Chapter 2

Probability review

2.1 Probability models and random variables

A probability model is used to describe random phenomenon that can have non-deterministic **outcomes**, where we call the set of all outcomes as the **sample space** and the "undetermine" object itself as a **random variable** (r.v.). If the sample space is continuous, the random variable is continuous. Otherwise, the random variable is discrete.

The probability of an **outcome** depicts the relative chance of getting that outcome. For a random variable X, we often denote the probability of X taking outcome a as $Pr(X = a)$. By convention, a probability is always non-negative and a probability of zero means that the outcome will never happen. On the other hand, a probability of one indicates that the outcome will certainly happen. So by definition, all possible outcomes should sum up to one as at least one of the outcome will certainly happen.

Example: Coin toss

Let's try to model a coin toss with a probability model. Let's denote the random variable as X and the outcomes **head** and **tail** as H and T, respectively.
Then the sample space is $\{H, T\}$. The probability of the entire space should sum up to 1. Thus $Pr(X = H) + Pr(X = T) = 1$.

2.1.1 A word on convention

We often denote a r.v. using upper case (such as X) and its realization (what was actually observed) using lower case (such as x). Therefore, $Pr(x = x_0)$ is a bad notation since x is not random and $Pr(x = x_0) = 0$ in general unless x and x_0 turn out to be identical.

2.2 Probability distributions

We often call the probabilities of all outcomes from the sampling space the probability distribution. Note that strictly speaking, this definition is only true for discrete r.v.. Because for continuous r.v., the probability of any outcomes is generally zero and so such definition is meaningless. Please see below for clarification.

2.2.1 Probability mass function

We often call the probability distribution of a discrete r.v. as probability mass function (PMF). For example, for the coin toss example described earlier, we may have $Pr(X = H) = Pr(X = T) = 0.5$ for a fair coin. It is quite wordy to write with the notation $Pr(X = x)$. Instead, we often denote $p_X(x) \triangleq Pr(X = x)$. And when the context is clear, we often simply the notation further to just $p(x)$.

2.2.2 Probability density function

For a continuous r.v. X, the probability of X equal to any arbitrary value is generally zero as mentioned above. For example, consider X as continuous r.v. uniformly distributed between 0 and 1. $Pr(X = 0.5) = 0$ since we can always argue that we didn't get a 0.5 no matter how close X really was. Maybe, X is 0.500001 rather than 0.5. By this argument, $Pr(X = x) = 0$ for any x. A fix for this is instead of trying to define a function that maps to the probability of an outcome. We define a function where the area underneath the curve is the probability instead. More precisely, we define

$$f(x) = \frac{1}{\Delta} \lim_{\Delta \to 0^+} Pr(x \leq X \leq x + \Delta) \tag{2.1}$$

and $f(x)$ will then be known as the PDF.

And note that from the definition above,

$$Pr(x \leq X \leq x + \Delta) \approx f(x)\Delta$$

for a small Δ. And thus

$$Pr(a \leq X \leq b) = \int_a^b f(x)dx$$

where the integral is just the area underneath $f(x)$ between a and b. Moreover, since X has to take some value in the real axis,

$$Pr(-\infty \leq X \leq \infty) = \int_{-\infty}^{\infty} f(x)dx = 1 \tag{2.2}$$

Note that since we can interpret PDF just as a normalized probability function, $f(x) \geq 0$ just as the original definition of the probability.

Example: PDF of the uniform distribution

Let's take a r.v. X that is uniformly distributed between 0 and 1 as an example. The PDF is simply

$$P_X(x) = \begin{cases} 1, & 0 \leq x \leq 1, \\ 0, & \text{otherwise} \end{cases}$$

Note that the area underneath $P_X(x)$ is 1 as expected.

A remark on notation

In probability textbook, it often use $f_X(x)$ to denote the PDF of X. However, we will follow the classic text of Cover and Thomas and use $p_X(x)$ directly to denote the PDF as there should be little confusion on notation. And we can free up the common symbol $f(\cdot)$ for other places. We just need to remind ourselves that when X is continuous, $p_X(x) \neq Pr(X = x)$.

2.3 Expectation and summary statistics

Expectation is one fundamental concept in probability and we can use it to define commonly used summary statistics such as the mean and the variance.

2.3.1 Expectation

Consider a r.v. X and a deterministic function $g(\cdot)$, $g(X)$ is a r.v. as well. So each sample of $g(X)$ will be different. However, as we sample $g(X)$ multiple times, the empirical average will converge as we have more and more sample. The converged value is called the expectation of $g(X)$, and denoted by $E[g(X)]$. Mathematically,

$$E[g(X)] = \frac{1}{N} \lim_{N\to\infty} g(x_1) + g(x_2) + \cdots + g(x_N), \tag{2.3}$$

where x_i is the ith sample of X.

For discrete r.v., for any outcome x from the sample space $\mathcal{X}$, there will be $p(x)$ fraction of time that x occur. Therefore, we can rewrite (2.3) as

$$E[g(X)] = \sum_{x\in\mathcal{X}} p_X(x)g(x). \tag{2.4}$$

For continuous r.v., note that the PDF does not equal to the probability explicitly as explained in the last section. Consequently, the expression will become an integral instead. That is

$$E[g(X)] = \int p_X(x)g(x)dx \tag{2.5}$$

since

$$E[g(X)] = \lim_{\Delta\to 0^+} \sum_{n=-\infty}^{\infty} \underbrace{[p_X(n\Delta)\Delta]}_{Pr(n\Delta\le X\le(n+1)\Delta)} g(n\Delta), \tag{2.6}$$

which is just the definition of $\int p_X(x)g(x)dx$.

Expectation is linear

One most important property of expectation is that $E[\cdot]$ as an operation is linear. It means that for any two r.v.'s X and Y and two constants a and b, we have

$$E[aX + bY] = aE[X] + bE[Y]. \tag{2.7}$$

The above result can be verified readily because as we see from (2.4) and (2.5), the definitions of expectation for both discrete and continuous r.v.'s just involve

either a sum and an integral, and both of these operations are linear. Similar, for any r.v. X and constant C,

$$E[X + C] = E[X] + C. \tag{2.8}$$

We will leave the proofs of the above results as exercises.

2.3.2 Summary statistics

With a different function of $g(\cdot)$, we can compute $E[g(X)]$ as a summary description of the distribution $p_X(\cdot)$. Such description is known as a **summary statistics**. The most common summary statistics are the **mean** and the **variance**.

Mean

The mean of a r.v. X is simply the expected value of X itself, and that is equivalent to $E[X]$. Since we are taking expectation on X directly, we expected that variable X is numerical rather than categorical. For example, it wouldn't make much sense to compute the mean of a coin toss unless we pre-map the outcomes of head and tail to some values.

From (2.3), we see that that the empirical average of samples of X should converge to the mean. That is, given samples $x_1, x_2, \cdots, x_N$,

$$\frac{1}{N}(x_1 + x_2 + \cdots + x_N) \to E[X] \tag{2.9}$$

as N goes to the infinity.

Variance

The variance of X describe how much fluctuation of X from its mean. It is defined as $E[(X - \bar{X})^2]$, where $\bar{X} \triangleq E[X]$ is the mean of X. Note that the mean $\bar{X}$ is a constant despite that X is a r.v. As we expand $E[(X - \bar{X})^2]$, we

have

$$E[(X-\bar{X})^2] = E[X^2 - 2\bar{X}X + \bar{X}^2] \tag{2.10}$$
$$\stackrel{(a)}{=} E[X^2] - 2\bar{X}E[X] + \bar{X}^2 \tag{2.11}$$
$$= E[X^2] - 2\bar{X}\bar{X} + \bar{X}^2 \tag{2.12}$$
$$= E[X^2] - \bar{X}^2, \tag{2.13}$$

where it is usually more convenient to compute variance of X with the last expression and (a) is coming from the linear property of expectation.

2.4 Joint probabilities and conditional probabilities

Up to now we only consider a single scalar r.v. at a time. Let's consider multiple r.v.'s and how they interact with one another in this section.

2.4.1 Joint distributions and marginal distributions

Given two discrete r.v.'s X and Y, the joint PMF $p_{X,Y}(x,y) \triangleq Pr(X=x, Y=y)$ provides all the statistical information with respect to X and Y. Moreover, we can compute the probability of only one variable regardless the value of others. For example,

$$p_X(x) = \sum_{y \in \mathcal{Y}} p_{X,Y}(x,y) \tag{2.14}$$

The above procedure of summing out all the dummy variables from the joint probability is known as marginalization and the resulting probability $p_X(x)$ is known as a marginal distribution.

For continuous variables, the marginalization step is similar. Just the summation is replaced by integral. For example, for continuous r.v.'s X and Y,

$$p_X(x) = \int_y p_{X,Y}(x,y)dy \tag{2.15}$$

Example: Weather forecast

Let's denote P and W as the predicted weather and actual weather tomorrow, where both variables can take the outcomes of *sunny* or *rainy*. Assume the joint probability $p_{P,W}(\cdot,\cdot)$ is tabulated as below

PW	Sunny	Rainy
Sunny	0.6	0.06
Rainy	0.04	0.3

As a sanity check, first note that the joint probability should sum up to one (0.6+0.06+0.04+0.3=1). The probability of sunny tomorrow is

$$p_{P,W}(summy, sunny) + p_{P,W}(rainy, sunny) = 0.6 + 0.04 = 0.64$$

and the probability of rainy tomorrow is

$$p_{P,W}(sunny, rainy) + p_{P,W}(rainy, rainy) = 0.06 + 0.3 = 0.36,$$

which of course is just equal to 1-probability of sunny tomorrow=1-0.64.

2.4.2 Conditional probability, Bayes' rule, and the chain rule

The joint probability gives us the probability of all variables with the desired outcomes. For example, $p_{P,W}(sunny, sunny)$ in the example of last subsection gives us the probability of both predicted and actual weather is sunny tomorrow. Often, we are interested in finding the probability when some variables are already fixed and known. For example, what if we already predicted that the weather is sunny tomorrow, what is the probability that the actual weather is sunny as well?

Since only $p_{P,W}(sunny, sunny)$ and $p_{P,W}(sunny, rainy)$ correspond to sunny prediction, and among them we are interested in the case that the actual weather is also sunny, the probability should be

$$\frac{p_{P,W}(sunny, sunny)}{p_{P,W}(sunny, sunny) + p_{P,W}(sunny, rainy)}, \tag{2.16}$$

which is known to be the *conditional probability* of weather being sunny given prediction being sunny, and often is denoted as $p_{W|P}(sunny|sunny)$.

Note that by the marginalization rule described in the last subsection, the denominator in (2.16) is just $p_P(sunny)$ and so

$$p_{W|P}(sunny|sunny) = \frac{p_{P,W}(sunny, sunny)}{p_P(sunny)} \tag{2.17}$$

In general, we have

$$p_{X|Y}(x|y) = \frac{p_{X,Y}(x,y)}{p_Y(y)} \tag{2.18}$$

Therefore, we have $p_{X,Y}(x,y) = p_Y(y)p_{X|Y}(x|y)$. And by the same token, $p_{X,Y}(x,y) = p_X(x)p_{Y|X}(y|x)$. And thus $p_X(x)p_{Y|X}(y|x) = p_Y(y)p_{X|Y}(x|y)$ and $p_{Y|X}(y|x) = \frac{p_Y(y)p_{X|Y}(x|y)}{p_X(x)}$ or simply

$$p(y|x) = \frac{p(y)p(x|y)}{p(x)}. \tag{2.19}$$

The last expression is the famous Bayes' rule but it is just a straightforward result of (2.18).

Despite the fame of Bayes' rule, the following chain rule is much more useful and general in practice. Note that we can rewrite (2.18) into

$$p(x,y) = p(y)p(x|y) \tag{2.20}$$

and this can be generalized to the case even when the right hand size is conditioned. For example,

$$p(x,y|z) \overset{(a)}{=} \frac{p(x,y,z)}{p(z)} \tag{2.21}$$

$$\overset{(b)}{=} \frac{p(x,z)p(y|x,z)}{p(z)} \tag{2.22}$$

$$= \frac{p(x,z)}{p(z)}p(y|x,z) \tag{2.23}$$

$$\overset{(c)}{=} p(x|z)p(y|x,z), \tag{2.24}$$

where (a) is from (2.18) taking (x,y) as a single variable, and (b) is from (2.20) taking (x,z) as a single variable and (c) is simply from (2.18) again with y replaced by z.

Combining (2.20) and (2.24), we have the chain rule

$$\begin{aligned} p(x_1, x_2, \cdots, x_N) &= p(x_1)p(x_2, \cdots, x_N|x_1) \\ &= p(x_1)p(x_2|x_1)p(x_3, \cdots, x_N|x_1, x_2) \\ &= p(x_1)p(x_2|x_1)p(x_3|x_1, x_2)p(x_4, \cdots, x_N|x_1, x_2) \\ &= p(x_1)p(x_2|x_1)p(x_3|x_1, x_2)\cdots p(x_N|x_1, \cdots, x_{N-1}), \end{aligned}$$

where the above decomposing a joint probability into product of probabilities is known as the chain rule.

It is a good place to introduce another shorthand used throughout this book. For a list of variables, $x_k, x_{k+1}, x_{k+2}, \cdots, x_N$, we may shorthand them as x_k^N. And when $k = 1$, we may shorthand it further to x^N. For example, we can rewrite the above chain rule to

$$p(x^N) = p(x_1)p(x_2|x_1)p(x_3|x^2)\cdots p(x_N|x^{N-1}). \tag{2.25}$$

2.5 Independence and conditional independence

Just as the name suggest, we say two r.v.'s to be independent if they should have no effect on one another. For example, the outcomes of tossing two different dices should be independent. On the other hand, the forecast variable P should depend on the the actual weather W in our earlier example.

Conditional independence is a very similar concept. Just we want to see if two variables may have effect on one another given some third variable is known. It probably will be surprising to many who first encounter these concepts. Independence and conditional independence are "independent" concepts. One property does not imply the other property and vice versa.

2.5.1 Independent variables

Consider the joint probability of two r.v.'s X and Y. Given a x, we can consider the conditional probability $p_{Y|X}(y|x)$ as a function of y parameterized by x. So if we have $p_{Y|X}(y|x_1) = p_{Y|X}(y|x_2)$ for all $x_1, x_2 \in \mathcal{X}$, we must have X and Y to be independent. Because no matter what value X takes, the conditional

distribution does not change. Moreover, if X and Y are independent, then

$$p_Y(y) = \sum_{x \in \mathcal{X}} p_{X,Y}(x,y) = \sum_{x \in \mathcal{X}} p_X(x) p_{Y|X}(y|x) \tag{2.26}$$

$$\stackrel{(a)}{=} p_{Y|X}(y|x) \sum_{x \in \mathcal{X}} p_X(x) \tag{2.27}$$

$$\stackrel{(b)}{=} p_{Y|X}(y|x), \tag{2.28}$$

where (a) is because $p_{Y|X}(y|x)$ is the same for all x and (b) is due to $\sum_{x \in \mathcal{X}} p_X(x) = 1$.

Furthermore, we have

$$p_{X,Y}(x,y) \stackrel{(a)}{=} p_X(x) p_{Y|X}(y|x) \stackrel{(b)}{=} p_X(x) p_Y(y), \tag{2.29}$$

where (a) is from the chain rule and (b) is from (2.28). Note that (2.29) is usually used as the formal "definition" of independence for most probability textbooks. However, I think that (2.28) is way more natural and easier to understand.

Example: Tossing two coins

Let's denote X_1 and X_2 as the outcomes of tossing two coins. Let's also assume that the probabilities of getting a head for X_1 and X_2 are p_1 and p_2, respectively. Unless that the coins interact with some mysterious way, it is safe to assume that their outcomes should be independent. And probability of getting both heads will be $p_1 \cdot p_2$. We can tabulate the joint probability as below.

X_1X_2	Head	Tail
Head	$p_1 \cdot p_2$	$p_1(1-p_2)$
Tail	$p_2(1-p_1)$	$(1-p_1)(1-p_2)$

Let's also try to tabulate the conditional probability distribution $p_{X_2|X_1}$. Since $p_{X_2|X_1}(x_2|x_1) = \frac{p(x_1,x_2)}{p(x_1)}$, we can create a table of $p_{X_2|X_1}$ by simply dividing row of the above table by the respective $p(x_1)$. That is, p_1 for the first row, and $(1-p_1)$ for the second row. And so we get $p_{X_2|X_1}$ given by

X_1X_2	Head	Tail
Head	p_2	$1-p_2$
Tail	p_2	$1-p_2$

Note that each row of the above table is the same. That means that $p_{X_2|X_1}(\cdot|x_1)$ does not change with different x_1. Therefore, X_1 and X_2 are indeed independent.

Example: Weather forecast (con't)

Let's continue with the example from Section 2.4. Recall that the joint probability $p_{P,W}(\cdot,\cdot)$ is tabulated as below

PW	Sunny	Rainy
Sunny	0.6	0.06
Rainy	0.04	0.3

Let's try to tabulate the conditional probability $p_{W|P}$ instead. Just as in the coin toss example earlier, we should divide each rows of the above table by the respective marginal probabilities (0.6+0.06=0.66 and 0.04+0.3=0.34). Therefore, we have

PW	Sunny	Rainy
Sunny	0.91	0.09
Rainy	0.12	0.88

Note that the two rows are very different, meaning that $p_{W|P}(\cdot|p)$ changes significantly with different p. Therefore W and P must depend on one another.

To conclude this section, note that we often denote $X \perp\!\!\!\perp Y$ when X and Y are independent, i.e., (2.28) and (2.29) are satisfied.

2.5.2 Conditionally independent variables

Now, let us consider three variables X, Y, and Z. From now on, we simplify the notation by removing the subscript of p. For example, $p(x|y,z) \equiv p_{X|Y,Z}(x|y,z)$. We say that X and Y are conditionally independent given Z if

$$p(x|y_1,z) = p(x|y_2,z) \tag{2.30}$$

for any z, y_1 and y_2.

The condition should be self-evident. It states that given a fixed z, the conditional distribution of X given Y and z does not depend on the choice of Y. So given z, X and Y will be independent.

Moreover, we have

$$p(x|z) = \sum_{y\in\mathcal{Y}} p(x,y|z) \tag{2.31}$$

$$\stackrel{(a)}{=} \sum_{y\in\mathcal{Y}} p(y|z)p(x|y,z) \tag{2.32}$$

$$\stackrel{(b)}{=} p(x|y,z)\sum_{y\in\mathcal{Y}} p(y|z) \tag{2.33}$$

$$\stackrel{(c)}{=} p(x|y,z), \tag{2.34}$$

where (a) is coming from the chain rule (c.f. (2.25)), and (b) is from (2.30), and (c) is due to probability summing up to 1.

Note that the conditional joint probability $p(x,y|z)$ can now be expanded as

$$p(x,y|z) = p(y|z)p(x|y,z) = p(y|z)p(x|z), \tag{2.35}$$

where the last equality is due to (2.34). Like (2.29), (2.35) is often written as the "definition" of conditional independence in many probability textbooks. Even though (2.30) is more natural and easier to be interpreted and understood.

Example: Naïve Bayes classifier

Naïve Bayes is a simple machine learning algorithm to classify an object with some given features. A major assumption of Naïve Bayes is that the features are conditionally independent given the object class. Say if O denotes the object that $c(O)$ denotes the corresponding class. And let $f_1(O), f_2(O), \cdots, f_K(O)$ denote K features of the object. For simplicity, let's rewrite $c(O)$ as C and $f_i(O)$ as F_i. But it is important to realize that the "randomness" of $c(O), f_i(O)$ is originated from O.

$$\begin{aligned} p(c|f_1, \cdots, f_K) &= \frac{p(c, f_1, \cdots, f_K)}{p(f_1, \cdots, f_K)} = \frac{p(c)p(f_1, \cdots, f_K|c)}{p(f_1, \cdots, f_K)} \\ &= \frac{p(c)p(f_1|c)\cdots p(f_K|c)}{p(f_1, \cdots, f_K)} \\ &\propto p(c)p(f_1|c)\cdots p(f_K|c) \end{aligned}$$

To classify an object, we can simply compute the product $p(c)p(f_1|c)\cdots p(f_K|c)$ for each c and the output class should be the one with the maximum value.

To conclude this section, we like to point out that X and Y conditionally independent given Z is often denoted by $X \perp\!\!\!\perp Y|Z$, when (2.34) and (2.35) are satisfied.

2.5.3 Independence but not conditional independence

As we mentioned at the beginning of this section, a mistake beginners often make is to assume that independence would imply conditional independence or vice versa. It turns out that the two properties are totally "independent".

Let's map the outcomes of two coin tosses to zero (tail) and one (head) and denote them as X_1 and X_2. Without any magical correlation, X_1 and X_2 should be independent. Let say the probability of head for both X_1 and X_2 be p. The joint probability is given by

X_1X_2	1	0
1	p^2	$p(1-p)$
0	$p(1-p)$	$(1-p)^2$

One can verify that $p(x_1, x_2) = p(x_1)p(x_2)$ for all combination above, satis-

fying the independent condition given by (2.29).

Now, let's define $Y = X_1 \oplus X_2$, where $\oplus$ is the exclusive-or operator. Note that while $X_1 \perp\!\!\!\perp X_2$, $X_1 \perp\!\!\!\perp X_2|Y$ does not hold. Actually, note that $Y = X_1 \oplus X_2$ implies $X_2 = X_1 \oplus Y$, so given Y, we can compute X_2 deterministically from X_1. There is no way X_1 and X_2 are independent given Y.

To get a bit even more insight, let's tabulate the distributions $p_{X_2|X_1,Y}(\cdot|x_1, y)$ below

$Y = 0$

X_1X_2	1	0
1	1	0
0	0	1

$Y = 1$

X_1X_2	1	0
1	0	1
0	1	0

For $X_1 \perp\!\!\!\perp X_2|Y$, the rows in each of the table should be identical ($p(x_2|x_1, y) = p(x_2|y)$). The rows being so different suggests that X_1 and X_2 are very correlated given Y.

2.5.4 Conditional independence but not independence

Let's consider two noisy observation Y_1 and Y_2 of a r.v. X. For simplicity, let's assume all three variables are binary. And the noises $Z_1 = X \oplus Y_1$ and $Z_2 = X \oplus Y_2$ are independently generated from a binary symmetric source with probability of 1 equal to p. Let's also assume that the probability of $X = 1$ is q.

Since Y_1 and Y_2 are independent observations of X, we would expect that they will be independent given X. On the other hand, it is reasonable that Y_1 and Y_2 won't be independent (actually they should be very correlated). Let's first show $Y_1 \perp\!\!\!\perp Y_2|X$.

Let's tabulate the joint probability $p(y_1, y_2, x)$ as below

$X = 0$

Y_1Y_2	1	0
1	$(1-q)p^2$	$(1-q)(1-p)p$
0	$(1-q)(1-p)p$	$(1-q)(1-p)^2$

$X = 1$

Y_1Y_2	1	0
1	$q(1-p)^2$	$qp(1-p)$
0	$qp(1-p)$	qp^2

And from the table above, let's tabulate the conditional probability $p(y_2|y_1, x) = \frac{p(y_2,y_1,x)}{p(y_1,x)}$ below

$X = 0$

Y_1Y_2	1	0
1	p	$1-p$
0	p	$1-p$

$X = 1$

Y_1Y_2	1	0
1	$1-p$	p
0	$1-p$	p

Note that both rows in each table are the same. That means that $p(y_2|y_1, x) = p(y_2|x)$ and thus $Y_1 \perp\!\!\!\perp Y_2|X$.

On the other hand, let's tabulate the joint probability $p(y_1, y_2)as$

Y_1Y_2	1	0
1	$(1-q)p^2 + q(1-p)^2$	$(1-p)p$
0	$(1-p)p$	$(1-q)(1-p)^2 + qp^2$

It is apparent that generally we won't have $p(y_1, y_2) = p(y_1)p(y_2)$. So Y_1 and Y_2 are not independent.

2.6 Markov chain

Many sequential random variables have relatively local influence to each other. For example, if we consider the price of a stock each day as a sequence of r.v.'s, the stock price today is probably correlated more with the price yesterday than the price last month. To an extreme, we may assume that all historical information regarding today's price is summarized completely by the yesterday's price. Even though it definitely is not true, it would be a good approximation to start with. And we will say these price variables form a Markov chain.

Mathematically, let $X_1, \cdots, X_N$ be the sequence of price variables. We say the variables form a Markov chain if for any k and l ($l < k-1$), X_k is conditionally independent of X_l given X_{k-1}. We often denote the chain by $X_1 \leftrightarrow X_2 \leftrightarrow \cdots \leftrightarrow X_N$. Some textbook also uses one directional arrow for the notation. However, we want to use double-sided arrow to indicate that the definition is symmetric. That is if we have a chain $X_1 \leftrightarrow X_2 \leftrightarrow \cdots \leftrightarrow X_N$, we have $X_N \leftrightarrow X_{N-1} \leftrightarrow \cdots \leftrightarrow X_1$. Note that the Markov property implies that we can express the joint probability

$$\begin{aligned} p(x^N) &= p(x_1)p(x_2|x_1)p(x_3|x^2)\cdots p(x_N|x^{N-1}) \\ &= p(x_1)p(x_2|x_1)p(x_3|x_2)\cdots p(x_N|x_{N-1}) \end{aligned} \tag{2.36}$$

As an EXERCISE, show that $X_1 \leftrightarrow X_2 \leftrightarrow \cdots \leftrightarrow X_N$ implies $X_N \leftrightarrow X_{N-1} \leftrightarrow \cdots \leftrightarrow X_1$.

2.7 Probabilistic inference

One of the most common problem we encounter in probability is to estimate some latent variables based on observations. The latent variables in turn will depend on the model that we choose, which is often decided by some model parameters.

2.7.1 ML vs MAP vs Bayesian inference

MAP

Let o, θ, and z are the observed variable, the model parameter, and the latent variable. Given the model parameter θ and the observation o, it is natural to estimate z as

$$\hat{z} = \arg\max_z p(z|\theta, o) \tag{2.37}$$

Note that we can rewrite $p(z|\theta, o) = \frac{p(z,o|\theta)}{p(o|\theta)} = \frac{p(o|z,\theta)p(z|\theta)}{p(o|\theta)}$. Since the denominator does not depend on x, (2.37) can be rewritten as

$$\hat{z} = \arg\max_z p(o|z, \theta)p(z|\theta) \tag{2.38}$$

Often time, the observation does not depend on the model parameter anymore when the laten variable z is given. Therefore, we can simplify (2.38) to

$$\hat{z} = \arg\max_z \underbrace{p(o|z)}_{likelihood} \underbrace{p(z|\theta)}_{prior}, \tag{2.39}$$

where $p(o|z)$ is known as the likelihood function and $p(z|\theta)$ is the prior. And (2.37)-(2.39) describes the so-called Maximum a-posteriori (MAP) estimator.

ML

It is sometimes impossible to figure the prior $p(z|\theta)$. In that case, the best we can do is simply ignore $p(z|\theta)$ and assume it to be a constant. Hence, (2.39) will become

$$\hat{z} = \arg\max_z p(o|z), \tag{2.40}$$

and this is known as the maximum likelihood (ML) estimation.

Bayesian estimation

In both MAP and ML, we just estimate z from the mode of some functions ($p(z|o,\theta)$ and $p(o|z)$). However, it probably is a waste to discard all model information besides the winner. The Bayesian estimation is more conservative and try to leverage all possible z by computing a weighted sum of them as the estimate. More precisely, we have

$$\hat{z} = \sum_{x \in \mathcal{Z}} z p(z|\theta, o) \tag{2.41}$$

Note that often time we are not fundamentally interested in z but instead some function say f that depends on z. Then, we will create an estimate as

$$\sum_{x \in \mathcal{Z}} f(z) p(z|\theta, o). \tag{2.42}$$

In contrast, we will simply output $f(\hat{z})$ for MAP and ML, when $\hat{z}$ is latent variable maximizes the a posteriori or the likelihood function.

To consolidate the idea, let's consider a simple toy example below.

Example: Three types of coins

Let say we have three types of identically looking coins but only the first type is fair. And the second type is heavily biased towards head with $p(Head) = 0.8$ and the third type is biased towards tail with $p(Head) = 0.4$.
We have an unknown number of these coins put into a jar. Then, we randomly draw a coin from the jar and toss the coin three times. Let say we got two tails for the first two tosses. What is the probability of getting a head for the last toss?

Our estimated result heavily relies on what estimation method does we use. Let's try to tackle the problem separately by ML, MAP, and Bayesian inference.

Example: Solving the coin problem with ML

Let's denote $Z \in \{1, 2, 3\}$ as the type of the coin that was actually picked. And let $x(z)$ be the probability of getting a head when type-z coin is picked. Then,

$$x(z) = \begin{cases} 0.5, & z = 1, \\ 0.8, & z = 2, \\ 0.4 & z = 3 \end{cases}$$

For ML, we assume no prior knowledge of Z and the best estimate of Z is simply

$$\hat{z} = \arg \max_{z \in \{1,2,3\}} p(o|z),$$

where the observation o is $(T)ail, (T)ail$. Thus,

$$p(o|z) = \begin{cases} 0.5 \cdot 0.5 = 0.25, & z = 1, \\ 0.2 \cdot 0.2 = 0.04, & z = 2, \\ 0.6 \cdot 0.6 = 0.36, & z = 3. \end{cases}$$

Since $p(o|z)$ is largest for $z = 3$. We will estimate $\hat{z} = 3$, the predicted probability of head for the last toss is 0.4.

Example: Solving the coin problem with MAP

When using ML, we do not assume any prior knowledge of Z. Let's assume that there are two type-1 coins, 7 type-2 coins, but only one type-3 coin in the jar. Thus, we have

$$p(z) = \begin{cases} 0.2, & z = 1, \\ 0.7, & z = 2, \\ 0.1, & z = 3. \end{cases}$$

For MAP, we compute the best estimate of z as

$$\hat{z} = \arg \max_{z \in \{1,2,3\}} p(z|o) \overset{(a)}{=} \arg \max_{z \in \{1,2,3\}} \frac{p(z)p(o|z)}{p(o)} \overset{(b)}{=} \arg \max_{z \in \{1,2,3\}} p(z)p(o|z),$$

where (a) is due to Bayes' rule and (b) is because $p(o)$ is a constant w.r.t. to z. Since

$$p(z)p(o|z) = \begin{cases} 0.2 \cdot 0.25 = 0.05, & z = 1, \\ 0.7 \cdot 0.04 = 0.028, & z = 2, \\ 0.1 \cdot 0.36 = 0.036, & z = 3, \end{cases}$$

the best estimate $\hat{z} = 1$, and so the predicted probability of head is 0.5 for the last toss.

Example: Solving the coin problem with Bayesian estimation

Rather than picking a single best model parameter in MAP, Bayesian estimation tries to leverage all models and makes prediction as the weighted average of estimates from all models. That is, we will estimate x as

$$\hat{x} = \sum_{z \in \{1,2,3\}} x(z)p(z|o).$$

Note that $p(z|o) \propto p(z)p(o|z)$ and should normalize to 1, therefore we can compute $p(z|o)$ as

$$p(z|o) = \begin{cases} \frac{0.05}{0.05+0.028+0.036} = 0.4386, & z = 1, \\ \frac{0.028}{0.05+0.028+0.036} = 0.2456, & z = 2, \\ \frac{0.036}{0.05+0.028+0.036} = 0.3158, & z = 3. \end{cases}$$

Therefore,

$$\hat{x} = 0.4386 \cdot 0.5 + 0.2456 \cdot 0.8 + 0.3158 \cdot 0.4 = 0.5421.$$

2.7.2 Conjugate prior

In the example given in the last section, we have exactly three types of coins and we know precisely the probability of head for each type. In many real problem, the prior knowledge can be more vague. What if we don't know about the probability of head for the coin but we tend to believe that we are more likely to have a fair coin (probability of head close to 0.5) than an unfair coin. In this case, we may impose a prior similar to that as shown in Figure 2.7.2.

There are many ways we can parametrize a prior as shown in Figure 2.7.2. The real problem is which function we should choose. Note that the likelihood function $p(o|x)$ is given by $(1-x)^2$. More generally, if we have u heads and v tails in $u+v$ tosses,

$$p(o|x) = x^u(1-x)^v \tag{2.43}$$

To estimate x with MAP, we want

$$\hat{x} = \arg\max_x p(o|x)p(x) = \arg\max_x x^u(1-x)^v p(x)$$

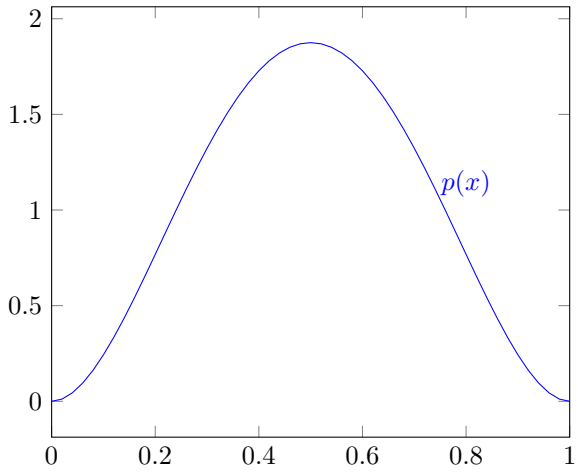

Figure 2.1: A desired continuous prior

Different people may have different opinions on the choice of $p(x)$. However, if we select $p(x)$ of a form $p(x) \propto x^{u'}(1-x)^{v'}$, then the resulting posterior distribution with the same form as before. This choice is often chosen for practical purposes, and a prior with same "form" as its likelihood (and thus posterior) is known as the conjugate prior.

It turns out the conjugate prior with $p(x) \propto x^{u'}(1-x)^{v'}$ is the Beta distribution. Here, we will just state a few facts useful for our discussion here. A reader may find more detail of the Beta distribution in the Appendix. A Beta distribution have two parameters (a, b) and its PDF is denoted by $Beta(x; a, b) \propto x^{a-1}(1-x)^{b-1}$. The mode of its PDF is given by

$$\begin{cases} \frac{a-1}{a+b-2}, & a, b > 1, \\ [0, 1], & a = b = 1, \\ 0 \text{ or } 1, & \text{otherwise.} \end{cases} \tag{2.44}$$

And the mean of Beta is given by $\frac{a}{a+b}$.

If a prior $Beta(a, b)$ is chosen, and u Heads and v Tails are observed when tossing a coin $u + v$ times, the posteriori probability

$$
\begin{aligned}
p(x|o) &= K_1 \cdot Beta(x; a, b)p(o|x) && (2.45)\\
&= K_1 \cdot Beta(x; a, b)x^u(1-x)^v && (2.46)\\
&= K_2 \cdot K_1 \cdot x^{a-1}(1-x)^{b-1}x^u(1-x)^v && (2.47)\\
&= K_2 \cdot K_1 \cdot x^{u+a-1}(1-x)^{v+b-1} && (2.48)\\
&= \underbrace{K_3 \cdot K_2 \cdot K_1}_{1} Beta(x; u+a, v+b), && (2.49)
\end{aligned}
$$

where K_1, K_2, and K_3 are some normalization factors and note that the product $K_1K_2K_3 = 1$ since both $p(x|o)$ and $Beta(x; u+a, v+b)$ normalize to 1 as we integrate over all x. In summary, the posteriori probability after observing u heads and v tails is simply another Beta but with parameters changed to $u+a$ and $v+b$.

Example: Revisit coin problem with $Beta(3,3)$ prior

Let's revisit the coin tossing problem but we do not restrict to any specific types of coin. Instead, we will simply assume the prior probability of head is $Beta(3,3)$, which is the one actually shown in Figure 2.7.2.
After observing two tails, the posteriori probability has a distribution $Beta(3, 3+2) = Beta(3,5)$.
If we are going to estimate the probability of Head with MAP, we should pick the mode of $Beta(x; 3, 5)$, which will be $\frac{3-1}{3+5-2} = \frac{1}{3}$.
If we tried to estimate the probability using Bayesian inference, the estimate should be

$$\int_x xp(x|o)dx = \int_x xBeta(x; 3, 5)dx = \frac{3}{3+5} = \frac{3}{8}.$$

In the above example, we assume that a prior shown in Figure 2.7.2 was used. What if we don't have any prior knowledge, it seems that it is reasonable to use a uniform prior, i.e., constant everywhere. Recall that $Beta(x; a, b) \propto x^{a-1}(1-x)^{b-1}$. Thus, we have a uniform prior if we pick $a = 1$ and $b = 1$.

Example: Revisit coin problem with uniform ($Beta(1,1)$) prior

Let's repeat the last example but just pick a uniform prior $Beta(1,1)$. Thus after observing two tails, the posteriori probability has a distribution $Beta(1,1+2) = Beta(1,3)$.

If we are going to estimate the probability of Head with MAP, we should pick the mode of $Beta(x;1,3)$, which will be $\frac{1-1}{1+3-2} = 0$. Note that the result is rather extreme as it essentially rules out the possibility of getting a head for the next toss.

Note that as the prior $Beta(1,1)$ is really a constant, the MAP estimation with such prior is actually just the ML estimate.

Instead, if we tried to estimate the probability using Bayesian inference, the estimate will be the mean of $Beta(x;1,3)$, which is

$$\int_x xp(x|o)dx = \int_x xBeta(x;1,3)dx = \frac{1}{1+3} = \frac{1}{4}.$$

It may seem surprising at first that the ML estimation (or MAP with uniform prior) result is so extreme. But without additional information, the best guess of the probability is from statistically counting. And the estimate of zero head probability seems reasonable from that perspective as none out of two historical tosses were head.

When we impose a non-uniform prior such as $Beta(3,3)$ as in our example. It introduces some "regularization" effect that makes the estimate less extreme. Just by inspection, we can see that the $Beta(3,3)$ prior can be interpreted as some prior experiment has been performed before our observations. In particular, the prior experiment included $4 = (3+3-2)$ tosses and out of that, $2 = (3-1)$ were head. Even though for $Beta(u,v)$ with non-integer u and v, it would be much more difficult to interpret the physical meaning of such prior.

Another interesting observation from the above example is that Bayesian inference includes some free regularization even when the non-informative uniform prior is used. The estimated probability of head is $\frac{1}{4}$ rather than 0 as we just consider the most probably model in MAP or ML. The averaging effect over many model parameters will create a less extreme estimate and so offers some regularization effect.