# 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 which values of $X$ are more likely and which are less likely. Based on this information, intelligent decisions can be made.

# Inference Example 1: Multi-armed 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 of 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 $Y_j$ be the indicator r.v. for the event that the $j$th flip of this randomly chosen coin comes up Heads. Since we don't know which coin we have chosen, $X$ is the hidden quantity. The $Y_j$'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 $Y_1$:

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

[Note: We are abusing notation here by writing "$Y_1 = H$" rather than "$Y_1 = 1$" for the event that the first coin toss comes up Heads. We are doing this to make things clearer, even though strictly speaking a random variable should take on only real values.]

Applying (**??**), we get:

$$\Pr[Y_1 = H] = \sum_{i=1}^{n} \Pr[X = i, Y_1 = H] = \frac{1}{n} \sum_{i=1}^{n} p_i. \tag{2}$$

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

$$\Pr[Y_1 = H] = \sum_{i=1}^{n} \Pr[X = i] \Pr[Y_1 = H | X = i]. \tag{3}$$

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 down equation (3).

### Predicting the second flip after observing the first

Now, given that we observed $Y_1 = H$, we have learned something about the randomly chosen coin $X$. This knowledge is captured by the conditional distribution

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

using eqns. (1) and (2).

Note that when we substitute eqn. (1) into the above equation, we are in effect using:

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

This is just Bayes' rule for events applied to random variables. Just as 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 $X$; $\{(i, \Pr[X = i|Y_1 = H]) : i = 1, \ldots, n\}$ 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 same coin in the second flip:

$$\Pr[Y_2 = H|Y_1 = H] = \sum_{i=1}^{n} \Pr[X = i|Y_1 = H]\Pr[Y_2 = H|X = i, Y_1 = H]. \tag{4}$$

This can be interpreted as the total probability rule (3) but in a new probability space *with all the probabilities under the additional condition* $Y_1 = H$. You should try to verify this formula from first principles.

Now let us calculate the various probabilities on the right hand side of (4). The probability $\Pr[X = i|Y_1 = H]$ is just the posterior distribution of $X$ given the observation, which we have already calculated above. What about the probability $\Pr[Y_2 = H|X = i, Y_1 = H]$? There are two conditioning events: $X = i$ and $Y_1 = H$. But here is the thing: once we know that the unknown coin is coin $i$, 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 $p_i$. In other words,

$$\Pr[Y_2 = H|X = i, Y_1 = H] = \Pr[Y_2 = H|X = i] = p_i. \tag{5}$$

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

**Definition 18.1 (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 *conditionally independent* given a third random variable $Z$ if for every $a, b, c$,

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

Going back to our coin example, note that the r.v.'s $Y_1$ and $Y_2$ are definitely *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 (5) into (4), we get the probability of winning using this coin in the second flip:

$$\Pr[Y_2 = H|Y_1 = H] = \sum_{i=1}^{n} \Pr[X = i|Y_1 = H]\Pr[Y_2 = H|X = i] = \frac{\sum_{i=1}^{n} p_i^2}{\sum_{i=1}^{n} p_i}.$$

It can be shown (using the Cauchy-Schwarz inequality) that $n\sum_i p_i^2 \geq (\sum_i p_i)^2$, which implies that

$$\Pr[Y_2 = H|Y_1 = H] = \frac{\sum_{i=1}^{n} p_i^2}{\sum_{i=1}^{n} p_i} \geq \frac{\sum_{i=1}^{n} p_i}{n} = \Pr[Y_1 = H].$$

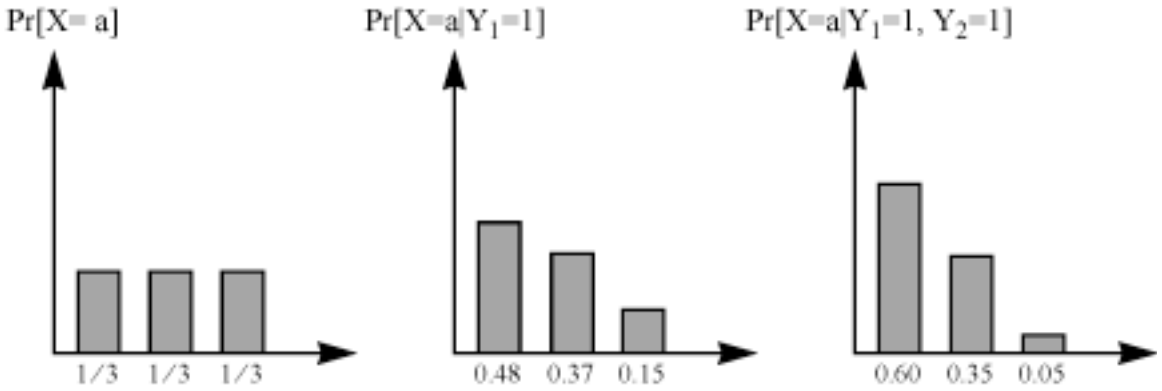


Figure 1: The conditional distributions of $X$ given no observations, 1 Heads, and 2 Heads.

Thus our observation of a Heads on the first flip increases the probability that the second toss is Heads. This, of course, is intuitively reasonable, because the posterior distribution puts larger weight on the coins with larger values of $p_i$.

**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:

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

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

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

Again, it can be shown that $\frac{\sum_{i=1}^n p_i^3}{\sum_{i=1}^n p_i^2} \geq \frac{\sum_{i=1}^n p_i^2}{\sum_{i=1}^n p_i}$, so the probability of seeing another Heads on the next flip has again increased. If we continue this process further (conditioning on having seen more and more Heads), the probability of Heads on the next flip will keep increasing towards the limit $p_{\max} = \max_i p_i$.

As a numerical illustration, suppose $n = 3$ and the three coins have Heads 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

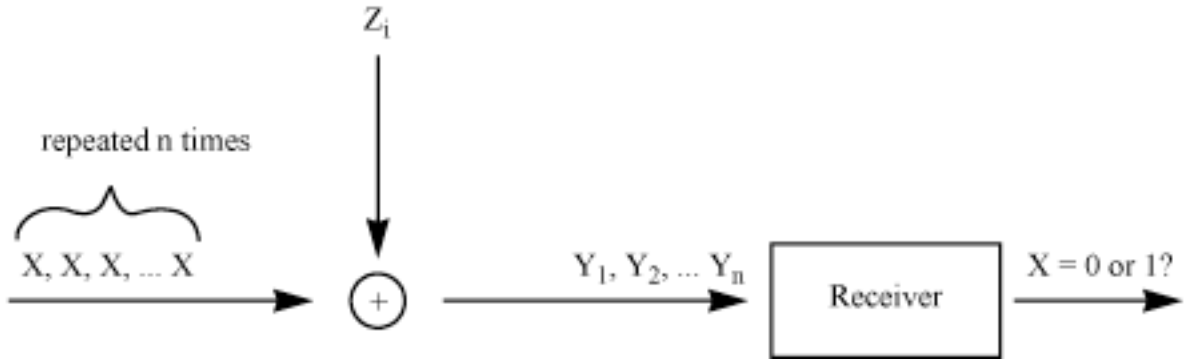


Figure 2: The system diagram for the communication problem.

are shown in Figure 1. 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 best coin. The corresponding probabilities of winning on the next flip after observing no flip, one 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 (approaching 2/3 in the limit).

# 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?

**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 impose a bound on the *probability* that each bit is flipped (so that the *expected* number of bits flipped is $np$). 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*, e.g., probability of error $< 0.01$.

### Modeling

The situation is shown in Figure 2.

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 (this is the prior). The received symbol on the $i$th 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 repetitions 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 (conditionally) independent since they then only depend on the noise $Z_i$.

**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 reformulate this rule so that it looks cleaner. By Bayes' rule, we have

$$\Pr[X=0|Y_1=b_1,\ldots Y_n=b_n] = \frac{\Pr[X=0]\Pr[Y_1=b_1,\ldots,Y_n=b_n|X=0]}{\Pr[Y_1=b_1,\ldots,Y_n=b_n]} \quad (6)$$

$$= \frac{\Pr[X=0]\Pr[Y_1=b_1|X=0]\Pr[Y_2=b_2|X=0]\ldots\Pr[Y_n=b_n|X=0]}{\Pr[Y_1=b_1,\ldots,Y_n=b_n]} \quad (7)$$

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

$$\Pr[X=1|Y_1=b_1,\ldots Y_n=b_n] = \frac{\Pr[X=1]\Pr[Y_1=b_1,\ldots,Y_n=b_n|X=1]}{\Pr[Y_1=b_1,\ldots,Y_n=b_n]} \quad (8)$$

$$= \frac{\Pr[X=1]\Pr[Y_1=b_1|X=1]\Pr[Y_2=b_2|X=1]\ldots\Pr[Y_n=b_n|X=1]}{\Pr[Y_1=b_1,\ldots,Y_n=b_n]} \quad (9)$$

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 (or equal to) 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 (7) and (9), and recalling that we are assuming $\Pr[X=1] = \Pr[X=0]$, the likelihood ratio $L$ is:

$$L = \prod_{i=1}^{n} \frac{\Pr[Y_i=b_i|X=0]}{\Pr[Y_i=b_i|X=1]}. \quad (10)$$

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 $L$ is greater than 1 if and only if the number of 0's is greater than the number of 1'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 at least as large as the number of 1's, otherwise guess that a 1 was transmitted.

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 as an exercise (see HW10).

**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 $n/2$. So the error probability of our majority rule is:

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

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 application 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 to bound the error probability:

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

using the fact that $\mathrm{Var}(S) = n\mathrm{Var}(Z_i) = np(1-p)$. 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

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

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

In HW10, you are asked to compare the bound with the actual error probability. You will see that the bound is rather pessimistic, and actually one can repeat many fewer times to get an error probability of 0.01. In an upper-division course such as CS 174 or EECS 126, you can learn about much better bounds on error probabilities like this.

[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.