Contents

Chapter 1

Probability

1.1 References

- F. Reif, *Fundamentals of Statistical and Thermal Physics* (McGraw-Hill, 1987)
  This has been perhaps the most popular undergraduate text since it first appeared in 1967, and with good reason.
- E. T. Jaynes, *Probability Theory* (Cambridge, 2007)
  The bible on probability theory for physicists. A strongly Bayesian approach.
- C. Gardiner, *Stochastic Methods* (Springer-Verlag, 2010)
  Very clear and complete text on stochastic mathematics.

1.2 A Statistical View

1.2.1 Distributions for a 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

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

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

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

This is a normalized discrete probability distribution of the type discussed in section 1.4 below. The multivariate distribution for all the steps is then

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

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:

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

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:

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

Here we have used

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

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

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

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[1], but the RMS

[1] The exception is the unbiased case $p = q = \frac{1}{2}$, where $\langle X \rangle = 0$.

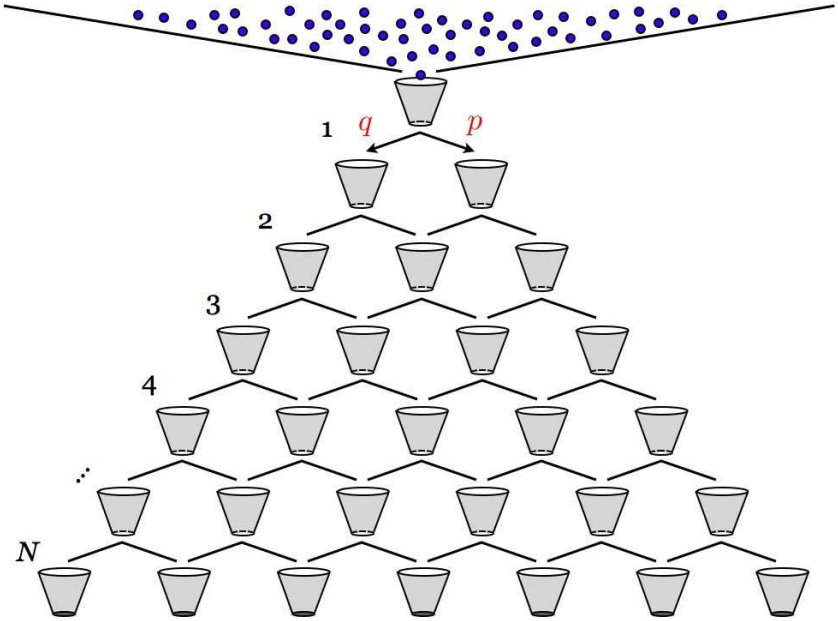

Figure 1.1: The falling ball system, which mimics a one-dimensional random walk.

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 §1.4.2 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

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

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

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

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

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

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 $\ln P_{N,X}$ using Stirling's asymptotic expansion

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

We then have

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

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

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

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

$$P_{N,X} = \mathcal{C}\,e^{-Nf(X/N)}\ , \tag{1.14}$$

where $\mathcal{C}$ is a normalization constant[2]. 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

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

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

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

We also have

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

so invoking Taylor's theorem,

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

Putting it all together, we have

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

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

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

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 §1.4.2 below.

[2]The origin of $\mathcal{C}$ lies in the $\mathcal{O}(\ln N)$ and $\mathcal{O}(N^0)$ terms in the asymptotic expansion of $\ln N!$. We have ignored these terms here. Accounting for them carefully reproduces the correct value of $\mathcal{C}$ in eqn. 1.20.

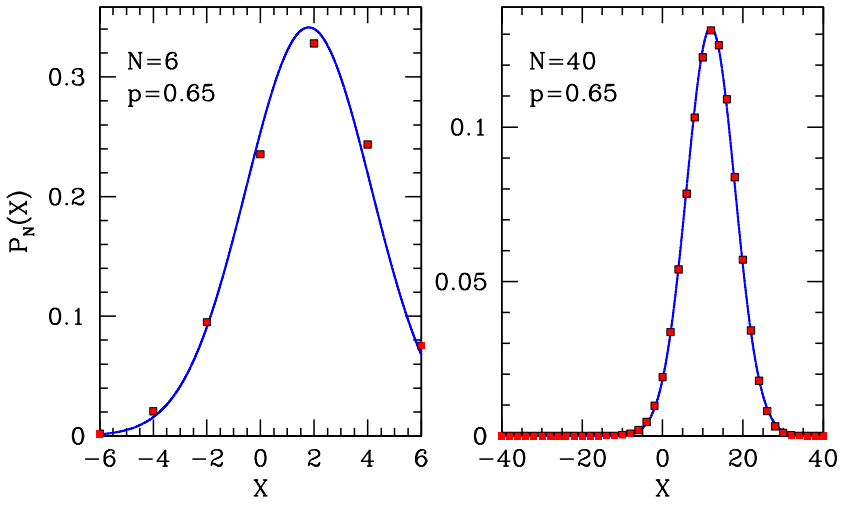

Figure 1.2: Comparison of exact distribution of eqn. 1.10 (red squares) with the Gaussian distribution of eqn. 1.19 (blue line).

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

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

The function $S(N,x) \equiv Ns(x)$ is analogous to the *statistical entropy* of our system[3]. We have

$$S(N,x) = Ns(x) = \ln\binom{N}{N_{\rm R}} = \ln\binom{N}{\frac{1}{2}N(1+x)}\ . \tag{1.22}$$

Thus, *the statistical entropy is the logarithm of the number of ways the system can be configured so as to yield the same value of X (at fixed N)*. The second contribution to $f(x)$ is the energy term. We write

$$E(N,x) = Ne(x) = -\tfrac{1}{2}N\ln(pq) - \tfrac{1}{2}Nx\ln(p/q)\ . \tag{1.23}$$

The energy term 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 $p < q$ (*i.e.* $p < \frac{1}{2}$), the energy is minimized by taking x as small as possible (meaning as negative as possible). The smallest possible allowed value of $x = X/N$ is $x = -1$. Conversely, when $p > q$ (*i.e.* $p > \frac{1}{2}$), the energy is minimized by taking x as large as possible, which means $x = 1$. 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.

In actual thermodynamic systems, as we shall see, 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.

[3]The function $s(x)$ is the *specific entropy*.

1.2.4 Entropy and information theory

It was shown in the classic 1948 work of Claude Shannon that entropy is in fact a measure of *information*[4]. 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

$$I(p) = -A\, \ln p\ , \tag{1.24}$$

where A is an arbitrary constant that can be absorbed into the base of the logarithm, since $\log_b x = \ln x / \ln b$. We will take $A = 1$ and use e as the base, so $I(p) = -\ln p$. Another common choice 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 $I(0) = \infty$. This means that the observation of an extremely rare event carries a great deal of information.

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 $-\ln p_n$, we have that the average information per observation is

$$S = \frac{\langle I_N \rangle}{N} = -\sum_n p_n \ln p_n\ , \tag{1.25}$$

which is known as the entropy of the distribution. Thus, maximizing S is equivalent to maximizing the *information* content per observation.

Consider, for example, the information content of course grades. As we have seen, 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 binary entropy is then $S = \log_2 \Gamma$, 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

$$g(N) = \frac{N!}{\prod_{n=1}^K N_n!}\ . \tag{1.26}$$

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

$$\log_2 g(N) = \log_2(N!) - \sum_{n=1}^K \log_2(N_n!) \approx -N \sum_{n=1}^K p_n \log_2 p_n\ , \tag{1.27}$$

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

where we have invoked Stirling's approximation. If the distribution is uniform, then we have $p_n = \frac{1}{K}$ for all $n \in \{1, \ldots, K\}$, and $\log_2 g(N) = N \log_2 K$.

1.3 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.

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 $\langle X \rangle = (2p-1)N$. 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

$$\sum_{\{\sigma_j\}} P(\sigma_1, \ldots, \sigma_N) = 1 \,, \tag{1.28}$$

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

$$P(\sigma_1, \ldots, \sigma_N) = \prod_{j=1}^{N} \left(p\, \delta_{\sigma_j, 1} + q\, \delta_{\sigma_j, -1} \right) = \prod_{j=1}^{N} p^{(1+\sigma_j)/2}\, q^{(1-\sigma_j)/2} \,, \tag{1.29}$$

corresponding to N independent steps?

1.3.1 The principle of maximum entropy

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

$$S = -\sum_n p_n \ln p_n \,, \tag{1.30}$$

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

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

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

$$S^*\big(\{p_n\}, \lambda\big) = -\sum_n p_n \ln p_n - \lambda \Big(\sum_n p_n - 1 \Big) \tag{1.31}$$

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

$$\frac{\partial S^*}{\partial p_n} = -\big(\ln p_n + 1 + \lambda\big) = 0 \tag{1.32}$$

as well as

$$\frac{\partial S^*}{\partial \lambda} = \sum_n p_n - 1 = 0\ . \tag{1.33}$$

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

$$\sum_n p_n = e^{-(1+\lambda)} \cdot \sum_n 1 = \Gamma\, e^{-(1+\lambda)}\ , \tag{1.34}$$

where $\Gamma \equiv \sum_n 1$ is the total number of possible events. Thus,

$$p_n = \frac{1}{\Gamma}\ , \tag{1.35}$$

which says that all events are equally probable.

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

$$S^*\big(\{p_n\}, \lambda_0, \lambda_1\big) = -\sum_n p_n \ln p_n - \lambda_0\Big(\sum_n p_n - 1\Big) - \lambda_1\Big(\sum_n X_n\, p_n - X\Big)\ . \tag{1.36}$$

We then have

$$\frac{\partial S^*}{\partial p_n} = -\big(\ln p_n + 1 + \lambda_0 + \lambda_1\, X_n\big) = 0\ , \tag{1.37}$$

which yields the two-parameter distribution

$$p_n = e^{-(1+\lambda_0)}\, e^{-\lambda_1 X_n}\ . \tag{1.38}$$

To fully determine the distribution $\{p_n\}$ we need to 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:

$$e^{-(1+\lambda_0)} \sum_n e^{-\lambda_1 X_n} = 1 \tag{1.39}$$

$$e^{-(1+\lambda_0)} \sum_n X_n\, e^{-\lambda_1 X_n} = X\ . \tag{1.40}$$

General formulation

The generalization to K extra pieces of information (plus normalization) is immediately apparent. We have

$$X^a = \sum_n X^a_n\, p_n\ , \tag{1.41}$$

and therefore we define

$$S^*\big(\{p_n\}, \{\lambda_a\}\big) = -\sum_n p_n \ln p_n - \sum_{a=0}^{K} \lambda_a\Big(\sum_n X^a_n\, p_n - X^a\Big)\ , \tag{1.42}$$

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

$$\begin{aligned} p_n &= \exp\left\{ -1 - \sum_{a=0}^{K} \lambda_a \, X_n^a \right\} \\ &= \frac{1}{Z} \exp\left\{ - \sum_{a=1}^{K} \lambda_a \, X_n^a \right\} , \end{aligned} \tag{1.43}$$

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. 1.41.

Example

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

$$\sum_{\sigma_1} \cdots \sum_{\sigma_N} P(\sigma_1, \ldots, \sigma_N) = 1 \tag{1.44}$$

$$\sum_{\sigma_1} \cdots \sum_{\sigma_N} P(\sigma_1, \ldots, \sigma_N) \sum_{j=1}^{N} \sigma_j = X \; . \tag{1.45}$$

Here the discrete label n from §1.3.1 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

$$P(\sigma_1, \ldots, \sigma_N) = \mathcal{C} \, \exp\left\{ - \lambda \sum_j \sigma_j \right\} = \mathcal{C} \prod_{j=1}^{N} e^{-\lambda \, \sigma_j} \; , \tag{1.46}$$

where $\mathcal{C} \equiv e^{-(1+\lambda_0)}$ and $\lambda \equiv \lambda_2$. Normalization then requires

$$\mathsf{Tr} \; P \equiv \sum_{\{\sigma_j\}} = \mathcal{C} \left(e^{\lambda} + e^{-\lambda} \right)^N \; , \tag{1.47}$$

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

$$P(\sigma_1, \ldots, \sigma_N) = \prod_{j=1}^{N} \frac{e^{-\lambda \sigma_j}}{e^{\lambda} + e^{-\lambda}} = \prod_{j=1}^{N} \left(p \, \delta_{\sigma_j, 1} + q \, \delta_{\sigma_j, -1} \right) , \tag{1.48}$$

where

$$p = \frac{e^{-\lambda}}{e^{\lambda} + e^{-\lambda}} \quad , \quad q = 1 - p = \frac{e^{\lambda}}{e^{\lambda} + e^{-\lambda}} \; . \tag{1.49}$$

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

1.3.2 Continuous probability distributions

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

$$X^a = \int\limits_{\Omega} d\mu \, X^a(\boldsymbol{\varphi}) \, P(\boldsymbol{\varphi}) \; , \tag{1.50}$$

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

$$S^*\big[P(\boldsymbol{\varphi}), \{\lambda_a\}\big] = \int\limits_\Omega d\mu\, P(\boldsymbol{\varphi}) \ln P(\boldsymbol{\varphi}) - \sum_{a=0}^{K} \lambda_a \left(\int\limits_\Omega d\mu\, P(\boldsymbol{\varphi})\, X^a(\boldsymbol{\varphi}) - X^a \right) \tag{1.51}$$

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:

$$\ln P(\boldsymbol{\varphi}) = -1 - \sum_{a=0}^{K} \lambda_a\, X^a(\boldsymbol{\varphi}) \ . \tag{1.52}$$

The $K+1$ Lagrange multipliers $\{\lambda_a\}$ are then determined from the $K+1$ constraint equations in eqn. 1.50.

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

$$\int\limits_{-\infty}^{\infty} dx\, P(x) = 1 \quad , \quad \int\limits_{-\infty}^{\infty} dx\, x\, P(x) = \mu \quad , \quad \int\limits_{-\infty}^{\infty} dx\, x^2\, P(x) = \mu^2 + \sigma^2 \ . \tag{1.53}$$

Extremizing the entropy, we then obtain

$$P(x) = \mathcal{C}\, e^{-\lambda_1 x - \lambda_2 x^2} \ , \tag{1.54}$$

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

$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-(x-\mu)^2/2\sigma^2} \ . \tag{1.55}$$

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

1.4 General Aspects of Probability Distributions

1.4.1 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

$$\sum_{n \in \mathcal{C}} P_n = 1 \ . \tag{1.56}$$

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

$$\langle A \rangle = \sum_{n \in \mathcal{C}} P_n\, A_n \ . \tag{1.57}$$

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. 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

$$d\mu = W(\varphi_1, \ldots, \varphi_n)\, d\varphi_1\, d\varphi_2 \cdots d\varphi_n \ . \tag{1.58}$$

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

$$d\mu = \mathcal{C} \prod_{\sigma=1}^{r} dq_\sigma\, dp_\sigma \ , \tag{1.59}$$

where $\mathcal{C}$ is a constant we shall discuss later on below[5].

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

$$\int\limits_{\Omega} d\mu\, P(\boldsymbol{\varphi}) = 1 \ . \tag{1.60}$$

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

$$\langle A \rangle = \int\limits_{\Omega} d\mu\, P(\boldsymbol{\varphi})\, A(\boldsymbol{\varphi}) \ . \tag{1.61}$$

For example, consider the Gaussian distribution

$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-(x-\mu)^2/2\sigma^2} \ . \tag{1.62}$$

From the result[6]

$$\int\limits_{-\infty}^{\infty} dx\, e^{-\alpha x^2}\, e^{-\beta x} = \sqrt{\frac{\pi}{\alpha}}\, e^{\beta^2/4\alpha} \ , \tag{1.63}$$

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

$$\begin{aligned} \langle x \rangle &= \mu \\ \langle x^2 \rangle - \langle x \rangle^2 &= \sigma^2 \ . \end{aligned} \tag{1.64}$$

We call μ the *mean* and σ the *standard deviation* of the distribution, eqn. 1.62.

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 infinitesimal – 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 $\frac{1}{10}$.

[5]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*. See the discussion in appendix II of chapter 4.

[6]Memorize this!

If x is distributed according to $P_1(x)$, then the probability distribution on the product space $(x_1\ ,\ x_2)$ is simply the product of the distributions:

$$P_2(x_1, x_2) = P_1(x_1)\, P_1(x_2)\ . \tag{1.65}$$

Suppose we have a function $\phi(x_1, \ldots, x_N)$. How is it distributed? Let $Q(\phi)$ be the distribution for ϕ. We then have

$$\begin{aligned}\mathcal{P}(\phi) &= \int\limits_{-\infty}^{\infty} dx_1 \cdots \int\limits_{-\infty}^{\infty} dx_N\, P_N(x_1, \ldots, x_N)\, \delta\Big(\phi(x_1, \ldots, x_N) - \phi\Big) \\ &= \int\limits_{-\infty}^{\infty} dx_1 \cdots \int\limits_{-\infty}^{\infty} x_N\, P_1(x_1) \cdots P_1(x_N)\, \delta\Big(\phi(x_1, \ldots, x_N) - \phi\Big)\ ,\end{aligned} \tag{1.66}$$

where the second line is appropriate if the $\{x_j\}$ are themselves distributed independently. Note that

$$\int\limits_{-\infty}^{\infty} d\phi\, \mathcal{P}(\phi) = 1\ , \tag{1.67}$$

so $\mathcal{P}(\phi)$ is itself normalized.

1.4.2 Central limit theorem

In particular, consider the distribution function of the sum

$$X = \sum_{i=1}^{N} x_i\ . \tag{1.68}$$

We will be particularly interested in the case where N is large. For general N, though, we have

$$\mathcal{P}_N(X) = \int\limits_{-\infty}^{\infty} dx_1 \cdots \int\limits_{-\infty}^{\infty} dx_N\, P_1(x_1) \cdots P_1(x_N)\, \delta\big(x_1 + x_2 + \ldots + x_N - X\big)\ . \tag{1.69}$$

It is convenient to compute the Fourier transform[7] of $\mathcal{P}(X)$:

$$\begin{aligned}\hat{\mathcal{P}}_N(k) &= \int\limits_{-\infty}^{\infty} dX\, \mathcal{P}_N(X)\, e^{-ikX} \\ &= \int\limits_{-\infty}^{\infty} dX \int\limits_{-\infty}^{\infty} dx_1 \cdots \int\limits_{-\infty}^{\infty} x_N\, P_1(x_1) \cdots P_1(x_N)\, \delta\big(x_1 + \ldots + x_N - X\big)\, e^{-ikX} \\ &= \big[\hat{P}_1(k)\big]^N\,,\end{aligned} \tag{1.70}$$

where

$$\hat{P}_1(k) = \int\limits_{-\infty}^{\infty} dx\, P_1(x)\, e^{-ikx} \tag{1.71}$$

is the Fourier transform of the single variable distribution $P_1(x)$. The distribution $\mathcal{P}_N(X)$ 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 $\hat{P}_1(k)$

$$\begin{aligned}\hat{P}_1(k) &= \int\limits_{-\infty}^{\infty} dx\, P_1(x)\, \big(1 - ikx - \tfrac{1}{2}\, k^2 x^2 + \tfrac{1}{6}\, i\, k^3\, x^3 + \ldots\big) \\ &= 1 - ik\langle x\rangle - \tfrac{1}{2}\, k^2 \langle x^2\rangle + \tfrac{1}{6}\, i\, k^3 \langle x^3\rangle + \ldots\ .\end{aligned} \tag{1.72}$$

Thus,

$$\ln \hat{P}_1(k) = -i\mu k - \tfrac{1}{2}\sigma^2 k^2 + \tfrac{1}{6}\, i\, \gamma^3\, k^3 + \ldots\ , \tag{1.73}$$

where

$$\begin{aligned}\mu &= \langle x\rangle \\ \sigma^2 &= \langle x^2\rangle - \langle x\rangle^2 \\ \gamma^3 &= \langle x^3\rangle - 3\,\langle x^2\rangle\,\langle x\rangle + 2\,\langle x\rangle^3\end{aligned} \tag{1.74}$$

We can now write

$$\big[\hat{P}_1(k)\big]^N = e^{-iN\mu k}\, e^{-N\sigma^2 k^2/2}\, e^{iN\gamma^3 k^3/6} \cdots \tag{1.75}$$

[7] 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. Source: http://www.robertnowlan.com/pdfs/Fourier,%20Joseph.pdf

Now for the inverse transform. In computing $\mathcal{P}_N(X)$, 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

$$\begin{aligned}\mathcal{P}_N(X) &= \int\limits_{-\infty}^{\infty}\frac{dk}{2\pi}\,e^{ik(X-N\mu)}\,e^{-N\sigma^2k^2/2}\left\{1+\tfrac{1}{6}\,i\,N\gamma^3k^3+\ldots\right\}\\ &= \left(1-\tfrac{1}{6}N\gamma^3\,\frac{\partial^3}{\partial X^3}+\ldots\right)\frac{1}{\sqrt{2\pi N\sigma^2}}\,e^{-(X-N\mu)^2/2N\sigma^2}\\ &= \frac{1}{\sqrt{2\pi N\sigma^2}}\,e^{-(X-N\mu)^2/2N\sigma^2}\quad(N\to\infty)\,.\end{aligned}\tag{1.76}$$

In going from the second line to the third, we have written $X=\sqrt{N}\,\xi$, in which case $N\frac{\partial^3}{\partial X^3}=N^{-1/2}\,\frac{\partial^3}{\partial\xi^3}$, which gives a subleading contribution which vanishes in the $N\to\infty$ limit. We have just proven the *central limit theorem*: in the limit $N\to\infty$, the distribution of a sum of N independent random variables x_i is 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 moment 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.

1.4.3 Multidimensional Gaussian integral

Consider the multivariable Gaussian distribution,

$$P(\boldsymbol{x})\equiv\left(\frac{\det A}{(2\pi)^n}\right)^{1/2}\exp\Big(-\tfrac{1}{2}\,x_i\,A_{ij}\,x_j\Big)\,,\tag{1.77}$$

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

$$Z(\boldsymbol{b})=\left(\frac{\det A}{(2\pi)^n}\right)^{1/2}\int\limits_{-\infty}^{\infty}dx_1\cdots\int\limits_{-\infty}^{\infty}dx_n\;\exp\Big(-\tfrac{1}{2}\,x_i\,A_{ij}\,x_j+b_i\,x_i\Big)=\exp\Big(\tfrac{1}{2}\,b_i\,A^{-1}_{ij}\,b_j\Big)\,.\tag{1.78}$$

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

$$\begin{aligned}\langle\,x_{j_1}\cdots x_{j_{2k}}\,\rangle &= \int d^nx\;P(\boldsymbol{x})\,x_{j_1}\cdots\,x_{j_{2k}}\\ &= \frac{\partial^nZ(\boldsymbol{b})}{\partial b_{j_1}\cdots\,\partial b_{j_{2k}}}\bigg|_{\boldsymbol{b}=0}\\ &= \sum_{\text{contractions}}A^{-1}_{j_{\sigma(1)}j_{\sigma(2)}}\cdots A^{-1}_{j_{\sigma(2k-1)}j_{\sigma(2k)}}\,.\end{aligned}\tag{1.79}$$

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 $C_{2k}=(2k)!/2^kk!$ possible such contractions. To obtain this result for C_k, 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

$$C_{2k}=(2k-1)\cdot(2k-3)\cdots3\cdot1=\frac{(2k)!}{2^kk!}\,.\tag{1.80}$$

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

$$\langle\, x_{j_1} x_{j_2} x_{j_3} x_{j_4} \,\rangle = A^{-1}_{j_1 j_2} A^{-1}_{j_3 j_4} + A^{-1}_{j_1 j_3} A^{-1}_{j_2 j_4} + A^{-1}_{j_1 j_4} A^{-1}_{j_2 j_3} \ . \tag{1.81}$$

1.5 Appendix : Bayesian Statistics

Let the probability of a discrete event A be $P(A)$. We now introduce two additional probabilities. The *joint probability* for events A and B together is written $P(A \cap B)$. The *conditional probability* of B given A is $P(B|A)$. We can compute the joint probability $P(A \cap B) = P(B \cap A)$ in two ways:

$$P(A \cap B) = P(A|B) \cdot P(B) = P(B|A) \cdot P(A) \ . \tag{1.82}$$

Thus,

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \ , \tag{1.83}$$

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

$$P(B) = \sum_i P(B|A_i) \cdot P(A_i) \ . \tag{1.84}$$

We then have

$$P(A_i|B) = \frac{P(B|A_i) \cdot P(A_i)}{\sum_j P(B|A_j) \cdot P(A_j)} \ , \tag{1.85}$$

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

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B|A) \cdot P(A) + P(B|\neg A) \cdot P(\neg A)} \ . \tag{1.86}$$

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[8]. 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 $P(A|B)$. Applying eqn. 1.86 with

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

we find $P(A|B) = \frac{1}{2}$. 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.

[8]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. In our example in the text, the sensitivity and specificity are both 0.9999.

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

$$P(y) = \int dx\, P(y|x) \cdot P(x) \tag{1.87}$$

and

$$P(x|y) = \frac{P(y|x) \cdot P(x)}{\int dx'\, P(y|x') \cdot P(x')}\ . \tag{1.88}$$

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 $P(A_i|B)$ 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 obtaining the prior probabilities $P(A_i)$ is to obtain them from a maximum entropy construction.