8.333 Fall 2025 Recitation 1-2: Probability Theory

Jessica Metzger
jessmetz@mit.edu | Office hours: Tuesday 4-5pm (8-???)

Statistical mechanics differs from other disciplines (e.g. classical mechanics) by taking a *probabilistic* approach to describing physical systems. We will thus use the first two recitations as an overview of probability theory.

These notes are largely a conglomeration of the previous years' recitation notes by Julien Tailleur, Amer Al-Hiyasat, and Sara Dal Cengio.

References. All the essential information in these recitations can be found in Chapter 2 of Mehran Kardar's *Statistical Physics of Particles.* For a more extensive reference on probability theory/statistics, I recommend *Statistical Inference* (Casella & Berger).

I. PROBABILITY: GENERAL NOTIONS

Whenever we talk about probabilities, we implicitly invoke the precise mathematical notion of a probability space:

Definition I.1: Probability Space

A **probability space** $(\mathcal{S}, P)$ consists of a space of outcomes $\mathcal{S}$ and a probability measure P. For any event $A \subseteq \mathcal{S}$, the probability measure defines its probability $P(A)$, which must satisfy the following:

1. Positivity: for all $A \subseteq \mathcal{S}$, we have $P(A) \geq 0$.
2. Additivity: If $A \cap B = 0$, then $P(A \cup B) = P(A) + P(B)$.
3. Normalization: $P(\mathcal{S}) = 1$.

These 3 conditions are known as the "Kolmogorov axioms of probability". They have some elementary and intuitive consequences, such as:

- The empty set $\varnothing$ has probability zero: $P(\varnothing) = 0$.
- For any event A, its complement obeys $P(\mathcal{S} \setminus A) = 1 - P(A)$.
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ for *arbitrary* $A, B \subseteq \mathcal{S}$.

It's usually easy enough to prove these and similar statements using venn diagrams. Note that for discrete probability distributions, we will often use the shorthand notation for the probability of a singleton set $P(\{x\}) = P(x)$.

The outcomes of a probability space are called random variables:

Definition I.2: Random Variable (RV)

A **random variable (RV)** x belonging to a probability space $(\mathcal{S}, P)$ is an object that takes on values from $\mathcal{S}$ with probability P.

An important remark:

Remark I.1: Random variables vs. their outcomes

Mathematicians make a distinction between a random variable (which they usually denote by a capital letter like X) and a particular *realization/outcome* of that random variable (which they usually denote by a lowercase letter like x). In most disciplines of physics, there is usually no need to make this distinction, so we don't, and call both a random variable and its realizations x.

On a similar note, one must be careful with the notion of random variable. We can have two random variables x and y corresponding to the *same* probability space that have *different* realizations. These are then called **identically-distributed**. However, they can have any degree of correlation or decorrelation with each other.

A. Conditional probabilities

The notion of **statistical independence** is precisely defined as follows:

Definition I.3: Statistical independence

Two events A and B are called **statistically independent** if

$$P(A \cap B) = P(A)P(B)\ . \tag{1}$$

Sometimes we are interested in the probability of some event B, given that another event A will also occur. This is represented by the **conditional probability**.

Definition I.4: Conditional probability

Suppose P is a probability measure on a space of outcomes $\mathcal{S}$. For any events $A, B \subseteq \mathcal{S}$, the probability of B conditioned on A is given by

$$P(B|A) \equiv \frac{P(A \cap B)}{P(A)} \ . \tag{2}$$

In particular, if A and B are statistically independent, then $P(B|A) = P(B)$; i.e. A has no bearing on the likelihood of B.

One straightforward consequence of the definition of conditional probability is **Bayes' theorem.**

Theorem I.1: Bayes' theorem

Consider a probability space $(\mathcal{S}, P)$. For any events $A, B \subseteq \mathcal{S}$,

$$P(A|B)P(B) = P(B|A)P(A) \ . \tag{3}$$

B. Change of variables

If we have a function mapping $\mathcal{S}$ to some other space, what are the statistics of its image, the random variable $y \equiv f(x)$? The answer is given by this relatively straightforward lemma:

Lemma I.1: Change of variable

Consider the probability space $(\mathcal{S}_1, P_1)$ with a random variable x. Suppose there is an arbitrary set $\mathcal{S}_2$ and a function $f : \mathcal{S}_1 \to \mathcal{S}_2$. Define the new probability measure P_2 such that for all subsets $A \subseteq \mathcal{S}_2$

$$P_2(A) = P_1(f^{-1}(A)) \ . \tag{4}$$

Then $f(x)$ is a random variable for the probability space $(\mathcal{S}_2, P_2)$.

(The preimage $f^{-1}(A)$ of A under f is defined as all the elements $x \in \mathcal{S}$ such that $f(x) \in A$.)

We are now done with probability in a generic setting. Before restricting ourselves to probabilities over the real numbers, we will look at a few examples.

C. Examples

Here is an example of a continuous probability space:

Example I.1: Darts

You are a darts player who is not very good. Suppose you always land the dart on the board, but within the board, the placement is completely random. Then we can let $\mathcal{S}$ be the board, and define P such that for any patch of the board A, $P(A) = \text{Area}(A)$. Then, your dart throw is a random variable on the probability space $(\mathcal{S}, P)$.

And here is an example of a discrete probability space:

Example I.2: Two interacting spins

Now consider two spins s_1, s_2 which can be either up $(+)$ or down $(-)$. Then the outcome space is

$$\mathcal{S} = \big\{(s_1, s_2) \;\big|\; s_1, s_2 \in \{-, +\}\big\} = \{(-,-), (-,+), (+,-), (+,+)\} \,. \tag{5}$$

Suppose that the spins have *aligning* interactions: configurations where the spins are aligned a factor of w^2 more likely than configurations where they are anti-aligned. If we define the probability measure P as

$$P\big((+,+)\big) = P\big((-,-)\big) = \frac{w}{2(w+w^{-1})}, \quad P\big((+,-)\big) = P\big((-,+)\big) = \frac{w^{-1}}{2(w+w^{-1})} \,, \tag{6}$$

then the state (s_1, s_2) is a random variable belonging to the probability space $(\mathcal{S}, P)$.

Conditional probabilities. The two spins are not statistically independent: if we let $A_1 = \{(+,+), (+,-)\}$ be the event that s_1 is up and $A_2 = \{(+,+), (-,+)\}$ be the event that s_2 is up, we have

$$P(A_1 \cap A_2) = P\big((+,+)\big) = \frac{w}{2(w+w^{-1})} \,, \quad \text{while} \tag{7}$$

$$P(A_1)P(A_2) = \left(\frac{w}{2(w+w^{-1})} + \frac{w^{-1}}{2(w+w^{-1})}\right)^2 = \left(\frac{1}{2}\right)^2 = \frac{1}{4} \neq P(A_1 \cap A_2) \,. \tag{8}$$

Thus A_1 and A_2 do not satisfy the definition of statistical independence I.3, and we can say that the spins are not independent.

We can calculate the conditional probability that spin 1 is up given spin 2 is up using Def. I.4:

$$P(A_1|A_2) = \frac{P(A_1 \cap A_2)}{P(A_2)} = \frac{P\big((+,+)\big)}{P\big((+,+)\big) + P\big((-,+)\big)} = \frac{w}{w+w^{-1}} \,. \tag{9}$$

If $w > 1$ (correlated spins), then $P(A_1|A_2) > 1/2$; i.e. knowing spin 2 is up means spin 1 is more likely to be up.

II. INTEGER RANDOM VARIABLE

Consider the discrete probability space $(\mathbb{Z}, P)$, where $\mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$ is the set of integers, with random variable n. We will use this simple setting to define notions of probability like moments, cumulants, etc. before we move onto continuous probability spaces in Sec. III.

A. Moments

You are probably familiar with the notion of **expectation value** of a random variable:

Definition II.1: Expectation value in discrete probability spaces

Suppose $(\mathbb{Z}, P)$ is a probability space with random variable n, and $F : \mathbb{Z} \to \mathbb{R}$ is some function. Then the **expectation value** of the random variable $F(n)$ is defined as

$$\langle F(n) \rangle = \sum_{n \in \mathbb{Z}} P(n) F(n) \,. \tag{1}$$

One commonly-used expectation value is the **moments** of a random variable:

Definition II.2: Moments of a discrete random variable

Suppose $(\mathbb{Z}, P)$ is a probability space with random variable n. Let $\ell > 0$ be an integer. Then, the **ℓth moment** μ_ℓ of n is the expectation value of n^ℓ:

$$\mu_\ell \equiv \langle n^\ell \rangle = \sum_{n\in\mathbb{Z}} P(n) n^\ell \ . \tag{2}$$

These can often be easily calculated using a probability space's **characteristic function**

Definition II.3: Characteristic (moment-generating) function of a discrete probability space

Suppose $(\mathbb{Z}, P)$ is a probability space with random variable n. Its **characteristic function**, or **moment-generating function**, $\tilde{P}$ is

$$\tilde{P}(k) \equiv \langle e^{-ikn} \rangle = \sum_{n\in\mathbb{Z}} P(n) e^{-ikn} \ . \tag{3}$$

Here, k is simply a dummy variable. Why is $\tilde{P}$ called the moment-generating function? Expanding its definition, we have

$$\tilde{P}(k) = \sum_{\ell=0}^{\infty} \frac{(-ik)^\ell}{\ell!} \sum_{n\in\mathbb{Z}} P(n) n^\ell = \sum_{n=0}^{\infty} \frac{(-ik)^\ell}{\ell!} \mu_\ell \quad \Longrightarrow \quad \boxed{\mu_\ell = \left(\frac{\partial}{\partial(-ik)} \right)^\ell \tilde{P}(k) \Big|_{k=0}} \ . \tag{4}$$

Note that determining the characteristic function $\tilde{P}(k)$ is sufficient to determine $P(n)$, and vice versa. If two probability measures have the same characteristic function, then they are equal.

B. Cumulants

The **cumulant** is a similar concept that will be very useful in stat mech:

Definition II.4: Cumulant of a discrete probability space

Suppose $(\mathbb{Z}, P)$ is a probability space with characteristic function $\tilde{P}(k)$. The **ℓth cumulant** κ_ℓ is defined as

$$\kappa_m \equiv \langle x^\ell \rangle_c \equiv \left(\frac{\partial}{\partial(-ik)} \right)^\ell \ln \tilde{P}(k) \Big|_{k=0} \ . \tag{5}$$

For this reason, $\ln \tilde{P}(k)$ is called the **cumulant-generating function**. Through a somewhat long proof (which we include in Appendix A), we can show that the moments can be written as sums of the cumulants via a diagrammatic shorthand. To calculate the ℓth moment as a sum of cumulants, consider ℓ points and the different ways of grouping them into p_1 bins of size 1, p_2 bins of size 2, etc. For each such grouping, add $\prod_m \kappa_m^{p_m} \frac{\ell!}{(m!)^{p_m} p_m!}$. The combinatorial factor represents the the degeneracy of this grouping. This has a straightforward diagrammatic representation, which is shown in Fig. 1 for the first four moments.

These equations can then be inverted to give the cumulants in terms of the moments. For future reference, here are the first four cumulants of a random variable x:

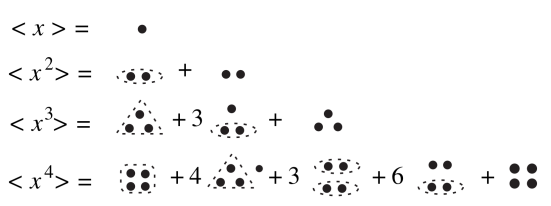
$$\langle x \rangle_c = \langle x \rangle \qquad \text{(mean)} \tag{6}$$

$$\langle x^2 \rangle_c = \langle x^2 \rangle - \langle x \rangle^2 \qquad \text{(variance)} \tag{7}$$

$$\langle x^3 \rangle_c = \langle x^3 \rangle - 3\langle x^2 \rangle \langle x \rangle + 2\langle x \rangle^3 \qquad \text{(skewness)} \tag{8}$$

$$\langle x^4 \rangle_c = \langle x^4 \rangle - 4\langle x^3 \rangle \langle x \rangle - 3\langle x^2 \rangle^2 + 12\langle x^2 \rangle \langle x \rangle^2 - 6\langle x \rangle^4 \qquad \text{(kurtosis)} \ . \tag{9}$$

Now we will look at an example of a discrete probability distribution, the binomial distribution.



$$\langle x\rangle = \langle x\rangle_c\,,$$
$$\langle x^2\rangle = \langle x^2\rangle_c + \langle x\rangle_c^2\,,$$
$$\langle x^3\rangle = \langle x^3\rangle_c + 3\langle x^2\rangle_c\langle x\rangle_c + \langle x\rangle_c^3\,,$$
$$\langle x^4\rangle = \langle x^4\rangle_c + 4\langle x^3\rangle_c\langle x\rangle_c + 3\langle x^2\rangle_c^2 + 6\langle x^2\rangle_c\langle x\rangle_c^2 + \langle x\rangle_c^4$$

FIG. 1:**Left:** the diagrammatic representation of the moments of a random variable x as sums of its cumulants (proof given in Appendix A). **Right:** the corresponding algebraic expressions. [Fig. 2.5 of Kardar's *Statistical Physics of Particles*]

C. Example: the binomial distribution

The binomial distribution is a discrete distribution that models the outcome of multiple (biased) coin flips. Precisely, consider the probability space $(\{0,1,2,\ldots,N\},P)$ where N is a positive integer, and P is defined as

$$P(n) = \binom{N}{n} q^n (1-q)^{N-n} \tag{10}$$

where $q\in[0,1]$, and $\binom{N}{n} = \frac{N!}{n!(N-n)!}$ is the binomial coefficient. This models the outcome of N flips of a biased coin with probability q of landing on heads: the number of "heads" is a random variable in this probability space.

The characteristic function is then

$$\tilde{P}(k) = \sum_{n=0}^{N}\binom{N}{n} q^n(1-q)^{N-n}e^{-ikn} = \sum_{n=0}^{N}\binom{N}{n}(1-q)^{N-n}(qe^{-ik})^n/n'[; = \left(1-q+qe^{-ik}\right)^N\,. \tag{11}$$

so the cumulant generating function is $\ln\tilde{P}(k) = N\ln\left(1+q(e^{-ik}-1)\right)$. Thus, we can calculate the first few cumulants:

$$\langle n\rangle_c = Nq\,,\qquad \langle n^2\rangle_c = Nq(1-q)\,,\qquad \ldots \tag{12}$$

The mean and variance are both proportional to N. Thus the **standard deviation** goes like $\sqrt{N}$. However, the **standard error of the mean** (or **relative uncertainty**) decreases with N.

Definition II.5: Variance, standard deviation, and relative uncertainty

The **variance** σ^2 of a random variable is its first cumulant: $\sigma^2 = \langle x^2\rangle_c$.

The **standard deviation** σ of a random variable is the square root of its variance: $\sigma = \sqrt{\langle x^2\rangle_c}$.

The **relative uncertainty** is the standard deviation divided by the mean $\sigma/\langle x\rangle = \sqrt{\langle x^2\rangle_c}/\langle x\rangle_c$.

Thus, for instance, when you flip very many fair coins (where $q=1/2$), even though its variance increases like N, the proportion of "heads" approaches $1/2$.

Example II.1: Binomial distribution

For a positive integer N and $q\in[0,1]$, the binomial distribution P over the space of outcomes $\{0,1,2,\ldots,N\}$ and its characteristic function $\tilde{P}$ are

$$P(n) = \binom{N}{n} q^n(1-q)^{N-n},\qquad \tilde{P}(k) = \left(1+q(e^{-ik}-1)\right)^N\,. \tag{13}$$

Its first two cumulants are

$$\langle n\rangle_c = Nq\,,\qquad \langle n^2\rangle_c = Nq(1-q)\,. \tag{14}$$

[End of recitation 1]

III. ONE CONTINUOUS RANDOM VARIABLE

Now we will restrict ourselves to probability spaces whose outcomes are the real numbers; i.e. $\mathcal{S} = \mathbb{R}$. Because summation is not defined on this space, it is necessary to define the notion of probability density function in order to calculate moments, cumulants, etc. using integration.

A. Definitions

Definition III.1: Cumulative distribution function (CDF)

For a probability space $(\mathbb{R}, P)$, the **cumulative distribution function (CDF)** $Q : \mathbb{R} \to \mathbb{R}$ is defined as

$$Q(x) = P\big((-\infty, x]\big) \ . \tag{1}$$

i.e. $Q(x)$ is the probability that a random variable from this probability space is $\leq x$.

Definition III.2: Probability density function

For a probability space $(\mathbb{R}, P)$ with cumulative distribution function Q and RV x, the **probability density function (PDF)** $\rho : \mathbb{R} \to \mathbb{R}$ is defined as

$$\rho(x) = \frac{dQ(x)}{dx} \ , \qquad \text{i.e.} \quad Q(x) = \int_{-\infty}^{x} \rho(y) dy \ . \tag{2}$$

There are some important things to understand about the probability density ρ:

Remark III.1: Understanding ρ

- **Relation to P.** The probability *density* ρ is related to the probability *measure* P via the relationship defined on sets:

$$A \subseteq \mathbb{R} \quad \Longrightarrow \quad \int_A \rho(x) dx = P(A) \ . \tag{3}$$

  It's important to understand that they are different, and defined on different domains (P is a function on the subsets of $\mathbb{R}$; ρ is a function on $\mathbb{R}$).
- **Approximate pointwise relationship to P.** However, for a very small interval $[x_0, x_0 + \Delta x]$, we can write the approximate relation (following from Defs. III.1-III.2)

$$\rho(x_0)\Delta x \approx P([x_0, x_0 + \Delta x]) \ . \tag{4}$$

- **Values.** While the probability measure P only takes values ≤ 1, the probability density ρ may be > 1. As long as $\rho(x)$ is positive for all x, and the total integral is normalized to $\int_{\mathbb{R}} \rho(x) dx = 1$, anything else is allowed.
- **Units.** While the probability measure P is unitless, the probability density ρ has units $[1/x]$.

The expectation value on the real numbers is given by:

Definition III.3: Expectation value over real numbers

Consider a random variable x from the probability space $(\mathbb{R}, P)$, and a function $F : \mathbb{R} \to \mathbb{R}$. The expectation value of the random variable $F(x)$ can in fact be written as

$$\langle F(x) \rangle = \int F(x) \rho(x) dx \ . \tag{5}$$

This is analogous to the discrete version, Def. II.1.

B. Change of variables

The change of variables definition from before (Lemma. I.1) tells us how the probability *measure* transforms. But how does the probability *density* transform? First we consider one-to-one maps.

Lemma III.1: Change of variable for the probability density

Consider a probability space $(\mathbb{R}, P_1)$ with probability density ρ_1 and random variable x. Consider a *one-to-one* function $f : \mathbb{R} \to \mathbb{R}$, which defines a new random variable $y \equiv f(x)$ belonging to the probability space $(\mathbb{R}, P_2)$, with P_2 defined in Lem. I.1. The corresponding probability density of y is then given by

$$\rho_2(y) = \rho_1\big(f^{-1}(y)\big)\big|(f^{-1})'(y)\big| = \frac{\rho_1\big(f^{-1}(y)\big)}{|f'\big(f^{-1}(y)\big)|} \,. \qquad \text{Or,} \quad \rho_2\big(f(x)\big) = \frac{\rho_1(x)}{|f'(x)|} \,. \tag{6}$$

If f weren't one-to-one, then $f^{-1}(y)$ would be a set of points, not a single point. Here's a proof of Lem. III.1:

$$\rho_2(y) \approx \frac{P_2([y, y+\Delta y])}{\Delta y} = \frac{P_1\big(f^{-1}([y, y+\Delta y])\big)}{\Delta y} \approx \frac{\rho_1\big(f^{-1}(y)\big)\big|f^{-1}(y+\Delta y) - f^{-1}(y)\big|}{\Delta y} \tag{7}$$

$$\approx \rho_1\big(f^{-1}(y)\big)\big|(f^{-1})'(y)\big| = \frac{\rho_1\big(f^{-1}(y)\big)}{|f'\big(f^{-1}(y)\big)|} \,. \tag{8}$$

This can be written in "shorthand" as follows: conservation of probability tells us that $\rho_1(x)\Delta x = \rho_2(y)\Delta y$. Then, $\rho_2(y) = \rho_1(x)|\Delta x/\Delta y|$. Using $\Delta x \approx |(f^{-1})'(y)\Delta y|$ gives the result (6). This is visualized in Fig. 2.

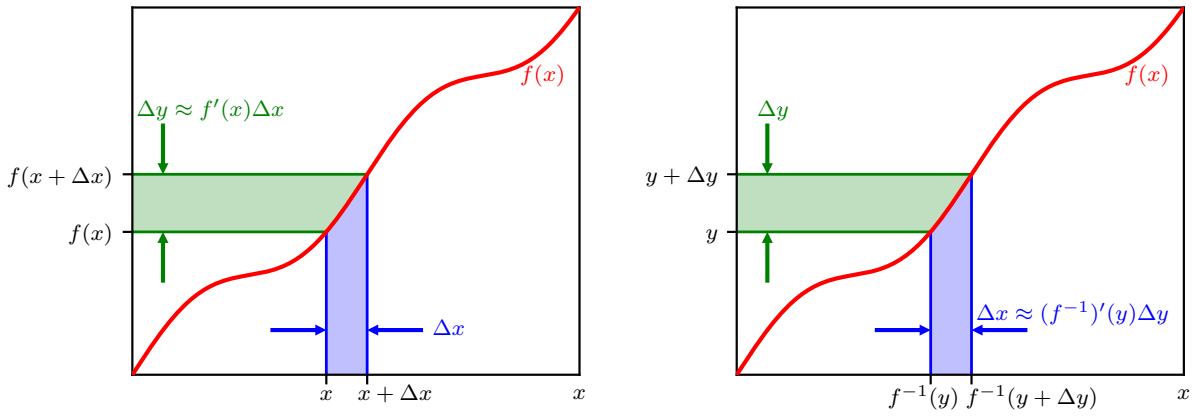


FIG. 2:Visualization of the change of variables for a one-to-one function $f : \mathbb{R} \to \mathbb{R}$, in both x coordinates (left) and y coordinates (right). Because the probability of x falling in the blue region equals the probability of $f(x)$ falling in the green region, the probability density ρ_2 of y must obey $\rho_2(f(x)) = \rho_1(x)/|f'(x)|$ (Lem. III.1).

Remark III.2: Transformations that are not one-to-one

If the change-of-variables transformation f is not one-to-one, the preimage $f^{-1}(y)$ is a set of points, rather than a single point. Thus it is necessary to modify the rule (6) to the following:

$$\rho_2(y) = \sum_{x \in f^{-1}(y)} \frac{\rho_1(x)}{|f'(x)|} \,. \tag{9}$$

This is visualized in Fig. 3.

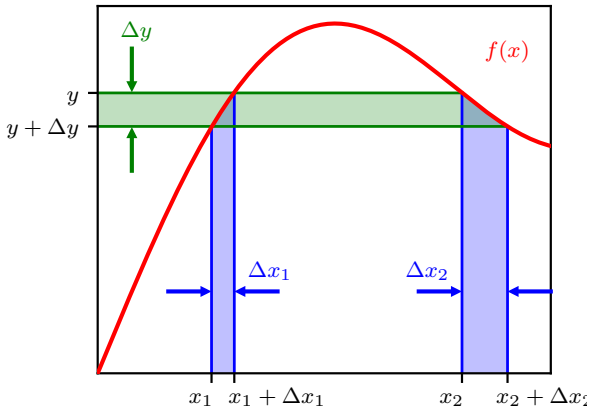


FIG. 3:Visualization of the change of variables for a function $f : \mathbb{R} \to \mathbb{R}$ which is *not* one-to-one. The total probability of x falling in the blue regions equals the probability of $f(x)$ falling in the green region. Thus computing $\rho_2(y)$ requires summing $\rho_1(x_i)/|f'(x_i)|$ for all $x_i \in f^{-1}(y)$ (Eq. (9)).

C. Moments and cumulants

Consider a probability space $(\mathbb{R}, P)$ with random variable x and probability density function ρ. Just as in the discrete case (Def. II.2), the nth moment μ_n is given by the expectation value of x^n. It can be found using the characteristic function, which has a definition analogous to the discrete version (Def. II.3), but involving the probability density:

Definition III.4: Characteristic (moment-generating) function

The **characteristic function** $\tilde{\rho}(k)$ of a probability density ρ is defined as

$$\tilde{\rho}(k) = \int \rho(x) e^{-ikx} dx = \langle e^{-ikx} \rangle \,. \tag{10}$$

As in the discrete case, once we know the characteristic function, we can determine the moments using

$$\mu_n = i^n \frac{\partial^n}{\partial k^n} \tilde{\rho}(k) \bigg|_{k=0} \,. \tag{11}$$

As in the discrete case, determining the $\tilde{\rho}(k)$ is sufficient to determine $\rho(x)$, and vice versa, via the continuous version of the inversion formula:

Lemma III.2: Inversion formula

The probability density function $\rho(x)$ can be found using the characteristic function $\tilde{\rho}(k)$ using the inverse Fourier transform

$$\rho(x) = \int_{-\infty}^{\infty} \frac{dk}{2\pi} \tilde{\rho}(k) e^{ikx} \,. \tag{12}$$

As a consequence, two probability distributions with the same characteristic function are equal.

As in the discrete case, the cumulant-generating function $\ln \tilde{\rho}(k) = \langle e^{-ikn} \rangle$ can be used to determine the cumulants κ_n using

$$\kappa_n \equiv \langle x^n \rangle_c \equiv \left(\frac{\partial}{\partial(-ik)} \right)^n \ln \tilde{\rho}(k) \bigg|_{k=0} \,. \tag{13}$$

Now we will look at an example of a 1d continuous probability space, the 1-dimensional Gaussian distribution.

D. Example: the 1-dimensional normal (Gaussian) distribution

In 1 dimension, the normal (Gaussian) distribution with mean λ and variance σ^2 has probability density

$$\rho(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\lambda)^2}{2\sigma^2} \right] . \tag{14}$$

It is useful to know the integral of a Gaussian function

$$\int_{-\infty}^{\infty} e^{-ax^2+bx} = \sqrt{\frac{2\pi}{a}} e^{b^2/2a} . \tag{15}$$

The Gaussian distribution has characteristic function

$$\tilde{\rho}(k) = \int_{-\infty}^{\infty} \frac{dx}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\lambda)^2}{2\sigma^2} - ikx \right] = \int_{-\infty}^{\infty} \frac{dx}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\lambda+ik\sigma^2)^2}{2\sigma^2} - ik\lambda - \frac{k^2\sigma^2}{2} \right] \tag{16}$$

$$= \exp\left[-ik\lambda - \frac{k^2\sigma^2}{2} \right] . \tag{17}$$

The characteristic function of a Gaussian distribution with mean λ and variance σ^2 is another Gaussian. The cumulant generating function is then

$$\psi(k) = \ln \tilde{\rho}(k) = -ik\lambda - \frac{k^2\sigma^2}{2} \tag{18}$$

which, using Eq. (13), allows us all to calculate all the cumulants

$$\langle x \rangle_c = \lambda , \quad \langle x^2 \rangle_c = \sigma^2 , \quad \langle x^3 \rangle_c = \langle x^4 \rangle_c = \ldots = 0 . \tag{19}$$

All cumulants above the variance are zero. The converse is also true: when an arbitrary probability distribution has zero cumulants at order ≥ 3, it is a normal distribution. This will prove useful later.

Its moments can be found using the diagrammatic approach (Fig. 1), which is greatly simplified by the fact that higher cumulants are zero. For example,

$$\langle x \rangle = \lambda , \quad \langle x^2 \rangle = \sigma^2 + \lambda^2 , \quad \langle x^3 \rangle = 3\sigma^2\lambda + \lambda^3 , \quad \ldots . \tag{20}$$

Example III.1: Summary: 1-dimensional normal distribution

The normal (Gaussian) distribution ρ with mean λ and variance σ^2, and its characteristic function $\tilde{\rho}$, are given by

$$\rho(x) = \frac{\exp\left[-(x-\lambda)^2/2\sigma^2 \right]}{\sqrt{2\pi\sigma^2}} , \qquad \tilde{\rho}(k) = \exp\left[-ik\lambda - \frac{k^2\sigma^2}{2} \right] . \tag{21}$$

Its cumulants are

$$\langle x \rangle_c = \lambda , \qquad \langle x^2 \rangle_c = \sigma^2 , \qquad \langle x^3 \rangle_c = \langle x^4 \rangle_c = \ldots = 0 \tag{22}$$

and its moments are

$$\langle x \rangle = \lambda , \qquad \langle x^2 \rangle = \sigma^2 + \lambda^2 , \quad \langle x^3 \rangle = 3\sigma^2\lambda + \lambda^3 , \quad \ldots , \qquad \ldots . \tag{23}$$

IV. $N \geq 1$ CONTINUOUS RANDOM VARIABLES

We will quickly review probability spaces whose space of outcomes is the space of N-dimensional vectors; i.e. $\mathcal{S} = \mathbb{R}^N$, with random variable $\vec{x} \equiv (x_1, x_2, \ldots, x_N)$. Much of this is straightforward generalizations of the $N = 1$ case.

A. Definitions

Definition IV.1: Notation for integral over real vectors

Consider the function $F : \mathbb{R}^N \to \mathbb{R}$. Then, we denote the integral of F over $\mathbb{R}^N$ as

$$\int d^N\vec{x} F(\vec{x}) \equiv \int_{-\infty}^{\infty} dx_1 \int_{-\infty}^{\infty} dx_2 \ldots \int_{-\infty}^{\infty} dx_N F(\vec{x}) \,. \tag{1}$$

Definition IV.2: Joint cumulative distribution function

For a probability space $(\mathbb{R}^N, P)$, the **joint cumulative distribution function** (CDF) $Q : \mathbb{R}^N \to \mathbb{R}$ is defined as

$$Q(\vec{x}) = P\big((-\infty, x_1] \times (-\infty, x_2] \times \ldots \times (-\infty, x_N]\big) \,. \tag{2}$$

In other words, it is the probability that the first variable is $\leq x_1$, the second variable is $\leq x_2$, etc. Note that if the different coordinates are independent, then Q factorizes: $Q(x_1, \ldots, x_N) = Q(x_1) \cdot \ldots \cdot Q(x_N)$.

Definition IV.3: Joint probability density function

For a probability space $(\mathbb{R}^N, P)$, the **joint probability density function** $\rho : \mathbb{R}^N \to \mathbb{R}$ is defined as

$$\rho(\vec{x}) = \frac{\partial^n Q}{\partial x_1 \partial x_2 \ldots \partial x_N} \,. \tag{3}$$

Similar to the 1-dimensional case, this can be related to the probability measure P by the approximate relation

$$P\Big([x_1, x_1 + \Delta x_1] \times [x_2, x_2 + \Delta x_2] \times \ldots \times [x_N, x_N + \Delta x_N]\Big) \approx \rho(\vec{x}) \Delta x_1 \Delta x_2 \ldots \Delta x_N \,. \tag{4}$$

The expectation value of an observable $F(\vec{x})$ can also be found using integration against $\rho(\vec{x})$, i.e. $\langle F(\vec{x}) \rangle = \int d^N\vec{x} F(\vec{x}) \rho(\vec{x})$.

I'll skip the change of variable formula for $N > 1$ dimension, but make a quick remark:

Remark IV.1: Change of variables for N-dimensional probability distribution

For a single variable, the change of variable formula is given by Lem. I.1, a consequence of the conservation of the probability within an interval. In $N > 1$ dimensions, probability is now conserved within a *volume*. So for a small box $B(\vec{x})$ located at $\vec{x}$, we can say "$\rho(\vec{x}) \mathrm{Vol}\big[B(\vec{x})\big] \approx \rho\big(\vec{f}(\vec{x})\big) \mathrm{Vol}\big[\vec{f}\big(B(\vec{x})\big)\big]$". This results in an analogous formula where the derivative becomes the Jacobian determinant of $\vec{f}$:

$$\rho\big(\vec{f}(\vec{x})\big) = \frac{\rho(\vec{x})}{|\det \mathcal{J}_{\vec{f}}(\vec{x})|} \,. \tag{5}$$

However, this will likely not be necessary for this course.

B. Marginal and conditional probabilities

Sometimes, we are interested in the probability density function for only a few of the coordinates, e.g. $(x_1, x_2, \ldots, x_M)$ with $M < N$. This requires defining a new probability space with the following PDF:

Definition IV.4: Marginal or unconditional probability density function

Let $\vec{x} = (x_1, x_2, \ldots, x_N)$ be a random variable from the probability space $(\mathbb{R}^N, P)$ with probability density ρ. Consider an integer $M < N$. The **marginal** or **unconditional probability density function** (PDF) for the first M coordinates is given by

$$\rho_M(x_1, \ldots, x_M) \equiv \int dx_{M+1} \ldots dx_N \rho(\vec{x}) \, . \tag{6}$$

Suppose we know the values of the first M coordinates, and are interested in the probability density for the last $N - M$ coordinates. For this, we define the **conditional** probability density function:

Definition IV.5: Conditional probability density function

Let $\vec{x} = (x_1, x_2, \ldots, x_N)$ be a random variable from the probability space $(\mathbb{R}^N, P)$ with probability density ρ. Consider an integer $M < N$. The **conditional probability density function** for the last $N - M$ coordinates is given by

$$\rho_{(N-M)|M}(x_{M+1}, \ldots, x_N | x_1, \ldots, x_M) = \frac{\rho(x_1, \ldots, x_N)}{\rho_M(x_1, \ldots, x_M)} \, . \tag{7}$$

Note the normalization: $\int dx_{M+1} \ldots dx_N \rho_{(N-M)|M} = 1$. This definition makes sense in light of the definition of the conditional probability *measure* (Def. I.4), but adapted for probability *densities*:

$$\rho(A|B) = \frac{\rho(A, B)}{\rho(B)} \, . \tag{8}$$

Often, you will hear about two random variables being **independent** of each other, which is defined as

Definition IV.6: Independent random variables

For a probability space $(\mathbb{R}^2, P)$ with random variable $\vec{x} = (x_1, x_2)$ and probability density ρ, the variables x_1 and x_2 are said to be **independent**, and ρ is said to **factorize**, if it is possible to write

$$\rho(x_1, x_2) = \rho_1(x_1)\rho_2(x_2) \, , \quad \text{where} \quad \rho_i(x_i) \equiv \int dx_{j \neq i} \rho(\vec{x}) \, . \tag{9}$$

C. Moments and cumulants

Just like the 1-dimensional case, moments and cumulants can be calculated using the N-dimensional characteristic function, $\tilde{\rho}(\vec{k}) = \langle e^{-i\vec{k}\cdot\vec{x}} \rangle$. The moments are given by:

Definition IV.7: N-dimensional moments

The moments of an N-dimensional probability distribution are are given by:

$$\langle x_1^{n_1} x_2^{n_2} \ldots x_N^{n_N} \rangle = \left[i \frac{\partial}{\partial k_1} \right]^{n_1} \left[i \frac{\partial}{\partial k_2} \right]^{n_2} \ldots \left[i \frac{\partial}{\partial k_N} \right]^{n_N} \tilde{\rho}(\vec{k}) \Big|_{\vec{k}=0} \, , \quad \text{where} \quad \tilde{\rho}(\vec{k}) \equiv \langle e^{-i\vec{k}\cdot\vec{x}} \rangle \, . \tag{10}$$

This is evident once we Taylor expanding the moment-generating function:

$$\langle e^{-i\vec{k}\cdot\vec{x}} \rangle = \sum_{n_1=0}^{\infty} \sum_{n_2=0}^{\infty} \ldots \sum_{n_N=0}^{\infty} \frac{(-ik_1)^{n_1}}{n_1!} \frac{(-ik_2)^{n_2}}{n_2!} \ldots \frac{(-ik_N)^{n_N}}{n_N!} \langle x_1^{n_1} x_2^{n_2} \ldots x_N^{n_N} \rangle \, . \tag{11}$$

Likewise, we define the cumulants using the cumulant-generating function $\ln \tilde{\rho}(\vec{k})$ as follows:

Definition IV.8: N-dimensional cumulants

The cumulants of an N-dimensional probability distribution are are given by:

$$\langle x_1^{n_1} * x_2^{n_2} * \ldots * x_N^{n_N} \rangle_c = \left[i\frac{\partial}{\partial k_1} \right]^{n_1} \left[i\frac{\partial}{\partial k_2} \right]^{n_2} \ldots \left[i\frac{\partial}{\partial k_N} \right]^{n_N} \ln \tilde{\rho}(\vec{k}) \Big|_{\vec{k}=0} . \tag{12}$$

The relationship between moments and cumulants can then be written as

$$\exp\left(\sum_{m_1=1}^{\infty} \ldots \sum_{m_N=1}^{\infty} \frac{k_1^{m_1}}{m_1!} \ldots \frac{k_N^{m_N}}{m_N!} \langle x_1^{m_1} * \ldots * x_N^{m_N} \rangle_c \right) = \sum_{n_1=1}^{\infty} \ldots \sum_{n_N=1}^{\infty} \frac{k_1^{n_1}}{n_1!} \ldots \frac{k_N^{n_N}}{n_N!} \langle x_1^{n_1} \ldots x_N^{n_N} \rangle . \tag{13}$$

This allows calculating the moments as sums and products of the cumulants. This relationship can again be simplified with a graphical representation (see Fig. 1), but we will skip it here. It is useful, however, to know the 2nd-order relationship:

$$\langle x_n * x_m \rangle_c = \langle x_n x_m \rangle - \langle x_n \rangle \langle x_m \rangle \equiv \mathrm{Corr}[x_n, x_m] . \tag{14}$$

This is easy enough to check using Eq. (13). This defines the notion of **correlation** between variables x_n and x_m. If x_n and x_m are independent random variables, $\mathrm{Corr}[x_n, x_m] = 0$.

Now we will look at an example of a N-dimensional continuous probability distribution: the Gaussian distribution.

D. Example: the N-dimensional normal (Gaussian) distribution

In $N \geq 1$ dimensions, the Gaussian distribution with mean $\vec{\lambda}$ and symmetric, positive-definite correlation matrix $\mathcal{C}$ takes the form

$$\rho(\vec{x}) = \frac{\exp\Big[-(\vec{x}-\vec{\lambda})^T \mathcal{C}^{-1} (\vec{x}-\vec{\lambda})/2 \Big]}{\sqrt{(2\pi)^N |\det \mathcal{C}|}} = \frac{1}{\sqrt{(2\pi)^N \det \mathcal{C}}} \exp\left[-\frac{1}{2} \sum_{m,n} (\mathcal{C}^{-1})_{mn} (x_m - \mu_m)(x_n - \mu_n) \right] . \tag{15}$$

We can understand the normalization as follows: $\mathcal{C}$ can be diagonalized by a unitary matrix, which we denote by $\mathcal{U}$. Write the diagonal matrix of its eigenvalues as Δ, so that $\mathcal{C} = \mathcal{U}\Delta\mathcal{U}^T$ and $\mathcal{C}^{-1} = \mathcal{U}\Delta^{-1}\mathcal{U}^T$. We also have $\det \mathcal{C} = \det \Delta$. Then, define $\vec{y} \equiv \vec{f}(\vec{x}) = \mathcal{U}^T(\vec{x} - \vec{\lambda})$. This transformation has unit Jacobian determinant, the change of variables is simply $\rho(\vec{x}) = \rho_{\vec{y}}(\vec{f}(\vec{x}))$, where

$$\rho_{\vec{y}}(\vec{y}) = \frac{\exp\Big[-\vec{y}^T \Delta^{-1} \vec{y}/2 \Big]}{\sqrt{(2\pi)^N |\det \Delta|}} = \prod_{n=1}^{N} \frac{e^{-y_n^2/2\Delta_{nn}}}{\sqrt{2\pi\Delta_{nn}}} . \tag{16}$$

This is the product of N independent, properly-normalized Gaussians; thus its integral with respect to $\vec{y}$ is 1. Then, the integral of $\rho(\vec{x}) = \rho_{\vec{y}}(\vec{f}(\vec{x}))$ with respect to $\vec{x}$ must also be 1, again because the change of variables from $\vec{y}$ to $\vec{x}$ is unitary.

The characteristic function can be found via a similar computation:

$$\tilde{\rho}(\vec{k}) = \int d^d\vec{x} \frac{\exp\Big[-(\vec{x}-\vec{\lambda})^T \mathcal{C}^{-1} (\vec{x}-\vec{\lambda})/2 - i\vec{k}\cdot\vec{x} \Big]}{\sqrt{(2\pi)^N |\det \mathcal{C}|}} = \int d^d\vec{y} \frac{\exp\Big[-\vec{y}^T \Delta^{-1} \vec{y}/2 - i\vec{k}^T(\mathcal{U}\vec{y} + \vec{\lambda}) \Big]}{\sqrt{(2\pi)^N |\det \Delta|}} . \tag{17}$$

Defining $\vec{q} \equiv \mathcal{U}^T\vec{k}$, we have

$$\begin{aligned} \tilde{\rho}(\vec{k}) &= \int d^N\vec{y} \frac{\exp\Big[-\vec{y}^T \Delta^{-1} \vec{y}/2 - i\vec{q}\cdot\vec{y} - i\vec{k}\cdot\vec{\lambda} \Big]}{\sqrt{(2\pi)^N |\det \Delta|}} = \prod_{n=1}^{N} \int dy_n \frac{\exp\Big[-y_n^2/2\Delta_{nn} - iq_n y_n - ik_n\lambda_n \Big]}{\sqrt{2\pi\Delta_{nn}}} \\ &= \exp\left[-i\vec{k}\cdot\vec{\lambda} - \sum_{n=1}^{N} \frac{q_n^2 \Delta_{nn}}{2} \right] = \exp\left[-i\vec{k}\cdot\vec{\lambda} - \sum_{n,m,\ell} \frac{\mathcal{U}_{mn}\mathcal{U}_{\ell n} k_m k_\ell \Delta_{nn}}{2} \right] = \exp\left[-i\vec{k}\cdot\vec{\lambda} - \sum_{n,m,\ell} \frac{k_n \mathcal{C}_{nm} k_m}{2} \right] , \end{aligned} \tag{18}$$

giving finally

$$\tilde{\rho}(\vec{k}) = \exp\left[-i\vec{k}\cdot\vec{\lambda} - \frac{\vec{k}^T\mathcal{C}\vec{k}}{2}\right] . \tag{19}$$

We can thus easily find the joint cumulants of the variables

$$\langle x_n\rangle_c = \lambda_n \ , \qquad \langle x_n * x_m\rangle_c = \mathcal{C}_{mn} \ . \tag{20}$$

The elements of the correlation matrix are exactly the correlations between x_n and x_m. Again, all other higher cumulants are zero.

The moments can be found using the diagrammatic method (which we didn't discuss for $N > 1$), or by hand starting with Eq. (13). However, in the zero-mean case where $\vec{\lambda} = \vec{0}$, there is a useful formula which can be deduced diagrammatically or from Eq. (13).

Theorem IV.1: Wick's theorem

Suppose $\vec{x}$ is a zero-mean Gaussian random variable with correlation matrix $\mathcal{C}$. The moments are given by

$$\langle x_{n_1}x_{n_2}\dots x_{n_m}\rangle = \sum_{p_m}\prod_{(i,j)\in p_m}\mathcal{C}_{n_in_j} \ , \tag{21}$$

where p_m is the set of all possible pairings of $\{1,2,\dots,m\}$.

For example,

$$\langle x_1x_2x_3x_4\rangle = \mathcal{C}_{12}\mathcal{C}_{34} + \mathcal{C}_{13}\mathcal{C}_{24} + \mathcal{C}_{14}\mathcal{C}_{23} \ . \tag{22}$$

These definitions and results are summarized below:

Example IV.1: N-dimensional Gaussian distribution

In $N \geq 1$ dimensions, the Gaussian distribution ρ with mean $\vec{\lambda}$ and correlation matrix $\mathcal{C}$, and its characteristic function $\tilde{\rho}$, are given by

$$\rho(\vec{x}) = \frac{\exp\left[-(\vec{x}-\vec{\lambda})^T\mathcal{C}^{-1}(\vec{x}-\vec{\lambda})/2\right]}{\sqrt{(2\pi)^N|\det\mathcal{C}|}} \ , \qquad \tilde{\rho}(\vec{k}) = \exp\left[-i\vec{k}\cdot\vec{\lambda} - \frac{\vec{k}^T\mathcal{C}\vec{k}}{2}\right] . \tag{23}$$

Its cumulants are

$$\langle x_n\rangle_c = \lambda_n \ , \qquad \langle x_n * x_m\rangle_c = \mathcal{C}_{nm} \ , \qquad \langle x_n * x_m * x_\ell \dots\rangle_c = 0 \ . \tag{24}$$

If $\vec{\lambda} = \vec{0}$, the moments can be found using Wick's theorem (Thm. IV.1).

V. SUMS OF CONTINUOUS RANDOM VARIABLES

Consider a 2-dimensional probability space $(\mathbb{R}^2, P)$, with probability density ρ and random variable $\vec{x} = (x_1, x_2)$. Consider a new random variable $X \equiv x_1 + x_2$. What is the probability distribution ρ_X of X?

For this, we can use the helpful identity

$$y = f(\vec{x}) \quad \Longrightarrow \quad \rho_y(y) = \int d\vec{x}\delta\big(y - f(\vec{x})\big)\rho_{\vec{x}}(\vec{x}) = \langle\delta(y - f(\vec{x}))\rangle_{\vec{x}} \tag{1}$$

which gives

$$\rho_X(X) = \int dx_1 \int dx_2\delta(x_1 + x_2 - X)\rho(x_1, x_2) = \int dx_1\rho(x_1, X - x_1) \ . \tag{2}$$

The characteristic function takes on the nice form

$$\tilde{\rho}_X(k) = \langle e^{-ikX} \rangle = \int dX \rho_X(X) e^{-ikX} = \int dX \left(\int dx_1 \int dx_2 \delta(x_1 + x_2 - X) \rho(x_1, x_2) \right) e^{-ikX} \tag{3}$$

$$= \int dx_1 \int dx_2 \rho(x_1, x_2) e^{-ik(x_1+x_2)} = \tilde{\rho}(k, k) \; . \tag{4}$$

It's simply the characteristic function of the original distribution, with each entry evaluated at the same k.

[End of recitation 2]

Now consider a probability space of arbitrary dimension N, $(\mathbb{R}^N, P)$ with probability density ρ and random variable $\vec{x} = (x_1, \ldots, x_N)$. Consider the random variable $X \equiv \sum_{i=1}^N x_i$. Its probability distribution ρ_X and characteristic function $\tilde{\rho}_X$ are then, similarly,

$$\rho_X(X) = \int d^N \vec{x} \rho(\vec{x}) \delta\left(X - \sum_{i=1}^N x_i \right) = \int \left(\prod_{i=1}^{N-1} dx_i \right) \rho(x_1, \ldots, x_{N-1}, X - x_1 - \ldots - x_{N-1}) \tag{5}$$

$$\tilde{\rho}_X(k) = \int dX \rho_X(X) e^{-iKX} = \int dX \int d^N \vec{x} \rho(\vec{x}) \delta\left(X - \sum_{i=1}^N x_i \right) e^{-ikX} = \int d^N \vec{x} \rho(\vec{x}) e^{-ik \sum_i x_i} = \tilde{\rho}(k, k, \ldots, k) \; . \tag{6}$$

Thus, we can use the cumulant generating function

$$\ln \tilde{\rho}_X(k) = \ln \tilde{\rho}(k, k, \ldots, k) = -ik \sum_{i=1}^N \langle x_i \rangle_c + \frac{(-ik)^2}{2!} \sum_{i,j=1}^N \langle x_i * x_j \rangle_c + \ldots \tag{7}$$

to calculate the cumulants of X, such as

$$\langle X \rangle_c = \sum_{i=1}^N \langle x_i \rangle_c \; , \qquad \langle X^2 \rangle_c = \sum_{i,j=1}^N \langle x_i * x_j \rangle_c \; , \qquad \langle X^3 \rangle_c = \sum_{i,j,k=1}^N \langle x_i * x_j * x_k \rangle_c \; , \qquad \ldots \; . \tag{8}$$

A. Sums of independent random variables

A common case is when the random variables x_i are independent, i.e. $\rho(\vec{x}) = \prod_{i=1}^N \rho_i(x_i)$ with characteristic function $\tilde{\rho}(\vec{k}) = \prod_{i=1}^N \tilde{\rho}_i(k_i)$. Then, cumulants linking different variables are zero, and the cumulants of their sum X (Eq (8)) reduce to

$$x_i \text{ independent} \qquad \Longrightarrow \qquad \langle X^n \rangle_c = \sum_{i=1}^N \langle x_i^n \rangle_c \; . \tag{9}$$

Suppose the variables x_i are, moreover, identically distributed. So, we say they are **independent and identically distributed (iid)**. Then, we have

$$x_i \text{ iid} \qquad \Longrightarrow \qquad \langle X^n \rangle_c = N \langle x_i^n \rangle_c \; . \tag{10}$$

This generalizes the result we found for the binomial distribution: The mean and variance are both proportional to N. Consider a new random variable, Y:

$$Y \equiv \frac{X - \langle X \rangle_c}{\sqrt{N}} \; . \tag{11}$$

All cumulants of Y scale like N, and therefore the nth cumulant of Y scales like $N/(\sqrt{N})^n = N^{1-n/2}$. In the limit as $N \to \infty$, only the variance remains; all other cumulants approach zero. But the only probability distribution with zero cumulants above 2 is the normal distribution! We have just proven a weaker version of the **central limit theorem**:

Theorem V.1: Central limit theorem

Consider N identically distributed random variables $\{x_i\}$. Suppose their correlations are weak enough so that

$$\sum_{i_1=1}^{N} \dots \sum_{i_m=1}^{N} \langle x_{i_1} * \dots * x_{i_m} \rangle_c \ll \mathcal{O}(N^{m/2}) \,. \tag{12}$$

Then, the probability density of their sum approaches a Gaussian distribution:

$$\lim_{N\to\infty} \rho\left(y \equiv \frac{\sum_{i=1}^{N} x_i - N\langle x\rangle_c}{\sqrt{N}} \right) = \frac{1}{\sqrt{2\pi\langle x^2\rangle_c}} \exp\left[-\frac{y^2}{2\langle x^2\rangle_c} \right] . \tag{13}$$

Appendix A: Proof: diagrammatic shorthand for cumulants

Here we prove the diagrammatic shorthand for writing the moments of any probability distribution in terms of its cumulants, first defined in Sec. II B.

The cumulants can be written in terms of the moments by exponentiating its Taylor series $\psi(k) = \sum_m \frac{(-i)^m}{m!} k^m \kappa_m$ and absorbing $-i$ into k by replacing $k \to -ik$, to find

$$\exp\left[\sum_{m=1}^{\infty} \frac{k^m}{m!} \kappa_m \right] = \exp\left[\psi(k)\right] = \hat{\rho}(k) = \sum_{n=0}^{\infty} \frac{k^n}{n!} \mu_n \,. \tag{A1}$$

We can further re-write the exponential

$$\sum_{n=0}^{\infty} \frac{k^n}{n!} \mu_n = \exp\left[\sum_{m=1}^{\infty} \frac{k^m}{m!} \kappa_m \right] = \prod_{m=1}^{\infty} \exp\left[\frac{k^m}{m!} \kappa_m \right] = \prod_{m=1}^{\infty} \sum_{\ell=0}^{\infty} \frac{1}{\ell!} \left(\frac{k^m}{m!} \kappa_m \right)^{\ell} . \tag{A2}$$

For each n, define the set $\mathcal{I}_n$, which represents all the unordered partitions of $\{1, \dots, n\}$:

$$\mathcal{I}_n = \left\{ \{\ell_1, \ell_2, \dots\} \;\middle|\; \sum_{m=1}^{\infty} m\ell_m = n \right\} . \tag{A3}$$

Here are some examples for small n:

$$\mathcal{I}_1 = \big\{ \{1, 0, 0, \dots\} \big\} \tag{A4}$$

$$\mathcal{I}_2 = \big\{ \{0, 1, 0, 0, \dots\}, \{2, 0, 0, \dots\} \big\} \tag{A5}$$

$$\mathcal{I}_3 = \big\{ \{0, 0, 1, 0, 0, \dots\}, \{1, 1, 0, 0, \dots\}, \{3, 0, 0, \dots\} \big\} \tag{A6}$$

$$\mathcal{I}_4 = \big\{ \{0, 0, 0, 1, 0, 0, \dots\}, \{1, 0, 1, 0, 0, \dots\}, \{0, 2, 0, 0, \dots\}, \{2, 1, 0, 0, \dots\}, \{4, 0, 0, \dots\} \big\} \,. \tag{A7}$$

Think of each set in $\mathcal{I}_n$ as representing a different way to put n points into ℓ_1 bins of size 1, ℓ_2 bins of size 2, etc. Then,

$$\sum_{n=0}^{\infty} \frac{k^n}{n!} \mu_n = \sum_{n=0}^{\infty} k^n \sum_{\{\ell_m\}\in\mathcal{I}_n} \prod_{m=1}^{\infty} \frac{\kappa_m^{\ell_m}}{(m!)^{\ell_m} \ell_m!} = \sum_{n=0}^{\infty} \frac{k^n}{n!} \sum_{\{\ell_m\}\in\mathcal{I}_n} \prod_{m=1}^{\infty} \kappa_m^{\ell_m} \frac{n!}{\prod_{m=1}^{\infty} (m!)^{\ell_m} \ell_m!} \tag{A8}$$

$$\implies \mu_n = \sum_{\{\ell_m\}\in\mathcal{I}_n} \prod_{m=1}^{\infty} \kappa_m^{\ell_m} \frac{n!}{\prod_{m=1}^{\infty} (m!)^{\ell_m} \ell_m!} \,. \tag{A9}$$

The combinatorial factor gives the number of ways to put n items in ℓ_1 bins of size 1, ℓ_2 bins of size 2, etc. without caring about the order within each bin (hence the $1/m!$) or the order of equal-sized bins (hence the $1/\ell_m!$). This brings us to our diagrammatic shorthand, shown in Fig. 1 for the first four moments.