Information Theory

Vishal Raman

We present expository notes on *Elements of Information Theory* by Cover and Thomas. Solutions to some exercises are presented, but many topics are currently left out. Any typos or mistakes are my own - please redirect them to my email.

Contents

1 Entropy, Relative Entropy, Mutual Information

Let X be a discrete random variable with alphabet $\mathcal{X}$ and probability mass function $p(x) = \Pr\{X = x\}$, $x \in \mathcal{X}$.

1.1 Entropy

Definition 1.1 (Entropy)**.** The entropy $H(X)$ of a discrete random variable X is defined by

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

We will sometimes denote this as $H(p)$, and note that log denotes $\log_2$. The unit of entropy is bits, and we use the convention that $0 \lg 0 = 0$. We denote $H_a(X)$ to be the same corresponding entropy with base a for the logarithm.

We can also express this in terms of expected value as

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

> **Remark 1.2.** It is possible to axiomatically obtain the definition of entropy, which we will do in a future exercise. Instead, we show that it answers various questions that are important to us.

First, we state some obvious proprties.

- $H(X) \geq 0$.
- $H_b(X) = (\log_b a) H_a(X)$.

There is a connection between entropy and the expected number of binary questions to answer a given question. In particular, we will show that this number lies between $H(X)$ and $H(X) + 1$.

1.2 Joint Entropy and Conditional Entropy

Definition 1.3 (Joint Entropy)**.** Given a pair of discrete random variables (X, Y) with a joint distribution $p(x, y)$, the joint entropy $H(X, Y)$ is defined as

$$H(X, Y) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x, y) = -E \log p(X, Y).$$

Definition 1.4 (Conditional Entropy)**.** If $(X, Y) \sim p(x, y)$, the conditional entropy $H(X \mid Y)$ is defined as

$$H(Y \mid X) = -E \log p(Y \mid X).$$

Proposition 1.5 (Chain rule)
$H(X, Y) = H(X) + H(Y|X)$ and $H(X, Y|Z) = H(X|Z) + H(Y|X, Z)$.

Proof. Note that $\log p(X, Y) = \log p(X) + \log p(Y|X)$ and take expectations of both sides. The second result is a corollary of the first. □

Remark 1.6. Note that $H(X|Y) \neq H(Y|X)$ in general, but we do have

$$H(X) - H(X|Y) = H(Y) - H(Y|X).$$

1.3 Relative Entropy and Mutual Information

Relative entropy is a measure of distance between two distributions. It arises naturally in statistics as the expected logarithm of the likelihood ratio.

Definition 1.7 (Kullback-Leibler Distance)**.** The relative entropy or Kullback-Leibler distance between two mass functions p, q is defined as

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

One interpretation of this is the following: if we knew p for some random variable, we could construct a code with average description length $H(p)$. If we used a code for distribution q instead, then it would take $H(p) + D(p||q)$ bits on average to describe the random variable.

Remark 1.8. Note that this is not a norm since it is not symmetric and does not satisfy the triangle inequality. But it does satisfy the positive definite property of a norm.

Definition 1.9 (Mutual information)**.** Consider two random variables X, Y with joint PMF $p(x, y)$ and marginal PMFs $p(x), p(y)$. The mutual information $I(X;Y)$ is the relative entropy between the joint distribution and the product distribution $p(x)p(y)$:

$$I(X;Y) = D(p(x,y)||p(x)p(y)) = E_{p(x,y)} \log \frac{p(X,Y)}{p(X)p(Y)}.$$

Proposition 1.10
$I(X;Y) = H(X) - H(X|Y)$.

Proof. We can rewrite the definition as

$$\begin{aligned} I(X;Y) &= E_{p(x,y)} \log \frac{p(X,Y)}{p(X)p(Y)} \\ &= E_{p(x,y)} \log \frac{p(X|Y)}{p(X)} \\ &= H(X) - H(X|Y). \end{aligned}$$

□

Hence, the mutual information is the reduction in uncertainty of X due to the knowledge of Y. Note that by symmetry, we also have that $I(X;Y) = H(Y) - H(Y|X)$, so X says as much about Y as Y says about X.

We also have the following representations of mutual information summarized below:

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

- $I(X;Y) = H(X) + H(Y) - H(X,Y)$.
- $I(X;Y) = I(Y;X)$.
- $I(X;X) = H(X)$.

We have the following chain rules for entropy, mutual information, and KL-divergence.

Theorem 1.11 (Chain rule for entropies)
Let $X_1, X_2, \ldots, X_n$ be drawn according to $p(x_1, \ldots, x_n)$. Then

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

Define $I(X;Y|Z) = H(X|Z) - H(X|Y,Z)$.

Theorem 1.12 (Chain rule for mutual information)

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

Theorem 1.13 (Chain rule for KL-divergence)

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

1.4 Jensen's inequality

Recall the famous Jensen's inequality:

Theorem 1.14 (Jensen's inequality)
If f is a convex function and X is a random variable,

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

Moreover, if f is strictly convex, the equality case implies that $X = EX$ with probability 1.

Jensen's inequality has immediate consequences in information theory.

Theorem 1.15 (Information inequality)
Let $p(x), q(x), x \in \mathcal{X}$ be two PMFs. Then

$$D(p||q) \geq 0$$

with equality if and only if $p(x) = q(x)$ for all x.

Proof. Let $A = \{x : p(x) > 0\}$. Then,

$$\begin{aligned}
-D(p||q) &= -\sum_{x\in A} p(x) \log \frac{p(x)}{q(x)} \\
&= \sum_{x\in A} p(x) \log \frac{p(x)}{q(x)} \\
&\leq \log \sum_{x\in A} p(x) \frac{q(x)}{p(x)} \\
&= \log \sum_{x\in A} q(x) \\
&\leq \log \sum_{x\in \mathcal{X}} q(x) \\
&= \log 1 = 0
\end{aligned}$$

where we used the concavity of log in the second step. □

As a corollary of this, note that we have

$$0 \leq I(X;Y) = H(X) - H(X|Y),$$

so it follows that

$$H(X|Y) \leq H(X).$$

In other words, information doesn't hurt us on average.

As a corollary of the corollary, we also have

$$H(X_1, \ldots, X_n) \leq \sum_{i=1}^{n} H(X_i).$$

Theorem 1.16

$H(X) \leq \log |\mathcal{X}|$, where $|\mathcal{X}|$ denotes the number of elements in the range of X, with equality if and only if X has a uniform distribution over $\mathcal{X}$.

Proof. Take $u(x) = \frac{1}{|\mathcal{X}|}$ and apply the information inequality to $D(p||u)$. □

1.5 Log-sum inequality

Theorem 1.17 (Log-sum inequality)

For non-negative $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$,

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

with equality if and only if a_i/b_i is a constant.

This follows from the strict convexity of $f(t) = t \log t$ and Jensen's. First, note that The log-sum inequality provides another proof for the information inequality. As a corollary, we have the following:

Corollary 1.18

$D(p||q)$ is convex in (p, q). That is,

$$D(\lambda p_1 + (1-\lambda)p_2||\lambda q_1 + (1-\lambda)q_2) \leq \lambda D(p_1||q_1) + (1-\lambda)D(p_2||q_2)$$

for all $\lambda \in [0, 1]$.

Since we can write $H(p) = \log|\mathcal{X}| - D(p||u)$, we obtain that H is a concave function. Finally, we also obtain a similar result for mutual information:

Theorem 1.19

Take $(X, Y) \sim p(x, y)$. The mutual information $I(X;Y)$ is a concave function of $p(x)$ for fixed $p(y|x)$ and a concave function of $p(y|x)$ for fixed $p(x)$.

1.6 Data-processing inequality, Sufficient statistics

Theorem 1.20

If $X \to Y \to Z$ form a Markov chain , then $I(X;Y) \geq I(X;Z)$.

Proof. Note that $I(X;Z|Y) = 0$ since $X \perp\!\!\!\perp_Y Z$. By the chain rule,

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

and $I(X;Y|Z) \geq 0$, which implies the result. □

Remark 1.21. We only have equality when $I(X;Y|Z) = 0$, or $X \to Z \to Y$ is Markov.

As corollaries of this result, we have

- If $Z = g(Y)$, then $I(X;Y) \geq I(X;g(Y))$
- If $X \to Y \to Z$ then $I(X;Y|Z) \leq I(X;Y)$.

The data-processing inequality is important in the context of statistics. Suppose we have a family of PMFs $\{f_\theta(x)\}$ and X a sample from the distribution. Let $T(X)$ be a statistic of the distirbution. Then, $\theta \to X \to T(X)$, so by the data-processing inequality

$$I(\theta;T(X)) \leq I(\theta;X)$$

for any distribution on θ. But no information is lost if equality holds. This leads to the definition of a sufficient statistic:

Definition 1.22 (Sufficient statistic)**.** A function $T(X)$ is said to be sufficient relative to $\{f_\theta(x)\}$ if X is independent of θ given $T(X)$ for any distribution on θ, i.e. $\theta \to T(X) \to X$ forms a Markov chain.

1.7 Fano's Inequality

Suppose we know Y and want to guess the value of a correlated variable X. Fano's inequality will give us a relation of X to its conditional entropy $H(X|Y)$. This will be crucial to proving the converse of the Shannon channel capacity theorem.

Theorem 1.23 (Fano's inequality)

For any estimator $\widehat{X}$ with $X \to Y \to \widehat{X}$ and $P_e = \Pr(X \neq \widehat{X})$, we have

$$H(P_e) + P_e \log |\mathcal{X}| \geq H(X|\widehat{X}) \geq H(X|Y).$$

This can be weakened to

$$P_e \geq \frac{H(X|Y) - 1}{\log |\mathcal{X}|}.$$

Proof. Define $E = 1_{\widehat{X} \neq X}$. By the chain rule for entropies written in two ways,

$$\begin{aligned} H(E, X|\widehat{X}) &= H(X|\widehat{X}) + H(E|X, \widehat{X}) \\ &= H(E|\widehat{X}) + H(X|E, \widehat{X}). \end{aligned}$$

Then, $H(E|\widehat{X}) \leq H(E) = H(P_e)$ and $H(E|X, \widehat{X}) = 0$, since E is a function of $X, \widehat{X}$. Finally, note that

$$H(X|E, \widehat{X}) = E_E H(X|\widehat{X}, E) \leq (1 - P_e)0 + P_e \log |\mathcal{X}|.$$

It follows that

$$H(P_e) + P_e \log |\mathcal{X}| \geq H(X|\widehat{X}).$$

Furthermore, by the data-processing inequality, we have $I(X; \widehat{X}) \leq I(X; Y)$ since $X \to Y \to \widehat{X}$ is a Markov chain, and therefore, $H(X|\widehat{X}) \geq H(X|Y)$. Thus,

$$H(P_e) + P_e \log |\mathcal{X}| \geq H(X \mid \widehat{X}) \geq H(X \mid Y).$$

□

Note that if $g(Y)$ takes values in $\mathcal{X}$, then we can obtain a slightly stronger inequality:

$$H(P_e) + P_e \log(|\mathcal{X}| - 1) \geq H(X \mid Y),$$

since the number of possible outcomes is now given by $|\mathcal{X}| - 1$.

Example 1.24 (Sharpness of Fano)

Suppose there is no knowledge of Y. Then, X must be guessed without any information. Let $X \in \{1, \ldots, m\}$ and $p_1 \geq \cdots \geq p_m$. The best guess is then $\widehat{X} = 1$, so Fano's inequality becomes

$$H(P_e) + P_e \log(m - 1) \geq H(X).$$

The PMF

$$(p_1, \ldots, p_m) = \left(1 - P_e, \frac{P_e}{m-1}, \ldots, \frac{P_e}{m-1}\right)$$

acheives equality, proving the sharpness of Fano's inequality.

Lemma 1.25

If X, X' are iid with entropy $H(X)$, then

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

with equality if and only if X is uniform.

Proof. Suppose $X \sim p(x)$. By Jensen's

$$2^{E \log p(X)} \leq Ep(x),$$

which implies that

$$2^{-H(X)} \leq \sum p(x)p(x) = \sum p^2(x) = P(X \neq X').$$

□

Corollary 1.26

Let X, X' be independent with $X \sim p(x), X' \sim r(x)$, $x, x' \in \mathcal{X}$. Then

$$\Pr(X = X') \geq 2^{-H(p)-D(p||r)}$$

$$\Pr(X = X') \geq 2^{-H(r)-D(r||p)}$$

2 Asymptotic Equipartition Property

Theorem 2.1 (AEP)

If $X_1, X_2, \ldots \overset{\text{i.i.d.}}{\sim} p(x)$, then

$$-\frac{1}{n}\log p(X_1, \ldots, X_n) \xrightarrow{P} H(X)$$

Proof. It essentially follows from the weak law of large numbers. Since functions of independnent random variables are independent, it follows that $\log p(X_i)$ are independent. Then,

$$\begin{aligned} -\frac{1}{n}\log p(X_1, \ldots, X_n) &= -\frac{1}{n}\log p(X_i) \\ &\xrightarrow{P} -E\log p(X) \\ &= H(X). \end{aligned}$$

□

We can use this to characterize "typical" sets where the sample entropy is close to the true entropy, and "nontypical" sets for the other sequences.

Definition 2.2 (Typical Set)**.** The typical set $A_\epsilon^{(n)}$ with respet to $p(x)$ is the set of sequences $(x_1, \ldots, x_n) \in \mathcal{X}^n$ with the property

$$2^{-n(H(X)+\epsilon)} \leq p(x_1, \ldots, x_n) \leq 2^{-n(H(X)-\epsilon)}.$$

As a consequence of the AEP, we have the following properties.

1. If $(x_1, \ldots, x_n) \in A_\epsilon^{(n)}$, then $H(X) - \varepsilon \leq -\frac{1}{n}\log p(x_1, \ldots, x_n) \leq H(X) + \varepsilon$.
2. $\Pr(A_\epsilon^{(n)}) \geq 1 - \epsilon$ for n sufficiently large.
3. $|A_\epsilon^{(n)}| \leq 2^{n(H(X)+\epsilon)}$.
4. $|A_\epsilon^{(n)}| \geq (1-\epsilon)2^{n(H(X)-\epsilon)}$.

All of these essentially follow from the definition of a typical set and through properties of convergence in probability.

2.1 Data Compression

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} p(x)$. We will divide all sequences in $\mathcal{X}^n$ into the disjoint union $A_\epsilon^{(n)}$ and its complement.

Now, we present a method to generate a short description for such sequences of random variables. Order all elements in each set according to some order. Then, we represent each sequence of $A_\epsilon^{(n)}$ by giving the index of the sequence in the set. This requires no more than $n(H+\epsilon)+1$ bits by the previous theorem. Then, we add a prefix bit of 0 for the sequences in $A_\epsilon^{(n)}$. For the other sequences, we use no more than $n\log|\mathcal{X}|+1$ bits and prefix by 1.

- This code is one-to-one and easily decodable.

- We have used a brute-force enumeration of $C(A_\epsilon^{(n)})$ without taking into account the number of elements is less than the number of elements in $\mathcal{X}^n$.
- The typical sequences have short description lengths.

Theorem 2.3
Let $X^n \overset{\text{i.i.d.}}{\sim} p(x)$ and let $\epsilon > 0$. There exists a code that maps sequences x^n of length n into binary strings such that the mapping is one-to-one so that

$$E\left[\frac{1}{n}\ell(X^n)\right] \leq H(X) + \epsilon$$

for n sufficiently large, where $\ell(x^n)$ represents the length of the codeword corresponding to x^n.

Proof. If n is large enough so that $\Pr(A_\epsilon^{(n)}) \geq 1 - \epsilon$, then

$$\begin{aligned}
E(\ell(X^n)) &= \sum_{x^n} p(x^n)\ell(x^n) \\
&= \sum_{x^n \in A_\epsilon^{(n)}} p(x^n)\ell(x^n) + \sum_{x^n \notin A_\epsilon^{(n)}} p(x^n)\ell(x^n) \\
&\leq \sum_{x^n \in A_\epsilon^{(n)}} p(x^n)(n(H+\epsilon)+2) + \sum_{x^n \notin A_\epsilon^{(n)}} p(x^n)(n\log|\mathcal{X}|+2) \\
&\leq n(H+\epsilon) + \epsilon n(\log|\mathcal{X}|) + 2 \\
&= n(H+\epsilon'),
\end{aligned}$$

where $\epsilon' = \epsilon + \epsilon\log|\mathcal{X}| + \frac{2}{n}$ can be made arbitrarily small. □

2.2 High-probability sets and the typical set

Definition 2.4 (High-probability set)**.** For each $n \in \mathbb{Z}$, let $B_\delta^{(n)} \subset \mathcal{X}^n$ be the smallest set with

$$\Pr(B_\delta^{(n)}) \geq 1 - \delta.$$

Theorem 2.5
Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} p(x)$. For $\delta < 1/2$ and $\delta' > 0$, if $\Pr(B_\delta^{(n)}) > 1 - \delta$, then

$$\frac{1}{n}\log|B_\delta^{(n)}| > H - \delta'.$$

From this theorem, it follows that $B_\delta^{(n)}$ has at least 2^{nH} elements, so $A_\epsilon^{(n)}$ has about the same size as the smallest high-probability set.

We will use the following notation to charaterize first order equality in the exponent:

$$a \equiv b \iff \frac{1}{n}\log\frac{a_n}{b_n} \to 0.$$

Using this notation, we can restate the result as follows: if $\delta_n \to 0$ and $\epsilon_n \to 9$, then

$$|B_{\delta_n}^{(n)}| \equiv |A_{\epsilon_n}^{(n)}| \equiv 2^{nH}$$

3 Entropy Rates of a Stochastic Process

We assume basic familiarity with Markov chains and their associated terminology.

3.1 Entropy Rate

Definition 3.1 (Entropy rate)**.** The entropy of a stochastic process $\{X_i\}$ is defined by

$$H(\mathcal{X}) = \lim_{n\to\infty} \frac{1}{n} H(X_1, \ldots, X_n)$$

when the limit exists.

Some simple examples of stochastic processes with their entropy rates are as follows:

- Typewriter. Consider a typewriter that has m equally likely output letters. The typewriter can produce m^n sequences of length n, all of them equally likely. Hence $H(X_1, \ldots, X_n) = \log m^n$ and the entropy rate is $\log m$ bits per symbol.
- $X_1, X_2, \ldots$ are i.i.d. random variables. Then $H(\mathcal{X}) = H(X_1)$.
- Sequence of independent variables. In this case,

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

  which may or may not exist.

A related quantity is

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

$H(\mathcal{X})$ corresponds to the per symbol entropy of the n random variables, while $H'(\mathcal{X})$ is the conditional entropy of the last variable given the past.

We have the following result for stationary proesses:

Theorem 3.2

For a stationary stochastic process, $H(\mathcal{X}) = H'(\mathcal{X})$ and both always exist.

Proof. We first prove that $H'(\mathcal{X})$ exists. Note that

$$H(X_{n+1} | X_1, \ldots, X_n) \leq H(X_{n+1} | X_n, \ldots, X_2) = H(X_n | X_{n-1}, \ldots, X_1).$$

It follows that $H(X_n | X_{n-1}, \ldots, X_1)$ is a decreasing sequence of nonnegative numbers so it has a limit, $H'(\mathcal{X})$.

Now, note that

$$\begin{aligned} H(\mathcal{X}) &= \lim \frac{H(X_1, \ldots, X_n)}{n} \\ &= \lim \frac{1}{n} \sum_{i=1}^{n} H(X_i | X_{i-1}, \ldots, X_1) \\ &= \lim H(X_n | X_{n-1}, \ldots, X_1) \\ &= H'(\mathcal{X}), \end{aligned}$$

where we used the Cesáro mean theorem in the third equality. □

3.1.1 Markov Chain Entropy Rate

For a stationary Markov chain, the entropy rate is given by

$$H(\mathcal{X}) = H'(\mathcal{X}) = \lim H(X_n|X_{n-1}) = H(X_2|X_1).$$

We can calculate the conditional entropy using the stationary distribution. In particular, note that

$$H(\mathcal{X}) = H(X_2|X_1) = \sum_i \mu_i (\sum_j -P_{ij} \log P_{ij}).$$

3.2 Second Law of Thermodynamics

In thermodynamics, the basic laws of physics states that the entropy of an isolated system is nondecreasing. We now explore the relationship between the second law and the entropy function.

We model the isolated system as a Markov hain with transitions obeying the physical laws governing the system. There are many different interpretations of the second law:

1. Relative entropy $D(\mu_n||\mu'_n)$ decreases with n. Let μ_n and μ'_n be two probability distributions on the state space of a Markov chain at time n, with corresponding joint mass functions p, q. Then, $p(x_n, x_{n+1}) = p(x_n)r(x_{n+1}|x_n)$ and $q(x_n, x_{n+1}) = q(x_n)r(x_{n+1}|x_n)$ where $r(\cdot|\cdot)$ is the probability transition function for the Markov chain. By the relative entropy chain rule,

$$D(p(x_n, x_{n+1})||q(x_n, x_{n+1})) = D(p(x_n)||q(x_n)) + D(p(x_{n+1}|x_n)||q(x_{n+1}|x_n))$$

$$D(p(x_n, x_{n+1})||q(x_n, x_{n+1})) = D(p(x_{n+1})||q(x_{n+1})) + D(p(x_n|x_{n+1})||q(x_n|x_{n+1}))$$

   Since p and q are from the same Markov chain, $p(x_{n+1}|x_n) = q(x_{n+1}|x_n) = r(x_{n+1}|x_n)$ so it follows that $D(p(x_{n+1}|x_n)||q(x_{n+1}|x_n)) = 0$. Using the information inequality, it follows that

$$D(\mu_n||\mu'_n) \geq D(\mu_{n+1}|\mu'_{n+1}).$$

2. Relative entropy $D(\mu_n||\mu)$ between a distribution μ_n on the states at time n anda stationary distirbution μ decreases with n. This is because if μ'_n is a stationary μ, then μ'_{n+1} is also μ, so

$$D(\mu_n||\mu) \geq D(\mu_{n+1}||\mu)$$

   which implies that any state distribution gets closer to the stationary distribution as time passes.

3. Entropy increases if the stationary distirbution is uniform, since we can express it as

$$D(\mu_n||\mu) = \log|\mathcal{X}| - H(X_n).$$

4. The conditional entropy increases with $H(X_n \mid X_1)$ increases with n for a stationary Markov process. This follows from the dataa processing inequality applied to

$$X_1 \to X_{n-1} \to X_n \implies I(X_1; X_{n-1}) \geq I(X_1; X_n).$$

   Then, we expand in terms of entropies accountd noting that $H(X_{n-1}) = H(X_n)$, we obtain the desired result.

5. Shuffling increases entropy - we will show this in a future exercise.

3.3 Functions of Markov Chains

Suppose we have a stationary Markov chain $(X_n)_{n\geq 0}$ and $Y_i = \varphi(X_i)$ for some function of the corresponding state. Our goal will be to compute $H(\mathcal{Y})$. We already know that $H(Y_n|Y_{n-1},\dots,Y_1)$ converges monotonically to $H(\mathcal{Y})$ from above.

Theorem 3.3

If $(X_n)_{n\geq 0}$ forms a stationary Markov chain, and $Y_i = \varphi(X_i)$, then

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

and

$$\lim H(Y_n|Y_{n-1},\dots,Y_1,X_1) = H(\mathcal{Y}) = \lim H(Y_n|Y_{n-1},\dots,Y_1).$$

Proof. First, we prove the lower bound. Note that

$$\begin{aligned}
H(Y_n|Y_{n-1},\dots,Y_2,X_1) &= H(Y_n \mid Y_{n-1},\dots,Y_2,Y_1,X_1) \\
&= H(Y_n \mid Y_{n-1},\dots,Y_1,X_1,X_0,X_{-1},\dots,X_{-k}) \\
&= H(Y_n \mid Y_{n-1},\dots,Y_1,X_1,X_0,X_{-1},\dots,X_{-k},Y_0,\dots,Y_{-k}) \\
&\leq H(Y_n \mid Y_{n-1},\dots,Y_1,Y_0,\dots,Y_{-k}) \\
&= H(Y_{n+k+1} \mid Y_{n+k},\dots,Y_1)
\end{aligned}$$

Since this inequality holds for all k, it follows in the limit.

Next, we show that the interval between the upper and lower bounds decreases in length. It can be written as

$$H(Y_n \mid Y_{n-1},\dots,Y_1) - H(Y_n \mid Y_{n-1},Y_1,X_1) = I(X_1;Y_n|Y_{n-1},\dots,Y_1)$$

It is clear that $I(X_1;Y_1,\dots,Y_n) \leq H(X_1)$ and it also increases with n. It follows that the limit above is bounded by $H(X_1)$. Then, note that

$$\begin{aligned}
H(X) &\geq \lim_{n\to\infty} I(X_1;Y_1,\dots,Y_n) \\
&= \lim_{n\to\infty} \sum_{i=1}^{n} I(X_1;Y_i \mid Y_{i-1},\dots,Y_1) \\
&= \sum_{i=1}^{\infty} I(X_1;Y_i \mid Y_{i-1},\dots,Y_1)
\end{aligned}$$

Since the sum is finite and the terms are non-negative, they must tend to zero in the limit, which proves the result. □

4 Data Compression

We use the definition of entropy in order to establish fundamental limits for the compression of information.

4.1 Examples of Codes

Definition 4.1 (Source Code)**.** A source code C is a map from $\mathcal{X} = \text{range}(X) \to \mathcal{D}^*$, the set of finite-length strings of symbols from a D-ary alphabet. $C(x)$ denotes the codeword correspond to x and $\ell(x)$ denotes the length of $C(x)$.

Definition 4.2 (Expected Length)**.** The expected length $L(C)$ of a source code $C(x)$ for a random variable X with p.m.f. $p(x)$ is given by

$$L(C) = \sum_{x \in \mathcal{X}} p(x)\ell(x).$$

Let $x^n = (x_1, \ldots, x_n)$.

Definition 4.3 (Nonsingular Code)**.** A code is said to be nonsingular if every element of the $\mathcal{X}$ maps into a different string in $\mathcal{D}^*$, i.e. C is injective.

Definition 4.4 (Code Extension)**.** The extension C^* of a code C is the mapping from finite-length strings of $\mathcal{X}$ to finite-Length strings of $\mathcal{D}$, defined by

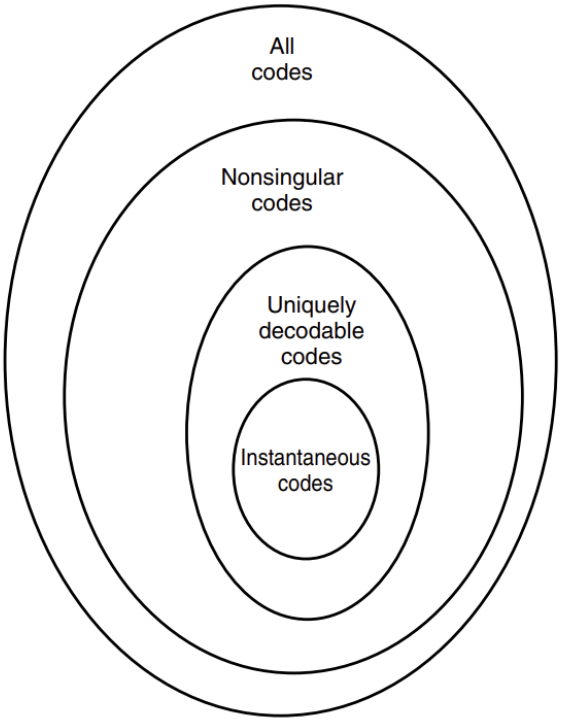
$$C(x_1 \ldots x_n) = C(x_1) \ldots C(x_n)$$

where multiplication corresponds to concatenation.

Definition 4.5 (Uniquely Decodable)**.** A code is uniquely decodable if its extension is nonsingular.

Definition 4.6 (Prefix Code)**.** A code is called a prefix code or an instantaneous code if no codeword is a prefix of any other codeword.

Note the following heirarchy of codes:



4.2 Kraft Inequality

The set of codeword lengths possible for instantaneous codes is limited by the following inequality:

Theorem 4.7 (Kraft Inequality)
For any instantaneous code over an alphabet of size D, the codeword lengths $\ell_1, \ldots, \ell_m$ must satisfy the inequality

$$\sum_i D^{-\ell_i} \leq 1.$$

Conversely, given a set of codeword lengths that satisfy the inequalty, there exists an instantaneous code with these word lengths.

Proof. Consider a D-ary tree where each node has D children. The branches of the tree represent the symbols of the codeword, and the corresponding words are given by leaves of the tree. The path from the root traces out the symbols of the codeword. By the prefix condition, no codeword is an ancestor of any other codeword on the tree, which implies that each codeword eliminates all its descendents as codewords.

Let ℓ_{max} by the length of the longest codeword of the set of codewords. Consider the nodes of the tree at level ℓ_{max}. A codeword at ℓ_i has $D^{\ell_{max}-\ell_i}$ descendents at level ℓ_{max}. Also, note that each of these descendent sets must be disjoint. Finally, the total number of nodes is at most $D^{\ell_{max}}$. Combining all of these, we obtain

$$\sum D^{\ell_{max}-\ell_i} \implies \sum D^{\ell_i} \leq 1.$$

For the converse, note that given lengths $\ell_1, \ldots, \ell_m$, we can construct a D-ary tree as above. Label the first node of depth ℓ_1 as 1 and removes the descendents. Then, label the first remaining node of depth ℓ_2 as 2 and remove descendents. We can repeat this procedure to construct a prefix code with the specified $\ell_1, \ldots, \ell_m$. □

Theorem 4.8 (Extended Kraft Inequality)
For any countably infinite set of codewords that form a prefix code, the codeworth lengths satisfy the extended Kraft inequality,

$$\sum_{i \geq 1} D^{-\ell_i} \leq 1.$$

4.3 Optimal Codes

Now, we address the problem of finding the prefix code with minimum expected length. This is equivalent to finding a set of lengths that satisfy the Kraft inequality with minimal expected length $L = \sum p_i \ell_i$. This is a standard optimization problem:

$$\text{argmin} \quad L = \sum p_i \ell_i$$

$$\ell_1, \ldots, \ell_m \in \mathbb{Z}_{>0}, \quad \sum D^{-\ell_i} \leq 1.$$

If we remove the integer constraint, we can solve this using Lagrange multipliers obtaining the result

$$\ell_i^* = -\log_D p_i, \implies L^* = H_D(X).$$

However, we only have integer lengths. The optimality among integers is verified in by the following theorem:

Theorem 4.9

The expected length L of any instantaneous D-ary code for a random variable X is greater than or equal to the entropy $H_D(X)$, with equality if and only if $D^{-\ell_i} = p_i$.

Proof. Note that

$$\begin{aligned} L - H_D(X) &= \sum p_i \ell_i - \sum p_i \log_D \frac{1}{p_i} \\ &= -\sum p_i \log_D D^{-\ell_i} + \sum p_i \log_D p_i \\ &= \sum p_i \log \frac{p_i}{r_i} - \log_D c \\ &= D(p||r) + \log_D \frac{1}{c} \geq 0 \end{aligned}$$

where we define $r_i = D^{-\ell_i} / \sum_j D^{-\ell_j}$, $c = \sum D^{-\ell_j}$. □

Definition 4.10 (D-adic)**.** A probability distribution is called D-adic if each of the probabilities is equal to D^{-n} for some n. Therefore, we have equality in the previous theorem if and only if the distribution of X is D-adic.

Remark 4.11. The above results indicate procedures for findind an optimal code, but finding a closest D-adic distribution is not an easy problem. We will provide suboptimal procedures (Shannon-Fano coding) and optimal procedures (Huffman Coding) to actually obtain the optimal code.

Theorem 4.12

Let $\ell_1^*, \ldots, \ell_m^*$ be the optimal codeword lengths for a source distribution p and a D-ary alphabet, and let L^* be the associated expected length of an optimal code. Then

$$H_D(X) \leq L^* < H_D(X) + 1.$$

Proof. Take $\ell_i = \left\lceil \log_D \frac{1}{p_i} \right\rceil$ to obtain the result. □

In order to reduce the overhead of 1 bit, we can spread it our over many symbols. In particular, we draw n symbols i.i.d. from $p(x)$. Define L_n to be the expected codeword length per input symbol:

$$L_n = \frac{1}{n} El(X_1, \ldots, X_n).$$

We can apply the bound from before to obtain

$$H(X_1, \ldots, X_n) \leq El(X_1, \ldots, X_n) < H(X_1, \ldots, X_n) + 1$$

. Finally, since $X_1, \ldots, X_n$ are i.i.d., it follows that $H(X_1, \ldots, X_n) = nH(X)$, so we can divide by n to obtain

$$H(X) \leq L_n < H(X) + \frac{1}{n}.$$

If the symbols are not i.i.d, then we still obtain

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

If the stochastic process is stationary, then $H(X_1, \ldots, X_n)/n \to H(\mathcal{X})$, so the expected description length tends to the entropy rate as $n \to \infty$.

Finally, we show a result that describes the expected length when we choose a wrong code.

Theorem 4.13 (Wrong Code)

The expected length under $p(x)$ of the code assignment $\ell(x) = \left\lceil \log \frac{1}{q} \right\rceil$ satisfies

$$H(p) + D(p||q) \leq E_p \ell(x) < H(p) + D(p||q) + 1.$$

Proof.

$$\begin{aligned}
E\ell(X) &= \sum_x p(x) \left\lceil \log \frac{1}{q(x)} \right\rceil \\
&\leq \sum_x p(x) \left(\log \frac{1}{q(x)} + 1 \right) \\
&= \sum_x p(x) \log \frac{p(x)}{q(x)} \frac{1}{p(x)} + 1 \\
&= \sum_x p(x) \log \frac{p(x)}{q(x)} + \sum_x p(x) \log \frac{1}{p(x)} + 1 \\
&= D(p||q) + H(p) + 1.
\end{aligned}$$

The lower bound is derived similarly. □

We now show that the class of uniquely decodable codes does not offer any further possibilities for the set of codeword lengths than do instantaneous codes.

Theorem 4.14 (McMillan)

The codeword lengths of any uniquely decodable D-ary code must satisfy the Kraft inequality:

$$\sum D^{\ell_i} \leq 1.$$

Conversely, given a set of codeword lengths that satisfy this inequality, it is possible to construct a set of uniquely decodable code with these codeword lengths.

Proof. Consider C^k, the kth extension of a code (given by concatenating k reptitions of the given uniquely decodable code C). By definition, C^k is nonsingular. Furthermore, there are only D^n different D-ary strings of length n, so unique decodability implies that the number of code sequences of length n in C^k must be no greater than D^n.

Now, we prove Kraft's inequality. Let the codeword length of $x \in \mathcal{X}$ be denoted by $l(x)$.

Note that

$$\begin{aligned}\left(\sum_{x\in\mathcal{X}} D^{-l(x)}\right)^k &= \sum_{x_1\in\mathcal{X}} \cdots \sum_{x_k\in\mathcal{X}} D^{-\sum_{i=1}^k l(x_i)} \\ &= \sum_{x^k\in\mathcal{X}^k} D^{-l(x^k)} \\ &= \sum_{m=1}^{kl_{max}} a(m) D^{-m} \\ &\leq \sum_{m=1}^{kl_{max}} D^m D^{-m} \\ &= kl_{max}\end{aligned}$$

which implies that

$$\sum_j D^{-l_j} \leq (kl_{max})^{1/k} \xrightarrow{k\to\infty} 1.$$

The converse follows from the converse of Kraft's inequality, since instantaneous codes are also uniquely decodable. □

Corollary 4.15

A uniquely decodable code for an infinite source alphabet $\mathcal{X}$ also satisfies the Kraft inequality.

4.4 Huffman Codes

In this special case, the integer programming problem can be solved exactly (which is exceedingly rare). This is given by the **Huffman coding algorithm**, which is essentially to recursively merge the smallest probability remaining pair of symbols, and label each branch with a 0 or a 1. The D-ary version requires combining the smallest D probabilities at a time.

Now, we prove optimality. Without loss of generality, we assume the probability masses are ordered, so that $p_1 \geq \cdots \geq p_m$.

Lemma 4.16

For any distribution, there exists an optimal instantaneous code that satisfies the following properties:

- The lengths are ordered inversely with the probabilities.
- The two longest codewords have the same length.
- Two of the longest codewords differ only in the last bit and correspond to the two least likely symbols.

Proof. We prove each of the parts as follows:

- If not, interchange $l(x)$ and $l(y)$ for a better code.
- If not, we can reduce the length of the longest codeword by removing a bit from the end, which gives a better code.
- If the sibling of a longest length codeword is not present, we can reduce the length of that codeword by removing the last bit. We can guarantee these are the smallest probability symbols by relabeling.

□

Theorem 4.17

Huffman coding is optimal; that is, if C^* is a Huffman code and C' is any other uniquely decodable code, $L(C^*) \leq L(C')$

Proof. Easily follows from induction using the above lemma. □

4.5 Shannon-Fano-Elias Coding

Without loss of generality, take $\mathcal{X} = [m]$. Assume that $p(x) > 0$ for all x. The cumulative distribution function $F(x)$ is defined as

$$F(x) = \sum_{a \leq x} p(a).$$

Consider the modified CDF,

$$\overline{F}(x) = \sum_{a < x} p(a) + \frac{1}{2} p(x).$$

Since all the probabilities are positive, we can determine x if we know $\overline{F}(x)$, so it provides a code for x. But this is a real number in general, so we use an approximate value in general.

If we truncate to $l(x)$ bits, then we have

$$\overline{F}(x) - \left\lfloor \overline{F}(x) \right\rfloor_{l(x)} < 2^{-l(x)}.$$

If we choose $l(x) = \left\lceil \log \frac{1}{p(x)} \right\rceil + 1$, then we have

$$2^{-l(x)} < \frac{p(x)}{2} = \overline{F}(x) - F(x-1).$$

It follows that the truncated value lies within the step corresponding to x, so $l(x)$ bits suffice to describe x. It is easy to verify that this code is prefix free.

Moreover, the expected code length is given by

$$L = \sum_x p(x) \left(\left\lceil \log \frac{1}{p(x)} \right\rceil + 1 \right) < H(X) + 2.$$

Remark 4.18. The arithmetic coding scheme is based on updating the above procedure as we go.

4.6 Solutions to selected problems

Exercise 4.19 (5.1)**.** Let $L = \sum_{i=1}^{m} p_i l_i^{100}$. Let $L_1 = \min L$ over all instantaneous codes and $L_2 = \min L$ over all uniquely decodable codes. What inequality exists between L_1 and L_2?

Proof. $L_1 = L_2$. It is obvious that $L_2 \leq L_1$ but we also have $L_1 \leq L_2$ since the codelengths acheiving the minimum for L_2 will satisfy the Kraft inequality, and we can thus construct an instantaneous codeword with the same lengths, giving the same minimum. □

Remark 4.20. When it comes to lengths, instantaneous and uniquely decodable codes are pretty much equivalent due to the Kraft-McMillan inequality.

Exercise 4.21 (5.3, Slackness in the Kraft inequality)**.** An instantaneous code has lengths $l_1, \ldots, l_m$, which satisfy

$$\sum_{i=1}^{m} D^{-l_i} < 1.$$

The code alphabet is $\mathcal{D} = [D]$. Show that there exist arbitrarily long sequences of code symbols in $\mathcal{D}^*$ which cannot be decoded into sequences of codewords.

Proof. Without loss of generality, suppose $l_1 \leq \cdots \leq l_m$. Since the codewords satisfy the Kraft inequality, we can construct a corresponding D-ary tree, where we take each edge with probability $1/D$. But since the inequality is strict, it implies that some branches of the tree do not correspond to any of the codewords. Since we can branch infinitely, there are arbitrarily long sequences that cannot be decoded.

Alternatively, let $l = \max(l_1, \ldots, l_m)$. There are D^l total sequences and of these sequences D^{l-l_i} start with the ith codeword. By the prefix condition, no two sequences can start with the same codeword. It follows that the total number of sequences that start with some codeword is given by

$$\sum_{i=1}^{m} D^{l-l_i} < D^l$$

which implies there are some sequences that do not start with any codeword. □

5 Channel Capacity

Channel capacity is an important concept dealing with the maximum mutual information between two communicating channels. We describe the concept formally below.

Definition 5.1 (Discrete channel)**.** A discrete channel is a system consisting of an input alphabet $\mathcal{X}$ and an output alphabet $\mathcal{Y}$ with a probability transition matrix $p(y \mid x)$ that expresses the probability of observing the output symbol y given that we send the symbol x. The channel is said to be memoryless if the probability distribution of the output depends only on the input at that time and is conditionally independent of the previous channel inputs or outputs.

Definition 5.2 (Channel capacity)**.** The "information" channel capacity of a discrete memoryless channel is given by

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

where the maximum is taken over all possible input distributions $p(x)$.

> **Remark 5.3.** We specify the "information" channel capacity because there is also an operational definition, which is the highest rate in bits per channel at which information can be sent with arbitrarily low error probability. But Shannon's second theorem establishes that the two notions are equivalent.

Some properties of channel capacity are as follows:

- $C \geq 0$.
- $C \leq \max(\log |\mathcal{X}|, \log |\mathcal{Y}|)$.
- $I(X;Y)$ is a continuous function of $p(x)$.
- $I(X;Y)$ is a concave function of $p(x)$.

5.1 Examples

We now present several examples. There are corresponding diagrams present in Cover and Thomas in section 7.1.

- Noiseless binary channel: a channel whose binary input is reproduced exactly at the output. In this case, any transmitted bit is received without error. It is clear in this case that $C = \max I(X;Y) = 1$, acheived by $p(x) = (1/2, 1/2)$.
- Noisy channel with nonoverlapping outputs: each input channel in this case has two possible output channels. But in reality, this is not noisy at all because we can determine the input from the output since they are nonoverlapping. So again $C = 1$.
- Noisy Typewriter: In this case, the channel input is either received unchanged or transformed into the next letter with probability 1/2. We can transmit every alternate symbol in this case to return to the nonoverlapping outputs case. This gives

  $$C = \max I(X;Y) = \max(H(Y) - H(Y \mid X)) = \max H(Y) - 1 = \log(26) - 1 = \log 13.$$

- Binary symmetric channel: our first interesting example. This is a binary channel where the input symbols are complemented with probability p. We can bound the mutual information as follows:

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y \mid X) \\ &= H(Y) - \sum p(x) H(Y \mid X = x) \\ &= H(Y) - \sum p(x) H(p) \\ &= H(Y) - H(p) \\ &\leq 1 - H(p) \end{aligned}$$

  Equality is acheived when the input distribution is uniform, so $C = 1 - H(p)$.

- Binary Erasure channel: this is an analog where a fraction α of the bits are erased, and the receiver knows which bits have been erased. The capacity is given by

$$C = \max_{p(x)} H(Y) - H(\alpha).$$

  Letting $E = \{Y = e\}$, we obtain:

$$H(Y) = H(Y, E) = H(E) + H(Y \mid E)$$

  and letting $\Pr(X = 1) = \pi$, we obtain

$$H(Y) = H(\alpha) + (1 - \alpha) H(\pi)$$

  which gives $C = 1 - \alpha$, where the capacity is acheived by $\pi = 1/2$.

5.2 Symmetric Channels

Definition 5.4 (Symmetric channel)**.** A channel is said to be symmetric if the rows of the channel transition matrix $p(y \mid x)$ are permutations of each other. A channel is said to be weakly symmetric if every row of the transition matrix $p(\cdot \mid x)$ is a permutation of every other row and all the column sums $\sum_x p(y \mid x)$ are equal.

Theorem 5.5

For a weakly symmetric channel,

$$C = \log |\mathcal{Y}| - H(\text{row of transition matrix})$$

and this is achieved by a uniform distribution on the input alphabet.

Proof. It is clear that

$$I(X;Y) = H(Y) - H(r) \leq \log |\mathcal{Y}| - H(r)$$

and we can see that for $p(x) = 1/|\mathcal{X}|$, we obtain

$$p(y) = \sum_{x \in \mathcal{X}} p(y \mid x) p(x) = \frac{c}{|\mathcal{X}|} = \frac{1}{|\mathcal{Y}|}$$

where c is the sum of entries in one column of the transition matrix. □

5.3 Channel Coding Theorem

5.3.1 Basic setup

We will analyze a communication system, which has the following setup. A message W is drawn from the index set $[M]$ and results in the signal $X^n(W)$, which is received by the receiver as a random sequence $Y^n \sim p(y^n \mid x^n)$. The receiver then guesses the index W by an appropriate decoding rule $\widehat{W} = g(Y^n)$. The receiver makes an error if $\widehat{W} \neq W$.

Definition 5.6. The nth extension of the descrete memoryless channel is the channel $(\mathcal{X}^n, p(y^n \mid x^n), \mathcal{Y}^n)$, where

$$p(y_k \mid x^k, y^{k-1}) = p(y_k \mid x_k)$$

Remark 5.7. If the channel is used without feedback (the input symbols do not depend on the past output symbols), the channel transition function for the nth extension of the discrete memoryless channel reduces to

$$p(y^n \mid x^n) = \prod_{i=1}^{n} p(y_i \mid x_i).$$

Definition 5.8 ((M, n)-code)**.** An (M, n)-code for the channel $(\mathcal{X}, p(y \mid x), \mathcal{Y})$ consists of the following:

1. An index set $[M]$.
2. An encoding function $X^n : [M] \to \mathcal{X}^n$ yielding codewords $x^n(1), \ldots, x^n(M)$. The set of codewords is called the codebook.
3. A decoding function $g : \mathcal{Y}^n \to [M]$, which is a deterministic rule that assigns a guess to each possible received vector.

Definition 5.9 (Conditional error probability)**.** Let

$$\lambda_i = \Pr(g(Y^n) \neq i | X^n = x^n(i)) = \sum_{y^n} p(y^n \mid x^n(i)) I(g(y^n) \neq i)$$

be the conditional probability of error given that index i was sent, where $I(\cdot)$ is the indicator function.

Definition 5.10 (Maximal probability of error)**.** The maximal error probability $\lambda^{(n)}$ for an (M, n)-code is defined as

$$\lambda^{(n)} = \max_{i \in [M]} \lambda_i$$

Definition 5.11 (Average probability of error)**.** The arithmetic error probability $P_e^{(n)}$ for an (M, n)-code is defined as

$$P_e^{(n)} = \frac{1}{M} \sum_{i=1}^{M} \lambda_i$$

Definition 5.12 (Code rate)**.** The rate R of an (M, n)-code is

$$R = \frac{\log M}{n}$$

with units bits per transmission.

Definition 5.13 (Acheivable rates)**.** A rate R is said to be acheivable if there exists a sequence of $(\lceil 2^{nR} \rceil, n)$ codes such that the $\lambda^{(n)}$ tends to 0 as $n \to \infty$.

Definition 5.14 (Capacity)**.** The capacity of a channel is the supremum of all achievable rates.

5.4 Jointly Typical Sequences

Definition 5.15 (Jointly Typical Sequences)**.** The set $A_\epsilon^{(n)}$ of jointly typical sequences $\{(x^n, y^n)\}$ with respect to the distribution $p(x, y)$ is the set of n-sequences with empirical entropies ϵ-close to the true entropies:

$$A_\epsilon^{(n)} = \{(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n : |H(x^n) - H(X)|, |H(y^n) - H(Y)|, |H(x^n, y^n) - H(X, Y)| < \epsilon\}$$

where

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

Theorem 5.16 (Joint AEP)

Let $(X^n, Y^n) \overset{\text{i.i.d.}}{\sim} p(x^n, y^n)$. Then:

1. $\Pr((X^n, Y^n) \in A_\epsilon^{(n)}) \to 1$ as $n \to \infty$.
2. $|A_\epsilon^{(n)}| \leq 2^{(n(H(X,Y)+\epsilon))}$
3. If $(X^n, Y^n) \sim p(x^n)p(y^n)$, then

$$\Pr((X^n, Y^n) \in A_\epsilon^{(n)}) \leq 2^{-n(I(X;Y)-3\epsilon)}$$

and for sufficiently large n,

$$\Pr((X^n, Y^n) \in A_\epsilon^{(n)}) \geq (1 - \epsilon)2^{-n(I(X;Y)+3\epsilon)}$$

The proof is simple, pretty much following the same as the AEP proof.

5.5 Channel Coding Theorem

Theorem 5.17 (Channel Coding Theorem)

For a DMC, all rates below capacity C are acheivable. Specifically, for every rate $R < C$, there exists a sequence of $(2^{nR}, n)$ codes with maximum probability of error $\lambda^{(n)} \to 0$.

Conversely, any sequence of $(2^{nR}, n)$ codes with $\lambda^{(n)} \to 0$ must have $R \leq C$.

Proof. Fix $p(x)$. Generate a $(2^{nR}, n)$ code at random according to the distribution $p(x)$. Consider the corresponding random codebook:

$$\mathcal{C} = \begin{bmatrix} x_1(1) & x_2(1) & \dots & x_n(1) \\ \vdots & \vdots & \ddots & \vdots \\ x_1(2^{nR}) & x_2(2^{nR}) & \dots & x_n(2^{nR}) \end{bmatrix}$$

Each entry in the matrix is generated i.i.d. according to $p(x)$, so the probability we generate a code $\mathcal{C}$ is given by

$$\Pr(C) = \prod_{w=1}^{2^{nR}} \prod_{i=1}^{n} p(x_i(w)).$$

We will consider the following sequence of events:

1. A random code $\mathcal{C}$ is generated according to $p(x)$.
2. The code $\mathcal{C}$ is revealed to the sender and receiver. They are both assumed to know the channel transition matrix $p(y \mid x)$.
3. A message W is chosen according to a uniform distribution $W \sim \text{Unif}([2^{nR}])$.
4. The wth codeword $X^n(w)$ is sent over the channel.
5. The receiver recieves a sequence Y^n according to the distribution $P(y^n|x^n(w))$.
6. The receiver guesses which message was sent. Although the optimum procedure is maximal likelihood encoding, this is difficult to analyze, so we choose jointly typical decoding which is asymptotically optimal. In this procedure, the receiver declares that the index $\widehat{W}$ was sent if the following conditions are satisfied:
   - $(X^n(\widehat{W}), Y^n)$ is jointly typical.
   - There is no other index $W' \neq \widehat{W}$ such that $(X^n(W'), Y^n) \in A_\epsilon^{(n)}$.

   If no such $\widehat{W}$ is declared or more than one exists, an error is declared.
7. There is a decoding error if $\widehat{W} \neq W$. Let $\mathcal{E} = \{\widehat{W}(Y^n) \neq W\}$.

Now, we evaluate the error probability. Note that

$$\begin{aligned}
\Pr(\mathcal{E}) &= \sum_{\mathcal{C}} \Pr(\mathcal{C}) P_e^{(n)}(\mathcal{C}) \\
&= \sum_{\mathcal{C}} \Pr(\mathcal{C}) 2^{-nR} \sum_{w=1}^{2^{nR}} \lambda_w(\mathcal{C}) \\
&= 2^{-nR} \sum_{w=1}^{2^{nR}} \sum_{\mathcal{C}} \Pr(\mathcal{C}) \lambda_w(\mathcal{C}),
\end{aligned}$$

where $P_e^{(n)}(\mathcal{C})$ is defined for jointly typical decoding. By symmetry, the average probability does not depend on w, so we obtain

$$\Pr(\mathcal{E}) = \sum_{\mathcal{C}} \Pr(\mathcal{C}) \lambda_1(\mathcal{C}) = \Pr(\mathcal{E}|W = 1).$$

Define the events

$$E_i = \{(X^n(i), Y^n) \in A_\epsilon^{(n)}\}, i \in [2^{nR}]$$

where E_i is the event that the ith codeword and Y^n are jointly typical.

It follows that

$$\begin{aligned}\Pr(\mathcal{E}|W=1) &= P(E_1^c \cup E_2 \cup \cdots \cup E_{2^{nR}}|W=1) \\ &\leq P(E_1^c|W=1) + \sum_{i=2}^{2^{nR}} P(E_i|W=1) \\ &\leq \epsilon + (2^{nR}-1)2^{-n(I(X;Y)-3\epsilon)} \\ &\leq 2\epsilon\end{aligned}$$

if we choose n sufficiently large and $R < I(X;Y) - 3\epsilon$. Therefore, if we have $R < I(X;Y)$, we can choose n and ϵ so that the average error probability is less than ϵ.

We can strengthen the conclusion to low maximal error probability by a series of code selections:

1. Choose $p(x)$ to be the distribution on X that acheives capacity. Thus, the condition can be replaced with $R < C$.

2. Remove the average over codebooks. Note that there exists at least one codebook $\mathcal{C}^*$ with small average error probability, so $\Pr(\mathcal{E}|\mathcal{C}^*) \leq 2\epsilon$. This can be acheived by an exhaustive search over all $(2^{nR}, n)$ codes. Then, note that

$$\Pr(\mathcal{E}|\mathcal{C}^*) = 2^{-nR} \sum_{i=1}^{2^{nR}} \lambda_i(\mathcal{C}^*)$$

   since we have chosen $\widehat{W}$ according to a uniform distribution.

3. Throw away the worst half of codewords in the best codebook. Since $P_e^{(n)}(\mathcal{C}^*)$ for this code is less than 2ϵ, we have

$$\Pr(\mathcal{E}|\mathcal{C}^*) \leq 2^{-nR} \sum \lambda_i(\mathcal{C}^*) \leq 2\epsilon$$

   which implies that at least half the indices and their corresponding codewords have conditional erroro probability of error λ_i less than 4ϵ. Therefore, the best half of codewords have a maximal error probability of 4ϵ. We can reindex so that these are 2^{nR-1} and throwing out half the codewords changes the rate from R to $R - 1/n$ which is negligible for large n.

□

Before proving the converse, we show the result in the case of zero-error codes. In particular, we show that $P_e^{(n)} = 0$ implies that $R \leq C$. Assume that we have a $(2^{nR}, n)$ code with zero error probability. Then, the input index W is determined by the output sequence. To obtain a strong bound, assume that $W \sim \text{Unif}([2^{nR}])$, so that $H(W) = nR$. Then, we can write

$$\begin{aligned}nR &= H(W) \\ &= H(W|Y^n) + I(W;Y^n) \\ &= I(W;Y^n) \\ &\leq I(X^n;Y^n) \\ &\leq \sum_{i=1}^{n} I(X_i;Y_i) \\ &\leq nC\end{aligned}$$

where we will prove the second-to-last inequality in a later lemma. The one before that follows from the data-processing inequality.

Now, we show the converse. First, recall that $P_e^{(n)} = \Pr(W \neq \widehat{W})$ and by Fano's inequality applied to W,

$$H(W|\widehat{W}) \leq 1 + P_e^{(n)} nR.$$

Lemma 5.18

Let Y^n be the result of passing X^n through a DMC of capacity C. Then

$$I(X^n; Y^n) \leq nC$$

for all $p(x^n)$.

Proof.

$$\begin{aligned}
I(X^n; Y^n) &= H(Y^n) - H(Y^n|X^n) \\
&= H(Y^n) - \sum_{i=1}^{n} H(Y_i|Y_1, \ldots, Y_{i-1}, X^n) \\
&= H(Y^n) - \sum_{i=1}^{n} H(Y_i|X_i) \\
&\leq \sum_{i=1}^{n} H(Y_i) - H(Y_i|X_i) \\
&= \sum_{i=1}^{n} I(X_i; Y_i) \\
&\leq nC.
\end{aligned}$$

□

Finally, we can prove the converse: that any sequence of $(2^{nR}, n)$ codes with $\lambda^{(n)} \to 0$ must have $R \leq C$.

Proof. Note that $\lambda^{(n)} \to 0$ implies that $P_e^{(n)} \to 0$. For a fixed encoding rule $X^n(\cdot)$ and a fixed decoding rule $\widehat{W} = g(Y^n)$, we have $W \to X^n(W) \to Y^n \to \widehat{W}$. Let $W \sim \text{Unif}([2^{nR}])$. We have

$$\begin{aligned}
nR &= H(W) \\
&= H(W|\widehat{W}) + I(W; \widehat{W}) \\
&\leq 1 + P_e^{(n)} nR + I(W; \widehat{W}) \\
&\leq 1 + P_e^{(n)} nR + I(X^n; Y^n) \\
&\leq 1 + P_e^{(n)} nR + nC.
\end{aligned}$$

Dividing by n, we obtain

$$R \leq P_e^{(n)} R + \frac{1}{n} + C$$

and the first two terms ten to 0 which implies that $R \leq C$.

Furthermore, we can rewrite

$$P_e^{(n)} \geq 1 - \frac{C}{R} - \frac{1}{nR}$$

which shows that if $R > C$, the probability of error is bounded away from 0 for sufficiently large n. Hence, we cannot acheive an arbitrarily low probability of error at rates above capacity. □

Remark 5.19. This above result is called the weak converse to the channel coding theorem. It is possible to also prove a strong converse, which states that for rates above capacity, the probability of error goes exponentially to 1.

5.6 Feedback Capacity

Definition 5.20 (Feedback code)**.** We define a $(2^{nR}, n)$ feedback code as a sequence of mappings $x_i(W, Y^{i-1})$ where each x_i is a function only of the message $W \in 2^{nR}$ and previous received values, $Y_1, \ldots, Y_{i-1}$, and a sequence of decoding functions $g : \mathcal{Y}^n \to [2^{nR}]$.

Definition 5.21 (Capacity with feedback)**.** The capacity with feedback C_{FB} of a DMC is the supremum of all rates acheivable by feedback codes.

Theorem 5.22 (Feedback capacity)

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

Proof. It is clear that $C_{FB} \geq C$. Let $W \sim \text{Unif}([2^{nR}])$. Then $\Pr(W \neq \widehat{W}) = P_e^{(n)}$ and

$$\begin{aligned} nR &= H(W) \\ &= H(W|\widehat{W}) + I(W;\widehat{W}) \\ &\leq 1 + P_e^{(n)} nR + I(W;\widehat{W}) \\ &\leq 1 + P_e^{(n)} nR + I(W;Y^n), \end{aligned}$$

by Fano and the data-processing inequalities. Now, note that

$$\begin{aligned} I(W;Y^n) &= H(Y^n) - H(Y^n|W) \\ &= H(Y^n) - \sum_{i=1}^{n} H(Y_i|Y_1, \ldots, Y_{i-1}, W) \\ &= H(Y^n) - \sum_{i=1}^{n} H(Y_i|Y_1, \ldots, Y_{i-1}, W, X_i) \\ &= H(Y^n) - \sum_{i=1}^{n} H(Y_i|X_i) \end{aligned}$$

since X_i is a function of $Y_1, \ldots, Y_{i-1}, W$, and conditioned on X_i, Y_i, is independent of W and past samples of Y. It follows that

$$\begin{aligned} I(W; Y^n) &= H(Y^n) - \sum_{i=1}^{n} H(Y_i|X_i) \\ &\leq \sum_{i=1}^{n} H(Y_i) - H(Y_i|X_i) \\ &= \sum_{i=1}^{n} I(X_i; Y_i) \\ &\leq nC. \end{aligned}$$

Putting this together we obtain the same bound as before allowing us to conclude that $R \leq C$. □

5.7 Source-Channel Separation Theorem

Consider a source V that generates symbols from an alphabet $\mathcal{V}$. We only make the assumption that it is from a finite alphabet and satisfies AEP. We wish to send the sequence of symbols $V^n = V_1, \ldots, V_n$ over the channel so that the receiver can reconstruct the sequence. To do this, we map the sequence onto a codeword $X^n(V^n)$ and send the codeword over the channel. The receiver looks at the received sequence and makes an estimate $\widehat{V}^n$.

Theorem 5.23 (Source-channel coding theorem)

If $V_1, V_2, \ldots V_n$ is a finite alphabet stochastic process that satisfies the AEP and $H(\mathcal{V}) < C$, there exists a source-channel code with error probability $\Pr(\widehat{V}^n \neq V^n) \to 0$. Conversely, for any statioanry stochastic process, if $H(\mathcal{V}) > C$, the probability of error is bounded away from zero, and it is not possible to send the process over the channel with arbitrarily low probability of error.

Proof. Since we assumed the process satisfies the AEP, there exists a typical set $A_\epsilon^{(n)}$ of size at most $2^{n(H(\mathcal{V})+\epsilon)}$ which contains most of the probability. We only encode sequences belonging to this set - the other erroring sequences contribute at most ϵ to the error probability.

Index all the sequences belonging to $A_\epsilon^{(n)}$ using at most $n(H+\epsilon)$ bits. We can transmit the desired index with at most ϵ error probability if $H(\mathcal{V}) + \epsilon = R < C$.

Now,

$$\Pr(V^n \neq \widehat{V}^n) \leq P(V^n \notin A_e^{(n)}) + P(g(Y^n) \neq V^n | V^n \in A_\epsilon^{(n)}) < 2\epsilon$$

for sufficiently large n. Hence, we can reconstruct the error probability for n if $H(\mathcal{V}) < C$.

Conversely, we wish to show that $\Pr(\widehat{V}^n \neq V^n) \to 0$ implies that $H(\mathcal{V}) \leq C$ for any sequence of source-channel codes $X^n(V^n) : \mathcal{V}^n \to \mathcal{X}^n, g_n(Y^n) : \mathcal{Y}^n \to \mathcal{V}^n$.

By Fano's inequality,

$$H(V^n|\widehat{V}^n) \leq 1 + \Pr(\widehat{V}^n \neq V^n) \log |\mathcal{V}^n| = 1 + \Pr(\widehat{V}^n \neq V^n) n \log |\mathcal{V}|.$$

Therefore,

$$\begin{aligned}
H(\mathcal{V}) &\leq \frac{H(V_1,\ldots,V_n)}{n} \\
&= H(V^n)/n \\
&= \frac{1}{n}H(V^n|\widehat{V}^n) + \frac{1}{n}I(V^n;\widehat{V}^n) \\
&\leq \frac{1}{n}(1+\Pr(\widehat{V}^n \neq V^n)n\log|\mathcal{V}|) + \frac{1}{n}I(V^n:\widehat{V}^n) \\
&\leq \frac{1}{n}(1+\Pr(\widehat{V}^n \neq V^n)n\log|\mathcal{V}|) + \frac{1}{n}I(X^n:Y^n) \\
&\leq \frac{1}{n} + \Pr(\widehat{V}^n \neq V^n)\log|\mathcal{V}| + C
\end{aligned}$$

Then, $\Pr(\widehat{V}^n \neq V^n) \to 0$ as $n \to \infty$ which gives the desired result. □

5.8 Solutions to selected problems

Exercise 5.24 (7.3, *Channels with memory have higher capacity*)**.** Consider a binary symmetric channel $Y_i = X_i \oplus Z_i$ where $\oplus$ is addition mod 2 and $X_i, Y_i \in \{0,1\}$. Suppose that $Z_i \sim \text{Ber}(p)$, but they are not necessarily independent. Assume Z^n is independent of X^n. Let $C = 1 - H(p)$. Show that

$$\max_{p(x_1,\ldots,x_n)} I(X_1,\ldots,X_n;Y_1,\ldots,Y_n) \geq nC.$$

Proof. Choose $X_i \overset{\text{i.i.d.}}{\sim} \text{Ber}(1/2)$. Note that

$$\begin{aligned}
I(X_1,\ldots,X_n;Y_1,\ldots,Y_n) &= H(X_1,\ldots,X_n) - H(X_1,\ldots,X_n|Y_1,\ldots,Y_n) \\
&= H(X_1,\ldots,X_n) - H(Z_1,\ldots,Z_n|Y_1,\ldots,Y_n) \\
&\geq H(X_1,\ldots,X_n) - H(Z_1,\ldots,Z_n) \\
&\geq H(X_1,\ldots,X_n) - nH(Z_1) \\
&= n - nH(p) \\
&= nC.
\end{aligned}$$

□

Remark 5.25. The intuitive explanation according to the official solutions is that the correlation between the noise decreases the effective noise.

Exercise 5.26 (7.5, **Using two channels at once**)**.** Consider two DMCs $(\mathcal{X}_i, p(y_i \mid x_i), \mathcal{Y}_i)$ with capacities C_i. Consider the channel given by taking the product of the two channels: $\mathcal{X}_1 \times \mathcal{X}_2 \to \mathcal{Y}_1 \times \mathcal{Y}_2$. Find the capcity of this channel.

Proof. Note that

$$\begin{aligned}
I(X_1,X_2;Y_1,Y_2) &= H(Y_1,Y_2) - H(Y_1,Y_2|X_1,X_2) \\
&= H(Y_1,Y_2) - H(Y_1|X_1) - H(Y_2|X_2) \\
&\leq H(Y_1) + H(Y_2) - H(Y_1|X_1)H(Y_2|X_2) \\
&= I(X_1;Y_1) + I(X_2;Y_2)
\end{aligned}$$

with equality if Y_1, Y_2 are independent, which happens iff X_1, X_2 are independent. Taking the distribution $p^*(x_1, x_2) = p^*(x_1)p^*(x_2)$ as the maximizing distributions for the individual capacities, we obtain

$$\max_{p(x_1,x_2)} I(X_1, X_2; Y_1; Y_2) = \max_{p(x_1)} I(X_1; Y_1) + \max_{p(x_2)} I(X_2, Y_2) = C_1 + C_2.$$

□

Exercise 5.27 (7.7, Cascade of binary symmetric channels)**.** SHow that a cascade of n identical independent binary symmetric channels, each with raw error probability p is equivalent to a single BSC with error probability $1/2(1 - (1 - 2p)^n)$. No encoding or decoding takes places at the intermediate terminals.

Proof. It is clear that we essentially have a product of transition matrices P^n, which is easy to compute using the eigendecomposition of P (it is a real symmetric matrix). □

6 Differential Entropy

Definition 6.1 (Differential entropy)**.** The differential entropy $h(X)$ of a continuous random variable X with density $f(x)$ is defined as

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

where $S = \{x : f(x) > 0\}$ is the support set of f.

6.1 AEP for continuous random variables

Theorem 6.2 (AEP)

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} f(x)$. Then,

$$-\frac{1}{n} \log f(X_1, \ldots, X_n) \to E[-\log f(x)] = h(X)$$

in probability.

This leads to the definition of a typical set.

Definition 6.3 (Typical set)**.** For $\epsilon > 0$ and any n, we define the typical set $A_\epsilon^{(n)}$ with respect to $f(x)$ as follows:

$$A_\epsilon^{(n)} = \left\{ (x_1, \ldots, x_n) \in S^n : \left| -\frac{1}{n} \log f(x_1, \ldots, x_n) - h(X) \right| \leq \epsilon \right\}$$

The analog of cardinality in the discrete case is volume in the continuous case.

Definition 6.4 (Volume)**.** The volume $\text{Vol}(A)$ of a set $A \in \mathbb{R}^n$ is defined as $\text{Vol}(A) = \int_A dx_1 \ldots dx_n$.

The typical set has the following properties:

1. $\Pr(A_\epsilon^{(n)}) > 1 - \epsilon$ for n sufficiently large.
2. $\text{Vol}(A_\epsilon^{(n)}) \leq 2^{n(h(X)+\epsilon)}$ for all n.
3. $\text{Vol}(A_\epsilon^{(n)}) \geq (1-\epsilon) 2^{n(h(X)-\epsilon)}$ for n sufficiently large.

6.2 Relation between continuous and discrete entropy

Consider the quantized random variable X^Δ defined by

$$X^\Delta = x_i, \quad i\Delta \leq X \leq (i+1)\Delta,$$

where x_i satisfies

$$f(x_i)\Delta = \int_{i\Delta}^{(i+1)\Delta} f(x)\, dx$$

(such a value exists by the mean vlaue theorem). Using the definition of Riemann integrability, we can show the following result relating discrete and differential entropy.

Theorem 6.5

If the density $f(x)$ of the random variable X is Riemann integrable, then

$$H(X^{\Delta}) + \log \Delta \xrightarrow{\Delta \to 0} h(f) = h(X).$$

Thus, the entropy of an n-bit quantization of a continuous random variable X is approximately $h(X) + n$.

We also have several definitions analogous to the discrete and single variable cases.

Definition 6.6 (Differential entropy of a set)**.** The differential entropy of a set $X_1, \ldots, X_n \sim f(x_1, \ldots, x_n)$ is defined as $h(X_1, \ldots, X_n) = -\int f(x^n) \log f(x^n)\, dx^n$

Definition 6.7 (Conditional differential entropy)**.** If $X, Y \sim f(x, y)$, we can define the conditional differential entropy as

$$h(X \mid Y) = -\int f(x, y) \log f(x|y)\, dx\, dy$$

Theorem 6.8 (Entorpy of a multivariate normal)

Let $X_1, \ldots, X_n \sim \mathcal{N}(\mu, K)$. Then,

$$h(X_1, \ldots, X_n) = h(\mathcal{N}_n(\mu, K)) = \frac{1}{2} \log_2 (2\pi e)^n \det(K)$$

.

Proof. Noting the probability density

$$f(x) = \frac{1}{\sqrt{c} \exp(-\frac{1}{2}(x-\mu)^{\intercal} K^{-1} (x-\mu))},$$

we have

$$\begin{aligned}
h(f) &= \frac{1}{2} E\left[\sum_{i,j} (X_i - \mu_i)(K^{-1})_{ij}(X_j - \mu_j)\right] + \frac{1}{2} \ln(2\pi)^n \det(K) \\
&= \frac{1}{2} \sum_{i,j} E[(X_j - \mu_j)(X_i - \mu_i)](K^{-1})_{ij} + \frac{1}{2} \ln(2\pi)^n \det(K) \\
&= \frac{1}{2} \sum_{j} (K K^{-1})_{ij} + \frac{1}{2} \ln(2\pi)^n \det(K) \\
&= \frac{n}{2} + \frac{1}{2} \ln(2\pi)^n \det(K) \\
&= \frac{1}{2} \ln(2\pi e)^n \det(K) \\
&= \frac{1}{2} \log(2\pi e)^n \det(K)
\end{aligned}$$

□

6.3 Relative Entropy and Mutual Information

Definition 6.9 (Relative Entropy)**.** The relative entropy of KL-divergence $D(f\|g)$ between two densities is defined as

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

Definition 6.10 (Mutual Information)**.** The mutual information $I(X;Y)$ between two random variables with joint density $f(x,y)$ is defined as

$$I(X;Y) = \int f(x,y) \log \frac{f(x,y)}{f(x)f(y)}\, dx\, dy.$$

6.4 Properties of Differential Entropy

First, note the following results:

Theorem 6.11

$$D(f\|g) \geq 0$$

with equality iff $f = g$ almost everywhere.

As corollaries, we have

- $I(X;Y) \geq 0$ with equality iff X and Y are independent.
- $h(X \mid Y) \leq h(X)$ with equality iff X and Y are independent.

Theorem 6.12 (Chain rule for differential entropy)

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

As a corollary we have the usual indepedence inequality:

$$h(X_1, \ldots, X_n) \leq \sum h(X_i)$$

with equality if and only if $X_1, \ldots, X_n$ are independent. An interesting application is Hadamard's inequality:

Theorem 6.13 (Hadamard's Inequality)
Let $X \sim \mathcal{N}(0, K)$. We have $\det(K) \leq \prod_{i=1}^{n} K_i i$

Proof. Compute both sides of the independence inequality. □

Some interesting scaling properties of differential entropy are as follows:

- $h(X + c) = h(X)$
- $h(aX) = h(X) + \log |a|$
- $h(AX) = h(X) + \log |\det(A)|$

Theorem 6.14

Let the random vector $X \in \mathbb{R}^n$ have zero mean and covariance $K = EXX^\intercal$. Then

$$h(X) \leq \frac{1}{2} \log(2\pi e)^n \det(K)$$

with equality if and only if $X \sim \mathcal{N}(0, K)$.

Proof. Let $g(x)$ be a density satisfying $\int g(x) x_i x_j \, dx = K_{ij}$ for all i, j. Let φ_K be the density of a $\mathcal{N}(0, K)$ vector, where we set $\mu = 0$. Note that $\log \varphi_K(x)$ is a quandratic form with $\int x_i x_j \varphi_K(x) \, dx = K_{ij}$. Then, we have

$$\begin{aligned} 0 \leq D(g \| \varphi_K) & \\ &= \int g \log \frac{g}{\varphi_K} \\ &= -h(g) - \int g \log \varphi_K \qquad\qquad = -h(g) - \int \varphi_K \log \varphi_K \\ &= -h(g) + h(\varphi_K) \end{aligned}$$

where $\int g \log \varphi_K = \int \varphi_K \log \varphi_K$ since they both yield the same moments of the quadriatic form $\log \varphi_K(x)$. □

Using the fact that the Gaussian distribution maximizes entropy over all distributions with the same variance, we can to estimation that is analogous to Fano's inequality.

Theorem 6.15 (Estimation error)

For any random variable X and estimator $\widehat{X}$,

$$E(X - \widehat{X})^2 \geq \frac{1}{2\pi e} \exp(2h(X)).$$

6.5 Solutions to selected problems

Exercise 6.16 (8.1, **Differential entropy**)**.** Evaluate the differential entropy for the following:

1. The exponential density, $f(x) = \lambda e^{-\lambda x}$, $x \geq 0$,
2. The Laplace density, $f(x) = \frac{1}{2}\lambda e^{-\lambda|x|}$
3. The sum of X_1 and X_2 where X_1, X_2 are independent normal random variables with means μ_i and variances σ_i^2

1. Note that

$$\begin{aligned} h(f) &= -\int_{x\geq 0} f(x)\ln f(x) \\ &= -\int_{x\geq 0} \lambda e^{-\lambda x}\ln(\lambda e^{-\lambda x})\,dx \\ &= -\ln\lambda + 1 \\ &= \ln(e/\lambda). \end{aligned}$$

2. Note that we can write

$$h(f_{\text{Laplace}}) = h(f_{\text{Exponential}}) + \ln 2 = \ln(2e/\lambda).$$

3. It follows immediately from the fact that $X_1 + X_2 \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

Exercise 6.17 (8.2, **Ky Fan Inequality**)**.** Let K_1, K_2 be two symmetric nonnegative definite $n \times n$ matrices. Prove the Ky Fan inequality:

$$|\lambda K_1 + \overline{\lambda} K_2| \geq |K_1|^{\lambda}|K_2|^{\overline{\lambda}}$$

where $\lambda \in [0,1]$, $\overline{\lambda} = 1 - \lambda$.

Proof. Following the hint, we let $Z = X_\theta$ where $X_1 \sim \mathcal{N}(0, K_1), X_2 \sim \mathcal{N}(0, K_2)$ and $\theta = \text{Ber}(\lambda)$. By the fact that conditioning decreases entropy, we have $h(Z \mid \theta) \leq h(Z)$. Now, note that

$$h(Z \mid \theta) = \frac{\lambda}{2}\log(2\pi e)^n|K_1| + \frac{\overline{\lambda}}{2}\log(2\pi e)^n|K_2|.$$

Now, note that

$$E(ZZ^\intercal) = E(E(ZZ^\intercal|\theta)) = \lambda K_1 + \overline{\lambda}K_2 = K_Z.$$

It follows that

$$h(Z) \leq \max_{EZZ^\intercal = K_Z} h(Z) = \frac{1}{2}\log(2\pi e)^n|\lambda K_1 + \overline{\lambda}K_2|.$$

Simplifying the logarithms completes the proof. □