DS-GA 1002 Lecture notes 3 September 28, 2015

Probability (continued)

1 Expectation

As we have seen in the previous section, we can characterize the behavior of a random variable completely thorough its cdf, its pmf or its pdf (depending on whether the variable is discrete or continuous). However, it is often useful to work with quantities that describe the behavior of a random variable more parsimoniously. This is the purpose of the expected-value operator, which maps functions of a random variable to a single number.

1.1 Definition

The expected value or expectation of a function of a random variable is an average of its values weighted by their probabilities. The expected value of a function of several random variables is defined analogously.

Definition 1.1 (Expectation for discrete random variables)**.** *Let X be a discrete random variable with range R. The expected value of any real-valued function $g : \mathbb{R} \to \mathbb{R}$ is*

$$\operatorname{E}(g(X)) = \sum_{x \in R} g(x)\, p_X(x). \tag{1}$$

If X, Y are discrete random variables defined on the same probability space with range R', the expected value of a function $h : \mathbb{R}^2 \to \mathbb{R}$ is

$$\operatorname{E}(h(X,Y)) = \sum_{(x,y) \in R'} h(x,y)\, p_{X,Y}(x,y). \tag{2}$$

Definition 1.2 (Expectation for continuous random variables)**.** *Let X be a continuous random variable. The expected value of any real-valued function $g : \mathbb{R} \to \mathbb{R}$ is*

$$\operatorname{E}(g(X)) = \int_{x=-\infty}^{\infty} g(x)\, f_X(x)\, dx. \tag{3}$$

If X, Y are continuous random variables defined on the same probability space, the expected value of a function $h : \mathbb{R}^2 \to \mathbb{R}$ is

$$\operatorname{E}(h(X,Y)) = \int_{x=-\infty}^{\infty} \int_{y=-\infty}^{\infty} h(x,y)\, f_{X,Y}(x,y)\, dx\, dy. \tag{4}$$

In the case of quantities that depend on both continuous and discrete random variables, the product of the marginal and conditional distributions (see (78) in Lecture Notes 2) plays the role of the joint pdf or pmf.

Definition 1.3 (Expectation for quantities that depend on continuous and discrete random variables)**.** *If C is a continuous random variable and D a discrete random variable with range R_D defined on the same probability space, the expected value of a function $h : \mathbb{R}^2 \to \mathbb{R}$ is*

$$\mathrm{E}\left(h\left(C,D\right)\right) = \int_{c=-\infty}^{\infty} \sum_{d \in R_D} h\left(c,d\right) f_C\left(c\right) p_{D|C}\left(d|c\right)\, dc \tag{5}$$

$$= \sum_{d \in R_D} \int_{c=-\infty}^{\infty} h\left(c,d\right) p_D\left(d\right) f_{C|D}\left(c|d\right)\, dc. \tag{6}$$

Note that the expected value of a certain quantity may be infinite or not even exist if the corresponding sum or integral tends towards infinity or has an undefined value. This is illustrated by Examples 1.4 and 1.8 below.

Example 1.4 (St Petersburg paradox)**.** A casino offers you the following game. You will flip an unbiased coin until it lands on heads and the casino will pay you 2^k dollars where k is the number of flips. How much are you willing to pay in order to play?

You compute the expected gain. You model the number of flips X as a geometric random variable such that $p_X\left(k\right) = 1/2^k$. The gain is 2^X so

$$\mathrm{E}\left(\text{Gain}\right) = \sum_{k=1}^{\infty} 2^k \cdot \frac{1}{2^k} = \infty. \tag{7}$$

The expected gain is ∞, but since you only get to play once, perhaps you should not pay an arbitrarily large amount of money. This is known as the St Petersburg paradox.

A fundamental property of the expectation operator is that it is linear.

Theorem 1.5 (Linearity of expectation)**.** *For any constants a and b and any functions $g_1, g_2 : \mathbb{R}^2 \to \mathbb{R}$*

$$\mathrm{E}\left(a\, g_1\left(X,Y\right) + b\, g_2\left(X,Y\right)\right) = a\,\mathrm{E}\left(g_1\left(X,Y\right)\right) + b\,\mathrm{E}\left(g_2\left(X,Y\right)\right). \tag{8}$$

Proof. The theorem follows immediately from the linearity of sums and integrals. □

If two random variables are independent, then the expectation of the product factors into a product of expectations.

Theorem 1.6 (Expectation of functions of independent random variables). *If X, Y are independent random variables defined on the same probability space, and $g, h : \mathbb{R} \to \mathbb{R}$ are univariate real-valued functions, then*

$$\mathrm{E}\left(g\left(X\right)h\left(Y\right)\right) = \mathrm{E}\left(g\left(X\right)\right)\mathrm{E}\left(h\left(Y\right)\right). \tag{9}$$

Proof. The result follows directly from Definitions 2.20 and 2.21 in Lecture Notes 2. We will prove the result for continuous random variables, but the proof for discrete random variables is almost the same.

$$\begin{aligned}
\mathrm{E}\left(g\left(X\right)h\left(Y\right)\right) &= \int_{x=-\infty}^{\infty}\int_{y=-\infty}^{\infty} g\left(x\right)h\left(y\right)f_{X,Y}\left(x,y\right)\,\mathrm{d}x\,\mathrm{d}y && (10)\\
&= \int_{x=-\infty}^{\infty}\int_{y=-\infty}^{\infty} g\left(x\right)h\left(y\right)f_{X}\left(x\right)f_{Y}\left(y\right)\,\mathrm{d}x\,\mathrm{d}y \quad \text{by independence} && (11)\\
&= \mathrm{E}\left(g\left(X\right)\right)\mathrm{E}\left(h\left(Y\right)\right). && (12)
\end{aligned}$$

□

1.2 Mean and variance

The expected value or **mean** of a random variable has special significance. It is the center of mass of the probability measure of the random variable.

Definition 1.7 (Mean). *The mean or first moment of X is the expected value of X:* $\mathrm{E}\left(X\right)$.

If the distribution of a random variable is very *heavy tailed*, which means that the probability of the random variable taking large values decays slowly, the mean of a random variable may be infinite. This was the case of the random variable representing the gain in Example 1.4. The following example shows that the mean may not exist if the value of the corresponding sum or integral is not defined.

Example 1.8 (Cauchy random variable). The pdf of the Cauchy random variable is given by

$$f_X(x) = \frac{1}{\pi(1+x^2)}.$$

By the definition of expected value,

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} \frac{x}{\pi(1+x^2)}\,\mathrm{d}x = \int_{0}^{\infty} \frac{x}{\pi(1+x^2)}\,\mathrm{d}x - \int_{0}^{\infty} \frac{x}{\pi(1+x^2)}\,\mathrm{d}x.$$

Now, by the change of variables $t = x^2$,

$$\int_{0}^{\infty} \frac{x}{\pi(1+x^2)}\,\mathrm{d}x = \int_{0}^{\infty} \frac{1}{2\pi(1+t)}\mathrm{d}t = \lim_{t\to\infty} \frac{\log(1+t)}{2\pi} = \infty,$$

so $\mathrm{E}(X)$ does not exist, as it is the difference of two limits that tend to infinity.

The expected value of the square of a random variable quantifies its expected *energy*.

Definition 1.9 (Second moment)**.** *The mean square value or second moment of X is the expected value of X^2:* $\mathrm{E}\left(X^2\right)$.

The **variance** of a random value quantifies its deviation from the mean. Figure 1 shows the pdfs of Gaussian random variables with different variances. Intuitively, a larger variance implies a larger spread of the values of the random variable from the mean.

Definition 1.10 (Variance and standard deviation)**.** *The variance of X is the mean square deviation from the mean*

$$\begin{aligned} \mathrm{Var}\left(X\right) &:= \mathrm{E}\left(\left(X - \mathrm{E}\left(X\right)\right)^2\right) && (13) \\ &= \mathrm{E}\left(X^2\right) - \mathrm{E}^2\left(X\right). && (14) \end{aligned}$$

The standard deviation σ_X of X is the square root of the variance

$$\sigma_X := \sqrt{\mathrm{Var}\left(X\right)}. \tag{15}$$

The variance operator is not linear, but it is straightforward to determine the variance of a linear function of a random variable.

Lemma 1.11 (Variance of linear functions)**.** *For any constants a and b*

$$\mathrm{Var}\left(a\,X + b\right) = a^2\,\mathrm{Var}\left(X\right). \tag{16}$$

Proof.

$$\begin{aligned} \mathrm{Var}\left(a\,X + b\right) &= \mathrm{E}\left(\left(a\,X + b - \mathrm{E}\left(a\,X + b\right)\right)^2\right) && (17) \\ &= \mathrm{E}\left(\left(a\,X + b - a\mathrm{E}\left(X\right) - b\right)^2\right) && (18) \\ &= a^2\,\mathrm{E}\left(\left(X - a\mathrm{E}\left(X\right)\right)^2\right) && (19) \\ &= a^2\,\mathrm{Var}\left(X\right). && (20) \end{aligned}$$

□

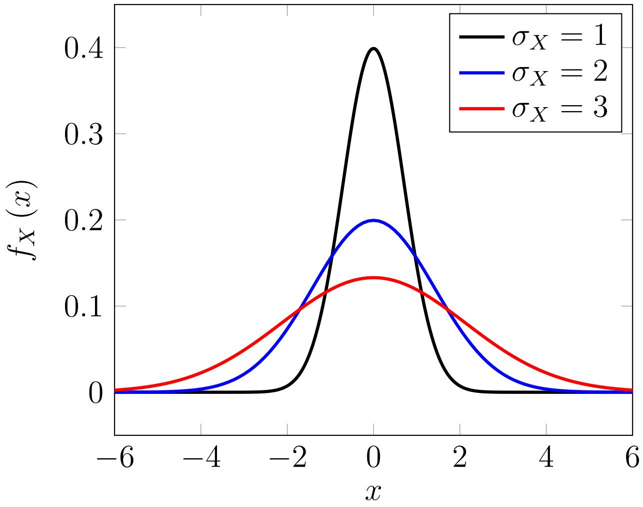

Figure 1: Gaussian random variable with different standard deviations.

We have compiled the means and variances of some important random variables in Table 1. The derivation can be found in Section A of the appendix.

1.3 Bounding probabilities using expectations

It is often much more tractable to estimate the mean and the variance of a random variable than its whole distribution. The following inequalities allow to characterize the behavior of a random valuable to some extent just from its mean and variance.

Theorem 1.12 (Markov's inequality)**.** *Let X be a nonnegative random variable. For any positive constant $a > 0$,*

$$\mathrm{P}\left(X > a\right) \leq \frac{\mathrm{E}\left(X\right)}{a} \tag{21}$$

Proof. Consider the indicator variable $1_{X>a}$. Clearly the random variable

$$X - a\,1_{X>a} \geq 0. \tag{22}$$

So in particular its expectation is non-negative (as it is the sum or integral of a non-negative quantity over the positive real line). By linearity of expectation and the fact that $1_{X>a}$ is a Bernouilli random variable with expectation $\mathrm{P}\left(X > a\right)$ we have

$$\mathrm{E}\left(X\right) \geq a\,\mathrm{E}\left(1_{X>a}\right) = a\,\mathrm{P}\left(X > a\right). \tag{23}$$

□

Random variable	**Parameters**	**Mean**	**Variance**
Bernouilli	p	p	$p(1-p)$
Geometric	p	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Binomial	n, p	np	$np(1-p)$
Poisson	λ	λ	λ
Uniform	a, b	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exponential	λ	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Gaussian	μ, σ	μ	σ^2

Table 1: Means and variance of common random variables, derived in Section A of the appendix.

Example 1.13 (Age of students)**.** You hear that the mean age of NYU students is 20 years, but you know quite a few students that are older than 30. You decide to apply Markov's inequality to bound the fraction of students above 30 by modeling age as a nonnegative random variable A.

$$P(A > 30) \leq \frac{E(A)}{30} = \frac{2}{3},$$

So at most two thirds of the students are over 30.

As you can see from Example 1.16, Markov's inequality is rather loose. The reason is that it barely uses any information about the distribution of the random variable. Chebyshev's inequality takes into account the variance, allowing us to control the deviation of the random variable from its mean.

Theorem 1.14 (Chebyshev's inequality)**.** *For any positive constant $a > 0$ and any random variable X with bounded variance,*

$$P(|X - E(X)| > a) \leq \frac{\operatorname{Var}(X)}{a^2}. \tag{24}$$

Proof. Applying Markov's inequality to the random variable $Y = (X - \mathrm{E}(X))^2$ yields the result. □

An interesting corollary to Chebyshev's inequality shows that if the variance of a random variable is zero, then the random variable is a constant or, to be precise, the probability that it deviates from its mean is zero.

Corollary 1.15. *If* $\mathrm{Var}(X) = 0$ *then* $\mathrm{P}(X \neq \mathrm{E}(X)) = 0$*.*

Proof. Take any $\epsilon > 0$, by Chebyshev's inequality

$$\mathrm{P}(|X - \mathrm{E}(X)| > \epsilon) \leq \frac{\mathrm{Var}(X)}{\epsilon^2} = 0. \tag{25}$$

□

Example 1.16 (Age of students (continued))**.** You are not very satisfied with your bound on the number of students above 30. You ask around and discover that the standard deviation of student age is actually just 3 years. You apply Chebyshev's inequality to this information and obtain

$$\mathrm{P}(|A - \mathrm{E}(A)| > 10) \leq \frac{\mathrm{Var}(A)}{100} = \frac{9}{100}.$$

So actually at least 91% of the students are under 30 (and above 10).

1.4 Correlation and covariance

The covariance of two random variables provides a first-order characterization of the joint behavior of two random variables. It measures whether the two random variables deviate similarly from their respective means.

Definition 1.17 (Covariance)**.** *The* ***covariance*** *of* X *and* Y *is*

$$\begin{aligned} \mathrm{Cov}(X, Y) &:= \mathrm{E}((X - \mathrm{E}(X))(Y - \mathrm{E}(Y))) \qquad (26) \\ &= \mathrm{E}(XY) - \mathrm{E}(X)\,\mathrm{E}(Y). \qquad (27) \end{aligned}$$

If $\mathrm{Cov}(X, Y) = 0$*,* X *and* Y *are* ***uncorrelated****.*

The variance of the sum of two random variables can be expressed in terms of their individual variances and their covariance. As a result, their fluctuations reinforce each other if the covariance is positive and cancel each other if it is negative.

Theorem 1.18 (Variance of the sum of two random variables)**.**

$$\operatorname{Var}(X+Y) = \operatorname{Var}(X) + \operatorname{Var}(Y) + 2\operatorname{Cov}(X,Y). \tag{28}$$

Proof.

$$\begin{aligned}\operatorname{Var}(X+Y) &= \operatorname{E}\left((X+Y-\operatorname{E}(X+Y))^2\right) \qquad (29)\\ &= \operatorname{E}\left((X-\operatorname{E}(X))^2\right) + \operatorname{E}\left((Y-\operatorname{E}(Y))^2\right) + 2\operatorname{E}\left((X-\operatorname{E}(X))(Y-\operatorname{E}(Y))\right) \qquad (30)\\ &= \operatorname{Var}(X) + \operatorname{Var}(Y) + 2\operatorname{Cov}(X,Y). \qquad (31)\end{aligned}$$

□

Corollary 1.19. *If X and Y are uncorrelated, then*

$$\operatorname{Var}(X+Y) = \operatorname{Var}(X) + \operatorname{Var}(Y). \tag{32}$$

The following lemma and example show that independence implies uncorrelation, but uncorrelation does not always imply independence.

Lemma 1.20 (Independence implies uncorrelation)**.** *If two random variables are independent, then they are uncorrelated.*

Proof. By Theorem 1.6, if X and Y are independent

$$\operatorname{Cov}(X,Y) = \operatorname{E}(XY) - \operatorname{E}(X)\operatorname{E}(Y) = \operatorname{E}(X)\operatorname{E}(Y) - \operatorname{E}(X)\operatorname{E}(Y) = 0. \tag{33}$$

□

Example 1.21 (Uncorrelation does not imply independence)**.** Let X and Y be two independent Bernouilli random variables with parameter 1/2. Consider the random variables

$$\begin{aligned}U &= X+Y, \qquad (34)\\ V &= X-Y. \qquad (35)\end{aligned}$$

Note that

$$p_U(0) = \mathrm{P}(X = 0, Y = 0) = \frac{1}{4}, \tag{36}$$

$$p_V(0) = \mathrm{P}(X = 1, Y = 1) + \mathrm{P}(X = 0, Y = 0) = \frac{1}{2}, \tag{37}$$

$$p_{U,V}(0, 0) = \mathrm{P}(\{X = 0, Y = 0\}) = \frac{1}{4} \neq p_U(0)\, p_V(0) = \frac{1}{8}, \tag{38}$$

so U and V are not independent. However, they are uncorrelated as

$$\begin{aligned} \mathrm{Cov}(U, V) &= \mathrm{E}(UV) - \mathrm{E}(U)\,\mathrm{E}(V) && (39) \\ &= \mathrm{E}((X + Y)(X - Y)) - \mathrm{E}(X + Y)\,\mathrm{E}(X - Y) && (40) \\ &= \mathrm{E}\left(X^2\right) - \mathrm{E}\left(Y^2\right) - \mathrm{E}^2(X) + \mathrm{E}^2(Y) = 0. && (41) \end{aligned}$$

The final equality holds because X and Y have the same distribution.

The covariance does not take into account the magnitude of the variances of the random variables involved. The **Pearson correlation coefficient** is obtained by normalizing the covariance using the standard deviations of both variables.

Definition 1.22 (Pearson correlation coefficient)**.** *The Pearson correlation coefficient of two random variables X and Y is*

$$\rho_{X,Y} := \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}. \tag{42}$$

Although it might not be immediately obvious, the magnitude of the correlation coefficient is bounded by one. A useful interpretation of this coefficient is that it quantifies to what extent X and Y are linearly related. In fact, if it is equal to 1 or -1 then one of the variables is a linear function of the other! This follows from the notorious Cauchy-Schwarz inequality.

Theorem 1.23 (Cauchy-Schwarz inequality)**.** *For any random variables X and Y defined on the same probability space*

$$|\mathrm{E}(XY)| \leq \sqrt{\mathrm{E}(X^2)\,\mathrm{E}(Y^2)}. \tag{43}$$

Assume $\mathrm{E}(X^2) \neq 0$*,*

$$\mathrm{E}(XY) = \sqrt{\mathrm{E}(X^2)\,\mathrm{E}(Y^2)} \iff Y = \sqrt{\frac{\mathrm{E}(Y^2)}{\mathrm{E}(X^2)}}X, \tag{44}$$

$$\mathrm{E}(XY) = -\sqrt{\mathrm{E}(X^2)\,\mathrm{E}(Y^2)} \iff Y = -\sqrt{\frac{\mathrm{E}(Y^2)}{\mathrm{E}(X^2)}}X. \tag{45}$$

Proof. If $\mathrm{E}\left(X^2\right) = 0$ then $X = 0$ by Corollary 1.15 $X = 0$ with probability one, which implies $\mathrm{E}\left(XY\right) = 0$ and consequently that equality holds in (43). The same is true if $\mathrm{E}\left(Y^2\right) = 0$.

Now assume that $\mathrm{E}\left(X^2\right) \neq 0$ and $\mathrm{E}\left(Y^2\right) \neq 0$. Let us define the constants $a = \sqrt{\mathrm{E}\left(Y^2\right)}$ and $b = \sqrt{\mathrm{E}\left(X^2\right)}$. By linearity of expectation,

$$\mathrm{E}\left((aX + bY)^2\right) = a^2\mathrm{E}\left(X^2\right) + b^2\mathrm{E}\left(Y^2\right) + 2\,a\,b\,\mathrm{E}\left(XY\right) \tag{46}$$

$$= 2\left(\mathrm{E}\left(X^2\right)\mathrm{E}\left(Y^2\right) + \sqrt{\mathrm{E}\left(X^2\right)\mathrm{E}\left(Y^2\right)}\mathrm{E}\left(XY\right)\right), \tag{47}$$

$$\mathrm{E}\left((aX - bY)^2\right) = a^2\mathrm{E}\left(X^2\right) + b^2\mathrm{E}\left(Y^2\right) - 2\,a\,b\,\mathrm{E}\left(XY\right) \tag{48}$$

$$= 2\left(\mathrm{E}\left(X^2\right)\mathrm{E}\left(Y^2\right) - \sqrt{\mathrm{E}\left(X^2\right)\mathrm{E}\left(Y^2\right)}\mathrm{E}\left(XY\right)\right). \tag{49}$$

The expectation of a non-negative quantity is nonzero because the integral or sum of a non-negative quantity is non negative. Consequently, the left-hand side of (46) and (48) is non-negative, so (47) and (49) are both non-negative, which implies (43).

Let us prove (45) by proving both implications.

($\Rightarrow$). Assume $\mathrm{E}\left(XY\right) = -\sqrt{\mathrm{E}\left(X^2\right)\mathrm{E}\left(Y^2\right)}$. Then (47) equals zero, so

$$\mathrm{E}\left(\left(\sqrt{\mathrm{E}\left(X^2\right)}X + \sqrt{\mathrm{E}\left(X^2\right)}Y\right)^2\right) = 0, \tag{50}$$

which by Corollary 1.15 means that $\sqrt{\mathrm{E}\left(Y^2\right)}X = -\sqrt{\mathrm{E}\left(X^2\right)}Y$ with probability one.

($\Leftarrow$). Assume $Y = -\frac{\mathrm{E}\left(Y^2\right)}{\mathrm{E}\left(X^2\right)}X$. Then one can easily check that (47) equals zero, which implies $\mathrm{E}\left(XY\right) = -\sqrt{\mathrm{E}\left(X^2\right)\mathrm{E}\left(Y^2\right)}$.

The proof of (44) is almost identical (using (46) instead of (47)). □

Corollary 1.24. *The Pearson correlation coefficient of two random variables X and Y satisfies*

$$|\rho_{X,Y}| \leq 1, \tag{51}$$

with equality if and only if there is a linear relationship between X and Y

$$|\rho_{X,Y}| = 1 \iff Y = c\,X + d. \tag{52}$$

where

$$c := \begin{cases} \frac{\sigma_Y}{\sigma_X} & \text{if } \rho_{X,Y} = 1, \\ -\frac{\sigma_Y}{\sigma_X} & \text{if } \rho_{X,Y} = -1, \end{cases} \qquad b := \mathrm{E}\left(Y\right) - c\,\mathrm{E}\left(X\right). \tag{53}$$

Proof. The result follows from defining

$$U := X - \mathrm{E}\left(X\right), \tag{54}$$

$$V := Y - \mathrm{E}\left(Y\right). \tag{55}$$

From the definition of the variance and the correlation coefficient,

$$\mathrm{E}\left(U^2\right) = \mathrm{Var}\left(X\right), \tag{56}$$

$$\mathrm{E}\left(V^2\right) = \mathrm{Var}\left(Y\right) \tag{57}$$

$$\rho_{X,Y} = \frac{\mathrm{E}\left(UV\right)}{\sqrt{\mathrm{E}\left(U^2\right)\mathrm{E}\left(V^2\right)}}. \tag{58}$$

The result now follows from applying Theorem 1.23 to U and V. □

Figure 2 shows the joint pdf of two Gaussian random variables for different values of the correlation coefficient. If the coefficient is zero, then the joint pdf has a spherical form. If the coefficient approaches 1, then the joint pdf becomes skewed so that the two variables have similar values. If the coefficient approaches -1, then the same happens, only now the random variables will tend to have similar values with opposite sign.

1.5 Conditional expectation

The expectation of a function of two random variables conditioned on one of them taking a certain value can be computed using the conditional pmf or pdf.

$$\mathrm{E}\left(g\left(X,Y\right)|X=x\right) = \sum_{y \in R} g\left(x,y\right) p_{Y|X}\left(y|x\right), \tag{59}$$

if Y is discrete and has range R, whereas

$$\mathrm{E}\left(g\left(X,Y\right)|X=x\right) = \int_{y=-\infty}^{\infty} g\left(x,y\right) f_{Y|X}\left(y|x\right)\,\mathrm{d}y, \tag{60}$$

if Y is continuous.

Note that $\mathrm{E}\left(g\left(X,Y\right)|X=x\right)$ can actually be interpreted as a *function of x* since it maps every value of x to a real number. We can then define the **conditional expectation** of $g\left(X,Y\right)$ given X as follows.

Definition 1.25 (Conditional expectation)**.** *The conditional expectation of $g\left(X,Y\right)$ given X is*

$$\mathrm{E}\left(g\left(X,Y\right)|X\right) := h\left(X\right), \tag{61}$$

where

$$h\left(x\right) := \mathrm{E}\left(g\left(X,Y\right)|X=x\right). \tag{62}$$

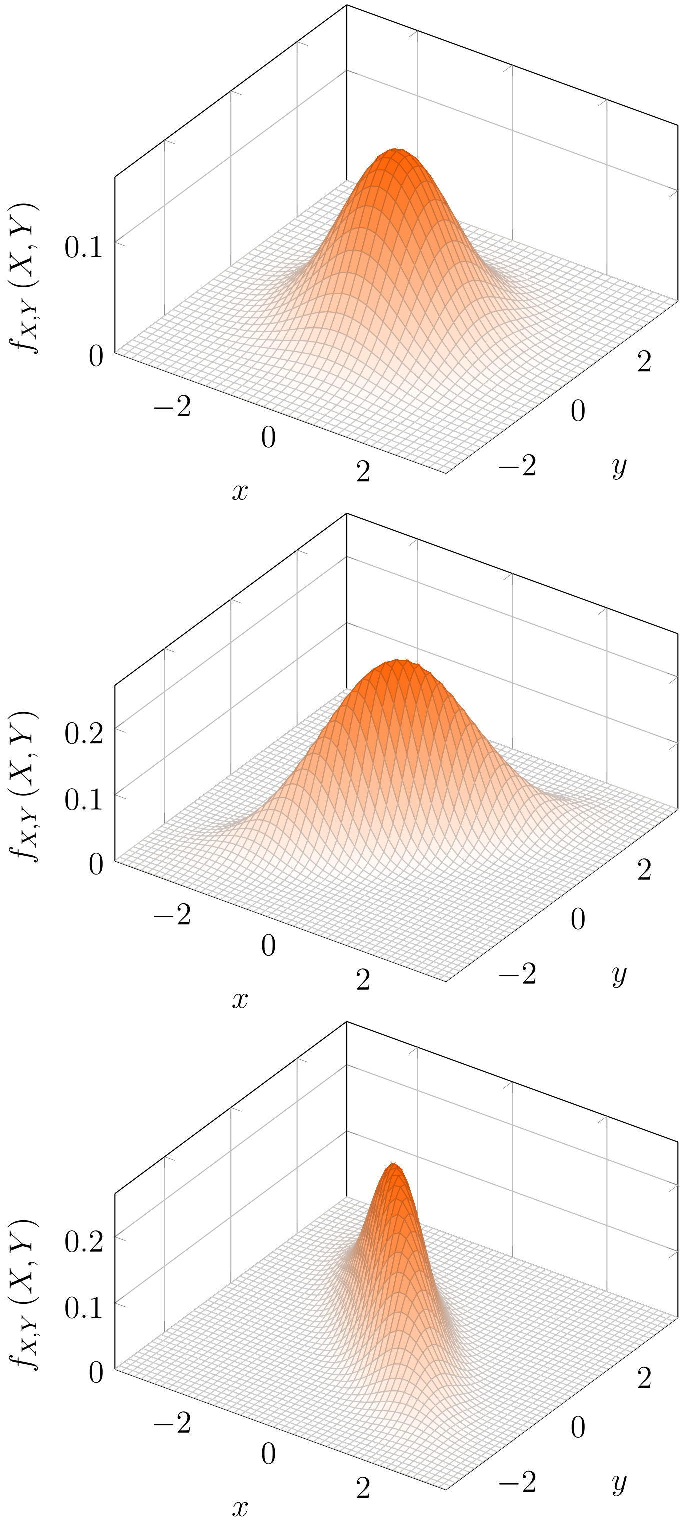

Figure 2: Joint pdf of a bivariate Gaussian random variable X, Y such that X and Y have zero mean and unit variance for $\rho = 0$ (top), $\rho = 0.8$ (center) and $\rho = -0.8$ (bottom).

Beware the confusing definition, the conditional expectation is actually a random variable!

Iterated expectation is a useful tool to compute the expected value of a certain quantity that depends on several random variables. The idea is that the expected value can be obtained as the expectation of the conditional expectation.

Theorem 1.26 (Iterated expectation)**.** *For any random variables X and Y and any function $g : \mathbb{R}^2 \to \mathbb{R}$*

$$\mathrm{E}\left(g\left(X,Y\right)\right) = \mathrm{E}\left(\mathrm{E}\left(g\left(X,Y\right)|X\right)\right). \tag{63}$$

Proof. We prove the result for continuous random variables, the proof for discrete random variables, and for quantities that depend on both continuous and discrete random variables, is almost identical. To make the explanation clearer, we define

$$\begin{align} h\left(x\right) &:= \mathrm{E}\left(g\left(X,Y\right)|X=x\right) \tag{64}\\ &= \int_{y=-\infty}^{\infty} g\left(x,y\right) f_{Y|X}\left(y|x\right) \,\mathrm{d}y. \tag{65}\end{align}$$

Now,

$$\begin{align} \mathrm{E}\left(\mathrm{E}\left(g\left(X,Y\right)|X\right)\right) &= \mathrm{E}\left(h\left(X\right)\right) \tag{66}\\ &= \int_{x=-\infty}^{\infty} h\left(x\right) f_X\left(x\right) \,\mathrm{d}x \tag{67}\\ &= \int_{x=-\infty}^{\infty}\int_{y=-\infty}^{\infty} f_X\left(x\right) f_{Y|X}\left(y|x\right) g\left(x,y\right) \,\mathrm{d}y\,\mathrm{d}x \tag{68}\\ &= \mathrm{E}\left(g\left(X,Y\right)\right). \tag{69}\end{align}$$

□

Example 1.27 (Lost cellphones)**.** Chris loses his cellphone quite often when he is traveling. He calculates that the probability that he loses it on any given trip is between 0.1 and 0.2. He decides to model his uncertainty over that probability as a uniform random variable in $[0.1, 0.2]$. If he takes n trips next year, what is the expected number of times he will lose his cellphone? Assume that he never learns to be careful (i.e. the event that he loses his cellphone on a trip is independent of whether he loses it in other trips).

Let L be the probability that Chris loses his cellphone in a given trip and C the number of lost cellphones. Under the independence assumption, the number of lost cellphones is distributed as a binomial random variable with parameter L. This implies that

$$\mathrm{E}\left(C|L=l\right) = nl, \tag{70}$$

so the conditional expectation of C given L is the random variable nL. By iterated expectation,

$$\mathrm{E}\left(C\right) = \mathrm{E}\left(\mathrm{E}\left(C|L\right)\right) = \int_{0.1}^{0.2} \frac{nl}{0.1} \mathrm{d}l = 0.15\, n. \tag{71}$$

According to the probabilistic model on average he will lose $0.15\, n$ phones.

---

2 Random vectors

In this section, we extend the definitions in Section 2 of Lecture Notes 2 concerning bivariate distributions to multiple random variables that are jointly distributed. When several random variables are defined on the same probability space, we group them into a vector.

Definition 2.1 (Random vector)**.** *Let $X_1, X_2, \ldots, X_n$ be n random variables defined on the same probability space. We define a random vector as*

$$\boldsymbol{X} := \begin{bmatrix} X_1 \\ X_2 \\ \cdots \\ X_n \end{bmatrix}. \tag{72}$$

2.1 Joint distribution of a random vector

The distribution of a discrete random vector is completely determined by its joint probability mass function.

Definition 2.2 (Joint probability mass function)**.** *Let $\boldsymbol{X} : \Omega \to \mathbb{R}^n$ be a random vector of dimension n on a probability space $(\Omega, \mathcal{F}, \mathrm{P})$. The joint pmf of $\boldsymbol{X}$ is defined as*

$$p_{\boldsymbol{X}}\left(\boldsymbol{x}\right) := \mathrm{P}\left(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\right). \tag{73}$$

In words, $p_{\boldsymbol{X}}\left(\boldsymbol{x}\right)$ is the probability of $\boldsymbol{X}$ being equal to $\boldsymbol{x}$.

As in the case of two-dimensional random variables, we characterize the distribution of random vectors defined on continuous spaces by restricting them to belong to multidimensional Borel sets, which can be thought of as unions, intersections, etc. of Cartesian products of intervals. As a result, the distribution of a continuous random vector is completely determined by its joint cumulative distribution function and, if it exists, its joint probability density function.

Definition 2.3 (Joint cumulative distribution function)**.** *Let $\boldsymbol{X} : \Omega \to \mathbb{R}^n$ be a random vector of dimension n on a probability space $(\Omega, \mathcal{F}, \mathrm{P})$. The joint cdf of $\boldsymbol{X}$ is defined as*

$$F_{\boldsymbol{X}}(\boldsymbol{x}) := \mathrm{P}\left(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n\right). \tag{74}$$

In words, $F_{\boldsymbol{X}}(\boldsymbol{x})$ is the probability that $X_i \leq x_i$ for all $i = 1, 2, \ldots, n$.

Definition 2.4 (Joint probability density function)**.** *If the joint cdf of a random vector $\boldsymbol{X}$ is differentiable, then its joint pdf is defined as*

$$f_{\boldsymbol{X}}(\boldsymbol{x}) := \frac{\partial^n F_{\boldsymbol{X}}(\boldsymbol{x})}{\partial x_1 \, \partial x_2 \cdots \partial x_n}. \tag{75}$$

In order to ease the exposition, we introduce some notation.

Definition 2.5 (Subvector)**.** *A subvector of a vector $\boldsymbol{x}$ is the vector formed by a subset of its entries, indexed by a set $\mathcal{I} = \{i_1, i_2, \ldots, i_m\} \subseteq \{1, 2, \ldots, n\}$,*

$$\boldsymbol{x}_{\mathcal{I}} := \begin{bmatrix} x_{i_1} \\ x_{i_2} \\ \cdots \\ x_{i_m} \end{bmatrix}. \tag{76}$$

Definition 2.6 (Summing and integrating over a vector)**.** *Given a subset of the entries of a vector $\mathcal{I} = \{i_1, i_2, \ldots, i_m\} \subseteq \{1, 2, \ldots, n\}$ we denote the operation of summing or integrating over those entries by*

$$\sum_{\mathcal{I}} g(\boldsymbol{x}) := \sum_{i_1 \in R_1} \sum_{i_2 \in R_2} \cdots \sum_{i_m \in R_m} g(x_1, x_2, \ldots, x_n), \tag{77}$$

$$\int_{\mathcal{I}} g(\boldsymbol{x}) \, \mathbf{d}\boldsymbol{x}_{\mathcal{I}} := \int_{i_1=-\infty}^{\infty} \int_{i_2=-\infty}^{\infty} \cdots \int_{i_m=-\infty}^{\infty} g(x_1, x_2, \ldots, x_n) \; dx_{i_1} \, dx_{i_2} \cdots dx_{i_m}, \tag{78}$$

where R_i represents the range of X_i for $1 \leq i \leq n$.

As in the case of bivariate distributions, it follows from the axioms of probability measures that

$$p_{\mathbf{X}}(\mathbf{x}) \geq 0, \tag{79}$$

$$f_{\mathbf{X}}(\mathbf{x}) \geq 0, \tag{80}$$

$$\sum_{\{1,2,\ldots,n\}} p_{\mathbf{X}}(\mathbf{x}) = 1, \tag{81}$$

$$\int_{\{1,2,\ldots,n\}} f_{\mathbf{X}}(\mathbf{x}) \, \boldsymbol{d}\mathbf{x} = 1, \tag{82}$$

and for any set $\mathcal{S} \subseteq \mathbb{R}^n$

$$\mathrm{P}\left(\mathbf{X} \in \mathcal{S}\right) = \sum_{\mathbf{x} \in \mathcal{S}} p_{\mathbf{X}}\left(\mathbf{x}\right) \tag{83}$$

in the case of discrete random vectors and

$$\mathrm{P}\left(\mathbf{X} \in \mathcal{S}\right) = \int_{\mathbf{x} \in \mathcal{S}} f_{\mathbf{X}}\left(\mathbf{x}\right) \boldsymbol{d}\mathbf{x} \tag{84}$$

in the case of continuous random vectors.

Generalizing the concept of marginalization, to obtain the joint distribution of a subset of the variables in the vector we either sum or integrate over the rest of the variables.

Theorem 2.7 (Marginalization)**.** *The marginal pmf of a discrete random subvector $\boldsymbol{X}_{\mathcal{I}}$, $\mathcal{I} \subseteq \{1, 2, \ldots, n\}$, is obtained by summing over the components not included in the subvector*

$$p_{\boldsymbol{X}_{\mathcal{I}}}\left(\boldsymbol{x}_{\mathcal{I}}\right) = \sum_{\{1,2,\ldots,n\}/\mathcal{I}} p_{\boldsymbol{X}}\left(\boldsymbol{x}\right). \tag{85}$$

Similarly, the marginal pdf is obtained by integrating over the rest of the components,

$$f_{\boldsymbol{X}_{\mathcal{I}}}\left(\boldsymbol{x}_{\mathcal{I}}\right) = \int_{\{1,2,\ldots,n\}/\mathcal{I}} f_X\left(\boldsymbol{x}\right) \mathbf{d}\boldsymbol{x}_{\{1,2,\ldots,n\}/\mathcal{I}}. \tag{86}$$

Proof. Let $\{1, 2, \ldots, n\}/\mathcal{I} = \{j_1, j_2, \ldots, j_{n-m}\}$. For discrete random vectors, (85) follows from applying (83) to the event

$$\{X_{i_1} = x_{i_1}, \ldots, X_{i_m} = x_{i_m}\} \tag{87}$$

$$= \bigcup_{x_{j_1} \in R_1} \cdots \bigcup_{x_{j_{n-m}} \in R_{n-m}} \left\{X_{i_1} = x_{i_1}, \ldots, X_{i_m} = x_{i_m}, X_{j_1} = x_{j_1}, \ldots, X_{j_{n-m}} = x_{j_{n-m}}\right\}, \tag{88}$$

where R_i represents the range of X_i for $1 \leq i \leq n$. For continuous random vectors, (86) follows from applying (84) to the event

$$\{X_{i_1} \leq x_{i_1}, \ldots, X_{i_m} \leq x_{i_m}\} \tag{89}$$

$$= \bigcup_{x_{j_1}=-\infty}^{\infty} \cdots \bigcup_{x_{j_{n-m}}=-\infty}^{\infty} \left\{X_{i_1} \leq x_{i_1}, \ldots, X_{i_m} \leq x_{i_m}, X_{j_1} = x_{j_1}, \ldots, X_{j_{n-m}} = x_{j_{n-m}}\right\} \tag{90}$$

in order to obtain the marginal cdf and then differentiating. □

Applying the same ideas as in the bivariate case, we define the conditional distribution of a subvector given the rest of the random vector.

Definition 2.8 (Conditional pmf and pdf)**.** *The conditional pmf of a discrete random subvector* $\boldsymbol{X}_{\mathcal{I}}$*,* $\mathcal{I} \subseteq \{1, 2, \ldots, n\}$*, given the subvector* $\boldsymbol{X}_{\{1,\ldots,n\}/\mathcal{I}}$ *is*

$$p_{\boldsymbol{X}_{\mathcal{I}}|\boldsymbol{X}_{\{1,\ldots,n\}/\mathcal{I}}}\left(\boldsymbol{x}_{\mathcal{I}}|\boldsymbol{x}_{\{1,\ldots,n\}/\mathcal{I}}\right) := \frac{p_{\boldsymbol{X}}\left(\boldsymbol{x}\right)}{p_{\boldsymbol{X}_{\{1,\ldots,n\}/\mathcal{I}}}\left(\boldsymbol{x}_{\{1,\ldots,n\}/\mathcal{I}}\right)}. \tag{91}$$

The conditional pdf of a discrete random subvector $\boldsymbol{X}_{\mathcal{I}}$*,* $\mathcal{I} \subseteq \{1, 2, \ldots, n\}$*, given the subvector* $\boldsymbol{X}_{\{1,\ldots,n\}/\mathcal{I}}$ *is*

$$f_{\boldsymbol{X}_{\mathcal{I}}|\boldsymbol{X}_{\{1,\ldots,n\}/\mathcal{I}}}\left(\boldsymbol{x}_{\mathcal{I}}|\boldsymbol{x}_{\{1,\ldots,n\}/\mathcal{I}}\right) := \frac{f_{\boldsymbol{X}}\left(\boldsymbol{x}\right)}{f_{\boldsymbol{X}_{\{1,\ldots,n\}/\mathcal{I}}}\left(\boldsymbol{x}_{\{1,\ldots,n\}/\mathcal{I}}\right)}. \tag{92}$$

It is often useful to represent the joint pmf or pdf of a random vector by factoring it into conditional pmfs or pdfs using the **chain rule**.

Lemma 2.9 (Chain rule for random vectors)**.** *The joint pmf of a random vector* $\boldsymbol{X}$ *can be decomposed into*

$$\begin{align} p_{\boldsymbol{X}}\left(\boldsymbol{x}\right) &= p_{X_1}\left(x_1\right) p_{X_2|X_1}\left(x_2|x_1\right) \ldots p_{X_n|X_1,\ldots,X_{n-1}}\left(x_n|x_1,\ldots,x_{n-1}\right) \tag{93}\\ &= \prod_{i=1}^{n} p_{X_i|\boldsymbol{X}_{\{1,\ldots,i-1\}}}\left(x_i|\boldsymbol{x}_{\{1,\ldots,i-1\}}\right). \tag{94}\end{align}$$

The joint pdf of a random vector $\boldsymbol{X}$ *can be decomposed into*

$$\begin{align} f_{\boldsymbol{X}}\left(\boldsymbol{x}\right) &= f_{X_1}\left(x_1\right) f_{X_2|X_1}\left(x_2|x_1\right) \ldots f_{X_n|X_1,\ldots,X_{n-1}}\left(x_n|x_1,\ldots,x_{n-1}\right) \tag{95}\\ &= \prod_{i=1}^{n} f_{X_i|\boldsymbol{X}_{\{1,\ldots,i-1\}}}\left(x_i|\boldsymbol{x}_{\{1,\ldots,i-1\}}\right). \tag{96}\end{align}$$

Where the order is ***arbitrary****, you can reorder the components of the vector in any way you like.*

Proof. The result follows from applying the definitions of conditional pmf and pdf recursively. □

Example 2.10 (Desert)**.** Dani and Felix are traveling through the desert in Arizona. They become concerned that their car might break down and decide to build a probabilistic model to evaluate the risk. They model the time that will pass until the car breaks down as an exponential random variable with a parameter that depends on the state of the motor and the state of the road. These three quantities form a random vector

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix}, \tag{97}$$

where X_1 is the time until the car breaks down, X_2 is the state of the motor and X_3 is the state of the road. Unfortunately they have no idea what the state of the motor is so they assume that it is uniform between 0 (no problem with the motor) and 1 (the motor is almost dead). Similarly, they have no information about the road, so they also assume that its state is a uniform random variable between 0 (no problem with the road) and 1 (the road is terrible). In addition, they assume that the states of the road and the car are independent and that the parameter of the exponential random variable that represents the time in hours until there is a breakdown is equal to $X_2 + X_3$.

To find the joint distribution of the random vector, we apply the chain rule to obtain,

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}) &= f_{X_2}(x_2)\, f_{X_3|X_2}(x_3|x_2)\, f_{X_1|X_2,X_3}(x_1|x_2,x_3) && (98)\\ &= f_{X_2}(x_2)\, f_{X_3}(x_3|x_2)\, f_{X_1|X_2,X_3}(x_1|x_2,x_3) \quad \text{(by independence of } X_2 \text{ and } X_3\text{)} && (99)\\ &= \begin{cases} (x_2+x_3)\, e^{-(x_2+x_3)x_1} & \text{for } x_1 \geq 0,\ 0 \leq x_2 \leq 1,\ 0 \leq x_3 \leq 1, \\ 0 & \text{otherwise.} \end{cases} && (100) \end{aligned}$$

Note that we start with X_2 and X_3 because we know their marginal distribution, whereas we only know the conditional distribution of X_1 given X_2 and X_3.

After 15 minutes, the car breaks down. The road seems OK, about a 0.2 in the scale they defined for the value of X_3, so they naturally wonder about the state the motor was in. This is given by the conditional distribution of X_2 given X_1 and X_3.

To compute it, we first need to compute the distribution of X_1 and X_3 by marginalizing over X_2. In order to simplify the computations, we use the following simple lemma.

Lemma 2.11. *For any constant $c > 0$,*

$$\int_0^1 e^{-cx}\, dx = \frac{1-e^{-c}}{c}, \tag{101}$$

$$\int_0^1 x e^{-cx}\, dx = \frac{1-(1+c)\,e^{-c}}{c^2}. \tag{102}$$

Proof. Equation (101) is obtained using the antiderivative of the exponential function (itself), whereas integrating by parts yields (102). □

We have

$$\begin{aligned} f_{\mathbf{X}_{\{1,3\}}}\left(\mathbf{x}_{\{1,3\}}\right) &= e^{-x_1x_3}\left(\int_{x_2=0}^{1} x_2 e^{-x_1x_2}\,\mathrm{d}x_2 + x_3\int_{x_2=0}^{1} e^{-x_1x_2}\,\mathrm{d}x_2\right) && (103)\\ &= e^{-x_1x_3}\left(\frac{1-(1+x_1)\,e^{-x1}}{x_1^2} + \frac{x_3\,(1-e^{-x_1})}{x_1}\right) \quad \text{by (101) and (102)} && (104)\\ &= \frac{e^{-x_1x_3}}{x_1^2}\left(1+x_1x_3 - e^{-x1}\,(1+x_1+x_1x_3)\right), && (105) \end{aligned}$$

for $x_1 \geq 0,\ 0 \leq x_3 \leq 1$.

The conditional pdf of X_2 given X_1 and X_3 is

$$f_{X_2|\mathbf{X}_{\{1,3\}}}\left(x_2|\mathbf{x}_{\{1,3\}}\right) = \frac{f_{\mathbf{X}}\left(\mathbf{x}\right)}{f_{\mathbf{X}_{\{1,3\}}}} \tag{106}$$

$$= \frac{\left(x_2 + x_3\right) e^{-(x_2+x_3)x_1}}{\frac{e^{-x_1 x_3}}{x_1^2}\left(1 + x_1 x_3 - e^{-x1}\left(1 + x_1 + x_1 x_3\right)\right)} \tag{107}$$

$$= \frac{\left(x_2 + x_3\right) x_1^2 e^{-x_1 x_2}}{1 + x_1 x_3 - e^{-x1}\left(1 + x_1 + x_1 x_3\right)}, \tag{108}$$

for $x_1 \geq 0,\ 0 \leq x_2 \leq 1,\ 0 \leq x_3 \leq 1$. Plugging in the observed values, the conditional pdf is equal to

$$f_{X_2|\mathbf{X}_{\{1,3\}}}\left(x_2 \,\middle|\, \begin{bmatrix}0.25\\0.2\end{bmatrix}\right) = \frac{\left(x_2 + 0.2\right) 0.25^2 e^{-0.25 x_2}}{1 + 0.25 \cdot 0.2 - e^{-0.25}\left(1 + 0.25 + 0.25 \cdot 0.2\right)} \tag{109}$$

$$\approx 1.66\left(x_2 + 0.2\right) e^{-0.25 x_2}. \tag{110}$$

for $0 \leq x_2 \leq 1$ and to zero otherwise. The pdf is plotted in Figure 3. According to the model, it seems quite likely that the state of the motor was not good.

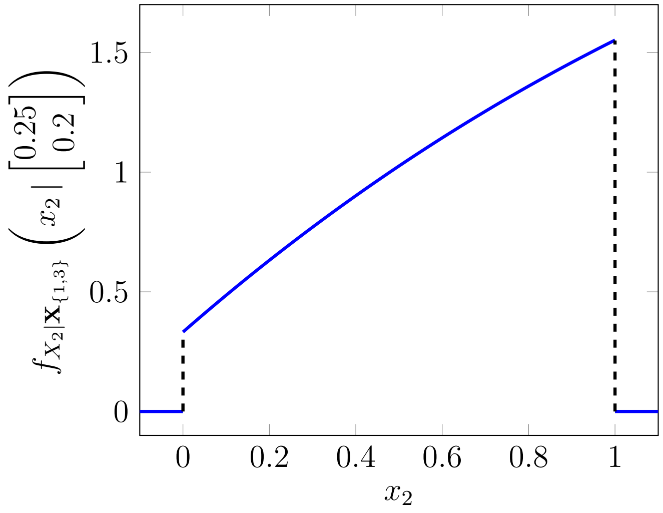

Figure 3: Conditional pdf of X_2 given $X_1 = 0.25$ and $X_3 = 0.2$ in Example 2.10.

2.2 Independence of random vectors

The components of a random vector are **mutually independent** if their joint pmf or pdf factors into the individual marginal pmfs or pdfs.

Definition 2.12 (Jointly independent random variables). *The components of a random vector $\boldsymbol{X}$ are mutually independent if*

$$p_{\boldsymbol{X}}(\boldsymbol{x}) = \prod_{i=1}^{n} p_{X_i}(x_i) \tag{111}$$

in the case of discrete random vectors and if

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \prod_{i=1}^{n} f_{X_i}(x_i) \tag{112}$$

in the case of continuous random vectors.

The definition of conditional independence is very similar.

Definition 2.13 (Mutually conditionally independent random variables). *The components of a subvector $\boldsymbol{X}_{\mathcal{I}}$, $\mathcal{I} \subseteq \{1, 2, \ldots, n\}$ are mutually conditionally independent given another subvector $\boldsymbol{X}_{\mathcal{J}}$, $\mathcal{J} \subseteq \{1, 2, \ldots, n\}$ and $\mathcal{I} \cap \mathcal{J} = \emptyset$ if*

$$p_{\boldsymbol{X}_{\mathcal{I}}|\boldsymbol{X}_{\mathcal{J}}}(\boldsymbol{x}_{\mathcal{I}}|\boldsymbol{x}_{\mathcal{J}}) = \prod_{i \in \mathcal{I}} p_{X_i|\boldsymbol{X}_{\mathcal{J}}}(x_i|\boldsymbol{x}_{\mathcal{J}}) \tag{113}$$

in the case of discrete random vectors and

$$f_{\boldsymbol{X}_{\mathcal{I}}|\boldsymbol{X}_{\mathcal{J}}}(\boldsymbol{x}_{\mathcal{I}}|\boldsymbol{x}_{\mathcal{J}}) = \prod_{i \in \mathcal{I}} f_{X_i|\boldsymbol{X}_{\mathcal{J}}}(x_i|\boldsymbol{x}_{\mathcal{J}}) \tag{114}$$

in the case of continuous random vectors.

As we saw in Examples 4.3 and 4.4 of Lecture Notes 1, independence does not imply conditional independence or vice versa. The following example shows that pairwise independence does not imply mutual independence.

Example 2.14 (Pairwise independence does not imply joint independence). Let X_1 and X_2 be the outcomes of independent unbiased coin flips. Let X_3 be the indicator of the event X_1 *and* X_2 *have the same outcome*,

$$X_3 = \begin{cases} 1 & \text{if } X_1 = X_2, \\ 0 & \text{if } X_1 \neq X_2. \end{cases} \tag{115}$$

The pmf of X_3 is

$$\begin{aligned} p_{X_3}(1) &= \mathrm{P}\left(\{X_1 = 1, X_2 = 1\} \cup \{X_1 = 0, X_2 = 0\}\right) && (116)\\ &= \mathrm{P}\left(X_1 = 1, X_2 = 1\right) \mathrm{P}\left(X_1 = 0, X_2 = 0\right) = \frac{1}{2}, && (117)\\ p_{X_3}(0) &= \mathrm{P}\left(\{X_1 = 0, X_2 = 1\} \cup \{X_1 = 0, X_2 = 1\}\right) && (118)\\ &= \mathrm{P}\left(X_1 = 0, X_2 = 1\right) \mathrm{P}\left(X_1 = 0, X_2 = 1\right) = \frac{1}{2}. && (119) \end{aligned}$$

X_1 and X_2 are independent by assumption. X_1 and X_3 are independent because

$$\begin{aligned} p_{X_1,X_3}(0,0) &= \mathrm{P}\left(X_1 = 0, X_2 = 1\right) = \frac{1}{4} = p_{X_1}(0)\, p_{X_3}(0), && (120)\\ p_{X_1,X_3}(1,0) &= \mathrm{P}\left(X_1 = 1, X_2 = 0\right) = \frac{1}{4} = p_{X_1}(1)\, p_{X_3}(0), && (121)\\ p_{X_1,X_3}(0,1) &= \mathrm{P}\left(X_1 = 0, X_2 = 0\right) = \frac{1}{4} = p_{X_1}(0)\, p_{X_3}(1), && (122)\\ p_{X_1,X_3}(1,1) &= \mathrm{P}\left(X_1 = 1, X_2 = 1\right) = \frac{1}{4} = p_{X_1}(1)\, p_{X_3}(1). && (123) \end{aligned}$$

X_2 and X_3 are independent too (the reasoning is the same).

However, are X_1, X_2 and X_3 mutually independent?

$$p_{X_1,X_2,X_3}(1,1,1) = \mathrm{P}\left(X_1 = 1, X_2 = 1\right) = \frac{1}{4} \neq p_{X_1}(1)\, p_{X_2}(1)\, p_{X_3}(1) = \frac{1}{8}. \qquad (124)$$

They are not.

2.3 Mean and covariance of a random vector

The expected value of a function $g : \mathbb{R}^n \to \mathbb{R}$ that depends on the value of a vector is defined in the same way as in the case of bivariate distributions.

Definition 2.15 (Expected value of a function of a random vector)**.** *The expected value of a function $g : \mathbb{R}^n \to \mathbb{R}$ applied to a random vector $\boldsymbol{X}$ is defined as*

$$\mathrm{E}\left(g\left(\boldsymbol{X}\right)\right) = \sum_{\{1,\ldots,n\}} g\left(\boldsymbol{x}\right) p_{\boldsymbol{X}}\left(\boldsymbol{x}\right) \qquad (125)$$

if X is discrete and

$$\mathrm{E}\left(g\left(\boldsymbol{X}\right)\right) = \int_{\{1,\ldots,n\}} g\left(\boldsymbol{x}\right) f_{\boldsymbol{X}}\left(\boldsymbol{x}\right) \, \mathrm{d}\boldsymbol{x} \qquad (126)$$

if X is continuous.

The **mean** of a random vector is the vector of the means of its components.

Definition 2.16 (Mean of a random vector)**.** *The mean of a random vector $\boldsymbol{X}$ is defined as*

$$\mathrm{E}(\boldsymbol{X}) := \begin{bmatrix} \mathrm{E}(X_1) \\ \mathrm{E}(X_2) \\ \cdots \\ \mathrm{E}(X_n) \end{bmatrix}. \tag{127}$$

Similarly, if the entries of a matrix are random variables, we define the mean of the matrix as the matrix of the means of the entries,

$$\mathrm{E}\left(\begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & \cdots & A_{nn} \end{bmatrix} \right) := \begin{bmatrix} \mathrm{E}(A_{11}) & \mathrm{E}(A_{12}) & \cdots & \mathrm{E}(A_{1n}) \\ \mathrm{E}(A_{21}) & \mathrm{E}(A_{22}) & \cdots & \mathrm{E}(A_{2n}) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{E}(A_{n1}) & \mathrm{E}(A_{n2}) & \cdots & \mathrm{E}(A_{nn}) \end{bmatrix}. \tag{128}$$

It follows immediately from the linearity of the expectation operator in one dimension that the mean operator is linear.

Theorem 2.17 (Mean of linear transformation of a random vector)**.** *For any random vector $\boldsymbol{X}$ of dimension n, any matrix $A \in \mathbb{R}^{m \times n}$ and $\boldsymbol{b} \in \mathbb{R}^m$*

$$\mathrm{E}(A\boldsymbol{X} + \boldsymbol{b}) = A\,\mathrm{E}(\boldsymbol{X}) + \boldsymbol{b}. \tag{129}$$

Proof.

$$\mathrm{E}(A\mathbf{X} + \mathbf{b}) = \begin{bmatrix} \mathrm{E}\left(\sum_{i=1}^n A_{1i} X_i + b_1\right) \\ \mathrm{E}\left(\sum_{i=1}^n A_{2i} X_i + b_2\right) \\ \cdots \\ \mathrm{E}\left(\sum_{i=1}^n A_{mi} X_i + b_n\right) \end{bmatrix} \tag{130}$$

$$= \begin{bmatrix} \sum_{i=1}^n A_{1i} \mathrm{E}(X_i) + b_1 \\ \sum_{i=1}^n A_{2i} \mathrm{E}(X_i) + b_2 \\ \cdots \\ \sum_{i=1}^n A_{mi} \mathrm{E}(X_i) + b_n \end{bmatrix} \quad \text{by linearity of expectation} \tag{131}$$

$$= A\,\mathrm{E}(\mathbf{X}) + \mathbf{b}. \tag{132}$$

□

The **covariance matrix** of a random vector captures the interaction between the components of the vector. It contains the variance of each variable in the diagonal and the covariances between the different variables in the off diagonals.

Definition 2.18. *The covariance matrix of a random vector $\boldsymbol{X}$ is defined as*

$$\Sigma_{\boldsymbol{X}} := \begin{bmatrix} \operatorname{Var}(X_1) & \operatorname{Cov}(X_1, X_2) & \cdots & \operatorname{Cov}(X_1, X_n) \\ \operatorname{Cov}(X_2, X_1) & \operatorname{Var}(X_2) & \cdots & \operatorname{Cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{Cov}(X_n, X_2) & \operatorname{Cov}(X_n, X_2) & \cdots & \operatorname{Var}(X_n) \end{bmatrix} \tag{133}$$

$$= \operatorname{E}\left(\boldsymbol{X}\boldsymbol{X}^T\right) + \operatorname{E}(\boldsymbol{X})\operatorname{E}(\boldsymbol{X})^T. \tag{134}$$

From Theorem 2.17 we obtain a simple expression for the covariance matrix of the linear transformation of a random vector.

Theorem 2.19 (Covariance matrix after a linear transformation)**.** *Let $\boldsymbol{X}$ be a random vector of dimension n with covariance matrix Σ. For any matrix $A \in \mathbb{R}^{m \times n}$ and $\boldsymbol{b} \in \mathbb{R}^m$,*

$$\Sigma_{A\boldsymbol{X}+\boldsymbol{b}} = A\Sigma_{\boldsymbol{X}}A^T. \tag{135}$$

Proof.

$$\Sigma_{A\mathbf{X}+\mathbf{b}} = \operatorname{E}\left((A\mathbf{X}+\mathbf{b})(A\mathbf{X}+\mathbf{b})^T\right) - \operatorname{E}(A\mathbf{X}+\mathbf{b})\operatorname{E}(A\mathbf{X}+\mathbf{b})^T \tag{136}$$

$$= \operatorname{E}\left(A\mathbf{X}\mathbf{X}^TA^T + \mathbf{b}\mathbf{X}^TA^T + A\mathbf{X}\mathbf{b}^T + \mathbf{b}\mathbf{b}^T\right) - (A\operatorname{E}(\mathbf{X}) + \mathbf{b})(A\operatorname{E}(\mathbf{X}) + \mathbf{b})^T \tag{137}$$

$$= A\operatorname{E}\left(\mathbf{X}\mathbf{X}^T\right)A^T + \mathbf{b}\operatorname{E}(\mathbf{X})^TA^T + A\operatorname{E}(\mathbf{X})\mathbf{b}^T + \mathbf{b}\mathbf{b}^T \tag{138}$$

$$- A\operatorname{E}(\mathbf{X})\operatorname{E}(\mathbf{X})^TA^T - A\operatorname{E}(\mathbf{X})\mathbf{b}^T - \mathbf{b}\operatorname{E}(\mathbf{X})^TA^T - \mathbf{b}\mathbf{b}^T \tag{139}$$

$$= A\left(\operatorname{E}\left(\mathbf{X}\mathbf{X}^T\right) - \operatorname{E}(\mathbf{X})\operatorname{E}(\mathbf{X})^T\right)A^T \tag{140}$$

$$= A\Sigma_{\mathbf{X}}A^T. \tag{141}$$

□

2.4 Gaussian random vectors

Gaussian random vectors are a multidimensional generalization of Gaussian random variables. They are parametrized by their mean and covariance matrix. Figure 2 shows examples of the joint pdf of a two-dimensional Gaussian random vector.

Definition 2.20 (Gaussian random vector)**.** *A Gaussian random vector $\boldsymbol{X}$ is a random vector with joint pdf*

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^n|\Sigma|}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) \tag{142}$$

where the mean vector $\boldsymbol{\mu} \in \mathbb{R}^n$ and the covariance matrix Σ, which is symmetric and positive definite, parametrize the distribution.

A fundamental property of Gaussian random vectors is that performing linear transformations on them always yields vectors with joint distributions that are also Gaussian.

Theorem 2.21 (Linear transformations of Gaussian random vectors are Gaussian)**.** *Let* $\boldsymbol{X}$ *be a Gaussian random vector of dimension* n *with mean* $\boldsymbol{\mu}$ *and covariance matrix* Σ*. For any matrix* $A \in \mathbb{R}^{m\times n}$ *and* $\boldsymbol{b} \in \mathbb{R}^{m}$ $\boldsymbol{Y} = A\boldsymbol{X} + \boldsymbol{b}$ *is a Gaussian random vector with mean* $A\boldsymbol{\mu} + \boldsymbol{b}$ *and covariance matrix* $A\Sigma A^T$*.*

Proof. By Theorems 2.17 and 2.19 $A\mathbf{X} + \mathbf{b}$ has mean $A\boldsymbol{\mu} + \mathbf{b}$ and covariance matrix $A\Sigma A^T$. We still need to prove that the joint distribution of $\mathbf{Y}$ is Gaussian. We will show that this is the case in the univariate case (the argument can be extended to the vector case). Let $Y = aX + b$,

$$F_Y(y) = \mathrm{P}(aX + b \le y) \tag{143}$$

$$= \begin{cases} \mathrm{P}\left(X \le \frac{y-b}{a}\right) = F_X\left(\frac{y-b}{a}\right) & \text{if } a \ge 0, \\ \mathrm{P}\left(X \ge \frac{y-b}{a}\right) = 1 - F_X\left(\frac{y-b}{a}\right) & \text{if } a < 0. \end{cases} \tag{144}$$

To obtain the pdf of Y we differentiate with respect to y. If $a \ge 0$,

$$f_Y(y) = \frac{\mathrm{d}F_Y(y)}{\mathrm{d}y} \tag{145}$$

$$= \frac{\mathrm{d}F_X\left(\frac{y-b}{a}\right)}{\mathrm{d}y} \tag{146}$$

$$= \frac{f_X\left(\frac{y-b}{a}\right)}{a} \tag{147}$$

$$= \frac{1}{\sqrt{2\pi}a\sigma} e^{-\frac{(y-b-a\mu)^2}{2a^2\sigma^2}}. \tag{148}$$

Similarly, if $a < 0$

$$f_Y(y) = \frac{\mathrm{d}\left(1 - F_X\left(\frac{y-b}{a}\right)\right)}{\mathrm{d}y} \tag{149}$$

$$= -\frac{f_X\left(\frac{y-b}{a}\right)}{a} \tag{150}$$

$$= \frac{1}{\sqrt{2\pi}\,|a|\,\sigma} e^{-\frac{(y-b-a\mu)^2}{2a^2\sigma^2}}. \tag{151}$$

The pdf of Y is a Gaussian pdf with mean $a\mu$ and variance $a^2\sigma^2$. The multivariate case can be established in a similar way. □

A corollary of this result is that the joint pdf of a subvector of a Gaussian random vector is also a Gaussian vector.

Corollary 2.22 (Marginals of Gaussian random vectors are Gaussian)**.** *The joint pdf of any subvector of a Gaussian random vector is Gaussian. Without loss of generality, assume that the subvector* $\boldsymbol{X}$ *consists of the first* m *entries of the Gaussian random vector,*

$$\boldsymbol{Z} := \begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \end{bmatrix}, \tag{152}$$

where

$$\Sigma_{\boldsymbol{Z}} = \begin{bmatrix} \Sigma_{\boldsymbol{X}} & \Sigma_{\boldsymbol{XY}} \\ \Sigma_{\boldsymbol{XY}}^T & \Sigma_{\boldsymbol{Y}} \end{bmatrix}, \qquad \Sigma_{\boldsymbol{XY}} := \mathrm{E}\left(XY^T\right) - \mathrm{E}\left(X\right)\mathrm{E}\left(Y\right)^T. \tag{153}$$

Then $\boldsymbol{X}$ *is a Gaussian random vector with mean* $\mathrm{E}\left(\boldsymbol{X}\right)$ *and covariance matrix* $\Sigma_{\boldsymbol{X}}$*.*

Proof. Note that

$$\mathbf{X} = \begin{bmatrix} I_m & 0_{m\times n-m} \\ 0_{n-m\times m} & 0_{n-m\times n-m} \end{bmatrix} \begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix} = \begin{bmatrix} I_m & 0_{m\times n-m} \\ 0_{n-m\times m} & 0_{n-m\times n-m} \end{bmatrix} \mathbf{Z}, \tag{154}$$

where $I \in \mathbb{R}^{m\times m}$ is an identity matrix and $0_{c\times d}$ represents a matrix of zeros of dimensions $c \times d$. The result then follows from Theorem 2.21. □

Interestingly, as a result of the dependence of the Gaussian joint pdf on the covariance matrix, in the case of Gaussian random vectors uncorrelation implies mutual independence.

Lemma 2.23 (Uncorrelation implies mutual independence for Gaussian random vectors)**.** *If all the components of a Gaussian random vector are uncorrelated, then they are also mutually independent.*

Proof. The parameter Σ of the joint pdf of a Gaussian random vector is its covariance matrix (one can verify this by applying the definition of covariance and integrating). If all the components are uncorrelated then

$$\Sigma_{\mathbf{X}} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}, \tag{155}$$

where σ_i is the standard deviation of the ith component. Now, the inverse of this diagonal matrix is just

$$\Sigma_{\mathbf{X}}^{-1} = \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_2^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sigma_n^2} \end{bmatrix}, \tag{156}$$

and its determinant is $|\Sigma| = \prod_{i=1}^{n} \sigma_i^2$ so that from (142)

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})\,\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) \tag{157}$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{(2\pi)}\sigma_i} \exp\left(-\frac{(x_i-\mu_i)^2}{2\sigma_i^2}\right) \tag{158}$$

$$= \prod_{i=1}^{n} f_{X_i}(x_i)\,. \tag{159}$$

□

A Derivation of means and variances in Table 1

A.1 Bernoulli

$$\begin{aligned} \mathrm{E}\left(X\right) &= p_X\left(1\right) = p, && (160)\\ \mathrm{E}\left(X^2\right) &= p_X\left(1\right), && (161)\\ \mathrm{Var}\left(X\right) &= \mathrm{E}\left(X^2\right) - \mathrm{E}^2\left(X\right) = p\left(1-p\right). && (162) \end{aligned}$$

A.2 Geometric

To compute the mean of a geometric random variable, we apply Lemma B.3:

$$\begin{aligned} \mathrm{E}\left(X\right) &= \sum_{k=1}^{\infty} k\, p_X\left(k\right) && (163)\\ &= \sum_{k=1}^{\infty} k\, p\left(1-p\right)^{k-1} && (164)\\ &= \frac{p}{1-p}\sum_{k=1}^{\infty} k\left(1-p\right)^{k} && (165)\\ &= \frac{1}{p}. && (166) \end{aligned}$$

To compute the mean squared value we apply Lemma B.4:

$$\begin{aligned} \mathrm{E}\left(X^2\right) &= \sum_{k=1}^{\infty} k^2\, p_X\left(k\right) && (167)\\ &= \sum_{k=1}^{\infty} k^2\, p\left(1-p\right)^{k-1} && (168)\\ &= \frac{p}{1-p}\sum_{k=1}^{\infty} k^2\left(1-p\right)^{k} && (169)\\ &= \frac{2-p}{p^2}. && (170)\\ \mathrm{Var}\left(X\right) &= \mathrm{E}\left(X^2\right) - \mathrm{E}^2\left(X\right) = \frac{1-p}{p^2}. && (171) \end{aligned}$$

A.3 Binomial

By Lemma 5.8 in Lecture Notes 1, if we define n Bernouilli random variables with parameter p we can write a binomial random variable with parameters n and p as

$$X = \sum_{i=1}^{n} B_i, \tag{172}$$

where $B_1, B_2, \ldots$ are mutually independent Bernouilli random variables with parameter p. Since the mean of all the Bernouilli random variables is p, by linearity of expectation

$$\mathrm{E}(X) = \sum_{i=1}^{n} \mathrm{E}(B_i) = np. \tag{173}$$

Note that $\mathrm{E}(B_i^2) = p$ and $\mathrm{E}(B_i B_j) = p^2$ by independence, so

$$\begin{aligned} \mathrm{E}\left(X^2\right) &= \mathrm{E}\left(\sum_{i=1}^{n}\sum_{j=1}^{n} B_i B_j\right) && (174) \\ &= \sum_{i=1}^{n} \mathrm{E}\left(B_i^2\right) + 2\sum_{i=1}^{n-1}\sum_{i=j+1}^{n} \mathrm{E}(B_i B_j) = np + n(n-1)p^2. && (175) \\ \mathrm{Var}(X) &= \mathrm{E}\left(X^2\right) - \mathrm{E}^2(X) = np(1-p). && (176) \end{aligned}$$

A.4 Poisson

From calculus we have

$$\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{\lambda}, \tag{177}$$

which is the Taylor series expansion of the exponential function. Now we can establish that

$$\begin{aligned} \mathrm{E}(X) &= \sum_{k=1}^{\infty} k p_X(k) && (178) \\ &= \sum_{k=1}^{\infty} \frac{\lambda^k e^{-\lambda}}{(k-1)!} && (179) \\ &= e^{-\lambda} \sum_{m=0}^{\infty} \frac{\lambda^{m+1}}{m!} && (180) \\ &= \lambda, && (181) \end{aligned}$$

and

$$\begin{aligned}
\mathrm{E}\left(X^2\right) &= \sum_{k=1}^{\infty} k^2 p_X(k) && (182)\\
&= \sum_{k=1}^{\infty} \frac{k\lambda^k e^{-\lambda}}{(k-1)!} && (183)\\
&= e^{-\lambda}\left(\sum_{k=1}^{\infty} \frac{(k-1)\,\lambda^k}{(k-1)!} + \frac{k\lambda^k}{(k-1)!}\right) && (184)\\
&= e^{-\lambda}\left(\sum_{m=1}^{\infty} \frac{\lambda^{m+2}}{m!} + \sum_{m=1}^{\infty} \frac{\lambda^{m+1}}{m!}\right) && (185)\\
&= \lambda^2 + \lambda. && (186)\\
\mathrm{Var}(X) &= \mathrm{E}\left(X^2\right) - \mathrm{E}^2(X) = \lambda. && (187)
\end{aligned}$$

A.5 Uniform

We apply the definition of expected value for continuous random variables to obtain

$$\begin{aligned}
\mathrm{E}(X) &= \int_{-\infty}^{\infty} x f_X(x)\,\mathrm{d}x = \int_a^b \frac{x}{b-a}\mathrm{d}x && (188)\\
&= \frac{b^2-a^2}{2(b-a)} = \frac{a+b}{2}. && (189)
\end{aligned}$$

Similarly,

$$\begin{aligned}
\mathrm{E}\left(X^2\right) &= \int_a^b \frac{x^2}{b-a}\mathrm{d}x && (190)\\
&= \frac{b^3-a^3}{3(b-a)} && (191)\\
&= \frac{a^2+ab+b^2}{3}. && (192)\\
\mathrm{Var}(X) &= \mathrm{E}\left(X^2\right) - \mathrm{E}^2(X) && (193)\\
&= \frac{a^2+ab+b^2}{3} - \frac{a^2+2ab+b^2}{4} = \frac{(b-a)^2}{12}. && (194)
\end{aligned}$$

A.6 Exponential

Applying integration by parts,

$$\mathrm{E}\left(X\right) = \int_{-\infty}^{\infty} x f_X\left(x\right) \mathrm{d}x \tag{195}$$

$$= \int_{0}^{\infty} x\lambda e^{-\lambda x} \mathrm{d}x \tag{196}$$

$$= xe^{-\lambda x}]_0^{\infty} + \int_0^{\infty} e^{-\lambda x} \mathrm{d}x \tag{197}$$

$$= \frac{1}{\lambda}. \tag{198}$$

Similarly,

$$\mathrm{E}\left(X^2\right) = \int_0^{\infty} x^2 \lambda e^{-\lambda x} \mathrm{d}x \tag{199}$$

$$= x^2 e^{-\lambda x}]_0^{\infty} + 2\int_0^{\infty} x e^{-\lambda x} \mathrm{d}x \tag{200}$$

$$= \frac{2}{\lambda^2}. \tag{201}$$

$$\mathrm{Var}\left(X\right) = \mathrm{E}\left(X^2\right) - \mathrm{E}^2\left(X\right) = \frac{1}{\lambda^2}. \tag{202}$$

A.7 Gaussian

We apply the change of variables $t = \left(x - \mu\right)/\sigma$.

$$\mathrm{E}\left(X\right) = \int_{-\infty}^{\infty} x f_X\left(x\right) \mathrm{d}x \tag{203}$$

$$= \int_{-\infty}^{\infty} \frac{x}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \mathrm{d}x \tag{204}$$

$$= \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t e^{-\frac{t^2}{2}} \mathrm{d}t + \frac{\mu}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{t^2}{2}} \mathrm{d}t \tag{205}$$

$$= \mu, \tag{206}$$

where the last step follows from the fact that the integral of a bounded odd function over a symmetric interval is zero and from Lemma 5.21 in Lecture Notes 1.

Applying the change of variables $t = (x-\mu)/\sigma$ and integrating by parts, we obtain that

$$\begin{aligned}
\mathrm{E}\left(X^2\right) &= \int_{-\infty}^{\infty} x^2 f_X(x)\,\mathrm{d}x && (207)\\
&= \int_{-\infty}^{\infty} \frac{x^2}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,\mathrm{d}x && (208)\\
&= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t^2 e^{-\frac{t^2}{2}}\,\mathrm{d}t + \frac{2\mu\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t e^{-\frac{t^2}{2}}\,\mathrm{d}t + \frac{\mu^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{t^2}{2}}\,\mathrm{d}t && (209)\\
&= \frac{\sigma^2}{\sqrt{2\pi}} \left(t^2 e^{-\frac{t^2}{2}}]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} e^{-\frac{t^2}{2}}\,\mathrm{d}t \right) + \mu^2 && (210)\\
&= \sigma^2 + \mu^2. && (211)\\
\mathrm{Var}(X) &= \mathrm{E}\left(X^2\right) - \mathrm{E}^2(X) = \sigma^2. && (212)
\end{aligned}$$

B Geometric series

Lemma B.1. *For any $\alpha \neq 0$ and any integers n_1 and n_2*

$$\sum_{k=n_1}^{n_2} \alpha^k = \frac{\alpha^{n_1} - \alpha^{n_2+1}}{1-\alpha}. \tag{213}$$

Corollary B.2. *If $0 < \alpha < 1$*

$$\sum_{k=0}^{\infty} \alpha^k = \frac{\alpha}{1-\alpha}. \tag{214}$$

Proof. We just multiply the sum by the factor $(1-\alpha)/(1-\alpha)$ which obviously equals one,

$$\begin{aligned}
\alpha^{n_1} + \alpha^{n_1+1} + \cdots + \alpha^{n_2-1} + \alpha^{n_2} &= \frac{1-\alpha}{1-\alpha}\left(\alpha^{n_1} + \alpha^{n_1+1} + \cdots + \alpha^{n_2-1} + \alpha^{n_2}\right) && (215)\\
&= \frac{\alpha^{n_1} - \alpha^{n_1+1} + \alpha^{n_1+1} + \cdots - \alpha^{n_2} + \alpha^{n_2} - \alpha^{n_2+1}}{1-\alpha} && (216)\\
&= \frac{\alpha^{n_1} - \alpha^{n_2+1}}{1-\alpha} && (217)
\end{aligned}$$

□

Lemma B.3. *For $0 < \alpha < 1$*

$$\sum_{k=1}^{\infty} k\,\alpha^k = \frac{1}{(1-\alpha)^2}. \tag{218}$$

Proof. By Corollary B.2,

$$\sum_{k=0}^{\infty} \alpha^k = \frac{1}{1-\alpha}. \tag{219}$$

Since the left limit converges, we can differentiate on both sides to obtain

$$\sum_{k=0}^{\infty} k\alpha^{k-1} = \frac{1}{1-\alpha}. \tag{220}$$

□

Lemma B.4. *For* $0 < \alpha < 1$

$$\sum_{k=1}^{\infty} k^2 \, \alpha^k = \frac{\alpha\,(1+\alpha)}{(1-\alpha)^3}. \tag{221}$$

Proof. By Lemma B.3,

$$\sum_{k=1}^{\infty} k^2\alpha^k = \frac{\alpha\,(1+\alpha)}{(1-\alpha)^3}. \tag{222}$$

Since the left limit converges, we can differentiate on both sides to obtain

$$\sum_{k=1}^{\infty} k^2\alpha^{k-1} = \frac{1+\alpha}{(1-\alpha)^3}. \tag{223}$$

□