Document: EEC 266 (Spring 2011)
Professor: Zhao
Latest Update: April 2, 2012
Author: Jeff Irion
http://www.math.ucdavis.edu/~jlirion

Contents

0 Important

0.1 Key Formulas

- Entropy:

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

- Entropy Change of Base Formula:

$$H_b(X) = \log_b a H_a(X)$$

- Joint Entropy:

$$\begin{aligned} H(X, Y) &= \sum_x \sum_y p(x, y) \log \frac{1}{p(x, y)} \\ &= H(X) + H(Y|X) = H(Y) + H(X|Y) \end{aligned}$$

- Conditional Entropy:

$$\begin{aligned} H(Y|X) &= \sum_x p(x) \sum_y p(y|x) \log \frac{1}{p(y|x)} \\ &= H(X, Y) - H(X) \end{aligned}$$

- Relative Entropy:

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

  - $D(p||q) \geq 0$, with equality iff $p = q$

- Mutual Information:

$$\begin{aligned} I(X; Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) = I(Y; X) \end{aligned}$$

- Conditional Mutual Information:

$$\begin{aligned} I(X; Y|Z) &= H(X|Z) - H(X|Y, Z) \\ &= H(Y|Z) - H(Y|X, Z) \end{aligned}$$

- Chain Rules
  - Entropy:

$$H(X_1, \ldots, X_n) = H(X_1) + H(X_2|X_1) + \ldots + H(X_n|X_1, \ldots, X_n)$$

- Information:

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

- Information Can't Hurt:

$$H(X) \geq H(X|Y)$$

  - Corollary - Independence Bound on Entropy:

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

- Bound on Entropy:
  - $H(X) \leq \log |\mathcal{X}| \quad \Leftrightarrow \quad$ for a fixed alphabet size, the uniform distribution has the largest entropy.
- Weak Law of Large Numbers:

$$\frac{1}{n} \sum_{i=1}^{n} X_i \to \mathbb{E}[X]$$

- Entropy Rate:

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

- Kraft Inequality

$$\sum D^{-l_i} \leq 1$$

- Channel Capacity:

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

  - Capacity of a Weakly Symmetric Channel:

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

- Differential Entropy:

$$h(X) = \int_S f(x) \log \frac{1}{f(x)} \, dx$$

- Uniform Distribution: $x \sim \mu(0, a) \quad \Rightarrow \quad h(X) = \log a \quad$ (See Example 8.2)
- Normal (Gaussian) Distribution: $x \sim \mathcal{N}(\mu, \sigma^2) \quad \Rightarrow \quad h(X) = \frac{1}{2} \log 2\pi e \sigma^2 \quad$ (See Example 8.3)

- Capacity of a Gaussian Channel:

$$C = \frac{1}{2} \log \left(1 + \frac{P}{N}\right)$$

where P is the power constraint and N is the noise variance.

1 Introduction and Preview

Remark 1.1. ***2 Main Questions of Information Theory***
page 1 and Notes 3/28/11

1. What is the ultimate data compression? (Answer: the entropy H)
2. What is the ultimate transmission rate of communication? (Answer: the channel capacity C)

Remark 1.2. ***3 Main Concepts***
Notes 3/28/11

1. Entropy
2. Relative Entropy
3. Mutual Information

Remark 1.3.
Notes 3/28/11

How do we measure information?

- Reduction of uncertainty
  - Flip a coin, heads shows up
  - Roll a die, it is an even number

How do we measure uncertainty?

Remark 1.4. ***Notation***
Notes 3/28/11

Rather than writing $p_X(x)$ and $p_Y(y)$, the terms $p(x)$ and $p(y)$ shall be used.

Unless otherwise stated, logs are base 2. (Recall: $\log_b(x) = \frac{\log_a(x)}{\log_a(b)}$)

Capital letters denote variables, lowercase letters denote realizations.

The units of entropy are bits.

2 Entropy, Relative Entropy, and Mutual Information

2.1 Entropy

Definition 2.1. ***Entropy***
page 13 and Notes 3/28/11

Entropy is a measure of the uncertainty of a random variable. Let X be a discrete random variable with alphabet $\mathcal{X}$ and *probability mass function* $p(x)$. The entropy is defined as

$$H(X) = -\sum_{x\in\mathcal{X}} p(x)\log_2 p(x) = \mathbb{E}_p \log\frac{1}{p(x)} = -\mathbb{E}_p \log p(x)$$

where $\mathbb{E}(g(x)) = \sum_x p(x)g(x)$. If the base of the entropy is $b \neq 2$, then we write $H_b(X)$.

Remark 2.2.
pages 14 & 15 and Notes 3/28/11

1. We use the convention that $0\log 0 \equiv 0$. (Note: $\lim_{\epsilon\to 0}\epsilon\log\epsilon = 0$.) This means that adding any terms of zero probability does not change the entropy.
2. Entropy is a function of the distribution of X. It does not depend on the values taken by X.
3. $H(X) \geq 0$
4. $H_b(X) = \log_b a \; H_a(X)$

Example 2.3.
page 15 and Notes 3/28/11

Let

$$X = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p \end{cases}$$

Then

$$H(X) = -p\log p - (1-p)\log(1-p) \equiv H(p)$$

In particular, when $p = \frac{1}{2}$ then $H(X) = 1$ bit.

Example 2.4.
page 15 and Notes 3/28/11

Let

$$X = \begin{cases} a & \text{with probability } \frac{1}{2} \\ b & \text{with probability } \frac{1}{4} \\ c & \text{with probability } \frac{1}{8} \\ d & \text{with probability } \frac{1}{8} \end{cases}$$

Then

$$H(X) = \frac{7}{4} \text{ bits}$$

$\frac{7}{4}$ is the minimum expected number of binary questions required to determine the value of X. This scheme could be stored as

$$a \leftrightarrow 0 \qquad b \leftrightarrow 10 \qquad c \leftrightarrow 110 \qquad d \leftrightarrow 111$$

Note that $-\log p(x)$ is approximately the number of bits we want to assign to x.

2.2 Joint Entropy and Conditional Entropy

Definition 2.5. ***Joint Entropy***
page 16 and Notes 3/28/11

The *joint entropy* $H(X, Y)$ of a pair of discrete random variables (X, Y) with a joint distribution $p(x, y)$ is defined as

$$H(X, Y) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x, y) = -\mathbb{E}_p \log \frac{1}{p(x, y)}$$

Definition 2.6. ***Conditional Entropy***
page 17 and Notes 3/28/11

If $(X, Y) \sim p(x, y)$, the *conditional entropy* $H(Y|X)$ is defined as

$$\begin{aligned} H(Y|X) &= \sum_{x \in \mathcal{X}} p(x) H(Y|X = x) \\ &= -\sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y|x) \log p(y|x) \\ &= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(y|x) \\ &= -E_{p(x,y)} \log p(Y|X) \end{aligned}$$

Theorem 2.7. ***Chain Rule***
page 17 and Notes 3/28/11

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

Remark 2.8.
page 18 and Notes 3/28/11

$$H(X|Y) \neq H(Y|X)$$
$$H(X) - H(X|Y) = H(Y) - H(Y|X)$$

The second line says that the reduction in the uncertainty (achieved via correlation) is the same.

2.3 Relative Entropy and Mutual Information

Definition 2.9. ***Relative Entropy***
page 19 and Notes 3/28/11

Relative entropy is a measure of the distance between two distributions. Specifically, the relative entropy $D(p||q)$ is a measure of the inefficiency of assuming that the distribution is q when the true distribution is p. It is also known as the *Kullback-Leibler distance/divergence*. It is given by

$$D(p||q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} = E_p \log \frac{p(X)}{q(X)}$$

Remark 2.10.
Notes 3/28/11

The number of bits is on the order of $\sum_{x \in \mathcal{X}} p(x) \log \frac{1}{q(x)}$ based on the incorrect coding scheme q.

$$\sum_{x \in \mathcal{X}} p(x) \log \frac{1}{q(x)} = \sum_{x \in \mathcal{X}} p(x) \log \frac{1}{p(x)} + D(p||q)$$

Remark 2.11.
page 19 and Notes 3/28/11

1. $p \log \frac{p}{0} = \infty$. If there is any x such that $p(x) > 0$ but $q(x) = 0$ then $D(p||q) = \infty$.

   Next class we will show:

2. $D(p||q) \geq 0$ with equality iff $p = q$.
3. Relative entropy is not a true distance function between distributions because $D(p||q) \neq D(q||p)$, and it also doesn't satisfy the triangle inequality.

Definition 2.12. ***Conditional Relative Entropy***
Notes 3/28/11

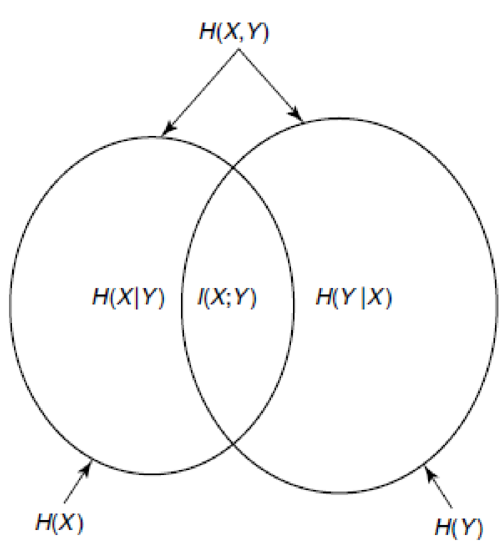
Given $p(x,y)$ and $q(x,y)$, the *conditional relative entropy* $D\big(p(y|x)||q(y|x)\big)$ is the average entropy between $p(y|x)$ and $q(y|x)$ averaged over $p(x)$.

$$D\big(p(y|x)||q(y|x)\big) = \sum_x p(x) \sum_y p(y|x) \log \frac{p(y|x)}{q(y|x)} = \sum_x \sum_y p(x,y) \log \frac{p(y|x)}{q(y|x)}$$

Definition 2.13. ***Mutual Information***
page 19 and Notes 3/28/11

Consider 2 random variables X and Y with a joint probability mass function $p(x,y)$ and marginal probability mass functions $p(x)$ and $p(y)$. The *mutual information* $I(X,Y)$ is the relative entropy between the joint distribution $p(x,y)$ and the product distribution $p(x)p(y)$.

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



2.4 Relationship Between Entropy and Mutual Information

Remark 2.14.
page 21 and Notes 3/28/11

We can prove that:

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

This last identity is why entropy is sometimes called *self-information*.

2.5 Chain Rules for Entropy, Relative Entropy, and Mutual Information

Theorem 2.15. ***Chain Rule for Entropy***
page 22 and Notes 3/30/11

Given: $X_1, \ldots, X_n \sim p(x_1), \ldots, p(x_n)$
Then:

$$\begin{aligned} H(X_1, \ldots, X_n) &= H(X_1) + H(X_2|X_1) + H(X_3|X_1, X_2) + \ldots + H(X_n|X_1, \ldots, X_n) \\ &= \sum_{i=1}^{n} H(X_i|X_1, \ldots, X_{i-1}) \end{aligned}$$

Definition 2.16. ***Conditional Mutual Information***
page 23

The *conditional mutual information* of random variables X and Y given Z is

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

Theorem 2.17. ***Chain Rule for Information***
page 24 and Notes 3/30/11

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

Proof.

$$
\begin{aligned}
I(X_1,\dots,X_n;Y) &= H(X_1,\dots,X_n) - H(X_1,\dots,X_n|Y) \\
&= \sum_{i=1}^{n} H(X_i|X_1,\dots,X_{i-1}) - \sum_{i=1}^{n} H(X_i|X_1,\dots,X_{i-1},Y) \\
&= \sum_{i=1}^{n} I(X_i;Y|X_1,\dots,X_{i-1}
\end{aligned}
$$

□

Theorem 2.18. ***Chain Rule for Relative Entropy***
page 24 and Notes 3/30/11

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

2.6 Jensen's Inequality and Consequences

Definition 2.19. ***Convex, Concave***
page 25 and Notes 3/30/11

A function f is *convex* if

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

i.e. the function lies below every chord. If the inequality is strict then it is *strictly convex*. A function g is *concave* if $-g$ is convex.

Theorem 2.20. ***Jensen's Inequality***
page 27 and Notes 3/30/11

If f is convex, then

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$$

If f is strictly convex then X is a constant, i.e. $X = \mathbb{E}[X]$.

If f is concave, then

$$\mathbb{E}[f(X)] \leq f(\mathbb{E}[X])$$

Theorem 2.21. ***Information Inequality***
page 28 and Notes 3/30/11

$D(p||q) \geq 0$, with equality iff $p = q$.

Proof.

$$\begin{aligned}
-D(p||q) &= -\sum_x p(x) \log \frac{p(x)}{q(x)} \\
&= \sum_x \log \frac{q(x)}{p(x)} \\
&\le \log \sum_x p(x) \frac{q(x)}{p(x)} \qquad (2.1) \\
&\le \log 1 \le 0
\end{aligned}$$

where (2.1) follows from Jensen's Inequality (Theorem 2.20), since log is concave. □

Corollary 2.22. ***Nonnegativity of Mutual Information***
page 28 and Notes 3/30/11

$I(X;Y) \ge 0$, with equality iff X and Y are *independent* $\Rightarrow p(x,y) = p(x)p(y)$.

Theorem 2.23. ***Conditioning Reduces Entropy $\Leftrightarrow$ Information Can't Hurt***
page 29 and Notes 3/30/11

$$H(X|Y) \le H(X)$$

with equality iff X and Y are independent.

Proof. $0 \le I(X;Y) = H(X) - H(X|Y)$ □

Remark 2.24.
page 30 and Notes 3/30/11

$H(X|Y = y)$ may actually be bigger than $H(X)$. For example, consider

Y \ X	1	2
1	0	$\frac{3}{4}$
2	$\frac{1}{8}$	$\frac{1}{8}$

$$H(X) = H\left(\frac{1}{8}, \frac{1}{8}\right) = 0.544$$
$$H(X|Y=2) = 1$$
$$H(X|Y=1) = 0$$
$$H(X|Y) = \frac{3}{4}\cdot 0 + \frac{1}{4}\cdot 1 = \frac{1}{4} < H(X)$$

Theorem 2.25. ***Independence Bound on Entropy***
page 30 and Notes 3/30/11

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

Proof. By the chain rule for entropies (Theorem 2.15),

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

□

Remark 2.26.
Notes 3/30/11

For a fixed alphabet size, the uniform distribution has the largest entropy. Given X with a finite alphabet $\mathcal{X}$, then $H(X) \le \log|\mathcal{X}|$ and

$$0 \le D(p||u) = \sum_x p(x) \log \frac{p(x)}{\frac{1}{|\mathcal{X}|}} = \sum_x p(x) \log p(x) + \log|\mathcal{X}| = \log|\mathcal{X}| - H(X)$$

2.7 Log Sum Inequality and its Applications

Theorem 2.27. ***Log Sum Inequality***
page 31 and Notes 3/30/11

For nonnegative numbers $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$,

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

with equality if $a_i = cb_i$ for some constant c.

The proof of this uses Jensen's Inequality (Theorem 2.20).

Theorem 2.28. ***Convexity of Relative Entropy***
page 32 and Notes 3/30/11

$D(p||q)$ is convex in the pair (p, q). That is, if (p_1, q_1) and (p_2, q_2) are two pairs of probability mass functions, then

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

Proof. Applying the log sum inequality (Theorem 2.27) to the LHS of the above equation, we get

$$\big(\lambda p_1(x) + (1-\lambda)p_2(x)\big) \log \frac{\lambda p_1(x) + (1-\lambda)p_2(x)}{\lambda q_1(x) + (1-\lambda)q_2(x)} \leq \lambda p_1(x) \log \frac{\lambda p_1(x)}{\lambda q_1(x)} + (1-\lambda)p_2(x) \log \frac{(1-\lambda)p_2(x)}{(1-\lambda)q_2(x)}$$

Summing over all x, we get the desired result. □

Theorem 2.29. ***Concavity of Entropy***
page 32 and Notes 4/4/11

$H(p)$ is a concave function of p.

Proof.

$$H(p) = \log |\mathcal{X}| - D(p||u)$$

This is because

$$\begin{aligned} D(p||u) &= \sum_x p(x) \log \frac{p(x)}{u(x)} = \sum_x p(x) \log |\mathcal{X}| + \sum_x p(x) \log p(x) \\ &= \log |\mathcal{X}| - H(X) \end{aligned}$$

$D(p||u)$ is convex in p, so the negative makes $H(p)$ concave. □

Example 2.30.
Notes 4/4/11

Let $p_1 = \{\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}\}$ and $p_2 = \{\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\}$.
Then $H(p_1) = \frac{7}{4}$ and $H(p_2) = 2$
If we take $\lambda = \frac{1}{4}$, then

$$H(\lambda p_1 + (1-\lambda)p_2) \geq \lambda H(p_1) + (1-\lambda)H(p_2)$$

2.8 Data-Processing Inequality

Definition 2.31. ***Markov Chain***
page 34 and Notes 4/4/11

Random variables X, Y, Z are said to form a *Markov chain*, denoted $X \to Y \to Z$, if

$$p(x, y, z) = p(x)p(y|x)p(z|y)$$

Remark 2.32.
page 34 and Notes 4/4/11

1. $X \to Y \to Z$ iff X and Z are conditionally independent given Y
2. If $X \to Y \to Z$ then $Z \to Y \to X$
3. If $Z = f(Y)$, then $X \to Y \to Z$
4. If $X \to Y \to Z$, then $I(X;Z|Y) = 0$

Theorem 2.33. ***Data Processing Inequality***
page 34 and Notes 4/4/11

If $X \to Y \to Z$, then $I(X;Y) \geq I(X;Z)$

Proof. By the chain rule,

$$\begin{aligned} I(X;Y|Z) &= I(X;Z) + \underbrace{I(X;Y|Z)}_{\geq 0} \\ &= I(X;Y) + \underbrace{I(X;Z|Y)}_{=0} \end{aligned}$$

where $I(X;Z|Y) = 0$ because X and Z are conditionally independent given Y. Since $I(X;Y|Z) \geq 0$, we have

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

with equality iff $I(X;Y|Z) = 0$, i.e. $X \to Z \to Y$ forms a Markov chain. □

Corollary 2.34.
page 35 and Notes 4/4/11

If $Z = f(Y)$ then $I(X;Y) \geq I(X;f(Y))$

Remark 2.35.
page 35 and Notes 4/4/11

It is possible that $I(X;Y|Z) > I(X;Y)$ when X, Y, Z do not form a Markov chain. For example, let X and Y be independent binary random variables and set $Z = X + Y$. Then $I(X;Y) = 0$ and

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z) = H(X|Z) = P(Z=1)H(X|Z=1) = \frac{1}{2} \text{ bit}$$

2.9 Sufficient Statistics

2.10 Fano's Inequality

Theorem 2.36. ***Fano's Inequality***
page 38 and Notes 4/4/11

Suppose that we want to estimate the value of a random variable X using a correlated random variable Y. Let $\hat{X} = f(Y)$. We define the *probability error*

$$P_e = \Pr[\hat{X} \neq X]$$

Fano's Inequality tells us that for any estimator $\hat{X}$ such that $X \to Y \to \hat{X}$, with $P_e = \Pr[\hat{X} \neq X]$, we have

$$\begin{aligned} H(P_e) + P_e \log|\mathcal{X}| &\geq H(X|Y) \quad \text{if } \hat{\mathcal{X}} \neq \mathcal{X} \\ H(P_e) + P_e \log(|\mathcal{X}| - 1) &\geq H(X|Y) \quad \text{if } \hat{\mathcal{X}} = \mathcal{X} \end{aligned}$$

and thus

$$P_e \geq \frac{H(X|Y) - 1}{\underbrace{\log|\mathcal{X}|}_{\text{or } \log(|\mathcal{X}|-1)}}$$

Proof. Let

$$E = \begin{cases} 1 & \text{if } \hat{X} \neq X \\ 0 & \text{if } \hat{X} = X \end{cases}$$

Then $\Pr[E = 1] = P_e$ and

$$\begin{aligned} H(E, X|\hat{X}) &= H(X|\hat{X}) + \underbrace{H(E|X, \hat{X})}_{=0} \\ &= \underbrace{H(E|\hat{X})}_{\leq H(P_e)} + \underbrace{H(X|E, \hat{X})}_{\leq P_e \log|\mathcal{X}|} \end{aligned}$$

We can show that

$$H(X|\hat{X}) \leq H(P_e) + P_e \log|\mathcal{X}|$$

and it follows from the data-processing inequality that

$$H(X|\hat{X}) \geq H(X|Y)$$

□

Remark 2.37.
Notes 4/4/11

Fano's Inequality is sharp, as seen in these 2 cases:

1. If $X = g(Y)$ then $H(X|Y) = 0$ and $P_e = 0$ because $\hat{X} = g(Y)$
2. No observation (no knowledge of Y)
   $X \in \{1, \ldots, m\},\ p_1 \geq p_2 \geq \ldots \geq p_m$
   $\hat{X} = 1,\ P_e = 1 - p_1$, and equality in Fano's Inequality is achieved when the probabilities are $\left(p, \frac{1-p}{m-1}, \ldots, \frac{1-p}{m-1}\right)$
   This is found by setting $H(P_e) + P_e \log(m-1) = H(X)$

Remark 2.38. ***Review of Key Concepts***
Notes 4/6/11

$$\begin{aligned}
H(X) &= H(p) = -\mathbb{E}[\log p(X)] = \sum_x p(x) \log \frac{1}{p(x)} \\
D(p||q) &= \sum_x p(x) \log \frac{p(x)}{q(x)} \\
I(X;Y) &= D\big(p(x,y)||p(x)p(y)\big) = H(X) - H(X|Y) = H(Y) - H(Y|X)
\end{aligned}$$

Jensen's Inequality: If f is convex, then $\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$.
It follows that $D(p||q) \geq 0,\ I(X;Y) \geq 0,\ H(X|Y) \leq H(X),\ H(X) \leq \log|\mathcal{X}|,\ H(X_1, \ldots, X_n) \leq \sum\limits_i H(X_i)$.

Log-Sum Inequality:

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

$D(p||q)$ is convex, $H(p)$ is concave, $I(X;Y)$ is concave in $p(x)$ for fixed $p(y|x)$ and convex in $p(y|x)$ for fixed $p(x)$.

Data Processing Inequality:

$$\text{If } X \to Y \to Z, \text{ then } I(X;Y) \geq I(X;Z)$$

Fano's Inequality: For any estimator $\hat{X}$ such that $X \to Y \to \hat{X}$, we have

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

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

Lemma 2.39.
page 40 and Notes 4/6/11

Let X, X' be two independent random variables, $X \sim p,\ X' \sim p'$. Then

$$\left.\begin{array}{l}\Pr\left[X = X'\right] \geq 2^{-H(p)-D(p||p')} \\ \Pr\left[X = X'\right] \geq 2^{-H(p')-D(p'||p)}\end{array}\right\}\text{not necessarily the same value}$$

If X and X' are *independent identically distributed* random variables ($i.i.d.$), meaning that $p = p'$, then

$$\Pr\left[X = X'\right] \geq 2^{-H(p)}$$

Proof.

$$\begin{aligned} 2^{-H(p)-D(p||p')} &= 2^{\sum_x p(x)\log p(x) - \sum_x p(x)\log\frac{p(x)}{p'(x)}} \\ &= 2^{\sum_x p(x)\log p'(x)} \\ &= 2^{\mathbb{E}[\log p'(x)]} \\ &\leq \mathbb{E}_p\left[2^{\log p'(x)}\right] = \mathbb{E}_p[p'(x)] = \sum_x p(x)p'(x) = \Pr\left[X = X'\right] \end{aligned}$$

□

3 Asymptotic Equipartition Property

3.1 Asymptotic Equipartition Property Theorem

Theorem 3.1. ***Weak Law of Large Numbers***
Notes 4/6/11

If $X_1, X_2, \ldots$ are i.i.d. random variables drawn from p, then

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to \mathbb{E}_p[X] \text{ in probability}$$

($X_n \overset{\text{in prob}}{\longrightarrow} X$ means that $\Pr\left[|X_n - X| > \epsilon\right] \to 0$.)

Theorem 3.2. ***Asymptotic Equipartition Property (AEP) Theorem***
page 58 and Notes 4/6/11

If $X_1, \ldots, X_n$ are i.i.d. $\sim p(x)$, then

$$-\frac{1}{n}\log p(X_1, \ldots, X_n) \to H(X) \quad \text{in probability}$$

Proof. The LHS:

$$-\frac{1}{n}\sum_i \log p(X_i) \to -\mathbb{E}[\log p(X)] = H(X)$$

□

Definition 3.3. ***Typical Set***
page 59 and Notes 4/6/11

For any $\epsilon > 0$, the *typical set* $A_\epsilon^{(n)}$ with respect to $p(x)$ is the set of all sequences $(x_1, \ldots, x_n)$ satisfying

$$2^{-n[H(X)+\epsilon]} \le p(x_1, \ldots, x_n) \le 2^{-n[H(X)-\epsilon]}$$

Properties of $A_\epsilon^{(n)}$:

1. $\Pr\left[A_\epsilon^{(n)}\right] > 1 - \epsilon$ for n sufficiently large
2. $|A_\epsilon^{(n)}| \le 2^{n[H(X)+\epsilon]}$
3. $|A_\epsilon^{(n)}| \ge (1-\epsilon) \cdot 2^{n[H(X)-\epsilon]}$

Remark 3.4. ***Number of Typical Sequences***
Notes 4/6/11

The number of typical sequences $\approx \binom{n}{np} \sim 2^{nH(X)}$.

To see this, recall Stirling's formula: $n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$

$$M = \binom{n}{np} \sim \frac{\sqrt{2\pi n}\left(\frac{n}{e}\right)^n}{\sqrt{2\pi np}\left(\frac{np}{e}\right)^{np}\sqrt{2\pi nq}\left(\frac{nq}{e}\right)^{nq}} = \frac{1}{\sqrt{2\pi npq}p^{np}q^{nq}}$$

$$\log M \sim -\frac{1}{2}\log(2\pi npq) - np\log p - nq\log q$$

$$\sim n\left[H(X) - \frac{\frac{1}{2}\log(2\pi npq)}{n}\right]$$

3.2 Consequences of the AEP: Data Compression

Remark 3.5. ***Code Word Length***
Notes 4/6/11

For sequences in $A_\epsilon^{(n)}$, the code word length is $n(H+\epsilon)+2$ bits.

For atypical sequences, the code word length is $n\log|\mathcal{X}|+2$ bits.

Theorem 3.6. ***Average Code Word Length***
page 61 and Notes 4/6/11

$$\begin{aligned} L &= \sum_{x_1^n \in A_\epsilon^{(n)}} p(x_1^n) l_1 + \sum_{x_1^n \notin A_\epsilon^{(n)}} p(x_1^n) l_2 \\ &= n(H+\epsilon)\sum_{x_1^n \in A_\epsilon^{(n)}} p(x_1^n) + n\log|\mathcal{X}| \sum_{x_1^n \notin A_\epsilon^{(n)}} p(x_1^n) + 2 \\ &\le n(H+\epsilon) + n\log|\mathcal{X}|\epsilon + 2 \\ &\le n[H(X)+\epsilon'] \end{aligned}$$

where $\epsilon' = \epsilon + \epsilon\log|\mathcal{X}| + \frac{2}{n}$.

Example 3.7.
Notes 4/11/11

Consider a biased coin with $p(\text{heads}) = 0.9$. The Asymptotic Equipartition Property (Theorem 3.2) says that if we flip it enough times then

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

Definition 3.8. ***High-Probability Set***
page 62 and Notes 4/11/11

For each $n = 1, 2, \ldots$, define the *high-probability set* $B_\delta^{(n)} \subset \mathcal{X}^n$ to be the smallest set with

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

Remark 3.9. ***Typical Sequence $\neq$ Most Likely Sequence***
Notes 4/11/11

(From Example 3.7) Typical sequences have 90% heads. The most likely sequence is all heads.

Theorem 3.10.
page 63 and Notes 4/11/11

Let $X_1, \ldots, X_n$ be i.i.d. $\sim p(x)$. Then for every $\delta' > 0$,

$$\begin{aligned}\frac{1}{n}\log|B_\delta^{(n)}| &> H - \delta'\\ |B_\delta^{(n)}| &> 2^{n(H-\delta')}\end{aligned}$$

Proof.

$$\begin{aligned}\Pr\ \{A_\epsilon^{(n)} \cap B_\delta^{(n)}\} &= \sum_{x_1^n \in A_\epsilon^{(n)} \cap B_\delta^{(n)}} \Pr\ (x_1^n) = \sum_{x_1^n \in A_\epsilon^{(n)}} p(x_1^n) + \sum_{x_1^n \in B_\delta^{(n)}} p(x_1^n) - \sum_{x_1^n \in A_\epsilon^{(n)} \cup B_\delta^{(n)}} p(x_1^n)\\ &> (1-\epsilon) + (1-\delta) - 1\\ &> 1 - \epsilon - \delta \end{aligned} \tag{3.1}$$

We also get

$$\begin{aligned}\Pr\ \{A_\epsilon^{(n)} \cap B_\delta^{(n)}\} &= \sum_{x_1^n \in A_\epsilon^{(n)} \cap B_\delta^{(n)}} \Pr\ (x_1^n)\\ &\leq \sum_{x_1^n \in A_\epsilon^{(n)} \cap B_\delta^{(n)}} 2^{-n(H-\epsilon)} = |A_\epsilon^{(n)} \cap B_\delta^{(n)}| 2^{-n(H-\epsilon)}\\ &\leq |B_\delta^{(n)}| 2^{-n(H-\epsilon)}\end{aligned} \tag{3.2}$$

Combining (3.1) and (3.2) gives

$$\begin{aligned}
|B_\delta^{(n)}|2^{-n(H-\epsilon)} &\geq 1-\epsilon-\delta \\
|B_\delta^{(n)}| &\geq 2^{n(H-\epsilon)}(1-\epsilon-\delta) \\
\frac{1}{n}\log|B_\delta^{(n)}| > H - \underbrace{\epsilon + \frac{\log(1-\epsilon-\delta)}{n}}_{\delta'} &= H-\delta'
\end{aligned}$$

□

Remark 3.11. ***Notation:*** $\doteq$
page 63 and Notes 4/11/11

$a_n \doteq b_n$ denotes that a_n and b_n are equal to the first order exponent. That is,

$$\lim_{n\to\infty}\frac{1}{n}\log\frac{a_n}{b_n} = 0$$

For example:

$$a_n = 2^{n\left(H+\frac{\sqrt{n}}{n}\right)}, \qquad b_n = 2^{n\left(H+\frac{\log n}{n}\right)}, \qquad c_n = 2^{nH}$$

It is easily seen that $a_n \doteq b_n \doteq c_n$.

4 Entropy Rates of a Stochastic Process

4.1 Markov Chains

Definition 4.1. ***Stochastic Process, Stationary***
page 71 and Notes 4/11/11

A *stochastic process* $\{X_i\}$ is an indexed sequence of random variables that is characterized by the joint distribution $p(x_1, x_2, \ldots, x_n)$. A stochastic process is said to be *stationary* if it is invariant with respect to shifts in the time index; that is,

$$\Pr\{X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\} = \Pr\{X_{1+l} = x_1, X_{2+l} = x_2, \ldots, X_{n+l} = x_n\}$$

4.2 Entropy Rate

Definition 4.2. ***Entropy Rate***
page 74 and Notes 4/11/11

The *entropy rate* of a stochastic process is

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

provided the limit exists. A second definition is given by

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

provided the limit exists.

Example 4.3. ***Entropy Rate Examples***
Notes 4/11/11

1. Given: $X_1, X_2, \ldots, X_n$ are i.i.d. random variables. Then $H(\mathcal{X}) = H(X) = H'(\mathcal{X})$.
2. Given: X_i are binary random variables with $p_i = \Pr[X_i = 1]$ independent.

$$p_i = \begin{cases} 0.5 & \text{if } \lceil \log i \rceil \text{ is odd} \Rightarrow H(X_i) = 1 \\ 0 & \text{if } \lceil \log i \rceil \text{ is even} \Rightarrow H(X_i) = 0 \end{cases}$$

i	1	2	3	4	5	6	7	8	9
$H(X_i)$	0	1	0	0	1	1	1	1	0

$$H(X_{2^{r-1}+1}) = H(X_{2^r}) = \begin{cases} 1 & \text{if } r \text{ odd} \\ 0 & \text{if } r \text{ even} \end{cases}$$

$$\sum_{i=1}^{2^r} H(X_i) = \begin{cases} 1 + 2^2 + 2^4 + \ldots + 2^{r-1} = \frac{2^{r+1}-1}{3} & r \text{ odd} \\ 1 + 2^2 + \ldots + 2^r = \frac{2^{r-1}-1}{3} & r \text{ even} \end{cases}$$

$$\frac{\sum_{i=1}^{2^r} H(X_i)}{2^r} = \begin{cases} \frac{2}{3} - \frac{1}{3 \cdot 2^r} & r \text{ odd} \\ \frac{1}{3} - \frac{1}{3 \cdot 2^r} & r \text{ even} \end{cases} \Rightarrow \text{ no limit}$$

$$H'(\mathcal{X}) = \lim_{n \to \infty} H(X_n | X_1, \ldots, X_n) \Rightarrow \text{ does not exist}$$

Theorem 4.4.
page 75 and Notes 4/11/11

For a stationary stochastic process, $H(\mathcal{X})$ and $H'(\mathcal{X})$ are defined and equal.

Proof. First show $H'(\mathcal{X})$ is defined.

$$H(X_n | X_1, \ldots, X_{n-1}) \leq H(X_n | X_2, \ldots, X_{n-1}) = H(X_{n-1} | X_1, \ldots, X_{n-2})$$

because it is stationary. The sequence is nonincreasing and nonnegative, so the limit exists. Computing $H(\mathcal{X})$ we get that

$$\frac{1}{n} H(X_1, \ldots, X_n) = \frac{1}{n}(H(X_1) + H(X_2|X_1) + \ldots + H(X_n | X_1, \ldots, X_{n-1}) \to H'(\mathcal{X})$$

by the Cesáro Mean Theorem (Theorem 4.5). □

Theorem 4.5. ***Cesáro Mean***
page 76 and Notes 4/11/11

If $a_n \to a$ and $b_n = \frac{1}{n} \sum_{i=1}^{n} a_i$, then $b_n \to a$.

Theorem 4.6. ***Shannon-McMillan-Breiman Theorem (AEP)***
page 77 and Notes 4/11/11

For any stationary ergodic process, we have

$$-\frac{1}{n}\log p(X_1,\ldots,X_n)\xrightarrow{\text{i.p.}}H(\mathcal{X})$$

with probability 1. The proof uses the law of large numbers for ergodic processes.

Example 4.7. ***Markov Chain, Time-Invariant, Probability Transition Matrix, Irreducible, Aperiodic, Stationary Distribution***
page 73 and Notes 4/11/11

Consider a Markov chain $X_1,\ldots,X_n$. Each random variable depends only on the one preceding it and is conditionally independent of all the other preceding random variables; that is,

$$\Pr\left[X_n|X_1,\ldots,X_{n-1}\right]=\Pr\left[X_n|X_{n-1}\right]$$

If $\Pr\left[X_n|X_{n-1}\right]=$ constant for all n, then the Markov chain is *time-invariant* and we write

$$\Pr\left[X_n|X_{n-1}\right]\equiv P_{i,j}$$

We form the *probability transition matrix* $P=[P_{ij}],\ i,j\in\{1,2,\ldots,m\}$ by setting

$$P_{ij}=\Pr\left[X_n=j|X_{n-1}=i\right]$$

If it is possible to go with positive probability from any state of the Markov chain to any other state in a finite number of steps then the Markov chain is said to be *irreducible*. If the largest common factor of the lengths of different paths from a state to itself is 1, the Markov chain is *aperiodic*.

If there exists a state $\pi=[P_1,\ldots,P_n]$ such that the distribution at the next time step is identical, i.e. $\pi=P\pi$, then π is a *stationary distribution*. If $\Pr\left[X_1\right]=\pi$ then we will stay there forever and the Markov chain is a stationary process, and

$$\begin{aligned}H(\mathcal{X})&=\lim_{n\to\infty}H(X_n|X_1,\ldots,X_{n-1})\\&=\lim_{n\to\infty}H(X_n|X_{n-1})\\&=H(X_2|X_1)\\&=\sum_{i=1}^{M}\pi_iH(X_2|X_1=i)\\&=\sum_{i=1}^{M}\pi_i\sum_{j=1}^{M}P_{ij}\log\frac{1}{P_{ij}}\end{aligned}$$

In other words, we have (at least for a 2 state Markov chain, see HW3 Problem 4.7)

$$H(\mathcal{X})=\mu_1H(\mathbb{P}_{\text{row }1})+\mu_2H(\mathbb{P}_{\text{row }2}).$$

If we have a finite, irreducible Markov chain with finite space, then it has a limiting distribution (the unique stationary distribution).

5 Data Compression

5.1 Examples of Codes

Definition 5.1. ***Source Code***
page 103 and Notes 4/13/11

A *source code* C for a random variable X is a mapping from $\mathcal{X}$ to $\mathcal{D}^*$, the set of finite-length strings from a D-ary alphabet. Let $C(x)$ denote the codeword corresponding to x and let $l(x)$ denote the length of $C(x)$.

Definition 5.2. ***Expected Length***
page 104 and Notes 4/13/11

The *expected length* $L(C)$ of $C(x)$ is given by

$$L(C) = \sum_x p(x)l(x)$$

Definition 5.3. ***Nonsingular***
page 105 and Notes 4/13/11

A code is nonsingular if every element in $\mathcal{X}$ is mapped to a different codeword. In other words, $x \neq x'$ implies that $C(x) \neq C(x')$.

Definition 5.4. ***Extension, Uniquely Decodable***
page 105 and Notes 4/13/11

The *extension* C^* of a code C is the mapping from finite-length strings of $\mathcal{X}$ to finite-length strings in D^* given by

$$C(x_1x_2 \ldots x_n) = C(x_1)C(x_2) \ldots C(x_n)$$

A code is *uniquely decodable* if its extension is nonsingular.

Definition 5.5. ***Instantaneous Code, Prefix Code***
page 106 and Notes 4/13/11

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

Remark 5.6.
page 106 and Notes 4/13/11

All codes $\supset$ Nonsingular $\supset$ Uniquely Decodable $\supset$ Instantaneous

Example 5.7.
page 107 and Notes 4/13/11

X	Singular	Nonsingular, not uniquely decodable	Uniquely decodable, not instantaneous	Instantaneous
1	0	0	10	0
2	0	010	00	10
3	0	01	11	110
4	0	10	110	111

5.2 Kraft Inequality

Theorem 5.8. ***Kraft Inequality***
page 107 and Notes 4/13/11

For any prefix code over an alphabet of size $D \geq 2$, the codeword lengths $l_1, l_2, \ldots, l_m$ must satisfy

$$\sum_i D^{-l_i} \leq 1$$

Conversely, given a set of codeword lengths satisfying this inequality, there exists a prefix code with those codeword lengths.

Theorem 5.9. ***Extended Kraft Inequality***
page 109 and Notes 4/13/11

For any countably infinite set of codewords that form a prefix code (or a uniquely decodable code), the codeword lengths satisfy

$$\sum_{i=1}^{\infty} D^{-l_i} \leq 1$$

Conversely, given any $l_1, l_2, \ldots$ satisfying the above inequality, we can construct a prefix code with these codeword lengths.

Theorem 5.10. ***Kraft Inequality (McMillan)***
page 116 and Notes 4/18/11

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

$$\sum D^{-l_i} \leq 1$$

Proof. Consider C^k, the kth extension of the code. By the definition of unique decodability, the kth extension of the code is nonsingular. Then

$$\begin{aligned}\left(\sum_{x\in\mathcal{X}} D^{-l(x)}\right)^k &= \sum_{x_1\in\mathcal{X}}\sum_{x_2\in\mathcal{X}}\cdots\sum_{x_k\in\mathcal{X}} D^{-l(x_1)}D^{-l(x_2)}\dots D^{-l(x_k)}\\ &= \sum_{x_1,x_2,\dots,x_k\in\mathcal{X}^k} D^{-l(x_1)}D^{-l(x_2)}\dots D^{-l(x_k)}\\ &= \sum_{x^k\in\mathcal{X}^k} D^{-l(x^k)}\end{aligned}$$

and somehow this leads to the desired result. □

5.3 Optimal Codes

Remark 5.11.
page 110 and Notes 4/18/11

We want to minimize

$$L = \sum p_i l_i$$

while satisfying

$$\sum D^{-l_i} \le 1.$$

We do this using Lagrange multipliers. We set

$$\begin{aligned} J &= \sum p_i l_i + \lambda\left(\sum d^{-l_i}\right)\\ \frac{\partial J}{\partial l_i} &= p_i - \lambda D^{-l_i}\log_e D = 0\\ D^{-l_i} &= \frac{p_i}{\lambda \log_e D}\\ \lambda &= \frac{1}{\log_e D}\\ p_i &= D^{-l_i}\\ l_i^* &= -\log_D p_i\end{aligned}$$

where l_i^* is the optimal code length for x_i.

Theorem 5.12.
page 111 and Notes 4/18/11

The expected length L of any prefix D-ary code for a random variable X satisfies

$$L \ge H_D(X)$$

with equality iff $\log_D \frac{1}{p_i}$ is an integer for all i.

Proof.

$$\begin{aligned} L - H_D(X) &= \sum p_i l_i - \sum p_i \log_D \frac{1}{p_i} \\ &= -\sum p_i \log_D D^{-l_i} + \sum p_i \log_D p_i \end{aligned}$$

Let

$$c = \sum D^{-l_i} \qquad \text{and} \qquad r_i = \frac{D^{-l_i}}{\sum D^{-l_i}} = \frac{D^{-l_i}}{c}$$

Then continuing from above, we have

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

□

Definition 5.13. ***D-adic***
page 112 and Notes 4/18/11

A probability distribution is D-*adic* if each probability equals D^{-n} for some integer n.

5.4 Bounds on the Optimal Code Length

Definition 5.14. ***Shannon-Fano Coding***
page 112 and Notes 4/18/11

Choose code lengths by

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

This is a prefix code because

$$\sum_i D^{-l_i} = \sum_i D^{-\left\lceil \log_D \frac{1}{p_i} \right\rceil} \leq \sum_i D^{-\log_D \frac{1}{p_i}} = \sum p_i = 1$$

We can bound the expected codeword length by

$$L = \sum_i p_i \left\lceil \log_D \frac{1}{p_i} \right\rceil \leq \sum_i p_i \left(\log_D \frac{1}{p_i} + 1 \right) = H_D(X) + 1$$

Theorem 5.15.
page 113 and Notes 4/18/11

Let L^* be the associated expected length of the optimal prefix code. Then

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

Remark 5.16. ***Approaching the Entropy***
page 113 and Notes 4/18/11

Let L_n be the expected codeword length per input symbol; that is,

$$L_n = \frac{1}{n} \sum_{(x_1,\ldots,x_n)\in\mathcal{X}^n} p(x_1,\ldots,x_n) l(x_1,\ldots,x_n)$$

Then by Theorem 5.15,

$$H_D(X_1,\ldots,X_n) \leq nL_n \leq H_D(X_1,\ldots,X_n) + 1$$

Because $X_1,\ldots,X_n$ are i.i.d., $H(X_1,\ldots,X_n) = \sum H(X_i) = nH(X)$. Thus, we get

$$H_D(X) \leq L_n \leq H_D(X) + \frac{1}{n}$$

If we have a stochastic process that is stationary, then

$$L_n \to H(\mathcal{X})$$

Theorem 5.17.
page 114 and Notes 4/18/11

The minimum expected codeword length per symbol satisfies

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

Moreover, if $X_1,\ldots,X_n$ is a stationary stochastic process then

$$L_n^* \to H(\mathcal{X})$$

Theorem 5.18. ***Wrong Code***
page 115 and Notes 4/18/11

If the true distribution is $p(x)$ and our code is designed for $q(x)$ with $l(x) = \left\lceil \log \frac{1}{q(x)} \right\rceil$, then

$$H(p) + D(p||q) \leq \mathbb{E}_p l(X) \leq H(p) + D(p||q) + 1$$

Proof.

$$\mathbb{E}_p l(X) = \sum_x p(x) \left\lceil \log \frac{1}{q(x)} \right\rceil$$

$$< \sum_x p(x) \left(\log \frac{1}{q(x)} + 1 \right) = \sum_x p(x) \log \frac{1}{q(x)} \cdot \frac{p(x)}{p(x)} + 1$$

$$< \sum_x p(x) \log \frac{p(x)}{q(x)} + \sum_x p(x) \log \frac{1}{p(x)} + 1$$

$$< H(p) + D(p||q) + 1$$

□

5.6 Huffman Codes

Example 5.19. ***Huffman Code*** *($D = 2$)*
page 118 and Notes 4/20/11

Construction of Huffman code for $D = 2$, $\mathcal{X} = \{1, 2, 3, 4, 5\}$, $p = \{0.25, 0.25, 0.2, 0.15, 0.15\}$

1	$0.25 \Rightarrow 01$	$0.3 \Rightarrow 00$	$0.45 \Rightarrow 1$	$0.55 \Rightarrow 0$
2	$0.25 \Rightarrow 10$	$0.25 \Rightarrow 01$	$0.25 \Rightarrow 10$	$0.2 \Rightarrow 11$
3	$0.2 \Rightarrow 11$	$0.25 \Rightarrow 10$	$0.25 \Rightarrow 01$	
4	$0.15 \Rightarrow 000$	$0.2 \Rightarrow 11$		
5	$0.15 \Rightarrow 001$			

Example 5.20. ***Huffman Code*** *($D = 3$)*
page 119 and Notes 4/20/11

Construction of Huffman code for $D = 2$, $\mathcal{X} = \{1, 2, 3, 4, 5\}$, $p = \{0.25, 0.25, 0.2, 0.15, 0.15\}$

1	0.25	$0.5 \Rightarrow 0$
2	0.25	$0.25 \Rightarrow 1$
3	0.2	$0.2 \Rightarrow 2$
4	0.15	
5	0.15	

Example 5.21. ***Huffman Code*** *($D = 4$)*
page 119 and Notes 4/20/11

Construction of Huffman code for $D = 2$, $\mathcal{X} = \{1, 2, 3, 4, 5\}$, $p = \{0.25, 0.25, 0.2, 0.15, 0.15\}$

$1 \Rightarrow 1$	0.25	$0.3 \Rightarrow 0$
$2 \Rightarrow 2$	0.25	$0.25 \Rightarrow 1$
$3 \Rightarrow 3$	0.2	$0.25 \Rightarrow 2$
$4 \Rightarrow 00$	0.15	0.2
$5 \Rightarrow 01$	0.15	
6	0	
7	0	

Remark 5.22.
page 119 and Notes 4/20/11

- The total number of symbols should be $1 + k(D-1)$
- It is possible to have 2 optimal codes with different codeword lengths, but the same expected codeword length
- The codeword lengths of optimal codes are not unique

Example 5.23.
Notes 4/20/11

Let $D = 2,\ \mathcal{X} = \{1,2,3,4\},\ p = \left\{\frac{1}{3}, \frac{1}{3}, \frac{1}{4}, \frac{1}{12}\right\}$.

$1 \Rightarrow 1$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{2}{3}$
$2 \Rightarrow 00$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$
$3 \Rightarrow 010$	$\frac{1}{4}$	$\frac{1}{3}^*$	
$4 \Rightarrow 011$	$\frac{1}{12}$		

$1 \Rightarrow 00$	$\frac{1}{3}$	$\frac{1}{3}^*$	$\frac{2}{3}$
$2 \Rightarrow 01$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$
$3 \Rightarrow 10$	$\frac{1}{4}$	$\frac{1}{3}$	
$4 \Rightarrow 11$	$\frac{1}{12}$		

5.7 Some Comments on Huffman Codes

Remark 5.24. ***Huffman vs. Shannon***
page 122 and Notes 4/20/11

For Shannon code, $\left\lceil \log \frac{1}{p_i} \right\rceil$, choose p_i small, e.g. $p = \{0.999, 0.001\}$. Then for Huffman code,

$$l_i \le \left\lceil \log \frac{1}{p_i} \right\rceil$$

5.8 Optimality of Huffman Codes

Lemma 5.25.
page 123 and Notes 4/20/11

For any distribution, there exists an optimal prefix code that satisfies

1. the lengths of the codeword are ordered inversely with probability, i.e. $p_j \ge p_k \Rightarrow l_j \le l_k$.
2. the two longest codewords have the same length.
3. two of the longest codewords differ only in the last bit

Proof. Consider C' with codewords j and k interchanged from C^*. Then

$$\begin{aligned} L(C') - L(C^*) &= p_j l_k + p_k l_j - p_j l_j - p_k l_k \\ &= \underbrace{(p_j - p_k)}_{\geq 0}(l_k - l_j) \end{aligned}$$

□

Definition 5.26. ***Canonical Codes***
page 125 and Notes 4/20/11

Canonical codes are codes that satisfy the 3 properties in Lemma 5.25.

Definition 5.27. ***Huffman Reduction***
page 125 and Notes 4/20/11

$$\begin{aligned} |\mathcal{X}| = m,\ \mathbb{P} = (p_1, \ldots, p_m) &\text{ with } p_1 \geq p_2 \geq \cdots \geq p_m \\ |\mathcal{X}'| = m-1,\ \mathbb{P} = (p_1, \ldots, p_{m-2}, p_{m-1} + p_m) \end{aligned}$$

Remark 5.28.
Notes 4/20/11

Let $C^*_{m-1}(\mathbb{P}')$ be the optimal code for $\mathbb{P}'$.
Let $C^*_m(\mathbb{P})$ be the optimal code for $\mathbb{P}$.
From $C^*_{m-1}(\mathbb{P}')$ we can construct an extension code for $|\mathcal{X}| = m$. To do this, take the codeword in C^*_{m-1} for $p_{m-1} + p_m$ and extend it by adding 1 more bit at the end. The average length $\sum_i l_i p_i$ is:

$$L(\mathbb{P}) = L^*(\mathbb{P}') + p_{m-1} + p_m$$

Start from a canonical code for $|\mathcal{X}| = m$. We can construct a code for $\mathbb{P}'$ by throwing away the last bit of the two codewords for p_{m-1} and p_m. Then we have

$$\begin{aligned} L(\mathbb{P}') &= L^*(\mathbb{P}) - p_{m-1} - p_m \qquad \big(L^*(\mathbb{P}) = p_{m-1} l_{\max} + p_m l_{\max}\big) \\ L(\mathbb{P}) + L(\mathbb{P}') &= L^*(\mathbb{P}) + L^*(\mathbb{P}') \\ \underbrace{[L(\mathbb{P}') - L^*(\mathbb{P}')]}_{0} + \underbrace{[L(\mathbb{P}) - L^*(\mathbb{P})]}_{0} &= 0 \end{aligned}$$

7 Channel Capacity

7.1 Examples of Channel Capacity

Definition 7.1. ***Discrete Channel***
page 183 and Notes 4/25/11

A *discrete channel* consists of

- A discrete alphabet $\mathcal{X}$ (input alphabet)
- A discrete alphabet $\mathcal{Y}$ (output alphabet)
- A conditional probability $p(y^n|x^n)$ for each n

$$x^n = (x_1, \ldots, x_n) \in \mathcal{X}^n$$
$$y^n = (y_1, \ldots, y_n) \in \mathcal{Y}^n$$

Definition 7.2. ***Memoryless Channnel***
page 184 and Notes 4/25/11

A *memoryless channel* satisfies

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

Remark 7.3.
Notes 4/25/11

A channel can be given by a matrix, $\mathbb{P}$, with rows corresponding to x and columns corresponding to y.

Definition 7.4. ***Operational Channel Capacity***
page 184 and Notes 4/25/11

Operational channel capacity is the highest rate at which information can be sent (with arbitrarily low probability of error).

Definition 7.5. ***Information Channel Capacity***
page 184 and Notes 4/25/11

We define the *information channel capacity* as

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

Example 7.6. ***Noisy Channel with Nonoverlapping Outputs***
page 185 and Notes 4/25/11

$0 \mapsto 0$
$1 \mapsto 1, 2$ with equal probability
$2 \mapsto 3$

$$\mathbb{P} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

There is no ambiguity (nonoverlapping output).

$$\begin{aligned} C &= \max_{p(x)} I(X;Y) = \max_{p(x)} H(X) - H(X|Y) = \max_{p(x)} H(X) \\ &= \log 3 \end{aligned}$$

Example 7.7. ***Noisy Typewriter***
page 186 and Notes 4/25/11

$A \mapsto A, B$ with equal probability, $B \mapsto B, C$ with equal probability, ..., $Z \mapsto Z, A$ with equal probability.

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) = H(Y) - 1 \\ &\leq \log 26 - 1 \\ C &= \max_{p(x)} H(Y) - 1 = \log 26 - 1 \\ &= \log 13 \end{aligned}$$

Example 7.8. ***Binary Symmetric Channel***
page 187 and Notes 4/25/11

$$\mathbb{P} = \begin{bmatrix} 1-p & p \\ p & 1-p \end{bmatrix}$$

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) = H(Y) - H(p) \\ &\leq 1 - H(p) \\ C &= 1 - H(p), \qquad \text{achieved when } p(x) \text{ is uniform} \end{aligned}$$

Example 7.9. ***Binary Erasure***
page 188 and Notes 4/25/11

$$0 \mapsto \begin{cases} 0 & \text{with probability } 1-\alpha \\ e & \text{with probability } \alpha \end{cases}$$

$$1 \mapsto \begin{cases} e & \text{with probability } \alpha \\ 1 & \text{with probability } 1-\alpha \end{cases}$$

Define

$$E = \begin{cases} 0 & \text{if } Y = e \\ 1 & \text{if } Y \neq e \end{cases}$$

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) = H(Y) - H(\alpha) \\ H(Y) &= H(Y,E) = H(E) + H(Y|E) = H(\alpha) \\ H(Y|E) &= \Pr\,[E=0]H(Y|E=0) \\ &\quad + \Pr\,[E=1]H(Y|E=1) \\ &\le 1-\alpha \\ C &= \max_{p(x)} [H(E) + H(Y|E) - H(\alpha)] \\ &= 1-\alpha \end{aligned}$$

Example 7.10.
Notes 4/25/11

$$\mathbb{P} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0.8 & 0.2 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Define a probability distribution for X: $p(0,1,2,3) \sim (p_0, p_1, p_2, p_3)$.

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ H(X|Y) &= \sum_y H(X|Y=y)p(y) = H(X|Y=3)\Pr\,(Y=3) \\ &= \cancel{(p_2+p_3)} \left[\frac{p_2}{\cancel{p_2+p_3}} \log \frac{p_2+p_3}{p_2} + \frac{p_3}{\cancel{p_2+p_3}} \log \frac{p_2+p_3}{p_3} \right] \\ &= p_2 \log \frac{p_2+p_3}{p_2} + p_3 \log \frac{p_2+p_3}{p_3} \\ I(X;Y) &= p_0 \log \frac{1}{p_0} + p_1 \log \frac{1}{p_1} + p_2 \log \frac{1}{p_2} + p_3 \log \frac{1}{p_3} - p_2 \log \frac{p_2+p_3}{p_2} - p_3 \log \frac{p_2+p_3}{p_3} \\ &= p_0 \log \frac{1}{p_0} + p_1 \log \frac{1}{p_1} + (p_2+p_3) \log \frac{1}{p_2+p_3} \\ C &= \log 3, \qquad \text{achieved with } p_0 = p_1 = p_2 + p_3 \end{aligned}$$

7.2 Symmetric Channels

Definition 7.11. ***Weakly Symmetric***
page 190 and Notes 4/27/11

A channel is *weakly symmetric* if the rows of $\mathbb{P}$ are permutations of each other and all the column sums are equal.

Definition 7.12. ***Symmetric***
page 190 and Notes 4/27/11

A channel is *symmetric* if the rows and columns are permutations of each other.

Theorem 7.13.
page 191 and Notes 4/27/11

For a weakly symmetric channel $(\mathcal{X}, \mathbb{P}, \mathcal{Y})$,

$$C = \max_{p(x)} I(X;Y) = \log|\mathcal{Y}| - H(\text{row of transition matrix})$$

Proof.

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) = H(Y) - H(\text{row of } \mathbb{P}) \\ \max_{p(x)} I(X;Y) &= \log|\mathcal{Y}| - H(\text{row of } \mathbb{P}) \end{aligned}$$

which is achieved for $p(x) =$ uniform distribution. □

7.3 Properties of Channel Capacity

Remark 7.14.
page 191 and Notes 4/27/11

1. $C \geq 0$ (since mutual information is nonnegative)
2. $C \leq \log|\mathcal{X}|$
3. $C \leq \log|\mathcal{Y}|$
4. $I(X;Y)$ is a continuous and concave function of $p(x)$, so $C = \max_{p(x)} I(X;Y)$, and a local maximum is a global maximum

7.5 The Communication System

Definition 7.15. ***The Communication System***
page 193 and Notes 4/27/11

$$\xrightarrow{W(\text{message})} \text{Encoder} \xrightarrow{X^n} \text{Channel } p(y|x) \xrightarrow{Y^n} \text{Decoder} \xrightarrow{\hat{W}(\text{estimate of message})}$$

A message W, drawn from $\{1, 2, \ldots, M\}$, results in the signal $X^n(W)$. $X^n(i)$ denotes the codeword for message i.

The receiver receives the message as $Y^n \sim p(y^n|x^n)$.

The receiver guesses the message using a decoding rule $\hat{W} = g(Y^n)$.

If $\hat{W} \neq W$ then the receiver has made an error.

Definition 7.16. (M, n) ***Codebook***
page 193 and Notes 4/27/11

An (M, n) code for the channel $(\mathcal{X}, p(y|x), \mathcal{Y})$ consists of the following:

1. An index set $\{1, 2, \ldots, M\}$.
2. An encoding function $X^n : \{1, 2, \ldots, M\} \to \mathcal{X}^n$. The set of codewords $x^n(1), x^n(2), \ldots, x^n(M)$ is called the *codebook*.
3. A decoding function $g : \mathcal{Y}^n \to \{1, 2, \ldots, M\}$.

Definition 7.17. ***Conditional Probability of Error***
page 194 and Notes 4/27/11

The *conditional probability of error* given that message i is sent is

$$\lambda_i = \Pr\ \left[g(Y^n) \neq i \;\middle|\; x^n = x^n(\lambda)\right]$$

Definition 7.18. ***Maximal Probability of Error***
page 194 and Notes 4/27/11

The *maximal probability of error* is

$$\lambda^{(n)} = \max_{i=1,\ldots,M} \lambda_i$$

Definition 7.19. ***Average Probability of Error***
page 194 and Notes 4/27/11

The *average probability of error* is

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

Definition 7.20. ***Rate, Achievable***
page 195 and Notes 4/27/11

The *rate* R of an (M, n) code is

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

A rate is said to be *achievable* if there exists a sequence of $\left(\lceil 2^{nR}\rceil, n\right)$ codes such that the max probability of error $\lambda^{(n)} \to 0$.

7.6 Jointly Typical Sequences

Definition 7.21. ***Jointly Typical Sequence***
page 195 and Notes 4/27/11

Let n be a positive integer and set $\epsilon > 0$. The set $A_\epsilon^{(n)}$ of *jointly typical sequences* with respect to $p(x, y)$ is given by

$$A_\epsilon^{(n)} = \Big\{(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n \;\Big|\; \left|1 - \frac{1}{n}\log p(x^n) - H(X)\right| < \epsilon, \\ \left|1 - \frac{1}{n}\log p(y^n) - H(Y)\right| < \epsilon, \\ \left|1 - \frac{1}{n}\log p(x^n, y^n) - H(X,Y)\right| < \epsilon\Big\}$$

Theorem 7.22. ***Joint AEP Theorem***
page 196 and Notes 4/27/11

Let X^n, Y^n be sequences of length n drawn according to $p(x^n, y^n) = \prod p(x_i, y_i)$.

1. $\Pr\left[(X^n, Y^n) \in A_\epsilon^{(n)}\right] \to 1$ as $n \to \infty$
2. $|A_\epsilon^{(n)}| \le 2^{n[H(X,Y)+\epsilon]}$
3. $|A_\epsilon^{(n)}| \ge 2^{n[H(X,Y)-\epsilon]}$
4. If $(\tilde{X}^n, \tilde{Y}^n) \sim p(x^n)p(y^n)$, then

$$\Pr\left[(X^n, Y^n) \in A_\epsilon^{(n)}\right] \le 2^{-n[I(X;Y)-3\epsilon]}$$
$$\Pr\left[(X^n, Y^n) \in A_\epsilon^{(n)}\right] \ge 2^{-n[I(X;Y)-3\epsilon]}$$

Proof. By the weak law of large numbers,

$$-\frac{1}{n}\log p(X^n) \to -\mathbb{E}[\log p(X)] = H(X)$$
$$-\frac{1}{n}\log p(Y^n) \to H(Y)$$
$$-\frac{1}{n}\log p(X^n, Y^n) \to H(X,Y)$$

For n large,

$$\Pr\left[\left|-\frac{1}{n}\log p(X^n) - H(X)\right| \geq \epsilon\right] < \frac{\epsilon}{3}$$
$$\Pr\left[\left|-\frac{1}{n}\log p(Y^n) - H(Y)\right| \geq \epsilon\right] < \frac{\epsilon}{3}$$
$$\Pr\left[\left|-\frac{1}{n}\log p(X^n, Y^n) - H(X,Y)\right| \geq \epsilon\right] < \frac{\epsilon}{3}$$

For the rest of the proof see pages 197 and 198. □

7.7 Channel Coding Theorem

Theorem 7.23. ***Channel Coding Theorem***
page 200 and Notes 5/2/11

For a discrete memoryless channel, all rates below capacity C are achievable. 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 $R < C$.

(See the Channel Coding Theorem Converse, Theorem 7.27.)

Proof. Fix $p(x) = p^*(x)$ that minimizes $I(X;Y)$. Generate each codebook according to $p(x)$. Fix $R < C$. Our $(2^{nR}, n)$ codebook is a $w^{nR} \times n$ matrix:

$$\begin{bmatrix} X^n(1) \\ X^n(2) \\ \vdots \\ X^n(2^{nR}) \end{bmatrix} = \begin{bmatrix} X_1(1), & X_2(1), & \ldots, & X_n(1) \\ X_1(2), & X_2(2), & \ldots, & X_n(2) \\ \vdots & \vdots & \ddots & \vdots \\ X_1(2^{nR}), & X_2(2^{nR}), & \ldots, & X_n(2^{nR}) \end{bmatrix}$$

All $2^{nR} \times n$ elements are i.i.d. $\sim p(x)$.

Assume: all messages are equally likely.

Optimal decoder: $\hat{W} = \arg\max \Pr\left[Y^n | X^n(i)\right],\ X^n(i) \in$ codebook.

We consider the jointly typical decoder: when we receive a sequence Y^n, if there exists a unique codeword $X^n(i)$ that is jointly typical with Y^n, then $\hat{W} = i$.

$$
\begin{aligned}
\Pr(\varepsilon) &= \sum_{\mathcal{C}\ \text{(codebooks)}} \Pr(\mathcal{C}P_e^{(n)}(\mathcal{C}) \\
&= \sum_{\mathcal{C}} \Pr(\mathcal{C}) \cdot \frac{1}{2^{nR}} \sum_{W=1}^{2^{nR}} \lambda_W(\mathcal{C}) \qquad (W \text{ is the index of the message}) \\
&= \frac{1}{2^{nR}} \sum_{W=1}^{2^{nR}} \sum_{\mathcal{C}} \Pr(\mathcal{C})\lambda_W(\mathcal{C}) \\
&= \Pr[\varepsilon | W = 1]
\end{aligned}
$$

Define the event $E_i,\ i = 1, 2, \ldots, 2^{nR}$, as

$$
E_i = \left\{ (X^n(i), Y^n) \in A_\epsilon^{(n)} \right\}
$$

where Y^n is generated by $X^n(1)$. Then

$$
\begin{aligned}
\varepsilon &= E_1^C \cup E_2 \cup E_3 \cup \cdots \cup E_{2^{nR}} \\
\Pr[\varepsilon | W = 1] &= \Pr[E_1^C \cup E_2 \cup \cdots \cup E_{2^{nR}} | W = 1] \\
&\le \Pr[E_1^C] + \sum_{i=2}^{2^{nR}} \Pr[E_i] \\
\Pr[E_1^C] &\le \epsilon \text{ for } n \text{ sufficiently large}
\end{aligned}
$$

To bound $\Pr[E_i]$,

$$
\Pr[E_i] \le 2^{-n[I(X;Y) - 3\epsilon]}
$$

$$
\begin{aligned}
\Pr[\varepsilon] &= \Pr[E | W = 1] \\
&\le \epsilon + \sum_{i=1}^{2^{nR}} 2^{-n[I(X;Y) - 3\epsilon]} \\
&\le \epsilon + \left(2^{nR} - 1\right) \cdot 2^{-n[I(X;Y) - 3\epsilon]} \\
&\le \epsilon + 2^{-n[I(X;Y) - R]} \cdot 2^{3n\epsilon} \\
&\le 2\epsilon \text{ for } n \text{ sufficiently large}
\end{aligned}
$$

Make $C - R > 3\epsilon \ \Rightarrow\ \epsilon < \frac{C-R}{3} \ \Rightarrow\ I(X;Y) - R - 3\epsilon > 0$. There exists a codebook $\mathcal{C}^*$ with average probability of error $P_e^{(n)}(\mathcal{C}^*) \le 2\epsilon$, i.e.

$$
P_e^{(n)}(\mathcal{C}^*) = \frac{1}{2^{nR}} \underbrace{\sum_{i=1}^{2^{nR}} \lambda_i(\mathcal{C}^*)}_{\le 2^{nR} \cdot 2^{\epsilon}} \le 2\epsilon
$$

At least half of the messages have $\lambda_i(\mathcal{C}^*) \le 4\epsilon$. Consider a codebook containing only these "good" codewords. We have $2^{nR-1} = 2^{nR'}$ codewords, where $R' = R - \frac{1}{n}$, each with probability of error $\le 4\epsilon$. □

7.8 Zero-Error Codes

Remark 7.24.
Notes 5/4/11

For any $(2^{nR}, n)$ code with zero probability of error, we have $R < C$.

$$\Pr\,[\hat{W} = W] = 1 \quad \Rightarrow \quad H(W|Y^n) = 0$$

Assume W is uniformly distributed.

$$\begin{aligned} nR = H(W) &= \underbrace{H(W|Y^n)}_{0} + I(W;Y^n) \\ &\le I(X^n;Y^n) \\ &\le nC \qquad R \le C \end{aligned}$$

$$\begin{gathered} W \to X^n \to Y^n \\ Y^n \to X^n \to W \end{gathered}$$

Recall Fano's Inequality (Theorem 2.36): If $\hat{X}$ is an estimate of X based on Y (i.e. $\hat{X} = g(Y)$), then $P_e \equiv \Pr\,[\hat{X} \neq X]$.

$$\begin{gathered} P_e = \Pr\,[\hat{X} \neq X] \le 1 + P_e \log|\mathcal{X}| \\ H(W|Y^n) \le 1 + P_e^{(n)} \log 2^{nR} = 1 + nRP_e^{(n)} \end{gathered}$$

where $P_e^{(n)}$ is the average probability of error.

7.9 Fano's Inequality and the Converse to the Coding Theorem

Lemma 7.25. ***Fano's Inequality***
page 206

For a discrete memoryless channel, we have

$$H(W|\hat{W}) \le 1 + P_e^{(n)} nR$$

Lemma 7.26.
page 206 and Notes 5/4/11

For a discrete memoryless channel,

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

Proof.

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

□

Theorem 7.27. ***Converse of the Channel Coding Theorem***
page 207 and Notes 5/4/11

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

(See the Channel Coding Theorem, Theorem 7.23.)

Proof. $\lambda^{(n)} \to 0$, so $P_e^{(n)} \to 0$ for any distribution of W. Consider the uniform distribution for W.

$$
\begin{aligned}
nR = H(W) &= H(W|Y^n) + I(W;Y^n) \\
&\le 1 + nRP_e^{(n)} + I(X^n;Y^n) && \text{(Fano's \& data-processing inequalities)} \\
&\le 1 + nRP_e^{(n)} + nC && \text{(Lemma 7.26} \\
P_e^{(n)} &\ge \frac{nR - nC - 1}{nR} = 1 - \frac{C}{R} - \frac{1}{nR}
\end{aligned}
$$

If $R > C$ then $P_e^{(n)} \not\to 0$ as $n \to \infty$. □

7.10 5-9-11

Theorem 7.28. ***Converse to Channel Coding Theorem (Review)***

If we have $(2^{nR}, n)$ codes with $\lambda^{(n)} \to 0$, then $R \leq C$.

Proof. Assume W is uniformly distributed over these 2^{nR} possible messages.
$W \to X^n \to Y^n \to \hat{W}$.

$$\begin{aligned} nR = H(W) &= \underbrace{H(W|\hat{W})}_{\substack{\text{bound} \\ \text{by Fano}}} + I(W;\hat{W}) \\ &\leq 1 + P_e^{(n)} nR + I(X^n;Y^n) \qquad \text{(by Data Processing Inequality)} \\ nR &\leq 1 + P_e^{(n)} nR + nC \\ P_e^{(n)} &\geq 1 - \frac{C}{R} - \frac{1}{nR} \end{aligned}$$

□

Remark 7.29.

So far our channel has looked like:

$$\xrightarrow{W}\rightarrow \text{Encoder} \xrightarrow{X^n} p(y|x) \xrightarrow{Y^n} \text{Decoder} \xrightarrow{\hat{W}}$$

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

What if our channel has feedback? In other words, the receiver can communicate with the transmitter. Feedback is always immediate and error-free. Can we transmit at a higher rate than without feedback?

With feedback, out channel looks like:

$$\xrightarrow{W}\rightarrow \underbrace{\text{Encoder} \xrightarrow{X_i(W,Y^{i-1})} p(y|x) \xrightarrow{Y_i}}_{\leftarrow} - \text{Decoder} \xrightarrow{\hat{W}}$$

$(2^{nR}, n)$ *feedback code*: a sequence of mapping $x_i(W, Y^{i-1})$ for each $i = 1, \ldots, n$.
Decoder: $g : y^n \rightarrow \{1, 2, \ldots, 2^{nR}\}$
Probability of Error: $P_e^{(n)} = \Pr\left[g(Y^n) \neq W\right]$

Direct: there exists a sequence of $(2^{nR}, n)$ codes ...
Converse:

$$\begin{aligned}
nR = H(W) &= H(W|\hat{W}) + I(W;\hat{W}) \\
&\leq 1 + P_e^{(n)} nR + I(W;\hat{W}) && \text{(Fano's Inequality)} \\
&\leq 1 + P_e^{(n)} nR + I(W;Y^n) && W \rightarrow X^n \rightarrow Y^n \rightarrow \hat{W} \\
I(W;Y^n) &= H(Y^n) - H(Y^n|W) \\
&= H(Y^n) - \sum_{i=1}^{n} H(Y_i|Y_1, \ldots, Y_{i-1}, W) \\
&= H(Y^n) - \sum_{i=1}^{n} H(Y_i|Y_1, \ldots, Y_{i-1}, W, X_i) \\
&= H(Y^n) - \sum_{i=1}^{n} H(Y_i|X_i) \\
&\leq \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i|X_i) \stackrel{?}{=} I(X;Y) \leq nC
\end{aligned}$$

This says that for a discrete memoryless channel, feedback doesn't get you anything extra.

Remark 7.30.

$$\underbrace{\text{Source, } V}_{\substack{\text{stationary,}\\ \text{ergodic}}} \to \underbrace{H(V)}_{R \geq H(V)}$$

We have $nH(V)$ messages and $2^{nH(V)}$ codes. We can transmit a source provided that $H(V) < C$.

$$\text{Source, } V \to \text{Encoder} \to p(y|x) \to$$

$$n \text{ outputs} \to \text{Source Code} \to \text{Channel Code}$$

Theorem 7.31. ***Source-Channel Coding Theorem***

If $V_1, V_2, \ldots, V_n$ is a finite alphabet stochastic process satisfying AEP (stationary and ergodic) with $H(V) < C$, then there exists a source-channel code with

$$\text{Pr }[\hat{V}^n \neq V^n] \to 0$$

Conversely, for any source with $H(V) > C$, the probability of error is bounded away from zero.

Definition 7.32. ***Source-Channel Code***

$$\xrightarrow{v^n=\{V_1,\ldots,V_n\}} \text{Source Coding} \to \text{Channel Coding} \xrightarrow{x^n(V^n)} p(y|x) \xrightarrow{Y^n} \text{Channel Coding} \to \text{Source Coding} \xrightarrow{\hat{V}^n}$$

$$\xrightarrow{V^n=\{V_1,\ldots,V_n\}} \text{Encoder} \xrightarrow{x^n(V^n)} p(y|x) \xrightarrow{Y^n} \text{Decoder} \xrightarrow{\hat{V}^n}$$

Remark 7.33.

Need to show:

$$\text{Pr }[\hat{V}^n \neq V^n] \to 0 \quad \text{implies} \quad H(V) \leq C$$

$x^n(V^n)$ can be viewed as a function:

$$x^n(V^n) : V^n \to \mathcal{X}^n$$

From Fano's Inequality we know the following:

$$\begin{aligned}
H(v^n|\hat{V}^n) &\leq 1 + \text{Pr }[\hat{V}^n \neq V^n] n \log |\mathcal{V}| \\
H(\mathcal{V}) &= \lim_{n\to\infty} \frac{H(V_1, \ldots, V_n)}{n} = \lim_{n\to\infty} H(V_n|V_1, \ldots, V_{n-1}) \\
&\leq \frac{H(V_1, \ldots, V_n)}{n} = \frac{H(V^n)}{n} = \frac{H(V^n|\hat{V}^n) + I(V^n; \hat{V}^n)}{n} \\
&\leq \frac{1}{n}\left(1 + P_e n \log |\mathcal{V}|\right) + \frac{1}{n} \qquad\qquad V^n \to X^n \to Y^n \to \hat{V}^n \\
H(V) &\leq \frac{1}{n} n + P_e \log |\mathcal{V}| + C \quad \to \quad P_e \log |\mathcal{V}| \geq H(V) - C - \frac{1}{n}
\end{aligned}$$

7.11 5-11-11

Example 7.34.

\# of information bits: 4
\# of parity check bits: 3

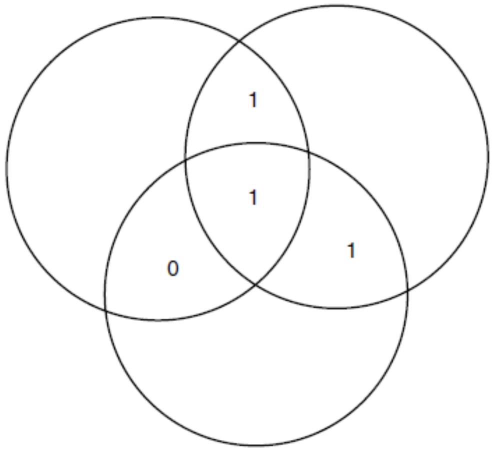


FIGURE 7.10. Venn diagram with information bits.

Definition 7.35. ***Hamming Codes***

Codeword length: $n = 2^m - 1$
\# of information bits: $k = 2^m - m - 1$
\# of parity check bits: $m = n - k$
Error correcting capability: $t = 1$ (regardless of m)
Coding rate: $\frac{k}{n} = \frac{2^m - m - 1}{2^m - 1}$
$\Rightarrow$ enlarging m gives a higher rate, but you can't correct as effectively

Example 7.36.

$m = 3,\ n = 2^3 - 1 = 7,\ k = 4$
The parity check matrix:

$$H = \begin{bmatrix} 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{bmatrix}$$

A codeword $C = [C_1\ C_2\ \ldots\ C_7]^T$ is one satisfying

$$HC = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \quad \text{modulo 2}$$

number of codewords: $2^4 = 16$
List of the codewords:

0000000	0001111	0010110	0011001
0100101	0101010	0110011	0111100
1000011	1001100	1010101	1011010
1100110	1101001	1110000	1111111

The first 4 bits are the information bits, and the last 3 are the parity check bits.

Note that every codeword (except 0000000) has at least 3 ones. Thus, the *minimum weight* = 3. We cannot have 1 or 2 ones because all of the columns of H are different, and thus no two columns can add up to $[0\ 0\ 0]^T$. The *minimum distance* (the # of bits that differ) between any two codewords is $d = 3$. Note that the distance between any 2 codewords is also a codeword:

$$HC_1 = 0$$
$$HC_2 = 0$$
$$H(C_1 - C_2) = 0$$

Suppose that a codeword c is transmitted with an error:

$$c \to r = c + e_i \qquad \text{where } e_i = [0\ \cdots\ \underbrace{1}_{i}\ 0\ \cdots\ 0]$$

$$Hr = \cancel{Hc} + He_i = i\text{th column of } H$$

The column of H that we end up with corresponds to the location of the error.

8 Differential Entropy

8.1 5-11-11

Definition 8.1. ***Differential Entropy***

For a discrete r.v. X, $H(X) = -\sum_x p(x) \log p(x)$
For a continuous r.v. with PDF $f(x)$,

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

where $S = \{x \mid f(x) > 0\} = \text{supp } x$

Example 8.2. ***Uniform Distribution***

A random variable distributed uniformly from 0 to a, $X \sim \mu(0, a)$, is given by

$$f(x) = \begin{cases} \frac{1}{a} & x \in (0, a) \\ 0 & \text{otherwise.} \end{cases}$$

Its entropy is given by

$$h(x) = -\int_0^a \frac{1}{a} \log \frac{1}{a}\, dx = \log a.$$

Example 8.3. ***Normal (Gaussian) Distribution***

A normally distributed random variable is given by

$$X \sim \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} = \phi(x).$$

We calculate its entropy as

$$\begin{aligned} h(x) &= -\int_{-\infty}^{\infty} \phi(x) \ln \phi(x)\, dx = -\int_{-\infty}^{\infty} \phi(x) \left(-\frac{x^2}{2\sigma^2} - \ln \sqrt{2\pi\sigma} \right) dx \\ &= \int_{-\infty}^{\infty} \phi(x) \frac{x^2}{2\sigma^2}\, dx + \ln \sqrt{2\pi\sigma^2} \int_{-\infty}^{\infty} \phi(x)\, dx \\ &= \frac{1}{2} + \ln \sqrt{2\pi\sigma^2} \\ &= \frac{1}{2} \ln 2\pi\sigma^2 e \text{ nats} \\ &= \frac{1}{2} \log 2\pi\sigma^2 e \text{ bits.} \end{aligned}$$

Remark 8.4.

For a fixed variance, a Gaussian distribution has the largest differential entropy.

8.2 5-18-11

Definition 8.5. ***Differential Entropy (Review)***

$x \sim f$, support $S \subset \mathbb{R}$ such that $f(x) > 0$

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

Uniform Distribution: $x \sim \mu(0, a) \quad \Rightarrow \quad h(X) = \log a$
Normal Distribution: $x \sim \mathcal{N}(\mu, \sigma^2) \quad \Rightarrow \quad h(X) = \frac{1}{2}\log(2\pi e \sigma^2)$

Theorem 8.6. ***AEP for Continuous Random Variables***

Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables $\sim f$. By the weak law of large numbers,

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

Definition 8.7. ***Typical Set*** $A_\epsilon^{(n)}$

For $\epsilon > 0$ and n, the *typical set* is

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

where $f(x_1, \ldots, x_n) = f(x_1) \cdots f(x_n)$.

Theorem 8.8.

The typical set has the following properties:

1. Pr $(A_\epsilon^{(n)}) > 1 - \epsilon$ for n sufficiently large
2. Vol $(A_\epsilon^{(n)}) \equiv \int_{A_\epsilon^{(n)}} dx_1 \cdots dx_n \leq 2^{n[h(X)+\epsilon]}$ for all n (this is the *volume* of the typical set)
3. Vol $(A_\epsilon^{(n)}) \geq (1-\epsilon)2^{n[h(X)-\epsilon]}$ for n sufficiently large

Theorem 8.9.

The set $A_\epsilon^{(n)}$ is the smallest volume set with probability $> 1 - \epsilon$ to the first order in the exponent (i.e. the $nh(X)$ term).

Remark 8.10.

Differential entropy can be negative. For example, $x \sim \mu(0, a),\ a < 0$.

Remark 8.11.

The sequences in $A_\epsilon^{(n)}$ are roughly equally likely, i.e. uniformly distributed.

Remark 8.12.

The differential entropy can be thought of as the log of the side length of the n-dimensional cube that is the typical set, where the volume of the typical set is

$$(2^{h(X)})^n \approx 2^{nh(X)}$$

Remark 8.13. ***Relationship Between Differential Entropy and Discrete Entropy***

We can quantize a differential random variable by dividing the range of X into intervals of length Δ. By the Mean Value Theorem, there exists $x_i \in [i\Delta, (i+1)\Delta]$ such that

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

Consider the quantized random variable x^Δ defined as

$$x^\Delta = x_i \quad \text{if } x \in [i\Delta, (i+1)\Delta]$$

Then $\Pr\left[x^\Delta = x_i\right] = \int_{i\Delta}^{(i+1)\Delta} f(x)\, dx = f(x_i)\Delta$.

$$\begin{aligned} H(X^\Delta) &= -\sum_{i=-\infty}^{\infty} p_i \log p_i = \sum_i f(x_i)\Delta \log(f(x_i)\Delta) = -\sum_i f(x_i)\Delta \log f(x_i) - \sum_i f(x_i)\Delta \log \Delta \\ &\xrightarrow{\Delta \to 0} -\int_x f(x) \log f(x)\, dx - \sum_i \left(\int_{i\Delta}^{(i+1)\Delta} f(x)\, dx \right) \log \Delta \\ &= h(X) - \log \Delta \\ h(X) &\approx H(X^\Delta) + \log \Delta \end{aligned}$$

Definition 8.14. ***Joint Entropy***

Given $X_1, \dots, X_n \sim f(x_1, \dots, x_n)$, the *joint entropy* is

$$h(X_1, \dots, X_n) = -\int f(x_1, \dots, x_n) \log f(x_1, \dots, x_n)\, dx_1 \dots dx_n$$

Definition 8.15. ***Conditional Differential Entropy***

Given $p(x|Y = y)$,

$$\begin{aligned} h(X|Y = y) &= -\int_y f(y) \int_x f(x|y) \log f(x|y)\, dx \\ &= -\int_{(x,y)} f(x, y) \log f(x|y)\, dx\, dy \end{aligned}$$

Definition 8.16. ***Relative Entropy (K-L Divergence)***

$$D(f||g) = \int_x f(x) \log \frac{f(x)}{g(x)}\, dx$$

Definition 8.17. ***Mutual Information***

$$\begin{aligned} I(X;Y) &= D(f(x, y)||f(x)f(y)) \\ &= \int f(x, y) \log \frac{f(x, y)}{f(x)f(y)}\, dx\, dy \\ &= h(Y) - h(Y|X) \\ &= \lim_{\Delta \to 0} I(X^\Delta, Y^\Delta) \\ &= \sup_{P,Q} I([X]_P; [Y]_Q) \end{aligned}$$

Example 8.18. ***Mutual Information between 2 Gaussian r.v.'s***

$(X,Y) \sim \mathcal{N}(0, \mathbf{k})$ where

$$\mathbf{k} = \begin{bmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{bmatrix}$$

Then

$$\begin{aligned} I(X;Y) &= h(X) + h(Y) - h(X,Y) \\ h(X) &= \frac{1}{2}\log 2\pi e\sigma^2 = h(Y) \\ h(X,Y) &= \frac{1}{2}\log(2\pi e)^2|\mathbf{k}| \\ &= \frac{1}{2}\log 2\pi e\sigma^2 + \frac{1}{2}\log 2\pi e\sigma^2 - \frac{1}{2}(2\pi e)^2\sigma^4(1-\rho^2) \\ &= -\frac{1}{2}\log(1-\rho^2) \end{aligned}$$

Proposition 8.19.

Properties:

- $D(f||q) \geq 0$
- $I(X;Y) \geq 0$ with equality iff X, Y are independent
- $h(X_1, \ldots, X_n) = \sum_{i=1}^{n} h(X_i|X_1, \ldots, X_{i-1}) \leq \sum_{i=1}^{n} h(X_i)$
- $h(X + c) = h(X)$
- $h(\alpha X) = h(X) + \log|\alpha|$
- $h(\mathbf{A}X) = h(X) + \log|\det \mathbf{A}|$

Definition 8.20. ***Jointly Gaussian***

$X_1, \ldots, X_n$ are *jointly Gaussian* if

$$f(x_1, \ldots, x_n) = \frac{1}{(\sqrt{2\pi})^n|\mathbf{k}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T\mathbf{K}^{-1}(\mathbf{x}-\mu)}$$

where

$$\mu = [\mu_1 \ \cdots \ \mu_n]^T = [\mathbb{E}(x_1) \ \cdots \ \mathbb{E}(x_n)]^T$$

and

$$\mathbf{K} = \mathbb{E}[(\mathbf{x}-\mu)(\mathbf{x}-\mu)^T] = \{K_{i,j}\}_{1 \leq i,j \leq n}$$

where $K_{i,j} = \mathbb{E}[(x_i - \mu_i)(x_j - \mu_j)]$.

Theorem 8.21.

$$h(\mathcal{N}(\mu, \mathbf{k})) = \frac{1}{2}\log((2\pi e)^n|\mathbf{k}|)$$

Proof.

$$\begin{aligned}
\mathcal{N}(\mu, \mathbf{k})) &= -\int f(\mathbf{x})\log f(\mathbf{x})\, d\mathbf{x} \\
&= \int f(\mathbf{x})\left(\frac{1}{2}(\mathbf{x}-\mu)^T\mathbf{k}^{-1}(\mathbf{x}-\mu)\right)\, d\mathbf{x} + \log\left((\sqrt{2\pi})^n|\mathbf{k}|^{1/2}\right) \\
&= \frac{1}{2}\mathbb{E}\left[(\mathbf{X}-\mu)^T\mathbf{k}^{-1}(\mathbf{X}-\mu)\right] \\
&= \frac{1}{2}\mathbb{E}\left[\sum_{i,j}(x_i-\mu_i)(\mathbf{k}^{-1})_{i,j}(x_j-\mu_j)\right] + \log((\sqrt{2\pi})^n|\mathbf{k}|^{1/2}) \\
&= \frac{1}{2}\sum_{i,j}\mathbb{E}\left[(x_i-\mu_i)(x_j-\mu_j)\right](\mathbf{k}^{-1})_{i,j} + \log((\sqrt{2\pi})^n|\mathbf{k}|^{1/2}) \\
&= \frac{1}{2}\sum_{i,j}(\mathbf{k})^{-1}_{i,j} + \log((\sqrt{2\pi})^n|\mathbf{k}|^{1/2}) \\
&= \frac{1}{2}\sum_j\sum_i \mathbf{k}_{j,i}(\mathbf{k}^{-1})_{i,j} + \log((\sqrt{2\pi})^n|\mathbf{k}|^{1/2}) \\
&= \frac{1}{2}\sum_j(\mathbf{k}\mathbf{k}^{-1})_{jj} + \log((\sqrt{2\pi})^n|\mathbf{k}|^{1/2}) \\
&= \frac{n}{2} + \log((\sqrt{2\pi})^n|\mathbf{k}|^{1/2}) \\
&= \frac{1}{2}\log\left((2\pi e)^n|\mathbf{k}|\right)
\end{aligned}$$

□

Remark 8.22. ***Connection to Linear Algebra***

Hadamad's Inequality tells us that

$$|\mathbf{k}| \leq \prod_{i=1}^n k_{i,i}$$

Proof.

$$\begin{aligned}
h(X_1, \ldots, X_n) &= \frac{1}{2}\log((2\pi e)^n|\mathbf{k}|) \\
&\leq \sum_{i=1}^n h(X_i) = \sum_i \frac{1}{2}\log 2\pi e k_{i,i} \\
|\mathbf{k}| &\leq \sum_i k_{i,i}
\end{aligned}$$

□

Theorem 8.23.

The Gaussian distribution maximizes entropy over all densities with the same variance. Specifically, if we have an n-dimensional vector $\mathbf{x}$ with $\mu, \mathbf{k}$, then

$$h(X) \leq \frac{1}{2}\log((2\pi e)^n|\mathbf{k}|)$$

with equality iff $x \sim \mathcal{N}_n(\mu, \mathbf{k})$.

Proof. Let $\mathbf{x} \sim g,\ \phi \sim \mathcal{N}(\mu, \|)$. Then

$$\int g(\mathbf{x}) \log \phi(\mathbf{x})\, d\mathbf{x} = \int \phi(\mathbf{x}) \log \phi(\mathbf{x})\, d\mathbf{x}$$

We compute the K-L divergence between g and ϕ:

$$\begin{aligned} 0 \leq D(g||\phi) &= \int g \log \frac{g}{\phi}\, d\mathbf{x} \\ &= -h(g) - \int g \log \phi\, dx \\ &= -h(g) + h(\phi) \\ h(g) &\leq h(\phi) \end{aligned}$$

□

9 Gaussian Channel

9.1 5-23-11

Definition 9.1. ***Gaussian Channel***

The *Gaussian channel* accepts a sequence $X_1, X_2, \ldots$ of real numbers and produces and output of Y_i's.

$$Y_i = X_i + Z_i, \qquad Z_i \sim \mathcal{N}(0, N)$$

Z_i's are independent of each other and X_i's.

Remark 9.2. ***Power Constraint***

For any codeword $(X_1, X_2, \ldots, X_n)$ transmitted over the channel,

$$\frac{1}{n}\sum_{i=1}^{n} x_i^2(w) \leq P$$

Example 9.3. ***One Way To Use Gaussian Channel***

$$x = \begin{cases} \sqrt{p} & \Pr\ \frac{1}{2} \\ -\sqrt{p} & \Pr\ \frac{1}{2} \end{cases}, \qquad \hat{x} = \begin{cases} \sqrt{p} & Y > 0 \\ -\sqrt{p} & Y < 0 \end{cases}$$

$$\begin{aligned} \Pr\text{ (error)} &= \frac{1}{2}\Pr\left\{Y \leq 0 \mid x = \sqrt{p}\right\} + \frac{1}{2}\Pr\left\{Y \geq 0 \mid x = -\sqrt{p}\right\} \\ &= \frac{1}{2}\Pr\left\{Z \leq -\sqrt{p}\right\} + \frac{1}{2}\Pr\left\{Z \geq \sqrt{p}\right\} \\ &= \Pr\left\{Z \geq \sqrt{p}\right\} \\ &= 1 - \Phi\left(\sqrt{\frac{p}{n}}\right) \end{aligned}$$

where

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt$$

Definition 9.4. ***Capacity (Continuous)***

The *capacity (continuous)* of the Gaussian channel with power constraint P is

$$C = \max_{f_x(\cdot), \mathbb{E}\cdot x^2 \le P} I(X;Y)$$

where

$$\begin{aligned}
I(X;Y) &= h(Y) - h(Y|X) = h(Y) - h(\underbrace{Y - X}_{Z}|X) \\
&= h(Y) - h(Z|X) \\
&= h(Y) - h(Z) \\
h(Z) &= \frac{1}{2}\log(2\pi e N) \\
\mathbb{E}Y^2 &= \mathbb{E}(X+Z)^2 = \mathbb{E}X^2 + 2\cancel{\mathbb{E}(XZ)} + \underbrace{\mathbb{E}Z^2}_{N} \le P + N \\
I(X;Y) &\le \frac{1}{2}\log(2\pi e(P+N)) - \frac{1}{2}\log(2\pi e N) \\
&\le \frac{1}{2}\log\left(\frac{P+N}{N}\right) \\
&= \frac{1}{2}\log\left(1 + \frac{P}{N}\right)
\end{aligned}$$

Thus,

$$\begin{aligned}
C &= \max_{f_x, \mathbb{E}X^2 \le P} I(X;Y) \\
&= \frac{1}{2}\log\left(1 + \frac{P}{N}\right)
\end{aligned}$$

Definition 9.5.

An (M, n) code for the Gaussian channel with power constraint P consists of

- An encoding function $x : \{1, 2, \ldots, M\} \to \mathbb{R}^n$ yielding codewords $X^n(1), X^n(2), \ldots, X^n(M)$ satisfying the power constraint P, i.e. for every $x^n(w) = (x_1(w), \ldots, x_n(w))$,

$$\frac{1}{n}\sum_{i=1}^{n} x_1^2(w) \le P, \qquad w = 1, 2, \ldots, M$$

- A decoding function $g : \mathbb{R}^n \to \{1, 2, \ldots, M\}$. The *rate* of the code is

$$R = \frac{\log M}{n} \text{ bits per transmission}$$

The *probability of error* given message W is

$$\lambda_w = \Pr\left\{g(Y^n) \neq W \;\middle|\; X^n = X^n(w)\right\}$$

The *average probability of error* is

$$P_e(n) = \frac{1}{n}\sum_{w=1}^{M} \lambda_w$$

The *maximum probability of error* is

$$\lambda^{(n)} = \max_{w=1,2,\ldots,M} \lambda_w$$

Definition 9.6. ***Achievable***

The rate R is *achievable* if there exists a sequence of $(2^{nR}, n)$ codes such that

$$\lambda^{(n)} \xrightarrow{n\to\infty} 0$$

Theorem 9.7. ***Capacity of a Gaussian Channel***

The capacity of a Gaussian channel with power constraint P and noise variance N is:

$$C = \frac{1}{2}\log\left(1 + \frac{P}{N}\right) \text{ bits per transmission}$$

Proof. (Achievability)

Given $\epsilon > 0$, we have the jointly typical set $A_\epsilon^{(n)}$ with respect to the density of $f(x,y)$:

$$A_\epsilon^{(n)} = \Big\{(x^n, y^n) \in \mathbb{R}^n \times \mathbb{R}^n \ : \ \left|-\frac{1}{n}\log f_{X^n}(x^n) - h(X)\right| < \epsilon \quad \left|-\frac{1}{n}\log f_{Y^n}(y^n) - h(Y)\right| < \epsilon \quad \left|-\frac{1}{n}\log f_{X^n,Y^n}(x^n,y^n) - h(X,Y)\right| < \epsilon\Big\}$$

where $f_{X^n,Y^n}(x^n,y^n) = \prod_{i=1}^{n} f(x_i,y_i)$.

Let $\mathcal{C}$ be a $(2^{nR}, n)$ code, and $X^n(W) = (X_1(W), \ldots, X_n(W))$ be the codeword corresponding to message W. If Y is received and there is a unique W^* for which $(X^n(W^*), Y^n) \in A_\epsilon^{(n)}$, then the decoder's estimate is W^*. An error occurs if:

- $X^n(W)$ does not satisfy the power constraint P
- $(X^n(W), Y^n)$ is not jointly typical
- $(X^n(W^*), Y^n)$ is jointly typical and $W^* \neq W$

We define the events

$$E_0 = \left\{\frac{1}{n}\sum_{i=1}^{n} x_i^2(1) > P\right\}$$
$$E_W = \left\{(X^n(W), Y^n) \in A_\epsilon^{(n)}\right\}$$

Thus, the average probability of error is

$$P_e = \Pr\ \left\{E_0 \cup E_1^C \cup E_2 \cup \cdots \cup E_{2^{nR}}\right\}$$

By the Law of Large Numbers, for large n we have that

$$P(E_0) \leq \epsilon$$

where $X_1^2(1), X_2^2(1), \ldots, X_n^2(1)$ are i.i.d. with mean $P - \epsilon$ if we choose $X_i(W) \sim \mathcal{N}(0, P-\epsilon)$. By property (1) of $A_\epsilon^{(n)}$, we have that $\Pr\ \{E_1^C\} \leq \epsilon$ for large n. ($\Pr\ \{E_1\} \geq 1-\epsilon$, Theorem 7.69.) By property (2) of $A_\epsilon^{(n)}$,

$$P(E_W) \leq 2^{-n[I(X;Y)-3\epsilon]}, \qquad w \geq 2$$

Thus,

$$\begin{aligned} P_e^{(n)} &\leq \epsilon + \epsilon + \sum_{w=2}^{2^{nR}} 2^{-n[I(X;Y)-3\epsilon]} \\ &\leq 2\epsilon + (2^{nR}-1)2^{-n[I(X;Y)-3\epsilon] \to -n[I(X;Y)-R-3\epsilon]} \\ &\leq 2\epsilon + (2^{nR}-1)2^{-n[I(X;Y)-R-3\epsilon]} \end{aligned}$$

This probability will go to zero if

$$\begin{aligned} -(R+3\epsilon) + I(X;Y) &> 0 \\ R &< I(X;Y) - 3\epsilon \\ R &< I(X;Y) \end{aligned}$$

Thus, $R < I(X;Y) \ \Rightarrow \ P_e^{(n)} \to 0$.

To show that the maximum probability of error, we use the "throw half of the codes away" trick that we have used in the past. □

9.2 5-25-11

Continuing from last time, we want to prove that if $R > C$ then $P_e^{(n)} \not\to 0$. Equivalently, we want to prove that $P_e^{(n)} \to 0$ implies that $R \leq C$.

Proof. Assume that we have a $(2^{nR}, n)$ codebook that satisfies the power constraint:

$$\frac{1}{n}\sum_{i=1}^{n} x_i^2(u) \leq P \;\forall\; w$$

Our scheme looks like:

$$W \to X^n(W) \to Y^n(W) \to \hat{W}$$

Fano's Inequality gives us that

$$H(W|\hat{W}) \leq 1 + nRP_e^{(n)} = n\epsilon_n$$

where $\epsilon_n \to 0$ because $P_e^{(n)} \to 0$.

$$\begin{aligned}
nR = H(W) &= I(W;\hat{W}) + H(W|\hat{W}) \\
&\leq I(W;\hat{W}) + n\epsilon_n \\
&\leq I(W;Y^n) + n\epsilon_n \\
&\leq I(X^n;Y^n) + n\epsilon_n \\
&= h(Y^n) - h(Y^n|X^n) + n\epsilon_n \\
&= h(Y^n) - h(Z^n) + n\epsilon_n \\
&\leq \sum_{i=1}^{n}(h(Y_i) - h(Z_i)) + n\epsilon_n
\end{aligned}$$

We have that

$$P_i = \mathbb{E}x_i^2 = \frac{1}{2^{nR}}\sum_{w=1}^{2^{nR}} x_i^2(w)$$

Also,

$$\frac{1}{n}\sum P_i \leq P$$

We compute the expectation value of Y_i^2:

$$\begin{aligned}
\mathbb{E}Y_i^2 &= \underbrace{\mathbb{E}X_i^2}_{\to P_i} + \cancel{2\mathbb{E}X_iZ_i} + \underbrace{\epsilon Z^2}_{\to N} \\
&= P_i + N \\
nR &\leq \sum_{i=1}^{n}\left(\frac{1}{2}\log\left(1+\frac{P_i}{N}\right)\right) + n\epsilon_n \\
R &\leq \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{2}\log\left(1+\frac{P_i}{N}\right)\right) + \epsilon_n \qquad (9.1)
\end{aligned}$$

The power constraint is that:

$$\begin{aligned}
\mathbb{E}_i X^2 &< P \;\forall\; W \\
\mathbb{E}_W \mathbb{E}_i X^2 &< P \\
\mathbb{E}_i \underbrace{E_W X^2}_{P_i} &< P
\end{aligned}$$

Continuing from (9.1), we have

$$R \leq \frac{1}{2}\log\left(1+\frac{1}{n}\sum_{i=1}^{n}\frac{P_i}{N}\right)+\epsilon_n$$
$$\leq \underbrace{\frac{1}{2}\log\left(1+\frac{P}{N}\right)}_{C}+\epsilon_n$$

Thus, $R \leq C+\epsilon_n$. Therefore, if $\epsilon_n \to 0$ then $R \leq C$. □

9.2.1 Shannon Limit for Gaussian Channel

Definition 9.8. ***SNR for a Code Symbol***

$$\frac{P}{2N} \triangleq \text{SNR for a Code Symbol}$$
$$\gamma_G(R) = \frac{P}{2NR} = \text{Source-bit SNR}$$

Remark 9.9.

For reliable communication, we know that

$$R \leq C = \frac{1}{2}\log\left(1+\frac{P}{N}\right)$$
$$= \frac{1}{2}\log\left(1+2R\gamma_G\right)$$
$$R \leq \frac{1}{2}\log\left(1+2R\gamma_G\right)$$
$$\gamma_G \geq \frac{2^{2R}-1}{2R}$$

9.2.2 Parallel Gaussian Channels

Remark 9.10.

$$
\begin{aligned}
Y_j &= X_j + Z_j, \qquad j = 1, 2, \ldots, k, \qquad Z_j \sim \mathcal{N}(0, N_j) \\
\mathbb{E}\sum_{j=1}^{k} X_j^2 &\leq P \\
C &= \max_{f(\cdot)\mathbb{E}X^2 \leq P} I(X_1, \ldots, X_k; Y_1, \ldots, Y_k) \\
&= h(Y_1, \ldots, Y_k) - h(Y_1, \ldots, Y_k | X_1, \ldots, X_k) \\
&= h(Y_1, \ldots, Y_k) - h(Z_1, \ldots, Z_k) \\
&\leq \sum_{i=1}^{k} h(Y_i) - h(Z_i) \\
&\leq \sum_{i=1}^{k} \frac{1}{2} \log\left(1 + \frac{P_i}{N_i}\right)
\end{aligned}
$$

where $P_i = \mathbb{E}X_i^2$ and $\sum_{i=1}^{k} P_i \leq P$ (power constraint). For the optimization problem, Lagrangian multipliers give us

$$
\begin{aligned}
J(P_1, \ldots, P_k) &= \sum_{i=1}^{k} \frac{1}{2} \log\left(1 + \frac{P_i}{N_i}\right) + \lambda\left(\sum_{i=1}^{k} P_i - P\right) \\
\frac{1}{2}\frac{1}{P_i + N_i} + \lambda = 0 \\
P_i &= \nu - N_i
\end{aligned}
$$

This is sometimes referred to as *water-filling*.

Definition 9.11. ***Kuhn-Tucker Conditions***

The *Kuhn-Tucker conditions* can be used to verify that

$$P_i = (\nu \cdot N_i)^+$$

is the solution that maximizes capacity (where the superscript "+" denotes nonnegative), with ν chosen so that

$$\sum_{i=1}^{k} (\nu - N_i)^+ = P.$$

This means that we favor channels with lower noise (see Figure 9.4 on page 277 (303)).

Remark 9.12.

Consider the following optimization problem: maximize $f(\mathbf{x})$ subject to $g_j(\mathbf{x}) \leq 0,\ j = 1, \ldots, k$, where $f : \mathbb{R}^n \to \mathbb{R}$ is concave and $g_j : \mathbb{R}^n \to \mathbb{R}$ is convex.

Theorem 9.13. ***The Lagrangian***

$$L(\mathbf{x}) = f(\mathbf{x}) - \sum_{j=1}^{k} \lambda_j g_j(\mathbf{x})$$

Let x^* be a feasible point (satisfies the constraint g). Suppose $\lambda_1, \ldots, \lambda_k$:

$$\nabla L(x^*) = 0$$

$\lambda_j \geq 0\ \forall\ j$ and $\lambda_j = 0$ if $g_j(x^*) < 0$. Then x^* solves the maximization problem.

Lemma 9.14.

If $f : \mathbb{R}^n \to \mathbb{R}$ is concave and $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^n$, then

$$f(\mathbf{x}) \leq f(\mathbf{y}) + \nabla f(\mathbf{y})(\mathbf{x} - \mathbf{y})^T$$

For a convex function g, we have

$$g(\mathbf{x}) \geq g(\mathbf{y}) + \nabla g(\mathbf{y})(\mathbf{x} - \mathbf{y})^T$$

Proof. (of Theorem 9.13)
Assume $\mathbf{x}$ is a feasible point, i.e. $g(\mathbf{x}) \leq 0\ \forall\ j$. Then from Lemma 9.14,

$$\begin{aligned}
f(\mathbf{x}) &\leq f(\mathbf{x}^*) + \nabla f(\mathbf{x}^*)(\mathbf{x} - \mathbf{x}^*)^T \\
g_j(\mathbf{x}) &\geq g_j(\mathbf{x}^*) + \nabla g(\mathbf{x}^*)(\mathbf{x} - \mathbf{x}^*)^T \\
L(\mathbf{x}^*) &= f(\mathbf{x}) - \sum \lambda_j g_j(\mathbf{x}^*) \\
\nabla L(\mathbf{x}^*) &= \mathbf{0} \\
\nabla f(\mathbf{x}^*) &= \sum \lambda_j \nabla g_j(\mathbf{x}^*) \\
f(\mathbf{x}) &\leq f(\mathbf{x}^*) + \nabla f(\mathbf{x}^*)(\mathbf{x} - \mathbf{x}^*)^T \\
&\leq f(\mathbf{x}^*) + \sum \lambda_j (g_j(\mathbf{x}) - g_j(\mathbf{x}^*)) \\
&\leq f(\mathbf{x}^*) - \sum \underbrace{\lambda_j}_{\searrow 0} g_j(\mathbf{x}^*) \leq f(\mathbf{x}^*)
\end{aligned}$$

□

Remark 9.15.

$$
\begin{aligned}
f(\mathbf{P}) &= \frac{1}{2}\sum \log\left(1+\frac{P_i}{N}\right) \\
g_0(\mathbf{P}) &= \sum P_j - P \leq 0 \\
g_j(\mathbf{P}) &= -P_j \leq 0, \quad j = 1, \ldots, k
\end{aligned}
$$

9.3 6-1-11

Remark 9.16. ***Course & Final Info***

We can pick up the homework on Friday outside her office.

Office hours Tuesday 5-6.

2.5 standard problems (capacity, entropy, Huffman code, etc.), 1.5 tricky problems.

Remark 9.17. ***Review of the Gaussian System***

$$Y = X + Z, \qquad Z \sim \mathcal{N}(0, N)$$

For the problem to be well-posed, we have the constraint

$$\mathbb{E}[X^2] \le P$$

We know that the capacity is

$$C = \frac{1}{2} \log \left(1 + \frac{P}{N}\right)$$

$\frac{P}{N}$ = SNR = Signal to Noise Ratio

Remark 9.18. ***Review of Parallel Gaussian Channels***

We have k independent channels:

$$Y_1 = X_1 + Z_1, \ \cdots, \ Y_k = X_k + Z_k, \qquad Z_i \sim \mathcal{N}(0, N_i)$$

The power constraint here is

$$\mathbb{E} \sum_{i=1}^{k} X_i^2 \leq P$$

For any given power allocation $P_1, \dots, P_k$ with $P_1 + \cdots + P_k = P$, then

$$C(P_1, \dots, P_k) = \sum_{i=1}^{k} \frac{1}{2} \log \left(1 + \frac{P_i}{N_i}\right)$$

We want to maximize $C(P_1, \dots, P_k)$ subject to the constraint $\sum P_i \leq P$. We can do this with Lagrange multipliers:

$$\begin{aligned}
J(P_1, \dots, P_k) &= \sum_{i=1}^{k} \frac{1}{2} \log \left(1 + \frac{P_i}{N_i}\right) + \lambda \sum_{i=1}^{k} P_i \\
\frac{\partial J}{\partial P_i} &= 0 \\
0 &= \frac{1}{2} \cdot \frac{1}{P_i + N_i} + \lambda \\
P_i + N_i &= \nu \\
P_i &= (\nu - N_i)^+
\end{aligned}$$

Definition 9.19. ***Bandlimited Channel***

A *bandlimited channel* cuts out all frequencies greater than its *bandwidth*, W.

$$\underbrace{X(t)}_{P\text{ Watts}} \rightarrow \overbrace{\oplus}^{Z(t)} \rightarrow \underbrace{H(f)}_{\substack{\text{bandpass}\\\text{filter}}} \rightarrow Y(t)$$

We can model the bandpass filter as a convolution with $h(t)$, giving us:

$$\underbrace{Y(t)}_{\substack{\text{bandlimited}\\\text{time-limited in }T}} = (X(t)+Z(t)) * h(t) = \underbrace{X(t) * h(t)}_{\substack{\text{bandlimited}\\\text{time-limited in }T}} + \underbrace{Z(t) * h(t)}_{\substack{\text{bandlimited}\\\text{time-limited in }T}}$$

We can convert this to a discrete signal with $2WT$ samples (Nyquist). Thus, we have

$$Y_i = X_i + N_i$$

$$\frac{1}{2}\log\left(1 + \frac{P_{\text{sample}}}{N_{\text{sample}}}\right)$$

where

$$\begin{aligned}
P_{\text{sample}} &= \frac{PT}{2TW} = \frac{P}{2W}\\
N_{\text{sample}} &= \frac{N_0 WT}{2TW} = \frac{N_0}{2}\\
\text{power spectral density} &\triangleq \frac{N_0}{2}\text{ watts/hertz}\\
\text{bandwidth} &\triangleq W\text{ hertz}
\end{aligned}$$

So the capacity of a bandlimited channel is

$$\begin{aligned}
C &= \frac{P}{N_0}\frac{WN_0}{P}\log\left(1 + \frac{P}{N_0 W}\right)\\
&= W\log\left(1 + \frac{P}{N_0 W}\right)\text{ bits/second}
\end{aligned}$$

Index