Elements of Information Theory
2006
Thomas M. Cover and Joy A. Thomas

Chapter 1. Introduction.

Information theory answers two fundamental questions in communication theory: What is the ultimate data compression (answer: the entropy H), and what is the ultimate transmission rate of communication (answer: the channel capacity C). For this reason some consider information theory to be a subset of communication theory. We argue that it is much more. Indeed, it has fundamental contributions to make in statistical physics (thermodynamics), computer science (Kolmogorov complexity or algorithmic complexity), statistical inference (Occam's Razor: "The simplest explanation is best"), and to probability and statistics (error exponents for optimal hypothesis testing and estimation).

Information theory is related to physics (statistical mechanics), mathematics (probability theory), electrical engineering (communication theory), and computer science (algorithmic complexity).

Chapter 2. Entropy, Relative Entropy, and Mutual Information

Entropy is a measure of the uncertainty of a random variable; it is a measure of the amount of information required on the average to describe the random variable.

Definition: The *entropy* H(X) of a discrete random variable X is defined by:

$$H(X) = \sum_{x \in X} p(x) \log p(x)$$

The entropy of X can also be interpreted as the expected value of the random variable $\log \frac{1}{p(X)}$, where X is drawn according to a mass function p(x). Thus

$$H(X) = E_p \log \frac{1}{p(X)}$$

Properties of H

1. $H(X) \geq 0$
2. $H_b(X) = (\log_b a) H_a(X)$
3. (Conditioning reduces entropy) For any two random variables, X and Y, we have $H(X|Y) \leq H(X)$ with equality iff X and Y are independent.
4. $H(X_1, X_2, \ldots, X_n) \leq \sum_{(i=1)}^{n} H(X_i)$ with equality if and only if the Xi are independent.
5. $H(X) \leq \log|\chi|$ With equality if and only if X is distributed uniformly over X
6. H(p) is concave in p.

A function is ***convex*** if it always lies below any chord.

Definition: A function f(x) is said to be *convex* over an interval (a,b) if for every $x_1, x_2 \in (a,b)$ and 0 ≤ λ ≤ 1

$$f(\lambda x_1 + (1-\lambda) x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$$

A function is said to be strictly convex if equality holds only for λ = 0 or λ = 1

Relative entropy is a measure of the distance between two distributions; it is a measure of the inefficiency of assuming that the distribution is q when the true distribution is p. The *relative entropy* or Kullback-Leiber distance is not a true distance between distributions since it is not symmetric and does not satisfy the triangle inequality.

Definition: The *relative entropy* D(p||q) of the

probability mass function p with respect to the probability mass function q is defined by:

$$D(p||q)=\sum_x p(x)\log\frac{p(x)}{q(x)}$$

Mutual Information is a measure of the amount of information that one random variable contains about another random variable; it is a reduction in the uncertainty of one random variable due to the knowledge of another.

Definition: The *mutual information* between two random variables X and Y is defined as:

$$I(X;Y)=\sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}$$

The *mutual information* of a random variable with itself is the entropy of the random variable. This is the reason that entropy is sometimes referred to as self-information.

Alternative expressions.

$$H(X)=E_p\log\frac{1}{p(X)}$$

$$H(X,Y)=E_p\log\frac{1}{p(X,Y)}$$

$$H(X|Y)=E_p\log\frac{1}{p(X|Y)}$$

$$I(X;Y)=E_p\log\frac{p(X,Y)}{p(X)p(Y)}$$

$$D(X||Y)=E_p\log\frac{p(X)}{q(X)}$$

Properties of D and I

I(X;Y) = H(X) – H(X|Y)
= H(Y) – H(Y|X)
= H(X) + H(Y) – H(X,Y)

D(p||q) ≥ 0
with equality if and only if p(x) = q(x), for all x ϵ X

I(X;Y) = D(p(x,y) || p(x)q(y)) ≥ 0,
with equality iff p(x,y) = p(x)p(y) (i.e. X and Y are independent)

If |X|=m, and u is the uniform distribution over X, then D(p||u) = log m – H(p)

D(p||q) is convex in the pair (p,q)

Chain Rules

Entropy:

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

Mutual Information:

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

Relative entropy:

$$D(p(x,y)||q(x,y))=D(p(x)||q(x))+D(p(x|y)||q(y|x))$$

Jensen's inequality. Jensen's inequality is one of the most widely used in mathematics and one that underlies many of the basic results in information theory. If f is a convex function, then

$$Ef(X)\geq f(EX)$$

Log sum inequality. For n positive numbers, a_1, a_2, . . . , a_n and b_1, b_2, . . . , b_n,

$$\sum_{i=1}^n a_i\log\frac{a_i}{b_i}\geq\left(\sum_{i=1}^n a_i\right)\log\frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n b_i}$$

with equality iff $\frac{a_i}{b_i}=constant$

Data processing inequality. The data-processing inequality can be used to show that no cleaver manipulation of the data can improve inferences that can be made from the data.
If X → Y → Z forms a Markov chain,

$$I(X;Y)\geq I(X;Z)$$

Sufficient Statistic. A statistic T(X) is called sufficient for Θ if it contains all the information in X about Θ.
T(X) is sufficient relative to Θ iff I(Θ;X) = I(Θ;T(X)) for all distributions on Θ

A statistic T(X) is a *minimal sufficient statistic* relative to $\{f_\Theta(x)\}$ if it is a function of every other sufficient statistic U.

Fano's inequality. Relates the probability of error in guessing the random variable X to its conditional entropy H(X|Y). Let $P_e=Pr\{\hat{X}(Y)\neq X\}$
Then,

$$H(P_e)+P_e\log|X|\geq H(X|Y)$$

Inequality. If X and X' are independent and identically distributed then

$$PR(X=X')\geq 2^{-H(X)}$$

Chapter 3. Asymptotic Equipartition Property.

AEP "Almost all event are almost equally surprising." Specifically, if X_1, X_2, . . . are i.d.d. ~ p(x), then

$$\frac{-1}{n}\log p(X_{1,}X_{2,}\ldots,X_n)\to H(X)$$

in probability.

Definition: The sequence X1, X2, ... converges to a random variable X:

- *In probability* if for every ϵ>0, $\Pr\{|X_n-X|>\epsilon\} \to 0$
- *In mean square* if $E(X_n-X)^2 \to 0$
- *With probability 1* if $\Pr\{\lim_{n\to\infty} X_n=X\} = 1$

Definition: The typical set $A_\epsilon^{(n)}$ is the set of sequences $x_1, x_2, \ldots, x_n$ satisfying

$$2^{-n(H(X)+\epsilon)}\le p(x_{1,}x_{2,}\ldots,x_n)\le 2^{-n(H(X)-\epsilon)}$$

Properties of the typical set.

If $(x_{1,}x_{2,}\ldots,x_n)\in A_\epsilon^{(n)}$ then $p(x_{1,}x_{2,..},x_n)=2^{-n(H\pm\epsilon)}$

$Pr\{A_\epsilon^{(n)}\}>1-\epsilon$ for n sufficiently large

$|A_\epsilon^{(n)}|\le 2^{n(H(X)+\epsilon)}$, where |A| denotes the number of elements in set A

Thus the typical set has probability nearly 1, all elements of the typical set are nearly equiprobable, and the number of elements in the typical set is nearly 2^{nH}

Definition:

$a_n = b_n$ means that $\frac{1}{n}\log\frac{a_n}{b_n}\to 0$ as $n\to\infty$

Smallest probable set. Let $X_1, X_2, \ldots, X_n$ be i.d.d. ~ p(x), and for δ<½, let $B_\delta^{(n)}\subset X^n$ be the smallest subset such that $Pr\{B_\delta^{(n)}\}\ge 1-\delta$ Then $|B_\delta^{(n)}|=2^{nH}$

The typical set has essentially the same number of elements as the smallest set, to first order in the exponent.

Note that for a binary random variable with Pr(0)=0.1, Pr(1)=0.9, the typical sequences will have proportions close to 1:9 but this does not include the most likely single sequence of straight 1's.

Chapter 4 Entropy Rates of a Stochastic Process.

A stochastic process is said to be ***stationary*** if the joint distribution of any subset of the sequence of random variables is invariant with respect to shifts in the time index.

A ***Markov process*** is a stochastic process in which each random variable depends only on the one preceding it and is conditionally independent of all the other preceding random variables.

A Markov chain is said to be ***time invariant*** if the conditional probability does not depend on n – the position in the chain.

A ***stationary distribution*** on the states of a Markov process is the same for n and n+1.

$$\mu_j=\sum_i \mu_i P_{ij}$$

Entropy Rate. Two definitions of entropy rate for a stochastic process are:

$$H(X)=\lim_{n\to\infty}\frac{1}{n}H(X_{1,}X_{2,}\ldots,X_n)$$

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

The first is the per symbol entropy of the n random variables, and the second is the conditional entropy of the last random variable given the past.

For a stationary stochastic process:

H(X) = H'(X)

Entropy rate of a stationary Markov chain

$$H(X)=-\sum_{ij}\mu_i P_{ij}\log P_{ij}$$

Second law of thermodynamics. For a Markov chain:

1. Relative entropy $D(\mu_n||\mu'_n)$ decreases with time
2. Relative entropy $D(\mu_n||\mu)$ between a distribution and the stationary distribution decreases with time
3. Entropy $H(X_n)$ increases if the stationary distribution is uniform.
4. The conditional entropy $H(X_n|X_1)$ increases with time for a stationary Markov chain.
5. The conditional entropy $H(X_0|X_n)$ of the initial condition X_0 increases for any Markov chain.

Functions of a Markov chain. If X_1, X_2, ... X_n form a stationary Markov chain and Y_i $\Phi(X_i)$, then:

$$H(Y_n|Y_{n-1},\dots Y_1,X_1) \le H(Y) \le H(Y_n|Y_{n-1},\dots Y_1)$$

$$\lim_{n\to\infty} H(Y_n|Y_{n-1},\dots Y_1,X_1) = H(Y) = \lim_{n\to\infty} H(Y_n|Y_{n-1},\dots Y_1)$$

Chapter 5 Data Compression.

A code is ***nonsingular*** if every element of the range of X maps onto a different string in D*. However, delimiter is required to transmit a sequence.

The ***extension*** C* of a code C is the mapping from finite length strings of X to finite-length strings of D where each element is concatenated.

A code is called ***uniquely decodable*** if its extension is non-singular (ie every encoded string in a uniquely decodable code has only one possible source string producing it)

A code is called a ***prefix code*** or an ***instantaneous code*** if no codeword is a prefix of any other codeword.

Kraft inequality. The set of codeword lengths possible for instantaneous codes is limited by:

$$instantaneous\,codes \Leftrightarrow \sum D^{-l_i} \le 1$$

A probability distribution is called ***D-adic*** if each of the probabilities is equal to D^{-n} for some n.

McMillan inequality. The set of codeword lengths possible for uniquely decodable codes is limited by:

$$Uniquely\,decodable\,codes \Leftrightarrow \sum D^{-l_i} \le 1$$

Entropy bound on data compression:

$$L \triangleq \sum p_i l_i \ge H_D(X)$$

Shannon code:

$$l_i = \lceil \log_D \frac{1}{p_i} \rceil$$

$$H_D(X) \le L < H_D(X) + 1$$

where $\lceil x \rceil$ is the smallest integer ≥ x

Huffman code: The Hauffman code is constructed by recursively combining the symbols with the lowest probability into a single source symbol until the problem has reduced to one symbol, then assigning codewords to the symbols.

Code word	X	Probability				
01	1	0.25	0.3	0.45	0.55	1
10	2	0.25	0.25	0.3	0.45	
11	3	0.2	0.25	0.25		
000	4	0.15	2			
001	5	0.15				

$$L^{¿} = \min_{\sum D^{-l_i} \le 1} \sum p_i l_i$$

$$H_D(X) \le L^{¿} < H_D(X) + 1$$

Hauffman coding:

- is optimal
- is equivalent to “20 questions”
- can code weighted codewords
- can be used for slice codes (alphabetic codes)
- can be sub-optimum when used with codeword lengths $\lceil \log \frac{1}{p_i} \rceil$ (Shannon coding)
- is similar to fano coding which divides the symbols into groups of nearly equal combined probability

Wrong code:

$$X \approx p(x), l(x) = \lceil \log \frac{1}{q}(x) \rceil, L = \sum p(x) l(x):$$

$$H(p) + D(p||q) \le L < H(p) + D(p||q) + 1$$

Shannon-Fano-Elias coding assigns codewords to the midpoint of each discrete step in the cumulative distribution function. Unfortunately S-F-E coding calculations grow exponentially with block length and precision grows linearly. Arithmetic coding is an extension which resolves these issues.

Stochastic processes: The number of fair coin flips required to generate a random variable with X drawn according to a specified probability mass function is equal to the Entropy.

$$\frac{H(X_1, X_2, .., X_n)}{n} \le L_n < \frac{H(X_1, X_2, .., X_n)}{n} + \frac{1}{n}$$

Stationary processes $L_n \to H(X)$

Competitive optimality Shannon code $l(x) = \lceil \log \frac{1}{p}(x) \rceil$ versus any other code l'(x):

$$Pr(l(X) \ge l'(X) + c) \le \frac{1}{2^{c-1}}$$

Chapter 6 Gambling and Data Compression.

Doubling Rate: The doubling rate is the rate at which wealth grows when each outcome has probability = p_k, the bet placed on each outcome = b_k and the payout on each outcome is $o_k b_k$

$$W(b, p) = E(\log S(X)) = \sum_{k=1}^{m} p_k \log b_k o_k$$

Optimal doubling rate W*(p) = $\max_b W(b,p)$

Proportional gambling is log optimal. Where the bets placed are in proportion with the probability of the outcome.

$$W*(p) = \max_b W(b, p) = \sum p_i \log o_i - H(p)$$

is achieved by b*=p

Growth rate. Wealth grows as $S_n = 2^{nW*(p)}$

The doubling rate is equal to the difference between the distance of the bookies estimate from the true distribution and the distance of the gamblers estimate from the true distribution

W(b,p) = D(p||r) - D(p||b)

Conservation law. For uniform odds, H(p) + W*(p) = log m

Side Information. In a horse race X, the increase ΔW in doubling rate due to side information Y is $\Delta W = I(X;Y)$

Any sequence on which a gambler makes a large amount of money is also a sequence that can be compressed by a large factor.

The entropy of English is approx 1.3 bits.

Chapter 7 Channel Capacity.

Channel capacity: The logarithm of the number of distinguishable inputs is given by:

$$C = \max_{p(x)} I(X;Y)$$

During compression, we remove all the redundancy in the data to form the most compressed version possible, whereas during data transmission, we add redundancy in a controlled fashion to combat errors in the channel.

Examples

- Binary symmetric channel C=1-H(p)
- Binary erasure channel C=1-α
- Symmetric channel C=log|Y| - H(row of transition matrix)

Properties of C

- $0 \le C \le \min\{\log|X|, \log|Y|\}$
- I(X;Y) is a continuous concave function of p(x)

Joint Typicality. We decode a channel output Y^n as the i^{th} index if the codeword $X^n(i)$ is "jointly typical" with the received signal Y^n.

The set $A_\epsilon^{(n)}$ of jointly typical sequences $\{(x^n, y^n)\}$ with respect to the distribution p(x,y) is given by

$$\overbrace{\begin{aligned} &\left|-\frac{1}{n}\log p(x^n) - H(X)\right| < \epsilon, \\ &\left|-\frac{1}{n}\log p(y^n) - H(Y)\right| < \epsilon, \\ &\underbrace{\left|-\frac{1}{n}\log p(x^n, y^n) - H(X,Y)\right| < \epsilon}_{\text{where } p(x^n, y^n) = \prod_{i=1}^n p(x_i, y_i)} \end{aligned}}^{A_\epsilon^{(n)} = (x^n, y^n) \in X^n \times Y^n :}$$

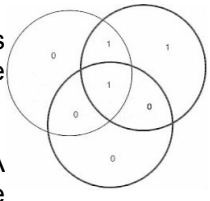
Joint AEP. Let (X_n, Y_n) be sequences of length n drawn i.i.d according to

$p(x^n, y^n) = \prod_{i=1}^n p(x_i, y_i)$ then:

- $Pr((X^n, Y^n) \in A_\epsilon^{(n)}) \to 1 \text{ as } n \to \infty$
- $\left|A_\epsilon^{(n)}\right| \le 2^{n(h(X,Y)+\epsilon)}$
- $If\ (\tilde{X}^n, \tilde{Y}^n) \sim p(x^n)p(y^n),$ $then\ Pr((\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)}) \le 2^{-n(I(X;Y)-3\epsilon)}$

Channel coding theorem. All rates below capacity C are achievable, and all rates above capacity are not; that is for all rates R < C, there exists a sequence of $(2^{nR}, n)$ codes with probability of error $\lambda(n) \to 0$. Conversely, for rates R > C, λ(n) is bounded away from 0.

A **Hamming code** is an example of a parity check code where one or more parity bits are added to the end of the transmitted sequence which depend on various subsets of the information bits. Hamming codes can be visualised as a Venn diagram where each region of overlap represents one information bit and each region belonging to just one set is a parity bit. The location of an error can be inferred from the location of the parity bit.



Feedback capacity. Feedback does not increase capacity for discrete memoryless channels (i.e., $C_{FB} = C$)

Source-channel theorem. A stochastic process with entropy rate H cannot be sent reliably over a discrete memoryless channel if H > C. Conversely, if the process satisfies the AEP, the source can be transmitted reliably if H < C.

"The result - that a two-stage process is as good as any one - stage process-seems so obvious that it may be appropriate to point out that it is not always true. There are examples of multi-user channels where the decomposition breaks down. We also consider two simple situations where the theorem appears to be misleading. A simple example is that of sending English text over an erasure channel. We can look for the most efficient binary representation of the text and send it over the channel. But the errors will be very difficult to decode. If, however, we send the English text directly over the channel, we can lose up to about half the letters and yet be able to make sense out of the message. Similarly, the human ear has some unusual properties that enable it to distinguish speech under very high noise levels if the noise is white. In such cases, it may be appropriate to send the uncompressed speech over the noisy channel rather than the compressed version. Apparently, the redundancy in the source is suited to the channel." pg 219

"The data compression theorem is a consequence of the AEP, which shows that there exists a "small" subset (of size 2^{nH}) of all possible source sequences that contain most of the probability and that we can therefore represent the source with a small probability of error using H bits per symbol. The data transmission theorem is based on the joint AEP; it uses the fact that for long block lengths, the output sequence of the channel is very likely to be jointly typical with the input codeword, while any other codeword is jointly typical with probability $\approx 2^{-nI}$. Hence, we can use about 2^{nI} codewords and still have negligible probability of error. The source-channel separation theorem

shows that we can design the source code and the channel code separately and combine the results to achieve optimal performance." pg 222

Chapter 8 Differential Entropy

Differential entropy is the entropy of a continuous random variable. Differential entropy can be negative but the volume of the support set must be non-negative.

$$h(X)=h(f)=-\int_S f(x)\,logf(x)\,dx$$

if it exists.

AEP for continuous random variables.

$$f(X^n)=2^{-nh(X)}$$

Properties of the typical set $A_\epsilon^{(n)}$ parallel those for the discrete random variable.

$$Vol(A_\epsilon^{(n)})=2^{nh(X)}$$

The entropy of a n-bit quantisation of a continuous random variable X is, on average, the number of bits required to describe X to n-bit accuracy.

$$H([X]_{2^{-n}})\approx h(X)+n$$

The entropy of the Normal distribution with mean μ and variance σ:

$$h(N(0,\sigma^2))=\frac{1}{2}\log 2\pi e\sigma^2 \text{ bits}$$

The entropy of a multivariate Normal distribution with mean μ and covariance matrix K:

$$h(N_n(\mu,K))=\frac{1}{2}\log(2\pi e)^n|K| \text{ bits}$$

The relative entropy (or Kullback-Leibler distance) between two densities is defined by:

$$D(f\|g)=\int f\log\frac{f}{g}\geq 0$$

Chain rule for differential entropy:

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

$h(X|Y)\leq h(X)$ with equality iff X and Y are independent.

$$h(aX)=h(X)+\log|a|$$

Mutual Information:

$$I(X|Y)=\int f(x,y)\log f\frac{(x,y)}{f(x)f(y)}\geq 0$$

The multivariate normal distribution maximises the entropy over all distributions with the same covariance:

$$\max_{EXX^t=K} h(X)=\frac{1}{2}\log(2\pi e)^n|K|$$

Given side information Y and estimator $\hat{X}(Y)$, it follow that:

$$E(X-\hat{X}(Y))^2\geq\frac{1}{2\pi e}e^{2h(X|Y)}$$

$2^{nH(X)}$ is the effective alphabet size for a discrete random variable.

$2^{nh(X)}$ is the effective support set size for a continuous random variable.

2^C is the effective alphabet size of a channel of capacity C.

Chapter 9 Gaussian Channel.

The most important continuous alphabet channel is the Gaussian channel. It is modelled as a time discrete channel with output Y_i at time i, where Y_i is the sum of the input X_i and the noise Z_i. The noise is drawn i.i.d. From a Gaussian distribution. The most common limitation on the input is an energy or power constraint.

The information capacity of the Gaussian channel is: $C=\max_{f(x):EX^2\leq P} I(X;Y)$

Maximum entropy.

$$\max_{EX^2=\alpha} h(X)=\frac{1}{2}\log(2\pi e\alpha)$$

Gaussian Channel.

$$Y_i=X_i+Z_i;$$
$$Z_i\sim N(0,N);$$
$$Power\ constraint\ \ \frac{1}{n}\sum_{i=1}^{n}x_i^2\leq P;\text{ and}$$
$$C=\frac{1}{2}\log(1+\frac{P}{N})\ \ bits\ per\ transmission$$

Constructing (2nC,n) codes with a low probability of error is analogous to sphere packing. The received vector is normally distributed around the mean of the codeword and with noise variance. It is contained within a sphere of $\sqrt{n(N+\epsilon)}$ with high probability.

Bandwidth additive white Gaussian noise channel.

$$Bandwidth\ W;$$
$$two-sided\ power\ spectral\ density\ N_0/2;$$
$$signal\ power\ P;\text{ and}$$
$$C=W\log(1+\frac{P}{N_0W})$$

One of the most famous formulas in information theory.

Water-filling (k parallel Gaussian channels)
Considers the case where the noise on the parallel channels is independent. The objective is to distribute the total power among the channels so as to maximise the capacity.

$$Y_j = X_j + Z_j,\ j = 1,2,\ldots,k;$$
$$Z_j \sim N(0, N_j);$$
$$\sum_{j=1}^{n} X_j^2 \le P;\ \text{and}$$
$$C = \sum_{i=1}^{k} \frac{1}{2}\log\left(1 + \frac{(v - N_i)^+}{N_i}\right)$$
$$\text{where } v \text{ is chosen so that } \sum (v - N_i)^+ = nP$$

As the signal power is increased from zero, we allot the power to the channels with the lowest noise.

Additive non-white Gaussian noise channel.
Considers the case where the noise on the channels is dependent (parallel channels or channels with memory). Let K_Z be the covariance matrix of the noise, and K_X be the input covariance matrix.

$$Y_i = X_i + Z_i;$$
$$Z^n \sim N(0, K_Z);\ \text{and}$$
$$C = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2}\log\left(1 + \frac{(v - \lambda_i)^+}{\lambda_i}\right)$$
$$\text{where } \lambda_1, \lambda_2, \ldots, \lambda_n \text{ are the eigenvalues of } K_Z$$
$$\text{and } v \text{ is chosen so that } \sum (v - N_i)^+ = P$$

In this case, the above water-filling argument translates to water-filling in the spectral domain. For channels in which the noise forms a stationary stochastic process, the input signal should be chosen to be a Gaussian process with a spectrum that is large at frequencies where the noise spectrum is small.

As in the discrete case, feedback does not increase capacity for memoryless Gaussian channels. However, for channels with memory, where the noise is correlated from time instant to time instant, feedback does increase capacity.

Capacity without feedback

$$C_n = \max_{tr(K_X) \le nP} \frac{1}{2n}\log\frac{|K_X + K_Z|}{|K_Z|}$$

Capacity with feedback

$$C_{n,FB} = \max_{tr(K_X) \le nP} \frac{1}{2n}\log\frac{|K_{X+Z}|}{|K_Z|}$$

feedback bounds

$$C_{n,FB} \le C_n + \frac{1}{2} \quad \text{bits per transmission.}$$
$$C_{n,FB} \le 2C_n$$

Chapter 10 Rate Distortion Theory.

A continuous random source requires infinite precision to represent it exactly, so a we cannot represent it exactly with a finite rate code. The question is then to find the best possible representation for any given data rate. Given a source distribution and a distortion measure, what is the minimum expected distortion achievable at a particular rate. Interestingly, joint descriptions are more efficient than individual descriptions, even for independent random variables.

Quantization. The Lloyd algorythm constructs a good quantization by starting with a set of reconstruction points, finding the optimal set of construction regions (nearest neighbours with respect to the distortion measure), then find the optimal reconstruction points for these regions and interate.

Rate distortion. The information rate distortion function for a source X ~ p(x) and distortion measure $d(x, \hat{x})$ is:

$$R(D) = \min_{p(\hat{x}|x):\ \sum_{(x,\hat{x})} p(x)\,p(\hat{x}|x)\,d(x,\hat{x}) \le D} I(X; \hat{X})$$

where the minimisation is over all conditional distributions $p(x, \hat{x})$ for which the joint distribution $p(x, \hat{x}) = p(x)\,p(\hat{x}|x)$ satisfies the expected distortion constraint.

Rate distortion theorem. If R>R(D), there exists a sequence of codes $\hat{X}^n(X^n)$ with the number of codewords $|\hat{X}^n(\cdot)| \le 2^{nR}$ with $Ed(\hat{X}, \hat{X}^n(X^n)) \to D$. If R<R(D), no such codes exist.

Bernoulli source. For a Bernoulli source with Hamming distortion,
R(D) = H(p) – H(D)

Gaussian Source. For a Gaussian source with squared-error distortion.

$$R(D) = \frac{1}{2}\log\frac{\sigma^2}{D}$$

Each bit of description reduces the expected distortion by a factor of 4. With a 1-bit description, the best expected square error is $\sigma^2/4$. A 1-bit quantization of N(0,σ2) random variable usine two regions corresponding to the +ve and -ve real lines and reproduction points as the centroids, the expected distortion is $\frac{(\pi - 2)}{\pi}\sigma^2 = 0.3633\sigma^2$

Hence, we can achieve a lower distortion by considering several distortion problems in

succession (long block lengths) than can be achieved by considering them separately

Source-channel separation. A source with rate distortion R(D) can be sent over a channel of capacity C and recovered with distortion D if and only if R(D)<C.

Multivariate Gaussian Source. The rate distortion function for a multivariate normal vector with Euclidean mean-squared-error distortion is given by reverse water-filling on the eigenvalues. We choose a constant λ and only describe those random variables with variances greater than λ.

We can transform a good code for channel transmission into a good code for rate distortion. The essential idea is to fill the space of source sequences: In channel transmission, we want to find the largest set of code words that have a large minimum distance between codewords, whereas in rate distortion, we wish to find the smallest set of codewords that covers the entire space.

Chapter 11. Information Theory and Statistics

The method of types is a powerful method in deviation theory which considers sequences which have the same empirical distribution.

Definition: The *type* P_x (or empirical probability distribution) of a sequence $x_1, x_2,..., x_n$ is the relative proportion of occurrences of each symbol X (i.e., $P_x(a) = N(a|\mathbf{x})/n$ for all aєX, where N(a|x) is the number of times the symbol a occurs in the sequence $\mathbf{x}$єX_n.

Basic identities

$$|P_n| = (n+1)^{|X|}$$
$$Q^n(x) = 2^{-n(D(P_x\|Q)+H(P_x))}$$
$$|T(P)| = 2^{nH(P)}$$
$$Q^n(T(P)) = 2^{-nD(P\|Q)}$$

These equations state that there is only a polynomial number of types and that there are an exponential number of sequences of each type. There is an exact formula for the probability of any sequence of type P under distribution Q and an approximate formula for the probability of a type class.

The crucial point is that it follows that at least one type has exponentially many sequences in its type class. In fact, the largest type class has essentially the same number of elements as the entire set of sequences, to the first order in the exponent.

Since the probability of each type class depends exponentially on the relative entropy distance between the type P and the distribution Q, type classes that are far away from the true distribution have exponentially smaller distribution.

Universal Data Compression

Huffman coding compresses an i.i.d source with a known distribution p(x) to its entropy limit H(X). However if the code is designed for some incorrect distribution q(x), a penalty of D(p||q) is incurred. Surprisingly, there is a universal code of rate R, say, that suffices to describe every i.i.d source with entropy H(X)<R.

$$P_e^{(n)} \le 2^{-nD(P_R^*\|Q)} \text{ for all } Q$$
$$\text{where } D(P_R^*) = \min_{P:H(P)\ge R} D(P\|Q)$$

However, universal source codes need a longer block length to obtain the same performance as a code defined specifically for the probability distribution.. We pay a penalty for this increase in block length by the increased complexity of the

encoder and decoder.

Large deviations (Sanov's theorem)
Estimates the probability of a set of non-typical types. The probability that a sample will show a large deviation from the expected outcome is exponentially small. (We can estimate the exponent using the central limit theorem but this is a poor approximation for more than a few standard deviations).

$$Q^n(E) = Q^n(E \cap P_n) \leq (n+1)^{|X|} 2^{-nD(P^* \| Q)}$$
$$D(P^* \| Q) = \min_{P \in E} D(P \| Q)$$

If E is the closure of its interior, then
$$Q^n(E) = 2^{-nD(P^* \| Q)}$$

L_1 bound on relative entropy.
Convergence in relative entropy implies convergence in the L1 norm.

$$D(P_1 \| P_2) \geq \frac{1}{2\ln 2} \| P_1 - P_2 \|_1^2$$

Pythagorean theorem.
Many of the intuitive properties of distance are not valid for D(P||Q) which behaves like the square of the Euclidean distance.

If E is a convex set of types, distribution
$Q \notin E$, and P* achieves,
$D(P^* \| Q) = min_{P \in E} D(P \| Q)$ we have
$D(P \| Q) \geq D(P \| P^*) + D(P^* \| Q)$ for all $P \in E$

Conditional Limit Theorem.
The conditional limit theorem implies that there is a very high probability that the type observed is close to P* where P* is probability of the closest type. This is established by considering the empirical distribution of the sequence of outcomes given that the type is in a particular set of distributions E. The probability of E is essentially determined by D(P||Q), the distance of the closest element of E to Q, and the conditional type is essentially P*, and the probability of other types, that are far away from P* is negligible.

If $X_1, X_2, ..., X_n$ i.i.d. ~ Q, then
$Pr(X_1 = a | P_{X^n} \in E) \rightarrow P^*(a)$ in probability,
where P* minimizes D(P‖Q) over $P \in E$.
In particular,
$$Pr\left\{ X_1 = a | \frac{1}{n} \sum_{i=1}^{n} X_i \geq \alpha \right\} \rightarrow \frac{Q(a) e^{\lambda a}}{\sum_x Q(x) e^{\lambda x}}$$

Neyman-Pearson lemma. The optimal test between two densities P_1 and P_2 has a decision region of the form
$$accept\ P = P_1 \ \ if \ \ \frac{P_1(x_1, x_2, ..., x_n)}{P_2(x_1, x_2, ..., x_n)} > T$$

Chernoff-Stein lemma.
The Chernoff-Stein lemma considers hypothesis testing in the case where one of the probabilities of error is held fixed and the other is made as small as possible. The other prorbability of error is exponentially small, with an exponential rate equal to the relative entropy between the two distributions. The best achievable error exponent
$$\beta_n^\epsilon \ \text{if} \ \ \alpha_n \leq \epsilon \ : \ \ \beta_n^\epsilon = \min_{\substack{A_n \subseteq X^n \\ \alpha_n \leq \epsilon}} \beta_n$$
$$\lim_{(n \to \infty)} \frac{1}{n} \log \beta_n^\epsilon = -D(P_1 \| P_2)$$

Chernoff information.
The Bayesian approach assigns prior probabilities to both hypotheses and minimises the overall probability of error given by the weighted sum of the individual probabilities of error. The best achievable exponent for a Bayesian probability of error is:

$$D^* = D(P_{\lambda^*} \| P_1) = D(P_{\lambda^*} \| P_2) \ \ where$$
$$P_\lambda = \frac{P_1^\lambda(x) P_2^{1-\lambda}(x)}{\sum_{a \in X} P_1^\lambda(x) P_2^{1-\lambda}(x)} \ \ with$$
$$\lambda = \lambda^* \ chosen \ so \ that$$
$$D(P_\lambda \| P_1) = D(P_\lambda \| P_2)$$

Fisher information.
A standard problem in statistical estimation is to determine a parameter θ of a distribution from a sample of data drawn from that distribution. An estimator T is meant to approximate the value of the parameter. The bias is the expected value of the error in the estimator. However, a bias = 0 does not guarantee that the error is low with high probability. A loss function of the error is required ; most commonly the expected square of the error.

$$J(\theta) = E_\theta \left[\frac{\partial}{\partial \theta} \ln f(x; \theta) \right]^2$$

The Fisher information is a measure of the amount of “information” about θ that is present in the data. It is a lower bound on estimating θ from the data.

Cramer-Rao inequality.
The Cramer-Rao inequality is a lower bound on the variance of all unbiased estimators. For and unbiased estimator T of θ,

$$E_\theta(T(X)-\theta)^2 = var(T) \geq \frac{1}{J(\theta)}$$

Fisher information is defined with respect to a family of parametric distributions, unlike entropy, which is defined for all distributions. Entropy is related to the volume of the typical set and Fisher information to the surface area of the typical set.

Chapter 12 Maximum Entropy

The temperature of a gas corresponds to the average kinetic energy of the molecules in the gas. What can we say about the distribution of velocities in the gas at a given temperature? We know from physics that this distribution is the maximum entropy distribution under the temperature constraint, otherwise known as the Maxwell-Boltzmann distribution. The maximum entropy distribution corresponds to the macro state (as indexed by the empirical distribution) that has the most micro states (the individual gas velocities). Implicit in the use of maximum entropy methods in physics is a sort of AEP which says that all microstates are equally probable.

Maximum Entropy Distribution. Let f be a probability density satisfying the constraints

$$\int_S f(x) r_i(x) = \alpha_i \quad for\ 1 \leq i \leq m$$

Let $f^*(x) = f_\lambda(x) = e^{\lambda_0 \sum_{i=1}^m \lambda_i r_i(x)}, x \in S$ and let $\lambda_0,...,\lambda_m$ be chosen so that f* satisfies the constraints. Then f* uniquely maximises h(f) overall f satisfying these constraints. F is a density on support set S meeting certain moment constraints $\alpha_1, \alpha_2,..., \alpha_m$

Example: dice, no constraints. Let S = {1,2,3,4,5,6}. The distribution that maximises the entropy is the uniform distribution, p(x)=1/6 for $s \in S$

Example used by Boltzman - dice with $EX = \sum ip_i = \alpha$ That is, n dice are thown which sum to nα then what proportion of the dice are showing each face?

Example: S=[0,∞] and EX=μ. The distribution of the height of molecules in the atmosphere.

Maximum entropy spectral density estimation. Burg assumed a process was stationary and Gaussian and found that the process that maximises the entropy subject to the correlation constraints is an autoregressive Gaussian process of the appropriate order.

The entropy rate of a stochastic process subject to autocorrelation constraints $R_0, R_1,..., R_p$ is maximised by the p^{th} order zero-mean Gauss-Markov process satisfying these constraints. The maximum entropy rate is

$$h^* = \frac{1}{2}\log(2\pi e)\frac{|K_p|}{|K_{p-1}|}$$

and the maximum entropy spectral density is

$$S(\lambda)=\frac{\sigma^2}{\left|1+\sum_{k=1}^{p} a_k e^{-ik\lambda}\right|^2}$$

Chapter 13 Universal Source Coding

If the probability distribution underlying the source is unknown, then we cannot apply the methods of Chapter 5 directly (e.g. Huffman coding) unless two passes of the data are made. However, there are online algorithms that use the probability distribution of the inbound data to guide the compression, and these do well for any distribution within a class of distributions. In an individual sequence (i.e. text and music) there is no underlying probability distribution. We compare our performance to that achievable by optimal word assignments with respect to Bernoulli distributions or k^{th}-order Markov processes. The ultimate compression of an individual sequence is the Kolmogorov complexity.

Ideal word length. If the distribution is known.

$$l^*(x)=\log\frac{1}{p}(x)$$

Average description length. As a basis for comparison.

$$E_p l^*(x)=H(p)$$

Estimated probability distribution $\hat{p}(x)$ increases the word length by an amount equal to the relative entropy between the estimate an the actual.

$$If \quad \hat{l}(x)=\log\frac{1}{\hat{p}(x)}, \quad then$$
$$E_p \hat{l}(x)=H(p)+D(p\|\hat{p})$$

Average redundancy of using universal coding.

$$R_p=E_p l(X)-H(p)$$

Minimax redundancy.

$$For \quad X\sim p_\theta(x), \theta\in\theta$$
$$D^*=\min_{l}\max_{p} R_p=\min_{q}\max_{\theta} D(p_\theta\|q)$$

This minimax redundancy is achieved by a distribution q that is at the 'center' of the information ball containing the distributions p_θ, that is, the distribution q whose maximum distance from any of the distributions p_θ is minimised.

Minimax theorem. D*=C, where C is the capacity of the channel $\{\theta, p_\theta(x), X\}$.

This is a channel with the rows of the transition matrix equal to the different p_θ's, the possible distributions of the source. The minimax redundancy is equal to the capacity of this channel and the corresponding optimal coding distribution is the output distribution of this channel induced by

the capacity-achieving input distribution.

Bernoulli sequences. For X^n ~ Bernoulli(θ), the redundancy is

$$D_n^* = \min_q \max_\theta D(p_\theta(x^n) \| q(x^n)) \approx \frac{1}{2}\log n + o(\log n$$

That is, the cost of describing the sequence is about ½log(n) bits above the optimal cost with the Shannon code for a Bernoulli distribution corresponding to k/n.

Arithmetic coding. The objective of the arithmetic coding algorithm is to represent a sequence of random variables by a subinterval in [0,1]. As the algorithm observes more input symbols the length of the subinterval corresponding to the input sequence decreases. As the top and bottom ends of the interval get closer they begin to agree in the first few bits and they can be output. The process continues on the remaining subinterval until the whole sequence is output. The procedure achieves an average block length within 2 bits of the entropy for any block-length.

nH bits of $F(x^n)$ reveal approximately n bits of x^n.

Lempel-Ziv coding. The key idea of the Lempel-Ziv algorithm is to parse the string into phrases and to replace phrases by pointers to where the same string has occurred in the past. The differences between the algorithms is based on differences in the set of possible match locations (and match lengths) the algorithm allows.

Lempel-Ziv coding (recurrence time coding). Let $R_n(X^n))$ be the last time in the past that we have seen a block of n symbols X^n, Then $\frac{1}{n}\log R_n \to H(X)$, and encoding by describing the recurrence time is asymptotically optimal. Used in zip and gzip implementations.

Lempel-Ziv coding (sequence parsing). If a sequence is parsed into the shortest phrases not seen before (e.g. 011011101 is parsed to 0,1,10,11,101,...) and $l(x^n)$ is the description length of the parsed sequence, then,

$$\limsup \frac{1}{n} l(X^n) \le H(X) \quad with\ probability\ 1$$

for every stationary ergodic process (X_i). Used in compress in Uni, modems and the GIFF format.

Chapter 14 Kolmogorov Complexity

Definition: The *Kolmogorov complexity* K(x) of a string x is:

$$K(x) = \min_{p:U(p)=x} l(p)$$

$$K(x|l(x)) = \min_{p:U(p,l(x))=x} l(p)$$

Kolmogorov complexity is the minimum length over all programs that print x and halt.

Universality of Kolmogorov complexity. There exists a universal computer U such that for any other computer A,

$$K_u(x) \le K_a(x) + c_A$$

for any string x, where the constant c_A does not depend on x. If U and A are universal $|K_U(x) - K_A(x)| < c$ for all x.

Upper bound on Kolmogorov complexity.

$$K(x|l(x)) \le l(x) + c$$

$$K(x) \le K(x|l(x)) + 2\log l(x) + c$$

Komologrov complexity and entropy. If X_1, X_2,..., X_n are i.d.d. Integer-valued random variables with entropy H, there exists a constant c such that for all n,

$$H \le \frac{1}{n} EK(X^n|n) \le H + |X| \frac{\log n}{n} + \frac{c}{n}$$

Lower bound on Kolmogorov complexity. There are no more than 2^k strings x with complexity K(x) < k. If X_1, X_2,..., X_n are drawn according to Bernoulli(½) process.

$$Pr(K(X_1, X_2 \ldots X_n|n) \le n - k) 2^{-k}$$

Definition. A sequence x is said to be incompressible if $\frac{K(x_1, x_2 \ldots x_n|n)}{n} \to 1$

Strong law of large numbers for incompressible sequences

$$\frac{K(x_1, x_2, \ldots, x_n)}{n} \to 1 \Rightarrow \frac{1}{n}\sum_{i=1}^{n} x_i \to \frac{1}{2}$$

Definition. The universal probability of a string x is

$$P_U(x) = \sum_{p:U(p)=x} 2^{-l(p)} = Pr(U(p)=x)$$

This is the probability that a randomly drawn program will print out the string x.

Most sequences of length n have complexity close to n. Shorter programs are much more probable than longer ones. That is, there are not enough programs to go around.

The halting problem and the non-computability of Kolmogorov complexity. Both these are related to Godells incompleteness theorem and all three are bases on self-referential ideas like “This sentence is false.”

Universality of $P_U(x)$. For every computer A, $P_U(x) \geq c_A P_A(x)$ for every string $x \in \{0,1\}^*$ where the constant c_A depends only on U and A.

Definition.

$$\Omega = \sum_{p:U(p)halts} 2^{-l(p)} = Pr(U(p)halts)$$

is the probability that the computer halts and the input p to the computer is a binary string drawn according to a Bernoulli(½) process.

Properties of Ω

- Ω is not computable. The halting problem.
- Ω is a “philosopher's stone”. Knowing Ω to an accuracy of n bits will enable us to decide the truth of any provable of finitely refutable mathematical theorem that can be written in less than n bits.
- Ω is algorithmically random (incompressible).

Universal Gambling. Is based on $P_U(x)$ and does asymptotically as well as a scheme that uses of the true distribution.

Equivalence of K(x) and $\log(\frac{1}{P_U(x)})$. There exists a constant c independent of x such that,

$$|\log \frac{1}{P_U(x)} - K(x)| \leq c$$

for all strings x. Thus the universal probability of a string x is essentially determined by its Kolmogorov complexity.

Notice that the ideal Shannon code length assignment $l(x) = \log(\frac{1}{p(x)})$ achieves an *average* description length H(X), while in Kolmogorov complexity theory, the ideal description length $\log(\frac{1}{P_U(x)})$ is *almost* equal to K(X). Thus $\log(\frac{1}{p(x)})$ is the natural notion of descriptive complexity of x in algorithmic as well as probabilistic settings.

Definition. The *Kolmogorov structure function* $K_k(x_n|n)$ of a binary string $x^n \in \{0,1\}^n$ is defined as,

$$K_k(x^n|n) = \min_{\substack{p:l(p)\leq k \\ U(p,n)=S \\ x\in S}} \log|S|$$

Definition. Let k* be the least k such that,

$$K_{k^*}(x^n|n) + k^* = K(x^n|n)$$

Let S** be the corresponding set and let p** be the program that prints out the indicator function of S**. Then p** is the *Kolmogorov minimal sufficient statistic* for x.

Chapter 15 Network Information Theory

I have not read this chapter – it did not appear relevant to my interests (and it was also long)

Chapter 16 Information Theory and Portfolio Theory

I have not read this chapter.

Chaper 17 Inequalities in Information Theory

This chapter is a brutal summary of the key aspects of the previous chapters with a similarly brutal introduction of some additional results.

There may be important aspects here if attempting to construct new proofs – particularly in relation to Fisher Information.

Further Reading

- pg 508 “A non-technical introduction to the various measures of complexity can be found in a thought provoking book by Pagels, H. The Dreams of Reason: the Computer and the Rise of the Sciences of Complexity. Simon and Schuster, New York 1988.
- pg 171 non-technical introduction to the estimation of various information sources including English is Lucky , R. W. (1989) Silicone Dreams: Information, Man and Machine. St Martins Press, New York.