \sigma_Y = 1$, with $\rho = 0$ (uncorrelated), $\rho = .5$ (positively correlated), and $\rho = -.5$ (negatively correlated), respectively.

- $\mathrm{Cov}(aX + bY + c, dZ + e) = ad\,\mathrm{Cov}(X, Z) + bd\,\mathrm{Cov}(Y, Z)$

△

Exercise 0.25. Find the expected values and variances of X and Y from Exercise 0.8. Find $\mathrm{Cov}(X, Y)$. △

0.4.2 Independence

Recall that two events A and B are independent iff (if and only if) $\Pr(A \cap B) = \Pr(A)\Pr(B)$. Two random variables X and Y are independent if $\{X \in S_1\}$ and $\{Y \in S_2\}$ are independent for all sets S_1 and S_2. This implies that

$$p(x, y) = p(x)p(y). \tag{0.7}$$

For two independent random variables, we have

$$\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y] \tag{0.8}$$

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

Exercise 0.26. Prove (0.8) using (0.7). △

Exercise 0.27. For two independent RVs X and Y, find $\mathrm{Var}[X + Y]$ and $\mathbb{E}[(X - Y)^2 + 3XY + 5]$ in terms of means and variances of X and Y. △

A collection $X_1, \ldots, X_n$ of random variables that are independent from each other but have the same distribution are called **independent and identically distributed (iid)**. We have

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i). \tag{0.9}$$

Exercise 0.28. For iid RVs $X_1, \ldots, X_n$, let $S_n = \sum_{i=1}^{n} X_i$. Show that

$$\mathrm{Var}(S_n) = \sum_{i=1}^{n} \mathrm{Var}(X_i). \tag{0.10}$$

△

Exercise 0.29. For iid RVs $X_1, \ldots, X_n$, suppose $\mathbb{E}[X_i] = \mu$ and $\mathrm{Var}[X_i] = \sigma^2$, and let $\bar{X}$ be their average. Show that

$$\mathbb{E}[\bar{X}] = \mu, \qquad\qquad \mathrm{Var}[\bar{X}] = \frac{\sigma^2}{n}. \tag{0.11}$$

△

0.4.3 Conditional probability and conditional distributions

For two discrete variables X and Y, the conditional probability distribution of Y given X is given by

$$p_{Y|X}(y|x) = \Pr(Y = y|X = x) = \frac{\Pr(Y = y, X = x)}{\Pr(X = x)} = \frac{p_{X,Y}(x, y)}{p_X(x)}.$$

For continuous RVs, we also have $p_{Y|X}(y|x) = \frac{p_{X,Y}(x,y)}{p_X(x)}$. In this case, however, we interpret the conditional density as

$$p_{Y|X}(y|x) \simeq \frac{\Pr(y - \epsilon/2 \le Y \le y + \epsilon/2 | x - \epsilon/2 \le X \le x + \epsilon/2)}{\epsilon},$$

for small positive ϵ. This essentially says to find $p_{Y|X}(y|x)$, we first assume that X is in a narrow strip around x and then find the density for Y given this assumption.

Law of total probability. Let $A_1, A_2, \dots, A_n$ be a partition of the sample space. That is, $\cup_{i=1}^n A_i = \Omega$ and for all $i \neq j$, we have $A_i \cap A_j = \varnothing$. For an event B_i, we have

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

In particular, if X can take on $\{1, 2, \dots, n\}$, then for another RV Y,

$$p_Y(y) = \sum_{x=1}^n p_{Y|X}(y|x)p_X(x).$$

Chain rule of probability. For events $A_1, \dots, A_n$, we have

$$\Pr(A_1 \cap A_2 \cap \dots \cap A_n) = \Pr(A_1)\Pr(A_2|A_1)\Pr(A_3|A_1, A_2) \cdots \Pr(A_n|A_1, \dots, A_{n-1}),$$

which can be easily proven by induction. A similar rule holds for random variables $X_1, \dots, X_n$:

$$p(x_1, \dots, x_n) = p(x_1)p(x_2|x_1)p(x_3|x_1, x_2) \dots p(x_n|x_1, \dots, x_{n-1}).$$

Conditional expectations are defined based on conditional distributions, e.g.,

$$\mathbb{E}[X|Y = y] = \sum_x x p_{X|Y}(x|y).$$

Exercise 0.30. Suppose the joint pmf is given as

$p_{X,Y}(x, y)$	$x = 0$	$x = 1$
$y = 0$	0.25	0
$y = 1$	0.5	0.25

Find $p(y|x)$, $p(x|y)$, $\mathbb{E}[Y|X = 0]$, $\mathbb{E}[Y|X = 1]$, $\mathbb{E}[X|Y = 0]$, $\mathbb{E}[X|Y = 1]$. △

Exercise 0.31. A point is chosen uniformly at random in a triangle with vertices on $(0, 0), (1, 0), (1, 1)$. Let X and Y determine the x and y coordinates of the chosen point. Find $p(x|y)$, $p(y|x)$, $\mathbb{E}[X|Y = y]$, $\mathbb{E}[Y|X = x]$. △

0.4.3.1 Law of iterated expectations.

Consider a random variable X and a function $g(x)$. We can now obtain $g(X)$ by replacing the deterministic value for x with a random one. Note that $g(X)$ is a random variable. For example, if $X \sim \text{Uni}(-1, 1)$ and $g(x) = |x|$, then $g(X)$ is a random variable with distribution $\text{Uni}(0, 1)$.

Now let $g(x) = \mathbb{E}[Y|X = x]$. This is, of course, a well-defined function. We define $\mathbb{E}[Y|X] = g(X)$, which is as discussed a random variable. Now that we have a random variable, we can compute its expectation, i.e., $\mathbb{E}[\mathbb{E}[Y|X]]$.

Exercise 0.32. A die is rolled, showing X. A coin is then flipped X times resulting in Y heads. Find $\mathbb{E}[Y]$, $\mathbb{E}[Y|X = x]$, the pmf of $\mathbb{E}[Y|X]$, and $\mathbb{E}[\mathbb{E}[Y|X]]$. △

It can be shown that

$$\mathbb{E}[\mathbb{E}[Y|X]] = \mathbb{E}[Y], \qquad \mathbb{E}[\mathbb{E}[Y|X, Z]|Z] = \mathbb{E}[Y|Z]. \tag{0.12}$$

0.4.4 Bayes' rule

In Exercise 0.32, the conditional distribution $p(y|x)$ is readily available as

$$p(y|x) = \binom{x}{y} 2^{-x}.$$

But what if we are interested in $p(x|y)$? Since $p(x|y) = \frac{p(x,y)}{p(y)}$ and $p(x,y) = p(y|x)p(x)$, we have

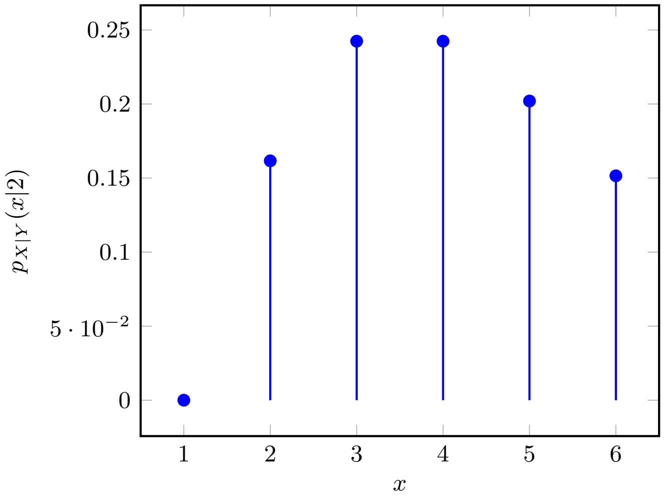
$$p(x|y) = \frac{p(y|x)p(x)}{p(y)} = \frac{p(y|x)p(x)}{\sum_{x'} p(y|x')p(x')},$$

which is called the **Bayes rule**.

Example 0.33. In Exercise 0.32, we can use the Bayes rule to find $p(x|y)$,

$$p(x|y) = \frac{\binom{x}{y}2^{-x}(1/6)}{\sum_{x'=y}^{6}\binom{x'}{y}2^{-x'}(1/6)} = \frac{\binom{x}{y}2^{-x}}{\sum_{x'=y}^{6}\binom{x'}{y}2^{-x'}}$$

We may ask for example, what is the likeliest value for X if $Y = 2$. Below, $p_{X|Y}(x|2)$, i.e., the conditional distribution of X given $Y = 2$. We can see that the likeliest values for X are $3, 4$.



△

Bayes' rule is used in *evidential reasoning*, examples of which we will see in the next chapter. In this setting, the goal is to find the probabilities of different causes based on the evidence.

Bayesian inference takes its name from Bayes rule. In this setting, it is often the case that we know the distribution of data given the parameters. But what we actually have is data and need to find the distribution of the parameters. The Bayes rule allows us to find this conditional distribution, a topic we will discuss in detail later.

0.5 Inequalities and limits

0.5.1 Inequalities

0.5.1.1 Markov inequality

Suppose the average length of a blue whale is 22m and we do not know anything else about the distribution of the lengths of blue whales. Can we say anything about the probability that the length of a randomly chosen blue whale is ≥ 30m? For example, is it possible that this probability is 0.8 or larger? No, since in that case, the average would be $\geq 0.8 \times 30\text{m} = 24\text{m}$. So only knowing the mean enables us to say something about the extremes of the probability distribution.

This observation is formalized via the **Markov inequality**. For a *non-negative* random variable X, we have

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

Exercise 0.34. Prove the Markov inequality. △

A special case of this occurs when X counts something, i.e., it only takes non-negative integer values. Then,

$$\Pr(X \geq 1) = \Pr(X > 0) \leq \mathbb{E}\, X, \qquad \Pr(X = 0) \geq 1 - \mathbb{E}\, X.$$

In particular, if the mean $\mathbb{E}\, X$ is small, then there is a large probability that $X = 0$.

Exercise 0.35 (†). Provide a bound on the probability that in a random binary sequence of length n, there exists a run (consecutive occurrences) of 1s of length at least $2\log_2 n$? (The result will tell you that this is unlikely for large n.) △

0.5.1.2 Chebyshev inequality

If in addition to the mean, we also have the variance, we can use the Chebyshev bound. For a random variable X with mean μ and variance σ^2,

$$\Pr\left(\left|\frac{X-\mu}{\sigma}\right| \geq a\right) \leq \frac{1}{a^2}.$$

Exercise 0.36. Prove the Chebyshev bound using the Markov bound. △

Example 0.37. The Chebyshev bound tells us that being k standard deviations away from the mean has probability at most $1/k^2$.

k	2	3	4	5	6	7	8	9	10
Probability of deviating more than $k \times$ std is $\leq$	25%	11.1%	6.25 %	4%	2.78%	2.04%	1.56%	1.23%	1%

In particular, being 10 standard deviations away from the mean has probability at most 1%. △

0.5.2 Limits

Limits in probability provide a way to understand what happens when the number of experiments grows or many random effects accumulate. Limit theorems are beneficial given that we often deal with large volumes of data. The following limit theorems will be helpful to us later in the course.

0.5.2.1 Law of large numbers

Let $X_1, \ldots, X_n$ be random variables with mean μ and variance $\leq \sigma^2$ and suppose that for each i and j, X_i and X_j are uncorrelated (in particular, independent). Also, let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Then, for any $\epsilon > 0$,

$$\Pr\bigl(|\bar{X}_n - \mu| \geq \epsilon\bigr) \leq \frac{\sigma^2}{n\epsilon^2}. \tag{0.13}$$

As n becomes large the right side becomes smaller and smaller. So for large n the probability of $\bar{X}_n$ being too far from the mean is very small. This is referred to as the **Law of Large Numbers** (LLN). In other words, if we take n independent samples from a random variable X, then the average of those samples will be close to the mean $\mathbb{E}\, X$,

$$\frac{1}{n}(x_1 + x_2 + ... + x_n) \simeq \mathbb{E}[X],$$

which is what we used to motivate expected value.

Exercise 0.38. Use the Chebyshev inequality to prove LLN when random variables are independent and all have the same variance σ^2. △

Example 0.39. Suppose $X_i \sim \text{Poi}(2)$, $1 \le i \le 500$, and let $\bar{X}_n$ be the average of the first n X_is. Figure 2 shows the plot for $\bar{X}_n$ for a realization of X_is obtained via computer simulation. It is observed that for large values of n, $\bar{X}_n$ is close to 2, the mean of the Poisson distribution. △

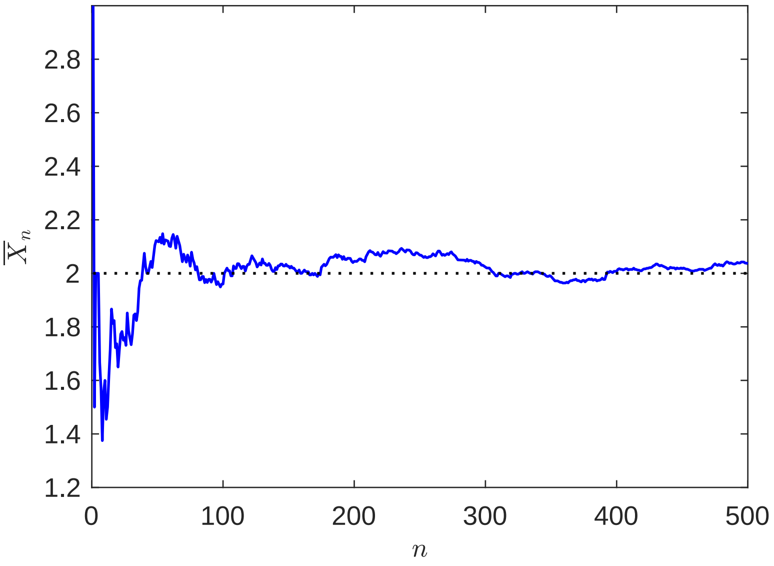


Figure 2: $\bar{X}_n$ based on $X_i \sim \text{Poi}(2)$ as a function of n.

0.5.2.2 Central limit theorem

Let $X_1, X_2, \ldots$ be iid random variables with mean μ and variance σ^2 and let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. As $n \to \infty$. The **Central Limit Theorem (CLT)** states that

$$\text{distribution of } \sqrt{n}(\bar{X}_n - \mu) \quad \to \quad \mathcal{N}(0, \sigma^2). \tag{0.14}$$

That is, the distribution of $\sqrt{n}(\bar{X}_n - \mu)$ approaches the distribution of a normal random variable with mean 0 and variance σ^2.

Loosely speaking, the CLT also means $S_n = \sum_{i=1}^{n} X_i$ has distribution $\mathcal{N}(n\mu, n\sigma^2)$.

Example 0.40. Let $X_i \sim \text{Uni}(0,1), 1 \le i \le n = 10$. We produce $50,000$ samples of $\bar{X}_n$ (and S_n), and plot the normalized histograms for $\sqrt{n}(\bar{X}_n - \mu)$ and the pdf of $\mathcal{N}(0, \sigma^2)$ and the normalized histogram for S_n and the pdf of $\mathcal{N}(n\mu, n\sigma^2)$ in Figure 3. △

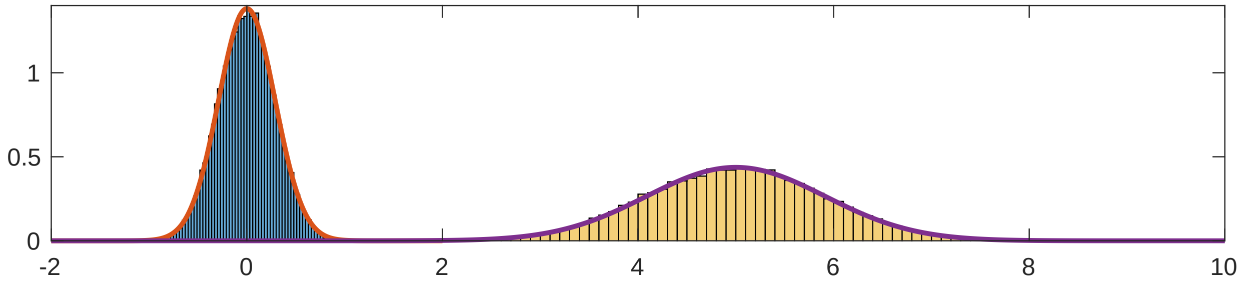


Figure 3: The normalized histograms for $\sqrt{n}(\bar{X}_n - \mu)$ and the pdf of $\mathcal{N}(0, \sigma^2)$ (on the left) and the normalized histogram for S_n and the pdf of $\mathcal{N}(n\mu, n\sigma^2)$ (on the right) for uniform X_i with $\mu = 1/2$ and $\sigma^2 = 1/12$ and with $n = 10$.

0.6 Random vectors

A **random vector** is a vector of random variables.[5] Consider the random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$

$$\boldsymbol{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_m \end{pmatrix}, \quad \boldsymbol{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}. \tag{0.15}$$

The **expected value** of $\boldsymbol{X}$ is

$$\mathbb{E}\,\boldsymbol{X} = \begin{pmatrix} \mathbb{E}\,X_1 \\ \vdots \\ \mathbb{E}\,X_m \end{pmatrix}. \tag{0.16}$$

The **correlation matrix** of $\boldsymbol{X}$ and $\boldsymbol{Y}$ is the $m \times n$ matrix $\mathbb{E}[\boldsymbol{X}\boldsymbol{Y}^T]$, whose i, jth element is $\mathbb{E}[X_i Y_j]$. The **cross-covariance matrix** $\mathrm{Cov}(\boldsymbol{X}, \boldsymbol{Y})$ of $\boldsymbol{X}$ and $\boldsymbol{Y}$ is the matrix $\mathbb{E}[(\boldsymbol{X} - \mathbb{E}\,\boldsymbol{X})(\boldsymbol{Y} - \mathbb{E}\,\boldsymbol{Y})^T]$, whose i, jth element is $\mathrm{Cov}(X_i, Y_j)$. The covariance of a vector $\boldsymbol{X}$ is $\mathrm{Cov}(\boldsymbol{X}) = \mathrm{Cov}(\boldsymbol{X}, \boldsymbol{X})$. The **conditional expectation** $\mathbb{E}[\boldsymbol{X}|\boldsymbol{Y}]$ of $\boldsymbol{X}$ given $\boldsymbol{Y}$ is a vector whose ith element is $\mathbb{E}[X_i|\boldsymbol{Y}]$.

If the elements of $\boldsymbol{X}$ are uncorrelated, then $\mathrm{Cov}(X_i, X_j) = 0$ for $i \neq j$ and the covariance matrix becomes diagonal. If, in addition, $\mathrm{Cov}(X_i, X_i) = \mathrm{Var}(X_i) = \sigma^2$, i.e., all elements of $\boldsymbol{X}$ have the same variance σ^2, then $\mathrm{Cov}(\boldsymbol{X}) = \sigma^2 I$.

0.6.1 Properties of expectation and covariance

For deterministic matrices A, B, deterministic vectors $\boldsymbol{a}, \boldsymbol{b}$, and random vectors $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{Z}$, we have [1]

1. $\mathbb{E}[\mathsf{A}\boldsymbol{X} + \boldsymbol{a}] = \mathsf{A}\,\mathbb{E}\,\boldsymbol{X} + \boldsymbol{a}$
2. $\mathrm{Cov}(\boldsymbol{X}, \boldsymbol{Y}) = \mathbb{E}[\boldsymbol{X}(\boldsymbol{Y} - \mathbb{E}\,\boldsymbol{Y})^T] = \mathbb{E}[(\boldsymbol{X} - \mathbb{E}\,\boldsymbol{X})\boldsymbol{Y}^T] = \mathbb{E}[\boldsymbol{X}\boldsymbol{Y}^T] - \mathbb{E}\,\boldsymbol{X}\,\mathbb{E}\,\boldsymbol{Y}^T$
3. $\mathbb{E}[(\mathsf{A}\boldsymbol{X})(\mathsf{B}\boldsymbol{Y})^T] = \mathsf{A}\,\mathbb{E}[\boldsymbol{X}\boldsymbol{Y}^T]\mathsf{B}^T$
4. $\mathrm{Cov}(\mathsf{A}\boldsymbol{X} + \boldsymbol{a}, \mathsf{B}\boldsymbol{Y} + \boldsymbol{b}) = \mathsf{A}\,\mathrm{Cov}(\boldsymbol{X}, \boldsymbol{Y})\mathsf{B}^T$
5. $\mathrm{Cov}(\mathsf{A}\boldsymbol{X} + \boldsymbol{a}) = \mathsf{A}\,\mathrm{Cov}(\boldsymbol{X})\mathsf{A}^T$
6. $\mathrm{Cov}(\boldsymbol{W} + \boldsymbol{X}, \boldsymbol{Y} + \boldsymbol{Z}) = \mathrm{Cov}(\boldsymbol{W}, \boldsymbol{Y}) + \mathrm{Cov}(\boldsymbol{W}, \boldsymbol{Z}) + \mathrm{Cov}(\boldsymbol{X}, \boldsymbol{Y}) + \mathrm{Cov}(\boldsymbol{X}, \boldsymbol{Z})$

[5] We use lowercase bold letters to denote deterministic vectors, uppercase bold letters to denote random vectors, and uppercase sans serif letters, such as A, to denote matrices.

Example 0.41. For a random vector $\boldsymbol{X}$ and constants $a, \boldsymbol{b}$, from property 5, we have $\operatorname{Cov}(a\boldsymbol{X} + \boldsymbol{b}) = a^2 \operatorname{Cov}(\boldsymbol{X})$. We also prove this using the other properties. The relevant properties are given in each step.

$$\begin{aligned}
\operatorname{Cov}(a\boldsymbol{X} + \boldsymbol{b}) &= \operatorname{Cov}(a\boldsymbol{X} + \boldsymbol{b}, a\boldsymbol{X} + \boldsymbol{b}) && (0.17)\\
&\overset{6}{=} \operatorname{Cov}(a\boldsymbol{X}, a\boldsymbol{X}) + \operatorname{Cov}(a\boldsymbol{X}, \boldsymbol{b}) + \operatorname{Cov}(\boldsymbol{b}, a\boldsymbol{X}) + \operatorname{Cov}(\boldsymbol{b}, \boldsymbol{b}) && (0.18)\\
&\overset{2}{=} a^2 \operatorname{Cov}(\boldsymbol{X}, \boldsymbol{X}) + 0 + 0 + 0 && (0.19)
\end{aligned}$$

△

References

[1] Bruce Hajek. *Random Processes for Engineers*. Illinois, 2014. URL: http://hajek.ece.illinois.edu/Papers/randomprocJuly14.pdf (visited on 01/30/2017).