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:

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

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

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

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

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

1.2 Lecture 2

1.2.1 Probability Fundamentals and Random Variables

We begin probability by defining a set Ω called the sample space. Elements of the sample space are termed outcomes. Subsets of Ω 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 $X : \omega \to B$ 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

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

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
- $\text{Var}(X) = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2$
- $\text{Var}(aX) = a^2\text{Var}(X)$ 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 $\mathbb{E}\left[X^k\right]$ as the kth 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 $\epsilon > 0$:

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

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

Proof Apply Markov's inequality:

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

Note that

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

This means that

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

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 $X_1, X_2, X_3, \ldots$ are independent random variables with the same expectation μ and the same variance σ^2, and define $Y_n = \frac{(X_1 + X_2 + \cdots + X_n)}{n}$ Then, we have that for any constant $\epsilon > 0$.

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

This can be shown by Chebyshev's inequality, namely note that the expression in the limit is bounded by $\frac{\mathrm{Var}(Y_n)}{\epsilon^2} = \frac{n\sigma^2}{n^2\epsilon^2} \to 0$.
In words, this means the probability of the sample mean being within ϵ of the true mean approaches 1.

1.3.2 Covariance and Estimation

Definition 1.9 (Covariance)

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

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

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

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 Y.

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

Theorem 1.6 (Linear Least Squares Estimate) The LLSE

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

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:

$$A_1 \subseteq A_2 \subseteq \ldots$$

Then, $\mathbb{P}\left[A_n\right] \to \mathbb{P}\left[A\right]$.
Furthermore, consider some set B where $B = \bigcap_{n=1}^{\infty} B_n$ where:

$$B_1 \supseteq B_2 \supseteq \ldots$$

Then, $\mathbb{P}\left[B_n\right] \to \mathbb{P}\left[A\right]$.

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 $\mathbb{P}[B_n] \to \mathbb{P}[B]$ as $n \to \infty$. Thus, we must compute:

$$\mathbb{P}[B] = \lim_{n\to\infty} \mathbb{P}[B_n]$$
$$\mathbb{P}[B_n] \leq \sum_{m=n}^{\infty} \mathbb{P}[A_m] \to 0$$
$$\mathbb{P}[B] = 0$$

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 $\lim_{n\to\infty} \sum_{m=n}^{\infty} a_m \to 0$.
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 $\{A_n$ 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:

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

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.

Now there is one more question. Does $\mathbb{P}[A_n \text{ i.o.}] = 0 \implies \{A_n \text{i.o.}\} = 0$? The answer is no. In this case, $\mathbb{P}[A_n \text{i.o.}] = 0$, but consider the outcome ω_n where the nth flip onwards is a heads; these are all in the infinitely often set, so it actually has infinite cardinality!

2.2 Lecture 4

2.2.1 The Laws of Large Numbers, Revisited

Here is a recap of the two different laws of large numbers.

First, we define two different types of convergence:

Definition 2.1 (Almost Sure Convergence) A random variable A almost surely converges to constant b if

$$\mathbb{P}[A \to b] = 1$$

as $n \to \infty$.

Definition 2.2 (Convergence in Probability) A random variable A converges in probability to constant b if

$$\mathbb{P}[|A - b| < \epsilon \to 1]$$

for any real number $\epsilon > 0$.

Theorem 2.3 (Strong Law of Large Numbers) Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed (iid) random variables. Define:

$$Y_n = \frac{X_1 + \cdots + X_n}{n}$$
$$Y = \mathbb{E}[X_1]$$

Y_n converges to Y almost surely.

Note the contrast with the weak law of natural numbers. The weak law had only convergence in probability. A key thing to note is that the strong law **implies** the weak law.

2.2.2 Independence

Let us now refine the notion of Independence.

Definition 2.3 (Pairwise Independence) Consider events A_j with $j \in J$.
The events are pairwise independent if for any $j, k \in J$,

$$\mathbb{P}\left[A_j \cap A_k\right] = \mathbb{P}\left[A_j\right] \mathbb{P}[A_k]$$

Definition 2.4 (Mutual Independence) Consider events A_j with $j \in J$.
The events are mutually independent if

$$\mathbb{P}\left[\bigcap_{j \in K} A_j\right] = \prod_{j \in K} \mathbb{P}\left[A_j\right], \forall K \subseteq J$$

Note that pairwise independence does not imply mutual independence. Here is an example of that edge case:

Example 2.2 Take probability space $\Omega = \{1, 2, 3, 4\}$, all equally likely. Consider the events: $A = \{1, 2\}, B = \{1, 3\}, C = \{1,4\}$.
Note that $\mathbb{P}[A \cap B] = \frac{1}{4} = \mathbb{P}[A]\,\mathbb{P}[B]$,
but $\mathbb{P}[A \cap B \cap C] = \frac{1}{4} \neq \frac{1}{8} = \mathbb{P}[A]\,\mathbb{P}[B]\,\mathbb{P}[C]$.

Now with independence, we can find that the converse of Borel-Cantelli is often true:

Theorem 2.4 (Converse of Borel-Cantelli Theorem) Let A_n be a collection of mutually indpendent events such that $\sum_{n=1}^{\infty} \mathbb{P}[A_n] = \infty$. Then, $\mathbb{P}[A_n \text{ infinitely often}] = 1$.

Let us use another example to understand this:

Example 2.3 Let A_n be the same event as the other example (the nth flip is heads) and assign:

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

where all the A_n are mutually independent.
Since $\sum_{n=1}^{\infty} \frac{1}{n} = \infty$ and thus by the converse of Borel-Cantelli: $\mathbb{P}[A_n \text{ io}] = 1$.

Example 2.4 (Glued Coins) Suppose you have n coins that are all glued together, i.e. the only two outcomes are $HHH\ldots$ or $TTT\ldots$. Then let A_n = the nth coin is heads. Note that

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

Theorem 2.5 (Kolmogorov's 0-1 theorem) If you have a set of events $\{A_n\}_{n=1}^{\infty}$ that all independent, then

$$\mathbb{P}[A_n \text{ infinitely often}] = 0 \text{ or } 1$$

2.2.3 Conditional Probability

Now we refine conditional probability for many events.

Definition 2.5 (Conditional Probability) Let A and B be two events, and assume $\mathbb{P}[B] > 0$. Then the conditional probability of A given B is:

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

Theorem 2.6 (Chain Rule) For two events we had $\mathbb{P}[A \cap B] = \mathbb{P}[A \mid B]\,\mathbb{P}[B]$. For n events A_i, we have:

$$\mathbb{P}[A_1 \cap A_2 \cap \cdots \cap A_n] = \mathbb{P}[A_1]\,\mathbb{P}[A_2 \mid A_1]\,\mathbb{P}[A_3 \mid A_1 \cap A_2]\ldots\mathbb{P}[A_n \mid A_1 \cap A_2 \cap \cdots \cap A_{n-1}]$$

if $\mathbb{P}[A_1 \cap A_2 \cap \ldots A_{n-1}] > 0$.

The generalized result above can be shown by induction, taking the case of two events as the base case and then inducting on n. Now we will bring in some of the most powerful tools.

Theorem 2.7 (Law of Total Probability) Let $A, B_1, \dots B_n$ be events where B_i's are disjoint and $\bigcup_{i=1}^n B_i = \Omega$. Then,

$$\mathbb{P}[A] = \sum_{i=1}^{n} \mathbb{P}[A \cap B_i]$$

Theorem 2.8 (Bayes' Rule) Let $A, B_1, \dots B_n$ be events where B_i's are disjoint and $\bigcup_{i=1}^n B_i = \Omega$.

$$\mathbb{P}[B_i \mid A] = \frac{\mathbb{P}[A \mid B_i]\,\mathbb{P}[B_i]}{\sum_{j=1}^{n} \mathbb{P}\left[A \mid B_j\right] \mathbb{P}\left[B_j\right]}$$

Proof Note that we can use the initial definition to expand the left side:

$$\begin{aligned}
\mathbb{P}[B_i \mid A] &= \frac{\mathbb{P}[A \cap B_i]}{\mathbb{P}[A]} \\
&= \frac{\mathbb{P}[A \mid B_i]\,\mathbb{P}[B_i]}{\sum_{j=1}^{n} \mathbb{P}\left[A \cap B_j\right]} \\
&= \frac{\mathbb{P}[A \mid B_i]\,\mathbb{P}[B_i]}{\sum_{j=1}^{n} \mathbb{P}\left[A \mid B_j\right] \mathbb{P}\left[B_j\right]}
\end{aligned}$$

where the summation in the denominator comes from the law of total probability.

Often the B_j's are termed the prior probabilities, and A is considered the posterior probability.

For an event $B \subseteq R$, $\mathbb{P}[X \in B] = \mathbb{P}\left[(X^{-1}(B))\right]$ where

$$X^{-1}(B) = \{\omega \in \Omega \mid X(\omega) \in B\}$$

.

We can define the following for a random variable to the reals.

Definition 2.6 (Cumulative Distribution Function (CDF)) The Cumulative Distribution Function $F_X(x)$ of random variable X is defined by:

$$F_X(x) = \mathbb{P}[X \in (-\infty, x]] = \mathbb{P}[X \leq x]$$

Here are some properties of the CDF:

- F_X is non-decreasing.
- F_X is right-continuous.
- $F_X \to 0$ as $x \to -\infty$ and $F_X \to 1$ as $x \to \infty$.

Example 2.5 (CDF of an Indicator) Consider the following random variable:

$$I = \begin{cases} 0 \text{ with probability } 1 - p \\ 1 \text{ with probability } p \end{cases}$$

Then the $F_I(i)$ is a step function: TODO Add figure

2.3 Lecture 5

Definition 2.7 (Discrete Random Variable) A discrete random variable X can be described fully by:

$$\{(x_n, p_n), n = 1, \ldots, N\}$$

where $p_n = \mathbb{P}[X = x_n]$. This is called the probability mass function (PMF) of X.

We can write the expectation as follows:

$$\mathbb{E}[X] = \sum_{n=1}^{N} x_n p_n$$

With $N = \infty$, the expectation may not be defined.

Definition 2.8 (Function of a Random Variable) Calling $h(X)$ means changing to another random variable with the following PMF:

$$(h(x_n), p_n), n = 1, \ldots, N$$

The expectation of this is as follows:

$$\mathbb{E}[h(X)] = \sum_{n=1}^{N} h(x_n) p_n$$

Definition 2.9 (Coefficient of Variation) The coefficient of variation c of X is defined:

$$c = \sigma_X / \mathbb{E}[X]$$

2.4 Common Discrete Distributions

Bernoulli random variables model situations like individual coin flips.

Definition 2.10 (Bernoulli Random Variables) If $X =_D B(p)$ with $p \in [0, 1]$, then the PMF of X is:

$$\{(0, 1 - p), (1, p)\}$$

Furthermore, $\mathbb{E}[X] = p$ and $\text{Var}(X) = p(1 - p)$.

Geometric random variables model the situation where you count the number of coin flips until you get "heads".

Definition 2.11 (Geometric Random Variable) If $X =_D G(p)$ with $p \in [0, 1]$, then the PMF of X is:

$$\mathbb{P}[X = n] = (1 - p)^{n-1} p$$

Furthermore, $\mathbb{E}[X] = \frac{1}{p}$ and $\text{Var}(X) = \frac{1-p}{p^2}$.

The CDF can also be derived as $\mathbb{P}[X \leq n] = 1 - (1 - p)^n$, since it's the complement of failing n times. The CCDF (Complementary CDF) is thus $\mathbb{P}[X < n] = (1 - p)^n$.

Note 2.1 (Memoryless Property) The geometric distribution is memoryless, i.e. if $X =_D G(p)$, then

$$\mathbb{P}[X > m + n \mid X > m] = \mathbb{P}[X > n]$$

Binomial random variables model the situation of doing n coin flips and counting the heads, or the sum of n i.i.d. Bernoulli random variables.

Definition 2.12 (Binomial Random Variable) If $X =_D B(N, p)$ with $p \in [0, 1]$ and $N \geq 1$, then the PMF of X is:

$$\mathbb{P}[X = n] = \binom{N}{n} p^n (1-p)^{N-n}$$

Furthermore, $\mathbb{E}[X] = Np$ and $\mathrm{Var}(X) = Np(1-p)$

The mode of the binomial distribution (the maximum probability) is at $n = \lfloor p(N+1) \rfloor$.

Poisson random variables are the limit of the binomials as the rate of coin flips goes to infinity. This represents the number of successes in an interval during a continuous process.

Definition 2.13 (Poisson Random Variable) If $X =_D P(\lambda)$ with $\lambda > 0$, then the PMF of X is:

$$\mathbb{P}[X = n] = \frac{e^{-\lambda} \lambda^n}{n!}$$

Furthermore, $\mathbb{E}[X] = \lambda$ and $\mathrm{Var}(X) = \lambda$.

In fact, we can make this limit more precise.

Theorem 2.9 (Binomial Converges to Poisson) We have, setting $Np = \lambda$, where λ is fixed,

$$B(N, \lambda/N) \to P(\lambda)$$

2.5 Multiple Discrete Random Variables

Consider a pair of random variables (X, Y).

Definition 2.14 (Joint PMF) The joint distribution is given by:

$$p_{i,j} = \mathbb{P}\left[X = x_i, Y = y_j\right]$$

To find the PMF of one of the variables from the joint distribution, we can

Note 2.2 (Marginal PMF from JPMF)

$$\mathbb{P}[X = x_i] = \sum_j \mathbb{P}\left[X = x_i, Y = y_j\right]$$

Furthermore,

Theorem 2.10 (Independence for Random Variables) X and Y are independent if and only if

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

If you have a function of multiple random variables, you can apply it similarly to the one variable case.

$$\mathbb{E}[h(X, Y)] = \sum_i \sum_j h(x_i, y_j) \mathbb{P}\left[X = x_i, Y = y_j\right]$$

First we extend the idea of conditioning to random variables.

Definition 2.15 (Conditional PMF) We call the conditional distribution of Y given X as:

$$\mathbb{P}\left[Y = y_j \mid X = x_i\right] = \frac{\mathbb{P}\left[X = x_i, Y = y_j\right]}{\mathbb{P}\left[X = x_i\right]}$$

Definition 2.16 (Conditional Expectation) The expectation of Y given X (i.e. the best guess of Y given X) is denoted $\mathbb{E}\left[Y \mid X\right]$ and is a function of X

Furthermore, if we want to use a function, we can compute it as follows:

$$\mathbb{E}\left[h(Y) \mid X = x_i\right] = \sum_j h(y_j)\mathbb{P}\left[Y = y_j \mid X = x_i\right]$$

Theorem 2.11 (Properties of Conditional Expectation) For two random variables X, Y,

$$\begin{aligned}
\mathbb{E}\left[\mathbb{E}\left[Y \mid X\right]\right] &= \mathbb{E}\left[Y\right] \\
\mathbb{E}\left[h(X)Y \mid X\right] &= h(X)\mathbb{E}\left[Y \mid X\right] \\
\mathbb{E}\left[Y \mid X\right] &= \mathbb{E}\left[Y\right] \text{ if } X \text{ and } Y \text{ are independent} \\
\mathbb{E}\left[h_1(Y) + h_2(Y) \mid X\right] &= \mathbb{E}\left[h_1(Y) \mid X\right] + \mathbb{E}\left[h_2(Y) \mid X\right]
\end{aligned}$$