AEP 4230: Probability

Contents

I. Definitions 2

II. Functions of one random variable 4

III. Important probability distributions 7

A. Gaussian distribution 7

B. Binomial distribution 9

C. Poisson distribution 10

IV. Multivariate probability distributions 11

V. Multivariate Gaussian distribution 13

VI. Central limit theorem 15

A. Stirling's approximation 17

VII. Information, entropy, and estimation 17

Now that we have covered thermodynamics, which tells us about transformations of state independent of any microscopic description, we now wish to predict these thermodynamic properties and understand them in terms of microscopic laws of mechanics and quantum mechanics. This is the subject of *statistical mechanics*. It will turn out that in statistical mechanics, a central quantity will be the *probability* that a system is in a given *microstate*. From these microstate probabilities, we can work out the most likely *macrostate* of a system, connecting to thermodynamics.

Thus, we devote one unit of the course to discussing some important elements of probability.

I. DEFINITIONS

We are interested in describing the *outcomes* $S \equiv \{x_1, x_2, \dots\}$ of a *random variable* x. These outcomes can be *discrete* in the case of a dice $\{1, 2, 3, 4, 5, 6\}$ where x is the number of dots on the side of the dice facing up. The outcomes can also be

continuous, such as the location of a molecule of gas in a container $\{\mathbf{r}\}$ where $\mathbf{r}$ is any position in the container. An *event* is any subset of the outcomes $E \subset S$, and is assigned a *probability*. For example, the probability of the dice roll yielding an odd value is $p(\{1, 3, 5\}) = 1/2$.

We take as fundamental axioms of probability the following:

1. *Positivity*: $p(E) \geq 0$.
2. *Additivity*: For disjoint or disconnected events, $p(A \text{or} B) = p(A) + p(B)$.
3. *Normalization*: The probability of any event in the set of outcomes is $p(S) = 1$.

There are two main ways to *assign* probabilities: based on objective or subjective probabilities.

Objective probabilities are based on direct experimental measurement of the frequency of different outcomes. Suppose we observe N outcomes of some process, and we find that outcome A happens N_A times. Then we would say that the objective probability of A occurring is $p(A) = \lim_{N\to\infty} N_A/N$. The limit must be taken because *fluctuations*, which are larger when event numbers are smaller, can yield fractions that are very different from their probabilities. For example, if we flip a coin two times, you would not be surprised to see heads appear twice. In fact the probability of such for an unbiased coin is $1/4$. If we use the objective probability formula, we find that the probability of flipping heads is 1, which is obviously wrong. If we flip a coin 1000 times instead, the number of heads you get will be very close to 500. We will make this more quantitative when we discuss the central limit theorem.

Subjective probabilities are probabilities derived by theoretical estimation or argumentation. For example, if we have a coin, and we *think* is unbiased, we would say that there is no way for the coin to prefer an outcome, and so we assign probability $1/2$ to the two outcomes (heads, tails). Such subjective probabilities have to be checked against experiments. In statistical mechanics, we will use similar ideas of *a priori equal probabilities*, not unlike our fair coin.

II. FUNCTIONS OF ONE RANDOM VARIABLE

Consider a random variable x whose outcomes are any real number $-\infty < x < \infty$. We define the following functions:

1. The *probability density function* $p(x)$ is the function which tells us the probability of finding x to be in an infinitesimal window $[x, x + dx)$. In particular $P(x \in [x, x + dx)) = p(x)dx$. The positivity condition on probabilities tells us that $p(x) \geq 0$ for any x. The normalization condition for probabilities tells us that $\int_{-\infty}^{\infty} dx p(x) = 1$.

2. The *cumulative probability function* $P(X)$ is the probability that $x < X$. The additivity condition for probabilities tells us that $P(X) = \int_{-\infty}^{X} dx p(x)$. To see this, note that if we chop up the real line into segments of length dx (with left endpoint x) that these subsets of the real line are disjoint from each other and so the probabilities add.

3. The *expectation value* or *average value* of any function $f(x)$ is given by

$$\langle f \rangle = \int dx\, f(x)p(x), \tag{1}$$

where if I do not write the limits of integration, it is intended to span the entire domain of event space (here the real line).

We can consider the function f itself to be a random variable and write a probability density for values of f. To arrive at such a probability density, we need only evaluate the probability that we observe a value f_0. That is simply the sum probability of finding any x such that $f(x) = f_0$. We may express that constraint as follows:

$$p(f_0) = \int dx\, \delta(f(x) - f_0)p(x) = \sum_k \frac{p(x_k)}{\left|\frac{df}{dx}\right|_{x_k}}, \tag{2}$$

where k labels over the number of solutions to the equation $f(x_k) = x_0$ (there may be more than one). This distribution is normalized and positive. Normalization can be checked as follows $\int df_0 p(f_0) = \int df_0 p(f_0) \int dx \delta(f(x) - f_0)p(x) = \int dx p(x) = 1$.

As an example, consider $p(x) = \lambda e^{-\lambda|x|}/2$ and consider $f(x) = x^2$. Then the probability of a particular value of $f = f_0$ is given from our formula as

$$p(f) = \frac{\lambda}{2\sqrt{f}} e^{-\lambda\sqrt{f}}. \tag{3}$$

Note that this is only for $f > 0$. I encourage you to check that this distribution is normalized.

4. *Moments* of the probability distribution are defined as expectation values of powers of x. The nth moment is given by $m_n = \langle x^n \rangle = \int dx\, x^n p(x)$.

5. The *characteristic function* $\tilde{p}(k)$ of the PDF is the Fourier transform of the probability distribution:

$$\tilde{p}(k) = \int dx e^{-ikx} p(x) = \langle e^{-ikx} \rangle = \sum_{n=0}^{\infty} \frac{(-ik)^n}{n!} \langle x^n \rangle. \tag{4}$$

As can be seen by the Taylor expansion of the exponential, the characteristic function encodes the moments of the distribution. The moments are in fact just Taylor expansion coefficients. The PDF itself is just given by the inverse Fourier transform of the characteristic function: $p(x) = \int \frac{dk}{2\pi} e^{ikx} \tilde{p}(k)$.

It is also useful to generate moments of the distribution around a reference point x_0, e.g., $\langle (x - x_0)^n \rangle$. You can show that these appear in the Taylor expansion of $e^{ikx_0}\tilde{p}(k)$.

6. The *cumulant generating function* is the logarithm of the characteristic function. We say that its expansion generates *cumulants* of the distribution. In particular, we define:

$$\ln \tilde{p}(k) = \sum_{n=1}^{\infty} \frac{(-ik)^n}{n!} \langle x^n \rangle_c. \tag{5}$$

We may relate the cumulants to moments via exponentiating this definition and comparing it with the Taylor expansion of the characteristic function as defined earlier:

$$\tilde{p}(k) = \exp\left[\sum_{n=1}^{\infty} \frac{(-ik)^n}{n!} \langle x^n \rangle_c \right] = \sum_{n=0}^{\infty} \frac{(-ik)^n}{n!} \langle x^n \rangle. \tag{6}$$

The exponent may be written in a Taylor series as

$$\exp\left[\sum_{n=1}^{\infty}\frac{(-ik)^n}{n!}\langle x^n\rangle_c\right] = 1+\sum_{n=1}^{\infty}\frac{1}{n!}\sum_{m_1\cdots m_n=1}^{\infty}\frac{(-ik)^{m_1+m_2\cdots+m_n}}{m_1!m_2!\cdots m_n!}\langle x^{m_1}\rangle_c\langle x^{m_2}\rangle_c\cdots\langle x^{m_n}\rangle_c. \tag{7}$$

Let us now compare terms order-by-order in matching Eq. (6) with Eq. (7). Let's start by matching the $(-ik)^1$ term in Eq. (6). In Eq. (7), the only way that is possible is if $n = 1$ because the minimum value of *any* m_i is 1. Therefore, we can write

$$\frac{(-ik)^1}{1!}\langle x^1\rangle = \frac{(-ik)^1}{1!}\langle x^1\rangle_c \implies \langle x\rangle = \langle x\rangle_c. \tag{8}$$

The second term in the characteristic function ($(-ik)^2\langle x^2\rangle/2!$) is also simple because the only terms that contribute are the $n = 1$ term and the $n = 2$ term. The $n = 1$ term gives $\frac{1}{1!}\frac{(-ik)^2}{2!}\langle x^2\rangle_c$, while the $n = 2$ term gives $\frac{1}{2!}\frac{(-ik)^{1+1}}{1!1!}\langle x\rangle_c^2$. Therefore, we have

$$\langle x^2\rangle = \langle x^2\rangle_c + \langle x\rangle^2 \implies \langle x^2\rangle_c = \langle x^2\rangle - \langle x\rangle^2 \equiv \sigma_x^2. \tag{9}$$

The second cumulant is just the variance of the distribution. Let us also evaluate the third moment: the $n = 1$ term gives $\frac{(-ik)^3}{3!}\langle x^3\rangle_c$, the $n = 2$ term gives $2\times\frac{1}{2!}\frac{(-ik)^{2+1}}{2!1!}\langle x^2\rangle_c\langle x\rangle$ (the 2 comes from the equal terms: $m_1 = 2, m_2 = 1$, $m_1 = 1, m_2 = 2$), and the $n = 3$ term gives $\frac{1}{3!}\frac{(-ik)^{1+1+1}}{1!1!1!}\langle x\rangle_c^3$. Putting it altogether gives

$$\langle x^3\rangle = \langle x^3\rangle_c + 3\langle x^2\rangle_c\langle x\rangle_c + \langle x\rangle_c^3 \implies \langle x^3\rangle_c = \langle x^3\rangle - 3\langle x^2\rangle\langle x\rangle + 2\langle x\rangle^3. \tag{10}$$

Similar lines of reasoning for the fourth moment yield

$$\langle x^4\rangle = \langle x^4\rangle_c + 4\langle x^3\rangle_c\langle x\rangle_c + 3\langle x^2\rangle_c^2 + 6\langle x^2\rangle_c\langle x\rangle_c^2 + \langle x\rangle_c^4, \tag{11}$$

which would determine the fourth cumulant as

$$\langle x^4\rangle_c = \langle x^4\rangle - 4\langle x^3\rangle\langle x\rangle - 3\langle x^2\rangle^2 + 12\langle x^2\rangle\langle x\rangle^2 - 6\langle x\rangle^4. \tag{12}$$

It is worth mentioning that there is a powerful pictorial technique that enables determining the relation between moments and cumulants. Let us imagine that we want the relationship between the fourth moment of a distribution and

cumulants. You already can tell from the power-matching above that we can write the fourth moment as

$$\langle x^4 \rangle = a_1 \langle x \rangle_c^4 + a_2 \langle x^2 \rangle_c^2 + a_3 \langle x^2 \rangle_c \langle x \rangle_c^2 + a_4 \langle x^3 \rangle_c \langle x \rangle_c + a_6 \langle x^4 \rangle_c. \tag{13}$$

The terms are dictated by all of the ways we can add some number of positive integers to get the power of the moment in question (in this case, that power is four). The pictorial way to get the coefficients is to draw four indistinguishable balls and ask for each term, how many ways can I put those four balls in M bags where M is the number of monomials present in each term. For example, the first term we want to find how many ways there are to put 4 bags each containing one ball (this is 1). The second term is asking: how many ways can we put 2 bags, each which contain two balls (that is 3). The third: how many ways can we put 3 bags around 4 balls, with one containing 2 and two containing one (this is 6). The fourth: two bags, one containing three and one containing one (that is 4). The fifth: one bag containing four balls (that is one). So we have $a_1 = 1, a_2 = 3, a_3 = 6, a_4 = 4, a_5 = 1$.

III. IMPORTANT PROBABILITY DISTRIBUTIONS

A. Gaussian distribution

The normal or Gaussian distribution is associated with the following normalized probability density function for a continuous random variable on the real line:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-x_0)^2/2\sigma^2}. \tag{14}$$

Before evaluating the moments of the distribution, it is useful to cover some basic results related to integrating Gaussians. The first is that

$$\int_{-\infty}^{\infty} dx\, e^{-\alpha x^2} = \sqrt{\pi/\alpha}. \tag{15}$$

To show this, define $I = \int_{-\infty}^{\infty} dx\, e^{-\alpha x^2}$ and compute

$$I^2 = \int_{-\infty}^{\infty} dx\, e^{-\alpha x^2} \int_{-\infty}^{\infty} dy\, e^{-\alpha y^2}, \tag{16}$$

and perform a change of variables into polar coordinates, so that

$$I^2 = \int d\theta \int_0^{\infty} dr\, re^{-\alpha r^2} = 2\pi \frac{1}{2} \int_0^{\infty} du\, e^{-\alpha u} = \frac{\pi}{\alpha} \implies I = \sqrt{\frac{\pi}{\alpha}}. \tag{17}$$

Another useful identity has to do with even moments of gaussians (odd moments vanish by symmetry). They can be generated from the integral we evaluated by repeated differentiation:

$$(-1)^n \frac{d^n}{d\alpha^n} I(\alpha) = \int dx\, x^{2n} e^{-\alpha x^2} = \frac{(2n-1)!!}{2^n} \sqrt{\pi} \alpha^{-(2n+1)/2}. \tag{18}$$

The Fourier transform of the Gaussian gives us the characteristic function, which we may expand to get cumulants. To evaluate the Fourier transform, we can do the following [1]:

$$\int_{-\infty}^{\infty} dx\, e^{-ikx} e^{-\alpha x^2} = \sum_{n=0}^{\infty} \frac{(-ik)^{2n}}{(2n)!} \int dx\, x^{2n} e^{-\alpha x^2}. \tag{19}$$

I have already used the fact that odd moments vanish. Next, we can use the formula we just derived for even moments, writing

$$\sum_{n=0}^{\infty} \frac{(-ik)^{2n}}{(2n)!} \frac{(2n-1)!!}{2^n} \sqrt{\frac{\pi}{\alpha}} \frac{1}{\alpha^n}. \tag{20}$$

Using the fact that $(2n-1)!!/(2n)! = 2^{-n}/n!$, we can write

$$\sqrt{\frac{\pi}{\alpha}} \sum_{n=0}^{\infty} \frac{\left(-\frac{k^2}{4\alpha}\right)^n}{n!} = \sqrt{\frac{\pi}{\alpha}} e^{-\frac{k^2}{4\alpha}}, \tag{21}$$

which implies that the characteristic function of the normalized Gaussian is

$$\tilde{p}(k, x_0 = 0) = \int_{-\infty}^{\infty} dx\, e^{-ikx} \frac{e^{-x^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}} = e^{-\frac{1}{2}\sigma^2 k^2}, \tag{22}$$

[1] The standard "derivation" typically presented involves completing the square and using the Gaussian integral formula derived above. It turns out the formula is valid when α is complex but with negative real part, but this is not apparent from the above derivation of the Gaussian integral, and so I would like to present a more "above-board" approach.

The characteristic function for the off-centered Gaussian is simply:

$$\tilde{p}(k, x_0) = e^{-ikx_0 - \frac{1}{2}\sigma^2 k^2}, \tag{23}$$

which is immediate from a change of variables in the Fourier transform $x \to x + x_0$. We may immediately calculate cumulants as follows:

$$\ln \tilde{p}(k) = -ikx_0 - \frac{1}{2}\sigma^2 k^2 = (-ik)x_0 + \frac{(-ik)^2}{2!}\sigma^2, \tag{24}$$

which implies that $\langle x \rangle_c = x_0$, $\langle x^2 \rangle_c = \sigma^2$, and $\langle x^n \rangle_c = 0$ if $n > 2$. This is a very unique property to the Gaussian distribution. This may be used with the formulae in the previous section to compute arbitrary-order moments of the Gaussian.

B. Binomial distribution

Consider a process with two outcomes, A and B with probabilities p_A and $p_B = 1 - p_A$. Suppose we repeat this process $N.$ times. The probability that outcome A occurs N_A times is given by the binomial distribution corresponding to the $\binom{N}{N_A}$ ways that this event can occur with probability $p_A^{N_A}(1 - p_A)^{N - N_A}$.

$$p(N_A) = \binom{N}{N_A} p_A^{N_A}(1 - p_A)^{N - N_A}. \tag{25}$$

The characteristic function for this distribution is given by $\langle e^{-ikN_A} \rangle$

$$\tilde{p}(k) = \sum_{N_A=0}^{N} \binom{N}{N_A} (p_A e^{-ik})^{N_A}(1 - p_A)^{N - N_A} = \left(p_A e^{-ik} + (1 - p_A)\right)^N, \tag{26}$$

where the second equality follows from the binomial theorem: $(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}$. The cumulant generating function is therefore

$$\ln \tilde{p}(k) = N \ln(p_A e^{-ik} + (1 - p_A)) = N \ln p_1(k), \tag{27}$$

where $p_1(k)$ is the characteristic function of the single-trial ($N = 1$) binomial distribution. We may calculate the cumulants of the single-trial distribution as follows. The expectation value $\langle N_A^\ell \rangle = p_A$ (assuming we map outcome A to the value 1), and so we can immediately say that for the N-trial distribution:

$$\langle N_A \rangle = N p_A, \langle N_A^2 \rangle_c = N\left(p_A - p_A^2\right) = N p_A p_B. \tag{28}$$

C. Poisson distribution

The Poisson distribution describes the probability of a number of events occurring after some time T, when the events occur at some average rate r. The Poisson distribution is in fact a limiting case of the binomial distribution. To see that, chop up the time interval from 0 to T into $N_t = T/dt$ time steps of length dt. At each time step of length dt, there is a probability rdt that the event occurs and a probability $1 - rdt$ that the event does not occur. The probability of two events in a timespan dt is negligible. The probability that $n \ll N_t$ events occur in our time interval is

$$p(n) = \lim_{N_t \to \infty} \binom{N_t}{n} (rdt)^n (1-rdt)^{N_t - n} = \frac{1}{n!} e^{-rT} \frac{N_t!}{(N_t - n)!} (rdt)^n (1-rdt)^{-n} \approx \frac{(rT)^n}{n!} e^{-rT}. \tag{29}$$

In simplifying, I have used that $(1 - rdt)^{N_t} = (1 - rT/N_t)^{N_t} \to e^{-rt}$ as $N_t \to \infty$ and that for any reasonable n, $nrdt \ll 1$ and that $n \ll N_t \implies \frac{N_t!}{(N_t - n)!} \approx N_t^n$.

The Poisson distribution's characteristic function can be evaluated as

$$\tilde{p}(k) = e^{rT(e^{-ik} - 1)}. \tag{30}$$

The cumulant generating function is

$$\ln \tilde{p}(k) = rT(e^{-ik} - 1) = rT \sum_{n=1}^{\infty} \frac{(-ik)^n}{n!}, \tag{31}$$

which implies that $\langle n^k \rangle_c = rT$ for all k. We may therefore express various moments as $\langle n \rangle = rT$, $\langle n^2 \rangle_c = \langle n^2 \rangle - \langle n \rangle^2 = rT$, etc.

As an example of the Poisson distribution, let us ask: assuming that stars are uniformly distributed within the galaxy with density n, what is the probability that the nearest one is at a distance R away from the center of the galaxy? This statement is the statement that the volumetric distribution of stars is Poissonian, with spatial "rate" n. The probability that there is one star in a volume V is thus $p(1) = (nV)e^{-nV}$ with $V = \frac{4}{3}\pi R^3$. The probability that the star is in a spherical shell centered at R with width dR is set by the volume ratio, since the star is equally likely to show up at any spatial location. Therefore, we can write as the radial probability $p(R) = dp/dR = 4\pi R^2 n e^{-nV}$.

IV. MULTIVARIATE PROBABILITY DISTRIBUTIONS

We generalize the ideas to functions of many random variables. Now, an outcome is specified by N variables $\mathbf{x} = \{x_1, x_2, \ldots, x_N\}$. As an example, suppose we observe the position and velocity of a molecule in a gas: that requires specifying six variables (three positions, three momenta) and so the position-momentum PDF is a function of six variables. The probability density $p(\mathbf{x})$ is defined such that

$$\int d^N x p(\mathbf{x}) = 1. \tag{32}$$

Expectation values are specified by

$$\langle f(\mathbf{x}) \rangle = \int d^N x \; f(\mathbf{x}) p(\mathbf{x}). \tag{33}$$

The *unconditional PDF* of a subset of random variables $p(x_1 \cdots x_m)$ $(m < N)$ is the probability of a subset of variables having certain values, with *no constraint* on the values of the unmeasured variables. That is given by

$$p(x_1 \cdots x_m) = \int dx_{m+1} \cdots dx_N \; p(\mathbf{x}). \tag{34}$$

A related concept is the *conditional PDF* $p(x_1 \cdots x_m | x_{m+1} \cdots x_N)$ which is the probability of observing $x_1 \cdots x_m$ *given* that the other variables have the values $x_{m+1} \cdots x_N$. The PDF, properly normalized, is

$$p(x_1 \cdots x_m | x_{m+1} \cdots x_N) = \frac{p(x_1 \cdots x_m, x_{m+1} \cdots x_N)}{p(x_{m+1} \cdots x_N)}. \tag{35}$$

This result is known as Bayes' theorem.

Bayes' theorem is extremely important and has often subtle consequences. Let us explore one called the Monty Hall problem. The statement of the problem is:

Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 3, which has a goat. He then says to you, "Do you want to pick door No. 2?" Is it to your advantage to switch your choice?'.

The answer is that it is advantageous to stick to the door that you chose. To show this, let us first rephrase the problem as follows:

Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what's behind the doors, opens some other door with a goat, arbitrarily. *What is the probability that the goat is behind your door versus some other door?'*.

From Bayes' theorem, we may write

$$p(\text{your door}|\text{some other door is opened}) = \frac{p(\text{your door})}{p(\text{some other door is opened})} = 1/3, \tag{36}$$

where the two probabilities on the RHS are unconditional probabilities. The denominator is 1 by construction. The alternative probability is

$$p(\text{not your door}|\text{some other door is opened}) = \frac{p(\text{not your door})}{p(\text{some other door is opened})} = 2/3. \tag{37}$$

So you should switch [2].

Joint moments and cumulants of a multidimensional probability distribution can be computed in terms of the characteristic function

$$p(\mathbf{k}) = \langle \exp(-i\mathbf{k}\cdot\mathbf{x})\rangle, \tag{38}$$

with $\mathbf{k}\cdot\mathbf{x} = k_1x_1 + \cdots k_Nx_N$. The cumulant generating function is once again $\ln p(\mathbf{k})$. Taylor expanding the exponent immediately shows that moments and also cumulants are given as

$$\langle x_1^{n_1}x_2^{n_2}\cdots x_N^{n_N}\rangle = \left(\frac{\partial}{\partial(-ik_1)}\right)^{n_1}\left(\frac{\partial}{\partial(-ik_2)}\right)^{n_2}\cdots\left(\frac{\partial}{\partial(-ik_N)}\right)^{n_N}\Big|_{\mathbf{k}=0} p(\mathbf{k}) \tag{39}$$

[2] This derivation may appear a bit mysterious. And indeed, it obscures a subtle point, so you may find it more convincing to compute the conditional probability by listing the set of possible outcomes. The solution through that approach is as follows. The outcomes are $100, D_2$, $100, D_3$, $010, D_3$ and $001, D_2$, where the $D_{2,3}$ is denoting which door is open. The probability of the first two outcomes are each $1/6$: the probability of the 100 outcome multiplied by the $1/2$ probability that either D_2 or D_3 is opened. The probability of the other two outcomes is each $1/3$ (the probability of D_3 being open if we have 010 is 1). Therefore, we may evaluate

$$p(\text{your door}|\text{some other door is opened}) = 1/6 + 1/6 = 1/3,$$

and similarly

$$p(\text{not your door}|\text{some other door is opened}) = 1/3 + 1/3 = 1/6.$$

and

$$\langle x_1^{n_1} * x_2^{n_2} * \cdots * x_N^{n_N} \rangle_c = \left(\frac{\partial}{\partial(-ik_1)}\right)^{n_1} \left(\frac{\partial}{\partial(-ik_2)}\right)^{n_2} \cdots \left(\frac{\partial}{\partial(-ik_N)}\right)^{n_N} \Big|_{\mathbf{k}=0} \ln p(\mathbf{k}). \quad (40)$$

The graphical relation described for univariate distributions holds here as well, except that now if we want a moment like $\langle x^2 y \rangle_c$, there are two types of balls: x-balls and y-balls. You can convince yourself that this is given by

$$\langle x^2 y \rangle = \langle x \rangle_c^2 \langle y \rangle_c + \langle x^2 \rangle_c \langle y \rangle_c + 2\langle x * y \rangle_c \langle x \rangle_c + \langle x^2 * y \rangle_c. \quad (41)$$

The first term corresponds to the one way to put three bags around these three balls (with the order of the bags not being important). The second and third terms are associated with the ways of putting one bag around two balls and a second around the third ball. The last term is associated with the one way of putting a bag around all three balls.

Another example of this, which is simpler, and extremely important, is for the so-called *connected correlation* or covariance of a distribution. You should show that

$$\langle xy \rangle = \langle x \rangle_c \langle y \rangle_c + \langle xy \rangle_c \implies \langle xy \rangle_c = \langle xy \rangle - \langle x \rangle \langle y \rangle. \quad (42)$$

This vanishes when the unconditional probability distribution $p(x, y) = p(x)p(y)$ (in other words, when it is separable).

V. MULTIVARIATE GAUSSIAN DISTRIBUTION

By far the most important multivariate probability distribution in physics is the multivariate Gaussian distribution. It is given as

$$p(\mathbf{x}) = \mathcal{N} \exp\left[-\frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^T M (\mathbf{x} - \mathbf{x}_0)\right], \quad (43)$$

where $\mathbf{v}^T C \mathbf{v}$ denotes a vector-matrix-vector product, written in repeated-index notation as $v_i C_{ij} v_j$. The way to simplify this is to diagonalize M (M in fact needs to be diagonalizable in order for this distribution to make sense). Denote the diagonalization of M as $M = S^T D S$ where S is an orthogonal matrix ($S S^T = S^T S = 1$)

and T represents the transpose. The matrix $D = \text{diag}(\lambda_1 \cdots \lambda_N)$ is a diagonal matrix where the λ represent the eigenvalues of M and N is the dimensionality of $\mathbf{x}$. Then we write

$$(\mathbf{x} - \mathbf{x}_0)^T M(\mathbf{x} - \mathbf{x}_0) = (\mathbf{x} - \mathbf{x}_0)^T S^T DS(\mathbf{x} - \mathbf{x}_0) \equiv (\mathbf{v} - \mathbf{v}_0)^T D(\mathbf{v} - \mathbf{v}_0), \tag{44}$$

where $\mathbf{v} = S\mathbf{x}$, and $\mathbf{v}_0 = S\mathbf{x}_0$. Denoting the individual components of $\mathbf{v}$ as v_i and defining $\lambda_i = \frac{1}{\sigma_i^2}$ (which requires that the λ are positive), we may then we-write the distribution as

$$p(\mathbf{v}) = \mathcal{N} \exp\left[-\sum_i \frac{(v - v_0)^2}{2\sigma_i^2}\right], \tag{45}$$

which allows us to read off the normalization constant by integrating in v space (there is no Jacobian to worry about because the transformation is enacted by an orthogonal matrix, which is volume-preserving. Stated differently $|\det(S)| = 1$). We may then write

$$p(\mathbf{v}) = \sqrt{\frac{\det(\mathrm{S})}{(2\pi)^N}} \exp\left[-\sum_i \frac{(v - v_0)^2}{2\sigma_i^2}\right]. \tag{46}$$

It is straightforward to see that by re-expressing $\mathbf{v}$ in terms of $\mathbf{x}$, we have

$$p(\mathbf{x}) = \sqrt{\frac{\det(\mathrm{S})}{(2\pi)^N}} \exp\left[-\frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^T M(\mathbf{x} - \mathbf{x}_0)\right]. \tag{47}$$

We can determine the characteristic function via

$$p(\mathbf{k}) = \sqrt{\frac{\det(\mathrm{S})}{(2\pi)^N}} \int d\mathbf{v}\, e^{-i(S\mathbf{k})\cdot\mathbf{v}} \exp\left[-\sum_i \frac{(v - v_0)^2}{2\sigma_i^2}\right]. \tag{48}$$

Therefore, we have

$$p(\mathbf{k}) = \exp\left[-i(S\mathbf{k}) \cdot \mathbf{v}_0 - \frac{1}{2}\sum_{i=1}^{N} \sigma_i^2 (S\mathbf{k})_i (S\mathbf{k})_i\right]. \tag{49}$$

Since $S\mathbf{k} \cdot \mathbf{v}_0 = \mathbf{k} \cdot S^T\mathbf{v}_0 = \mathbf{k} \cdot \mathbf{x}_0$ and $\sum_{i=1}^{N} \sigma_i^2 (S\mathbf{k})_i (S\mathbf{k})_i = \mathbf{k}^T S^T D^{-1} S\mathbf{k} = \mathbf{k}M^{-1}\mathbf{k}$, we may write

$$p(\mathbf{k}) = \exp\left[-i\mathbf{k} \cdot \mathbf{x}_0 - \frac{1}{2}\mathbf{k}^T M^{-1}\mathbf{k}\right]. \tag{50}$$

Its logarithm, which is the cumulant generating function, is still quadratic and so cumulants of third and higher-order vanish in the multidimensional case. That vanishing, combined with the pictorial rule for relating multidimensional moments

to multidimensional cumulants, allows us to write down even-order moments of a multidimensional Gaussian in terms of quadratic-order moments. This connection is sometimes called Wick's theorem. For example, let us evaluate for a multivariate (centered) Gaussian ($\mathbf{x}_0 = 0$): $\langle x_1 x_2 x_3 x_4 \rangle$, where $x_{1,2,3,4}$ refer to different variables. According to our pictorial rule, we can write this moment as a sum of (1) a term associated with a fourth-order cumulant, (2) a term with third and first-order cumulants, (3) a term with two second-order cumulants, (4) a term with one second-order cumulant and two first-order cumulants, and (5) a term with four first order-cumulants. All but (3) vanish because we assumed that the Gaussian is centered, so the first-order cumulant vanishes, and the third- and fourth-order cumulants vanish as a result of the distribution being Gaussian.

Now we ask, if we have four distinct types of balls, how many ways may we put them into two bags, each of with two balls. There are three, and so we have

$$\langle x_1 x_2 x_3 x_4 \rangle = \langle x_1 * x_2 \rangle_c \langle x_3 * x_4 \rangle_c + \langle x_1 * x_3 \rangle_c \langle x_2 * x_4 \rangle_c + \langle x_1 * x_4 \rangle_c \langle x_2 * x_3 \rangle_c, \quad (51)$$

which can be simplified to

$$\langle x_1 x_2 x_3 x_4 \rangle = \langle x_1 x_2 \rangle \langle x_3 x_4 \rangle + \langle x_1 x_3 \rangle \langle x_2 x_4 \rangle + \langle x_1 x_4 \rangle \langle x_2 x_3 \rangle. \quad (52)$$

VI. CENTRAL LIMIT THEOREM

With multivariate probability distributions in place, we may now show a powerful result, which is that for a wide variety of univariate probability distributions $p(x)$, the probability distribution for the sum of N instances of x approaches a Gaussian, as N gets large. The central limit theorem is going to help us understand the transition from microscopic behaviors to macroscopic thermodynamic quantities. With the apparatus in place, we may prove it as follows. Consider the variable

$$X = x_1 + \cdots + x_N, \quad (53)$$

where we're saying we draw N random numbers from the probability distribution $p(x)$ and evaluate X. Then we do that again and again, and construct a distribution

$p(X)$. The probability distribution in question is given as

$$p(X) = \int dx_1 \cdots dx_N \; p(x_1) \cdots p(x_N) \delta\left(X - \sum_{i=1}^{N} x_i\right). \tag{54}$$

Let us calculate its characteristic function (if something is going to reduce to a Gaussian, a way to do that it to show that the cumulants beyond second-order vanish). The characteristic function is simply

$$p(k) = \int dX \; e^{-ikX} \int dx_1 \cdots dx_N \; p(x_1) \cdots p(x_N) \delta\left(X - \sum_{i=1}^{N} x_i\right) = p_1(k)^N, \tag{55}$$

where p_1 is the characteristic function of $p(x)$. Now, the cumulants are generated by

$$\ln p(k) = N \ln p_1(k) = N\left((-ik)\langle x\rangle_c + \frac{(-ik)^2}{2!}\langle x^2\rangle_c + \frac{(-ik)^3}{3!}\langle x^3\rangle_c + \dots\right). \tag{56}$$

Of course, we could have written also

$$\ln p(k) = \left((-ik)\langle X\rangle_c + \frac{(-ik)^2}{2!}\langle X^2\rangle_c + \frac{(-ik)^3}{3!}\langle X^3\rangle_c + \dots\right). \tag{57}$$

Therefore, we have

$$\langle X\rangle = N\langle x\rangle, \langle X^2\rangle_c = N\langle x^2\rangle_c, \langle X^3\rangle_c = N\langle x^3\rangle_c, \dots. \tag{58}$$

This equation suggests that the higher cumulants get successfully less important in relative magnitude. Why? Because we expect that typically for a centered distribution, if $\langle x^2\rangle_c = \sigma^2$, then $\langle x^n\rangle_c \sim \sigma^n$ and in such a case, all cumulants are relevant in describing the distribution. According to Eq. (58), the higher-order cumulants grow much more slowly (we expect them to grow as N^n). Therefore, we drop terms of third and higher-order as $N \to \infty$, leading to

$$\ln p(k) \approx (-ik)N\langle x\rangle_c + \frac{(-ik)^2}{2!}N\sigma^2, \tag{59}$$

which implies that the distribution is Gaussian

$$p(X) = \frac{1}{\sqrt{2\pi N\sigma^2}} \exp\left[-\frac{(X - N\langle x\rangle)^2}{2N\sigma^2}\right]. \tag{60}$$

A. Stirling's approximation

We take a small break to introduce a mathematical approximation which features very many times in this course. It has to do with the approximation of $\ln N$ for $N \gg 1$. It may be approximated by

$$\ln N! = \sum_{n=1}^{N} \ln n \approx \int_{1}^{N} dx \ln x = (x \ln x - x)\Big|_{1}^{N} = N \ln N - N + 1 \approx N \ln N - N. \quad (61)$$

VII. INFORMATION, ENTROPY, AND ESTIMATION

Consider a random variable with a discrete set of outcomes (here, called symbols) $\{x_i\}$ with the associated probability distribution $p(i)$ where $i = 1, \ldots, M$. An important concept is the *information content* of a probability distribution. Let us imagine that we have received a message of length N composed of symbols drawn from the probability distribution p. If you wish, we could imagine that letters are drawn from the alphabet (26 symbols) and some probability distribution. One may ask: how many bits of information do I need to encode a message without error?

Let us start with a simple example. Suppose that we have an alphabet with two characters A and B and they each have probability $1/2$ to appear. Then the four messages AA, AB, BA, BB are equally likely to occur with probability $1/4$. I will need $\log_2 4 = 2$ bits of information to encode these possible messages faithfully. In binary, I could represent these messages by 00, 01, 10, 11. If I use less than 2 bits, I cannot represent all the messages uniquely. For example, if I encode both AA, AB as 0 and BA, BB as 1, then if you receive a 0 you cannot be sure what the message really is. Similarly, if I use more than 2 bits of information to encode these four messages, that is simply redundant and we want to send messages using the least numbers of symbols or bits or data as possible. For example, I could decide to use three bits and assign: AA = 00 or 100, AB = 01 or 101, BA = 10 or 110, BB = 11 or 111. If I send you either 11 or 111, you can be sure of what I'm sending, but it is not necessary.

Now if I have two symbols drawn with equal probability and a message of length

N, then there are 2^N possible messages and I need N bits to encode all the possible messages uniquely (it's just the binary representation of each message). Similarly, if I have M symbols of equal probability, then I need $N \log_2 M$ bits to encode the M^N equally likely messages.

What happens if some symbols are more likely than others? The number of bits needed always decreases relative to the case where bits are equally likely to occur. This is because you can encode the less likely messages in a compressed representation with a smaller number of bits. Let's start by describing the extreme case in which we have two symbols, A and B, but $p_A = 1, p_B = 0$. The only messages I can form of length N are $A_1 \cdots A_N$. I don't even need a whole bit to encode the possible messages because there is only one that can be sent.

Now let's consider the two-symbol case in which we suppose A is much more likely than B (but $p_B \neq 0$). The basic idea is to use a smaller number of bits to encode the more likely messages and a larger number of bits to encode the less likely messages. Since the less likely messages rarely show up, on average you're using less bits.

Let us talk about the general case now. Suppose we have a message of length $N \gg M$ where M is the number of symbols. The length N should be imagined as going to the infinite limit. We know that the probability of a *general* message $s_1 s_2 \cdots s_N$ where s_i denotes the value of the ith symbol, is given by the *multinomial coefficient*

$$p(s_1 \cdots s_N) = N! \prod_{i=1}^{M} \frac{p_i^{n_i}}{n_i!}, \tag{62}$$

with the constraint that $\sum n_i = N$. In the limit as N becomes large, the central limit theorem kicks in, and this distribution converges to a Gaussian which eventually becomes infinitely tight (relative to the mean). Therefore, one can convince themselves that they expect to rarely see messages in which we see n_i being very different from Np_i where very different means by more than $\sqrt{Np_i}$. The number of *typical* messages then becomes the number of messages where $n_i = Np_i$ times a factor which scales like the product of the n_i (this would come from, for example, integrating the probability distribution to capture most of the probability density).

The number of messages where $n_i = Np_i$ is

$$g = N! \prod_{i=1}^{M} \frac{1}{(Np_i)!} \tag{63}$$

So we say that the number bits would be like $\log_2 g(n_1 \cdots n_M) = \log_2 g + \zeta M \log_2 N$, where ζ is some proportionality constant that we do not need to evaluate because as $N \gg m$, we know that $N \gg \log N$ and so therefore, we may just neglect the second term altogether. We shall define the average number of bits needed (the number of bits corresponding to typical messages, divided by the message length) as the *entropy*, S of the probability distribution. Since we have made no connection to thermodynamic entropy yet, we could call this the *information theoretic entropy*. In what follows, we will replace the base-2 logarithm by the natural logarithm (since we will not care too much about using this as a coding scheme in what follows, we stick with the easiest base to use). The entropy S is given by

$$NS = \ln g = \ln N! - \sum_{i=1}^{M} \ln(Np_i)! \approx N \ln N - N - \sum_{i=1}^{M} (Np_i \ln Np_i - Np_i), \tag{64}$$

where we have used Stirling's approximation for the logarithm of a factorial. The entropy further simplifies as

$$NS = N \ln N - \sum_{i=1}^{M} Np_i (\ln N + \ln p_i) = - \sum_{i=1}^{M} Np_i \ln p_i, \tag{65}$$

or simply

$$S = - \sum_{i=1}^{M} p_i \ln p_i. \tag{66}$$

The entropy has the interpretation that it is the mean number of bits needed per symbol to communicate a message efficiently and accurately. The number of typical messages can be calculated given the entropy as e^{NS}. If we take all typical messages to be equally probable (which is the case for typical messages), then we can say that the probability of a message is $p(s_1 \cdots s_N) = e^{-NS}$. This property that for a large string, all messages are equally likely is called asymptotic equipartition and will come up in a few lectures when we describe the distribution of microstates that a large physical system can have.

The entropy has a few important properties that are useful in what follows. One is that it is maximized for the uniform probability distribution. We gave the intuition, but now let's prove it because it introduces another important tool. We want to maximize the entropy given with a *constraint* that the probabilities are normalized. Therefore, we introduce a Lagrange multiplier λ and maximize

$$S(p, \lambda) = -\sum_{i=1}^{M} p_i \ln p_i - \lambda(\sum_{i=1}^{M} p_i - 1). \tag{67}$$

Therefore:

$$\frac{\partial S(p, \lambda)}{\partial p_n} = 0 \implies -\ln p_n - 1 - \lambda = 0 \implies \ln p_n = -\lambda - 1 \implies p_n = e^{-(1+\lambda)}. \tag{68}$$

We also have that

$$\frac{\partial S(p, \lambda)}{\partial \lambda} = 0 \implies \sum_{i=1}^{M} p_i = 1, \tag{69}$$

which further requires that

$$Me^{-(1+\lambda)} = 1 \implies p_n = 1/M. \tag{70}$$

One can also verify that this is a maximum by taking second derivatives.

What about the minimum entropy state? As can be seen from the expression, entropy is positive definite: probabilities are positive, and negative logarithms of probabilities are also positive. The minimum value of this expression is zero, and it is realized when all but one probability is zero. If we have $p_1 = 1$ and $p_{2,\ldots,M} = 0$, the entropy is zero ($x \ln x \to 0$ as $x \to 0$). You can see from these two cases that the entropy of a probability distribution is measuring something like the dispersion or disorder of the distribution.

Before we move on to statistical mechanics, there is one more technique which is worth showing. Suppose we want to find an unbiased estimate of a distribution whose mean we know for some quantities $f_1, f_2, \ldots, f_n$. In other words:

$$f_\alpha = \sum_i p_i F_{\alpha,i}, \tag{71}$$

where $F_{\alpha,i}$ is the function that, given a state i, returns the value of property F_α. As a concrete example, the property in question could be energy, and it would depend

on the state of a set of particles, denoted i. An unbiased estimate is one where we set all probabilities to be the same (with this constraint). That is the same thing as maximizing entropy with this constraint. Therefore, we should maximize

$$S(p, \lambda_1 \cdots \lambda_n) = -\sum_{i=1}^{M} p_i \ln p_i - \lambda_0 \left(\sum_{i=1}^{M} p_i - 1 \right) - \sum_{\alpha=1}^{n} \lambda_\alpha \left(\sum_{i=1}^{M} p_i F_{\alpha,i} - f_\alpha \right). \quad (72)$$

Taking the derivatives with respect to the p_i and setting them to zero, yield:

$$\ln p_i + 1 + \lambda_0 + \sum_{\alpha} \lambda_\alpha F_{\alpha,i} = 0, \quad (73)$$

with solution

$$p_i = \exp\left[-1 - \lambda_0 - \sum_{\alpha} \lambda_\alpha F_{\alpha,i} \right]. \quad (74)$$

The normalization constraint implies that we can write p_i as

$$p_i = \frac{\exp\left[-\sum_{\alpha} \lambda_\alpha F_{\alpha,i} \right]}{\sum_{i=1}^{M} \exp\left[-\sum_{\alpha} \lambda_\alpha F_{\alpha,i} \right]}. \quad (75)$$