Notes for Lecture 4

More is different

We have already argued that the maximum entropy principle is central to the statistical understanding of thermodynamics. In short, for a closed system, the entropy will be maximized for the equilibrium state, since the entropy will tend to increase as the system progresses. Indeed, this is a major way in which we can "understand" the arrow of time.

The arrow of time is mysterious. If the microscopic laws that underlie the macroscopic phenomena are invariant under time reversal, why should the macroscopic systems exhibit a definite directionality in time at all? One answer is that the events tend to increase the entropy of the Universe. Such a process is irreversible, defining the arrow of time.

Even if one might accept the maximum entropy principle as the fundamental axiom, a question still lingers. How does an irreversible process occur when the microscopic laws are time-reversal invariant? This is one of the most central questions of statistical mechanics, and research on this question is still very active. One thing that is certain is that the irreversibility has to do with the many-particle nature of the system. In fact, this applies to almost everything that we do in statistical mechanics. Not just the notion of the irreversible process. For example, a phase transition is a sure thing that we can repeatedly observe again and again, if we apply the same set of conditions, thanks to large number of molecules involved. More mundane things such as the pressure and the temperature—the certainty of being able to measure and quantify those things is also all due to the same large number of molecules involved. We know that underneath it all, every molecule is very uncertain and jittery at all times. But, in the limit of large number of molecules, all of these quantities assume very certain values. Also, more surprising "laws" such as the irreversibility and the

maximum entropy principle emerge[1]. This is how we get this physical principle that we are so sure of to call it statistical "mechanics."

So, the statistic mechanics is a great example of a now famous maxim: *more is different*[2].

Here, we summarize some more essential topics related to the probability, and explore some properties of large numbers.

4.1 Probability for many random variables

Naturally, a many body system involves many random variables, and they require extending our discussion of the last lecture.

The PDF now becomes $p(x_1, x_2, ..., x_N)$, assuming that there are N random variables. In short, we may write $p(\vec{x})$, keeping in mind that $\vec{x}$ is an N dimensional vector. We call this PDF a **joint PDF**.

4.1.1 Independence

If we consider the previous example of a coin toss, then this probability can be taken to mean throwing N coins at the same time. Assuming normal coins, it stands to reason that the result of one coin has no bearing on the result of another coin. In this case, the join PDF can be decomposed simply.

$$p(x_1, x_2, ..., x_N) = p(x_1)p(x_2)\cdots p(x_N) \qquad \text{independent variables} \tag{4.1}$$

Note here that the same symbol p is used both for the joint PDF and for the single variable PDF. This is simply for the economy of notation, and by no means imply that they are identical functions. We just use $p(\star)$ to mean the PDF of $\star$. In particular, note that each of $p(x_i)$'s does not have to be the same PDF in general. In the coin toss example, they will be the same, if all coins are made the same way. If all coins are rigged differently, somehow, then they will be represented by different functions.

[1] Another example of a law that emerge due to many particles is Newton's laws.
[2] P. W. Anderson, Science 177, 393 (1972).

In physics, one often neglects the inter-particle interaction and builds an independent particle model to get a quick start on any problem. Within this approximation, the probability that governs each particle's degree of freedom can be considered independent. Sometimes, this is a good approximation, but, sometimes, it is not. Clearly, an independent particle approximation makes the problem very easy.

While we will often use independent particle models in this course, it is important to know the limitations of such models. In the following discussions, we do not assume independence of the above random variables, unless noted otherwise explicitly.

4.1.2 Bayes theorem

For a given N random variables, $x_1, ..., x_N$, suppose that one has actually *measured* a few random variables, say, $x_{m+1}, ..., x_N$. Then, one can ask, given these partial outcomes, what is the probability distribution for the rest of random variables, $x_1, ..., x_m$? This probability is called a **conditional probability**, and it is given by the famous **Bayes theorem**.

$$p(x_1, ..., x_m | x_{m+1}, ..., x_N) = \frac{p(x_1, ..., x_N)}{p(x_{m+1}, ..., x_N)} \qquad \textbf{conditional PDF; Bayes} \qquad (4.2)$$

In contrast, one can ask the question, what is the probability distribution for $x_1, ..., x_m$, regardless of the outcomes of $x_{m+1}, ..., x_N$? To obtain this probability, we simply need to integrate over all possible outcomes of $x_{m+1}, ..., x_N$. The resulting PDF is called the **unconditional PDF**.

$$p(x_1, ..., x_m) = \int dx_{m+1} ... \int dx_N \, p(x_1, ..., x_N) \qquad \textbf{unconditional PDF} \qquad (4.3)$$

Bayesian inference

Using symbols $A = x_1, ..., x_m$ and $B = x_{m+1}, ..., x_N$, we can writ the above Bayes theorem as $p(A|B) = p(A, B)/p(B)$, with A and B, now representing any two general random variables, each of which can be multi-dimensional. Clearly, one can also write $p(B|A) = p(A, B)/p(A)$. Combining, we get $p(A|B)p(B) = p(B|A)p(A)$, which is a well-known form, equivalent to Eq. 4.2, in which Bayes theorem is written. **Bayesian inference** is directly based on this theorem:

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}. \qquad \text{Bayesian inference} \qquad (4.4)$$

In this expression, the right hand side is the input, and the left hand side is the output, and this inference is usually used iteratively. In this Bayesian inference scheme, $p(A)$ is **the prior probability**, and $p(A|B)$ is **the posterior probability**, which is the updated probability based on evidence B.

4.1.3 Joint characteristic functions, Cluster expansion

The generalization of the characteristic functions to the multi-variable case is straightforward.

$$\tilde{p}(\vec{k}) \;=\; \left\langle e^{-i\vec{k}\cdot\vec{x}} \right\rangle \tag{4.5}$$

where, again, you are reminded that each vector here is an N dimensional vector.

Joint moments and **joint cumulants** are defined much the same way, i.e. as the expansion coefficients for the characteristic functions.

$$\begin{aligned}
\tilde{p}(\vec{k}) \;&=\; \left\langle \prod_{j=1}^{N} e^{-ik_j x_j} \right\rangle \\
&=\; \left\langle \prod_{j=1}^{N} \sum_{n_j=0}^{\infty} \frac{(-ik_j)^{n_j}}{n_j!} (x_j)^{n_j} \right\rangle \\
&=\; \sum_{n_1,n_2,\ldots,n_N} \left\langle \prod_{j} \frac{(-ik_j)^{n_j}}{n_j!} (x_j)^{n_j} \right\rangle \\
&=\; \sum_{n_1,n_2,\ldots,n_N} \frac{(-ik_1)^{n_1}(-ik_2)^{n_2}\cdots(-ik_N)^{n_N}}{n_1!n_2!\cdots n_N!} \left\langle x_1^{n_1} x_2^{n_2} \cdots x_N^{n_N} \right\rangle
\end{aligned} \tag{4.6}$$

The last line defines **joint moments**, $\langle x_1^{n_1} x_2^{n_2} \cdots x_N^{n_N} \rangle$. It is then straightforward to define **joint cumulants** as

$$\log \tilde{p}(\vec{k}) \;=\; \sum_{n_1,n_2,\ldots,n_N} \frac{(-ik_1)^{n_1}(-ik_2)^{n_2}\cdots(-ik_N)^{n_N}}{n_1!n_2!\cdots n_N!} \left\langle x_1^{n_1}, x_2^{n_2}, \cdots, x_N^{n_N} \right\rangle_c \tag{4.7}$$

The commas in the definition of joint cumulants are significant. Without such commas, an ambiguity in interpretation can arise: $\langle x_1 x_2 \rangle_c$ can be taken to mean a joint cumulant between x_1 and x_2, or as the first cumulant of a derived variable $x_1 x_2$—the two mean quite different things. To avoid this confusion, inserting a comma as in $\langle x_1, x_2 \rangle_c$, makes it clear that we are not multiplying two variables before taking the cumulant.

From the above, we can write

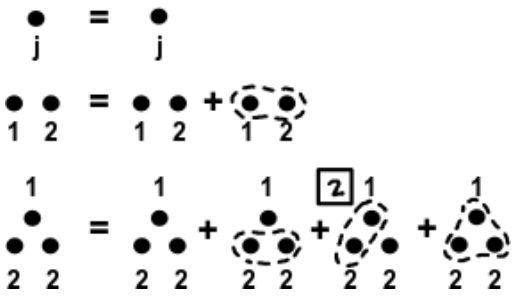
$$\langle x_1^{n_1} x_2^{n_2} \cdots x_N^{n_N} \rangle = \left(\frac{\partial}{\partial(-ik_1)} \right)^{n_1} \cdots \left(\frac{\partial}{\partial(-ik_N)} \right)^{n_N} \tilde{p}(\vec{k}) \Big|_{\vec{k}=0} \tag{4.8}$$

$$\langle x_1^{n_1}, x_2^{n_2}, \cdots, x_N^{n_N} \rangle_c = \left(\frac{\partial}{\partial(-ik_1)} \right)^{n_1} \cdots \left(\frac{\partial}{\partial(-ik_N)} \right)^{n_N} \log \tilde{p}(\vec{k}) \Big|_{\vec{k}=0} \tag{4.9}$$

Note that both joint moment and joint cumulant are invariant upon permutation of indices $j = 1, ..., N$: i.e., the order in which $x_1^{n_1}$, ..., $x_N^{n_N}$ appear on the left hand side of the above relations is only conventional, and unimportant.

The **cluster expansion** remains valid for this multi-variable case. You are welcome to prove it, if you are so inclined. For the current multi-variable case, one important difference is that not all dots are identical. They must be labeled with the index $j = 1, ..., N$, and dots with different indices must be treated differently. Here, studying some examples suffices.

Here are two graphical equations, where the multiplicity is now noted in a square, to avoid confusion with dot labels.



They correspond to the following equations.

$$\langle x_j \rangle = \langle x_j \rangle_c \tag{4.10}$$

$$\langle x_1 x_2 \rangle = \langle x_1 \rangle_c \langle x_2 \rangle_c + \langle x_1, x_2 \rangle_c \tag{4.11}$$

$$\left\langle x_1 x_2^2 \right\rangle = \langle x_1 \rangle_c \langle x_2 \rangle_c^2 + \langle x_1 \rangle_c \left\langle x_2^2 \right\rangle_c + 2 \langle x_1, x_2 \rangle_c \langle x_2 \rangle_c + \left\langle x_1, x_2^2 \right\rangle_c \tag{4.12}$$

The first equation seems trivial, but not necessarily. It should be properly understood that the first cumulant of a single variable is the same as the first moment of that variable, *even if* all variables $x_1, ..., x_N$ are strongly correlated, i.e. even when they are far from being independent. In other words, the expectation value for the first equation is the integral over the unconditional PDF (Eq. 4.3 with $m = 1$ for $j = 1$), *not* a PDF of a lone single variable. Similar comments apply to all single variable cluster expansion formulae that we studied in the previous lecture, which remain valid in the current case as well.

Now, the second equation is very much worth noting. It means

$$\langle x_1, x_2 \rangle_c \;=\; \langle x_1 x_2 \rangle - \langle x_1 \rangle \langle x_2 \rangle \,. \qquad \text{covariance, correlation} \tag{4.13}$$

This is the so-called **covariance** or the **correlation** between two variables x_1 and x_2. Some authors define the **correlation** as a properly scaled dimensionless version of this: $\frac{\langle x_1, x_2 \rangle_c}{\sigma_{x_1} \sigma_{x_2}}$, where σ's are standard deviations. See similar discussion at the bottom of page 5 in LN 3.

4.1.4 Independence, revisited

If we assume that x_j's are independent (Eq. 4.1), then

$$\langle x_1^{n_1}, x_2^{n_2}, \cdots, x_N^{n_N} \rangle_c \;=\; 0 \qquad \text{if at least two of } n_j\text{'s are non-zero.} \tag{4.14}$$

This means that **cumulants that involve different variables measure an interdependence**, since any cumulants that involve two or more variables are zero for independent variables. To prove the above result, one can note that Eq. 4.1 means that

$$\tilde{p}(\vec{k}) \;=\; \tilde{p}_1(k_1)\tilde{p}_2(k_2)\cdots\tilde{p}_N(k_N), \qquad \text{independence} \tag{4.15}$$

$$\log\tilde{p}(\vec{k}) \;=\; \log\tilde{p}_1(k_1) + \log\tilde{p}_2(k_2) + \cdots \log\tilde{p}_N(k_N). \qquad \text{independence} \tag{4.16}$$

Here, subscripts are used for $\tilde{p}$'s so that it is clear beyond any doubt that each single variable PDF is a separate one, consistent with our assumption all along, as mentioned in the early part of this LN. Upon differentiating the second equation successively by any two (or more) variables, then, we get zero, which proves our assertion, since cumulants involve such successive differentiation (Eq. 4.9).

4.2 Central limit theorem

We noted in the previous lecture that the binomial distribution approaches the Gaussian distribution as the number of trials goes to infinity. This is not a coincidence at all. It is due to the **central limit theorem**, which is one of the most celebrated theorems of statistics, due to its broad application. Here, we will study this theorem.

It is quite instructive to re-visit the coin toss problem. While we treated this problem as a single variable probability problem in the previous lecture, we can also view this problem as a many variable problem: instead of doing N coin tosses of a single coin, we can make N exact copies of the same coin, and then toss them all at

the same time. Let x_1, x_2, ..., x_N, be the outcome of N coins. By construction, these are independent variables, and each variable is governed by the same PDF

$$p(x_j) = \begin{cases} h\delta(x_j - 1) \\ (1-h)\delta(x_j) \end{cases} \tag{4.17}$$

where we assigned value $x_j = 1$ for head (probability $0 \le h \le 1$) and $x_j = 0$ for tail. And, the joint PDF is given by $p(x_1, ..., x_N) = \prod_j p(x_j)$, by Eq. 4.1.

Let us ask the question: what is the probability distribution for variable x?

$$x \equiv x_1 + x_2 + ... + x_N \tag{4.18}$$

Note that x counts the number of heads out of N simultaneous tosses, and so is the same central quantity that we investigated in the previous lecture.

We will now answer this question in a general way, not only for this coin toss problem, to establish the central limit theorem. The only **overarching assumption** that we will make is that **all variables, x_1, x_2, ..., x_N, are equivalent**, whether or not they are independent. This means that the unconditional probability for each variable x_i ($i = 1, \ldots, N$) is one and the same; in particular, in the case of independence, all functions $\tilde{p}_i$'s in Eqs. 4.15, 4.16 are *now* all the same. Physically, what we are assuming is that x_i's describe the same properties of the same type of particle, for instance.

Note that the PDF for x would be an unconditional PDF for x (Eq. 4.3). To be more precise, we can transform variables $x_1, x_2, ..., x_N$ to x, x_2, x_3, ..., x_N (or x plus any $N-1$ variables from the original set), and then integrate the PDF over x_2, ..., x_N, to obtain the PDF for x. However, we do not need to bother to go through this process, if we study the characteristic function of x.

$$\tilde{p}(k) = \left\langle e^{-ikx} \right\rangle \tag{4.19}$$

where the right hand side is given by $\int dx_1 ... dx_N \, e^{-ikx} p(x_1, ..., x_N)$. Comparing this with Eq. 4.5, we recognize this characteristic function for x as

$$\tilde{p}(k) = \tilde{p}(\vec{k})\Big|_{k_1 = k_2 = ... = k_N = k} \tag{4.20}$$

Then, using Eq. 4.7, we get

$$\begin{aligned} \log \tilde{p}(k) &= \sum_{n_1, n_2, ..., n_N} \frac{(-ik)^{n_1 + n_2 + ... + n_N}}{n_1! n_2! \cdots n_N!} \left\langle x_1^{n_1}, x_2^{n_2}, \cdots, x_N^{n_N} \right\rangle_c \\ &= \sum_n \frac{(-ik)^n}{n!} \sum_{n_1, ..., n_N; \sum_j n_j = n} \frac{n!}{n_1! n_2! \cdots n_N!} \left\langle x_1^{n_1}, x_2^{n_2}, \cdots, x_N^{n_N} \right\rangle_c . \end{aligned}$$

So, quite generally, we get the result that the n-th cumulant for x is given by the sum of joint cumulants, weighted by multinomial coefficients (cf. Eq. 3.35).

$$\langle x^n \rangle_c = \sum_{n_1,\ldots,n_N;\sum_j n_j=n} \frac{n!}{n_1! n_2! \cdots n_N!} \langle x_1^{n_1}, x_2^{n_2}, \cdots, x_N^{n_N} \rangle_c \tag{4.21}$$

Consider two cases.

1. **If x_j's are independent**, like in the coin toss example, then we can use Eq. 4.14. The above equation simplifies to

$$\langle x^n \rangle_c = \sum_j \langle {x_j}^n \rangle_c = N \langle {x_1}^n \rangle_c \qquad n \geq 1 \tag{4.22}$$

while, of course, the zeroth ($n = 0$) cumulant is 1, by probability sum rule. Note that in the last step, the equivalence of all x_j's has been used. This means that all cumulants of x are of order N, assuming that cumulants $\langle {x_1}^n \rangle_c$ is well-defined (i.e., not divergent). We kind of knew this already in the previous lecture, while the current proof is much more general! For instance, Eqs. 3.32, 3.33, 3.40, 3.41, and 3.54 are all "examples" of the principle that we just proved. Let us, again, appreciate the fact that part of our general result is that the mean value of order N, and the standard deviation is of order $\sqrt{N}$. It then makes a lot of sense to examine a new variable y, defined as follows.

$$y \equiv \frac{x - \bar{x}}{\sigma}. \qquad \bar{x} = N\bar{x}_1, \ \sigma = \sqrt{N}\sigma_{x_1}. \tag{4.23}$$

The key observation to make here is that $\bar{x}_1$ and σ_{x_1} (and more generally, $\langle {x_1}^n \rangle_c$) are quantities defined by the single variable PDF (generally, an unconditional PDF; for an independent case, Eq. 4.17 gives an example) and thus are independent of N. Note that y has 0 mean and 1 standard deviation. Now, using results of Sections 3.1.5 and 3.1.6, we can quickly figure out how various cumulants of y scale as a function of N, for higher orders ($n > 2$).

$$\langle y^n \rangle_c = O\left(N^{\frac{2-n}{2}}\right) \qquad n > 2 \tag{4.24}$$

So, all higher cumulants vanish in the limit of large N. By Eq. 3.41, we see that we get a Gaussian distribution for y. This is the so-called **central limit theorem** for independent equivalent variables. As long as the single variable PDF is reasonably compact so that its moments are well-defined, we get a Gaussian distribution, if we repeat the measurements again and again and average over the results. This is of course very non-trivial, since the single variable PDF can be quite arbitrary, even within this assumption. It could be a uniform distribution, or a set of delta functions (like in Eq. 4.17, or anything in between with

any strange skewness and what not—all of these details do not matter, when the system is sampled a lot of times! Note that sampling a lot of times can be accomplished by tossing the same coin many times or by tossing many coins (assuming that they are manufactured to the same specs) at the same time.

2. **If x_j's are dependent,** then the above argument must be examined further. In Eq. 4.24, we recognize two contributions for the exponent of N: 1 from $\langle x^n \rangle_c$ and $-n/2$ from the scaling of x by $\sigma = O(\sqrt{N})$. In a general case, none of these two conditions need be true. For instance, σ can be as large as $O(N)$, if all correlations are of similar magnitude and non-negligible compared with the single variable variance (Eq. 4.21). We shall limit ourselves to those cases where $\sigma / \langle x \rangle \to 0$ as $N \to \infty$. Then, we must assume $\langle x^2 \rangle_c = O(N^b)$ where $b < 2$. Then, the sufficient condition for the central limit theorem is that $\langle x^n \rangle_c = O(N^c)$ where $c - bn/2 < 0$. These conditions can be re-stated as

$$\sum_{n_1,\ldots,n_N;\sum_j n_j=n} \frac{n!}{n_1! n_2! \cdots n_N!} \langle x_1^{n_1}, x_2^{n_2}, \cdots, x_N^{n_N} \rangle_c = \begin{cases} O(N^b) & b < 2, n = 2 \\ O(N^c) & c < \frac{bn}{2}, n > 2 \end{cases} \quad (4.25)$$

Note that, for the Lorentzian distribution, cumulants diverge, and therefore, the central limit theorem does not apply at all. In fact, one can show that any sum of independent variables, each satisfying Lorentzian statistics, also satisfies the Lorentzian statistics, where both the positions and the widths (Γ) of the constituent distributions simply add up, respectively.

4.3 Strange world of fantastically large numbers

The central limit theorem is one way that the effect of big numbers shows up in a universal way. Now, we shall see that big numbers have some other very interesting rules, as well.

4.3.1 Summation

If $\varepsilon_i > 0$ and $\sim O(\exp(N\phi_i))$ where ϕ_i is a number (regardless of sign) which does not depend on N, then

$$\mathcal{S} \equiv \sum_{i=1}^{\mathcal{N}} \varepsilon_i \sim \varepsilon_{max} \quad (4.26)$$

when N is a large integer, $\mathcal{N}$ is some polynomial of N, and ε_{max} is the maximum of ε_i. Notation: here, $\mathcal{S}$ means a "sum," and does not denote the entropy.

Note the use of the symbol ~ above. Usually, this symbol means "is on the order of." Here, its meaning is a bit looser, in pure mathematical terms: it means "has a similar order of magnitude as." However, as we discuss below, physically the meaning of the symbol ~ here is as good as the meaning of the symbol ≈, as far as big numbers are concerned. So, in future lectures, we will use these two symbols interchangeably.

The proof is quite elementary. $\varepsilon_{max} \le \mathcal{S} \le \mathcal{N}\varepsilon_{max}$. Taking the logarithm, we get $\log \varepsilon_{max} \le \log \mathcal{S} \le \log \mathcal{N} + \log \varepsilon_{max}$. This means that $\log S \approx \log \varepsilon_{max}$ for large N, since $\log \varepsilon_{max} = N\phi_{i_{max}} \gg n \log N \gg \log \mathcal{N}$ ($n = m + 1$, where $\mathcal{N}$ is the m-th order polynomial of N). Thus, $\log S \approx \log \varepsilon_{max}$.

Now, since $\log \mathcal{S} \approx \log \varepsilon_{max}$, it seems to make sense to write $\mathcal{S} \approx \varepsilon_{max}$. However, such an expression could be viewed as something of a bad taste, from a mathematical point of view, since the approximate equality of logarithms really ensures the approximate equality of the order of magnitude only. Indeed, a statement such as $N \exp(N) \approx \exp(N)$ for a large N would not make much sense since the difference between $N \exp(N)$ and $\exp(N)$ is about $N \exp(N)$, i.e. the relative error for the above statement is 100 %! So, we might settle for a slightly weaker notation: $\mathcal{S} \sim \varepsilon_{max}$.

However, this mathematical point is quite moot in most *real* physical cases when N is given by the number of particles or a related large number that is subject to a large fluctuation $\sim \sqrt{N}$ (which is nevertheless small, compared to N). In such a case, one *can* view $N \exp(N)$ and $\exp(N)$ as physically equivalent to each other, since the additional multiplicative factor due to the uncertainty $\sim \exp(\sqrt{N}) \gg N$ or any polynomial of N. Besides, consider also the fact that in characterizing the real systems, we always suffer from uncertainty of the measurement. Such uncertainty, which may be a finite fraction of N in a typical situation, would introduce an even greater range of values for $\exp(N)$.

So, with this in mind, we can freely use $\mathcal{S} \approx \varepsilon_{max}$ for $\mathcal{S} \sim \varepsilon_{max}$.

4.3.2 Integral

Consider the integral

$$\mathcal{J} = \int dx \exp(N\phi(x)), \tag{4.27}$$

where $\int$ is a definite integral over a finite or infinite range. Here, ϕ is a function that is independent of N. When N is large

$$\mathcal{J} \sim \exp(N\phi(x_{max})) \tag{4.28}$$

where the symbol $\sim$ is used in the same meaning as in the previous section, and x_{max} is the x value at which the function $\exp(N\phi(x))$ is the maximum[3].

Note that the above statement is somewhat more disturbing than that of the last section, since, if x has a physical dimension, then $\mathcal{J}$ and $\exp(N\phi(x_{max}))$ are clearly *different* in terms of their dimensions. However, we will see below that this does *not* pose any problem.

To prove the above statement, let us note that, if the integral exists, then it means that

$$\mathcal{J} \approx h\sum_{i=1}^{M}\exp(N\phi(x_i)). \tag{4.29}$$

Here, $h = (x_u - x_l)/M$, where x_u is the upper limit of integration and x_l is the lower limit of integration. Note that the original integral might have an infinite range of integration. Even then, the finite sum approximation must be possible where x_u and x_l are finite numbers and M is a sufficiently large positive integer.

We do not know a priori what the value of M is. Here, we will simply assume that for a given value of N, M is of the order of polynomial of N at the most[4]. Then, by the result of the previous section, we have $\mathcal{J} \sim h\exp(N\phi(x_{max}))$.

Taking the log of $h\exp(N\phi(x_{max}))$ we get

$$\log h + N\phi(x_{max}) = \log(x_u - x_l) - \log M + N\phi(x_{max})$$

For a large N, the first two terms are of the order of $\log N$ at the most, by our assumption, and so they can be ignored, thus proving Eq. 4.28.

The last step of the proof also shows why even the unit of x does not matter. That is, the dimension mismatch between Eqs. 4.27 and 4.28 is *not* important, if we define the unit change transformation as $x = A \cdot 10^p x'$, where p is a physically reasonable integer $|p| \lesssim 50$ and A is a real number with the property $|A| < 10$. By such a transformation, all that will happen above is $\log(x_u - x_l) \to p\log_{10} A + \log(x'_u - x'_l)$.

[3]Or, one of the degenerate maxima, as long as the number of degenerate maxima is of a polynomial order of N.

[4]The grid on which finite sums are evaluated can be made logarithmic, also. This would be a sensible thing to do, if the above integral is evaluated numerically, in practice. The net result of such a sensible operation is a great reduction of the number of summands.

Assuming that $p \ll N$ such a unit change will have no effect to our result here. Note that p is an integer whose magnitude is, at the most about 50, while usually on the order of 10 or 20, for all purposes that we can meaningfully discuss, while N is on the order of 10^{23} for a macroscopic sample. That is, $10^{10^{23}}$ seconds $\sim$ $10^{10^{23}}$ years $\sim$ $10^{10^{23}}$ ages of the Universe[5], if one notes that the age of the Universe is *merely* $\approx 4 \times 10^{17}$ seconds, and one year $\approx 3 \times 10^7$ seconds.

It is helpful to consider the above integral in more details, assuming that there is only one maximum[6]. Now, consider the above integral as approximated as

$$\mathcal{J} = \int dx \exp\left(N\phi_m - \frac{N}{2}|\phi''_m| \cdot (x - x_{max})^2 + \frac{N}{2}\phi'''_m \cdot (x - x_{max})^3 + ... \right) \tag{4.30}$$

where $\phi_m = \phi(x_{max}), \phi''_m = \phi''(x_{max})$, etc. Factoring out the first part outside the integral and ignoring all other terms give the approximation of Eq. 4.28. However, one can obtain a more systematic expansion by changing the variable to $y = \sqrt{\frac{N|\phi''_m|}{2}}(x - x_{max})$. Then, $\mathcal{J} = \exp(N\phi_m)\sqrt{\frac{2}{N|\phi''_m|}} \int dy \exp(-y^2 + \sum_{n=3}^{\infty} a_n y^n)$. An important point is that $a_n = O(N^{1-n/2})$.

$$\begin{aligned} \mathcal{J} &= \exp(N\phi_m)\sqrt{\frac{2}{N|\phi''_m|}} \int dy \exp\left(-y^2 + \sum_{n=3}^{\infty} a_n y^n\right) \\ &\approx \exp(N\phi_m)\sqrt{\frac{2}{N|\phi''_m|}} \int dy \exp(-y^2)\left(1 + a_3 y^3 + ...\right) \end{aligned}$$

Noting that the integral range is given by $y_{min} = -O(\sqrt{N})$ and $y_{max} = O(\sqrt{N})$, it can be thought of as $-\infty$ to ∞ due to the Gaussian function $\exp(-y^2)$. Since the integral $\int_{-\infty}^{\infty} dy \exp(-y^2) = \sqrt{\pi}$ and any integral $\int_{-\infty}^{\infty} dy \exp(-y^2) y^n$ is finite (in particular zero for any odd n due to parity), we get a very useful result

$$\mathcal{J} = \sqrt{\frac{2\pi}{N|\phi''_m|}} \exp(N\phi_m)\left(1 + O\left(\frac{1}{N}\right)\right) \tag{4.31}$$

where the $O(1/N)$ term comes from the fact that a_4 is of that order, while the a_3 term integrates to zero due to parity. We make use of this result in the next section.

[5]Note that here, one can also say that $e^{10^{23}}$ seconds $\sim$ $e^{10^{23}}$ years $\sim$ $e^{10^{23}}$ ages of the Universe. Base e is more directly relevant to the current discussion, while the point made is the same no matter which base one uses.

[6]If there are multiple maxima, then one can break the integral into many finite pieces, when possible to do so. Our result here will apply to each maximum separately, and they will need to be summed up.

4.3.3 Stirling's approximation

$$N! = \Gamma(N+1) = \int_0^\infty dx x^N \exp(-x) \tag{4.32}$$

$$= \int_0^\infty dx \exp(N \log x - x) \tag{4.33}$$

So, $\phi(x) = \log x - x/N$, which has the maximum $\log N - 1$ at $x = N$ and $\phi'' = -1/N^2$ at $x = N$. Applying Eq. 4.31

$$N! = \sqrt{2\pi N} \exp(N \log N - N) \left(1 + O\left(\frac{1}{N}\right)\right) \tag{4.34}$$

In other words,

$$\log N! = N \log N - N + \frac{1}{2} \log(2\pi N) + O\left(\frac{1}{N}\right) \qquad \text{Stirling's formula} \tag{4.35}$$

which is the well-known "Stirling's formula" or "Stirling's approximation" for the factorial/gamma function.