Part I

Information measures in simple coding problems

1 Source coding and hypothesis testing; information measures

A (discrete) *source* is a sequence $\{X_i\}_{i=1}^{\infty}$ of random variables (RVs) taking values in a finite set X called the *source alphabet*. If the X_i's are independent and have the same distribution P, we speak of a *discrete memoryless source* (DMS) with *generic distribution* P.

A k-to-n binary *block code* is a pair of mappings

$$f : \mathsf{X}^k \to \{0,1\}^n, \quad \varphi : \{0,1\}^n \to \mathsf{X}^k.$$

For a given source, the *probability of error* of the code (f, φ) is

$$e(f,\varphi) \triangleq \Pr\{\varphi(f(X^k)) \neq X^k\},$$

where X^k stands for the k-length initial string of the sequence $\{X_i\}_{i=1}^{\infty}$. We are interested in finding codes with small ratio n/k and small probability of error. → 1.1

More exactly, for every k let $n(k,\varepsilon)$ be the smallest n for which there exists a k-to-n binary block code satisfying $e(f,\varphi) \leqq \varepsilon$; we want to determine $\lim_{k\to\infty} \frac{n(k,\varepsilon)}{k}$. → 1.2

THEOREM 1.1 For a DMS with generic distribution $P = \{P(x) : x \in \mathsf{X}\}$

$$\lim_{k\to\infty} \frac{n(k,\varepsilon)}{k} = H(P) \quad \text{for every} \quad \varepsilon \in (0,1), \tag{1.1}$$

where $H(P) \triangleq -\sum_{x\in\mathsf{X}} P(x)\log P(x)$. ○

COROLLARY 1.1

$$0 \leqq H(P) \leqq \log|\mathsf{X}|. \tag{1.2}$$

○

Proof The existence of a k-to-n binary block code with $e(f,\varphi) \leqq \varepsilon$ is equivalent to the existence of a set $\mathsf{A} \subset \mathsf{X}^k$ with $P^k(\mathsf{A}) \geqq 1-\varepsilon$, $|\mathsf{A}| \leqq 2^n$ (let A be the set of those sequences $\mathbf{x} \in \mathsf{X}^k$ which are reproduced correctly, i.e., $\varphi(f(\mathbf{x})) = \mathbf{x}$). Denote by $s(k,\varepsilon)$ the minimum cardinality of sets $\mathsf{A} \subset \mathsf{X}^k$ with $P^k(\mathsf{A}) \geqq 1-\varepsilon$. It suffices to show that

$$\lim_{k\to\infty} \frac{1}{k}\log s(k,\varepsilon) = H(P) \quad (\varepsilon \in (0,1)). \tag{1.3}$$

To this end, let $\mathsf{B}(k,\delta)$ be the set of those sequences $\mathbf{x} \in \mathsf{X}^k$ which have probability

$$\exp\{-k(H(P)+\delta)\} \leqq P^k(\mathbf{x}) \leqq \exp\{-k(H(P)-\delta)\}.$$

We first show that $P^k(\mathsf{B}(k,\delta)) \to 1$ as $k \to \infty$, for every $\delta > 0$. In fact, consider the real-valued RVs

$$Y_i \triangleq -\log P(X_i);$$

these are well defined with probability 1 even if $P(x) = 0$ for some $x \in \mathsf{X}$. The Y_i's are independent, identically distributed and have expectation $H(P)$. Thus by the weak law of large numbers

$$\lim_{k\to\infty} \Pr\left\{\left|\frac{1}{k}\sum_{i=1}^{k} Y_i - H(P)\right| \leqq \delta\right\} = 1 \quad \text{for every} \quad \delta > 0.$$

As $X^k \in \mathsf{B}(k,\delta)$ iff $|\frac{1}{k}\sum_{i=1}^{k} Y_i - H(P)| \leqq \delta$, the convergence relation means that

$$\lim_{k\to\infty} P^k(\mathsf{B}(k,\delta)) = 1 \quad \text{for every} \quad \delta > 0, \tag{1.4}$$

as claimed. The definition of $\mathsf{B}(k,\delta)$ implies that

$$|\mathsf{B}(k,\delta)| \leqq \exp\{k(H(P)) + \delta)\}.$$

Thus (1.4) gives for every $\delta > 0$

$$\overline{\lim_{k\to\infty}}\ \frac{1}{k}\log s(k,\varepsilon) \leqq \overline{\lim_{k\to\infty}}\ \frac{1}{k}\log|\mathsf{B}(k,\delta)| \leqq H(P) + \delta. \tag{1.5}$$

On the other hand, for every set $\mathsf{A} \subset \mathsf{X}^k$ with $P^k(\mathsf{A}) \geqq 1 - \varepsilon$, (1.4) implies

$$P^k(\mathsf{A} \cap \mathsf{B}(k,\delta)) \geqq \frac{1-\varepsilon}{2}$$

for sufficiently large k. Hence, by the definition of $\mathsf{B}(k,\delta)$,

$$\begin{aligned}|\mathsf{A}| \geqq |\mathsf{A} \cap \mathsf{B}(k,\delta)| &\geqq \sum_{\mathbf{x}\in\mathsf{A}\cap\mathsf{B}(k,\delta)} P^k(\mathbf{x})\exp\{k(H(P)-\delta)\} \\ &\geqq \frac{1-\varepsilon}{2}\exp\{k(H(P)-\delta)\},\end{aligned}$$

proving that for every $\delta > 0$

$$\underline{\lim_{k\to\infty}}\ \frac{1}{k}\log s(k,\varepsilon) \geqq H(P) - \delta.$$

This and (1.5) establish (1.3). The corollary is immediate. □

For intuitive reasons expounded in the Introduction, the limit $H(P)$ in Theorem 1.1 is interpreted as a measure of the information content of (or the uncertainty about) a RV X with distribution $P_X = P$. It is called the *entropy* of the RV X or of the distribution P:

$$H(X) = H(P) \triangleq -\sum_{x\in\mathsf{X}} P(x)\log P(x).$$

This definition is often referred to as *Shannon's formula*.

The mathematical essence of Theorem 1.1 is formula (1.3). It gives the asymptotics for the minimum size of sets of large probability in X^k. We now generalize (1.3) for the case when the elements of X^k have unequal weights and the size of subsets is measured by total weight rather than cardinality.

Let us be given a sequence of positive-valued "mass functions" $M_1(x), M_2(x), \ldots$ on X and set

$$M(\mathbf{x}) \triangleq \prod_{i=1}^{k} M_i(x_i) \quad \text{for} \quad \mathbf{x} = x_1 \cdots x_k \in \mathsf{X}^k.$$

For an arbitrary sequence of X-valued RVs $\{X_i\}_{i=1}^{\infty}$ consider the minimum of the M-mass

$$M(\mathsf{A}) \triangleq \sum_{\mathbf{x} \in \mathsf{A}} M(\mathbf{x})$$

of those sets $\mathsf{A} \subset \mathsf{X}^k$ which contain X^k with high probability: let $s(k, \varepsilon)$ denote the minimum of $M(\mathsf{A})$ for sets $\mathsf{A} \subset \mathsf{X}^k$ of probability

$$P_{X^k}(\mathsf{A}) \geqq 1 - \varepsilon.$$

The previous $s(k, \varepsilon)$ is a special case obtained if all the functions $M_i(x)$ are identically equal to 1.

THEOREM 1.2 If the X_i's are independent with distributions $P_i \triangleq P_{X_i}$ and $|\log M_i(x)| \leqq c$ for every i and $x \in \mathsf{X}$ then, setting

$$E_k \triangleq \frac{1}{k} \sum_{i=1}^{k} \sum_{x \in \mathsf{X}} P_i(x) \log \frac{M_i(x)}{P_i(x)},$$

we have for every $0 < \varepsilon < 1$

$$\lim_{k \to \infty} \left(\frac{1}{k} \log s(k, \varepsilon) - E_k \right) = 0.$$

More precisely, for every $\delta, \varepsilon \in (0, 1)$,

$$\left| \frac{1}{k} \log\, s(k, \varepsilon) - E_k \right| \leqq \delta \quad \text{if} \quad k \geqq k_0 = k_0(|\mathsf{X}|, c, \varepsilon, \delta). \tag{1.6}$$

○

Proof Consider the real-valued RVs

$$Y_i \triangleq \log \frac{M_i(X_i)}{P_i(X_i)}.$$

Since the Y_i's are independent and $E\left(\frac{1}{k}\sum_{i=1}^{k} Y_i\right) = E_k$, Chebyshev's inequality gives for any $\delta' > 0$

$$\Pr\left\{ \left| \frac{1}{k} \sum_{i=1}^{k} Y_i - E_k \right| \geqq \delta' \right\} \leqq \frac{1}{k^2 \delta'^2} \sum_{i=1}^{k} \operatorname{var}(Y_i) \leqq \frac{1}{k\delta'^2} \max_i \operatorname{var}(Y_i).$$

This means that for the set

$$\mathsf{B}(k,\delta') \triangleq \left\{\mathbf{x} : \mathbf{x} \in \mathsf{X}^k,\ E_k - \delta' \leqq \frac{1}{k}\log\frac{M(\mathbf{x})}{P_{X^k}(\mathbf{x})} \leqq E_k + \delta'\right\}$$

we have

$$P_{X^k}(\mathsf{B}(k,\delta')) \geqq 1 - \eta_k, \quad \text{where} \quad \eta_k \triangleq \frac{1}{k\delta'^2}\max_i \operatorname{var}(Y_i).$$

Since by the definition of $\mathsf{B}(k,\delta')$

$$M(\mathsf{B}(k,\delta')) = \sum_{\mathbf{x}\in\mathsf{B}(k,\delta')} M(\mathbf{x}) \leqq \sum_{\mathbf{x}\in\mathsf{B}(k,\delta')} P_{X^k}(\mathbf{x})\exp[k(E_k+\delta')] \leqq \exp[k(E_k+\delta')],$$

it follows that

$$\frac{1}{k}\log s(k,\varepsilon) \leqq \frac{1}{k}\log M(\mathsf{B}(k,\delta')) \leqq E_k + \delta' \quad \text{if} \quad \eta_k \leqq \varepsilon.$$

On the other hand, we have $P_{X^k}(\mathsf{A}\cap\mathsf{B}(k,\delta')) \geqq 1-\varepsilon-\eta_k$ for any set $\mathsf{A}\subset\mathsf{X}^k$ with $P_{X^k}(\mathsf{A}) \geqq 1-\varepsilon$. Thus for every such A, again by the definition of $\mathsf{B}(k,\delta')$,

$$\begin{aligned} M(\mathsf{A}) \geqq M(\mathsf{A}\cap\mathsf{B}(k,\delta')) &\geqq \sum_{\mathbf{x}\in\mathsf{A}\cap\mathsf{B}(k,\delta')} P_{X^k}(\mathbf{x})\exp\{k(E_k-\delta')\} \\ &\geqq (1-\varepsilon-\eta_k)\exp[(E_k-\delta')], \end{aligned}$$

implying

$$\frac{1}{k}\log s(k,\varepsilon) \geqq \frac{1}{k}\log(1-\varepsilon-\eta_k) + E_k + \delta'.$$

Setting $\delta' \triangleq \delta/2$, these results imply (1.6) provided that

$$\eta_k = \frac{4}{k\delta^2}\max_i \operatorname{var}(Y_i) \leqq \varepsilon \quad \text{and} \quad \frac{1}{k}\log(1-\varepsilon-\eta_k) \geqq -\frac{\delta}{2}.$$

By the assumption $|\log M_i(x)| \leqq c$, the last relations hold if $k \geqq k_0(|\mathsf{X}|, c, \varepsilon, \delta)$. □

An important corollary of Theorem 1.2 relates to *testing statistical hypotheses*. Suppose that a probability distribution of interest for the statistician is given by either $P = \{P(x) : x \in \mathsf{X}\}$ or $Q = \{Q(x) : x \in \mathsf{X}\}$. She or he has to decide between P and Q on the basis of a *sample* of size k, i.e., the result of k independent drawings from the unknown distribution. A (non-randomized) test is characterized by a set $\mathsf{A}\subset\mathsf{X}^k$, in the sense that if the sample $X_1\ldots X_k$ belongs to A, the statistician accepts P and else accepts Q. In most practical situations of this kind, the role of the two hypotheses is not symmetric. It is customary to prescribe a bound ε for the tolerated probability of wrong decision if P is the true distribution. Then the task is to minimize the probability of a wrong decision if hypothesis Q is true. The latter minimum is

→ 1.3

→ 1.4

$$\beta(k,\varepsilon) \triangleq \min_{\substack{\mathsf{A}\subset\mathsf{X}^k \\ P^k(\mathsf{A})\geqq 1-\varepsilon}} Q^k(\mathsf{A}).$$

COROLLARY 1.2 For any $0 < \varepsilon < 1$,

$$\lim_{k\to\infty} \frac{1}{k} \log \beta(k, \varepsilon) = -\sum_{x\in\mathsf{X}} P(x) \log \frac{P(x)}{Q(x)}. \qquad \bigcirc$$

Proof If $Q(x) > 0$ for each $x \in \mathsf{X}$, set $P_i \triangleq P$, $M_i \triangleq Q$ in Theorem 1.2. If $P(x) > Q(x) = 0$ for some $x \in \mathsf{X}$, the P-probability of the set of all k-length sequences containing this x tends to 1. This means that $\beta(k, \varepsilon) = 0$ for sufficiently large k, so that both sides of the asserted equality are $-\infty$. □

It follows from Corollary 1.2 that the sum on the right-hand side is non-negative. It measures how much the distribution Q differs from P in the sense of statistical distinguishability, and is called *informational divergence* or *I-divergence*:

$$D(P||Q) \triangleq \sum_{x\in\mathsf{X}} P(x) \log \frac{P(x)}{Q(x)}.$$

Another common name given to this quantity is *relative entropy*. Intuitively, one can say that the larger $D(P||Q)$ is, the more information for discriminating between the hypotheses P and Q can be obtained from one observation. Hence $D(P||Q)$ is also called the *information for discrimination*. The amount of information measured by $D(P||Q)$ is, however, conceptually different from entropy, since it has no immediate coding interpretation.

On the space of infinite sequences of elements of X one can build up product measures both from P and Q. If $P \neq Q$, the two product measures are mutually orthogonal; $D(P||Q)$ is a (non-symmetric) measure of how fast their restrictions to k-length strings approach orthogonality.

REMARK Both entropy and informational divergence have a form of expectation:

$$H(X) = E(-\log P(X)), \quad D(P||Q) = E \log \frac{P(X)}{Q(X)},$$

where X is a RV with distribution P. It is convenient to interpret $-\log P(x)$, resp. $\log P(x)/Q(x)$, as a measure of the amount of information, resp. the weight of evidence in favor of P against Q provided by a particular value x of X. These quantities are important ingredients of the mathematical framework of information theory, but have less direct operational meaning than their expectations. $\bigcirc$

The entropy of a pair of RVs (X, Y) with finite ranges X and Y needs no new definition, since the pair can be considered a single RV with range $\mathsf{X} \times \mathsf{Y}$. For brevity, instead of $H((X, Y))$ we shall write $H(X, Y)$; similar notation will be used for any finite collection of RVs.

The intuitive interpretation of entropy suggests to consider as further information measures certain expressions built up from entropies. The difference $H(X, Y) - H(X)$ measures the additional amount of information provided by Y if X is already known.

It is called the *conditional entropy* of Y given X:

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

Expressing the entropy difference by Shannon's formula we obtain

$$H(Y|X) = -\sum_{x\in\mathsf{X}}\sum_{y\in\mathsf{Y}} P_{XY}(x,y)\log\frac{P_{XY}(x,y)}{P_X(x)} = \sum_{x\in\mathsf{X}} P_X(x)H(Y|X=x), \quad (1.7)$$

where

$$H(Y|X=x) \stackrel{\triangle}{=} -\sum_{y\in\mathsf{Y}} P_{Y|X}(y|x)\log P_{Y|X}(y|x).$$

Thus $H(Y|X)$ is the expectation of the entropy of the conditional distribution of Y given $X = x$. This gives further support to the above intuitive interpretation of conditional entropy. Intuition also suggests that the conditional entropy cannot exceed the unconditional one.

→ 1.5

LEMMA 1.3

$$H(Y|X) \leqq H(Y). \qquad \bigcirc$$

Proof

$$H(Y) - H(Y|X) = H(Y) - H(X,Y) + H(X)$$
$$= \sum_{x\in\mathsf{X}}\sum_{y\in\mathsf{Y}} P_{XY}(x,y)\log\frac{P_{XY}(x,y)}{P_X(x)P_Y(y)} = D(P_{XY}\|P_X\times P_Y) \geqq 0. \qquad \square$$

REMARK For certain values of x, $H(Y|X=x)$ may be larger than $H(Y)$. $\bigcirc$

The entropy difference in the preceding proof measures the decrease of uncertainty about Y caused by the knowledge of X. In other words, it is a measure of the amount of information about Y contained in X. Note the remarkable fact that this difference is symmetric in X and Y. It is called *mutual information*:

$$I(X\wedge Y) \stackrel{\triangle}{=} H(Y) - H(Y|X) = H(X) - H(X|Y) = D(P_{XY}\|P_X\times P_Y). \quad (1.8)$$

Of course, the amount of information contained in X about itself is just the entropy:

$$I(X\wedge X) = H(X).$$

Mutual information is a measure of stochastic dependence of the RVs X and Y. The fact that $I(X\wedge Y)$ equals the informational divergence of the joint distribution of X and Y from what it would be if X and Y were independent reinforces this interpretation. There is no compelling reason other than tradition to denote mutual information by a different symbol than entropy. We keep this tradition, although our notation $I(X\wedge Y)$ differs slightly from the more common $I(X;Y)$.

Discussion

Theorem 1.1 says that the minimum number of binary digits needed – on average – to represent one symbol of a DMS with generic distribution P equals the entropy $H(P)$. This fact – and similar ones discussed later on – are our basis for interpreting $H(X)$ as a measure of the amount of information contained in the RV X, resp. of the uncertainty about this RV. In other words, in this book we adopt an operational or *pragmatic approach* to the concept of information. Alternatively, one could start from the intuitive concept of information and set up certain postulates which an information measure should fulfil. Some representative results of this *axiomatic approach* are treated in Problems 1.11–1.14.

Our starting point, Theorem 1.1, has been proved here in the conceptually simplest way. The key idea is that, for large k, all sequences in a subset of X^k with probability close to 1, namely $\mathsf{B}(k, \delta)$, have "nearly equal" probabilities in an exponential sense. This proof easily extends also to non-DM cases (not in the scope of this book).

On the other hand, in order to treat DM models at depth, another – purely combinatorial – approach will be more suitable. The preliminaries to this approach will be given in Chapter 2.

Theorem 1.2 demonstrates the intrinsic relationship of the basic source coding and hypothesis testing problems. The interplay of information theory and mathematical statistics goes much further; its more substantial examples are beyond the scope of this book. ○

Problems

1.1. (a) Check that the problem of determining $\lim_{k\to\infty} \frac{1}{k} n(k, \varepsilon)$ for a discrete source is just the formal statement of the LMTR problem (see the Introduction) for the given source and the binary noiseless channel, with the probability of error fidelity criterion.

(b) Show that for a DMS and a noiseless channel with arbitrary alphabet size m the LMTR is $H(P)/\log m$, where P is the generic distribution of the source.

1.2. Given an encoder $f : \mathsf{X}^k \to \{0, 1\}^n$, show that the probability of error $e(f, \varphi)$ is minimized iff the decoder $\varphi : \{0, 1\}^n \to \mathsf{X}^k$ has the property that $\varphi(\mathbf{y})$ is a sequence of maximum probability among those $\mathbf{x} \in \mathsf{X}^k$ for which $f(\mathbf{x}) = \mathbf{y}$.

1.3. A *randomized test* introduces a chance element into the decision between the hypotheses P and Q in the sense that if the result of k successive drawings is $\mathbf{x} \in \mathsf{X}^k$, one accepts the hypothesis P with probability $\pi(\mathbf{x})$, say. Define the analog of $\beta(k, \varepsilon)$ for randomized tests and show that it still satisfies Corollary 1.2.

1.4. (*Neyman–Pearson lemma*) Show that for any given bound $0 < \varepsilon < 1$ on the probability of wrong decision if P is true, the best randomized test is given by

$$\pi(\mathbf{x}) = \begin{cases} 1 & \text{if } P^k(\mathbf{x}) > c_k Q^k(\mathbf{x}) \\ \gamma_k & \text{if } P^k(\mathbf{x}) = c_k Q^k(\mathbf{x}) \\ 0 & \text{if } P^k(\mathbf{x}) < c_k Q^k(\mathbf{x}), \end{cases}$$

where c_k and γ_k are appropriate constants. Observe that the case $k = 1$ contains the general one, and there is no need to restrict attention to independent drawings.

1.5. (a) Let $\{X_i\}_{i=1}^{\infty}$ be a sequence of independent RVs with common range X but with arbitrary distributions. As in Theorem 1.1, denote by $n(k, \varepsilon)$ the smallest n for which there exists a k-to-n binary block code having probability of error $\leqq \varepsilon$ for the source $\{X_i\}_{i=1}^{\infty}$. Show that for every $\varepsilon \in (0, 1)$ and $\delta > 0$

$$\left| \frac{n(k, \varepsilon)}{k} - \frac{1}{k} \sum_{i=1}^{k} H(X_i) \right| \leqq \delta \quad \text{if} \quad k \geqq k_0(|\mathsf{X}|, \varepsilon, \delta).$$

Hint Use Theorem 1.2 with $M_i(x) = 1$.

(b) Let $\{(X_i, Y_i)\}_{i=1}^{\infty}$ be a sequence of independent replicas of a pair of RVs (X, Y) and suppose that X^k should be encoded and decoded in the knowledge of Y^k. Let $\tilde{n}(k, \varepsilon)$ be the smallest n for which there exists an encoder $f : \mathsf{X}^k \times \mathsf{Y}^k \to \{0, 1\}^n$ and a decoder $\varphi : \{0, 1\}^n \times \mathsf{Y}^k \to \mathsf{X}^k$ such that the probability of error is $\Pr\{\varphi(f(X^k, Y^k), Y^k) \neq X^k\} \leqq \varepsilon$.

Show that

$$\lim_{k \to \infty} \frac{\tilde{n}(k, \varepsilon)}{k} = H(X|Y) \quad \text{for every} \quad \varepsilon \in (0, 1).$$

Hint Use part (a) for the conditional distributions of the X_i's given various realizations $\mathbf{y}$ of Y^k.

1.6. (*Random selection of codes*) Let $\mathcal{F}(k, n)$ be the class of all mappings $f : \mathsf{X}^k \to \{0, 1\}^n$. Given a source $\{X_i\}_{i=1}^{\infty}$, consider the class of codes (f, φ_f), where f ranges over $\mathcal{F}(k, n)$ and $\varphi_f : \{0, 1\}^n \to \mathsf{X}^k$ is defined so as to minimize $e(f, \varphi)$; see Problem 1.2. Show that for a DMS with generic distribution P we have

$$\frac{1}{|\mathcal{F}(k, n)|} \sum_{f \in \mathcal{F}(k,n)} e(f, \varphi_f) \to 0,$$

if k and n tend to infinity, so that

$$\inf \frac{n}{k} > H(P).$$

Hint Consider a random mapping F of X^k into $\{0, 1\}^n$, assigning to each $\mathbf{x} \in \mathsf{X}^k$ one of the 2^n binary sequences of length n with equal probabilities 2^{-n}, independently of each other and of the source RVs. Let $\Phi : \{0, 1\}^n \to \mathsf{X}^k$ be the random mapping taking the value φ_f if $F = f$. Then

$$\begin{aligned} \frac{1}{|\mathcal{F}(k, n)|} \sum_{f \in \mathcal{F}(k,n)} e(f, \varphi_f) &= \Pr\{\Phi(F(X^k)) \neq X^k\} \\ &= \sum_{\mathbf{x} \in \mathsf{X}^k} P^k(\mathbf{x}) \Pr\{\Phi(F(\mathbf{x})) \neq \mathbf{x}\}. \end{aligned}$$