Appendix C

Probability

Probability is critical to deep learning. In supervised learning, deep networks implicitly rely on a probabilistic formulation of the loss function. In unsupervised learning, generative models aim to produce samples that are drawn from the same probability distribution as the training data. Reinforcement learning occurs within Markov decision processes, and these are defined in terms of probability distributions. This appendix provides a primer for probability as used in machine learning.

C.1 Random variables and probability distributions

A random variable x denotes a quantity that is uncertain. It may be *discrete* (take only certain values, for example integers) or *continuous* (take any value on a continuum, for example real numbers). If we observe several instances of a random variable x, it will take different values, and the relative propensity to take different values is described by a *probability distribution* $Pr(x)$.

For a discrete variable, this distribution associates a *probability* $Pr(x=k) \in [0,1]$ with each potential outcome k, and the sum of these probabilities is one. For a continuous variable, there is a non-negative *probability density* $Pr(x=a) \geq 0$ associated with each value a in the domain of x, and the integral of this probability density function (PDF) over this domain must be one. This density can be greater than one for any point a. From here on, we assume that the random variables are continuous. The ideas are exactly the same for discrete distributions but with sums replacing integrals.

C.1.1 Joint probability

Consider the case where we have two random variables x and y. The *joint distribution* $Pr(x,y)$ tells us about the propensity that x and y take particular combinations of values (figure C.1a). Now there is a non-negative probability density $Pr(x=a, y=b)$ associated with each pair of values $x=a$ and $y=b$ and this must satisfy:

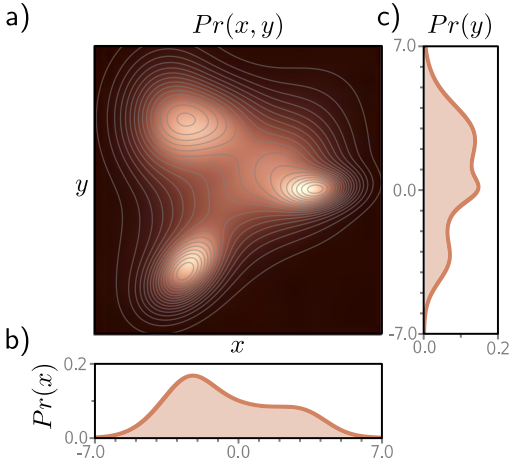

Figure C.1 Joint and marginal distributions. a) The joint distribution $Pr(x, y)$ captures the propensity of variables x and y to take different combinations of values. Here, the probability density is represented by the color map, so brighter positions are more probable. For example, the combination $x=6, y=6$ is much less likely to be observed than the combination $x = 5, y = 0$. b) The marginal distribution $Pr(x)$ of variable x can be recovered by integrating over y. c) The marginal distribution $Pr(y)$ of variable y can be recovered by integrating over x.

$$\iint Pr(x, y) \cdot dxdy = 1. \tag{C.1}$$

This idea extends to more than two variables, so the joint density of x, y, and z is written as $Pr(x, y, z)$. Sometimes, we store multiple random variables in a vector $\mathbf{x}$, and we write their joint density as $Pr(\mathbf{x})$. Extending this, we can write the joint density of all of the variables in two vectors $\mathbf{x}$ and $\mathbf{y}$ as $Pr(\mathbf{x}, \mathbf{y})$.

C.1.2 Marginalization

If we know the joint distribution $Pr(x, y)$ over two variables, we can recover the *marginal* distributions $Pr(x)$ and $Pr(y)$ by integrating over the other variable (figure C.1b-c):

$$\begin{aligned} \int Pr(x, y) \cdot dx &= Pr(y) \\ \int Pr(x, y) \cdot dy &= Pr(x). \end{aligned} \tag{C.2}$$

This process is called *marginalization* and has the interpretation that we are computing the distribution of one variable *regardless* of the value the other one took. The idea of marginalization extends to higher dimensions, so if we have a joint distribution $Pr(x, y, z)$, we can recover the joint distribution $Pr(x, z)$ by integrating over y.

C.1.3 Conditional probability and likelihood

The *conditional probability* $Pr(x|y)$ is the probability of variable x taking a certain value, assuming we know the value of y. The vertical line is read as the English word "given,"

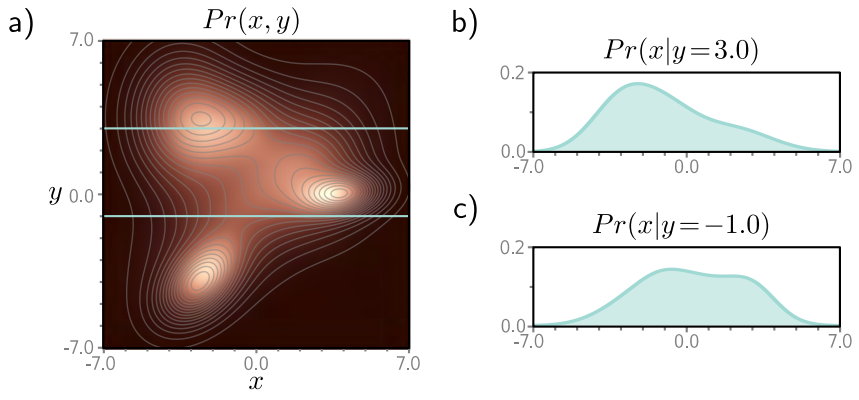

Figure C.2 Conditional distributions. a) Joint distribution $Pr(x, y)$ of variables x and y. b) The conditional probability $Pr(x|y = 3.0)$ of variable x, given that y takes the value 3.0, is found by taking the horizontal "slice" $Pr(x, y = 3.0)$ of the joint probability (top cyan line in panel a), and dividing this by the total area $Pr(y = 3.0)$ in that slice so that it forms a valid probability distribution that integrates to one. c) The joint probability $Pr(x, y=-1.0)$ is found similarly using the slice at $y=-1.0$.

so $Pr(x|y)$ is the probability of x given y. The conditional probability $Pr(x|y)$ can be found by taking a slice through the joint distribution $Pr(x, y)$ for a fixed y. This slice is then divided by the probability of that value y occurring (the total area under the slice) so that the conditional distribution sums to one (figure C.2):

$$Pr(x|y) = \frac{Pr(x, y)}{Pr(y)}. \tag{C.3}$$

Similarly,

$$Pr(y|x) = \frac{Pr(x, y)}{Pr(x)}. \tag{C.4}$$

When we consider the conditional probability $Pr(x|y)$ as a function of x, it must sum to one. When we consider the same quantity $Pr(x|y)$ as a function of y, it is termed the *likelihood* of x given y and does not have to sum to one.

C.1.4 Bayes' rule

From equations C.3 and C.4, we get two expressions for the joint probability $Pr(x, y)$:

$$Pr(x, y) = Pr(x|y)Pr(y) = Pr(y|x)Pr(x), \tag{C.5}$$

which we can rearrange to get:

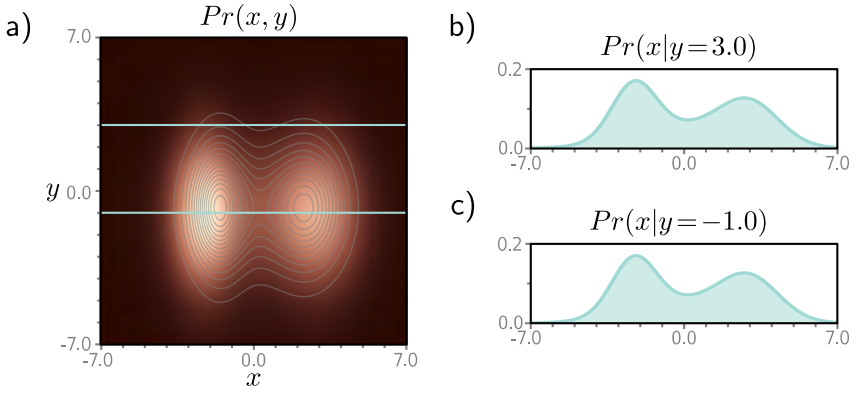

Figure C.3 Independence. a) When two variables x and y are independent, the joint distribution factors into the product of marginal distributions, so $Pr(x,y) = Pr(x)Pr(y)$. Independence implies that knowing the value of one variable tells us nothing about the other. b–c) Accordingly, all of the conditional distributions $Pr(x|y=\bullet)$ are the same and are equal to the marginal distribution $Pr(x)$.

$$Pr(x|y) = \frac{Pr(y|x)Pr(x)}{Pr(y)}. \tag{C.6}$$

This expression relates the conditional probability $Pr(x|y)$ of x given y to the conditional probability $Pr(y|x)$ of y given x and is known as *Bayes' rule*.

Each term in this Bayes' rule has a name. The term $Pr(y|x)$ is the *likelihood* of y given x, and the term $Pr(x)$ is the *prior probability* of x. The denominator $Pr(y)$ is known as the *evidence*, and the left-hand side $Pr(x|y)$ is termed the *posterior probability* of x given y. The equation maps from the prior $Pr(x)$ (what we know about x before observing y) to the posterior $Pr(x|y)$ (what we know about x after observing y).

C.1.5 Independence

If the value of the random variable y tells us nothing about x and vice-versa, we say that x and y are *independent*, and we can write $Pr(x|y) = Pr(x)$ and $Pr(y|x) = Pr(y)$. It follows that all of the conditional distributions $Pr(y|x=\bullet)$ are identical, as are the conditional distributions $Pr(x|y=\bullet)$.

Starting from the first expression for the joint probability in equation C.5, we see that the joint distribution becomes the product of the marginal distributions:

$$Pr(x,y) \quad = \quad Pr(x|y)Pr(y) = Pr(x)Pr(y) \tag{C.7}$$

when the variables are independent (figure C.3).

C.2 Expectation

Consider a function $\mathrm{f}[x]$ and a probability distribution $Pr(x)$ defined over x. The *expected value* of a function $\mathrm{f}[\bullet]$ of a random variable x with respect to the probability distribution $Pr(x)$ is defined as:

$$\mathbb{E}_x\big[\mathrm{f}[x]\big] = \int \mathrm{f}[x] Pr(x) dx. \tag{C.8}$$

As the name suggests, this is the expected or average value of $\mathrm{f}[x]$ after taking into account the probability of seeing different values of x. This idea generalizes to functions $\mathbf{f}[\bullet, \bullet]$ of more than one random variable:

$$\mathbb{E}_{x,y}\big[\mathrm{f}[x,y]\big] = \iint \mathrm{f}[x,y] Pr(x,y) dx dy. \tag{C.9}$$

An expectation is always taken with respect to a distribution over one or more variables. However, we don't usually make this explicit when the choice of distribution is obvious and write $\mathbb{E}[\mathrm{f}[x]]$ instead of $\mathbb{E}_x[\mathrm{f}[x]]$.

If we drew a large number I of samples $\{x_i\}_{i=1}^I$ from $Pr(x)$, calculated $\mathrm{f}[x_i]$ for each sample and took the average of these values, the result would approximate the expectation $\mathbb{E}[\mathrm{f}[x]]$ of the function:

$$\mathbb{E}_x\big[\mathrm{f}[x]\big] \approx \frac{1}{I}\sum_{i=1}^{I} \mathrm{f}[x_i]. \tag{C.10}$$

C.2.1 Rules for manipulating expectations

There are four rules for manipulating expectations:

$$\begin{aligned}
\mathbb{E}\big[k\big] &= k \\
\mathbb{E}\big[k \cdot \mathrm{f}[x]\big] &= k \cdot \mathbb{E}\big[\mathrm{f}[x]\big] \\
\mathbb{E}\big[\mathrm{f}[x] + \mathrm{g}[x]\big] &= \mathbb{E}\big[\mathrm{f}[x]\big] + \mathbb{E}\big[\mathrm{g}[x]\big] \\
\mathbb{E}_{x,y}\big[\mathrm{f}[x] \cdot \mathrm{g}[y]\big] &= \mathbb{E}_x\big[\mathrm{f}[x]\big] \cdot \mathbb{E}_y\big[\mathrm{g}[y]\big] \qquad \text{if } x, y \text{ independent,}
\end{aligned} \tag{C.11}$$

where k is an arbitrary constant. These are proven below for the continuous case.

Rule 1: The expectation $\mathbb{E}[k]$ of a constant value k is just k.

$$\begin{aligned}
\mathbb{E}\big[k\big] &= \int k \cdot Pr(x) dx \\
&= k \cdot \int Pr(x) dx \\
&= k.
\end{aligned}$$

Rule 2: The expectation $\mathbb{E}[k \cdot \mathrm{f}[x]]$ of a constant k times a function of the variable x is k times the expectation $\mathbb{E}[\mathrm{f}[x]]$ of the function:

$$\begin{aligned}\mathbb{E}\big[k \cdot \mathrm{f}[x]\big] &= \int k \cdot \mathrm{f}[x] Pr(x) dx \\ &= k \cdot \int \mathrm{f}[x] Pr(x) dx \\ &= k \cdot \mathbb{E}\big[\mathrm{f}[x]\big].\end{aligned}$$

Rule 3: The expectation of a sum $\mathbb{E}[\mathrm{f}[x] + \mathrm{g}[x]]$ of terms is the sum $\mathbb{E}[\mathrm{f}[x]] + \mathbb{E}[\mathrm{g}[x]]$ of the expectations:

$$\begin{aligned}\mathbb{E}\big[\mathrm{f}[x] + \mathrm{g}[x]\big] &= \int (\mathrm{f}[x] + \mathrm{g}[x]) \cdot Pr(x) dx \\ &= \int \big(\mathrm{f}[x] \cdot Pr(x) + \mathrm{g}[x] \cdot Pr(x)\big) dx \\ &= \int \mathrm{f}[x] \cdot Pr(x) dx + \int \mathrm{g}[x] \cdot Pr(x) dx \\ &= \mathbb{E}\big[\mathrm{f}[x]\big] + \mathbb{E}\big[\mathrm{g}[x]\big].\end{aligned}$$

Rule 4: The expectation of a product $\mathbb{E}[\mathrm{f}[x] \cdot \mathrm{g}[y]]$ of terms is the product $\mathbb{E}[\mathrm{f}[x]] \cdot \mathbb{E}[\mathrm{g}[y]]$ if x and y are independent.

$$\begin{aligned}\mathbb{E}\big[\mathrm{f}[x] \cdot \mathrm{g}[y]\big] &= \int\int \mathrm{f}[x] \cdot \mathrm{g}[y] Pr(x, y) dx dy \\ &= \int\int \mathrm{f}[x] \cdot \mathrm{g}[y] Pr(x) Pr(y) dx dy \\ &= \int \mathrm{f}[x] \cdot Pr(x) dx \int \mathrm{g}[y] \cdot Pr(y) dy \\ &= \mathbb{E}\big[\mathrm{f}[x]\big] \mathbb{E}\big[\mathrm{g}[y]\big],\end{aligned}$$

where we used the definition of independence (equation C.7) between the first two lines.

The four rules generalize to the multivariate case:

$$\begin{aligned}\mathbb{E}\big[\mathbf{A}\big] &= \mathbf{A} \\ \mathbb{E}\big[\mathbf{A} \cdot \mathbf{f}[\mathbf{x}]\big] &= \mathbf{A}\mathbb{E}\big[\mathbf{f}[\mathbf{x}]\big] \\ \mathbb{E}\big[\mathbf{f}[\mathbf{x}] + \mathbf{g}[\mathbf{x}]\big] &= \mathbb{E}\big[\mathbf{f}[\mathbf{x}]\big] + \mathbb{E}\big[\mathbf{g}[\mathbf{x}]\big] \\ \mathbb{E}_{\mathbf{x},\mathbf{y}}\big[\mathbf{f}[\mathbf{x}]^T \mathbf{g}[\mathbf{y}]\big] &= \mathbb{E}_{\mathbf{x}}\big[\mathbf{f}[\mathbf{x}]\big]^T \mathbb{E}_{\mathbf{y}}\big[\mathbf{g}[\mathbf{y}]\big] \qquad \text{if } \mathbf{x}, \mathbf{y} \text{ independent,} \qquad (C.12)\end{aligned}$$

where now $\mathbf{A}$ is a constant matrix and $\mathbf{f}[\mathbf{x}]$ is a function of the vector $\mathbf{x}$ that returns a vector, and $\mathbf{g}[\mathbf{y}]$ is a function of the vector $\mathbf{y}$ that also returns a vector.

C.2.2 Mean, variance, and covariance

For some choices of function f[•], the expectation is given a special name. These quantities are often used to summarize the properties of complex distributions. For example, when f[x] = x, the resulting expectation $\mathbb{E}[x]$ is termed the *mean*, μ. It is a measure of the center of a distribution. Similarly, the expected squared deviation from the mean $\mathbb{E}[(x-\mu)^2]$ is termed the *variance*, σ^2. This is a measure of the spread of the distribution. The *standard deviation* σ is the positive square root of the variance. It also measures the spread of the distribution but has the merit that it is expressed in the same units as the variable x.

As the name suggests, the *covariance* $\mathbb{E}[(x-\mu_x)(y-\mu_y)]$ of two variables x and y measures the degree to which they co-vary. Here μ_x and μ_y represent the mean of the variables x and y, respectively. The covariance will be large when the variance of both variables is large and when the value of x tends to increase when the value of y increases.

If two variables are independent, then their covariance is zero. However, a covariance of zero does not imply independence. For example, consider a distribution $Pr(x,y)$ where the probability is uniformly distributed on a circle of radius one centered on the origin of the x, y plane. There is no tendency on average for x to increase when y increases or vice-versa. However, knowing the value of $x = 0$ tells us that y has an equal chance of taking the values ± 1, so the variables cannot be independent.

The covariances of multiple random variables stored in a column vector $\mathbf{x} \in \mathbb{R}^D$ can be represented by the $D \times D$ *covariance matrix* $\mathbb{E}[(\mathbf{x}-\boldsymbol{\mu}_x)(\mathbf{x}-\boldsymbol{\mu}_x)^T]$, where the vector $\boldsymbol{\mu}_x$ contains the means $\mathbb{E}[\mathbf{x}]$. The element at position (i,j) of this matrix represents the covariance between variables x_i and x_j.

C.2.3 Variance identity

The rules of expectation (appendix C.2.1) can be used to prove the following identity that allows us to write the variance in a different form:

$$\mathbb{E}\big[(x-\mu)^2\big] = \mathbb{E}\big[x^2\big] - \mathbb{E}\big[x\big]^2. \tag{C.13}$$

Proof:

$$\begin{aligned}
\mathbb{E}\big[(x-\mu)^2\big] &= \mathbb{E}\big[x^2 - 2\mu x + \mu^2\big] \\
&= \mathbb{E}\big[x^2\big] - \mathbb{E}\big[2\mu x\big] + \mathbb{E}\big[\mu^2\big] \\
&= \mathbb{E}\big[x^2\big] - 2\mu \cdot \mathbb{E}\big[x\big] + \mu^2 \\
&= \mathbb{E}\big[x^2\big] - 2\mu^2 + \mu^2 \\
&= \mathbb{E}\big[x^2\big] - \mu^2 \\
&= \mathbb{E}\big[x^2\big] - \mathbb{E}\big[x\big]^2,
\end{aligned} \tag{C.14}$$

where we have used rule 3 between lines 1 and 2, rules 1 and 2 between lines 2 and 3, and the definition $\mu = \mathbb{E}[x]$ in the remaining two lines.

C.2.4 Standardization

Setting the mean of a random variable to zero and the variance to one is known as *standardization*. This is achieved using the transformation:

$$z = \frac{x-\mu}{\sigma}, \tag{C.15}$$

where μ is the mean of x and σ is the standard deviation.

Proof: The mean of the new distribution over z is given by:

$$\begin{aligned}
\mathbb{E}[z] &= \mathbb{E}\left[\frac{x-\mu}{\sigma}\right] \\
&= \frac{1}{\sigma}\mathbb{E}\big[x-\mu\big] \\
&= \frac{1}{\sigma}\left(\mathbb{E}\big[x\big]-\mathbb{E}\big[\mu\big]\right) \\
&= \frac{1}{\sigma}\left(\mu-\mu\right) = 0,
\end{aligned} \tag{C.16}$$

where again, we have used the four rules for manipulating expectations. The variance of the new distribution is given by:

$$\begin{aligned}
\mathbb{E}\big[(z-\mu_z)^2\big] &= \mathbb{E}\big[(z-\mathbb{E}[z])^2\big] \\
&= \mathbb{E}\big[z^2\big] \\
&= \mathbb{E}\left[\left(\frac{x-\mu}{\sigma}\right)^2\right] \\
&= \frac{1}{\sigma^2}\cdot\mathbb{E}[(x-\mu)^2] \\
&= \frac{1}{\sigma^2}\cdot\sigma^2 = 1.
\end{aligned} \tag{C.17}$$

By a similar argument, we can take a standardized variable z with mean zero and unit variance and convert it to a variable x with mean μ and variance σ^2 using:

$$x = \mu + \sigma z. \tag{C.18}$$

In the multivariate case, we can standardize a variable $\mathbf{x}$ with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ using:

$$\mathbf{z} = \boldsymbol{\Sigma}^{-1/2}(\mathbf{x}-\boldsymbol{\mu}). \tag{C.19}$$

The result will have a mean $\mathbb{E}[\mathbf{z}] = \mathbf{0}$ and an identity covariance matrix $\mathbb{E}[(\mathbf{z}-\mathbb{E}[\mathbf{z}])(\mathbf{z}-\mathbb{E}[\mathbf{z}])^T] = \mathbf{I}$. To reverse this process, we use:

$$\mathbf{x} = \boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\mathbf{z}. \tag{C.20}$$

C.3 Normal probability distribution

Probability distributions used in this book include the Bernoulli distribution (figure 5.6), categorical distribution (figure 5.9), Poisson distribution (figure 5.15), von Mises distribution (figure 5.13), and mixture of Gaussians (figures 5.14 and 17.1). However, the most common distribution in machine learning is the normal or Gaussian distribution.

C.3.1 Univariate normal distribution

A univariate normal distribution (figure 5.3) over scalar variable x has two parameters, the mean μ and the variance σ^2, and is defined as:

$$Pr(x) = \text{Norm}_x[\mu, \sigma^2] = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]. \tag{C.21}$$

Unsurprisingly, the mean $\mathbb{E}[x]$ of a normally distributed variable is given by the mean parameter μ and the variance $\mathbb{E}[(x - \mathbb{E}[x])^2]$ by the variance parameter σ^2. When the mean is zero and the variance is one, we refer to this as a *standard normal distribution.*

The shape of the normal distribution can be inferred from the following argument. The term $-(x-\mu)^2/2\sigma^2$ is a quadratic function that falls away from zero when $x = \mu$ at a rate that increases when σ becomes smaller. When we pass this through the exponential function (figure B.1), we get a bell-shaped curve, which has a value of one at $x = \mu$ and falls away to either side. Dividing by the constant $\sqrt{2\pi\sigma^2}$ ensures that the function integrates to one and is a valid distribution. It follows from this argument that the mean μ control the position of the center of the bell curve, and the square root σ of the variance (the standard deviation) controls the width of the bell curve.

C.3.2 Multivariate normal distribution

The multivariate normal distribution generalizes the normal distribution to describe the probability over a vector quantity $\mathbf{x}$ of length D. It is defined by a $D \times 1$ *mean vector* $\boldsymbol{\mu}$ and a symmetric positive definite $D \times D$ *covariance matrix* $\boldsymbol{\Sigma}$:

$$\text{Norm}_{\mathbf{x}}[\boldsymbol{\mu}, \boldsymbol{\Sigma}] = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left[-\frac{(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}{2}\right]. \tag{C.22}$$

The interpretation is similar to the univariate case. The quadratic term $-(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})/2$ returns a scalar that decreases as $\mathbf{x}$ grows further from the mean $\boldsymbol{\mu}$, at a rate that depends on the matrix $\boldsymbol{\Sigma}$. This is turned into a bell-curve shape by the exponential, and dividing by $(2\pi)^{D/2}|\Sigma|^{1/2}$ ensures that the distribution integrates to one.

The covariance matrix can take spherical, diagonal, and full forms:

$$\boldsymbol{\Sigma}_{spher} = \begin{bmatrix}\sigma^2 & 0\\ 0 & \sigma^2\end{bmatrix} \quad \boldsymbol{\Sigma}_{diag} = \begin{bmatrix}\sigma_1^2 & 0\\ 0 & \sigma_2^2\end{bmatrix} \quad \boldsymbol{\Sigma}_{full} = \begin{bmatrix}\sigma_{11}^2 & \sigma_{12}^2\\ \sigma_{21}^2 & \sigma_{22}^2\end{bmatrix}. \tag{C.23}$$

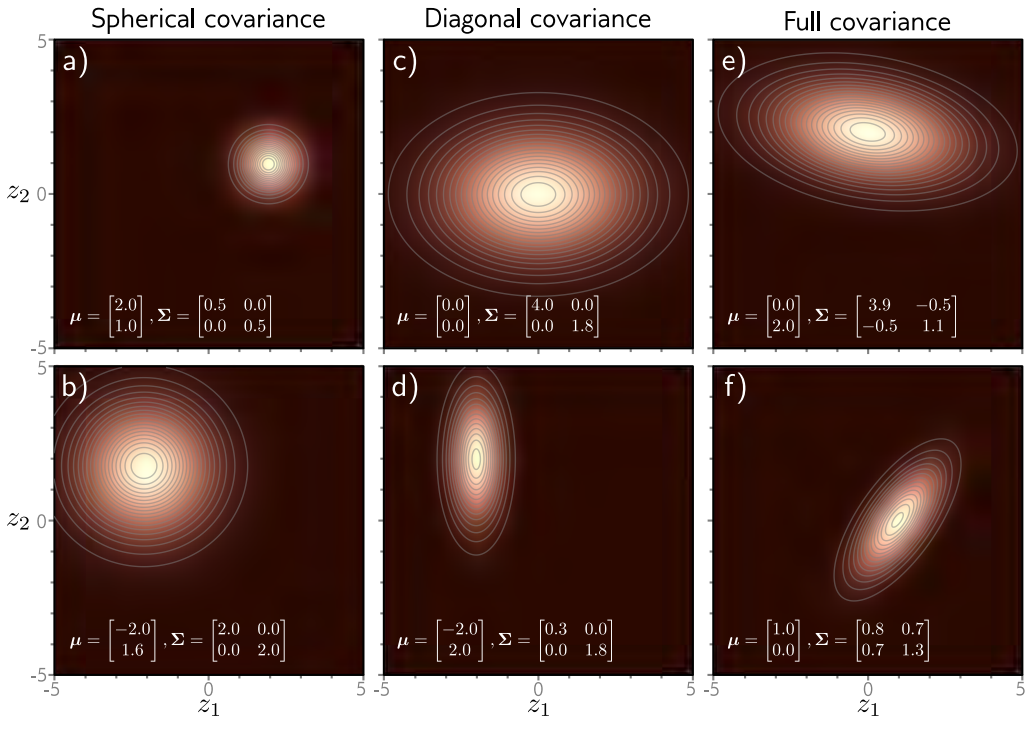

Figure C.4 Bivariate normal distribution. a–b) When the covariance matrix is a multiple of the diagonal matrix, the isocontours are circles, and we refer to this as spherical covariance. c–d) When the covariance is an arbitrary diagonal matrix, the isocontours are axis-aligned ellipses, and we refer to this as diagonal covariance e–f) When the covariance is an arbitrary symmetric positive definite matrix, the iso-contours are general ellipses, and we refer to this as full covariance.

In two dimensions (figure C.4), spherical covariances produce circular iso-density contours, and diagonal covariances produce ellipsoidal iso-contours that are aligned with the coordinate axes. Full covariances produce general ellipsoidal iso-density contours. When the covariance is spherical or diagonal, the individual variables are independent:

$$\begin{aligned} Pr(x_1, x_2) &= \frac{1}{2\pi\sqrt{|\boldsymbol{\Sigma}|}} \exp\left[-0.5 \begin{pmatrix} x_1 & x_2 \end{pmatrix} \boldsymbol{\Sigma}^{-1} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\right] \\ &= \frac{1}{2\pi\sigma_1\sigma_2} \exp\left[-0.5 \begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} \sigma_1^{-2} & 0 \\ 0 & \sigma_2^{-2} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\right] \\ &= \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left[-\frac{x_1^2}{2\sigma_1^2}\right] \cdot \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left[-\frac{x_2^2}{2\sigma_2^2}\right] \\ &= Pr(x_1) \cdot Pr(x_2). \end{aligned} \tag{C.24}$$

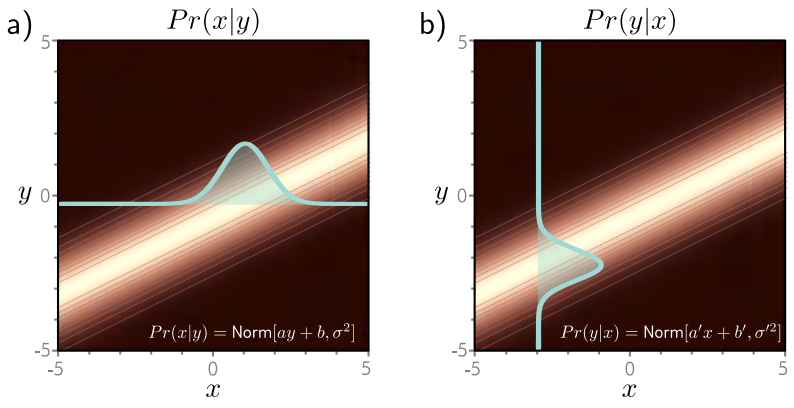

Figure C.5 Change of variables. a) The conditional distribution $Pr(x|y)$ is a normal distribution with constant variance and a mean that depends linearly on y. Cyan distribution shows one example for $y = -0.2$. b) This is proportional to the conditional probability $Pr(y|x)$, which is a normal distribution with constant variance and a mean that depends linearly on x. Cyan distribution shows one example for $x = -3$.

C.3.3 Product of two normal distributions

The product of two normal distributions is proportional to a third normal distribution according to the relation:

$$\text{Norm}_{\mathbf{x}}[\mathbf{a}, \mathbf{A}]\text{Norm}_{\mathbf{x}}[\mathbf{b}, \mathbf{B}] \;\propto\; \text{Norm}_{\mathbf{x}}\Big[(\mathbf{A}^{-1} + \mathbf{B}^{-1})^{-1}(\mathbf{A}^{-1}\mathbf{a} + \mathbf{B}^{-1}\mathbf{b}), (\mathbf{A}^{-1} + \mathbf{B}^{-1})^{-1}\Big].$$

This is easily proved by multiplying out the exponential terms and completing the square (see problem 18.5).

C.3.4 Change of variable

When the mean of a multivariate normal in $\mathbf{x}$ is a linear function $\mathbf{A}\mathbf{y} + \mathbf{b}$ of a second variable $\mathbf{y}$, this is proportional to another normal distribution in $\mathbf{y}$, where the mean is a linear function of $\mathbf{x}$:

$$\text{Norm}_{\mathbf{x}}\left[\mathbf{A}\mathbf{y} + \mathbf{b}, \boldsymbol{\Sigma}\right] \propto \text{Norm}_{\mathbf{y}}[(\mathbf{A}^T\boldsymbol{\Sigma}^{-1}\mathbf{A})^{-1}\mathbf{A}^T\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \mathbf{b}), (\mathbf{A}^T\boldsymbol{\Sigma}^{-1}\mathbf{A})^{-1}]. \quad \text{(C.25)}$$

At first sight, this relation is rather opaque, but figure C.5 shows the case for scalar x and y, which is easy to understand. As for the previous relation, this can be proved by expanding the quadratic product in the exponential term and completing the square to make this a distribution in $\mathbf{y}$. (see problem 18.4).

C.4 Sampling

To sample from a univariate distribution $Pr(x)$, we first compute the cumulative distribution $\mathrm{F}[x]$ (the integral of $Pr(x)$). Then we draw a sample z^* from a uniform distribution over the range $[0, 1]$ and evaluate this against the inverse of the cumulative distribution, so the sample x^* is created as:

$$x^* = \mathrm{F}^{-1}[z^*]. \tag{C.26}$$

C.4.1 Sampling from normal distributions

The method above can be used to generate a sample x^* from a univariate standard normal distribution. A sample from a normal distribution with mean μ and variance σ^2 can then be created using equation C.18. Similarly, a sample $\mathbf{x}^*$ from a D-dimensional multivariate standard distribution can be created by independently sampling D univariate standard normal variables. A sample from a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ can be then created using equation C.20.

C.4.2 Ancestral sampling

When the joint distribution can be factored into a series of conditional probabilities, we can generate samples using *ancestral sampling*. The basic idea is to generate a sample from the root variable(s) and then sample from the subsequent conditional distributions based on this instantiation. This process is known as *ancestral sampling* and is easiest to understand with an example. Consider a joint distribution over three variables, x, y, and z, where the distribution factors as:

$$Pr(x, y, z) = Pr(x)Pr(y|x)Pr(z|y). \tag{C.27}$$

To sample from this joint distribution, we first draw a sample x^* from $Pr(x)$. Then we draw a sample y^* from $Pr(y|x^*)$. Finally, we draw a sample z^* from $Pr(z|y^*)$.

C.5 Distances between probability distributions

Supervised learning can be framed in terms of minimizing the distance between the probability distribution implied by the model and the discrete probability distribution implied by the samples (section 5.7). Unsupervised learning can often be framed in terms of minimizing the distance between the probability distribution of real examples and the distribution of data from the model. In both cases, we need a measure of distance between two probability distributions. This section considers the properties of several different measures of distance between distributions (see also figure 15.8 for a discussion of the Wasserstein or earth mover's distance).

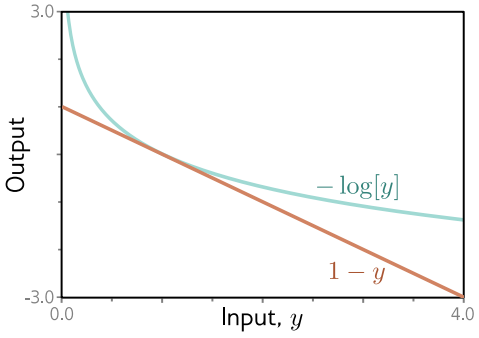

Figure C.6 Lower bound on negative logarithm. The function $1-y$ is always less than the function $-\log[y]$. This relation is used to show that the Kullback-Leibler divergence is always greater than or equal to zero.

C.5.1 Kullback-Leibler divergence

The most common measure of distance between probability distributions $p(x)$ and $q(x)$ is the *Kullback-Leibler* or KL divergence and is defined as:

$$D_{KL}\big[p(x)||q(x)\big] = \int p(x)\log\left[\frac{p(x)}{q(x)}\right]dx. \tag{C.28}$$

This distance is always greater than or equal to zero, which is easily demonstrated by noting that $-\log[y] \geq 1-y$ (figure C.6) so:

$$\begin{aligned} D_{KL}\Big[p(x)\Big|\Big|q(x)\Big] &= \int p(x)\log\left[\frac{p(x)}{q(x)}\right]dx \\ &= -\int p(x)\log\left[\frac{q(x)}{p(x)}\right]dx \\ &\geq \int p(x)\left(1-\frac{q(x)}{p(x)}\right)dx \\ &= \int p(x)-q(x)dx \\ &= 1-1=0. \end{aligned} \tag{C.29}$$

The KL divergence is infinite if there are places where $q(x)$ is zero but $p(x)$ is non-zero. This can lead to problems when we are minimizing a function based on this distance.

C.5.2 Jensen-Shannon divergence

The KL divergence is not symmetric (i.e., $D_{KL}[p(x)||q(x)] \neq D_{KL}[q(x)||p(x)]$). The Jensen-Shannon divergence is a measure of distance that is symmetric by construction:

$$D_{JS}\Big[p(x)\Big|\Big|q(x)\Big] = \frac{1}{2}D_{KL}\left[p(x)\Big|\Big|\frac{p(x)+q(x)}{2}\right] + \frac{1}{2}D_{KL}\left[q(x)\Big|\Big|\frac{p(x)+q(x)}{2}\right]. \tag{C.30}$$

It is the mean divergence of $p(x)$ and $q(x)$ to the average of the two distributions.

C.5.3 Fréchet distance

The Fréchet distance D_{FR} between two distributions $p(x)$ and $q(x)$ is given by:

$$D_{Fr}\Big[p(x)\Big|\Big|q(y)\Big] = \sqrt{\min_{\pi(x,y)}\left[\iint \pi(x,y)|x-y|^2 dxdy\right]}, \tag{C.31}$$

where $\pi(x,y)$ represents the set of joint distributions that are compatible with the marginal distributions $p(x)$ and $q(y)$. The Fréchet distance can also be formulated as a measure of the maximum distance between the cumulative probability curves.

C.5.4 Distances between normal distributions

Often we want to compute the distance between two multivariate normal distributions with means $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ and covariances $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$. In this case, various measures of distance can be written in closed form.

The KL divergence can be computed as:

$$\begin{aligned} D_{KL}\Big[\mathrm{Norm}[\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1]\Big|\Big|\mathrm{Norm}[\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2]\Big] \;&=\; \\ &\frac{1}{2}\left(\log\left[\frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|}\right] - D + \mathrm{tr}\left[\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\right] + (\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)\right). \end{aligned} \tag{C.32}$$

where $tr[\bullet]$ is the trace of the matrix argument. The Fréchet/2-Wasserstein distance is given by:

$$D^2_{Fr/W_2}\Big[\mathrm{Norm}[\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1]\Big|\Big|\mathrm{Norm}[\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2]\Big] = |\boldsymbol{\mu}_1-\boldsymbol{\mu}_2|^2 + \mathrm{tr}\left[\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2-2\left(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2\right)^{1/2}\right]. \tag{C.33}$$