1 Mathematical Preliminaries

1.1 Introduction

In this chapter we will discuss some mathematical concepts that are useful to familiarize yourself with before you start reading Parts II and III of the book. In particular, we will discuss various aspects of probability theory, including fundamental concepts such as Bayes' rule, and information-theoretic concepts such as entropy. We will introduce you to the concept of symmetries in physics, such as the fundamental concept of time-reversal symmetry. Next, we will provide a comprehensive introduction to convex optimization where you will learn how Lagrange multipliers can help you enforce constraints. Finally, we will say a few words about the difference between the Eulerian and Lagrangian views, which describes the difference between fixed and comoving frames of reference.

1.2 Probabilities

1.2.1 Probability Mass and Density Functions

Probability distributions for a random variable X come in two flavors: probability mass functions (PMFs) and probability density functions (PDFs). The former is defined on discrete states $x = i,\ i = 1, \ldots, N$, while the latter is defined on a continuous space: $x \in \mathbb{R}$. A PMF is defined by a set of non-negative numbers that sum to 1:

$$p(X = x) \geq 0 \text{ for all } x, \quad \sum_x p(X = x) = 1. \tag{1.1}$$

A PDF is a more complicated object due to the presence of a volume element. It consists of a non-negative function that integrates to 1:

$$p(X = x) \geq 0 \text{ for all } x, \quad \int \mathrm{d}x\, p(X = x) = 1. \tag{1.2}$$

Note that for a PMF, all $p(x)$ must be less than or equal to 1; otherwise, they cannot sum to 1 as required by (1.1). This constraint, however, does not necessarily apply to a PDF due to the volume element $\mathrm{d}x$ in (1.2). In fact, we can have $p(x) > 1$ for a PDF as long as the integral over the entire domain equals 1. We will see an example of this shortly when we discuss Gaussian distributions.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 03:28:49, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009709071.003

For simplicity, we will often write $p(X = x)$ using subscript notation as $p_X(x)$. However, if the context is clear, we will omit specifying the random variable entirely in the following chapters.

A PMF is often described by a categorical distribution, which assigns probabilities to each discrete state x:

$$p(x = i) = \pi_i, \quad \pi_i \geq 0 \text{ for all } i, \quad \sum_i \pi_i = 1. \tag{1.3}$$

Here, we consider π_i to be parameters of the model. To enforce the positivity constraint, we can parameterize the categorical distribution using an exponential form:

$$\pi_i = \frac{1}{Z} \exp(-\varepsilon_i), \quad Z = \sum_i \exp(-\varepsilon_i), \tag{1.4}$$

where ε_i can be interpreted as the energy for state i, and Z is the normalization constant.

There are many examples of PDFs, but one of the most commonly used is the Gaussian (or normal) distribution. We denote a Gaussian distribution with mean μ and standard deviation σ by $G(\mu, \sigma)$, which has the probability density function:

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right). \tag{1.5}$$

This distribution is illustrated in Figure 1.1. One can check that

$$\mathbb{E}_p[x] = \mu, \quad \mathbb{E}_p[(x - \mu)^2] = \sigma^2, \tag{1.6}$$

$$Z = \int \mathrm{d}x \, \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right) = \sqrt{2\pi}\sigma, \tag{1.7}$$

where $\mathbb{E}_p[\cdot]$ denotes the expectation with respect to the distribution p, defined as

$$\mathbb{E}_p[f(x)] = \int \mathrm{d}x \, p(x) f(x), \tag{1.8}$$

for any function $f(x)$. In the limit of $\sigma \to 0^+$, the Gaussian distribution $G(\mu, \sigma)$ tends to a Dirac delta function,

$$\lim_{\sigma \to 0^+} G(\mu, \sigma) = \delta(x - \mu), \tag{1.9}$$

which has the property that

$$\int \mathrm{d}x \, \delta(x - \mu) f(x) = f(\mu). \tag{1.10}$$

The value of $\delta(x - \mu)$ approaches infinity at $x = \mu$ and zero elsewhere in such a way that the integral remains well defined and equals 1. This confirms our earlier assertion following from (1.2) that for PDFs, the values of $p(x)$ can indeed exceed 1!

The high-dimensional generalization of the Gaussian distribution, (1.5), known as the multivariate Gaussian (or normal) distribution, is given by

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2} |\det(\boldsymbol{\Sigma})|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right), \tag{1.11}$$

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 03:28:49, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009709071.003

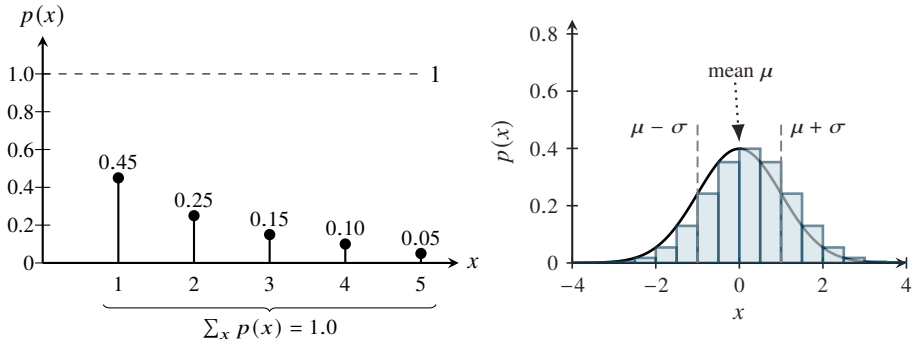


Figure 1.1 (a) A probability mass function assigns probabilities to discrete values that must sum to 1. Each vertical line with a dot represents the probability of a specific outcome, with explicit probability values shown. (b) A Gaussian probability density function with standard deviation $\sigma = 1$. Vertical dashed guides at $\mu - \sigma$ and $\mu + \sigma$ mark the interval that contains approximately 68% of the probability mass, and the histogram bars illustrate how discrete samples from a Gaussian distribution approximate the continuous probability density function.

with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The moments of the multivariate Gaussian distribution are

$$\mathbb{E}_p[\mathbf{x}] = \boldsymbol{\mu}, \quad \mathbb{E}_p[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top] = \boldsymbol{\Sigma}, \tag{1.12}$$

$$Z = \int \mathrm{d}\mathbf{x} \, \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right) = (2\pi)^{d/2} |\det(\boldsymbol{\Sigma})|^{1/2}. \tag{1.13}$$

The above examples are part of a larger family of distributions with nice properties, known as the exponential family (Wainwright and Jordan, 2007). It can be defined as

$$p(x) = \frac{1}{Z} \exp\left(-\sum_k \theta_k f_k(x)\right), \tag{1.14}$$

where θ_k are parameters and f_k are features or sufficient statistics (functions of the random variable x). The function $F(\boldsymbol{\theta}) = -\log Z(\boldsymbol{\theta})$ plays a special role in the exponential family, as it generates the centralized moments (or cumulants) of the distribution:

$$\nabla_{\theta_k} F(\boldsymbol{\theta}) = \mathbb{E}_p[f_k], \tag{1.15}$$

$$\nabla_{\theta_j} \nabla_{\theta_k} F(\boldsymbol{\theta}) = \mathbb{E}_p[f_j]\mathbb{E}_p[f_k] - \mathbb{E}_p[f_j f_k] = -\operatorname*{Cov}_p[f_j, f_k], \tag{1.16}$$

and so on. Many other common distributions, such as the Bernoulli and Poisson, also belong to the exponential family. For example, the Bernoulli distribution, which models binary random variables, can be written as

$$p(x) = \mu^x (1 - \mu)^{1-x} = \exp\left(x \log \frac{\mu}{1 - \mu} + \log(1 - \mu)\right), \tag{1.17}$$

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 03:28:49, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009709071.003

where μ is the probability of success (i.e., $x = 1$). Comparing (1.17) with (1.14), we can identify $f(x) = x$, $\theta = -\log \frac{\mu}{1-\mu}$, and $\log Z(\theta) = \log(1 + e^{-\theta})$.

1.2.2 Conditional Distributions and Bayes' Rule

Focusing on PDFs, the probability of two events occurring simultaneously can be defined as a non-negative function of two arguments that integrates to 1 over all variables:

$$p(x, y) \geq 0 \text{ for all } x, \text{ for all } y, \qquad \iint \mathrm{d}x\, \mathrm{d}y\; p(x, y) = 1. \tag{1.18}$$

This is referred to as the *joint distribution* of X and Y. The previously discussed high-dimensional generalization of the Gaussian distribution is an example of this, where the different dimensions represent different variables. We can retrieve the single-variable PDF by integrating over all other variables:

$$p(x) = \int \mathrm{d}y\; p(x, y). \tag{1.19}$$

When derived from a joint distribution, $p(x)$ is referred to as the marginal distribution of $p(x, y)$. The marginal distribution still obeys the constraint $\int \mathrm{d}x\; p(x) = 1$.

Instead of marginalizing out y, we can also consider the case where the value of y is fixed. We refer to this as the *conditional distribution* of x given y, obtained as

$$p(x|y) = \frac{p(x, y)}{p(y)}. \tag{1.20}$$

The conditional distribution is also a valid PDF and thus satisfies $\int \mathrm{d}x\; p(x|y) = 1$ for all y.

Bayes' Rule

We can relate the conditional distributions for x given y and y given x as

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}, \tag{1.21}$$

which is known as *Bayes' theorem* or Bayes' rule (named after Thomas Bayes).

Bayes' rule is an example of a (detailed) balance equation, which we will encounter in various forms throughout the later chapters of this book. If we write it as

$$p(y|x)p(x) = p(x|y)p(y), \tag{1.22}$$

it states that the total flow of probability between any two points x and y is the same, and as such, the marginal probabilities p_X and p_Y will remain the same. The conditional distributions $p(y|x)$ and $p(x|y)$ act as transition probabilities for these flows. We can say that Bayes' rule expresses that the systems described by p_X and p_Y are in relative *equilibrium*.

Bayes' theorem plays a crucial role in machine learning, where it is used to update our beliefs given new evidence. In this context, we refer to $p(y)$ as our *prior probability*, reflecting our original beliefs about a system without having observed any data.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 03:28:49, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009709071.003

Correspondingly, $p(y|x)$ is referred to as the *posterior distribution*, representing our belief about the system after observing x. The denominator in Bayes' theorem,

$$p(x) = \int \mathrm{d}y \; p(x|y)p(y), \tag{1.23}$$

is known as the *evidence*. Because it requires marginalization over the entire space of y, this can be computationally very expensive.

(Conditional) Independence Conditions

Two random variables x and y are said to be *independent* if observing one does not affect the probability of observing the other. For example, consider flipping two fair coins. Observing heads on the first coin does not affect the probability of heads on the second coin, and as such, we can consider these two events to be independent of each other.

More formally, we state that two variables x and y are independent (denoted as $x \perp y$) if and only if their joint probability distribution factorizes; that is,

$$p(x, y) = p(x)p(y), \tag{1.24}$$

which also implies $p(x|y) = p(x)$ and $p(y|x) = p(y)$. Moreover, $\mathbb{E}[xy] = \mathbb{E}[x]\mathbb{E}[y]$.

Additionally, two random variables x and y are said to be *conditionally independent* given a third random variable z (written $x \perp y|z$) if x and y are independent given z,

$$p(x, y|z) = p(x|z)p(y|z). \tag{1.25}$$

As an example, consider flipping two fair coins. If we are told that at least one of the coins landed on heads, the outcomes of the two coins become conditionally dependent given this information. For instance, knowing the first coin is tails now makes it certain the second coin is heads (to satisfy the condition). Knowing the outcome of one coin provides information about the other under this condition. In this scenario, the two coin flips are *marginally independent* but *conditionally dependent* given the information about the combined outcome.

1.2.3 Entropy

An important property of a probability distribution, beyond its cumulants, that we will consider in this book is its entropy. In information-theoretic terms, entropy defines an average measure of surprise or uncertainty encoded by the distribution. Formally, it is defined as

$$\mathcal{S}[p] = -\int \mathrm{d}\mathbf{x} \; p(\mathbf{x}) \log p(\mathbf{x}), \tag{1.26}$$

where $S[p](\mathbf{x}) = -\log p(\mathbf{x})$ is known as the *surprisal* (or *unexpectedness*) value for an individual value $\mathbf{x}$. Depending on the context, the base of the logarithm can vary. For example, in information theory, base 2 is often used, in which case the entropy has units of *bits*. In the remainder of this book, however, we will use base e and use *nats* as our unit (Winter, 2021). One nat is $\log_2 e \approx 1.44$ bits.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 03:28:49, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009709071.003

In the case of a joint distribution of two independent variables, the entropy is given by the sum of their marginal entropies

$$\mathcal{S}[p_{X,Y}] = \mathcal{S}[p_X] + \mathcal{S}[p_Y]. \tag{1.27}$$

For continuous distributions supported on the entire real line, the Gaussian distribution introduced earlier serves as a canonical example. Its entropy can be computed as

$$\begin{aligned}
\mathcal{S}[p] &= \mathbb{E}_p[-\log p(x)] && (1.28)\\
&= \mathbb{E}_p\left[-\log\left(\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)\right)\right] && (1.29)\\
&= \log(\sqrt{2\pi}\sigma) + \frac{1}{2\sigma^2}\mathbb{E}_p[(x-\mu)^2] && (1.30)\\
&= \log(\sqrt{2\pi}\sigma) + \frac{1}{2}. && (1.31)
\end{aligned}$$

This shows that the entropy of the Gaussian distribution depends only on the standard deviation and not on the mean of the distribution.

1.2.4 Comparing Distributions

The Kullback–Leibler (KL) divergence is a nonsymmetric measure of the difference between two probability distributions. Given two different distributions $p(x)$ and $q(x)$, the KL divergence is defined as

$$D_{\mathrm{KL}}[p\|q] = \int \mathrm{d}x\; p(x)\log\frac{p(x)}{q(x)}. \tag{1.32}$$

Intuitively, the KL divergence represents the amount of information lost when expressing a system described by p in terms of the probability distribution q. This information loss interpretation becomes clear when expressing the KL divergence as

$$\begin{aligned}
D_{\mathrm{KL}}[p\|q] &= \int \mathrm{d}x\; p(x)\left(\log p(x) - \log q(x)\right) && (1.33)\\
&= -\mathcal{S}[p] + \mathcal{S}[p\|q], && (1.34)
\end{aligned}$$

where $\mathcal{S}[p\|q] = -\int \mathrm{d}x\; p(x)\log q(x)$ is known as the cross-entropy. In this interpretation, the KL divergence is also known as the relative entropy. When using the KL divergence, it is important to note that it is only properly defined when q has the same support as p; that is, $q(x) \neq 0$ for all x, where $p(x) \neq 0$.

As an example, consider two univariate Gaussian distributions $p(x) = G(\mu_1, \sigma_1)$ and $q(x) = G(\mu_2, \sigma_2)$. The KL divergence between them is given by

$$D_{\mathrm{KL}}[p\|q] = \log\frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2}. \tag{1.35}$$

As mentioned, the KL divergence is nonsymmetric ($D_{\mathrm{KL}}[p\|q] \neq D_{\mathrm{KL}}[q\|p]$) and therefore not a true metric. This nonsymmetry has important implications in machine learning regarding which direction of divergence to use, as illustrated in Figure 1.2.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 03:28:49, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009709071.003

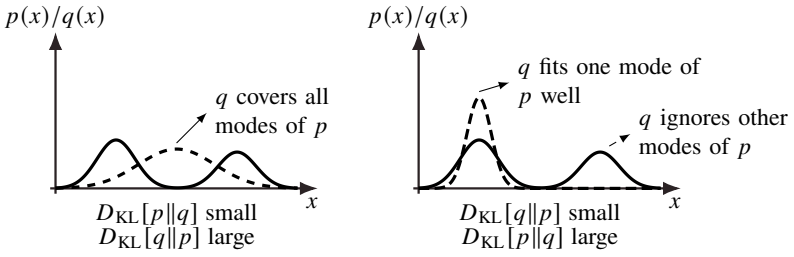


Figure 1.2 Asymmetry of the KL divergence. Let p be the true (bimodal) distribution and q be the approximation. Minimizing the "forward" KL divergence $D_{\mathrm{KL}}[p\|q]$ encourages q to be nonzero wherever p is nonzero, leading to a broad, mode-covering approximation (left). Minimizing the "reverse" KL divergence $D_{\mathrm{KL}}[q\|p]$ penalizes q for being nonzero where p is zero, leading to a focused, mode-seeking approximation that may ignore some modes of p (right).

Furthermore, it does not necessarily satisfy the triangle inequality. However, it is non-negative and satisfies $D_{\mathrm{KL}}[p\|q] = 0$ only if $p = q$ almost everywhere, and can therefore be considered a *divergence* (hence the name).

A useful decomposition of a KL divergence over two (or more) random variables X and Y is given by

$$D_{\mathrm{KL}}[p_{X,Y}\|q_{X,Y}] = D_{\mathrm{KL}}[p_X\|q_X] + \mathbb{E}_{p_X}\left[D_{\mathrm{KL}}[p_{Y|X}\|q_{Y|X}]\right]. \tag{1.36}$$

One can think of this as "Bayes' rule" for KL divergences.

While we will primarily consider the KL divergence in the rest of this book, various other divergences and approaches for measuring the difference between distributions are available. Most notable among these are the χ^2 divergence, mutual information, and the Wasserstein distance (Cover and Thomas, 2006).

1.3 Approximating Expectations

As we can see, many problems surrounding probabilities resolve to solving integrals in the form of expectations,

$$\mathbb{E}_p[f(x)] = \int \mathrm{d}x\; p(x) f(x). \tag{1.37}$$

Achieving this efficiently will be a central theme of this book and, as such, it makes sense to spend some time on the basics of this.

As seen, in the case of a Gaussian this can often be done tractably, but this is not the case in general. Therefore, approximations are required.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 03:28:49, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009709071.003

1.3.1 Monte Carlo Approximation

Monte Carlo methods are widely used for estimating probabilities and expectations by drawing samples from a target distribution. In regular sampling, we draw N independent and identically distributed (i.i.d.) samples from the target distribution $p(\mathbf{x})$:

$$\mathbf{x}_n \sim p(\mathbf{x}), \quad n = 1, \ldots, N, \tag{1.38}$$

where the notation $\mathbf{x}_n \sim p(\mathbf{x})$ means that each sample $\mathbf{x}_n$ is drawn randomly from the distribution $p(\mathbf{x})$. Given these samples, the expected value of a function $f(\mathbf{x})$, (1.8), can be approximated by the Monte Carlo estimate

$$\mathbb{E}_p[f(\mathbf{x})] \simeq \langle f \rangle = \frac{1}{N} \sum_{n=1}^{N} f(\mathbf{x}_n). \tag{1.39}$$

The law of large numbers ensures that as N increases, the Monte Carlo estimate $\langle f \rangle$ converges to the true expected value $\mathbb{E}_p[f(\mathbf{x})]$. This property is known as *consistency*.

The Monte Carlo estimate $\langle f \rangle$ is also an *unbiased* estimator of $\mathbb{E}_p[f(\mathbf{x})]$:

$$\mathbb{E}[\langle f \rangle] = \frac{1}{N} \sum_{n=1}^{N} \mathbb{E}_p[f(\mathbf{x}_n)] = \mathbb{E}_p[f(\mathbf{x})], \tag{1.40}$$

where we used the linearity of expectation and the fact that each sample is drawn from the same distribution $p(\mathbf{x})$.

Estimator Variance

The expected error of the Monte Carlo estimate can be quantified by its variance, defined as

$$\mathbb{V}[x] = \mathbb{E}\left[(x - \mathbb{E}[x])^2\right]. \tag{1.41}$$

Using this definition, and noting that the samples are independent (so $\mathrm{Cov}(f(\mathbf{x}_n), f(\mathbf{x}_m)) = 0$ for $n \neq m$), we can derive the variance of $\langle f \rangle$ as follows:

$$\begin{aligned}
\mathbb{V}[\langle f \rangle] &= \mathbb{E}\left[\left(\langle f \rangle - \mathbb{E}_p[f(\mathbf{x})]\right)^2\right] = \mathbb{V}\left(\frac{1}{N} \sum_{n=1}^{N} f(\mathbf{x}_n)\right) \\
&= \frac{1}{N^2} \sum_{n=1}^{N} \sum_{m=1}^{N} \mathrm{Cov}\left(f(\mathbf{x}_n), f(\mathbf{x}_m)\right) \\
&= \frac{1}{N^2} \sum_{n=1}^{N} \mathbb{V}\left(f(\mathbf{x}_n)\right) \\
&= \frac{1}{N} \mathbb{V}_p\left(f(\mathbf{x})\right).
\end{aligned} \tag{1.42}$$

The key insight from this derivation is that the variance of the Monte Carlo estimate decreases as $1/N$, where N is the number of samples. This means that increasing the number of samples reduces the expected error of the estimate, with the standard deviation of the error decreasing as $O(1/\sqrt{N})$.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 03:28:49, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009709071.003

Importance Sampling

As seen, doing a Monte Carlo approximation requires drawing i.i.d. samples from the target distribution $p(\mathbf{x})$. In the next section, as well as at various other points in this book, we will be exploring different methods for doing this, but it is important to note that in many cases this remains intractable within a given compute budget. Importance sampling (IS) in this scenario provides a useful alternative.

Importance sampling uses the identity that we can relate two distributions $p(\mathbf{x})$ and $q(\mathbf{x})$ through the ratio $w(\mathbf{x}) = p(\mathbf{x})/q(\mathbf{x})$. This allows us to rewrite the expectation of the target distribution $p(\mathbf{x})$ in terms of an alternative distribution $q(\mathbf{x})$ as

$$\mathbb{E}_p[f(\mathbf{x})] = \int \mathrm{d}\mathbf{x}\, p(\mathbf{x}) f(\mathbf{x}) = \int \mathrm{d}\mathbf{x}\, q(\mathbf{x}) \frac{p(\mathbf{x})}{q(\mathbf{x})} f(\mathbf{x}) = \mathbb{E}_q\left[w(\mathbf{x}) f(\mathbf{x})\right]. \tag{1.43}$$

If we define the alternative distribution, often referred to as the *proposal distribution*, $q(\mathbf{x})$ such that we can easily draw samples from it, we can use the weights $w(\mathbf{x})$ to obtain an unbiased and consistent estimate of the expectation of the target distribution $p(\mathbf{x})$ as

$$\mathbb{E}_p[f(\mathbf{x})] \simeq \langle f \rangle_{\mathrm{IS}} = \frac{1}{N} \sum_{n=1}^{N} w(\mathbf{x}_n) f(\mathbf{x}_n), \qquad \text{where } \mathbf{x}_n \sim q(\mathbf{x}). \tag{1.44}$$

Both unbiasedness and consistency can similarly be proven using the linearity of expectation and the fact that the samples are i.i.d., as done for the Monte Carlo estimate.

Self-Normalized Importance Sampling

However, while importance sampling allows us to obtain an unbiased estimate of the expectation without requiring direct samples from the distribution, it does require us to be able to compute the ratio $p(\mathbf{x})/q(\mathbf{x})$ for all $\mathbf{x}$. This is often not possible because the normalization constant Z_p of the target distribution $p(\mathbf{x})$ is unknown. If we let $p(x) = \frac{1}{Z_p}\tilde{p}(x)$, then we can rewrite the importance sampling estimate as

$$\mathbb{E}_p[f(\mathbf{x})] = \int \mathrm{d}\mathbf{x}\, \frac{1}{Z_p} \tilde{p}(\mathbf{x}) f(\mathbf{x}) = \frac{1}{Z_p} \int \mathrm{d}\mathbf{x}\, q(\mathbf{x}) \frac{\tilde{p}(\mathbf{x})}{q(\mathbf{x})} f(\mathbf{x}) = \frac{1}{Z_p} \mathbb{E}_q\left[\tilde{w}(\mathbf{x}) f(\mathbf{x})\right], \tag{1.45}$$

where we now have the unnormalized weights $\tilde{w}(\mathbf{x}) = \frac{\tilde{p}(\mathbf{x})}{q(\mathbf{x})}$. Fortunately, it turns out that if we now take the expectation of the unnormalized weights over the proposal distribution $q(\mathbf{x})$, we obtain

$$\mathbb{E}_q[\tilde{w}(\mathbf{x})] = \int \mathrm{d}\mathbf{x}\, q(\mathbf{x}) \tilde{w}(\mathbf{x}) = \int \mathrm{d}\mathbf{x}\, q(\mathbf{x}) \frac{\tilde{p}(\mathbf{x})}{q(\mathbf{x})} = \int \mathrm{d}\mathbf{x}\, \tilde{p}(\mathbf{x}) = Z_p. \tag{1.46}$$

As such, we can combine these to construct the *self-normalized importance sampling* (SNIS) estimator:

$$\mathbb{E}_p[f(\mathbf{x})] \simeq \langle f \rangle_{\mathrm{SNIS}} = \frac{\frac{1}{N}\sum_{n=1}^{N} \tilde{w}(\mathbf{x}_n) f(\mathbf{x}_n)}{\frac{1}{N}\sum_{n=1}^{N} \tilde{w}(\mathbf{x}_n)}. \tag{1.47}$$

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 03:28:49, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009709071.003

While this is a very useful result, it is important to note that the SNIS estimate is no longer unbiased (i.e., $\mathbb{E}_q[\langle f \rangle_{\text{SNIS}}] \neq \mathbb{E}_p[f(\mathbf{x})]$) due to the fact that the ratio of two expectations is not equal to the expectation of the ratio. It is, however, both asymptotically unbiased and consistent: the bias vanishes ($\mathbb{E}_q[\langle f \rangle_{\text{SNIS}}] \to \mathbb{E}_p[f(\mathbf{x})]$) and $\langle f \rangle_{\text{SNIS}} \to \mathbb{E}_p[f(\mathbf{x})]$ as $N \to \infty$.

We will further discuss importance sampling at various points throughout the book. For example, we will see that it plays a central role in defining the ELBO in the chapter on VAEs (Chapter 7), and in Chapter 12 we will extend it further to be defined over an extended space of distributions as part of the annealed importance sampling framework. This chapter is also where we will discuss the variance of the IS and SNIS estimators in more detail.

1.3.2 Sampling Methods

In the previous section, we saw that obtaining a Monte Carlo estimation of an expectation $\mathbb{E}_p[f(\mathbf{x})]$ requires drawing N independent and identically distributed samples from the distribution $p(\mathbf{x})$. This, however, is in itself a nontrivial problem; some chapters in the book are dedicated to methods that try to achieve this; for example, Chapter 11 uses Markov chain Monte Carlo methods. We now highlight two general approaches, based on either using a deterministic transformation to transform a simple distribution into a target distribution, or by accepting or rejecting samples from a proposal distribution.

Invertible Transformation Methods

When a distribution $p(\mathbf{x})$ is linked to some other less complex distribution $q(\mathbf{z})$ through some deterministic transformation $\mathbf{x} = g(\mathbf{z})$, we can exploit this relationship to define a sampling method for $p(\mathbf{x})$ by sampling from $q(\mathbf{z})$ and then transforming the samples to $\mathbf{x}$ using g.

Notably, the exact relationship between $p(\mathbf{x})$ and $q(\mathbf{z})$ in this case is given by the change of variables formula:

$$p(\mathbf{x}) = q(g^{-1}(\mathbf{x})) \left| \det \frac{\partial g^{-1}(\mathbf{x})}{\partial \mathbf{x}} \right|. \tag{1.48}$$

The Jacobian determinant $\left| \det \frac{\partial g^{-1}(\mathbf{x})}{\partial \mathbf{x}} \right|$ represents how infinitesimal volume elements from the simpler distribution $q(\mathbf{z})$ are scaled when transformed into samples from $p(\mathbf{x})$ via g. This formula will appear again in detail in the chapter on normalizing flows (Chapter 8).

Example: Box–Muller Transform

A classic example of this is the Box–Muller transform, which transforms a pair of independent uniform random variables $\mathbf{u} = (u_1, u_2)$ into a pair of independent standard normal random variables $\mathbf{x} = (x_1, x_2)$ through the transformation

$$x_1 = \sqrt{-2 \log u_1} \cos(2\pi u_2), \qquad x_2 = \sqrt{-2 \log u_1} \sin(2\pi u_2). \tag{1.49}$$

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 03:28:49, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009709071.003

Example: Inverse CDF

Formally, a PDF is characterized by its cumulative distribution function (CDF) $F(x) = \int_{-\infty}^{x} \mathrm{d}x'\, p(x')$. If this CDF is invertible, we can use it to define a sampling method for $p(x)$ by drawing a uniform random number $u \sim \mathrm{Unif}(0, 1)$ and setting $x = F^{-1}(u)$. This simple approach is limited to distributions that have an invertible CDF and gets complicated for high-dimensional distributions.

Rejection Sampling

A second approach, which does not rely on the existence of an invertible deterministic transformation, is accept–reject sampling. Given our target distribution $p(\mathbf{x})$, accept–reject sampling uses a proposal distribution $q(\mathbf{x})$ to sample from, but instead of transforming the samples $\mathbf{x}$, it accepts them with a probability proportional to the ratio $p(\mathbf{x})/(Mq(\mathbf{x}))$, where M is a constant such that $p(\mathbf{x}) \leq Mq(\mathbf{x})$ for all $\mathbf{x}$. The collection of accepted samples is then distributed according to $p(\mathbf{x})$.

The efficiency of this method is determined by the choice of proposal distribution $q(\mathbf{x})$ and the constant M. The closer the proposal distribution $q(\mathbf{x})$ is to the target distribution $p(\mathbf{x})$, the more efficient the sampling will be. For high-dimensional distributions, accept–reject sampling is often not practical as the acceptance rate is typically very low.

1.4 Symmetries

Symmetries play a very important role throughout physics. They have unified electric and magnetic fields through Lorentz transformations between observers, and they are at the root of general relativity by formulating all physical laws covariant with respect to arbitrary invertible coordinate transformations, which led to the unification of gravity and acceleration as two sides of the same coin. They also organize elementary particles into families related to each other by symmetry transformations.

In this book, we will encounter one very fundamental symmetry in particular: *time reversal symmetry*. This is the idea that the laws of physics should remain the same if we run the "movie" forward or backward in time. If you watch a bunch of molecules jumping around, you cannot tell whether the movie is being played normally or in reverse. This is in sharp contrast to everyday life, where you will never see an omelet unscramble itself. We say that time reversal symmetry is broken at the macroscopic scale for thermodynamical systems that are not in equilibrium. Time reversal symmetry is restored in equilibrium.

The importance of time reversal symmetry (or its breaking) is that concepts such as heat production and entropy production, both macroscopic concepts that have no real analogy at the microscopic level, can be directly tied to it. In loose terms, the total amount of asymmetry between forward and backward dynamics determines the amount of entropy production in the system. This is made more precise in the famous fluctuation theorems of Crooks and Jarzynski (Jarzynski, 2011; Seifert, 2012).

In this section, we will provide some background on the mathematics of symmetries.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 03:28:49, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009709071.003

Let us assume we have a vector field $\mathbf{v}(\mathbf{x})$ and we want to understand what happens if we rotate the vector field. To describe the rotated vector field, we add a prime: $\mathbf{v}'(\mathbf{x})$. We now ask, how is the new vector field related to the old vector field? After some thinking, we find

$$\mathbf{v}'(\mathbf{x}) = R\mathbf{v}(r^{-1}\mathbf{x}), \tag{1.50}$$

where we use the representation r of the rotation group to rotate the location $\mathbf{x}$ and the representation R to rotate the tangent vector $\mathbf{v}$. These can be different in general, but in this example, you may assume that they are the same. In words, this equation says that in order to get the new vector $\mathbf{v}'$ at position $\mathbf{x}$, you need to rotate $\mathbf{x}$ back to the position $r^{-1}\mathbf{x}$ before the rotation, and then rotate that vector by an amount R. A similar rule applies to matrices and higher-order tensors:

$$\mathbf{M}'(\mathbf{x}) = R\mathbf{M}(r^{-1}\mathbf{x})R^\top \;\Rightarrow\; M'_{i,j}(\mathbf{x}) = \sum_{i'j'} R_{i,i'}R_{j,j'}M_{i',j'}(r^{-1}\mathbf{x}), \tag{1.51}$$

$$T'(\mathbf{x}) = R \otimes R \otimes \cdots \otimes R\, T(r^{-1}\mathbf{x}) \;\Rightarrow$$

$$T'_{i_1,i_2,\ldots,i_d}(\mathbf{x}) = \sum_{i'_1,i'_2,\ldots,i'_d} R_{i_1,i'_1}R_{i_2,i'_2}\cdots R_{i_d,i'_d}\, T_{i'_1,i'_2,\ldots,i'_d}(r^{-1}\mathbf{x}). \tag{1.52}$$

These transformations often have the structure of a group. A group consists of elements that compose into new group elements: a sequence of two rotations is another rotation. Other properties of group transformations include that there is an identity transformation, an inverse transformation for every group element, and that group transformations are associative.

The particular form in which they act on a vector space depends on the space's dimension. This form is called the representation. Representations need to respect the group structure, in the sense that if $M(g)$ is a representation of a group element g, then $M(g) \cdot M(h) = M(g \circ h)$. We call this a homomorphism.

The group structure for time inversion is very simple indeed; it is an *involution*. Namely, applying time reversal Ω twice gets you back where you started:

$$\Omega^2 = \mathbb{1}. \tag{1.53}$$

Physical variables transform under certain "symmetry transformations." For involutions there are two types of such variables, even- and odd-parity variables, where the latter switch sign under time reversal. For a trajectory $(\mathbf{x}(\tau), \mathbf{p}(\tau))$ over a time interval $[0, \tau_f]$, the time-reversed trajectory, denoted by a tilde, is defined as

$$\tilde{\mathbf{x}}(\tau) = \Omega\mathbf{x}(\tilde{\tau}) = \mathbf{x}(\tau_f - \tau), \tag{1.54}$$

$$\tilde{\mathbf{p}}(\tau) = \Omega\mathbf{p}(\tilde{\tau}) = -\mathbf{p}(\tau_f - \tau). \tag{1.55}$$

For both cases we indeed have that applying time reversal twice will give us back the original variables.

A symmetry is now "something" that is invariant under group transformations. A circle is invariant under rotations of the circle. We are interested in probability distributions that are invariant and in physical laws that are invariant. At the microscopic

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 03:28:49, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009709071.003

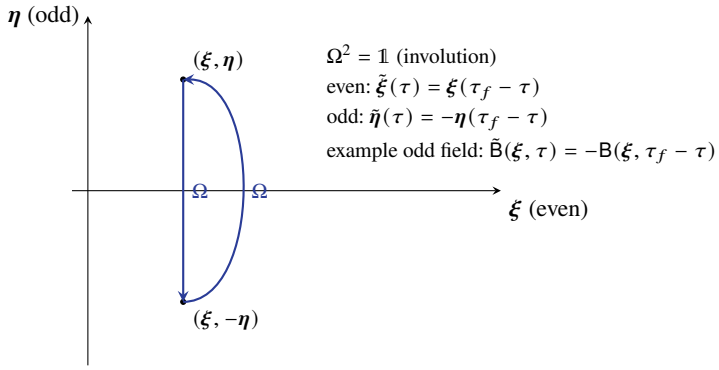


Figure 1.3 Time reversal parity in phase space. Positions (even variables) remain unchanged under time reversal, while momenta (odd variables) flip sign: $\Omega\colon\ (\boldsymbol{\xi}, \boldsymbol{\eta}) \mapsto (\boldsymbol{\xi}, -\boldsymbol{\eta})$. The operator Ω is an involution: $\Omega^2 = \mathbb{1}$.

level all physical laws are believed to be invariant under joint charge–parity–time reversal transformations. In this book we consider parity and time reversal. For instance, Hamilton's equations are invariant under time reversal, in the sense that a trajectory that starts at $\mathbf{x}(0), \mathbf{p}(0)$ and ends in the state $\mathbf{x}(\tau_f), \mathbf{p}(\tau_f)$ has, under Hamilton's equations of motion, a dual path that starts at $\mathbf{x}(\tau_f), -\mathbf{p}(\tau_f)$ and evolves under the same equations back to the initial state (with flipped momenta) $\mathbf{x}(0), -\mathbf{p}(0)$; see Figure 1.3.

In thermodynamics, the situation is more complex. The stochastic Brownian-like motion is not invertible in general because, thanks to collisions, we are losing information about the system. That means that the usual stochastic equations of motion are not time-reversal symmetric. The degree by which we lose information is measured by the entropy increase in the system. When the system is in equilibrium, time reversal symmetry is restored and no more entropy is produced (we have lost all the information there is to lose that is consistent with macroscopic variables such as average energy, temperature, and pressure).

1.5 Essentials of Convex Optimization

1.5.1 The Primal Problem

A constrained optimization problem can be formally expressed as

$$\min_{\mathbf{x}} \quad f_0(\mathbf{x})$$
$$\text{subject to: } \begin{cases} f_i(\mathbf{x}) \le 0, & \text{for all } i \\ h_j(\mathbf{x}) = 0, & \text{for all } j. \end{cases} \tag{1.56}$$

Here, we have an objective function $f_0(\mathbf{x})$ that we want to minimize, subject to inequality constraints $f_i(\mathbf{x}) \le 0$ and equality constraints $h_j(\mathbf{x}) = 0$. The goal is to find

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 03:28:49, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009709071.003

the optimal solution $\mathbf{x}^*$ that satisfies all the constraints while minimizing the objective function.

To gain intuition about this problem, consider the analogy of a ball rolling down a surface defined by the objective function $f_0(\mathbf{x})$. The ball's motion is constrained by walls, represented by the inequality constraints $f_i(\mathbf{x}) \leq 0$, and tracks, represented by the equality constraints $h_j(\mathbf{x}) = 0$. The optimal solution $\mathbf{x}^*$ corresponds to the position where the ball comes to rest, satisfying all the constraints while minimizing its potential energy.

In convex optimization, (1.56) is known as the primal problem. We now write the primal Lagrangian of this problem, which will be helpful in the following development:

$$\mathcal{L}_{\mathrm{P}}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = f_0(\mathbf{x}) + \sum_i \lambda_i f_i(\mathbf{x}) + \sum_j \nu_j h_j(\mathbf{x}), \tag{1.57}$$

where $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots)$ and $\boldsymbol{\nu} = (\nu_1, \nu_2, \ldots)$ are Lagrange multipliers associated with the inequality and equality constraints, respectively. The Lagrange multipliers can be interpreted as the "price" or "cost" of violating the constraints. In the ball analogy, λ_i represents the reaction force exerted by the wall on the ball when it is in contact with the wall, and ν_j represents the force exerted by the track on the ball to keep it on the track. We will assume in the following that $\lambda_i \geq 0$ for all i.

1.5.2 The Dual Problem

The primal problem, (1.56), can be transformed into an alternative formulation called the dual problem. From the primal Lagrangian, we can define the dual function as

$$\mathcal{L}_{\mathrm{D}}(\boldsymbol{\lambda}, \boldsymbol{\nu}) = \inf_{\mathbf{x}} \mathcal{L}_{\mathrm{P}}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}). \tag{1.58}$$

Here, "inf" denotes the infimum, which is the greatest lower bound of a set. For optimization problems, we often use inf instead of min when we cannot guarantee that the minimum is actually achieved within the feasible set. Similarly, "sup" denotes the supremum, which is the least upper bound of a set. In this context, it represents the lowest value of the primal Lagrangian $\mathcal{L}_{\mathrm{P}}$ over all possible values of $\mathbf{x}$, for fixed values of $\boldsymbol{\lambda}$ and $\boldsymbol{\nu}$. This objective can actually become $-\infty$ for certain values of its arguments. We will call parameters $\boldsymbol{\lambda} \geq 0, \boldsymbol{\nu}$ for which $\mathcal{L}_{\mathrm{D}}(\boldsymbol{\lambda}, \boldsymbol{\nu}) > -\infty$ *dual feasible*.

It is important to note that the dual Lagrangian, (1.58), is a concave function of $\boldsymbol{\lambda}, \boldsymbol{\nu}$ because it is a pointwise infimum of a family of linear functions in $\boldsymbol{\lambda}, \boldsymbol{\nu}$. Hence, even if the primal problem, (1.56), is not convex, the dual is always concave (Boyd and Vandenberghe, 2004)!

It is not hard to show that

$$\mathcal{L}_{\mathrm{D}}(\boldsymbol{\lambda}, \boldsymbol{\nu}) \leq p^*, \tag{1.59}$$

where p^* is the primal optimal value. This follows because for any primal feasible point $\mathbf{x}$, we have $f_i(\mathbf{x}) \leq 0$ and $h_j(\mathbf{x}) = 0$. With $\lambda_i \geq 0$, the sum $\sum_i \lambda_i f_i(\mathbf{x}) + \sum_j \nu_j h_j(\mathbf{x}) \leq 0$. Therefore, $\mathcal{L}_{\mathrm{P}}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) \leq f_0(\mathbf{x})$ for any feasible $\mathbf{x}$. Taking the infimum over all $\mathbf{x}$ on the left and the minimum over feasible $\mathbf{x}$ on the right gives $\mathcal{L}_{\mathrm{D}}(\boldsymbol{\lambda}, \boldsymbol{\nu}) \leq p^*$.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 03:28:49, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009709071.003

Thus, the dual problem always provides a lower bound to the primal problem, as seen in (1.59). The tightest lower bound can be found by solving the dual problem,

$$\begin{aligned} \max_{\boldsymbol{\lambda},\boldsymbol{\nu}} \quad & \mathcal{L}_{\mathrm{D}}(\boldsymbol{\lambda},\boldsymbol{\nu}) \\ \text{subject to:} \quad & \lambda_i \geq 0 \text{ for all } i, \end{aligned} \tag{1.60}$$

which is therefore a convex optimization problem.

1.5.3 Weak and Strong Duality

If we denote by d^* the dual optimal value, we always have

$$d^* = \mathcal{L}_{\mathrm{D}}(\boldsymbol{\lambda}^*,\boldsymbol{\nu}^*) \leq p^*, \tag{1.61}$$

where $\boldsymbol{\lambda}^*$ and $\boldsymbol{\nu}^*$ are the optimal solutions to the dual problem. This property is known as *weak duality*. The difference between the primal and dual optimal values, $p^* - d^*$, is called the *duality gap*.

Strong duality holds when $p^* = d^*$; that is, when the duality gap is zero. Strong duality is particularly useful, especially if we can express the primal solution $\mathbf{x}^*$ in terms of the dual solution $\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*$. In such cases, we can simply solve the dual problem, (1.60), and convert the answer to the primal domain, knowing that the solution must be optimal. Often, the dual problem is easier to solve.

So when does strong duality hold? Up to some mathematical details, the answer is: *if the primal problem,* (1.56)*, is convex and the equality constraints are affine* (Boyd and Vandenberghe, 2004). This means that $f_0(\mathbf{x})$ and $\{f_i(\mathbf{x})\}$ are convex functions and $h_j(\mathbf{x}) = A_j \cdot \mathbf{x} - b_j$.

The primal problem, (1.56), can be written as

$$p^* = \inf_{\mathbf{x}} \sup_{\boldsymbol{\lambda}\geq 0,\boldsymbol{\nu}} \mathcal{L}_{\mathrm{P}}(\mathbf{x},\boldsymbol{\lambda},\boldsymbol{\nu}). \tag{1.62}$$

This can be understood by noting that $\sup_{\boldsymbol{\lambda}\geq 0,\boldsymbol{\nu}} \mathcal{L}_{\mathrm{P}}(\mathbf{x},\boldsymbol{\lambda},\boldsymbol{\nu}) = f_0(\mathbf{x})$ when $\mathbf{x}$ is feasible but ∞ otherwise. To see this, first observe that by violating one of the constraints, you can find a choice of $\boldsymbol{\lambda},\boldsymbol{\nu}$ that makes the Lagrangian, (1.57), infinite. Also, when all the constraints are satisfied, the best we can do is maximize the additional terms to be zero, which is always possible. For instance, we can simply set all $\boldsymbol{\lambda},\boldsymbol{\nu}$ to zero, even though this is not necessary if the constraints themselves vanish.

The dual problem by definition is given by

$$d^* = \sup_{\boldsymbol{\lambda}\geq 0,\boldsymbol{\nu}} \inf_{\mathbf{x}} \mathcal{L}_{\mathrm{P}}(\mathbf{x},\boldsymbol{\lambda},\boldsymbol{\nu}). \tag{1.63}$$

Hence, the "sup" and "inf" can be interchanged if strong duality holds, so the optimal solution is a saddle point. It is important to realize that the order of maximization and minimization matters for arbitrary functions (but not for convex–concave functions, as the minimax theorem guarantees). Imagine a "V-shaped" valley running diagonally across the coordinate system. If we first maximize over one direction, keeping the other direction fixed, and then minimize the result, we end up with the lowest point on the rim. If we reverse the order, we end up with the highest point in the valley.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 03:28:49, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009709071.003

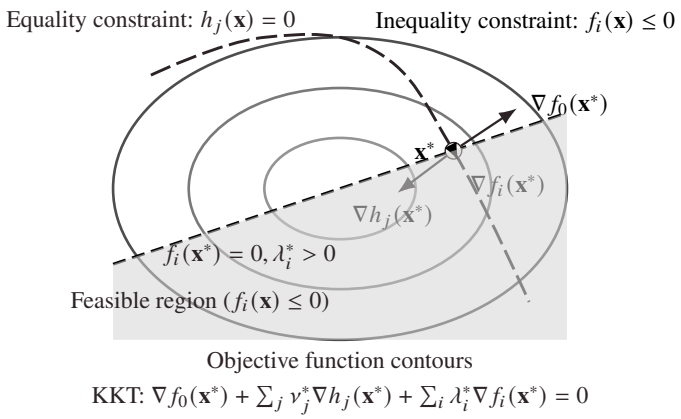


Figure 1.4 Geometric interpretation of constrained optimization and KKT conditions. The optimal point $\mathbf{x}^*$ lies where the long-dashed equality contour $h_j(\mathbf{x}) = 0$ intersects the dashed inequality-constraint boundary $f_i(\mathbf{x}) \leq 0$. The gradient vectors at $\mathbf{x}^*$ illustrate the stationarity condition: $\nabla f_0(\mathbf{x}^*)$ is perpendicular to the objective contour, while the gradients of the constraints oppose motion into infeasible regions. Their weighted sum balances to zero, and the hatched wedge marks the feasible set.

1.5.4 Karush–Kuhn–Tucker Conditions

The Karush–Kuhn–Tucker (KKT) conditions are a set of necessary conditions that hold for problems with zero duality gap. These conditions turn out to be sufficient for convex optimization problems (Boyd and Vandenberghe, 2004). The KKT conditions, illustrated in Figure 1.4, are given by

$$\nabla f_0(\mathbf{x}^*) + \sum_i \lambda_i^* \nabla f_i(\mathbf{x}^*) + \sum_j \nu_j^* \nabla h_j(\mathbf{x}^*) = 0, \tag{1.64}$$

$$f_i(\mathbf{x}^*) \leq 0, \tag{1.65}$$

$$h_j(\mathbf{x}^*) = 0, \tag{1.66}$$

$$\lambda_i^* \geq 0, \tag{1.67}$$

$$\lambda_i^* f_i(\mathbf{x}^*) = 0. \tag{1.68}$$

The first condition, (1.64), known as the stationarity condition, can be easily derived from (1.62), which states that $p^* = \inf_{\mathbf{x}} \mathcal{L}_{\mathrm{P}}(\mathbf{x}, \boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$. This implies that all the derivatives of the Lagrangian with respect to $\mathbf{x}$ must vanish at the optimal point $\mathbf{x}^*$. This condition has a nice interpretation as a "balancing of forces." Imagine a ball rolling down a surface defined by $f_0(\mathbf{x})$ (i.e., you are doing gradient descent to find the minimum). The ball gets blocked by a wall, which is the constraint surface. If the surface and constraint are convex, then at the point where the ball does not move, we have reached the optimal solution. At that point, the forces on the ball must balance. The first term represents the force of the ball against the wall due to gravity (the ball is still on a slope). The second term represents the reaction force of the wall in the opposite direction. The λ_i^* represents the magnitude of the reaction force, which needs

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 03:28:49, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009709071.003

to be higher if the surface slopes more. We say that this constraint is "active". Other constraints which do not exert a force are "inactive" and have $\lambda_i^* = 0$.

The final KKT condition, (1.68), known as *complementary slackness*, states that either $f_i(\mathbf{x}^*) = 0$ (the constraint is saturated and hence active), in which case λ_i^* is allowed to take on a nonzero value, or if the constraint is inactive ($f_i(\mathbf{x}^*) < 0$), so λ_i^* must vanish.

Complementary slackness, (1.68), can be derived as follows:

$$f_0(\mathbf{x}^*) = \mathcal{L}_\mathrm{D}(\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*) \tag{1.69}$$

$$= \inf_{\mathbf{x}} \left(f_0(\mathbf{x}) + \sum_i \lambda_i^* f_i(\mathbf{x}) + \sum_j \nu_j^* h_j(\mathbf{x}) \right) \tag{1.70}$$

$$\leq f_0(\mathbf{x}^*) + \sum_i \lambda_i^* f_i(\mathbf{x}^*) + \sum_j \nu_j^* h_j(\mathbf{x}^*) \tag{1.71}$$

$$\leq f_0(\mathbf{x}^*). \tag{1.72}$$

Here, (1.69) follows from strong duality, (1.70) is the definition of the dual Lagrangian, (1.71) holds because the infimum is always less than or equal to the value at any point $\mathbf{x}^*$, and (1.72) follows from the feasibility of $\mathbf{x}^*$ – (1.65) and (1.66) – and the non-negativity of λ_i^*, (1.67). Since the first and final expressions in this derivation are equal, all the inequalities must be tight, which implies that each term $\lambda_i^* f_i(\mathbf{x}^*)$ must be zero, yielding the complementary slackness condition.

By solving the KKT conditions, we can find the optimal primal and dual solutions, $\mathbf{x}^*$ and $(\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$, respectively.

1.5.5 Canonical Ensemble from the Maximum Entropy Principle

A principle proposed by E. T. Jaynes for both physical systems and information systems is that in the absence of information (measured by entropy), the best solution to a problem with some given observations is the one that is consistent with the observations but otherwise maximizes entropy over the remaining degrees of freedom (Jaynes, 1957a,b). In some sense, the system assumes as little as possible about things it does not know. This is sometimes called Occam's razor in Bayesian statistics (Jaynes, 2003).

For a canonical ensemble, we assume that the average energy of the system is known and equal to some value E. The maximum entropy (MaxEnt) problem can then be formulated as

$$\rho^\mathrm{eq} = \operatorname*{argmax}_{\rho} \ \mathcal{L}_\mathrm{P}(\rho, \beta, \lambda) \tag{1.73}$$

$$= \operatorname*{argmax}_{\rho} \left[\mathcal{S}[\rho] - \beta \left(\int \mathrm{d}\boldsymbol{\zeta}\, \rho(\boldsymbol{\zeta}) \mathcal{H}(\boldsymbol{\zeta}) - E \right) - \lambda \left(\int \mathrm{d}\boldsymbol{\zeta}\, \rho(\boldsymbol{\zeta}) - 1 \right) \right], \tag{1.74}$$

where $\rho(\boldsymbol{\zeta})$ is the probability distribution over microstates $\boldsymbol{\zeta}$, $\mathcal{S}[\rho]$ is the entropy functional, and $\mathcal{H}(\boldsymbol{\zeta})$ is the Hamiltonian of the system. Here, we have two constraints enforced by the Lagrange multipliers β and λ, ensuring that the average energy of the system is equal to E and the probability distribution is normalized. Taking (functional)

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 03:28:49, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009709071.003

derivatives of the primal Lagrangian $\mathcal{L}_\mathrm{P}$ with respect to ρ and setting it to zero gives the stationarity condition in the KKT conditions:

$$\frac{\delta \mathcal{L}_\mathrm{P}}{\delta \rho} = -\log \rho(\boldsymbol{\zeta}) - 1 - \beta \mathcal{H}(\boldsymbol{\zeta}) - \lambda = 0, \tag{1.75}$$

which is solved by

$$\rho^{\mathrm{eq}}(\boldsymbol{\zeta}) = \exp(-\beta \mathcal{H}(\boldsymbol{\zeta}) - \lambda - 1) = \frac{1}{Z} \exp(-\beta \mathcal{H}(\boldsymbol{\zeta})), \quad \lambda = \log Z - 1. \tag{1.76}$$

The Lagrange multiplier λ is easily solved in terms of the normalization constant Z. The other Lagrange multiplier, β, is identified with the inverse temperature $1/(k_\mathrm{B}T)$, with k_B being the Boltzmann constant and T the temperature of the heat bath. Note that we could have added a separate inequality constraint to enforce $\rho \geq 0$, but thanks to the exponential form of the solution, this constraint is automatically satisfied. We set $k_\mathrm{B} = 1$ in what follows.

The next step is to insert the solution (1.76) of the KKT condition (1.75) back into the primal Lagrangian, (1.73), giving the dual Lagrangian:

$$\mathcal{L}_\mathrm{D}(\beta, \lambda) = \beta E + \lambda + \int \mathrm{d}\boldsymbol{\zeta} \ \exp(-\beta \mathcal{H}(\boldsymbol{\zeta}) - \lambda - 1). \tag{1.77}$$

Defining $\tilde{\lambda} = \lambda + 1$ and solving for $\tilde{\lambda}$ to normalize the distribution, we get

$$\tilde{\lambda}^* = \log Z(\beta) = \log \int \mathrm{d}\boldsymbol{\zeta} \ \exp(-\beta \mathcal{H}(\boldsymbol{\zeta})), \tag{1.78}$$

which leads to

$$\mathcal{L}_\mathrm{D}(\beta, \lambda^*) = \beta E + \log Z(\beta) = \beta(E - F(\beta)) = S(\beta), \tag{1.79}$$

where we have used the definition of the equilibrium free energy as $F = -T \log Z$, and the definition of the entropy from $F = E - TS$. We note that this is the entropy at equilibrium as it is a function of β only and not of ρ. This is a nice result! The dual Lagrangian, (1.77), of the (nonequilibrium) maximum entropy problem, (1.73), is an expression for the entropy of the Boltzmann distribution at equilibrium, which we can now maximize over β. Since we know the dual Lagrangian is always concave, this should be a tractable optimization problem. Computing derivatives of $\mathcal{L}_\mathrm{D}$ with respect to β, we get

$$\frac{\partial \mathcal{L}_\mathrm{D}}{\partial \beta} = E - \mathbb{E}_{\rho^{\mathrm{eq}}(\beta)}[\mathcal{H}]. \tag{1.80}$$

We can perform iterative gradient ascent updates on $\mathcal{L}_\mathrm{D}$ until we reach the maximum. The resulting β^* is the inverse temperature of the system that corresponds to an average energy of E.

Since the original problem, (1.73), was convex, we have strong duality which means the primal and dual optimal values are equal; that is, $p^* = d^*$. This ensures that our solution is indeed the global optimum.

To conclude, Jaynes' maximum entropy principle neatly provides us with a derivation of the canonical ensemble, (1.76), or the Boltzmann distribution, through the

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 03:28:49, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009709071.003

method of Lagrange multipliers with the interpretation of the inverse temperature β as the Lagrange multiplier that enforces the average energy constraint.

1.6 The Lagrangian versus Eulerian View

In both stochastic thermodynamics and machine learning, we often need to track how probability distributions evolve over time. There are two complementary perspectives for performing such calculations: the Eulerian view, where we observe the system from a fixed reference frame in space, and the Lagrangian view, where we follow individual particles (or samples) as they move. In the fixed Eulerian reference frame, we define space-time coordinates $(\zeta, \tau) = (\boldsymbol{\xi}, \boldsymbol{\eta}, \tau)$ in phase space and fields as functions of these fixed coordinates; that is, we may have a Hamiltonian $\mathcal{H}(\zeta, \tau)$. In general we will write $\mathbf{h}(\zeta, \tau)$ for such functions, which can be scalars, vectors ("fields") or even matrices and higher-order tensors. The particle density is described by the density field $\rho(\zeta, \tau)$, which satisfies the following continuity equation:

$$\partial_\tau \rho(\zeta, \tau) = -\boldsymbol{\nabla}_\zeta \cdot (\rho(\zeta, \tau)\mathbf{v}(\zeta, \tau)) = -\mathbf{div}(\rho(\zeta, \tau)\mathbf{v}(\zeta, \tau)), \tag{1.81}$$

for some vector field $\mathbf{v}$. This is just the continuum expression for what is known as a push-forward (deterministic) transformation of some density along the vector field $\mathbf{v}$. When we compute an average we compute

$$\int \mathrm{d}\zeta \; \rho(\zeta, \tau) \; \mathbf{h}(\zeta, \tau). \tag{1.82}$$

In the above we note that $\rho(\zeta, \tau)$ is a different (transformed) distribution than $\rho(\zeta, 0)$. If this density transforms under a deterministic (push-forward) transformation $\zeta_0 \to \zeta_\tau = g(\zeta_0, \tau)$ for all ζ_0, then we have the following relation:

$$\rho(\zeta_\tau, \tau) = \rho(g(\zeta_0, \tau), \tau) = \rho(\zeta_0, 0) \; \left| \det \left(\frac{\partial g(\zeta_0, \tau)}{\partial \zeta_0} \right) \right|^{-1}. \tag{1.83}$$

Note that the measure will remain invariant:

$$\rho(\zeta_\tau, \tau) \, \mathrm{d}\zeta_\tau = \rho(\zeta_0, 0) \, \mathrm{d}\zeta_0. \tag{1.84}$$

When we compute an average we can thus do the following:

$$\begin{aligned}
&\int \mathrm{d}\zeta \; \rho(\zeta, \tau) \; \mathbf{h}(\zeta, \tau) \\
&\quad = \int \mathrm{d}\zeta_\tau \; \rho(\zeta_\tau, \tau) \mathbf{h}(\zeta_\tau, \tau) \quad (\zeta \text{ is a dummy integration variable}) \\
&\quad = \int \mathrm{d}\zeta_0 \; \rho(\zeta_0, 0) \; \mathbf{h}(g(\zeta_0, \tau), \tau) \quad (\text{invariance of measure}) \\
&\quad = \iint \mathrm{d}\zeta_0 \; \mathrm{d}\zeta_\tau \; \rho(\zeta_0, 0) \delta(\zeta_\tau - g(\zeta_0, \tau)) \; \mathbf{h}(\zeta_\tau, \tau),
\end{aligned} \tag{1.85}$$

where we can define

$$\rho(\zeta_\tau, \zeta_0) = \rho(\zeta_\tau | \zeta_0) \rho(\zeta_0) = \rho(\zeta_0, 0) \delta(\zeta_\tau - g(\zeta_0, \tau)). \tag{1.86}$$

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 03:28:49, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009709071.003

In the first two lines we consider $\boldsymbol{\zeta}_0$ as a random variable over which we need to integrate. Moreover, both the density $\rho(\boldsymbol{\zeta}, \tau)$ and the function $\mathbf{h}(\boldsymbol{\zeta}, \tau)$ are considered functions of τ. We call this the Eulerian view as the random variables are fixed. In the third line, however, we consider a fixed measure $\rho(\boldsymbol{\zeta}_0, 0)\, \mathrm{d}\boldsymbol{\zeta}_0$, which is the initial measure from which we sample ensemble points. Now the function $\mathbf{h}(\boldsymbol{\zeta}_\tau, \tau)$ depends on time in two ways: first explicitly, but also implicitly through $\boldsymbol{\zeta}_\tau = g(\boldsymbol{\zeta}_0, \tau)$. We call this the Lagrangian view. This result is also exploited in the reparameterization trick in VAEs.

Let us compute the time derivative of the expression above:

$$
\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}\tau} \int \mathrm{d}\boldsymbol{\zeta}\; \rho(\boldsymbol{\zeta}, \tau)\; \mathbf{h}(\boldsymbol{\zeta}, \tau) \\
&= \int \mathrm{d}\boldsymbol{\zeta}\; \partial_\tau \rho(\boldsymbol{\zeta}, \tau)\; \mathbf{h}(\boldsymbol{\zeta}, \tau) + \rho(\boldsymbol{\zeta}, \tau)\; \partial_\tau \mathbf{h}(\boldsymbol{\zeta}, \tau) \\
&= \int \mathrm{d}\boldsymbol{\zeta}\; \boldsymbol{\nabla}_{\boldsymbol{\zeta}} \cdot (-\rho(\boldsymbol{\zeta}, \tau)\mathbf{v}(\boldsymbol{\zeta}, \tau))\; \mathbf{h}(\boldsymbol{\zeta}, \tau) + \rho(\boldsymbol{\zeta}, \tau)\; \partial_\tau \mathbf{h}(\boldsymbol{\zeta}, \tau) \\
&= \int \mathrm{d}\boldsymbol{\zeta}\; \rho(\boldsymbol{\zeta}, \tau)\; (\mathbf{v}(\boldsymbol{\zeta}, \tau) \cdot \boldsymbol{\nabla}_{\boldsymbol{\zeta}} \mathbf{h}(\boldsymbol{\zeta}, \tau) + \partial_\tau \mathbf{h}(\boldsymbol{\zeta}, \tau)) \\
&= \int \mathrm{d}\boldsymbol{\zeta}_\tau\; \rho(\boldsymbol{\zeta}_\tau, \tau)\; (\mathbf{v}(\boldsymbol{\zeta}_\tau, \tau) \cdot \boldsymbol{\nabla}_{\boldsymbol{\zeta}_\tau} \mathbf{h}(\boldsymbol{\zeta}_\tau, \tau) + \partial_\tau \mathbf{h}(\boldsymbol{\zeta}_\tau, \tau)) \\
&= \int \mathrm{d}\boldsymbol{\zeta}_0\; \rho(\boldsymbol{\zeta}_0, 0)\; (\mathbf{v}(\boldsymbol{\zeta}_\tau, \tau) \cdot \boldsymbol{\nabla}_{\boldsymbol{\zeta}_\tau} \mathbf{h}(\boldsymbol{\zeta}_\tau, \tau) + \partial_\tau \mathbf{h}(\boldsymbol{\zeta}_\tau, \tau)); \quad \boldsymbol{\zeta}_\tau = g(\boldsymbol{\zeta}_0, \tau) \\
&= \int \mathrm{d}\boldsymbol{\zeta}_0\; \rho(\boldsymbol{\zeta}_0, 0)\; \frac{\mathrm{d}}{\mathrm{d}\tau} \mathbf{h}(\boldsymbol{\zeta}_\tau, \tau); \quad \boldsymbol{\zeta}_\tau = g(\boldsymbol{\zeta}_0, \tau) \\
&= \frac{\mathrm{d}}{\mathrm{d}\tau} \int \mathrm{d}\boldsymbol{\zeta}_0\; \rho(\boldsymbol{\zeta}_0, 0)\; \mathbf{h}(\boldsymbol{\zeta}_\tau, \tau); \quad \boldsymbol{\zeta}_\tau = g(\boldsymbol{\zeta}_0, \tau).
\end{aligned}
\tag{1.87}
$$

The equality of the first line with the final line is consistent with the results above on the Eulerian and Lagrangian views of computing an average. In the first line we need to take two partial derivatives because both ρ and $\mathbf{h}$ depend on τ. However, $\boldsymbol{\zeta}_0$ is just an integration variable and does not depend on τ. In the last line, we take the time derivative in the Lagrangian view, where ρ is now fixed (the initial distribution) while $\mathbf{h}$ depends on τ in two ways, explicitly and implicitly through $\boldsymbol{\zeta}_\tau$. The total derivative of $\mathbf{h}$ is sometimes called the "material" or "comoving" derivative. We note that to connect the two views we needed the continuity equation and partial integration.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 17 Nov 2026 at 03:28:49, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781009709071.003