Mathematical methods in communication July 5th, 2009

Appendix A : Introduction to Probability and stochastic processes

Lecturer: Haim Permuter *Scribe: Shai Shapira and Uri Livnat*

"The probability of an event is the ratio of the number of cases favorable to it, to the number of all cases possible when nothing leads us to expect that any one of these cases should occur more than any other, which renders them, for us, equally possible."

Pierre Simon Laplace , 1812

I. Basic Concepts of Probability.

Definition I.1 (Probability Space: (Ω, F, P)) Probability space formalizes three interrelated ideas by three mathematical notions.

- Ω - Sample space : set of all possible outcomes ω of a particular random experiment.
  Where, $\{\omega : \omega \in \Omega\}$.
- F - Collection of all events, subsets of Ω.
- P - Probability measure.

Properties I.1 (Probability Space) 1) $\mathrm{P} : F \to [0, 1]$

2) Empty Event - $\emptyset$. $\mathrm{P}(\emptyset) = 0$
3) Deterministic Event - Ω. $\mathrm{P}(\Omega) = 1$
4) Disjoint Events - $A \bigcap B = \emptyset$
5) Complementary Events - $A \bigcap A^c = \emptyset,\ A \bigcup A^c = \Omega$
6) For each collection of disjoint events $\{A_n, A_n \in F\}, A_i \bigcap A_j = \emptyset\ \forall i \neq j$
   exists: $\mathrm{P}\left(\bigcup_n A_n\right) = \sum_n \mathrm{P}(A_n)$

Conclusions I.1 (Probability Space) Conclusions arising from the above properties:

1) $\mathrm{P}(A^c) = 1 - \mathrm{P}(A)$
2) For any arbitrary eventrs $\{A, B\}$: $\mathrm{P}(A \bigcup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \bigcap B)$

Example 1 (Fair coin flip) If the space concerns one flip of a fair coin, then the outcomes are heads and tails: $\Omega = \{H, T\}$. $F = 2^\Omega$ contains $2^2 = 4$ events, namely, $\{H\}$: heads, $\{T\}$: tails, $\{\}$: neither heads nor tails, and $\{H, T\}$: heads or tails. So, $F = \{\{\}, \{H\}, \{T\}, \{H, T\}\}$. There is a fifty percent chance of tossing either heads or tail: thus $P(\{H\}) = P(\{T\}) = 0.5$. The chance of tossing neither is zero: $P(\{\}) = 0$, and the chance of tossing one or the other is one: $P(\{H, T\}) = 1$.

A. Conditional Probability.

Definition I.2 (Conditional probability) is the probability of some event A, given the occurrence of some other event B. Conditional probability is written $\mathrm{P}(A|B)$, and is read "the probability of A, given B". Intersection events and conditional events are related by the formula:

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(A \bigcap B)}{\mathrm{P}(B)} \tag{I.1}$$

which can also be written as,

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(A, B)}{\mathrm{P}(B)} \tag{I.2}$$

B. Statistically Independent Probabilities.

Definition I.3 (Statistically Independent Probabilities) Events A and B are Statistically Independent iff:

$$\mathrm{P}(A|B) = \mathrm{P}(A), \tag{I.3}$$

or similarly,

$$\mathrm{P}\left(A \bigcap B\right) = \mathrm{P}(A)\,\mathrm{P}(B) \tag{I.4}$$

which can also be written as,

$$\mathrm{P}(A, B) = \mathrm{P}(A)\,\mathrm{P}(B) \tag{I.5}$$

C. Bayes' Theorem.

Bayes' theorem relates the conditional and marginal probabilities of events A and B, where B has a non-vanishing probability:

Theorem I.1 (Bayes' theorem)

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(A)\,\mathrm{P}(B|A)}{\mathrm{P}(B)} \tag{I.6}$$

Each term in *Bayes' theorem* has a conventional name.

- $\mathrm{P}(A)$ is the prior probability or marginal probability of A. It is "prior" in the sense that it does not take into account any information about B.
- $\mathrm{P}(A|B)$ is the conditional probability of A, given B. It is also called the posterior probability because it is derived from or depends upon the specified value of B.
- $\mathrm{P}(B|A)$ is the conditional probability of B given A.
- $\mathrm{P}(B)$ is the prior or marginal probability of B, and acts as a normalizing constant.

Intuitively, Bayes' theorem in this form describes the way in which one's beliefs about observing $'A'$ are updated by having observed $'B'$.

II. Random Variables.

A random variable is a variable whose possible values are numerical outcomes of a stochastic phenomenon. There are two types of random variables, *discrete* and *continuous*.

A. Continuous Random Variables.

Definition II.1 (Continuous Random Variables) is a function $X : \Omega \to \mathbb{R}$ such that for any real number a the set $\{\omega : X(\omega) \leq a\}$ is an event. According to that definition:

1) The Cumulutive Distribution Function (CDF) of a continuous random variable:
   $F(\alpha) = \mathrm{P}(\{X(\omega) \leq \alpha\})$
2) The Probability Density Function (PDF) of a continuous random variable where $F_X(\alpha)$ is a differentiable function $f_X(\alpha) = \frac{\partial F_X(\alpha)}{\partial \alpha}$
   Note that the derivative $\frac{\partial F_X(\alpha)}{\partial \alpha}$ might not always exist, but in our course we will deal only with continuous random variable that the PDF exists

Properties II.1 (Continuous Random Variables) 1) Its CDF is monotonically rising and right continuous.
2) $F_X(\infty) = 1, \ F_X(-\infty) = 0$
3) if $a_2 > a_1$ then $\mathrm{P}(a_1 < X \leq a_2) = F_X(a_2) - F_X(a_1)$

Conclusions II.1 (Continuous Random Variables) 1) $F_X(\beta) = \int_{-\infty}^{\beta} f_X(\alpha) d\alpha$
2) $\int_{-\infty}^{\infty} f_X(\alpha) d\alpha = F_X(\infty) = 1$
3) $f_X(\alpha) \geq 0$

B. Discrete Random Variables.

Definition II.2 (Discrete Random Variables) Let $\{x_0, x_1, \dots\}$ be the values a discrete r.v can take with non-zero probability. $\Omega_i = \{\omega : X(\omega) = x_i\}$, $i \in \mathbb{N}$

1) A discrete r.v has a Probability Mass Function (PMF) (instead of a PDF such as for a continuous r.v). It is a function that gives the probability that a discrete r.v is exactly equal to some value.

$$P_X(x) = \mathrm{P}(X = x) = \mathrm{P}(\{\omega \in \Omega : X(\omega) = x\}) \qquad \text{where} \qquad \{\omega \in \Omega\} \tag{II.1}$$

2) Since the image of X is countable, the probability mass function $P_X(x)$ is zero for all but a countable number of values of X. The discontinuity of probability mass functions reflects the fact that the CDF of a discrete r.v is also discontinuous. Where it is differentiable, the derivative is zero, just as the probability mass function is zero at all such points.

Properties II.2 (Discrete Random Variables) 1) $\mathrm{P}_X(x) \geq 0$

2) $\sum_i \mathrm{P}(X = x_i) = \sum_i \mathrm{P}_X(x_i) = 1$

C. Expectation.

Definition II.3 (Expected value of a continuous random variable)

$$\mu = \mathrm{E}[X] = \int_{-\infty}^{\infty} \alpha f_X(\alpha) d\alpha \qquad \text{where} \qquad \left\{ \int_{-\infty}^{\infty} f_X(\alpha) d\alpha < \infty \right\} \tag{II.2}$$

Definition II.4 (Expected value of a discrete random variable)

$$\begin{aligned} \mathrm{E}[X] &= \sum_{x \in \mathcal{X}} x \, \mathrm{P}(X = x) \\ &= \sum_i x_i \, \mathrm{P}(X = x_i) \end{aligned} \tag{II.3}$$

Properties II.3 (Expectation) 1) For a deterministic variable: $\mathrm{E}[c] = c$

2) Linearity : $\mathrm{E}[cX + dY] = \mathrm{E}[cX] + \mathrm{E}[dY]$ where X, Y are r.v and c, d are constants.

3) Monotonicity : If X and Y are random variables such that $X(\omega) \geq Y(\omega)$ then $\mathrm{E}[X] \geq \mathrm{E}[Y]$.

D. Variance.

The *variance* of a random variable is a measure of statistical dispersion, averaging the squares of the deviations of its possible values from its expected value.

$$\begin{aligned} \mathrm{Var}[X] &= \mathrm{E}\left[(X - \mathrm{E}[X])^2\right] \\ &= \mathrm{E}\left[X^2\right] - (\mathrm{E}[X])^2 \end{aligned} \tag{II.4}$$

Definition II.5 (Variance of a continuous random variable)

$$\mathrm{Var}(X) = \int_{-\infty}^{\infty} (\alpha - \mathrm{E}[X])^2 f_x(\alpha) d\alpha \tag{II.5}$$

Definition II.6 (Variance of a discrete random variable)

$$\mathrm{Var}(X) = \sum_{x \in \mathcal{X}} (x - \mathrm{E}[X])^2 \, \mathrm{P}(X = x) \tag{II.6}$$

$$= \sum_{x \in \mathcal{X}} x^2 \, \mathrm{P}(X = x) - \mathrm{E}^2[X] \tag{II.7}$$

Properties II.4 (Variance) 1) For a deterministic variable: $\mathrm{Var}[c] = 0$

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

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

E. Covariance.

Definition II.7 (Covariance) Covariance is a measure of how much two variables change together.

$$\mathrm{Cov}(X,Y) = \mathrm{E}\left[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])\right] \tag{II.8}$$

which can also be written as,

$$\mathrm{Cov}(X,Y) = \mathrm{E}[XY] - \mathrm{E}[X]\,\mathrm{E}[Y] \tag{II.9}$$

Properties II.5 (Covariance) Let X, Y be real valued r.v and a, b are constants.

1) If X and Y are independent their covariance is zero and they are called uncorrelated.

$$\begin{aligned}\mathrm{Cov}(X,Y) &= \mathrm{E}[XY] - \mathrm{E}[X]\,\mathrm{E}[Y] \\ &= \mathrm{E}[X]\,\mathrm{E}[Y] - \mathrm{E}[X]\,\mathrm{E}[Y] \\ &= 0\end{aligned}$$

2) $\mathrm{Cov}(X,a) = 0$
3) $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$
4) $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$
5) $\mathrm{Cov}(aX,bY) = ab\,\mathrm{Cov}(X,Y)$
6) $\mathrm{Cov}(X+a,Y+b) = \mathrm{Cov}(X,Y)$

Example 2 (Fair coin flip) For a coin toss, the possible events are heads or tails. The number of heads appearing in one fair coin toss can be described using the following random variable:

$$X = \begin{cases} 1, & \text{if heads,} \\ 0, & \text{if tails.} \end{cases}$$

with probability mass function given by:

$$f_X(x) = \begin{cases} \frac{1}{2}, & \text{if } x = 0, \\ \frac{1}{2}, & \text{if } x = 1, \\ 0, & \text{otherwise.} \end{cases}$$

Let's calculate the *Expectetaion* of the r.v :

$$\begin{aligned}\mathrm{E}[X] &= \sum_i x_i\,\mathrm{P}(X = x_i) \\ &= \frac{1}{2}\cdot 0 + \frac{1}{2}\cdot 1 \\ &= \frac{1}{2}\end{aligned}$$

And the *Variance*:

$$\begin{aligned}\mathrm{Var}(X) &= \sum_{i=1}^{2} {x_i}^2 \, \mathrm{P}(X=x) - \mathrm{E}^2[X] \\ &= \frac{1}{2}\cdot 0^2 + \frac{1}{2}\cdot 1^2 - \left(\frac{1}{2}\right)^2 \\ &= \frac{1}{4}\end{aligned}$$

III. Random Vectors.

Definition III.1 (Random Vectors) Let $X_1, X_2 \ldots X_n$ Random Variables of the same probability space. Then, Their columnwise order will compose a Random Vector in $\mathbb{R}^n$. Hence, $X = \underline{X} = \begin{Bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{Bmatrix}$

A. CDF of Random Vector:

Definition III.2 (CDF of Random Vector)

$$F_X(\alpha_1, \alpha_2 \ldots \alpha_n) = P(X_1 \le \alpha_1, X_2 \le \alpha_2, \ldots, X_n \le \alpha_n) \tag{III.1}$$

- if $F_X(\alpha_1, \alpha_2 \ldots \alpha_n) = \prod_{i=1}^{n} f_{X_i}(\alpha_i)$ then $X_1, X_2 \ldots X_n$ are statistically independent, or mutually independent.

IV. Independent and Identically-Distributed Random Variables.

Definition IV.1 (i.i.d.) A collection of random variables is *independent and identically distributed (i.i.d.)* if each random variable has the same probability distribution as the others and all are mutually independent

V. Gaussian Distribution.

A. Gaussian Random Variable

Definition V.1 (Gaussian Random Variable) The random variable X is said to be a Gaussian random variable (or normal random variable) if its PDF has the form :

$$f_X(\alpha) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(\alpha-\mu)^2}{2\sigma^2}} \qquad \text{where} \qquad \left\{\mu = E[X],\ \sigma^2 = Var(X)\right\} \tag{V.1}$$

Hence, a Gaussian r.v is characterized by 2 parameters : $\left\{\mu, \sigma^2\right\}$ and its common notation is : $X \sim N(\mu, \sigma^2)$

B. Gaussian Random Vector

Definition V.2 (Gaussian Random Vector) The random variables $\{X_i\}_{i=1}^n$ are called Jointly Gaussian random variables, or similarly
$X = (X_1, X_2 \ldots X_n)^T$ is called Gaussian random vector if for every collection of non random constants $\{a_i\}_{i=1}^n$ such that $\sum_{i=1}^n a_i X_i$ is a Gaussian random vector.

Definition V.3 (Gaussian Random Vector) An equivalent definition: A random vector $X \in \mathbb{R}^n$ is called a Gaussian random vector if it is continuous with the joint PDF :

$$f_{(\underline{X})}\left(\underline{\alpha}\right)= \frac{1}{(2\pi)^{\frac{n}{2}} |\Lambda|^{\frac{1}{2}}}\, e^{-\frac{1}{2}\left[\left(\underline{X}-\underline{\mu}\right)^T \Lambda^{-1}\left(\underline{X}-\underline{\mu}\right)\right]} \tag{V.2}$$

where

$$\underline{\mu} = E\left[\underline{X}\right] = [E[X_1], E[X_2], \ldots, E[X_n]]$$
$$\Lambda = E\left[\left(\underline{X}-\underline{\mu}\right)\left(\underline{X}-\underline{\mu}\right)^T\right]$$

Properties V.1 (Gaussian Random Vector)
1) If $\underline{X}$ is a Gaussian random vector, each of its components is a Gaussian random variable.
2) If $\underline{X}$ is a random vector with independent components and each one is a Gaussian r.v then $\underline{X}$ is a Gaussian random vector.
3) Linear transformation of a Gaussian random vector $\underline{Y} = \mathbf{A}\underline{X} + \underline{b}$ is a Gaussian random vector.
4) The Gaussian distribution is uniquely determined by the 2 first moments: $\left\{\underline{\mu}, \Lambda\right\}$.

VI. MMSE Estimator.

MMSE - Minimum Mean Square Error

In statistics and signal processing, an *MMSE estimator* describes the approach which minimizes the mean square error (MSE), which is a common measure of estimator quality.

A. MMSE Estimator.

Definition VI.1 (MMSE Estimator) Let X be an unknown random variable/vector, and let Y be a known random variable - the measurements of X. An estimator $\hat{X}(Y)$ is any function of the measurements Y, and its MSE is given by :

$$\mathrm{MSE} = \mathrm{E}[\epsilon^2]\Big|_{\epsilon=\hat{X}(Y)-X} = \mathrm{E}\left[\left(\hat{X}(Y) - X\right)^2\right] \tag{VI.1}$$

$$\hat{X}^{\mathrm{MMSE}}(Y) \;=\; \arg\min_{\hat{X}^{\mathrm{MMSE}}(Y)\in\forall\hat{X}(Y)}\{\mathrm{MSE}\} \quad \text{where} \quad \left\{\hat{X}(Y) = g(Y)\right\} \tag{VI.2}$$

The MMSE estimator is then defined as the estimator achieving minimal MSE.

B. Linear MMSE Estimator.

- The linear MMSE estimator is the estimator achieving minimum MSE among all estimators of the form $\mathbf{A}\underline{Y} + \underline{b}$. If the measurement Y is a random vector, **A** is a matrix and $\underline{\text{b}}$ is a vector.
- If X and Y are jointly Gaussian, then the MMSE estimator is linear. As a consequence, to find the MMSE estimator, it is sufficient to find the linear MMSE estimator.

$$\mathrm{MSE} = \mathrm{E}[\epsilon^2]\Big|_{\epsilon=\hat{X}_{\text{linear}}(Y)-X} = \mathrm{E}\left[\left(\hat{X}_{\text{linear}}(Y) - X\right)^2\right] \tag{VI.3}$$

$$\hat{X}^{\mathrm{MMSE}}_{\text{linear}}(Y) \;=\; \arg\min_{\hat{X}^{\mathrm{MMSE}}_{\text{linear}}(Y)\in\forall\hat{X}_{\text{linear}}(Y)}\{\mathrm{MSE}\} \quad \text{where} \quad \left\{\hat{X}_{\text{linear}} = \mathbf{A}Y + \underline{b}\right\} \tag{VI.4}$$

For the *Scalar* case :

$$\hat{X}^{\mathrm{MMSE}}_{\text{linear}}(Y) = \mathrm{E}[X] + \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(Y)}\left(Y - \mathrm{E}[Y]\right) \tag{VI.5}$$

The estimation error in the *Scalar* case :

$$\mathrm{MSE} = \mathrm{E}\left[\epsilon^2_{\text{linear}}\right] = \mathrm{Var}(X) - \frac{\mathrm{Cov}^2(X,Y)}{\mathrm{Var}(Y)} \tag{VI.6}$$

For the *Vectorial* case :

$$\underline{\hat{X}}^{\mathrm{MMSE}}_{\text{linear}}(Y) = \mathrm{E}[\underline{X}] + \Lambda_{\underline{XY}}\Lambda^{-1}_{\underline{YY}}\left(\underline{Y} - \mathrm{E}[\underline{Y}]\right) \tag{VI.7}$$

The estimation error *Covariance* matrix for the *Vectorial* case :

$$\Lambda_{\underline{\epsilon\epsilon}} = \mathrm{E}\left[\left(\underline{X} - \underline{\hat{X}}_{linear}\right)\left(\underline{X} - \underline{\hat{X}}_{linear}\right)^T\right] \tag{VI.8}$$

$$= \Lambda_{\underline{XX}} - \Lambda_{\underline{XY}}\Lambda^{-1}_{\underline{YY}}\Lambda_{\underline{YX}} \tag{VI.9}$$

Where,

$$\Lambda_{\underline{XY}} = \mathrm{E}\left[\left(\underline{X} - \mathrm{E}[\underline{X}]\right)\left(\underline{Y} - \mathrm{E}[\underline{Y}]\right)^T\right]$$

Properties VI.1 (MMSE Estimator) 1) $\hat{X}(Y)$ is an *unbiased* estimator of $\mathrm{E}[X]$. In other words, X and $\hat{X}(Y)$ have the same expectation:

$$\mathrm{E}[\hat{X}] = \mathrm{E}[X] \tag{VI.10}$$

Since the expectation of the error is 0.

$$\mathrm{E}[\epsilon] = 0 \qquad \text{where} \qquad \left\{\epsilon = \left(\hat{X} - X\right)\right\} \tag{VI.11}$$

2) The estimator $\hat{X}$ and the error ϵ are *orthogonal:*

$$\mathrm{E}[\hat{X}\cdot\epsilon] = 0 \qquad \text{(Orthogonality Principle)} \tag{VI.12}$$

This will imply that the estimator $\hat{X}$ and the error ϵ are *uncorrelated*