2.1 Entropy

Supplement: it is easy to verify that \( \frac{dH(p)}{dp} = \log \frac{1-p}{p} \). Then, the slope is infinite both at \( p = 0 \) and \( p = 1 \).

Example: three equiprobable symbols. \( H(X) = \log 3 = 1.585 \). Does this mean that 1.585 is the minimum number of bits per message, when we encode this source? Yes, as we shall see soon.

Additional example: \( p(x) = \frac{1}{4} \quad x = 1, \ldots, 4 \). \( H(X) = 2 \). This is sound: it is like concatenating two binary symbols.

What if we concatenate two binary, non equiprobable, symbols? We obtain \( H(X) < 2 \). Is this true for any quaternary probability distribution? In general, what happens for an alphabet of \( M \) symbols?

Theorem 2.6.4 (with a different proof)

\[
\log_e z \leq z - 1 \text{ with equality iff } z = 1; \text{ then } \log z \leq (z - 1) \log e
\]

\[
H(X) - \log M = \sum p(x) \log \frac{1}{M_\text{p}(x)} \leq \log e \sum p(x) \left( \frac{1}{M_\text{p}(x)} - 1 \right) = 0 \text{ with equality iff } P(x) = 1/M \ \forall x.
\]

5.1 Examples of codes (skip Example 5.1.3)

5.2 Kraft inequality (skip Theorem 5.2.2)

Let \( D = 2 \) everywhere (the generalization is straightforward)

5.5 Kraft inequality for uniquely decodable codes

Skip the corollary

Theorem 5.3.1 (with a different proof)

\[
H(X) - L = \sum p_i \log \frac{1}{p_i} - \sum p_i l_i = \sum p_i \log \frac{2^{-l_i}}{p_i} \leq \log e \sum p_i \left( \frac{2^{-l_i}}{p_i} - 1 \right) = 0, \text{ with equality iff } p_i = 2^{-l_i}, \text{ i.e., } l_i = \log \frac{1}{p_i} \ \forall i.
\]

Skip the definition of \( D \)-adic probability distribution

5.4 Bounds on the optimal code length

Comment: \( l_i = \left\lceil \log \frac{1}{p_i} \right\rceil \) is not optimal, but suffices to prove the theorem. In many cases there are better codes. E.g., if \( p = 0.45, 0.45, 0.1 \) we obtain \( l = 2, 2, 4 \) instead of 1,2,2

Skip Theorem 5.4.1

Skip from page 89 on (for the time being)
Additional remark: actually, there is another way to encode the source. For simplicity, let us consider a binary source with entropy \( H(p) \). We can encode variable-length blocks like 0,10,110,1110,1111 (this is called run-length encoding). It takes some time to show that the entropy of this ensemble is equal to \( H(p) \) times the average number of source bits encoded in one pass. Then the minimum average length of the encoded sequence, divided by the average number of source symbols is unchanged. We can not escape the fundamental limit: \( L \geq H(X) \), where \( L \) is the average number of encoded bits per source message. Possible practical advantages of disadvantages of run-length encoding are another story, of course.

To prove this result we need the following property of entropy:

\[
H(p_1, p_2, p_3, \ldots, p_M) = H(p_1 + p_2, p_3, \ldots, p_M) + (p_1 + p_2)H\left(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2}\right)
\]

This property can be generalized to arbitrary groupings, e.g.,

\[
H(p_1, p_2, p_3, \ldots, p_M) = H(p_1 + \ldots + p_m, p_{m+1} + \ldots + p_M) + (p_1 + \ldots + p_m)H\left(\frac{p_1}{p_1 + \ldots + p_m}, \ldots, \frac{p_m}{p_1 + \ldots + p_m}\right) + (p_{m+1} + \ldots + p_M)H\left(\frac{p_{m+1}}{p_{m+1} + \ldots + p_M}, \ldots, \frac{p_M}{p_{m+1} + \ldots + p_M}\right)
\]

2.2 Joint entropy and conditional entropy

Addendum: we can also define the conditional entropy \( H(Y|x) \), conditioned on a specific value of \( X \): \( H(Y|x) = \sum p(y|x) \log p(y|x) \) (the sum is over \( y \); \( x \) is fixed). Then we have \( H(Y|X) = \sum p(x)H(Y|x) \). In other words: first we average over \( y \), then over \( x \). This can be useful since in many cases \( H(Y|x) \) does not depend on \( x \). In these cases \( H(Y|X) = H(Y|x) \) can be evaluated just once, for a single (arbitrary) \( x \).

Theorem 2.6.5 (with a different proof)

\[
H(X|Y) - H(X) = \sum \sum p(x,y) \log \frac{p(x)}{p(x|y)} \leq \log \sum \sum p(x,y) \left( \frac{p(x)}{p(x|y)} - 1 \right) = 0
\]

2.5 Chain rules ... (skip from eq. (2.59) on, for the time being)

4.2 Entropy rate (many people call it simply Entropy)

Skip Examples 1,2,3

Skip from eq. (4.23) on

Note that \( H(X_n|X_{n-1}, \ldots, X_1) \) converges to the source entropy faster than \( \frac{1}{n} H(X_1, \ldots, X_n) = \frac{1}{n} H(X_1) + H(X_2|X_1) + \ldots + H(X_n|X_{n-1}, \ldots, X_1) \)

4.1 Markov chains

Eq. (4.25)

Note that for a first-order Markov source \( \frac{1}{n} H(X_1, \ldots, X_n) = \frac{1}{n} H(X_1) + \frac{n-1}{n} H(X_2|X_1) \) converges to the source entropy only for \( n \to \infty \), while \( H(X_2|X_1) \) gives the source entropy
Comment: we could define the source entropy as $\lim_{n \to \infty} \frac{1}{n} H(X_{m+n}, \ldots, X_m | X_{m-1}, \ldots, X_1)$. This suggests that we encode blocks of $n$ symbols with $M^n$ different codes, chosen according to the previous $m$-tuple (for instance, $m = 1$ for a first-order Markov source).

Now back to page 88 (last three lines) and page 89 (up to eq. (5.41)), for sources with memory

5.6 Huffman codes (just the first page)

5.8 Optimality of Huffman codes (read at your own risk)

And now some exercises (First exercises - Source coding)

Other exercises, from Cover, Thomas, 1st ed.: 2.1, 2.5, 4.2, 4.4, 5.8, 5.12

Other exercises, from Cover, Thomas, 2nd ed.: 2.29a, 2.29c, 5.15, 5.44

5.10 Arithmetic coding (you can give a look also to 5.9 Shannon-Fano Elias Coding)

Arithmetic coding is discussed also in 13.3, 2nd ed.

Universal coding (Sect. 12.3 is hardly useful; 13.2, 2nd ed., is better)

12.10 Lempel-Ziv coding (skip the "simple" proof; it is not that simple)

13.4, 2nd ed., gives more details on Lempel-Ziv coding

13.5, 2nd ed., gives the proofs of optimality (read at your own risk)

2.3 Mutual information (skip relative entropy) (forget about Fig. 2.2)

Comment: $I(X; Y)$ is often indicated also by $I(X, Y)$; in these cases, $I(X, Y; Z)$ is written as $I(XY, Z)$

2.5 Chain rules for mutual information

Example: when in doubt, expand e.g. as follows:

$I(X; Y, Z) = E \log \frac{p(X, Y, Z)}{p(X)p(Y|Z)} = E \log \frac{p(X,Y)p(Z|X,Y)}{p(X)p(Y)p(Z|Y)} = I(X; Y) + I(X; Z|Y)$

(2.91) Proof: $I(X; Y) = H(X) - H(X|Y) \geq 0$ with equality iff $X$ and $Y$ are independent

2.8 Data processing inequality (note that eq. (2.121) should read $= I(X; Y) + I(X; Z|Y)$)

Comment: you can show that $I(X; Z|Y) = 0$ this way: $I(X; Z|Y) = E \log \frac{p(Z|X,Y)}{p(Z|Y)} = 0$

Exercises 2.29b, 2.29d

Ch. 8, Introduction

8.1.4 Binary Symmetric Channel

8.1.5 Binary Erasure Channel
8.2 Symmetric channels

8.3 Properties of channel capacity

Addendum: evaluation of channel capacity

For any input symbol \( x \), the partial derivative with respect to \( p(x) \) of the mutual information
\[ I(X, Y) = \sum_{x'} p(x') \sum_y p(y|x') \log \frac{p(y|x')}{p(y)}, \]
where \( p(y) = \sum_{x''} p(x'')p(y|x'') \), is
\[ \sum_y p(y|x) \log \frac{p(y|x)}{p(y)} - \log e \sum_{x'} p(x') \sum_y p(y|x') \frac{p(y|x)}{p(y)} = \sum_y p(y|x) \log \frac{p(y|x)}{p(y)} - \log e \]

We have also the equality constraint \( \sum p(x) = 1 \), and the inequality constraints \( p(x) \geq 0 \).
Let \( \lambda \) be a Lagrange multiplier. Deriving \( I(X, Y) + \lambda \sum p(x) \) we obtain
\[ \sum_y p(y|x) \log \frac{p(y|x)}{p(y)} - \log e + \lambda = 0 \]
i.e.,
\[ \sum_y p(y|x) \log \frac{p(y|x)}{p(y)} = C \]
for all \( x \) such that \( p(x) > 0 \). If \( p(x) = 0 \) for some \( x \), the maximum is on the boundary, and we have
\[ \sum_y p(y|x) \log \frac{p(y|x)}{p(y)} \leq C \]

In any case, multiplying by \( p(x) \) and summing over all \( x \) we see that the constant \( C \) is the channel capacity.

As an example, evaluate the capacity of the binary \( Z \) channel, with transition probabilities \( p(0|0) = 1 \), \( p(0|1) = \varepsilon \) and \( p(1|1) = 1 - \varepsilon \), and verify that the channel capacity is
\[ C = \log(1 + 2^{-H(\varepsilon)/(1-\varepsilon)}) = \log(1 + (1 - \varepsilon)\varepsilon/(1-\varepsilon)) \]

In many cases the channel capacity can be obtained only numerically. The expressions of the derivatives with respect to \( p(x) \) guide the iterative search towards the optimal input distribution \( p(x) \).

9.1 Differential entropy

9.3 Differential entropy vs discrete entropy

9.4 Joint and conditional differential entropy

9.5 Mutual information

\(^1\) do not confuse the dummy summation indices \( x' \) and \( x'' \) for \( x \)
Give a look to 9.6 Properties of differential entropy

In Th. 9.6.5 do not mind about $D(g∥\Phi_K)$. Use $\int g \log(\Phi_K/g) \leq \log e \int g \left(\frac{\Phi_K}{g} - 1\right) = 0$, with equality iff $g = \Phi_K$.

Ch. 10, Introduction

10.1 The Gaussian channel (up to page 242)

10.3 Band-limited channels (give a look; probably you are already familiar with digital transmission; the only novelty should be eq. (10.60))

Some exercises (Exercises - Channel capacity)

Other exercises, from Cover, Thomas, 1st ed.: 8.9, 8.10

Other exercises, from Cover, Thomas, 2nd ed.: 7.12, 7.16, 7.18, 7.20, 7.23, 7.27, 7.34a, 7.34b, 7.35, 7.36

8.7 Channel coding theorem (with a different proof; see channel\_coding\_theorem.pdf on this web page; you can give a look also to 2.6 Jensen’s inequality)

2.11 Fano’s inequality (note that if $X$ is binary we have $H(P_e) \geq H(X|Y)$)

Converse to the channel coding theorem (see channel\_coding\_theorem.pdf on this web page)

Pages 213-214 Capacity with feedback

Do not mind Fano’s inequality. The key point is that $I(W;Y^n) \leq nC$ even with feedback (eq.s (8.131) to (8.138))

10.4 Parallel Gaussian channels

Ch. 3 Asymptotic equipartition property (skip 3.3)

8.4 Preview of the channel coding theorem

8.5 Definitions

8.6 Jointly typical sequences (do not be fussy with all the details; we only want to grasp the flavor of proofs like these)

8.7 The channel coding theorem (do not be fussy with all the details)

Ch. 13 Introduction

13.1 Quantization

13.2 Definitions

13.3 Calculation of the rate distortion function (have only a look at 13.3.3)
13.4 Converse to the rate distortion theorem (We need Theorem 2.7.4, whose proof is a little tricky. A simpler proof is the following).

Let \( p(y|x) = \lambda p_1(y|x) + (1 - \lambda)p_2(y|x) \). Let \( Z \) be a random variable, independent of \( X \), and let \( P(Z = 1) = \lambda, P(Z = 2) = 1 - \lambda \). We have \( I(X; Y, Z) = I(X; Y) + I(X; Z|Y) \). Therefore \( I(X; Y) \geq I(X; Y|Z) \), i.e., \( \lambda I_1(X; Y) + (1 - \lambda)I_2(X; Y) \geq I(X; Y) \).

13.5 Achievability of the rate distortion function (on this matter, the examples that follow will suffice)

Example (Hamming distortion). We want to encode (with loss) a binary source without memory, with \( P(0) = P(1) = 1/2 \), using \( R < 1 \) bits per symbol. We must accept some (Hamming) distortion \( p \). Pick a capacity achieving code for the BSC with crossover probability \( p \). The code rate \( R = k/n = C = 1 - H(p) \). Note that \( R \) is also equal to the rate distortion function. To encode an \( n \)-tuple, send it to the decoder. The \( k \) output bits represent the \( n \) source bits. We are spending \( R \) bits per source symbol. To retrieve the source, send the \( k \) bits to the encoder. The output vector differs from the source vector in about \( np \) positions, i.e., the average (Hamming) distortion is \( p \).

The same reasoning holds if we want to transmit \( R \) bits of information per channel use on a channel whose capacity \( C \) is less than \( R \). First, compress the source by the factor \( C/R \), then transmit (without errors) through the channel, and finally expand to the original rate. The probability of error \( p \) is determined by \( 1 - H(p) = C/R \), i.e., \( H(p) = (R - C)/R \). Note that this agrees with the bound stated by the converse to the channel coding theorem.

Example (Quantization of a Gaussian random vector). Send the \( n \)-tuple \( Y^n \) to the decoder for a capacity achieving code for the Gaussian channel, with rate \( R \). The \( n \)-tuple is represented with \( nR \) bits. To retrieve the source \( n \)-tuple (with some distortion), encode the \( nR \) bits. Let \( X^n \) be the output. Setting \( D = \sigma_Y^2 \) and \( \sigma_X^2 = \sigma_Y^2 + \sigma_N^2 \), we get \( R = \frac{1}{2} \log \left( 1 + \frac{\sigma_X^2}{\sigma_N^2} \right) = \frac{1}{2} \log \frac{\sigma_Y^2}{D} \), i.e., \( D = \sigma_Y^2 2^{-2R} \).

Exercises from Cover, Thomas, 1st ed.: 13.2, 13.7, 13.9

Exercises from Blahut: 6.1, 6.2

Exercise from Gallager: 9.2

Exercise from Cover, Thomas, 2nd ed.: 10.5, 10.15

Ch. 14 Network information theory (Introduction)

14.1.3 The Gaussian broadcast channel

Theorem 14.3.1 Multiple access channel capacity (no proof; only eq.s 14.56, 14.57, 14.58)

Example 14.3.3 Binary erasure multiple access channel (note that channel \( X_2 \) must be decoded, and then subtracted, to get \( X_1 \))

14.3.2 Comments on the capacity region
14.3.6 Gaussian multiple access channel (start from the Definition on page 405)

Duality of multiple access and broadcast channels:

http://www.ece.umn.edu/~nihar/papers/mac_bc_allerton01_final.pdf (no proofs)

or

http://www.ece.umn.edu/~nihar/papers/mac_bc_ittrans.pdf (with all proofs)

Finally, some exercises:

Exercises from Cover, Thomas, 1st ed.: 14.1, 14.2, 14.3, 14.6

Exercises from Cover, Thomas, 2nd ed.: 15.20, 15.23, 15.25, 15.27, 15.28, 15.30, 15.31, 15.33

Then, our course will be over.

THANK YOU FOR HAVING ATTENDED