Lecture 2: Math Review

These notes are to supplement those released by Prof. Wang, and do not cover all the material covered in class this day.

1 Probability

What is a random variable?

- A random variable is a variable that can take on different values randomly.
- Takes on values in an alphabet $\mathcal{X}$. E.g., $\{0,1,2,3\}$, $\mathbb{R}$, $[0,1]^d$
- Can be continuous or discrete.

Probability distributions $p(x)$ must satisfy:

$$\text{DISCRETE:} \quad \sum_{x\in\mathcal{X}} p(x) = 1 \quad \text{and} \quad 0 \le p(x) \le 1 \text{ for all values of } x.$$

$$\text{CONTINUOUS:} \quad \int_{\mathcal{X}} p(x) = 1 \quad \text{and} \quad p(x) \ge 0 \text{ for all values of } x.$$

The **expected value** of a random variable $x \sim p(x)$ is:[1]

$$\mathbb{E}[x] \triangleq \int_{\mathcal{X}} p(x)x dx.$$

Similarly, we can write the expected value of a conditional random variable $x \sim p(x \mid y)$ as

$$\mathbb{E}[x \mid y] = \int_{\mathcal{X}} p(x \mid y) dx.$$

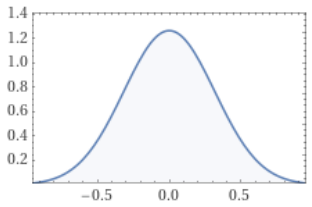

Figure 1: Probability densities can be greater than 1, as shown by a $\mathcal{N}(\mu = 0, \sigma = 0.1)$ distribution.

Brief note about wording. Typically, we will use "probability mass function" to refer to probability distributions over discrete x and "probability density function" for continuous x. For continuous random variables, we can have $p(x) > 0$ (Fig. 1). Strictly speaking, the *probability* of a continuous random variable taking a certain value is 0: $\mathbb{P}(x = 0.3) = 0$. Here's one way to think about this. Imagine that you have a needle sticking up in a bin, and drop a grain on salt randomly inside that bin. The probability that the grain of salt will end up balanced on the point of the needle is zero – it will never happen. However, the *probability density* is still non-zero – if the grain of salt is dropped uniformly at random inside the bin, then the probability density of it landing at any point in the bin is $p(x) = \frac{1}{\text{BinVolume}}$.

1.1 Joint distributions.

A joint distribution $p(x, y)$ can be viewed as an single-variate distibution over the alphabet $\mathcal{X} \times \mathcal{Y}$, satisfying the usual properties. For example, if x and y are both continuous, then the joint distribution must satisfy $\int_{\mathcal{X}\times\mathcal{Y}} p(x, y)dxdy = 1$. Note that we can also have joint distributions where x and y have different *types*. For example, when modeling the state of a car, x might be an image and y might be a LIDAR scan.

[1] I will use the symbol "$\triangleq$" to denote "is defined to be equal to," in the same way that computer scientists sometimes use ":=". Please let me know if/when notation is unclear. The site `https://detexify.kirelabs.org/` is also sometimes useful for looking up unfamiliar notation.

For a joint distribution $p(x,y)$, we can define the *marginal distributions* as $p(x) = \int_{\mathcal{Y}} p(x,y)dy$ and $p(y) = \int_{\mathcal{X}} p(x,y)dx$.[2] The marginal distributions tells you about the state of the world in the absence of any information about one of the variables (e.g., if a car's LIDAR breaks).

The conditional distribution $p(x \mid y)$[3] is an updated belief about the state of the world, given some observation. Let's say that we observe a concrete value $y = 4$. Then the conditional distribution satisfies

$$p(x \mid y = 4) = \frac{p(x, y = 4)}{p(y = 4)} = \frac{p(x, y = 4)}{\int p(x, y = 4)dx}.$$

We can think about conditional distributions as a table, with $\mathcal{X}$ on one axis and $\mathcal{Y}$ on the other axis (see Table 1). The cells of the table sum to 1. Conditioning corresponds to selecting one row (or column) of the table, and then re-normalizing that row so that it sums to one.

	$\mathcal{X}$				
	0.1	0.03	0.05	0.0	0.06
	0.0	0.0	0.08	0.01	0.1
	0.01	0.02	0.04	0.04	0.1
$y = 4$	0.01	0.01	0.0	0.03	0.2
	0.08	0.0	0.0	0.02	0.05

Table 1: Visualizing a conditional distribution for discrete random variables.

This same "table" analogy works in higher dimensions, as shown in Fig. 2.

There can be some notational confusion about conditional distributions: is $p(x \mid y)$ a function of one input or two inputs? This really depends on the problem setting, on whether you're just analyzing this function mathematically, or really have observed a certain value of y. We will use notation like $p(x \mid y = 4)$ in instances when we are only considering varying x.

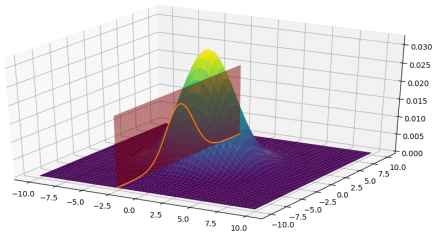

Figure 2: Visualizing a conditional distibution of continuous random variables. Animation. Note the similarity with Table 1.

We say that random variables x, y are **conditionally independent** if observing one of them doesn't tell you anything about the other. Formally, we might write this as

$$p(x \mid y) = p(x) \quad \text{for all values of } x, y. \tag{1}$$

1.2 Chain Rule

The chain rule lets you write a joint distribution as a product of certain conditional distributions:

$$p(x,y,z) = p(x \mid y,z)p(y \mid z)p(z). \tag{2}$$

The same rule applies if everything is conditioned on some additional variable (say) a:

$$p(x,y,z \mid a) = p(x \mid y,z,a)p(y \mid z,a)p(z). \tag{3}$$

The process of writing a joint distribution in terms of conditional distributions is known as **factoring**. Factoring becomes interesting when certain random variables are independent of one another – that allows you to exclude certain terms on the RHS of the "|". For example, suppose that z is conditionally independent of x given y. That is, $p(z \mid x,y) = p(z \mid y)$ for all values of x, y, z. Then we can factor the joint distribution as $p(x,y,z) = p(x)p(y \mid x)p(z \mid y)$.

[2] Moving forward I'm going to assume everything is continuous, unless explicitly stated. The math still works for discrete objects.

[3] Use `\mid` to typeset conditional distributions.

1.3 Bayes' Rule

Bayes' Rule allows us to use observations of one random variable to make guesses about the values of another random variable. Often we have some sense of how observations are generated. For example, consider the GPS on your phone, which gets noisy measurements y of distances to satellites orbiting above, which depend on your true (unknown) GPS location, x. If we have an estimate for what the distances y will look like given your current location x, then Bayes' Rule tells us how to use that information to estimate your current location:

$$p(x \mid y) = \frac{p(y \mid x)p(x)}{p(y)}. \tag{4}$$

If you ever forget Bayes' Rule, it's easy to rederive from the following identity:

$$p(x, y) = p(y \mid x)p(x) = p(x \mid y)p(y). \tag{5}$$

Bayes' Rule will be incredibly useful in this course because it will allow us to infer the solutions to decision making problems. Precisely, we will use it to answer the counterfactual question: if you start at A and want to get to B, what states might you visit or what actions might you take? Often we have data or some model that can tell us the likelihood of getting to B if we take some action (or visit some state); plugging this into Bayes' Rule allows us to identify the best actions for arriving at B:

$$p(\text{action} \mid \text{start} = A, \text{end} = B) = \frac{p(\text{end} = B \mid \text{action}, \text{start} = A)p(\text{action} \mid \text{start} = A)}{p(\text{end} = B \mid \text{start} = A)}.$$

Note that in this application of Bayes' rule, each of the four terms is conditioned on the same event "start $= A$".

1.4 Important distributions

In this course we will frequently make use of the geometric distribution $\textsc{Geom}(1-\gamma)$ and the $\textsc{Bernoulli}(1-\gamma)$ distribution. The Bernoulli distribution with parameter p is a coin toss:

$$p(x) = \begin{cases} 1 & \text{with probability } 1-\gamma \\ 0 & \text{with probability } \gamma. \end{cases}$$

The expected value of a Bernoulli random variable is $\mathbb{E}[x] = 1-\gamma$.

The geometric distribution corresponds to the number of times you have to toss that coin until it lands on heads:

$$p(x) = (1-\gamma)\gamma^x \qquad \text{for } x = 0, 1, \cdots.$$

The expected value of a geometric distribution is $\mathbb{E}[x] = \frac{1}{1-\gamma}$. This tells you, on average, how many times you have to toss the coin until it lands on heads. I have defined these distributions with a parameter $1-\gamma$, rather than γ, because it will highlight connections with reinforcement learning in later lectures.

References