2.1: Discrete Probability Theory

Discrete Random Variables

Consider a trial $\mathcal{T}$ where the observation is a measurement of the z component $\hbar m_S$ of spin angular momentum of a spin $S = 5/2$. There are just six possible outcomes (events) that can be labeled with the magnetic spin quantum number m_S or indexed by integer numbers 1, 2, ... 6. In general, the probabilities of the six possible events will differ from each other. They will depend on preparation and may depend on evolution time before the observation. To describe such situations, we define a set of elementary events

$$A = \{a_j\} \ , \tag{2.1.1}$$

where in our example index j runs from 1 to 6, whereas in general it runs from 1 to the number N_A of possible events. Each of the events is assigned a probability $0 \leq P(a_j) \leq 1$. Impossible events (for a given preparation) have probability zero and a certain event has probability 1. Since one and only one of the events must happen in each trial, the probabilities are normalized, $\sum_j^{N_A} P(a_j) = 1$. A simplified model of our example trial is the rolling of a die. If the die is fair, we have the special situation of a uniform probability distribution, i.e., $P(a_j) = 1/6$ for all j.

A set of random events with their associated probabilities is called a *random variable*. If the number of random events is countable, the random variable is called *discrete*. In a computer, numbers can be assigned to the events, which makes the random variable a *random number*. A series of trials can then be simulated by generating a series of $\mathcal{N}$ pseudo-random numbers that assign the events observed in the $\mathcal{N}$ trials. Such simulations are called *Monte Carlo* simulations. Pseudo-random numbers obtained from a computer function need to be adjusted so that they reproduce the given or assumed probabilities of the events. [concept:random_variable]

Using the Matlab function `rand` , which provides uniformly distributed random numbers in the open interval $(0, 1)$, write a program that simulates throwing a die with six faces. The outer function should have trial number $\mathcal{N}$ as an input and a vector of the numbers of encountered ones, twos, ... and sixes as an output. It should be based on an inner function that simulates a single throw of the die. Test the program by determining the difference from the expectation $P(a_j) = 1/6$ for ever larger numbers of trials.

Multiple Discrete Random Variables

For two sets of events A and B and their probabilities, we define a *joint probability* $P(a_j, b_k)$ that is the probability of observing both a_j and b_k in the same trial. An example is the throwing of two dice, one black and one red, and asking about the probability that the black die shows a 2 and the red die a 3. A slightly more complicated example is the measurement of the individual z components of spin angular momentum of two coupled spins $S_{\mathrm{A}} = 5/2$ and $S_{\mathrm{B}} = 5/2$. Like individual probabilities, joint probabilities fall in the closed interval $[0, 1]$. Joint probabilities are normalized,

$$\sum_a \sum_b P(a, b) = 1 \ . \tag{2.1.2}$$

Note that we have introduced a brief notation that suppresses indices j and k. This notation is often encountered because of its convenience in writing.

If we know the probabilities $P(a, b)$ for all $N_A \cdot N_B$ possible combinations of the two events, we can compute the probability of a single event, for instance a,

$$P_A(a) = \sum_b P(a, b) \ , \tag{2.1.3}$$

where $P_A(a)$ is the *marginal probability* of event a.

The unfortunate term 'marginal' does not imply a small probability. Historically, these probabilities were calculated in the margins of probability tables .

Another quantity of interest is the *conditional probability* $P(a|b)$ of an event a, provided that b has happened. For instance, if we call two cards from a full deck, the probability of the second card being a Queen is conditional on the first card having been a Queen. With the definition for the conditional probability we have

$$P(a, b) = P(a|b)P_B(b) \tag{2.1.4}$$

$$= P(b|a)P_A(a) \ . \tag{2.1.5}$$

Theorem 2.1.1: Bayes' theorem

If the marginal probability of event b is not zero, the conditional probability of event a given b is

$$P(a|b) = \frac{P(b|a)P_A(a)}{P_B(b)} \; . \tag{2.1.6}$$

Bayes' theorem is the basis of Bayesian inference, where the probability of proposition a is sought given prior knowledge (short: the prior) b. Often Bayesian probability is interpreted subjectively, i.e., different persons, because they have different prior knowledge b, will come to different assessments for the probability of proposition a. This interpretation is incompatible with theoretical physics, where, quite successfully, an objective reality is assumed. Bayesian probability theory can also be applied with an objective interpretation in mind and is nowadays used, among else, in structural modeling of biomacromolecules to assess agreement of a model (the proposition) with experimental data (the prior).

In experimental physics, biophysics, and physical chemistry, Bayes' theorem can be used to assign experimentally informed probabilities to different models for reality. For example assume that a theoretical modeling approach, for instance an MD simulation, has provided a set of conformations $A = \{a_j\}$ of a protein molecule and associated probabilities $P_A(a_j)$. The probabilities are related, *via* the Boltzmann distribution, to the free energies of the conformations (this point is discussed later in the lecture course). We further assume that we have a measurement B with output b_k and we know the marginal probability $P_B(b)$ of encountering this output for a random set of conformations of the protein molecule. Then we need only a physical model that provides the conditional probabilities $P(b_k|a_j)$ of measuring b_k given the conformations a_j and can compute the probability $P(a_j|b_k)$ that the true conformation is a_j, given the result of our measurement, via Bayes' theorem. Equation 2.1.6). This procedure can be generalized to multiple measurements. The required $P(b_k|a_j)$ depend on measurement errors. The approach allows for combining possibly conflicting modeling and experimental results to arrive at a 'best estimate' for the distribution of conformations.

The events associated with two random variables can occur completely independent of each other. This is the case for throwing two dice: the number shown on the black die does not depend on the number shown on the red die. Hence, the probability to observe a 2 on the black and a 3 on the red die is $(1/6) \cdot (1/6) = 1/36$. In general, joint probabilities of *independent* events factorize into the individual (or marginal) probabilities, which leads to huge simplifications in computations. In the example of two coupled spins $S_{\mathrm{A}} = 5/2$ and $S_{\mathrm{B}} = 5/2$ the two random variables $m_{S,\mathrm{A}}$ and $m_{S,\mathrm{B}}$ may or may not be independent. This is decided by the strength of the coupling, the preparation of trial $\mathcal{T}$, and the evolution time t before observation.

If two random variables are *independent*, the joint probability of two associated events is the product of the two marginal probabilities,

$$P(a,b) = P_A(a)P_B(b) \; . \tag{2.1.7}$$

As a consequence, the conditional probability $P(a|b)$ equals the marginal probability of a (and *vice versa*),

$$P(a|b) = P_A(a) \; . \tag{2.1.8}$$

[concept:independent_variables]

For a set of more than two random variables two degrees of independence can be established, a weak type of *pairwise independence* and a strong type of *mutual independence*. The set is mutually independent if the marginal probability distribution in any subset, i.e. the set of marginal probabilities for all event combinations in this subset, is given by the product of the corresponding marginal distributions for the individual events.[2] This corresponds to complete independence. Weaker pairwise independence implies that the marginal distributions for any pair of random variables are given by the product of the two corresponding distributions. Note that even weaker independence can exist within the set, but not throughout the set. Some, but not all pairs or subsets of random variables can exhibit independence.

Another important concept for multiple random variables is whether or not they are distinguishable. In the example above we used a black and a red die to specify our events. If both dice would be black, the event combinations (a_2, b_3) and (a_3, b_2) would be indistinguishable and the corresponding composite event of observing a 2 and a 3 would have a probability of $1/18$, i.e. the product of the probability $1/36$ of the basic composite event with its multiplicity 2. In general, if n random variables are indistinguishable, the multiplicity equals the number of permutations of the n variables, which is $n! = 1 \cdot 2 \cdots (n-1) \cdot n$.

Functions of Discrete Random Variables

We consider an event g that depends on two other events a and b. For example, we ask for the probability that the sum of the numbers shown by the black and red die is g, where g can range from 2 to 12, given that we know the probabilities $P(a,b)$, which in our example all have the value 1/36. In general, the probability distribution of random variable G can be computed by

$$P_G(g) = \sum_a \sum_b \delta_{g,G(a,b)} P(a,b) \; , \tag{2.1.9}$$

where $G(a,b)$ is an arbitrary function of a and b and the Kronecker delta $\delta_{g,G(a,b)}$ assumes the value one if $g = G(a,b)$ and zero otherwise. In our example, $g = G(a,b) = a+b$ will assume the value of 5 for the event combinations $(1,4), (2,3), (3,2), (4,1)$ and no others. Hence, $P_G(5) = 4/36 = 1/9$. There is only a single combination for $g = 2$, hence $P_G(2) = 1/36$, and there are 6 combinations for $g = 7$, hence $P_G(7) = 1/6$. Although the probability distributions for the individual random numbers A and B are uniform, the one for G is not. It peaks at the value of $g = 7$ that has the most realizations. Such peaking of probability distributions that depend on multiple random variables occurs very frequently in statistical mechanics. The peaks tend to become the sharper the larger the number of random variables that contribute to the sum. If this number N tends to infinity, the distribution of the sum g is so sharp that the distribution width (to be specified below) is smaller than the error in the measurement of the mean value g/N (see Section [section:prob_dist_sum]). This effect is the very essence of statistical thermodynamics: Although quantities for a single molecule may be broadly distributed and unpredictable, the mean value for a large number of molecules, let's say 10^{18} of them, is very well defined and perfectly predictable.

In a numerical computer program, Equation 2.1.9) for only two random variables can be implemented very easily by a loop over all possible values of g with inner loops over all possible values of a and b. Inside the innermost loop, $G(a,b)$ is computed and compared to loop index g to add or not add $P(a,b)$ to the bin corresponding to value g. Note however that such an approach does not carry to large numbers of random variables, as the number of nested loops increases with the number of random variables and computation time thus increases exponentially. Analytical computations are simplified by the fact that $\delta_{g,G(a,b)}$ usually deviates from zero only within certain ranges of the summation indexes j (for a) and k (for b). The trick is then to find the proper combinations of index ranges.

Compute the probability distribution for the sum g of the numbers shown by two dice in two ways. First, write a computer program using the approach sketched above. Second, compute the probability distribution analytically by making use of the uniform distribution for the individual events ($P(a,b) = 1/36$ for all a, b. For this, consider index ranges that lead to a given value of the sum g.[3]

Discrete Probability Distributions

In most cases random variables are compared by considering the *mean values* and widths of their probability distributions. As a measure of the width, the *standard deviation* σ of the values from the mean value is used, which is the square root of the *variance* σ^2. The concept can be generalized by considering functions $F(A)$ of the random variable. In the following expressions, $F(A) = A$ provides the mean value and standard deviation of the original random variable A.

Theorem 2.1.1: Mean value and standard deviation

For any function $F(A)$ of a random variable A, the mean value $\langle F \rangle$ is given by,

$$\langle F \rangle = \sum_a F(a) P_A(a) \; . \tag{2.1.10}$$

The standard deviation, which characterizes the width of the distribution of the function values $F(a)$, is given by,

$$\sigma = \sqrt{\sum_a \left(F(a) - \langle F \rangle \right)^2 P_A(a)} \; . \tag{2.1.11}$$

The mean value is the first moment of the distribution, with the n^{th} moment being defined by

$$\langle F^n \rangle = \sum_a F^n(a) P_A(a) \; . \tag{2.1.12}$$

The n^{th} central moment is

$$\langle (F - \langle F \rangle)^n \rangle = \sum_a (F(a) - \langle F \rangle)^n P_A(a) \, . \tag{2.1.13}$$

For the variance, which is the second central moment, we have

$$\sigma^2 = \langle F^2 \rangle - \langle F \rangle^2 \, . \tag{2.1.14}$$

Assume that we know the mean values for functions $F(A)$ and $G(B)$ of two random variables as well as the mean value $\langle FG \rangle$ of their product, which we can compute if the joint probability function $P(a, b)$ is known. We can then compute a *correlation function*

$$R_{FG} = \langle FG \rangle - \langle F \rangle \langle G \rangle \, , \tag{2.1.15}$$

which takes the value of zero, if F and G are independent random numbers.

Exercise 2.1.1

Compute the probability distribution for the normalized sum g/M of the numbers obtained on throwing M dice in a single trial. Start with $M = 1$ and proceed via $M = 10, 100, 1000$ to $M = 10000$. Find out how many Monte Carlo trials $\mathcal{N}$ you need to guess the converged distribution. What is the mean value $\langle g/M \rangle$? What is the standard deviation σ_g? How do they depend on $\mathcal{N}$?

Probability Distribution of a Sum of Random Numbers

If we associate the random numbers with N molecules, identical or otherwise, we will often need to compute the sum over all molecules. This generates a new random number

$$S = \sum_{j=1}^{N} F_j \, , \tag{2.1.16}$$

whose mean value is the sum of the individual mean values,

$$\langle S \rangle = \sum_{j=1}^{N} \langle F_j \rangle \, . \tag{2.1.17}$$

If motion of the individual molecules is uncorrelated, the individual random numbers F_j are independent. It can then be shown that the variances add ,

$$\sigma_S^2 = \sum_{j=1}^{N} \sigma_j^2 \tag{2.1.18}$$

For identical molecules, all random numbers have the same mean $\langle F \rangle$ and variance σ_F^2 and we find

$$\langle S \rangle = N \langle F \rangle \tag{2.1.19}$$

$$\sigma_S^2 = N \sigma_F^2 \tag{2.1.20}$$

$$\sigma_S = \sqrt{N} \sigma_F \, . \tag{2.1.21}$$

This result relates to the concept of *peaking* of probability distributions for a large number of molecules that was introduced above on the example of the probability distribution for sum of the numbers shown by two dice. The width of the distribution normalized to its mean value,

$$\frac{\sigma_S}{\langle S \rangle} = \frac{1}{\sqrt{N}} \frac{\sigma_F}{\langle F \rangle} \, , \tag{2.1.22}$$

scales with the inverse square root of N. For 10^{18} molecules, this *relative width* of the distribution is one billion times smaller than for a single molecule. Assume that for a certain physical quantity of a single molecule the standard deviation is as large as the mean value. No useful prediction can be made. For a macroscopic sample, the same quantity can be predicted with an accuracy better than the precision that can be expected in a measurement.

Binomial Distribution

We consider the measurement of the z component of spin angular momentum for an ensemble of N spins $S = 1/2$.[4] The random number associated with an individual spin can take only two values, $-\hbar/2$ or $+\hbar/2$. Additive and multiplicative constants can be taken care of separately and we can thus represent each spin by a random number A that assumes the value $a = 1$ (for $m_S = +1/2$) with probability P and, accordingly, the value $a = 0$ (for $m_S = -1/2$) with probability $1 - P$. This is a very general problem, which also relates to the second postulate of Penrose (see Section [Penrose_postulates]). A simplified version with $P = 1 - P = 0.5$ is given by N flips of a fair coin. A fair coin or a biased coin with $P \neq 0.5$ can be easily implemented in a computer, for instance by using `a = floor(rand+P)` in Matlab. For the individual random numbers we find $\langle A \rangle = P$ and $\sigma_A^2 = P(1-P)$, so that the relative standard deviation for the ensemble with N members becomes $\sigma_S / \langle S \rangle = \sqrt{(1-P)/(N \cdot P)}$.[5]

To compute the explicit probability distribution of the sum of the random numbers for the whole ensemble, we realize that the probability of a subset of n ensemble members providing a 1 and $N - n$ ensemble members providing a 0 is $P^n (1-P)^{N-n}$. The value of the sum associated with this probability is n.

Now we still need to consider the phenomenon already encountered for the sum of the numbers on the black and red dice: Different numbers n have different multiplicities. We have $N!$ permutations of the ensemble members. Let us assign a 1 to the first n members of each permutation. For our problem, it does not matter in which sequence these n members are numbered and it does not matter in which sequence the remaining $N - n$ members are numbered. Hence, we need to divide the total number of permutations $N!$ by the numbers of permutations in each subset, $n!$ and $(N-n)!$ for the first and second subset, respectively. The multiplicity that we need is the number of combinations of N elements to the n^{th} class, which is thus given by the binomial coefficient,

$$\binom{N}{n} = \frac{N!}{n!(N-n)!} , \tag{2.1.23}$$

providing the probability distribution

$$P_S(n) = \binom{N}{n} P^n (1-P)^{N-n} . \tag{2.1.24}$$

For large values of N the binomial distribution tends to a Gaussian distribution,

$$G(s) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(s - \langle s \rangle)^2}{2\sigma^2} \right] . \tag{2.1.25}$$

As we already know the mean value $\langle s \rangle = \langle n \rangle = NP$ and variance $\sigma_S^2 = NP(1-P)$, we can immediately write down the approximation

$$P_S(n) \approx \frac{1}{\sqrt{2\pi P(1-P)N}} \exp\left[-\frac{(n - PN)^2}{2P(1-P)N} \right] = G(n) . \tag{2.1.26}$$

As shown in Figure 2.1.1 the Gaussian approximation of the binomial distribution is quite good already at $N = 1000$.

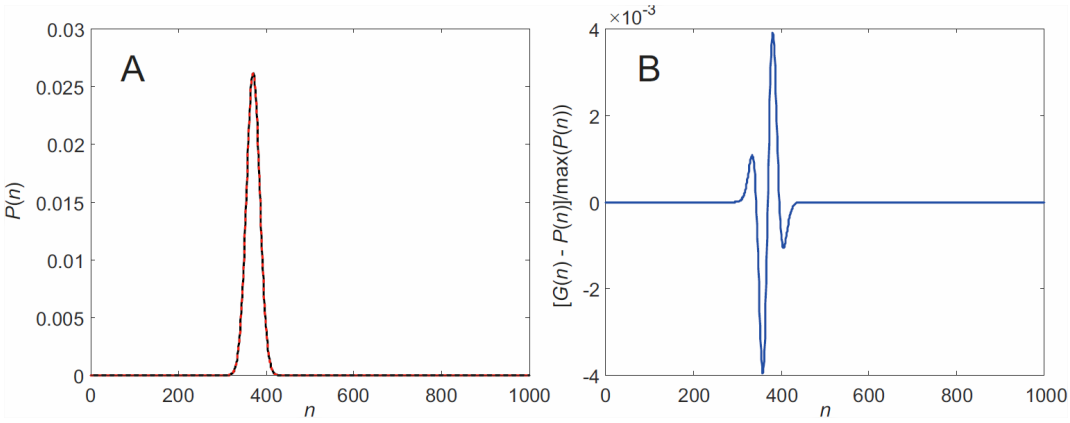

Figure 2.1.1: Gaussian approximation of the binomial distribution. (A) Gaussian approximation (red dashed line) and binomial distribution (black solid line) for $P = 0.37$ and $N = 1000$. (B) Error of the Gaussian approximation relative to the maximum value of the binomial distribution.

In fact, the Gaussian (or normal) distribution is a general distribution for the arithmetic mean of a large number of independent random variables:

Suppose that a large number N of observations has been made with each observation corresponding to a random number that is independent from the random numbers of the other observations. According to the *central limit theorem*, the mean value $\langle S \rangle / N$ of the sum of all these random numbers is approximately normally distributed, regardless of the probability distribution of the individual random numbers, as long all the probability distributions of all individual random numbers are identical.[6] The central limit theorem applies, if each individual random variable has a well-defined mean value (expectation value) and a well-defined variance. These conditions are fulfilled for statistically regular trials $\mathcal{T}$. [concept:central_limit_theorem]

Stirling's Formula

The number $N!$ of permutations increases very fast with N, leading to numerical overflow in calculators and computers at values of N that correspond to nanoclusters rather than to macroscopic samples. Even binomial coefficients, which grow less strongly with increasing ensemble size, cannot be computed with reasonable precision for $N \gg 1000$. Furthermore, the factorial $N!$ is difficult to handle in calculus. The scaling problem can be solved by taking the logarithm of the factorial,

$$\ln N! = \ln\left(\prod_{n=1}^{N} n\right) = \sum_{n=1}^{N} \ln n \, . \tag{2.1.27}$$

For large numbers N the natural logarithm of the factorial can be approximated by *Stirling's formula*

$$\ln N! \approx N \ln N - N + 1 \, , \tag{2.1.28}$$

which amounts to the approximation

$$N! \approx N^N \exp(1 - N) \tag{2.1.29}$$

for the factorial itself. For large numbers N it is further possible to neglect 1 in the sum and approximate $\ln N! \approx N \ln N - N$.

The absolute error of this approximation for $N!$ looks gross and increases fast with increasing N, but because $N!$ grows much faster, the relative error becomes insignificant already at moderate N. For $\ln N!$ it is closely approximated by $-0.55/N$. In fact, an even better approximation has been found by Gosper ,

$$\ln N! \approx N \ln N - N + \frac{1}{2} \ln\left[\left(2N + \frac{1}{3}\right)\pi\right] \, . \tag{2.1.30}$$

Gosper's approximation is useful for considering moderately sized systems, but note that several of our other assumptions and approximations become questionable for such systems and much care needs to be taken in interpreting results. For the macroscopic systems, in which we are mainly interested here, Stirling's formula is often sufficiently precise and Gosper's is not needed.

Slightly better than Stirling's original formula, but still a simple approximation is

$$N! \approx \sqrt{2\pi N}\left(\frac{N}{e}\right)^N . \tag{2.1.31}$$