AP286
Notes
10/15/2022

Name: *Ziyuan Zhao*

Information Theory, Statistical Mechanics, and Thermodynamics

1 Preamble

Why do we need statistical mechanics? Well, the typical argument goes like this: imagine we have a system of N particles, then it has $6N$ degrees of freedom, including positions and momentums of each particle. Given an equation of motion from classical mechanics or its equivalent, theoretically we can fully predict the trajectory and the property of our system. However, when N is exceedingly large, it may not be possible in reality to measure each degree of freedom or to proper describe how they the state will evolve. In practice we only have access to the macroscopic properties of the system, such as its volume and the average kinetic energy of the particles over time. Well, the subject of statistical mechanics is primarily about making this link between our knowledge of the microscopic parameters with the macroscopic properties that we can measure, so that the variation of macroscopic properties under different conditions can be used to infer the microscopic parameters, and, conversely, our knowledge of the microscopic parameters can be used to predict the macroscopic properties.

Now here's my personal opinion after having briefly studied thermodynamics and statistical mechanics in school. The subject can have many different flavors depending how you approach it, which mirrors the historical progression of ideas developed by Clausius, Boltzmann, Gibbs, Shannon, and, lately, Jaynes. Today I want to talk about perhaps the mathematically purer approach via information theory. I really liked this because we can build the statistical framework with minimal physical assumptions, making it very general and potentially applicable to other systems beyond the traditional focuses of equilibrium statistical mechanics. Then as we apply our framework to physical problems, we can inject assumptions along the way as needed.

So Shannon came along with idea of 'entropy' of a discrete probability distribution, defined as

$$H(p) = -\sum_{i=1}^{n} p_i \log p_i,$$

and the quantities has a myriad of other names, such as information, uncertainty, ignorance, etc. Shannon's entropy is indeed a useful measure of our ignorance, if we assume that our belief about the potential outcomes is encoded in the distribution (adopting the subjectivist view of probability). For H is maximized when we have a uniform distribution over all outcomes, which corresponds to the case when we are maximally ambivalent of what the outcome might be. On the other hand, H is nonnegative and minimized when $p_i = 1$ for some outcome i, and that corresponds to our belief that the outcome will certainly occur. In this worldview, gaining knowledge means 'sharpening' our probability distribution over outcomes and reducing H.

Note: for now we will assume that all outcomes are distinguishable. We will deal with the tricky issue of indistinguishability later.

Another useful interpretation of the H quantity is that it's the log of the volume of our phase space. We will probably not use this interpretation so I'd not pursue it further.

2 (just enough) Max Entropy Method

Once we have a objective quantity that measures our subjective ignorance, a natural question to ask is that, given some knowledge about the statistics of the outcomes that constrains our probability distribution, what is a distribution with highest H, or uncertainty in our belief, that we can come up with? Such a distribution is the most parsimonious representation of our belief in the sense it satisfies what we know but assumes no more (under our H measure). This is the philosophy of the Max Entropy (MaxEnt) principle, first propounded by Jaynes in 1957.

We will start simple. Assume we have a discrete random variable X with a support of $\{x_i\}_{i\in I}$, where I is the index set we use to label the different outcomes, and we assume that x_i and x_j are physically distinguishable when $i \neq j$. We can define a probability measure that encodes our belief in how likely the different outcomes will occur, we denote the probability for outcome x_i as p_i.

According to the MaxEnt principle, we have an optimization problem, that is, we want to maximize $H(X)$ subject to some constraints. Of course we need to ensure that $\sum_i p_i = 1$, that is, the probabilities are normalized. Additionally let us assume we can observe some simple statistics, i.e., expectations of some real valued functions $f_1, \ldots, f_n$ with respect to the distribution of X. A contrived case to illustrate why this might be of interest could be, imagine we are playing a slot machine, although we have no idea what's the probability different patterns can occur on each roll, we might be told by an engineer that the expected payoff of the machine is set to be some definite value. Anyway, let us define $\mu_j = \sum_{i\in I} p_i f_j(x_i) = \langle f_j(X)\rangle$.

To maximize $H(X)$ subject to the constraints, we can use the Lagrange multipliers method, and introduce multipliers λ_0 for the normalization constraint, and $\lambda_1, \ldots, \lambda_n$ for the "fixed expectation constraints".

Chugging through the maths we can obtain a series of general and very important results that I beg you to memorize:

$$Z(\mu_1, \ldots, \mu_n) = \sum_{i\in I} \exp\left(-\sum_{j=1}^{n} \lambda_j f_j(x_i)\right). \tag{1}$$

$$p_i(\mu_1, \ldots, \mu_n) = Z(\mu_1, \ldots, \mu_n)^{-1} \exp\left(-\sum_{j=1}^{n} \lambda_j f_j(x_i)\right). \tag{2}$$

$$H_{max}(\mu_1, \ldots, \mu_n) = \lambda_0 + \sum_{j=1}^{n} \lambda_j \mu_j. \tag{3}$$

$$\lambda_0(\mu_1, \ldots, \mu_n) = \log Z(\mu_1, \ldots, \mu_n). \tag{4}$$

For absolute clarity and in anticipation of what's to come the next, I have explicitly spelled out the independent variables as $\mu_1, \ldots, \mu_n$ because we have promised to use the knowledge of these statistics to constrain our optimization problem. Now the astute reader will realize I haven't provided an explicit solution for $\lambda_1, \ldots, \lambda_j$ yet. These are constrained by our choice of f_j's and

μ_j's, via the expectations:

$$\mu_j = \sum_{i \in I} p_i f_j(x_i) = \frac{\sum_{i \in I} \exp\left(-\sum_{j=1}^{n} \lambda_j f_j(x_i)\right) f_j(x_i)}{\sum_{i \in I} \exp\left(-\sum_{j=1}^{n} \lambda_j f_j(x_i)\right)}. \tag{5}$$

I bet you don't want to solve these implicit functions by yourself, neither do I! Note, of course, if we know H_{max}, which we write as H from now on, then the Lagrange multipliers ($j \geq 0$) is simply given by

$$\lambda_j = \frac{\partial H}{\partial \mu_j}. \tag{6}$$

Note, by Legendre transform, dually, we have

$$\mu_j = -\frac{\partial \lambda_0}{\partial \lambda_j} \tag{7}$$

where we hold all multipliers but λ_j constant.

3 Connecting to Statistical Thermodynamics

Well, physicists will note that the mathematical form of the results in the previous sections are exactly the same as the standard results from statistical mechanics. Here I'm just trying to make that connection.

First, we let $\{x_i\}$ be the collection of (distinguishable) microstates of the system we are interested in such that at every moment the system is in exactly one of these states, and that's what we can interpret as "outcomes" of the system. Let us choose f_j's to report the extensive properties of the system, such as internal energy, volume, and particle number of the k-th kind. So we make the identifications $f_1 \leftrightarrow E$, $f_2 \leftrightarrow V$, $f_{k+2} \leftrightarrow n_k$. We also assume that the energy of each state can depend on external parameters such as volume, strain tensor, applied electric or magnetic fields, etc, so we write $E(x_i, \alpha_1, \dots)$ and abbreviate this as just E_i.

Now we assume that our system is large, in the thermodynamic sense, such that the fluctuations around the expectation values of f_j's are small, so it is sufficient to observe the average values experimentally and compare it with what our theory predicts, then we make the identifications $\mu_1 \leftrightarrow U$, the internal energy of the system, $\mu_2 \leftrightarrow \langle V \rangle =: V$, the thermodynamic volume as a function of state, in contrast to the volume of microstates, and $\mu_{k+2} \leftrightarrow N_k$, the particle numbers as a function of state, rather than the particle number in each microstate. We additionally note that E_i could depend on N_k's if we allow interactions between particles beyond the dilute regime.

The last thing we need to do is to declare the significance of the max entropy distribution. Well, let us assume that corresponds to the equilibrium of the system, if we make the identification between Shannon's entropy with Gibbs entropy, upto a multiplicative factor of k_B (we will comment on this later), so $k_B H \leftrightarrow S$. Our claim is not too unreasonable in the light of the quantum H-theorem.

With all these identifications made, we can recast the MaxEnt method as a information theoretic approach to determine the properties of a macroscopic system at thermodynamic equilibrium. Given some system with its microstates well defined and observations of macroscopic state variables U, V, N_k, we have the following relationships:

$$Z(U, V, \{N_k\}) = \sum_{i \in I} \exp\left(-\lambda_1 E_i - \lambda_2 V_i - \sum_{k=1} \lambda_{k+2} N_{ki}\right). \tag{8}$$

$$p_i(U, V, \{N_k\}) = Z(U, V, \{N_k\})^{-1} \exp\left(-\lambda_1 E_i - \lambda_2 V_i - \sum_{k=1} \lambda_{k+2} n_{ki}\right). \tag{9}$$

$$S(U, V, \{N_k\})/k_B = \lambda_0 + \lambda_1 U + \lambda_2 V + \sum_{k=1} \lambda_{k+2} N_k. \tag{10}$$

$$\lambda_0(U, V, \{N_k\}) = \log Z(U, V, \{N_k\}). \tag{11}$$

This already looks suspiciously familiar. But what are these λ's? Well, if you have studied thermodynamics you probably immediately recognize them as related to the temperature, pressure, and chemical potential of the system. We will make that connection via the fundamental relation in thermodynamics:

$$dU = TdS - PdV + \sum_k \mu_k dN_k.$$

We will make use of the equation 9 which describes the entropy of our system, which we arrange to

$$U = -\lambda_0/\lambda_1 + S/(k_B\lambda_1) - (\lambda_2/\lambda_1)V - \sum(\lambda_{k+1}/\lambda_1)N_k$$

where we have suppressed the dependence of S and λ's on the choice of U, V, N_k. Now going back to the fundamental relation, if we assume that entropy is first order homogeneous in $U, V, \{N_k\}$, then by Euler's homogeneous function theorem (maybe I will do the proof in the appendix), the integral form of the fundamental relation is

$$U = TS - PV + \sum_{\mu_k} N_k$$

All that remains is to match up the coefficients, which leads to the following identifications

$$\lambda_1 \leftrightarrow \frac{1}{k_B T} =: \beta; \quad \lambda_2 \leftrightarrow \frac{p}{k_B T}; \quad \lambda_{k+2} \leftrightarrow -\frac{\mu_k}{k_B T}$$

Sidenote: β is known as thermodynamic coldness and, although not intuitive, provides a connection between information theory and thermodynamics and could be useful in some cases (to add). Also, note that an implication of this is that λ_0 vanishes in a macroscopic homogeneous system.

Look at all these awful factors of k_B, known as the Boltzmann factor with a dimension of energy over temperature (and what the heck is that), just because, by historical coincidence,

Shannon formulated his entropy later than the physicists who worked with all that thermodynamics stuff. In the parallel world where we had k_B set to unity, β will have the dimension of reciprocal energy and T the dimension of energy, so things could have been so simple.

Thus, we made our first step to connect with statistical mechanics. And the result in all its g(l)ory, as of now, is shown below:

$$Z(U, V, \{N_k\}) = \sum_{i \in I} \exp\left(-\frac{1}{k_B T} E_i - \frac{p}{k_B T} V_i + \sum_{k=1} \frac{\mu_k}{k_B T} N_{ki} \right). \tag{12}$$

$$p_i(U, V, \{N_k\}) = Z^{-1} \exp\left(-\frac{1}{k_B T} E_i - \frac{p}{k_B T} V_i + \sum_{k=1} \frac{\mu_k}{k_B T} n_{ki} \right). \tag{13}$$

$$S(U, V, \{N_k\})/k_B = \log Z + \frac{1}{k_B T} U + \frac{p}{k_B T} V - \sum_{k=1} \frac{\mu_k}{k_B T} N_k. \tag{14}$$

These are similar to the results derived from the grand canonical ensemble, except that I added one more pressure-volume term instead of letting the system volume be an external parameter (we talk about this later). Note, we shall make the identification of ensemble average with what we refer to as the expectation values given to derive the max entropy distribution, i.e., $\mu = \langle f(x) \rangle$. Rearranging the final equation gives us a nice equation about the state of the system

$$\Phi = U - TS + pV - \sum_k \mu_k N_k \tag{15}$$

where $\Phi = -k_B T \log Z$ is of the same form as the grand potential (except the additional pV term) and is clearly minimized at equilibrium, when the ensemble averages are fixed or, in other words, we fix our macrostate, because the entropy term S is maximized. A curious fact is that $\Phi = 0$ if the system is macroscopic and homogeneous, which is similar to the fact that the grand potential is $-pV$ in such situations.

There are already some other handy identities we can derive, for example, if somehow we are given the partition we can derive all these ensemble averages. We will talk about how to derive these in a later section. For now, these may not self-evident (well except for the last three, you can write out the expectation formulation of U, V, N_k and try relate that to Z)

$$S = -\frac{\partial \Phi}{\partial T}; \quad U = -\frac{\partial \log Z}{\partial \beta}; \quad V = -\frac{\partial \log Z}{\partial (\beta p)}; \quad N_k = \frac{\partial \log Z}{\partial (\beta \mu_k)}. \tag{16}$$

So now we can understand why our teachers and professors really mean it when they say that the partition function is a really powerful object and gives all information about the system!

Need to talk about ensembles, and then talk about how it's the same to control the expectation or the lagrange multiplier in Jaynesian systems.

4 Phase space and conjugate variables

Thus far our approach to describing the collective microstates of a system is pretty cumbersome. We required an arbitrarily indexed set of microstates with the readout functions f_j's that measures extensive properties of the microstates. These minimal assumptions are rarely useful in

practice. Now is the time to introduce the concept of phase space. Assume that the index set can be decomposed into a product $I = \prod_j I_j$ so now every microstate is index by a tuple of indices, each from the corresponding component index set. As a physical example, for a system of N particles, we can consider I as a $6N$ dimensional space, so the microstate of the system is described by $3N$ positions and $3N$ momentums of the particles. For now let us quantize the system and assume each I_j is finite. Quantum physically, we can imagine as if working with a a set of compatible observables of our system, each with a discrete spectra. Then we can call this I our phase space as it includes all possible states of our system.

To further simplify our counting of states, assume that, despite all being distinguishable, we do not care any more about the differences between any two microstates beyond what the f_j's tell us about them. For example, if we can only describe the energy, volume, and particle numbers for each microstate, then we shall say that two microstates with equal energy, volume, and particle numbers are equivalent for our purpose, even if they could be different in other aspects. Then we can map our full phase space as described previously into a more coarse grained phase space that I will denote with $\Omega = \prod_j \Omega_j$ (abusing notation from statmech) where each Ω_j is the range of f_j over our phase space. Our mapping from I to Ω naturally defined by $x_{\{i_j\}} \mapsto (\dots, f_k(x_{\{i_j\}}), \dots)$. As a physical example, consider mapping our aforementioned $6N$ dimensional phase space, taking $N = 2$ for simplicity, into Ω where we are only interested in the kinetic energy of each particle. Then Ω is the set of all possible kinetic energies the two particles can have, and the mapping simply reports the kinetic energies of the particle 1 and 2.

Note that distinguishable states indexed by I may no longer be distinguishable in Ω, and this is the important concept of eigenstate degeneracy in quantum mechanics. To compute the partition function, we need to sum over all possible microstates, but with the idea of degeneracy with respect to our chosen observables, we simply the counting by summing over Ω and counting the degeneracy at each point in Ω. To explain this more clearly, let's look at the formula for partition function we have found:

$$Z(U, V, \{N_k\}) = \sum_{i \in I} \exp\left(-\frac{1}{k_B T} E_i - \frac{p}{k_B T} V_i + \sum_{k=1} \frac{\mu_k}{k_B T} N_{ki} \right).$$

But with Ω defined, we can write this as

$$Z(U, V, \{N_k\}) = \sum_{(E_i, V_j, \{N_{kl}\}) \in \Omega} N(E_i, V_i, \{N_{ki}\}) \exp\left(-\frac{1}{k_B T} E_i - \frac{p}{k_B T} V_j + \sum_{k=1} \frac{\mu_k}{k_B T} N_{kl} \right),$$

where $N(E_i, V_j, \{N_{kl}\})$ simply counts the number of microstates with the same energy, volume, and particle numbers.

Now if we choose discretize Ω by equally spaced volume elements $\Delta\Omega$, we can write the number of states in each volume element as $\Delta N = \frac{\Delta N}{\Delta \Omega} \Delta\Omega$. Passing $\Delta\Omega$ to 0, we define $g(x) = \frac{dN}{d\Omega}|_x$ to be the density of state at $x \in \Omega$. Physically, when Ω is described by energy alone, we recover the density of state as defined in solid state and condensed matter physics. But in general if we write $d\Omega = dE dV dN_1 \dots dN_k$, then the density of state is a higher-order derivative with respect to these variables (check, for examples when this is useful). The definition of density of state motivates us to rewrite the above partition function as an integral over Ω:

$$Z = \frac{1}{\Delta\Omega} \int_\Omega dE dV dN_1 \dots g(E, V, N_1, \dots)\, p(E, V, N_1, \dots), \tag{17}$$

where $p(\cdot)$ is the exponential factor, also called the Boltzmann factor, that regulates state occupancy. In a similar manner we can compute expectation values with respect to the Boltzmann/MaxEnt distribution by evaluating integrals over Ω.

So far we have assumed that we can observe extensive variables such as U, V, N_k and get all other thermodynamic quantities. We've also talked about how the Lagrange multipliers are somehow coupled to our choice of U, V, N_k. This is most clearly seen from equation 13, where the conjugate variables β, p, μ_k always appear in a product with the extensive variables (hence the name, conjugate, just have some imagination and take it for granted!) We can differentiate the maximum entropy with respect to the extensive variables to get:

$$\frac{1}{T} = \left(\frac{\partial S}{\partial U}\right)_{V,\{N_k\}} ; \quad \frac{p}{T} = \left(\frac{\partial S}{\partial V}\right)_{U,\{N_k\}} ; \quad -\frac{\mu_k}{T} = \left(\frac{\partial S}{\partial N_k}\right)_{U,V,\{N_{-k}\}}, \tag{18}$$

where in the subscripts of partial derivatives I have made clear which independent variables we hold constant. Now going back to equation 15, it might look a bit shady now...we said that our independent variables are U, V, N_k but then proceeds to compute the partial derivatives with respect to the conjugate variables, what is going on, what are we holding constant there? The next section deals with this and more.

5 Legendre Transformation, Thermodynamical Potentials, and Massieu Functions

So far, we have been able to express max entropy of a system as a function of extensive and independent variables $U, V, \{N_k\}$, i.e., $S(U, V, \{N_k\})$. By the monotonicity of $S(U)$ from physical considerations we can invert the function to express the energy at max entropy (which is actually the minimal possible energy at this entropy, proof to include in appendix or later) as a function of independent variables S, V, N_k, i.e., $U(S, V, \{N_k\})$. These are both fundamental thermodynamic relations for they tells us how to completely specify the state of the system at equilibrium.

From the end of the last section, and also from how physicists perform actual experiments, we realize it will be useful to somehow use conjugate variables in the fundamental relations above. Well, you might say, why not just express S as a function of $T, p, \{\mu_k\}$, personally I think it's fine, and you can write out 2^{2+k} different sets of independent variables by swapping each variable by its conjugate, appealing to the monotonicity of each independent variable as a function of its conjugate and vice versa (see appendix for physical justification). Note, some people will comment that I'm wrong on that we can write $S(T, p, \{\mu_k\})$, due to Gibbs-Duhem equation, but that somehow assumes macroscopic, homogeneous system in the derivation and the equation does not apply to all thermodynamic systems. I'm not going to argue with anyone on that, anyways. Similarly, For the energy functions, $U(S, V, \{N_k\})$ we can write 2^{2+k} different sets of independent variables by the same swapping to conjugates trick.

For simplicity now let's assume the system has constant particle number, so we write $U(S, V)$ and $S(U, V)$ as the starting fundamental relations. Based on our argument above we can as well write $U(T, p)$ or $S(T, p)$, bu then the problem is, what's the use of doing so? We have chosen our independent variables, so we can manipulate them to express the change in entropy as a response

of small changes in the independent variables:

$$dU = \left(\frac{\partial U}{\partial T}\right)_p dT + \left(\frac{\partial U}{\partial p}\right)_T dp.$$

Note this total derivative is somewhat desirable because we can experimentally control temperature and pressure better than entropy (how do we even do that?) or volume. The only thing that stops us is that we cannot relate either partial derivative back to the existing thermodynamic variables we have defined. The first is called heat capacity at constant pressure and the second is related to a jumble of other properties.

With that in mind, that wiggling the conjugate variables give us a proportional change in energy, but the coefficient may not be in an immediately recognizable form, we can ask, is it possible to construct an energy like quantity such that its partial derivative with respect to the conjugate variables immediately give us their conjugates, so we get back to the ensemble averages/extensive parameters. Such functions are useful because they invert the control-response relation to allow us measure the extensive parameters by manipulating the intensive ones. This is achieve by a mathematical technique called Legendre transformation (for simplicity we will work with single-phase homogeneous system to dull down the math needed). The geometrical insight is such: for a real, properly differentiable convex function in Euclidean space, we can equivalently represent the function as the collection of all points described by the argument vector and the function value. Alternative, we can describe each point by a gradient of the hyperplane tangent to the surface along with its intersect with the value axis (z-axis). In such way we can obtain various equivalent representation of a multivariable function by swapping out a variable by the partial derivative with respect to that variable (the conjugate). After cranking through the math (maybe to include in Appendix), we obtain the result that for a function $F(x, x')$ where x is a vector, we can write it as $G(s, x')$ such that

$$F(x, x') - G(s, x') = s \cdot x. \tag{19}$$

The notation implies that we can transform just a subset of the original variables. Also, note that we have used minus sign because of the convention in thermodynamics. It really doesn't matter mathematically.

We now apply Legendre transform to the internal energy written as $U(S, V)$ to generate other thermodynamical potentials. They are called potentials because they have units of energy and are minimized at equilibrium (next section). We first recognize that, $\left(\frac{\partial U}{\partial S}\right)_V = T$ is conjugate to S, so we have $U(S, V) - F(T, V) = TS$, so $F(T, V) = U(S, V) - TS = U - TS$. This is the Helmholtz potential, and by comparison to equation 14, we can infer that it corresponds to the canonical potential, i.e., we just regulate internal energy / temperature and allow volume and particle numbers to vary as mechanical variables. Alternatively we can swap out V for $p = -\left(\frac{\partial U}{\partial V}\right)_T$ and get enthalpy $U(S, V) - H(S, p) = -pV$ so $H(S, p) = U(S, V) + pV = U + pV$. This does not match any well-known ensemble. We can transform both variable and get Gibbs free energy of $G(T, p) = U + pV - TS$ which corresponds to the isothermal-isobaric ensemble, by converting volume / pressure from a mechanical variable to a dynamical variable to be regulated. If we start with $U(S, V, N_k)$ and transform S and all N_k, we'd get the grand potential corresponding to the grand canonical ensemble.

We can repeat the above procedures but starting with entropy as $S(U, V)$. The derived functions all have the same units as entropy rather than energy, so these are more suitable for statistical mechanics than thermodynamics. These are referred to Massieu functions (see Callen).

We note that the Legendre transform $G(s, x')$ by definition has the property that its partial derivative at s_i yields $-x_i(s_i)$, i.e.,

$$\frac{\partial F}{\partial s_i} - \frac{\partial G}{\partial s_i} = x_i + s\frac{\partial x_i}{s_i}$$

which rearranges to

$$x_i + \frac{\partial G}{\partial s_i} = \frac{\partial F}{\partial s_i} - \frac{\partial F}{x_i}\frac{\partial x_i}{s_i} = 0.$$

Sorry that I've sloppily written all this and suppressed the dependence of s on x.

The takeaway is that when we differentiate thermo potentials or Massieu functions with respect to the conjugate variables we should get back the original variable, and this is what we wanted to achieve in the beginning. For example, we have

$$-\left(\frac{\partial G(T,p)}{\partial T}\right)_p = S; \quad \left(\frac{\partial G(T,p)}{\partial p}\right)_T = V. \tag{20}$$

Thus, the idea is that if the temperature of a system varies by ΔT, then its Gibbs free energy varies by $S\Delta T$, so if we can measure the change in Gibbs free energy in response to us manipulating the temperature, that gives us a handle on the entropy. Similar argument goes with the pV conjugate pair. We are quickly getting to the applications and the messy end of the theory, so let us stop here.

6 Convexity and Extremization Principles

TBD.

7 Linear Response Coefficients

A linear response function describes the linear portion of the input-output response of a system. Such concept can be applied to thermo-statistic systems described either by Jaynes' MaxEnt approach or classical thermodynamics. The central question is, given a change in all extensive variables / ensemble averages, how do we characterize the corresponding change in the intensive, conjugate variables and vice versa. In the MaxEnt framework which is admittedly more abstract, we can ask about the relationship between the constraint expectations and the Lagrange multipliers. We will see in this section that the linear response to perturbations is described by a matrix of thermodynamic derivatives that can be related to material properties in one way or another. In the next section, we use the abstract formulation to investigate information geometric property of the MaxEnt distributions.

First we discuss the abstract formulation. Recall from equation 6 that the multipliers are partial derivatives of the maximum entropy with respect to the constraint, while holding all others

constant. We take another derivative to get

$$-g_{ji} := \left(\frac{\partial \lambda_i}{\partial \mu_j}\right)_{-\mu_j} = \left(\frac{\partial H}{\partial \mu_i \partial \mu_j}\right)_{-\mu_i,-\mu_j} = \left(\frac{\partial H}{\partial \mu_j \partial \mu_i}\right)_{-\mu_i,-\mu_j} = \left(\frac{\partial \lambda_j}{\partial \mu_i}\right)_{-\mu_i} =: -g_{ij}.$$

This tells us many things. Firstly, this equation is derived from purely within the MaxEnt framework (and note H is Shannon entropy, effective the same as Gibbs entropy but with $k_B = 1$), we make no further thermodynamic assumptions to derive it. Secondly, if we can manipulate μ_j's and measure the changes in λ_i, then these partial derivatives are indeed linear response coefficients. In the subscripts labelled the set of independent variable we need to hold constant while performing partial differentiation, the minus sign just indicates all set of variables except those indicated. The third thing we note is that we have related these coefficients to the curvature of H by the second derivatives and thus could endow them with geometric meanings. This anticipates our discussion in the next section. Finally, we note that by way of exchanging the order of performing the differentiations, we can equate pairs of coefficients that are in symmetric positions in the Hessian of H, i.e., $-g_{ji} = -g_{ij}$ in $D^2H(\{\mu_k\})$. Such relations are known as Maxwell relations in thermodynamics.

After the discussion of Legendre transformation in the thermodynamics setting, not surprisingly, we can apply the same idea in the MaxEnt framework and get another matrix of linear response coefficients. From equation 3 we see that

$$H(\{\mu_k\}) - \lambda_0(\{\lambda_k\}) = \sum_j \lambda_j \mu_j = \lambda \cdot \mu.$$

So the log partition function is exactly the Legendre transform of the entropy by swapping all constraints by the multipliers, as they are conjugate in the entropy function. From the Legendre transform we can define

$$\gamma_{ji} := -\left(\frac{\partial \mu_i}{\partial \lambda_j}\right)_{-\lambda_j} = \left(\frac{\partial^2 \lambda_0}{\partial \lambda_i \partial \lambda_j}\right)_{-\lambda_i,-\lambda_j} = \left(\frac{\partial^2 \lambda_0}{\partial \lambda_j \partial \lambda_i}\right)_{-\lambda_j,-\lambda_i} = -\left(\frac{\partial \mu_j}{\partial \lambda_i}\right)_{-\lambda_i} =: \gamma_{ij}.$$

Additionally, we note one more intepretation of γ as the variance-covariance matrix of the microstate properties f_j's, which is straightforward to verify by further differentiating equation 5:

$$\gamma_{ji} = \gamma_{ij} = \langle f_i f_j \rangle - \langle f_i \rangle \langle f_j \rangle = Cov(f_i, f_j).$$

So if we interpret γ_{ji}'s as susceptibility coefficients that are widely used in thermodynamics, we can realize they are related to the fluctuations of microstate properties in the max entropy distribution. Further derivatives give higher central moments of f_j's and produce a hierarchy of fluctuation laws, as Jaynes observed in his 1962 lecture.

By chain rule $\frac{\partial \mu_i}{\partial \mu_j} = \sum_l \frac{\partial \mu_i}{\partial \lambda_l} \frac{\partial \lambda_l}{\partial \mu_j}$ it's immediate that the two linear coefficient matrices are inverses of each other:

$$g = \gamma^{-1}. \tag{21}$$

Now proceed to explain the connection to thermodynamics... if I have time.

8 Generalized Heat and Work

Typically, the system we are interested in may have additional variational parameters that we do not put an expectation constraint on. Let's call these parameters $\alpha_1, \ldots, \alpha_r$ and these goes into the microstate property functions f_j's so now we write $f_j(x_i; \alpha_1, \ldots, \alpha_r)$ to denote our their dependence on the additional parameters. As a result the partition function shall also have the additional parameters, so we write $Z(\mu_1, \ldots; \alpha_1, \ldots, \alpha_r)$. For notational simplicity we consider only having α for now. A physical example is that in the canonical ensemble, we regard volume as an mechanical variable we can vary externally, rather than a state variable of the system (this is an important distinction).

We can easily verify that the expected derivative of microstate properties to α satisfies the following relation:

$$\sum_k \lambda_k \left\langle \frac{\partial f_k}{\partial \alpha} \right\rangle = -\frac{\partial}{\partial \alpha} \log Z. \tag{22}$$

We now consider a infinitestimal variation in our system description, such that the expectation values and mechanical parameters change by small amounts of $\delta\mu_k$ and $\delta\alpha$, respectively. δ denotes a path dependent differential or also called an inexact differential, because these integrals are path dependent. From 3 the maximum entropy will change by a small amount

$$\delta H = \sum_k \frac{\partial \log Z}{\partial \lambda_k} \delta\lambda_k + \frac{\partial \log Z}{\partial \alpha} \delta\alpha + \sum_k \mu_k \delta\lambda_k + \sum_k \lambda_k \delta\mu_k.$$

The first and third term cancels by equation 7, and using equation 22 we obtain

$$\delta H = -\sum_k \lambda_k \left\langle \frac{\partial f_k}{\partial \alpha} \right\rangle \delta\alpha + \sum_k \lambda_k \delta\mu_k. \tag{23}$$

We then define the infinestimal variation in the properties function as $\delta f_k = \frac{\partial f_k}{\partial \alpha}\delta\alpha$ so we can write

$$\delta H = \sum_k \lambda_k (\delta\langle f_k \rangle - \langle \delta f_k \rangle). \tag{24}$$

We then proceed to define the generalized heat of the k-th kind as $\delta q_k = \delta\langle f_k \rangle - \langle \delta f_k \rangle$ so that

$$\delta H = \sum_k \lambda_k \delta q_k. \tag{25}$$

Similar to our intuition from thermodynamics, δq_k is not exact, i.e., no function is the antiderivative of δq_k wrt the state variables and mechanical parameters. However, max entropy is a state function, so the entropy differential is exact. Equation 24 shows that a weighted linear combination of δq_k's is exact, as the Lagrange multipliers provide integrating factors for the corresponding differentials. This generalizes Clausius' definition of entropy from $dS = \delta Q/T$ from classical thermodynamics.

Now we observe that $\delta\langle f_k \rangle = d(\sum_i p_i f(x_i)) = \sum_i d(p_i f(x_i))$ is the total change in the expected value of the distribution and is analogous to change in internal energy. Then $\langle \delta f_k \rangle = \sum_i p_i \delta f_k(i)$ is the part of change due to the change in f_k, and this is analogous to shifted energy

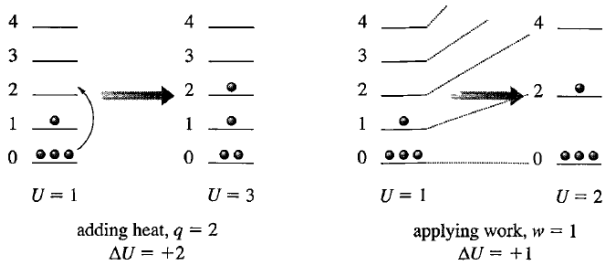

Figure 4.1
Two totally different ways to add energy to a system and thus raise its temperature. In the first way, 2 units of heat are added, and a particle moves up in energy (starts moving more violently). In the second way, 1 unit of work is added, carrying the particle up to higher energy without changing its level. It is the whole level in that case that moves; the particle just goes along for the ride.

Figure 1: Heat and work associated with a process are two ways to change the energy of a quantum system

level when we vary the mechanical parameter of the system. We can call this term generalized work. Applying product rule to the differential $\delta(p_i f(x_i))$, we can see that the generalized heat can be written as $\delta q = \sum_i f(x_i) dp_i$, this is the part of the change in expectation value due to change in the max entropy distribution, and it's analogous to altering the relative occupancy of different energy levels of a quantum system, so the term is judiciously named as generalized work. From the discussion we see that the total change can be decomposed into the generalized heat and generalized work components. This is exactly analogous to the first law of thermodynamics. Figure 1 from my freshman chemistry textbook illustrates the idea of heat and work of a process from the statistical mechanical perspective and it's in the same rein as our discussion above, although our formulation is more general.

9 Differential Geometry Fun

Since γ is the variance-covariance matrix of the microstate properties, it's symmetric nonnegative semidefinite, and thus $g = \gamma^{-1}$ is also symmetric nonnegative semidefinite. We can thus use them define metric tensors in the Euclidean space where the hypersurface parametrized by $\{\mu_k\}$ or $\{\lambda_k\}$ along with H lives. We can then define the line element either in terms of the entropy or the conjugate free energy, by using g or γ.

The upshot is that, given a curve (note, we now switch to tensor algebra index convention) $\mu^i(t)$ parametrized by t, where $\mu^i(t_0)$ and $\mu^i(t_1)$ are the beginning and end states of the system, and if we assume we transform the system from the beginning state to the end state along the curve by a quasistatic process, i.e., slow enough that at every point along the trajectory the system is regarded as in equilibrium, then we can compute the entropy change associated with the process, which is just the line integral on the equilibrium hypersurface:

$$\Delta s = \int_{t_0}^{t_1} \sqrt{g_{ij} \frac{d\mu^i}{dt} \frac{d\mu^j}{dt}} dt. \tag{26}$$

Dually, we can compute the free energy ($\Phi = -\lambda_0$) change by

$$\Delta \Phi = \int_{t_0}^{t_1} \sqrt{\gamma_{ij} \frac{d\lambda^i}{dt} \frac{d\lambda^j}{dt}} dt. \tag{27}$$

Given two states on the hypersurface, we can also ask what is the shortest path between them using the distance measure for paths as above, and such shortest path is known as the geodesic. When I have enough time, I will work on this part that shows you can derive a simple form of least action bound for Jayesian systems. Such bound provides a lower bound for the cost of a process in entropy units. It emerge from the Riemannian geometry as described.

10 Yes, you asked for applications, so why not

TBD.

11 Appendix

TBD.