Indian Institute of Technology Bombay
Department of Electrical Engineering

Handout 4 EE 764 Wireless and Mobile Communication
Lecture Notes 2 Feb 10, 2015

1 Conditional Probability

As the name implies, there are at least two random variables involved, let us denote it as X and Y. Let us start with a familiar idea of probabilities of the joint events of the form i.e. $\{X \leq x\}$ and $\{Y \leq y\}$. We will use this later to define conditional probabilities. We know that

$$P(X \leq x | Y \leq y) = \frac{P(X \leq x, Y \leq y)}{P(Y \leq y)} \mathbb{1}_{\{P(Y \leq y) \neq 0\}}.$$

In the all-discrete case, we used $P(Y = y)$ instead of $P(Y \leq y)$ in the above expression, to define the conditional cumulative distribution. However, this cannot be done if Y is absolutely continuous, as $P(Y = y) = 0$ uniformly for all y. So we will define four separate versions of conditional probability depending on whether X and Y are continuous or not. Though we say different versions, all of them have the same theme, with appropriate replacement of distributions by densities. The first case is when X and Y are discrete, which is already familiar.

1. $X-$ discrete, $Y-$ discrete

$$P(X = x | Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} \mathbb{1}_{\{P(Y=y)>0\}}.$$

2. $X-$ continuous, $Y-$ discrete

Here we will replace the event $\{X = x\}$ by $\{X \leq x\}$

$$P(X \leq x | Y = y) = \int_{-\infty}^{x} f_{X|Y}(u|y) du \mathbb{1}_{\{P(Y=y)>0\}},$$

where we used the notation $f_{X|Y}(x|y)$ to denote the density function of X given $Y = y$. Notice that

$$\begin{aligned} P(X \leq x | Y = y) &= \frac{P(X \leq x, Y = y)}{P(Y = y)} \mathbb{1}_{\{P(Y=y)>0\}} \\ &= \mathbb{1}_{\{P(Y=y)>0\}} \frac{1}{P(Y = y)} \sum_{v=y} \int_{-\infty}^{x} f_{XY}(u, v) du \\ &= \int_{-\infty}^{x} \frac{1}{P(Y = y)} f_{XY}(u, y) \mathbb{1}_{\{P(Y=y)>0\}} du \end{aligned}$$

In this case, it makes sense to write

$$f_{X|Y}(u|y) = \frac{f_{XY}(u, y)}{P(Y = y)} \mathbb{1}_{\{P(Y=y)>0\}}. \tag{1}$$

Since the above case occurs frequently in the examples of electrical engineering, we will conveniently denote $f_{X|Y}(u|k)$ as $f_k(x)$, i.e. the conditional density of X given $Y = k$. To check the consistency of our definition, observe that

$$\begin{aligned}\int_{\mathbb{R}} f_{X|Y}(u|y)du &= \int \frac{f_{XY}(u,y)}{P(Y=y)}\mathbb{1}_{\{P(Y=y)>0\}}du \\ &= \frac{\mathbb{1}_{\{P(Y=y)>0\}}}{P(Y=y)}\int f_{XY}(u,y)du \\ &= \frac{\mathbb{1}_{\{P(Y=y)>0\}}}{P(Y=y)}P(Y=y) \\ &= \mathbb{1}_{\{P(Y=y)>0\}}.\end{aligned}$$

Thus, our definition gives a valid probability measure for all $y : P(Y = y) > 0$. Also, the marginal distribution of X becomes

$$P(X \le x) = \sum_k P(Y=k)\int_{-\infty}^{x} f_k(x)dx.$$

3. $X-$ discrete, $Y-$continuous

We will reverse engineer a consistent definition of conditional distribution from the last two cases. Specifically, let us define

$$P(X=i|Y=y) = \frac{f_{XY}(i,y)}{f_Y(y)}\mathbb{1}_{\{f_Y(y)>0\}}. \tag{2}$$

Clearly

$$\begin{aligned}\sum_i P(X=i|Y=y) &= \sum_i \frac{f_{XY}(i,y)}{f_Y(y)}\mathbb{1}_{\{f_Y(y)>0\}} \\ &= \frac{f_Y(y)}{f_Y(y)}\mathbb{1}_{\{f_Y(y)>0\}} \\ &= \mathbb{1}_{\{f_Y(y)>0\}}.\end{aligned}$$

Thus, our definition gives a valid probability measure for all $y : f_Y(y) > 0$. The conditional probability that we defined also takes the convenient form,

$$.\frac{f_{XY}(i,y)}{f_Y(y)} = \frac{P(X=i)f_i(y)}{f_Y(y)} = \frac{P(X=i)f_i(y)}{\sum_i P(X=i)f_i(y)}, \tag{3}$$

whenever the quantities are well-defined. Recall the definition of the conditional density $f_i(y)$ from the previous case.

4. $X-$ continuous, $Y-$ continuous

Let us define

$$f_{X|Y}(x|y) = \frac{f_{XY}(x,y)}{f_Y(y)}\mathbb{1}_{\{f_Y(y)>0\}}. \tag{4}$$

It is simple to check that

$$\int_{\mathbb{R}} f_{X|Y}(x|y)dx = \mathbb{1}_{\{f_Y(y)>0\}}.$$

Thus, it is a valid density function.

1.1 Conditional Distribution of Random Vectors

The definitions that we made are equally applicable to random vectors, which are of the mixed type, i.e. they may have some elements discrete and rest continuous. In each case, we can modify the appropriate definition and get consistent results.

2 Conditional Expectation

Expectation with respect to a conditional distribution is known as conditional expectation. Since we defined the conditional distribution for 4 separate cases, the conditional expectation has to be evaluated accordingly in these cases. For generality, we will denote the conditional distribution that we introduced in the last section by $\Pi(x|y)$. We will mention the general framework here.

Definition 1. *Consider a function $g : \mathbb{R}\times\mathbb{R} \to \mathbb{R}$, which is either non-negative or integrable. The function*

$$\Psi(y) = \mathbb{E}[g(X,Y)|Y = y]$$

is known as the conditional expectation of $g(X,Y)$ given $Y = y$, where the expectation is evaluated with respect to $\Pi(x|y)$.

Observe that this definition can be easily specialized to each of the cases that we dealt. The most important case in our discussion is the last one, where it reads,

$$\Psi(y) = \int_{\mathbb{R}} g(x,y)\Pi(x|y)dx$$

in which $\Pi(x|y)$ is same as the function $f_{X|Y}(x|y)$ given in (4). On the other hand, when X is discrete and Y continuous, we can write

$$\Psi(y) = \sum_i g(i,y)\Pi(i|y),$$

where $\Pi(i|y)$ is taken as per (2), or the more convenient form in (3).

Example 1. *Let (X_1, X_2) be a zero mean jointly Gaussian random vector with covariance matrix*

$$K = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}.$$

Find $\mathbb{E}[X_1|X_2]$ and $\mathbb{E}[X_1^2|X_2]$.

Solution: Notice that we need to specialize our definition of conditional distribution. In particular, since both RVs are continuous, let us look at their joint density.

$$f_{X_1,X_2}(x_1,x_2) = \frac{1}{\sqrt{\det(2\pi K)}} e^{-\frac{1}{2}x^T K^{-1} x},$$

where $x = (x_1, x_2)^T$. The marginal distribution of X_2 is a Gaussian (since the given vector is jointly Gaussian), which we can identify as

$$f_{X_2}(x_2) = \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-\frac{x_2^2}{2\sigma_2^2}}.$$

The conditional density becomes

$$\frac{f_{X_1,X_2}(x_1,x_2)}{f_{X_2}(x_2)} = \frac{1}{\sqrt{2\pi(1-\rho^2)\sigma_1^2}} e^{-\frac{1}{2(1-\rho^2)\sigma_1^2}(x_1-\rho\frac{\sigma_1}{\sigma_2}x_2)^2}.$$

Observe that the conditional density is nothing but a Gaussian density function with mean $\rho\frac{\sigma_1}{\sigma_2}x_2$ and variance $(1-\rho^2)\sigma_1^2$. Thus, it is easy to identify

$$\mathbb{E}[X_1|X_2] = \rho\frac{\sigma_1}{\sigma_2}x_2 \text{ and } \mathbb{E}[X_1^2|X_2] = (1-\rho^2)\sigma_1^2 + \rho^2\frac{\sigma_1^2}{\sigma_2^2}x_2^2.$$

■

3 Other results

It is clear that $\Psi(Y)$ defined above is a random variable taking values in $\bar{\mathbb{R}}$ whenever $\mathbb{E}|g(X,Y)| < \infty$, and then it is meaningful to talk about its expectation. We have performed such computations for the discrete case. Many results from the conditional expectation for discrete cases have exact analogs in the general case. Here we list a few of them, whose proofs are straightforward, whenever the quantities involved are finite or non-negative. For example

1.
$$\mathbb{E}\left[\mathbb{E}[g(X,Y)|Y=y]\right] = \mathbb{E}g(X,Y) \tag{5}$$

2.
$$\mathbb{E}[g(X)|X] = g(X) \tag{6}$$

3. If $X \perp\!\!\!\perp Y$,
$$\mathbb{E}_y\mathbb{E}_x[g(X)|Y] = \mathbb{E}g(X).$$

4. Wald's Inequality.

Exercise 1. *Using the definition of conditional probabilities, prove each of these expressions.*

4 Detection Theory

The conditional probability and expectation that we learned have many applications, particularly in communication and signal processing. We present a motivating application, known as *detection*, or *Hypothesis testing* in general statistical literature. In the detection problem, there is a finite set of indices or parameters that we are interested in. Let us denote it as $S = \{\theta_1, \cdots, \theta_K\}$. In communication problems, these parameters typically take the form of messages. In statistical literature, one says that there are K possible hypothesis, $H_0, \cdots, H_{K-1}$. In both cases, our job is to pick the correct hypothesis or parameter. What is meant by *correct*?.

Imagine a parameter of interest θ taking values in S. If we can observe θ, certainly we will always have the correct answer for θ. Unfortunately, in many real life situations we

may only have some noisy observations of what θ is, i.e. we observe some corrupted version of the parameter. This can be due to random *perturbations*, where a random vector $Y \in \mathbb{R}^n$ is observed, which is a function of θ and some other unobservable random phenomenon $Z \in \mathbb{R}^m$. Thus

$$Y = g(\theta, Z)\,,\ g : \mathbb{R} \times \mathbb{R}^m \to \mathbb{R}^n.$$

For example, a scalar output $Y = \theta + Z$ is a popular model.

In general, we have to design a device, called the *detector*, which predicts θ after observing Y. The output of the detector is $\hat{\theta}(Y)$, where $\hat{\theta}(Y) \in S$.

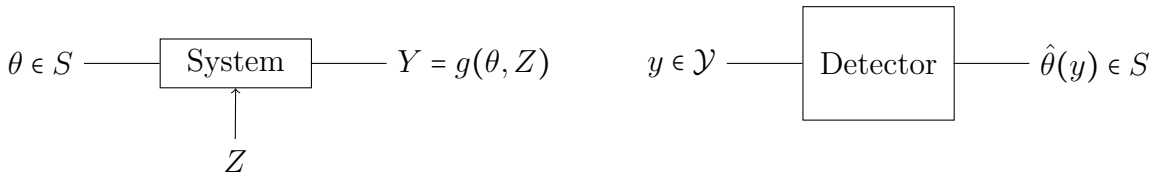

Figure 1: Observation and Detection: Block representation

The system demands that the detector should unambiguously output a value in S for every observed Y. Let $\mathcal{Y}$ be the space where Y takes value in. We will partition this into K regions. Denoting a disjoint union by $\sum$, we have

$$\mathcal{Y} = \sum A_i,$$

where $A_i = \{y \in \mathcal{Y} : \hat{\theta}(y) = i\}$. In other words, whenever y falls in the region A_i, the detector will output i as the value of $\hat{\theta}$. The region A_i is also known as the decoding region for parameter i.

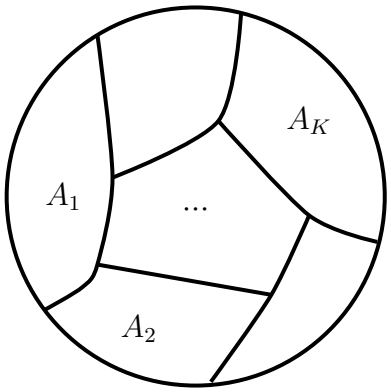

Figure 2: The space $\mathcal{Y}$ is partitioned to K regions at the detector

4.1 Detection Error

We wish the detector to be correct most of the times, if not all the time. However, since random phenomenons are involved we may not be able to guarantee the correctness every time we run the experiment. So, we should target to reduce the average probability of error. In a physical sense, one imagines repeating the experiment of predicting θ, and looks for the empirical average the event $\mathbb{1}_{\{\hat{\theta}(Y) \neq \theta\}}$. There are two kinds of randomness involved in this outcome.

1. The way θ is chosen every time. We assume independent experiments with the probability law $P(\theta = i) = \pi(i), 1 \le i \le K$. For example, if only a subset of S is chosen every time, our detector need not consider the rest for possible outputs.

2. The noise phenomenon Z. As an example, in communication systems, depending on the conditions, additive noise may *over-power* the signal, and nothing will be received correctly.

We define the error probability as

$$P(E) = P(\hat{\theta}(Y) \neq \theta).$$

We will use $\hat{\theta}$ instead of $\hat{\theta}(Y)$ from now onward. Using Baye's rule,

$$P(E) = \sum_{i=1}^{K} P(\theta = i, \theta \neq \hat{\theta}) \tag{7}$$

$$= \sum_{i=1}^{K} \pi(i) P(Y \notin A_i | \theta = i). \tag{8}$$

We wish to minimize $P(E)$, which is same as maximizing $1 - P(E)$.

$$1 - P(E) = 1 - \sum_{i=1}^{K} \pi(i) P(Y \notin A_i | \theta = i) \tag{9}$$

$$= \sum \pi(i) \Big(1 - P(Y \notin A_i | \theta = i)\Big) \tag{10}$$

$$= \sum \pi(i) \int_{A_i} f_i(y) dy, \tag{11}$$

where $f_i(y)$ is the conditional density $f_{Y|\theta}(y|i)$ (see previous section). Let us now try to maximize the probability of correct detection.

$$\sum_{i=1}^{K} \pi(i) \int_{A_i} f_i(y) dy = \sum_{i=1}^{K} \pi(i) \int \mathbb{1}_{\{y \in A_i\}} f_i(y) dy \tag{12}$$

$$= \int \sum_{i=1}^{K} \pi(i) f_i(y) \mathbb{1}_{\{y \in A_i\}} dy \tag{13}$$

$$\le \int \sum_{i=1}^{K} \Big(\max_{1 \le j \le K} \pi(j) f_j(y) \Big) \mathbb{1}_{\{y \in A_i\}} dy \tag{14}$$

$$= \int \Big(\max_{1 \le j \le K} \pi(j) f_j(y) \Big) \Big(\sum_{i=1}^{K} \mathbb{1}_{\{y \in A_i\}} \Big) dy \tag{15}$$

$$= \int \max_{1 \le j \le K} \pi(j) f_j(y) dy \tag{16}$$

We will now introduce the best decoder which minimizes the probability of error, known as the **MAP** decoder (Maximum Aposteriori Probability Decoder). To this end, for $1 \le k \le K$, consider the sets given by

$$\hat{A}_k = \{y : \pi(k) f_k(y) > \pi(j) f_j(y), \forall j \neq k\}. \tag{17}$$

Clearly $\hat{A}_k$ are non-overlapping subsets of $\mathcal{Y}$. Imagine a detector employing the sets $\hat{A}_k, 1 \le k \le K$ to decide $\hat{\theta}$. Let us also assign

$$\hat{\theta}(y) = 1 \text{ if } \{y \notin \hat{A}_k, 1 \le k \le K\},$$

to take care of the remaining elements in $\mathcal{Y}$, which does not fall into any $\hat{A}_k$. This is due to the strict inequality in (17), but will not impact our probability computations[1]. The probability of correct detection in such a detector is, by (11)

$$1 - P(E) = \sum_{k=1}^{K} \int_{\hat{A}_k} \pi(k) f_k(y) dy \tag{18}$$

$$= \sum_{k=1}^{K} \int_{\hat{A}_k} \big(\max_{1 \le j \le K} \pi(j) f_j(y) \big) dy \tag{19}$$

$$= \int \big(\max_{1 \le j \le K} \pi(j) f_j(y) \big) dy. \tag{20}$$

The second equality follows since $\pi(k) f_j(y) \ge \pi(j) f_j(y), \forall y \in \hat{A}_k$. Comparing this result to (16), it is clear that the probability of correct detection by any other scheme can at best be the same as that of a decoder which uses $\hat{A}_k, 1 \le k \le K$ as the decoding regions. Such a decoder is called the MAP decoder.

Theorem 1. *The MAP decoding scheme minimizes the error probability $P(\hat{\theta} \neq \theta)$.*

Example 2. *Let $\theta \in \{-3, -1, +1, +3\}$ and the observation be given by $Y = \theta + Z$, where $Z \perp\!\!\!\perp \theta$. Let $Z \sim \mathcal{N}(0, \sigma^2)$. If each parameter is equally likely to be chosen, find the minimum probability of error in detection.*

[1] feel free to ignore those elements outside $\bigcup \hat{A}_k$, whenever Y admits a density