Information Theory and Predictability
Lecture 5: Differential entropy and continuous outcome random variables

1. CONTINUOUS LIMIT

Above we consider random variables with countable outcomes. This may be generalized to the case where outcomes are from a continuum. Here the relevant probability object is the *probability density function* (pdf) which needs to be integrated over a particular interval in the continuum to become a probability in the usual sense of the word. There is a natural limiting process between these two concepts provided by the usual Riemann sum of integration. In the one dimensional case this can be expressed as follows:

$$\begin{array}{ccc} P_i & = & p(x_i^*)(x_{i+1}-x_i) \\ p(x_i^*) & = \frac{1}{(x_{i+1}-x_i)} & \int_{x=x_i}^{x=x_{i+1}} p(x)dx \end{array}$$

where$\{x_i\}$ is a particular partitioning of the one dimensional continuum under consideration.

The entropic functionals considered previously can be generalized to so called *differential* entropic functionals however some care needs to be exercised in interpreting them as limits of their discrete analogs. Thus for example the differential entropy $h(X)$ is defined as:

$$h(X) \equiv -\int_S p(x)\log(p(x))dx$$

where S is the continuous outcome set for X. If we convert this to a Riemann sum we obtain

$$h(X) \sim -\sum_{i\in\Lambda} p(x_i^*)\log(p(x_i^*))\Delta = -\sum_{i\in\Lambda} P_i \log P_i + \log\Delta = H(\tilde{X}) + \log\Delta \tag{1.1}$$

where Δ is the (assumed constant) volume element for the Riemann sum partitioning Λ chosen. Clearly as this approaches zero, the second term approaches $-\infty$ and the differential entropy is finite only because $H(\tilde{X})$ diverges to $+\infty$. This latter divergence occurs because the larger the size of the index set Λ the larger the entropy since there is increased choice in outcomes (recall Shannon's second axiom). One can overcome this rather awkward limiting process by restricting one's attention to entropy differences of different random variables in which case the $\log\Delta$ term cancels.

By contrast the differential relative entropy is a straightforward limit of the ordinary relative entropy:

$$\begin{array}{lll} D(p||q) & \equiv \int_S p(x)\log(p(x)/q(x))dx & \sim \sum_{i\in\Lambda} p(x_i^*)\log(p(x_i^*)/q(x_i^*))\Delta \\ & & = -\sum_{i\in\Lambda} P_i \log(P_i/Q_i) = D(P||Q) \end{array} \tag{1.2}$$

Note the cancellation of Δ in the third expression here.

This "cancellation" effect is also important to the transformational properties of the differential functionals. In particular suppose we have the following general non-linear change of variables:

$$\mathbf{y} = \mathbf{r}(\mathbf{x})$$

The transformed probability density $p^{'}(\mathbf{y})$ is well known to be given by

$$p^{'}(\mathbf{y}) = p(\mathbf{r^{-1}}(\mathbf{y})) \left| \det \left\{ \frac{\partial \mathbf{r^{-1}}(\mathbf{y})}{\partial \mathbf{y}} \right\} \right|$$

and the change of variables formula for integration has it that

$$d\mathbf{x} = \mathbf{dy} \left| \det \left\{ \frac{\partial \mathbf{r^{-1}}(\mathbf{y})}{\partial \mathbf{y}} \right\} \right| \equiv d\mathbf{y} \left| \det J \right|$$

where J is the so-called Jacobian of the transformation. So we have

$$\begin{array}{ll} D(p^{'}||q^{'}) = \int_{S'} p^{'}(\mathbf{y}) \log(p^{'}(\mathbf{y})/q^{'}(\mathbf{y})) d\mathbf{y} & = \int_S p(\mathbf{x}) \log((p(\mathbf{x}) \left| \det J \right|) / (q(\mathbf{x}) \left| \det J \right|)) d\mathbf{x} \\ & = D(p||q) \end{array}$$

providing the determinant of the transformation does not vanish which is a condition for the non-singularity of the transformation. Notice that this proof does not work for the differential entropy because of the lack of cancellation. The *difference* of the differential entropy of two random variables will be invariant under *linear* transformations because then $\det J$ is constant and the integral of the probability density is also constant (unity). For linear transformations one can easily establish in using the above arguments that

$$h(A\mathbf{x}) = h(\mathbf{x}) + \log \left| \det A \right|$$

Defining the continuous entropic functionals as the limits of countable functionals allows one to establish (providing of course that the limits exist) that all the important properties we established in Lecture 2 carry over to the continous case. In particular the non-negativity theorem for relative entropy; the chain rules for regular entropy as well as the relation between mutual information and conditional entropy all hold. Note that it becomes important to check that in fact the limits do actually exist. In practice many standard distributions used in mathematical statistics do have well defined differential entropies.

2. CONTINUOUS STOCHASTIC PROCESSES

2.1. **Fokker Planck Equation.** A particularly important class of stochastic processes is provided by solutions to the Fokker Plank Equation (FPE). They are often referred to as solutions of stochastic differential equations (see [1], for a good introduction). The FPE is an evolution equation for probability density functions (pdfs) defined on an N dimensional space of the form

$$\partial_t p = -\sum_{i=1}^{N} \partial_i \left[A_i(\mathbf{x}, t) p \right] + \frac{1}{2} \sum_{i,j=1}^{N} \partial_i \partial_j \left\{ \left[\mathbf{B}(\mathbf{x}, t) \mathbf{B}^t(\mathbf{x}, t) \right]_{ij} p \right\} \tag{2.1}$$

This equation has the intuitive interpretation that it represents the collection of possible evolutions of a dynamical system which is stochastically forced by Gaussian noise which is white in time (i.e. completely uncorrelated in the time dimension). The dynamical system has the form

$$\frac{\partial x_i}{\partial t} = A_i(\mathbf{x}, t) \tag{2.2}$$

and it's Gaussian forcing has a covariance matrix given by $C_{ij} = \left[\mathbf{B}(\mathbf{x},t)\mathbf{B}^{\mathbf{t}}(\mathbf{x},\mathbf{t})\right]_{ij}$. Note that this interpretation only works when C is non-negative definite and symmetric which we assume for the remainder of this section. More rigorously the above equation is associated with the Ito stochastic differential equation

$$dx_i = A_i(\boldsymbol{x},t)dt + B_{ij}(\boldsymbol{x},t)dW_j$$

where W_i is an uncorrelated vector of Wiener processes and we are assuming that repeated indices are summed.

Three interesting results are available regarding the evolution of the ordinary and relative entropy within such a system. The first two apply to systems without stochastic forcing

Theorem 1. *Suppose we have a realization of a stochastic process obeying equation (2.1) with* $\mathbf{B} = 0$

then the ordinary (differential) entropy satisfies the evolution equation

$$h_t = \int p\nabla \cdot \mathbf{A}dx = \langle \nabla \cdot \mathbf{A} \rangle_p \tag{2.3}$$

Proof. Let the realization of the process have pdf f then it follows that

$$\begin{aligned} -(f\ln(f))_t &= -f_t\left(\ln f + 1\right) \\ &= \nabla \cdot (\mathbf{A}f)(\ln f + 1) \\ &= f\nabla \cdot \mathbf{A}(\ln f + 1) + \mathbf{A}\cdot(\nabla f)(\ln f + 1) \\ &= f\nabla \cdot \mathbf{A}(\ln f + 1) + \nabla \cdot (\mathbf{A}f\ln f) - \nabla \cdot \mathbf{A}f\ln f \\ &= f\nabla \cdot \mathbf{A} + \nabla \cdot (\mathbf{A}f\ln f) \end{aligned}$$

The second term is in the form of a divergence so when integrated over all space contributes nothing to the entropy evolution due to Gauss's theorem (argument given in class). We are then left with equation (2.3). □

Notice the importance of $\nabla\cdot\mathbf{A}$ to the entropy evolution. This also measures the rate at which an infinitesimal volume element expands or contracts in the dynamical system. When it vanishes the system is sometimes said to satisfy a Liouville condition. Many inviscid (frictionless) fluids satisfy such a condition. We shall use equation (2.3) in a central way later when we consider the concept of information flow.

The relative entropy on the other hand is conserved in all systems with $\mathbf{B} = 0$:

Theorem 2. *Suppose we have two realizations of a stochastic process obeying equation (2.1) which have the additional condition that* $\mathbf{B}(\mathbf{x},t) = 0$ *then the relative entropy of the two realizations (if defined) is time invariant.*

Proof. Let the two realizations of the process have pdfs f and g then it follows that

$$\begin{aligned} (f\ln(f/g))_t &= f_t ln(f/g) + f_t - g_t(f/g) = f_t(\ln(f/g)+1) - g_t(f/g) \\ &= -\nabla \cdot (\mathbf{A}f)(\ln(f/g)+1) + \nabla \cdot (\mathbf{A}g)(f/g) \\ &= \left[-f\nabla \cdot \mathbf{A} - \mathbf{A}\cdot(\nabla f)\right]\left[\ln(f/g)+1\right] + \left[g\nabla \cdot \mathbf{A} + \mathbf{A}\cdot(\nabla g)\right](f/g) \\ &= -\nabla \cdot \mathbf{A}f\ln(f/g) - \mathbf{A}\cdot(\nabla f\left[\ln(f/g)+1\right] - \nabla g(f/g)) \\ &= -\nabla \cdot \mathbf{A}f\ln(f/g) - \mathbf{A}\cdot\nabla(f\ln(f/g)) \\ &= -\nabla \cdot (\mathbf{A}f\ln(f/g)) \end{aligned}$$

In this case the entire right hand side of the evolution equation is in the form of a divergence and as argued in the previous theorem this implies that the global integral of the left hand side vanishes by Gauss's theorem. □

In many classical systems with $\mathbf{B} = 0$ if one calculates the relative entropy with respect to a particular finite partitioning of state space rather in the limit of infinitesimal partitioning then the conservation property no longer holds and in nearly all interesting cases it declines with time instead and the system equilibriates. This reflects the fact that as time increases the difference in the distributions tends to occur on the unresolved scales which are not measured by the second relative entropy calculation. This *coarse graining* effect is often related to the next result.

In the stochastically forced case we have:

Theorem 3. *Suppose we have two distinct*[1] *stochastic processes obeying (2.1) with* $C = \mathbf{B}(\mathbf{x}, t)\mathbf{B}^{\mathbf{t}}(\mathbf{x}, \mathbf{t})$ *positive definite almost everywhere then the relative entropy strictly declines.*

Proof. As in the previous therem we consider the relative entropy "density" function $r = f \ln f/g$. Clearly the proof of this shows we need only consider the time rate of change in this function due to C since that due to $\mathbf{A}$ leads to no change in time of the global integral of r. The change in r due to C is easily calculated using equation (2.1):

$$(r_c)_t = (\ln(f/g) + 1)\, \partial_i \partial_j \left(C_{ij} f\right) - \frac{f}{g} \partial_i \partial_j \left(C_{ij} g\right) \tag{2.4}$$

where we are using the summation convention for repeated latin indices. Now it is easy to see that

$$\partial_i \partial_j (C_{ij} uw) = w \partial_i \partial_j (C_{ij} u) + 2(\partial_j (C_{ij} u))(\partial_i w) + C_{ij} u \partial_i \partial_j (w) \tag{2.5}$$

where we are using the symmetry of C. Writing $g = f(g/f)$ and applying the last relation we derive that the second term of equation (2.4) is

$$(2) = -\frac{f}{g}\left[\frac{g}{f}\partial_i \partial_j (C_{ij} f) + 2\partial_i (C_{ij} f)\partial_i \left(\frac{g}{f}\right) + C_{ij} f \partial_i \partial_j \left(\frac{g}{f}\right)\right]$$

combining this with the first term we get a cancellation of the first term of (2) with part of the first term of equation (2.4) and so

$$(r_c)_t = \ln(f/g)\partial_i \partial_j \left(C_{ij} f\right) - 2\left(\frac{f}{g}\right)\partial_j (C_{ij} f)\partial_i \left(\frac{g}{f}\right) - \left(\frac{f}{g}\right) C_{ij} f \partial_i \partial_j \left(\frac{g}{f}\right)$$

Now to this equation we add and subtract the terms

$$2\partial_i (\ln \frac{f}{g})\partial_j (C_{ij} f) + C_{ij} f \partial_i \partial_j \left(\ln \frac{f}{g}\right)$$

and use equation (2.5) to deduce that

$$(r_c)_t = \partial_i \partial_j (C_{ij} r) - \left(C_{ij} f\left[\frac{f}{g}\partial_i \partial_j \left(\frac{g}{f}\right) + \partial_i \partial_j \left(\ln \frac{f}{g}\right)\right]\right)$$

where we are using the definition of r as well as cancelling two terms involving $\partial_j (C_{ij} f)$. It is straightforward (albeit tedious) to simplify the expression in the square brackets and obtain finally

$$(r_c)_t = \partial_i \partial_j (C_{ij} r) - f C_{ij} \partial_i (\ln \frac{f}{g})\partial_j (\ln \frac{f}{g}) \tag{2.6}$$

The first term on the right is of the form of a divergence and so as usual does not contribute to the evolution of the global integral of r_C. Actually the positive

[1] In other words differing on a set of measure greater than zero.

definite nature of C shows that it is purely diffusive of the density r. The second term is negative almost everywhere due to the fact that C is positive definite almost everywhere and that f and g differ almost everywhere. Thus in that situation if we take the global integral of r_C we conclude that the relative entropy declines strictly with time. □

This third theorem shows the central role of stochastic forcing in causing relative entropy to decline.

In stochastic modeling of dynamical systems it is common to separate the state space into fast and slow components and model the former with noise terms and dissipation of the slow modes. Presumably in this case if such a model works well for the total unforced system then the last two theorems imply that there is a "leakage" of relative entropy from the slow to the fast components of the system.

In practical systems such as the atmosphere, we find that if a slow subspace is chosen which represents a large fraction of the variability of the system then the subspace relative entropy will always show a monotonic decline which is suggestive that stochastic models of this system may work well.

It is possible to extend the Fokker Planck equation to include discontinuous jump processes and then this equation becomes the more general Chapman-Kolmogorov equation. The additional terms are often referred to (on their own) as the Master equation. It is then possible by similar arguments to those given above to conclude that the jump processes result in an additional strict monotonic decline in relative entropy. The interested reader is referred to Chapter 3 of [1] for a sketch proof and more information and references.

References

[1] C. W. Gardiner. *Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences*, volume 13 of *Springer Series in Synergetics*. Springer, 2004.