Probability and Information Theory Review

Akhil Vasvani

March 2019

1 Questions

Exercise 1. Compare "Frequentist probability" vs. "Bayesian probability"?

Proof. **Frequentist probability** is related directly to the rates at which events occur. For example, 50% of drawing from a deck results in a red car. These kinds of events are often repeatable. When we say that an outcome has a probability p of occurring, it means that if we repeated the experiment (e.g., draw a hand of cards) infinitely many times, then proportion p of the repetitions would result in that outcome. **Bayesian probability**, on the other hand, is related to qualitative levels of certainty. For example, a doctor analyzes a patient and says that the patient has a 40% chance of having the flu. The probability represents a degree of belief, with 1 indicating absolute certainty that the patient has the flu and 0 indicating absolute certainty that the patient does not have the flu.

Both frequentist probability and Bayesian probability follow the same set of axioms. □

Exercise 2. What is a random variable?

Proof. A **random variable** is a variable that can take on different values randomly. On its own, a random variable is just a description of the states that are possible; it must be coupled with a probability distribution that specifies how likely each of these states are.

Random variables may be discrete or continuous. A discrete random variable is one that has a finite or countably infinite number of states. Note that these states are not necessarily the integers; they can also just be named states that are not considered to have any numerical value. A continuous random variable is associated with a real value. □

Exercise 3. What is a probability distribution?

Proof. A **probability distribution** is a description of how likely a random variable or set of random variables is to take on each of its possible states. The way we describe probability distributions depends on whether the variables are discrete or continuous. □

Exercise 4. What is a probability mass function?

Proof. A probability distribution over discrete variables may be described using a **probability mass function** (PMF). The probability mass function maps from a state of a random variable to the probability of that random variable taking on that state. The probability that $\mathrm{x} = x$ is denoted as $P(\mathrm{x})$ with a probability of 1 indicating that $\mathrm{x} = x$ is certain and a probability of 0 indicating that $\mathrm{x} = x$ is impossible. □

Exercise 5. What is a probability density function?

Proof. When working with continuous random variables, we describe probability distributions using a **probability density function** (PDF) rather than a probability mass function.

Note. A probability density function $p(x)$ does not give the probability of a specific state directly, instead the probability of landing inside an infinitesimal region with volume is given by $p(x)$. We can integrate the density function to find the actual probability mass of a set of points—called the **cumulative density function (CDF)**. Specifically, the probability that x lies in some set $\mathbb{Q}$ is given by the integral of $p(x)$ over that set. There is also a CDF for probability mass functions as well.

□

Exercise 6. What is a joint probability distribution?

Proof. Probability mass functions can act on many variables at the same time. Such a probability distribution over many variables is known as a **joint probability distribution**. $P(\mathrm{x} = x, \mathrm{y} = y)$ denotes that $\mathrm{x} = x$ and $\mathrm{y} = y$ simultaneously. We may also write $P(x, y)$ for brevity □

Exercise 7. What are the conditions for a function to be a probability mass function?

Proof. To be a probability mass function on a random variable x, a function P must satisfy the following properties:

1) The domain of P must be the set of all possible states of x.

2) $\in \mathrm{x}$, $0 \leq P(x) \leq 1$. An impossible event has probability 0 and no state can be less probable than that. Likewise, an event that is guaranteed to happen has probability 1 , and no state can have a greater chance of occurring.

3) $\sum_{x \in \mathrm{x}} P(x) = 1$. We refer to this property as being normalized. Without this property, we could obtain probabilities greater than one by computing the probability of one of many events occurring. □

Exercise 8. What are the conditions for a function to be a probability density function?

Proof. To be a probability density function, a function p must satisfy the following properties:

1) The domain of p must be the set of all possible states of x.

2) $\in \mathrm{x}$, $p(x) \geq 0$. Note that we do not require $p(x) \leq 1$.

3) $\int p(x) dx = 1$. □

Exercise 9. What is a marginal probability? Given the joint probability function, how will you calculate it?

Proof. Sometimes we know the probability distribution over a set of variables and we want to know the probability distribution over just a subset of them. The probability distribution over the subset is known as the **marginal probability distribution**.

For discrete random variables x and y, we can find $P(\mathrm{x})$ with the sum rule:

$$\forall x \in \mathrm{x}, P(\mathrm{x} = x) = \sum_{y} P(\mathrm{x} = x, \mathrm{y} = y). \tag{1}$$

For continuous random variables x and y, we use integration instead of summation:

$$p(x) = \int p(x, y)dy. \tag{2}$$

□

Exercise 10. What is conditional probability? Given the joint probability function, how will you calculate it?

Proof. In many cases, we are interested in the probability of some event, given that some other event has happened. This is called a conditional probability. We denote the **conditional probability** that $\mathrm{y} = y$ given $\mathrm{x} = x$ as $P(\mathrm{y} = y \mid \mathrm{x} = x)$. This conditional probability can be computed with the formula:

$$P(\mathrm{y} = y \mid \mathrm{x} = y) = \frac{P(\mathrm{y} = y, \mathrm{x} = x)}{P(\mathrm{x} = x)}. \tag{3}$$

Note. The conditional probability is only defined when $P(\mathrm{x} = x) > 0$. We cannot compute the conditional probability conditioned on an event that never happens.

Example. Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be two jointly continuous random variables with joint PDF

$$f_{XY}(x, y) = \begin{cases} \frac{x^2}{4} + \frac{y^2}{4} + \frac{xy}{6} & 0 \le x \le 1, 0 \le y \le 2 \\ 0 & \text{otherwise} \end{cases}$$

For $0 \le y \le 2$, let's find the conditional PDF of $\boldsymbol{X}$ given $\boldsymbol{Y} = y$:

$$\begin{aligned} f_{\boldsymbol{Y}(y)} &= \int_0^1 \frac{x^2}{4} + \frac{y^2}{4} + \frac{xy}{6}\ dx \\ &= \frac{3y^2 + y + 1}{12}, \qquad \text{for } 0 \le y \le 2. \end{aligned}$$

Thus, for $0 \le y \le 2$, we have

$$\begin{aligned} f_{\boldsymbol{X}|\boldsymbol{Y}}(x|y) &= \frac{f_{\boldsymbol{XY}}(x, y)}{f_{\boldsymbol{Y}(y)}} \\ &= \frac{3x^2 + 3y^2 + 2xy}{3y^2 + y + 1}, \qquad \text{for } 0 \le x \le 1. \end{aligned}$$

$$\rightarrow f_{\boldsymbol{X}|\boldsymbol{Y}}(x|y) = \begin{cases} \frac{3x^2+3y^2+2xy}{3y^2+y+1} & 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

□

Exercise 11. State the Chain rule of conditional probabilities.

Proof. Any joint probability distribution over many random variables may be decomposed into conditional distributions over only one variable (called Chain Rule or Product Rule):

$$P(\mathrm{x}^{(1)}, ..., \mathrm{x}^{(n)}) = P(\mathrm{x}^{(1)}) \prod_{i=2}^{n} P(\mathrm{x}^{(i)} | \mathrm{x}^{(1)}, ..., \mathrm{x}^{(i-1)}) \tag{4}$$

□

Exercise 12. What are the conditions for independence and conditional independence of two random variables?

Proof. Two random variables x and y are **independent** if their probability distribution can be expressed as a product of two factors, one involving only x and one involving only y:

$$\forall x \in \mathrm{x}, y \in \mathrm{y}, p(\mathrm{x} = x, \mathrm{y} = y) = p(\mathrm{x} = x) p(\mathrm{y} = y). \tag{5}$$

Two random variables x and y are **conditionally independent** given a random variable z if the conditional probability distribution over x and y factorizes in this way for every value of z:

$$\forall x \in \mathrm{x}, y \in \mathrm{y}, z \in \mathrm{z}, p(\mathrm{x} = x, \mathrm{y} = y | \mathrm{z} = z) = p(\mathrm{x} = x | \mathrm{z} = z) p(\mathrm{y} = y | \mathrm{z} = z). \tag{6}$$

□

Exercise 13. What are expectation, variance and covariance?

Proof. The **expectation** or expected value of some function $f(x)$ with respect to a probability distribution $P(\mathrm{x})$ is the average or mean value that f takes on x is drawn from P. For discrete variables this can be computed with a summation:

$$\mathbb{E}_{\mathrm{x} \sim P}[f(x)] = \sum_{x} P(x) f(x), \tag{7}$$

while for continuous variables it is computed with an integral:

$$\mathbb{E}_{\mathrm{x} \sim P}[f(x)] = \int p(x) f(x) dx. \tag{8}$$

The **variance** gives a measure of how much the values of a function of a random variable x vary as we sample different values of x from its probability distribution:

$$\mathrm{Var}(f(x)) = \mathbb{E}[(f(x) - \mathbb{E}[f(x)]^2]. \tag{9}$$

Note. When the variance is low, the values of $f(x)$ cluster near their expected value. The square root of the variance is known as the standard deviation.

The **covariance** gives some sense of how much two values are linearly related to each other, as well as the scale of these variables:

$$\mathrm{Cov}(f(x), g(y)) = \mathbb{E}[(f(x) - \mathbb{E}[f(x)])(g(y) - \mathbb{E}[g(y)])] \tag{10}$$

High absolute values of the covariance mean that the values change very much and are both far from their respective means at the same time. If the sign of the covariance is positive, then both variables tend to take on relatively high values simultaneously. If the sign of the covariance is negative, then one variable tends to take on a relatively high value at the times that the other takes on a relatively low value and vice versa.

□

Exercise 14. Compare covariance and independence.

Proof. The notions of covariance and dependence are related, but are in fact distinct concepts. They are related because two variables that are independent have zero covariance, and two variables that have non-zero covariance are dependent.

Note. If two variables are uncorrelated, then this DOES NOT imply that the two variables are independent.

□

Exercise 15. What is the covariance for a vector of random variables?

Proof. The **covariance matrix** of a random vector $\boldsymbol{x} \in \mathbb{R}^n$ is an $n \times n$ matrix, such that:

$$\mathrm{Cov}(\mathbf{x})_{i,j} = \mathrm{Cov}(\mathrm{x}_i, \mathrm{x}_j). \tag{11}$$

The diagonal elements of the covariance give the variance:

$$\mathrm{Cov}(\mathrm{x}_i, \mathrm{x}_i) = \mathrm{Var}(\mathrm{x}_i). \tag{12}$$

□

Exercise 16. What is a Bernoulli distribution? Calculate the expectation and variance of a random variable that follows Bernoulli distribution

Proof. The **Bernoulli distribution** is a distribution over a single binary random variable. It is controlled by a single parameter $\phi \in [0, 1]$ which gives the probability of the random variable being equal to 1. It has the following properties:

$$P(\mathrm{x} = 1) = \phi \tag{13}$$

$$P(\mathrm{x} = 0) = 1 - \phi \tag{14}$$

$$P(\mathrm{x} = x) = \phi^x (1 - \phi)^{1-x} \tag{15}$$

$$\mathbb{E}_{\mathrm{x}}[\mathrm{x}] = \phi \tag{16}$$

$$\mathrm{Var}_{\mathrm{x}}[\mathrm{x}] = \phi(1 - \phi) \tag{17}$$

□

Exercise 17. What is a multinoulli distribution?

Proof. The **multinoulli** or **categorical distribution** is a distribution over a single discrete variable with k different states, where k is finite.

Note. The multinoulli distribution is a special case of the **multinomial** distribution. A multinomial distribution is the distribution over vectors in $0, ..., n^k$ representing how many times each of the k categories is visited when n samples are drawn from a multinoulli distribution. Many texts use the term "multinomial" to refer to multinoulli distributions without clarifying that they refer only to the $n = 1$ case.

The multinoulli distribution is parametrized by a vector $\boldsymbol{p} \in [0, 1]^{k-1}$, where p_i gives the probability of the i-th state. The final k-th state's probability is given by $1 - \mathbf{1}^\top \boldsymbol{p}$. Note that we must constrain $\mathbf{1}^\top \boldsymbol{p} \leq 1$.

Multinoulli distributions are often used to refer to distributions over categories of objects, so we do not usually assume that state 1 has numerical value 1, etc. □

Exercise 18. What is a normal distribution?

Proof. The most commonly used distribution over real numbers is the **normal distribution** also known as the Gaussian distribution:

$$\mathcal{N}(x; \mu, \sigma^2) = \sqrt{\frac{1}{2\pi\sigma^2}} \exp\Big(-\frac{1}{2\sigma^2}(x - \mu)^2 \Big). \tag{18}$$

The two parameters $\mu \in \mathbb{R}$ and $\sigma \in (0, \infty)$ control the normal distribution. The parameter μ gives the coordinate of the central peak. This is also the mean of the distribution: $\mathbb{E}[\mathrm{x}] = \mu$. The standard deviation of the distribution is given by σ, and the variance by σ^2.

When we evaluate the PDF, we need to square and invert σ (let's call it β). When we need to frequently evaluate the PDF with different parameter values, a more efficient way of parametrizing the distribution is to use the β parameter $\in (0, \infty)$ to control precision or inverse variance of the distribution:

$$\mathcal{N}(x; \mu, \beta^{-1}) = \sqrt{\frac{\beta}{2\pi}} \exp\Big(-\frac{1}{2}\beta(x - \mu)^2 \Big). \tag{19}$$

The normal distribution generalizes to $\mathbb{R}^n$, in which case it is known as the **multivariate normal distribution**. It may be parametrized with a positive definite symmetric matrix $\mathbf{\Sigma}$:

$$\mathcal{N}(x; \boldsymbol{\mu}, \mathbf{\Sigma}) = \sqrt{\frac{1}{(2\pi)^n \det(\mathbf{\Sigma})}} \exp\Big(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \Big). \tag{20}$$

The parameter $\boldsymbol{\mu}$ still gives the mean of the distribution, though now it is vector-valued. The parameter $\mathbf{\Sigma}$ gives the covariance matrix of the distribution. As in the univariate case, when we wish to evaluate the PDF several times for many different values of the parameters, the covariance is not a computationally efficient way to parametrize the distribution, since we need to invert $\mathbf{\Sigma}$ to evaluate the PDF. We can instead use a **precision matrix** $\boldsymbol{\beta}$:

$$\mathcal{N}(x; \boldsymbol{\mu}, \boldsymbol{\beta}^{\mathbf{-1}}) = \sqrt{\frac{\det(\boldsymbol{\beta})}{(2\pi)^n}} \exp\Big(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\beta} (\boldsymbol{x} - \boldsymbol{\mu}) \Big). \tag{21}$$

We often fix the covariance matrix to be a diagonal matrix. An even simpler version is the isotropic Gaussian distribution, whose covariance matrix is a scalar times the identity matrix.

□

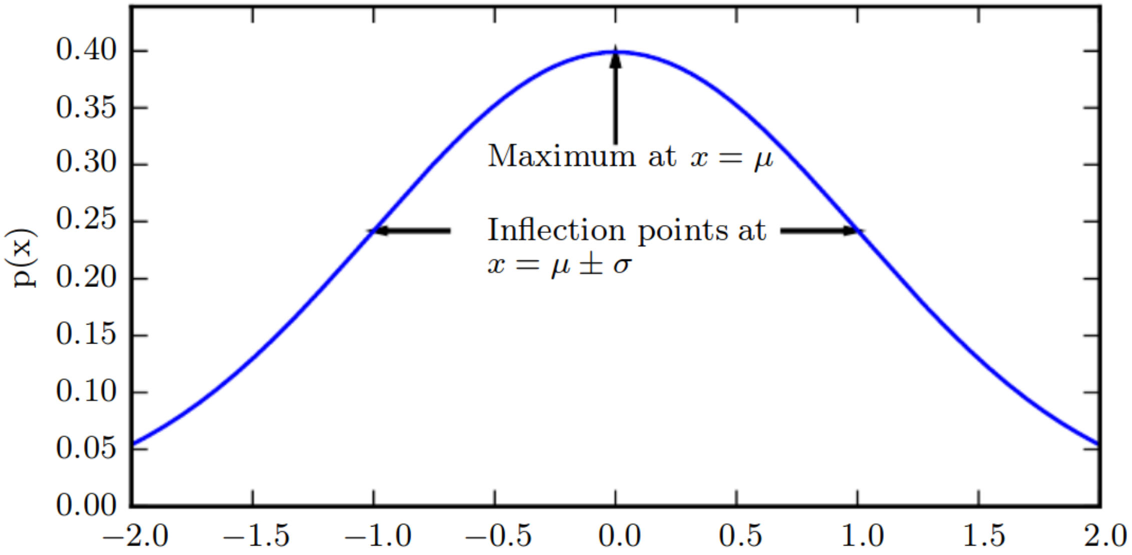

Figure 1: **The normal distribution**: The normal distribution $\mathcal{N}(x; \mu, \sigma^2)$ exhibits a classic “bell curve” shape, with the x coordinate of its central peak given by μ, and the width of its peak controlled by σ. In this example, we depict the standard normal distribution, with $\mu = 0$ and $\sigma = 1$.

Exercise 19. Why is the normal distribution a default choice for a prior over a set of real numbers?

Proof. Normal distributions are a sensible choice for many applications. In the absence of prior knowledge about what form a distribution over the real numbers should take, the normal distribution is a good default choice for two major reasons.

First, many distributions we wish to model are truly close to being normal distributions. The central limit theorem shows that the sum of many independent random variables is approximately normally distributed. This means that in practice, many complicated systems can be modeled successfully as normally distributed noise, even if the system can be decomposed into parts with more structured behavior.

Second, out of all possible probability distributions with the same variance, the normal distribution encodes the maximum amount of uncertainty over the real numbers. We can thus think of the normal distribution as being the one that inserts the least amount of prior knowledge into a model.

□

Exercise 20. What is the central limit theorem?

Proof. The central limit theorem shows that the sum of many independent random variables is approximately normally distributed.

Or more formally put: given a population with a finite mean μ and a finite non-zero variance σ^2, the sampling distribution of the mean approaches a normal distribution with a mean of μ and a variance of $\frac{\sigma^2}{N}$ as N, the sample size, increases. □

Exercise 21. What are exponential and Laplace distribution?

Proof. In the context of deep learning, we often want to have a probability distribution with a sharp point at $x = 0$. To accomplish this, we can use the **exponential distribution**:

$$p(x; \lambda) = \lambda \mathbf{1}_{x \geq 0} \exp(-\lambda x). \tag{22}$$

The exponential distribution uses the indicator function $\mathbf{1}_{x \geq 0}$ to assign probability zero to all negative values of x.

A closely related probability distribution that allows us to place a sharp peak of probability mass at an arbitrary point μ is the **Laplace distribution**

$$\text{Laplace}(x; \mu, \gamma) = \frac{1}{2\gamma} \exp\Big(- \frac{|x - \mu|}{\gamma} \Big). \tag{23}$$

□

Exercise 22. What are Dirac distribution and Empirical distribution?

Proof. In some cases, we wish to specify that all of the mass in a probability distribution clusters around a single point. This can be accomplished by defining a PDF using **the Dirac delta function**, $\delta(x)$:

$$p(x) = \delta(x - \mu). \tag{24}$$

The Dirac delta function is defined such that it is zero-valued everywhere except 0, yet integrates to 1. The Dirac delta function is not an ordinary function that associates each value x with a real-valued output, instead it is a different kind of mathematical object called a generalized function that is defined in terms of its properties when integrated. By defining $p(x)$ to be δ shifted by $-\mu$ we obtain an infinitely narrow and infinitely high peak of probability mass where $x = \mu$.

A common use of the Dirac delta distribution is as a component of an **empirical distribution**,

$$\hat{p}(\boldsymbol{x}) = \frac{1}{m} \sum_{i=1}^{m} \delta(\boldsymbol{x} - \boldsymbol{x}^{(i)}) \tag{25}$$

Example. Let $\boldsymbol{X}$ be a random variable with the following CDF:

$$F_{\boldsymbol{X}}(x) = \begin{cases} \frac{1}{2} + \frac{1}{2}(1 - e^{-x}) & x \geq 1 \\ \frac{1}{4} + \frac{1}{2}(1 - e^{-x}) & 0 \leq x < 1 \\ 0 & x < 0 \end{cases}$$

Now we want to find the (generalized) PDF of $\boldsymbol{X}$. To find the PDF, we need to differentiate the CDF. We must be careful about the points of discontinuity. In particular, we have two jumps: one at $x = 0$ and one at $x = 1$. The size of the jump for both points is equal to $\frac{1}{4}$. Thus, the CDF has two delta functions: $\frac{1}{4}\delta(x) + \frac{1}{4}\delta(x - 1)$. The continuous part of the CDF can be written as $\frac{1}{2}e^{-x}u(x)$ for $x > 0$. Therefore, we can conclude:

$$f_{\boldsymbol{X}}(x) = \frac{1}{4}\delta(x) + \frac{1}{4}\delta(x - 1) + \frac{1}{2}e^{-x}u(x)$$

.

□

Exercise 23. What is mixture of distributions?

Proof. It is also common to define probability distributions by combining other simpler probability distributions. One common way of combining distributions is to construct a **mixture distribution**. A mixture distribution is made up of several component distributions. On each trial, the choice of which component distribution generates the sample is determined by sampling a component identity from a multinoulli distribution:

$$P(\mathrm{x}) = \sum_i P(\mathrm{c} = i) P(\mathrm{x} \mid \mathrm{c} = i) \tag{26}$$

where $P(\mathrm{c})$ is the multinoulli distribution over component identities. □

Exercise 24. Name two common examples of mixture of distributions?

Proof. We have already seen one example of a mixture distribution: the empirical distribution over real-valued variables is a mixture distribution with one Dirac component for each training example.

A very powerful and common type of mixture model is the **Gaussian mixture model**, in which the components $p(\mathbf{x} \mid \mathrm{c} = i)$ are Gaussians. Each component has a separately parametrized mean $\boldsymbol{\mu}^{(i)}$ and covariance $\boldsymbol{\Sigma}^{(i)}$. Some mixtures can have more constraints. For example, the covariances could be shared across components via the constraint $\boldsymbol{\Sigma}^{(i)} = \boldsymbol{\Sigma}, \forall i$. As with a single Gaussian distribution, the mixture of Gaussians might constrain the covariance matrix for each component to be diagonal or isotropic.

In addition to the means and covariances, the parameters of a Gaussian mixture specify the **prior probability** $\alpha_i = P(\mathrm{c} = i)$ given to each component i. The word "prior" indicates that it expresses the model's beliefs about c before it has observed $\mathbf{x}$. By comparison, $P(\mathrm{c} \mid \boldsymbol{x})$ is a posterior probability, because it is computed after observation of $\mathbf{x}$.

Note. The mixture model is one simple strategy for combining probability distributions to create a richer distribution. The mixture model allows us to briefly glimpse a concept that will be of paramount importance later—the latent variable. A latent variable is a random variable that we cannot observe directly. The component identity variable c of the mixture model provides an example. Latent variables may be related to x through the joint distribution, in this case, $P(\mathrm{x,c}) = P(\mathrm{x}|\mathrm{c})P(\mathrm{c})$. The distribution $P(\mathrm{c})$ over the latent variable and the distribution $P(\mathrm{x}|\mathrm{c})$ relating the latent variables to the visible variables determines the shape of the distribution $P(\mathrm{x})$ even though it is possible to describe $P(\mathrm{x})$ without reference to the latent variable.

□

Exercise 25. Is Gaussian mixture model a universal approximator of densities?

Proof. A Gaussian mixture model is a universal approximator of densities, in the sense that any smooth density can be approximated with any specific, non-zero amount of error by a Gaussian mixture model with enough components.

□

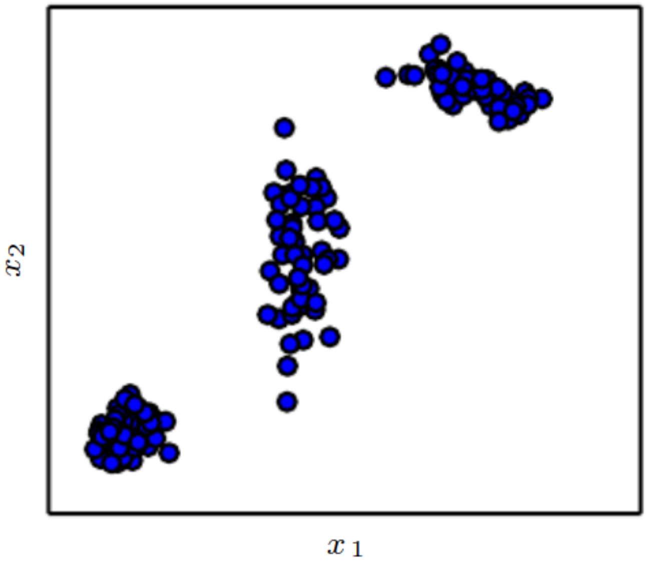

Figure 2: Samples from a Gaussian mixture model. In this example, there are three components. From left to right, the first component has an isotropic covariance matrix, meaning it has the same amount of variance in each direction. The second has a diagonal covariance matrix, meaning it can control the variance separately along each axis-aligned direction. This example has more variance along the x_2 axis than along the x_1 axis. The third component has a full-rank covariance matrix, allowing it to control the variance separately along an arbitrary basis of directions.

Exercise 26. Write the formulae for logistic and softplus function.

Proof. Certain functions arise often while working with probability distributions, especially the probability distributions used in deep learning models. One of these functions is the **logistic sigmoid**:

$$\sigma(x) = \frac{1}{1 + \exp(-x)}. \tag{27}$$

The logistic sigmoid is commonly used to produce the ϕ parameter of a Bernoulli distribution because its range is (0,1), which lies within the valid range of values for the ϕ parameter. The sigmoid function **saturates** when its argument is very positive or very negative, meaning that the function becomes very flat and insensitive to small changes in its input.

Another commonly encountered function is the **softplus function**:

$$\xi(x) = \log(1 + \exp(x)) \tag{28}$$

The softplus function can be useful for producing the β or σ parameter of a normal distribution because its range is $(0, \infty)$. It also arises commonly when manipulating expressions involving sigmoids. The name of the softplus function comes from the fact that it is a smoothed or “softened” version of

$$x^+ = \max(0, x). \tag{29}$$

The following properties are all useful enough that you may wish to memorize them:

$$\sigma(x) = \frac{\exp(x)}{\exp(x) + \exp(0)} \tag{30}$$

$$\frac{d}{dx}\sigma(x) = \sigma(x)(1 - \sigma(x)) \tag{31}$$

$$1 - \sigma(x) = \sigma(-x) \tag{32}$$

$$\log \sigma(x) = -\xi(-x) \tag{33}$$

$$\frac{d}{dx}\xi(x) = \sigma(x) \tag{34}$$

$$\forall x \in (0, 1), \sigma^{-1}(x) = \log\left(\frac{x}{1 - x}\right) \tag{35}$$

$$\forall x > 0, \xi^{-1}(x) = \log\left(\exp(x) - 1\right)) \tag{36}$$

$$\xi(x) = \int_{-\infty}^{x} \sigma(y)dy \tag{37}$$

$$\xi(x) - \xi(-x) = x \tag{38}$$

The function $\sigma^{-1}(x)$ is called the **logit** in statistics, but this term is more rarely used in machine learning.

Equation 36 provides extra justification for the name “softplus.” The softplus function is intended as a smoothed version of the **positive part function**, $x^+ = \max\{0, x\}$. The positive part function is the counterpart of the **negative part** function, $x^- = \max\{0, x\}$. To obtain a smooth function that is analogous to the negative part, one can use $\xi(-x)$. Just as x can be recovered from its positive part and negative part via the identity $x^+ x = x$, it is also possible to recover x using the same relationship between $\xi(x)$ and $\xi(-x)$, as shown in equation 38.

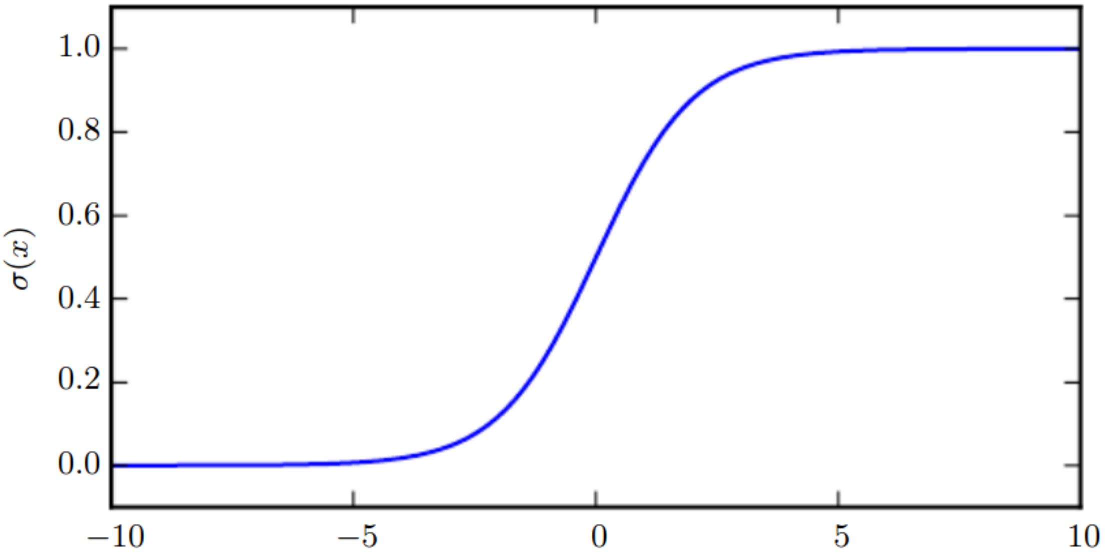

Figure 3: The logistic sigmoid function

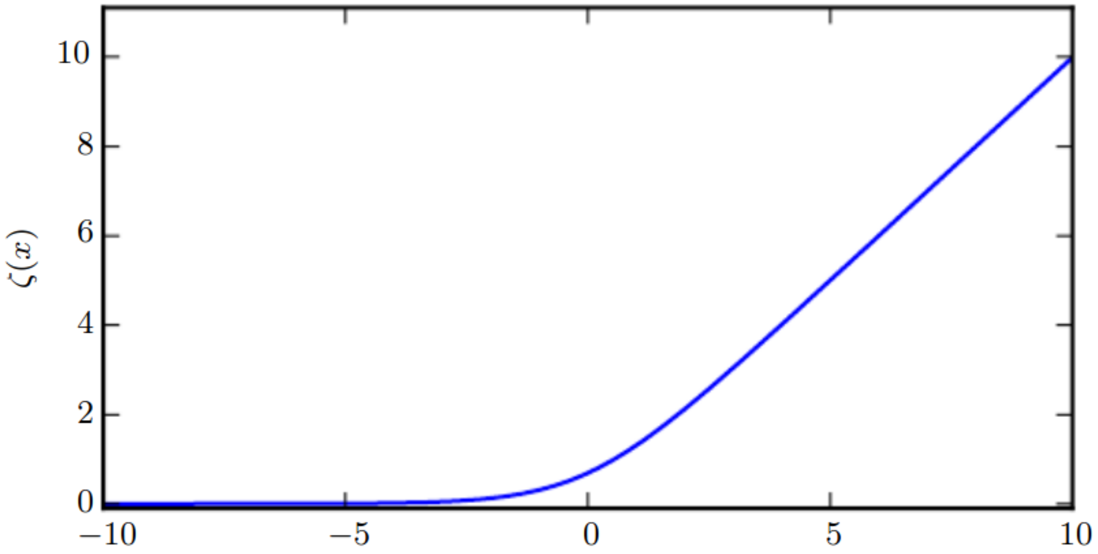

Figure 4: The softplus function

□

Exercise 27. Write the formulae for Bayes rule

Proof. We often find ourselves in a situation where we know $P(\mathrm{y} \mid \mathrm{x})$ and need to know $P(\mathrm{x} \mid \mathrm{y})$. Fortunately, if we also know $P(\mathrm{x})$, we can compute the desired quantity using **Bayes' rule**:

$$P(\mathrm{x} \mid \mathrm{y}) = \frac{P(\mathrm{x})P(\mathrm{y}|\mathrm{x})}{P(\mathrm{y})}. \tag{39}$$

Note that while $P(\mathrm{y})$ appears in the formula, it is usually feasible to compute $P(\mathrm{y}) = \sum_x P(\mathrm{y} \mid x)P(x)$, so we do not need to begin with knowledge of $P(\mathrm{y})$. □

Exercise 28. What do you mean by measure zero and almost everywhere?

Proof. For our purposes, measure theory is more useful for describing theorems that apply to most points in R n but do not apply to some corner cases. Measure theory provides a rigorous way of describing that a set of points is negligibly small. Such a set is said to have **measure zero**. For our purposes, it is sufficient to understand the intuition that a set of measure zero occupies no volume in the space we are measuring.

Example. Within $\mathbb{R}^2$, a line has measure zero, while a filled polygon has positive measure. Likewise, an individual point has measure zero. Any union of countably many sets that each have measure zero also has measure zero (so the set of all the rational numbers has measure zero, for instance).

Another useful term from measure theory is **almost everywhere**. A property that holds almost everywhere holds throughout all of space except for on a set of measure zero. Because the exceptions occupy a negligible amount of space, they can be safely ignored for many applications. Some important results in probability theory hold for all discrete values but only hold "almost everywhere" for continuous values.

□

Exercise 29. If two random variables are related in a deterministic way, how are the PDFs related?

Proof. For real-valued vectors $\boldsymbol{x}$ and $\boldsymbol{y}$,

$$p_x(\boldsymbol{x}) = p_y(g(\boldsymbol{x}))\left|\det\left(\frac{\partial g(\boldsymbol{x})}{\partial \boldsymbol{x}}\right)\right|. \tag{40}$$

Note. Deterministic way means that the successive states are completely determined by the preceding states. in this case, the pdf $(p_{\boldsymbol{x}})$ is dependent on the pdf $(p_{\boldsymbol{y}})$.

□

Exercise 30. Define self-information. What are its units?

Proof. The basic intuition behind information theory is that learning that an unlikely event has occurred is more informative than learning that a likely event has occurred. A message saying “the sun rose this morning” is so uninformative as to be unnecessary to send, but a message saying “there was a solar eclipse this morning” is very informative.

We would like to quantify information in a way that formalizes this intuition. Specifically,

• Likely events should have low information content, and in the extreme case, events that are guaranteed to happen should have no information content whatsoever.

• Less likely events should have higher information content.

• Independent events should have additive information. For example, finding out that a tossed coin has come up as heads twice should convey twice as much information as finding out that a tossed coin has come up as heads once.

In order to satisfy all three of these properties, we define the **self information** of an event $\mathrm{x} = x$ to be

$$I(x) = -\log P(x). \tag{41}$$

Note. log is the natural logarithm with base e. Our definition of $I(x)$ is therefore written in units of **nats**. One nat is the amount of information gained by observing an event of probability $\frac{1}{e}$.

Note. Other texts use base-2 logarithms and units called **bits** or **shannons**; information measured in bits is just a rescaling of information measured in nats.

When x is continuous, we use the same definition of information by analogy, but some of the properties from the discrete case are lost. For example, an event with unit density still has zero information, despite not being an event that is guaranteed to occur.

Example. With 1 bit, we can represent 2 different facts (i.e. **information**), either as a 1 or a 0 (i.e. True or False).

Let’s say you are a commander in World War II in 1945. Your telegrapher told you if the Nazis surrender, then he will send you a ’1’. Otherwise, if they do not, then he will send you a ’0’.

Now, in 2018, you can send the exact same information on a smartphone typing:

”The war is over” (instead of 1 bit, we use 8 bits * 15 characters = **120 bits**)

”The war is not over” (8 bits * 19 characters = **152 bits**)

So we are using more than a 100 bits to send a message **that could be reduced to just one bit**.

Now, let’s say there are four possible war outcomes tomorrow instead of two:

1) Germany and Japan both surrender.

2) Germany surrenders but Japan does not.

3) Japan surrenders but Germany does not.

4) Both do not surrender.

Now your telegrapher would need 2 bits (00, 01, 10, 11) to encode this message. In the same way, he would need only 8 bits if there were 256 different scenarios.

So, think of x (the variable) as the **news from the telegrapher**. The news can be anything— it does not have to be 4 states, 256 states, etc. In real life, news can be millions of different facts.

In terms of the equation above, $I(x)$ is the information content of x. $I(x)$ itself is **a random variable**, which, in this example, are the possible outcomes of the War.

□

Exercise 31. What are Shannon entropy and differential entropy?

Proof. Self-information deals only with a single outcome. We can quantify the amount of uncertainty in an entire probability distribution using the **Shannon entropy**:

$$H(\mathrm{x}) = \mathbb{E}_{\mathrm{x}\sim P}[I(x)] = -\,\mathbb{E}_{\mathrm{x}\sim P}[\log P(x)]. \tag{42}$$

also denoted $H(P)$. In other words, the Shannon entropy of a distribution is the expected amount of information in an event drawn from that distribution. It gives a lower bound on the number of bits (if the logarithm is base 2, otherwise the units are different) needed on average to encode symbols drawn from a distribution P. Distributions that are nearly deterministic (where the outcome is nearly certain) have low entropy; distributions that are closer to uniform have high entropy.

When x is continuous, the Shannon entropy is known as the **differential entropy**.

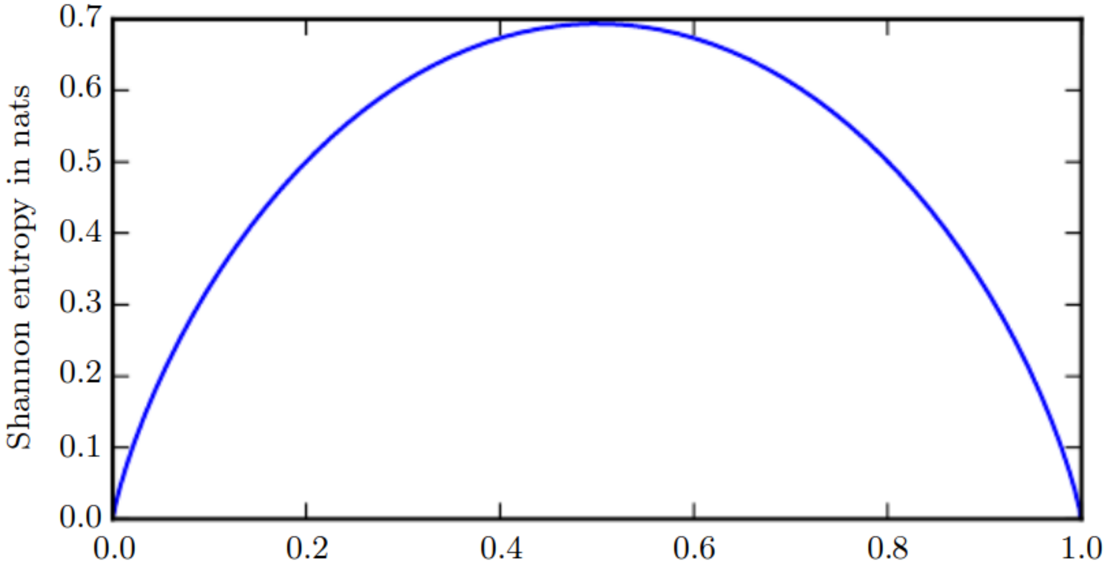

Figure 5: This plot shows how distributions that are closer to deterministic have low Shannon entropy while distributions that are close to uniform have high Shannon entropy. On the horizontal axis, we plot p, the probability of a binary random variable being equal to 1. The entropy is given by $(p-1)\log(1-p) - p\log p$. When p is near 0, the distribution is nearly deterministic, because the random variable is nearly always 0. When p is near 1, the distribution is nearly deterministic, because the random variable is nearly always 1. When $p = 0.5$, the entropy is maximal, because the distribution is uniform over the two outcomes.

Example. In laymen's terms, the **the entropy of a variable** is the "amount of information" contained in the variable. More formally put, the entropy for a probability distribution, $H(\mathrm{x})$, is **the expected value of every possible information** or the **total amount of information** in an **entire** probability distribution.

Using the definition of expected value, we can rewrite the above equation as:

$$H(\mathrm{x}) = -\mathbb{E}_{\mathrm{x}\sim P}[\log P(x)] = -\sum_x P(x)[\log P(x)]$$

$$= \sum_x P(x) \log \left(\frac{1}{P(x)}\right)$$

Now, we know that $P(x)$ is the probability of a case x occurring, which means that $\frac{1}{P(x)}$ is **the information of each case** (winning the war, losing the war, etc.).

So then, why is $\frac{1}{P(x)}$ the amount of information?

Say there is 50-50 chance that the Nazis would surrender ($p = \frac{1}{2}$). Then if your telegrapher tells you that they did surrender, you can eliminate the uncertainty of total 2 events (both surrender and not surrender), which is the reciprocal of ($p = \frac{1}{2}$).

In short, when your events are all equally likely to happen and you know one event happened, you can eliminate the possibility of every other event (total $\frac{1}{p}$ events) happening. For example, if there are four events and they are all equally likely to happen ($p = \frac{1}{4}$), then when one even happens the other three did not happen. Thus, we know what happened to total of 4 events.

But what about events that are not equally likely?

Say there is a 75% chance that the Nazis will surrender and there is a 25% chance that they will not.

How much information does the event "surrender" have? ($\log_2\left(\frac{1}{0.75}\right) = \log_2(1.333) = 0.41$)

How much information does the event "not surrender" have? ($\log_2\left(\frac{1}{0.25}\right) = \log_2(4) = 2$)

As you can see, the unlikely event has a higher entropy.

Let's show a diagram explaining this more intuitively about why **information is the reciprocal of the probability**:

Figure 6: Black dot denotes: Nazis won ($\frac{1}{4}$), White dots denote: America won ($\frac{3}{4}$)

Say the Nazis won World War II. The black dot is the news. By knowing the black dot, we can eliminate the other 3 white dots at the same time. This is a total of ($\frac{1}{\frac{1}{4}}$ =) 4 dots (total amount of information) burst.

The total amount of information you can burst = information content in each EACH news.

Now, let's look at the flip side:

Figure 7: Black dots denote: America won ($\frac{3}{4}$), White dot denotes: Nazis won ($\frac{1}{4}$)

Now, by knowing America won the war, we can represent the 3 black dots as ONE black dot. So how many TOTAL dots can we burst?

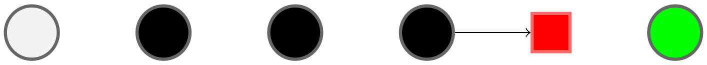

Figure 8: Red Square denotes: $\frac{1}{3}$ of black dot, Green dot denotes: ONE black dot

We can eliminate a total of 1 and $\frac{1}{3} = \frac{4}{3}$ dots, which is **the reciprocal of** $\frac{3}{4}$.

Thus, **the information in EVERY possible news** is: $\frac{1}{4}\log(4) + \frac{3}{4}\log_2\left(\frac{4}{3}\right) = 0.81$ (Via's Shannon's entropy formula).

This explains where $\frac{1}{p}$ comes from. Shannon thought that the information content of anything can be measured in bits. To write a number N in bits, we need to take a $\log_2 N$. Hence, why we use $\log_2$ or log.

TL;DR 1. If we have $P(\text{win}) = 1$, then the entropy $= 0$. It has 0 bits of uncertainty ($-\log_2 1 = 0$). Notice, that in the example, with the "equally likely" messages, the entropy is higher (2 bits) than the "not equally likely" messages (0.81 bits). This is because there is less uncertainty in "not equally likely" messages. One event is more likely to come up than the other. This reduces the uncertainty.

Note. The thermodynamic "entropy" and the "entropy" in information theory both capture increasing randomness.

□

Exercise 32. What is Kullback-Leibler (KL) divergence?

Proof. If we have two separate probability distributions $P(\mathrm{x})$ and $Q(\mathrm{x})$ over the same random variable x, we can measure how different these two distributions are using the **Kullback-Leibler (KL) divergence**:

$$D_{KL}(P||Q) = \mathbb{E}_{\mathrm{x}\sim P}\left[\log\frac{P(x)}{Q(x)}\right] = \mathbb{E}_{\mathrm{x}\sim P}[\log P(x) - \log Q(x)] \tag{43}$$

In the case of discrete variables, it is the extra amount of information (measured in bits if we use the base 2 logarithm, but in machine learning we usually use nats and the natural logarithm) needed to send a message containing symbols drawn from probability distribution P, when we use a code that was designed to minimize the length of messages drawn from probability distribution Q. □

Exercise 33. Can KL divergence be used as a distance measure?

Proof. The KL divergence has many useful properties, most notably that it is non-negative. The KL divergence is 0 if and only if P and Q are the same distribution in the case of discrete variables, or equal "almost everywhere" in the case of continuous variables. Because the KL divergence is non-negative and measures the difference between two distributions, it is often conceptualized as measuring some sort of distance between these distributions. However, it is not a true distance measure because it is not symmetric: $D_{KL}(P||Q) \neq D_{KL}(Q||P)$ for some P and Q. This asymmetry means that there are important consequences to the choice of whether to use $D_{KL}(P||Q)$ or $D_{KL}(Q||P)$.

□

Divergence.png

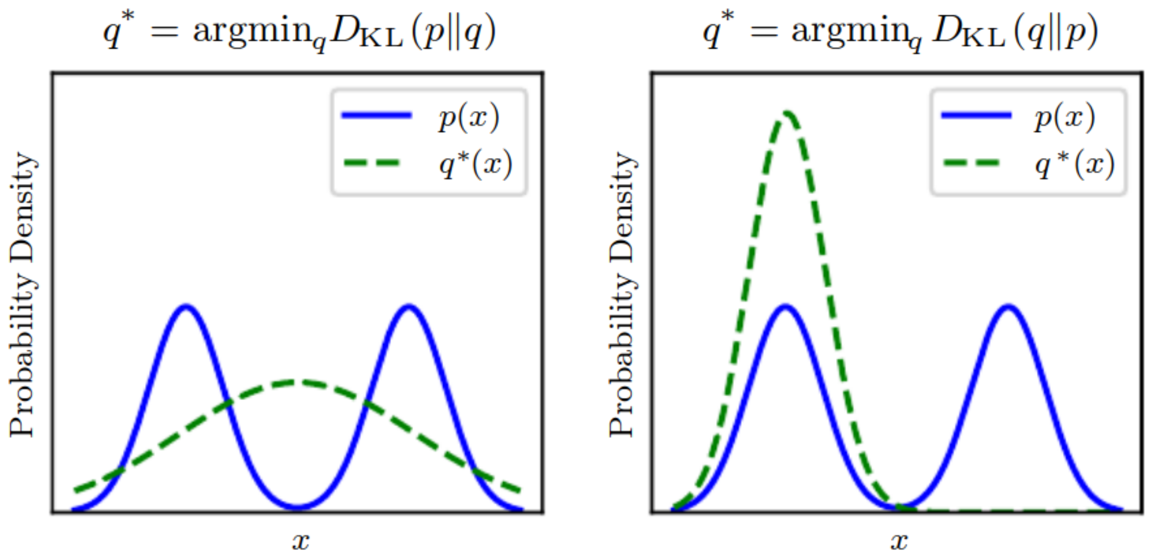

Figure 9: The KL divergence is asymmetric. Suppose we have a distribution $p(x)$ and wish to approximate it with another distribution $q(x)$. We have the choice of minimizing either $D_{KL}(p||q)$ or $D_{KL}(q||p)$. We illustrate the effect of this choice using a mixture of two Gaussians for p, and a single Gaussian for q. The choice of which direction of the KL divergence to use is problem-dependent. Some applications require an approximation that usually places high probability anywhere that the true distribution places high probability, while other applications require an approximation that rarely places high probability anywhere that the true distribution places low probability. The choice of the direction of the KL divergence reflects which of these considerations takes priority for each application. *(Left)* The effect of minimizing$D_{KL}(p||q)$ In this case, we select a q that has high probability where p has high probability. When p has multiple modes, q chooses to blur the modes together, in order to put high probability mass on all of them. *(Right)* The effect of minimizing $D_{KL}(q||p)$. In this case, we select a q that has low probability where p has low probability. When p has multiple modes that are sufficiently widely separated, as in this figure, the KL divergence is minimized by choosing a single mode, in order to avoid putting probability mass in the low-probability areas between modes of p. Here, we illustrate the outcome when q is chosen to emphasize the left mode. We could also have achieved an equal value of the KL divergence by choosing the right mode. If the modes are not separated by a sufficiently strong low probability region, then this direction of the KL divergence can still choose to blur the modes.

Exercise 34. Define cross-entropy.

Proof. A quantity that is closely related to the KL divergence is the **cross-entropy** $H(P, Q) = H(P) + D_{KL}(P||Q)$, which is similar to the KL divergence but lacking the term on the left:

$$H(P, Q) = -\mathbb{E}_{\mathrm{x}\sim P} \log Q(x). \tag{44}$$

Minimizing the cross-entropy with respect to Q is equivalent to minimizing the KL divergence, because Q does not participate in the omitted term.

Note. When computing many of these quantities, it is common to encounter expressions of the form $0 \log 0$. By convention, in the context of information theory, we treat these expressions as $\lim_{x\to} x \log x = 0$.

□

Exercise 35. What are structured probabilistic models or graphical models?

Proof. Machine learning algorithms often involve probability distributions over a very large number of random variables. Often, these probability distributions involve direct interactions between relatively few variables. Using a single function to describe the entire joint probability distribution can be very inefficient (both computationally and statistically).

Instead of using a single function to represent a probability distribution, we can split a probability distribution into many factors that we multiply together.

Example. Suppose we have three random variables: a, b and c. Suppose that a influences the value of b and b influences the value of c, but that a and c are independent given b. We can represent the probability distribution over all three variables as a product of probability distributions over two variables:

$$p(\mathrm{a}, \mathrm{b}, \mathrm{c}) = p(\mathrm{a})p(\mathrm{b} \,|\, \mathrm{a})p(\mathrm{c} \,|\, \mathrm{b}). \tag{45}$$

These factorizations can greatly reduce the number of parameters needed to describe the distribution. Each factor uses a number of parameters that is exponential in the number of variables in the factor. This means that we can greatly reduce the cost of representing a distribution if we are able to find a factorization into distributions over fewer variables

We can describe these kinds of factorizations using graphs. Here we use the word "graph" in the sense of graph theory: a set of vertices that may be connected to each other with edges. When we represent the factorization of a probability distribution with a graph, we call it a **structured probabilistic model** or **graphical model**. □

Exercise 36. In the context of structured probabilistic models, what are directed and undirected models? How are they represented? What are cliques in undirected structured probabilistic models?

Proof. There are two main kinds of structured probabilistic models: directed and undirected. Both kinds of graphical models use a graph $\mathcal{G}$ in which each node in the graph corresponds to a random variable, and an edge connecting two random variables means that the probability distribution is able to represent direct interactions between those two random variables.

Directed models use graphs with directed edges, and they represent factorizations into conditional probability distributions, as in the example above. Specifically, a directed model contains one factor for every random variable x_i in the distribution, and that factor consists of the conditional distribution over x_i given the parents of x_i , denoted $Pa_{\mathcal{G}}(\mathrm{x}_\mathrm{i})$:

$$p(\mathbf{x}) = \prod_i p(\mathrm{x}_i | Pa_{\mathcal{G}}(\mathrm{x}_\mathrm{i})). \tag{46}$$

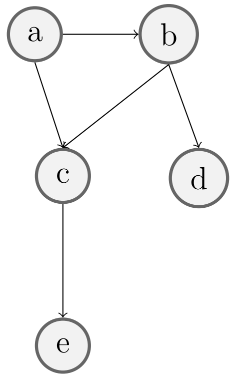

Figure 10: A directed graphical model over random variables a, b, c, d and e. This graph corresponds to probability distributions that can be factored as

$$p(\mathrm{a}, \mathrm{b}, \mathrm{c}, \mathrm{d}, \mathrm{e}) = p(\mathrm{a})p(\mathrm{b} \mid \mathrm{a})p(\mathrm{c} \mid \mathrm{a}, \mathrm{b})p(\mathrm{d} \mid \mathrm{b})p(\mathrm{e} \mid \mathrm{c}) \tag{47}$$

This graph allows us to quickly see some properties of the distribution. For example, a and c interact directly, but a and e interact only indirectly via c.

Undirected models use graphs with undirected edges, and they represent factorizations into a set of functions; unlike in the directed case, these functions are usually not probability distributions of any kind. Any set of nodes that are all connected to each other in $\mathcal{G}$ is called a clique. Each clique $\mathcal{C}^{(i)}$ in an undirected model is associated with a factor $\phi^{(i)}(\mathcal{C}^{(i)})$. These factors are just functions, not probability distributions. The output of each factor must be non-negative, but there is no constraint that the factor must sum or integrate to 1 like a probability distribution.

The probability of a configuration of random variables is **proportional** to the product of all of these factors—assignments that result in larger factor values are more likely. Of course, there is no guarantee that this product will sum to 1. We therefore divide by a normalizing constant Z , defined to be the sum or integral over all states of the product of the ϕ functions, in order to obtain a normalized probability distribution:

$$p(\mathbf{x}) = \frac{1}{Z} \prod_i \phi^{(i)}(\mathcal{C}^{(i)}). \tag{48}$$

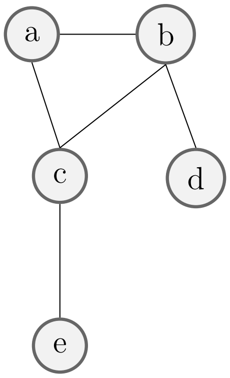

Figure 11: An undirected graphical model over random variables a, b, c, d and e. This graph corresponds to probability distributions that can be factored as

$$p(\mathrm{a}, \mathrm{b}, \mathrm{c}, \mathrm{d}, \mathrm{e}) = \frac{1}{Z}\phi^{(1)}(\mathrm{a}, \mathrm{b}, \mathrm{c})\phi^{(2)}(\mathrm{b}, \mathrm{d})\phi^{(3)}(\mathrm{c}, \mathrm{e}) \quad (49)$$

This graph allows us to quickly see some properties of the distribution. For example, a and c interact directly, but a and e interact only indirectly via c.

Keep in mind that these graphical representations of factorizations are a language for describing probability distributions. They are not mutually exclusive families of probability distributions. Being directed or undirected is not a property of a probability distribution; it is a property of a particular **description** of a probability distribution, but any probability distribution may be described in both ways. □