PROBABILITY THEORY

- Probability, loosely speaking, concerns the study of uncertainty. Probability can be thought of as the fraction of times an event occurs, or as a degree of belief about an event. Quantifying uncertainty requires the idea of a random variable, which is a function that maps outcomes of random experiments to a set of properties that we are interested in. Associated with the random variable is a function that measures the probability that a particular outcome (or a set of outcomes) will occur; this is called the probability distribution

(1.) PROBABILITY AND RANDOM VARIABLES

↳ There are three distinct ideas that are often confused when discussing probabilities. First is the idea of a probability space, which allows us to quantify the idea of a probability. However, we mostly do not work directly with this basic probability space. Instead, we work with random variables (the second idea), which transfers the probability to a more convenient (often numerical) space. The third idea is the idea of a distribution or law associated with a random variable.

Modern probability is based on a set of axioms proposed by Kolmogorov (Grimstead and Snell, 1997 ; Jaynes, 2003) that introduce the three concepts of sample space, event space, and probability measure. The probability space models a real-world process (referred to as an experiment) with random outcomes.

▶ SAMPLE SPACE : The sample space is the set of all possible outcomes of the experiment, usually denoted by Ω. For example, two successive coin tosses have a sample space of $\{hh, ht, th, tt\}$, where "h" denotes "heads" and "t" denotes "tails".

▶ EVENT SPACE : The event space is the space of potential results of the experiment. A subset A of the sample space Ω is in the event space $\mathcal{A}$ if at the end of the experiment we can observe whether a particular outcome $\omega \in \Omega$ is in A. The event space $\mathcal{A}$ is obtained by considering the collection of subsets of Ω, and for discrete probability distributions $\mathcal{A}$ is often the power set of Ω.

▶ PROBABILITY : With each event $A \in \mathcal{A}$, we associate a number $\mathbb{P}(A)$ or $P(A)$ that measures the probability or degree of belief that event will occur. Also, $\mathbb{P}(A)$ is called the probability of A.

↳ The probability of a single event must be in the interval $[0, 1]$, and the total probability over all outcomes in the sample space Ω must be 1, i.e., $\mathbb{P}(\Omega) = 1$. Given a probability space $(\Omega, \mathcal{A}, P)$, we want to use it to model some real-world phenomenon.

↳ In machine learning, we often avoid explicitly referring to the probability space, but instead refer to probabilities on quantities of interest, which we denote by $\mathcal{T}$. In this notes, we refer to $\mathcal{T}$ as the target space and refer to elements of $\mathcal{T}$ as states. We introduce a function $X : \Omega \to \mathcal{T}$ that takes an element of Ω (an outcome) and returns a particular quantity of interest x, a value in $\mathcal{T}$. This association/mapping from Ω to $\mathcal{T}$ is called a random variable.

For example, in the case of tossing two coins and counting the number of heads, a random variable X maps to the three possible outcomes: $X(hh) = 2$, $X(ht) = 1$, $X(th) = 1$, and $X(tt) = 0$. In this particular case, $\mathcal{T} = \{0, 1, 2\}$, and it is the probabilities on elements of $\mathcal{T}$ that we are interested in. For a finite sample space Ω and finite $\mathcal{T}$, the function corresponding to a random variable is essentially a lookup table. For any subset $S \subseteq \mathcal{T}$, we associate $\mathbb{P}_X(S) \in [0, 1]$ (the probability) to a particular event occuring corresponding to the random variable X.

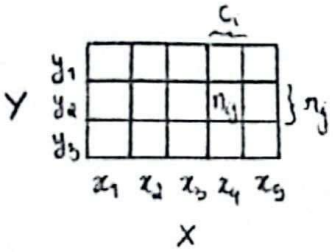
*) REMARK[1] : The aforementioned sample space Ω unfortunately is referred to by different names in different books. Another common name for Ω is "state space" (Jacob and Protter, 2004), but state space is sometimes reserved for referring to states in a dynamical systems (Haselblatt and Katok, 2003). Other names sometimes used to describe Ω are: "sample description space", "possibility space", and "event space". ◇

(2.) DISCRETE PROBABILITIES

↳ When the target space is discrete, we can imagine the probability distribution of multiple random variables as filling out a (multidimensional) array of numbers. The target space of the joint probability is the Cartesian product of the target spaces of each of the random variables. We define the joint probability as the entry of both values jointly:

$$\mathbb{P}(X = x_i, Y = y_j) = \frac{n_{ij}}{N} \,, \tag{3.1}$$

where n_{ij} is the number of events with state x_i and y_j and N is the total number of events. The joint probability of the intersection of both events, that is, the condition $\mathbb{P}(X = x_i, Y = y_j) = \mathbb{P}(X = x_i \cap Y = y_j)$.



④ This Figure illustrates the probability mass function (pmf) of a discrete probability distribution. For two random variables X and Y, the probability that $X = x$ and $Y = y$ is (lazily) written as $p(x, y)$ and is called the joint probability.

↳ One can think of a probability as a function that takes state x and y and returns a real number, which is the reason we write $p(x, y)$. The marginal probability that X takes the value x irrespective of the value of random variable Y is (lazily) written as $p(x)$. We write $X \sim p(x)$ to denote that the random variable X is distributed according to $p(x)$. If we consider only the instances where $X = x$, then the fraction of instances (the conditional probability) for which $Y = y$ is written (lazily) as $p(x|y)$.

(3.) Continuous Distributions

↳ We consider real-valued random variables in this notes, i.e., we consider target spaces that are intervals of the real line $\mathbb{R}$. In this notes, we pretend that we can perform operations on real random variables as if we have discrete probability spaces with finite spaces. However, this simplification is not precise for two situations: when we repeat something infinitely often, and when we want to draw a point from an interval. This first situation arises when we discuss generalization error in machine learning. The second situation arises when we want to discuss continuous distributions, such as the Gaussian. For our purposes, the lack of precision allows for a briefer introduction to probability.

*) Remark[2]: In continuous spaces, there are two additional technicalities, which are counter-intuitive. First, the set of all subsets (used to define the event space $\mathcal{A}$) is not well behaved enough. Also, $\mathcal{A}$ needs to be restricted to behave well under set complements, set intersections, and set unions. Second, the size of a set (which in discrete spaces can be obtained by counting the elements) turns out to tricky. The size of a set is called its measure. For example, the cardinality of discrete sets, the length of an interval in $\mathbb{R}$, and the volume of a region in $\mathbb{R}^d$ are all measures. Sets that behave well under set operations and additionally have a topology are called a Borel σ-algebra.

▶ Definition[1] (Probability Density Function): A function $f: \mathbb{R}^D \to \mathbb{R}$ is called a probability density function (pdf) if:

1. $\forall \underline{x} \in \mathbb{R}^D : f(\underline{x}) \geq 0$;
2. Its integral exists and:

$$\underbrace{\int_{\mathbb{R}} \cdots \int_{\mathbb{R}}}_{D \text{ times}} f(\underline{x})\, d\underline{x} \equiv \int_{\mathbb{R}^D} f(\underline{x})\, d\underline{x} = 1 . \tag{3.2}$$

For probability mass functions (pmf) of discrete random variables, the integral in Eq. (3.2) is replaced with a sum given by:

$$\sum_{i=1}^{N} \mathbb{P}(X = x_i) = 1 \quad \text{and} \quad \sum_{j=1}^{N} \mathbb{P}(Y = y_j) = 1 . \tag{3.3}$$

↳ Observe that the probability density function is any function f that is non-negative and integrates to one. We associate a random variable X with this function f by:

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f(x)\, dx , \tag{3.4}$$

where $a, b \in \mathbb{R}$ and $x \in \mathbb{R}$ are outcomes of the continuous random variable X. The states $\underline{x} \in \mathbb{R}^D$ are defined analogously by considering a vector of $x \in \mathbb{R}$. This association (3.4) is called the law or distribution of the random variable X.

*) REMARK[3] : In contrast to discrete random variables, the probability of a continuous random variable X taking a particular value $\mathbb{P}(X = x)$ is zero. This is like trying to specify an interval in Eq. (3.4) where $a = b$.

◊

▶ DEFINITION[2] (Cumulative Distribution Function) : A cumulative distribution function (cdf) of a multivariate real valued random variable X with states $\underline{x} \in \mathbb{R}^D$ is given by:

$$F_X(\underline{x}) = \mathbb{P}(X_1 \leq x_1, \dots, X_D \leq x_D), \tag{3.5}$$

where $X = (X_1, \dots, X_D)^T$, $\underline{x} = (x_1, \dots, x_D)^T$, and the right-hand side represents the probability that random variable X_i takes the value smaller than or equal to x_i.

↳ The cdf can be expressed also as the integral of the probability density function $f(\underline{x})$ so that:

$$F_X(\underline{x}) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_D} f(z_1, \dots, z_D)\, dz_1 \cdots dz_D. \tag{3.6}$$

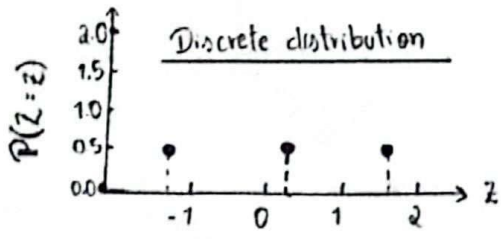
Also, there are cdf's which do not have corresponding pdf's.

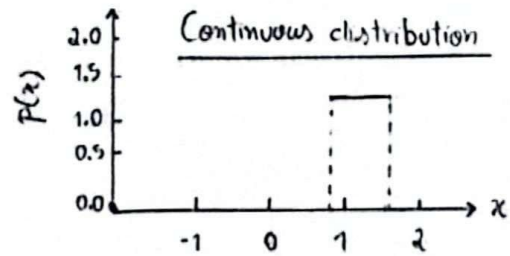
*) REMARK[4] : We reiterate that there are in fact two distinct concepts when talking about distributions. First is the idea of an pdf [denoted by $f(x)$], which is a non-negative function that sums to one. Second is the law of a random variable X, that is, the association of a random variable X with the pdf $f(x)$.

◊

(4.) CONTRASTING DISCRETE AND CONTINUOUS DISTRIBUTIONS

↳ Recall that probabilities are positive and the total probability sums up to one. For discrete random variables [Eq. (3.3)], this implies that the probability of each state must lie in the interval $[0, 1]$. However, for continuous random variables the normalization [see Eq. (3.2)] does not imply that the value of density is less than or equal to 1 for all values.





TYPE	"POINT PROBABILITY"	"INTERVAL PROBABILITY"
Discrete	$\mathbb{P}(X = x)$ Probability mass function	Not applicable
Continuous	$p(x)$ Probability density function	$\mathbb{P}(X \leq x)$ Cumulative distribution function

(5.) Sum Rule, Product Rule, and Bayes Theorem

↳ Recall from Eq. (3.1) that $p(\underline{x}, \underline{y})$ is the joint distribution of the two random variables $\underline{x}$, $\underline{y}$. The distributions $p(\underline{x})$ and $p(\underline{y})$ are the corresponding marginal distributions, and $p(\underline{y}|\underline{x})$ is the conditional distribution of $\underline{y}$ given $\underline{x}$. Given the definitions of the marginal and conditional probability for discrete and continuous random variables, we can now present the two fundamental rules in probability theory.

↝ The first rule, the sum rule, states that:

$$p(\underline{x}) = \begin{cases} \sum_{\underline{y} \in \mathcal{Y}} p(\underline{x}, \underline{y}) & \text{, if } \underline{y} \text{ is discrete,} \\ \int_{\mathcal{Y}} p(\underline{x}, \underline{y}) \, d\underline{y} & \text{, if } \underline{y} \text{ is continuous,} \end{cases} \tag{3.7}$$

where $\mathcal{Y}$ are the states of the target space of a random variable Y. This means that we sum out (or integrate out) the sets of states $\underline{y}$ of the random variable Y. The sum rule is also known as the marginalization property. The sum rule relates the joint distribution to a marginal distribution. In general, when the joint distribution contains more than two random variables, the sum rule can be applied to any subset of the random variables, resulting in a marginal distribution of potentially more than one random variable. More concretely, if $\underline{x} = (x_1, \ldots, x_D)^T$, we obtain the marginal:

$$p(x_i) = \int p(x_1, \ldots, x_D) \, d\underline{x}_{\setminus i}, \tag{3.8}$$

by repeated application of the sum rule where we integrate / sum out all random variables except x_i, which is indicated by $\setminus i$, which reads "all except i".

*) Remark[5]: Many of the computational challenges of probabilistic modelling are due to the application of the sum rule. When there are many variables or discrete variables with many states, the sum rule boils down to performing a high-dimensional sum or integral. Performing high-dimensional sums or integrals is generally computationally hard, in the sense that there is no known polynomial-time algorithm to calculate them exactly. ◇

↝ The second rule, known as the product rule, relates the joint distribution to the conditional distribution via:

$$p(\underline{x}, \underline{y}) = p(\underline{y}|\underline{x}) \, p(\underline{y}). \tag{3.9}$$

The product rule can be interpreted as the fact that every joint distribution of two random variables can be factorized (written as a product) of other two distributions. The two factors are the marginal distribution of the first random variable $p(\underline{x})$, and the conditional distribution of the second random variable given the first $p(\underline{y}|\underline{x})$. Since the ordering of random variables is arbitrary in $p(\underline{x}, \underline{y})$, the product rule also implies $p(\underline{x}, \underline{y}) = p(\underline{x}|\underline{y}) \, p(\underline{y})$. To be precise, Eq. (3.9) is expressed in terms of the probability mass functions for discrete random variables. For continuous random variables, the product rule is expressed in terms of the probability density functions.

↪ In machine learning and Bayesian statistics, we are often interested in making inferences of unobserved (latent) random variables given that we have observed other random variables. Let us assume we have some prior knowledge $p(\underline{x})$ about an unobserved random variable $\underline{x}$ and some relationship $p(\underline{y}|\underline{x})$ between $\underline{x}$ and a second random variable $\underline{y}$, which we can observe. If we observe $\underline{y}$, we can use Bayes' theorem to draw some conclusions about $\underline{x}$ given the observed values of $\underline{y}$. Bayes' theorem (also Bayes' rule or Bayes' law):

$$\underbrace{p(\underline{x}|\underline{y})}_{\text{posterior}} = \frac{\overbrace{p(\underline{y}|\underline{x})}^{\text{likelihood}}\overbrace{p(\underline{x})}^{\text{prior}}}{\underbrace{p(\underline{y})}_{\text{evidence}}}, \tag{3.10}$$

is a direct consequence of the product rule in Eq. (3.9) since:

$$p(\underline{x},\underline{y}) = p(\underline{x}|\underline{y})p(\underline{y}) = p(\underline{y}|\underline{x})p(\underline{y}) \iff p(\underline{x}|\underline{y}) = \frac{p(\underline{y}|\underline{x})p(\underline{x})}{p(\underline{y})}. \tag{3.11}$$

- In Eq. (3.10), $p(\underline{x})$ is the prior, which encapsulates our subjective prior knowledge of the unobserved (latent) variable $\underline{x}$ before observing any data. We can choose any prior that makes sense to us, but it is critical to ensure that the prior has a non-zero pdf (or pmf) on all plausible $\underline{x}$, even if they are very rare.

- The likelihood $p(\underline{y}|\underline{x})$ describe how $\underline{x}$ and $\underline{y}$ are related, and in the case of discrete probability distributions, it is the probability of the data $\underline{y}$ if we were to know the latent variable $\underline{x}$. Note that the likelihood is not a distribution in $\underline{x}$, but only in $\underline{y}$. We call $p(\underline{y}|\underline{x})$ either the "likelihood of $\underline{x}$ (given $\underline{y}$)" or the "probability of $\underline{y}$ given $\underline{x}$" but never the likelihood of $\underline{y}$.

- The posterior $p(\underline{x}|\underline{y})$ is the quantity of interest in Bayesian statistics because it expresses what we are interested in, i.e., what we know about $\underline{x}$ after having observed $\underline{y}$. The quantity:

$$p(\underline{y}) := \int p(\underline{y}|\underline{x})p(\underline{x})\,d\underline{x} = \mathbb{E}_{\underline{x}}[p(\underline{y}|\underline{x})], \tag{3.12}$$

is the marginal likelihood / evidence. The right-hand side of Eq. (3.12) uses the expectation operator. By definition, the marginal likelihood integrates the numerator of Eq. (3.10) with respect to the latent variable $\underline{x}$. Therefore, the marginal likelihood is independent of $\underline{x}$, and it ensures that the posterior $p(\underline{x}|\underline{y})$ is normalized. The marginal likelihood can also be interpreted as the expected likelihood where we take the expectation with respect to the prior $p(\underline{x})$. Beyond normalization of the posterior, the marginal likelihood also plays an important role in Bayesian model selection. In general, the evidence is often hard to compute.

↪ The Bayes' theorem (3.10) allows us to invert the relationship between $\underline{x}$ and $\underline{y}$ given by the likelihood. Therefore, Bayes' theorem is sometimes called the probabilistic inverse.

*) REMARK[6]: In Bayesian statistics, the posterior distribution is the quantity of interest as it encapsulates all available information from the prior and the data. Instead of carrying the posterior around, it is possible to focus on some statistics of the posterior, such as the maximum of the posterior. However, focusing on some statistic of the posterior leads to loss of information.

6. Means and Covariances

↳ Mean and (co)variance are often useful to describe properties of probability distributions (expected value and spread). There is a useful family of distributions (called the exponential family), where the statistics of the random variable capture all possible information. The concept of the expected value is central to machine learning, and the foundational concepts of probability itself can be derived from the expected value.

▶ **Definition 3** (Expected Value) : The expected value of a function $g: \mathbb{R} \to \mathbb{R}$ of a univariate continuous random variable $X \sim p(x)$ is given by :

$$\mathbb{E}_X[g(x)] = \int_{\mathcal{X}} g(x)\, p(x)\, dx \,. \tag{5.13}$$

Correspondingly, the expected value of a function g of a discrete random variable $X \sim p(x)$ is given by :

$$\mathbb{E}_X[g(x)] = \sum_{x \in \mathcal{X}} g(x)\, p(x), \tag{5.14}$$

where $\mathcal{X}$ is the set of possible outcomes (the target space) of the random variable X.

*) **Remark 7** : We consider multivariate random variables X as a finite vector of univariate random variables $(X_1, \dots, X_D)^T$. For multivariate random variables, we define the expected value element wise :

$$\mathbb{E}_X[g(x)] = \begin{pmatrix} \mathbb{E}_{X_1}[g(x_1)] \\ \vdots \\ \mathbb{E}_{X_D}[g(x_D)] \end{pmatrix} \in \mathbb{R}^D \,, \tag{5.15}$$

where the subscript $\mathbb{E}_{X_d}$ indicates that we are taking the expected value with respect to the d^{th} element of the vector $\underline{x}$.

↳ The **Definition 3** defines the meaning of the notation $\mathbb{E}_X$ as the operator indicating that we should take the integral with respect to the probability density (for continuous distributions) or the sum over all states (for discrete distributions). The definition of the mean is a special case of the expected value, obtained by choosing g to be the identity function.

▶ **Definition 4** (Mean) : The mean of a random variable X with state $\underline{x} \in \mathbb{R}^D$ is an average and is defined as :

$$\mathbb{E}_X[\underline{x}] = \begin{pmatrix} \mathbb{E}_{X_1}[x_1] \\ \vdots \\ \mathbb{E}_{X_D}[x_D] \end{pmatrix} \in \mathbb{R}^D, \quad \text{where :} \quad \mathbb{E}_{X_d}[x_d] := \begin{cases} \displaystyle\int_{\mathcal{X}} x_d\, p(x_d)\, dx_d \,, & \text{if } X \text{ is a continuous random variable} \,, \\ \displaystyle\sum_{x_i \in \mathcal{X}} x_i\, p(x_d = x_i) \,, & \text{if } X \text{ is a discrete random variable} \,, \end{cases}$$

for $d = 1, \dots, D$, where the subscript d indicates the corresponding dimension of $\underline{x}$. The integral and sum are over the state $\mathcal{X}$ of the target space of the random variable X.

↳ In one dimension, there are two other intuitive notions of "average", which are the median and the mode. The median is the "middle" value if we sort the values, i.e., 50% of the values are greater than the median and 50% are smaller than the median. This idea can be generalized to continuous values by considering the value where the cdf is 0.5. For distributions, which are asymmetric or have long tails, the median provides an estimate of a typical value that is closer to human intuition than the mean value. Furthermore, the median is more robust to outliers than the mean. The generalization of the median to higher dimensions is non-trivial as there is no obvious way to "sort" in more than one dimension.

↳ The mode is the most frequently occurring value. For a discrete random variable, the mode is defined as the value of x having the highest frequency of occurence. For a continuous random variable, the mode is defined as a peak in the density $p(\underline{x})$. A particular density $p(\underline{x})$ may have more than one mode, and furthermore there may be a very large number of modes in high-dimensional distributions. Therefore, finding all the modes of a distribution can be computationally challenging.

*) REMARK[8] : The expected value is a linear operator. For example, given a real-valued function $f(\underline{x}) = ag(\underline{x}) + bh(\underline{x})$ where $a, b \in \mathbb{R}$ and $\underline{x} \in \mathbb{R}^D$, we obtain:

$$\mathbb{E}_{\underline{x}}[f(\underline{x})] = \int f(\underline{x})\, p(\underline{x})\, d\underline{x} = \int [ag(\underline{x}) + bh(\underline{x})]\, p(\underline{x})\, d\underline{x} = a\int g(\underline{x})\, p(\underline{x})\, d\underline{x} + b\int h(\underline{x})\, p(\underline{x})\, d\underline{x}\,, \text{ so :}$$

$$\therefore \quad \mathbb{E}_{\underline{x}}[f(\underline{x})] = a\,\mathbb{E}_{\underline{x}}[g(\underline{x})] + b\,\mathbb{E}_{\underline{x}}[h(\underline{x})]\,. \tag{3.16}$$

◇

• For two random variables, we may wish to characterize their correspondence to each other. The covariance intuitively represents the notion of how dependent random variables are to one another.

▶ DEFINITION[5] [Covariance (Univariate)] : The covariance between two univariate random variables $x, y \in \mathbb{R}$ is given by the expected product of their deviations from their respective means, i.e.,

$$\mathrm{Cov}_{x,y}[x,y] := \mathbb{E}_{x,y}\big[(x - \mathbb{E}_x[x])(y - \mathbb{E}_y[y])\big]\,. \tag{3.17}$$

*) REMARK[9] : When the random variable associated with the expectation or covariance is clear by its arguments, the subscript is often expressed (for example, $\mathbb{E}_x[x]$ is often written as $\mathbb{E}[x]$).

◇

↳ By using the linearity of expectations, the expression in DEFINITION[5] can be rewritten as the expected value of the product minus the product of the expected values, i.e.,

$$\mathrm{Cov}[x,y] = \mathbb{E}[xy] - \mathbb{E}[x]\,\mathbb{E}[y]\,. \tag{3.18}$$

The covariance of a variable with itself $\mathrm{Cov}[x,x]$ is called the variance and is denoted by $\mathbb{V}_x[x]$. The square root of the variance is called the standard deviation and is often denoted by $\sigma(x)$. The notion of covariance can be generalized to multivariate random variables.

▶ DEFINITION[6] [Covariance (Multivariate)] : If we consider two multivariate random variables X and Y with states $\underline{x} \in \mathbb{R}^D$ and $\underline{y} \in \mathbb{R}^E$ respectively, the covariance between X and Y is defined as :

$$\mathrm{Cov}[\underline{x}, \underline{y}] = \mathbb{E}[\underline{x}\underline{y}^T] - \mathbb{E}[\underline{x}]\mathbb{E}[\underline{y}]^T = \mathrm{Cov}[\underline{y}, \underline{x}^T] \in \mathbb{R}^{D \times E} . \tag{3.19}$$

↳ The DEFINITION[6] can be applied with the same multivariate random variable in both arguments, which results in a useful concept that intuitively captures the "spread" of a random variable. For a multivariate random variable, the variance describes the relation between individual dimensions of the random variable.

▶ DEFINITION[7] (Variance) : The variance of a random variable X with states $\underline{x} \in \mathbb{R}^D$ and a mean vector $\underline{\mu} \in \mathbb{R}^D$ is defined as :

• $$\mathbb{V}_X[\underline{x}] = \mathrm{Cov}_X[\underline{x}, \underline{x}] = \mathbb{E}_X[(\underline{x} - \underline{\mu})(\underline{x} - \underline{\mu})^T] = \mathbb{E}_X[\underline{x}\underline{x}^T] - \mathbb{E}_X[\underline{x}]\mathbb{E}_X[\underline{x}]^T , \text{ so that :}$$

$$\therefore \quad \mathbb{V}_X[\underline{x}] = \begin{pmatrix} \mathrm{Cov}[x_1, x_1] & \mathrm{Cov}[x_1, x_2] & \cdots & \mathrm{Cov}[x_1, x_D] \\ \mathrm{Cov}[x_2, x_1] & \mathrm{Cov}[x_2, x_2] & \cdots & \mathrm{Cov}[x_2, x_D] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}[x_D, x_1] & \mathrm{Cov}[x_D, x_2] & \cdots & \mathrm{Cov}[x_D, x_D] \end{pmatrix}_{D \times D} . \tag{3.20}$$

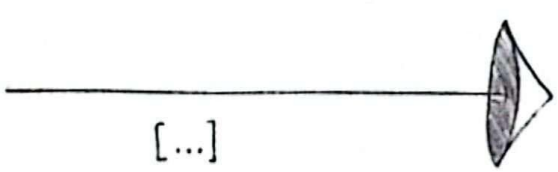
The $D \times D$ matrix in Eq. (3.20) is called the covariance matrix of the multivariate random variable X. The covariance matrix is symmetric and positive semidefinite and tell us something about the spread of the data. On its diagonal the covariance matrix contains the variance of the marginals :

$$p(x_i) = \int p(x_1, \ldots, x_D)\, dx_{\setminus i} , \tag{3.21}$$

where "$\setminus i$" denotes "all variables but i". The off-diagonal entries are the cross-covariance terms $\mathrm{Cov}[x_i, x_j]$ for $i, j = 1, \ldots, D$, with $i \neq j$.

*) REMARK[10] : In this notes, we generally assume that covariance matrices are positive definite to enable better intuition. We therefore do not discuss corner cases that result in positive semidefinite (low-rank) covariance matrices. ◇

↝ When we want to compare the covariances between different pairs of random variables, it turns out that the variance of each random variable affects the value of the covariance. The normalized version of covariance is called the correlation.



[...]

▶ **Definition 8** (Correlation): The correlation between two random variables X, Y is given by:

$$\text{corr}[x,y] = \frac{\text{Cov}[x,y]}{\sqrt{\mathbb{V}[x]\mathbb{V}[y]}} \in [-1,1] \,. \tag{3.22}$$

↳ The correlation matrix is the covariance matrix of standardized random variables, $x/\sigma(x)$. In other words, each random variable is divided by its standard deviation (the square root of variance) in the correlation matrix. The covariance (and correlation) indicate how two random variables are related. Positive correlation $\text{corr}[x,y]$ means that when x grows, then y is also expected to grow. Negative correlation means that as x increases, then the random variable y decreases.

7. Empirical Means and Covariances

↳ The definitions in Section 6. are often also called the population mean and covariance, as it refers to the true statistics for the population. In machine learning, we need to learn from empirical observations of data. Consider a random variable X. There are two conceptual steps to go from population statistics to the realization of empirical statistics. First, we use the fact that we have a finite dataset (of size N) to construct an empirical statistics that is a function of a finite number of identical random variables, $X_1, \dots, X_N$. Second, we observe the data, that is, we look at the realization $x_1, \dots, x_N$ of each of the random variables and apply the empirical statistic.

Specifically, for the mean, given a particular dataset we can obtain an estimate of the mean, which is called the empirical mean or sample mean. The same holds for the empirical covariance.

▶ **Definition 9** (Empirical Mean and Covariance): The empirical mean vector is the arithmetic average of the observations for each variable, and it is defined as:

$$\bar{\underline{x}} := \frac{1}{N}\sum_{n=1}^{N} \underline{x}_n \,, \quad \text{where: } \underline{x}_n \in \mathbb{R}^D \,. \tag{3.23}$$

Similar to the empirical mean, the empirical covariance matrix is a $D \times D$ matrix:

$$\underline{\Sigma} := \frac{1}{N}\sum_{n=1}^{N} (\underline{x}_n - \bar{\underline{x}})(\underline{x}_n - \bar{\underline{x}})^T \,. \tag{3.24}$$

↳ To compute the statistics for a particular dataset, we would use the realizations (observations) $\underline{x}_1, \dots, \underline{x}_N$ and use the Eqs. (3.23) and (3.24). Empirical covariance matrices are symmetric, positive and semidefinite.

(8.) Sums and Transformations of Random Variables

↳ We may want to model a phenomenon that cannot be well explained by textbooks distributions, and hence may perform simple manipulations of random variables (such as adding two random variables). Consider two random variables X, Y with states $\underline{x}, \underline{y} \in \mathbb{R}^D$. Then:

$$\bullet \begin{cases} \mathbb{E}[\underline{x}+\underline{y}] = \mathbb{E}[\underline{x}] + \mathbb{E}[\underline{y}] , \\ \mathbb{E}[\underline{x}-\underline{y}] = \mathbb{E}[\underline{x}] - \mathbb{E}[\underline{y}] , \\ \mathbb{V}[\underline{x}+\underline{y}] = \mathbb{V}[\underline{x}] + \mathbb{V}[\underline{y}] + \mathrm{Cov}[\underline{x},\underline{y}] + \mathrm{Cov}[\underline{y},\underline{x}] , \\ \mathbb{V}[\underline{x}-\underline{y}] = \mathbb{V}[\underline{x}] + \mathbb{V}[\underline{y}] - \mathrm{Cov}[\underline{x},\underline{y}] - \mathrm{Cov}[\underline{y},\underline{x}] . \end{cases} \tag{3.25}$$

Mean and (co)variance exhibit some useful properties when it comes to affine transformations of random variables. Consider a random variable X with mean $\underline{\mu}$ and covariance $\underline{\Sigma}$ and a (deterministic) affine transformation $\underline{y} = \underline{A}\underline{x} + \underline{b}$ of $\underline{x}$. Then $\underline{y}$ is itself a random variable whose mean vector and covariance matrix are given by:

$$\bullet \begin{cases} \mathbb{E}_Y[\underline{y}] = \mathbb{E}_X[\underline{A}\underline{x}+\underline{b}] = \underline{A}\,\mathbb{E}_X[\underline{x}] + \underline{b} = \underline{A}\underline{\mu} + \underline{b} , \\ \mathbb{V}_Y[\underline{y}] = \mathbb{V}_X[\underline{A}\underline{x}+\underline{b}] = \mathbb{V}_X[\underline{A}\underline{x}] = \underline{A}\,\mathbb{V}_X[\underline{x}]\underline{A}^T = \underline{A}\underline{\Sigma}\underline{A}^T , \end{cases} \tag{3.26}$$

respectively. Furthermore:

$$\begin{aligned} \mathrm{Cov}[\underline{x},\underline{y}] &= \mathbb{E}\left[\underline{x}(\underline{A}\underline{x}+\underline{b})^T\right] - \mathbb{E}[\underline{x}]\,\mathbb{E}\left[\underline{A}\underline{x}+\underline{b}\right]^T \\ &= \mathbb{E}[\underline{x}]\underline{b}^T + \mathbb{E}[\underline{x}\underline{x}^T]\underline{A}^T - \underline{\mu}\underline{b}^T - \underline{\mu}\underline{\mu}^T\underline{A}^T \\ &= \underline{\mu}\underline{b}^T - \underline{\mu}\underline{b}^T + \left(\mathbb{E}[\underline{x}\underline{x}^T] - \underline{\mu}\underline{\mu}^T\right)\underline{A}^T \\ &\underset{(3.20)}{=} \underline{\Sigma}\,\underline{A}^T , \end{aligned} \tag{3.27}$$

where $\underline{\Sigma} = \mathbb{E}[\underline{x}\underline{x}^T] - \underline{\mu}\underline{\mu}^T$ is the covariance of X.

(9.) Statistical Independence

▸ Definition[10] (Independence): Two random variables X, Y are statistically independent if and only if:

$$p(\underline{x},\underline{y}) = p(\underline{x})\,p(\underline{y}) . \tag{3.28}$$

↳ Intuitively, two random variables X and Y are independent if the value of $\underline{y}$ (once known) does not add any additional information about $\underline{x}$ (and vice versa). If X, Y are (statistically) independent, then:

$$\begin{cases} p(\underline{y}|\underline{x}) = p(\underline{y}) \,, \\ p(\underline{x}|\underline{y}) = p(\underline{x}) \,, \\ \mathbb{V}_{X,Y}[\underline{x}+\underline{y}] = \mathbb{V}_X[\underline{x}] + \mathbb{V}_Y[\underline{y}] \,, \\ \mathrm{Cov}_{X,Y}[\underline{x},\underline{y}] = \underline{0} \end{cases} \tag{3.29}$$

The last result may not hold in converse, i.e., two random variables can have covariance zero but are not statistically independent. To understand why, recall that covariance measures only linear dependence. Therefore, random variables that are not nonlinearly dependent could have covariance zero.

↳ In machine learning, we often consider problems that can be modeled as independent and identically distributed (i.i.d.) random variables, $X_1, \ldots, X_N$. For more than two random variables, the word "independent" given in DEFINITION[10] usually refers to mutually independent random variables, where all subsets are independent. The phrase "identically distributed" means that all the random variables are from the same distribution.

▶ DEFINITION[11] (Conditional Independence): Two random variables X and Y are conditionally independent given Z if and only if:

$$p(\underline{x}, \underline{y}|\underline{z}) = p(\underline{x}|\underline{z})\, p(\underline{y}|\underline{z}) \,, \quad \forall\, \underline{z} \in \mathcal{Z} \,, \tag{3.30}$$

where $\mathcal{Z}$ is the set of states of random variable Z. We write $X \perp\!\!\!\perp Y | Z$ to denote that X is conditionally independent of Y given Z.

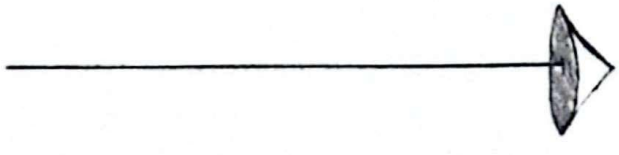
↳ DEFINITION[11] requires that the relation in Eq. (3.30) must hold true for every value of $\underline{z}$. The interpretation of Eq. (3.30) can be understood as "given knowledge about $\underline{z}$, the distribution of $\underline{x}$ and $\underline{y}$ factorizes". Independence can be cast as a special case of conditional independence if we write $X \perp\!\!\!\perp Y | \emptyset$. By using the product rule of probability given by Eq. (3.9), we can expand the left-hand side of Eq. (3.30) to obtain:

$$p(\underline{x}, \underline{y}|\underline{z}) = p(\underline{x}|\underline{y}, \underline{z})\, p(\underline{y}|\underline{z}) \,. \tag{3.31}$$

By comparing the right-hand side of Eq. (3.30) with Eq. (3.31), we see that $p(\underline{y}|\underline{z})$ appears in both sides of them so that:

$$p(\underline{x}|\underline{y}, \underline{z}) = p(\underline{x}|\underline{z}) \,. \tag{3.32}$$

Eq (3.32) provides an alternative definition of conditional independence, i.e., $X \perp\!\!\!\perp Y | Z$. This alternative presentation provides the interpretation "given that we know $\underline{z}$, knowledge about $\underline{y}$ does not change our knowledge of $\underline{x}$".



10. Gaussian Distribution

↳ The Gaussian distribution is the most well-studied probability distribution for continuous-valued random variables. It is also referred to as the normal distribution. Its importance originates from the fact that it has many computationally convenient properties. In particular, we will use it to define the likelihood and prior for linear regression, and consider a mixture of Gaussians for density estimation.

There are many other areas of machine learning that also benefit from using a Gaussian distribution, for example Gaussian processes, variational inference, and reinforcement learning. It is also widely used in other application areas such as signal processing (e.g., Kalman filter), control (e.g., linear quadratic regulator), and statistics (e.g., hypothesis testing).

↳ For a univariate random variable, the Gaussian distribution has a density that is given by:

$$p(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] . \qquad (3.33)$$

The multivariate Gaussian distribution is fully characterized by a mean vector $\underline{\mu}$ and a covariance matrix $\underline{\Sigma}$ and defined as:

$$p(\underline{x}|\underline{\mu},\underline{\Sigma}) = (2\pi)^{-D/2}|\underline{\Sigma}|^{-1/2} \exp\left[-\frac{1}{2}(\underline{x}-\underline{\mu})^T\underline{\Sigma}^{-1}(\underline{x}-\underline{\mu})\right] , \text{ where : } \underline{x}\in\mathbb{R}^D . \qquad (3.34)$$

We write $p(\underline{x}) = \mathcal{N}(\underline{x}|\underline{\mu},\underline{\Sigma})$ or $X \sim \mathcal{N}(\underline{\mu},\underline{\Sigma})$. The special case of the Gaussian with zero mean and identity covariance, that is, $\underline{\mu}=\underline{0}$ and $\underline{\Sigma}=\underline{I}$, is referred to as the standard normal distribution.

↳ Gaussians are widely used in statistical estimation and machine learning as they have closed-form expressions for marginal and conditional distributions. A major advantage of modeling with Gaussian random variables is that variable transformations are often not needed. Since the Gaussian distribution is fully specified by its mean and covariance, we often can obtain the transformed distribution by applying the transformation to the mean and covariance of the random variable.