1. Introduction

The major focus later in this course will be on statistical predictability of dynamical systems. In that context the primary issue we will consider is the evolution of uncertainty within a system. One of the biggest issues in practical prediction concerns proper definition of the initial conditions for a deterministic prediction using a (usually very complex) dynamical model. For obvious practical reasons such conditions are imperfectly known mainly because the system under consideration can only be measured at a particular time to within a certain precision. Consequently it makes conceptual sense to regard the dynamical variables of such systems as random variables and to study in detail the temporal evolution of such variables.

The generic name for such an evolution is a “stochastic process”. There exists a large and extensive literature in applied mathematics on such processes which we (briefly) review here from the perspective of information theory.

**Definition 1.** A stochastic process is a set of random variables \{X(\alpha)\} with \(\alpha \in A\) an ordered set.

Usually we shall take \(A\) to be the reals or integers and give this ordered set the interpretation of time.

If the probability functions defined on a stochastic process satisfy particular relations then the process is said to be of a particular (restrictive) kind. Among the more important of these are the following:

**Definition 2.** A stochastic process is said to be **time-homogeneous (or time invariant)** if \(A = R\) and the conditional probabilities satisfy

\[
p(x(t)|x(t-a)) = p(x(s)|x(s-a)) \quad \forall t, s \in R \text{ and } \forall a > 0
\]

The case where \(A = Z\) is similar. This has an attractive (physical) interpretation in terms of statistical causality since it says that if the random variable is known exactly at a particular time then the later probability of the variable is specified uniquely. Note the introduction of an “arrow in time” since we require that \(a > 0\).

Some time-homogeneous processes may be “reversible” in the sense that we also have equation (1.1) holding for \(a < 0\). We will consider such processes later.

A stochastic process is called discrete if \(A = Z\). It is written \(\{X_j\}\). Often we consider only positive integers \((A = N)\) to allow for an “initial condition” at \(j = 1\). In that case

**Definition 3.** A discrete stochastic process is called **Markov** if it satisfies

\[
p(x_{j+1}|x_j, x_{j-1}, x_{j-2}, \ldots, x_1) = p(x_{j+1}|x_j) \quad \forall j
\]

It is easily seen that repeated use of this property with increasing \(j\) allows us to write for all \(n > 1\):

\[
p(x_1, x_2, \ldots, x_n) = p(x_1)p(x_2|x_1)p(x_3|x_2) \ldots p(x_n|x_{n-1})
\]
showing that the initial condition probability and the conditional probability functions $p(x_j|x_{j-1})$ are all that are required to describe a Markov process. Intuitively a Markov process is one in which the probability at a given step depends only on the previous step and not on earlier steps.

In most discussion of Markov processes it is normal to also assume that they are time-homogeneous. Note that these conditions are independent in the sense that there exist time-homogeneous non-Markov and non-time-homogeneous Markov processes. For more detail see Chapter 1 from [2].

In the normal time-homogeneous Markov case mentioned it is clear that the conditional probabilities $p(x_j|x_{j-1})$ serve (with the initial condition probability) to describe the complete set of probability functions for the system. They are referred to as the transition functions (or matrix when $X_j$ has a countable set of values).

In general if some of the transition functions are zero then it will not be possible to pass from a particular state to another in one time step. However it may be possible to do so in more than one steps i.e. indirectly. This motivates the following properties of some Markov processes:

Definition 4. A Markov process in which it is possible to pass from any state to another in a finite number of time steps is called irreducible. If the possible path lengths (number of time-steps) of such a transition has a common factor of 1 then the process is further termed aperiodic.

Often as time increases the probability function of a Markov process will tend to converge. In particular

Definition 5. A time-homogeneous Markov process for which $p(x_{n+1}) = p(x_n)$ is called stationary.

It is proven in standard texts on Markov processes that if the process is irreducible and aperiodic then asymptotically (i.e. as time increases) it converges to a unique stationary process. This stationary process can be found for the case that there are a finite number of states as an eigenvector of the transition matrix with eigenvalue one.

Later in this course we will consider practical stochastic processes such as those applicable to atmospheric prediction. We shall assume that here there exists a unique asymptotic process which we will refer to as the equilibrium or climatological process. In the real system it is not stationary but instead varies periodically with the annual cycle (to a very good approximation anyway). Such a process is sometimes termed cyclo-stationary.

The concept of probabilistic causality introduced in the context of time homogeneous stochastic processes can be used to define a similarity relation on stochastic processes:

Definition 6. Suppose we have two stochastic processes $X(t)$ and $Y(t)$ defined on the ordered set $R$ with associated probability functions $p$ and $q$ and the same outcome sets $\{\mathcal{X}(t)\}$. We say that the two processes are causally similar if

$$p(\mathcal{X}(t)|\mathcal{X}(t-a)) = q(\mathcal{X}(t)|\mathcal{X}(t-a)) \quad \forall t \text{ and } \forall a > 0$$

It is obvious that all time translated time-homogeneous processes are causally similar. This condition expresses intuitively the notion that the physical system giving rise to the two processes is identical and that the set of values for outcomes at a particular time are sufficient to determine the future evolution of probabilities.
This is intuitively the case for a closed physical system however not the case for an open system since in this second case other external variables may influence dynamical evolution. If two processes are time homogenous and Markov and share the same transition matrix it is also easy to show that they are causally similar. Finally if two Markov processes (not necessarily time homogenous) share the same transition matrix at the same time step then they are also causally similar.

2. Entropy Rates

For simplicity we consider $A = \mathbb{N}$. As time increases and new random variables are added then the entropy of the whole system evidently increases. The average entropy of the whole sequence $\{X_n\}$ may however converge to a particular value which we call the entropy rate $H(\chi)$ of the process:

$$H(\chi) \equiv \lim_{n \to \infty} \frac{1}{n} H(X_1, X_2, \ldots, X_n)$$

A related concept is the modified entropy rate $H'(\chi)$ which asks for the asymptotic entropy of the last variable given everything that has happened previously:

$$H'(\chi) \equiv \lim_{n \to \infty} H(X_n | X_{n-1}, X_{n-2}, \ldots, X_1)$$

Obviously if the last variable is not independent of the previous variables then this will be less than the limiting entropy of each additional variable. In interesting dynamical systems (including Markov processes) such dependency is obviously usual.

If all the $X_j$ are identically and independently distributed (i.i.d.) then it easy to see that in fact $H(\chi) = H'(\chi) = H(X_j)$. For a more general stationary stochastic process we have the result

**Theorem 7.** For a stationary stochastic process $H(\chi) = H'(\chi)$

**Proof.** Since additional conditioning reduces entropy (see previous lecture) we have

$$H(X_{n+1} | X_n, \ldots, X_1) \leq H(X_{n+1} | X_n, X_2)$$

and if the process is stationary then the right hand side is equal to $H(X_n | X_{n-1}, \ldots, X_1)$. This shows that as $n$ increases then $H(X_{n+1} | X_n, \ldots, X_1)$ does not increase and since it is positive (and thus bounded below) the monotonic convergence theorem of sequences shows the limit as $n \to \infty$ must exist and by our definition be $H'(\chi)$. A fairly standard result in sequence theory (Cesaro mean result) states that if we have a sequence $\{a_n\}$ with $\lim_{n \to \infty} a_n = a$ then the sequence of averages $b_n \equiv \frac{1}{n} \sum_{i=1}^{n} a_i$ converges also to $a$. Now the chain rule for entropy from the previous lecture shows that

$$\frac{1}{n} H(X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} H(X_i | X_{i-1}, \ldots, X_1)$$

Since we have already established that the limit of the conditional entropies in the sum on the right hand side approaches the particular value $H'(\chi)$ it follows by the Cesaro mean result that the limit of the average of these also approaches this value. The limit of the left hand side is of course our definition for the entropy rate $H(\chi)$.\qed
For a stationary time-homogeneous Markov chain these rates are easily calculated since

\[ H(\chi) = H'(\chi) = \lim_{n \to \infty} H(X_n|X_{n-1}, X_{n-2}, \ldots, X_1) \]

and the Markov property implies that the expression in the limit is equal to \( H(X_n|X_{n-1}) \) (exercise) and the time-homogeneous property implies all conditional entropies of this type are equal to \( H(X_2|X_1) \) which is therefore equal to both entropy rates.

We consider a very simple example to illustrate some of the concepts introduced: Suppose we have a time-homogeneous Markov process consisting of two states and a two by two transition matrix (we apply it from the left to probability vectors) given by

\[ P = p(\overrightarrow{x_j}|\overrightarrow{x_{j-1}}) \equiv \begin{pmatrix} 1 - r & s \\ r & 1 - s \end{pmatrix} \]

Note that this matrix preserves the required property that the sum of the components of the probability vector is unity (Exercise: What property do we need to ensure \( 0 \leq p_i \leq 1 \)?). The matrix is also easily shown to have the characteristic equation \((\lambda - 1)(\lambda + r + s - 1)\) which shows that there exists a stationary probability vector (the eigenvector with eigenvalue 1) given by

\[ \begin{pmatrix} \frac{s}{r+s} \\ \frac{r}{r+s} \end{pmatrix} \]

which if taken as the first member of the Markov chain will define a stationary process. The entropy of the stationary process at any step is easily calculated as

\[ H(X_n) = -\frac{r}{r+s} \log \left( \frac{r}{r+s} \right) - \frac{s}{r+s} \log \left( \frac{s}{r+s} \right) \]

while the entropy rate can be shown to be (using the definition of conditional entropy)

\[ H(\chi) = H(X_j|X_{j-1}) = -\frac{s}{r+s} [r \log r + (1 - r) \log (1 - r)] - \frac{r}{r+s} [s \log s + (1 - s) \log (1 - s)] \]

As we pointed out above the entropy rate is the additional uncertainty to the whole chain introduced by adding another random variable and in general this will be less than the entropy of this particular variable by itself since there is dependency between the random variable then and earlier random variables in the chain. The ratio of these two quantities is plotted below for ranges of some allowable values of \( r \) and \( s \).
3. A GENERALIZED SECOND LAW OF THERMODYNAMICS

The information theoretic concept of entropy introduced here has been compared in detail to that used in statistical mechanics (and hence thermodynamics) by Jaynes [1] so a natural question that arises in the context of stochastic processes is whether entropy is non-decreasing\(^1\). Rather surprisingly the answer is no and the reason is rather simple: If the asymptotic process probability is non-uniform while the initial condition probability is uniform, it follows from the maximal property of the entropy of a uniform probability function that at some point in time the entropy must decrease. In the molecular dynamics context of statistical mechanics it turns out that the asymptotic (equilibrium) distribution is actually uniform (but constrained by energy conservation) and so this counter argument does not hold.

Despite the non-monotonic temporal nature of regular entropy there is a simple generalization using relative entropy which allows us to recover the usual second law in the case of an asymptotic uniform probability function. In particular we have the following theorem

**Theorem 8.** Suppose we have two causally similar stochastic processes with probability functions at time \(t\) of \(p(\mathcal{X}_t)\) and \(q(\mathcal{X}_t)\). Then

\[
D(p(\mathcal{X}_t)||q(\mathcal{X}_t)) \leq D(p(\mathcal{X}_s)||q(\mathcal{X}_s)) \quad \text{when } t > s
\]

\(\text{In statistical mechanics there exist processes called reversible in which entropy is conserved. The other processes are called irreversible.}\)
Proof. Consider the joint distributions $p(\vec{x}_t, \vec{x}_s)$ and $q(\vec{x}_t, \vec{x}_s)$ between random variables at times $t$ and $s$ then the chain rule of relative entropy (Theorem 3.3 previous lecture) shows that

$$D(p(\vec{x}_t, \vec{x}_s)||q(\vec{x}_t, \vec{x}_s)) = D(p(\vec{x}_t)||q(\vec{x}_t)) + D(p(\vec{x}_s|\vec{x}_t)||q(\vec{x}_s|\vec{x}_t))$$

Consider now the second terms on both lines. From the definition of causally similar processes and that of conditional relative entropy it follows that one or the other must vanish depending on the ordering of $t$ and $s$. Of course it is possible that both may vanish in which case relative entropy is conserved but nevertheless the inequality of the theorem must hold.

Now suppose for a given closed physical system that there exists an equilibrium stochastic process at all times then it is reasonable to assume that such a process is causally similar to any other stochastic process derived from the same physical system and so the relative entropy between the two processes is non-increasing. Note also that the arrow of time follows directly from an assumption of causality similarity not from a Markov property per se. This makes intuitive and philosophical sense since a direction in time should arise from a very basic physical principle i.e. here cause preceding effect. Remember though that Markov processes sharing the same transition matrix at a given time are causally similar.

In the case that the asymptotic function $q$ is uniform then it is easy to check that the relative entropy reduces to minus the regular entropy plus the (constant) entropy of the uniform probability function. Thus for such a special situation (as holds in molecular dynamics) then the regular entropy is non-decreasing in time.

There is one kind of stochastic process for which the conditional regular entropy is actually non-decreasing:

**Theorem 9.** In a stationary Markov process the entropy conditioned on the initial condition is non-decreasing.

Proof. The conditional entropy is $H(X_n|X_1)$ which is greater than or equal to $H(X_n|X_1, X_2)$ since further conditioning reduces entropy/uncertainty. By the Markov and stationarity assumption successively we have

$$H(X_n|X_1, X_2) = H(X_n|X_2) = H(X_n-1|X_1)$$

which shows that $H(X_n|X_1)$ is non-decreasing.

In the field of statistical predictability in general only asymptotic processes are stationary so this result has limited application in that particular context.

**References**
