Information Theory and Predictability
Lecture 6: Maximum Entropy Techniques

1 Philosophy

Often with random variables of high dimensional systems it is difficult to deduce the appropriate probability distribution from the given observations of outcomes. We shall confront this problem in full force in the final lecture on predictability later in the course. One approach to this due to Jaynes [2] and later Mead and Papanicolou [5] is to first assume certain highly plausible properties of the random variable and then posit maximum uncertainty with respect to all other properties of the distribution. Such an ansatz then often allows the deduction of the desired distribution and is usually referred to as a maximum entropy principle. Jaynes first used this principle in the context of statistical mechanics and the assumed properties of the distribution were associated with conservation principles connected with such quantities as energy and angular momentum. Such ideas have also been applied to turbulent fluids (for geophysical applications see Salmon [6]) with some notable successes. Mead and Papanicolou on the other hand considered the case where certain low order moments were assumed known from sampling and then the distribution consistent with maximal ignorance concerning higher order moments deduced. We consider this case first. We shall refer to the assumed properties as constraints and the maximization of uncertainty shall be achieved via the entropy functional. From a practical viewpoint the problem is one of constrained optimization.

2 Moment Constraints

Consider a finite set of polynomials $\{r_i(x)\}$ where x is an element of the continuous outcome space with (unspecified) probability density $p(x)$ and $i \leq N$. Assume that the following relations hold

$$\int p(x) r_i(x) dx = M_i \tag{1}$$

where the M_i are called the moments. For convenience and completeness we take $r_0(x) \equiv 1$ and $M_0 = 1$ to incorporate the further needed constraint regarding $p(x)$ being a probability density.

The problem then becomes to find $p(x)$ subject to the requirement that the associated differential entropy $h(X)$ is maximized since this reflects maximal ignorance or assumption concerning all other moments.

The obvious method for solving this constrained optimization problem is to use Lagrange multipliers for each of the constraints and maximize the augmented functional with respect to $p(x)$

$$J \equiv -\int p(x)\ln(p(x))dx + \sum_{i=0}^{N} \lambda_i \left(\int p(x) r_i(x) dx - M_i \right)$$

Taking the functional derivative with respect to $p(x)$ and setting to zero we get

$$\frac{\delta J}{\delta p} = -\ln\left(p(x)\right) + 1 + \sum_{i=0}^{N} \lambda_i r_i(x) = 0$$

which implies that the maximum entropy distribution must have the form

$$p(x) = C(\overrightarrow{\lambda}) \exp\left(\sum_{i=1}^{N} \lambda_i r_i(x) \right) \tag{2}$$

where the C are undetermined normalization "constants" which are also equal to $\exp(1+\lambda_0)$. The Lagrange coefficients λ_i need to be determined by use of the moment constraints which is in general a non-trivial task which we consider below. The family of distributions of the form given by (2) is known as the exponential family. Notice that as a subset it contains the Gaussian distributions but is obviously much larger. We now show that this distribution determines a unique maximum entropy distribution f of all distributions q satisfying the moment constraints (1). Consider the differential entropy of q

$$\begin{aligned}
h(q) &= -\int q \ln q \\
&= -\int q \ln(\tfrac{q}{f} f) \\
&= -D(q||f) - \int q \ln f \\
&\leq -\int q \ln f \\
&= -\int q \left(\sum_{i=0}^{N} \lambda_i r_i + 1 \right) \\
&= -\int f \left(\sum_{i=0}^{N} \lambda_i r_i + 1 \right) \\
&= -\int f \ln f = h(f)
\end{aligned}$$

where the fourth line follows from the non-negativity of relative entropy; the fifth because f is exponential and the sixth because q satisfies the same moment constraint (1) as f does. Note that not only does this show exponential distributions maximize the entropy (subject to the moment constraint) they are unique in this since the inequality in this argument only becomes an equality when $q = f$ exactly again by the relative entropy theorem.

3 Simple examples

3.1 Gaussian distributions

Suppose we restrict ourselves to polynomials $r_i(x)$ of order two or less. In such a case all moments can be written in terms of the means and covariances of a vector random variable. Additionally it is obvious from the generic form of the exponential family restricted to second order polynomials that they cover Gaussian distributions only. The Lagrange multipliers can be obtained from the general form of a Gaussian which is

$$p(\overrightarrow{x}) = \frac{1}{\sqrt{(2\pi)^n det(\sigma^2)}} \exp\left(-\frac{1}{2}(\overrightarrow{x} - \overrightarrow{\mu})^t \left[\sigma^2\right]^{-1} (\overrightarrow{x} - \overrightarrow{\mu})\right)$$

where the σ^2 is the covariance matrix and μ the mean vector. The moment constraints are easily converted to mean and covariance constraints. Insertion in the above form then gives the appropriate λ_i. Note that the normalization constant $C(\overrightarrow{\lambda})$ is also determined by the same sort of procedure. The maximum entropy formulation for Gaussian distributions is revealing as it says that maximal ignorance with regard to moments higher than two implies the Gaussian distribution.

3.2 One sided one dimensional exponential distribution

Suppose we restrict ourselves in one dimension to non-negative x for our random variable X and further suppose we constrain the mean of X to always be μ. The maximum entropy distribution in such a case must have the form $\lambda \exp(-\lambda x)$ (for proper normalization). Substituting into the moment constraint shows that the Lagrange multiplier must be $1/\mu$. An interesting interpretation of this result is as a vertical column of gas in the Earth's atmosphere. The potential energy (PE) of a small parcel of air at height z with cross sectional area unity is

$$\Delta PE = \rho(z) g z \Delta z$$

where ρ is the density and g is the acceleration due to gravity. The integrated PE of the whole column is thus

$$PE = g \int_0^\infty \rho(z) z dz$$

Now if we further assume that the density is proportional to the probability density of an individual gas molecule as seems reasonable we see that an assumption of a fixed potential energy for the column amounts to the first moment constraint just discussed. Thus one might expect the density to exhibit the exponential fall off with height derived above which indeed it does approximately. In order to get a more precise description of the gas we need to include further moment constraints. In particular the mean kinetic energy which is proportional to local temperature could be used as a constraint of molecule momentum (as opposed to height). We discuss this particular constraint further below.

4 The Legendre transformation

4.1 Information geometry

The exponential family of distributions we met above is an example of a parameterized set of distributions where the Lagrange multipliers play the role of the parameter vector. Another example is the class of mixture distributions. Suppose we select a set of N fixed distributions p_i (for example a series of Gaussians with different means and covariances). Now form the following parametrized family of distributions:

$$q(\overrightarrow{\alpha}) \equiv \sum_{i=1}^{N} \alpha_i p_i$$

with $\alpha_i \geq 0$ and $\sum_{i=1}^{N} \alpha_i = 1$ in order to ensure $q(\overrightarrow{\alpha})$ is a probability density. This family is called a set of mixtures and is parametrized by the vector $\overrightarrow{\alpha}$. Consider now a general class of parametrized distributions $p(x, \overrightarrow{\theta})$. The relative entropy was identified in earlier lectures as a "distance" function on distributions. Let us now make this a little more precise. Suppose we consider a small perturbation to a particular parameter set and compute the resulting relative entropy between the perturbed and unperturbed distributions. Write

$$\theta_i' = \theta_i + \varepsilon v_i$$

and assume ε small. Now expand $\ln p(x, \boldsymbol{\theta'})$ as a Taylor expansion in the small parameter ε

$$\begin{aligned} \ln p(x, \boldsymbol{\theta'}) &= \ln p(x, \boldsymbol{\theta}) + \varepsilon v_i \tfrac{1}{p(x,\boldsymbol{\theta})} \tfrac{\partial p(x,\boldsymbol{\theta})}{\partial \theta_i} \\ &+ \tfrac{\varepsilon^2}{2} v_i v_j \frac{\frac{\partial^2 p(x,\boldsymbol{\theta})}{\partial\theta_i \partial\theta_j} p(x,\boldsymbol{\theta}) - \frac{\partial p(x,\boldsymbol{\theta})}{\partial\theta_i}\frac{\partial p(x,\boldsymbol{\theta})}{\partial\theta_j}}{p^2(x,\boldsymbol{\theta})} + O(\varepsilon^3) \end{aligned}$$

where we are using the summation convention for Latin indices. Now substitute this into the expression for $D(p(\boldsymbol{\theta})||p(\boldsymbol{\theta'}))$ to obtain

$$\begin{aligned} D(p(\boldsymbol{\theta})||p(\boldsymbol{\theta'})) &= -\varepsilon v_i \tfrac{\partial}{\partial\theta_i} \textstyle\int p(x,\boldsymbol{\theta}) dx - \tfrac{\varepsilon^2}{2} v_i v_j \times \\ &\left[\tfrac{\partial^2}{\partial\theta_i \partial\theta_j} \textstyle\int p(x,\boldsymbol{\theta}) dx - \int p(x,\boldsymbol{\theta}) \tfrac{\partial \ln p(x,\boldsymbol{\theta})}{\partial\theta_i} \tfrac{\partial \ln p(x,\boldsymbol{\theta})}{\partial\theta_j} dx \right] + O(\varepsilon^3) \end{aligned}$$

The first two terms vanish due to the integral property of probability densities and the third is a symmetric billinear form involving the so-called symmetric Fisher information matrix $G = (g_{ij})$:

$$\begin{aligned} D(p(\boldsymbol{\theta})||p(\boldsymbol{\theta'})) &= \tfrac{\varepsilon^2}{2} v_i v_j g_{ij}(\boldsymbol{\theta}) + O(\varepsilon^3) \\ g_{ij}(\boldsymbol{\theta}) &\equiv \textstyle\int p(x,\boldsymbol{\theta}) \tfrac{\partial \ln p(x,\boldsymbol{\theta})}{\partial\theta_i} \tfrac{\partial \ln p(x,\boldsymbol{\theta})}{\partial\theta_j} dx \end{aligned}$$

The Fisher information matrix plays a central role in mathematical statistics where it is crucial to the theory of fitting of parametrized distribution families to data (it provides a lower bound on error estimates for such parameters). Here we see it also plays a crucial role in defining a local distance function on parametrized distributions. In fact because of the above expansion it may be identified as the metric tensor in the usual differential geometric sense. A change of parameters describing the distributions corresponds with a change of coordinates in the usual differential geometric sense and the Fisher information matrix transforms as a second rank (metric) tensor under this change. Considerable more detail on this elegant formalism may be found in the book by Amari and Nagoaka [1].

4.2 Legendre Potential

We consider now the family of exponential distributions which have a set of coordinates (in the sense of the previous subsection) given by the vector of Lagrange multipliers. Among this family is the desired maximum entropy distribution which will be specified exactly if we find the appropriate vector $\overrightarrow{\lambda}$. To achieve this objective we introduce what we term the Legendre potential:

$$\Gamma(\overrightarrow{m}, \overrightarrow{\lambda}) \equiv \ln Z(\overrightarrow{\lambda}) - (\overrightarrow{m}, \overrightarrow{\lambda})$$

where Z is referred to as the partition function because of the connection of this formalism with statistical mechanics (see below). It is defined by

$$Z(\overrightarrow{\lambda}) = C^{-1}(\overrightarrow{\lambda}) = \int \exp\left(\sum_{i=1}^{N} \lambda_i r_i(x)\right) dx \tag{3}$$

Differentiation of Γ with respect to the Lagrange multipliers λ_i shows that is has an extrema when the moments of the exponential distribution associated with λ_i are equal to m_i. By considering the second derivatives one may also show that this stationary point is convex and thus corresponds to a minimum (a proof of this assertion can be found[1] in Mead and Papanicolou [5]). Further at this minimum Γ is easily seen to be the differential entropy of the corresponding exponential distribution. We therefore have an algorithm for finding the maximum entropy distribution:

1. In the Legendre potential set $m_i = M_i$ the desired moments.
2. Choose a particular (arbitrary) set of λ_i which because they have different moments will not correspond to a minimum value for the potential.

[1] Actually it easily checked that the Hessian matrix $\frac{\partial^2 \Gamma}{\partial \lambda_i \partial \lambda_j} = \langle r_i(x) r_j(x) \rangle - \langle r_i(x) \rangle \langle r_j(x) \rangle$ at the critical point where the angle brackets denote expectation with respect to the extrema (maxent) distribution which demonstrates that there is usually a minimum since the covariance matrix is non-negative definite. Only degenerate situations avoid this conclusion. Note also that this Hessian is actually just twice the Fisher information matrix/metric discussed previously for the exponential family of distributions under consideration.

3. Iterate the λ_i using a "steepest descent" or related numerical algorithm.

4. The nature of second derivative Γ ensures that usually a unique minimum will be (eventually) found which will be the maximum entropy exponential distribution and the minimum value will be that entropy.

It is also easy to check that the discrepancy in Legendre potential between that for any particular choice of λ_i and the minimum value is actually the relative entropy $D(p||q)$ where p is the maxent distribution while q is the exponential distribution with parameters λ_i. This also establishes that the maxent distribution is a global minimum for the Legendre potential.

This problem will only be well behaved in general if there is sufficient convexity in all directions of parameter space. Note also that the method requires the calculation of Z which is a multidimensional integral. For high dimensions this calculation can have inaccuracy making the optimization problem badly behaved.

Finally observe that this optimization implies a unique correspondence between λ_i and M_i and so the moments rather than the Lagrange multipliers could be used as "coordinates" in the space of exponential distributions. This coordinate transformation is called a Legendre transformation and can be dealt with using the information geometry formalism introduced above. Such transformations are very important in equilibrium statistical mechanics.

5 Connection with equilibrium statistical mechanics

Statistical mechanics is the study of very large collections of particles. In general only the statistics of such particles are known experimentally and these are referred to generically as macro or thermodynamical properties as opposed to micro properties that apply to individual particles. Statistical mechanics is a theory for deriving macro properties from the molecular properties using statistical methods. Entropy can be made a central organizing principle in deriving such a theory. This is done by saying that the macro properties of the system should be taken as compulsory but then that one should assume as little as possible regarding the remaining microvariables of the system which make up the complete dynamical description. This amounts, of course, to some kind of constrained maximum entropy problem. Such an ansatz may be shown to work well in the case of macroscopic equilibrium where it forms the basis for conventional equilibrium statistical mechanics. Note that it also implies that a system not in equilibrium will not be in a state of maximum entropy so will have smaller entropy than the equilibrium state.

In order to compute probability distributions of interest in statistical mechanics one considers various system "setups" which impose the above mentioned macro constraints on the variables of the system in different ways. The set of all configurations of a dynamical system consistent with such constraints is termed

an ensemble. One such ensemble consists of a closed system subject to a series of conservation principles. The best known such principle is energy but several others such as momentum and angular momentum are possible as well. The conservation principle then functionally constrains the system variables. For example the total energy of all molecules within a gas is assumed fixed. The ensemble of configurations consistent with these constraints is called microcanonical. In general many of the macrovariables correspond with conserved quantities within such a system. If one restricts the variables of the system so that the conservation constraints are met then the maximum entropy ansatz above corresponds with assuming a uniform distributions for the remaining microvariables. Put another way, the probability distribution of the microcanonical ensemble members is uniform. It is possible to rigorously prove that this conclusion[2] is true in certain dynamical systems which are termed ergodic. An important example being a set of colliding hard balls enclosed within a box with hard walls. Results of this type called billard dynamical systems have had a number of ergodic results proven in recent times beginning in the the 1960s with Sinai see [7]. In a system which satisfies the Liouville condition $\nabla \bullet \boldsymbol{A} = 0$ (see previous lecture) one can also show that this uniform distribution is time invariant i.e. an equilibrium distribution. A particularly clear introduction to this large area of ergodic theory may be found in [3].

Computing relevant probability distributions in such a way can be challenging because one needs to take into account the geometry of the conserved variable constraints. For this reason and for reasons of physical realism open rather than closed setups are more typically considered. Here one assumes that the system exchanges specified conserved quantities with the environment and is in equilibrium with it. Clearly now unlike the closed case the exchanged conserved variables may not be constant and thus cannot be used directly as thermodynamical variables however if the concept of equilibrium is to make statistical sense then the mean values of such variables should be fixed. By a mean here it is meant with respect to the ensemble of identically prepared systems which could be a series of realisations over time. Such means then are functionally related to important thermodynamical or macrovariables. When energy alone is allowed to exchange in this way the set of configurations is termed a canonical ensemble. If in addition the total number of particles is also allowed to exchange then the set of configurations is termed a grand canonical ensemble. Notice for the open type of configuration the constraints on the system are statistical rather than absolute as they are in the closed case. The open setup also evidently corresponds directly with the mathematical case considered above. Thus an invocation of the maximum entropy principle when the

[2]One needs to be rather careful from a mathematically technical viewpoint in stating this result. Indeed there can exist members of the microcanonical ensemble that are never visited in any equilibrium configuration. These have however measure zero. The hypothesis that all microcanonical ensemble members are "equally likely" was first stated by Boltzmann and was called the ergodic hypothesis. It was shown to be strictly untrue around 1913 but was replaced by the quasi-ergodic hypthosesis which very roughly means any coarse graining of the submanifold with equal measure will have equal probability measure. This was proven by Birkhoff and Von Neumann in the 1930s. More details in the review [3].

conserved variables are polynomials of state variables will lead to a variety of exponential family distributions. Interestingly in the limit of a very large number of particles it can often be shown (often numerically) that the probability distributions for both the open and closed configurations converge and this is actually a general assumption in statistical physics. The canonical and grand canonical ensembles were first introduced by J. Willard Gibbs over 100 years ago and the equilibrium probability distributions/measures are commonly called Gibbs measures. More details on this from a physics perspective can be found in standard texts on statistical mechanics such as [4].

Let us consider a specific very simple example. Suppose we have a collection of hard colliding classical particles with momentum p_i which is at rest on average. This is called an ideal gas. The total energy of such a system of N particles[3] is given by its kinetic energy.

$$E = \frac{1}{2m}\sum_{i=1}^{N} p_i^2 \tag{4}$$

As mentioned previously for the canonical ensemble one then assumes that in equilibrium the *mean* energy should remain fixed. Note that for the microcanonical ensemble all states must be restricted to lie on a momentum hypersphere with radius $\sqrt{2mE}$. It is relatively easy to establish then that the maximum entropy distribution is uniform on that hypersphere and the ergodic results of Sinai and co-workers mentioned above allow this to be made rigorous. Converting such a hypersphere uniform distribution to a marginal distribution for the momentum of an individual particle is however a complicated exercise.

Now suppose instead one wished to deduce the probability distribution for all particles momentum $q(p_1, p_2, p_3, \ldots, p_N)$ for the canonical ensemble. The maximum entropy principle can be used subject to the constraint that the sum of second moments of momentum is fixed at some value ($2m$ times the mean energy of the open system). One can deduce using our arguments earlier that this distribution should be Gaussian in momentum and of a particular form which turns out to be the so-called Boltzmann distribution. Associated with this constraint is a Lagrange multiplier which turns out to be proportional to the inverse absolute temperature of the gas. Notice that the thermodynamical quantity is the temperature rather than the total energy which fluctuates. Naturally the form of the energy in equation (4) can be generalized to more complex situations in which the particles interact with external potentials and with themselves as well as have different masses (i.e. they are different molecules). We saw above an example of when there is an external gravitational potential. The maximum entropy principle means that the resulting distribution will belong to the exponential family. The integral of this (unnormalized) exponential function can be identified with the usual partition function of traditional statistical mechanics (see equation (3) above) and the derivatives of this with respect to the parameters describing energy or the associated Lagrange multipliers will

[3] N is usually huge for common applications and is related to the well known Avagadro's number which is of order 10^{24} .

then give us many thermodynamical properties of interest for the system. Other macro quantities associated with conservation laws will give us more elaborate maximum entropy problems and other thermodynamical quantities associated with the new Lagrange mulipliers. One example is particle number and the grand canonical ensemble explained above. The associated new Lagrange muliplier is then proportional to the quotient of the so-called chemical potential and absolute temperature. The conclusions derived from this approach have been extensively tested and verified in a very large variety of different physical systems.

This methodology for equilibrium statistical mechanics can also be applied successfully to the study of turbulent fluids (as a starting point see Salmon [6]). Certain fluids have conservation principles not just associated with energy but also other quantities such as enstrophy which is associated with fluid angular momentum and so one may add additional statistical constraints which results in a richer probability distribution structure. There are widespread applications in the literature.

References

[1] S. Amari and H. Nagaoka. *Methods of Information Geometry.* Translations of Mathematical Monographs, AMS, Oxford University Press, New York, 2000.

[2] Jaynes E. T. Information theory and statistical mechanics. *Phys. Rev.*, 106:620, 1957.

[3] J.L. Lebowitz and O. Penrose. Modern ergodic theory. *Physics Today*, 26(2):23–29, 1973.

[4] G.F. Mazenko. *Equilibrium statistical mechanics.* Wiley, 2000. 630pp.

[5] L. R. Mead and N. Papanicolaou. Maximum entropy in the problem of moments. *J. Math. Phys.*, 25:2404–2417, 1984.

[6] R. Salmon. *Lectures on Geophysical Fluid Dynamics.* Oxford Univ. Press, New York, 1998.

[7] Ya. G. Sinai. Ergodicity of Boltzmann's Equations. *Sov. Math.-Dokl.*, 4:1818, 1963.