Chapter 3
Entropy

Abstract The development of the idea of entropy of random variables and processes by Claude Shannon provided the beginnings of information theory and of the modern age of ergodic theory. Entropy and related information measures will be shown to provide useful descriptions of the long term behavior of random processes and this behavior is a key factor in developing the coding theorems of information theory. Here the various notions of entropy for random variables, vectors, processes, and dynamical systems are introduced and and there fundamental properties derived. In this chapter the case of finite-alphabet random processes is emphasized for simplicity, reflecting the historical development of the subject. Occasionally we consider more general cases when it will ease later developments.

3.1 Entropy and Entropy Rate

There are several ways to introduce the notion of entropy and entropy rate. The difference between the two concepts is that entropy is relevant to a single random variable or random vector or, equivalently, to a partition of the sample space, while entropy rate describes a limiting entropy per time unit is we look at sample vectors with increasing dimensions or iterates of a partition. We take some care at the beginning in order to avoid redefining things later. We also try to use definitions resembling the usual definitions of elementary information theory where possible. Let $(\Omega, \mathcal{B}, P, T)$ be a dynamical system. Let $f : \Omega \to A$ be a finite-alphabet measurement (a simple function) defined on Ω and define the random process $f_n = fT^n; n \in \mathbb{T}$. For the moment we focus on one sided processes with $\mathbb{T} = \{0, 1, 2, \ldots\}$. If the transformation T is invertible, we can extend the definition to all integer n and obtain a two-sided process with $\mathbb{T} = \mathbb{Z}$. This process can be viewed as a *coding* of the original space, that

R.M. Gray, *Entropy and Information Theory*, DOI 10.1007/978-1-4419-7970-4_3,

is, one produces successive coded values by transforming (e.g., shifting) the points of the space, each time producing an output symbol using the same rule or mapping. If Ω is itself a sequence space and T is a shift, then f is a sliding-block code as considered in Section 2.6 and it induces a stationary sequence code $\overline{f} = \{fT^n; n \in \mathbb{T}\}$.

In the usual way we can construct an equivalent Kolmogorov model of this process. Let $A = \{a_1, a_2, \ldots, a_{\|A\|}\}$ denote the finite alphabet of f and let $(A^{Z_+}, \mathcal{B}_A^{Z_+})$ be the resulting one-sided sequence space, where $\mathcal{B}_A$ is the power set. We abbreviate the notation for this sequence space to $(A^\infty, \mathcal{B}_A^\infty)$. Let T_A denote the shift on this space and let X denote the time zero sampling or coordinate function and define $X_n(x) = X(T_A^n x) = x_n$. Let m denote the process distribution induced by the original space and the fT^n, i.e., $m = P_{\overline{f}} = P\overline{f}^{-1}$ where $\overline{f}(\omega) = (f(\omega), f(T\omega), f(T^2\omega), \ldots)$.

Observe that by construction, shifting the input point yields an output sequence that is also shifted, that is,

$$\overline{f}(T\omega) = T_A \overline{f}(\omega).$$

Sequence-valued measurements of this form are called *stationary* or *invariant* codings (or *time-invariant* or *shift-invariant* codings in the case of the shift) since the coding commutes with the transformations. Stationary codes will play an important role throughout this book and are discussed in some detail in Chapter 2. If the input space Ω is itself a sequence space and T is a shift, then the code is also called a *sliding-block code* to reflect the fact that the code operates by shifting the input sequence (sliding) and applying a common measurement or mapping to it. Both the sequence-to-symbol mapping f and the sequence-to-sequence mapping $\overline{f}$ are referred to as a sliding-block code, each implies the other.

The entropy and entropy rates of a finite-alphabet measurement depend only on the process distributions and hence are usually more easily stated in terms of the induced directly given model and the process distribution. For the moment, however, we point out that the definition can be stated in terms of either system. Later we will see that the entropy of the underlying system is defined as a supremum of the entropy rates of all finite-alphabet codings of the system.

The *entropy* of a discrete alphabet random variable f defined on the probability space $(\Omega, \mathcal{B}, P)$ is defined by

$$H_P(f) = -\sum_{a \in A} P(f = a) \ln P(f = a). \tag{3.1}$$

We define $0 \ln 0$ to be 0 in the above formula. We shall often use logarithms to the base 2 instead of natural logarithms. The units for entropy are "nats" when the natural logarithm is used and "bits" for base 2 logarithms. The natural logarithms are usually more convenient for mathe-

matics while the base 2 logarithms provide more intuitive descriptions. The subscript P can be omitted if the measure is clear from context. Be forewarned that the measure will often not be clear from context since more than one measure may be under consideration and hence the subscripts will be required. A discrete alphabet random variable f has a probability mass function (PMF), say p_f, defined by $p_f(a) = P(f = a) = P(\{\omega : f(\omega) = a\})$ and hence we can also write

$$H(f) = -\sum_{a \in A} p_f(a) \ln p_f(a).$$

It is often convenient to consider the entropy not as a function of the particular outputs of f but as a function of the partition that f induces on Ω. In particular, suppose that the alphabet of f is $A = \{a_1, a_2, \ldots, a_{\|A\|}\}$ and define the partition $\mathcal{Q} = \{Q_i; i = 1, 2, \ldots, \|A\|\}$ by $Q_i = \{\omega : f(\omega) = a_i\} = f^{-1}(\{a_i\})$. In other words, $\mathcal{Q}$ consists of disjoint sets which group the points in Ω together according to what output the measurement f produces. We can consider the entropy as a function of the partition and write

$$H_P(\mathcal{Q}) = -\sum_{i=1}^{\|A\|} P(Q_i) \ln P(Q_i). \tag{3.2}$$

Clearly different mappings with different alphabets can have the same entropy if they induce the same partition. Both notations will be used according to the desired emphasis. We have not yet defined entropy for random variables that do not have discrete alphabets; we shall do that later.

Return to the notation emphasizing the mapping f rather than the partition. Defining the random variable $P(f)$ by $P(f)(\omega) = P(\lambda : f(\lambda) = f(\omega))$ we can also write the entropy as

$$H_P(f) = E_P(-\ln P(f)).$$

Using the equivalent directly given model we have immediately that

$$H_P(f) = H_P(\mathcal{Q}) = H_m(X_0) = E_m(-\ln m(X_0)). \tag{3.3}$$

At this point one might ask why we are carrying the baggage of notations for entropy in both the original space and in the sequence space. If we were dealing with only one measurement f (or X_n), we could confine interest to the simpler directly-given form. More generally, however, we will be interested in different measurements or codings on a common system. In this case we will require the notation using the original system. Hence for the moment we keep both forms, but we shall often focus on the second where possible and the first only when necessary.

The *nth order entropy of a discrete alphabet measurement f with respect to T* is defined as

$$H_P^{(n)}(f) = n^{-1}H_P(f^n)$$

where $f^n = (f, fT, fT^2, \ldots, fT^{n-1})$ or, equivalently, we define the discrete alphabet random process $X_n(\omega) = f(T^n\omega)$, then

$$f^n = X^n = X_0, X_1, \ldots, X_{n-1}.$$

As previously, this is given by

$$H_m^{(n)}(X) = n^{-1}H_m(X^n) = n^{-1}E_m(-\ln m(X^n)).$$

This is also called the entropy (per-coordinate or per-sample) of the random vector f^n or X^n. We can also use the partition notation here. The partition corresponding to f^n has a particular form: Suppose that we have two partitions, $\mathcal{Q} = \{Q_i\}$ and $\mathcal{P} = \{P_i\}$. Define their *join* $\mathcal{Q} \bigvee \mathcal{P}$ as the partition containing all nonempty intersection sets of the form $Q_i \bigcap P_j$. Define also $T^{-1}\mathcal{Q}$ as the partition containing the atoms $T^{-1}Q_i$. Then f^n induces the partition

$$\bigvee_{i=0}^{n-1} T^{-i}\mathcal{Q}$$

and we can write

$$H_P^{(n)}(f) = H_P^{(n)}(\mathcal{Q}) = n^{-1}H_P(\bigvee_{i=0}^{n-1} T^{-i}\mathcal{Q}).$$

As before, which notation is preferable depends on whether we wish to emphasize the mapping f or the partition $\mathcal{Q}$.

The *entropy rate* or *mean entropy* of a discrete alphabet measurement f with respect to the transformation T is defined by

$$\begin{aligned}\overline{H}_P(f) &= \limsup_{n\to\infty} H_P^{(n)}(f) \\ &= \overline{H}_P(\mathcal{Q}) = \limsup_{n\to\infty} H_P^{(n)}(\mathcal{Q}) \\ &= \overline{H}_m(X) = \limsup_{n\to\infty} H_m^{(n)}(X).\end{aligned}$$

Given a dynamical system $(\Omega, \mathcal{B}, P, T)$, the *entropy* $H(P,T)$ of the system (or of the measure with respect to the transformation) is defined by

$$H(P,T) = \sup_f \overline{H}_P(f) = \sup_{\mathcal{Q}} \overline{H}_P(\mathcal{Q}), \tag{3.4}$$

where the supremum is over all finite-alphabet measurements (or codings) or, equivalently, over all finite measurable partitions of Ω. (We emphasize that this means alphabets of size M for all finite values of M.) The entropy of a system is also called the *Kolmogorov-Sinai invariant* of the system because of the generalization by Kolmogorov [102] and Sinai [168] of Shannon's entropy rate concept to dynamical systems and the demonstration that equal entropy was a necessary condition for two dynamical systems to be isomorphic.

Note that the entropy rate is well-defined for a continuous-alphabet random process as the supremum over the entropy rates over all finite-alphabet codings of the process. Such an entropy rate is usually infinite, but it is defined.

Suppose that we have a dynamical system corresponding to a finite-alphabet random process $\{X_n\}$, then one possible finite-alphabet measurement on the process is $f(x) = x_0$, that is, the time 0 output. In this case clearly $\overline{H}_P(f) = \overline{H}_P(X)$ and hence, since the system entropy is defined as the supremum over *all* simple measurements,

$$H(P,T) \geq \overline{H}_P(X). \tag{3.5}$$

We shall later see in Theorem 6.1 that (3.5) holds with equality for finite alphabet random processes and provides a generalization of entropy rate for processes that do not have finite alphabets.

3.2 Divergence Inequality and Relative Entropy

Many of the basic properties of entropy follow from a simple result known as the *divergence inequality*. A slight variation is well-known as the log-sum inequality). The divergence or relative entropy is a variation on the idea of entropy and it crops up often as a useful tool for proving and interpreting results and for comparing probability distributions. In this section several fundamental definitions and results are collected together for use in the next section in developing the properties of entropy and entropy rate.

Lemma 3.1. *Given two probability mass functions $\{p_i\}$ and $\{q_i\}$, that is, two countable or finite sequences of nonnegative numbers that sum to one, then*

$$\sum_i p_i \ln \frac{p_i}{q_i} \geq 0$$

with equality if and only if $q_i = p_i$, all i.

Proof: The lemma follows easily from the elementary inequality for real numbers

$$\ln x \le x - 1 \tag{3.6}$$

(with equality if and only if $x = 1$) since

$$\sum_i p_i \ln \frac{q_i}{p_i} \le \sum_i p_i \left(\frac{q_i}{p_i} - 1\right) = \sum_i q_i - \sum_i p_i = 0$$

with equality if and only if $q_i/p_i = 1$ all i. Alternatively, the inequality follows from Jensen's inequality [74] since ln is a convex $\bigcap$ function:

$$\sum_i p_i \ln \frac{q_i}{p_i} \le \ln\left(\sum_i p_i \frac{q_i}{p_i}\right) = 0$$

with equality if and only if $q_i/p_i = 1$, all i. □

The inequality has a simple corollary that we record now for later use.

Corollary 3.1. *The function $x \ln(x/y)$ of real positive x, y is convex in (x, y).*

Proof. Let (x_i, y_i), $i = 1, 2$, be pairs of real positive numbers, $0 \le \lambda \le 1$, and define $x = \lambda x_1 + (1-\lambda)x_2$ and $y = \lambda y_1 + (1-\lambda)y_2$. Apply Lemma 3.1 to the probability mass functions p and q defined by

$$\begin{aligned} p_1 &= \frac{\lambda x_1}{\lambda x_1 + (1-\lambda)x_2} \\ p_2 &= \frac{(1-\lambda)x_2}{\lambda x_1 + (1-\lambda)x_2} \\ q_1 &= \frac{\lambda y_1}{\lambda y_1 + (1-\lambda)y_2} \\ q_2 &= \frac{(1-\lambda)y_2}{\lambda y_1 + (1-\lambda)y_2} \end{aligned}$$

yields

$$0 \le \frac{\lambda x_1}{\lambda x_1 + (1-\lambda)x_2} \ln\left(\frac{\frac{\lambda x_1}{\lambda x_1 + (1-\lambda)x_2}}{\frac{\lambda y_1}{\lambda y_1 + (1-\lambda)y_2}}\right) + \frac{(1-\lambda)x_2}{\lambda x_1 + (1-\lambda)x_2} \ln\left(\frac{\frac{(1-\lambda)x_2}{\lambda x_1 + (1-\lambda)x_2}}{\frac{(1-\lambda)y_2}{\lambda y_1 + (1-\lambda)y_2}}\right).$$

Cancelling the positive denominator and rearranging

$$\lambda x_1 \ln \frac{x_1}{y_1} + (1-\lambda)\lambda x_2 \ln \frac{x_2}{y_2} \ge (\lambda x_1 + (1-\lambda)x_2) \ln\left(\frac{\lambda x_1 + (1-\lambda)x_2}{\lambda y_1 + (1-\lambda)y_2}\right),$$

proving the claimed convexity. □

The quantity used in Lemma 3.1 is of such fundamental importance that we introduce another notion of information and recast the inequality in terms of it. As with entropy, the definition for the moment is only for finite-alphabet random variables. Also as with entropy, there are a variety of ways to define it. Suppose that we have an underlying measurable space $(\Omega, \mathcal{B})$ and two measures on this space, say P and M, and we have a random variable f with finite alphabet A defined on the space and that $\mathcal{Q}$ is the induced partition $\{f^{-1}(a); a \in A\}$. Let P_f and M_f be the induced distributions and let p and m be the corresponding probability mass functions, e.g., $p(a) = P_f(\{a\}) = P(f = a)$. Define the *relative entropy* of a measurement f with measure P with respect to the measure M by

$$H_{P\|M}(f) = H_{P\|M}(\mathcal{Q}) = \sum_{a\in A} p(a) \ln \frac{p(a)}{m(a)} = \sum_{i=1}^{\|A\|} P(Q_i) \ln \frac{P(Q_i)}{M(Q_i)}.$$

Observe that this only makes sense if $p(a)$ is 0 whenever $m(a)$ is, that is, if P_f is absolutely continuous with respect to M_f or $M_f \gg P_f$. Define $H_{P\|M}(f) = \infty$ if P_f is not absolutely continuous with respect to M_f. The measure M is referred to as the *reference measure*. Relative entropies will play an increasingly important role as general alphabets are considered. In the early chapters the emphasis will be on ordinary entropy with similar properties for relative entropies following almost as an afterthought. When considering more abstract (nonfinite) alphabets later on, relative entropies will prove indispensible.

Analogous to entropy, given a random process $\{X_n\}$ described by two process distributions p and m, if it is true that

$$m_{X^n} \gg p_{X^n};\ n = 1, 2, \ldots,$$

then we can define for each n the *nth order relative entropy* $n^{-1}H_{p\|m}(X^n)$ and the *relative entropy rate*

$$\overline{H}_{p\|m}(X) \equiv \limsup_{n\to\infty} \frac{1}{n} H_{p\|m}(X^n).$$

When dealing with relative entropies it is often the measures that are important and not the random variable or partition. We introduce a special notation which emphasizes this fact. Given a probability space $(\Omega, \mathcal{B}, P)$, with Ω a finite space, and another measure M on the same space, we define the *divergence of P with respect to M* as the relative entropy of the identity mapping with respect to the two measures:

$$D(P\|M) = \sum_{\omega\in\Omega} P(\omega) \ln \frac{P(\omega)}{M(\omega)}.$$

Thus, for example, given a finite-alphabet measurement f on an arbitrary probability space $(\Omega, \mathcal{B}, P)$, if M is another measure on $(\Omega, \mathcal{B})$ then

$$H_{P\|M}(f) = D(P_f\|M_f).$$

Similarly,

$$H_{p\|m}(X^n) = D(P_{X^n}\|M_{X^n}),$$

where P_{X^n} and M_{X^n} are the distributions for X^n induced by process measures p and m, respectively. The theory and properties of relative entropy are therefore determined by those for divergence.

There are many names and notations for relative entropy and divergence throughout the literature. The idea was introduced by Kullback for applications of information theory to statistics (see, e.g., Kullback [106] and the references therein) and was used to develop information theoretic results by Perez [145] [147] [146], Dobrushin [32], and Pinsker [150]. Various names in common use for this quantity are discrimination, discrimination information, Kullback-Leibler (KL) number , directed divergence, informational divergence, and cross entropy.

The lemma can be summarized simply in terms of divergence as in the following theorem, which is commonly referred to as the divergence inequality. The assumption of finite alphabets will be dropped later.

Theorem 3.1. Divergence Inequality *Given any two probability measures P and M on a common finite-alphabet probability space, then*

$$D(P\|M) \geq 0 \tag{3.7}$$

with equality if and only if $P = M$.

In this form the result is known as the *divergence inequality*. The fact that the divergence of one probability measure with respect to another is nonnegative and zero only when the two measures are the same suggest the interpretation of divergence as a "distance" between the two probability measures, that is, a measure of how different the two measures are. It is not a true distance or metric in the usual sense since it is not a symmetric function of the two measures and it does not satisfy the triangle inequality. The interpretation is, however, quite useful for adding insight into results characterizing the behavior of divergence and it will later be seen to have implications for ordinary distance measures between probability measures.

The divergence plays a basic role in the family of information measures all of the information measures that we will encounter — entropy, relative entropy, mutual information, and the conditional forms of these information measures — can be expressed as a divergence.

There are three ways to view entropy as a special case of divergence. The first is to permit M to be a general measure instead of requiring it to

be a probability measure and have total mass 1. In this case entropy is minus the divergence if M is the counting measure, i.e., assigns measure 1 to every point in the discrete alphabet. If M is not a probability measure, then the divergence inequality (3.7) need not hold. Second, if the alphabet of f is A_f and has $\|A_f\|$ elements, then letting M be a uniform PMF assigning probability $1/\|A\|$ to all symbols in A yields

$$D(P\|M) = \ln \|A_f\| - H_P(f) \geq 0$$

and hence the entropy is the log of the alphabet size minus the divergence with respect to the uniform distribution. Third, we can also consider entropy a special case of divergence while still requiring that M be a probability measure by using product measures and a bit of a trick. Say we have two measures P and Q on a common probability space $(\Omega, \mathcal{B})$. Define two measures on the product space $(\Omega \times \Omega, \mathcal{B}(\Omega \times \Omega))$ as follows: Let $P \times Q$ denote the usual product measure, that is, the measure specified by its values on rectangles as $P \times Q(F \times G) = P(F)Q(G)$. Thus, for example, if P and Q are discrete distributions with PMF's p and q, then the PMF for $P \times Q$ is just $p(a)q(b)$. Let P' denote the "diagonal" measure defined by its values on rectangles as $P'(F \times G) = P(F \bigcap G)$. In the discrete case P' has PMF $p'(a, b) = p(a)$ if $a = b$ and 0 otherwise. Then

$$H_P(f) = D(P'\|P \times P).$$

Note that if we let X and Y be the coordinate random variables on our product space, then both P' and $P \times P$ give the same marginal probabilities to X and Y, that is, $P_X = P_Y = P$. P' is an extreme distribution on (X, Y) in the sense that with probability one $X = Y$; the two coordinates are deterministically dependent on one another. $P \times P$, however, is the opposite extreme in that it makes the two random variables X and Y independent of one another. Thus the entropy of a distribution P can be viewed as the relative entropy between these two extreme joint distributions having marginals P.

3.3 Basic Properties of Entropy

For the moment fix a probability measure m on a measurable space $(\Omega, \mathcal{B})$ and let X and Y be two finite-alphabet random variables defined on that space. Let A_X and A_Y denote the corresponding alphabets. Let P_{XY}, P_X, and P_Y denote the distributions of (X, Y), X, and Y, respectively.

First observe that since $P_X(a) \leq 1$, all a, $-\ln P_X(a)$ is positive and hence

$$H(X) = -\sum_{a \in A} P_X(a) \ln P_X(a) \geq 0. \tag{3.8}$$

From (3.7) with M uniform as in the second interpretation of entropy above, if X is a random variable with alphabet A_X, then

$$H(X) \le \ln \|A_X\|.$$

Since for any $a \in A_X$ and $b \in A_Y$ we have that $P_X(a) \ge P_{XY}(a,b)$, it follows that

$$\begin{aligned} H(X,Y) &= -\sum_{a,b} P_{XY}(a,b) \ln P_{XY}(a,b) \\ &\ge -\sum_{a,b} P_{XY}(a,b) \ln P_X(a) = H(X). \end{aligned}$$

Using Lemma 3.1 we have that since P_{XY} and $P_X P_Y$ are probability mass functions,

$$H(X,Y) - (H(X) + H(Y)) = \sum_{a,b} P_{XY}(a,b) \ln \frac{P_X(a)P_Y(b)}{P_{XY}(a,b)} \le 0.$$

This proves the following result.

Lemma 3.2. *Given two discrete alphabet random variables X and Y defined on a common probability space, we have*

$$0 \le H(X) \tag{3.9}$$

and

$$\max(H(X), H(Y)) \le H(X,Y) \le H(X) + H(Y) \tag{3.10}$$

where the right hand inequality holds with equality if and only if X and Y are independent. If the alphabet of X has $\|A_X\|$ symbols, then

$$H_X(X) \le \ln \|A_X\|. \tag{3.11}$$

There is another proof of the left hand inequality in (3.10) that uses an inequality for relative entropy that will be useful later when considering codes. The following lemma gives the inequality. First we introduce a definition. A partition $\mathcal{R}$ is said to *refine* a partion $\mathcal{Q}$ if every atom in $\mathcal{Q}$ is a union of atoms of $\mathcal{R}$, in which case we write $\mathcal{Q} < \mathcal{R}$.

Lemma 3.3. *Suppose that P and M are two measures defined on a common measurable space $(\Omega, \mathcal{B})$ and that we are given a finite partitions $\mathcal{Q} < \mathcal{R}$. Then*

$$H_{P\|M}(\mathcal{Q}) \le H_{P\|M}(\mathcal{R})$$

and

$$H_P(\mathcal{Q}) \le H_P(\mathcal{R}).$$

Comments: The lemma can also be stated in terms of random variables and mappings in an intuitive way: Suppose that U is a random variable with finite alphabet A and $f : A \to B$ is a mapping from A into another finite alphabet B. Then the composite random variable $f(U)$ defined by $f(U)(\omega) = f(U(\omega))$ is also a finite alphabet random variable. If U induces a partition $\mathcal{R}$ and $f(U)$ a partition $\mathcal{Q}$, then $\mathcal{Q} < \mathcal{R}$ (since knowing the value of U implies the value of $f(U)$). Thus the lemma immediately gives the following corollary.

Corollary 3.2. *If $M \gg P$ are two measures describing a random variable U with finite alphabet A and if $f : A \to B$, then*

$$H_{P\|M}(f(U)) \le H_{P\|M}(U)$$

and

$$H_P(f(U)) \le H_P(U).$$

Since $D(P_f\|M_f) = H_{P\|M}(f)$, we have also the following corollary which we state for future reference.

Corollary 3.3. *Suppose that P and M are two probability measures on a discrete space and that f is a random variable defined on that space, then*

$$D(P_f\|M_f) \le D(P\|M).$$

The lemma, discussion, and corollaries can all be interpreted as saying that taking a measurement on a finite-alphabet random variable lowers the entropy and the relative entropy of that random variable. By choosing U as (X, Y) and $f(X, Y) = X$ or Y, the lemma yields the promised inequality of the previous lemma.

Proof of Lemma: If $H_{P\|M}(\mathcal{R}) = +\infty$, the result is immediate. If $H_{P\|M}(\mathcal{Q}) = +\infty$, that is, if there exists at least one Q_j such that $M(Q_j) = 0$ but $P(Q_j) \neq 0$, then there exists an $R_i \subset Q_j$ such that $M(R_i) = 0$ and $P(R_i) > 0$ and hence $H_{P\|M}(\mathcal{R}) = +\infty$. Lastly assume that both $H_{P\|M}(\mathcal{R})$ and $H_{P\|M}(\mathcal{Q})$ are finite and consider the difference

$$\begin{aligned} H_{P\|M}(\mathcal{R}) - H_{P\|M}(\mathcal{Q}) &= \sum_i P(R_i) \ln \frac{P(R_i)}{M(R_i)} - \sum_j P(Q_j) \ln \frac{P(Q_j)}{M(Q_j)} \\ &= \sum_j \left[\sum_{i:R_i \subset Q_j} P(R_i) \ln \frac{P(R_i)}{M(R_i)} - P(Q_j) \ln \frac{P(Q_j)}{M(Q_j)} \right]. \end{aligned}$$

We shall show that each of the bracketed terms is nonnegative, which will prove the first inequality. Fix j. If $P(Q_j)$ is 0 we are done since then also $P(R_i)$ is 0 for all i in the inner sum since these R_i all belong to Q_j. If $P(Q_j)$ is not 0, we can divide by it to rewrite the bracketed term as

$$P(Q_j)\left(\sum_{i:R_i\subset Q_j}\frac{P(R_i)}{P(Q_j)}\ln\frac{P(R_i)/P(Q_j)}{M(R_i)/M(Q_j)}\right),$$

where we also used the fact that $M(Q_j)$ cannot be 0 since then $P(Q_j)$ would also have to be zero. Since $R_i\subset Q_j, P(R_i)/P(Q_j)=P(R_i\bigcap Q_j)/P(Q_j)$ $=P(R_i|Q_j)$ is an elementary conditional probability. Applying a similar argument to M and dividing by $P(Q_j)$, the above expression becomes

$$\sum_{i:R_i\subset Q_j}P(R_i|Q_j)\ln\frac{P(R_i|Q_j)}{M(R_i|Q_j)}$$

which is nonnegative from Lemma 3.1, which proves the first inequality. The second inequality follows similarly. Consider the difference

$$\begin{aligned}H_P(\mathcal{R})-H_P(\mathcal{Q}) &= \sum_j\left[\sum_{i:R_i\subset Q_j}P(R_i)\ln\frac{P(Q_j)}{P(R_i)}\right]\\ &= \sum_j P(Q_j)\left[-\sum_{i:R_i\subset Q_j}P(R_i|Q_j)\ln P(R_i|Q_j)\right]\end{aligned}$$

and the result follows since the bracketed term is nonnegative since it is an entropy for each value of j (Lemma 3.2). □

Concavity of Entropy

The next result provides useful inequalities for entropy considered as a function of the underlying distribution. In particular, it shows that entropy is a concave (or convex $\bigcap$) function of the underlying distribution. The concavity follows from Corollary 3.5, but for later use we do a little extra work to obtain an additional property. Define the binary entropy function (the entropy of a binary random variable with probability mass function $(\lambda, 1-\lambda)$) by

$$h_2(\lambda) = -\lambda\ln\lambda-(1-\lambda)\ln(1-\lambda).$$

Lemma 3.4. *Let m and p denote two distributions for a discrete alphabet random variable X and let $\lambda\in(0,1)$. Then for any $\lambda\in(0,1)$*

$$\begin{aligned}\lambda H_m(X)+(1-\lambda)H_p(X) &\le H_{\lambda m+(1-\lambda)p}(X)\\ &\le \lambda H_m(X)+(1-\lambda)H_p(X)+h_2(\lambda).\end{aligned}\quad(3.12)$$

Proof: Define the quantities

$$I = -\sum_x m(x) \ln(\lambda m(x) + (1-\lambda)p(x))$$

and

$$J = H_{\lambda m+(1-\lambda)p}(X) = -\lambda \sum_x m(x) \ln(\lambda m(x) + (1-\lambda)p(x)) - (1-\lambda) \sum_x p(x) \ln(\lambda m(x) + (1-\lambda)p(x)).$$

First observe that

$$\lambda m(x) + (1-\lambda)p(x) \geq \lambda m(x)$$

and therefore applying this bound to both m and p

$$\begin{aligned} I &\leq -\ln\lambda - \sum_x m(x) \ln m(x) = -\ln\lambda + H_m(X) \\ J &\leq -\lambda \sum_x m(x) \ln m(x) - (1-\lambda) \sum_x p(x) \ln p(x) + h_2(\lambda) \\ &= \lambda H_m(X) + (1-\lambda) H_p(X) + h_2(\lambda). \end{aligned} \tag{3.13}$$

To obtain the lower bounds of the lemma observe that

$$\begin{aligned} I &= -\sum_x m(x) \ln m(x) (\lambda + (1-\lambda)\frac{p(x)}{m(x)}) \\ &= -\sum_x m(x) \ln m(x) - \sum_x m(x) \ln(\lambda + (1-\lambda)\frac{p(x)}{m(x)}). \end{aligned}$$

Using (3.6) the rightmost term is bound below by

$$-\sum_x m(x)((\lambda + (1-\lambda)\frac{p(x)}{m(x)} - 1) = -\lambda - 1 + \lambda \sum_{a \in A} p(X=a) + 1 = 0.$$

Thus for all n

$$I \geq -\sum_x m(x) \ln m(x) = H_m(X). \tag{3.14}$$

and hence also

$$\begin{aligned} J &\geq -\lambda \sum_x m(x) \ln m(x) - (1-\lambda) \sum_x p(x) \ln p(x) \\ &= \lambda H_m(X) + (1-\lambda) H_p(X). \end{aligned}$$

□

Convexity of Divergence

Relative entropy possesses a useful convexity property with respect to the two probability measures, as described in the following lemma.

Lemma 3.5. *$D(P\|M)$ is convex in (P, M) for probability measures P, M on a common finite-alphabet probability space, that is, if (P_i, M_i), $i = 1, 2$ are pairs of probability measures, all of which are on a common finite-alphabet probability space, and $(P, M) = \lambda(P_1, M_1) + (1-\lambda)(P_2, M_2)$, then*

$$D(P\|M) \le \lambda D(P_1\|M_1) + (1-\lambda)D(P_2\|M_2).$$

Proof. The result follows from the convexity of $a\ln(a/b)$ in (a, b) from Corollary 3.1. □

Entropy and Binomial Sums

The next result presents an interesting connection between combinatorics and binomial sums with a particular entropy. We require the familiar definition of the binomial coefficient:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

Lemma 3.6. *Given $\delta \in (0, \frac{1}{2}]$ and a positive integer M, we have*

$$\sum_{i \le \delta M} \binom{M}{i} \le e^{Mh_2(\delta)}. \tag{3.15}$$

If $0 < \delta \le p \le 1$, then

$$\sum_{i \le \delta M} \binom{M}{i} p^i (1-p)^{M-i} \le e^{-Mh_2(\delta\|p)}, \tag{3.16}$$

where

$$h_2(\delta\|p) = \delta \ln \frac{\delta}{p} + (1-\delta)\ln \frac{1-\delta}{1-p}.$$

Proof: We have after some simple algebra that

$$e^{-h_2(\delta)M} = \delta^{\delta M}(1-\delta)^{(1-\delta)M}.$$

If $\delta < 1/2$, then $\delta^k(1-\delta)^{M-k}$ increases as k decreases (since we are having more large terms and fewer small terms in the product) and hence

if $i \le M\delta$,

$$\delta^{\delta M}(1-\delta)^{(1-\delta)M} \le \delta^i(1-\delta)^{M-i}.$$

Thus we have the inequalities

$$1 = \sum_{i=0}^{M} \binom{M}{i} \delta^i(1-\delta)^{M-i} \ge \sum_{i \le \delta M} \binom{M}{i} \delta^i(1-\delta)^{M-i}$$
$$\ge e^{-h_2(\delta)M} \sum_{i \le \delta M} \binom{M}{i}$$

which completes the proof of (3.15). In a similar fashion we have that

$$e^{Mh_2(\delta\|p)} = (\frac{\delta}{p})^{\delta M}(\frac{1-\delta}{1-p})^{(1-\delta)M}.$$

Since $\delta \le p$, we have as in the first argument that for $i \le M\delta$

$$(\frac{\delta}{p})^{\delta M}(\frac{1-\delta}{1-p})^{(1-\delta)M} \le (\frac{\delta}{p})^{i}(\frac{1-\delta}{1-p})^{M-i}$$

and therefore after some algebra we have that if $i \le M\delta$ then

$$p^i(1-p)^{M-i} \le \delta^i(1-\delta)^{M-i}e^{-Mh_2(\delta\|p)}$$

and hence

$$\sum_{i \le \delta M} \binom{M}{i} p^i(1-p)^{M-i} \le e^{-Mh_2(\delta\|p)} \sum_{i \le \delta M} \binom{M}{i} \delta^i(1-\delta)^{M-i}$$
$$\le e^{-nh_2(\delta\|p)} \sum_{i=0}^{M} \binom{M}{i} \delta^i(1-\delta)^{M-i} = e^{-Mh_2(\delta\|p)},$$

which proves (3.16). □

The following is a technical but useful property of sample entropies. The proof follows Billingsley [16].

Lemma 3.7. *Given a finite-alphabet process $\{X_n\}$ (not necessarily stationary) with distribution m, let $X_k^n = (X_k, X_{k+1}, \ldots, X_{k+n-1})$ denote the random vectors giving a block of samples of dimension n starting at time k. Then the random variables $n^{-1} \ln m(X_k^n)$ are m-uniformly integrable (uniform in k and n).*

Proof: For each nonnegative integer r define the sets

$$E_r(k,n) = \{x : -\frac{1}{n} \ln m(x_k^n) \in [r, r+1)\}$$

and hence if $x \in E_r(k,n)$ then

$$r \le -\frac{1}{n} \ln m(x_k^n) < r + 1$$

or

$$e^{-nr} \ge m(x_k^n) > e^{-n(r+1)}.$$

Thus for any r

$$\begin{aligned}\int_{E_r(k,n)} \left(-\frac{1}{n} \ln m(X_k^n)\right) dm &< (r+1) m(E_r(k,n)) \\ &= (r+1) \sum_{x_k^n \in E_r(k,n)} m(x_k^n) \le (r+1) \sum_{x_k^n} e^{-nr} \\ &= (r+1) e^{-nr} \|A\|^n \le (r+1) e^{-nr},\end{aligned}$$

where the final step follows since there are at most $\|A\|^n$ possible n-tuples corresponding to thin cylinders in $E_r(k,n)$ and by construction each has probability less than e^{-nr}.

To prove uniform integrability we must show uniform convergence to 0 as $r \to \infty$ of the integral

$$\begin{aligned}\gamma_r(k,n) &= \int_{x:-\frac{1}{n}\ln m(x_k^n) \ge r} (-\frac{1}{n} \ln m(X_k^n))\, dm = \sum_{i=0}^{\infty} \int_{E_{r+i}(k,n)} (-\frac{1}{n} \ln m(X_k^n))\, dm \\ &\le \sum_{i=0}^{\infty} (r+i+1) e^{-n(r+i)} \|A\|^n \le \sum_{i=0}^{\infty} (r+i+1) e^{-n(r+i-\ln\|A\|)}.\end{aligned}$$

Taking r large enough so that $r > \ln \|A\|$, then the exponential term is bound above by the special case $n = 1$ and we have the bound

$$\gamma_r(k,n) \le \sum_{i=0}^{\infty} (r+i+1) e^{-(r+i-\ln\|A\|)}$$

a bound which is finite and independent of k and n. The sum can easily be shown to go to zero as $r \to \infty$ using standard summation formulas. (The exponential terms shrink faster than the linear terms grow.) □

Variational Description of Divergence

Divergence has a variational characterization that is a fundamental property for its applications to large deviations theory [182] [31]. Although this theory will not be treated here, the basic result of this section provides an alternative description of divergence and hence of relative entropy that has intrinsic interest. The basic result is originally due to Donsker and Varadhan [35].

Suppose now that P and M are two probability measures on a common discrete probability space, say $(\Omega, \mathcal{B})$. Given any real-valued random variable Φ defined on the probability space, we will be interested in the quantity

$$E_M e^{\Phi}. \tag{3.17}$$

which is called the *cumulant generating function* of Φ with respect to M and is related to the characteristic function of the random variable Φ as well as to the moment generating function and the operational transform of the random variable. The following theorem provides a variational description of divergence in terms of the cumulant generating function.

Theorem 3.2.

$$D(P\|M) = \sup_{\Phi} \left(E_P \Phi - \ln(E_M(e^{\Phi})) \right). \tag{3.18}$$

Proof: First consider the random variable Φ defined by

$$\Phi(\omega) = \ln(P(\omega)/M(\omega))$$

and observe that

$$E_P \Phi - \ln(E_M(e^{\Phi})) = \sum_{\omega} P(\omega) \ln \frac{P(\omega)}{M(\omega)} - \ln(\sum_{\omega} M(\omega) \frac{P(\omega)}{M(\omega)})$$

$$= D(P\|M) - \ln 1 = D(P\|M).$$

This proves that the supremum over all Φ is no smaller than the divergence.

To prove the other half observe that for any bounded random variable Φ,

$$E_P \Phi - \ln E_M(e^{\Phi}) = E_P \left(\ln \frac{e^{\Phi}}{E_M(e^{\Phi})} \right) = \sum_{\omega} P(\omega) \left(\ln \frac{M^{\Phi}(\omega)}{M(\omega)} \right),$$

where the probability measure M^{Φ} is defined by

$$M^{\Phi}(\omega) = \frac{M(\omega) e^{\Phi(\omega)}}{\sum_x M(x) e^{\Phi(x)}}.$$

We now have for any Φ that

$$D(P\|Q) - \left(E_P \Phi - \ln(E_M(e^{\Phi})) \right) =$$

$$\sum_{\omega} P(\omega) \left(\ln \frac{P(\omega)}{M(\omega)} \right) - \sum_{\omega} P(\omega) \left(\ln \frac{M^{\Phi}(\omega)}{M(\omega)} \right) =$$

$$\sum_{\omega} P(\omega) \left(\ln \frac{P(\omega)}{M^{\Phi}(\omega)} \right) \geq 0$$

using the divergence inequality. Since this is true for any Φ, it is also true for the supremum over Φ and the theorem is proved. □

3.4 Entropy Rate

For simplicity we focus on the entropy rate of a directly given finite-alphabet random process $\{X_n\}$. We also will emphasize stationary measures, but we will try to clarify those results that require stationarity and those that are more general.

As a reminder, we recall the underlying structure. Let A be a finite set. Let $\Omega = A^{\mathbf{Z}_+}$ and let $\mathcal{B}$ be the sigma-field of subsets of Ω generated by the rectangles. Since A is finite, $(A, \mathcal{B}_A)$ is standard, where $\mathcal{B}_A$ is the power set of A. Thus $(\Omega, \mathcal{B})$ is also standard by Lemma 2.4.1 of [55] or Lemma 3.7 of [58]. In fact, from the proof that cartesian products of standard spaces are standard, we can take as a basis for $\mathcal{B}$ the fields $\mathcal{F}_n$ generated by the finite dimensional rectangles having the form $\{x : X^n(x) = x^n = a^n\}$ for all $a^n \in A^n$ and all positive integers n. (Members of this class of rectangles are called *thin cylinders*.) The union of all such fields, say $\mathcal{F}$, is then a generating field.

Again let $\{X_n; n = 0, 1, \ldots\}$ denote a finite-alphabet random process and apply Lemma 3.2 to vectors and obtain

$$H(X_0, X_1, \ldots, X_{n-1}) \le$$
$$H(X_0, X_1, \ldots, X_{m-1}) + H(X_m, X_{m+1}, \ldots, X_{n-1});\ 0 < m < n. \quad (3.19)$$

Define as usual the random vectors $X_k^n = (X_k, X_{k+1}, \ldots, X_{k+n-1})$, that is, X_k^n is a vector of dimension n consisting of the samples of X from k to $k + n - 1$. If the underlying measure is stationary, then the distributions of the random vectors X_k^n do not depend on k. Hence if we define the sequence $h(n) = H(X^n) = H(X_0, \ldots, X_{n-1})$, then the above equation becomes

$$h(k + n) \le h(k) + h(n);\ \text{all } k, n > 0.$$

Thus $h(n)$ is a subadditive sequence as treated in Section 7.5 of [55] or Section 8.5 of [58]. A basic property of subadditive sequences is that the limit $h(n)/n$ as $n \to \infty$ exists and equals the infimum of $h(n)/n$ over n. (See, e.g., Lemma 7.5.1 of [55] or Lemma 8.5.3 of [58].) This immediately yields the following result.

Lemma 3.8. *If the distribution m of a finite-alphabet random process $\{X_n\}$ is stationary, then*

$$\overline{H}_m(X) = \lim_{n\to\infty} \frac{1}{n} H_m(X^n) = \inf_{n\ge 1} \frac{1}{n} H_m(X^n).$$

Thus the limit exists and equals the infimum.

The next two properties of entropy rate are primarily of interest because they imply a third property, the ergodic decomposition of entropy rate, which will be described in Theorem 3.3. They are also of some independent interest. The first result is a continuity result for entropy rate when considered as a function or functional on the underlying process distribution. The second property demonstrates that entropy rate is actually an affine functional (both convex $\bigcup$ and convex $\bigcap$) of the underlying distribution, even though finite order entropy was only convex $\bigcap$ and not affine.

We apply the distributional distance described in Section 1.7 to the standard sequence measurable space $(\Omega, \mathcal{B}) = (A^{Z_+}, \mathcal{B}_A^{Z_+})$ with a σ-field generated by the countable field $\mathcal{F} = \{F_n; n = 1, 2, \ldots\}$ generated by all thin rectangles.

Corollary 3.4. *The entropy rate $\overline{H}_m(X)$ of a discrete alphabet random process considered as a functional of stationary measures is upper semicontinuous; that is, if probability measures m and m_n, $n = 1, 2, \ldots$ have the property that $d(m, m_n) \to 0$ as $n \to \infty$, then*

$$\overline{H}_m(X) \geq \limsup_{n\to\infty} \overline{H}_{m_n}(X).$$

Proof: For each fixed n

$$H_m(X^n) = -\sum_{a^n \in A^n} m(X^n = a^n) \ln m(X^n = a^n)$$

is a continuous function of m since for the distance to go to zero, the probabilities of all thin rectangles must go to zero and the entropy is the sum of continuous real-valued functions of the probabilities of thin rectangles. Thus we have from Lemma 3.8 that if $d(m_k, m) \to 0$, then

$$\begin{aligned}\overline{H}_m(X) &= \inf_n \frac{1}{n} H_m(X^n) = \inf_n \frac{1}{n} \lim_{k\to\infty} H_{m_k}(X^n) \\ &\geq \limsup_{k\to\infty} \left(\inf_n \frac{1}{n} H_{m_k}(X^n) \right) = \limsup_{k\to\infty} \overline{H}_{m_k}(X).\end{aligned}$$

□

The next lemma uses Lemma 3.4 to show that entropy rates are affine functions of the underlying probability measures.

Lemma 3.9. *Let m and p denote two distributions for a discrete alphabet random process $\{X_n\}$. Then for any $\lambda \in (0, 1)$,*

$$\begin{aligned}\lambda H_m(X^n) + (1-\lambda) H_p(X^n) &\leq H_{\lambda m + (1-\lambda)p}(X^n) \\ &\leq \lambda H_m(X^n) + (1-\lambda) H_p(X^n) + h_2(\lambda), \quad (3.20)\end{aligned}$$

and

$$\limsup_{n\to\infty}(-\int dm(x)\frac{1}{n}\ln(\lambda m(X^n(x)) + (1-\lambda)p(X^n(x))))$$
$$= \limsup_{n\to\infty} -\int dm(x)\frac{1}{n}\ln m(X^n(x)) = \overline{H}_m(X). \quad (3.21)$$

If m and p are stationary then

$$\overline{H}_{\lambda m+(1-\lambda)p}(X) = \lambda\overline{H}_m(X) + (1-\lambda)\overline{H}_p(X) \quad (3.22)$$

and hence the entropy rate of a stationary discrete alphabet random process is an affine function of the process distribution.

Comment: Eq. (3.20) is simply Lemma 3.4 applied to the random vectors X^n stated in terms of the process distributions. Eq. (3.21) states that if we look at the limit of the normalized log of a mixture of a pair of measures when one of the measures governs the process, then the limit of the expectation does not depend on the other measure at all and is simply the entropy rate of the driving source. Thus in a sense the sequences produced by a measure are able to select the true measure from a mixture.

Proof: Eq. (3.20) is just Lemma 3.4. Dividing by n and taking the limit as $n \to \infty$ proves that entropy rate is affine. Similarly, take the limit supremum in expressions (3.13) and (3.14) and the lemma is proved. □

We are now prepared to prove one of the fundamental properties of entropy rate, the fact that it has an ergodic decomposition formula similar to property (c) of Theorem 1.5 when it is considered as a functional on the underlying distribution. In other words, the entropy rate of a stationary source is given by an integral of the entropy rates of the stationary ergodic components. This is a far more complicated result than property (c) of the ordinary ergodic decomposition because the entropy rate depends on the distribution; it is not a simple function of the underlying sequence. The result is due to Jacobs [80].

Theorem 3.3. *The Ergodic Decomposition of Entropy Rate*

Let $(A^{Z_+}, \mathcal{B}(A)^{Z_+}, m, T)$ be a stationary dynamical system corresponding to a stationary finite alphabet source $\{X_n\}$. Let $\{p_x\}$ denote the ergodic decomposition of m. If $\overline{H}_{p_x}(X)$ is m-integrable, then

$$\overline{H}_m(X) = \int dm(x)\overline{H}_{p_x}(X).$$

Proof: The theorem follows immediately from Corollary 3.4 and Lemma 3.9 and the ergodic decomposition of semi-continuous affine funtionals as in Theorem 8.9.1 of [55] or Theorem 8.5 of [58]. □

3.5 Relative Entropy Rate

The properties of relative entropy rate are more difficult to demonstrate. In particular, the obvious analog to (3.19) does not hold for relative entropy rate without the requirement that the reference measure be memoryless, and hence one cannot immediately infer that the relative entropy rate is given by a limit for stationary sources. The following lemma provides a condition under which the relative entropy rate is given by a limit. The condition, that the dominating measure be a kth order (or k-step) Markov source will occur repeatedly when dealing with relative entropy rates. A discrete alphabet source is kth order Markov or k-step Markov (or simply Markov if k is clear from context) if for any n and any $N \geq k$

$$P(X_n = x_n | X_{n-1} = x_{n-1}, \ldots, X_{n-N} = x_{n-N}) \\ = P(X_n = x_n | X_{n-1} = x_{n-1}, \ldots, X_{n-k} = x_{n-k});$$

that is, conditional probabilities given the infinite past depend only on the most recent k symbols. A 0-step Markov source is a memoryless source. A Markov source is said to have *stationary transitions* if the above conditional probabilities do not depend on n, that is, if for any n

$$P(X_n = x_n | X_{n-1} = x_{n-1}, \ldots, X_{n-N} = x_{n-N}) = \\ P(X_k = x_n | X_{k-1} = x_{n-1}, \ldots, X_0 = x_{n-k}).$$

Lemma 3.10. *If p is a stationary process and m is a k-step Markov process with stationary transitions, then*

$$\overline{H}_{p\|m}(X) = \lim_{n\to\infty} \frac{1}{n} H_{p\|m}(X^n) = -\overline{H}_p(X) - E_p[\ln m(X_k|X^k)],$$

where $E_p[\ln m(X_k|X^k)]$ is an abbreviation for

$$E_p[\ln m(X_k|X^k)] = \sum_{x^{k+1}} p_{X^{k+1}}(x^{k+1}) \ln m_{X_k|X^k}(x_k|x^k).$$

Proof: If for any n it is not true that $m_{X^n} \gg p_{X^n}$, then $H_{p\|m}(X^n) = \infty$ for that and all larger n and both sides of the formula are infinite, hence we assume that all of the finite dimensional distributions satisfy the absolute continuity relation. Since m is Markov,

$$m_{X^n}(x^n) = \prod_{l=k}^{n-1} m_{X_l|X^l}(x_l|x^l) m_{X^k}(x^k).$$

Thus

$$\begin{aligned}\frac{1}{n}H_{p\|m}(X^n) &= -\frac{1}{n}H_p(X^n) - \frac{1}{n}\sum_{x^n} p_{X^n}(x^n)\ln m_{X^n}(x^n)\\ &= -\frac{1}{n}H_p(X^n) - \frac{1}{n}\sum_{x^k} p_{X^k}(x^k)\ln m_{X^k}(x^k)\\ &\quad - \frac{n-k}{n}\sum_{x^{k+1}} p_{X^{k+1}}(x^{k+1})\ln m_{X_k|X^k}(x_k|x^k).\end{aligned}$$

Taking limits then yields

$$\overline{H}_{p\|m}(X) = -\overline{H}_p - \sum_{x^{k+1}} p_{X^{k+1}}(x^{k+1})\ln m_{X_k|X^k}(x_k|x^k),$$

where the sum is well defined because if $m_{X_k|X^k}(x_k|x^k) = 0$, then so must $p_{X^{k+1}}(x^{k+1}) = 0$ from absolute continuity. □

Combining the previous lemma with the ergodic decomposition of entropy rate yields the following corollary.

Corollary 3.5. *The Ergodic Decomposition of Relative Entropy Rate*
Let $(A^{Z_+}, \mathcal{B}(A)^{Z_+}, p, T)$ be a stationary dynamical system corresponding to a stationary finite alphabet source $\{X_n\}$. Let m be a kth order Markov process for which $m_{X^n} \gg p_{X^n}$ for all n. Let $\{p_x\}$ denote the ergodic decomposition of p. If $\overline{H}_{p_x\|m}(X)$ is p-integrable, then

$$\overline{H}_{p\|m}(X) = \int dp(x)\overline{H}_{p_x\|m}(X).$$

3.6 Conditional Entropy and Mutual Information

We now turn to other notions of information. While we could do without these if we confined interest to finite-alphabet processes, they will be essential for later generalizations and provide additional intuition and results even in the finite alphabet case. We begin by adding a second finite-alphabet measurement to the setup of the previous sections. To conform more to information theory tradition, we consider the measurements as finite-alphabet random variables X and Y rather than f and g. This has the advantage of releasing f and g for use as functions defined on the random variables: $f(X)$ and $g(Y)$. It should be kept in mind, however, that X and Y could themselves be distinct measurements on a common random variable. This interpretation will often be useful when comparing codes.

Let $(\Omega, \mathcal{B}, P, T)$ be a dynamical system. Let X and Y be finite-alphabet measurements defined on Ω with alphabets A_X and A_Y. Define the *conditional entropy* of X given Y by

$$H(X|Y) \equiv H(X,Y) - H(Y).$$

The name conditional entropy comes from the fact that

$$\begin{aligned} H(X|Y) &= -\sum_{x,y} P(X=a, Y=b) \ln P(X=a|Y=b) \\ &= -\sum_{x,y} p_{X,Y}(x,y) \ln p_{X|Y}(x|y) \\ &= -\sum_{y} p_Y(y) \left[\sum_{x} p_{X|Y}(x \mid y) \ln p_{X|Y}(x|y) \right], \end{aligned}$$

where $p_{X,Y}(x,y)$ is the joint PMF for (X,Y) and

$$p_{X|Y}(x|y) = p_{X,Y}(x,y)/p_Y(y)$$

is the conditional PMF. Defining

$$H(X|Y=y) = -\sum_{x} p_{X|Y}(x|y) \ln p_{X|Y}(x|y)$$

we can also write

$$H(X|Y) = \sum_{y} p_Y(y) H(X|Y=y).$$

Thus conditional entropy is an average of entropies with respect to conditional PMF's. We have immediately from Lemma 3.2 and the definition of conditional entropy that

$$0 \le H(X|Y) \le H(X). \tag{3.23}$$

The inequalities could also be written in terms of the partitions induced by X and Y. Recall that according to Lemma 3.2 the right hand inequality will be an equality if and only if X and Y are independent.

Define the *average mutual information* between X and Y by

$$\begin{aligned} I(X;Y) &= H(X) + H(Y) - H(X,Y) \\ &= H(X) - H(X|Y) = H(Y) - H(Y|X). \end{aligned}$$

In terms of distributions and PMF's we have that

$$\begin{aligned} I(X;Y) &= \sum_{x,y} P(X=x, Y=y) \ln \frac{P(X=x, Y=y)}{P(X=x)P(Y=y)} \\ &= \sum_{x,y} p_{X,Y}(x,y) \ln \frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)} = \sum_{x,y} p_{X,Y}(x,y) \ln \frac{p_{X|Y}(x|y)}{p_X(x)} \\ &= \sum_{x,y} p_{X,Y}(x,y) \ln \frac{p_{Y|X}(y|x)}{p_Y(y)}. \end{aligned}$$

Note also that mutual information can be expressed as a divergence by

$$I(X;Y) = D(P_{X,Y} \| P_X \times P_Y),$$

where $P_X \times P_Y$ is the product measure on X, Y, that is, a probability measure which gives X and Y the same marginal distributions as P_{XY}, but under which X and Y are independent. Entropy is a special case of mutual information since

$$H(X) = I(X;X).$$

We can collect several of the properties of entropy and relative entropy and produce corresponding properties of mutual information. We state these in the form using measurements, but they can equally well be expressed in terms of partitions.

Lemma 3.11. *Suppose that X and Y are two finite-alphabet random variables defined on a common probability space. Then*

$$0 \le I(X;Y) \le \min(H(X), H(Y)).$$

Suppose that $f : A_X \to A$ and $g : A_Y \to B$ are two measurements. Then

$$I(f(X); g(Y)) \le I(X;Y).$$

Proof: The first result follows immediately from the properties of entropy. The second follows from Lemma 3.3 applied to the measurement (f, g) since mutual information is a special case of relative entropy. □

The next lemma collects some additional, similar properties.

Lemma 3.12. *Given the assumptions of the previous lemma,*

$$\begin{aligned} H(f(X)|X) &= 0, \\ H(X, f(X)) &= H(X), \\ H(X) &= H(f(X)) + H(X|f(X)), \\ I(X; f(X)) &= H(f(X)), \\ H(X|g(Y)) &\geq H(X|Y), \\ I(f(X); g(Y)) &\leq I(X; Y), \\ H(X|Y) &= H(X, f(X, Y))|Y), \end{aligned}$$

and, if Z is a third finite-alphabet random variable defined on the same probability space,

$$H(X|Y) \geq H(X|Y, Z).$$

Comments: The first relation has the interpretation that given a random variable, there is no additional information in a measurement made on the random variable. The second and third relationships follow from the first and the definitions. The third relation is a form of chain rule and it implies that given a measurement on a random variable, the entropy of the random variable is given by that of the measurement plus the conditional entropy of the random variable given the measurement. This provides an alternative proof of the second result of Lemma 3.3. The fifth relation says that conditioning on a measurement of a random variable is less informative than conditioning on the random variable itself. The sixth relation states that coding reduces mutual information as well as entropy. The seventh relation is a conditional extension of the second. The eighth relation says that conditional entropy is nonincreasing when conditioning on more information.

Proof: Since $g(X)$ is a deterministic function of X, the conditional PMF is trivial (a Kronecker delta) and hence $H(g(X)|X = x)$ is 0 for all x, hence the first relation holds. The second and third relations follow from the first and the definition of conditional entropy. The fourth relation follows from the first since $I(X; Y) = H(Y) - H(Y|X)$. The fifth relation follows from the previous lemma since

$$H(X) - H(X|g(Y)) = I(X; g(Y)) \leq I(X; Y) = H(X) - H(X|Y).$$

The sixth relation follows from Corollary 3.3 and the fact that

$$I(X; Y) = D(P_{X,Y} \| P_X \times P_Y).$$

The seventh relation follows since

$$\begin{aligned} H(X, f(X, Y))|Y) &= H(X, f(X, Y)), Y) - H(Y) \\ &= H(X, Y) - H(Y) = H(X|Y). \end{aligned}$$

The final relation follows from the second by replacing Y by Y, Z and setting $g(Y, Z) = Y$. □

In a similar fashion we can consider conditional relative entropies. Suppose now that M and P are two probability measures on a common space, that X and Y are two random variables defined on that space, and that $M_{XY} \gg P_{XY}$ (and hence also $M_X \gg P_Y$). Analagous to the definition of the conditional entropy we can define

$$H_{P\|M}(X|Y) \equiv H_{P\|M}(X, Y) - H_{P\|M}(Y).$$

Some algebra shows that this is equivalent to

$$\begin{aligned} H_{P\|M}(X|Y) &= \sum_{x,y} p_{X,Y}(x, y) \ln \frac{p_{X|Y}(x|y)}{m_{X|Y}(x|y)} \\ &= \sum_{y} p_Y(y) \left(\sum_{x} p_{X|Y}(x|y) \ln \frac{p_{X|Y}(x|y)}{m_{X|Y}(x|y)} \right). \end{aligned} \tag{3.24}$$

This can be written as

$$H_{P\|M}(X|Y) = \sum_{y} p_Y(y) D(p_{X|Y}(\cdot|y) \| m_{X|Y}(\cdot|y)),$$

an average of divergences of conditional PMF's, each of which is well defined because of the original absolute continuity of the joint measure. Manipulations similar to those for entropy can now be used to prove the following properties of conditional relative entropies.

Lemma 3.13. *Given two probability measures M and P on a common space, and two random variables X and Y defined on that space with the property that $M_{XY} \gg P_{XY}$, then the following properties hold:*

$$\begin{aligned} H_{P\|M}(f(X)|X) &= 0, \\ H_{P\|M}(X, f(X)) &= H_{P\|M}(X), \\ H_{P\|M}(X) &= H_{P\|M}(f(X)) + H_{P\|M}(X|f(X)), \end{aligned} \tag{3.25}$$

If $M_{XY} = M_X \times M_Y$ (that is, if the PMFs satisfy $m_{X,Y}(x, y) = m_X(x) m_Y(y)$), then

$$H_{P\|M}(X, Y) \geq H_{P\|M}(X) + H_{P\|M}(Y)$$

and

$$H_{P\|M}(X|Y) \geq H_{P\|M}(X).$$

Eq. (3.25) is a chain rule for relative entropy which provides as a corollary an immediate proof of Lemma 3.3. The final two inequalities resemble inequalities for entropy (with a sign reversal), but they do not hold for all reference measures.

The above lemmas along with Lemma 3.3 show that all of the information measures thus far considered are reduced by taking measurements or by coding. This property is the key to generalizing these quantities to nondiscrete alphabets.

We saw in Lemma 3.4 that entropy was a convex $\bigcap$ function of the underlying distribution. The following lemma provides similar properties of mutual information considered as a function of either a marginal or a conditional distribution.

Lemma 3.14. *Let μ denote a PMF on a discrete space A_X, $\mu(x) = \Pr(X = x)$, and let ν be a conditional PMF, $\nu(y|x) = \Pr(Y = y|X = x)$. Let $\mu\nu$ denote the resulting joint PMF $\mu\nu(x, y) = \mu(x)\nu(y|x)$. Let $I_{\mu\nu} = I_{\mu\nu}(X;Y)$ be the average mutual information. Then $I_{\mu\nu}$ is a convex $\bigcup$ function of ν; that is, given two conditional PMF's ν_1 and ν_2, a $\lambda \in [0,1]$, and $\nu = \lambda\nu_1 + (1-\lambda)\nu_2$, then*

$$I_{\mu\nu} \le \lambda I_{\mu\nu_1} + (1-\lambda) I_{\mu\nu_2},$$

and $I_{\mu\nu}$ is a convex $\bigcap$ function of μ, that is, given two PMF's μ_1 and μ_2, $\lambda \in [0,1]$, and $\mu = \lambda\mu_1 + (1-\lambda)\mu_2$,

$$I_{\mu\nu} \ge \lambda I_{\mu_1\nu} + (1-\lambda) I_{\mu_2\nu}.$$

Proof. First consider the case of a fixed source μ and consider conditional PMFs ν_1, ν_2, and for $0 \le \lambda \le 1$ define $\nu = \lambda\nu_1 + (1-\lambda)\nu_2$. Denote the corresponding input/output pair processes by $p_i = \mu\nu_i$, $i = 1, 2$, and $p = \lambda p_1 + (1-\lambda)p_2$ Let η (respectively, η_1, η_2, η) denote the PMF for Y resulting from ν (respectively ν_1, ν_2, ν), that is, $\eta(y) = \Pr(Y = y) = \sum_x \mu(x)\nu(y|x)$. Note that p_1, p_2, and p all have a common input marginal PMF μ. We have that

$$\mu \times \eta = \lambda\mu \times \eta_1 + (1-\lambda)\mu \times \eta_2$$

so that from Lemma 3.5

$$\begin{aligned} I_{\mu\nu} &= D(\mu\nu||\mu \times \eta) = D(\lambda p_1 + (1-\lambda)p_2 \| \lambda\mu \times \eta_1 + (1-\lambda)\mu \times \eta_2) \\ &\le \lambda D(p_1 \| \mu \times \eta_1) + (1-\lambda) D(p_2 \| \mu \times \eta_2) \\ &= \lambda I_{\mu\nu_1} + (1-\lambda) I_{\mu\nu_2}, \end{aligned}$$

proving the convexity of mutual information with respect to the channel. The author is indebted to T. Linder for suggesting this proof, which is much simpler than the one in the first edition.

Similarly, let $\mu = \lambda\mu_1 + (1-\lambda)\mu_2$ and let η_1, η_2, and η denote the corresponding induced output PMF's. Then

$$
\begin{aligned}
I_{\mu\nu} &= \lambda \sum_{x,y} \mu_1(x)\nu(y|x) \log \left(\frac{\nu(y|x)}{\eta(y)} \frac{\eta_1(y)}{\nu(y|x)} \frac{\nu(y|x)}{\eta_1(y)} \right) \\
&\quad + (1-\lambda) \sum_{x,y} \mu_2(x)\nu(y|x) \log \left(\frac{\nu(y|x)}{\eta(y)} \frac{\eta_2(y)}{\nu(y|x)} \frac{\nu(y|x)}{\eta_2(y)} \right) \\
&= \lambda I_{\mu_1\nu} + (1-\lambda) I_{\mu_2\nu} - \lambda \sum_{x,y} \mu_1(x)\nu(y|x) \log \frac{\eta(y)}{\eta_1(y)} \\
&\quad - (1-\lambda) \sum_{x,y} \mu_2(x)\nu(y|x) \log \frac{\eta(y)}{\eta_2(y)} \\
&= \lambda I_{\mu_1\nu} + (1-\lambda) I_{\mu_2\nu} + \lambda D(\eta_1 \| \eta) + (1-\lambda) D(\eta_2 \| \eta) \\
&\geq \lambda I_{\mu_1\nu} + (1-\lambda) I_{\mu_2\nu} \qquad (3.26)
\end{aligned}
$$

from the divergence inequality. □

We consider one other notion of information: Given three finite-alphabet random variables X, Y, Z, define the *conditional mutual information* between X and Y given Z by

$$I(X;Y|Z) = D(P_{XYZ} \| P_{X\times Y|Z}) \qquad (3.27)$$

where $P_{X\times Y|Z}$ is the distribution defined by its values on rectangles as

$$P_{X\times Y|Z}(F \times G \times D) = \sum_{z\in D} P(X \in F|Z=z)P(Y \in G|Z=z)P(Z=z). \qquad (3.28)$$

$P_{X\times Y|Z}$ has the same conditional distributions for X given Z and for Y given Z as does P_{XYZ}, but now X and Y are conditionally independent given Z. Alternatively, the conditional distribution for X, Y given Z under the distribution $P_{X\times Y|Z}$ is the product distribution $P_X|Z \times P_Y|Z$. Thus

$$
\begin{aligned}
I(X;Y|Z) &= \sum_{x,y,z} p_{XYZ}(x,y,z) \ln \frac{p_{XYZ}(x,y,z)}{p_{X|Z}(x|z)p_{Y|Z}(y|z)p_Z(z)} \\
&= \sum_{x,y,z} p_{XYZ}(x,y,z) \ln \frac{p_{XY|Z}(x,y|z)}{p_{X|Z}(x|z)p_{Y|Z}(y|z)}. \qquad (3.29)
\end{aligned}
$$

Since

$$\frac{p_{XYZ}}{p_{X|Z}p_{Y|Z}p_Z} = \frac{p_{XYZ}}{p_X p_{YZ}} \times \frac{p_X}{p_{X|Z}} = \frac{p_{XYZ}}{p_{XZ}p_Y} \times \frac{p_Y}{p_{Y|Z}}$$

we have the first statement in the following lemma.

Lemma 3.15.

$$I(X;Y|Z) + I(Y;Z) = I(Y;(X,Z)), \qquad (3.30)$$

$$I(X;Y|Z) \geq 0, \qquad (3.31)$$

with equality if and only if X and Y are conditionally independent given Z, that is, $p_{XY|Z} = p_{X|Z}p_{Y|Z}$. Given finite valued measurements f and g,

$$I(f(X); g(Y)|Z) \leq I(X; Y|Z).$$

Proof: The second inequality follows from the divergence inequality (3.7) with $P = P_{XYZ}$ and $M = P_{X\times Y|Z}$, i.e., the PMF's p_{XYZ} and $p_{X|Z}p_{Y|Z}p_Z$. The third inequality follows from Lemma 3.3 or its corollary applied to the same measures. □

Comments: Eq. (3.30) is called *Kolmogorov's formula*. If X and Y are conditionally independent given Z in the above sense, then we also have that $p_{X|YZ} = p_{XY|Z}/p_{Y|Z} = p_{X|Z}$, in which case $Y \to Z \to X$ forms a *Markov chain* — given Z, X does not depend on Y. Note that if $Y \to Z \to X$ is a Markov chain, then so is $X \to Z \to Y$. Thus the conditional mutual information is 0 if and only if the variables form a Markov chain with the conditioning variable in the middle. One might be tempted to infer from Lemma 3.3 that given finite valued measurements f, g, and r

$$I(f(X); g(Y)|r(Z)) \overset{(?)}{\leq} I(X; Y|Z).$$

This does not follow, however, since it is not true that if $\mathcal{Q}$ is the partition corresponding to the three quantizers, then $D(P_{f(X),g(Y),r(Z)}\|P_{f(X)\times g(Y)|r(Z)})$ is $H_{P_{X,Y,Z}\|P_{X\times Y|Z}}(f(X), g(Y), r(Z))$ because of the way that $P_{X\times Y|Z}$ is constructed; e.g., the fact that X and Y are conditionally independent given Z implies that $f(X)$ and $g(Y)$ are conditionally independent given Z, but it does not imply that $f(X)$ and $g(Y)$ are conditionally independent given $r(Z)$. Alternatively, if M is $P_{X\times Z|Y}$, then it is not true that $P_{f(X)\times g(Y)|r(Z)}$ equals $M(fgr)^{-1}$. Note that if this inequality were true, choosing $r(z)$ to be trivial (say 1 for all z) would result in $I(X; Y|Z) \geq I(X; Y|r(Z)) = I(X; Y)$. This cannot be true in general since, for example, choosing Z as (X, Y) would give $I(X; Y|Z) = 0$. Thus one must be careful when applying Lemma 3.3 if the measures and random variables are related as they are in the case of conditional mutual information.

We close this section with an easy corollary of the previous lemma and of the definition of conditional entropy. Results of this type are referred to as *chain rules* for information and entropy.

Corollary 3.6. *Given finite-alphabet random variables $Y, X_1, X_2, \ldots, X_n$,*

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i|X_1, \ldots, X_{i-1})$$

$$H_{p\|m}(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H_{p\|m}(X_i|X_1, \ldots, X_{i-1})$$

$$I(Y;(X_1,X_2,\ldots,X_n)) = \sum_{i=1}^{n} I(Y;X_i|X_1,\ldots,X_{i-1}).$$

3.7 Entropy Rate Revisited

The chain rule of Corollary 3.6 provides a means of computing entropy rates for stationary processes. We have that

$$\frac{1}{n}H(X^n) = \frac{1}{n}\sum_{i=0}^{n-1} H(X_i|X^i).$$

First suppose that the source is a stationary kth order Markov process, that is, for any $m > k$

$$\begin{aligned}\Pr(X_n = x_n|X_i = x_i;\ i = 0,1,\ldots,n-1)\\ = \Pr(X_n = x_n|X_i = x_i;\ i = n-k,\ldots,n-1).\end{aligned}$$

For such a process we have for all $n \geq k$ that

$$H(X_n|X^n) = H(X_n|X_{n-k}^k) = H(X_k|X^k),$$

where $X_i^m = X_i,\ldots,X_{i+m-1}$. Thus taking the limit as $n \to \infty$ of the nth order entropy, all but a finite number of terms in the sum are identical and hence the Cesàro (or arithmetic) mean is given by the conditional expectation. We have therefore proved the following lemma.

Lemma 3.16. *If $\{X_n\}$ is a stationary kth order Markov source, then*

$$\overline{H}(X) = H(X_k|X^k).$$

If we have a two-sided stationary process $\{X_n\}$, then all of the previous definitions for entropies of vectors extend in an obvious fashion and a generalization of the Markov result follows if we use stationarity and the chain rule to write

$$\frac{1}{n}H(X^n) = \frac{1}{n}\sum_{i=0}^{n-1} H(X_0|X_{-1},\ldots,X_{-i}).$$

Since conditional entropy is nonincreasing with more conditioning variables ((3.23) or Lemma 3.12), $H(X_0|X_{-1},\ldots,X_{-i})$ has a limit. Again using the fact that a Cesàro mean of terms all converging to a common limit also converges to the same limit we have the following result.

Lemma 3.17. *If $\{X_n\}$ is a two-sided stationary source, then*

$$\overline{H}(X) = \lim_{n\to\infty} H(X_0|X_{-1},\ldots,X_{-n}).$$

It is tempting to identify the above limit as the conditional entropy given the infinite past, $H(X_0|X_{-1},\ldots)$. Since the conditioning variable is a sequence and does not have a finite alphabet, such a conditional entropy is not included in any of the definitions yet introduced. We shall later demonstrate that this interpretation is indeed valid when the notion of conditional entropy has been suitably generalized.

The natural generalization of Lemma 3.17 to relative entropy rates unfortunately does not work because conditional relative entropies are not in general monotonic with increased conditioning and hence the chain rule does not immediately yield a limiting argument analogous to that for entropy. The argument does work if the reference measure is a kth order Markov, as considered in the following lemma.

Lemma 3.18. *If $\{X_n\}$ is a source described by process distributions p and m and if p is stationary and m is kth order Markov with stationary transitions, then for $n \ge k$ $H_{p\|m}(X_0|X_{-1},\ldots,X_{-n})$ is nondecreasing in n and*

$$\begin{aligned}\overline{H}_{p\|m}(X) &= \lim_{n\to\infty} H_{p\|m}(X_0|X_{-1},\ldots,X_{-n})\\ &= -\overline{H}_p(X) - E_p[\ln m(X_k|X^k)].\end{aligned}$$

Proof: For $n \ge k$ we have that

$$\begin{aligned}&H_{p\|m}(X_0|X_{-1},\ldots,X_{-n}) =\\ &\qquad - H_p(X_0|X_{-1},\ldots,X_{-n}) - \sum_{x^{k+1}} p_{X^{k+1}}(x^{k+1}) \ln m_{X_k|X^k}(x_k|x^k).\end{aligned}$$

Since the conditional entropy is nonincreasing with n and the remaining term does not depend on n, the combination is nondecreasing with n. The remainder of the proof then parallels the entropy rate result. □

It is important to note that the relative entropy analogs to entropy properties often require kth order Markov assumptions on the reference measure (but not on the original measure).

3.8 Markov Approximations

Recall that the relative entropy rate $\overline{H}_{p\|m}(X)$ can be thought of as a distance between the process with distribution p and that with distribution m and that the rate is given by a limit if the reference measure m is Markov. A particular Markov measure relevant to p is the distribution

$p^{(k)}$ which is the *kth order Markov approximation to p* in the sense that it is a kth order Markov source and it has the same kth order transition probabilities as p. To be more precise, the process distribution $p^{(k)}$ is specified by its finite dimensional distributions

$$p^{(k)}_{X^k}(x^k) = p_{X^k}(x^k)$$
$$p^{(k)}_{X^n}(x^n) = p_{X^k}(x^k) \prod_{l=k}^{n-1} p_{X_l|X_{l-k}^k}(x_l|x_{l-k}^k);\ n = k, k+1, \ldots$$

so that

$$p^{(k)}_{X_k|X^k} = p_{X_k|X^k}.$$

It is natural to ask how good this approximation is, especially in the limit, that is, to study the behavior of the relative entropy rate $\overline{H}_{p\|p^{(k)}}(X)$ as $k \to \infty$.

Theorem 3.4. *Given a stationary process p, let $p^{(k)}$ denote the kth order Markov approximations to p. Then*

$$\lim_{k\to\infty} \overline{H}_{p\|p^{(k)}}(X) = \inf_k \overline{H}_{p\|p^{(k)}}(X) = 0.$$

Thus the Markov approximations are asymptotically accurate in the sense that the relative entropy rate between the source and approximation can be made arbitrarily small (zero if the original source itself happens to be Markov).

Proof: As in the proof of Lemma 3.18 we can write for $n \geq k$ that

$$\begin{aligned} &H_{p\|p^{(k)}}(X_0|X_{-1}, \ldots, X_{-n}) \\ &= -H_p(X_0|X_{-1}, \ldots, X_{-n}) - \sum_{x^{k+1}} p_{X^{k+1}}(x^{k+1}) \ln p_{X_k|X^k}(x_k|x^k) \\ &= H_p(X_0|X_{-1}, \ldots, X_{-k}) - H_p(X_0|X_{-1}, \ldots, X_{-n}). \end{aligned}$$

Note that this implies that $p^{(k)}_{X^n} \gg p_{X^n}$ for all n since the entropies are finite. This automatic domination of the finite dimensional distributions of a measure by those of its Markov approximation will not hold in the general case to be encountered later, it is specific to the finite alphabet case. Taking the limit as $n \to \infty$ gives

$$\begin{aligned} \overline{H}_{p\|p^{(k)}}(X) &= \lim_{n\to\infty} H_{p\|p^{(k)}}(X_0|X_{-1}, \ldots, X_{-n}) \\ &= H_p(X_0|X_{-1}, \ldots, X_{-k}) - \overline{H}_p(X). \end{aligned}$$

The corollary then follows immediately from Lemma 3.17. □

Markov approximations will play a fundamental role when considering relative entropies for general (nonfinite-alphabet) processes. The ba-

sic result above will generalize to that case, but the proof will be much more involved.

3.9 Relative Entropy Densities

Many of the convergence results to come will be given and stated in terms of relative entropy densities. In this section we present a simple but important result describing the asymptotic behavior of relative entropy densities. Although the result of this section is only for finite alphabet processes, it is stated and proved in a manner that will extend naturally to more general processes later on. The result will play a fundamental role in the basic ergodic theorems to come.

Throughout this section we will assume that M and P are two process distributions describing a random process $\{X_n\}$. Denote as before the sample vector $X^n = (X_0, X_1, \ldots, X_{n-1})$, that is, the vector beginning at time 0 having length n. The distributions on X^n induced by M and P will be denoted by M_n and P_n, respectively. The corresponding PMF's are m_{X^n} and p_{X^n}. The key assumption in this section is that for all n if $m_{X^n}(x^n) = 0$, then also $p_{X^n}(x^n) = 0$, that is,

$$M_n \gg P_n \text{ for all } n. \tag{3.32}$$

If this is the case, we can define the relative entropy density

$$h_n(x) \equiv \ln \frac{p_{X^n}(x^n)}{m_{X^n}(x^n)} = \ln f_n(x), \tag{3.33}$$

where

$$f_n(x) \equiv \begin{cases} \frac{p_{X^n}(x^n)}{m_{X^n}(x^n)} & \text{if } m_{X^n}(x^n) \neq 0 \\ 0 & \text{otherwise} \end{cases} \tag{3.34}$$

Observe that the relative entropy is found by integrating the relative entropy density:

$$\begin{aligned} H_{P\|M}(X^n) = D(P_n\|M_n) &= \sum_{x^n} p_{X^n}(x^n) \ln \frac{p_{X^n}(x^n)}{m_{X^n}(x^n)} \\ &= \int \ln \frac{p_{X^n}(X^n)}{m_{X^n}(X^n)}\, dP \end{aligned} \tag{3.35}$$

Thus, for example, if we assume that

$$H_{P\|M}(X^n) < \infty, \text{ all } n, \tag{3.36}$$

then (3.32) holds.

The following lemma will prove to be useful when comparing the asymptotic behavior of relative entropy densities for different probability measures. It is the first almost everywhere result for relative entropy densities that we consider. It is somewhat narrow in the sense that it only compares limiting densities to zero and not to expectations. We shall later see that essentially the same argument implies the same result for the general case (Theorem 7.4), only the interim steps involving PMF's need be dropped. Note that the lemma requires neither stationarity nor asymptotic mean stationarity.

Lemma 3.19. *Given a finite-alphabet process $\{X_n\}$ with process measures P, M satisfying (3.32), Then*

$$\limsup_{n\to\infty} \frac{1}{n} h_n \le 0, \ M - a.e. \tag{3.37}$$

and

$$\liminf_{n\to\infty} \frac{1}{n} h_n \ge 0, \ P - a.e.. \tag{3.38}$$

If in addition $M \gg P$, then

$$\lim_{n\to\infty} \frac{1}{n} h_n = 0, \ P - a.e.. \tag{3.39}$$

Proof: First consider the probability

$$M(\frac{1}{n} h_n \ge \epsilon) = M(f_n \ge e^{n\epsilon}) \le \frac{E_M(f_n)}{e^{n\epsilon}},$$

where the final inequality is Markov's inequality. But

$$\begin{aligned} E_M(f_n) &= \int dM f_n = \sum_{x^n:\, m_{X^n}(x^n)\neq 0} m_{X^n}(x^n) \frac{p_{X^n}(x^n)}{m_{X^n}(x^n)} \\ &= \sum_{x^n:\, m_{X^n}(x^n)\neq 0} p_{X^n}(x^n) \le 1 \end{aligned}$$

and therefore

$$M(\frac{1}{n} h_n \ge \epsilon) \le 2^{-n\epsilon}$$

and hence

$$\sum_{n=1}^{\infty} M(\frac{1}{n} h_n > \epsilon) \le \sum_{n=1}^{\infty} e^{-n\epsilon} < \infty.$$

From the Borel-Cantelli Lemma (e.g., Lemma 4.6.3 of [55] or Lemma 5.17 of [58]) this implies that $M(n^{-1} h_n \ge \epsilon \text{ i.o.}) = 0$ which implies the first equation of the lemma.

Next consider

$$P(-\frac{1}{n}h_n > \epsilon) = \sum_{x^n:-\frac{1}{n}\ln p_{X^n}(x^n)/m_{X^n}(x^n)>\epsilon} p_{X^n}(x^n)$$
$$= \sum_{x^n:-\frac{1}{n}\ln p_{X^n}(x^n)/m_{X^n}(x^n)>\epsilon \text{ and } m_{X^n}(x^n)\neq 0} p_{X^n}(x^n)$$

where the last statement follows since if $m_{X^n}(x^n) = 0$, then also $p_{X^n}(x^n) = 0$ and hence nothing would be contributed to the sum. In other words, terms violating this condition add zero to the sum and hence adding this condition to the sum does not change the sum's value. Thus

$$P(-\frac{1}{n}h_n > \epsilon)$$
$$= \sum_{x^n:-\frac{1}{n}\ln p_{X^n}(x^n)/m_{X^n}(x^n)>\epsilon \text{ and } m_{X^n}(x^n)\neq 0} \frac{p_{X^n}(x^n)}{m_{X^n}(x^n)} m_{X^n}(x^n)$$
$$= \int_{f_n<e^{-n\epsilon}} dM f_n \leq \int_{f_n<e^{-n\epsilon}} dM e^{-n\epsilon}$$
$$= e^{-n\epsilon} M(f_n < e^{-n\epsilon}) \leq e^{-n\epsilon}.$$

Thus as before we have that $P(n^{-1}h_n > \epsilon) \leq e^{-n\epsilon}$ and hence that $P(n^{-1}h_n \leq -\epsilon \text{ i.o.}) = 0$ which proves the second claim. If also $M \gg P$, then the first equation of the lemma is also true P-a.e., which when coupled with the second equation proves the third. □