Contents

Contents i

List of Tables ii

List of Figures iii

1 Fundamentals of Probability 1

1.1 References . . . 1

1.2 Statistical Properties of Random Walks . . . 2

1.2.1 One-dimensional random walk . . . 2

1.2.2 Thermodynamic limit . . . 4

1.2.3 Entropy and energy . . . 5

1.3 Basic Concepts in Probability Theory . . . 6

1.3.1 Fundamental definitions . . . 6

1.3.2 Bayesian statistics . . . 7

1.3.3 Random variables and their averages . . . 8

1.4 Entropy and Probability . . . 10

1.4.1 Entropy and information theory . . . 10

1.4.2 Rényi entropy . . . 11

1.4.3 Kullback-Leibler divergence . . . 12

1.4.4 Probability distributions from maximum entropy . . . 12

1.4.5 Continuous probability distributions . . . 16

1.5 General Aspects of Probability Distributions 17
1.5.1 Discrete and continuous distributions 17
1.5.2 Central limit theorem 19
1.5.3 Moments and cumulants 21
1.5.4 Multidimensional Gaussian integral 22
1.6 Appendix : Bayesian Statistical Inference 22
1.6.1 Frequentists and Bayesians 22
1.6.2 Updating Bayesian priors 23
1.6.3 Hyperparameters and conjugate priors 25
1.6.4 The problem with priors 26

List of Tables

List of Figures

1.1 The falling ball novelty 3
1.2 Approaching the Gaussian distribution as $N \to \infty$ 5

Chapter 1

Fundamentals of Probability

1.1 References

- C. Gardiner, *Stochastic Methods* (4^{th} edition, Springer-Verlag, 2010)
  Very clear and complete text on stochastic methods with many applications.
- J. M. Bernardo and A. F. M. Smith, *Bayesian Theory* (Wiley, 2000)
  A thorough textbook on Bayesian methods.
- D. Williams, *Weighing the Odds: A Course in Probability and Statistics* (Cambridge, 2001)
  A good overall statistics textbook, according to a mathematician colleague.
- E. T. Jaynes, *Probability Theory* (Cambridge, 2007)
  An extensive, descriptive, and highly opinionated presentation, with a strongly Bayesian approach.
- A. N. Kolmogorov, *Foundations of the Theory of Probability* (Chelsea, 1956)
  The *Urtext* of mathematical probability theory.

1.2 Statistical Properties of Random Walks

1.2.1 One-dimensional random walk

Consider the mechanical system depicted in fig. 1.1, a version of which is often sold in novelty shops. A ball is released from the top, which cascades consecutively through N levels. The details of each ball's motion are governed by Newton's laws of motion. However, to predict where any given ball will end up in the bottom row is difficult, because the ball's trajectory depends sensitively on its initial conditions, and may even be influenced by random vibrations of the entire apparatus. We therefore abandon all hope of integrating the equations of motion and treat the system statistically. That is, we assume, at each level, that the ball moves to the right with probability p and to the left with probability $q = 1 - p$. If there is no bias in the system, then $p = q = \frac{1}{2}$. The position X_N after N steps may be written

$$X_N = \sum_{j=1}^{N} \sigma_j \quad , \tag{1.1}$$

where $\sigma_j = +1$ if the ball moves to the right at level j, and $\sigma_j = -1$ if the ball moves to the left at level j. At each level, the probability for these two outcomes is given by

$$P_\sigma = p\,\delta_{\sigma,+1} + q\,\delta_{\sigma,-1} = \begin{cases} p & \text{if } \sigma = +1 \\ q & \text{if } \sigma = -1 \end{cases} \quad . \tag{1.2}$$

This is a normalized discrete probability distribution of the type discussed in section 1.5 below. The multivariate distribution for all the steps is then

$$P(\sigma_1\,,\,\ldots\,,\sigma_N) = \prod_{j=1}^{N} P(\sigma_j) \quad . \tag{1.3}$$

Our system is equivalent to a one-dimensional *random walk*. Imagine an inebriated pedestrian on a sidewalk taking steps to the right and left at random. After N steps, the pedestrian's location is X.

Now let's compute the average of X:

$$\langle X \rangle = \Big\langle \sum_{j=1}^{N} \sigma_j \Big\rangle = N\langle\sigma\rangle = N \sum_{\sigma=\pm1} \sigma\, P(\sigma) = N(p-q) = N(2p-1) \quad . \tag{1.4}$$

This could be identified as an *equation of state* for our system, as it relates a measurable quantity X to the number of steps N and the local bias p. Next, let's compute the average of X^2:

$$\langle X^2 \rangle = \sum_{j=1}^{N} \sum_{j'=1}^{N} \langle \sigma_j \sigma_{j'} \rangle = N^2(p-q)^2 + 4Npq \quad . \tag{1.5}$$

Here we have used

$$\langle \sigma_j \sigma_{j'} \rangle = \delta_{jj'} + \big(1 - \delta_{jj'}\big)(p-q)^2 = \begin{cases} 1 & \text{if } j = j' \\ (p-q)^2 & \text{if } j \neq j' \end{cases} \quad . \tag{1.6}$$

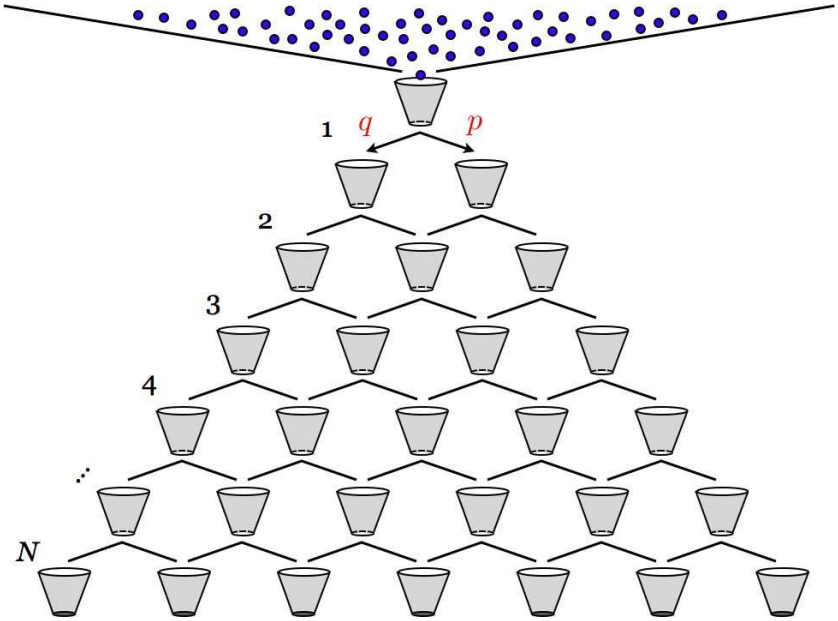


Figure 1.1: The falling ball system, which mimics a one-dimensional random walk.

Note that $\langle X^2\rangle \geq \langle X\rangle^2$, which must be so because

$$\mathrm{Var}(X) = \langle (\Delta X)^2\rangle \equiv \big\langle \big(X - \langle X\rangle\big)^2\big\rangle = \langle X^2\rangle - \langle X\rangle^2 \quad . \tag{1.7}$$

This is called the *variance* of X. We have $\mathrm{Var}(X) = 4Npq$. The *root mean square* deviation, $\Delta X_{\rm rms}$, is the square root of the variance: $\Delta X_{\rm rms} = \sqrt{\mathrm{Var}(X)}$. Note that the mean value of X is linearly proportional to N (for all cases other than $p = q = \frac{1}{2}$), but the RMS fluctuations $\Delta X_{\rm rms}$ are proportional to $N^{1/2}$. In the limit $N \to \infty$ then, the ratio $\Delta X_{\rm rms}/\langle X\rangle$ vanishes as $N^{-1/2}$. This is a consequence of the central limit theorem (see §1.5.2 below), and we shall meet up with it again on several occasions.

We can do even better. We can find the complete probability distribution for X. It is given by

$$P_{N,X} = \binom{N}{N_{\rm R}} p^{N_{\rm R}}\, q^{N_{\rm L}} \quad , \tag{1.8}$$

where $N_{\rm R/L}$ are the numbers of steps taken to the right/left, with $N = N_{\rm R} + N_{\rm L}$, and $X = N_{\rm R} - N_{\rm L}$. There are many independent ways to take $N_{\rm R}$ steps to the right. For example, our first $N_{\rm R}$ steps could all be to the right, and the remaining $N_{\rm L} = N - N_{\rm R}$ steps would then all be to the left. Or our final $N_{\rm R}$ steps could all be to the right. For each of these independent possibilities, the probability is $p^{N_{\rm R}}\, q^{N_{\rm L}}$. How many possibilities are there? Elementary combinatorics tells us this number is

$$\binom{N}{N_{\rm R}} = \frac{N!}{N_{\rm R}!\, N_{\rm L}!} \quad . \tag{1.9}$$

Note that $N \pm X = 2N_{\rm R/L}$, so we can replace $N_{\rm R/L} = \frac{1}{2}(N \pm X)$. Thus,

$$P_{N,X} = \frac{N!}{\left(\frac{N+X}{2}\right)!\,\left(\frac{N-X}{2}\right)!}\, p^{(N+X)/2}\, q^{(N-X)/2} \quad . \tag{1.10}$$

1.2.2 Thermodynamic limit

Consider the limit $N \to \infty$ but with $x \equiv X/N$ finite. This is analogous to what is called the *thermodynamic limit* in statistical mechanics. Since N is large, x may be considered a continuous variable. We evaluate $\log P_{N,X}$ using Stirling's asymptotic expansion

$$\log N! \simeq N \log N - N + \mathcal{O}(\log N) \quad . \tag{1.11}$$

We then have

$$\begin{aligned} \log P_{N,X} &\simeq N \log N - N - \tfrac{1}{2}N(1+x)\log\Big[\tfrac{1}{2}N(1+x)\Big] + \tfrac{1}{2}N(1+x) \\ &\quad - \tfrac{1}{2}N(1-x)\log\Big[\tfrac{1}{2}N(1-x)\Big] + \tfrac{1}{2}N(1-x) + \tfrac{1}{2}N(1+x)\,\log p + \tfrac{1}{2}N(1-x)\,\log q \\ &= -N\Big[\big(\tfrac{1+x}{2}\big)\log\big(\tfrac{1+x}{2}\big) + \big(\tfrac{1-x}{2}\big)\log\big(\tfrac{1-x}{2}\big)\Big] + N\Big[\big(\tfrac{1+x}{2}\big)\log p + \big(\tfrac{1-x}{2}\big)\log q\Big] \quad . \end{aligned} \tag{1.12}$$

Notice that the terms proportional to $N \log N$ have all cancelled, leaving us with a quantity which is linear in N. We may therefore write $\log P_{N,X} = -N f(x) + \mathcal{O}(\log N)$, where

$$f(x) = \Big[\big(\tfrac{1+x}{2}\big)\log\big(\tfrac{1+x}{2}\big) + \big(\tfrac{1-x}{2}\big)\log\big(\tfrac{1-x}{2}\big)\Big] - \Big[\big(\tfrac{1+x}{2}\big)\log p + \big(\tfrac{1-x}{2}\big)\log q\Big] \quad . \tag{1.13}$$

We have just shown that in the large N limit we may write

$$P_{N,X} \approx \mathcal{C}\, e^{-N f(X/N)} \quad , \tag{1.14}$$

where $\mathcal{C}$ is a normalization factor[1]. Since N is by assumption large, the function $P_{N,X}$ is dominated by the minimum (or minima) of $f(x)$, where the probability is maximized. To find the minimum of $f(x)$, we set $f'(x) = 0$, where

$$f'(x) = \tfrac{1}{2}\log\left(\frac{q}{p}\cdot\frac{1+x}{1-x}\right) \quad . \tag{1.15}$$

Setting $f'(x) = 0$, we obtain

$$\frac{1+x}{1-x} = \frac{p}{q} \quad \Rightarrow \quad \bar{x} = p - q \quad . \tag{1.16}$$

We also have

$$f''(x) = \frac{1}{1-x^2} \quad , \tag{1.17}$$

so invoking Taylor's theorem,

$$f(x) = f(\bar{x}) + \tfrac{1}{2}f''(\bar{x})\,(x-\bar{x})^2 + \ldots \quad . \tag{1.18}$$

Putting it all together, we have

$$P_{N,X} \approx \mathcal{C}\,\exp\left[-\frac{N(x-\bar{x})^2}{8pq}\right] = \mathcal{C}\,\exp\left[-\frac{(X-\bar{X})^2}{8Npq}\right] \quad , \tag{1.19}$$

[1]The origin of $\mathcal{C}$ lies in the $\mathcal{O}(\log N)$ and $\mathcal{O}(N^0)$ terms in the asymptotic expansion of $\log N!$ – these terms will in general also depend on $x = X/N$. We have ignored these terms here. Accounting for them carefully reproduces the correct value of $\mathcal{C}$ in eqn. 1.20.

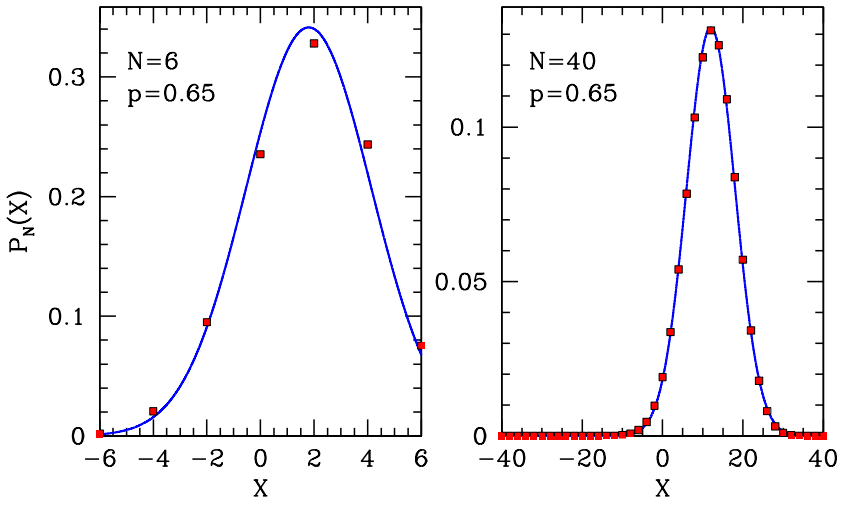


Figure 1.2: Comparison of exact distribution of eqn. 1.10 (red squares) with the Gaussian distribution of eqn. 1.19 (blue line).

where $\bar{X} = \langle X \rangle = N(p-q) = N\bar{x}$. The constant $\mathcal{C}$ is determined by the normalization condition,

$$\sum_{X=-\infty}^{\infty} P_{N,X} \approx \tfrac{1}{2}\mathcal{C} \int_{-\infty}^{\infty} dX \ \exp\left[-\frac{(X-\bar{X})^2}{8Npq} \right] = \sqrt{2\pi Npq}\,\mathcal{C} \quad , \tag{1.20}$$

and thus $\mathcal{C} = 1/\sqrt{2\pi Npq}$. Why don't we go beyond second order in the Taylor expansion of $f(x)$? We will find out in §1.5.2 below.

1.2.3 Entropy and energy

The function $f(x)$ can be written as a sum of two contributions, $f(x) = e(x) - s(x)$, where

$$\begin{aligned} s(x) &= -\left(\tfrac{1+x}{2}\right)\log\left(\tfrac{1+x}{2}\right) - \left(\tfrac{1-x}{2}\right)\log\left(\tfrac{1-x}{2}\right) \\ e(x) &= -\tfrac{1}{2}(1+x)\log p - \tfrac{1}{2}(1-x)\log q \quad . \end{aligned} \tag{1.21}$$

The function $S(N,x) \equiv Ns(x)$ is analogous to the *statistical entropy* of our system, and $E(N,x) \equiv Ne(x)$ to the *energy* of the system[2]. The statistical entropy is the logarithm of the number of ways, at fixed N, that the system can be configured so as to yield the same value of X. The energy biases the probability $P_{N,X} = \exp(S-E)$ so that low energy configurations are more probable than high energy configurations. For our system, we see that when $p < \frac{1}{2}$ the energy is minimized by taking x as small as possible, *i.e.* $x = -1$. Conversely, when $p > \frac{1}{2}$ the energy is minimized by taking x as large as possible, *i.e.* $x = +1$. The average value of x, as we have computed explicitly, is $\bar{x} = p - q = 2p - 1$, which falls somewhere in between these two extremes.

[2]The functions $s(x)$ and $e(x)$ are the *specific entropy* and *specific energy*, respectively.

In actual thermodynamic systems, entropy and energy are not dimensionless. What we have called S here is really $S/k_{\rm B}$, which is the entropy in units of Boltzmann's constant. And what we have called E here is really $E/k_{\rm B}T$, which is energy in units of Boltzmann's constant times temperature.

1.3 Basic Concepts in Probability Theory

Here we recite the basics of probability theory.

1.3.1 Fundamental definitions

The natural mathematical setting is set theory. *Sets* are generalized collections of *objects*. The basics: $\omega \in A$ is a binary relation which says that the object ω is an *element* of the set A. Another binary relation is *set inclusion*. If all members of A are in B, we write $A \subseteq B$. The *union* of sets A and B is denoted $A \cup B$ and the *intersection* of A and B is denoted $A \cap B$. The *Cartesian product* of A and B, denoted $A \times B$, is the set of all ordered pairs (a, b) where $a \in A$ and $b \in B$.

Some details: If ω is not in A, we write $\omega \notin A$. Sets may also be objects, so we may speak of sets of sets, but typically the sets which will concern us are simple discrete collections of numbers, such as the possible rolls of a die {1,2,3,4,5,6}, or the real numbers $\mathbb{R}$, or Cartesian products such as $\mathbb{R}^N$. If $A \subseteq B$ but $A \neq B$, we say that A is a *proper subset* of B and write $A \subset B$. Another binary operation is the *set difference* $A \backslash B$, which contains all ω such that $\omega \in A$ and $\omega \notin B$.

In probability theory, each object ω is identified as an *event*. We denote by Ω the set of all events, and $\emptyset$ denotes the set of no events. There are three basic axioms of probability:

i) To each set A is associated a non-negative real number $P(A)$, which is called the probability of A.

ii) $P(\Omega) = 1$.

iii) If $\{A_i\}$ is a countable collection of disjoint sets, *i.e.* if $A_i \cap A_j = \emptyset$ for all $i \neq j$, then

$$P\Big(\bigcup_i A_i\Big) = \sum_i P(A_i) \quad . \tag{1.22}$$

From these axioms follow a number of conclusions. Among them, let $\neg A = \Omega \backslash A$ be the *complement* of A, *i.e.* the set of all events *not* in A. Then since $A \cup \neg A = \Omega$, we have $P(\neg A) = 1 - P(A)$. Taking $A = \Omega$, we conclude $P(\emptyset) = 0$.

The operational definition of $P(A)$ is that if events ω are chosen from Ω *at random*[3], then the relative frequency for $\omega \in A$ approaches $P(A)$ as the number of trials tends to infinity. But what do we mean by 'at random'? One meaning we can impart to the notion of randomness is that a process is random if its outcomes can be accurately modeled using the axioms of probability. This entails the identification

[3]We assume independent and identically distributed trials, and that the law of large numbers holds.

of a *probability space* Ω as well as a *probability measure* P. For example, in the microcanonical ensemble of classical statistical physics, the space Ω is the collection of phase space points $\boldsymbol{\varphi} = \{q_1, \ldots, q_n, p_1, \ldots, p_n\}$ and the probability measure is $d\mu = \Sigma^{-1}(E) \prod_{i=1}^n dq_i\, dp_i\, \delta\big(E - H(q,p)\big)$, so that for $A \subseteq \Omega$ the probability of A is $P(A) = \int d\mu\, \chi_{_A}(\boldsymbol{\varphi})$, where $\chi_{_A}(\boldsymbol{\varphi}) = 1$ if $\boldsymbol{\varphi} \in A$ and $\chi_{_A}(\boldsymbol{\varphi}) = 0$ if $\boldsymbol{\varphi} \notin A$ is the *characteristic function* of A. The quantity $\Sigma(E)$ is determined by normalization: $\int d\mu = 1$.

1.3.2 Bayesian statistics

We now introduce two additional probabilities. The *joint probability* for sets A and B together is written $P(A \cap B)$. That is, $P(A \cap B) = \mathsf{Prob}[\omega \in A \text{ and } \omega \in B]$. For example, A might denote the set of all politicians, B the set of all American citizens, and C the set of all living humans who are allergic to cats. Then $A \cap B$ would be the set of all politicians who are also American citizens, *etc.*

The *conditional probability* of B *given* A is written $P(B \,|\, A)$. The joint probability $P(A \cap B) = P(B \cap A)$ may be expressed in two ways:

$$P(A \cap B) = P(A \,|\, B) \cdot P(B) = P(B \,|\, A) \cdot P(A) \quad . \tag{1.23}$$

Thus,

$$P(A \,|\, B) = \frac{P(B \,|\, A)\, P(A)}{P(B)} \quad , \tag{1.24}$$

a result known as *Bayes' theorem*. Now suppose the 'event space' is partitioned as $\{A_i\}$. Then

$$P(B) = \sum_i P(B|A_i)\, P(A_i) \quad . \tag{1.25}$$

We then have

$$P(A_i \,|\, B) = \frac{P(B \,|\, A_i)\, P(A_i)}{\sum_j P(B \,|\, A_j)\, P(A_j)} \quad , \tag{1.26}$$

a result sometimes known as the *extended form of Bayes' theorem*. When the event space is a 'binary partition' $\{A, \neg A\}$, we have

$$P(A \,|\, B) = \frac{P(B \,|\, A)\, P(A)}{P(B \,|\, A)\, P(A) + P(B \,|\, \neg A)\, P(\neg A)} \quad . \tag{1.27}$$

Note that $P(A \,|\, B) + P(\neg A \,|\, B) = 1$ (which follows from $\neg\neg A = A$).

As an example, consider the following problem in epidemiology. Suppose there is a rare but highly contagious disease A which occurs in 0.01% of the general population. Suppose further that there is a simple test for the disease which is accurate 99.99% of the time. That is, out of every 10,000 tests, the correct answer is returned 9,999 times, and the incorrect answer is returned only once. Now let us administer the test to a large group of people from the general population. Those who test positive are quarantined. Question: what is the probability that someone chosen at random from the quarantine group actually has the disease? We use Bayes' theorem with the binary partition $\{A, \neg A\}$. Let B denote the event that an individual tests positive. Anyone from the quarantine group has tested positive. Given

this datum, we want to know the probability that that person has the disease. That is, we want $P(A \mid B)$. Applying eqn. 1.27 with

$$P(A) = 0.0001 \quad , \quad P(\neg A) = 0.9999 \quad , \quad P(B \mid A) = 0.9999 \quad , \quad P(B \mid \neg A) = 0.0001 \quad ,$$

we find $P(A \mid B) = \frac{1}{2}$. That is, there is only a 50% chance that someone who tested positive actually has the disease, despite the test being 99.99% accurate! The reason is that, given the rarity of the disease in the general population, the number of false positives is statistically equal to the number of true positives.

In the above example, we had $P(B \mid A) + P(B \mid \neg A) = 1$, but this is not generally the case. What is true instead is $P(B \mid A) + P(\neg B \mid A) = 1$. Epidemiologists define the *sensitivity* of a binary classification test as the fraction of actual positives which are correctly identified, and the *specificity* as the fraction of actual negatives that are correctly identified. Thus, $\mathsf{se} = P(B \mid A)$ is the sensitivity and $\mathsf{sp} = P(\neg B \mid \neg A)$ is the specificity. We then have $P(B \mid \neg A) = 1 - P(\neg B \mid \neg A)$. Therefore,

$$P(B \mid A) + P(B \mid \neg A) = 1 + P(B \mid A) - P(\neg B \mid \neg A) = 1 + \mathsf{se} - \mathsf{sp} \quad . \tag{1.28}$$

In our previous example, $\mathsf{se} = \mathsf{sp} = 0.9999$, in which case the RHS above gives 1. In general, if $P(A) \equiv f$ is the fraction of the population which is afflicted, then

$$P(\mathsf{infected} \mid \mathsf{positive}) = \frac{f \cdot \mathsf{se}}{f \cdot \mathsf{se} + (1-f) \cdot (1 - \mathsf{sp})} \quad . \tag{1.29}$$

For continuous distributions, we speak of a probability *density*. We then have

$$P(y) = \int dx \, P(y \mid x) \, P(x) \tag{1.30}$$

and

$$P(x \mid y) = \frac{P(y \mid x) \, P(x)}{\int dx' \, P(y \mid x') \, P(x')} \quad . \tag{1.31}$$

The range of integration may depend on the specific application.

The quantities $P(A_i)$ are called the *prior distribution*. Clearly in order to compute $P(B)$ or $P(A_i \mid B)$ we must know the priors, and this is usually the weakest link in the Bayesian chain of reasoning. If our prior distribution is not accurate, Bayes' theorem will generate incorrect results. One approach to approximating prior probabilities $P(A_i)$ is to derive them from a *maximum entropy construction*.

1.3.3 Random variables and their averages

Consider an abstract probability space $\mathcal{X}$ whose elements (*i.e.* events) are labeled by x. A *random variable* X can take values $x \in \mathcal{X}$.[4] The average of a function $f(X)$ of the random variable X is denoted as $\mathbb{E}f(X)$ or $\langle f(X) \rangle$, and is defined for discrete sets as

$$\mathbb{E}f(X) = \langle f(X) \rangle = \sum_{x \in \mathcal{X}} f(x) \, P(x) \quad , \tag{1.32}$$

[4] Formally, a random variable X is a measurable function $X : \Omega \to \mathcal{X}$ from the underlying sample space Ω to a value space $\mathcal{X}$

where $P(x) = \mathsf{Prob}(X = x)$. For continuous sets, we have

$$\mathbb{E}f(X) = \langle f(X)\rangle = \int\limits_{\mathcal{X}} dx\; f(x)\, P(x) \quad , \tag{1.33}$$

where now $P(x)$ is the *probability density* at x. We then have that

$$P(x)\, dx = \mathsf{Prob}\big(X \in [x, x + dx]\big) \quad . \tag{1.34}$$

Typically for continuous sets we have $\mathcal{X} = \mathbb{R}$ or $\mathcal{X} = \mathbb{R}_{\geq 0}$. While it is formally useful to use symbols such as X and Y for random variables, at times we may slip and loosely speak of x and y as random variables.

When there are two random variables $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$, we have $\Omega = \mathcal{X} \times \mathcal{Y}$ is the product space, and

$$\mathbb{E}f(X,Y) = \langle f(X,Y)\rangle = \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} f(x,y)\, P(x,y) \quad , \tag{1.35}$$

with the obvious generalization to continuous sets. This generalizes to higher rank product probability spaces, *i.e.* $\Omega = \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_N$ with $x_i \in \mathcal{X}_i$ for $i \in \{1, \ldots, N\}$. The *covariance* of X_i and X_j is defined as

$$C_{ij} \equiv \big\langle \big(X_i - \langle X_i\rangle\big)\big(X_j - \langle X_j\rangle\big)\big\rangle = \langle X_i X_j\rangle - \langle X_i\rangle\langle X_j\rangle \quad . \tag{1.36}$$

If $f(x)$ is a convex function then one has

$$\mathbb{E}f(X) \geq f(\mathbb{E}X) \quad . \tag{1.37}$$

If a function $f(x)$ is twice differentiable on the interval $[a, b]$, it is convex if and only if $f''(x) \geq 0$ everywhere[5]. In such cases, then for $x_{1,2} \in [a, b]$ we must have

$$f\big(\lambda x_1 + (1-\lambda)x_2\big) \leq \lambda f(x_1) + (1-\lambda) f(x_2) \quad , \tag{1.38}$$

where $\lambda \in [0, 1]$. This is easily generalized to

$$f\Big(\sum_n p_n\, x_n\Big) \leq \sum_n p_n\, f(x_n) \quad , \tag{1.39}$$

where $p_n = P(x_n)$, a result known as *Jensen's theorem*. For continuous random variables,

$$f\left(\int\limits_{\mathcal{X}} dx\, P(x)\, x\right) \leq \int\limits_{\mathcal{X}} dx\, P(x)\, f(x) \quad , \tag{1.40}$$

[5] A function $g(x)$ is *concave* if $-g(x)$ is convex.

1.4 Entropy and Probability

1.4.1 Entropy and information theory

It was shown in the classic 1948 work of Claude Shannon that entropy is in fact a measure of *information*[6]. Suppose we observe that a particular event occurs with probability p. We associate with this observation an amount of information $I(p)$. The information $I(p)$ should satisfy certain desiderata:

1 Information is non-negative, *i.e.* $I(p) \geq 0$.

2 If two events occur independently so their joint probability is $p_1 p_2$, then their information is additive, *i.e.* $I(p_1 p_2) = I(p_1) + I(p_2)$.

3 $I(p)$ is a continuous function of p.

4 There is no information content to an event which is always observed, *i.e.* $I(1) = 0$.

From these four properties, it is easy to show that the only possible function $I(p)$ is

$$I(p) = -A \log p \quad , \tag{1.41}$$

where A is an arbitrary constant that can be absorbed into the base of the logarithm, since $\log_b x = \log x / \log b$. In statistical physics we will take $A = 1$ and use e as the base, so $I(p) = -\log p$. These are so-called *natural units* of information. Another common choice, typical in computer science, is to take the base of the logarithm to be 2, so $I(p) = -\log_2 p$. In this latter case, the units of information are known as *bits*. Note that, in either case, $I(0) = \infty$, which means that the observation of an extremely rare event carries a great deal of information[7].

Now suppose we have a set of events labeled by an integer n which occur with probabilities $\{p_n\}$. What is the expected amount of information in N observations? Since event n occurs an average of Np_n times, and the information content in p_n is $-\log p_n$, we have that the average information per observation is

$$S = \frac{\langle I_N \rangle}{N} = -\sum_n p_n \log p_n \quad , \tag{1.42}$$

which is known as the entropy of the distribution. Thus, maximizing S is equivalent to maximizing the *information* content per observation. S is known as the *Shannon entropy* of the distribution $\{p_n\}$. The base of the logarithm determines the units: bits for base 2, nats for base e, and hartleys or decimal digits for base 10.

Consider, for example, the information content of course grades. As we shall see, if the only constraint on the probability distribution is that of overall normalization, then S is maximized when all the probabilities p_n are equal. The Shannon entropy is then $S = \log_2 \Gamma$, where Γ is the total size of our discrete

[6] See 'An Introduction to Information Theory and Entropy' by T. Carter, Santa Fe Complex Systems Summer School, June 2011: http://astarte.csustan.edu/$\sim$tom/SFI-CSSS/info-theory/info-lec.pdf.

[7] My colleague John McGreevy refers to $I(p)$ as the *surprise* of observing an event occurring with probability p. I like this very much.

space of states, since $p_n = 1/\Gamma$. Thus, for pass/fail grading, the maximum average information per grade is $-\log_2(\frac{1}{2}) = \log_2 2 = 1$ bit. If only A, B, C, D, and F grades are assigned, then the maximum average information per grade is $\log_2 5 = 2.32$ bits. If we expand the grade options to include {A+, A, A-, B+, B, B-, C+, C, C-, D, F}, then the maximum average information per grade is $\log_2 11 = 3.46$ bits.

Equivalently, consider, following the discussion in vol. 1 of Kardar, a random sequence $\{n_1, n_2, \ldots, n_N\}$ where each element n_j takes one of K possible values. There are then K^N such possible sequences, and to specify one of them requires $\log_2(K^N) = N \log_2 K$ bits of information. However, if the value n occurs with probability p_n, then on average it will occur $N_n = Np_n$ times in a sequence of length N, and the total number of such sequences will be

$$g(N) = \frac{N!}{\prod_{n=1}^{K} N_n!} \quad . \tag{1.43}$$

In general, this is far less than the total possible number K^N, and the number of bits necessary to specify one from among these $g(N)$ possibilities is

$$\log_2 g(N) = \log_2(N!) - \sum_{n=1}^{K} \log_2(N_n!) \approx -N \sum_{n=1}^{K} p_n \log_2 p_n \quad , \tag{1.44}$$

up to terms of order $\log N$. Here we have invoked Stirling's approximation. If the distribution is uniform, then we have $p_n = K^{-1}$ for all $n \in \{1, \ldots, K\}$, and $\log_2 g(N) = N \log_2 K$.

A quantum mechanical *density matrix* $\hat{\varrho} = \sum_{j,k} \varrho_{jk} \, |j\rangle\langle k|$ is a Hermitian operator normalized such that $\mathsf{Tr}\ \hat{\varrho} = 1$ and subject to the condition that each of its eigenvalues p_n lies on the interval $[0, 1]$. Thus we have $\mathsf{Tr}\ \hat{\varrho} = \sum_n p_n = 1$. The $q^{\rm th}$ power of $\hat{\varrho}$ is then

$$\hat{\varrho}^q = \sum_{j,k} (\varrho^q)_{jk} \, |\, j \,\rangle\langle\, k \,| \quad , \tag{1.45}$$

where $(\varrho^q)_{jk}$ is the j, k matrix element of the matrix ϱ^q. We can write the entropy of the system as

$$S = -\,\mathsf{Tr}\,\big(\varrho \log \varrho\big) = -\sum_n p_n \log p_n \quad . \tag{1.46}$$

We define $\Omega \equiv \sum_n 1$ to be the total number of events. It may be the case that $\Omega = \infty$.

1.4.2 Rényi entropy

The Rényi entropy $S_q\big(\{p_n\}\big)$ is a one-parameter family of entropy functions parameterized by a real number $q \in [0, \infty)$, with

$$S_q = \frac{\log \mathsf{Tr}\,(\varrho^q)}{1-q} = \frac{1}{1-q} \log\Big(\sum_n (p_n)^q \Big) \quad . \tag{1.47}$$

With $q = 1 + \varepsilon$, we have $p_n^{1+\varepsilon} = p_n \cdot \exp(\varepsilon \log p_n) = p_n + \varepsilon\, p_n \log p_n + \mathcal{O}(\varepsilon^2)$ and thus in the limit $q \to 1$ we recover the Shannon entropy (up to a multiplicative constant depending on the base of the logarithm),

$$\lim_{q \to 1} S_q = -\sum_n p_n \log p_n \quad , \tag{1.48}$$

Note that the Rényi entropy is additive for all q. Thus, if ϱ^{A} and ϱ^{B} are density matrices for two independent systems, then $\varrho^{\mathsf{AB}} = \varrho^{\mathsf{A}} \otimes \varrho^{\mathsf{B}}$ and

$$S_q^{\mathsf{AB}} = S_q^{\mathsf{A}} + S_q^{\mathsf{B}} \quad . \tag{1.49}$$

Again, the Rényi entropy S_q is defined for all $q \in \mathbb{R}_{\geq 0}$. Some special cases:

- In the limit $q \to 0^+$ we have $S_{0^+} = \log \Omega'$, where Ω' is the number of events for which the probability of occurrence is nonzero, *i.e.* $\Omega' = \sum_n \Theta(p_n)$ where $\Theta(x)$ is the step function.
- For $q \in \mathbb{Z}_+$ a positive integer, we can think of the Rényi entropy in terms of *replicas*. Imagine we take q identical systems, each governed by the same probability distribution. Then the probability of observing the same event among each of the q replicas is $\sum_n (p_n)^q = e^{(1-q)S_q}$.
- As $q \to \infty$, the sum $\sum_n (p_n)^q$ is dominated by the maximum value of p_n. Let $p^* = \max_n\{p_n\}$. Then $S_\infty = -\log p^*$. One can check that the result is independent of the degeneracy g^* of the maximum probability[8].

The following inequality holds in general for $q > 1$:

$$S_\infty \leq S_q \leq \left(\frac{q}{q-1}\right) S_\infty \quad . \tag{1.50}$$

1.4.3 Kullback-Leibler divergence

Suppose we approximate a probability distribution $P(x)$ by another function $Q(x)$. The expected difference in the logarithms of the probabilities is then given by

$$\begin{aligned} D_{\mathsf{KL}}(P \parallel Q) &= -\sum_{x \in \mathcal{X}} P(x) \Big\{ \log Q(x) - \log P(x) \Big\} \\ &= \sum_{x \in \mathcal{X}} P(x) \log\left(\frac{P(x)}{Q(x)}\right) = \mathbb{E}_P \log\left(\frac{P(X)}{Q(X)}\right) \quad , \end{aligned} \tag{1.51}$$

called the *Kullback-Leibler divergence* or *relative entropy*. If $P(x) > 0$ and $Q(x) = 0$ for any event x then $D_{\mathsf{KL}}(P \parallel Q) = \infty$.

1.4.4 Probability distributions from maximum entropy

We have shown how one can proceed from a probability distribution and compute various averages. We now seek to go in the other direction, and determine the full probability distribution based on a knowledge of certain averages.

[8] Provided that g^* is finite, or more generally, that $\log(g^*)/q \to 0$ as $q \to \infty$.

At first, this seems impossible. Suppose we want to reproduce the full probability distribution for an N-step random walk from knowledge of the average $\langle X\rangle = (2p-1)N$, where p is the probability of moving to the right at each step (see §1.2 above). The problem seems ridiculously underdetermined, since there are 2^N possible configurations for an N-step random walk: $\sigma_j = \pm 1$ for $j = 1, \ldots, N$. Overall normalization requires

$$\sum_{\{\sigma_j\}} P(\sigma_1, \ldots, \sigma_N) = 1 \quad , \tag{1.52}$$

but this just imposes one constraint on the 2^N probabilities $P(\sigma_1, \ldots, \sigma_N)$, leaving $2^N - 1$ overall parameters. What principle allows us to reconstruct the full probability distribution

$$P(\sigma_1, \ldots, \sigma_N) = \prod_{j=1}^{N} \left(p\,\delta_{\sigma_j,1} + q\,\delta_{\sigma_j,-1} \right) = \prod_{j=1}^{N} p^{(1+\sigma_j)/2}\, q^{(1-\sigma_j)/2} \quad , \tag{1.53}$$

corresponding to N independent steps?

The principle of maximum entropy

The entropy of a discrete probability distribution $\{p_n\}$ is defined as

$$S = -\sum_n p_n \log p_n \quad , \tag{1.54}$$

where here we take e as the base of the logarithm. The entropy may therefore be regarded as a function of the probability distribution: $S = S\big(\{p_n\}\big)$. One special property of the entropy is the following. Suppose we have two independent normalized distributions $\{p_a^{\mathsf{A}}\}$ and $\{p_b^{\mathsf{B}}\}$. The joint probability for events a and b is then $P_{a,b} = p_a^{\mathsf{A}}\, p_b^{\mathsf{B}}$. The entropy of the joint distribution is then

$$\begin{aligned} S &= -\sum_a \sum_b P_{a,b} \log P_{a,b} = -\sum_a \sum_b p_a^{\mathsf{A}}\, p_b^{\mathsf{B}} \log\left(p_a^{\mathsf{A}}\, p_b^{\mathsf{B}}\right) = -\sum_a \sum_b p_a^{\mathsf{A}}\, p_b^{\mathsf{B}} \left(\log p_a^{\mathsf{A}} + \log p_b^{\mathsf{B}}\right) \\ &= -\sum_a p_a^{\mathsf{A}} \log p_a^{\mathsf{A}} \cdot \sum_b p_b^{\mathsf{B}} - \sum_b p_b^{\mathsf{B}} \log p_b^{\mathsf{B}} \cdot \sum_a p_a^{\mathsf{A}} = -\sum_a p_a^{\mathsf{A}} \log p_a^{\mathsf{A}} - \sum_b p_b^{\mathsf{B}} \log p_b^{\mathsf{B}} = S^{\mathsf{A}} + S^{\mathsf{B}} \quad . \end{aligned} \tag{1.55}$$

Thus, the entropy of a joint distribution formed from two independent distributions is additive.

Suppose all we knew about $\{p_n\}$ was that it was normalized. Then $\sum_n p_n = 1$. This is a constraint on the values $\{p_n\}$. Let us now extremize the entropy S with respect to the distribution $\{p_n\}$, but subject to the normalization constraint. We do this using Lagrange's method of undetermined multipliers. We define

$$S^*\big(\{p_n\}, \lambda\big) = -\sum_n p_n \log p_n - \lambda\Big(\sum_n p_n - 1\Big) \tag{1.56}$$

and we freely extremize S^* over all its arguments. Thus, for all n we have

$$\begin{aligned} 0 &= \frac{\partial S^*}{\partial p_n} = -\big(\log p_n + 1 + \lambda\big) \\ 0 &= \frac{\partial S^*}{\partial \lambda} = \sum_n p_n - 1 \quad . \end{aligned} \tag{1.57}$$

From the first of these equations, we obtain $p_n = e^{-(1+\lambda)}$, and from the second we obtain

$$\sum_n p_n = e^{-(1+\lambda)} \cdot \sum_n 1 = \Gamma\, e^{-(1+\lambda)} \quad , \tag{1.58}$$

where $\Gamma \equiv \sum_n 1$ is the total number of possible events. Thus, $p_n = \Gamma^{-1}$, which says that all events are equally probable.

Now suppose we know one other piece of information, which is the average value $X = \sum_n X_n\, p_n$ of some quantity. We now extremize S subject to two constraints, and so we define

$$S^*\big(\{p_n\}, \lambda_0, \lambda_1\big) = -\sum_n p_n \log p_n - \lambda_0\Big(\sum_n p_n - 1\Big) - \lambda_1\Big(\sum_n X_n\, p_n - X\Big) \quad . \tag{1.59}$$

We then have

$$\frac{\partial S^*}{\partial p_n} = -\big(\log p_n + 1 + \lambda_0 + \lambda_1\, X_n\big) = 0 \quad , \tag{1.60}$$

which yields the two-parameter distribution

$$p_n = e^{-(1+\lambda_0)}\, e^{-\lambda_1 X_n} \quad . \tag{1.61}$$

To fully determine the distribution $\{p_n\}$, we invoke the two equations $\sum_n p_n = 1$ and $\sum_n X_n\, p_n = X$, which come from extremizing S^* with respect to λ_0 and λ_1, respectively:

$$\begin{aligned} 1 &= e^{-(1+\lambda_0)} \sum_n e^{-\lambda_1 X_n} \\ X &= e^{-(1+\lambda_0)} \sum_n X_n\, e^{-\lambda_1 X_n} \quad . \end{aligned} \tag{1.62}$$

General formulation

The generalization to $K+1$ constraints is immediately apparent. We have

$$X^a = \sum_n X_n^a\, p_n \quad , \tag{1.63}$$

and therefore we define

$$S^*\big(\{p_n\}, \{\lambda_a\}\big) = -\sum_n p_n \log p_n - \sum_{a=0}^{K} \lambda_a\Big(\sum_n X_n^a\, p_n - X^a\Big) \quad , \tag{1.64}$$

with $X_n^{(a=0)} \equiv X^{(a=0)} = 1$. Then the optimal distribution which extremizes S subject to the $K+1$ constraints is

$$p_n = \exp\left\{-1 - \sum_{a=0}^{K} \lambda_a\, X_n^a\right\} = \frac{1}{Z} \exp\left\{-\sum_{a=1}^{K} \lambda_a\, X_n^a\right\} \quad , \tag{1.65}$$

where $Z = e^{1+\lambda_0}$ is determined by normalization: $\sum_n p_n = 1$. This is a $(K+1)$-parameter distribution, with $\{\lambda_0, \lambda_1, \ldots, \lambda_K\}$ determined by the $K+1$ constraints in eqn. 1.63.

Example

As an example, consider the random walk problem. We have two pieces of information:

$$\begin{aligned}\sum_{\sigma_1}\cdots\sum_{\sigma_N} P(\sigma_1,\ldots,\sigma_N) &= 1 \\ \sum_{\sigma_1}\cdots\sum_{\sigma_N} P(\sigma_1,\ldots,\sigma_N)\sum_{j=1}^N \sigma_j &= X \quad .\end{aligned} \tag{1.66}$$

Here the discrete label n from §1.4.4 ranges over 2^N possible values, and may be written as an N digit binary number $r_N\cdots r_1$, where $r_j = \frac{1}{2}(1+\sigma_j)$ is 0 or 1. Extremizing S subject to these constraints, we obtain

$$P(\sigma_1,\ldots,\sigma_N) = \mathcal{C}\,\exp\left\{-\lambda\sum_j \sigma_j\right\} = \mathcal{C}\prod_{j=1}^N e^{-\lambda\,\sigma_j} \quad , \tag{1.67}$$

where $\mathcal{C} \equiv e^{-(1+\lambda_0)}$ and $\lambda \equiv \lambda_1$. Normalization then requires

$$\mathsf{Tr}\;P \equiv \sum_{\{\sigma_j\}} P(\sigma_1,\ldots,\sigma_N) = \mathcal{C}\left(e^{\lambda}+e^{-\lambda}\right)^N \quad , \tag{1.68}$$

hence $\mathcal{C} = (2\cosh\lambda)^{-N}$. We then have

$$P(\sigma_1,\ldots,\sigma_N) = \prod_{j=1}^N \frac{e^{-\lambda\sigma_j}}{e^{\lambda}+e^{-\lambda}} = \prod_{j=1}^N \left(p\,\delta_{\sigma_j,1} + q\,\delta_{\sigma_j,-1}\right) \quad , \tag{1.69}$$

where

$$p = \frac{e^{-\lambda}}{e^{\lambda}+e^{-\lambda}} \quad , \quad q = 1-p = \frac{e^{\lambda}}{e^{\lambda}+e^{-\lambda}} \quad . \tag{1.70}$$

We then have $X = (2p-1)N$, which determines $p = \frac{1}{2}\big(1+\frac{X}{N}\big)$, and we have recovered the Bernoulli distribution.

Of course there are no miracles[9], and there are an infinite family of distributions for which $X = (2p-1)N$ that are not Bernoulli. For example, we could have imposed another constraint, such as $E = \sum_{j=1}^{N-1}\sigma_j\,\sigma_{j+1}$. This would result in the distribution

$$P(\sigma_1,\ldots,\sigma_N) = \frac{1}{Z}\exp\left\{-\lambda_1\sum_{j=1}^N \sigma_j - \lambda_2\sum_{j=1}^{N-1}\sigma_j\,\sigma_{j+1}\right\} \quad , \tag{1.71}$$

with $Z(\lambda_1,\lambda_2)$ determined by normalization: $\sum_{\boldsymbol{\sigma}} P(\boldsymbol{\sigma}) = 1$. This is the one-dimensional Ising chain of classical equilibrium statistical physics. Defining the transfer matrix $R_{ss'} = e^{-\lambda_1(s+s')/2}\,e^{-\lambda_2 ss'}$ with $s, s' = \pm 1$,

$$\begin{aligned}R &= \begin{pmatrix} e^{-\lambda_1-\lambda_2} & e^{\lambda_2} \\ e^{\lambda_2} & e^{\lambda_1-\lambda_2}\end{pmatrix} \\ &= e^{-\lambda_2}\cosh(\lambda_1)\,\mathbb{I} - e^{-\lambda_2}\sinh(\lambda_1)\,\tau^z + e^{\lambda_2}\,\tau^x \quad ,\end{aligned} \tag{1.72}$$

[9] See §10 of *An Enquiry Concerning Human Understanding* by David Hume (1748).

where τ^x and τ^z are Pauli matrices, we have that

$$Z_{\rm ring} = {\rm Tr}\left(R^N\right) \quad , \quad Z_{\rm chain} = {\rm Tr}\left(R^{N-1}S\right) \quad , \tag{1.73}$$

where $S_{ss'} = \exp\left(-\frac{1}{2}\lambda_1(s+s')\right)$, *i.e.*

$$S = \begin{pmatrix} e^{-\lambda_1} & 1 \\ 1 & e^{\lambda_1} \end{pmatrix} = \cosh(\lambda_1)\,\mathbb{I} - \sinh(\lambda_1)\,\tau^z + \tau^x \quad . \tag{1.74}$$

The appropriate case here is that of the chain, but in the thermodynamic limit $N \to \infty$ both chain and ring yield identical results, so we will examine here the results for the ring, which are somewhat easier to obtain. Clearly $Z_{\rm ring} = \zeta_+^N + \zeta_-^N$, where $\zeta_\pm$ are the eigenvalues of R:

$$\zeta_\pm = e^{-\lambda_2}\cosh\lambda_1 \pm \sqrt{e^{-2\lambda_2}\sinh^2\!\lambda_1 + e^{2\lambda_2}} \quad . \tag{1.75}$$

In the thermodynamic limit, the ζ_+ eigenvalue dominates, and $Z_{\rm ring} \simeq \zeta_+^N$. We now have

$$X = \Big\langle \sum_{j=1}^N \sigma_j \Big\rangle = -\frac{\partial \log Z}{\partial \lambda_1} = -\frac{N\sinh\lambda_1}{\sqrt{\sinh^2\!\lambda_1 + e^{4\lambda_2}}} \quad . \tag{1.76}$$

We also have $E = -\partial \log Z/\partial\lambda_2$. These two equations determine the Lagrange multipliers $\lambda_1(X,E,N)$ and $\lambda_2(X,E,N)$. In the thermodynamic limit, we have $\lambda_i = \lambda_i(X/N, E/N)$. Thus, if we fix $X/N = 2p-1$ alone, there is a continuous one-parameter family of distributions, parametrized by $\varepsilon = E/N$, which satisfy the constraint on X.

So what is it about the maximum entropy approach that is so compelling? Maximum entropy gives us a calculable distribution which is consistent with "maximum ignorance" given our known constraints[10]. In that sense, it is as unbiased as possible, from an information theoretic point of view. As a starting point, a maximum entropy distribution may be improved upon, using Bayesian methods for example (see §1.6.2 below).

1.4.5 Continuous probability distributions

Suppose we have a continuous probability density $P(\boldsymbol{\varphi})$ defined over some set Ω. We have observables

$$X^a = \int\limits_\Omega d\mu\, X^a(\boldsymbol{\varphi})\, P(\boldsymbol{\varphi}) \quad , \tag{1.77}$$

where $d\mu$ is the appropriate integration measure. We assume $d\mu = \prod_{j=1}^D d\varphi_j$, where D is the dimension of Ω. Then we extremize the functional

$$S^*\big[P(\boldsymbol{\varphi}), \{\lambda_a\}\big] = -\int\limits_\Omega d\mu\, P(\boldsymbol{\varphi})\log P(\boldsymbol{\varphi}) - \sum_{a=0}^K \lambda_a \left(\int\limits_\Omega d\mu\, P(\boldsymbol{\varphi})\, X^a(\boldsymbol{\varphi}) - X^a \right) \tag{1.78}$$

[10] This is conventional Jaynesian language, but more accurately the result depends critically on the chosen state space, reference measure, and constraints. There is no observer-independent definition of ignorance.

with respect to $P(\boldsymbol{\varphi})$ and with respect to $\{\lambda_a\}$. Again, $X^0(\boldsymbol{\varphi}) \equiv X^0 \equiv 1$. This yields the following result:

$$\log P(\boldsymbol{\varphi}) = -1 - \sum_{a=0}^{K} \lambda_a \, X^a(\boldsymbol{\varphi}) \quad . \tag{1.79}$$

The $(K+1)$ Lagrange multipliers $\{\lambda_a\}$ are then determined from the $(K+1)$ constraint equations in eqn. 1.77.

As an example, consider a distribution $P(x)$ over the real numbers $\mathbb{R}$. We constrain

$$\int\limits_{-\infty}^{\infty} dx \, P(x) = 1 \qquad , \qquad \int\limits_{-\infty}^{\infty} dx \, x \, P(x) = \mu \qquad , \qquad \int\limits_{-\infty}^{\infty} dx \, x^2 \, P(x) = \mu^2 + \sigma^2 \quad . \tag{1.80}$$

Extremizing the entropy, we then obtain

$$P(x) = \mathcal{C} \, e^{-\lambda_1 x - \lambda_2 x^2} \quad , \tag{1.81}$$

where $\mathcal{C} = e^{-(1+\lambda_0)}$. We already know the answer:

$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \, e^{-(x-\mu)^2/2\sigma^2} \quad . \tag{1.82}$$

In other words, $\lambda_1 = -\mu/\sigma^2$ and $\lambda_2 = 1/2\sigma^2$, with $\mathcal{C} = (2\pi\sigma^2)^{-1/2} \, \exp(-\mu^2/2\sigma^2)$.

1.5 General Aspects of Probability Distributions

1.5.1 Discrete and continuous distributions

Consider a system whose possible configurations $|\, n \,\rangle$ can be labeled by a discrete variable $n \in \mathcal{C}$, where $\mathcal{C}$ is the set of possible configurations. The total number of possible configurations, which is to say the *order* of the set $\mathcal{C}$, may be finite or infinite. Next, consider an ensemble of such systems, and let P_n denote the probability that a given random element from that ensemble is in the state (configuration) $|\, n \,\rangle$. The collection $\{P_n\}$ forms a *discrete probability distribution*. We assume that the distribution is *normalized*, meaning

$$\sum_{n \in \mathcal{C}} P_n = 1 \quad . \tag{1.83}$$

Now let A_n be a quantity which takes values depending on n. The average of A is given by

$$\langle A \rangle = \sum_{n \in \mathcal{C}} P_n \, A_n \quad . \tag{1.84}$$

Typically, $\mathcal{C}$ is the set of integers ($\mathbb{Z}$) or some subset thereof, but it could be any countable set. As an example, consider the throw of a single six-sided die. Then $P_n = \frac{1}{6}$ for each $n \in \{1, \ldots, 6\}$. Let $A_n = 0$ if

n is even and 1 if n is odd. We then find $\langle A \rangle = \frac{1}{2}$, *i.e.* on average half the throws of the die will result in an even number.

It may be that the system's configurations are described by several discrete variables $\{n_1, n_2, n_3, \ldots\}$. We can combine these into a vector $\boldsymbol{n}$ and then we write $P_{\boldsymbol{n}}$ for the discrete distribution, with $\sum_{\boldsymbol{n}} P_{\boldsymbol{n}} = 1$.

Another possibility is that the system's configurations are parameterized by a collection of continuous variables, $\boldsymbol{\varphi} = \{\varphi_1, \ldots, \varphi_n\}$. We write $\boldsymbol{\varphi} \in \Omega$, where Ω is the phase space (or configuration space) of the system. Let $d\mu$ be a *measure* on this space. In general, we can write

$$d\mu = W(\varphi_1, \ldots, \varphi_n)\, d\varphi_1\, d\varphi_2 \cdots d\varphi_n \quad . \tag{1.85}$$

The phase space measure used in classical statistical mechanics gives equal weight W to equal phase space volumes:

$$d\mu = \mathcal{C} \prod_{\sigma=1}^{r} dq_\sigma\, dp_\sigma \quad , \tag{1.86}$$

where $\mathcal{C}$ is a constant we shall discuss below[11].

Any continuous probability distribution $P(\boldsymbol{\varphi})$ is normalized according to

$$\int\limits_{\Omega} d\mu\, P(\boldsymbol{\varphi}) = 1 \quad . \tag{1.87}$$

The average of a function $A(\boldsymbol{\varphi})$ on configuration space is then

$$\langle A \rangle = \int\limits_{\Omega} d\mu\, P(\boldsymbol{\varphi})\, A(\boldsymbol{\varphi}) \quad . \tag{1.88}$$

For example, consider the Gaussian distribution

$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-(x-\mu)^2/2\sigma^2} \quad . \tag{1.89}$$

From the result[12]

$$\int\limits_{-\infty}^{\infty} dx\, e^{-\alpha x^2}\, e^{-\beta x} = \sqrt{\frac{\pi}{\alpha}}\, e^{\beta^2/4\alpha} \quad , \tag{1.90}$$

we see that $P(x)$ is normalized. One can then compute

$$\langle x \rangle = \mu \qquad , \qquad \langle x^2 \rangle - \langle x \rangle^2 = \sigma^2 \quad . \tag{1.91}$$

We call μ the *mean* and σ the *standard deviation* of the distribution, eqn. 1.89.

[11] Such a measure is invariant with respect to canonical transformations, which are the broad class of transformations among coordinates and momenta which leave Hamilton's equations of motion invariant, and which preserve phase space volumes under Hamiltonian evolution. For this reason $d\mu$ is called an *invariant phase space measure*.

[12] Memorize this!

The quantity $P(\boldsymbol{\varphi})$ is called the *distribution* or *probability density*. One has

$$P(\boldsymbol{\varphi})\, d\mu = \text{probability that configuration lies within volume } d\mu \text{ centered at } \boldsymbol{\varphi}$$

For example, consider the probability density $P = 1$ normalized on the interval $x \in \big[0, 1\big]$. The probability that some x chosen at random will be *exactly* $\frac{1}{2}$, say, is *zero* – one would have to specify each of the infinitely many digits of x. However, we can say that $x \in \big[0.45, 0.55\big]$ with probability $\frac{1}{10}$.

If x is distributed according to $P_1(x)$, then if x_1 and x_2 are independent copies drawn from $P_1(x)$, the probability distribution on the product space $(x_1\,,\,x_2)$ is simply the product of the distributions: $P_2(x_1, x_2) = P_1(x_1)\, P_1(x_2)$. Suppose we have a function $\phi(x_1, \ldots, x_N)$. How is it distributed? Let $P(\phi)$ be the distribution for ϕ. We then have

$$\begin{aligned} P(\phi) &= \int\limits_{-\infty}^{\infty} dx_1 \cdots \int\limits_{-\infty}^{\infty} dx_N\, P_N(x_1, \ldots, x_N)\, \delta\Big(\phi(x_1, \ldots, x_N) - \phi\Big) \\ &= \int\limits_{-\infty}^{\infty} dx_1 \cdots \int\limits_{-\infty}^{\infty} dx_N\, P_1(x_1) \cdots P_1(x_N)\, \delta\Big(\phi(x_1, \ldots, x_N) - \phi\Big) \quad , \end{aligned} \tag{1.92}$$

where the second line is appropriate if the $\{x_j\}$ are themselves distributed independently. Note that the distribution $P(\phi)$ is normalized, *i.e.* $\int\limits_{-\infty}^{\infty} d\phi\, P(\phi) = 1$.

1.5.2 Central limit theorem

In particular, consider the distribution function of the sum $X = \sum_{i=1}^{N} x_i$. We will be particularly interested in the case where N is large. For general N, though, we have

$$P_N(X) = \int\limits_{-\infty}^{\infty} dx_1 \cdots \int\limits_{-\infty}^{\infty} dx_N\, P_1(x_1) \cdots P_1(x_N)\, \delta\big(x_1 + x_2 + \ldots + x_N - X\big) \quad . \tag{1.93}$$

It is convenient to compute the Fourier transform of $P(X)$:

$$\begin{aligned} \hat{P}_N(k) &= \int\limits_{-\infty}^{\infty} dX\, P_N(X)\, e^{-ikX} \\ &= \int\limits_{-\infty}^{\infty} dX \int\limits_{-\infty}^{\infty} dx_1 \cdots \int\limits_{-\infty}^{\infty} dx_N\, P_1(x_1) \cdots P_1(x_N)\, \delta\big(x_1 + \ldots + x_N - X\big)\, e^{-ikX} = \big[\hat{P}_1(k)\big]^N \quad , \end{aligned} \tag{1.94}$$

where

$$\hat{P}_1(k) = \int\limits_{-\infty}^{\infty} dx\, P_1(x)\, e^{-ikx} \tag{1.95}$$

is the Fourier transform[13] of the single variable distribution $P_1(x)$. The distribution $P_N(X)$ is a *convolution* of the individual $P_1(x_i)$ distributions. We have therefore proven that *the Fourier transform of a convolution is the product of the Fourier transforms*.

OK, now we can write for $\hat{P}_1(k)$

$$\begin{aligned}\hat{P}_1(k) &= \int_{-\infty}^{\infty} dx\, P_1(x)\left(1 - ikx - \tfrac{1}{2}\,k^2x^2 + \tfrac{1}{6}ik^3\,x^3 + \ldots\right) \\ &= 1 - ik\langle x\rangle - \tfrac{1}{2}\,k^2\langle x^2\rangle + \tfrac{1}{6}\,ik^3\langle x^3\rangle + \ldots \quad .\end{aligned} \tag{1.96}$$

Thus,

$$\log\hat{P}_1(k) = -i\mu k - \tfrac{1}{2}\sigma^2k^2 + \tfrac{1}{6}\,i\gamma^3\,k^3 + \ldots \quad , \tag{1.97}$$

where

$$\mu = \langle x\rangle \qquad , \qquad \sigma^2 = \langle x^2\rangle - \langle x\rangle^2 \qquad , \qquad \gamma^3 = \langle x^3\rangle - 3\,\langle x^2\rangle\,\langle x\rangle + 2\,\langle x\rangle^3 \tag{1.98}$$

We can now write

$$\left[\hat{P}_1(k)\right]^N = e^{-iN\mu k}\,e^{-N\sigma^2k^2/2}\,e^{iN\gamma^3k^3/6}\ldots \quad . \tag{1.99}$$

Now for the inverse transform. In computing $P_N(X)$, we will expand the term $e^{iN\gamma^3k^3/6}$ and all subsequent terms in the above product as a power series in k. We then have

$$\begin{aligned}P_N(X) &= \int_{-\infty}^{\infty}\frac{dk}{2\pi}\,e^{ik(X-N\mu)}\,e^{-N\sigma^2k^2/2}\left\{1 + \tfrac{1}{6}\,i\,N\gamma^3k^3 + \ldots\right\} \\ &= \left(1 - \frac{\gamma^3}{6}N\,\frac{\partial^3}{\partial X^3} + \ldots\right)\frac{1}{\sqrt{2\pi N\sigma^2}}\,e^{-(X-N\mu)^2/2N\sigma^2} \\ &= \left(1 - \frac{\gamma^3}{6}\,N^{-1/2}\,\frac{\partial^3}{\partial\xi^3} + \ldots\right)\frac{1}{\sqrt{2\pi N\sigma^2}}\,e^{-\xi^2/2\sigma^2} \quad .\end{aligned} \tag{1.100}$$

[13] Jean Baptiste Joseph Fourier (1768-1830) had an illustrious career. The son of a tailor, and orphaned at age eight, Fourier's ignoble status rendered him ineligible to receive a commission in the scientific corps of the French army. A Benedictine minister at the École Royale Militaire of Auxerre remarked, "Fourier, not being noble, could not enter the artillery, although he were a second Newton." Fourier prepared for the priesthood but his affinity for mathematics proved overwhelming, and so he left the abbey and soon thereafter accepted a military lectureship position. Despite his initial support for revolution in France, in 1794 Fourier ran afoul of a rival sect while on a trip to Orléans and was arrested and very nearly guillotined. Fortunately the Reign of Terror ended soon after the death of Robespierre, and Fourier was released. He went on Napoleon Bonaparte's 1798 expedition to Egypt, where he was appointed governor of Lower Egypt. His organizational skills impressed Napoleon, and upon return to France he was appointed to a position of prefect in Grenoble. It was in Grenoble that Fourier performed his landmark studies of heat, and his famous work on partial differential equations and Fourier series. It seems that Fourier's fascination with heat began in Egypt, where he developed an appreciation of desert climate. His fascination developed into an obsession, and he became convinced that heat could promote a healthy body. He would cover himself in blankets, like a mummy, in his heated apartment, even during the middle of summer. On May 4, 1830, Fourier, so arrayed, tripped and fell down a flight of stairs. This aggravated a developing heart condition, which he refused to treat with anything other than more heat. Two weeks later, he died. Fourier's is one of the 72 names of scientists, engineers and other luminaries which are engraved on the Eiffel Tower.

In going from the second line to the third, we have written $X = N\mu + \sqrt{N}\,\xi$, in which case $\partial_X = N^{-1/2}\,\partial_\xi$, and the non-Gaussian terms give a subleading contribution which vanishes in the $N \to \infty$ limit. We have just proven the *central limit theorem*: for large but finite N, the distribution of a sum of N independent random variables x_i is approximately a Gaussian with mean $N\mu$ and standard deviation $\sqrt{N}\,\sigma$. Our only assumptions are that the mean μ and standard deviation σ exist for the distribution $P_1(x)$. Note that $P_1(x)$ itself need not be a Gaussian – it could be a very peculiar distribution indeed, but so long as its first and second moments exist, where the $k^{\rm th}$ moment is simply $\langle x^k\rangle$, the distribution of the sum $X = \sum_{i=1}^N x_i$ is a Gaussian.

Note that we have assumed the existence of a finite third moment γ in the above. Without this assumption, we may still conclude

$$\hat{P}(k) = 1 - i\mu k - \tfrac{1}{2}\langle x^2\rangle\, k^2 + o(k^2) \quad . \tag{1.101}$$

1.5.3 Moments and cumulants

Consider a general multivariate distribution $P(x_1, \ldots, x_N)$ and define the multivariate Fourier transform

$$\hat{P}(k_1, \ldots, k_N) = \int\limits_{-\infty}^{\infty} dx_1 \cdots \int\limits_{-\infty}^{\infty} dx_N\, P(x_1, \ldots, x_N)\, \exp\left(-i \sum_{j=1}^N k_j x_j\right) \quad . \tag{1.102}$$

The inverse relation is

$$P(x_1, \ldots, x_N) = \int\limits_{-\infty}^{\infty} \frac{dk_1}{2\pi} \cdots \int\limits_{-\infty}^{\infty} \frac{dk_N}{2\pi}\, \hat{P}(k_1, \ldots, k_N)\, \exp\left(+i \sum_{j=1}^N k_j x_j\right) \quad . \tag{1.103}$$

Acting on $\hat{P}(\boldsymbol{k})$, the differential operator $i\,\frac{\partial}{\partial k_i}$ brings down from the exponential a factor of x_i inside the integral. Thus,

$$\left[\left(i\,\frac{\partial}{\partial k_1}\right)^{m_1} \cdots \left(i\,\frac{\partial}{\partial k_N}\right)^{m_N} \hat{P}(\boldsymbol{k})\right]_{\boldsymbol{k}=0} = \left\langle x_1^{m_1} \cdots x_N^{m_N}\right\rangle \quad . \tag{1.104}$$

Similarly, we can reconstruct the distribution from its moments, *viz.*

$$\hat{P}(\boldsymbol{k}) = \sum_{m_1=0}^{\infty} \cdots \sum_{m_N=0}^{\infty} \frac{(-ik_1)^{m_1}}{m_1!} \cdots \frac{(-ik_N)^{m_N}}{m_N!} \left\langle x_1^{m_1} \cdots x_N^{m_N}\right\rangle \quad . \tag{1.105}$$

The *cumulants* $\langle\!\langle x_1^{m_1} \cdots x_N^{m_N}\rangle\!\rangle$ are defined by the Taylor expansion of $\log \hat{P}(\boldsymbol{k})$:

$$\log \hat{P}(\boldsymbol{k}) = \sum_{m_1=0}^{\infty} \cdots \sum_{m_N=0}^{\infty} \frac{(-ik_1)^{m_1}}{m_1!} \cdots \frac{(-ik_N)^{m_N}}{m_N!} \left\langle\!\left\langle x_1^{m_1} \cdots x_N^{m_N}\right\rangle\!\right\rangle \quad . \tag{1.106}$$

There is no general form for the cumulants. It is straightforward to derive the following low order results:

$$\begin{aligned}
\langle\!\langle x_i\rangle\!\rangle &= \langle x_i\rangle \\
\langle\!\langle x_i x_j\rangle\!\rangle &= \langle x_i x_j\rangle - \langle x_i\rangle\langle x_j\rangle \\
\langle\!\langle x_i x_j x_k\rangle\!\rangle &= \langle x_i x_j x_k\rangle - \langle x_i x_j\rangle\langle x_k\rangle - \langle x_j x_k\rangle\langle x_i\rangle - \langle x_k x_i\rangle\langle x_j\rangle + 2\langle x_i\rangle\langle x_j\rangle\langle x_k\rangle \quad .
\end{aligned} \tag{1.107}$$

1.5.4 Multidimensional Gaussian integral

Consider the multivariable Gaussian distribution,

$$P(\boldsymbol{x}) \equiv \left(\frac{\det A}{(2\pi)^n}\right)^{1/2} \exp\Big(-\tfrac{1}{2}\, x_i\, A_{ij}\, x_j\Big) \quad , \tag{1.108}$$

where A is a real, symmetric, positive definite matrix of rank n. A mathematical result which is extremely important throughout physics is the following:

$$Z(\boldsymbol{b}) = \left(\frac{\det A}{(2\pi)^n}\right)^{1/2} \int\limits_{-\infty}^{\infty} dx_1 \cdots \int\limits_{-\infty}^{\infty} dx_n\ \exp\Big(-\tfrac{1}{2}\, x_i\, A_{ij}\, x_j + b_i\, x_i\Big) = \exp\left(\tfrac{1}{2}\, b_i\, A^{-1}_{ij}\, b_j\right) \quad . \tag{1.109}$$

Here, the vector $\boldsymbol{b} = (b_1\,,\,\ldots\,,\,b_n)$ is identified as a *source*. Since $Z(0) = 1$, we have that the distribution $P(\boldsymbol{x})$ is normalized. Now consider averages of the form

$$\begin{aligned} \langle\, x_{j_1} \cdots\, x_{j_{2k}} \,\rangle &= \int d^n\!x\; P(\boldsymbol{x})\; x_{j_1} \cdots\, x_{j_{2k}} = \left.\frac{\partial^{2k} Z(\boldsymbol{b})}{\partial b_{j_1} \cdots\, \partial b_{j_{2k}}}\right|_{\boldsymbol{b}=0} \\ &= \sum_{\text{contractions}} A^{-1}_{j_{\sigma(1)} j_{\sigma(2)}} \cdots A^{-1}_{j_{\sigma(2k-1)} j_{\sigma(2k)}} \quad . \end{aligned} \tag{1.110}$$

The sum in the last term is over all *contractions* of the indices $\{j_1\,,\,\ldots\,,\,j_{2k}\}$. A contraction is an arrangement of the $2k$ indices into k pairs. There are $C_{2k} = (2k)!/2^k k!$ possible such contractions. To obtain this result for C_{2k}, we start with the first index and then find a mate among the remaining $2k-1$ indices. Then we choose the next unpaired index and find a mate among the remaining $2k-3$ indices. Proceeding in this manner, we have

$$C_{2k} = (2k-1)\cdot(2k-3)\cdots 3\cdot 1 = \frac{(2k)!}{2^k k!} \quad . \tag{1.111}$$

Equivalently, we can take all possible permutations of the $2k$ indices, and then divide by $2^k k!$ since permutation within a given pair results in the same contraction and permutation among the k pairs results in the same contraction. For example, for $k = 2$, we have $C_4 = 3$, and

$$\langle\, x_{j_1} x_{j_2} x_{j_3} x_{j_4} \,\rangle = A^{-1}_{j_1 j_2} A^{-1}_{j_3 j_4} + A^{-1}_{j_1 j_3} A^{-1}_{j_2 j_4} + A^{-1}_{j_1 j_4} A^{-1}_{j_2 j_3} \quad . \tag{1.112}$$

If we define $b_i = ik_i$, we have

$$\hat{P}(\boldsymbol{k}) = \exp\Big(-\tfrac{1}{2}\, k_i\, A^{-1}_{ij}\, k_j\Big) \quad , \tag{1.113}$$

from which we read off the cumulants $\langle\!\langle x_i x_j \rangle\!\rangle = A^{-1}_{ij}$, with all higher order cumulants vanishing.

1.6 Appendix: Bayesian Statistical Inference

1.6.1 Frequentists and Bayesians

The field of statistical inference is roughly divided into two schools of practice: frequentism and Bayesianism. You can find several articles on the web discussing the differences in these two approaches. In both

cases we would like to model observable data $\boldsymbol{x}$ by a distribution. The distribution in general depends on one or more parameters $\boldsymbol{\theta}$. The basic worldviews of the two approaches are as follows:

> **Frequentism:** Data $\boldsymbol{x}$ are a random sample drawn from an infinite pool at some *frequency*. The underlying parameters $\boldsymbol{\theta}$, which are to be estimated, remain fixed during this process. There is no information prior to the model specification. The experimental conditions under which the data are collected are presumed to be controlled and repeatable. Results are generally expressed in terms of *confidence intervals* and *confidence levels*, obtained via *statistical hypothesis testing*. Probabilities have meaning only for data yet to be collected. Calculations generally are computationally straightforward.

> **Bayesianism:** The only data $\boldsymbol{x}$ which matter are those which have been observed. The parameters $\boldsymbol{\theta}$ are unknown and due to uncertainty are described probabilistically using a *prior distribution*, which is generally based on some available information but which also may be at least partially subjective. The priors are then to be *updated* based on observed data $\boldsymbol{x}$. Results are expressed in terms of *posterior distributions* and *credible intervals*. Calculations can be computationally intensive.

In essence, frequentists say *the data are random and the parameters are fixed*. While Bayesians say *the data are fixed and the parameters are random*[14]. Overall, frequentism has dominated over the past several hundred years, but Bayesianism has been coming on strong of late, and many physicists seem naturally drawn to the Bayesian perspective.

1.6.2 Updating Bayesian priors

Given *data* D and a *hypothesis* H, Bayes' theorem tells us

$$P(H \,|\, D) = \frac{P(D \,|\, H)\, P(H)}{P(D)} \quad . \tag{1.114}$$

Typically the data are in the form of a set of values $\boldsymbol{x} = \{x_1, \ldots, x_N\}$, and the hypothesis in the form of a set of parameters $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_K\}$. It is notationally helpful to express distributions of $\boldsymbol{x}$ and distributions of $\boldsymbol{x}$ conditioned on $\boldsymbol{\theta}$ using the symbol f, and distributions of $\boldsymbol{\theta}$ and distributions of $\boldsymbol{\theta}$ conditioned on $\boldsymbol{x}$ using the symbol π, rather than using the symbol P everywhere. We then have

$$\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x}) = \frac{f(\boldsymbol{x} \,|\, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta})}{\int\limits_{\Theta} d\boldsymbol{\theta}'\, f(\boldsymbol{x} \,|\, \boldsymbol{\theta}')\, \pi(\boldsymbol{\theta}')} \quad , \tag{1.115}$$

where $\Theta \ni \boldsymbol{\theta}$ is the space of parameters. Note that $\int_\Theta d\boldsymbol{\theta}\, \pi(\boldsymbol{\theta} \,|\, \boldsymbol{x}) = 1$. The denominator of the RHS is simply $f(\boldsymbol{x})$, which is independent of $\boldsymbol{\theta}$, hence $\pi(\boldsymbol{\theta} \,|\, \boldsymbol{x}) \propto f(\boldsymbol{x} \,|\, \boldsymbol{\theta})\, \pi(\boldsymbol{\theta})$. We call $\pi(\boldsymbol{\theta})$ the *prior* for

[14] "A frequentist is a person whose long-run ambition is to be wrong 5% of the time. A Bayesian is one who, vaguely expecting a horse, and catching glimpse of a donkey, strongly believes he has seen a mule." – Charles Annis.

$\boldsymbol{\theta}$, $f(\boldsymbol{x}\,|\,\boldsymbol{\theta})$ the *likelihood* of $\boldsymbol{x}$ given $\boldsymbol{\theta}$, and $\pi(\boldsymbol{\theta}\,|\,\boldsymbol{x})$ the *posterior* for $\boldsymbol{\theta}$ given $\boldsymbol{x}$. The idea here is that while our initial guess at the $\boldsymbol{\theta}$ distribution is given by the prior $\pi(\boldsymbol{\theta})$, after taking data, we should *update* this distribution to the posterior $\pi(\boldsymbol{\theta}\,|\,\boldsymbol{x})$. The likelihood $f(\boldsymbol{x}\,|\,\boldsymbol{\theta})$ is entailed by our model for the phenomenon which produces the data. We can use the posterior to find the distribution of new data points $\boldsymbol{y}$, called the *posterior predictive distribution*,

$$f(\boldsymbol{y}\,|\,\boldsymbol{x}) = \int\limits_{\Theta} d\boldsymbol{\theta}\, f(\boldsymbol{y}\,|\,\boldsymbol{\theta})\,\pi(\boldsymbol{\theta}\,|\,\boldsymbol{x}) \quad . \tag{1.116}$$

This is the update of the *prior predictive distribution*,

$$f(\boldsymbol{x}) = \int\limits_{\Theta} d\boldsymbol{\theta}\, f(\boldsymbol{x}\,|\,\boldsymbol{\theta})\,\pi(\boldsymbol{\theta}) \quad . \tag{1.117}$$

Example: coin flipping

Consider a model of coin flipping based on a standard Bernoulli distribution, where $\theta \in [0,1]$ is the probability for heads ($x=1$) and $1-\theta$ the probability for tails ($x=0$). That is,

$$\begin{aligned} f(x_1,\ldots,x_N\,|\,\theta) &= \prod_{j=1}^{N}\Big[(1-\theta)\,\delta_{x_j,0} + \theta\,\delta_{x_j,1}\Big] \\ &= \theta^X(1-\theta)^{N-X} \quad , \end{aligned} \tag{1.118}$$

where $X = \sum_{j=1}^N x_j$ is the observed total number of heads, and $N-X$ the corresponding number of tails. We now need a prior $\pi(\theta)$. We choose the Beta distribution,

$$\pi(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\mathsf{B}(\alpha,\beta)} \quad , \tag{1.119}$$

where $\mathsf{B}(\alpha,\beta) = \Gamma(\alpha)\,\Gamma(\beta)/\Gamma(\alpha+\beta)$ is the Beta function. One can check that $\pi(\theta)$ is normalized on the unit interval: $\int_0^1 d\theta\,\pi(\theta) = 1$ for all positive α,β. Even if we limit ourselves to this form of the prior, different Bayesians might bring different assumptions about the values of α and β. Note that if we choose $\alpha=\beta=1$, the prior distribution for θ is flat, with $\pi(\theta)=1$.

We now compute the posterior distribution for θ:

$$\pi(\theta\,|\,x_1,\ldots,x_N) = \frac{f(x_1,\ldots,x_N\,|\,\theta)\,\pi(\theta)}{\int_0^1 d\theta'\, f(x_1,\ldots,x_N\,|\,\theta')\,\pi(\theta')} = \frac{\theta^{X+\alpha-1}(1-\theta)^{N-X+\beta-1}}{\mathsf{B}(X+\alpha,N-X+\beta)} \quad . \tag{1.120}$$

Thus, we retain the form of the Beta distribution, but with updated parameters,

$$\begin{aligned} \alpha' &= X+\alpha \\ \beta' &= N-X+\beta \quad . \end{aligned} \tag{1.121}$$

The fact that the functional form of the prior is retained by the posterior is generally *not* the case in Bayesian updating. We can also compute the prior predictive,

$$\begin{aligned} f(x_1,\ldots,x_N) &= \int\limits_0^1 d\theta\; f(x_1,\ldots,x_N\,|\,\theta)\,\pi(\theta) \\ &= \frac{1}{\mathsf{B}(\alpha,\beta)}\int\limits_0^1 d\theta\;\theta^{X+\alpha-1}(1-\theta)^{N-X+\beta-1} = \frac{\mathsf{B}(X+\alpha,N-X+\beta)}{\mathsf{B}(\alpha,\beta)} \quad . \end{aligned} \tag{1.122}$$

The posterior predictive is then

$$\begin{aligned} f(y_1,\ldots,y_M\,|\,x_1,\ldots,x_N) &= \int\limits_0^1 d\theta\; f(y_1,\ldots,y_M\,|\,\theta)\,\pi(\theta\,|\,x_1,\ldots,x_N) \\ &= \frac{1}{\mathsf{B}(X+\alpha,N-X+\beta)}\int\limits_0^1 d\theta\;\theta^{X+Y+\alpha-1}(1-\theta)^{N-X+M-Y+\beta-1} \\ &= \frac{\mathsf{B}(X+Y+\alpha,N-X+M-Y+\beta)}{\mathsf{B}(X+\alpha,N-X+\beta)} \quad . \end{aligned} \tag{1.123}$$

1.6.3 Hyperparameters and conjugate priors

In the above example, θ is a *parameter* of the Bernoulli distribution, *i.e.* the likelihood, while quantities α and β are *hyperparameters* which enter the prior $\pi(\theta)$. Accordingly, we could have written $\pi(\theta\,|\,\alpha,\beta)$ for the prior. We then have for the posterior

$$\pi(\boldsymbol{\theta}\,|\,\boldsymbol{x},\boldsymbol{\alpha}) = \frac{f(\boldsymbol{x}\,|\,\boldsymbol{\theta})\,\pi(\boldsymbol{\theta}\,|\,\boldsymbol{\alpha})}{\int\limits_\Theta d\boldsymbol{\theta}'\,f(\boldsymbol{x}\,|\,\boldsymbol{\theta}')\,\pi(\boldsymbol{\theta}'\,|\,\boldsymbol{\alpha})} \quad , \tag{1.124}$$

replacing eqn. 1.115, *etc.*, where $\boldsymbol{\alpha} \in A$ is the vector of hyperparameters. The hyperparameters can also be distributed, according to a *hyperprior* $\rho(\boldsymbol{\alpha})$, and the hyperpriors can further be parameterized by *hyperhyperparameters*, which can have their own distributions, *ad nauseam*.

What use is all this? We've already seen a compelling example: when the posterior is of the same form as the prior, the Bayesian update can be viewed as an automorphism of the hyperparameter space A, *i.e.* one set of hyperparameters $\boldsymbol{\alpha}$ is mapped to a new set of hyperparameters $\widetilde{\boldsymbol{\alpha}}$.

Definition: A parametric family of distributions $\mathcal{P} = \big\{\pi(\boldsymbol{\theta}\,|\,\boldsymbol{\alpha}) \mid \boldsymbol{\theta}\in\Theta,\,\boldsymbol{\alpha}\in A\big\}$ is called a *conjugate family* for a family of distributions $\big\{f(\boldsymbol{x}\,|\,\boldsymbol{\theta}) \mid \boldsymbol{x}\in\mathcal{X},\,\boldsymbol{\theta}\in\Theta\big\}$ if, for all $\boldsymbol{x}\in\mathcal{X}$ and $\boldsymbol{\alpha}\in A$,

$$\pi(\boldsymbol{\theta}\,|\,\boldsymbol{x},\boldsymbol{\alpha}) \equiv \frac{f(\boldsymbol{x}\,|\,\boldsymbol{\theta})\,\pi(\boldsymbol{\theta}\,|\,\boldsymbol{\alpha})}{\int\limits_\Theta d\boldsymbol{\theta}'\,f(\boldsymbol{x}\,|\,\boldsymbol{\theta}')\,\pi(\boldsymbol{\theta}'\,|\,\boldsymbol{\alpha})} \in \mathcal{P} \quad . \tag{1.125}$$

That is, $\pi(\boldsymbol{\theta}\,|\,\boldsymbol{x},\boldsymbol{\alpha}) = \pi(\boldsymbol{\theta}\,|\,\widetilde{\boldsymbol{\alpha}})$ for some $\widetilde{\boldsymbol{\alpha}}\in A$, with $\widetilde{\boldsymbol{\alpha}} = \widetilde{\boldsymbol{\alpha}}(\boldsymbol{\alpha},\boldsymbol{x})$.

As an example, consider the conjugate Bayesian analysis of the Gaussian distribution. We assume a likelihood

$$f(\boldsymbol{x}\,|\,u,s) = (2\pi s^2)^{-N/2}\exp\left\{-\frac{1}{2s^2}\sum_{j=1}^{N}(x_j-u)^2\right\} \tag{1.126}$$

with s regarded as *fixed*. There is now a single parameter: $\boldsymbol{\theta} = \{u\}$. Now consider the prior distribution

$$\pi(u\,|\,\mu_0,\sigma_0) = (2\pi\sigma_0^2)^{-1/2}\,\exp\left\{-\frac{(u-\mu_0)^2}{2\sigma_0^2}\right\} \quad . \tag{1.127}$$

Note that the prior distribution is independent of the parameter s and only depends on u and the hyperparameters $\boldsymbol{\alpha} = (\mu_0,\sigma_0)$. We now compute the posterior:

$$\begin{aligned}\pi(u\,|\,\boldsymbol{x},s,\mu_0,\sigma_0) &\propto f(\boldsymbol{x}\,|\,u,s)\,\pi(u\,|\,\mu_0,\sigma_0)\\ &= \exp\left\{-\left(\frac{1}{2\sigma_0^2}+\frac{N}{2s^2}\right)u^2+\left(\frac{\mu_0}{\sigma_0^2}+\frac{N\langle x\rangle}{s^2}\right)u-\left(\frac{\mu_0^2}{2\sigma_0^2}+\frac{N\langle x^2\rangle}{2s^2}\right)\right\} \quad ,\end{aligned} \tag{1.128}$$

with $\langle x\rangle = N^{-1}\sum_{j=1}^N x_j$ and $\langle x^2\rangle = N^{-1}\sum_{j=1}^N x_j^2$. This is also a Gaussian distribution for u, and after supplying the appropriate normalization one finds

$$\pi(u\,|\,\boldsymbol{x},s,\mu_0,\sigma_0) = (2\pi\sigma_1^2)^{-1/2}\exp\left\{-\frac{(u-\mu_1)^2}{2\sigma_1^2}\right\} \quad , \tag{1.129}$$

which is normalized over u, with

$$\mu_1 = \mu_0 + \frac{N\big(\langle x\rangle-\mu_0\big)\sigma_0^2}{s^2+N\sigma_0^2} \qquad , \qquad \sigma_1^2 = \frac{s^2\sigma_0^2}{s^2+N\sigma_0^2} \quad . \tag{1.130}$$

Thus, the posterior is among the same family as the prior, and we have derived the update rule for the hyperparameters $(\mu_0,\sigma_0) \to (\mu_1,\sigma_1)$. Note that $\sigma_1 < \sigma_0$, so the updated Gaussian prior is sharper than the original. The updated mean μ_1 shifts in the direction of $\langle x\rangle$ obtained from the data set.

1.6.4 The problem with priors

We might think that, for the coin flipping problem, the flat prior $\pi(\theta) = 1$ is an appropriate initial one, since it does not privilege any value of θ. This prior therefore seems 'objective' or 'unbiased', also called 'uninformative'. But suppose we make a change of variables, mapping the interval $\theta \in [0,1]$ to the entire real line according to $\zeta = \log\big[\theta/(1-\theta)\big]$. In terms of the new parameter ζ, we write the prior as $\tilde\pi(\zeta)$. Clearly $\pi(\theta)\,d\theta = \tilde\pi(\zeta)\,d\zeta$, so $\tilde\pi(\zeta) = \pi(\theta)\,d\theta/d\zeta$. For our example, find $\tilde\pi(\zeta) = \frac{1}{4}\mathrm{sech}^2(\zeta/2)$, which is not flat. Thus what was uninformative in terms of θ has become very informative in terms of the new parameter ζ. Is there any truly unbiased way of selecting a Bayesian prior?

One approach, advocated by E. T. Jaynes, is to choose the prior distribution $\pi(\boldsymbol{\theta})$ according to the principle of maximum entropy. For continuous parameter spaces, we must first define a parameter space metric so as to be able to 'count' the number of different parameter states. The entropy of a distribution $\pi(\boldsymbol{\theta})$ is then dependent on this metric: $S = -\int d\mu(\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})\,\log\pi(\boldsymbol{\theta})$.

Another approach, due to Jeffreys, is to derive a parameterization-independent prior from the likelihood $f(\boldsymbol{x}\,|\,\boldsymbol{\theta})$ using the so-called *Fisher information matrix*,

$$I_{ij}(\boldsymbol{\theta}) = -\mathbb{E}_\theta\left(\frac{\partial^2 \log f(\boldsymbol{x}\,|\,\boldsymbol{\theta})}{\partial\theta_i\,\partial\theta_j}\right) = -\int d\boldsymbol{x}\; f(\boldsymbol{x}\,|\,\boldsymbol{\theta})\;\frac{\partial^2 \log f(\boldsymbol{x}\,|\,\boldsymbol{\theta})}{\partial\theta_i\,\partial\theta_j} \quad . \tag{1.131}$$

The *Jeffreys prior* $\pi_{\rm J}(\boldsymbol{\theta})$ is defined as

$$\pi_{\rm J}(\boldsymbol{\theta}) \propto \sqrt{\det I(\boldsymbol{\theta})} \quad . \tag{1.132}$$

One can check that the Jeffreys prior is invariant under reparameterization. As an example, consider the Bernoulli process, for which $\log f(\boldsymbol{x}\,|\,\theta) = X\log\theta + (N-X)\log(1-\theta)$, where $X = \sum_{j=1}^N x_j$. Then

$$-\frac{d^2 \log p(\boldsymbol{x}\,|\,\theta)}{d\theta^2} = \frac{X}{\theta^2} + \frac{N-X}{(1-\theta)^2} \quad , \tag{1.133}$$

and since $\mathbb{E}_\theta\, X = N\theta$, we have

$$I(\theta) = \frac{N}{\theta(1-\theta)} \qquad \Rightarrow \qquad \pi_{\rm J}(\theta) = \frac{1}{\pi}\,\frac{1}{\sqrt{\theta(1-\theta)}} \quad , \tag{1.134}$$

which felicitously corresponds to a Beta distribution with $\alpha = \beta = \frac{1}{2}$. In this example the Jeffreys prior turned out to be a conjugate prior, but in general this is not the case.

We can try to implement the Jeffreys procedure for a two-parameter family where each x_j is normally distributed with mean μ and standard deviation σ. Let the parameters be $(\theta_1, \theta_2) = (\mu, \sigma)$. Then

$$-\log f(\boldsymbol{x}\,|\,\boldsymbol{\theta}) = N\log\sqrt{2\pi} + N\log\sigma + \frac{1}{2\sigma^2}\sum_{j=1}^N (x_j-\mu)^2 \quad , \tag{1.135}$$

and the Fisher information matrix is

$$I(\boldsymbol{\theta}) = -\frac{\partial^2 \log f(\boldsymbol{x}\,|\,\boldsymbol{\theta})}{\partial\theta_i\,\partial\theta_j} = \begin{pmatrix} N\sigma^{-2} & \sigma^{-3}\sum_j (x_j-\mu) \\ \sigma^{-3}\sum_j (x_j-\mu) & -N\sigma^{-2} + 3\sigma^{-4}\sum_j (x_j-\mu)^2 \end{pmatrix} \quad . \tag{1.136}$$

Taking the expectation value, we have $\mathbb{E}\,(x_j-\mu) = 0$ and $\mathbb{E}\,(x_j-\mu)^2 = \sigma^2$, hence

$$\mathbb{E}\,I(\boldsymbol{\theta}) = \begin{pmatrix} N\sigma^{-2} & 0 \\ 0 & 2N\sigma^{-2} \end{pmatrix} \tag{1.137}$$

and the Jeffreys prior is $\pi_{\rm J}(\mu,\sigma) \propto \sigma^{-2}$. This is problematic because if we choose a flat metric on the (μ,σ) upper half-plane, the Jeffreys prior is not normalizable. Note also that the Jeffreys prior no longer resembles a Gaussian, and hence is not a conjugate prior.