Appendix A

Probability and Statistics Notions

In this chapter we review some useful notions from Probability and Statistics literature to help the reader not familiar with these mathematical tools required to understand the developments in our thesis. The intent is to focus on intuition and usefulness rather than strict rigor in order to keep notation and explanations simple. For the interested reader, we provide references that contain a more rigorous treatment. Throughout this introduction we assume that the reader is familiar with elementary notions of set theory and elementary calculus.

A.1 Basic Probability Notions

In this section we give a brief overview of useful notions from Probability Theory. A rigorous introduction can be found, for example, in (Resnick, 1999), but this overview should suffice for the purpose of this thesis.

A.1.1 Probabilities and Random Variables

We first introduce the notion of probability and random variable and their conditional counterparts, then we introduce variance and covariance and give some of their useful properties.

Probability

Let Ω be some set and $\mathcal{F}$ be a set of subsets of Ω that contains $\emptyset$ and Ω and it is closed under union, intersection and complementation with respect to Ω (i.e. the intersection, union and complement of sets in $\mathcal{F}$ gives sets in $\mathcal{F}$). The pair $(\Omega, \mathcal{F})$ is called a *probability space*, and any element $A \in \mathcal{F}$ is called an *event*. If an event does not contain any other event, it is called an *elementary event*. We call Ω the *probability space* and $\mathcal{F}$ the *set of measurable sets*. With this, a mapping $P : \mathcal{F} \to [0, 1]$ from the set of measurable sets to the real numbers between 0 and 1 is called a *probability function*, in short probability, if the following properties hold:

$$P[\Omega] = 1 \tag{A.1}$$

$$\forall A, B,\ A \cap B = \emptyset,\ P[A \cup B] = P[A] + P[B] \tag{A.2}$$

where $A, B \in \mathcal{F}$ are two measurable sets.

These properties are enough to show that the following properties also hold:

$$P[\emptyset] = 0 \tag{A.3}$$

$$P[A] \leq P[B], \text{ if } A \subset B \tag{A.4}$$

$$P[\overline{A}] = 1 - P[A] \tag{A.5}$$

$$P[A - B] = P[A] - P[A \cap B] \tag{A.6}$$

$$P[A \cup B] = P[A] + P[B] - P[A \cap B] \tag{A.7}$$

where we denoted by $\overline{A}$ the complement of event A. $P[A \cap B]$ is usually replaced by the simpler notation $P[A, B]$, the probability that events A and B happen together.

Two types of probabilities are interesting for the purpose of understanding this thesis: discrete probabilities and continuous probabilities. We briefly take a look at each, deferring further discussion until random variables are introduced.

If set Ω is a finite set and we take $\mathcal{F} = 2^{\Omega}$ – the powerset of Ω, i.e. the set of all the possible subsets – any probability over Ω is fully specified by the probabilities of the elementary events, which are nothing else than the elements of Ω. We call such a probability a *discrete probability*.

If we take $\Omega = \mathbb{R}$, with $\mathbb{R}$ the set of all real numbers, and $\mathcal{F}$ to be the transitive closure under intersection and complement of the compact intervals over the real numbers (the so called Borel set), any probability defined over Ω is called a *continuous probability*. The notion of continuous probability is also extended to vector spaces over the real numbers in the natural manner. We will see examples of continuous probabilities in the next section. A continuous probability function P can be specified by its density function $p(x)$. Intuitively, $p(x)dx$ is the probability to see value x. Obviously for any $x \in \mathbb{R}$ this probability is 0, but this allows

the specification of the probabilities of intervals, that are the elementary events of continuous probabilities:

$$P[[a,b]] = \int_a^b p(x)dx \tag{A.8}$$

where $[a,b]$ is a compact interval of $\mathbb{R}$.

Independent Events

Events A and B are called independent if:

$$P[A,B] = P[A] \cdot P[B] \tag{A.9}$$

The notion of independent events is very important because of this factorization property of the probability, factorization that greatly simplifies the analysis.

Conditional Probability and

The *conditional probability* that event B happens given that event A happened, denoted by $P[B|A]$, is defined as:

$$P[B|A] = \frac{P[A,B]}{P[A]} \tag{A.10}$$

The conditional probability has the following useful properties:

$$P[A|\Omega] = P[A] \tag{A.11}$$

$$P[B|A] = 1, \text{ if } A \subset B \tag{A.12}$$

$$P[B|A] = \frac{P[B] \cdot P[A|B]}{P[A]} \tag{A.13}$$

The last formula is called *Bayes rule*.

Also, conditional probabilities have all the properties normal probabilities have.

Random Variables

A mapping $X : \Omega \rightarrow \mathbb{R}$ is called a *random variable* with respect to the probability space $(\Omega, \mathcal{F})$ if it has the property that:

$$\forall a \in \mathbb{R},\ \{\omega \in \Omega | X(\omega) < a\} \in \mathcal{F} \tag{A.14}$$

For discrete probability spaces, any mapping is a random variable. For continuous spaces, it is enough to require the mapping to be continuous everywhere except a finite number of points. Moreover, by combining random variables using continuous functions, random variables are also obtained. What this amounts to is the fact that all mapping we have to deal with in our thesis are random variables.

A random variable defined over a discrete or continuous probability space is called *discrete random variable* or *continuous random variable*, respectively. To specify a discrete random variable, it is enough to specify the value of the random variable for each elementary event. For continuous random variables, we have to specify the values of the random variable for each real number. We will see examples of random variables in the next section.

A very important notion with respect to random variables is the notion of *expectation*. Intuitively, the expectation of a random variable is its average value with respect to a probability function. We denote the expectation of a random variable X by $E_P[X]$. If the probability function is understood from the context, we use the simpler notation $E[X]$.

For discrete random variables, the expectation is defined as:

$$E[X] = \sum_{\omega \in \Omega} X(\omega) P[\omega] \tag{A.15}$$

For convenience, we also use the alternative notation X_ω instead of $X(\omega)$.

For continuous variables, the expectation of random variable X with respect to the probability function P with density $p(x)$ is defined as:

$$E[X] = \int_{-\infty}^{\infty} X(x)p(x)dx \tag{A.16}$$

A probability space together with a probability function are usually called a distribution. Discrete distributions are usually specified by the probability of the elementary events and continuous distributions by the density function $p(x)$. We say that a random variable X is distributed according to the distribution D, denoted by $X \sim D$, if the probability is specified by the distribution and X is the identity function. This means that for discrete distributions, Ω is a subset of $\mathbb{R}$ in general but a subset of $\mathbb{N}$ or $\mathbb{Z}$ most often.

Important properties of expectation are:

1. expectation of a constant:

$$E(a) = a$$

2. linearity of expectation:

$$E[aX] = aE[X]$$

$$E[X + Y] = E[X] + E[Y]$$

3. expected value of sum (no independence required):

$$E\left[\sum_i X_i\right] = \sum_i E[X_i] \tag{A.17}$$

Independent Random Variables

Two random variables X and Y, defined over the same probability space $\Omega, \mathcal{F}$ are independent if and only if for all $x, y \in \mathbb{R}$, the events $\{\omega \in \Omega | X(w) < x\}$ and

$\{\omega \in \Omega | Y(w) < y\}$ are independent. In this case it can be shown that:

$$E[XY] = E[X]\,E[Y] \tag{A.18}$$

which is one the most useful properties of expectation.

Variance and Covariance

Variance is an important property of distributions since it indicates how spread (or localized) the distribution is. It is defined as:

$$\begin{aligned} \mathrm{Var}(X) &= E[(X - E[X])] \\ &= E\left[X^2\right] - (E[X])^2 \end{aligned} \tag{A.19}$$

The covariance of two random variables X and Y is defined as:

$$\mathrm{Cov}(X,Y) = E[XY] - E[X]\,E[Y]$$

and gives an idea of how much random variables X and Y influence each-other. Notice that if X and Y are independent, $\mathrm{Cov}(X,Y) = 0$.

Some of the useful properties of variance are:

1. variance of a constant:

$$\mathrm{Var}(a) = 0$$

2. scalar multiplication:

$$\mathrm{Var}(aX) = a^2 \mathrm{Var}(X)$$

3. variance of sum of random variables:

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y)$$

or in general

$$\mathrm{Var}\left(\sum_i X_i\right) = \sum_i \mathrm{Var}(X_i) + \sum_i \sum_{i' \neq i} \mathrm{Cov}(X_i, X_{i'})$$

4. variance of sum of independent random variables:

$$\mathrm{Var}\left(\sum_i X_i\right) = \sum_i \mathrm{Var}\left(X_i\right)$$

A very useful property of covariance is the fact that it is bilinear:

$$\begin{aligned}
\mathrm{Cov}\left(aX, Y\right) &= a\mathrm{Cov}\left(X, Y\right) \\
\mathrm{Cov}\left(X, aY\right) &= a\mathrm{Cov}\left(X, Y\right) \\
\mathrm{Cov}\left(X_1 + X_2, Y\right) &= \mathrm{Cov}\left(X_1, Y\right) + \mathrm{Cov}\left(X_2, Y\right) \\
\mathrm{Cov}\left(X, Y_1 + Y_2\right) &= \mathrm{Cov}\left(X, Y_1\right) + \mathrm{Cov}\left(X, Y_2\right) \\
\mathrm{Cov}\left(\sum_i X_i, \sum_j Y_j\right) &= \sum_i \sum_j \mathrm{Cov}\left(X_i, Y_j\right)
\end{aligned}$$

Also, the covariance is commutative:

$$\mathrm{Cov}\left(X, Y\right) = \mathrm{Cov}\left(Y, X\right)$$

Conditional Expectation

Conditional expectation generalizes the notion of conditional expectation. For random variable X defined over the discrete probability $(\Omega, 2^\Omega, P)$, and an event $A \in 2^\Omega$, the conditional expectation is defined as:

$$E\left[X|A\right] = \frac{\sum_{\omega \in A} X(\omega) P[\omega]}{P[A]} \tag{A.20}$$

For a continuous probability with density $p(x)$, the conditional expectation is defined as:

$$E[X|A] = \frac{\int_A X(x) p(x) dx}{P[A]} \tag{A.21}$$

Conditional expectation has all the properties normal expectation has. Moreover, since the notion of variance is entirely based on the notion of expectation,

we can define *conditional variance* in terms of the conditional expectation as:

$$\mathrm{Var}\left(X|A\right) = E\left[X^2|A\right] - \left(E\left[X|A\right]\right)^2$$

Random Vectors

The notion of random variable can be extended to vectors and, more generally, to matrices. If $\mathbf{X} = [X_1, \ldots, X_n]$ is a random vector – a vector of random variables – its expectation is the vector of expectations of components:

$$E\left[\mathbf{X}\right] = \left[E\left[X_1\right], \ldots, E\left[X_n\right]\right]$$

With this, the variance of random vector $\mathbf{X}$ is a matrix, called the covariance matrix:

$$\begin{aligned}\mathrm{Var}\left(\mathbf{X}\right) &= E\left[\mathbf{X}^T\mathbf{X}\right] - E\left[\mathbf{X}\right]^T E\left[\mathbf{X}\right] \\ &= \begin{pmatrix} \mathrm{Var}\left(X_1\right) & \mathrm{Cov}\left(X_1, X_2\right) & \ldots & \mathrm{Cov}\left(X_1, X_n\right) \\ \mathrm{Cov}\left(X_2, X_1\right) & \mathrm{Var}\left(X_2\right) & \ldots & \mathrm{Cov}\left(X_2, X_n\right) \\ \ldots & \ldots & \ldots & \ldots \\ \mathrm{Cov}\left(X_n, X_1\right) & \mathrm{Cov}\left(X_n, X_2\right) & \ldots & \mathrm{Var}\left(X_n\right) \end{pmatrix}\end{aligned} \tag{A.22}$$

A.2 Basic Statistical Notions

In this section we introduce some useful statistical notions. More information can be found, for example, in (Wilks, 1962; Pratt et al., 1995; Shao, 1999).

P-Value

The p-value of an observed value x of a random variable X is the probability that the random value X would take a value as high or higher than x. Mathematically, the p-value is $P[X > x]$. Intuitively, a very small p-value is statistical proof that x is not a sample of the random variable X.

A.2.1 Discrete Distributions

Binomial Distribution

Binomial distribution is the distribution of the number of times one sees the head in N flips of an asymmetric coin that has probability of tossing head p. If X is a random variable binomially distributed with parameters N and p it can be shown that:

$$E(X) = Np$$

$$\mathrm{Var}\,(X) = Np(1-p)$$

The p-value of the binomial distribution is:

$$P[X > x] = 1 - I(p; x+1, N-x)$$

where

$$I(x; a, b) = \int_{-\infty}^{x} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} t^{a-1}(1-t)^{b-1} dt$$

is the incomplete regularized beta function.

Multinomial Distribution

The multinomial distribution generalizes the binomial distribution to multiple dimensions. It has as parameters N, the number of trials, and $(p_1, \ldots, p_n)$, the probabilities of an n face coin. The multinomial distribution is the distribution of number of times each of the faces is observed out of N trials. If we let $\mathbf{X} \sim \mathrm{Multinomial}(N, p_1, \ldots, p_n)$, and denote by X_i the i-th component of $\mathbf{X}$ we

have:

$$E[X_i] = Np_i$$

$$\mathrm{Var}(X_i) = Np_i(1-p_i)$$

$$\mathrm{Cov}(X_i, X_j) = -Np_ip_j$$

A.2.2 Continuous Distributions

Normal (unidimensional Gaussian) Distribution

Normal distribution, denoted by $N(\mu, \sigma^2)$, has two parameters: the mean μ and variance σ^2. σ must always be a positive quantity. Given $X \sim N(\mu, \sigma^2)$,

$$E[X] = \mu \tag{A.23}$$

$$\mathrm{Var}(X) = \sigma^2 \tag{A.24}$$

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{A.25}$$

$$P[X > x] = \frac{1}{2}\left(1 - \mathrm{Erf}\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)\right) \tag{A.26}$$

where $\mathrm{Erf}(x) = \int_{-\infty}^{x} e^{t^2/2} dt$ is the standard error function.

Gaussian Distribution

Gaussian distribution, denoted by $N(\mu, \Sigma)$, has two parameters: the mean vector μ and the covariance matrix Σ. Σ has to be positive definite which means that it always has a Choleski decomposition $\Sigma = GG^T$ (Golub & Loan, 1996). For

$\mathbf{X} \sim N(\mu, \Sigma)$,

$$E[\mathbf{X}] = \mu \tag{A.27}$$

$$\mathrm{Var}(\mathbf{X}) = \Sigma \tag{A.28}$$

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma_i|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)} \tag{A.29}$$

Gamma Distribution

The gamma distribution (with parameters α and θ) is the distribution with density

$$p(x) = \frac{x^\alpha e^{-x/\theta}}{\Gamma(\alpha)\theta^\alpha}$$

and p-value

$$P[X > x] = 1 - \frac{\Gamma(\alpha, x/\theta)}{\Gamma(\alpha)} = 1 - Q(\alpha, x/\theta) \tag{A.30}$$

where $\Gamma(x)$ is the gamma function and $\Gamma(x, y)$ is the incomplete gamma function. $Q(x, y)$ is called the incomplete regularized gamma function.

Mean and variance or a random variable X with gamma distribution are:

$$E(X) = \alpha\theta \tag{A.31}$$

$$\mathrm{Var}(X) = \alpha\theta^2 \tag{A.32}$$

Normal distribution is a particular case of the gamma distribution.

Beta Distribution

The beta distribution has parameters α and β and density:

$$p(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

The p-value is $1 - I(x; \alpha, \beta)$ with I the incomplete regularized beta function.

χ^2-test and χ^2-distribution

Having a set of random variables X_i, the χ^2-test is defined as:

$$\chi^2 = \sum_{i=1}^{n} \frac{(X_i - E_i)^2}{E_i} \tag{A.33}$$

where E_i is the expected value of X_i under some hypothesis (that is tested using the χ^2-test).

It can be shown that asymptoticly χ^2 has a χ^2 distribution, that coincides with a gamma distribution with parameters $\alpha = 1/2r$ and $\theta = 2$, where r is the degrees of freedom (number of variables n minus number of constrains between variables).

The mean and the variance for the χ^2 distribution are:

$$E(\chi^2) = r \tag{A.34}$$

$$\mathrm{Var}(\chi^2) = 2r \tag{A.35}$$