CS 70 Discrete Mathematics and Probability Theory
Fall 2009 Satish Rao,David Tse Lecture 16

Multiple Random Variables and Applications to Inference

In many probability problems, we have to deal with *multiple* r.v.'s defined on the same probability space. We have already seen examples of that when we saw, for example, that computing the expectation and variance of a binomial r.v. X, it is easier to write it as a sum $X = \sum_{i=1}^{n} X_i$ where X_i represents the result of the i'th trial. In inference problems, where we observe certain quantities and use the information to infer about other hidden quantities, multiple r.v.'s arise naturally in the modeling of the situation. We will see some examples of such problems after we go through some of the basics in the handling of multiple r.v.'s.

Joint Distributions

Consider two random variables X and Y defined on the same probability space. By linearity of expectation, we know that $\mathrm{E}(X+Y) = \mathrm{E}(X) + \mathrm{E}(Y)$. Since $\mathrm{E}(X)$ can be calculated if we know the distribution of X and $\mathrm{E}(Y)$ can be calculated if we know the distribution of Y, this means that $\mathrm{E}(X+Y)$ can be computed knowing only the two individual distributions. No information is needed about the *relationship* between X and Y. This is not true if we need to compute, say, $\mathrm{E}((X+Y)^2)$, e.g. as when we computed the variance of a binomial r.v. This is because $\mathrm{E}((X+Y)^2) = \mathrm{E}(X^2) + 2\mathrm{E}(XY) + \mathrm{E}(Y^2)$, and $\mathrm{E}(XY)$ depends on the relationship between X and Y. How can we capture such a relationship?

Recall that the distribution of a single random variable X is the collection of the probabilities of all events $X = a$, for all possible values of a that X can take on. When we have two random variables X and Y, we can think of (X,Y) as a "two-dimensional" random variable, in which case the events of interest are $X = a \wedge Y = b$ for all possible values of (a,b) that (X,Y) can take on. Thus, a natural generalization of the notion of distribution to multiple random variables is the following.

Definition 16.1 (joint distribution): The joint distribution of two discrete random variables X and Y is the collection of values $\{(a,b,\Pr[X = a \wedge Y = b]) : (a,b) \in \mathscr{A} \times \mathscr{B}\}$, where $\mathscr{A}$ and $\mathscr{B}$ are the sets of all possible values taken by X and Y respectively.

This notion obviously generalizes to three or more random variables. Since we will write $\Pr[X = a \wedge Y = b]$ quite often, we will abbreviate it to $\Pr[X = a, Y = b]$.

Just like the distribution of a single random variable, the joint distribution is *normalized*, i.e.

$$\sum_{a \in \mathscr{A}, b \in \mathscr{B}} \Pr[X = a, Y = b] = 1.$$

This follows from noticing that the events $X = a \wedge Y = b$, $a \in \mathscr{A}, b \in \mathscr{B}$, partition the sample space.

The joint distribution between two random variables fully describe their statistical relationships, and provides enough information for computing any probabilities and expectations involving the two random variables. For example,

$$\mathrm{E}(XY) = \sum_c c \times \Pr[XY = c] = \sum_a \sum_b ab \times \Pr[X = a, Y = b].$$

Y \ X	0	1	2
0	0.1	0.2	0.15
1	0.05	0.05	0.2
2	0.1	0.1	0.05

Figure 1: A tabular representation of a joint distribution.

More generally, if f is any function on $\mathbf{R} \times \mathbf{R}$,

$$\mathrm{E}(f(X,Y)) = \sum_c c \times \Pr[f(X,Y) = c] = \sum_a \sum_b f(a,b) \times \Pr[X = a, Y = b].$$

Moreover, the individual distributions of X and Y can be recovered from the joint distribution as follows:

$$\Pr[X = a] = \sum_{b \in \mathscr{B}} \Pr[X = a, Y = b] \qquad \forall a \in \mathscr{A}, \tag{1}$$

$$\Pr[Y = b] = \sum_{a \in \mathscr{A}} \Pr[X = a, Y = b] \quad \forall b \in \mathscr{B}. \tag{2}$$

The first follows from the fact that the events $Y = b$, $b \in \mathscr{B}$, form a partition of the sample space Ω, and so the events $X = a \wedge Y = b$, $b \in \mathscr{B}$ are disjoint and their union yields the event $X = a$. Similar logic applies to the second fact.

Pictorially, one can think of the joint distribution values as entries filling a table, with the columns indexed by the values that X can take on and the rows indexed by the values Y can take on (Figure 1). To get the distribution of X, all one needs to do is to sum the entries in each of the columns. To get the distribution of Y, just sum the entries in each of the rows. This process is sometimes called *marginalization* and the individual distributions are sometimes called *marginal* distributions to differentiate them from the joint distribution.

Independent Random Variables

Independence for random variables is defined in analogous fashion to independence for events:

Definition 16.2 (independent r.v.'s): Random variables X and Y on the same probability space are said to be *independent* if the events $X = a$ and $Y = b$ are independent for all values a, b. Equivalently, the joint distribution of independent r.v.'s decomposes as

$$\Pr[X = a, Y = b] = \Pr[X = a] \Pr[Y = b] \quad \forall a, b.$$

Note that for independent r.v.'s, the joint distribution is fully specified by the marginal distributions.

Mutually independence of more than two r.v.'s is defined similarly. A very important example of independent r.v.'s is indicator r.v.'s for independent events. Thus, for example, if $\{X_i\}$ are indicator r.v.'s for the ith coin toss being Heads (as in example 2 in the last lecture note) then the X_i are mutually independent r.v.'s.

We saw that the expectation of a sum of r.v.'s is the sum of the expectations of the individual r.v.'s. This is not true in general for variance. However, it turns out to be true if the random variables are independent. To

see that, first we look at the expectation of a product of independent r.v.'s (which is a quantity that frequently shows up in variance calculations, as we have seen).

Theorem 16.1: *For independent random variables X,Y, we have* $\mathrm{E}(XY) = \mathrm{E}(X)\mathrm{E}(Y)$.

Proof: We have

$$\begin{aligned}
\mathrm{E}(XY) &= \sum_a \sum_b ab \times \Pr[X = a, Y = b] \\
&= \sum_a \sum_b ab \times \Pr[X = a] \times \Pr[Y = b] \\
&= \left(\sum_a a \times \Pr[X = a]\right) \times \left(\sum_b b \times \Pr[Y = b]\right) \\
&= \mathrm{E}(X) \times \mathrm{E}(Y),
\end{aligned}$$

as required. In the second line here we made crucial use of independence. □

For example, this theorem would have allowed us to conclude immediately in our random walk example at the beginning of the last lecture note that $\mathrm{E}(X_i X_j) = \mathrm{E}(X_i)\mathrm{E}(X_j) = 0$, without the need for a calculation.

We now use the above theorem to conclude a nice property of the variance of independent random variables.

Theorem 16.2: *For independent random variables X,Y, we have* $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.

Proof: From the alternative formula for variance in Theorem 15.1, we have, using linearity of expectation extensively,

$$\begin{aligned}
\mathrm{Var}(X+Y) &= \mathrm{E}((X+Y)^2) - \mathrm{E}(X+Y)^2 \\
&= \mathrm{E}(X^2) + \mathrm{E}(Y^2) + 2\mathrm{E}(XY) - (\mathrm{E}(X) + \mathrm{E}(Y))^2 \\
&= (\mathrm{E}(X^2) - \mathrm{E}(X)^2) + (\mathrm{E}(Y^2) - \mathrm{E}(Y)^2) + 2(\mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y)) \\
&= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2(\mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y)).
\end{aligned}$$

Now *because X,Y are independent*, by Theorem 16.1 the final term in this expression is zero. Hence we get our result. □

Note: The expression $\mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y)$ appearing in the above proof is called the *covariance* of X and Y, and is a measure of the dependence between X,Y. It is zero when X,Y are independent.

Theorem 16.2 can be used to simplify several of our variance calculations in the last lecture. E.g., in example 1 of the last lecture note, since the X_i are independent r.v.'s with $\mathrm{Var}(X_i) = 1$ for each i, we have $\mathrm{Var}(X) = \mathrm{Var}(\sum_{i=1}^n X_i) = \sum_{i=1}^n \mathrm{Var}(X_i) = n \times 1 = n$. And in example 2 the X_i are independent with $\mathrm{Var}(X_i) = p(1-p)$, so we have $\mathrm{Var}(X) = \mathrm{Var}(\sum_{i=1}^n X_i) = \sum_{i=1}^n \mathrm{Var}(X_i) = np(1-p)$. Note, however, that we don't get any simplification in example 4 because the X_i are *not* independent.

It is very important to remember that **neither** Theorem 16.1 **nor** Theorem 16.2 is true in general, without the assumption that X,Y are independent. As a simple example, note that even for a 0-1 r.v. X with $\Pr[X = 1] = p$, $\mathrm{E}(X^2) = p$ is not equal to $\mathrm{E}(X)^2 = p^2$ (because of course X and X are not independent!).

Note also that Theorem 16.2 does not quite say that variance is *linear* for independent random variables: it says only that variances sum. It is *not* true that $\mathrm{Var}(cX) = c\mathrm{Var}(X)$ for a constant c. In fact, the following is true:

Theorem 16.3: *For any random variable X and constant c,* $\mathrm{Var}(cX) = c^2\mathrm{Var}(X)$.

The proof is left as a straightforward exercise.

Conditional Distribution and Expectation

In an earlier lecture, we discussed the concept of conditional probability of an event A given an event B. This concept allows us to define a notion of *conditional distribution* of a random variable given another random variable.

Definition 16.3 (conditional distribution): The conditional distribution of X given $Y = b$ is the collection of values $\{(a, \Pr[X = a|Y = b]) : a \in \mathscr{A}\}$, where $\mathscr{A}$ is the set of all possible values taken by X.

The conditional distribution can be calculated from the joint and marginal distributions :

$$\Pr[X = a|Y = b] = \frac{\Pr[X = a, Y = b]}{\Pr[Y = b]}.$$

It follows from eqn. (2) that

$$\sum_{a \in \mathscr{A}} \Pr[X = a|Y = b] = 1,$$

so the conditional distribution is normalized, just like a (unconditional) distribution. Note that if X and Y are independent r.v.'s, $\Pr[X = a|Y = b] = \Pr[X = a]$ for every a, b, i.e. the conditional and unconditional distributions of X are the same.

One can also naturally talk about the conditional distribution of multiple random variables. For example, the conditional distribution of X and Y given $Z = c$ is simply given by

$$\{(a, b, \Pr[X = a, Y = b|Z = c]) : a \in \mathscr{A}, b \in \mathscr{B}\}.$$

Conditional distributions have exactly the same properties as (unconditional) distributions. Therefore whatever we do with distributions we can do with conditional distributions. For example, one can compute expectations of conditional distributions. This leads to the concept of *conditional expectation*.

Definition 16.4 (conditional expectation): Let X and Y be two r.v.'s defined on the same probability space. The conditional expectation of X given $Y = b$ is defined to be:

$$\mathrm{E}(X|Y = b) := \sum_{a \in \mathscr{A}} a \times \Pr[X = a|Y = b].$$

Conditional probabilities often help us to calculate probabilities of event by means of the total probability law. Similarly, conditional expectations are often useful to compute expectations via the *total expectation law*:

$$\begin{aligned}
\mathrm{E}(X) &= \sum_{a \in \mathscr{A}} a \Pr[X = a] \\
&= \sum_{a \in \mathscr{A}} a \sum_{b \in \mathscr{B}} \Pr[Y = b] \Pr[X = a|Y = b] \\
&= \sum_{b \in \mathscr{B}} \Pr[Y = b] \sum_{a \in \mathscr{A}} a \Pr[X = a|Y = b] \\
&= \sum_{b \in \mathscr{B}} \Pr[Y = b] \times \mathrm{E}(X|Y = b).
\end{aligned}$$

This formula is quite intuitive: to calculate the expectation of the r.v. X, first calculate the conditional expectation of X given each of the various values of Y. Then sum them, weighted by the probabilities Y takes on the various values. The formula is another instantiation of the *divide into cases* strategy.

For example, suppose in the coin flipping example above, we are interested in calculating the expectation of the number of flips (say Z) of the randomly chosen coin until we see the first Head. Using the total expectation rule,

$$\mathrm{E}(Z) = \sum_{i=1}^{n} \Pr[X = i]\mathrm{E}(Z|X = i) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{p_i}.$$

The last step follows from the fact that conditional on the identity of the coin $X = i$, Z is geometrically distributed with parameter p_i.

For a slightly more interesting example, let us use the total expectation law to give an alternative way of computing the expectation of a geometrically distributed r.v. X. (In lecture note 14, we did it one way.) Recall that X is the number of independent trials until we get our first success. Let Y be the indicator r.v. of the event that the first trial is successful. Using the total expectation law,

$$\mathrm{E}(X) = \Pr[Y = 1]\mathrm{E}(X|Y = 1) + \Pr[Y = 0]\mathrm{E}(X|Y = 0) = p\mathrm{E}(X|Y = 1) + (1 - p)\mathrm{E}(X|Y = 0). \quad (3)$$

Now, if $Y = 1$, the first trial is already successful, and $X = 1$ with probability 1. Hence, $\mathrm{E}(X|Y = 1) = 1$. What about if $Y = 0$? If the first trial is unsuccessful, we are back to square one and have to continue trying. Hence the number of additional trials after the first trial is another geometric r.v. with the same parameter p, and $\mathrm{E}(X|Y = 0) = 1 + \mathrm{E}(X)$. Substituting into (3), we get:

$$\mathrm{E}(X) = p + (1 - p)\{1 + \mathrm{E}(X)\}.$$

Upon solving this equation, we get $\mathrm{E}(X) = \frac{1}{p}$.

It is interesting to see that while linearity of expectation was a useful tool to compute the expectation of a binomially distributed r.v., the total expectation rule is a more natural tool to compute the expectation of a geometrically distributed r.v. Both are tools that allow us to compute expectations without directly computing distributions.

Inference

One of the major uses of probability is to provide a systematic framework to perform *inference under uncertainty*. A few specific applications are:

- **communications**: Information bits are sent over a noisy physical channel (wireless, DSL phone line, etc.). From the received symbols, one wants to make a decision about what bits are transmitted.
- **control**: A spacecraft needs to be landed on the moon. From noisy measurements by motion sensors, one wants to estimate the current position of the spacecraft relative to the moon surface so that appropriate controls can be applied.
- **object recognition**: From an image containing an object, one wants to recognize what type of object it is.
- **speech recognition**: From hearing noisy utterances, one wants to recognize what is being said.
- **investing**: By observing past performance of a stock, one wants to estimate its intrinsic quality and hence make a decision on whether and how much to invest in it.

All of the above problems can be modeled with the following ingredients:

- a random variable X representing the hidden quantity not directly observed but in which one is interested. X can be the value of an information bit in a communication scenario, position of the spacecraft in the control application, or the object class in the recognition problem.
- random variables $Y_1, Y_2, \ldots Y_n$ representing the observations. They may be the outputs of a noisy channel at different times, pixel values of an image, values of the stocks on successive days, etc.
- The distribution of X, called the *prior* distribution. This can be interpreted as the knowledge about X *before* seeing the observations.
- The conditional distribution of $Y_1, \ldots Y_n$ given X. This models the noise or randomness in the observations.

Since the observations are noisy, there is in general no hope of knowing what the *exact* value of X is given the observations. Instead, all knowledge about X can be summarized by the *conditional distribution* of X given the observations. We don't know what the exact value of X is, but the conditional distribution tells us what values of X is more likely and which are less likely. Based on this information, intelligent decisions can be made.

Inference Example 1: Multi-arm Bandits

Question: You walk into a casino. There are several slot machines (bandits). You know some have odds very favorable to you, some have less favorable odds, and some have very poor odds. However, you don't know which are which. You start playing on some them, and by observing the outcomes, you want to learn which is which so that you can intelligently figure out which machine to play on (or not play at all, which may be the most intelligent decision.)

Stripped-down version: Suppose there are n biased coins. Coin i has probability p_i of coming up Heads; however, you don't know which is which. You randomly pick one coin and flip it. If the coin comes up Heads you win \$1, and if it comes up Tails you lose \$1. What is the probability of winning? What is the probability of winning on the next flip given you have observed a Heads with this coin? Given you have observed two Heads in a row? Would you bet on the next flip?

Modeling using Random Variables

Let X be the coin randomly chosen, and Y_i be the indicator r.v. for the event that the ith flip of this randomly chosen coin comes up Head. Since we don't know which coin we have chosen, X is the hidden quantity. The Y_i's are the observations.

Predicting the First Flip

The first question asks for $\Pr[Y_1 = 1]$. First we calculate the joint distribution of X and Y_1:

$$\Pr[X = i, Y_1 = 1] = \Pr[X = i] \Pr[Y_1 = 1 | X = i] = \frac{p_i}{n} \tag{4}$$

Applying (2), we get:

$$\Pr[Y_1 = 1] = \sum_{j=1}^{n} \Pr[X = j, Y_1 = 1] = \frac{1}{n} \sum_{j=1}^{n} p_j. \tag{5}$$

Note that combining the above two equations, we are in effect using the fact that:

$$\Pr[Y_1 = 1] = \sum_{i=1}^{n} \Pr[X = i] \Pr[Y_1 = 1 | X = i]. \tag{6}$$

This is just the *total probability rule* for events applied to random variables. Once you get familiar with this type of calculation, you can bypass the intermediate calculation of the joint distribution and directly write this down.

Predicting the Second Flip after Observing the First

Now, given that we observed $Y_1 = 1$, we are learning something about the randomly chosen coin X. This knowledge is captured by the conditional distribution:

$$\Pr[X = i | Y_1 = 1] = \frac{\Pr[X = i, Y_1 = 1]}{\Pr[Y_1 = 1]} = \frac{p_i}{\sum_{j=1}^{n} p_j},$$

using eqns. (4) and (5).

Note that when we substitute eqn. (4) into the above equation, we are in effect using:

$$\Pr[X = i | Y_1 = 1] = \frac{\Pr[X = i] \Pr[Y_1 = 1 | X = i]}{\Pr[Y_1 = 1]}.$$

This is just Bayes' rule for events applied to random variables. Just like for events, this rule has the interpretation of updating knowledge based on the observation: $\{(i, \Pr[X = i]), i = 1, \ldots, n\}$ is the *prior distribution* of the hidden X; $\{(i, \Pr[X = i | Y_1 = 1]) : i = 1, \ldots, n\}$ is the *posterior* distribution of X given the observation. Bayes' rule updates the prior distribution to yield the posterior distribution

Now we can calculate the probability of winning using this coin in the second flip:

$$\Pr[Y_2 = 1 | Y_1 = 1] = \sum_{j=1}^{n} \Pr[X = j | Y_1 = 1] \Pr[Y_2 = 1 | X = j, Y_1 = 1]. \tag{7}$$

This can be interpreted as the total probability rule (6) but in a new probability space with all the probabilities under the additional condition $Y_1 = 1$. You are asked to verify this formula from first principles.

Now let us calculate the various probabilities on the right hand side of (7).The probability $\Pr[X = j | Y_1 = 1]$ is just the posterior distribution of X given the observation. We have already calculated it. What about the probability $\Pr[Y_2 = 1 | X = j, Y_1 = 1]$? There are two conditioning events: $X = j$ and $Y_1 = 1$. But here is the thing: once we know that the unknown coin is coin j, then knowing the first flip is a Head is redundant and provides no further statistical information about the outcome of the second flip: the probability of getting a Heads on the second flip is just p_j. In other words,

$$\Pr[Y_2 = 1 | X = j, Y_1 = 1] = \Pr[Y_2 = 1 | X = j] = p_j. \tag{8}$$

The events $Y_1 = 1$ and $Y_2 = 1$ are said to be independent *conditional* on the event $X = j$. Since in fact $Y_1 = a$ and $Y_2 = b$ are independent given $X = j$ for all a, b, j, we will say that the *random variables* Y_1 and Y_2 are independent given the random variable X.

Definition 16.5 (Conditional Independence): Two events A and B are said to be *conditionally* independent given a third event C if

$$\Pr[A \wedge B | C] = \Pr[A | C] \times \Pr[B | C].$$

Two random variables X and Y are said to be independent given a third random variable Z if for every a, b, c,

$$\Pr[X = a, Y = b | Z = c] = \Pr[X = a | Z = c] \times \Pr[Y = b | Z = c].$$

Note that the r.v.'s Y_1 and Y_2 are *not* independent. Knowing the outcome of Y_1 tells us some information about the identity of the coin (X) and hence allows us to infer something about Y_2. However, if we already know X, then the outcomes of the different flips Y_1 and Y_2 are independent.

Now substituting (8) into (7), we get the probability of winning using this coin in the second flip:

$$\Pr[Y_2 = 1 | Y_1 = 1] = \sum_{j=1}^{n} \Pr[X = j | Y_1 = 1] \Pr[Y_2 = 1 | X = j] = \frac{\sum_{j=1}^{n} p_j^2}{\sum_{j=1}^{n} p_j}.$$

Predicting the Third Flip After Observing the First Two

Using Bayes' rule and the total probability rule, we can compute the posterior distribution of X given that we observed two Heads in a row:

$$\begin{aligned}
\Pr[X = j | Y_1 = 1, Y_2 = 1] &= \frac{\Pr[X = j] \Pr[Y_1 = 1, Y_2 = 1 | X = j]}{\Pr[Y_1 = 1, Y_2 = 1]} \\
&= \frac{\Pr[X = j] \Pr[Y_1 = 1, Y_2 = 1 | X = j]}{\sum_{i=1}^{n} \Pr[X = i] \Pr[Y_1 = 1, Y_2 = 1 | X = i]} \\
&= \frac{\Pr[X = j] \Pr[Y_1 = 1 | X = j] \Pr[Y_2 = 1 | X = j]}{\sum_{i=1}^{n} \Pr[X = i] \Pr[Y_1 = 1 | X = i] \Pr[Y_2 = 1 | X = i]} \\
&= \frac{p_j^2}{\sum_{i=1}^{n} p_i^2}
\end{aligned}$$

The probability of getting a win on the third flip using the same coin is:

$$\begin{aligned}
\Pr[Y_3 = 1 | Y_1 = 1, Y_2 = 1] &= \sum_{j=1}^{n} \Pr[X = j | Y_1 = 1, Y_2 = 1] \Pr[Y_3 = 1 | X = j, Y_1 = 1, Y_2 = 1] \\
&= \sum_{j=1}^{n} \Pr[X = j | Y_1 = 1, Y_2 = 1] \Pr[Y_3 = 1 | X = j] \\
&= \frac{\sum_{j=1}^{n} p_j^3}{\sum_{j=1}^{n} p_j^2}
\end{aligned}$$

Suppose $n = 3$ and the three coins have biased probabilities $p_1 = 2/3, p_2 = 1/2, p_3 = 1/5$. The conditional distributions of X after observing no flip, one Heads and two Heads in a row are shown in Figure 2. Note that as more Heads are observed, the conditional distribution is increasingly concentrated on coin 1 with $p_1 = 2/3$: we are increasingly certain that the coin chosen is the good coin. The corresponding probabilities of winning on the next flip after observing no flip, 1 Heads and two Heads in a row are $0.46, 0.54$ and 0.58 respectively. The conditional probability of winning gets better and better.

Inference Example 2: Communication over a Noisy Channel

Question: I have one bit of information that I want to communicate over a noisy channel. The noisy channel flips each one of my transmitted symbols independently with probability $p < 0.5$. How much improvement in performance do I get by repeating my transmission n times?

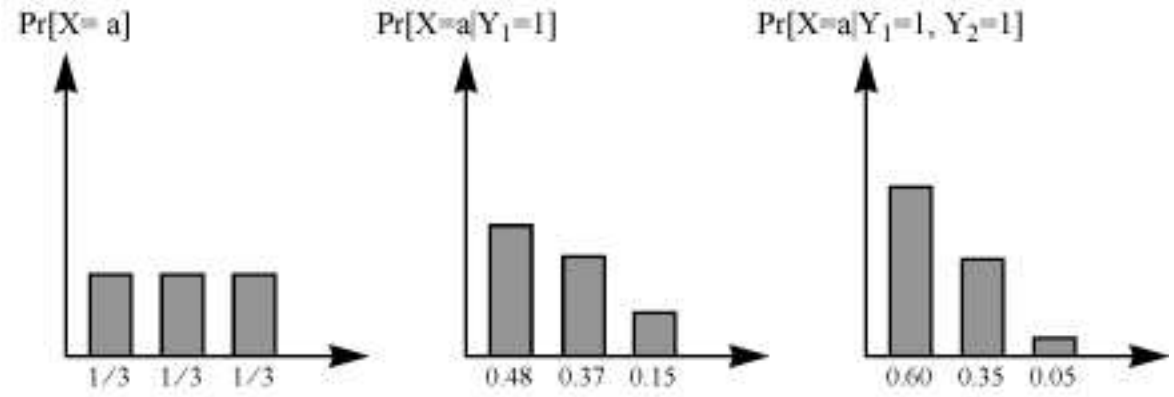

Figure 2: The conditional distributions of X given no observations, 1 Heads, and 2 Heads.

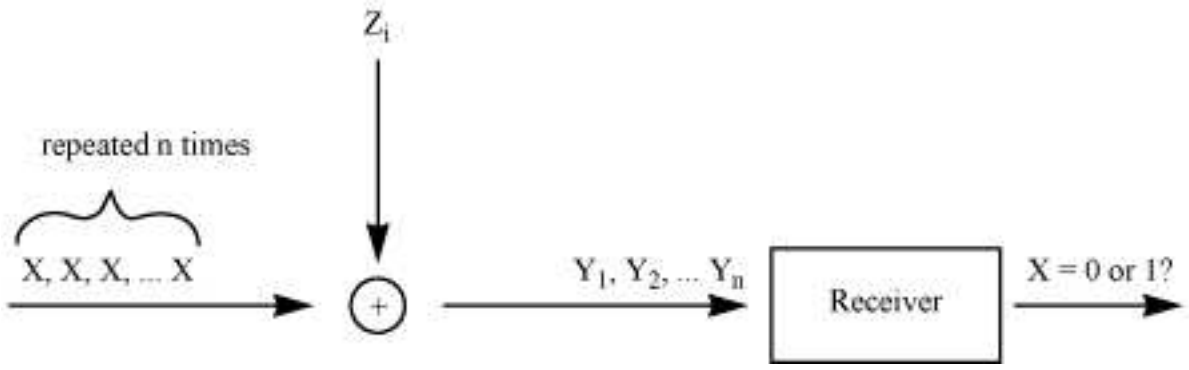

Figure 3: The system diagram for the communication problem.

Comment: In an earlier lecture note, we also considered a communication problem and gave some examples of error-correcting codes. However, the models for the communication channel are different. There, we put a bound on the maximum number of flips the channel can make. Here, we do not put such bounds *a priori* but instead imposes a probabilistic model. Since there is no bound on the maximum number of flips the channel can make, there is no guarantee that the receiver will always decode correctly. Instead, one has to be satisfied with being able to decode correctly with high probability, eg., probability of error < 0.01.

Modeling

The situation is shown in Figure 3.

Let X ($= 0$ or 1) be the value of the information bit I want to transmit. Assume that X is equally likely to be 0 or 1. The received symbol on the ith repetition of X is

$$Y_i = X + Z_i \mod 2, \qquad i = 1, 2, \ldots, n$$

with $Z_i = 1$ with probability p and $Z_i = 0$ with probability $1 - p$. Note that Y_i is different from X if and only if $Z_i = 1$. Thus, the transmitted symbol is flipped with probability p. The Z_i's are assumed to be mutually independent across different repetition of X and also independent of X. The Z_i's can be interpreted as *noise*.

Note that the received symbols Y_i's are not independent; they all contain information about the transmitted bit X. However, given X, they are independent since they then only depend on the noise Z_i's.

Decision Rule

First, we have to figure out what *decision rule* to use at the receiver, i.e. given each of the 2^n possible received sequences, $Y_1 = b_1, Y_2 = b_2, \ldots Y_n = b_n$, how should the receiver guess what value of X was transmitted?

A natural rule is the *maximum a posteriori* (MAP) rule: guess the value a^* for which the conditional probability of $X = a^*$ given the observations is the largest among all a. More explicitly:

$$a^* = \begin{cases} 0 & \text{if } \Pr[X=0|Y_1=b_1,\ldots,Y_n=b_n] \geq \Pr[X=1|Y_1=b_1,\ldots Y_n=b_n] \\ 1 & \text{otherwise} \end{cases}$$

Now, let's make some simplifications to this rule. By Bayes' rule,

$$\Pr[X=0|Y_1=b_1,\ldots Y_n=b_n] = \frac{\Pr[X=0]\Pr[Y_1=b_1,\ldots,Y_n=b_n|X=0]}{\Pr[Y_1=b_1,\ldots,Y_n=b_n]} \tag{9}$$

$$= \frac{\Pr[X=0]\Pr[Y_1=b_1|X=0]\Pr[Y_2=b_2|X=0]\ldots\Pr[Y_n=b_n|X=0]}{\Pr[Y_1=b_1,\ldots,Y_n=b_n]} \tag{10}$$

In the second step, we are using the fact that the observations Y_i's are conditionally independent given X. (Why?) Similarly,

$$\Pr[X=1|Y_1=b_1,\ldots Y_n=b_n] = \frac{\Pr[X=1]\Pr[Y_1=b_1,\ldots,Y_n=b_n|X=1]}{\Pr[Y_1=b_1,\ldots,Y_n=b_n]} \tag{11}$$

$$= \frac{\Pr[X=1]\Pr[Y_1=b_1|X=1]\Pr[Y_2=b_2|X=1]\ldots\Pr[Y_n=b_n|X=1]}{\Pr[Y_1=b_1,\ldots,Y_n=b_n]} \tag{12}$$

An equivalent way of describing the MAP rule is that it computes the ratio of these conditional probabilities and checks if it is greater than or less than 1. If it is greater than 1, then guess that a 0 was transmitted; otherwise guess that a 1 was transmitted. (This ratio indicates how likely a 0 is compared to a 1, and is called the *likelihood ratio*.) Dividing (10) and (12), the likelihood ratio L is:

$$L = \prod_{i=1}^{n} \frac{\Pr[Y_i=b_i|X=0]}{\Pr[Y_i=b_i|X=1]}. \tag{13}$$

Note that we didn't have to compute $\Pr[Y_1 = b_1,\ldots,Y_n = b_n]$, since it appears in both of the conditional probabilities and got canceled out when computing the ratio.

Now,

$$\frac{\Pr[Y_i=b_i|X=0]}{\Pr[Y_i=b_i|X=1]} = \begin{cases} \frac{p}{1-p} & \text{if } b_i = 1 \\ \frac{1-p}{p} & \text{if } b_i = 0 \end{cases}$$

In other words, L has a factor of $p/(1-p) < 1$ for every 1 received and a factor of $(1-p)/p > 1$ for every 0 received. So the likelihood ratio L isgreater than 1 if and only if the number of 1's is less than the number of 0's. Thus, the decision rule is simply a *majority* rule: guess that a 0 was transmitted if the number of 0's in the received sequence is more than the number of 1's and vice versa.

Note that in deriving this rule, we assumed that $\Pr[X = 0] = \Pr[X = 1] = 0.5$. When the prior distribution is not uniform, the MAP rule is no longer a simple majority rule. You are asked to derive the MAP rule in the general case in the exercises.

Error Probability Analysis

What is the probability that the guess is incorrect? This is just the event E that the number of flips by the noisy channel is greater than or equal to $n/2$. (This is a slight upper bound since one could be correct when there are $n/2$ flips under some model of how one guesses an answer if there is a tie.) So the error probability of our majority rule is:

$$\Pr[E] = \Pr[\sum_{i=1}^{n} Z_i \geq \frac{n}{2}] = \sum_{k=\lceil n/2 \rceil}^{n} \binom{n}{k} p^k (1-p)^{n-k},$$

recognizing that the random variable $S := \sum_{i=1}^{n} Z_i$ has a binomial distribution with parameters n and p.

This gives an expression for the error probability that can be numerically evaluated for given values of n. Given a target error probability of, say, 0.01, one can then compute the smallest number of repetitions needed to achieve the target error probability.[1]

As in the hashing and load balancing applications we looked at earlier in the course, we are interested in a more explicit relationship between n and the error probability to get a better intuition of the problem. The above expression is too cumbersome for this purpose. Instead, notice that $n/2$ is greater than the mean np of S and hence the error event is related to the tail of the distribution of S. One can therefore apply Chebyshev's inequality in the last lecture note to bound the error probability:

$$\Pr[S > \frac{n}{2}] < \Pr[|S - pn| > n(\frac{1}{2} - p)] \leq \frac{\mathrm{Var}(S)}{n^2(\frac{1}{2}-p)^2} = \frac{p(1-p)}{(\frac{1}{2}-p)^2} \cdot \frac{1}{n}$$

The important thing to note is that the error probability decreases with n, so indeed by repeating more times, the performance improves (as one would expect!). For a given target error probability of say 0.01, one needs to repeat no more than

$$100 \cdot \frac{p(1-p)}{(\frac{1}{2}-p)^2}$$

times. For $p = 0.25$, this evaluates to about 300.

In the exercises, you are asked to compare the bound with the actual error probability. You will see that the bound is not very good, and actually one can repeat much fewer times to get an error probability of 0.01. In an upper-division course like EECS 126, you can learn about much better bounds.

[1] Needless to say, one does not want to repeat more times than is necessary as we are using more time to transmit each information bit and the rate of communication is slowed down.