0 Notation and Terminology.

This course will be concerned with the applications of information theory concepts in statistics. Much of the course will be based on lectures given by Imre Csiszár at Maryland in 1989. Some recent results about dependent processes will also be given. It is assumed that the reader is familiar with basic information theory ideas as presented, for example, in the initial chapters of the Csiszár-Körner book, and with basic statistical concepts as presented, for example, in the book by Cox and Hinkley. Notation and terminology that will be used in these lectures will be introduced in this section.

The symbol $A = \{a_1, a_2, \ldots, a_{|A|}\}$ will denote a finite set of cardinality $|A|$ and x_m^n will denote the sequence $x_m, x_{m+1}, \ldots, x_n$, where each $x_i \in A$. The set of all n-length sequences x_1^n will be denoted by A^n, the set of all infinite sequences $x = x_1^\infty$, with $x_i \in A, i \geq 1$ will be denoted by A^∞, and the set of all finite sequences drawn from A will be denoted by A^*. If u and v are finite length sequences then their concatenation is denoted by uv, and $u^k = u^{k-1}u,\ k > 1$.

The entropy $H(P)$ of a probability distribution, $P = (P(a))$ on A, is defined by the formula

$$H(P) = -\sum_{a \in A} P(a) \log P(a),$$

where here, as elsewhere in these lectures, base two logarithms are used. Random variable notation is often used in this context, that is, $H(X)$ denotes the entropy of the distribution P of the random variable X. If P and Q are two distributions on A then their divergence or cross-entropy is defined by

$$D(P\|Q) = \sum_{a \in A} P(a) \log \frac{P(a)}{Q(a)}.$$

If P is the joint distribution of two random variables (X, Y) then their joint entropy is defined by

$$H(X, Y) = -\sum_{(a,b)} P(a, b) \log P(a, b),$$

while the conditional entropy $H(X|Y)$ and mutual information $I(X \wedge Y)$ are defined, respectively, by

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

Two types of codes will be of interest. A block code is a mapping $C\colon A^n \mapsto B^m$, while a variable-length code is a mapping $C\colon A^n \mapsto B^*$. The length function $L\colon A^n \mapsto \{1, 2, \ldots\}$ for a variable-length is defined by the formula

$$C(x_1^n) = b_1^{L(x_1^n)}.$$

Thus, in particular, a block code is just a variable length code whose length function is constant.

A block code C is invertible (or faithful) if it is one-to-one. A variable-length code is *uniquely decodable* if for any two distinct sequences, $u(1), u(2), \ldots, u(m)$ and $v(1), v(2), \ldots, v(k)$, where $u(i), v(j) \in A^n$, $\forall i, j$, the concatenations of the images, $C(u(1))C(u(2)) \cdots C(u(m))$ and $C(v(1))C(v(2)) \cdots C(v(k))$, are not equal. A condition that guarantees unique decodability is the prefix condition. A variable-length code C satisfies the prefix condition if

$$C(v) = C(u)w, u, v \in A^n, w \in B^* \Rightarrow w = \Lambda,\ u = v,$$

where Λ denotes the empty string.

In most cases of interest to us, the image alphabet will be binary, that is, $B = \{0, 1\}$. It is easy to see that the length function for a binary prefix code must satisfy the so-called Kraft inequality.

$$\sum_{x_1^n} 2^{-L(x_1^n)} \leq 1.$$

It can in fact be shown that a uniquely decodable binary code also satisfies the Kraft inequality, and that if L is a positive integer-valued function on A^n for which the Kraft inequality holds then there is a binary prefix code C whose length function is L. (Thus, in particular, for any uniquely decodable code C with length function L there is a prefix code $\tilde{C}$ whose length function is also L.) The reason for the connection between the Kraft inequality and prefix codes is the connection between the Kraft inequality and binary trees, a connection that we now sketch.

A (binary) tree is a directed graph (V, E), along with a distinguished vertex $r \in V$, called the root, such that the following properties hold.

1. The outdegree of each vertex is at most 2.
2. The indegree of the root is 0. The indegree of all other vertices is exactly 1.
3. Given any $v \in V - r$ there is a directed path from r to v.

It is easy to see from the above that there is only one path from r to any $v \neq r$; the length of this path is called

the depth $d(v)$ of v. A vertex is called an *outer node* if its outdegree is 0; otherwise it is an inner node. Let $\mathcal{O}$ denote the set of outer nodes. It is easy to see that the edges of the tree can be labeled by 0's and 1's so that for any vertex v whose outdegree is 2, the two edges leading out of v have different labels. Such a labeling assigns a binary sequence of length $d(v)$ to each outer node v such that distinct outer nodes are assigned distinct sequences. The labeling is therefore just a binary code on the set of outer nodes. Furthermore, the code is a prefix code, due to the simple fact that an outer node is not an inner node! It is clear that

$$\sum_{v\in\mathcal{O}} 2^{-v(x)} \le 1.$$

In summary, binary trees lead to binary prefix codes on their outer codes for which the Kraft inequality holds.

Now suppose L is a positive integer-valued function defined on a set A such that $\sum 2^{-L(a)} \le 1$. Our goal is to show that there is a prefix code C whose length function is L. Without loss of generality it can be assumed that A is labeled so that $L(a_i) \le L(a_{i+1})$, $i < |A|$. The code C is defined by setting $C(a_i) = w(i) \in B^*$, where $w(1)$ is a block of 0's of length $L(a_1)$, and $w(i)$, $i > 1$ is the first $L(a_i)$ bits in the binary expansion of $\sum_{j<i} 2^{-L(a_j)}$. It is left to the reader to show that this defines a prefix code. The code is known as the Shannon-Fano code, or simply the Shannon code. The following theorem summarizes this coding construction in a form that will be used later.

Theorem 1 Let P be a probability distribution on A and define $L(a) = \lceil -\log P(a) \rceil$, $a \in A$, where $\lceil\cdot\rceil$ denotes the least integer function. There is a binary prefix code for which the expected length satisfies $E(L) = \sum_a L(a)P(a) \le H(P) + 1$.

We shall also make use of a prefix code defined on the integers, a code that is essentially due to Elias. Let $b(n)$ be the usual binary representation of the integer $n \ge 0$, and let $\ell(n)$ denote the length of $b(n)$, so that $\ell(n) = \lceil \log_2(n+1) \rceil$. Let O^k denote a sequence of 0's of length k. The code is defined by

$$C(n) = 0^{\ell(\ell(n))} b(\ell(n)) b(n).$$

For example $b(12) = 1100$, so $b(\ell(12)) = b(4) = 100$ and $\ell(\ell(12)) = 3$. Thus $C(12) = 0001001100$. The decoding is as follows. The initial block 000 of 0's has length 3. This tells us to look in the next 3 places, where we see 100, the binary representation of 4, which in turn tells us to look in the next 4 places where we see 1100, the binary representation of 12. The code C is a prefix code; the codeword length is $\ell(n) + 2\ell(\ell(n))$, which, for large n, is approximately equal to

$$\log_2 n + 2\log_2\log_2 n.$$

1 Large Deviations.

One important application of information theory is to the theory of large deviations. A key to this application is the theory of types. The *n-type* of a sequence $x_1^n \in A^n$ is just another name for its empirical distribution $\hat{P} = \hat{P}_{x_1^n}$, that is, the distribution defined by

$$\hat{P}(a) = \frac{|\{i\colon x_i = a\}|}{n},\ a \in A.$$

Two sequences x_1^n and y_1^n are said to be equivalent if they have the same type; the equivalence classes will be called *type classes*. The type class of x_1^n will be denoted by $\mathcal{T}_P^n$, where $P = \hat{P}_{x_1^n}$. The proof of the following lemma is left to the student.

Lemma 1 The number of possible types is $\binom{n+|A|-1}{|A|-1}$.

Theorem 2 For any type P

$$\binom{n+|A|-1}{|A|-1}^{-1} 2^{nH(P)} \le |\mathcal{T}_P^n| \le 2^{nH(P)}.$$

Proof. Fix the type P and define $P^n(x_1^n) = \prod_i P(x_i)$. A simple calculation shows that if x_1^n has type P then $P^n(x_1^n) = 2^{-nH(P)}$. Since P^n is a probability distribution on A^n we must have $P^n(\mathcal{T}_P^n) \le 1$. This gives the desired upper bound since $P^n(\mathcal{T}_P^n) = |\mathcal{T}_P^n| 2^{-nH(P)}$.

The lower bound can be obtained as follows. Let $A = \{a_1, a_2, \ldots, a_t\}$, where $t = |A|$. By definition of types we can write $P(a_i) = k_i/n, i = 1, 2, \ldots, t$ with $k_1 + k_2 + \ldots + k_t = n$, where k_i is the number of times a_i appears in x_1^n for any fixed $x_1^n \in \mathcal{T}_P^n$. Thus we have

$$|\mathcal{T}_P^n| = \frac{n!}{k_1!k_2!\cdots k_t!},$$

so that

$$n^n = (k_1 + \ldots + k_t)^n = \sum \frac{n!}{j_1!\cdots j_t!} k_1^{j_1}\cdots k_t^{j_t},$$

where the sum is over all t-tuples $(j_1, \ldots, j_t)$ of nonnegative integers such that $j_1 + \ldots + j_t = n$. The number of terms is $\left(\begin{array}{c} n+|A|-1 \\ |A|-1 \end{array} \right)$, by Lemma **??**, and the largest term is

$$\frac{n!}{k_1! k_2! \cdots k_t!} k_1^{k_1} k_2^{k_2} \cdots k_t^{k_t},$$

for if $j_r > k_r$, $j_s < k_s$ then decreasing j_r by 1 and increasing j_s by 1 multiplies by

$$\frac{j_r}{k_r} \frac{k_s}{1+j_s} \geq \frac{j_r}{k_r} \geq 1.$$

This yields the lower bound.

The following corollary will be useful in a later section.

Corollary 1 The minimum number $\ell_{\min}^{(n)}$ of bits needed to encode sequences x_1^n of known type P, with codewords of a fixed length, satisfies

$$nH(P) - \log \left(\begin{array}{c} n+|A|-1 \\ |A|-1 \end{array} \right) \leq \ell_{\min}^{(n)} \leq \lceil nH(P) \rceil.$$

In particular, $(1/n)\ell^{(n)} \to H(P)$ as $n \to \infty$.

Our next result connects the theory of types with general probability theory.

Theorem 3 For any distribution P on A and any n-type Q

$$\begin{aligned} \left(\begin{array}{c} n+|A|-1 \\ |A|-1 \end{array} \right)^{-1} 2^{-nD(Q\|P)} &\leq P^n(\mathcal{T}_Q^n) \\ &\leq 2^{-nD(Q\|P)}, \end{aligned}$$

where P^n is the product measure defined by P on A^n.

Proof. If x_1^n has type Q then the number of times $x_i = a$ is just $nQ(a)$, and hence

$$P^n(x_1^n) = \prod_a P(a)^{nQ(a)} = 2^{n \sum_a Q(a) \log P(a)}.$$

Thus Lemma **??** yields the desired upper bound

$$\begin{aligned} P^n(\mathcal{T}_Q^n) &= \left| \mathcal{T}_Q^n \right| 2^{n \sum_a Q(a) \log P(a)} \\ &\leq 2^{-n \sum_a Q(a) \log \frac{P(a)}{Q(a)}} \\ &= 2^{-nD(Q\|P)}. \end{aligned}$$

A similar argument establishes the lower bound.

Let $X_1, X_2, \ldots$ be independent random variables taking values in X with common distribution P and let $\hat{P}_n$ be the n-type of the random sequence $X_1, \ldots, X_n$. The law of large numbers tells us that $\hat{P}_n \to P$ with probability 1 as $n \to \infty$. The next result is useful for estimating the (exponentially small) probability that $\hat{P}_n$ belongs to some set Π of distributions that does not contain the true distribution P. We use the notation $D(\Pi\|P) = \inf_{Q \in \Pi} D(Q\|P)$.

Theorem 4 (Sanov's Theorem.) Let Π be a set of distributions on A whose closure is equal to the closure of its interior. Then

$$-\frac{1}{n} \log P\left(\hat{P}_n \in \Pi\right) \to D(\Pi\|P).$$

Proof. Let $\mathcal{P}_n$ be the set of possible n-types and let $\Pi_n = \Pi \cap \mathcal{P}_n$. Theorem **??** implies that

$$P(\hat{P}_n \in \Pi_n) = P^n\left(\cup_{Q \in \Pi_n} \mathcal{T}_Q^n\right)$$

is upper bounded by

$$\left(\begin{array}{c} n+|A|-1 \\ |A|-1 \end{array} \right) 2^{-nD(\Pi_n\|P)}$$

and lower bounded by

$$\left(\begin{array}{c} n+|A|-1 \\ |A|-1 \end{array} \right)^{-1} 2^{-nD(\Pi_n\|P)}.$$

Since $D(Q\|P)$ is continuous in Q, the hypothesis on Π implies that $D(\Pi_n\|P)$ is arbitrarily close to $D(\Pi\|P)$ if n is large. Hence the theorem follows.

Example 1 Let f be a given function on A and set $\Pi = \{Q: \sum_a Q(a) f(a) > \alpha\}$ where $\alpha < \max_a f(a)$. The set Π is open and hence satisfies the hypothesis of Sanov's theorem. Note that $\hat{P}_n \in \Pi$ is equivalent to $(1/n) \sum_{x_i} f(x_i) > \alpha$, since $\sum_a \hat{P}_n(a) f(a) = (1/n) \sum_{x_i} f(x_i)$. Thus we obtain the classical large deviations result

$$-\frac{1}{n} \log P^n\left(\frac{1}{n} \sum_{i=1}^n f(x_i) > \alpha\right) \to D(\Pi\|P).$$

In this case, $D(\Pi\|P) = D(cl(\Pi)\|P) = \min D(Q\|P)$, where the minimum is over all Q for which $\sum Q(a) f(a) \geq \alpha$. In particular, for any $\alpha > \sum P(a) f(a)$ we have $D(\Pi\|P) > 0$, so that, the probability of $(1/n) \sum_1^n f(X_i) > \alpha$ goes to 0 exponentially fast.

It is instructive to see how to calculate the exponent $D(\Pi\|P)$ for the preceeding example. Consider the exponential family of distributions $\tilde{P}$ of the form $\tilde{P}(a) = cP(a)2^{tf(a)}$, where $c = (\sum_a P(a)2^{tf(a)})^{-1}$. Clearly $\sum_a \tilde{P}(a)f(a)$ is a continuous function of the parameter t and this function tends to $\max f(a)$ as $t \to \infty$. (Check!) As $t = 0$ gives $\tilde{P} = P$, it follows by the assumption

$$\sum_a P(a)f(a) < \alpha < \max_a f(a)$$

that there an element of the exponential family, with $t > 0$, such that $\sum \tilde{P}(a)f(a) = \alpha$. Denote this $\tilde{P}$ by Q^*, so that,

$$Q^*(a) = c^*P(a)2^{t^*f(a)},\ t^* > 0,\ \sum_a Q^*(a)f(a) = \alpha.$$

We claim that

$$D(\Pi\|P) = D(Q^*\|P) = \log c^* + t^*\alpha. \tag{1}$$

To show that $D(\Pi\|P) = D(Q^*\|P)$ it suffices to show that $D(Q\|P) > D(Q^*\|P)$ for every $Q \in \Pi$, i. e., for every Q for which $\sum_a Q(a)f(a) > \alpha$. A direct calculation gives

$$D(Q^*\|P) = \sum_a Q^*(a)\log\frac{Q^*(a)}{P(a)} = \sum_a Q^*(a)\left[\log c^* + t^*f(a)\right] = \log c^* + t^*\alpha \tag{2}$$

and

$$\sum_a Q(a)\log\frac{Q^*(a)}{P(a)} = \sum_a Q(a)\left[\log c^* + t^*f(a)\right] > \log c^* + t^*\alpha.$$

Hence

$$D(Q\|P) - D(Q^*\|P) > D(Q\|P) - \sum_a Q(a)\log\frac{Q^*(a)}{P(a)} = D(Q\|Q^*) > 0.$$

This completes the proof of (**??**).

Remark 1 Replacing P in (**??**) by any $\tilde{P}$ of the exponential family, i. e., $\tilde{P}(a) = cP(a)2^{tf(a)}$, we get that

$$D(Q^*\|\tilde{P}) = \log\frac{c^*}{c} + (t^* - t)\alpha = \log c^* + t^*\alpha - (\log c + t\alpha).$$

Since $D(Q^*\|\tilde{P}) > 0$ for $\tilde{P} \neq Q^*$, it follows that

$$\log c + t\alpha = -\log\sum_a P(a)2^{tf(a)} + t\alpha$$

attains its maximum at $t = t^*$. This means that the "large deviations exponent"

$$\lim_{n\to}\left[-\frac{1}{n}\log P^n\left\{\frac{1}{n}\sum_{i=1}^n f(X_i) > \alpha)\right\}\right]$$

can be represented also as

$$\max_{t\geq 0}\left[-\log\sum_a P(a)2^{tf(a)} + t\alpha\right].$$

This latter form is the one usually found in textbooks. Note that the restriction $t \geq 0$ is not needed when $\alpha > \sum_a P(a)f(a)$, because, as just seen, the unconstrained maximum is attained at $t^* > 0$. However, the restriction to $t \geq 0$ takes care also of the case when $\alpha \leq \sum_a P(a)f(a)$, when the exponent is equal to 0.

2 I-projections.

The *I-projection* of a distribution Q onto a closed, convex subset Π of distributions on A is the $P^* \in \Pi$ such that

$$D(P^*\|Q) = \min_{P\in\Pi} D(P\|Q).$$

In the sequel we suppose that $Q(a) > 0$ for all $a \in A$. The function $D(P\|Q)$ is then continuous and strictly convex in P, so that P^* exists and is unique.

The *support* of the distribution P is the set $S(P) = \{a\colon P(a) > 0\}$. Since Π is convex, among the supports of elements of Π there is one whose support contains all the others; this will be called the support of Π and denoted by $S(\Pi)$.

Theorem 5 $S(P^*) = S(\Pi)$ and $D(P\|Q) \geq D(P\|P^*) + D(P^*\|Q)$ for all $P \in \Pi$.

Proof. Of course, if the asserted inequality holds for some $P^* \in \Pi$ and all $P \in \Pi$ then P^* must be the I-projection of Q onto Π.

For arbitrary $P \in \Pi$, by the convexity of Π we have $P_t = (1-t)P^* + tP \in \Pi$, for $0 \leq t \leq 1$, hence for each $t \in (0,1)$,

$$0 \leq \frac{1}{t}\left[D(P_t\|Q) - D(P^*\|Q)\right] = \frac{d}{dt}D(P_t\|Q)\,|_{t=\tilde{t}}\,,$$

for some $\tilde{t} \in (0,t)$. But

$$\frac{d}{dt}D(P_t\|Q) = \sum_a (P(a) - P^*(a)) \log \frac{P_t(a)}{Q(a)},$$

and this converges (as $t \downarrow 0$) to $-\infty$ if $P^*(a) = 0$ for some $a \in S(P)$, and otherwise to

$$\sum_a (P(a) - P^*(a)) \log \frac{P^*(a)}{Q(a)}. \tag{3}$$

It follows that the first contigency is ruled out, proving that $S(P^*) \supset S(P)$, and also that the quantity (**??**) is nonnegative, proving the claimed inequality.

Now we examine some situations in which the inequality of Theorem **??** is actually an equality. For any given functions $f_1, f_2, \ldots, f_k$ on A and corresponding numbers $\alpha_1, \alpha_2, \ldots, \alpha_k$, the set

$$\mathcal{L} = \{P: \sum_a P(a) f_i(a) = \alpha_i, 1 \le i \le k\},$$

will be called a *linear family* of probability distributions. For any given functions $f_1, f_2, \ldots, f_k$ on A, the set $\mathcal{E}$ of all P such that

$$P(a) = cQ(a) \exp(\sum_1^k \theta_i f_i(a)), \text{ for some } \theta_1, \ldots, \theta_k,$$

will be called an *exponential family* of probability distributions; here Q is any given distribution and

$$c = c(\theta_1, \ldots, \theta_k) = \left(\sum_a Q(a) \exp(\sum_1^k \theta_i f_i(a)) \right)^{-1}.$$

We will assume that $S(Q) = A$; then $S(P) = A$ for all $P \in \mathcal{E}$. Note that $Q \in \mathcal{E}$. The family $\mathcal{E}$ depends on Q, of course, and but only in a weak manner, for any element of $\mathcal{E}$ could play the role of Q. If necessary to emphasize this dependence on Q we shall write $\mathcal{E} = \mathcal{E}_Q$.

Theorem 6 The I-projection P^* of Q onto a linear family $\mathcal{L}$ satisfies

$$D(P\|Q) = D(P\|P^*) + D(P^*\|Q), \ \forall P \in \mathcal{L}.$$

Further, if $S(\mathcal{L}) = A$ then $\mathcal{L} \cap \mathcal{E}_Q = \{P^*\}$.

Proof. By the preceeding theorem, $S(P^*) = S(\mathcal{L})$. Hence for every $P \in \mathcal{L}$ there is some $t < 0$ such that $P_t = (1-t)P^* + tP \in \mathcal{L}$. Therefore, we must have $(d/dt)D(P_t\|Q)|_{t=0} = 0$, that is, the quantity (**??**) in the preceeding proof is equal to 0, for all $P \in \mathcal{L}$. This gives the desired identity. Also we can equivalently write

$$\sum_a P(a) \left[\log \frac{P^*(a)}{Q(a)} - D(P^*\|Q) \right] = 0, \ P \in \mathcal{L}. \tag{4}$$

Now, by the definition of $\mathcal{L}$, the distributions $P \in \mathcal{L}$, regarded as $|A|$-dimensional vectors, are in the orthogonal complement of the subspace $\mathcal{F}$ spanned by the k vectors, $\{f_i(\cdot) - \alpha_i : 1 \le i \le k\}$. If $S(\mathcal{L}) = A$ then the distributions $P \in \mathcal{L}$ also span the orthogonal complement of $\mathcal{F}$, from Lemma **??**, below , and hence the identity (**??**) implies that the vector

$$\log \frac{P^*(\cdot)}{Q(\cdot)} - D(P^*\|Q)$$

must be in $\mathcal{F}$. This proves that $P^* \in \mathcal{E}_Q$.

Finally, if $\tilde{P} \in \mathcal{L} \cap \mathcal{E}_Q$ then it is easily checked that the identity (**??**) holds for $\tilde{P}$ in place of P^*. This implies that $\tilde{P}$ satisfies the Pythagorean identity in the role of P^*, and this, in turn, implies that $\tilde{P} = P^*$.

The proof of the theorem is finished, once the following linear algebra result is established.

Lemma 2 Suppose V is a the subspace of R^n such that there is a strictly positive vector $p \in V^\perp$, the orthogonal complement of V. Then $V^\perp$ is spanned by the probability vectors that belong to it.

Proof. Choose a basis for $V^\perp$ of the form $\{p, q_1, \ldots, q_\ell\}$ and determine $t_i \in (0,1)$, $1 \le i \le \ell$ such that $p_i = (1-t_i)p + t_i q_i$ is a nonnegative vector. The vectors $\{p, p_1, \ldots, p_\ell\}$ are easily seen to be a basis for $V^\perp$; each can be then be rescaled to obtain a basis for $V^\perp$ that consists of probability vectors. This completes the proof of the lemma.

If $S(\mathcal{L}) \neq A$ then no element of the exponential family $\mathcal{E} = \mathcal{E}_Q$ can belong to $\mathcal{L}$, but since $\mathcal{E}$ is not a closed set in general, some element of the closure, $cl(\mathcal{E})$ may be in $\mathcal{L}$. Indeed, if there is a $\tilde{P} \in \mathcal{L} \cap cl(\mathcal{E})$ then the Pythagorean identity still holds for $\tilde{P}$, and this implies that $\tilde{P} = P^*$. A sequence of elements converging to P^* can always be generated by the "generalized iterative scaling" algorithm, which will be discussed at the end of this section. Hence we always have $\mathcal{L} \cap cl(\mathcal{E}) = \{P^*\}$.

Suppose now that $\mathcal{L}_1, \ldots, \mathcal{L}_m$ are given linear families and generate a sequence of distributions P_n as follows: Set $P_0 = Q$ (any given distribution with $S(Q) = A$), let P_1 be the I-projection of P_0 onto $\mathcal{L}_1$, P_2 the I-projection of P_1 onto $\mathcal{L}_2$, and so on, where for $n > m$ we mean by $\mathcal{L}_n$ that $\mathcal{L}_i$ for which $i \equiv n \pmod m$; i. e., $\mathcal{L}_1, \ldots, \mathcal{L}_m$ is repeated cyclically.

Theorem 7 If $\cap_{i=1}^m \mathcal{L}_i = \mathcal{L} \neq \emptyset$ then $P_n \to P^*$, the I-projection of Q onto $\mathcal{L}$.

Proof. By the preceeding theorem, we have for every $P \in \mathcal{L}$ (even for $P \in \mathcal{L}_n$) that

$$D(P\|P_{n-1}) = D(P\|P_n) + D(P_n\|P_{n-1}), n = 1, 2, \ldots$$

Adding these equations for $1 \leq n \leq N$ we get that

$$D(P\|Q) = D(P\|P_0) = D(P\|P_N) + \sum_{n=1}^{N} D(P_n\|P_{n-1}).$$

By compactness there exists a subsequence $P_{N_k} \to P'$, say, and then from the preceeding inquality we get for $N_k \to \infty$ that

$$D(P\|Q) = D(P\|P') + \sum_{n=1}^{\infty} D(P_n\|P_{n-1}) \qquad (5)$$

Since this series is convergent we have $D(P_n\|P_{n-1}) \to 0$, and hence also $|P_n - P_{n-1}| \to 0$, where $|P_n - P_{n-1}|$ denotes the usual variational distance $\sum_a(|P_n(a) - P_{n-1}(a)|$. This implies that together with $P_{N_k} \to P'$ we also have

$$P_{N_k+1} \to P', P_{N_k+2} \to P', \ldots, P_{N_k+m} \to P'.$$

Since by the periodic construction, among the m consecutive elements, $P_{N_k}, P_{N_k+1}, \ldots, P_{N_k+m-1}$ there is one in each $\mathcal{L}_i, i = 1, 2, \ldots, m$, it follows that $P' \in \cap\mathcal{L}_i = \mathcal{L}$.

Since $P' \in \mathcal{L}$ it may be substituted for P in (**??**) to yield

$$D(P'\|Q) = \sum_{i=1}^{\infty} D(P_n\|P_{n-1}).$$

With this, in turn, (**??**) becomes

$$D(P\|Q) = D(P\|P') + D(P'\|Q),$$

which proves that P' equals the I-projection of Q onto $\mathcal{L}$. Finally, as P' was the limit of an arbitrary convergent subsequence of the sequence P_n, our result means that every convergent subsequence of P_n has the same limit P^*. Using compactness again, this proves that $P_n \to P^*$ and completes the proof of the theorem.

Now we discuss *iterative scaling*, a method for evaluating I-projections that is useful in the analysis of contigency tables, a subject to be discussed in the next section. Let $\mathcal{B} = \{B_1, B_2, \ldots, B_k\}$ be a partition of A and let P be a distribution on A. The distribution defined on $\{1, 2, \ldots, k\}$ by the formula

$$P^{\mathcal{B}}(i) = \sum_{a \in B_i} P(a),$$

is called the $\mathcal{B}$-*lumping* of P.

Fix nonnegative constants $\alpha_i, i \leq k$, whose sum is 1, and let $\mathcal{L} = \{P\colon P^{\mathcal{B}}(i) = \alpha_i, \forall i\}$. The I-projection of any Q onto $\mathcal{L}$ is obtained simply by "scaling":

$$P^*(a) = c_i Q(a),\ a \in B_i,\ \text{where } c_i = \frac{\alpha_i}{Q^{\mathcal{B}}(i)}. \qquad (6)$$

This follows from the fact that lumping does not increase divergence, that is,

$$D(P\|Q) \geq D(P^{\mathcal{B}}\|Q^{\mathcal{B}}).$$

The condition that $P \in \mathcal{L}$ is equivalent to the condition that $P(B_i) = \alpha_i,\ \forall i$. If $P^*(a) = \alpha_i Q(a)/Q(B_i),\ a \in B_i$ then

$$\begin{aligned} \sum_a P^*(a) \log \frac{P^*(a)}{Q(a)} &= \sum_i \sum_{a \in B_i} \frac{\alpha_i Q(a)}{Q(B_i)} \log \frac{\alpha_i}{Q(B_i)} \\ &= D(\alpha\|Q^B),\ \alpha = (\alpha_1, \ldots, \alpha_k). \end{aligned}$$

Thus, if $P \in \mathcal{L}$ then

$$D(P\|Q) \geq D(P^B\|Q^B) = D(\alpha\|Q^B),$$

which establishes (**??**).

Now, if $\mathcal{L}_1, \mathcal{L}_2, \ldots, \mathcal{L}_m$ are all of the preceeding form, then the iterated sequence of I-projections $P_1, P_2, \ldots,$ in Theorem **??** can all be obtained by iterative scaling, and the theorem gives that the so obtained sequence converges to the I-projection of Q onto the intersection $\mathcal{L} = \cap_{i=1}^m \mathcal{L}_i$. In particular, as we shall see in a later section, iterative scaling can be used to evaluate the I-projections that are needed in the analysis of contigency tables.

3 f-divergence and contigency tables.

Let $f(t)$ be a convex function defined for $t > 0$ with $f(1) = 0$. The *f-divergence* of a distribution P from Q is defined by

$$D_f(P\|Q) = \sum_a Q(x) f\left(\frac{P(x)}{Q(x)}\right).$$

Here we take $0f(\frac{0}{0}) = 0,\ f(0) = \lim_{t\to 0} f(t),\ 0f(\frac{a}{0}) = \lim_{t\to 0} tf(\frac{a}{t}) = a\lim_{u\to\infty} \frac{f(u)}{u}$.

Some examples include the following.

(1) $f(t) = t\log t \Rightarrow D_f(P\|Q) = D(P\|Q)$.

(2) $f(t) = -\log t \Rightarrow D_f(P\|Q) = D(Q\|P)$.

(3) $f(t) = (t-1)^2$
$\Rightarrow D_f(P\|Q) = \sum_a \frac{(P(a)-Q(a))^2}{Q(a)}$.

(4) $f(t) = 1-\sqrt{t}$
$\Rightarrow D_f(P\|Q) = 1 - \sum_a \sqrt{P(a)Q(a)}$.

(5) $f(t) = |t-1| \Rightarrow D_f(P\|Q) = |P-Q|$.

The expression $D_f(P\|Q) = \sum_a \frac{(P(a)-Q(a))^2}{Q(a)}$ will be denoted by $\chi^2(P,Q)$. The analogue of the log-sum inequality is

$$\sum_i b_i f\left(\frac{a_i}{b_i}\right) \geq bf\left(\frac{a}{b}\right),\ a = \sum a_i,\ b = \sum b_i.$$

Using this, many of the properties of the information divergence $D(P\|Q)$ extend to general f-divergences, in particular

Lemma 3 $D_f(P\|Q) \geq 0$ and if f is strictly convex at $t=1$ then $D_f(P\|Q) = 0$ only when $P = Q$. Further, $D_f(P\|Q)$ is a convex function of the pair (P,Q), and the partitioning property, $D_f(P\|Q) \geq D_f(P^{\mathcal{B}}\|Q^{\mathcal{B}})$ holds for any partition $\mathcal{B}$ of A.

A basic theorem about f-divergences is the following approximation property.

Theorem 8 If f is twice differentiable at $t=1$ and $f''(1) > 0$ then for any Q with $S(Q) = A$ and P " close" to Q we have

$$D_f(P\|Q) \sim \frac{f''(1)}{2}\chi^2(P,Q)$$

(Formally, $D_f(P\|Q)/\chi^2(P,Q) \to f''(1)/2$ as $\chi^2(P,Q) \to 0$.)

Proof. Since $f(1) = 0$, Taylor's expansion gives

$$f(t) = f'(1)(t-1) + \frac{f''(1)}{2}(t-1)^2 + \epsilon(t)(t-1)^2,$$

where $\epsilon(t) \to 0$ as $t \to 1$. Hence

$$Q(a)f\left(\frac{P(a)}{Q(a)}\right) =$$
$$f'(1)(P(a)-Q(a)) + \frac{f''(1)}{2}\frac{(P(a)-Q(a))^2}{Q(a)}$$
$$+\epsilon\left(\frac{P(a)}{Q(a)}\right)\frac{(P(a)-Q(a))^2}{Q(a)}.$$

Summing over $a \in A$ then establishes the theorem.

Remark 2 The same proof works even if Q is not fixed, provided that no $Q(a)$ can become arbitrarily small. However, the theorem (the "asymptotic equivalence" of f-divergences subject to the differentiability hypotheses) does not remain true if Q is not fixed and the probabilities of $Q(a)$ are not bounded away from 0.

Corollary 2 If f satisfies the hypotheses of the theorem and $\hat{P}$ is the empirical distribution (i. e., type) of a sample of size n drawn independently from the distribution Q, then $(2/f''(1))nD_f(\hat{P}\|Q)$ has an asymptotic χ^2 distribution, with $|A|-1$ degrees of freedom, as $n \to \infty$.

The χ^2 distribution with k degrees of freedom is defined as the distribution of the sum of squares of k independent random variables having the standard normal distribution. By this corollary, both $(2/\log e)nD(\hat{P}\|Q)$ and $(2/\log e)nD(Q\|\hat{P})$ are asymptotically χ^2 with $|A|-1$ degrees of freedom.

One property that distiguishes information divergence among f-divergences is transitivity of projections, as summarized in the following lemma. It can, in fact, be shown that the only f-divergence for which either of the two properties of the lemma holds is the informational divergence.

Lemma 4 Let P^* be the I-projection of Q onto a linear family $\mathcal{L}$. Then

(i) For any convex subfamily $\mathcal{L}' \subset \mathcal{L}$ the I-projections of Q and of P^* onto $\mathcal{L}'$ are the same.

(ii) For any "translate" $\mathcal{L}'$ of $\mathcal{L}$, the I-projections of Q and of P^* onto $\mathcal{L}'$ are the same, provided $S(P^*) = A$.

Proof. By the Pythagorean identity

$$D(P\|Q) = D(P\|P^*) + D(P^*\|Q),\ P \in \mathcal{L}.$$

It follows that on any subset of $\mathcal{L}$ the minimum of $D(P\|Q)$ and of $D(P\|P^*)$ are acheived by the same P. This establishes (i).

$\mathcal{L}'$ is called a translate of $\mathcal{L}$ if it is defined in terms of the same functions f_i, but possibly different α_i. Hence, the exponential family corresponding to $\mathcal{L}'$ is the same as it is for $\mathcal{L}$. Since $S(P^*) = A$, we know that P^* belongs to this exponential family. But every element of the exponential family has the same I-projection onto $\mathcal{L}'$, which establishes (ii).

Table 1: A 2-dimensional contingency table.

$x(0,0)$	$x(0,1)$	$\cdots$	$x(0,r_2)$	$x(0\cdot)$
$x(1,0)$	$x(1,1)$	$\cdots$	$x(1,r_2)$	$x(1\cdot)$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$x(r_1,0)$	$x(r_1,1)$	$\cdots$	$x(r_1,r_2)$	$x(r_1\cdot)$
$x(\cdot,0)$	$x(\cdot,1)$	$\cdots$	$x(\cdot,r_2)$	n

Now we apply some of these ideas to the analysis of contingency tables. A 2-dimensional contigency table is indicated in Figure **??**. The sample data have two features, with categories $0,\dots,r_1$ for the first feature and $0,\dots,r_2$ for the second feature. The cell counts

$$x(j_1,j_2),\ 0\le j_1\le r_1,\ 0\le j_2\le r_2$$

are nonnegative integers; thus in the sample there were $x(j_1,j_2)$ members that had category j_1 for the first feature and j_2 for the second. The table has two marginals with marginal counts

$$x(j_1\cdot)=\sum_{j_2=0}^{r_2}x(j_1,j_2),\ x(\cdot j_2)=\sum_{j_1=0}^{r_1}x(j_1,j_2).$$

The sum of all the counts is

$$n=\sum_{j_1}x(j_1\cdot)=\sum_{j_2}x(\cdot j_2)=\sum_{j_1}\sum_{j_2}x(j_1,j_2).$$

The term contigency table comes from this example, the cell counts being arranged in a table, with the marginal counts appearing at the margins. Other forms are also commonly used, e. g., the marginal empirical probabilities are indicated by replacing $x(j_1\cdot)$ by $\hat p(j_1\cdot)=x(j_1\cdot)/n$ and $x(\cdot j_2)$ by $\hat p(\cdot j_2)=x(\cdot j_2)/n$, and/or the counts are replaced by the relative counts, $\hat p(j_1,j_2)=x(j_1,j_2)/n$.

In the general case the sample has d *features* of interest, with the ith feature having categories $0,1,\dots,r_i$. The d-tuples $\omega=(j_1,\dots,j_d)$ are called *cells*; the corresponding *cell count* $x(\omega)$ is the number of members of the sample such that, for each i, the ith feature is in the j_ith category. The collection of possible cells will be denoted by Ω. The empirical distribution is defined by $\hat p(\omega)=x(\omega)/n$, where $n=\sum_\omega x(\omega)$ is the sample size. By a d-dimensional contingency table we mean either the aggregate of the cell counts $x(\omega)$, or the empirical distribution $\hat p$, or sometimes any distribution P on Ω (mainly when considered as a model for the "true distribution" from which the sample came.)

The *marginals* of a contingency table are obtained by restricting attention to those features i that belong to some given set $\gamma\subset\{1,2,\dots,d\}$. Formally, for $\gamma=(i_1,\dots,i_k)$ we denote by $\omega(\gamma)$ the γ-projection of $\omega=(j_1,\dots,j_d)$, that is, $\omega(\gamma)=(j_{i_1},j_{i_2},\dots,j_{i_k})$. The γ-marginal of the contingency table is given by the marginal counts

$$x(\omega(\gamma))=\sum_{\omega':\omega'(\gamma)=\omega(\gamma)}x(\omega')$$

or the corresponding empirical distribution $\hat p(\omega(\gamma))=x(\omega(\gamma))/n$. In general the γ-marginal of any distribution $P(\omega)\colon\omega\in\Omega$ is defined as the distribution P_γ defined by the marginal probabilities

$$P_\gamma(\omega(\gamma))=\sum_{\omega':\omega'(\gamma)=\omega(\gamma)}P(\omega').$$

In general a d-dimensional contigency table has d one-dimensional marginals, $d(d-1)/2$ two-dimensional marginals, etc., corresponding to the subsets of $\{1,\dots,d\}$ of one, two, etc., elements.

For contingency tables the most important linear families of distributions are those defined by fixing certain γ-marginals, for a family Γ of sets $\gamma\subset\{1,\dots,d\}$. Thus, denoting the fixed marginals by $\bar P_\gamma,\gamma\in\Gamma$, we consider

$$\mathcal{L}=\{P\colon P_\gamma=\bar P_\gamma,\gamma\in\Gamma\}.$$

The exponential family (through any given Q) that corresponds to this linear family $\mathcal{L}$ consists of all distributions that can be represented in product form as

$$P(\omega)=cQ(\omega)\prod_{\gamma\in\Gamma}a_\gamma(\omega(\gamma)).\tag{7}$$

In particular, if $\mathcal{L}$ is given by fixing the one-dimensional marginals (i. e., Γ consists of the one point subsets of $\{1,\dots,d\}$ then the corresponding exponential family consists of the distributions of the form

$$P(i_1,\dots,i_d)=cQ(i_1,\dots,i_d)a_1(i_1)\cdots a_d(i_d)$$

The family of all distributions of the form (**??**) is called the *log-linear family* with *interactions* $\gamma\in\Gamma$. In most applications, Q is chosen as the uniform distributions; often the name "log-linear family" is restricted to this case. Then (**??**) gives that the log of $P(\omega)$ is equal to a sum of terms, each representing an "interaction" $\gamma\in\Gamma$, for it depends on $\omega=(j_1,\dots,j_d)$ only through $\omega(\gamma)=(j_{i_1},\dots,j_{i_k})$, where $\gamma=(i_1,\dots,i_k)$.

A log-linear family is also called a *log-linear model*. It should be noted that the representation (**??**) is not

unique, because it corresponds to a representation in terms of linearly dependent functions. A common way of eachieving uniqueness is to postulate $a_\gamma(\omega(\gamma)) = 1$ whenever at least one component of $\omega(\gamma)$ is equal to 0. In this manner a unique representation of the form (**??**) is obtained, provided that with every $\gamma \in \Gamma$ also the subsets of γ are in Γ. Log-linear models of this form are also called *hierarchical models.*

Remark 3 The way we introduced log-linear models shows that restricting to the hierarchical ones is more a notational than a real restriction. Indeed, if some γ-marginal is fixed then so are the γ'-marginals for all $\gamma' \subset \gamma$.

In some cases of interest it is desirable to summarize the information content of a contingency table by its γ-marginals, $\gamma \in \Gamma$. In such cases it is natural to consider the linear family $\mathcal{L}$ constisting of those distributions whose γ-marginals equal those of the empirical distribution, $\hat{P}$. If a prior guess Q is available, then we accept the I-projection P^* of Q onto $\mathcal{L}$ as an estimate of the true distribution. By previous results, this P^* equals the intersections of the log-linear family (**??**), or its closure, with the linear family $\mathcal{L}$. Also, P^* equals the maximum likelihood estimate of the true distribution if it is assumed to belong to (**??**).

Again, by previous results, an asymptotically optimal test of the null-hypothesis that the true distribution belongs to the log-linear family $\mathcal{E}$ with interactions $\gamma \in \Gamma$ consists in accepting the null-hypothesis if

$$D(\hat{P}\|P^*) = \min_{p\in\mathcal{E}} D(\hat{P}\|P)$$

is "small." Unfortunately the numerical bounds obtained in our asymptotic calculation are too crude for most applications. Better bounds can be obtained from the following theorem (still asymptotic, but typically good for substantially smaller sample sizes than our exponential error bounds.)

Theorem 9 If the true distribution Q is in $\mathcal{E}$ then the terms on the right-hand side of the Pythagorean identity

$$D(\hat{P}\|Q) = D(\hat{P}\|P^*) + D(P^*\|Q)$$

are asymptotically independent and (after scaling) have χ^2 distributions with appropriate degrees of freedom.

Remark 4 The scaling is by $2n/\log e$ as in Corollary **??**. The degrees of freedom for $D(\hat{P}\|P^*)$ equals the number of (independent) constraints determining $\mathcal{L}$, and the degrees of freedom for $D(P^*\|Q)$ are determined from the condition that the total degrees of freedom is $|\Omega| - 1$.

The proof of Theorem **??** is omitted. Using this theorem, the null-hypothesis is rejected if $(2n/\log e)D(\hat{P}\|P^*)$ exceeds the threshold found in the table of the χ^2 distribution for the selected level of significance.

Now we look at the problem of *outliers.* A lack of fit (i. e., $D(\hat{P}\|P^*)$ "large") may be due not to the inadequacy of the model tested, but to outliers. A cell ω_0 is considered to be an outlier in the following case: Let $\mathcal{L}$ be the linear family determined by the γ-marginals of the empirical distribution $\hat{P}$, $(\gamma \in \Gamma)$ and let $\mathcal{L}'$ be the subfamily of $\mathcal{L}$ consisting of those $P \in \mathcal{L}$ that satisfy $P(\omega_0) = \hat{P}(\omega_0)$. Let P^{**} be the I-projection of P^* onto $\mathcal{L}'$. Ideally, we should consider ω_0 as an outlier if $D(P^{**}\|P^*)$ is "large", for if $D(P^{**}\|P^*)$ is close to $D(\hat{P}\|P^*)$ then $D(\hat{P}\|P^{**})$ will be small by the Pythagorean identity. Now by the partitioning inequality:

$$D(P^{**}\|P^*) \geq \hat{P}(\omega_0)\log\frac{\hat{P}(\omega_0)}{P^*(\omega_0)} + \left(1-\hat{P}(\omega_0\right)\log\frac{\hat{P}(\omega_0)}{P^*(\omega_0)},$$

and we declare ω_0 as an outlier if the right-hand side of this inequality is "large", that is, after scaling by $(2n/\log e)$, it exceeds the critical value of χ^2 with one degree of freedom.

If the above method produces only a few outliers, say $\omega_0, \omega_1, \ldots, \omega_\ell$, we consider the subset $\tilde{\mathcal{L}}$ of $\mathcal{L}$ consisting of those $P \in \mathcal{L}$ that satisfy $P(\omega_j) = \hat{P}(\omega_j)$ for $j = 0, \ldots, \ell$. If the I-projection of P^* onto $\tilde{\mathcal{L}}$ is already "close" to $\hat{P}$, we accept the model and attribute the original lack of fit to the outliers. Then the "outlier" cell counts $x(\omega_j), j = 0\ldots, \ell$ are deemed unreliable and they may be adjusted to $nP^*(\omega_j), j = 0\ldots, \ell$.

Similar techniques are applicable in the case when some cell counts are missing.

4 An iterative algorithm.

In this section an iterative algorithm to find the minimum divergence between two convex sets of distributions is presented. In this discussion the notation $x^* = \arg\min_{x\in X} f(x)$ is used to denote a member $x^* \in X$ at which the function f acheives its minimum, if such a minimum exists, otherwise arg min is undefined.

In the following lemma the sets $\mathcal{P}$ and $\mathcal{Q}$ and function $D(P\|Q)$ are completely arbitrary. In later applications $D(P\|Q)$ will be the divergence and $\mathcal{P}$ and $\mathcal{Q}$ will be convex sets of distributions on a finite set A.

Theorem 10 Let $D(P\|Q)$ be an arbitrary real-valued function defined for $P \in \mathcal{P}, Q \in \mathcal{Q}$ such that $P^* = P^*(Q) = \arg\min_{\mathcal{P}} D(P\|Q)$ exists for all $Q \in \mathcal{Q}$ and $Q^* = Q^*(P) = \arg\min_{\mathcal{Q}} D(P\|Q)$ exists for all $P \in \mathcal{P}$. Suppose further that there is a nonnegative function $\delta(P\|P')$ defined on $\mathcal{P} \times \mathcal{P}$ with the following "three-points property,"

$$\begin{aligned}&\delta(P\|P^*(Q)) + D(P^*(Q)\|Q)\\ &\quad\le D(P\|Q), \quad \forall P \in \mathcal{P}, Q \in \mathcal{Q},\end{aligned}$$

as well as the following "four-points property,"

$$\begin{aligned}&D(P'\|Q') + \delta(P'\|P)\\ &\quad\ge D(P'\|Q^*(P)), \quad \forall P, P' \in \mathcal{P}, Q' \in \mathcal{Q}.\end{aligned}$$

Let Q_0 be an arbitrary member of $\mathcal{Q}$ and recursively define

$$\begin{aligned}P_n &= \arg\min_{P\in\mathcal{P}} D(P\|Q_{n-1}),\\ Q_n &= \arg\min_{Q\in\mathcal{Q}} D(P_n\|Q).\end{aligned} \tag{8}$$

Then

$$\lim_{n\to\infty} D(P_n\|Q_n) = \inf_{P\in\mathcal{P},Q\in\mathcal{Q}} D(P\|Q).$$

If, in addition, (i) $\min_{Q\in\mathcal{Q}} D(P\|Q)$ is continuous in P, (ii) $\mathcal{P}$ is compact, and (iii) $\delta(P\|P_n) \to 0$ iff $P_n \to P$, then for the iteration (**??**) P_n will converge to some P^*, such that if $Q^* = \arg\min_{Q\in\mathcal{Q}} D(P^*\|Q)$ then $D(P^*\|Q^*) = \min_{P\in\mathcal{P},Q\in\mathcal{Q}} D(P\|Q)$ and, moreover, $\delta(P^*\|P_n) \downarrow 0$ and

$$D(P_n\|Q_n) - D(P^*\|Q^*) \le \delta(P^*\|P_{n-1}) - \delta(P^*\|P_n).$$

Proof. We have, by the three-points property,

$$\delta(P\|P_{n+1}) + D(P_{n+1}\|Q_n) \le D(P\|Q_n),$$

and, by the four-points property

$$D(P\|Q_n) \le D(P\|Q) + \delta(P\|P_n),$$

for all $P \in \mathcal{P}, Q \in \mathcal{Q}$. Hence

$$\delta(P\|P_{n+1}) \le D(P\|Q) - D(P_{n+1}\|Q_n) + \delta(P\|P_n) \tag{9}$$

The inequality (**??**) implies the desired basic limit result

$$\lim_{n\to\infty} D(P_n\|Q_n) = \inf_{P\in\mathcal{P},Q\in\mathcal{Q}} D(P\|Q).$$

Indeed, if this were false it would mean that there exist $P \in \mathcal{P}, Q \in \mathcal{Q}$ and $\epsilon > 0$ such that

$$\lim_{n\to\infty} D(P_n\|Q_n) = \lim_{n\to\infty} D(P_{n+1}\|Q_n) > D(P\|Q) + \epsilon.$$

Then (**??**) would give that $\delta(P\|P_{n+1}) \le \delta(P\|P_n) - \epsilon,\ n = 1, 2, \ldots$ which contradicts the assumption that δ is nonnegative.

Suppose assumptions (i)-(iii) hold. Pick a subsequence $P_{n_k} \to P^*$, as $k \to \infty$ and let $Q^* = \arg\min_{Q\in\mathcal{Q}} D(P^*\|Q)$. Our basic limit result and assumption (i) imply that (P^*, Q^*) achieves $\min_{P,Q} D(P\|Q)$. But it is easy to see that (**??**) implies that if (P, Q) achieves $\min_P \min_Q D(P\|Q)$ then $\delta(P\|P_{n+1}) \le \delta(P\|P_n)$ for every n. Thus $\delta(P^*\|P_n)$ must be nondecreasing, and, by assumption (iii), its limit must be 0. Using assumption (iii) once more, we conclude that $P_n \to P^*$. The final inquality in the statement of the theorem then follows from (**??**) by replacing (P, Q) by (P^*, Q^*). This completes the proof of the theorem.

Now we wish to apply the theorem to the case when $D(P\|Q)$ is the divergence and $\mathcal{P}$ and $\mathcal{Q}$ are convex, compact sets of nonnegative measures on A. No assumption that the measures are probability distributions is made at this point; hence, in particular, $D(P\|Q)$ may have negative values. Of course, if $\sum P(a) \ge \sum Q(a)$ then $D(P\|Q) \ge 0$. Furthermore, the quantity

$$\delta(P\|Q) = \sum_a \left[P(a)\log\frac{P(a)}{Q(a)} - (P(a) - Q(a))\log e\right],$$

is always nonnegative and vanishes iff $P = Q$. This δ satisfies assumption (iii) of the theorem as well as the three-points and four-points properties. We verify the four-points property and leave the verification of the other properties to the reader. Let $Q^* = \arg\min_{Q\in\mathcal{Q}}$, let Q' be an arbitrary member of $\mathcal{Q}$, and set $Q_t = (1-t)Q^* + tQ' \in \mathcal{Q}, 0 \le t \le 1$. Then

$$\begin{aligned}0 \le \frac{1}{t}\left[D(P\|Q_t) - D(P\|Q^*)\right] =\\ \frac{d}{dt}D(P\|Q_t)\,\big|_{t=\tilde{t}},\ 0 < \tilde{t} \le t.\end{aligned}$$

With $t \to 0$ it follows that

$$\begin{aligned}0 &\le \lim_{\tilde{t}\to 0}\sum_a P(a)\frac{(Q^*(a) - Q'(a))\log e}{(1-\tilde{t})Q^*(a) + \tilde{t}Q'(a)}\\ &= \sum_a P(a)\frac{Q^*(a) - Q'(a)}{Q^*(a)}\log e.\end{aligned} \tag{10}$$

If we then combine this with the fact that $\log t \geq (1 - 1/t)\log e)$ we obtain

$$\sum_a P'(a)\log\frac{P'(a)Q^*(a)}{Q'(a)P(a)} - (P'(a) - P(a))\log e \geq 0,$$

which is just a rewritten version of the four-points property.

Remark 5 Suppose we are given a convex family $\mathcal{F}$ of random variables defined on a finite probability space (Ω, P) and let X^* be a member of the family for which $E(\log X)$ is maximal. Then, letting X and X^* play the role of Q' and Q^*, respectively, the inequality (**??**) gives that

$$E\left(\frac{X^* - X}{X}\right) \geq 0, \text{ i. e., } E\left(\frac{X^*}{X}\right) \geq 1, \forall X \in \mathcal{F}.$$

The finiteness assumption is not really needed here, for all that is needed is that $\max E(\log X)$ is attained. This is known as *Cover's inequality.*

The result of Theorem **??** can be applied to the problem of minimizing divergence from a set of distributions that is the image of a "nice" set in some other space. Let $T\colon A \mapsto B$ be a given mapping and for any P on A write P^T for its image on B, that is, $P^T(b) = \sum_{a:Ta=b} P(a)$.

Problem 1. Given a set $\tilde{\mathcal{Q}} = \{Q^T\colon Q \in \mathcal{Q}\}$ of distributions on B for some set $\mathcal{Q}$ of distributions on A, minimize $D(\tilde{P}\|\tilde{Q})$, subject to $\tilde{Q} \in \tilde{\mathcal{Q}}$ for some given $\tilde{P}$ on B. Here it is assumed that to any $P \in \mathcal{P}$, a $Q \in \mathcal{Q}$ minimizing $D(P\|Q)$ can "easily" be found.

Problem 2. The same but with the role of P and Q interchanged.

The first problem is relevant for maximum likelihood estimation based on partially observed data, when estimation from the full data would be "easy." The two problems can be solved in similar ways; we concentrate on the first one.

Let $\mathcal{P}$ be the set of all P on A such that $P^T = \tilde{P}$. Here $\tilde{P}$ and the elements of $\tilde{\mathcal{Q}}$ are not necessarily probability distributions; indeed, either $\sum \tilde{P}(b)$ or $\sum \tilde{Q}(b)$ maybe less than, equal to, or greater than 1. Nevertheless the partitioning inequality gives $D(P\|Q) \geq D(P^T\|Q^T)$ with equality iff

$$\frac{P(a)}{Q(a)} = \frac{P^T(Ta)}{Q^T(Ta)}, \forall a \in A.$$

Hence $P^* \in \mathcal{P},\ Q^* \in \mathcal{Q}$ achieve $\min_{P,Q} D(P\|Q)$ iff $\tilde{Q}^* = Q^{*T}$ achieves $\min_{\tilde{Q}} D(\tilde{P}\|\tilde{Q})$.

Such (P^*, Q^*) can be achieved using Theorem **??**. Indeed, to $Q_{n-1} \in \mathcal{Q}$ we can find $P_n \in \mathcal{P}$ minimizing $D(P\|Q_{n-1})$ for $P \in \mathcal{P}$ merely by letting

$$P_n(a) = Q_{n-1}(a)\frac{\tilde{P}(Ta)}{Q^T_{n-1}(Ta)},$$

for by definition $P^T = \tilde{P}$, if $P \in \mathcal{P}$. The alternate step, finding $Q_n \in \mathcal{Q}$ minimizing $D(P_n\|Q)$ is 'easily" found, by assumption.

Now we apply the preceeding dicsussion to a mixture distribution problem. Let $\tilde{Q}$ be the set of all $\tilde{Q}$ of the form $\tilde{Q}(b) = \sum_{i=1}^k c_i\mu_i(b)$, where $c_i \geq 0, \sum c_i = 1$, and $\mu_i(b)$ are arbitrary nonnegative measures.

Goal: Find $(c_1^*, \ldots, c_k^*)$ achieving $\min_{\tilde{Q}} D(\tilde{P}\|\tilde{Q})$, for a given $\tilde{P}$.

Solution. Let A be the set of all pairs (i, b), $1 \leq i \leq k$, $b \in B$, and let $T(i, b) = b$. Define $\mathcal{P}$ and $\mathcal{Q}$ as above and apply the iteration scheme. Thus

$$\begin{aligned}\mathcal{P} &= \{P\colon \sum_{i=1}^k P(i,b) = \tilde{P}(b)\},\\ \mathcal{Q} &= \{Q\colon Q(i,b) = c_i\mu_i(b)\}.\end{aligned}$$

Start with an arbitrary $(c_1^0, \ldots, c_k^0)$ with *positive* components that sum to 1; this defines $Q_0(i, b) = c_i^0\mu_i(b)$. If $Q_{n-1}(i, b) = c_i^{n-1}\mu_i(b)$ is already defined let P_n be determined as above, that is,

$$\begin{aligned}P_n(i,b) &= Q_{n-1}(i,b)\frac{\tilde{P}(b)}{\tilde{Q}_{n-1}(b)}\\ &= c_i^{n-1}\mu_i(b)\frac{\tilde{P}(b)}{\sum_j c_j^{n-1}\mu_j(b)}.\end{aligned}$$

The next step is to find $Q_n \in \mathcal{Q}$ minimizing $D(P_n\|Q)$. To do this put $P_n(i) = \sum_b P_n(i, b)$, $P_n(b|i) = P_n(i, b)/P_n(i)$ and use the relation $Q(i, b) = c_i\mu_i(b)$ to write

$$D(P_n\|Q) = \sum_{i=1}^k \sum_b P_n(i,b)\log\frac{P_n(i,b)}{Q(i,b)}$$

in the form

$$D(P_n\|Q) = \sum_{i,b} P_n(i)P_n(b|i)\left[\log\frac{P_n(i)}{c_i} + \log\frac{P_n(b|i)}{\mu_i(b)}\right]. \tag{11}$$

Note that $\sum_i P_n(i) = \sum_b \tilde{P}_n(b)$, and hence $D(P_n\|Q)$ is minimized if in (**??**) we set $c_i = P_n(i)/\sum_b \tilde{P}(b)$ (using

the fact that $P_n(b|i)$ is a probability distribution for fixed i.) Thus the recursion for c_i^n will be

$$c_i^n = c_i^{n-1} \left[\frac{\sum_b \frac{\tilde{P}(b)\mu_i(b)}{\sum_j c_j^{n-1}\mu_j(b)}}{\sum_b \tilde{P}(b)} \right],$$

and by our general theorem, $c_i^n \to c_i^*$ achieving $\min_{\tilde{Q}} D(\tilde{P}\|\tilde{Q})$.

Remark 6 The finiteness of B is not essential for the convergence of this iteration. In particular, using the remark with Cover's inequality, Remark **??**, for positive valued random variables $X_1, \ldots, X_k$, the weights c_i^* maximizing $E(\log \sum_i c_i X_i)$ can be found by the same iteration, i. e.,

$$c_i^n = c_i^{n-1} E\left(\frac{X_i}{\sum_j c_j^{n-1} X_j} \right).$$

This is Cover's portfolio algorithm.

Remark 7 The "decomposition of mixtures" algorithm can be used also if the individual μ_i's depend on some parameter to be estimated, i. e., when

$$\tilde{\mathcal{Q}} = \left\{ \tilde{Q} \colon \tilde{Q}(b) = \sum_i c_i \mu(b|\theta_i) \right\}.$$

Then, from (**??**), θ_i^n is chosen to minimize the divergence

$$\sum_b P_n(b|i) \log \frac{P_n(y|i)}{\mu_i(y|\theta)}.$$

Unfortunately, the general theorem is not applicable to this case, because $\tilde{\mathcal{Q}}$ and $\mathcal{Q}$ are not convex. Indeed, the iteration may get stuck at a local mimimum and fail to find the global one.

5 Redundancy.

This and the next two sections are concerned with measuring the performance of codes. The symbol C_n will denote a binary prefix n-code with length function $L = L(C_n, n)$. The (pointwise) redundancy $R = R_P(C_n, n)$ of the code C_n relative to a distribution P on A^n is defined by

$$R(x_1^n) = L(x_1^n) - \log \frac{1}{P(x_1^n)}.$$

The expected redundancy is

$$\bar{R} = E(R) = \sum_{x_1^n} L(x_1^n) P(x_1^n) - \sum_{x_1^n} P(x_1^n) \log \frac{1}{P(x_1^n)}.$$

The Shannon code determined by the length function $L(x_1^n) = \lceil -\log P(x_1^n) \rceil$ produces essentially zero redundancy, and, is almost the optimal code for P in that it produces expected coding length within 1 bit of the minimal expected coding length. Thus, in general, redundancy gives an approximate measure of the cost in using the code C_n on P-sequences, rather than the optimal code.

Note that the expected redundancy $E(R) = E(L) - H(P)$ is always nonnegative, but the pointwise redundancy $R(x_1^n)$ can take negative values. We will show that for random processes the pointwise redundancy is essentially nonnegative. A random process is an infinite sequence $X_1, X_2, \ldots$ of A-valued random variables defined on probability space (Ω, P^*). The Kolmogorov representation of a process produces the measure P on the space A^∞ of infinite sequences drawn from A, which is defined by requiring that the value of P on cylinder sets $[a_1^n] = \{x \in A^\infty \colon x_1^n = a_1^n\}$ be given by the formula

$$P([a_1^n]) = \text{ Prob}\left(X_i = a_i, : 1 \le i \le n\right).$$

If P is the Kolmogorov measure determined by a process we shall write P^n for the measure on A^n determined by $P^n(a_1^n) = P([a_1^n])$. (In cases where n is clear from the context we write P in place of P^n.) Note that a process defines a sequence of distributions P^n, where P^n is defined on A^n. The key difference between the concept of process and the general concept of sequences $\{P_n\}$ of distributions is that P^{n+1} is required to be related to P^n by the (Kolmogorov consistency) formula

$$P^n(x_1^n) = \sum_{x_{n+1}} P^{n+1}(x_1^{n+1}).$$

In the remainder of this section, P will denote the Kolomogorov measure of a random process $\{X_n\}$ and

C_n will denote a binary prefix n-code, for $n = 1, 2, \ldots$. The word code will mean either the sequence $\{C_n\}$ or one member C_n of this sequence; the context will make clear which possiblity is being used. Our first result expresses the idea that for random processes the pointwise redundancy is essentially nonnegative, in that it is very unlikely to asymptotically take large negative values.

Theorem 11 Let $\{c_n\}$ be a sequence of positive numbers satisfying $\sum 2^{-c_n} < \infty$. Then $R(x_1^n) \geq -c_n$, eventually almost surely.

Proof. Let

$$A_n(c) = \{x_1^n : R(x_1^n) < -c\} = \{x_1^n : 2^{L(x_1^n)} P(x_1^n) < 2^{-c}\}.$$

Then

$$\begin{aligned} P(A_n(c)) &= \sum_{x_1^n \in A_n(c)} P(x_1^n) \\ &< 2^{-c} \sum_{x_1^n \in A_n(c)} 2^{-L(x_1^n)} \leq 2^{-c}, \end{aligned}$$

where we used the Kraft inequality. Hence

$$\sum_{n=1}^{\infty} \text{Prob}\,(R(X_1^n) < -c_n) = \sum_{n=1}^{\infty} P(A_n(c_n)) \leq \sum_{n=1}^{\infty} 2^{-c_n} < \infty.$$

The theorem now follows from the Borel-Cantelli principle.

A sharper lower bound can be obtained for the case when there is a process Q such that each C_n is a Shannon code for Q^n, or in the case when the sequence of codes $\{C_n\}$ satisfies the *strong prefix property*, that is, for $m \neq n$ the code word for x_1^m is not a prefix of the code word for x_1^n unless $m \leq n$ and x_1^m is a prefix of x_1^n. We state this as a corollary as its proof is a modification of the preceeding proof.

Corollary 3 For the Shannon code with respect to a process Q, or for a strongly prefix code, the pointwise redundancy $R(x_1^n)$ is bounded below by a random variable and $E(\inf_n R(x_1^n)) > -\log e$.

Proof. Let

$$B_n(c) = \{x_1^n : R(x_1^n) < -c,\ R(x_1^k) \geq -c, k < n\}.$$

As in the proof of the theorem,

$$P(B_n(c)) < 2^{-c} \sum_{x_1^n \in B_n(c)} 2^{-L(x_1^n)},$$

and it is sufficient to show that $\sum_{n=1}^{\infty} \sum_{x_1^n \in B_n(c)} 2^{-L(x_1^n)} \leq 1$.

If the code is a strong prefix code then

$$\sum_{n=1}^{\infty} \sum_{x_1^n \in B_n(c)} 2^{-L(x_1^n)} \leq 1,$$

and hence we are done. If the code is a Shannon code for a process Q then

$$\sum_{x_1^n \in B_n(c)} 2^{-L(x_1^n)} \leq \sum_{x_1^n \in B_n(c)} Q(x_1^n) = Q(\tilde{B}_n(c)),$$

where $\tilde{B}_n(c)$ is the union of the $[x_1^n]$ for which $x_1^n \in B_n(c)$. Since these sets are disjoint, the sum of their Q-measures cannot exceed 1 and we again reach the desired result that $\sum_{n=1}^{\infty} \sum_{x_1^n \in B_n(c)} 2^{-L(x_1^n)} \leq 1$. This completes the proof of the corollary.

If the sequences to be encoded are sample paths from some known random process P then we cannot do significantly better (in the sense of minimizing expected redundancy) than we can by using the Shannon code, which produces expected redundancy of at most 1. In many typical situations, however, the process P is unknown, although it may be known to belong to some parametric family. In such cases it is difficult to design codes for which the redundancy stays bounded. The following result shows that if the code is a Shannon code for some Q then the redundancy will indeed be unbounded, unless Q is already very nearly the same as P.

Theorem 12 If Q is singular with respect to P then the P-redundancy of the Shannon code with respect to Q goes to infinity with probability 1.

Proof. The redundancy equals $\log(P(x_1^n)/(Q(x_1^n))$, up to 1 bit, hence it suffices to show that $Z_n = Q(x_1^n)/P(x_1^n)$ goes to 0, with probability 1. Towards this end, let $\mathcal{F}_n$ be the smallest σ-algebra for which the sequences x_1^n are measurable, that is, the σ-algebra generated by the cylinder sets $[x_1^n]$, $x_1^n \in A^n$. Then $\{Z_n\}$ is a martingale with respect to the increasing sequence $\{\mathcal{F}_n\}$ and therefore converges almost surely to some random variable Z. It suffices to show that $Z = 0$.

Since Q is assumed to be singular with respect to P there is a measurable set $\tilde{A} \subset A^{\infty}$ such that $P(\tilde{A}) = 1$, $Q(\tilde{A}) = 0$. Let μ by the measure defined by

$$\mu(B) = P(B) + Q(B) + \int_B Z\, dP.$$

Since $\cup_n \mathcal{F}_n$ generates the entire σ-algebra, for every $\epsilon > 0$ there exists $\tilde{A}_m \in \mathcal{F}_m$, for sufficiently large m, such

that the symmetric difference between $\tilde{A}$ and $\tilde{A}_m$ has μ-measure less than ϵ. In particular,

$$P(\tilde{A}_m) > 1-\epsilon,\ Q(\tilde{A}_m) < \epsilon,\ \int_{\tilde{A}_m} Z\ dP > E(Z) - \epsilon.$$

But for $n \geq m$ the martingale property gives $\int_{\tilde{A}_m} Z_n\ dP = Q(\tilde{A}_m)$, and therefore Fatou's lemma gives

$$\int_{\tilde{A}_m} Z\ dP \leq \liminf_{n\to\infty} \int_{\tilde{A}_m} Z_n\ dP = Q(\tilde{A}_m) < \epsilon.$$

It follows that $E(Z) < 2\epsilon$ and hence that $E(Z) = 0$. Since $Z \geq 0$, we must have $Z = 0$, with probability 1, as claimed. This completes the proof of the theorem.

Good codes are those for which the P-redundancy grows slowly as $n \to \infty$. The following theorem gives a condition that guarantees the existence of such codes, under some restrictions about the process P. In this and later results, the limiting divergence-rate for processes is defined by

$$D^\infty(P\|Q) = \lim_{n\to\infty} \frac{1}{n} D(P^n\|Q^n),$$

provided this limit exists. The limit is known to exist for stationary P if Q is i.i.d. or finite-order Markov, but not necessarily otherwise.

Theorem 13 Suppose P is stationary ergodic and let Q be a mixture of stationary ergodic distributions, $Q = \int U\nu(dU)$, such that for every $\epsilon > 0$ the set of all finite-order Markov measures U with $D^\infty(P\|U) < \epsilon$ has positive ν-measure. Then for the Shannon code with respect to Q, the redundancy satisfies $R(x_1^n)/n \to 0$, almost surely.

Proof. We have to prove that for every $\epsilon > 0$, $\log(P(x_1^n)/Q(x_1^n)) < \epsilon n$, eventually almost surely, or, equivalently,

$$P(x_1^n) < 2^{\epsilon n} Q(x_1^n),\ \text{eventually a.s.}$$

Let N_ϵ be the set of finite-order Markov measures U for which $D^\infty(P\|U) < \epsilon$ and note that

$$Q(x_1^n) = \int U(x_1^n)\nu(dU) \geq \int_{N_\epsilon} U(x_1^n)\nu(dU),$$

so that

$$\frac{2^{\epsilon n} Q(x_1^n)}{P(x_1^n)} \geq \int_{N_\epsilon} \frac{2^{\epsilon n} U(x_1^n)}{P(x_1^n)} \nu(dU) \geq \int_{N_\epsilon} 2^{n\left(\epsilon - \log \frac{P(x_1^n)}{U(x_1^n)}\right)} \nu(dU). \tag{12}$$

The entropy theorem implies that

$$\frac{1}{n} \log \frac{P(x_1^n)}{U(x_1^n)} \to D^\infty(P\|U) < \epsilon,\ P \in N_\epsilon, \tag{13}$$

for P-almost all infinite sequences (the exceptional set may depend on U.) This means that the set of all pairs (x, U), where $x \in A^\infty$, for which (**??**) does not hold, has $P \times \nu$-measure 0; this in turn implies that for P-almost all x, the set of $U's$ not satisfying (**??**) has ν-measure 0 (in both cases, by Fubini's theorem.)

Thus, for P-almost all x the integrand in (**??**) goes to infinity for ν-almost all $P \in N_\infty$. It follows by Fatou's lemma that the integral itself goes to $+\infty$, which completes the proof of the theorem.

An important class of examples of codes that satisfy the hypotheses of the preceeding theorem are obtained as follows. Let Γ be a given (countable) list of stationary ergodic distributions, and let each $U \in \Gamma$ be assigned a "description length" $L(U)$, subject to the Kraft inequality, $\sum_{U\in\Gamma} 2^{-L(U)} \leq 1$. Then x_1^n can be encoded by a prefix code of length

$$\min_{U\in\Gamma} \left[L(U) + \log \frac{1}{U(x_1^n)} \right];$$

namely, choose $U \in \Gamma$ achieving this minimum, encode x_1^n by the Shannon code with respect to U, and add a preample of length $L(U)$ to identify U (here, the 1 bit error from dropping the upper integer part symbol is disregarded.) Let us call this the code generated by the list Γ.

Theorem 14 If to any $\epsilon > 0$ there is some finite-order Markov code in the list Γ with $D^\infty(P\|U) < \epsilon$, then the redundancy of the code generated by Γ satisfies $R(x_1^n)/n \to 0$, almost surely.

Proof. Set $Q = \sum_{U\in\Gamma} 2^{-2L(U)} U$; then Q satisfies the hypotheses of Theorem **??**, hence

$$\frac{1}{n} \log \frac{P(x_1^n)}{U(x_1^n)} \to 0, \quad \text{a.s.} \tag{14}$$

Now we want to show that $R(x_1^n)/n \to 0$, a.s., where R is the redundancy of the code defined by the list Γ. Towards this end, note that the condition $\sum_{U\in\Gamma} 2^{-L(U)} \leq 1$ implies that

$$Q(x_1^n) \leq \max_{U\in\Gamma} 2^{-L(U)} U(x_1^n),$$

from which it follows that

$$\log \frac{1}{Q(x_1^n)} \geq \min_{U\in\Gamma} \left[L(U) + \log \frac{1}{U(x_1^n)} \right],$$

which implies that $R(x_1^n) \le \log(P(x_1^n)/Q(x_1^n))$. This, combined with (**??**) implies our desired result that $R(x_1^n)/n \to 0$, a.s. This completes the proof of the theorem.

The following principle, called the *minimum description length (MDL) principle* has been suggested by Rissanen.

> *Principle.* The statistical information in data is best extracted when a possibly short description of the data is found. The distribution inferred from the data is the one that leads to the shortest description, taking into account that the inferred distribution *itself must be described.*

Let Γ be a given finite or countably infinite list of stationary ergodic processes on the space A^∞. Let to each $U \in \Gamma$ a codeword of length $L(U)$ be assigned as a description of U; these lengths must satisfy the Kraft inequality. Then, given a sample x_1^n, the MDL estimate $\hat{P}_n$ of the unknown distribution P is $\hat{P}_n = U$, where U achieves $\min_{U\in\Gamma}[L(U) - \log U(x_1^n)]$.

Theorem 15 If $P \in \Gamma$ then $\hat{P}_n = P$, eventually almost surely.

Proof. Let $Q = \sum_{U\in\Gamma-\{P\}} 2^{-L(U)}U$, and note that

$$Q(x_1^n) \ge \max_{U\in\Gamma-\{P\}} 2^{-L(U)}U(x_1^n),$$

that is,

$$\log\frac{1}{Q(x_1^n)} \le \min_{U\in\Gamma-\{P\}}\left[L(U) + \log\frac{1}{U(x_1^n)}\right]. \qquad (15)$$

Now, Q is singular with respect to P, since each stationary, ergodic $U \ne P$ is singular with respect to the stationary, ergodic process P, hence by Theorem **??** the redundancy of the Shannon code with respect to Q goes to $+\infty$, that is,

$$\log\frac{1}{Q(x_1^n)} - \log\frac{1}{P(x_1^n)} \to \infty, \text{ a.s.}$$

Using the bound (**??**) we therefore have

$$\min_{U\in\Gamma-\{P\}}\left[L(U) + \log\frac{1}{U(x_1^n)}\right] - \log\frac{1}{P(x_1^n)} \to \infty, \text{ a.s.},$$

hence, for sufficiently large n

$$\min_{U\in\Gamma-\{P\}}\left[L(U) + \log\frac{1}{U(x_1^n)}\right] > \log\frac{1}{P(x_1^n)} + L(P).$$

The preceeding inequality implies that $\hat{P}_n = P$ and completes the proof of the theorem.

Now let us be given a finite or countable list of parametric families of (stationary, ergodic) processes $\{P_\theta\colon \theta \in \Theta_\gamma$, where $\gamma \in \Gamma$, and to each family on the list, i. e., to each $\gamma \in \Gamma$ suppose there is assigned a codeword of length $L(\gamma)$ describing this family, such that the Kraft inequality holds. Further, let on each parameter set Θ_γ be given a "prior" ν_γ, i. e., ν_γ is a probability measure on Θ_γ. We also assume that the mixture distributions

$$Q_\gamma = \int_{\Theta_\gamma} P_\theta \nu_\gamma(d\theta),\ \gamma \in \Gamma$$

are mutually singular. (In particular, these mean that the families $\{P_\theta\colon \theta \in \Theta_\gamma$ are essentially *disjoint.*)

Theorem 16 There exists subsets $\tilde{\Theta}_\gamma \subset \Theta_\gamma$ of full measure 1, such that if $P \in \tilde{\Theta}_{\gamma^*}$, for some $\gamma^* \in \Gamma$, then

$$\min_{\gamma\in\Gamma}\left[L(\gamma) + \log\frac{1}{Q_\gamma(x_1^n)}\right]$$

is attained for $\gamma = \gamma^*$, eventually almost surely.

Proof. In other words, the family containing the true distribution will be found with probability 1, unless P is in a subset of this family having ν_γ-measure 0.

Exactly as in the proof of the preceeding theorem (replacing U by Q_γ and $L(U)$ by $L(\gamma)$) we obtain that for sufficiently large n,

$$\min_{\gamma\in\Gamma}\left[L(\gamma) + \log\frac{1}{Q_\gamma(x_1^n)}\right]$$

will be attained for $\gamma = \gamma^*$, with Q_{γ^*}-probability 1. Let F be the set of all $x \in A^\infty$ for which this "almost sure" statement is true, so that $Q_{\gamma^*}(F^c) = 0$. Since by definition

$$Q_{\gamma^*}(F^c) = \int_{\Theta_{\gamma^*}} P_\theta(F^c)\nu_{\gamma^*}(d\theta),$$

it follows that $\nu_{\gamma^*}(\{\theta\colon P_\theta(F^c) > 0\}) = 0$ and we can take

$$\tilde{\Theta}_{\gamma^*} = \Theta_{\gamma^*} - \{\theta\colon P_\theta(F^c) > 0\}.$$

This completes the proof of the theorem.

Remark 8 The hypotheses of Theorem **??** are fulfilled, in particular, when the parameter sets Θ_γ are subsets of Euclidean spaces of different dimensions and ν_γ is absolutely continuous with respect to the Lebesgue measure for the corresponding dimension.

6 Redundancy bounds.

Some techniques for obtaining bounds on redundancy for i.i.d processes will be discussed in this section. Consider the i.i.d. process with alphabet $A = \{1, \ldots, k\}$ with distribution P. We then have

$$P(x_1^n) = \prod_{i=1}^{k} P(i)^{n_i},$$

where n_i is the number of times i occurs in x_1^n. This probability is maximum if $P(i) = n_i/n$, hence the maximum likelihood estimate is given by

$$P_{\mathrm{ML}}(x_1^n) = \prod_{i=1}^{k} \left(\frac{n_i}{n}\right)^{n_i}.$$

When encoding with respect to an auxiliary distribution Q, the redundancy satisfies (disregarding at most 1 bit) the following simple bound

$$R(x_1^n) = \log \frac{P(x_1^n)}{Q(x_1^n)} \leq \log \frac{P_{\mathrm{ML}}(x_1^n)}{Q(x_1^n)}. \qquad (16)$$

Let us take for Q the mixture distribution $Q(x_1^n) = \int U(x_1^n)\nu(p)\, dp$, with a Dirichlet prior having density

$$\nu(p) = \frac{\Gamma\left(\sum_{i=1}^k \alpha_i + k\right)}{\prod_{i=1}^k \Gamma(\alpha_i + 1)} \prod_{i=1}^{k} p_i^{\alpha_i},\ p = (p_1, \ldots, p_k).$$

For $\alpha_1 = \ldots = \alpha_k = -1/2$ we will get a sharp upper bound on the redundancy (**??**), a bound not depending on the true distribution P nor x_1^n. Before we state and derive this bound we obtain a representation for Q that will be useful in constructing the Shannon code for Q..

For a Dirichlet prior with arbitrary $\alpha_i > -1, \forall i$, we have

$$\begin{aligned} Q(x_1^n) &= \int U(x_1^n)\nu(p)\, dp = \\ &= \int \prod_{i=1}^{k} p_i^{n_i+\alpha_i} dp \cdot \frac{\Gamma\left(\sum_{i=1}^k \alpha_i + k\right)}{\prod_{i=1}^k \Gamma(\alpha_i + 1)} \\ &= \frac{\Gamma\left(\sum_{i=1}^k \alpha_i + k\right)}{\Gamma\left(n + \sum \alpha_i + k\right)} \prod_{i+1}^{k} \frac{\Gamma(n_i + \alpha_i + 1)}{\Gamma(\alpha_i + 1)}. \end{aligned}$$

Using the functional equation $\Gamma(x+1) = x\Gamma(x)$ we see that $Q(x_1^n)$ is given by the ratio

$$\frac{\prod_{i=1}^k [(n_i + \alpha_i)(n_i - 1 + \alpha_i) \ldots (1 + \alpha_i]}{(n - 1 + \sum \alpha_i + k)(n - 2 + \sum \alpha_i + k) \ldots (\sum \alpha_i + k)}$$

or, equivalently,

$$Q(x_1^n) = \prod_{j=1}^{n} \frac{n(x_j|x_1^{j-1}) + 1 + \alpha_{x_j}}{j - 1 + \sum \alpha_i + k}. \qquad (17)$$

where $n(x_j|x_1^{j-1})$ is the number of occurences of the symbol x_j in the "past" x_1^{j-1}.

Theorem 17 If Q is defined by (**??**) with $\alpha_i = -1/2, \forall i$, the redundancy *always* satisfies

$$\begin{aligned} R(x_1^n) &\leq \log \frac{\Gamma(n + \frac{k}{2})\Gamma(\frac{1}{2})}{\Gamma(n + \frac{1}{2})\Gamma(\frac{k}{2})} \leq \\ &\leq \frac{k-1}{2} \log n - \log \frac{\Gamma(k/2)}{\Gamma(1/2)} + \epsilon_n \end{aligned}$$

where $\epsilon_n \to 0$ as $n \to \infty$.

Proof. The second inequality is a simple consequence of Stirling's formula for the Γ-function, so it is enough to prove the first inequality.

For $\alpha_i \equiv -1/2$ we have

$$\begin{aligned} Q(x_1^n) &= \frac{\Gamma(\frac{k}{2})}{\Gamma(n + \frac{k}{2})} \prod_{i=1}^{k} \frac{\Gamma(n_i + \frac{1}{2})}{\Gamma(\frac{1}{2})} = \\ &= \frac{\prod_{i=1}^k \left[(n_i - \frac{1}{2})(n_i - \frac{3}{2}) \cdots \frac{1}{2}\right]}{(n - 1 + \frac{k}{2})(n - 2 + \frac{k}{2}) \cdots \frac{k}{2}} \end{aligned} \qquad (18)$$

Note that, in particular, if x_1^n consists of identical symbols, say, $x_i \equiv a$, then

$$Q(x_1^n) = \frac{\Gamma(\frac{k}{2})\Gamma(n + \frac{1}{2})}{\Gamma(n + \frac{k}{2})\Gamma(\frac{1}{2})};$$

hence to prove Theorem **??** it is enough to show that $R(x_1^n) \leq \log(1/Q(x_1^n))$. The simple upper bound (**??**) then tells us that it is enough to show that

$$P_{\mathrm{ML}}(x_1^n) \leq \prod_{i=1}^{k} \left(\frac{n_i}{n}\right)^{n_i} \leq \frac{Q(x_1^n)}{Q(\tilde{x}_1^n)},$$

where $\tilde{x}_i \equiv a$. The identity (**??**) can then be used to see that it is enough to prove that

$$\prod_{i=1}^{k} \left(\frac{n_i}{n}\right)^{n_i} \leq \frac{\prod_{i=1}^k \left[(n_i - \frac{1}{2})(n_i - \frac{3}{2}) \cdots \frac{1}{2}\right]}{(n - \frac{1}{2})(n - \frac{3}{2}) \cdots \frac{1}{2}},$$

which can be converted to

$$\prod_{i=1}^{k} \left(\frac{n_i}{n}\right)^{n_i} \leq \frac{\prod_{i=1}^k [2n_i(2n_i - 1) \cdots (n_i + 1)]}{2n(2n - 1) \cdots (n + 1)} \qquad (19)$$

since

$$(n-\frac{1}{2})(n-\frac{3}{2})\cdots\frac{1}{2} = \frac{1}{n!}\left[n(n-\frac{1}{2})\cdots\frac{1}{2}\right] = \frac{(2n)!}{2^{2n}n!} = \frac{2n(2n-1)\cdots(n+1)}{2^{2n}}.$$

At last we have arrived at the assertion we shall prove, namely, (**??**). This will be proved if we show that it is possible to assign to each $\ell = 1,\ldots,n$ in a one-to-one mannner, a pair $(i,j),\ 1\le i\le k,\ 1\le j\le n$, such that

$$\frac{n_i}{n} \le \frac{n_i+j}{n+\ell} \tag{20}$$

Now, for any given ℓ and i, (**??**) holds iff $j \ge n_i\ell/n$. Hence the number of those $1\le j\le n_i$ that satisfy (**??**) is greater than $n_i - n_i\ell/n$, and the total number of pairs $(i,j),\ 1\le i\le k,\ 1\le j\le n$, satisfying (**??**) is greater than

$$\sum_{i=1}^{k}\left(n_i - \frac{n_i}{n}\ell\right) = n-\ell.$$

It follows that if we assign to $\ell = n$ any (i,j) satisfying (**??**) (i. e., i may be chosen arbitrarily and $j = n_i$), then recursively assign to each $\ell = n-1, n-2$, etc., a pair (i,j) satisfying (**??**) that were not assigned previously, we never get stuck; at each step there will be at least one "free" pair (i,j) (because the total number of pairs (i,j) satisfying (**??**) is greater than $n-\ell$, the number of pairs already assigned.) This completes the proof of the theorem.

Our next goal is to show that the result of the preceeding theorem is "best possible," even if we don't insist on a uniformly small redundancy (i. e., on a bound valid for every x_1^n), but want only the average redundancy $E(R)$ to be small.

Consider any prefix code. Without loss of generality (for the purpose of bounding the redundancy) we may assume that it satisfies the Kraft inequality with the equality sign, and therefore that is is a Shannon code with respect to some Q (not necessarily of mixture type.) Then

$$E(R(X_1^n)) = E\log\frac{P(X_1^n)}{Q(X_1^n)} = D(P^n\|Q).$$

Since P is unknown, we want to select Q in such a way that no matter what P is the average redundancy will be small, that is, we want Q to minimize

$$\sup_P E_P(R_P(X_1^n)) = \sup_P D(P^n\|Q).$$

Suppose we choose P at random with prior distribution ν; then the observation of x_1^n provides information about the unknown P, measured by the mutual information

$$I(\nu) = H(Q_\nu) - \int H(P^n)\nu(dP) = H(Q_\nu) - n\int H(P)\nu(dP),$$

where Q_ν is the mixture distribution, $Q_\nu = \int P(x_1^n)\nu(dP)$. Even though this mutual information appears to be unrelated to the previous average redundancy, the remarkable fact is that

$$\inf_Q\sup_P D(P^n\|Q) = \sup_\nu I(\nu).$$

Indeed, the following lemma holds in general.

Lemma 5 Consider any noisy channel with input alphabet $U = \{1,\ldots,\ell\}$ and output alphabet $V = \{1,\ldots,m\}$, given by the probability distributions P_i on V governing the output if the input is $i,\ i = 1,\ldots,\ell$. For any input distribution π, let Q_π denote the output distribution and let

$$I(\pi) = \sum_{i,j}\pi(i)P_i(j)\log\frac{P_i(j)}{Q_\pi(j)} = \sum_i \pi(i)D(P_i\|Q_\pi)$$

be the mutual information between input and output. Then

$$\max_\pi I(\pi) = \min_Q\max_{1\le i\le\ell} D(P_i\|Q).$$

Proof. The left-side is known as the channel capacity. The lemma states that it equals the "radius" of the smallest "divergence ball" that contains all the P_i's. To establish this relation first note that for any distribution Q on V,

$$I(\pi) = \sum_{i=1}^{\ell}\sum_{j=1}^{m}\pi(i)P_i(j)\left[\log\frac{P_i(j)}{Q(j)} + \log\frac{Q(j)}{Q_\pi(j)}\right] = \sum_{i=1}^{\ell}\pi(i)D(P_i\|Q) - D(Q_\pi\|Q).$$

This identity shows that for any fixed π

$$\min_Q\sum_{i=1}^{\ell}\pi(i)D(P_i\|Q) = I(\pi),$$

and hence

$$\max_\pi I(\pi) = \max_\pi\min_Q\sum_{i=1}^{\ell}\pi(i)D(P_i\|Q).$$

The *minimax theorem* asserts that if $f(x,y)$ is a continous function of two variables ranging over convex, compact sets, which is concave in x and convex in y then

$$\max_x \min_y f(x,y) = \min_y \max_x f(x,y).$$

In our case the theorem can be applied and we get

$$\max_\pi I(\pi) = \min_Q \max_\pi \sum_{i=1}^{\ell} \pi(i) D(P_i \| Q).$$

Since the inner maximum is clearly equal to the maximum of $D(P_i\|Q)$ over i, the proof of the lemma is completed.

The lemma is valid also in the case when the input alphabet X is infinite, providing the maximum with respect ot π is replaced by "supremum." We omit the proof of this more general case, even though this is what we need for lower bounding average redundancy. Indeed, using this result, we can state

Theorem 18 For any prefix code, the supremum for P of the expected redundancy is lower bounded by $I(\nu) = H(Q_\nu) - n\int H(P)\nu(dP)$, where ν is an arbitrarily chosen prior distribution.

Of course, the best bound is the supremum of $I(\nu)$, which is the channel capacity of the set of all possible distributions P on A considered as the input alphabet, A^n the output alphabet, and the distribution on A^n corresponding to the input P is P^n.

7 Rissanen's theorem.

Now we would like to establish the most general known lower bound on the redundancy of a prefix codes, a result due to Rissanen.

Theorem 19 Let $\{P_\theta\}_{\theta\in\Theta}$ be any family of random processes, not necessarily i.i.d., possibly not even stationary, where $\Theta \in R^k$. Suppose that for each $n \geq n_0$ there exists an estimator $\hat{\Theta}_n(x_1^n)$ with

$$E_\theta \|\hat{\theta} - \theta\|^2 \leq \frac{c(\theta)}{n}, \forall \theta \in \Theta. \tag{21}$$

Then, for every $\epsilon > 0$ there is a constant $K > 0$ such that for $n \geq n_0$ and for every probability density or mass function g we have,

$$E_\theta \log \frac{P_\theta(x_1^n)}{g(x_1^n)} \geq \frac{k}{2}\log n - K, \tag{22}$$

except possibly for a set of parameters θ of Lebesgue measure less than ϵ.

Proof. Suppose the set Θ_1 of those θ's for which (**??**) doesn't hold has Lebesgue measure at least ϵ. We will show that this supposition leads to a contradiction if K is suitably chosen. By Theorem **??** of the preceeding section, it suffices to show that for some distribution ν on Θ_1, we have $I(\nu) > (k\log n)/2 - K$. where $I(\nu)$ is the mutual information of the channel having input alphabet Θ_1 and transition probabilities P_θ. (In Theorem **??**, P_θ was i.i.d., but the proof is clearly valid in general.)

Now let c be so large that the subset Θ_2 of Θ_1 on which (**??**) holds with $c(\theta) = c$ has Lebesgue measure at least $\epsilon/2$. Let ν be the uniform distribution on Θ_2, let Z be a distribution chosen at random according to ν, and let $\hat{Z}$ be an estimator of Z such that $E\|Z - \hat{Z}\|^2 \leq c/n$. Then $I(\nu) \geq I(Z \wedge \hat{Z})$, by the data processing theorem.

Note that

$$\begin{aligned} I(Z \wedge \hat{Z}) &= H(Z) - H(Z|\hat{Z}) \\ &= H(Z) - H(Z - \hat{Z}|\hat{Z}) \\ &\geq \log\frac{e}{2} - H(Z - \hat{Z}). \end{aligned}$$

But the entropy of a k-dimensional random variable Y subject to $E(\|Y^2\|) \leq \alpha$ is maximized if its distribution is Gaussian with independent components of variance $\sigma^2 = \alpha/k$, and this maximum entropy equals $(k/2)\log(2\pi e\sigma^2)$. Applying this fact with $Y = Z - \hat{Z}, \alpha = c/n, \sigma^2 = c/(kn)$, it follows that

$$H(Z - \hat{Z}) \leq \frac{k}{2}\log(2\pi e\frac{c}{nk}) = -\frac{k}{2}\log n + B,$$

where B depends only on c and k. From this it follows that

$$I(\nu) \geq I(Z \wedge \hat{Z}) \geq \frac{k}{2}\log n - B + \log\frac{\epsilon}{2},$$

which proves Rissanen's theorem.

Corollary 4 If for some subfamily $\{P_\theta\}_{\theta\in\Theta_0}$ of a family of sources satisfying (**??**) there exist universal codes whose average redundancy grows slower than $(k/2)\log n$, i. e.,

$$\lim_{n\to\infty}\left(E_\theta R_\theta(X_1^n) - \frac{k}{2}\log n\right) = -\infty,\ \theta \in \Theta_0,$$

then, necessarily, Θ_0 has Lebesgue measure 0.

Proof. This is immediate, because, without restricting generality, any code can be supposed to be a Shannon code with respect to some distribution g.

The hypotheses of Rissanen's theorem are satisfied, in particular, if $\{P_\theta\}$ is the family of all i.i.d. distributions

on a finite alphabet $A = \{1, \ldots, k\}$. Then θ may be identified with the vector of the probabilities $(P_1, \ldots, P_k)$, and since these form a $(k-1)$-dimensional subspace we get that (**??**) holds for k replaced by $k-1$, thus proving that that universal codes constructed in the preceeding section have asymptotically optimal redundancy.

Our results extend beyond the i.i.d. case; in particular they extend to the Markov case. A Markov chain with transition matrix $P(j|i)$, $1 \le i, j \le k$, is given by the joint distributions

$$\text{Prob}\,(X_t = i_t, 0 \le t \le n) = P(i_0) \prod_{t=1}^{k} P(i_t|i_{t-1}).$$

We will suppose that the initial state i_0 is fixed, so that we can rewrite these probabilities in the form,

$$\text{Prob}\,(X_t = i_t, 0 \le t \le n) = \prod_{i=1}^{k} \prod_{j=1}^{k} P(j|i)^{n(i,j)}, \quad (23)$$

where $n(i,j)$ is the number of times the pair i, j occurs in adjacent places in x_0^n. Further, let $n(i) = \sum_j n(i,j)$ denote the number of occurences of i in the block x_0^{n-1} and note that the probability in (**??**) is maximized for $\hat{P}(j|i) = n(i,j)/n(i)$, that is

$$P_{\text{ML}}(x_0^n) = \prod_{i=1}^{k} \prod_{j=1}^{k} \hat{P}(j|i)^{n(i,j)}.$$

By analogy with the i.i.d. case we introduce the mixture distribution

$$Q(x_1^n) = \prod_{i=1}^{k} \int \prod_{j=1}^{k} P(j|i)^{n(i,j)} \nu(P(\cdot|i))dP,$$

where ν is the Dirichlet prior with $\alpha_i \equiv -1/2$. Thus $Q(x_1^n)$ is given by the product

$$\prod_{i=1}^{k} \left[\left(\prod_{j=1}^{k} \frac{\Gamma(n(i,j) + 1/2)}{\Gamma(1/2)} \right) \frac{\Gamma(k/2)}{\Gamma(n(i) + k/2)} \right],$$

which is, in turn, equal to the product

$$\prod_{i=1}^{k} \frac{\prod_{j=1}^{k} (n(i,j) - 1/2)(n(i,j) - 3/2) \ldots (1/2)}{(n(i) - 1 + k/2)(n(i) - 2 - k/2) \ldots (k/2)}.$$

The redundancy of the code based on the above auxiliary distribution can be bounded, using the corresponding i.i.d. result. It follows that

$$\begin{aligned} R(x_1^n) &= \sum_{i=1}^{k} \left[\frac{k-1}{2} \log n(i) + \text{ constant} \right] \\ &= \frac{k(k-1)}{2} \log n + \text{ constant}. \end{aligned}$$

Again, this result is asymptotically best possible (up to the constant term.) Indeed, on account of Risannen's theorem, even the average redundancy cannot be made significantly smaller than $(k(k-1)/2) \log n$ on a set of positive Lebesgue measure in the parameter space needed to describe Markov chain probabilities.

Rissanen has provided an interesting application of his theorem to a special class of processes, which we will call the chains with *finite context*. A process has finite context if there is a positive integer m and a function $f\colon A^m \mapsto S$ where S is some finite set) such that

$$\text{Prob}\,(X_t = i_t, 0 \le t \le n) = P(i_0) \prod_{t=1}^{k} P(i_t|f(i_{j-m}, \ldots, i_{j-1})),$$

where it is assumed here that $i_{-m+1}, \ldots, i_0$ is fixed. The elements of S are called "contexts" or "states" and $P(i|\ell)$ is interpreted as the "probability of the symbol i in the context ℓ." Of course, any source that is Markov of order m has finite memory, and conversely; the context idea emphasizes that the probability of occurence of the next symbol may be depend on something much simpler than the entire past of length m, namely $|S|$ may be much smaller than $|A^m| = k^m$, and it would be nice to take advantage of this fact in coding.

To obtain optimal bounds for processes with finite context we need make only a few changes in our preceeding discussion. Let us fix S and f and let $n(i,\ell), i < n, \ell \in S$ denote the number of pairs (i,ℓ) that occur among the pairs (i_t, s_{t-1}), where $s_{t-1} = f(i_{t-m}, \ldots, i_{t-1})$, for $t < n$. We then have

$$P(x_1^n) = \prod_{\ell \in S} \prod_{i=1}^{k} P(i|\ell)^{n(i,\ell)}$$

and the maximum likelihood probabilities

$$\begin{aligned} P_{ML}(x_1^n) &= \prod_{\ell \in S} \prod_{i=1}^{k} \hat{P}(i|\ell)^{n(i,\ell)}, \\ \hat{P}(i|\ell) &= n(i,\ell)/n(\ell),\ n(\ell) = \sum_i n(i,\ell). \end{aligned}$$

Again, as in the i.i.d. case, an asymptotically optimal universal code is the one based on the auxiliary mixture distribution (with the Dirichlet prior), as follows,

$$Q(x_1^n) = \prod_{\ell \in S} \int \prod_{i=1}^{k} P(i|\ell)^{n(i,\ell)} \nu(P(\cdot|\ell))dP$$

which is equal to

$$\prod_{\ell\in S}\frac{\prod_{i=1}^{k}(n(i,\ell)-1/2)(n(i,\ell)-3/2)\dots(1/2)}{(n(\ell)-1+k/2)(n(\ell)-2+k/2)\dots(k/2)},$$

and this code has redundancy

$$\begin{aligned} R(x_1^n) &\le \sum_{\ell\in S}\left[\frac{k-1}{2}\log n(\ell)+\text{ constant}\right] \\ &\le |S|\frac{k-1}{2}\log n+\text{ constant}. \end{aligned}$$

Furthermore, Rissanen's theorem implies that even the average redundancy cannot be substantially smaller the above bound, for any universal codes, except possibly for a vanishingly small set of parameter values, i. e., matrices $P(i\ell)$.

Remark 9 Before leaving this topic of redundancy bounds let us mention an aspect of our discussion which has some practical value in designing codes. In the preceeding section we derived the following formula (see (**??**)), valid for the i.i.d. case

$$Q(x_1^n)=\prod_{j=1}^{n}\frac{n(x_j|x_1^{j-1})+1+\alpha_{x_j}}{j-1+\sum\alpha_i+k}$$

where $n(x_j|x_1^{j-1})$ is the number of occurences of the symbol x_j in the "past" x_1^{j-1}. This formula suggests the conditional probabilities

$$Q(x_j|x_1^{j-1})=\frac{n(x_j|x_1^{j-1})+1+\alpha_{x_j}}{j-1+\sum\alpha_i+k}$$

The latter formula can be used as the specification of the conditional probabilities used in arithmetic coding, a (practical) sequential procedure that yields the same asymptotics as the Shannon coding procedure.

Likewise, the Markov discussion in this section leads to the conditional formula

$$Q(i_k|i_1,\dots,i_{k-1})= \frac{n_{k-1}(i,j)+1/2}{n_{k-1}(i)+k/2},\ \text{ if } i_{k-1}=i, i_k=j,$$

where $n_{k-1}(i,j)$ is the number of consecutive (i,j)'s in the sequence i_0^{k-1} and $n_{k-1}(i)=\sum_j n_{k-1}(i,j)$. These conditional probabilities are easily evaluated, because only simple updating is needed to go from $k-1$ to k; arithmetic coding can then be performed.

The corresponding finite context formula is

$$Q(i_k|i_1,\dots,i_{k-1})= \frac{n_{k-1}(i,\ell)+1/2}{n_{k-1}(\ell)+k/2},\ \text{ if } s_{k-1}=\ell, i_k=j.$$

These can then be used to do arithmetic coding in the finite context case; such coding will also yield the same asymptotics as the Shannon code. Rissanen's theorem implies that even the average redundancy can not be substantially smaller than the bound above, for any universal code, expect possibly for a vanishingly small set of parameters (i. e., matrices $P(i|\ell)$.)

8 Additions.

8.1 The scaling formula.

The scaling formula

$$P^*(a)=c_iQ(a),\ a\in B_i,\ \text{where } c_i=\frac{\alpha_i}{Q^{\mathcal{B}}(i)}. \qquad (24)$$

see (**??**) can be proved as follows. First, lumping does not increase divergence, that is,

$$D(P\|Q)\ge D(P^{\mathcal{B}}\|Q^{\mathcal{B}}).$$

The condition that $P\in\mathcal{L}$ is equivalent to the condition that $P(B_i)=\alpha_i,\ \forall i$. If $P^*(a)=\alpha_iQ(a)/Q(B_i),\ a\in B_i$ then

$$\begin{aligned} \sum_a P^*(a)\log\frac{P^*(a)}{Q(a)} &= \sum_i\sum_{a\in B_i}\frac{\alpha_iQ(a)}{Q(B_i)}\log\frac{\alpha_i}{Q(B_i)} \\ &= D(\alpha\|Q^B),\ \alpha=(\alpha_1,\dots,\alpha_k). \end{aligned}$$

Thus, if $P\in\mathcal{L}$ then

$$D(P\|Q)\ge D(P^B\|Q^B)=D(\alpha\|Q^B),$$

which establishes (**??**).

8.2 Pearson's χ^2.

The chi-square function was defined on page **??**. In the case when $P=\hat{P}$, the empirical distribution the formula can be rewritten as follows.

$$\begin{aligned} \chi^2(\hat{P},Q) &= \sum_a\frac{(\hat{P}(a)-Q(a))^2}{Q(a)} \\ &= \frac{1}{n}\sum\frac{(n\hat{P}(a)-nQ(a))^2}{nQ(a)} \\ &= \frac{1}{n}\chi^2_{k-1}, \end{aligned}$$

where $\chi^2_{k-1}=\sum\frac{(n\hat{P}(a)-nQ(a))^2}{nQ(a)}$ is Pearson's classical chi-square function. Here $n\hat{P}(a)$ gives the observed count, while $nQ(a)$ gives the expected count of the number of appearances of a.

8.3 Maximum entropy and Likelihood.

There is an important case when divergence minimization corresponds to maximum likelihood, namely, the case when the linear family contains the empirical distribution. Suppose we are given the corresponding linear and exponential families,

$$\begin{aligned}\mathcal{L} &= \{P: \sum_a P(a) f_i(a) = \alpha_i, 1 \le i \le k\} \\ \mathcal{E} &= \{P: P(a) = c(\theta) Q(a) \exp(\sum_1^k \theta_i f_i(a))\}.\end{aligned}$$

Theorem 20 *If $Q(a) > 0, \forall a$ and if the empirical distribution $\hat{P}$ belongs to $\mathcal{L}$ then the maximum likelihood estimate in $\mathcal{E}$ is the I-projection P^* of any member of $\mathcal{E}$ onto $\mathcal{L}$. Furthermore, the minimum value of $D(\hat{P}\|P)$ for $P \in \mathcal{E}$ is attained at P^*.*

Proof. Let $P^* = D(\mathcal{L}\|Q)$, so that $\mathcal{L} \cap \mathcal{E} = \{P^*\}$. If $P \in \mathcal{E}$ we can write

$$\begin{aligned}P &= cQ(a) \exp(\sum \theta_i f_i(a)), \\ P^* &= c^* Q(a) \exp(\sum \theta_i^* f_i(a)).\end{aligned}$$

Since $\sum P^*(a) f_i(a) = \alpha_i$ we have

$$0 \le D(P^*\|P) = \log c^* + \sum \theta_i^* \alpha_i - (\log c + \sum \theta_i \alpha_i),$$

so that

$$\log c^* + \sum \theta_i^* \alpha_i = \max_{P \in \mathcal{E}} (\log c + \sum \theta_i \alpha_i).$$

If $\hat{P} \in \mathcal{L}$, however, then $D(\hat{P}\|P) - D(\hat{P}\|P^*) = D(P^*\|P)$, since $\sum \hat{P}(a) f_i(a) = \alpha_i$. This proves that the minimum value of $D(\hat{P}\|P)$ for $P \in \mathcal{E}$ is attained at P^*. Furthermore, $P(x_1^n) = n \sum \hat{P}(a) \log P(a)$, so that if $P \in \mathcal{E}$ and $\hat{P} \in \mathcal{L}$ then

$$\log \frac{P^*(x_1^n)}{P(x_1^n)} = n \sum \hat{P}(a) \log \frac{P^*(a)}{P(a)} = D(P^*\|P) \ge 0,$$

so that P^* is indeed the MLE in $\mathcal{E}$.

The argument can be applied to any member of $\mathcal{E}$ in place of the given Q, since they all describe the same exponential family.

8.4 Redundancy for the LZ algorithm.

An upper bound on the reduncancy of the form $O(\log\log n / \log n)$ for the Lempel-Ziv (LZ) algorithm on the class of i.i.d. processes will now be established. Extensions of these results to the Markov and hidden Markov cases can also be obtained.

Let $c = c(x_1^n)$ be the number of commas in the LZ parsing of x_1^n. The final block, which may be empty, is coded by telling the first prior word that this block prefixes. Let $U_{LZ}(x_1^n)$ be the length of the resulting code. Each word, except the final word, can be encoded with at most $\lceil \log c \rceil$ bits to give the location of the prior occurence of all but its final symbol and $\lceil \log |A| \rceil$ to encode this final symbol. Thus we have the upper bound

$$U_{LZ}(x_1^n) \le (c+1)\lceil \log c \rceil + c \lceil \log |A| \rceil. \tag{25}$$

The next step in upper bounding the redundancy is to obtain a lower bound on $-\log P(x_1^n)$, stated here as the following lemma.

Lemma 6 *There is a positive number δ such that if P is an i.i.d. process then*

$$-\log P(x_1^n) \ge c \log c - c\delta + \frac{\log(n/c)}{n/c}.$$

Proof. Let $W = W(x_1^n)$ be the first c words in the LZ parsing of x_1^n, let $W_L = W_L(X_1^n)$ be the subset of W consisting of the words of length L, and let $c(L)$ be the cardinality of W_L. We then have

$$P(x_1^n) \le \prod_{L=1}^{L_{\max}} \prod_{w \in W_L} P(w),$$

so that

$$\begin{aligned}-\log P(x_1^n) &\ge -\sum_{L=1}^{L_{\max}} \sum_{w \in W_L} \log P(w) \\ &= -\sum_{L=1}^{L_{\max}} c(L) \sum_{w \in W_L} \frac{1}{c(L)} \log P(w) \\ &\ge -\sum_{L=1}^{L_{\max}} c(L) \log \sum_{w \in W_L} \frac{P(w)}{c(L)} \\ &\ge \sum_{L=1}^{L_{\max}} c(L) \log c(L)\end{aligned}$$

where the first inequality comes from Jensen's inequality, and the final inequality uses the fact that $\sum_{w \in W_L} P(w) \le 1$, which holds because the words in W_L are distinct and have fixed length L.

To obtain a suitable bound on $\sum c(L) \log c(L)$ set

$$\bar{L} = \frac{1}{c} \sum_{L=1}^{L_{\max}} c(L) \log c(L),$$

so that

$$\begin{aligned}
\sum c(L)\log c(L) &= -c\sum \tfrac{c(L)}{c}\log \tfrac{1}{c(L)} \\
&= -c\sum \tfrac{c(L)}{c}\log \tfrac{2^{-L/\bar L}}{c(L)} - c \\
&\overset{(a)}{\geq} -c + c\log c - c\log \textstyle\sum_1^{L_{\max}} 2^{-L/\bar L} \\
&\geq -c + c\log c - c\log \tfrac{2^{-1/\bar L}}{1-2^{-1/\bar L}} \\
&\overset{(b)}{\geq} -c + c\log c - c\log(2^{1/\bar L}-1) \\
&\overset{(c)}{\geq} -c + c\log c - c\log(\tfrac{\ln 2}{\bar L}) \\
&\geq c\log c - c\delta - c\log \tfrac{n}{c}.
\end{aligned}$$

where Jensen's inequality was used in (a) and the finite sum was replaced by the infinite sum (of a geometric series) to go to (b). The Taylor expansion of 2^x was used to obtain (c), while the final line used $\delta = 1 - \log(\ln 2)$ and the fact that $\bar L \leq n/c$. This completes the proof of Lemma **??**.

Taking the difference between the upper bound on the code length, (**??**), and the lower bound of Lemma **??**, then dividing by n, produces the redudancy bound

$$\frac{1}{n}R_{LZ}(x_1^n) \leq K\frac{c}{n} + \frac{\log n/c}{n/c}, \tag{26}$$

where K is a constant.

To complete the argument a simple bound for c/n will be needed, a bound that follows from the foct that the largest value of c is obtained when all short blocks occur. It is enough to consider the case when all blocks of length up to t occur, so that

$$c = \sum_1^t |A|^i \sim |A|^t,\ n = \sum_1^t i|A|^i \sim t|A|^t,$$

which gives the (asymptotic) bound $c/n = O(1/\log n)$. Since $\log x/x$ is decreasing in x for $x > e$, the desired result,

$$\frac{1}{n}R_{LZ}(x_1^n) = O\left(\frac{\log\log n}{\log n}\right),$$

follows easily from the bound (**??**).

8.5 Minimization for general measures.

The minimization result claimed in the paragraph following statement (10) on page 10 follows from a general result about nonnegative measures, a result that is a simple consequence of the log-sum inequality. Suppose P is an arbitrary nonnegative measure, suppose Q is a probability distribution, and set $Q^*(a) = P(a)/\sum P(b)$. The log-sum inequality then gives

$$D(P\|Q) \geq \Big(\sum P(a)\Big)\log\Big(\sum P(a)\Big) = D(P\|Q^*).$$

8.6 Cutting off the memory.

Let P be a stationary finite-alphabet process. The k-step Markoviztion of P is the k-step Markov process $P^{(k)}$ defined by the transition probabilities

$$P(x_{k+1}|x_1^k) = \frac{P(x_1^{k+1})}{P(x_1^k)}.$$

The following general result shows that the conditions stated in Theorems 13 and 14 often hold. For example, the set of all Markov types of all orders is a countable set for which the conditions of Theorem 14 hold for every ergodic process P.

Theorem 21 *$D^\infty(P\|P^{(k)}) \to 0$ as $k \to \infty$.*

Proof. We have

$$\log\frac{P(x_1^k)}{P^{(k)}(x_1^n)} = \sum_{i=k+1}^n \log\frac{P(x_{i+1}|x_1^i)}{P(x_{i+1}|x_{i-k+1}^i)}$$

so that taking expectations yields

$$\begin{aligned}
&E_P\left(\log\frac{P(x_1^k)}{P^{(k)}(x_1^n)}\right) = \\
&\qquad \sum_{i=k+1}^n \sum_{x_1^{i+1}} P(x_1^{i+1})\log\frac{P(x_{i+1}|x_1^i)}{P(x_{i+1}|x_{i-k+1}^i)}.
\end{aligned} \tag{27}$$

To see what this is we use the formula

$$I(X\wedge Y|Z) = \sum P(x,y,z)\log\frac{P(x|y,z)}{P(x|z)},$$

with $X = X_{i+1}$, $Y = X_1^i$, $Z = X_{i-k+1}^i$; the sum (**??**) then takes the form

$$\begin{aligned}
&\sum_{i=k+1}^n I(X_{i+1}\wedge X_1^i|X_{i-k+1}^i) = \\
&\qquad \sum_{i=k+1}^n I(X_1\wedge X_{-i+1}^0|X_{-k+1}^0)
\end{aligned}$$

where stationarity was used to obtain the final form. Now we pass to the limit in n, using the martingale theorem to obtain

$$D^\infty(P\|P^{(k)}) = I(X_1\wedge X_{-\infty}^0|X_{-k+1}^0),$$

whichs goes to 0 as $k \to \infty$, establishing the theorem.

8.7 Arithmetic coding.

An interesting idea due originally to Elias and later adapted in a useful form by Rissanen leads to sequential coding procedure known as arithmetic coding. Fix an integer n. An arithmetic code first assigns to each x_1^n a nonempty subinterval $J(x_1^n) = [\ell(x_1^n), r(x_1^n))$ of the unit interval $[0,1)$ such that the set $\{J(x_1^n)\}$ is a partition of the interval into disjoint subintervals. To obtain sequential codes it is required that

$$J(x_1^n) = \cup_s J(x_1^n s), \tag{28}$$

and to avoid trivialities it is required that $J(x_1^n)$ shrinks to a single point as $n \to \infty$. The code is then defined by setting $C(x_1^n) = z_1^m$ if the endpoints of $J(x_1^n)$ have binary expansions $.z_1 z_2 \dots z_m$ that agree in their first m places but no further, that is,

$$\ell(x_1^n) = .z_1 z_2 \dots z_m 0 \dots, \ r(x_1^n) = .z_1 z_2 \dots z_m 1 \dots .$$

Since the intervals are disjoint this is a prefix code. Furthermore, if $Q_n(x_1^n) = r(x_1^n) - \ell(x_1^n)$ then Q_n is a probability distribution; the condition (**??**) then implies that there is a process Q whose n-th order probabilties are given by Q_n. Note also that $2^{-L(x_1^n)-1} \geq Q(x_1^n)$ so that $L(x_1^n) \leq -\log Q(x_1^n)+1$ and hence the Q_n-expected code length is no more than $H(Q_n)+1$. The code operates sequentially in that the code word assigned to x_1^{n+1} is an extension of the word assigned to x_1^n.

In general, a process Q can be specified by giving its sequence of conditional probabilities $Q(x_n|x_1^n)$. These probabilities can then be used to specify subintervals of the unit interval in a sequential manner. Thus, we first partition $[0,1)$ into subintervals labeled $J(x_1), x_1 \in A$, then for each x_1 partition $J(x_1)$ into subintervals $J(x_1^2), x_2 \in A$. Proceeding in this manner the process Q defines a nested sequence of partitions $\{J(x_1^n)\colon x_1^n \in A^n\}$ that satisfy the compatibility condition (**??**), hence define an arithmetic code. The mixture processes introduced in Sections 6 and 7, thus lead to useful sequential codes, as noted in Remark 9 of the notes.

9 Further examples.

Example 2 A lot contains $n = 100$ defective items. Each item is tested but the test may fail with probability $p = 0.1$. Use the techniques of this course to estimate the probability that 20 or more defective items remain undetected.

Solution. Let X_i denote the outcome of testing the i-th defective item, $X_i = 1$, if detected, 0, otherwise. Let $\hat{P}_n$ denote the empirical distribution, and Π the set of all binary distributions Q with $Q(0) \geq 0.2$. We want to estimate tha probability that $\hat{P}_n \in \Pi$. Sanov's theorem gives

$$\text{Prob}(\hat{P}_n \in \Pi) \approx \exp(-nD(\Pi\|P)), \ P = (0.1, 0.9).$$

Now

$$\begin{aligned} D(\Pi\|P) &= \min_{Q\in\Pi} D(Q\|P) \\ &= 0.2 \log \frac{0.2}{0.1} + 0.8 \log \frac{0.8}{0.9} \approx 0.066, \end{aligned}$$

and $\exp(-nD(\Pi\|P)) \approx 0.01$. This number is suspect because the large deviations approximation is reliable for "very small" probabilities (however, since Π is convex, this number certainly gives an upper bound.) In our case, approximating the binomial distribution by the normal will be preferable, and its result is smaller by a factor of 10.

Example 3 The null-hypothesis P_1 is to be tested on the basis of an iid sample x_1^n, and the probability of first kind error is required to be no more than $\exp(-n(\gamma - o(1)))$. Prove that the test with critical region equal to $\{x_1^n\colon \hat{P}_{x_1^n} \notin \Pi\}$, with $\Pi = \Pi_\gamma = \{Q\colon D(Q\|P_1) \leq \gamma\}$, is asymptotically optimal against any alternative P_2 with $D(P_2\|P_1) > \gamma$, in the sense that for no test meeting the condition on the first kind error can the probability of second kind error go to 0 with a larger exponent.

Solution. The meaning of this problem is the following. Two distributions P_1 and P_2 are given such that $D(P_2\|P_1) > \gamma$. Based on a sample path x_1^n drawn from P_1^n or P_2^n, a decision is to be made as to which process it comes from. To make this decision, the set A^n of possible sample paths is partitioned into two disjoint sets H_1^n and H_2^n, and the decision rule is to choose P_i if $x_1^n \in H_i^n$. It is enough to specify the region H_2^n, which is called the critical region of the test, for we can take H_1^n to be its complement. The first part of the problem is to show that if $H_2^n = \{x_1^n\colon \hat{P}_{x_1^n} \notin \Pi\}$ then the probability of a first kind error, namely $P_1(H_2^n)$, satifies

$$P_1(H_2^n) \leq \exp(-n(\gamma - o(1))) \tag{29}$$

For this partition there will be a largest number $\delta > 0$ such that the probability of a second kind error, $P_2(H_1^n)$, satisfies

$$P_2(H_1^n) \leq \exp(-n(\delta - o(1))) \tag{30}$$

The second goal is to show that if $\{H_1^n, H_2^n\}$ is any sequence of partitions for which (**??**) holds then $P_2(H_1^n)$ cannot go to zero at a faster rate than (**??**).

Let us first show that for $H_2^n = \{x_1^n : \hat{P}_{x_1^n} \notin \Pi\}$, the two inequalities (**??**) and (**??**) both hold. Let P^* be the I-projection of P_2 onto Π and let $\delta = D(P^*\|P_2)$, which is necessarily positive. If $\hat{P}_{x_1^n} \notin \Pi$ then $D(\hat{P}_{x_1^n}\|P_1) > \gamma$, so that (**??**) holds. Sanov's theorem asserts that (**??**) holds.

To establish the second goal we first note that $D(\Pi_\gamma\|P_1)$ is continuous in γ, hence given $\epsilon > 0$ we can choose n so large that there is an n-type Q such that

$$D(Q\|P_1) < \gamma - \epsilon, \text{ and } D(Q\|P_2) < \delta + \epsilon.$$

If the critical region $C_n = H_2^n$ of some test contains at least half of $\mathcal{T}_Q^n$ then $P_1(C_n)$ is lower bounded by $P_1(\mathcal{T}_Q^n)/2$ which is turn lower bounded by

$$\frac{1}{2}\binom{n+|X|-1}{|X|-1}^{-1}\exp[-nD(Q\|P_1)] \geq \exp(-n(\gamma-\epsilon/2)),$$

for large enough n. Therefore if we require that (**??**) holds then, if n is large enough, at least half of $\mathcal{T}_Q^n$ is not in C_n and hence the second kind error $P_2(C_n^c)$ is lower bounded by

$$\frac{1}{2}\binom{n+|X|-1}{|X|-1}^{-1}\exp[-nD(Q\|P_2)] \geq \exp(-n(\delta+\epsilon/2),$$

for large enough n. Since ϵ is arbitrary this proves that the second kind error cannot go to 0 with a larger exponent than δ.

Example 4 Let a contingency table with 3 features, each with two categories, be given by the cell counts

$$\begin{array}{cccc} x_{111}=8, & x_{112}=10, & x_{121}=5, & x_{122}=7 \\ x_{211}=11, & x_{212}=1, & x_{221}=14, & x_{222}=4 \end{array}$$

Consider the log-linear models corresponding to (i) $\Gamma = \{\{1\},\{2\},\{3\}\}$ and (ii) $\Gamma = \{\{1,2\},\{1,3\}\}$, and determine the maximum likelihood estimate P^* for both models. Does either model fit the data?

Solution. Here $n = 60$ and the one-dimensional marginal counts are

$$x_{1\cdot\cdot} = x_{2\cdot\cdot} = x_{\cdot 1\cdot} = x_{\cdot 2\cdot} = 30,\ x_{\cdot\cdot 1} = 38,\ x_{\cdot\cdot 2} = 22.$$

The two-dimensional marginal counts needed in part (ii) are

$$\begin{array}{cccc} x_{11\cdot}=18, & x_{12\cdot}=12, & ,x_{21\cdot}=12, & x_{22\cdot}=18 \\ x_{1\cdot 1}=13, & x_{1\cdot 2}=17, & ,x_{2\cdot 1}=25, & x_{2\cdot 2}=5. \end{array}$$

(i) For $\Gamma = \{\{1\},\{2\},\{3\}\}$, P^* is the product of the empirical marginal distributions, that is, $P^*(i,j,k) = \frac{x_{i\cdot\cdot}}{n}\frac{x_{\cdot j\cdot}}{n}\frac{x_{\cdot\cdot k}}{n}$. This P^* clearly does not fit the sample.

(ii) For $\Gamma = \{\{1,2\},\{1,3\}\}$, P^* is of the form $P^*(i,j,k) = a(i,j)b(i,k)$, that is, under this model the second and third features are conditionally independent given the first feature. Hence $P^*(i,j,k)$ is obtained by multiplying the empirical marginal $(1/n)x_{i\cdot\cdot}$ by the conditional distributions evaluated from the $\{1,2\}$ and $\{1,3\}$ empirical marginals $x_{ij\cdot}/x_{i\cdot\cdot}$ and $x_{i\cdot k}/x_{i\cdot\cdot}$. Thus $P^*(i,j,k) = (x_{ij\cdot}x_{i\cdot k})/(nx_{i\cdot\cdot})$. For $x^*_{ijk} = nP^*(i,j,k)$ we get $x^*_{111} = 7.8$, $x^*_{112} = 10.2$, $x^*_{121} = 5.2$, $x^*_{122} = 6.8$, $x^*_{211} = 10$, $x^*_{212} = 2$, $x^*_{221} = 15$, $x^*_{222} = 3$, and

$$nD(\hat{P}\|P^*) = \sum_{i,j,k} x_{ijk}\log\frac{x_{ijk}}{x^*_{ijk}} = \frac{1.1}{2}\ln 2$$

The degress of freedom, that is, the dimensionality of L determined by the marginals, is 2, and $\text{Prob}(\chi^2 \geq 1.1) \approx 0.58$, hence the model fits well.

Example 5 A finite-valued random variable Y is ϵ-independent from a finite-valued X if

$$\sum_x P(X=x)\sum_y |P(Y=y|X=x) - P(Y=y)| < \epsilon.$$

Show that $I(X \wedge Y) \leq (\epsilon^2/2)\log e$ implies ϵ-independence.

Solution. Exercise 17, page 58, of the Csiszár-Körner book gives the bound $2D(P\|Q) \geq \log e|P-Q|^2$, where $|\cdot|$ denotes variational distance. Since $I(X \wedge Y) = D(P_{X,Y}\|P_X \times P_Y)$ the condition $I(X \wedge Y) \leq (\epsilon^2/2)\log e$ implies that $D(P_{X,Y}\|P_X \times P_Y) \leq \epsilon$, which is the condition for ϵ-independence.

Example 6 Consider an exponential family defined by densities $P_\theta(x) = c(\theta)\exp[\sum_{i=1}^k \theta_i f_i(x)]$, where $c(\theta) = (\int \exp[\sum_{i=1}^k \theta_i f_i(x)]dx)^{-1}$ and $\theta = \theta_1^k$. Let θ_{ML} be the value maximizing P_θ for a given x_0 (which we suppose exists.) Show that $-\log P_{\theta_{ML}}(x_0) = H(P_{\theta_{ML}})$.

Solution. Since

$$\begin{aligned} H(P_\theta) &= -\int P_\theta(x)\log P_\theta(x)\,dx \\ &= -\int \log c(\theta) P_\theta(x)\,dx \\ &\quad -\int (\log e)\sum_{i=1}^k \theta_i f_i(x) P_\theta(x)\,dx \\ &= -\log c(\theta) - (\log e)\sum_{i=1}^k \theta_i E_\theta f_i, \end{aligned}$$

it suffices to show that for $\theta = \theta_{\rm ML}$ we have $E_\theta f_i = f_i(x_0),\ 1 = 1, \ldots, k$. But this immediately follows by setting the derivatives $(\partial/\partial\theta_i) \log P_\theta(x_0)$ equal to 0. For this last step it is necessary to assume that $\theta_{\rm ML}$ is an interior point of the set of those $\theta' s$ for which P_θ is defined, that is, that the integral in the definition of $c(\theta)$ is finite.

Example 7 Let P^n and Q^n be n-dimensional distributions on A^n such that $n^{-1}D(P^n\|Q^n) \to 0$. Show that if for some sets $B_n \subset A^n$ we have $Q^n(B_n) < \exp(-\epsilon n)$ for some $\epsilon > 0$ that does not depend on n then $P^n(B_n) \to 0$. Is it also true that $P^n(B_n) < \exp(-\epsilon n)$ implies that $Q^n(B_n) \to 0$?

Solution. For an arbitrary set A,

$$\begin{aligned} &D(P^n\|Q^n) \\ &\quad \geq P^n(A) \log \frac{P^n(A)}{Q^n(A)} + P^n(A_n^c) \log \frac{P^n(A^c)}{Q^n(A^c)} \\ &\quad \geq P^n(A) \log P^n(A) + P^n(A_n^c) \log P^n(A_n^c) \\ &\qquad + P^n(A) \log \frac{1}{Q^n(A)} \\ &\quad \geq -1 + P^n(A) \log \frac{1}{Q^n(A)}. \end{aligned}$$

If here $Q^n(A) \leq \exp(-\epsilon n)$ then it follows that $D(P^n\|Q^n) \geq -1 + \epsilon n P^n(A)$, that is,

$$P^n(A) \leq \frac{1}{\epsilon}\left[\frac{1}{n} D(P^n\|Q^n) + \frac{1}{n}\right].$$

Example 8 Let X, Y, Z be real-valued random variables with unknown joint density $p(x, y, z)$ for which $E(X^2) + E(Y^2) + E(Z^2) = a$ and $E(XY) + E(YZ) = b$, where a and b are known. Show that the joint density achieving maximum entropy subject to these constraints is Gaussian with mean 0. Indicate how its covariance matrix could be determined (the actual computation is not required) and show that for this maximum entropy joint distribution $E(X^2) = E(Z^2) \neq E(Y^2)$ and $E(XY) = E(YZ) \neq E(XZ)$.

Solution. Let $f_1(x, y, z) = x^2 + y^2 + z^2$ and $f_2(x, y, z) = xy + yz$. Then the entropy $H(p) = -\int p(u) \log p(u)\, du$, where $u = (x, y, z)$, $du = dxdydz$, has to be maximized subject to the constraints $\inf f_1(u)p(u)\ du = a$, $\int f_2(u)p(u)\ du = b$. The maximizing density will be in the exponential family

$$p_\theta(u) = c(\theta) \exp[\theta_1 f_1(u) + \theta_2 f_2(u), \theta = (\theta_1, \theta_2),$$

provided this family has a member satisfying the constraints. Comparing the family with the 3-dimensional, mean 0, Gassian densities, that is, those of the form,

$$p(u) = \frac{(\det A)^{1/2})}{(2\pi)^{3/2}} \exp\left\{-\frac{1}{2} u A u^T\right\},$$

where A is symmetric and positive definite, we see that our exponential family is a subfamily of these Gaussians, with

$$\begin{bmatrix} -2\theta_0 & -\theta_1 & 0 \\ -\theta_1 & -2\theta_0 & -\theta_1 \\ 0 & -\theta_1 & -2\theta_0 \end{bmatrix}.$$

Computing the covariance matrix $\Sigma = A^{-1}$, the given moment constraints result in two equations for the unknowns θ_1 and θ_2. The solution of these equations is straightforward, but tedious. It is clear, however, from the form of A, that the first and last elements of the main diagonal of $\Sigma = A^{-1}$ will be equal and its middle element will be different from these (unless $\theta_1 = 0$, which occurs if $b = 0$, when the maximum entropy distribution is iid.) The remaining assertion of the problem also follows from the form of A without any further calculations.

10 Summary of Process Concepts.

A number of process concepts will be used in the discussion of redundancy. These concepts and the results to be used are summarized here.

A *(stochastic) process* is a sequence $\{X_n\}$ of random variables defined on a probability space, say (X, Σ, μ). We shall assume that all the random variables have values in a fixed finite set A, called the alphabet. For each n a process defines a probability measure on A^n, called the n-fold joint distribution, by the formula

$$P_n(x_1^n) = \text{ Prob}\,(X_i = x_i,\ 1 \leq i \leq n).$$

The sequence of measures $\{P_n\}$ is not completely arbitrary, for the consistency conditions,

$$P_n(x_1^n) = \sum_{x_{n+1}} P_{n+1}(x_1^{n+1}) \tag{31}$$

must hold.

The space (X, Σ, μ) on which the process is defined is not important; all that matters is the sequence of joint distributions, $\{P_n\}$. In fact, two processes are said to be *equivalent* if they have the same joint distributions; we are free to choose any convenient space and sequence of functions, as long as the joint distributions is not changed. The Kolmogorov model takes the space

to be the set A^∞ of infinite sequences drawn from A, and the functions to be the coordinate functions, defined by $\hat{X}_n(x) = x_n,\ x \in A^\infty$. The measure is the (unique) Borel measure P defined by the requirement that if

$$[a_1^n] = \{x: x_i = a_i, 1 \le i \le n\}$$

is the cylinder set defined by a_1^n, then $P([a_1^n]) = P_n(a_1^n)$.

In summary, the concept of process, that is, a sequence of measures $\{P_n\}$ that satisfy the consistency conditions, (**??**), is formally equivalent to the concept of Borel probability measure P on the sequence space A^∞. We usually take the latter as our definition of process; thus, when we say process we shall mean a Borel probability measure P on the sequence space A^∞. We shall use the notation $P(a_1^n)$ for $P([a_1^n])$, as well as sample path terminology. A sample path is a member of A^∞, while a finite sample path is a member of some A^n.

A process P is *stationary* if it is invariant under the shift T, which is the transformation on A^∞ defined by the formula $(Tx)_n = x_{n+1},\ x \in A^\infty, n = 1, 2, \ldots$. Thus P is stationary if and only if $P = P \circ T^{-1}$.

A stationary process is *ergodic* if almost every sample path is "typical" for the process. The concept of "typical" is defined as follows. The relative frequency of occurence of a_1^k in the sequence x_1^n is the distribution $\hat{P}_k = \hat{P}_k(\cdot|x_1^n)$ on A^k defined by

$$\hat{P}_k(a_1^k|x_1^n) = \frac{|\{i \in [0, n-k]: x_{i+1}^{i+k} = a_1^k\}|}{n-k+1}.$$

The measure $\hat{P}_k$ is also called the *empirical distribution of overlapping k-blocks in the sample path x_1^n*. The sequence x is said to be *typical* for the process P if for all k and all a_1^k, the following holds

$$P(a_1^k) = \lim_{n\to\infty} \hat{P}_k(a_1^k|x_1^n).$$

The set of sequences that are typical for P will be denoted by $\mathcal{T}(P)$. A stationary process P is ergodic if its set of typical sequences has measure 1, that is, if $P(\mathcal{T}(P)) = 1$.

The entropy (or entropy-rate) of a stationary process P is defined by $H(P) = \lim_n H_n/n$ where the n-th order entropy $H_n = H_n(P)$ is defined by

$$H_n = -\sum_{a_1^n} P(a_1^n) \log P(a_1^n).$$

The entropy theorem (also known as the Shannon-McMillan-Breiman Theorem) asserts that if P is an ergodic process of entropy H then

$$\frac{1}{n} \log \frac{1}{P(x_1^n)} = -\frac{1}{n} \log P(x_1^n) = H, \text{ a. s.}$$

For ergodic processes we also have that the entropy of the empirical distribution, $H(\hat{P}_k)$, converges almost surely to the theoretical entropy, H_k. Furthermore, if we define transition probabilities by the formula

$$\hat{P}_k(a_k|a_1^{k-1}) = \frac{\hat{P}_k(a_1^k)}{\sum_{a_k} \hat{P}_k(a_1^k)},$$

the entropy of the resulting Markov chain will converge almost surely, as sample path length $n \to \infty$, to the conditional entropy, $H(X_k|X_1^{k-1})$, which, in turn, converges as $k \to \infty$ to the entropy-rate $H(P)$.

A stationary process P is always a mixture of ergodic processes, that is, there is a probability space (Y, Σ, ν) and a family $U_y, y \in Y$, of ergodic processes such that for each a_1^n the function $U_y(x_1^n)$ is Σ-measurable and such that

$$P(a_1^n) = \int U_y(a_1^n)\nu(dy).$$

The process $\{X_n\}$ is *finite-state (hidden Markov)* if there is a finite alphabet process $\{S_n\}$ such that the process $Y_n = (X_n, S_n)$ is a Markov chain. Csiszár has shown (unpublished) that if Q is finite-state then for any stationary, ergodic process P the limiting divergence-rate

$$D^\infty(P\|Q) = \lim_n \frac{1}{n} \sum_{a_1^n} P(x_1^n) \log \frac{P(x_1^n)}{Q(x_1^n)}$$

exists, and, furthermore, $(1/n) \log P(x_1^n)/Q(x_1^n)$ converges, for P-almost all x, to the limit $D^\infty(P\|Q)$.

11 Homework # 1.

Due Date: Oktober 7-én.

1. Find the Shannon-Fano code for the distribution $P = (0.4, 0.35, 0.1, 0.1, 0.05)$. Determine the average length and compare it with the entropy $H(P)$. Can you improve this code by shortening some words, without losing the prefix property? Do you get an optimal code in this way?

2. Determine whether there exist binary prefix codes with the following codeword lengths and give such a code if the answer is yes.
   (a) 2,3,3,3,4,4,4,4,4,5,5,5 (b) 2,2,3,3,3,4,4,4,5,6,6

3. Determine which of the following bit sequences can be a code of some sequence of integers, using the prefix code given in the notes.
   (a) 00111100000011000100011101010101000111101
(b) 00001001110010000000100111100001011100010010000

4. Let $B \subset A^n$ and let $P = (1/|B|)\sum_{x\in B} P_x$, be the average type of the sequences in B.

   (a) Prove that $|B| \le \exp[nH(P)]$.

   (b) Prove that

$$\sum_{i=0}^{k} \binom{n}{i} \le 2^{nh(k/n)},\ k \le n/2$$

   where $h(p) = -p\log p - (1-p)\log(1-p)$.

   (c) Given a function f on a finite set X, show that for every α that is a possible value of $\sum_{i=1}^n f(x_i)/n$, we have

$$\left|\left\{x_1^n\colon \frac{1}{n}\sum_{i=1}^{n} f(x_i) = \alpha\right\}\right| \le \exp\left[n \max_{E(f(X))=\alpha} H(X)\right]$$

5. Prove that $H(Y|X)$ is a concave function of the joint distribution of (X,Y), that is, if $P_{XY} = \alpha P_{X_1Y_1} + (1-\alpha)P_{X_2Y_2}$ then $H(Y|X) \ge \alpha H(Y_1|X_1) + (1-\alpha)H(Y_2|X_2)$.

6. Let P_1 and P_2 be probability distributions on the finite set X such that $D(P_2\|P_1) > \gamma$. Prove that the I-projection P^* of P_2 on $\Pi = \{Q\colon D(Q\|P_1) \le \gamma\}$ is of the form $P^* = cP_1^\theta P_2^{1-\theta}$, where $c > 0$ and $0 < \theta < 1$ are determined by the requirements that $\sum P^*(x) = 1$ and $D(P^*\|P_1) = \gamma$. (Hint: first show that P^* is also the I-projection of P_2 on the linear family $L = \{Q\colon \sum Q(x)\log\frac{P_1(x)}{P_2(x)} = \delta - \gamma\}$, where $\delta = D(P^*\|P_2)$.)

7. Prove that $D(P\|Q) \le \chi^2(P,Q)\log e$.

8. Let $X_1, X_2, \ldots$ be an iid sequence of X-valued random variables with entropy H, and let $\hat{H}_n$ be the empirical entropy of X_1^n, that is, the entropy of the empirical distribution $\hat{P}_n$. Prove that

$$H - \frac{1}{n}\log\binom{n+|X|-1}{|X|-1} \le E(\hat{H}_n) \le H.$$

9. Given two strictly positive finite distributions P_1 and P_2 on X, determine γ such that there is exactly one P^* with $D(P^*\|P_1) = D(P^*\|P_2) = \gamma$. Show that

$$\gamma = -\log \min_{0\le\theta\le 1} \sum_x P_1^\theta(x)P_2^{1-\theta}(x).$$

12 Homework # 2.

1. For k simple hypotheses $P_1, \ldots, P_k$, and a classification rule consisting of the partition $A = (A_1, \ldots, A_k)$ of X^n such that P_i is accepted when the sample belongs to A_i, there are k error probabilities, $e_i = P_i^n(A_i^c), i = 1, \ldots, k$. Give a necessary and sufficient condition for the existence of classification rules such that all k error probabilities go to 0 with exponential rate at least some $\gamma > 0$, as the sample size n goes to infinity, that is, $e_i \le \exp(-n(\gamma + o(1)), i = 1, \ldots, k$.

2. Let $\mathcal{E}_1 \subset \mathcal{E}_2$ be exponential families of the form

$$\mathcal{E}_1 = \left\{Q\colon Q(x) = Q_0(x)c(\theta)\exp\left(\sum_{i=1}^{k_1}\theta_i f_i(x)\right)\right\}$$
$$\mathcal{E}_2 = \left\{Q\colon Q(x) = Q_0(x)c(\theta)\exp\left(\sum_{i=1}^{k_2}\theta_i f_i(x)\right)\right\},$$

   where $k_2 > k_1$. Given a sample with empirical distribution $\hat{P}$, let $P_i^* \in \mathcal{E}_i$, be the maximum likelihood estimate for the model $\mathcal{E}_i$ $i = 1, 2$. Prove that P_2^* is the I-projection of P_1^* onto $\mathcal{L} = \{P\colon \sum_{i=1}^{k_2} P(x)f_i(x) = \sum_{i=1}^{k_2}\hat{P}(x)f_i(x)\}$.

3. In a telephone network serving r cities, the incoming and outgoing calls were counted in each city on a given day. From these numbers, $x_{\text{in}}(k)$ and $x_{\text{out}}(k)$, $k = 1, \ldots, r$, the number of calls $x(i,j)$ from city i to city j are inferred by the method of maximum entropy, setting $x^*(i,j) = np^*(i,j)$; here n is the total number of calls and $P^* = \{p^*(i,j)\}$ is the maximum entropy distribution among those $P = \{p(i,j)\}$ that satisfy the marginal constraints

$$\sum_{j=1}^{r} p(k,j) = \frac{1}{n}x_{\text{out}}(k),\ \sum_{i=1}^{r} p(i,k) = \frac{1}{n}x_{\text{in}}(k),$$

   for $k = 1, \ldots, r$, and, in addition, $p(k,k) = 0$, $k = 1, \ldots, r$ (local calls were not counted.) Specify the exponential family for which this P^* is the maximum likelihood estimate, and suggest an iterative algorithm for determining P^*.

4. Suppose that for a 5×5 array of random variables X_{ij}, each taking values in a finite set X, the joint distributions of "neighboring pairs" $(X_{ij}, X_{i(j+1)})$ and $(X_{ij}, X_{(i+1)j})$ are known, where addition modulo 5 is used. Based on this information, the joint distribution of the whole array is estimated by

maximizing joint entropy subject to the constraints given. Interprete the maximum entropy joint distribution as an I-projection, and suggest a convergent iteration for computing it.

5. For binary sequences of length n let $Q(x_1^n)$ denote the uniform mixture of the iid probabilities $P(x_1^n) = p^{n_0}(1-p)^{n-n_0}$, where n_0 denotes the number of zeroes in x_1^n. Find an explicit formula for Q and determine the asymptotic behavior of the maximum redundancy $\log P_{ML}(x_1^n)/Q(x_1^n)$ as $n \to \infty$, for sequences of two kinds: (i) $n_0 \sim \alpha n$ for $0 < \alpha < 1$, and (ii) n_0 constant. Suggest a mixture distribution that is more appropriate for the latter case.

6. Determine the code-length for the sequence
   0001100010001110000100001100100000
   using the universal coding method discussed in class, supposing first that the sequence is iid, then second that it is Markov.

7. Let $\{P_\theta\}_{\theta\in\Theta}$ be an arbitrary family of distributions for a random process with finite alphabet A, and let r_n denote the smallest positive integer r for which there exists a prefix code with codeword lengths $L(x_1^n)$ such that the redundancy satisfies the uniform bound
   $$L(x_1^n) + logP_\theta(x_1^n) \le r, \forall x_1^n, \theta.$$
   Show that r_n equals $\log S_n$, up to 1 bit, where $S_n = \sum_{x_1^n} \sup_\theta P_\theta(x_1^n)$.

8. For an iid sequence of random variables with values in a finite set X let $\hat{H}_n = \hat{H}_{n,x_1^n}$ denote the empirical entropy of the sequence x_1^n, that is, $\hat{H}_n = H(\hat{P})$ where $\hat{P}$ is the type of x_1^n. Show that with probability 1
   $$n\hat{H}_n \le -\log P(x_1^n) \le n\hat{H}_n + \frac{|X|-1}{n}\log n + Z,$$
   where P is the true distribution and Z is a random variable, depending on n, such that $E(Z) < \infty$.

13 Solutions: Homework #1.

1. The i-th codeword, $c(i)$, is the first $\ell_i = \lceil -\log p_i \rceil$ binary digits of $a_i = \sum_{j<i} p_j$.

i	1	2	3	4	5
a_i	0	0.4	0.75	0.85	0.95
ℓ_i	2	2	4	4	5
$c(i)$	00	01	1100	1101	11110
opt(i)	00	01	10	110	111

   where the final line indicates the optimal (Huffman) code. Expected code length and entropy are
   $$L = \sum p_i\ell_i = 2.55,\ H = -\sum p_i \log p_i = 1.94$$
   while the optimal L is 2.15. Note that the last codeword can be shortened by deleting its final two bits, but the obtained code is still not optimal.

2. The Kraft inequality shows that there is no prefix code with length set (a), but there is one for (b).

3. The first bit sequence cannot be decoded, but it can if one more bit is added at the end.

4. Let $N(a, B)$ denote the number of occurrences of a in all the sequences $x_1^n \in B$, so that $P(a) = N(a, B)/n|B|$. Let $X_1, \ldots, X_n$ be random variables defined by $\text{Prob}(X_1^n = x_1^n) = 1/|B|$, $x_1^n \in B$ and 0, otherwise. But $P_i(a) = \text{Prob}(X_i = a)$ satisfies $\sum_i P_i(a) = N(a, B)/|B| = nP(a)$. Thus,
   $$\begin{aligned}\log|B| &= H(X_1^n) \le \sum_i H(P_i) \le \\ &\le nH\left(\frac{1}{n}\sum_i P_i\right) = nH(P).\end{aligned}$$
   This establishes part (a).

   For part (b) apply part (a) to the set B of binary sequences of length n that contain no more than k zeroes. For part (c), let P be the average type of the sequences x_1^n that belong to
   $$B = \left\{x_1^n : \frac{1}{n}\sum_{i=1}^n f(x_i) = \alpha\right\}.$$
   ¿From (a) we have $|B| \le \exp[nH(P)]$. But $x_1^n \in B$ means that its type $P_{x_1^n}$ satisfies $\sum_a P_{x_1^n}(a)f(a) = \alpha$; which therefore also holds for the average type P, that is, $\sum_a P(a)f(a) = \alpha$. Hence $H(P) \le \max_{E(f(X))=\alpha} H(X)$. A lower bound on $|B|$ cannot be given without additional assumptions.

5. It suffices to show that
   $$\begin{aligned}P_{XY}(x,y)\log\tfrac{P_{XY}(x,y)}{P_X(x)} \le \alpha P_{X_1Y_1}(x,y)\log\tfrac{P_{X_1Y_1}(x,y)}{P_{X_1}(x)} \\ +(1-\alpha)P_{X_2Y_2}(x,y)\log\tfrac{P_{X_2Y_2}(x,y)}{P_{X_2}(x)}.\end{aligned}$$
   This follows from
   $$a_1\log\frac{a_1}{b_1} + a_2\log\frac{a_2}{b_2} \ge (a_1+a_2)\log\frac{a_1+a_2}{b_1+b_2},$$
   with $a_1 = \alpha P_{X_1Y_1},\ a_2 = (1-\alpha)P_{X_2Y_2}(x,y),\ b_1 = \alpha P_{X_1}(x),\ b_2 = (1-\alpha)P_{X_2}(x)$.

6. Suppose the P_i are everywhere positive. We prove that the I-projection of P_2 on

$$L = \left\{ Q\colon \sum Q(x) \log \frac{P_1(x)}{P_2(x)} = \delta - \gamma \right\}$$

is that same P^* as the I-projection of P_2 on $\Pi = \{Q\colon D(Q\|P_1) \le \gamma\}$, where $\delta = D(P^*\|P_2)$. Indeed, $P^* \in L$; furthermore, for every $Q \in L$

$$D(Q\|P_2) - D(Q\|P_1) = \sum Q(x) \log \frac{P_1(x)}{P_2(x)} = \delta - \gamma.$$

Here if $D(Q\|P_2)$ were less than δ, also $D(Q\|P_1)$ would be less than γ, which would imply that $Q \in \Pi$, contradicting the hypothesis that the P^* with $D(P^*\|P_2) = \delta$ is the I-projection of P_2 on Π.

Now, the exponential family corresponding to L is

$$\begin{aligned} &\left\{ P_\theta\colon P_\theta(x) = c(\theta) P_2(x) \exp\left[\theta \log \frac{P_1(x)}{P_2(x)}\right] \right\} \\ &\quad = \left\{ P_\theta\colon P_\theta(x) = c(\theta) P_1^\theta(x) P_2^{1-\theta}(x) \right\} \end{aligned}$$

Since P_1 and P_2 belong to this family, and L speparates P_1 and P_2, there must be some θ^* with $P_{\theta^*} \in L$. Then, by the general theorem, $P^* = P_{\theta^*}$ is the I-projection of P_2 on L and therefore also on Π.

7. Using the inequality $\ln x \le x - 1$ we have

$$\begin{aligned} \sum_i p_i \ln \frac{p_i}{q_i} &\le \sum_i p_i \left(\frac{p_i}{q_i} - 1 \right) \\ &= \sum_i \frac{p_i^2}{q_i} - 1 = \sum_i \frac{(p_i - q_i)^2}{q_i} \end{aligned}$$

Multiplying both sides by $\log e$ produces the claimed inequality.

8. Since $\hat{H}_n = H(\hat{P}_n)$ and $E\hat{P}_n = P$, we have $E\hat{H}_n \le H(E\hat{P}_n) = H(P)$, by concavity. Here $H(\hat{P}_n)$ denotes the entropy of $\hat{P}_n$ as a distribution. On the other hand, $\hat{P}_n$ is also a random variable; $\bar{H}(\hat{P}_n)$ will denote the entropy of this random variable. Then we can write

$$\begin{aligned} nH(P) &= H(X_1^n) = H(X_1^n|\hat{P}_n) + \bar{H}(\hat{P}_n) \\ &= \sum_Q \Pr(\hat{P}_n = Q) \log |\mathcal{T}_Q| + \bar{H}(\hat{P}_n) \\ &\le \sum_Q \Pr(\hat{P}_n = Q) nH(Q) + \bar{H}(\hat{P}_n) \\ &= n\hat{H}_n + \bar{H}(\hat{P}_n) \end{aligned}$$

The result now follows from the fact that $\bar{H}(\hat{P}_n) \le \log \binom{n + |X| - 1}{|X| - 1}$.

9. From an earlier problem, the I-projection of P_2 on $\Pi = \{Q\colon D(Q\|P_1) \le \gamma\}$ is of the form

$$P^*(x) = cP_1^\theta(x) P_2^{1-\theta}(x),\ c = \left[\sum_x P_1^\theta(x) P_2^{1-\theta}(x) \right]^{-1}.$$

Therefore,

$$\begin{aligned} D(P^*\|P_1) &= \sum P^*(x) \log \frac{cP_2^{1-\theta}(x)}{P_1^{1-\theta}(x)} \\ &= \log c + (1-\theta) \sum_x P^*(x) \log \frac{P_2(x)}{P_1(x)} \\ D(P^*\|P_2) &= \sum P^*(x) \log \frac{cP_1^\theta(x)}{P_2^\theta(x)} \\ &= \log c - \theta \sum_x P^*(x) \log \frac{P_2(x)}{P_1(x)}. \end{aligned}$$

By assumption, $D(P^*\|P_1) = D(P^*\|P_2) = \gamma$, hence it follows that

$$\sum_x P^*(x) \log \frac{P_2(x)}{P_1(x)} = 0 \tag{32}$$

But this means exactly that $(d/d\theta) \sum_x P_1^\theta(x) P_2^{1-\theta}(x) = 0$, hence the θ for which (**??**) holds actually minimizes the convex function $\sum_x P_1^\theta(x) P_2^{1-\theta}(x)$. Since we also have $\gamma = \log c = -\log \sum_x P_1^\theta(x) P_2^{1-\theta}(x)$, we must have the desired result

$$\gamma = -\log \min_{0 \le \theta \le 1} \sum_x P_1^\theta(x) P_2^{1-\theta}(x).$$

14 Solutions: Homework.2.

1. The necessary and sufficient condition is that the "divergence balls" $\Pi_i = \{Q\colon D(Q\|P_i) < \gamma\}$ have to be disjoint, which follows from previous homework problems.

2. The assertion follows from the fact that the ML estimate from an exponential family equals the I-projection of any element of the exponential family onto the corresponding linear family (containing the empirical distribution), and from the transitivity property of I-projections.

3. In general, maximizing $H(P)$ is the same as minimizing $D(P\|Q_0)$ where Q_0 is the uniform distribution. In our case we must have $P(i,i) = 0$, for every i; therefore, maximizing $H(P)$ is equivalent to minimizing $D(P\|Q_0)$ where $Q(i,i) = 0, \forall i$, and $Q(i,j) =$ constant, $i \neq j$. This minimization can be performed by iteratively adjusting the marginals (iterative scaling).

   The exponential family will consist of all distributions of the form $P(i,j) = cQ(i,j)a(i)b(j)$ and the maximum entropy distribution will be the ML estimate for this family. In this case the exponential family through Q_0, the uniform distribution, is not appropriate because it does not intersect the set of feasible distributions, all of which have their diagonal elements equal to 0.

4. For each pair (i,j), $1 \leq i \leq 5, 1 \leq j \leq n$, let $\mathcal{L}_{ij}^{(1)}$ denote the set of all joint distributions on X^{25} whose two-dimensional marginal representing the joint distribution of X_{ij} and $X_{i(j+1)}$ equals the given one. Similarly, let $\mathcal{L}_{ij}^{(2)}$, $1 \leq i \leq n, 1 \leq j \leq 5$, be defined by the given joint distribution of X_{ij} and $X_{(i+1)j}$. Let $\mathcal{L}$ be the intersection of all these linear linear families and let P_0 be the uniform distribution on X^{25}. The required maximum entropy joint distribution will be the I-projection of P_0 on $\mathcal{L}$. It can be computed by iterative scaling, performing cyclically I-projections on the sets $\mathcal{L}_{ij}^{(1)}$ and $\mathcal{L}_{ij}^{(2)}$ (by adjusting the corresponding two-dimensional marginals.) Since $\mathcal{L}$ is the intersection of 40 sets $\mathcal{L}_{ij}^{(1)}$ and $\mathcal{L}_{ij}^{(2)}$, one cycle of the iteration will consist of 40 consecutive scalings.

5. From Section 6,

$$\nu(p) = \frac{\Gamma\left(\sum_{i=1}^k \alpha_i + k\right)}{\prod_{i=1}^k \Gamma(\alpha_i + 1)} \prod_{i=1}^k p_i^{\alpha_i},$$

   is a density (for every $\alpha_1, \ldots, \alpha_k$ greater than -1), hence its integral over the probability simplex is 1. Applying this with $k = 2$ and with n_0 and $n_1 = n - n_0$ in the role of α_1 and α_2, it follows that

$$\begin{aligned} Q(x_1^n) &= \int p^{n_0}(1-p)^{n_1} dp \\ &= \frac{\Gamma(n_0+1)\Gamma(n_1+1)}{\Gamma(n+2)} = \frac{n_0! n_1!}{(n+1)!} \\ &= \frac{1}{n+1}\frac{n_0! n_1!}{n!}. \end{aligned}$$

   If $n_0 \sim \alpha n$ then Stirling's formula, $k! \sim k^k e^{-k}\sqrt{2\pi k}$, gives

$$\frac{n_0! n_1!}{n!} \sim \frac{n_0^{n_0} n_1^{n_1}}{n^n}\sqrt{2\pi n\alpha(1-\alpha)}$$

   which implies that

$$\log \frac{P_{ML}(x_1^n)}{Q(x_1^n)} = \log\left(\frac{n_0^{n_0} n_1^{n_1}}{n^n} / Q(x_1^n)\right) \sim \frac{1}{2}\log n + \text{const}$$

   If, on the other hand, n_0 is a constant, then

$$Q(x_1^n) = \frac{1}{n+1}\frac{n_0! n_1!}{n!} = \frac{n_0!}{(n+1)n\cdots(n-n_0+1)}$$

   and

$$\begin{aligned} &\frac{P_{ML}(x_1^n)}{Q(x_1^n)} = \\ &\quad = \frac{n_0^{n_0}}{n_0!} \cdot \frac{(n-n_0)^{n-n_0}}{n^n} \cdot (n+1)n\cdots(n-n_0+1) \\ &\quad = \frac{n_0^{n_0}}{n_0!} \cdot (1-\frac{n_0}{n})^{n-n_0} \times \\ &\qquad \times \left[(1-\frac{1}{n})\cdots(1-\frac{n_0-1}{n})\right](n+1), \end{aligned}$$

   so that in this case,

$$\log \frac{P_{ML}(x_1^n)}{Q(x_1^n)} \sim \log n + \text{ const.}$$

   A better choice for Q is the mixture with respect to the Dirichlet prior with $\alpha_1 = \alpha_2 = -1/2$, i. e., with $\nu(p) = 1/(\pi\sqrt{p(1-p)})$, for then $\log \frac{P_{ML}(x_1^n)}{Q(x_1^n)}$ will be asymptotically $(1/2)\log n +$ constant, no matter what is x_1^n.

6. (i) From formula (17) in the lecture notes with $k = 2$ we have the auxiliary distribution

$$Q(x_1^n) = \frac{(n_0 - \frac{1}{2})(n_0 - \frac{3}{2})\cdots\frac{1}{2}\cdot(n_1 - \frac{1}{2})(n_1 - \frac{3}{2})\cdots\frac{1}{2}}{n!}$$

   With $n = 32, n_0 = 22, n_1 = 10$, we obtain $L(x_1^n) = \lceil -\log Q(x_1^n)\rceil = 32$.

   In the Markov case, the formula on page 16 of the notes yields

$$\begin{aligned} &Q(x_1^n) = \\ &\quad \frac{(n(0,0)-1/2)\ldots(1/2)\cdot(n(0,1)-1/2)\ldots(1/2)}{n_0!} \\ &\quad \times \frac{(n(1,0)-1/2)\ldots(1/2)\cdot(n(1,1)-1/2)\ldots(1/2)}{n_1!}. \end{aligned}$$

In our case, setting the unspecified initial state equal to 0, we have $n_{oo} = 16, n_{01} = 6, n_{10} = 6, n_{11} = 4$ and $n_0 = 22, n_1 = 10$. (Note that the present n_0 and n_1 equal those in part (i) only because the initial state has been set equal to the last bit of the sequence x_1^n; otherwise there would be a difference of 1.) Substituting these values we obtain $L(x_1^n) = \lceil -\log Q(x_1^n) \rceil = 33$.

Remark. The perhaps surprising result is that for the given sequence neither method leads to compression. This is so in spite of the fact that the first order empirical entropy $\hat{H}_n$ is clearly less than 1, and the second order empirical entropy $\hat{H}_n^{(2)}$ is clearly smaller than $\hat{H}_n$. The reason is that the true codelength is not $\hat{H}_n$ (or $\hat{H}_n^{(2)}$, respectively), rather, an additional term $(1/2)\log n$ (or $\log n$, respectively), has to be added which stands for the description of the ML distribution.

7. Let $r > 0$ be any number such that for some prefix code with word length function $L(x_1^n)$ we have

$$L(x_1^n) + \log P_\theta(x_1^n) \le r,\ \theta \in \Theta,\ x_1^n \in A^n.$$

Thus $\log P_\theta(x_1^n) \le -L(x_1^n) + r$, so that

$$\sup_{\theta\in\Theta} P_\theta(x_1^n) \le 2^{-L(x_1^n)} 2^r.$$

Summing over x_1^n and using the Kraft inequality then gives $S_n \le 2^r$, so that $r_n \ge \log S_n$.

On the other hand, for the Shannon code with respect to the auxiliary distribution $Q(x_1^n) = S_n^{-1} \sup_{\theta\in\Theta} P_\theta(x_1^n)$, we have

$$L(x_1^n) + \log P_\theta(x_1^n) \le \log \frac{P_\theta(x_1^n)}{Q(x_1^n)} + 1 \le \log S_n + 1,$$

which proves that $r_n \le \log S_n + 1$.

8. The first inequality is trivial because $n\hat{H}_n = -\log P_{ML}(x_1^n)$. Consider the mixture distribution Q with respect to the Dirichlet prior with $\alpha_i \equiv -1/2$. Theorem 17 gives

$$\log \frac{P_{ML}(x_1^n)}{Q(x_1^n)} \le \frac{k-1}{2}\log n + \text{ const.}$$

Combining this with $n\hat{H}_n = -\log P_{ML}(x_1^n)$ then yields

$$\log \frac{1}{Q(x_1^n)} \le n\hat{H}_n + \frac{k-1}{2}\log n + \text{ const.}$$

On the other hand, the (pointwise) redundancy of the Shannon code with respect to Q, though it might be negative for some x_1^n, is lower bounded by a random variable that has finite expectation, (Corollary 3). Thus $-\log Q(x_1^n) \ge -\log P^*(x_1^n) - Y$, where $E(Y)$ is finite. Putting this together with the preceeding inequality yields the bound

$$\log \frac{1}{P^*(x_1^n)} \le n\hat{H}_n + \frac{k-1}{2}\log n + \text{ const.} + Y,$$

which completes the proof.

Remark. It follows in a similar manner that if $X_1, X_2, \ldots,$ is an m-th order Markov chain (with arbitrarily specified states at times $0, -1, \ldots, -m+1$) then for the m-th order emprical (conditional) entropy $\hat{H}_n^m$ we have

$$\begin{aligned} n\hat{H}_n^m &\le -\log P^*(x_1^n) \\ &\le n\hat{H}_n^m + \frac{|A|^m(|A|-1)}{2}\log n + \text{ const.} + Z, \end{aligned}$$

where Z is a random variable not depending on n, whose expectation is finite. If $X_1, X_2, \ldots$ is Markov of order ℓ, then it is also Markov of order $m > \ell$ and hence

$$\begin{aligned} n\hat{H}_n^\ell &\le -\log P^*(x_1^n) \\ &\le n\hat{H}_n^m + \frac{|A|^m(|A|-1)}{2}\log n + \text{ const.} + Z_m, \end{aligned}$$

so that,

$$\hat{H}_n^\ell - \hat{H}_n^m \le \frac{|A|^m(|A|-1)}{2n}\log n + \text{ const.} + \frac{1}{n}Z_m.$$

This allows us to check if the Markov chain is of order $\ell < m$. Here Z_m can be positive or negative, but since it has finite expected value, we can use the Markov inequality to get bounds on the probability that $Z_m > \epsilon > 0$ and use this in the above.

14.1 Corrections.

Line $n+$ is the n-th line from the top and line $n-$ is the n-th line from the bottom.

1. Page 2, column 2, line 20+: Change j_{s+1} to $1 + j_s$.
2. Page 3, column 1, line 19+: Change $P(\hat{P}_n \in \Pi)$ to $P(\hat{P}_n \in \Pi_n)$.
3. Page 3, column 1, line 8-: Change $(1/n)\sum_a f(a) > \alpha$ to $(1/n)\sum_i f(x_i) > \alpha$.

4. Page 4, column 1, line 12+: The minimum should be over $P \in \Pi$, not $P^* \in \Pi$.

5. Page 7, column 2, line 18+: $\gamma = (j_1, \ldots, j_d)$ should be $\omega = (j_1, \ldots, j_d)$

6. Page 8, column 1, line 17-: $\log \hat{P}(\omega_0)/P(\omega_0)$ should be $\log \frac{1-\hat{P}(\omega_0)}{1-P(\omega_0)}$.

7. Page 11, column 1, lines 2- and 11-: Replace $\sum_{x_1^n \in B_n(c)} 2^{-L(x_1^n)}$ by $\sum_{n=1}^{\infty} \sum_{x_1^n \in B_n(c)} 2^{-L(x_1^n)}$.

8. Page 11, column 2, line 16+: Replace $Z_n = P(x_1^n)/Q(x_1^n)$ by $Z_n = Q(x_1^n)/P(x_1^n)$.

9. Page 11, column 2, line 16-: Replace $P(\tilde{A})$ by $P(\tilde{A}_m)$ and $Q(\tilde{A})$ by $Q(\tilde{A}_m)$.

10. Page 12, column 1, formula (11): In the integral exponent the logarithm should be multiplied by $1/n$.

11. Page 12, column 1, formula (12): Replace $P \in$ by $U \in$.

12. Page 12, column 1, line 19-: Replace $P \in N_\infty$ by $U \in N_\epsilon$.

13. Page 12, column 1, line 3-: Replace "code" by "process U".

14. Page 14, column 1, formula (14): Replace $i + 1$ by $i = 1$ in the product.

15. Page 15, column 2, line 14-: Replace $\log \frac{e}{2}$ by $\log \frac{\epsilon}{2}$.