

# 1 Elementary Probability

## 1.1 Lecture 1

This lecture I did not have a laptop yet, so I was unable to transcribe anything. Here is what I remember was discussed:

**Definition 1.1 (Conditional Probability)** For two events  $A, B$ , the probability of  $A$  given  $B$  is:

$$\mathbb{P}[A | B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}$$

**Definition 1.2 (MAP)** We say the Most likely A Posteriori (MAP) estimate of a random variable  $X$  given  $Y = y$ , is:

$$\operatorname{argmax}_x \mathbb{P}[X = x | Y = y]$$

**Definition 1.3 (MLE)** We say the Maximum Likelihood Estimate (MLE) of a random variable  $X$  given  $Y$ , is:

$$\operatorname{argmax}_x \mathbb{P}[Y = y | X = x]$$

## 1.2 Lecture 2

### 1.2.1 Probability Fundamentals and Random Variables

We begin probability by defining a set  $\Omega$  called the sample space. Elements of the sample space are termed outcomes. Subsets of  $\Omega$  are termed as events.

For some event  $A$ , we can define the probability of  $A$  as follows:

$$\mathbb{P}[A] = \sum_{\omega \in A} \mathbb{P}[\omega]$$

Probability maps events to  $[0, 1]$  in a consistent manner, satisfying the following axioms:

- $\mathbb{P}[\Omega] = 1$
- $\mathbb{P}[\emptyset] = 0$
- For two disjoint events  $A_1, A_2$ , we have  $\mathbb{P}[A_1 \cup A_2] = \mathbb{P}[A_1] + \mathbb{P}[A_2]$

This is what we term a probability space. Often it is more helpful to work with events than with individual sample points (especially in the case of an uncountably infinite amount of sample points).

**Definition 1.4 (Random Variable)** A random variable  $X : \omega \rightarrow B$  maps each outcome to elements of some other set (often  $\mathbb{R}$ ).  $X = x$  for some  $x$  is an event, with a well-defined probability.

**Definition 1.5 (Independence)** Two random variables  $X$  and  $Y$  are independent if

$$\mathbb{P}[X = x | Y = y] = \mathbb{P}[X = x]$$

i.e.

$$\mathbb{P}[X = x, Y = y] = \mathbb{P}[X = x] \mathbb{P}[Y = y]$$

**Example 1.1** Suppose you flip a coin 10 times. We will show that  $X$ , the amount of heads in the first 4 flips, and  $Y$ , the amount of heads in the last 6 flips, are independent.

Let  $a(x)$  be the amount of ways to get  $x$  heads in 4 flips and  $b(y)$  be the amount of ways to get  $y$  heads in 6 flips. Then,

$$\begin{aligned} \mathbb{P}[X = x] &= \frac{a(x) \cdot 2^6}{2^{10}} = \frac{a(x)}{2^4} \\ \mathbb{P}[Y = y] &= \frac{b(y) \cdot 2^4}{2^{10}} = \frac{b(y)}{2^6} \\ \mathbb{P}[X = x, Y = y] &= \frac{a(x) \cdot b(y)}{2^{10}} = \mathbb{P}[X = x] \cdot \mathbb{P}[Y = y] \end{aligned}$$

Thus, the random variables are independent.

**Definition 1.6 (Expectation)** The expectation of a (discrete) random variable is:

$$\mathbb{E}[X] = \sum_x x \mathbb{P}[X = x]$$

This is often called the mean or the average value.

**Theorem 1.1** Properties of expectation:

- $\mathbb{E}[a] = a$  for  $a \in \mathbb{R}$
- If the space is uniform, then  $\mathbb{E}[X] = \frac{1}{N} \sum_x x$
- $\mathbb{E}[\alpha X + \beta Y] = \alpha \mathbb{E}[X] + \beta \mathbb{E}[Y]$  for  $\alpha, \beta \in \mathbb{R}$
- $X \leq Y \Rightarrow \mathbb{E}[X] \leq \mathbb{E}[Y]$
- If  $X$  and  $Y$  are independent, then  $\mathbb{E}[XY] = \mathbb{E}[X] \mathbb{E}[Y]$
- $\mathbb{E}[XY] = \mathbb{E}[X] \mathbb{E}[Y] \nRightarrow X, Y$  independent

There are two ways of thus computing expectation. You can either sum over sample points, or take a lot of measurements of your random variable, then divide by the amount of measurements. The reason this works is because of property 2 above.

**Definition 1.7 (Variance and Standard Deviation)** The variance of a random variable  $X$  is defined as:

$$\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$$

The standard deviation of this random variable is:

$$\sigma_X = \sqrt{\text{Var}(X)}$$

The variance measures the spread away from the mean that a random variable may exhibit.

**Theorem 1.2** Properties of variance:

- $\text{Var}(X) \geq 0$ , with equality only if  $X$  is constant
- $\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$
- $\text{Var}(aX) = a^2 \text{Var}(X)$  for constant  $a$
- If  $X$  and  $Y$  are independent,  $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$
- In general,  $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) - 2\text{Cov}(X, Y)$

Often we term  $\mathbb{E}[X^k]$  as the  $k$ th moment of  $X$ , so the variance contains information about the second moment of  $X$ .

## 1.3 Lecture 3

### 1.3.1 Concentration Inequalities

**Definition 1.8 (Indicator Random Variable)**  $\mathbb{1}_A$  is the indicator for event  $A$ , i.e. a random variable with the following values:

$$\mathbb{1}_A = \begin{cases} 1 & \text{if sample point in event } A \\ 0 & \text{otherwise} \end{cases}$$

**Theorem 1.3 (Markov's Inequality)** Consider random variable  $X \geq 0$  and constant  $a > 0$ . Then,

$$\mathbb{P}[X \geq a] \leq \frac{\mathbb{E}[X]}{a}$$

**Proof** Let  $Y = \mathbb{1}_{X \geq a}$ . Then we know:

$$\begin{aligned} \mathbb{E}[Y] &= 0 \cdot \mathbb{P}[X < a] + 1 \cdot \mathbb{P}[X \geq a] = \mathbb{P}[X \geq a] \\ Y &\leq \frac{X}{a} \\ \mathbb{E}[Y] &\leq \mathbb{E}\left[\frac{X}{a}\right] = \frac{\mathbb{E}[X]}{a} \\ \mathbb{P}[X \geq a] &\leq \frac{\mathbb{E}[X]}{a} \end{aligned}$$

Markov's inequality tends to be a coarse bound, and  $X, a$  have to be non-negative.

**Theorem 1.4 (Chebyshev's Inequality)** For random variable  $X$  and  $\epsilon > 0$ :

$$\mathbb{P}[|X - \mathbb{E}[X]| \geq \epsilon] \leq \frac{\text{Var}(X)}{\epsilon^2}$$

Define  $Z = |X - \mathbb{E}[X]|^2$ ,  $a = \epsilon^2$ ,  $\epsilon > 0$ .

**Proof** Apply Markov's inequality:

$$\mathbb{P}[Z \geq \epsilon^2] \leq \frac{\mathbb{E}[Z]}{\epsilon^2}$$

Note that

$$\mathbb{E}[Z] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \text{Var}(X)$$

This means that

$$\mathbb{P}[\sqrt{Z} \geq \epsilon] \leq \frac{\text{Var}(X)}{\epsilon^2}$$

which is exactly the statement of Chebyshev's.

Chebyshev's is generally a tighter bound than Markov's.

**Theorem 1.5 (Weak Law of Large Numbers)** Assume  $X_1, X_2, X_3, \dots$  are independent random variables with the same expectation  $\mu$  and the same variance  $\sigma^2$ , and define  $Y_n = \frac{(X_1 + X_2 + \dots + X_n)}{n}$ . Then, we have that for any

constant  $\epsilon > 0$ .

$$\lim_{n \rightarrow \infty} \mathbb{P}[|Y_n - \mu| \geq \epsilon] \rightarrow 0$$

This can be shown by Chebyshev's inequality, namely note that the expression in the limit is bounded by  $\frac{\text{Var}(Y_n)}{\epsilon^2} = \frac{n\sigma^2}{n^2\epsilon^2} \rightarrow 0$ .

In words, this means the probability of the sample mean being within  $\epsilon$  of the true mean approaches 1.

### 1.3.2 Covariance and Estimation

#### Definition 1.9 (Covariance)

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

If  $X$  and  $Y$  are independent, then  $\text{Cov}(X, Y) = 0$ . If the latter is true, then  $X$  and  $Y$  are uncorrelated. We also define the coefficient of correlation:

$$\rho_{xy} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

This is handy because  $|\rho_{XY}| \leq 1$ .

Suppose we want to estimate a random variable  $Y$  by  $\hat{Y}$  given a correlated random variable  $X$ .

We want to minimize  $\mathbb{E}[(Y - \hat{Y})^2]$  but have a linear relationship. This yields LLSE:

#### Theorem 1.6 (Linear Least Squares Estimate) The LLSE

$$\hat{Y} = \mathbb{E}[Y] + \frac{\text{Cov}(X, Y)}{\text{Var}(X)}(X - \mathbb{E}[X])$$

is the best linear estimate of  $Y$  given  $X$ .

## 2 Basic Probability

### 2.1 Lecture 3, Continued

#### 2.1.1 Infinite Collections of Events and Borel-Cantelli

There are some important consequences of the axioms of probability.

**Theorem 2.1 (Infinite Sub-Events)** Consider some set  $A$  where  $A = \bigcup_{n=1}^{\infty} A_n$  where:

$$A_1 \subseteq A_2 \subseteq \dots$$

Then,  $\mathbb{P}[A_n] \rightarrow \mathbb{P}[A]$ .

Furthermore, consider some set  $B$  where  $B = \bigcap_{n=1}^{\infty} B_n$  where:

$$B_1 \supseteq B_2 \supseteq \dots$$

Then,  $\mathbb{P}[B_n] \rightarrow \mathbb{P}[B]$ .

**Theorem 2.2 (Borel-Cantelli Theorem)** Let  $\{A_n\}_{n=1}^{\infty}$  be a collection of events such that  $\sum_{n=1}^{\infty} \mathbb{P}[A_n] < \infty$ , then  $\mathbb{P}[A_n \text{ infinitely often}] = 0$ .  
 $\{A_n \text{ infinitely often}\}$  is the following event:

$$\{\omega \mid \exists N(\omega) \text{ such that } \forall n > N(\omega), \omega \notin A_n\}$$

In English, the set describes all outcomes where you can assign a number to that outcome such that after  $A_N$ , you have no set membership.

The theorem claims that you CAN assign such a number (max event) to any outcome with nonzero probability.

**Proof** Define

$$B_n = \bigcup_{m \geq n} A_m$$

Note that  $B_1 \supseteq B_2 \supseteq B_3 \dots$

Calling,

$$B = \bigcap_n B_n = B_n$$

Note that  $\omega \in \{A_n \text{ io}\}$  if and only if  $\omega \in B_n$  for all  $n$ . But this means that  $\omega \in B$ . This means that  $\{A_n \text{ io}\} = B$ . So we must calculate  $\mathbb{P}[B]$ . However, note that  $\mathbb{P}[B_n] \rightarrow \mathbb{P}[B]$  as  $n \rightarrow \infty$ . Thus, we must compute:

$$\begin{aligned} \mathbb{P}[B] &= \lim_{n \rightarrow \infty} \mathbb{P}[B_n] \\ \mathbb{P}[B_n] &\leq \sum_{m=n}^{\infty} \mathbb{P}[A_m] \rightarrow 0 \\ \mathbb{P}[B] &= 0 \end{aligned}$$

The second step is justified by the following result from analysis. For non-negative sequence  $a_n$ , if  $\sum_{i=1}^{\infty} a_n < \infty$ , then  $\lim_{n \rightarrow \infty} \sum_{m=n}^{\infty} a_m \rightarrow 0$ .

Our result shows that  $\mathbb{P}[A_n \text{ io}] = 0$  ■

Consider the following example for coin flips:

**Example 2.1 (Infinite Coin Flips)** Consider the experiment of flipping a coin infinitely many times. Let

$$A_n = \{n\text{th flip is heads}\}$$

Then, in this experiment, the event  $\{A_n \text{ infinitely often}\}$  (which we denote as  $\{A_n \text{ io}\}$ )

$$\{A_n \text{ io}\} = \{\omega \mid \text{heads never stop after some } N(\omega)\}$$

Here are some sequences that are in that event:

$$\begin{aligned} \omega &= 0, 0, 1, 1, 1, 1, \dots \\ \omega &= 0, 1, 0, 1, 0, 1, \dots \\ \omega &= 0, 0, \underbrace{\dots}_{1 \text{ million 0's}}, 1, 0, 0, \underbrace{\dots}_{1 \text{ million 0's}}, 1, \dots \end{aligned}$$

Now consider the assigning the following probabilities to each heads (instead of the normal, uniform probability space):

$$\mathbb{P}[A_n] = \frac{1}{n^2}$$

$\sum_{n=1}^{\infty} \frac{1}{n^2}$  converges, so by Borel-Cantelli,  $\mathbb{P}[A_n \text{ io}] = 0$ , i.e. the heads ALWAYS stop.