Lecture 2: Basic mathematical concepts from probability and information theory.

1 Basic probability concepts

Suppose we have a process (referred to as a random variable) where there are a finite number of possible outcomes $\{x_i : i = 1, \ldots, N\}$ to each trial of the system. If the system is observed for a large number of trials then the frequency of occurrence of each outcome will be approximately fixed and indeed as the limit of the number of trials approaches infinity we assume that this frequency divided by the number of trials approaches a value which we call the probability of a particular outcome written $p(x_i)$. Obviously the sum of probabilities by this definition is unity. Now if instead we wish to consider a process whose outcomes come from an infinite set of cardinality that of the continuum then considerable care needs to be exercised in defining probability which leads to the study of measure theory. For our purposes we consider the following simplified and less than rigorous approach relevant to many problems in physics. Suppose the continuum of outcomes can be mapped to R^n and define a countable number of intervals for each dimension of equal length Δx. We can now adopt the process for a finite outcome set and extend it to the countable sets of outcome interval volumes which will define a probability P_i for each interval volume in R^n. Associate now a probability density p_i with each interval by dividing P_i by each interval volume:

$$P_i = p_i \left(\Delta x\right)^n$$

and like the finite case we have the normalization condition

$$\sum_i p_i \left(\Delta x\right)^n = 1$$

We can now define a probability density $p(x)$ for any point $x \in R^n$ by taking the limit $\Delta x \to 0$ and then the normalization condition sum becomes an integral:

$$\int p(x_1, x_2, \ldots, x_n) dx_1 dx_2 \ldots dx_n = 1$$

and the probabilities of intervals can be recovered using definite integrals of the relevant density function.

2 Information theory

An important functional defined on random variables concerns the associated "uncertainty". Suppose firstly that the finite outcome case applies and that $p(x_i)$ is close to unity for a particular x_i then intuitively we say the random variable has very small uncertainty associated with it since the outcome of the trials is almost always the same. Alternatively if $p(x_i) = \frac{1}{N}$ then the reverse situation

applies since any outcome is equally as likely. Information theory provides an important measure of this uncertainty through the Shannon entropy H

$$H \equiv -\sum_{i=1}^{N} p(x_i) \log p(x_i) \tag{1}$$

It is easily checked that the first case above has $H \sim 0$ while the second case maximizes the entropy with a value of $H = \log N$. The entropy also has the property that if an outcome x_i is split into a set of sub-outcomes $y_{ij} : j = 1, \ldots, M$ then the uncertainty increases in an intuitive way viz

$$\begin{aligned} H_{new} &= H_{old} + p(x_i) H_i(y) \\ H_i(y) &\equiv -\sum_{j=1}^{M} q(y_{ij}) \log q(y_{ij}) \\ q(y_{ij}) &\equiv \frac{p(y_{ij})}{p(x_i)} \end{aligned}$$

Thus the entropy of the sub-outcomes created considered on their own is added to the original entropy weighted by the likelihood of the original outcome which was split. This property is easily verified from the definition and was actually used as an intuitively appealing axiom by Shannon to derive the formula (1).

Another important functional is the relative entropy which is a measure of the distance between probability distributions. It is defined as

$$D(p||q) \equiv \sum_{i=1}^{N} p(x_i) \log \frac{p(x_i)}{q(x_i)}$$

Intuitively it is the "inaccuracy induced" in assuming that a given random variable has a distribution q when in reality the distribution is p. An important property of this functional is that it vanishes if and only if p and q are identical. Notice that in general that $D(p||q) \neq D(q||p)$ a fact not inconsistent with the intuitive meaning.

It is possible to generalize these functionals to the continuum outcome case providing a little care is taken with the limiting process. Consider firstly the entropy: The formula above is easily applied to the subinterval formalism used in the limiting definition of a density. Thus we can write

$$H = -\sum_i P_i \log P_i = -\sum_i (\Delta x)^n \, p_i \log p_i - n \log \Delta x$$

In the limit $\Delta x \to 0$ the first term approaches the well defined limit

$$h \equiv -\int (p \log p) d^n x$$

however the second term diverges to $+\infty$. The first term is referred to as the differential entropy. Such a situation is curiously similar to the issue we faced in the previous lecture where we defined the statistical physics entropy as the differential entropy over phase space minus an additional constant which was related to the quantum lower limit on phase space resolution i.e. the equivalent of $(\Delta x)^n$.

The limiting process is more straightforward for the relative entropy:

$$D(P||Q) = \sum_i P_i \log \frac{P_i}{Q_i} = \sum_i (\Delta x)^n \, p_i \log \frac{p_i}{q_i}$$

The limit of the right hand term exists in general and we so we obtain

$$D(p||q) \equiv \int p \log \frac{p}{q} d^n x$$

As we have seen the differential entropy as defined above is directly applicable to the statistical physics case. As we shall see in Lecture 8, the relative entropy is useful is analyzing the equilibration process. The negative of the relative entropy is commonly used in the statistical physics literature where it is referred to as the conditional entropy. Notice that if we formally[1] set q as uniform then the negative of the relative entropy becomes the regular differential entropy.

3 Application to statistical physics

3.1 Liouville Equation

Consider a closed system that consists of a large number of still macroscopic subsystems that are interacting with each other. A concrete example would be a gas where various macroscopic quantities such as total energy can be measured for small but still macroscopic regions of the total gas volume. Given a short enough time scale[2] the number of particles in each subsystem that are in interaction with other subsystems will be small relative to the total number of particles within any subsystem[3]. Thus the macroscopic properties of each subsystem will evolve independently at least on this short time scale. The implication of this weak interaction of subsystems is that the probability densities of each of the subsystems will be statistically independent i.e.

$$\varrho_{12} = \varrho_1 \varrho_2$$

where ϱ_{12} is the joint probability density for the composite system of subsystem 1 and subsystem 2. ϱ_1 and ϱ_2 are the marginal distributions obtained by

[1] a uniform distribution only makes sense on a sub-manifold of the phase space so this formal exercise needs some care. It does make sense in the finite outcome case.

[2] This will be short compared to the thermal relaxation time of the whole closed system but much longer than the timescale of molecular motion and interaction i.e. the microscopic timescale

[3] Due to this the subsystem is said to be quasi-closed since only a very small part of it is interacting with it's surroundings.

integrating out variables from the other system. Consider now a macrovariable f of interest and let the value in the two subsystems be f_1 and f_2. As microscopic time proceeds within the short macroscopic time mentioned above these will both fluctuate about their mean values namely $\bar{f}_1$ and $\bar{f}_2$ where the means will be simply expectations with respect to the densities i.e. we will have

$$\bar{f}_i = \int f_i \varrho_{12\ldots i\ldots,N} = \int f_1 \varrho_1$$

where there are N subsystems. The statistical independence of subsystems means now that

$$\overline{f_1 f_2} = \bar{f}_1 \bar{f}_2$$

Consider now a subsystem on the timescale mentioned above. Clearly many different microscopic configurations could occur and would be consistent with the assumed density ϱ. Consider a large number of such microscopic configurations. Such a group is commonly referred to as a statistical ensemble. The density of such points in phase space of the ensemble will intuitively follow the density ϱ since they represent a sample of this density. Consider now the time evolution of such a cloud of phase space points or more precisely of the density ϱ. It may be regarded as a fluid in which matter is conserved since each point is conserved as time proceeds. The conservation of mass equation in a fluid in the $2N$ dimensional phase space reads

$$\frac{\partial \varrho}{\partial t} + \nabla \cdot (\varrho \dot{\mathbf{x}}) = 0$$

where the vector $\mathbf{x}$ is the phase space co-ordinates which in this case are the position and momenta of the N particles within the subsystem. Explicitly then we have

$$\frac{\partial \varrho}{\partial t} + \sum_{i=1}^{N} \left[\frac{\partial}{\partial q_i} (\varrho \dot{q}_i) + \frac{\partial}{\partial p_i} (\varrho \dot{p}_i) \right] = 0 \tag{2}$$

Now the laws of mechanical particles follow very particular types of equations namely those of Hamilton which read

$$\begin{aligned} \dot{q}_i &= \frac{\partial H}{\partial p_i} \\ \dot{p}_i &= -\frac{\partial H}{\partial q_i} \end{aligned}$$

where the function H is referred to as the Hamiltonian. Substituting this into (2) we get

$$\frac{\partial \varrho}{\partial t} + \sum_{i=1}^{N} \left[\frac{\partial \varrho}{\partial q_i} \dot{q}_i + \frac{\partial \varrho}{\partial p_i} \dot{p}_i \right] = \frac{d\varrho}{dt} = 0$$

so

$$\frac{\partial \varrho}{\partial t} + \sum_{i=1}^{N} \left[\frac{\partial \varrho}{\partial q_i} \frac{\partial H}{\partial p_i} - \frac{\partial \varrho}{\partial p_i} \frac{\partial H}{\partial q_i} \right] = 0$$

In Hamiltonian mechanics the second term on the left hand side is referred to as the Poisson bracket between the Hamiltonian and density ϱ and we write

$$\frac{d\varrho}{dt} = \frac{\partial \varrho}{\partial t} + \{\varrho, H\} = 0$$

This equation is referred to often as the Liouville equation for a Hamiltonian dynamical system and holds for any system whether in equilibrium or not. According to Hamiltonian mechanics the invariants of the dynamical system which are functions of the dynamical variables only, are characterised as those functions having a vanishing Poisson bracket with H. Furthermore if we assume the system has steady statistics which seems a reasonable definition of equilibrium in general then

$$\frac{\partial \varrho}{\partial t} = 0$$

and so we conclude that for that case the density ϱ regarded as a function of the $2N$ co-ordinates q_i and p_i, must be an invariant of the system. This is clearly also the case for any function of the density in particular the logarithm of the density i.e.

$$\log \varrho$$

Now this particular function has another important property due to the statistical independence of different subsystems: It is additive across the entire closed system since clearly

$$\log \varrho_{12} = \log \varrho_1 + \log \varrho_2$$

The theory of classical mechanics of non-spinning particles tells us that there are only seven independent such additive conserved quantities for a general mechanical system. Namely energy and the three components of momentum and angular momentum. We deduce therefore that the logarithm of the density must be a linear combination of these invariants which is essentially equation (1) of the previous lecture i.e. Gibb's equation/deduction.

Notice that we have only identified the general form of the distribution and still need to determine which linear combination of coefficients are appropriate given that the macrostate has certain mean values of the invariants. This problem is known as the Legendre transformation. This transformation gives the appropriate coefficients with a prescribed set of invariant mean values as input. Another interesting question is the physical signifcance of these coefficients. We address this in the next two lectures.

3.2 Maximum entropy distributions

Suppose we characterize the macrostate of a subsystem using the mean values of the invariants discussed above (see also the last section of Lecture 1). There is an

interesting relationship between this characterization and the Gibbs density just discussed. Consider the set of densities which have the desired set of mean values of the invariants and further suppose that we have identified the appropriate set of coefficients for the Gibbs density that produce these mean values. What is the status of the Gibbs density ϱ among the large set of all possible densities q with the given mean invariants? It turns out that it has the maximal entropy. We can prove this using information theory. First set the Gibbs density as follows

$$\varrho = \exp\left(\sum_{i=0}^{n} \alpha_i I_i\right)$$

where we set $I_0 = 1$ and choose α_0 so that the density is properly normalized i.e. the integral is unity. Consider now the following chain of equalities and inequalities:

$$\begin{aligned} h(q) &= -\int q \ln q \\ &= -\int q \ln(\tfrac{q}{\varrho}\varrho) \\ &= -D(q||\varrho) - \int q \ln \varrho \\ &\leq -\int q \ln \varrho \\ &= -\int q \left(\sum_{i=0}^{N} \alpha_i I_i\right) \\ &= -\int \varrho \left(\sum_{i=0}^{N} \alpha_i I_i\right) \\ &= -\int \varrho \ln \varrho = h(\varrho) \end{aligned}$$

where we have used the non-negativity of relative entropy and the fact that the expectation values (means) of the invariants are the same for both q and ϱ. This remarkably simple proof shows that the Gibbs density has the maximal entropy among all densities with the same mean invariants. Notice too that it is unique in that regard since the relative entropy vanishes if and only if its two arguments are the same density.