Electrical Engineering 229A Lecture Notes
Information Theory and Coding

Professor: Venkat Anantharam
Scribe: Daniel Raban

Contents

1 Introduction to Shannon Entropy

1.1 Shannon entropy

Information theory is unusual in that it originated from the work of one person, Claude Elwood Shannon, in the late 1950s.[1] Shannon's idea was how to numerically measure the "amount of (statistical) uncertainty" inherent in a probabilistic experiment.

Example 1.1 (Coin flipping)**.** The "uncertainty" in $(1/2, 1/2)$ is "more" than in $(3/4, 1/4)$, which is "more" than in $(99/100, 1/100)$.

Shannon developed a calculus to work with such quantities. This notion is called *entropy*.

Definition 1.1. Consider a probability distribution $(p(1), \ldots, p(d))$ on $\{1, \ldots, d\}$. The **Shannon entropy** of p is

$$H(p) = -\sum_{i=1}^{d} p(i) \log p(i).$$

Here, the log is base 2, which was Shannon's convention and the convention for engineers. In mathematics and statistical mechanics, the natural logarithm is used. We take the convention that $0 \log 0 = 0$ (which is $\lim_{x \downarrow 0} x \log x$).

Example 1.2. Note that

$$H\left(\frac{1}{2}, \frac{1}{2}\right) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = \log 2 = 1.$$

This is a kind of normalization.

1.2 Motivation for the formula of entropy

To motivate the actual formula, consider $d = 2$ and n independent copies of $\{1, 2\}$-valued random variables with probability distribution p. For a sequence x^n of 1s and 2s,

$$\begin{aligned} p(x^n) &= \prod_{i=1}^{n} p(x_i) \\ &= p(1)^{N(1|x^n)} p(2)^{N(2|x^n)} \\ &= 2^{n(N(1|x^n)/n \log p(1) + N(2|x^n)/n \log p(2))}, \end{aligned}$$

[1] Shannon lived from 1916-2001. His master's thesis is also considered a landmark. It introduced the boolean circuit view of computing. There is a 2017 movie about Shannon called *The Bit Player* and a book called *A Mind at Play*.

where $N(i \mid x^n)$ is the number of times i appears in x^n. But by the strong law of large numbers, $\frac{N(i|x^n)}{n} \to p(1)$ almost surely as $n \to \infty$. So

$$p(x^n) \approx (2^{p(1)\log p(1)+p(2)\log p(2)})^n.$$

This suggests that $-p(1)\log p(1) - p(2)\log p(2)$ represents the "uncertainty" in one toss.

1.3 Expectation formulation of entropy

If X is a random variable taking values in $\{1,\dots,d\}$ with probability distribution p, i.e. $\mathbb{P}(X=i) = p(i)$ for $1 \le i \le d$, we write $H(X)$ for $H(p)$. With this notation,

$$H(X) = \sum_{i=1}^{d} \mathbb{P}(X=i) \log \frac{1}{\mathbb{P}(X=i)} = \mathbb{E}[\log 1/p(X)].$$

1.4 Concavity of Shannon entropy and entropy of uniform distributions

Fix $d \ge 2$. The set of probability distributions on $\{1,\dots,d\}$ is called the **unit d-simplex** in $\mathbb{R}^d$. We can write it as $\{(p(1),\dots,p(n)) : p(i) \ge 0, \sum_{i=1}^d p(i) = 1\}$. This is a **convex** set, and H can be viewed as a function on this set.

Proposition 1.1. *H is a **concave function** on the (unit) d-simplex for each fixed d. That is, for all $p_0, p_1 \in \{1,\dots,d\}$ and $\lambda \in [0,1]$, if p_λ denotes $\lambda p_1 + (1-\lambda)p_0$, then*

$$H(p_\lambda) \ge \lambda H(p_1) + (1-\lambda)H(p_0).$$

Proof. Because $H(p) = -\sum_{i=1}^d p(i)\log p(i)$, we want to check that $x\log x$ is convex. This is twice differentiable, so it suffices to show that the second derivative is ≥ 0. Write

$$\begin{aligned}(x\log x)'' &= (\log_2 e)(x\log_e x)'' \\ &= (\log_2 e)(\log_e x + 1)' \\ &= (\log_2 e)\frac{1}{x} \\ &\ge 0.\end{aligned}$$

□

Corollary 1.1. *The uniform distribution on $\{1,\dots,d\}$ has the largest entropy among probability distributions on $\{1,\dots,d\}$.*

Proof. Let S_d denote the set of permutations of $\{1,\dots,d\}$. Then

$$(1/d,\dots,1/d) = \frac{1}{d!}\sum_{\sigma\in S_d} (p(\sigma(1)), p(\sigma(2)), \dots, p(\sigma(d))),$$

so by the concavity of H,

$$\begin{aligned} H(1/d, \ldots, 1/d) &\geq \frac{1}{d!} \sum_{\sigma \in S_d} H(p(\sigma(1)), p(\sigma(2)), \ldots, p(\sigma(d))) \\ &= H(p). \end{aligned}$$

□

1.5 Conditional entropy

The entropy calculus starts with the definition of "conditional entropy." Given a pair of random variables (X, Y), we consider $H(X, Y) - H(Y)$ and denote this $H(X \mid Y)$. This is known as the **conditional entropy of X given Y**. Next time, we will consider the information $I(X; Y) := H(X) - H(X \mid Y)$ and see that this is actually symmetric in X and Y.

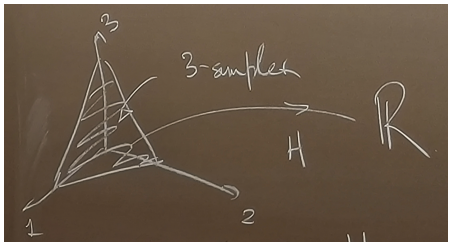
2 Entropic Quantities Relating Random Variables

2.1 The binary entropy function

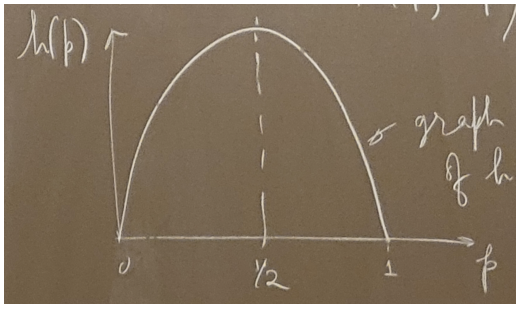
Suppose we have a probability distribution $p = (p_1, \ldots, p_d)$ on a finite set of $\mathscr{X}$ of size d, say $\mathscr{X} = \{1, \ldots, d\}$. We will use the notation $[d] = \{1, \ldots, d\}$. The function

$$H(p_1, \ldots, p_d) = -\sum_{j=1}^{d} p_i \log p_i$$

is called the entropy of the distribution p. Last lecture we saw that $H \geq 0$ and $H(p_1, \ldots, p_d) \leq H(1/d, \ldots, 1/d) = \log d$ as a consequence of the *concavity* of H as a function on the unit d-simplex. Concavity of H means that for $\lambda \in [0, 1]$, $H(\lambda p^{(1)} + (1-\lambda)p^{(0)}) \geq \lambda H(p^{(1)}) + (1-\lambda)H(p^{(0)})$.



Example 2.1. For $d = 2$, $H(p, 1-p) = -p \log p - (1-p)\log(1-p)$. We denote this as $h(p)$.



The function $h(p)$ is known as the **binary entropy function**. The graph is very steep near 0; all the derivatives approach ∞. $h(1/2) = 1$, and $h(p) = h(1-p)$. We can calculate

$$\begin{aligned} h'(p) &= \log_2 e(-\log_e p - 1 + \log_e(1-p) + 1) \\ &= \log \frac{1-p}{p}, \end{aligned}$$

which is $+\infty$ at $p = 0$ and $-\infty$ at $p = 1$. We can check

$$h''(p) = \log_2 e \left(-\frac{1}{1-p} - \frac{1}{p} \right),$$

which is $-\infty$ at $p = 0$ and $p = 1$.

2.2 Convexity and Jensen's inequality

Definition 2.1. A set $D \subseteq \mathbb{R}^n$ is **convex** if when $\lambda \in [0,1]$ and $x^{(0)}, x^{(1)} \in D$, $\lambda x^{(0)} + (1-\lambda)x^{(1)} \in D$, as well.

Definition 2.2. A function $f : D \to \mathbb{R}$ where $D \subseteq \mathbb{R}^n$ is a convex set is called a **convex function** if for all $\lambda \in [0,1]$ and $x^{(0)}, x^{(1)} \in D$, we have

$$f(\lambda x^{(1)} + (1-\lambda)x^{(0)}) \leq \lambda f(x^{(1)}) + (1-\lambda) f(x^{(0)}).$$

This implies that if for any $m \geq 1$, $x^{(1)}, x^{(2)}, \ldots, x^{(m)} \in D$ and any probability distribution $(\lambda_1, \ldots, \lambda_m)$ on $[m]$, we have

$$f\left(\sum_{i=1}^{m} \lambda_i x^{(i)}\right) \leq \sum_{i=1}^{m} \lambda_i f(x^{(i)}).$$

More generally, we have the following:

Theorem 2.1 (Jensen's inequality)**.** *For any random variable Z taking values in a convex set $D \subseteq \mathbb{R}^n$,*

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

2.3 Joint and conditional entropy

If X is a random variable taking values in $[d]$, we write $H(X)$ for $H(p_1, \ldots, p_d)$, where $p_i := \mathbb{P}(X = i)$. If X takes values in $\mathscr{X}$, then $H(X)$ denotes $H(p(x), x \in \mathscr{X})$, where $p(x) := \mathbb{P}(X = x)$. Now suppose X takes values in $\mathscr{X}$ and Y takes values in $\mathscr{Y}$, where $\mathscr{X}, \mathscr{Y}$ are finite sets. They have a joint probability distribution $(p(x,y), (x,y) \in \mathscr{X} \times \mathscr{Y})$.

Definition 2.3. The **joint entropy** of the pair (X, Y), which is just a random variable taking values in $\mathscr{X} \times \mathscr{Y}$, is denoted $H(X,Y)$ and equals

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

Definition 2.4. The difference $H(X,Y) - H(X)$, denoted $H(Y \mid X)$, is called the **conditional entropy** of Y given X.

Recall that the entropy is $H(X) = \mathbb{E}[\log 1/p(X)]$. The joint entropy can be written similarly:

$$H(X,Y) = \mathbb{E}\left[\log \frac{1}{p(X,Y)}\right].$$

We can also write the conditional entropy as

$$\begin{aligned} H(Y \mid X) &= \mathbb{E}\left[\log \frac{1}{p(Y \mid X)}\right] \\ &= \sum_{x,y} p(x,y) \log \frac{1}{p(y \mid x)} \\ &= \sum_{x} p(x) \sum_{y} p(y \mid x) \log \frac{1}{p(y \mid x)}. \end{aligned}$$

For each fixed $x \in \mathscr{X}$, $\sum_y p(y \mid x) \log \frac{1}{p(y|x)}$ is denoted $H(Y \mid X = x)$. It is the entropy of the conditional distribution of Y given that $X = x$. With this notation,

$$H(Y \mid X) = \sum_{x} p(x) H(Y \mid X = x).$$

Remark 2.1. This notation is not consistent with the rest of probability notation. $H(Y \mid X)$ is a number, rather than a random variable. This notation is widespread in information theory, however, because it was introduced by Shannon himself.

From this formula, we can see that $H(Y \mid X) \geq 0$.

2.4 Mutual information

We might hope that we "learn" about Y from observing X, i.e. the uncertainty in Y is reduced. That is, we hope that $H(Y) \geq H(Y \mid X)$. This is true.

Definition 2.5. $H(Y) - H(Y \mid X)$ is denoted $I(X;Y)$ (or sometimes denoted as $I(X \wedge Y)$) and is called the **mutual information** between X and Y.

We have

$$\begin{aligned} I(X;Y) &= \mathbb{E}\left[\log \frac{1}{p(Y)}\right] - \mathbb{E}\left[\log \frac{1}{p(Y \mid X)}\right] \\ &= \mathbb{E}\left[\log \frac{p(X,Y)}{p(X)p(Y)}\right] \\ &= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}. \end{aligned}$$

This is symmetric when X and Y are interchanged. That is, $I(X;Y) = I(Y;X)$.

2.5 Relative entropy

$I(X,Y) \geq 0$ because it is a relative entropy.

Definition 2.6. Given two probability distributions $(p(z), z \in \mathscr{Z})$ and $(q(z), z \in \mathscr{Z})$, we write

$$D(p \,||\, q) = \sum_{z \in \mathscr{Z}} p(z) \log \frac{p(z)}{q(z)},$$

which is called the **relative entropy** of p with respect to q. It is also called the **information distance/divergence** of p from q or the **Kullback-Leibler divergence**.

Remark 2.2. The relative entropy is *not* a distance; it is not symmetric in p and q and does not satisfy the triangle inequality.

We want to show that $D(p \,||\, q) \geq 0$. Note that

$$I(X;Y) = D(p(x,y) \,||\, p(x)p(y)),$$

where $p(x,y)$ is the joint distribution of (X,Y) and $p(x)p(y)$ is the distribution of $(\widetilde{X}, \widetilde{Y})$, where $\widetilde{X} \overset{d}{=} X$, $\widetilde{Y} \overset{d}{=} Y$, and $\widetilde{X}, \widetilde{Y}$ are independent. So we will get $I(X;Y) \geq 0$ if we can prove $D(p \,||\, q) \geq 0$ in general.

The relative entropy is a natural statistical quantity that measures how far p is from q. So the conceptual meaning of $I(X;Y)$ is that it measures how far apart the joint distribution of (X,Y) is from being a product distribution of independent X, Y.

Proposition 2.1. $D(p \,||\, q) \geq 0$.

Proof. Write

$$\begin{aligned} D(p \,||\, q) &= \sum_{z \in \mathscr{Z}} q(z) \frac{p(z)}{q(z)} \log \frac{p(z)}{q(z)} \\ &= \sum_{z \in \mathscr{Z}} q(z) \phi\left(\frac{p(z)}{q(z)}\right), \end{aligned}$$

where $\phi : \mathbb{R}_+ \to \mathbb{R}$ is given by $\phi(u) = u \log u$, which is convex (checked below). Using Jensen's inequality,

$$\begin{aligned} &\geq \phi\left(\sum_{z \in \mathscr{Z}} q(z) \frac{p(z)}{q(z)}\right) \\ &= \phi(1) \\ &= 0. \end{aligned}$$

To check that ϕ is convex, we have $\phi'(u) = \log_2 e(\log_e u + 1)$, so $\phi''(u) = \log_2 e \cdot \frac{1}{u} \geq 0$. □

Corollary 2.1. $I(X;Y) \geq 0$.

3 Entropy Over Countable Alphabets and Features of Conditional Entropy

3.1 Entropy of distributions over countable sets

Let's adjust our definitions to allow for distributions over countable sets. Let X be a random variable taking values in $\mathscr{X}$, a finite or countably infinite set, and let $(p(x), x \in \mathscr{X})$ be its probability distribution. Its **entropy** is

$$H(X) = H((p(x), x \in \mathscr{X})) = -\sum_x p(x) \log p(x).$$

This is well-defined, even if $\mathscr{X}$ is countably infinite, because all the terms have the same sign.

Remark 3.1. In general, to define $\sum_{x\in\mathscr{X}} a(x)$, where $\mathscr{X}$ is countably infinite, define it to be $(\sum_{x\in\mathscr{X}} a^+(x)) - (\sum_{x\in\mathscr{X}} a^-(x))$, where $a^+(x) := \max(a(x), 0)$ and $a^-(x) := \max(-a(x), 0)$. This definition makes sense when at least one of $\sum_{x\in\mathscr{X}} a^+(x), \sum_{x\in\mathscr{X}} a^-(x)$ is finite.

To avoid subtracting infinities when dealing with entropies over countable sets, proceed as follows: Given a pair of random variables X, Y taking values taking values in (finite or countably infinite) $\mathscr{X}, \mathscr{Y}$, respectively, for each $y \in \mathscr{Y}$, define $H(X \mid Y = y)$ to be the entropy of the conditional distribution of X given $Y = y$:

$$H(X \mid Y = y) = -\sum_{x\in\mathscr{X}} p(x \mid y) \log p(x \mid y).$$

We can alternatively express

$$H(X) = \mathbb{E}\left[\log \frac{1}{p(X)}\right], \qquad \mathbb{E}\left[\log \frac{1}{p(X \mid Y)} \mid Y = y\right],$$

as before.

Define the **conditional entropy** of X given Y to be $\sum_y p(y) H(X \mid Y = y)$, denoted $H(X \mid Y)$. So

$$H(X \mid Y) = \mathbb{E}\left[\log \frac{1}{p(X \mid Y)}\right].$$

Now $H(X, Y) = H(Y) + H(X \mid Y)$ becomes a theorem, called the chain rule for entropy.

Theorem 3.1 (Chain rule)**.**

$$H(X, Y) = H(Y) + H(X \mid Y).$$

Proof.

$$\mathbb{E}[\log \frac{1}{p(X,Y)}] = \mathbb{E}\left[\log \frac{1}{p(Y)}\right] + \mathbb{E}\left[\log \frac{1}{p(X \mid Y)}\right]. \qquad \square$$

We define $D(p \mid\mid q)$ for $(p(x), x \in \mathscr{X})$, $(q(x), x \in \mathscr{X})$ as

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

To see that this is well-defined, observe that

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

Then this is well-defined because the function $u \mapsto u \log u$ defined on $\mathbb{R}^+$ is bounded below.

Then, we can define $I(X;Y) := D(p(x,y) \mid\mid p(x)p(y))$, and our previous definition for mutual information becomes a theorem:

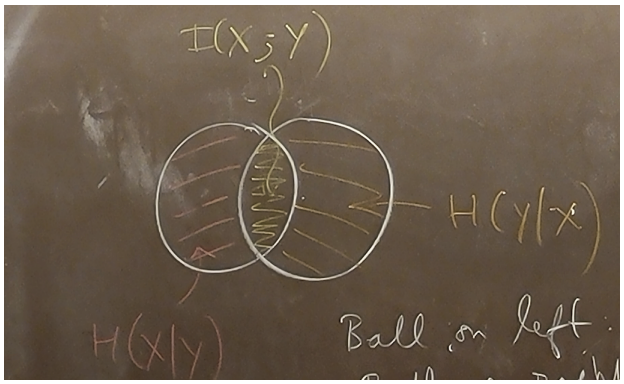
Theorem 3.2.

$$H(X) = I(X,Y) + H(X \mid Y).$$

Proof.

$$\mathbb{E}\left[\log \frac{1}{p(X)}\right] = \mathbb{E}\left[\log \frac{p(X,Y)}{p(X)p(Y)}\right] + \mathbb{E}\left[\log \frac{1}{p(X \mid Y)}\right]. \qquad \square$$

These "theorems" or (X,Y) can be schematically visualized via a Venn diagram.



3.2 Relationship between mutual information and independence

It is important to recognize that the condition for $I(X;Y) = 0$ is $p(x,y) = p(x)p(y)$ for all x, y, i.e. X, Y are independent (denoted $X \amalg Y$). Since $I(X;Y) = H(X) + H(Y) - H(X,Y)$ (inclusion-exclusion),

$$X \amalg Y \iff H(X,Y) = H(X) + H(Y).$$

3.3 General form of the chain rule

If we apply the chain rule twice, we get

$$\begin{aligned} H(X_1, X_2, X_3) &= H(X_1 \mid X_2, X_3) + H(X_2, X_3) \\ &= H(X_1 \mid X_2, X_3) + H(X_2 \mid X_3) + H(X_3). \end{aligned}$$

Similarly, using the notation X_1^n to denote $(X_1, \ldots, X_n)$, we get the general chain rule:

Theorem 3.3 (Chain rule, general form)**.**

$$H(X_1, \ldots, X_n) = H(X_1) + H(X_2 \mid X_1) + H(X_3 \mid X_1, X_2) + \cdots + H(X_n \mid X_1^{n-1}).$$

Example 3.1. Consider an urn[2] with 3 balls, two white and 1 red. Pull out all 3 balls in a random order. Let X_1 be the color of the first ball, let X_2 be the color of the second ball, and let X_3 be the color of the third ball. Then

$$H(X_1) = H(X_2) = H(X_3) = \frac{1}{3}\log 3 + \frac{2}{3}\log\frac{3}{2} = \log 3 - \frac{2}{3}.$$

We can also calculate the conditional entropies:

$$\begin{aligned} H(X_2 \mid X_1) &= \mathbb{P}(X_1 = \text{red})H(X_2 \mid X_1 = \text{red}) + \mathbb{P}(X_1 = \text{white})H(X_2 \mid X_1 = \text{white}) \\ &= \frac{2}{3}\log 2 \\ &= \frac{2}{3}. \end{aligned}$$

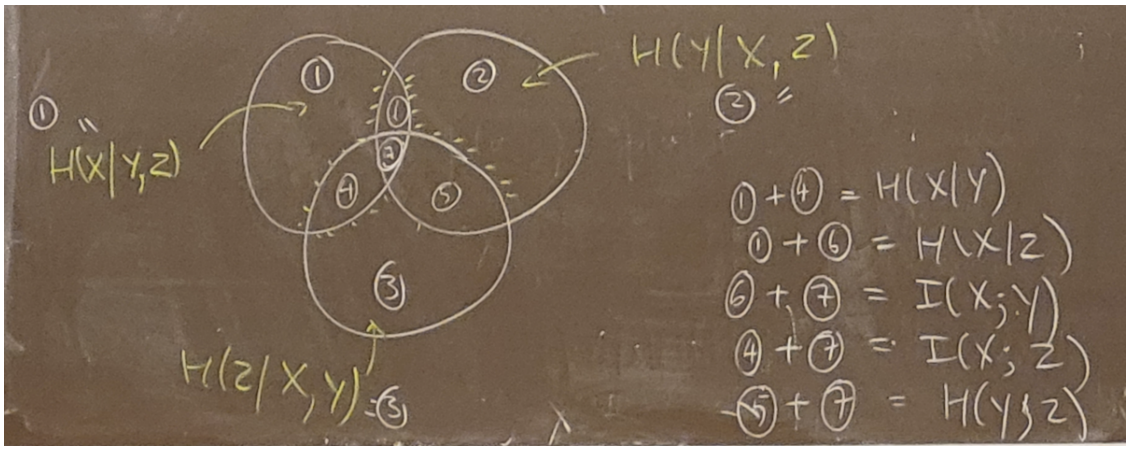
On the other hand, $H(X_3 \mid X_1, X_2) = 0$ because X_3 is determined by X_1, X_2. So the chain rule gives

$$\begin{aligned} H(X_1, X_2, X_3) &= H(X_1) + H(X_2 \mid X_1) + H(X_3 \mid X_1, X_2) \\ &= \log 3 - \frac{2}{3} + \frac{2}{3} + 0 \\ &= \log 3. \end{aligned}$$

[2] No one in the 21st century has ever seen an urn.

3.4 Problems with intuiting mutual information

Here is the Venn diagram for (X_1, X_2, X_3):



What does region 6 represent? This could be $I(X;Y \mid Z)$, the conditional relative entropy between the joint distribution (X, Y), conditioned on Z and the product distribution with the corresponding marginals, conditioned on Z. That is, region 6 is

$$H(X \mid Z) - H(X \mid Y, Z).$$

What does region 7 represent? This region is

$$I(X;Y) - I(X;Y \mid Z).$$

Here is a big problem, not for the math but for any hope of intuition: This can be *negative*. In particular, this says that in the presence of Z, Y can tell you more about X than it does alone.

Example 3.2. Let $X \amalg Y$, with $X \in \{1, -1\}$, $Y \in \{1, -1\}$, $\mathbb{P}(X = 1) = 1/2$, and $\mathbb{P}(Y = 1) = 1/2$. Let $Z = XY$ so $Z \in \{1, -1\}$ with $\mathbb{P}(Z = 1) = 1/2$. Then $Y \amalg Z$ and $X \amalg Z$, but X, Y, Z are not mutually independent. Since $X \amalg Y$, we have $I(X;Y) = 0$. However,

$$\begin{aligned} I(X;Y \mid Z) &= \mathbb{P}(Z = 1) I(X;Y \mid Z = 1) + \mathbb{P}(Z = -1) I(X;Y \mid Z = -1) \\ &= \mathbb{P}(Z = 1)(H(X \mid Z = 1) - H(X \mid Y, Z = 1)) \\ &\quad + \mathbb{P}(Z = -1)(H(X \mid Z = -1) - H(X \mid Y, Z = -1)) \end{aligned}$$

Since $X \amalg Z$, $H(X \mid Z = 1) = H(X \mid Z = -1) = H(X) = \log 2 = 1$. Also, $H(X \mid Y, Z = 1) = 0$ because $X = Y$ when $Z = 1$ and $H(X \mid Y, Z = 1) = 0$ because $X = -Y$ when $Z = -1$. So

$$= \frac{1}{2}(1 - 0) + \frac{1}{2}(1 - 0)$$

$$= 1.$$

This is strictly bigger than $I(X;Y)$.

Let's define $I(X;Y \mid Z)$ in a way that works for a countably infinite alphabet. We first define, given $p(x,y,z)$,

$$\sum_z p(z) D(p(x \mid z) \mid\mid p(y \mid z)),$$

denoted $D(p(x \mid z) \mid\mid p(y \mid z) \mid p(z))$ to be the conditional relative entropy of $p(x,z)$ with respect to $p(y,z)$ given z. Then $D(p(x,y \mid z) \mid\mid p(x \mid z)p(y \mid z) \mid p(z))$ would then be $I(X;Y \mid Z)$. That is,

$$\begin{aligned} I(X;Y \mid Z) &:= \sum_z p(z) \sum_{x,y} p(x,y \mid z) \log \frac{p(x,y \mid z)}{p(x \mid z)p(y \mid z)} \\ &= \mathbb{E}\left[\log \frac{p(X,Y \mid Z)}{p(X \mid Z)p(Y \mid Z)}\right] \\ &= H(X \mid Z) + H(Y \mid Z) - H(X,Y \mid Z). \end{aligned}$$

Then the chain rule gives

$$I(X;Y \mid Z) = H(X \mid Z) - H(X \mid Y,Z).$$

3.5 The chain rule for relative entropy

Theorem 3.4 (Chain rule for relative entropy)**.**

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

Proof.

$$\begin{aligned} D(p(x,y) \mid\mid q(x,y)) &= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{q(x,y)} \\ &= \mathbb{E}_p\left[\log \frac{p(X,Y)}{q(X,Y)}\right] \\ &= \mathbb{E}_p\left[\log \frac{p(X)}{q(X)}\right] + \mathbb{E}_p\left[\log \frac{p(Y \mid X)}{q(Y \mid X)}\right] \\ &= D(p(x) \mid\mid q(x)) + D(p(y \mid x) \mid\mid q(y \mid x) \mid p(x)). \end{aligned}$$

□

Similarly, there is a chain rule for mutual information

Theorem 3.5 (Chain rule for mutual information)**.**

$$I(X;Y_1,\ldots,Y_n) = I(X;Y_1) + I(X;Y_2 \mid Y_1) + \cdots + I(X;Y_n \mid Y_1^{n-1}).$$

4 Convexity of Relative Entropy and the Data Processing Inequality

4.1 Chain rules for entropy, relative entropy, and mutual information

The chain rule for entropy for two random variables says that

$$H(X_1, X_2) = H(X_1) + H(X_2 \mid X_1)$$

For n variables, we have

$$\begin{aligned} H(X_1^n) &= H(X_1^{n-1}, X_n) \\ &= H(X_1^{n-1}) + H(X_n \mid X_1^{n-1}) \\ &\vdots \\ &= H(X_1) + H(X_2 \mid X_1) + \cdots + H(X_n \mid X_1^{n-1}), \end{aligned}$$

which we can write as

$$= \sum_{\ell=1}^{n} H(X_\ell \mid X_1^{\ell-1}).$$

Here, the convention is that $X_1^{\ell-1}$ for $\ell = 1$ needs no conditioning.

This also comes from

$$\begin{aligned} H(X_1^n) &= \mathbb{E}\left[\log \frac{1}{\prod_{\ell=1}^n p(X_\ell \mid X_1^{\ell-1})}\right] \\ &= \sum_{\ell=1}^{n} \mathbb{E}\left[\log \frac{1}{p(X_\ell \mid X_1^{\ell-1})}\right] \\ &= \sum_{\ell=1}^{n} H(X_\ell \mid X_1^{\ell-1}). \end{aligned}$$

Similarly, we can obtain the chain rule for relative entropy from

$$\begin{aligned} D(p(x_1^n) \mid\mid q(x_1^n)) &= \mathbb{E}_p\left[\log \frac{p(X_1^n)}{q(X_1^n)}\right] \\ &= \mathbb{E}_p\left[\log \frac{\prod_{\ell=1}^n p(X_\ell \mid X_1^{\ell-1})}{\prod_{\ell=1}^n q(X_\ell \mid X_1^{\ell-1})}\right] \\ &= \sum_{\ell=1}^{n} \mathbb{E}_p\left[\log \frac{p(X_\ell \mid X_1^{\ell-1})}{p(X_\ell \mid X_1^{\ell-1})}\right] \end{aligned}$$

$$= \sum_{\ell=1}^{n} D(p(x_\ell \mid x_1^{\ell-1}) \mid\mid q(x_\ell \mid x_1^{\ell-1}) \mid p(x_1^{\ell-1})).$$

We can also obtain the chain rule for mutual information:

$$I(X; Y_1, Y_2) = I(X; Y_1) + I(X; Y_2 \mid Y_1).$$

This comes from

$$\begin{aligned}\mathbb{E}\left[\log \frac{p(X, Y_1, Y_2)}{p(X)p(Y_1, Y_2)}\right] &= \mathbb{E}\left[\frac{p(X, Y_1)}{p(X)p(Y_1)} \frac{p(X, Y_1, Y_2)p(Y_1)p(Y_1)}{p(Y_1)p(X, Y_1)p(Y_2, Y_1)}\right] \\ &= \mathbb{E}\left[\log \frac{p(X, Y_1)}{p(X)p(Y_1)} \frac{p(X, Y_2 \mid Y_1)}{p(X \mid Y_1)p(Y_2 \mid Y_1)}\right],\end{aligned}$$

More generally,

$$\begin{aligned}I(X; Y_1^n) &= I(X; Y_1^{n-1}, Y_n) \\ &= I(X; Y_1^{n-1}) + I(X; Y_n \mid Y_1^{n-1}) \\ &\vdots \\ &= I(X; Y_1) + I(X; Y_2 \mid Y_1) + \cdots + I(X; Y_n \mid Y_1^{n-1}),\end{aligned}$$

which we can write as

$$= \sum_{\ell=1}^{n} I(X; Y_\ell \mid Y_1^{\ell-1}).$$

4.2 Convexity of relative entropy and the log-sum inequality

An important property of relative entropy $D(p \mid\mid q)$ is that it is convex in the pair (p, q), where p denotes $(p(x), x \in \mathscr{X})$ and q denotes $(q(x), x \in \mathscr{X})$. That is for all $(p_0, q_0), (p_1, q_1)$ and $\lambda \in [0, 1]$, if we denote $p_\lambda = \lambda p_1 + (1 - \lambda)p_0$ and $q_\lambda = \lambda q_1 + (1 - \lambda)q_0$, then

$$D(p_\lambda \mid\mid q_\lambda) \leq \lambda D(p_1 \mid\mid q_1) + (1 - \lambda)D(p_0 \mid\mid q_0).$$

Remark 4.1. Note that $D(p \mid\mid q)$ can take the value $+\infty$.

This is a consequence of the **log-sum inequality**:

Lemma 4.1 (log-sum inequality)**.** *Suppose $a_i, b_i > 0$ for $i \in \mathscr{X}$.*

$$\sum_{i \in \mathscr{X}} a_i \log \frac{a_i}{b_i} \geq a \log \frac{a}{b},$$

where $a = \sum_{i \in \mathscr{X}} a_i$ and $b = \sum_{i \in \mathscr{X}} b_i$.

Proof. This comes from the convexity of $u \log u$ for $u \geq 0$. The left hand side is

$$\sum_{i \in \mathscr{X}} a_i \log \frac{a_i}{b_i} = b \sum_{i \in \mathscr{X}} \frac{b_i}{b} \left(\frac{a_i}{b_i} \log \frac{a_i}{b_i} \right)$$

Using Jensen's inequality,

$$\begin{aligned} &\geq b \left(\sum_i \frac{b_i}{b} \frac{a_i}{b_i} \right) \log \left(\sum_i \frac{b_i}{b} \frac{a_i}{b_i} \right) \\ &= a \log \frac{a}{b}. \end{aligned}$$

□

Corollary 4.1. *$D(p \,||\, q)$ is convex in the pair (p, q).*

Proof.

$$\begin{aligned} \lambda D(p_1 \,||\, q_1) + (1-\lambda) D(p_0 \,||\, q_0) &= \sum_x \lambda p_1(x) \log \frac{p_1(x)}{q_1(x)} + (1-\lambda) p_0(x) \log \frac{p_0(x)}{q_0(x)} \\ &= \sum_x \lambda p_1(x) \log \frac{\lambda p_1(x)}{\lambda q_1(x)} + (1-\lambda) p_0(x) \log \frac{(1-\lambda) p_0(x)}{(1-\lambda) q_0(x)} \end{aligned}$$

Using the log-sum inequality,

$$\begin{aligned} &\geq \sum_x (\lambda p_1(x) + (1-\lambda) p_0(x)) \log \frac{\lambda p_1(x) + (1-\lambda) p_0(x)}{\lambda q_1(x) + (1-\lambda) q_0(x)} \\ &= D(p_\lambda \,||\, q_\lambda). \end{aligned}$$

□

Remark 4.2. The inequality is still true if any of the terms $= +\infty$.

A good book on convex functions is the book by Rockafeller.

4.3 The data processing inequality

The data processing inequality says that if you are looking at the mutual information between X and Y and then you process Y in a way that does not use X, the mutual information can only decrease. How do we make this notion precise?

Definition 4.1. Given 3 random variables X, Y, Z, we write $Y - X - Z$ to indicate that Y and Z are conditionally independent given X. We may say that they form a **Markov chain** in this order. In probability notation, we may use the notation $Y \amalg_X Z$.

Recall that conditional independence says that $p(y, z \mid x) = p(y \mid x) p(z \mid x)$. Since

$$p(y, z \mid x) = p(y \mid x, z) p(z \mid x),$$

the assumed conditional independence gives

$$p(y \mid x, z) = p(y \mid x).$$

This argument can be run backwards, hence the "Markov" terminology.

Remark 4.3. Running the argument in the other direction gives $p(z \mid x, y) = p(z \mid x)$ if $Y - X - Z$.

Theorem 4.1 (Data processing inequality)**.** *Suppose $Y - X - Z$ form a Markov chain. Then*

$$I(Y; Z) \leq I(Y; X).$$

Proof. Use the chain rule in two different orders:

$$I(Y; X, Z) = I(Y; X) + I(Y; Z \mid X),$$

$$I(Y; X, Z) = I(Y; Z) + I(Y; X \mid Z).$$

Because $Y \amalg_X Z$, $I(Y; Z \mid X) = 0$. In fact, each $I(Y; Z \mid X = x)$ equals 0. So

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

as desired. □

Remark 4.4. The condition for equality is $I(Y; X \mid Z) = 0$, i.e. $Y \amalg_Z X$. This has interesting implications in statistics. Say we try to find an estimate for a random variable Θ (in a Bayesian framework) based on observations X. We might ask for some function $T(X)$ such that $\Theta - X - T(X)$. When is it true that $I(\Theta; T(X)) = I(\Theta; X)$? This happens precisely when $\Theta - T(X) - X$.

A typical example (not in a discrete context) is when Θ is the mean of the marginal, where each marginal is normal with variance 1. So conditioned on $\Theta = \theta$, each $X_i \sim N(0, 1)$ for $1 \leq i \leq n$. If $T(X) = \frac{1}{n} \sum_i X_i$, then $\Theta - T(X) - X$. By the data processing inequality, we should study $T(X)$ instead of X in a statistical context because it contains at least as much information as X in terms of estimating Θ.

5 Sufficient Statistics, Fano's Inequality, and the Asymptotic Equipartition Property

5.1 Sufficient statistics

Last time, we discussed the data processing inequality. Given $Y - X - Z$ (i.e. Y and Z are conditionally independent given X), the data processing inequality says that

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

The equality condition is when $I(X;Z \mid Y) = 0$, i.e. $X - Y - Z$.

We also discussed sufficient statistics. The idea is to think about learning about Θ by processing some observations X into $T(X)$, so $\Theta - X - T(X)$. Then $\Theta - T(X) - X$ if and only if $I(\Theta;X) = I(\Theta;T(X))$. Given $\Theta - T(X) - X$, we say that $T(X)$ is a **sufficient statistic** (for learning about Θ from X).

If $|\mathscr{X}| = d$, let $u(x) = \frac{1}{d}$ for $x \in \mathscr{X}$. Given $(p(x), x \in \mathscr{X})$, then

$$D(p \mid\mid u) = \sum_x p(x) \log \frac{p(x)}{u(x)} = \log d - H((p(x), x \in \mathscr{X})).$$

So it is difficult to define entropy in non-discrete settings. Regardless, here is a non-discrete example of a sufficient statistic.

Example 5.1. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\theta, 1)$ when $\Theta = \theta$, where $\Theta \in \{\theta_1, \ldots, \theta_d\}$ is a random variable. Then $\frac{1}{n}\sum_{i=1}^n X_i$ is a sufficient statistic for Θ. To check this, we need to show that $\Theta - \overline{X} - (X_1, \ldots, X_n)$, where $\overline{X} := \frac{1}{n}(X_1, \ldots, X_n)$. The conditional joint density is

$$\begin{aligned} f(x_1, \ldots, x_n \mid \theta) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-(x_i-\theta)^2/2} \\ &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-(x_i-\overline{x}+\overline{x}-\theta)^2/2} \\ &= \frac{1}{(2\pi)^{n/2}} e^{-\sum_{i=1}^n (x_i-\overline{x})^2} e^{-\frac{n}{2}(\overline{x}-\theta)^2} \underbrace{e^{-\sum_{i=1}^n (x_i-\overline{x})(\overline{x}-\theta)}}_{=1}. \end{aligned}$$

Now

$$\begin{aligned} f(x_1, \ldots, x_n \mid \overline{x}, \theta) &= \frac{f(x_1, \ldots, x_n, \theta, \overline{x})}{f(\theta, \overline{x})} \\ &= \frac{f(x_1, \ldots, x_n, \overline{x} \mid \theta)}{f(\overline{x} \mid \theta)} \end{aligned}$$

And $f(\overline{x} \mid \theta) = e^{-\frac{n}{2}(\overline{x}-\theta)^2}$ by integrating over $x_1, \ldots, x_n$, so

$$= f(x_1, \ldots, x_n \mid \overline{x}).$$

5.2 Fano's inequality

In the data processing inequality, we had $Y - X - \widehat{Y}$, where $\widehat{Y}$ is viewed as derived from X to learn about Y. Suppose Y and $\hat{Y}$ take values in the same set and our goal is to try to get small $\mathbb{P}(\widehat{Y} \neq Y)$. Fano's inequality gives us an lower bound on this probability using the conditional entropy of Y given X.

Theorem 5.1 (Fano's inequality)**.** *Suppose $Y - X - \widehat{Y}$, and let $p_e = \mathbb{P}(Y \neq \widehat{Y})$. Then*

$$H(Y \mid X) \leq H(Y \mid \widehat{Y}) \leq h(p_e) + p_e \log(|\mathscr{Y}| - 1),$$

where $h(p_e) = -p_e \log p_e - (1 - p_e) \log(1 - p_e)$ is the binary entropy function.

Proof. Because $I(X;Y) = H(Y) - H(Y \mid X)$ and $I(\widehat{Y};Y) = H(Y) - H(Y \mid \widehat{Y})$, the data processing inequality gives $H(Y \mid X) \leq H(Y \mid \widehat{Y})$. Now consider $H(Y, E \mid \widehat{Y})$, where $E = \mathbb{1}_{\{Y \neq \widehat{Y}\}}$ is a $\{0,1\}$-valued random variable. Write this as

$$H(Y, E \mid \widehat{Y}) = H(Y \mid \widehat{Y}) + \underbrace{H(E \mid Y, \widehat{Y})}_{=0}.$$

We can also write this as

$$\begin{aligned} H(Y, E \mid \widehat{Y}) &= H(E \mid \widehat{Y}) + H(Y \mid E, \widehat{Y}) \\ &\leq H(E) + p_e H(Y \mid E = 1, \widehat{Y}) \\ &\leq h(p_e) + p_e \log(|\mathscr{Y}| - 1). \end{aligned}$$

□

5.3 The asymptotic equipartition property

Given $(p(x), x \in \mathscr{X})$ with $\mathscr{X}$ finite, let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} p$. Then

$$\begin{aligned} p^n(x_1^n) &= \prod_{i=1}^n p(x_i) \\ &= \prod_{x \in \mathscr{X}} p(x)^{N(x|x_1^n)}, \end{aligned}$$

where $N(x \mid x_1^n) = \sum_{i=1}^n \mathbb{1}_{\{x_i = x\}}$ is the number of times x shows up in $x_1, \ldots, x_n$.

$$= 2^{\sum_{x \in \mathscr{X}} N(x|x_1^n) \log p(x)}.$$

The Strong Law of Large Numbers tells us that $\frac{1}{n} N(x \mid X_1^n) \to p(x)$ almost surely as $n \to \infty$. This suggests that for large n, the realizations that "matter" are those x_1^n for which each $N(x \mid x_1^n)$ is roughly $np(x)$. The asymptotic equipartition property formalizes this statement in a weak way via the weak law of large numbers. The "method of types" formalizes this more carefully.

The asymptotic equipartition property comes from applying the weak law of large numbers to the iid sequence of entropy densities, i.e. to the sequence $\log \frac{1}{p(x_1)}, \log \frac{1}{p(x_2)}, \ldots$.

Lemma 5.1 (Weak law of large numbers)**.** *For any real-valued iid sequence $Z_1, Z_2, \ldots$ with $\mathbb{E}[|Z_1|] < \infty$,*

$$\frac{1}{n}\sum_{i=1}^{n} Z_i \xrightarrow{p} \mathbb{E}[Z_1].$$

That is, for all $\varepsilon > 0$,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}[Z_1]\right| \leq \varepsilon\right) \to 1$$

as $n \to \infty$.

Theorem 5.2 (Asymptotic equipartition property)**.** *For all $\varepsilon > 0$,*

$$\mathbb{P}\left(\left|-\frac{1}{n}\log p(X_1, \ldots, X_n) - H(X)\right| \leq \varepsilon\right) \to 1$$

as $n \to \infty$.

This leads us to define the following.

Definition 5.1. The set of **ε-weakly typical sequences** is

$$A_\varepsilon^{(n)} := \left\{x_1^n : \left|-\tfrac{1}{n}\log p^n(x_1^n) - H(X)\right| \leq \varepsilon\right\}.$$

We can see

$$x_1^n \in A_\varepsilon^{(n)} \iff 2^{-nH(x)}2^{-n\varepsilon} \leq p^n(x_1^n) \leq 2^{-nH(X)}2^{n\varepsilon}.$$

Proposition 5.1.

$$|A_\varepsilon^{(n)}| \leq 2^{nH}2^{n\varepsilon}.$$

Proof. We must have $\mathbb{P}(X_1^n \in A_\varepsilon^{(n)}) \leq 1$. □

The AEP says that

$$\mathbb{P}(X_1^n \in A_\varepsilon^{(n)}) \to 1$$

as $n \to \infty$. The left hand side is equal to

$$\sum_{x_1^n \in A_\varepsilon^{(n)}} p^n(x_1^n).$$

Hence, for all $\varepsilon \to 0$, if n is large enough (how large depending on δ), $\mathbb{P}(X_1^n \in A_\varepsilon^{(n)}) \geq 1-\delta$. Hence,

$$|A_\varepsilon^{(n)}| \geq (1-\delta)2^{nH}2^{-n\varepsilon}$$

for all large enough n.

6 The Asymptotic Equipartition Property and Data Compression

6.1 The asymptotic equipartition property

Last time, we discussed the asymptotic equipartition property (AEP). Given an iid sequence of random variables $X_1, X_2, \cdots \sim (p(x), x \in \mathscr{X})$ with $\mathscr{X}$ finite, the weak law of large numbers applied to the sequence $\log \frac{1}{p(X_1)}, \log \frac{1}{p(X_2)}, \ldots$ tells us that for every $\varepsilon > 0$,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} \log \frac{1}{p(X_i)} - \mathbb{E}\left[\log \frac{1}{p(X)}\right]\right| < \varepsilon\right) \xrightarrow{n\to\infty} 1.$$

Note that $\frac{1}{n}\sum_{i=1}^{n} \log \frac{1}{p(X_i)} = \frac{1}{n} \log \frac{1}{p^n(X_1^n)}$ because $p^n(X_1^n) = \prod_{i=1}^{n} p(X_i)$ from the iid assumption. Also note that $\mathbb{E}[\log \frac{1}{p(X)}] = H(X)$. In other words,

$$\mathbb{P}\left(-\varepsilon < \frac{1}{n} \log \frac{1}{p(X_1^n)} - H(X) < \varepsilon\right) \xrightarrow{n\to\infty} 1.$$

We can also write this as

$$\mathbb{P}\left(2^{-nH}2^{-n\varepsilon} < p^n(X_1^n) < 2^{-nH}2^{n\varepsilon}\right) \xrightarrow{n\to\infty} 1.$$

We define the set of **ε-weakly typical sequences** $A_\varepsilon^{(n)} \subseteq \mathscr{X}^n$ as

$$A_\varepsilon^{(n)} := \{x_1^n \in \mathscr{X}^n : 2^{-nH}2^{-n\varepsilon} < p^n(x_1^n) < 2^{-nH}2^{n\varepsilon}\}.$$

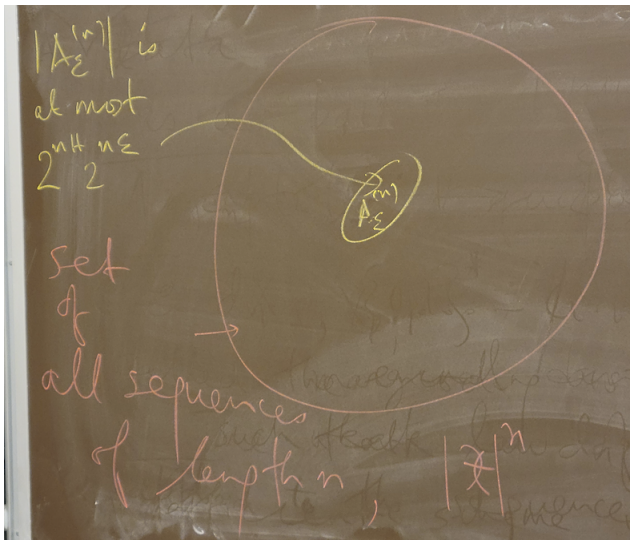
We learn that

1. For all $\varepsilon > 0$,
$$\mathbb{P}(X_1^n \in A_\varepsilon^{(n)}) \xrightarrow{n\to\infty} 1.$$

2. For all $\varepsilon > 0$, $|A_\varepsilon^{(n)}| \leq 2^{nH}2^{n\varepsilon}$ because
$$\mathbb{P}(X_1^n \in A_\varepsilon^{(n)}) = \sum_{x_1^n \in A_\varepsilon^{(n)}} p^n(x_1^n) \geq \sum_{x_1^n \in A_\varepsilon^{(n)}} 2^{-nH}2^{-n\varepsilon} = |A_\varepsilon^{(n)}|2^{-nH}2^{-n\varepsilon}.$$

3. For any $\varepsilon > 0$ and $\delta > 0$, for all sufficiently large n (how large depending on (ε, δ)),
$$|A_\varepsilon^{(n)}| > (1-\delta)2^{nH}2^{-n\varepsilon}$$
because if n is large enough,
$$1 - \delta < \mathbb{P}(X_1^n \in A_\varepsilon^{(n)}) = \sum_{x_1^n \in A_\varepsilon^{(n)}} p^n(x_1^n) \leq \sum_{x_1^n \in A_\varepsilon^{(n)}} 2^{-nH}2^{n\varepsilon} = |A_\varepsilon^{(n)}|2^{-nH}2^{n\varepsilon}.$$

Together, these three statements comprise the **asymptotic equipartition property**.



6.2 Data compression

From the point of view of data compression, the AEP says that there is a data compression scheme where you assign shorter length bit strings to more commonly occurring sequences. On average, you will end up compressing the data with such a scheme.

Definition 6.1. A **lossless data compression scheme at block length** n is a pair of maps (e_n, d_n) called the **encoding** and **decoding maps**

$$e_n : \mathcal{X}^n \to \{0,1\}^* \setminus \{\varnothing\}, \qquad d_n : \{0,1\}^* \setminus \{\varnothing\} \to \mathcal{X}^n$$

(with $\{0,1\}^*$ denoting the set of binary sequences of finite length) such that $d_n \circ e_n : \mathcal{X}^n \to \mathcal{X}^n$ is the identity map.

An efficient scheme will try to minimize $\mathbb{E}[\ell(e_n(X_1^n))]$, where $\ell : \{0,1\}^* \to \mathbb{N}$ denotes the length of the string and the expectation is for $X_1, \ldots, X_n \overset{\text{iid}}{\sim} (p(x), x \in \mathcal{X})$.

The AEP suggests the following scheme:

1. Use 1 bit to declare if $x_1^n \in A_\varepsilon^{(n)}$ or not.
2. If $x_1^n \in A_\varepsilon^{(n)}$, we can represent it by at most

$$\lceil \log |A_\varepsilon^{(n)}| \rceil \leq \lceil 2^{nH} 2^{n\varepsilon} \rceil \leq nH + n\varepsilon + 1$$

bits.

3. If $x_1^n \notin A_\varepsilon^{(n)}$, we can represent it by $\lceil \log |\mathscr{X}^n| \rceil \le n \log |\mathscr{X}| + 1$ bits.

With this data compression scheme,

$$\mathbb{E}[\ell(e_n(X_1^n))] \le 1 + \mathbb{P}(X_1^n \in A_\varepsilon^{(n)})(nH + n\varepsilon + 1) + (1 - \mathbb{P}(X_1^n \in A_\varepsilon^{(n)}))(n \log |\mathscr{X}| + 1),$$

so

$$\limsup_{n\to\infty} \frac{1}{n} \mathbb{E}[\ell(e_n(X_1^n))] \le H(X) + \varepsilon$$

because $\mathbb{P}(X_1^n \in A_\varepsilon^{(n)}) \to 1$. This scheme is lossless, as well.

6.3 Asymptotic optimality of the AEP compression scheme

It turns out that asymptotically compressing below $H(X) - \varepsilon$ bits per symbol via a lossless scheme is impossible for any $\varepsilon > 0$. To see this, let $B_\delta^{(n)} \subseteq \mathscr{X}^n$ be any set with $\mathbb{P}(X_1^n \in B_\delta^{(n)}) \ge 1 - \delta$. Then

$$\mathbb{P}(X_1^n \in B_\delta^{(n)} \cap A_\varepsilon^{(n)}) \ge 1 - 2\delta$$

for all large enough n because $\mathbb{P}(X_1^n \in A_\varepsilon^{(n)}) > 1 - \delta$ (and using a union bound). So

$$1 - 2\delta \le \sum_{x_1^n \in B_\delta^{(n)} \cap A_\varepsilon^{(n)}} p^n(x_1^n) \le |B_\delta^{(n)} \cap A_\varepsilon^{(n)}| 2^{-nH} 2^{n\varepsilon}.$$

This tells us that

$$|B_\delta^{(n)} \cap A_\varepsilon^{(n)}| \ge (1 - 2\delta) 2^{nH} 2^{-n\varepsilon}$$

for all large enough n.

Suppose we have a probability distribution on a finite set giving probability $2^{-nH}2^{n\varepsilon}$ to each of $\lfloor (1-2\delta)2^{nH}2^{-n\varepsilon} \rfloor$ elements of the set and giving an arbitrary distribution to the rest of the sequences. We claim that the expected length under any lossless binary encoding of such a distribution is "approximately" bounded below by $nH - n\varepsilon - 1$. To see this, consider a binary tree of depth L. The total number of nodes is $2 + 2^2 + \cdots + 2^L = 2^{L+1} - 2$. The total depth of all the nodes is

$$1 \cdot 2 + 2 \cdot 2^2 + 3 \cdot 2^3 + \cdots + L 2^L = (L-1)2^{L+1} + 2.$$

So the average depth is

$$\frac{(L-1)2^{L+1} + 2}{2^{L+1} - 2} \ge L - 1$$

The precise lower bound is

$$\log \left(\lfloor (1 - 2\delta) 2^{nH} 2^{n\varepsilon} \rfloor + 2 \right) - 2.$$

This is further lower bounded by

$$\log((1-2\delta)2^{nH}2^{-n\varepsilon}) - 2 = \log(1-2\delta) + n(H-\varepsilon) - 2.$$

So

$$\frac{1}{n}\text{expected depth} \geq \frac{1}{n}(\log(1-2\delta) - 2) + H - \varepsilon$$

A lossless compression scheme $\mathscr{X}^n \to \{0,1\}^* \backslash \{\varnothing\}$ must use at least this many bits/symbols because $\mathbb{P}(X_1^n \in B_\delta^{(n)} \cap A_\varepsilon^{(n)}) > 1 - 2\delta$ and each $x_1^n \in B_\delta^{(n)} \cap A_\varepsilon^{(n)}$ has $p^n(x_1^n) \leq 2^{-nH}2^{n\varepsilon}$.

7 Types, Typicality Sets, and Entropy Rate

7.1 Types

Let $\mathscr{X}$ be a finite set (called the alphabet). Given a sequence of symbols $x_1^n := (x_1, \ldots, x_n)$ taking values in $\mathscr{X}^n$ and $x \in \mathscr{X}$, let $N(x \mid x_1^n) = \sum_{i=1}^n \mathbb{1}_{\{x_i = x\}}$ be the number of times x shows up in x_1^n. Notice that $(\frac{N(x|x_1^n)}{n}, x \in \mathscr{X})$ is a probability distribution on $\mathscr{X}$ (which depends on $\mathscr{X}$).

Definition 7.1. The distribution $P_{x_1^n} = (\frac{N(x|x_1^n)}{n}, x \in \mathscr{X})$ is called the **type** of x_1^n in information theory and the **empirical distribution** of x_1^n more generally.

A type based on a sample of size n from $\mathscr{X}$ has to be of the form $(\frac{k_x}{n}, x \in \mathscr{X})$ for some integers $0 \le k_x \le n$ with $\sum_x k_x = n$. $\mathcal{P}_n$ denotes the set of all types based on samples of size n from $\mathscr{X}$.

Proposition 7.1.

$$|\mathcal{P}_n| \le (n+1)^{|\mathscr{X}|}.$$

So $|\mathcal{P}_n|$ grows only polynomially in n. Contrast this with the total number of sequences of length n, whose size is $|\mathscr{X}|^n$, exponential in n.

7.2 The scale of typicality sets

Definition 7.2. For $p \in \mathcal{P}_n$, the set $T(p) = \{x_1^n : P_{x_1^n} = p\} \subseteq \mathscr{X}^n$ is called the **typicality set** of type p.

Now note that given any probability distribution $(q(x), x \in \mathscr{X})$ and any sequence $x_1^n \in \mathscr{X}^n$, $q^n(x_1^n) = \prod_{i=1}^n q(x_i)$ is determined by $P_{x_1^n}$, the type of x_1^n, because

$$\begin{aligned} q^n(x_1^n) &= \prod_{x \in \mathscr{X}} q(x)^{N(x|x_1^n)} \\ &= \prod_{x \in \mathscr{X}} 2^{n P_{x_1^n}(x) \log q(x)} \\ &= 2^{n \sum_x P_{x_1^n}(x) \log q(x)}, \end{aligned}$$

which depends on x_1^n only through its type. But also note that

$$\sum_x P_{x_1^n}(x) \log q(x) = \sum_x P_{x_1^n}(x) \log \frac{q(x)}{P_{x_1^n}(x)} + \sum_x P_{x_1^n}(x) \log P_{x_1^n}(x),$$

so

$$q^n(x_1^n) = 2^{-n(H(P_{x_1^n}) + D(p_{x_1^n} || q))}.$$

This calculation implies the following:

Proposition 7.2. *For any $p \in \mathcal{P}_n$,*

$$|T(p)| \leq 2^{nH(p)}.$$

Proof. Take q to be p and consider x_1^n having $P_{x_1^n} = p$. This tells us that for all x_1^n with type $P_{x_1^n} = p$,

$$p^n(x_1^n) = 2^{-nH(p)}$$

because $D(p \,||\, p) = 0$.

But, given $p \in \mathcal{P}_n$,

$$\begin{aligned} 1 &= \sum_{x_1^n} p^n(x_1^n) \\ &\geq \sum_{x_1^n : P_{x_1^n} = p} p^n(x_1^n) \\ &= \sum_{x_1^n : P_{x_1^n} = p} 2^{-nH(p)} \\ &= |T(p)| 2^{-nH(p)}. \end{aligned}$$

□

We can also prove a lower bound:

Proposition 7.3. *For all $p \in \mathcal{P}_n$,*

$$|T(p)| \geq \frac{2^{nH(p)}}{(n+1)^{|\mathscr{X}|}}.$$

Proof. This comes from showing that for $p \in \mathcal{P}_n$, $p^n(T(p)) \geq p^n(T(\widehat{p}))$ for all $\widehat{p} \in \mathcal{P}_n$. The left hand side is

$$p^n(T(p)) = \sum_{x_1^n : P_{x_1^n} = p} p^n(x_1^n) = \sum_{x_1^n : P_{x_1^n} = p} 2^{-nH(p)} = |T(p)| 2^{-nH(p)},$$

while the right hand side is $|T(\widehat{p})| 2^{-n(H(\widehat{p}) + D(\widehat{p}||p))}$.

Substituting the exact values of $|T(p)|$ and $|T(\widehat{p})|$ using combinatorics, the left hand side is $\binom{n}{np(a_1), \ldots, np(a_d)} 2^{-nH(p)}$ (with $\mathscr{X} = \{a_1, \ldots, a_d\}$), while the right hand side is $\binom{n}{n\widehat{p}(a_1), \ldots, n\widehat{p}(a_d)} 2^{-n(H(\widehat{p}) + D(\widehat{p}||p))}$. So

$$\frac{p^n(T(p))}{p^n(T(\widehat{p}))} \geq \frac{n!}{np(a_1)! \cdots np(a_d)!} \frac{2^{n \sum_{i=1}^d p(a_i) \log p(a_i)}}{n!} \frac{n\widehat{p}(a_1)! \cdots n\widehat{p}(a_d)!}{2^{n \sum_{i=1}^d \widehat{p}(a_i) \log \widehat{p}(a_i)}}$$

Now observe that $\frac{m!}{\ell!} \geq \ell^{m-\ell}$ for all ℓ, m.

$$\geq \frac{\prod_{i=1}^n p(a_i)^{np(a_i)} (np(a_i))^{n\widehat{p}(a_i)}}{\prod_{i=1}^n \widehat{p}(a_i)^{n\widehat{p}(a_i)} (np(a_i))^{np(a_i)}}$$

$$= 1.$$

Finally, we have

$$\begin{aligned} 1 &= \sum_{\widehat{p} \in \mathcal{P}_n} p^n(T(\widehat{p})) \\ &\leq |\mathcal{P}_n| p^n(T(p)) \\ &\leq (n+1)^{|\mathscr{X}|} p^n(T(p)) \\ &= (n+1)^{|\mathscr{X}|} |T(p)| 2^{-nH(p)}. \end{aligned}$$

□

7.3 ε-typical sets in terms of types

For a probability distribution q on $\mathscr{X}$,

$$A_\varepsilon^{(n)} := \left\{ x_1^n : \left| -\frac{1}{n} \sum_{i=1}^n \log q(x_i) - H(q) \right| < \varepsilon \right\}.$$

Proposition 7.4.

$$A_\varepsilon^{(n)} = \{x_1^n : |D(P_{x_1^n} \,||\, q) + H(P_{x_1^n}) - H(q)| < \varepsilon\}.$$

Proof.

$$\begin{aligned} -\frac{1}{n} \sum_{i=1}^n \log q(x_i) &= -\frac{1}{n} \sum_x N(x \mid x_1^n) \log q(x) \\ &= -\sum_x p_{x_1^n}(x) \log q(x) \\ &= D(P_{x_1^n} \,||\, q) + H(P_{x_1^n}). \end{aligned}$$

So

$$A_\varepsilon^{(n)} = \{x_1^n : |D(P_{x_1^n} \,||\, q) + H(P_{x_1^n}) - H(q)| < \varepsilon\},$$

as claimed. □

7.4 Stationary sequences and entropy rate

Beyond iid sequences, we consider stationary random sequences.

Definition 7.3. As sequence of random variables $(X_k)_{k=-\infty}^{\infty}$ with $X_k \in \mathscr{X}$ is called **stationary** if

$$\mathbb{P}(X_\ell = x_0, X_{\ell+1} = x_1, \ldots, X_{\ell+L} = x_L) = \mathbb{P}(X_{\ell+m} = x_0, X_{\ell+m+1} = x_1, \ldots, X_{\ell+m+L} = x_L)$$

for all $\ell, m \in \mathbb{Z}$, $L \geq 0$, and $x_0, \ldots, x_L \in \mathscr{X}$.

For a stationary sequence,

$$H(X_2 \mid X_1) \leq H(X_2),$$

but $H(X_2) = H(X_1)$ by stationarity, so

$$H(X_2 \mid X_1) \leq H(X_1).$$

Similarly,

$$H(X_{L+2} \mid X_1, \ldots, X_{L+1}) \leq H(X_{L+1} \mid X_1, \ldots, X_L)$$

because the left hand side equals $H(X_{L+1} \mid X_0, \ldots, X_L)$ by stationarity.

This implies that for a stationary process,

$$\lim_{L \to \infty} H(X_{L+1} \mid X_1, \ldots, X_L)$$

exists and is called the **entropy rate** of the process. In fact, the chain rule says that this equals

$$\lim_{L \to \infty} \frac{1}{L} H(X_1, \ldots, X_L).$$

Definition 7.4. A stationary process is a **stationary Markov chain** if

$$\mathbb{P}(X_{L+1} = x_{L+1} \mid X_1 = x_1, \ldots, X_L = x_L) = \mathbb{P}(X_{L+1} = x_{L+1} \mid X_L = x_L)$$

for all $L \geq 1$ and $x_1, \ldots, x_{L+1}$.

So all that matters is the matrix $[p(j \mid i) : 1 \leq i, j \leq |\mathscr{X}|]$, where the **transition probabilities** $p(j \mid i) = \mathbb{P}(X_2 = j \mid X_1 = i)$. If we let $\pi(i) := \mathbb{P}(X_1 = i)$ for $i \in \mathscr{X}$ in a stationary Markov chain, then

$$\sum_i \pi(i) p(j, i) = \pi(j)$$

for all j. The entropy rate for a stationary markov chain will be $H(X_2 \mid X_1)$ because $H(X_2 \mid X_1, X_0) = H(X_2, X_1)$. So the entropy rate is

$$\sum_i \pi(i) \sum_j p(j \mid i) \log \frac{1}{p(j \mid i)}.$$

8 Entropy Rate, Markov Processes, and Data Compression for Sequences

8.1 Entropy rate

Last time, we introduced the entropy rate of a stationary stochastic process. If $\mathscr{X}$ is a finite or countably infinite set, a **stationary stochastic process** is a sequence of random variables $(X_k)_{k=-\infty}^{\infty}$ with the property that

$$\mathbb{P}(X_k = x_0, X_{k+1} = x_1, \dots, X_{k+t} = x_t)$$

does not depend on k (for all $t \geq 0$, $x_0, \dots, x_t$). The **entropy rate** of the process is the limit

$$\lim_{n\to\infty} \frac{1}{n} H(X_1, \dots, X_n),$$

where the limit exists because $H(X_1) \geq H(X_2 \mid X_1) \geq \cdots \geq H(X_n \mid X_1, \dots, X_{n-1})$, and in fact,

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

because of the chain rule, $H(X_1, \dots, X_n) = H(X_1) + H(X_2 \mid X_1) + \cdots + H(X_n \mid X_1^{n-1})$. We can think of the entropy rate as the asymptotic amount of information we learn from the next random variable in the sequence.

Observe that

$$p(x_1, \dots, x_t) = p(x_1)p(x_2 \mid x_1) \cdots p(x_t \mid x_1, \dots, x_{t-1}).$$

For large t and $k < t$,

$$p(x_1)p(x_2 \mid x_1) \cdots p(x_k \mid x_1, \dots, x_{k-1}) \prod_{j=1}^{t-k} p(x_{k+j} \mid x_{k+j-t-1}, \dots, x_{k+j-1})$$

might be a decent approximation from a modeling point of view. This defines a $(k-1)$-order stationary Markov process.

A **first order Markov process**[3] is defined by the transition probabilities $p(x_2 \mid x_1)$ for $x_1, x_2 \in \mathscr{X}$ and an initial distribution $(p(x), x \in \mathscr{X})$. A **stationary Markov process** is defined by the transition probabilities and a probability distribution $(\pi(x), x \in \mathscr{X})$ such that $\sum_{x\in\mathscr{X}} \pi(x)p(y \mid x) = \pi(y)$ for all $y \in \mathscr{X}$. This would mean that

$$p(x_1, \dots, x_t) = \pi(x_1)p(x_2 \mid x_1)p(x_3 \mid x_2) \cdots p(x_t \mid x_{t-1}).$$

[3]For finite or countable state spaces, these are often referred to as "Markov chains."

For a k-th order Markov process, we need $p(x_{k+1} \mid x_1, \ldots, x_k)$ with $x_i \in \mathscr{X}$ and $i = 1, \ldots, k+1$ and an initial distribution $(p(x_1, \ldots, x_k), x_1^k \in \mathscr{X}^k)$. For stationarity, we need a distribution $(\pi(x_1, \ldots, x_k), x_1^k \in \mathscr{X}^k)$ such that

$$\sum_{x_1} \pi(x_1, \ldots, x_k) p(x_{k+1} \mid x_1, \ldots, x_k) = \pi(x_2, \ldots, x_{k+1}).$$

The entropy rate for a stationary Markov process is $H(X_2 \mid X_1)$, while the entropy rate for a k-th order stationary Markov process is

$$H(X_{k+1} \mid X_1, \ldots, X_k) = \sum_{x_1, \ldots, x_k} \pi(x_1, \ldots, x_k) H((p(x_{k+1} \mid x_1, \ldots, x_k), x_{k+1} \in \mathscr{X})).$$

For $k = 1$, this is

$$H(X_2 \mid X_1) = -\sum_{x,y} \pi(x) p(y \mid x) \log p(y \mid x)$$

8.2 Time reversal and reversible Markov processes

An important class of examples is reversible stationary Markov processes.

Definition 8.1. A stationary Markov process is **reversible** if

$$\pi(x) p(y \mid x) = \pi(y) p(x \mid y) \qquad \forall x, y \in \mathscr{X}.$$

For a general stationary Markov chain depending on the transition probability matrix $[p(y \mid x)]_{x,y \in \mathscr{X}}$ and stationary distribution $(\pi(x), x \in \mathscr{X})$, one can define

$$\widetilde{p}(y \mid x) := \frac{\pi(y) p(x \mid y)}{\pi(x)}$$

(assuming $\pi(x) > 0$ for all $x \in \mathscr{X}$). Then

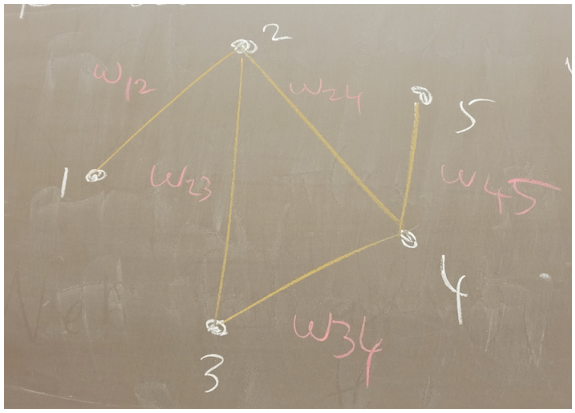
$$\sum_{y \in \mathscr{X}} \widetilde{p}(y \mid x) = \frac{\sum_{y \in \mathscr{X}} \pi(y) p(x \mid y)}{\pi(x)} = \frac{\pi(x)}{\pi(x)} = 1$$

and

$$\sum_{x \in \mathscr{X}} \pi(x) \widetilde{p}(y \mid x) = \sum_{x \in \mathscr{X}} \pi(y) p(x \mid y) = \pi(x),$$

so $[\widetilde{p}(y \mid x)]_{x,y \in \mathscr{X}}$ defines a transition probability matrix with stationary distribution $(\pi(x), x \in \mathscr{X})$. This is called the **time reversal** of the original process. A Markov process is time reversible if and only if its time reversal has the same joint distributions as as the original process.

Example 8.1. Stationary random walks on weighted graphs give rise to examples.



At any time t, X_t belongs to the set of vertices, and

$$\mathbb{P}(X_{t+1} = j \mid X_t = i) = \frac{w_{i,j}}{\sum_{k \in V} w_{i,k}}.$$

The stationary distribution will be

$$\pi(i) = \frac{\sum_{j \in V} w_{i,j}}{2 \sum_{i,j} w_{i,j}},$$

and this process is reversible.

This is of huge importance in algorithms, and it has connections to resistive network theory.[4]

8.3 Overview of data compression for sequences

The next 2-3 lectures will be about various schemes for lossless data compression. The goal is to represent observed data efficiently (using as few bits/symbols as possible). We have already seen, for example, that if $X_1, X_2, \ldots$ are iid with marginal distribution $(p(x), x \in \mathscr{X})$, there exists an encoding map $e_n : \mathscr{X}^n \to \{0,1\}^* \setminus \{\varnothing\}$ and a decompression map $d_n : \{0,1\}^* \setminus \{\varnothing\} \to \mathscr{X}^n$ (for each $n \geq 1$) such that $d_n \circ e_n$ is the identity map and

$$\frac{1}{n} \mathbb{E}[\text{length}(e_n(X_1, \ldots, X_n))] \leq H + \varepsilon.$$

Morover, we have also seen that for any $e_n : \mathscr{X}^n \to \{0,1\}^* \setminus \{\varnothing\}$ and $d_n : \{0,1\}^* \setminus \{\varnothing\} \to \mathscr{X}^n$ with $d_n \circ e_n = \text{id}$, for any $\varepsilon > 0$,

$$\frac{1}{n} \mathbb{E}[\text{length}(e_n(X_1, \ldots, X_n))] \geq H - \varepsilon.$$

[4]This is covered in a book by Doyle and Snell called *Random Walks and Electrical Networks*.

There is a book called *Handbook of Data Compression* by Salamon which discusses this.[5]

We would like a version of this for stationary processes. We'll see this as we go along, but here are some big picture facts related to this.

We cannot get an analog of the Strong Law of Large Numbers for stationary processes without assuming an additional condition called **ergodicity** which excludes examples like $p(\ldots, X_0 = 1, \ldots, X_t = 1) = \mathbb{P}(X_0 = 0, \ldots, X_t = 0) = 1/2$ for all $t \geq 0$.

For a stationary ergodic process, we have (under some conditions) the pointwise ergodic theorem:

Theorem 8.1 (Pointwise ergodic theorem, Birkhoff)**.** *Let $(X_k)_{k=-\infty}^{\infty}$ be a stationary, ergodic process with random variables taking values in $\mathscr{X}$. Given $f : \mathscr{X} \to \mathbb{R}$,*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} f(X_i) = \mathbb{E}[f(X_1)]$$

almost surely.

But even this is not enough for us to replace the Strong Law of Large Numbers applied to information densities. We need a further statement:

Theorem 8.2 (Shannon-McMillan-Breiman)**.** *If $(X_k)_{k=-\infty}^{\infty}$ is a stationary, ergodic process,*

$$\lim_{n\to\infty} -\frac{1}{n} \log p(X_1, \ldots, X_n) = \textit{entropy rate of process}$$

almost surely.

From a practical point of viewpoint, $e_n : \mathscr{X}^n \to \{0,1\}^* \setminus \{\varnothing\}$ needs to be construction from "smaller pieces." For example, start with $e : \mathscr{X} \to \{0,1\}^* \setminus \{\varnothing\}$ and define $e_n(x_1, \ldots, x_n) = e(x_1)e(x_2)\cdots e(x_n)$. This function e needs to be 1 to 1 for invertibility. But even if $e : \mathscr{X} \to \{0,1\}^* \setminus \{\varnothing\}$ is 1 to 1, e_n might not be.

Example 8.2. Let $\mathscr{X} = \{1, 2, 3\}$ with $e(1) = 0$, $e(2) = 00$, and $e(3) = 1$. Then

$$e_3(12) = 000, \qquad e_3(21) = 000.$$

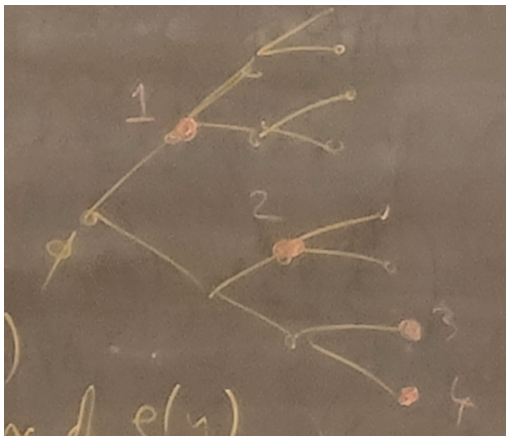
Definition 8.2. $e : \mathscr{X} \to \{0,1\}^* \setminus \{\varnothing\}$ is called **uniquely decodable** if each e_n is one to one.

One way to get this property is to make e **instantaneous** (or **prefix-free**) if no $e(x)$ is a prefix of $e(y)$ for $x \neq y$.

[5] The book is on the order of 1000 pages long.

Example 8.3. If $\mathscr{X} = \{1, 2, 3, 4\}$, we can take

$$e(1) = 1, \qquad e(2) = 01, \qquad e(3) = 001, \qquad e(4) = 000.$$



9 Uniquely Decodable Codes

9.1 Uniquely decodable and prefix-free codes

Last time, we talked about lossless data compression.

Definition 9.1. A **lossless data compression scheme** is a sequence $((e_n, d_n), n \geq 1)$ of maps $e_n : \mathscr{X}^n \to \{0,1\}^* \setminus \{\varnothing\}$ and $d_n : \{0,1\}^* \setminus \{\varnothing\} \to \mathscr{X}^n$ such that $d_n \circ e_n$ is the identity for each n.

Equivalently, we could specify $(e_n, n \geq 1)$ and insist that each e_n is one to one. Equivalently, we could specify $e_* : \mathscr{X}^* \setminus \{\varnothing\} \to \{0,1\}^* \setminus \varnothing$ which is one to one on each $\mathscr{X}^n$.

One practical way to create encoding maps is to specify $e : \mathscr{X} \to \{0,1\}^* \setminus \{\varnothing\}$ and define $e_n(x_1^n) = e(x_1)e(x_2)\cdots e(x_n)$.

Example 9.1. If $\mathscr{X} = \{1,2,3\}$ with

$$e(1) = 0, \qquad e(2) = 11, \qquad e(3) = 0110,$$

then we can encode larger words like

$$e(12) = 011.$$

We could also do this by specifying e on blocks of length r for some $r \geq 1$.

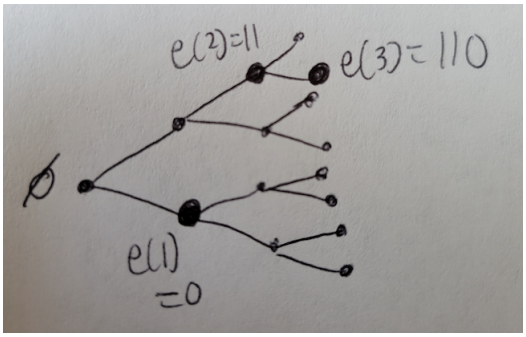
Definition 9.2. We'll say e is **uniquely decodable** if each e_n is one to one.

This is equivalent to requiring that e_* is one to one. One way to get unique decodability is if e is *instantaneous* or *prefix-free*.

Definition 9.3. The encoding map e is **instantaneous** or **prefix-free**[6] if for all $x \neq y \in \mathscr{X}$, $e(x)$ is not a prefix of $e(y)$.

It's easiest to think about this in terms of a binary tree. Prefix-free is equivalent to the requirement that no node $e(x)$ in the coding can lie on the path between the root and another node $e(y)$.

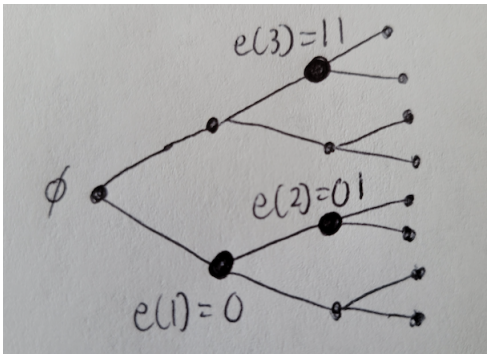
Example 9.2. The previous example is *not* prefix-free.



[6] Cover and Thomas call this a "prefix code," which is super confusing.

Example 9.3. The following example is uniquely decodable but not prefix-free:

$$e(1) = 0, \qquad e(2) = 01, \qquad e(3) = 11.$$



If the first received bit is 1 and the next bit is 1, we can parse 11 from the received sequence. If the first received bit is 0, then if the next bit is 0, we can parse off the first 0; if the next bit is 1, we need to wait to figure out the parity of the run of 1s.

9.2 Kraft's inequality

Theorem 9.1 (Kraft's inequality)**.** *For any prefix-free binary code e,*

$$\sum_{x \in \mathscr{X}} 2^{-\ell(e(x))} \leq 1.$$

Proof. A formal proof is in Cover and Thomas. The idea is to add the weights of the $e(x)$, where a node at depth d gets weight 2^{-d}. □

Remark 9.1. There is a version of this for prefix-free D-ary codes $e : \mathscr{X} \to \{0, \ldots, D-1\}$. In this case, Kraft's inequality says

$$\sum_{x \in \mathscr{X}} D^{-\ell(e(x))} \leq 1.$$

Here is a generalization by McMillan.

Theorem 9.2. *For every uniquely decodable code $e : \mathcal{X} \to \{0,1\}^* \setminus \{\varnothing\}$,*

$$\sum_{x \in \mathscr{X}} 2^{-\ell(e(x))} \leq 1.$$

Remark 9.2. There is also a version of this for uniquely decodable D-ary codes $e : \mathscr{X} \to \{0, \ldots, D-1\}$. In this case, Kraft's inequality says

$$\sum_{x \in \mathscr{X}} D^{-\ell(e(x))} \leq 1.$$

Let's prove the D-ary version:

Proof. We use a generating function technique. Consider the expression

$$\sum_{x_k \in \mathscr{X}} D^{-\ell(e(x_k))}$$

at time k. Then

$$\sum_{x_1^n \in \mathscr{X}^n} D^{-\ell(e_n(x_1^n))} = \sum_{x_1^n \in \mathscr{X}^n} \prod_{k=1}^{n} D^{-\ell(e(x_k))} = \prod_{k=1}^{n} \sum_{x_k \in \mathscr{X}} D^{-\ell(e(x_k))}.$$

Take n-th roots on both sides. The key observations are that on the left hand side,

1. The total number of terms that provide D^{-m} for any $m \geq 1$ is at most D^m (by unique decodability).
2. The largest m for which D^{-m} shows up in the left hand side is $n\ell_{\max}$, where $\ell_{\max} = \max_x \ell(e(x))$.

So this tells us that the left hand side is $\leq n\ell_{\max}$. Now observe that

$$(n\ell_{\max})^{1/n} = e^{\frac{1}{n}(\log n + \log \ell_{\max})} \xrightarrow{n\to\infty} 1. \qquad \square$$

9.3 Optimal compression as a linear programming problem

To optimize compression (in bits/symbol) for an iid source ($\mathscr{X}$-valued, marginal distribution $(p(x), x \in \mathscr{X})$), we want to solve

$$\text{minimize: } \sum_{x \in \mathscr{X}} p(x)\ell(e(x))$$

$$\text{subject to: } e \text{ is prefix-free.}$$

This suggests studying the problem

$$\text{minimize: } \sum_{x \in \mathscr{X}} p(x)\ell(e(x))$$

$$\text{subject to: } \sum_{x \in \mathscr{X}} 2^{-\ell(e(x))} \leq 1, \quad \ell(e(x)) \geq 1.$$

This, although looking like a weakening of the constraint, actually is equivalent because for every collection of lengths satisfying Kraft's inequality, there is a prefix-free code with those lengths.

Proposition 9.1. *Given any $(\ell(x), x \in \mathscr{X})$ with $\ell(x) \geq 1$ and $\sum_{x\in\mathscr{X}} 2^{-\ell(x)} = 1$, there exists a prefix-free $e : \mathscr{X} \to \{0,1\}^* \setminus \{\varnothing\}$ such that for all x, $\ell(e(x)) = \ell(x)$.*

Proof. Proceed by induction on $|\mathscr{X}|$. Merge the two smallest values $2^{-\ell(x)}$ to reduce the size of the alphabet by 1. □

This new linear program is an integer programming problem because the variables $\ell(x)$ are required. This is computationally difficult. We can relax this to get a tractable problem

$$\text{minimize: } \sum_{x\in\mathscr{X}} p(x)\ell(e(x))$$

$$\text{subject to: } \sum_{x\in\mathscr{X}} 2^{-\ell(e(x))} \leq 1,$$

where the $\ell(x)$ can be real-valued. The second condition implies $\ell(x) \geq 0$. We can also replace the inequality by an equality because this only improves the objective.

We can solve this using Lagrange multipliers. Consider the Lagrangian

$$\sum_x p(x)\ell(x) + \lambda\left(\sum_x 2^{-\ell(x)} - 1\right).$$

Differentiate in each $\ell(x)$, and set the derivative equal to 0 to get

$$p(x) - \lambda \log_e 2 \cdot 2^{-\ell(x)} = 0 \quad \forall x \in \mathscr{X}.$$

This requires

$$\ell(x) = -\log p(x) + k$$

for all $x \in \mathscr{X}$. The condition $\sum_x 2^{-\ell(x)} = 1$ gives us $k = 0$. So the optimal value of the objective is

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

We have just proven the following:

Theorem 9.3. *The expected length of a uniquely decodable binary code is $\geq H(p)$.*

Remark 9.3. Taking $\ell(x) = \lceil \log \frac{1}{p(x)} \rceil$ for $x \in \mathscr{X}$, we have

$$\sum_{x\in\mathscr{X}} 2^{-\ell(x)} \leq 1.$$

Hence, there is a prefix code with these lengths. Such a code is called a **Shannon code**. Its expected length is

$$\sum_x p(x)\left\lceil \log \frac{1}{p(x)} \right\rceil \leq H(p) + 1.$$

For a stationary process working at block length n, we can get lossless compression with bits/symbol at most

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

We will discuss this next time.

10 Shannon Codes, Huffman Codes, and Shannon-Fano-Elias Codes

10.1 Recap: Shannon codes

We have been discussing symbol by symbol codes $c : \mathcal{X} \to \{0,1\}^* \setminus \{\varnothing\}$, extended to $c : \mathcal{X}^* \setminus \{\varnothing\} \to \{0,1\}^* \setminus \{\varnothing\}$ by $c(x_1, \ldots, x_n) = c(x_1) \cdots c(x_n)$.

Definition 10.1. We say that c is **uniquely decodable** if its extension is one to one.

Definition 10.2. We say that c is **prefix-free** if $c(x)$ is not a prefix of $c(y)$ for $x \neq y \in \mathscr{X}$.

We showed the Kraft inequality for prefix-free codes:

$$\sum_{x \in \mathscr{X}} 2^{-\ell(c(x))} \leq 1.$$

We saw this by looking at the length associated to the codewords, viewed as sitting in a binary tree. We also proved an extension of this inequality by McMillan to uniquely decodable codes.

We saw that the minimization of the expected length of codewords $\sum_{x \in \mathscr{X}} p(x)\ell(c(x))$ over prefix-free codes is equivalent to the integer programming problem

$$\text{minimize: } \sum_{x \in \mathscr{X}} p(x)\ell(x)$$

$$\text{subject to: } \sum_{x \in \mathscr{X}} 2^{-\ell(x)} \leq 1,$$

where $\ell(x)$ is a positive integer for each $x \in \mathscr{X}$. This equivalence came from Kraft's inequality and from the fact that for every sequence of lengths $(\ell(x), x \in \mathscr{X})$ satisfying Kraft's inequality, there is a prefix code $c : \mathscr{X} \to \{0,1\}^* \setminus \{\varnothing\}$ with $\ell(c(x)) = \ell(x)$ for each $x \in \mathscr{X}$. Similarly, the Kraft-McMillan inequality says that this is equivalent to the same problem for uniquely decodable codes.

Using Lagrange multipliers, we saw that minimizing $\sum_{x \in \mathscr{X}} p(x)\ell(x)$ subject to $\sum_{x \in \mathscr{X}} 2^{-\ell(x)} \leq 1$ with real $\ell(x)$ has optimal solution

$$\ell(x) = \log \frac{1}{p(x)}, \qquad x \in \mathscr{X}$$

with optimal value

$$\sum_x p(x) \log \frac{1}{p(x)} = H(p).$$

This leads to the idea of a Shannon code.

Definition 10.3. A **Shannon code** is a code $c : \mathscr{X} \to \{0,1\}^* \setminus \{\varnothing\}$ with $\ell(c(x)) = \lceil \log \frac{1}{p(x)} \rceil$.

Such codes exist because

$$\sum_{x \in \mathscr{X}} 2^{-\lceil \log \frac{1}{p(x)} \rceil} \leq \sum_{x \in \mathscr{X}} 2^{-\log \frac{1}{p(x)}} = 1.$$

For a Shannon code,

$$H(X) \leq \sum_{x \in \mathscr{X}} p(x)\ell(c(x)) \leq H(X) + 1.$$

So if we create a Shannon code on blocks of length n,

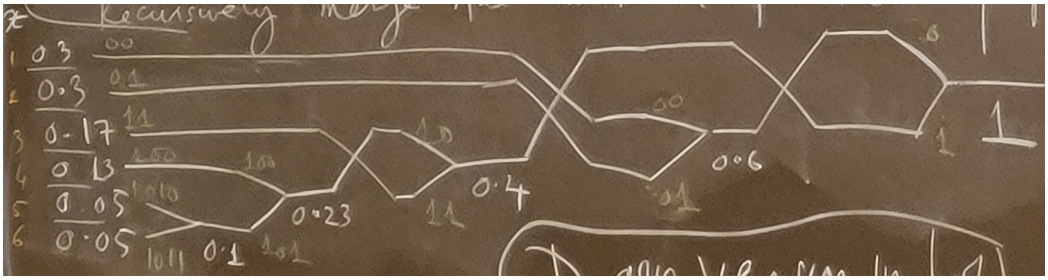
$$H(X_1, \ldots, X_n) \leq \sum_{x_1^n \in \mathscr{X}^n} p(x_1^n)\ell(c(x_1^n)) \leq H(X_1, \ldots, X_n) + 1,$$

with $X_i \sim p$. Dividing by n, the penalty is at most $\frac{1}{n}$ bits/symbol.

10.2 Huffman coding

It turns out that in this case, the integer programming problem can be solved exactly.[7] The **Huffman coding algorithm**, in one sentence, basically says to "Recursively merge the smallest probability pair of symbols."

Example 10.1. Draw the following diagram, successively merging the two smallest probabilities at each step:



Then label each branch with a 0 or a 1.

A D-ary version requires us to combine the smallest D probabilities at a time.

Theorem 10.1. *The Huffman code is optimal.*

Proof. Observe some properties that the solution to the integer problem must satisfy:

[7] This is unusual. Integer programming problems are generally very computationally difficult.

1. If $p(x) \geq p(y)$, then $\ell(x) \leq \ell(y)$.

   Proof. If not, interchange $\ell(x)$ and $\ell(y)$ to get a better code. □

2. There have to be at least two symbols $x, x' \in \mathscr{X}$ getting the longest length.

   Proof. If not, reducing the length of the longest codeword by removing a bit from the end gives a better code. □

Also, we can arrange the following:

3. There are two symbols $x, x' \in \mathscr{X}$ with longest length representations such that these representations differ only at the last bit, and these are the two smallest probability symbols.

   Proof. If the sibling of a longest length codeword is not present, we can reduce the length of that codeword by removing the last bit. To guarantee that these are the two smallest probability symbols, we can just relabel. □

This is enough to prove by induction that the Huffman coding algorithm is optimal. □

Example 10.2. A Shannon code can be worse than a Huffman code. One way to see this is to note that the expected length of a Huffman code is a continuous function on the probability simplex.

Now consider $\mathscr{X} = \{1, 2, 3, 4\}$ with

$$p(1) = \frac{1}{4}, \qquad p(2) = \frac{1}{4} + \varepsilon, \qquad p(3) = \frac{1}{4} - \varepsilon, \qquad p(4) = \frac{1}{4},$$

for small $\varepsilon > 0$. Here, the Shannon code is worse than the Huffman code because $\lceil \log \frac{1}{1/4-\varepsilon} \rceil = 3$.

10.3 Shannon-Fano-Elias Coding

Shannon-Fano-Elias coding is a precursor to **arithmetic coding** (which is widely used), allowing one to learn the statistics and "improve the code" as one goes along.[8] The idea comes from observing that if Z is a random variable with (invertible) CDF $F_Z(z) = \mathbb{P}(Z \leq z)$, then

$$\mathbb{P}(Z \leq F_Z^{-1}(u)) = F_Z(F_Z^{-1}(u)) = u, \qquad u \in [0, 1].$$

[8] This is similar to adaptive control in machine learning.

This shows that $F_Z(Z)$ has uniform distribution on $[0,1]$, when F_Z^{-1} is well-defined. This is not true in general but suggests creating codes based on F_X, where $X \in \mathscr{X}$ for a finite $\mathscr{X} \subseteq \mathbb{R}$.

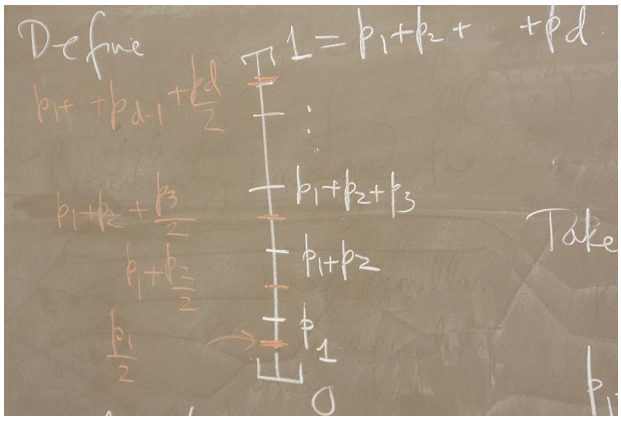
Suppose $|\mathscr{X}| = d$ with $\mathscr{X} = \{a_1, \ldots, a_d\}$, and write

$$\mathbb{P}(X = a_i) = p_i, \qquad 1 \le i \le d.$$

For $x \in \mathbb{R}$,

$$F_X(x) = \mathbb{P}(X \le x) = \sum_{i: a_i \le x} p_i.$$

Here is the idea of how the code works. Draw the values of the CDF as follows:



Now draw the midpoints between these points at $\frac{p_1}{2}, p_1 + \frac{p_2}{2}, \ldots$. Take the binary representation as a binary fraction of $p_1 + \cdots + p_{i-1} + \frac{p_i}{2}$. Then truncate it to get $\lceil \log \frac{1}{p_i} \rceil + 1$ bits. This will be prefix-free and within 2 bits of the entropy.

The arithmetic coding scheme is based on updating this procedure as we go.

11 Shanon-Fano-Elias, Arithmetic, and Lempel-Ziv Coding

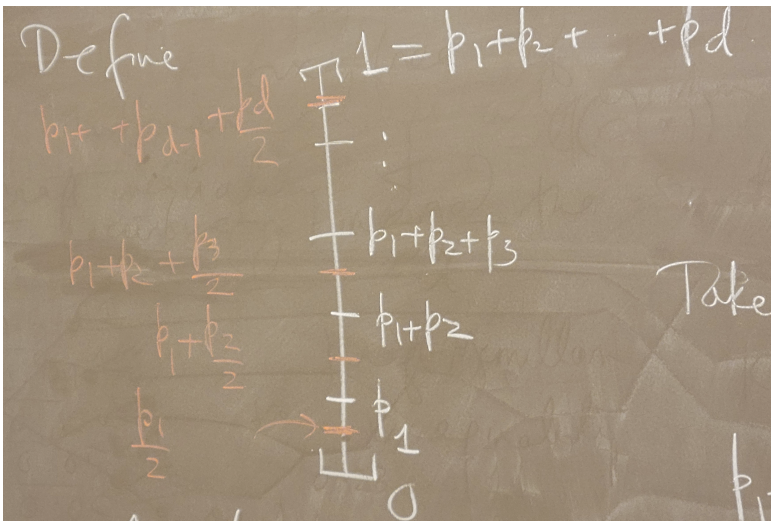
11.1 Shannon-Fano-Elias coding and arithmetic coding

Last time, we introduced the **Shannon-Fano-Elias coding** scheme. Suppose we have an alphabet $\mathscr{X}$ of size d, which we can assume are real numbers $a_1 < a_2 < \cdots < a_d$. If X is an $\mathscr{X}$-valued random variable, we can view it as a real-valued random variable. Then we have the Cumulative Distribution Function (CDF) of X:

$$F_X(x) = \mathbb{P}(X \leq x),$$

defined for all $x \in \mathbb{R}$.

Let $p_i := \mathbb{P}(X = a_i)$ for $i = 1, \ldots, d$, and label the values of the CDF on the interval $[0, 1]$ as follows.



Then label the midpoints of these values,

$$Q_1 = \frac{p_1}{2}, Q_2 = p_1 + \frac{p_2}{2}, \ldots, Q_d = \ldots, p_1 + \cdots + p_{d-1} + \frac{p_d}{2}.$$

Expand these in their binary representations, and truncate this representation to

$$\ell_i := \left\lceil \log \frac{1}{p_i} \right\rceil + 1$$

bits.

Example 11.1. Let $\mathscr{X} = \{1, 2, 3\}$ with $(p_1, p_2, p_3) = (1/4, 1/8, 5/8)$. Then the midpoints are

$$\frac{1}{8} = 0.001, \qquad \frac{3}{16} = 0.101, \qquad \frac{11}{16} = 0.1011.$$

Now we get the lengths

$$\ell_1 = \left\lceil \log \frac{1}{1/4} \right\rceil + 1 = 3,$$

$$\ell_2 = \left\lceil \log \frac{1}{1/8} \right\rceil + 1 = 3,$$

$$\ell_3 = \left\lceil \log \frac{1}{5/8} \right\rceil + 1 = 2.$$

So the Shannon-Fano-Elias code is

$$1 \mapsto 001, \qquad 2 \mapsto 010, \qquad 3 \mapsto 10.$$

Proposition 11.1. *This scheme gives a prefix-free code for* $(p_i, i \in \mathscr{X})$.

Proof. Think of the interval $[p_i + \cdots + p_{i-1}, p_1 + \cdots + p_i)$ as being *owned* by the symbol i, where the left endpoint is 0 for $i = 1$. Q_i is the midpoint of this interval. Let $(Q_i)_{\ell_i}$ denote the ℓ_i-truncation of the binary representation of Q_i. Observe that

$$Q_i - \sum_{j=1}^{i-1} p_j = \frac{p_i}{2} \geq \frac{1}{2^{\ell_i}}$$

because $\ell_i \geq \log \frac{1}{p_i} + 1$, so $2^{\ell_i - 1} \geq \frac{1}{p_i}$, which gives $\frac{p_i}{2} \geq \frac{1}{2^{\ell_i}}$.

Suppose that

$$(Q_i)_{\ell_i} = 0.b_1 b_2 \cdots b_{\ell_i}.$$

Consider

$$\left[0.b_1 b_2 \cdots b_{\ell_i}, 0.b_1 \cdots b_{\ell_i} + \frac{1}{2^{\ell_i}}\right).$$

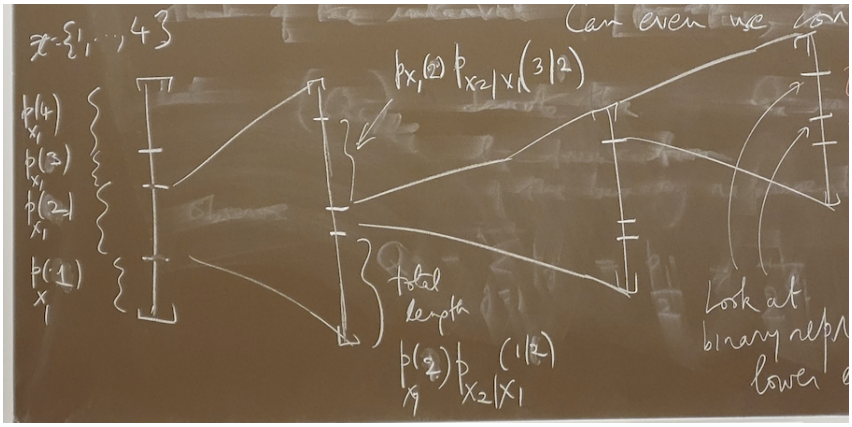
The binary representations of the real numbers in this interval are the continuations of the binary tree of the string $b_1 \cdots b_{\ell_i}$, i.e. those that might violate the prefix-free condition. But the inequality $Q_i - \sum_{j=1}^{i-1} p_j \geq 1/2^{\ell_i}$ tells us that this interval falls inside the interval $[p_i + \cdots + p_{i-1}, p_1 + \cdots + p_i)$ owned by the symbol i. Hence, no $(Q_k)_{\ell_k}$ for $k \neq i$ can have $(Q_i)_{\ell_i}$ as a prefix. □

Observe that we also get

$$\sum_{i=1}^{d} p_i \left\lceil \log \frac{1}{p_i} \right\rceil \leq H(X) + 2.$$

We can now amortize the 2 over multiple symbols (say n symbols) in the naive way by implementing a scheme on $\mathscr{X}^n$ (instead of $\mathscr{X}$). More interesting is to break $[\sum_{j-1}^{i-1} p_j, \sum_{j=1}^{i} p_j)$

itself into d subintervals and so on, iteratively. We can even use conditional probabilities for this.



Look at the binary representations of the upper and lower ends, and spit out the prefix they agree on.

11.2 Lempel-Ziv coding and comma-free coding of natural numbers

Imagine you've seen an infinite number of symbols from an ergodic source, and you see a new symbol. How will you compress the new symbol? We want to represent a symbol in context of its past. This is leading up to the Lempel-Ziv coding scheme from 1977.

First, we need a "comma free" representation of natural numbers. Imagine a transmitter has a natural number and needs to send a bit string from which the receiver can figure out this integer. Here is a solution to this problem due to Elias.[9]

The integer n can be written as a bit string of length $\tau(n)$ (so $2^{\tau(n)} \geq n$). Then $\tau(n)$ can be written as a bit string of length $\phi(n)$ (so $2^{\phi(n)} \geq \tau(n)$). We could go on like this for a few stages, but for us this is good enough. To send $\phi(n)$, send

$$\underbrace{000\cdots0}_{\phi(n) \text{ zeroes}}\,1.$$

The number of bits used in this scheme is

$$\begin{aligned} 2\phi(n) + 1 + \tau(n) &\leq \lceil \log n \rceil + 2\lceil \log \lceil \log n \rceil \rceil + 1 \\ &\leq \log n + 2 \log \log n + 5. \end{aligned}$$

The key idea in LZ'77 is to compress new samples in the context of the infinite past by transmitting how far back in the past to look to see the current samples in the context of

[9] There are many solutions. You can try to figure one out for yourself.

its own past. This works for a general stationary, ergodic process, but to understand why this works, we will consider the iid case, working with blocks. Suppose we have

$$(\dots, x_{-2}, x_{-1}, x_0, x_1, \dots, x_{k-1}, \dots),$$

where $x_{\leq -1}$ is to be shared with the receiver, and $x_0, x_1, \dots, x_{k-1}$ is to be conveyed. Find

$$\inf\{L \geq l : (x_0, \dots, x_{k-1}) = (x_{-L} x_{-L+1}, \dots, x_{-L+k-1})\}.$$

Let $p(a_0^{k-1}) := \mathbb{P}((X_0, \dots, X_{k-1}) = a_0^{k-1})$. We can show that the comma–free encoding needs $\leq (1+\varepsilon)H(X_1)$ bits. We will do this next time.

12 Lempel-Ziv Coding for Ergodic Processes

12.1 Intuition behind Lempel-Ziv coding

Last time, we discussed a comma-free binary representation of natural numbers using $\log n + 2\log\log n + k$ bits ($k = 5$). To send n, send $\lceil \log n \rceil$ bits (tells us $n \in 1, \ldots, 2^{\lceil \log n \rceil}$). To send $\lceil \log n \rceil$, send $\lceil \log \lceil \log n \rceil \rceil$ bits (same idea). Send $\lceil \log \lceil \log n \rceil \rceil$ as $\lceil \log \lceil \log n \rceil \rceil$ 0s, followed by a 1.

Example 12.1. To send $n = 17$, we have $\lceil \log n \rceil = 5$ and $\lceil \log 5 \rceil = 3$. Then transmit

$$0001 \quad \underbrace{101}_{\text{represents } 5} \quad 10001.$$

Example 12.2. To send $n = 14$, we have $\lceil \log n \rceil = 4$ and $\lceil \log 5 \rceil = 2$. Then transmit

$$00011001110,$$

which can be parsed as

$$0001 \quad 100 \quad 1110.$$

To motivate the LZ'77 scheme (which compresses to the entropy rate for any stationary ergodic process), let's consider i.i.d.

$$\ldots, X_{-2}, X_{-1}, X_0, X_1, X_2, \ldots$$

at the level of blocks of size L. The situation is that $\ldots, X_{-3}, X_{-2}, X_{-1}$ is common knowledge to the compressor and decompressor (or the transmitter and receiver). We need to send $(X_0, X_1, \ldots, X_{L-1})$. We do this by finding

$$\inf\{m \geq 1 : (X_0, X_1, \ldots, X_{L-1}) = (X_{-mL} X_{-mL+1}, \ldots, X_{-mL+L-1}\}$$

and sending m using the comma-free encoding of $\mathbb{N}$. Since the blocks of length L of the type $(X_{-jL}, X_{-jL+1}, \ldots, X_{-jL+L-1})$ are independent, m will be geometrically distributed, conditioned on $(X_0, X_1, \ldots, X_{L-1})$. Then

$$\mathbb{P}(m = j \mid (X_0, \ldots, X_{L-1}) = x_0^{L-1}) = p(x_0^{L-1})(1 - p(x_0^{L-1}))^{j-1}, \qquad j = 1, 2, \ldots.$$

So the conditional expectation on this event is

$$\mathbb{E}[m \mid (X_0, \ldots, X_{L-1}) = x_0^{L-1}] = \frac{1}{p(x_0^{L-1})}.$$

Also, for all x_0^{L-1},

$$\mathbb{P}(m > \widetilde{K} \frac{1}{p(x_0^{L-1})} \mid X_0^{L-1} = x_0^{L-1}) = \sum_{j = \lceil \widetilde{K} \frac{1}{p(x_0^{L-1})} \rceil} p(x_0^{L-1})(1 - p(x_0^{L-1}))^{j-1}$$

$$\leq (1 - p(x_0^{L-1}))^{\lceil \widetilde{K}(1/p(x_0^{L-1}))\rceil - 1}$$
$$\lesssim e^{-\widetilde{K}}$$

as $L \to \infty$.

The upshot is that we can, with probability close to 1, convey m with $\log \frac{\widetilde{K}}{p(x_0^{L-1})} + \log\log \frac{\widetilde{K}}{p(x_0^{L-1})} + k$ bits (conditioned on $X_0^{L-1} = x_0^{L-1}$) for any $\widetilde{K}$, as $L \to \infty$. Note that

$$\sum_{x_0^{L-1}} p(x_0^{L-1}) \left(\log \frac{\widetilde{K}}{p(x_0^{L-1})} + \log\log \frac{\widetilde{K}}{p(x_0^{L-1})} + k \right) \asymp H(X_0, \ldots, X_{L-1})$$

as $L \to \infty$.

12.2 Ergodicity and Kac's theorem

Definition 12.1. A two-sided process $(X_n, n \in \mathbb{Z})$ with $X_n \in \mathscr{X}$ for finite $\mathscr{X}$ is **ergodic** if

1. The process is stationary.
2. Every shift-invariant event should have probability 0 or probability 1.

By **shift-invariant**, we mean

$$\{(\ldots, X_{-1}, X_0, X_1, \ldots) \in A\} = \{(\ldots, X_{-2}, X_{-1}, X_0, \ldots) \in A\}.$$

Shift-invariant events can be very interesting.

Example 12.3. The event {there are infinitely many 1s in the sequence} is shift-invariant.

Example 12.4. The event {the lim sup of the sequence is 1} is shift-invariant.

Theorem 12.1 (Pointwise ergodic theorem, Birkhoff)**.** *If $(X_n, n \in \mathbb{Z})$ is ergodic and $\phi : \mathscr{X}^k \to \mathbb{R}$, then*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} \phi(X_t, X_{t+1}, \ldots, X_{t+k-1}) = \mathbb{E}[\phi(X_0, \ldots, X_{k-1})]$$

almost surely.

To look back in the past in the general ergodic case, we use the following theorem:

Theorem 12.2 (Kac)**.** *Let $(X_n, n \in \mathbb{Z})$ be an ergodic process with $X_n \in \mathscr{X}$ for all n, where $\mathscr{X}$ is finite. Let*

$$Q_b(i) = \mathbb{P}(X_{-i} = b, X_j \neq b \text{ for } -i+1 \leq j \leq -1 \mid X_0 = b).$$

Then

$$\sum_{i=1}^{\infty} iQ_b(i) = \frac{1}{\mathbb{P}(X_0 = b)}.$$

Proof. Fix $b \in \mathscr{X}$. Define the events

$$A_{j,k} := \{X_{-j} = b, X_{-j+1} \neq b, \ldots, X_{k-1} \neq b, X_k = b\}, \qquad k \geq 0, j \geq 1.$$

These events are disjoint. We claim that

$$\mathbb{P}\left(\bigcup_{j,k} A_{j,k}\right) = 1$$

if $\mathbb{P}(X_0 = b) > 0$. This is because b occurs some finite time in the future and some time in the past; we can see this from, for example, looking at the sample averages of the ergodic theorem with ϕ as the indicator of $\{b\}$.

Hence,

$$\sum_{j=1}^{\infty}\sum_{k=0}^{\infty} \mathbb{P}(A_{j,k}) = 1.$$

But this equals

$$\sum_{j=1}^{\infty}\sum_{k=0}^{\infty} \mathbb{P}(X_k = b)Q_b(j+k) = \mathbb{P}(X_0 = b)\sum_{i=0}^{\infty} iQ_b(i)$$

because $\mathbb{P}(X_k = b) = \mathbb{P}(X_0 = b)$ by stationarity and because the number of ways to get $j + k = i$ is i. □

Now for LZ'77, assume that $(X_n, n \in \mathbb{Z})$ is an ergodic process. For any fixed $L \geq 1$, define

$$R_L(X_0, X_1, \ldots, X_{L-1}) := \min\{j \geq 1 : (X_{-j}, X_{-j+1}, \ldots, X_{-j+L-1}) = (X_0, \ldots, X_{L-1})\}.$$

By Kac's theorem,

$$\mathbb{E}[R_L(X_0, X_1, \ldots, X_{L-1}) \mid X_0^{L-1} = x_0^{L-1}] = \frac{1}{p(x_0^{L-1})}.$$

The transmitter will send $R_L(X_0, X_1, \ldots, X_{L-1})$ by comma-free encoding (in order to convey X_0). Let

$$\lambda_L(x_0^{L-1}) = \log R_L(X_0^{L-1}) + \log\log R_L(X_0^{L-1}) + 5.$$

Next time, we will show that

$$\frac{1}{L}\,\mathbb{E}[\lambda_L(X_0^{L-1})] \xrightarrow{L\to\infty} H,$$

the entropy rate of the process.

13 Optimality of Lempel-Ziv Coding, The Burrows-Wheeler Transform, and Optimal Compression of IID Sequences

13.1 Asymptotic optimality of Lempel-Ziv coding

Last time, we were in the discussing LZ'77 for a general ergodic process $(X_n, n \in \mathbb{Z})$ with $X_n \in \mathscr{X}$ (finite). For any $L \geq 1$, we defined

$$R_L(X_0^{L-1}) := \min\{j \geq 1 : X_{-j}^{-j+L-1} = X_0^{L-1}\}.$$

The compressor conveys $R_L(X_0^{L-1})$ to the decompressor. The compressor knows $(X_n, n \leq -1)$ and X_0^{L-1}; the decompressor only knows the past, $(X_n, n \leq -1)$. This suffices for the decompressor to determine X_0^{L-1}. By comma-free encoding, it suffices to send

$$\log R_L(X_0^{L-1}) + \log\log R_L(X_0^{L-1}) + 5$$

many bits. Then

$$\begin{aligned}
\mathbb{E}[R_L(X_0^{L-1})] &\leq \log \mathbb{E}[R_L(X_0^{L-1})] \\
&= \log \frac{1}{p(X_0^{L-1})}, \\
&= H(X_0, \ldots, X_{L-1})
\end{aligned}$$

by Kac's lemma. So for fixed L,

$$\frac{1}{L}\,\mathbb{E}[\log R_L(X_0^{L-1})] \leq \frac{1}{L} H(X_0^{L-1}).$$

So

$$\limsup_{L\to\infty} \frac{1}{L}\,\mathbb{E}[\log R_L(X_0^{L-1})] \leq H,$$

the entropy rate. Also,

$$\begin{aligned}
\frac{1}{L}\,\mathbb{E}[\log\log R_L(X_0^{L-1})] &\leq \frac{1}{L}\log \mathbb{E}[\log R_L(X_0^{L-1})] \\
&\leq \frac{1}{L}\log H(X_0^{L-1}) \\
&\xrightarrow{L\to\infty} 0.
\end{aligned}$$

Finally, $5/L \to 0$ as $L \to \infty$, as well. So in total, we get

$$\frac{1}{L}\,\mathbb{E}[\log R_L(X_0^{L-1}) + \log\log R_L(X_0^{L-1}) + 5] \xrightarrow{L\to\infty} H.$$

13.2 The Burrows-Wheeler transform

Here is an algorithm that some people claim works better than the Lempel-Ziv coding scheme.

Example 13.1. To compress the string SHANNON, a string from the English alphabet, we'll consider all the cyclic permutations and lexicographically order them:

$$\begin{array}{ll} \text{SHANNON} & \text{ANNONSH} \\ \text{HANNONS} & \text{HANNONS} \\ \text{ANNONSH} & \text{NNONSHA} \\ \text{NNONSHA} \quad \mapsto & \text{NONSHAN} \\ \text{NONSHAN} & \text{NSHANNO} \\ \text{ONSHANN} & \text{ONSHANN} \\ \text{NSHANNO} & \text{SHANNON} \end{array}$$

Transmit the last column (in compressed form) and the number of the row that has the empirical string. The decompressor (after decompression) gets HSANONN and the number 7.

The decompressor can now recover the first column by lexicographically ordering the symbols (because each symbol in the last column shows up the same number of times it does in the original string). Then, the decompressor knows a list of pairs of symbols (the first and last symbol of each row). Using this, the decompressor can now figure out the second column by cyclically permuting these pairs and lexicographically ordering them, and so on. In this way, the decompressor can recover the original string.

Why does this compress the message? Compressing the last column can be done by e.g. arithmetic coding and works to compress down to the entropy rate for sequences from an ergodic process because (as the length of the sequence goes to infinity, and for each fixed L), the last column becomes piecewise iid with $|\mathscr{X}|^L$ pieces. The piece for x_0^{L-1} appears (asymptotically in n) $np(x_0^{L-1})$ times and has marginal with law $p(x \mid x_0^{L-1})$. Here,

$$H(X_L \mid X_0^{L-1}) = \sum_{x_0^{L-1}} p(x_0^{L-1}) H(p(x \mid x_0^{L-1}), x \in \mathscr{X})$$

is just the L-Markov approximation to the entropy rate. So we can compress to the entropy rate as $L \to \infty$.

13.3 Compression of iid sequences at rate R bits/symbol

Leading up to distributed data compression, we will first discuss the fixed length to fixed length (fixed-to-fixed) formulation of point to point data compression. To recognize the relevance of entropy, we need to allow for a probability of error in decompression (which becomes vanishingly small as the block length increases).

Definition 13.1. We'll say that compression can be done **at rate R bits/symbol** if there is a sequence of pairs of maps $((e_n, d_n) : n \geq 1)$ where $e_n : \mathscr{X}^n \to [M_n] := \{1, \ldots, M_n\}$ and $d_n : [M_n] \to \mathscr{X}^n$ such that

$$\limsup_{n\to\infty} \frac{1}{n} M_n \leq R$$

and

$$\lim_{n\to\infty} \mathbb{P}(d_n(e_n(X_1^n)) \neq X_1^n) = 0.$$

Theorem 13.1. *Let $X_1, X_2, \ldots$ be iid $\mathscr{X}$-valued with entropy rate H. Then compression can be done at rate H and cannot be done at any rate $< H$.*

Remark 13.1. This theorem is also true for arbitrary stationary sequences, but we will not prove that here.

Proof. Achievability: First, observe that it's enough to show that for all $\varepsilon > 0$, compression can be done at rate $H + \varepsilon$; this is because we can take $\varepsilon = 1/m$ for large enough n (depending on m). Recall that $A_\delta^{(n)}$ denotes the set of weakly ε-typical sequences. We know that $|A_\delta^{(n)}| \leq 2^{n(H+\delta)}$ and

$$\lim_{n\to\infty} \mathbb{P}(X_1^n \in A_\delta^{(n)}) = 1.$$

So if $e_n : \mathscr{X}^n \to [\lceil 2^{n(H+\delta)} \rceil + 1]$ gives a unique image to each element of $A_\delta^{(n)}$ and maps $(A_\delta^{(n)})^c$ to a single image, then

$$\limsup_{n\to\infty} \frac{1}{n} \log(\lceil 2^{n(H+\delta)} \rceil + 1) = H + \delta,$$

and

$$\mathbb{P}(d_n(e_n(X_1^n)) \neq X_1^n) \leq \mathbb{P}(X_1^n \in (A_\delta^{(n)})^c) \xrightarrow{n\to\infty} 0.$$

So take $\delta = \varepsilon$.

Converse: Given any $((e_n, d_n), n \geq 1)$, denote $W_n = e_n(X_1^n)$ and $\widehat{X}_1^n = d_n(W_n)$. Then we have the Markov chain $X_1^n - W_n - \widehat{X}_1^n$. We get from Fano's inequality that

$$H(X_1^n \mid W_n) \leq \underbrace{\mathbb{P}(d_n(e_n(X_1^n)) \neq \widehat{X}_1^n)}_{p_{\text{error}}^{(n)}} \log |\mathscr{X}| + \underbrace{h(p_{\text{error}}^{(n)})}_{\leq 1}$$

So if $p_{\text{error}}^{(n)} \to 0$ as $n \to \infty$, then

$$\frac{1}{n} H(X_1^n \mid W_n) \to 0$$

as $n \to \infty$. But

$$\frac{1}{n} H(X_1^n, W_n) = \frac{1}{n} H(W_n) + \frac{1}{n} H(X_1^n \mid W_n) \to 0,$$

and

$$\frac{1}{n}H(X_1^n, W_n) = \frac{1}{n}H(X_1^n) + \frac{1}{n}H(W_n \mid X_1^n).$$

So

$$\liminf_{n\to\infty} \frac{1}{n}H(W_n) \geq H(X_1),$$

which is the entropy rate of the iid sequence. Hence,

$$\liminf_{n} \frac{1}{n}\log M_n \geq H(X_1). \qquad \square$$

14 Joint ε-Weak Typicality and the Slepian-Wolf Theorem

14.1 Properties of joint ε-weak typicality

Suppose $(X_1, Y_1), (X_2, Y_2), \ldots$ are i.i.d. with $(X_i, Y_i) \in \mathscr{X} \times \mathscr{Y}$ finite and $(X_i, Y_i) \sim (p(x, y), x \in \mathscr{X}, y \in \mathscr{Y})$. We think of the X_is as being seen by Alice and the Y_is as being seen by Bob.

Definition 14.1 (Joint ε-weak typicality)**.** Define the set $A_\varepsilon^{(n)} \subseteq \mathscr{X}^n \times \mathscr{Y}^n$ to be the set of $(x_1^n, y_1^n) \in \mathscr{X}^n \times \mathscr{Y}^n$ such that

1. $| -\frac{1}{n} \log p(x_1^n) - H(X)| < \varepsilon$,
2. $| -\frac{1}{n} \log p(y_1^n) - H(Y)| < \varepsilon$,
3. $| -\frac{1}{n} \log p(x_1^n, y_1^n) - H(X, Y)| < \varepsilon$.

Here are some properties of this:

Theorem 14.1.

1.

$$\mathbb{P}((X_1^n, Y_1^n) \in A_\varepsilon^{(n)}) \xrightarrow{n\to\infty} 1.$$

Proof. Use the weak law of large numbers. □

2.

$$|A_\varepsilon^{(n)}| \leq 2^{nH(X,Y)} 2^{n\varepsilon}.$$

Proof. For all $(x_1^n, y_1^n) \in A_\varepsilon^{(n)}$,

$$p(x_1^n, y_1^n) \geq 2^{-nH(X,Y)} 2^{-n\varepsilon}$$

and

$$1 \geq \sum_{(x_1^n, y_1^n) \in A_\varepsilon^{(n)}} p(x_1^n, y_1^n).$$ □

3. For all large enough n,

$$|A_\varepsilon^{(n)}| \geq (1 - \delta) 2^{nH(X,Y)} 2^{-n\varepsilon}.$$

Proof. For all $(x_1^n, y_1^n) \in A_\varepsilon^{(n)}$,

$$p(x_1^n, y_1^n) \le 2^{-nH(X,Y)} 2^{n\varepsilon}$$

and, for all large enough n,

$$\sum_{(x_1^n, y_1^n) \in A_\varepsilon^{(n)}} p(x_1^n, y_1^n) \ge 1 - \delta. \qquad \square$$

4. *If* $(\widetilde{X}_1^n, \widetilde{Y}_1^n) \sim p(x_1^n)p(y_1^n)$, *then*

   (a)

   $$\mathbb{P}((\widetilde{X}_1^n, \widetilde{Y}_1^n) \in A_\varepsilon^{(n)}) \le 2^{-nI(X;Y)} 2^{3n\varepsilon}.$$

   Proof. The left hand side is

   $$\begin{aligned} \sum_{(x_1^n, y_1^n)} p(x_1^n)p(y_1^n) &\le |A_\varepsilon^{(n)}| 2^{-nH(X)} 2^{n\varepsilon} 2^{-nH(Y)} 2^{n\varepsilon} \\ &\le 2^{nH(X,Y)} 2^{-nH(X)} 2^{-nH(Y)} 2^{3n\varepsilon}. \end{aligned} \qquad \square$$

   (b) For all $\delta > 0$,

   $$\mathbb{P}((\widetilde{X}_1^n, \widetilde{Y}_1^n) \in A_\varepsilon^{(n)}) \ge (1 - \delta) 2^{-nI(X;Y)} 2^{-3n\varepsilon}.$$

   Proof. The left hand side is

   $$\begin{aligned} \sum_{(x_1^n, y_1^n)} p(x_1^n)p(y_1^n) &\ge |A_\varepsilon^{(n)}| 2^{-nH(X)} 2^{-n\varepsilon} 2^{-nH(Y)} 2^{-n\varepsilon} \\ &\ge (1 - \delta) 2^{nH(X,Y)} 2^{-nH(X)} 2^{-nH(Y)} 2^{-3n\varepsilon}. \end{aligned} \qquad \square$$

14.2 The Slepian-Wolf theorem on distributed lossless compression

In this section, lossless is interpreted in the sense of asymptotically vanishing error probability. The scenario is that Alice sees $X_1, \ldots, X_n$ and Bob sees $Y_1, \ldots, Y_n$. The pairs (X_i, Y_i) with $i = 1, \ldots, n$ are iid and $(X_i, Y_i) \sim (p(x, y), x \in \mathscr{X}, y \in \mathscr{Y})$. Alice compresses X_1^n, and Bob compresses Y_1^n. A fusion centor sees the compressed representations and wants to recover (X_1^n, Y_1^n) with small probability of error (going to 0 as $n \to \infty$). The problem is: What region of (Alice's bits/symbol, Bob's bits/symbol) is achievable?

Definition 14.2. We say that the pair of rates (R_1, R_2) is **achievable** if there is a sequence $((e_n^{(1)}, e_n^{(2)}, d_n), n \ge 1)$ where

$$e_n^{(1)} : \mathscr{X}^n \to [M_n^{(1)}] = \{1, \ldots, M_n^{(1)}\}, \quad \text{with} \quad \limsup_{n\to\infty} \frac{1}{n} \log M_n^{(1)} \le R_1,$$

$$e_n^{(2)} : \mathscr{X}^n \to [M_n^{(2)}], \quad \text{with} \quad \limsup_{n\to\infty} \frac{1}{n} \log M_n^{(2)} \le R_2,$$

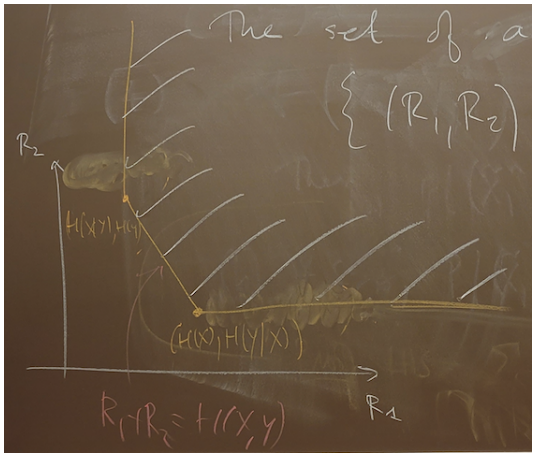
$$d_n : [M_n^{(1)}] \times [M_n^{(2)}] \to \mathscr{X}^n \times \mathscr{Y}^n,$$

such that

$$\mathbb{P}(d_n(e_n^{(1)}(X_1^n), e_n^{(2)}(Y_1^n)) \neq (X_1^n, Y_1^n)) \xrightarrow{n\to\infty} 0.$$

Theorem 14.2 (Slepian-Wolf)**.** *The set of achievable rate pairs is*

$$\{(R_1, R_2) : R_1 \ge H(X \mid Y), R_2 \ge H(Y \mid X), R_1 + R_2 \ge H(X, Y)\}.$$



We will prove the achievability using the *probabilistic method*; i.e. we will show that a suitable $((e_n^{(1)}, e_n^{(2)}, d_n), n \ge 1)$ exists without explicitly demonstrating it. Here is an example of the probabilistic method.

Example 14.1. Suppose that $f : [0,1] \to \mathbb{R}_+$. To show "there exists some x such that $f(x) > 10$," it's enough to show that $\mathbb{E}[f(Z)] > 10$ where $Z \sim \text{Unif}([0,1])$.

Proof. Achievability: It is enough to show that for all $\varepsilon > 0$, if (R_1, R_2) is such that $R_1 \ge H(X \mid Y) + \varepsilon$, $R_2 \ge H(Y \mid X) + \varepsilon$, and $R_1 + R_2 \ge H(X, Y) + \varepsilon$, then (R_1, R_2) is achievable. We use a "random binning" argument: $(e_n^{(1)}, e_n^{(2)}, d_n)$ will be random variables with

- $e_n^{(1)}$: randomly assign each $x_1^n \in \mathscr{X}^n$ to one of $M_n^{(1)}$ bins uniformly, independently over x_1^n,
- $e_n^{(2)}$: randomly assign each $y_1^n \in \mathscr{Y}^n$ to one of $M_n^{(2)}$ bins uniformly, independently over y_1^n

- $d_n(m_n^{(1)}, m_n^{(2)}) = (\widehat{x}_1^n, \widehat{x}_2^n)$ if there is exactly one $(\widehat{x}_1^n, \widehat{y}_1^n)$ with $e_n^{(1)}(\widehat{x}_1^n) = m_n^{(1)}$ and $e_n^{(2)}(\widehat{y}_1^n) = m_n^{(2)}$. Otherwise, $d_n(m_n^{(1)}, m_n^{(2)})$ can take any value.

Now we upper bound $\mathbb{P}(d_n(e_n^{(1)}(X_1^n), e_n^{(2)}(Y_1^n)) \neq (X_1^n, Y_1^n))$, where the randomness is in both (X_1^n, Y_1^n) and $(e_n^{(1)}, e_n^{(2)}, d_n)$. We have

$$\mathbb{P}(d_n(e_n^{(1)}(X_1^n), e_n^{(2)}(Y_1^n)) \neq (X_1^n, Y_1^n)) \leq \underbrace{\mathbb{P}(E_{0,n})}_{\xrightarrow{n\to\infty} 0} + \mathbb{P}(E_{1,n}) + \mathbb{P}(E_{2,n}) + \mathbb{P}(E_{12,n}).$$

Here,

- $E_{0,n} = \{(X_1^n, Y_1^n) \notin A_n^{(\delta)}\}$ for some $\delta > 0$, and the corresponding probability goes to 0 as $n \to \infty$.
- $E_{1,n} = \{\exists\, \widetilde{x}_1^n \neq X_1^n$ with $e_n^{(1)}(\widetilde{x}_1^n) = e_n^{(1)}(X_1^n)$ and $(\widetilde{x}_1^n, y_1^n) \in A_n^{(\delta)}\}$. Here,

$$\mathbb{P}(E_{1,n}) \leq \sum_{(x_1^n, y_1^n)} p(x_1^n, y_1^n) \sum_{\substack{\widetilde{x}_1^n \neq x_1^n \\ (\widetilde{x}_1^n, y_1^n) \in A_n^{(\delta)}}} \underbrace{\mathbb{P}(e_n^{(1)}(\widetilde{x}_1^n) = e_n^{(1)}(x_1^n))}_{=1/M_n^{(1)}}.$$

  Now $|\{\widetilde{x}_1^n : (\widetilde{x}_1^n, y_1^n) \in A_n^{(\delta)}\}| \leq 2^{nH(X|Y)} 2^{2n\delta}$ because

$$\begin{aligned} 1 &\geq \sum_{\widetilde{x}_1^n : (\widetilde{x}_1^n, y_1^n) \in A_n^{(\delta)}} p(\widetilde{x}_1^n \mid y_1^n) \\ &= \sum_{\widetilde{x}_1^n : (\widetilde{x}_1^n, y_1^n) \in A_n^{(\delta)}} \frac{p(\widetilde{x}_1^n, y_1^n)}{p(y_1^n)} \\ &\geq |\{\widetilde{x}_1^n : (\widetilde{x}_1^n, y_1^n) \in A_n^{(\delta)}\}| 2^{-nH(X|Y)} 2^{-2n\delta}. \end{aligned}$$

  So

$$\mathbb{P}(E_{1,n}) \leq \sum_{(x_1^n, y_1^n)} p(x_1^n, y_1^n) 2^{nH(X|Y)} 2^{2n\delta} 2^{-nR_1}.$$

  But $R_1 > H(X \mid Y) + \varepsilon$ by assumption, so if $2\delta < \varepsilon$, the right hand side goes to 0 as $n \to \infty$.

- $E_{2,n}$ is defined similarly to $E_{1,n}$, and $\mathbb{P}(E_{2,n}) \to 0$ as $n \to \infty$.

We are now left with $\mathbb{P}(E_{12,n})$, which we will examine next time. $\square$

15 Proof of the Slepian-Wolf Theorem and Introduction to Channel Coding

15.1 Proof of the Slepian-Wolf theorem

Last time, we were proving the Slepian-Wolf theorem. We had an iid sequence of pairs $(X_i, Y_i) \sim (p(x,y), x \in \mathscr{X}, y \in \mathscr{Y})$. Alice and Bob had respective encoding maps

$$e_n^{(1)} : \mathscr{X}^n \mapsto [M_n^{(1)}],$$

$$e_n^{(2)} : \mathscr{Y}^n \mapsto [M_n^{(2)}],$$

and a fusion center tries to decode the pairs of messages using the decoding maps

$$d_n : [M_n^{(1)}] \times [M_n^{(2)}] \to \mathscr{X}^n \times \mathscr{Y}^n.$$

We called the rate pair (R_1, R_2) **achievable** if there exist $((e_n^{(1)}, e_n^{(2)}, d_n), n \geq 1)$ such that

$$\limsup_n \frac{1}{n} \log M_n^{(1)} \leq R_1,$$

$$\limsup_n \frac{1}{n} \log M_n^{(2)} \leq R_2,$$

$$\lim_{n\to\infty} \mathbb{P}(d_n(e_n^{(1)}(X_1^n), e_n^{(2)}(Y_1^n)) \neq (X_1^n, Y_1^n)) = 0.$$

Theorem 15.1 (Slepian-Wolf)**.** *The set of achievable rate pairs is*

$$\{(R_1, R_2) : R_1 \geq H(X \mid Y), R_2 \geq H(Y \mid X), R_1 + R_2 \geq H(X,Y)\}.$$

We set up the proof of achievability using a random binning argument.

Proof. Achievability: By a diagonal-type argument, it suffices to consider (R_1, R_2) such that $R_1 > H(X \mid Y) + \varepsilon$, $R_2 > H(Y \mid X) + \varepsilon$, and $R_1 + R_2 > H(X,Y) + \varepsilon$. The idea is to let $M_n^{(1)} = \lceil 2^{nR_1} \rceil$ and $M_n^{(2)} = \lceil 2^{nR_2} \rceil$. Define random $e_n^{(1)}$ and $e_n^{(2)}$ via:

- $e_n^{(1)}$ randomly assigns each $x_1^n \in \mathscr{X}^n$ to one of $M_n^{(1)}$ bins uniformly, indepndently over x_1^n,
- $e_n^{(2)}$ randomly assigns each $y_1^n \in \mathscr{Y}^n$ to one of $M_n^{(2)}$ bins uniformly, indepndently over y_1^n
- $d_n(m_n^{(1)}, m_n^{(2)}) = (\widehat{x}_1^n, \widehat{x}_2^n)$ if there is exactly one $(\widehat{x}_1^n, \widehat{y}_1^n) \in A_\delta^{(n)}$ with $e_n^{(1)}(\widehat{x}_1^n) = m_n^{(1)}$ and $e_n^{(2)}(\widehat{y}_1^n) = m_n^{(2)}$. Otherwise, $d_n(m_n^{(1)}, m_n^{(2)})$ can take any value.

We have the probability (over randomness in (X_1^n, Y_1^n) and in $(e_n^{(1)}, e_n^{(2)})$)

$$\mathbb{P}(d_n(e_n^{(1)}(X_1^n), e_n^{(2)}(Y_1^n)) \neq (X_1^n, Y_1^n)) \leq \mathbb{P}(E_{0,n}) + \mathbb{P}(E_{1,n}) + \mathbb{P}(E_{2,n}) + \mathbb{P}(E_{12,n}),$$

where

$$E_{0,n} = \{(X_1^n, Y_1^n) \notin A_\delta^{(n)}\},$$
$$E_{1,n} = \{\exists\, \widetilde{x}_1^n \neq X_1^n \text{ with } e_n^{(1)}(\widetilde{x}_1^n) = e_n^{(1)}(X_1^n) \text{ and } (\widetilde{x}_1^n, y_1^n) \in A_n^{(\delta)}\},$$
$$E_{2,n} = \{\exists\, \widetilde{x}_1^n \neq X_1^n \text{ with } e_n^{(1)}(\widetilde{x}_1^n) = e_n^{(1)}(X_1^n) \text{ and } (\widetilde{x}_1^n, y_1^n) \in A_n^{(\delta)}\},$$
$$E_{12,n} = \{\exists\, \widetilde{y}_1^n \neq Y_1^n \text{ with } e_n^{(2)}(\widetilde{y}_1^n) = e_n^{(2)}(Y_1^n) \text{ and } (x_1^n, \widetilde{y}_1^n) \in A_n^{(\delta)}\},$$

$$E_{12,n} = \{\exists\, (\widetilde{x}_1^n, \widetilde{y}_1^n) \text{ s.t. } \widetilde{x}_1^n \neq X_1^n, \widetilde{y}_1^n \neq Y_1^n,$$
$$e_n^{(1)}(\widetilde{x}_1^n) = e_n^{(1)}(X_1^n), e_n^{(2)}(\widetilde{y}_1^n) = e_n^{(2)}(Y_1^n), (\widetilde{x}_1^n, \widetilde{x}_1^n) \in A_\delta^{(n)}\}.$$

We saw that the probabilities of the first three events goes 0 to as $n \to \infty$ if we pick $2\delta < \varepsilon$. It remains to show that $\mathbb{P}(E_{12,n}) \to 0$ as $n \to \infty$. Write

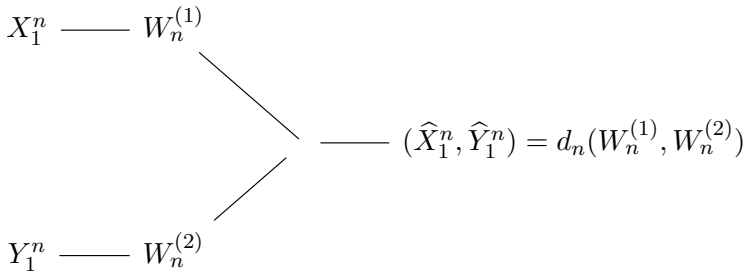
$$\mathbb{P}(E_{12,n}) = \mathbb{E}\left[\sum_{x_1^n, y_1^n} p(x_1^n, y_1^n) \sum_{\substack{\widetilde{x}_1^n \neq x_1^n \\ \widetilde{y}_1^n \neq y_1^n \\ (\widetilde{x}_1, \widetilde{y}_1^n) \in A_\delta^{(n)}}} \mathbb{1}_{\{e_n^{(1)}(\widetilde{x}_1^n) = e_n^{(1)}(x_1^n)\}} \mathbb{1}_{\{e_n^{(2)}(\widetilde{y}_1^n) = e_n^{(2)}(y_1^n)\}}\right]$$

Bring the expectation inside the sum, where the expectation of the inside is just a product of probabilities

$$= \sum_{x_1^n, y_1^n} p(x_1^n, y_1^n) \sum_{\substack{\widetilde{x}_1^n \neq x_1^n \\ \widetilde{y}_1^n \neq y_1^n \\ (\widetilde{x}_1, \widetilde{y}_1^n) \in A_\delta^{(n)}}} \mathbb{1}_{\{e_n^{(1)}(\widetilde{x}_1^n) = e_n^{(1)}(x_1^n)\}} \frac{1}{M_n^{(1)}} \frac{1}{M_n^{(2)}}$$
$$\leq \sum_{x_1^n, y_1^n} p(x_1^n, y_1^n) |A_\delta^{(n)}| \frac{1}{M_n^{(1)}} \frac{1}{M_n^{(2)}}$$
$$= |A_\delta^{(n)}| \frac{1}{M_n^{(1)}} \frac{1}{M_n^{(2)}}$$
$$\leq 2^{nH(X,Y)} 2^{n\delta} 2^{-nR_1} 2^{-nR_2}.$$

So if $\varepsilon > \delta$, this goes to 0 as $n \to \infty$ because $R_1 + R_2 > H(X, Y) + \varepsilon$ by assumption.

Converse: Consider any scheme $((e_n^{(1)}, e_n^{(2)}, d_n), n \geq 1)$ for which the error probability vanishes asymptotically. Letting $W_n^{(1)} = e_n^{(1)}(X_1^n)$ and $W_n^{(2)} = e_n^{(2)}(Y_1^n)$, we have



Let $p_e^{(n)} = \mathbb{P}((\widehat{X}_1^n, \widehat{Y}_1^n) \neq (X_1^n, Y_1^n))$. We have by Fano's inequality that

$$H(X_1^n, Y_1^n \mid W_n^{(1)}, W_n^{(2)}) \leq h(p_e^{(n)}) + p_e^{(n)}(\log|\mathscr{X}|^n + \log|\mathscr{Y}|^n),$$

so if $p_e^{(n)} \to 0$ then $H(X_1^n, Y_1^n \mid W_n^{(1)}, W_n^{(2)}) \leq n\varepsilon_n$ for some $\varepsilon_n \to 0$ as $n \to \infty$. Then, recalling that $R_1 = \frac{1}{n}\log M_n^{(1)}$ and $R_2 = \frac{1}{n}\log M_n^{(2)}$,

$$\begin{aligned} n(R_1 + R_2) &\geq H(W_n^{(1)}, W_n^{(2)}) \\ &= I(X_1^n, Y_1^n; W_n^{(1)}, W_n^{(2)}) + H(W_n^{(1)}, W_2^{(n)} \mid X_1^n, Y_1^n) \\ &= H(X_1^n, Y_1^n) - H(X_1^n, Y_1^n \mid W_n^{(1)}, W_n^{(2)}) \\ &\geq nH(X, Y) - n\varepsilon_n. \end{aligned}$$

But we also have

$$H(X_1^n \mid W_n(1), W_n^{(2)}, Y_1^n) \leq n\varepsilon_n,$$

which gives

$$\begin{aligned} nR_1 &\geq H(W_1^{(n)}) \\ &\geq H(W_n^{(1)} \mid Y_1^n) \\ &= I(X_1^n l W_1^{(n)} \mid Y_1^n) + H(W_1^{(n)} \mid X_1^n, Y_1^n) \\ &= H(X_1^n \mid Y_1^n) - H(X_1^{(n)} \mid W_n^{(1)}, Y_1^n, W_n^{(2)}, \end{aligned}$$

where we can throw $W_n^{(2)}$ in for free.

$$\geq nH(X \mid Y) - n\varepsilon_n.$$

Similarly, $R_2 \geq H(Y \mid X) - n\varepsilon_n$. Now divide by n and let $n \to \infty$ to get the lower bounds. This gives

$$\liminf_n \frac{1}{n}\log M_n^{(1)} + \frac{1}{n}\log M_n^{(2)} \geq H(X, Y),$$

$$\liminf_n \frac{1}{n} \log M_n^{(1)} \geq H(X \mid Y),$$
$$\liminf_n \frac{1}{n} \log M_n^{(2)} \geq H(Y \mid X). \qquad \square$$

15.2 The discrete memoryless channel model for data transmission

At each time, the transmitter sends a symbol $x \in \mathscr{X}$, and the receiver gets $y \in \mathscr{Y}$ according to the conditional probabilities $(p(y \mid X), x \in \mathscr{X}, y \in \mathscr{Y})$.

Example 15.1 (Binary symmetric channel)**.** The receival probability is $1 - p$, so

$$H(1 \mid 0) = p(0 \mid 1) = p, \qquad p(1 \mid 1) = p(0 \mid 0) = 1 - p.$$

Definition 15.1. A **communication scheme** is a sequence $((e_n, d_n), n \geq 1)$ such that

$$e_n : [M_n] \to \mathscr{X}^n, \qquad d_n : \mathscr{Y}^n \to [M_n].$$

Definition 15.2. Communication is possible **at rate** R if there exis $\mathrm{t}((e_n, d_n), n \geq 1)$ with

$$\liminf_n \frac{1}{n} \log M_n \geq R$$

and

$$\mathbb{P}(d_n(e_n(W_n)) \neq W_n) \xrightarrow{n \to \infty} 0,$$

where $W_n \sim \mathrm{Unif}([M_n])$.

Theorem 15.2 (Shannon's channel coding theorem)**.** *The supremum over all rates at which communication is possible is*

$$\sup_{(p(x), x \in \mathscr{X})} I(X;Y) = \sup_{(p(x), x \in \mathscr{X})} \sum_{x,y} p(x)p(y \mid x) \log \frac{p(y \mid x)}{p(x) \sum_{x'} p(x')p(y \mid x')}.$$

16 Discrete Memoryless Channels and Shannon's Channel Coding Theorem

16.1 Discrete memoryless channels

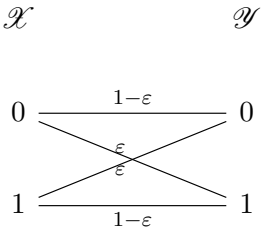
Shannon's **discrete memoryless channel** model of communication has 3 parts:

1. a finite set $\mathscr{X}$, called the **input alphabet**,
2. a finite set $\mathscr{Y}$, called the **output alphabet**,
3. a **channel matrix** of transition probabilities $[p(y \mid x)]_{x\in\mathscr{X},y\in\mathscr{Y}}$ with $p(y \mid x) \geq 0$ and $\sum_y p(y \mid x) = 1$ for all x.

Using the channel n times with inputs $x_1, x_2, \ldots, x_n$ results in outputs $y_1, y_2, \ldots, y_n$ with

$$p(y_1^n \mid y_1^n) = \prod_{i=1}^{n} p(y_i \mid x_i).$$

Example 16.1. The binary symmetric channel with crossover probability ε has $\mathscr{X} = \mathscr{Y} = \{0, 1\}$ and $p(0 \mid 1) = \varepsilon = p(1 \mid 0)$.



Here is the physical background: Fix time $T > 0$ (a real number) called the **symbol interval**. Suppose $(g_1(t), t \in [0, T]), (g_2(t), t \in [0, T])$ are orthonormal functions:

$$\int_0^T g_1^2(t)\,dt = 1, \qquad \int_0^T g_2^2(t)\,dt = 1, \qquad \int_0^T g_1(T)g_2(t)\,dt = 0.$$

Example 16.2. For example, we could take

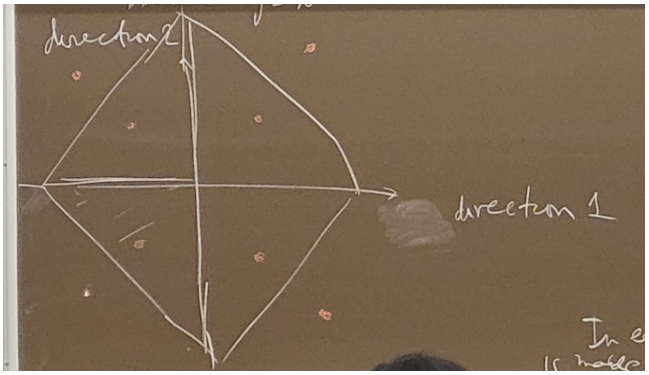
$$g_1(t) = \sqrt{\frac{2}{T}} \sin\left(\frac{2\pi t}{T}\right), \qquad g_2(t) = \sqrt{\frac{2}{T}} \cos\left(\frac{2\pi t}{T}.\right)$$

If we let $d = |\mathscr{X}|$ be the size of the input alphabet, then we have

$$\left\{ \begin{bmatrix} u_{1,1} \\ u_{1,2} \end{bmatrix}, \ldots, \begin{bmatrix} u_{d,1} \\ u_{d,2} \end{bmatrix} \right\},$$

where $g_i(t) = u_{i,1}g_1(t) + u_{i,2}g_2(t)$.

Now assume $\mathscr{Y} = \mathscr{X}$. The picture looks like this, called a **constellation**:



To send the sequence $i_1, \ldots, i_n$, what is physically sent is $\sum_{\ell=1}^{n} g_{i_\ell}(t - (\ell - 1)T)$. This is received in noise. In each interval, a decision is made as to what symbol was sent. For example, if the received output was in the bottom left triangle, we would make the decision that the dot in the center was the symbol sent.

16.2 Channel capacity and Shannon's channel coding theorem

Simple intuition suggests that in n uses of a DMC, we can hope to distinguish between a number of messages that is exponential in n. This motivates the Shannon formulation of "channel capacity."

Definition 16.1. Let

$$e_n : [M_n] \to \mathscr{X}^n, \qquad d_n : \mathscr{Y}^n \to [M_n].$$

We say that **communication is possible at rate** R if there is a sequence $((e_n, d_n), n \geq 1)$ such that $\mathbb{P}(d_n(e_n(W_n)) \neq W_n) \to 0$ as $n \to \infty$, where $W_n \sim \text{Unif}([M_n])$, and such that

$$\liminf_n \frac{1}{n} \log M_n \geq R.$$

Definition 16.2. Channel capacity is the supremum over rates at which communication is possible.

Theorem 16.1 (Shannon's channel coding theorem for a DMC)**.** *The channel capacity equals*

$$\sup_{(p(x), x \in \mathscr{X})} I(X;Y) = \sup_{(p(x), x \in \mathscr{X})} \sum_{x,y} p(x)p(y \mid x) \log \frac{p(y \mid x)}{\sum_{x' \in \mathscr{X}} p(y \mid x')p(x')}.$$

Remark 16.1. This is the maximum of a concave function. Often the maximizer is in the interior of the probability simplex.

Recall that $I(X;Y) = H(Y) - H(Y \mid X)$. Here, $H(Y \mid X) = \sum_x p(x) H(Y \mid X = x)$ is linear in $(p(x), x \in \mathscr{X})$ and $H(Y)$ is concave in $(p(y), y \in \mathscr{Y})$ and hence in $(p(y), x \in \mathscr{X})$.

Example 16.3 (Binary symmetric channel)**.** Suppose $p_X(1) = a = 1 - p_X(0)$. Then

$$\begin{aligned} I(X;Y) &= H(Y) - \underbrace{H(Y \mid X)}_{(1-a)H(Y|X=0)+aH(Y|X=1)} \\ &= h(a(1-\varepsilon) + (1-a)\varepsilon) - h(\varepsilon), \end{aligned}$$

to be optimized over a. This is maximized at $a = 1/2$. So the channel capacity is $1 - h(\varepsilon)$.

To get a feeling for why the theorem might be true, consider inputs to the channel $X_1, \ldots, X_n$ which are iid with $\mathbb{P}(X_1 = x) = p(x)$ for $x \in \mathscr{X}$. Then the outputs will be iid with marginals $(p(y), y \in \mathscr{Y})$, where $p(y) = \sum_x p(x) p(y \mid x)$. The inputs and outputs will be ε-jointly weakly typical with probability going to 1 as $n \to \infty$. The number of ε-weakly typical output sequences is $\geq (1-\varepsilon) 2^{nH(Y|X)} 2^{-n\varepsilon}$. The number of jointly ε-weakly typical output sequences with a specific ε-weakly typical input sequence is $\leq 2^{nH(Y|X)} 2^{2n\varepsilon}$. Then, using

$$1 = \sum_{y_1^n} p(y_1^n \mid x_1^n) = \sum_{x_1^n, y_1^n} \frac{p(x_1^n, y_1^n)}{p(x_1^n)},$$

we get

$$\frac{(1-\varepsilon) 2^{nH(Y)} 2^{-n\varepsilon}}{2^{nH(Y|X)} 2^{2n\varepsilon}} = (1-\varepsilon) 2^{nI(X;Y)} 2^{-3n\varepsilon}.$$

16.3 Proof of Shannon's channel coding theorem

Proof. For achievability, we need to show that for all rates $R < \max_{p(x), x \in \mathscr{X}} I(X;Y)$, we want to show that R is achievable. For the converse, we need to show that no $R > \max_{p(x), x \in \mathscr{X}} I(X;Y)$.

The achievability is given by a random coding argument.[10] We will take $M_n = \lceil 2^{nR} \rceil$, create random $e_n : [M_n] \to \mathscr{X}^n$ for each $n \geq 1$ and associated d_n and show that the error probability $\to 0$ as $n \to \infty$. Let

$$e_n(m) = (X_1(m), \ldots, X_n(m)), \qquad 1 \leq m \leq \lceil 2^{nR} \rceil = M_n$$

where $X_t(m) \sim (p(x), x \in \mathscr{X})$ is iid over $1 \leq t \leq n$ and $1 \leq m \leq M_n$. To define d_n, on receiving $y_1, \ldots, y_n$, find

$$\{1 \leq m \leq M_n \text{ s.t. } (x_1(m), \ldots, x_n(m)) \text{ is } \varepsilon\text{-jointly weakly typical with } (y_1, \ldots, y_n)\}.$$

[10] This is one of the historically earliest uses of the probabilistic method. It predates Erdös' widespread usage of the method.

If this set has exactly one member, return that member as $d_n(y_1, \ldots, y_n)$; otherwise, define $d_n(y_1, \ldots, y_n)$ arbitrarily.

Let $\mathcal{C}$ denote the (random) **codebook**

$$\begin{bmatrix} X)1(1) & \cdots & X_n(1) \\ \vdots & & \vdots \\ X_1(M_n) & \cdots & X_n(m) \end{bmatrix},$$

and let $P_e(\mathcal{C}) = \mathbb{P}(d_n(e_n(W_n) \neq W_n \mid \mathcal{C})$ denote the error probability conditioned on the codebook being $\mathcal{C}$. Then the expected error probability over the codebook is $\mathbb{P}(d_n(e_n(W_n)) \neq W_n)$. We have

$$\begin{aligned} \mathbb{P}(d_n(e_n(W_n)) \neq W_n) &= \sum_{\mathcal{C}} \mathbb{P}(\mathcal{C}) \mathbb{P}(d_n(e_n(W_n)) \neq W_n \mid \mathcal{C}) \\ &= \sum_{\mathcal{C}} P(\mathcal{C}) \sum_{m=1}^{M_n} \frac{1}{M_n} \lambda_m(\mathcal{C}), \end{aligned}$$

where $\lambda_m(c) = \mathbb{P}(d_n(e_n(m)) \neq m \mid \mathcal{C})$.

$$\begin{aligned} &= \sum_{m=1}^{M_n} \frac{1}{M_n} \sum_{\mathcal{C}} \mathbb{P}(\mathcal{C}) \lambda_m(\mathcal{C}) \\ &= \sum_{\mathcal{C}} \mathbb{P}(\mathcal{C}) \lambda_1(\mathcal{C}) \end{aligned}$$

by symmetry.

$$\begin{aligned} &\leq \mathbb{P}\left(E_0 \cup \left(\bigcup_{m=2}^{M_n} E_m \right) \right) \\ &\leq \mathbb{P}(E_0) + \sum_{m=2}^{M_n} \mathbb{P}(E_m). \end{aligned}$$

Note that E_0 is the event where $(Y_1, \ldots, Y_n)$ is not ε-jointly weakly typical with the sequence $(X_1(1), \ldots, X_n(1))$. For $m \geq 2$, E_m is the event where $(Y_1, \ldots, Y_n)$ is not ε-jointly weakly typical with $(X_1(m), \ldots, X_n(m))$. So

$$\mathbb{P}(d_n(e_n(W_N) \neq W_n) \leq \mathbb{P}(E_0^{(n)}) + (M_n - 1)\mathbb{P}(E_2^{(n)})$$

by symmetry. Now $\mathbb{P}(E_0^{(n)}) \to 0$ as $N \to \infty$, and $\mathbb{P}(E_2^{(n)}) \leq 2^{-nI(X;Y)} 2^{3n\varepsilon}$ So if $R < I(X;Y) - 3\varepsilon$, this goes to 0 as $n \to \infty$. □

Next time, we will prove the converse part of the theorem.

17 Upper Bound for Channel Capacity, Perfect Noiseless Feedback, and Joint Source Channel Coding

17.1 Upper bound for Shannon's channel coding theorem

Last time, we were proving Shannon's channel coding theorem for discrete memoryless channels. A DMC is given by a probability transition matrix $[p(y \mid x)]_{x\in\mathscr{X},y\in\mathscr{Y}}$, where $\mathscr{X},\mathscr{Y}$ are finite. Shannon's formulation uses block codes $((e_n,d_n),n\geq 1)$, where

$$e_n : [M_n] \to \mathscr{X}^n, \qquad d_n : \mathscr{Y}^n \to [M_n],$$

where M_n is exponentially growing in n. The "memoryless" part means

$$p(y_1^n \mid x_1^n) = \prod_{i=1}^n p(x_i,y_i).$$

Definition 17.1. We say **communication is possible at rate** R if there exist $((e_n,d_n),n\geq 1)$ such that

$$\mathbb{P}(d_n(e_n(W_n)) \neq W_n) \xrightarrow{n\to\infty} 0$$

and

$$\liminf_n \frac{1}{n}\log M_n \geq R.$$

Let

$$C := \max_{(p(x),x\in\mathscr{X})} I(X;Y).$$

Theorem 17.1 (Shannon's channel coding theorem)**.**

$$\sup\{R : \textit{can communicate at rate } R\} = C.$$

C is called the **Shannon capacity** of the channel. Let's finish the proof.

Proof. We have proved achievability: For $\varepsilon > 0$, we can communicate at rate $C-\varepsilon$. Now we prove the converse. Consider any $((e_n,d_n),n\geq 1)$. We have the Markov chain $W_n - X_1^n - \widehat{Y}_1^n - \widehat{W}_n$, where $W_n \sim \mathrm{Unif}([M_n])$, and $\widehat{W}_n = d_n(Y_1^n)$. In this notation, the error probability is $p_e^{(n)} = \mathbb{P}(\widehat{W}_n \neq W_n)$; let's assume $p_e^{(n)} \to 0$. We will prove that this implies $\limsup_n \frac{1}{n}H(W_n) \leq C$ as follows.

$$\begin{aligned} H(W_n) &= H(W_n \mid Y_1^n) + I(W_n;Y_1^n) \\ &\leq H(W_n \mid Y_1^n) + I(X_1^n;Y_1^n) \\ &\leq H(W_n \mid \widehat{W}_n) + I(X_1^n;Y_1^n) \end{aligned}$$

Fano's inequality says that $H(W_n \mid \widehat{W}_n) \leq h(p_e^{(n)}) + p_e^{(n)} \log(M_n - 1)$.

$$\leq h(p_e^{(n)}) + p_e^{(n)} \log M_n + I(X_1^n; Y_1^n)$$

To deal with the last term, use the chain rule to write

$$\begin{aligned} I(X_1^n; Y_1^n) &= H(Y_1^n) - H(Y_1^n, X_1^n) \\ &\leq \sum_{i=1}^n H(Y_i) - \sum_{i=1}^n \underbrace{H(Y_i \mid X_1^n, Y_1^n)}_{=H(Y_i|X_i)} \\ &= \sum_{i=1}^n H(Y_i \mid X_i) \\ &\leq nC. \end{aligned}$$

Our issue is now that $\log M_n$ looks like n. We can deal with this by noting that $H(W_n) = \log M_n$ on the left. So far, we have that

$$\log M_n \leq h(p_e^{(n)}) + p_e^{(n)} \log m_n + nC.$$

Hence,

$$(1 - p_e^{(n)}) \frac{\log M_n}{n} \leq \frac{h(p_e^{(n)})}{n} + C.$$

If $p_e^{(n)} \to 0$ as $n \to \infty$, this implies that $\limsup_n \frac{1}{n} \log M_n \leq C$. □

17.2 Communication with perfect noiseless feedback

Earlier, we had $X_i = e_{n,i}(m)$, where $e_n = (e_{n,1}, \ldots, e_{n,n})$ and $m \in [M_n]$ is a message.

Definition 17.2. Perfect noiseless feedback is when we have $X_i = e_{n,i}(m, Y_1, \ldots, Y_{i-1})$.

Theorem 17.2. *Perfect noiseless feedback cannot increase the rates at which communication is possible over a DMC.*

Proof. The achievability at rate $C - \varepsilon$ is the same as Shannon's coding theorem, since the encoder can ignore the feedback. But for the converse, we do not have a Markov chain. As before, write

$$\begin{aligned} \log M_n &= H(W_n) \\ &= H(W_n \mid Y_1^n) + I(W; Y_1^n) \\ &\leq h(p_e^{(n)}) + p_e^{(n)} \log M_n, \end{aligned}$$

where $p_e^{(n)} := P(d_n(Y_1^n) \neq W_n)$. Note that we can still use Fano's inequality because we have the Markov chain $W_n - Y_1^n - \widehat{W}_n$. Here, Y_i conditioned on $(X_1^{i-1}, Y_1^{i-1}, X_i = x_i)$ has the law $p(y_i \mid x_i)$. Observe that $p(m, x_1^n, y_1^n) = \frac{1}{M_n} \prod_{i=1}^n \mathbb{1}_{\{x_i = e_i(m, y_1^{i-1})\}} p(y_i \mid x_i)$.

The chain rule gives

$$\begin{aligned} I(W_n, Y_1^n) &= H(Y_1^n) - H(Y_1^n \mid W_n) \\ &\leq \sum_{i=1}^n H(Y_i) - \sum_{i=1}^n H(Y_i \mid W_n, Y_1^{i-1}) \end{aligned}$$

But $X_i = e_{n,i}(W_n, Y_1^{i-1})$, so $H(Y_i \mid W_n, Y^{i-1}) = H(Y_i \mid X_n, Y_1^{i-1}, X_i) = H(Y_i, X_i)$. So

$$I(W_n, Y_1^n) \leq nC,$$

and the rest of the proof proceeds as before. □

17.3 Joint source channel coding

Model a source as a random sequence $(V_k, k \in \mathbb{Z})$ (think stationary and ergodic) with $V_k \in \mathscr{V}$, where $\mathscr{V}$ is finite.

Definition 17.3. A **source channel code at block length** n is an encoding map

$$e_n : \mathscr{V}^{\ell_n} \to \mathscr{X}^n$$

and a decoding map

$$d_n : \mathscr{Y}^n \to \mathscr{V}^{\ell_n}.$$

Note that ℓ_n might be different from n. Here, Y_1^n results from X_1^n over a DMC.

$$V_1^{\ell_n} \xrightarrow{e_n} \mathscr{X}_1^n \xrightarrow{\text{DMC}} \mathscr{Y}_1^n \xrightarrow{d_n} \mathscr{V}_1^{\ell_n}.$$

Theorem 17.3 (Joint source channel coding theorem)**.** *If the source has entropy rate $H(V)$, then there exists $((e_n, d_n), n \geq 1)$ with $\mathbb{P}(d_n(e_n(V_1^{\ell_n})) \neq V_1^{\ell_n}) \to 0$ if and only if*

$$\limsup_n \frac{\ell_n H(V)}{nC} \leq 1.$$

Proof. Achievability: The idea is to compress the source and then use Shannon's channel coding theorem. Take $\ell_n = n$. If $H(V)/C \leq 1 - \delta$, we can compress V_1^n to $n(H(V) + \delta/2)$ bits with probability going to 0 as $n \to \infty$. Then send those bits over a DMC with error probability going to 0.

Converse: We have the Markov chain

$$V_1^{\ell_n} - X_1^n - Y_1^n - \widehat{V}_1^{\ell_n},$$

so Fano's inequality gives

$$H(V_1^{\ell_n} \mid \widehat{V}_1^{\ell_n}) \leq 1 + P_e^{(n)}(\ell_n \log |\mathscr{V}|).$$

We then have

$$\begin{aligned} H(V_1^n) &= H(V_1^{\ell_n} \mid \widehat{V}_1^{\ell_n}) + I(V_1^{\ell_n}; \widehat{V}_1^{\ell_n}) \\ &\leq 1 + p_e^{(n)}(\ell_n \log |\mathscr{V}|) + I(X_1^n; Y_1^n) \\ &\leq 1 + p_e^{(n)}(\ell n \log |\mathscr{V}|) + nC. \end{aligned}$$

Divide by ℓ_n and let $n \to \infty$ (we can assume without loss of generality that $\ell_n \to \infty$, otherwise we automatically have the limsup bounded by 1). We get

$$\frac{H(V_1^{\ell_n})}{\ell_n} \leq \frac{1}{\ell_n} + p_e^{(n)} \log |\mathscr{V}| + \frac{nC}{\ell_n}.$$

The left hand side converges to $H(V)$. The first term on the right goes to 0 because $\ell_n \to \infty$. The second term on the right goes to 0 because $p_e^{(n)} \to 0$ by assumption. $\square$

18 Differential Entropy and the Additive White Gaussian Noise Channel Model

18.1 Differential entropy

Let X be a real-valued random variable with density, i.e.

$$\mathbb{P}(X \in [a,b)) = \int_a^b f(x)\,dx$$

for some nonnegative function f.

Definition 18.1. The **differential entropy** of X is

$$h(f) := -\int_{-\infty}^{\infty} f(x)\log f(x)\,dx.$$

This need not be well-defined (an example is provided in Handout 7), so when we talk about $h(f)$, we will assume it exists.

Example 18.1. Let $X \sim \text{Unif}([a,b])$ with density

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

Then

$$h(f) = \int_a^b \frac{1}{b-a}\log(b-a)\,dx = \log(b-a).$$

Note that if $b-a<1$, this is *negative*. So $h(f)$ is very different from entropy.

Example 18.2. Let $X \sim N(\mu,\sigma^2)$ with density

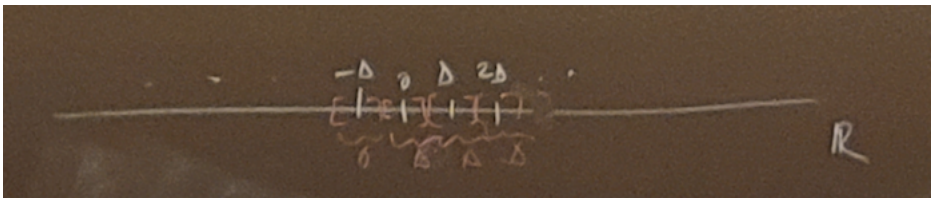
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma)}.$$

Then

$$\begin{aligned} h(f) &= (\log e)\int_{-\infty}^{\infty} f(x)\left[\frac{(x-\mu)^2}{2\sigma^2} + \frac{1}{2}\ln(2\pi\sigma^2)\right]dx \\ &= (\log e)\left[\frac{1}{2} + \frac{1}{2}\ln(2\pi\sigma^2)\right] \\ &= \frac{1}{2}\log(2\pi e\sigma^2). \end{aligned}$$

18.2 Connection to entropy

Here is the connection between differential entropy and an underlying entropy. Imagine quantizing $\mathbb{R}$ at scale Δ with $\Delta \to 0$.



We get a discrete probability distribution having probability

$$\int_{k\Delta-\frac{\Delta}{2}}^{k\Delta+\frac{\Delta}{2}} f(x)\,dx \quad \text{at } k$$

as an approximation to a random variable with density f. Think of the entropy of this approximation. This is

$$-\sum_{k\in\mathbb{Z}}(\Delta f(k\Delta)+o(\Delta))\log(\Delta f(k\Delta)+o(\Delta)) - \log\Delta - \Delta\sum_{k\in\mathbb{Z}} f(k\Delta)\log f(k\Delta) + o(\Delta).$$

So we can think of $h(f)$ as the amount of entropy of a quantized pproximation about $-\log\Delta$ as $\Delta \to 0$.

This $-\log\Delta$ is a problem because $-\log\Delta \to \infty$ as $\Delta \to 0$.

18.3 Relative entropy

However, this quantization problem does not show up for relative entropy.

Definition 18.2. Given two probability densities f and g, the **relative entropy** is

$$D(f \mid\mid g) := \int_{-\infty}^{\infty} f(x)\log\frac{f(x)}{g(x)}\,dx.$$

Note that in writing

$$\sum_{k\in\mathbb{Z}}(\Delta f(k\Delta)+o(\Delta))\log\frac{\Delta f(k\Delta)+o(\Delta)}{\Delta g f(k\Delta)+o(\Delta)},$$

the Δs in the log cancel.

The relative entropy will be nonnegative by convexity of $u \mapsto u\log u$ because it is

$$\int_{-\infty}^{\infty} g(x)\frac{f(x)}{g(x)}\log\frac{f(x)}{g(x)}\,dx.$$

18.4 Joint differential entropy

Definition 18.3. The **joint differential entropy** of $X_1, \ldots, X_n$ (real-valued random variables with a joint density f) is

$$h(X_1, \ldots, X_n) = -\mathbb{E}[\log f(X_1, \ldots, X_n)].$$

Example 18.3. The most important example is when $X_1, \ldots, X_n$ are jointly Gaussian random variables with invertible covariance matrix:

$$\begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m_1 \\ \vdots \\ m_n \end{bmatrix}, K \right),$$

where K is a symmetric, positive definite matrix. The joint density is

$$f(x_1, \ldots, x_n) = \frac{1}{(2\pi)^{n/2} (\det K)^{1/2}} e^{-\frac{1}{2}(x-m)^\top K^{-1}(x-m)}.$$

The joint differential entropy is

$$h(X_1, \ldots, X_n) = \frac{1}{2} \log((2\pi e)^n \det K).$$

This can be understood by diagonalizing K. $K = U^\top D U$, where $U^\top U = I$ and $D = \operatorname{diag}(\sigma_1^2, \ldots, \sigma_n^2)$. Then

$$h(X_1, \ldots, X_n) = \sum_{\ell=1}^{n} \frac{1}{2} \log(2\pi e \sigma_\ell^2).$$

18.5 Mutual information

If X and Y have joint density $f(x, y)$, then they will have marginal densities $f(x)$ and $f(y)$ respectively.

Definition 18.4. The **mutual information** is defined as

$$I(X; Y) = D(f(x, y) \mid\mid f(x) f(y)).$$

This will turn out to be

$$I(X; Y) = h(X) + h(Y) - h(X, Y),$$

when this expression makes sense. This will also be

$$I(X; Y) = h(X) - h(X \mid Y)$$

, if these quantities exist, where $h(X \mid Y)$ is the conditional differential entropy.

Definition 18.5. The **conditional differential entropy** is

$$h(X \mid Y) = \int_{-\infty}^{\infty} f(y) h(X \mid Y = y)\, dy,$$

where

$$h(X \mid Y = y) = -\int_{-\infty}^{\infty} f(x \mid y) \log f(x \mid y)\, dx.$$

18.6 Chain rules for differential entropy

We can write some chain rules.

Proposition 18.1 (Chain rule for differential entropy)**.** *When all these quantities make sense,*

$$h(X_1, \ldots, X_n) = h(X_1) + h(X_2 \mid X_1) + h(X_3 \mid X_1, X_2) + \cdots + h(X_n \mid X_1, \ldots, X_{n-1}).$$

Proposition 18.2 (Chain rule for mutual information)**.** *When $(X, Y_1, \ldots, Y_n)$ have a joint density,*

$$I(X; Y_1, \ldots, Y_n) = I(X; Y_1) + I(X; Y_2 \mid Y_1) + I(X; Y_3 \mid Y_1, Y_2) + \cdots + I(X; Y_n \mid Y_1, \ldots, Y_{n-1}).$$

18.7 Basic properties of differential entropy

Proposition 18.3. *For any constant c, $h(X + c) = h(X)$.*

Proof. Adding c just translates the density. □

Proposition 18.4. *If $c \neq 0$, then $h(cX) = h(X) - \log |c|$.*

Proof. The density of cX is $\frac{1}{|c|} f(x/c)$. So

$$\begin{aligned} h(cX) &= \int_{-\infty}^{\infty} \frac{1}{|c|} f(x/c) \log \frac{1}{|c|} f(x/c)\, dx \\ &= h(X) - \log |c|. \end{aligned}$$

□

Remark 18.1. This is consistent with $X \sim N(0, \sigma^2) \iff cX \sim 0, c^2\sigma^2)$. Here, $h(X) = \frac{1}{2} \log(2\pi e \sigma^2)$ and $h(cX) = \frac{1}{2} \log(2\pi e \sigma^2) + \log |c|$.

Proposition 18.5. *If $\mathbb{E}[X] = 0$ and $\mathbb{E}[X^2] = \sigma^2$, then*

$$h(X) \leq \frac{1}{2} \log(2\pi e \sigma^2).$$

This upper bound is the entropy of the Gaussian.

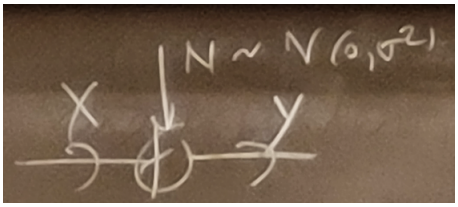
Proof. Let $\phi(x)$ denote the $N(0,\sigma^2)$ density, i.e. $\phi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-x^2/(2\sigma^2)}$. Write

$$\begin{aligned}
0 &\le D(f \,||\, \phi) \\
&= \int_{-\infty}^{\infty} f(x) \log \frac{f(x)}{\phi(x)} \\
&= -h(f) + (\log e) \int_{-\infty}^{\infty} f(x) \left[\frac{1}{2} \ln(2\pi\sigma^2) + \frac{x^2}{2\sigma^2} \right] dx \\
&= -h(f) + \frac{1}{2} \log(2\pi e \sigma^2).
\end{aligned}$$

□

18.8 The additive white Gaussian noise channel model

This is a discrete time model. At each channel use, the input is a real number, say $x \in \mathbb{R}$. The output is a real number Y. Conditioned on $X = x$, $Y \sim \mathcal{N}(x, \sigma^2)$, where σ^2 is the variance of the noise.



Consider an input power constrained scenario and block based communication: We have an encoding map

$$e_n : [M_n] \to \mathbb{R}$$

and a decoding map

$$d_n : \mathbb{R}^n \to [M_n], \qquad Y^n(m).$$

Here X_n is the output of e_n, and Y_n is the input of d_n.

Conditioned on $X^n(m) = x^n$,

$$Y^n \sim \mathcal{N}(x^n, \sigma^2 I),$$

i.e. the noise is iid over time. In other words,

$$f(y^n \mid x^n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y_i - x_i)/(2\sigma)}.$$

The **power constraint** P requires that each $X^n(m)$ satisfies

$$\sum_{i=1}^{n} (X_i(n))^2 \le nP.$$

Intuitively, M_n can be on the scale of

$$\frac{V_n(\sqrt{n(P+\sigma^2)})}{V_n(\sqrt{n\sigma^2})},$$

where $V_n(R)$ denotes the colume of the ball in $\mathbb{R}^n$ of radius R.

19 Capacity of an Additive White Gaussian Noise Channel

19.1 Shannon capacity of a additive white Gaussian noise channel

In the additive white Gaussian noise (AWGN) model, we send inputs real-valued X_i and receive real-valued outputs Y_i, where $Y_i = X_i + Z_i$, and $Z_1, Z_2, \ldots \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$. At block length n, we have an encoding map $e_n : [M_n] \to \mathbb{R}^n$ and a decoding map $d_n : \mathbb{R}^n \to [M_n]$. We assume an **input power constraint**, which, in Shannon's formulation, says that each codeword is required to have power at most P: If $X^n(m)$ denotes $e_n(m)$, then

$$\frac{1}{n}\sum_{i=1}^{n} X_i^n(m) \le P.$$

We want to find for which rates R we have

$$\liminf_n \frac{1}{n}\log M_n \ge R \quad \text{with} \quad \mathbb{P}(d_n(e_n(W_n)) \neq W_n) \to 0$$

for some sequence $((e_n, d_n), n \ge 1)$ with $W_n \sim \text{Unif}([M_n])$.

Theorem 19.1. *The supremum over rates for which communication is possible is*

$$\sup_{X \sim f : \int_{-\infty}^{\infty} x^2 f(x)\,dx \le P} I(X; X+Z),$$

which equals $\frac{1}{2}\log(1 + \frac{P}{\sigma^2})$ *and is achieved by* $X \sim N(0, P)$.

This quantity is called the **Shannon capacity**. The achievability part of the proof will use a random coding argument and requires the concept of ε-weakly typical sequences. The converse part of the proof involves Fano's inequality. Let's first see why the last claim is true:

Lemma 19.1. *If* $\mathbb{E}[X^2] \le P$, *then* $I(X; X+Z) \le \frac{1}{2}\log(1 + \frac{P}{\sigma^2})$, *with equality if and only if* $X \sim N(0, P)$.

Proof.

$$\begin{aligned} I(X; X+Z) &= h(X+Z) - h(X+Z \mid X) \\ &= h(X+Z) - h(Z) \\ &= h(X+Z) - \frac{1}{2}\log(2\pi e\sigma^2). \end{aligned}$$

Since $X \amalg Z$ and $\mathbb{E}[Z_1] = 0$, we also have

$$\mathbb{E}[(X+Z)^2] = \mathbb{E}[X^2] + \mathbb{E}[Z^2]$$

$$\leq P + \sigma^2.$$

So

$$h(X+Z) \leq \frac{1}{2}\log(2\pi e(P+\sigma^2))$$

with equality iff $X \sim N(0,P)$. So

$$\begin{aligned} I(X, X+Z) &\leq \frac{1}{2}\log\left(\frac{P+\sigma^2}{\sigma^2}\right) \\ &= \frac{1}{2}\log\left(1+\frac{P}{\sigma^2}\right). \end{aligned}$$

□

19.2 Weak-typicality for differential entropy

Definition 19.1. For $X \sim f$ with differential entropy $h(X)$ and $\varepsilon > 0$, the set of **ε-weakly typical sequences** for the density f sis

$$A_\varepsilon^n := \left\{ x^n \in \mathbb{R}^n : \left| -\frac{1}{n}\log\prod_{i=1}^n f(x_i) - h(X) \right| < \varepsilon \right\} \subseteq \mathbb{R}^n$$

By the weak law of large numbers,

$$\mathbb{P}(X^n \in A_\varepsilon^n) = 1$$

if $X_i \overset{\text{iid}}{\sim} f$. This is because $\mathbb{E}[\log \frac{1}{f(X)}] = h(X)$ when $X \sim f$.

Proposition 19.1. *For all n,*

$$\mathrm{Vol}(A_\varepsilon^n) \leq 2^{nh(X)}2^{n\varepsilon}.$$

Proof.

$$\begin{aligned} 1 &\geq \int_{A_\varepsilon^n} \prod_{i=1}^n f(x_i)dx^n \\ &\geq \int_{A_\varepsilon^n} 2^{-nh(X)}2^{-n\varepsilon}\, dx^n \\ &= \mathrm{Vol}(A_\varepsilon^n)2^{-nh(X)}2^{-n\varepsilon}. \end{aligned}$$

□

Proposition 19.2. *Given $\delta > 0$, for all sufficiently large n,*

$$\mathrm{Vol}(A_\varepsilon^n) \geq (1-\delta)2^{nh(X)}2^{-n\varepsilon}.$$

Proof. For sufficiently large n,

$$\begin{aligned}(1-\delta) &\le \int_{A_\varepsilon^n} \prod_{i=1}^n f(x_i)\, dx^n \\ &\le \int_{A_\varepsilon^n} 2^{-nh(X)} 2^{n\varepsilon}\, dx^n \\ &= \mathrm{Vol}(A_\varepsilon^n) 2^{-nh(X)} 2^{n\varepsilon}.\end{aligned}$$

□

Definition 19.2. Let $(X_1, Y_1), (X_2, Y_2), \dots$ be iid with $(X_i, Y_i) \sim f(x, y)$. The set of ε**-jointly weakly typical sequences** for f is

$$\begin{aligned}A_\varepsilon^n := \Bigg\{ (x^n, y^n) : &\left| -\frac{1}{n} \log \prod_{i=1}^n f(x_i) - h(X) \right| \le \varepsilon, \\ &\left| -\frac{1}{n} \log \prod_{i=1}^n f(y_i) - h(Y) \right| \le \varepsilon, \\ &\left| -\frac{1}{n} \log \prod_{i=1}^n f(x_i, y_i) - h(X,Y) \right| \le \varepsilon, \Bigg\}.\end{aligned}$$

With this definition in mind, we can show the following.

Lemma 19.2. *If $\widetilde{X}^n \overset{d}{=} X^n$, $\widetilde{Y}^n \overset{d}{=} Y^n$, and $\widetilde{X}^n \amalg \widetilde{Y}^n$, then*

$$(1-\delta) 2^{-nI(X;Y)} 2^{-3n\varepsilon} \le \mathbb{P}((\widetilde{X}^n, \widetilde{Y}^n) \in A_\varepsilon^n) \le 2^{-nI(X;Y)} 2^{3n\varepsilon}.$$

The upper bound holds for all n, and the lower bound holds for all sufficiently large n.

19.3 Proof of Shannon's channel coding theorem for an AWGN channel

Now we can prove the theorem.

Proof. Achievability: Generate a random codebook

$$\begin{bmatrix} X_1(1) & \cdots & X_n(1) \\ X_1(2) & \cdots & X_n(2) \\ \vdots & & \vdots \\ X_1(M_n) & \cdots & X_n(M_n) \end{bmatrix},$$

where each $X_n(i) \sim \mathcal{N}(0, P - \eta)$ is iid over i and n. Let $W_n \sim \mathrm{Unif}([M_n])$. The decoding rule is

$$d_n(Y^n) = \begin{cases} m & (X^n(m), Y^n) \text{ are } \varepsilon\text{-jointly weakly typical and for all } m' \neq m, \\ & \quad (X^n(m), Y^n) \text{ are not } \varepsilon\text{-jointly weakly typical} \\ \text{arbitrary} & \text{either no or } \ge 2\ X^n(m) \text{ are } \varepsilon\text{-jointly typical with } Y^n. \end{cases}$$

By symmetry,

$$\begin{aligned}\mathbb{P}(d_n(e_n(W_n)) \neq W_n) &= \mathbb{P}(d_n(e_n(1) \neq 1)) \\ &\leq P(E_{0,n}) + \sum_{m \neq 2} P(E_{m,n}),\end{aligned}$$

where $E_{0,n}$ is the event that $(X^n(1), Y^n)$ is not ε-jointly weakly typical and $E_{m,n}$ for $m \geq 2$ is the event that $(X^n(1), Y^n)$ is ε-jointly weakly typical. Then $\mathbb{P}(E_{0,n}) \to 0$ as $n \to \infty$, and for each $2 \leq m \leq M_n$, $\mathbb{P}(E_{m,n}) \leq 2^{-nI(X;Y)} 2^{3n\varepsilon}$. So if $M_n = 2^{nR}$ with $R < I(X;Y) - 3\varepsilon$, then $\mathbb{P}(d_n(e_n(W_n)) \neq W_n) \to 0$ as $n \to \infty$.

Converse: Consider any $((e_n, d_n), n \geq 1)$. We have $W_n \sim \text{Unif}([M_n])$ and the Markov chain $W_n - X^n - Y^n - \widehat{W}_n$ with $X^n = e_n(W_n)$, $Y = X + Z$, and $\widehat{W}_n = d_n(Y^n)$. Suppose $M_n = \lceil 2^{nR} \rceil$. The data-processing inequality gives

$$H(W_n \mid Y^n) \leq H(W_n \mid \widehat{W}_n).$$

Note that W_n is a discrete random variable, and Y^n is a continuous random variable. Here, we mean $H(W_n \mid Y^n) = \int_{-\infty}^{\infty} H(W_n \mid Y_n = y) f(y)\, dy$. If $p_e(n) := \mathbb{P}(\widehat{W}_n \neq W_n)$, then Fano's inequality gives

$$H(W_n \mid Y^n) \leq 1 + nRp_e(n).$$

Also, the data processing inequality gives

$$\begin{aligned}H(W_n) &= I(W_n; Y^n) + H(W_n \mid Y^n) \\ &\leq I(X^n; Y^n) + H(W_n \mid Y^n) \\ &= h(Y^n) - \sum_{i=1}^{n} h(Y_i \mid X^n, Y^{i-1}) + H(W_n \mid Y^n)\end{aligned}$$

Use $0 \leq D(f(y^n) \mid\mid \prod_{i=1}^n f(y_i)) = \int_{\mathbb{R}^n} f(y^n) \log \frac{f(y^n)}{\prod_{i=1}^n f(y_i)}\, dy^n = -h(Y^n) + \sum_{i=1}^n h(Y_i)$.

$$\leq \sum_{i=1}^{n} h(Y_i) - \sum_{i=1}^{n} h(Y_i \mid X^n, Y^{i-1}) + H(W_n \mid Y^n)$$

Use the Markov chain $Y_i - X_i - (X^{i-1}, X_{i+1}^n, Y^{i-1})$

$$\begin{aligned}&\leq \sum_{i=1}^{n} h(Y_i) - \sum_{i=1}^{n} h(Y_i \mid X_i) + H(W_n \mid Y^n) \\ &= \sum_{i=1}^{n} I(X_i; Y_i) + H(W_n \mid Y^n)\end{aligned}$$

Let $P_i := \mathbb{E}[X_i^2]$, and recall that $Y_i = X_i + Z_i$, where $Z \sim \mathcal{N}(0, \sigma^2)$ and $X_i \amalg Z_i$.

$$\leq \sum_{i=1}^{n} \frac{1}{2} \log\left(1 + \frac{P_i}{\sigma^2}\right) + H(W_n \mid Y^n)$$

$$\leq \frac{n}{2}\log\left(1+\frac{P}{\sigma^2}\right) + H(W_n \mid Y^n)$$
$$\leq \frac{n}{2}\log\left(1+\frac{P}{\sigma^2}\right) + 1 + (\log M_n)p_e(n).$$

Since $\frac{1}{n}\log M_n \to R$ if $p_e(n) \to 0$, this gives

$$\limsup_n \frac{1}{n}\log M_n \leq \frac{1}{2}\log\left(1+\frac{P}{\sigma^2}\right). \qquad \square$$

Why is this result interesting? Suppose the FCC assigns you a bandwidth of W Hertz, and you communicate over this channel for some time T at power constraint P (with units energy per unit time). One can show that if the noise that corrupts your waveform is additive white noise with power spectral density $\frac{N_0}{2}$, then the theoretical limit of the rate at which you can communicate is

$$W\log(1+\frac{P}{N_0 W}) \quad \text{bits/unit time.}$$

Studying the $W \to \infty$ limit and the $T \to \infty$ limit is interesting.

20 Capacity of Wide Sense Stationary Processes and Parallel Gaussian channels

20.1 Wide sense stationary processes

Definition 20.1. A **stationary stochastic process** $(X(t).t \in \mathbb{R})$ is a collection of random variables $X(t)$ such that

$$(X(t_1), \dots, X(t_d)) \stackrel{d}{=} (X(t_1 + s), \dots, X(t_d + s)).$$

The correct thing to study to understand spectral properties of such a process is the autocorrelation function.

Definition 20.2. The **autocorrelation function** is

$$R_{x,x}(s,t) := \mathbb{E}[X(t)X(s)] = R_x(t-s).$$

By stationarity, this only depends on $t - s$.

Definition 20.3. A **wide sense stationary (WSS) process** is a process for which $R_{x,x}(t,s)$ depends only on $t - s$ (and $\mathbb{E}[X(t)]$ is constant).

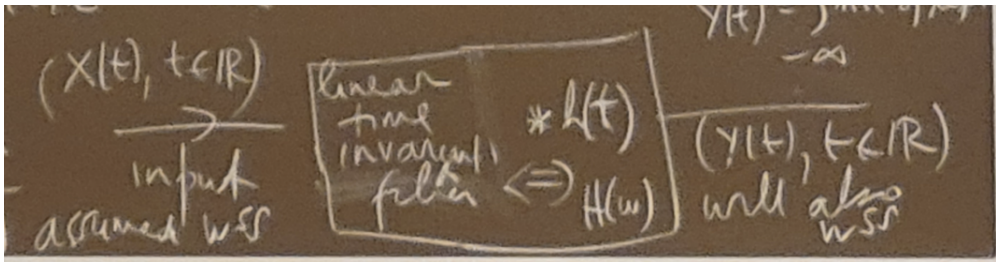
Definition 20.4. The **power spectral density** of the noise is

$$S_{x,x}(\omega) = \int_{-\infty}^{\infty} R_{x,x}(t) e^{-i\omega\tau}\, d\tau,$$

the Fourier transform of the autocorrelation function.[11]

If we input a WSS into a linear time invariant filter, which outputs a WSS, then we have the following magic formula:

$$S_{y,y}(\omega) = |H(\omega)|^2 S_{x,x}(\omega).$$



We should think of $S_{x,x}$ as telling us how much noise sits at each frequency.

[11] Professor Anantharam uses j instead of i, but I disagree.

Definition 20.5. If $S_{x,x}(\omega)$ is constant, then $(X(t), t \in \mathbb{R})$ is called **white noise**. If in addition, $(X(t), t \in \mathbb{R})$ is a Gaussian process, i.e. $(X(t_1), \ldots, X(t_d))$ is jointly Gaussian for all $t_1, \ldots, t_d$, we call this **white Gaussian noise**.

Assuming $\mathbb{E}[X(t)] = 0$ for all t, this is characterized by the properties

1.
$$\int_{-\infty}^{\infty} X(t)f(t)\,dt \sim N(0, \sigma^2, \quad \text{if } \int_{-\infty}^{\infty} f^2(t)\,dt,$$

2.
$$\int_{-\infty}^{\infty} f(t)g(t)\,dt = 0 \implies \int_{-\infty}^{\infty} X(t)f(t)\,dt \amalg \int_{-\infty}^{\infty} X(t)g(t)\,dt.$$

20.2 Connection between WSSs and AWGNs

Last time, we saw that the Shannon capacity of a Power-constrained AWGN is

$$\frac{1}{2}\log\left(1 + \frac{P}{\sigma^2}\right) \qquad \text{bits per use.}$$

This is interesting because it is a model for if you input a power-constrained waveform X (bandlimited to W Hz and time limited to T seconds) and the noise Z is additive and white Gaussian noise. Here, the output is $Y(t) = X(t) + Z(t)$.

The number of degrees of freedom, which represents the dimension of our input, is intuitively $2WT$. Nyquist sampling theory tells us that $2W$ samples per second is needed to recover a signal which is bandlimited to W. The Landau-Pollack paper makes this precise via prolate spheroidal functions.

The functions for which a fraction of at least $1 - \varepsilon_T^2$ of the entropy should be on $[-T/2, T/2]$ and which are bandlimited to W can be expressed in terms of $2WT + \text{constant}$ prolate spheroidal functions, capturing at least $1 - c\varepsilon_T^2$ of the energy. Here, $\varepsilon_T \to 0$ as $T \to \infty$.

The number of uses of the AWGN is replaced by $2WT$, and the power on a per use basis is replaced by power on a per degree of freedom basis. Let P denote power on a per unit time basis; then the power on a per degree of freedom basis is $\frac{P}{2W}$. The noise power σ^2 on a per use basis is replaced by the noise power per degree of freedom, $\frac{N_0}{2}$. The formula we get is

$$\frac{1}{T}\left(2WT\frac{1}{2}\log\left(1 + \frac{P/(2W)}{(N_0/2)}\right)\right) = W\log\left(1 + \frac{P}{N_0 W}\right) \quad \text{bits per unit time.}$$

Remark 20.1. Here is a practically important observation for space communication: For fixed P,

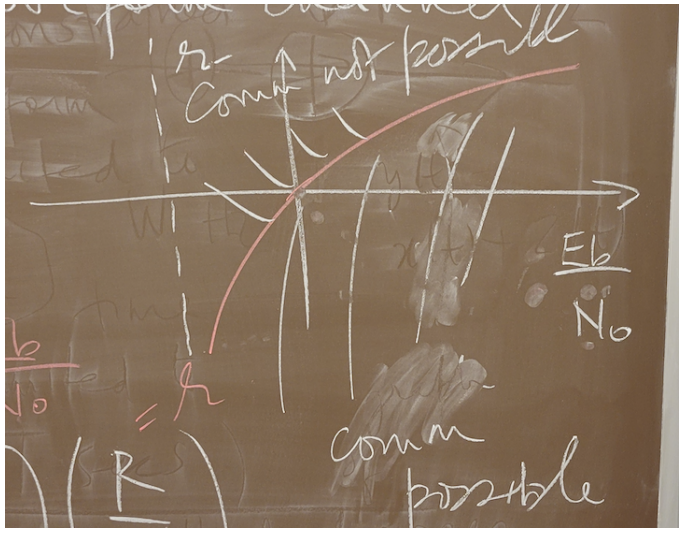
$$\lim_{W\to\infty} W\log\left(1 + \frac{P}{N_0 W}\right) = \frac{P}{N_0}\log_2 e \approx 1.44\frac{P}{N_0} \quad \text{bits per second.}$$

So even with infinite bandwidth, the communication rate is power-limited.

In situations where bandwidth is limited (e.g. terrestrial communication), we call R/W (denoted r) is called the **spectral efficiency** (bits/sec per Hz), and $P/(N_0R)$ (denoted E_b/N_0) is called the **signal to noise per bit**; here R is the rate of communication. Shannon's theorem for the white Gaussian noise waveform channel can be written as saying: We must have

$$r < \log\left(1 + \frac{E_b}{N_0} r\right).$$

This is considered a very insightful restatement of $R < W \log(1 + \frac{P}{N_0 W})$. Here is a graph (in a log-log scale) of the region in which communication is possible:



What is astonishing is that you need at least a minimum value of E_b/N_0 to communicate at all!

20.3 The Shannon capacity of a parallel Gaussian channel

Leading up to the waveform channel Shannon capacity over colored noise, we'll first study the **parallel Gaussian channel** model. At each channel use, say at time i, we have a vector of inputs $(X_i^{(1)}, \ldots, X_i^{(K)})$, each of which has some added independent Gaussian noise $Z_i^{(k)}$. We receive a vector of outputs $(Y_i^{(1)}, \ldots, Y_i^{(K)})$. Here, $Z_i^{(k)} \sim \mathcal{N}(0, \sigma_k^2)$ are independent over i and k for $k = 1, \ldots, K$ and $i = 1, 2, \ldots$.

When coding at block-length n, we require for each message $m \in [M_n]$ that

$$\sum_{i=1}^{n} \sum_{k=1}^{K} (x_i^{(k)}(m))^2 \leq nP.$$

where the term in the sum is the total energy in the codeword for message m.

Theorem 20.1. *In the parallel Gaussian channel model, the Shannon capacity is*

$$\sup_{\sum_{k=1}^{K} \mathbb{E}[(X^{(k)})^2] \leq P} I(X^{(1)}, \ldots, X^{(K)}; Y^{(1)}, \ldots, Y^{(K)})$$

We will discuss this further next time.

21 Shannon Capacity of the Parallel Gaussian Channel Model and Power-Constrained Waveform Channels with Colored Noise

21.1 Shannon capacity of the parallel Gaussian channel model

Last time, we began discussing the parallel Gaussian channel model. We are doing communication at times $i = 1, \dots, n$. At time i, we can send inputs $X_i^{(1)}, \dots, X_i^{(K)}$, and the the receiver receives $Y_i^{(1)}, \dots, Y_i^{(K)}$¡ where $Y_i^{(k)} = X_i^{(k)} + Z_i^{(k)}$, and the $Z_i^{(k)} \sim \text{iid}\,\mathcal{N}(0, \sigma_k^2)$. The power constraint is that for each message $m \in [M_n]$, the codeword

$$\begin{bmatrix} x_1^{(1)}(m) & \cdots & x_n^{(1)}(m) \\ \vdots & & \vdots \\ x_1^{(K)}(m) & \cdots & x_n^{(K)}(m). \end{bmatrix}$$

must satisfy

$$\sum_{i=1}^{n} \sum_{k=1}^{K} (x_i^{(k)})^2 \le nP.$$

Theorem 21.1. *The Shannon capacity is*

$$\sup_{\sum_{k=1}^{K} \mathbb{E}[(X^{(k)})^2] \le nP} I(X^{(1)}, \dots, X^{(K)}; Y^{(1)}, \dots, Y^{(K)}).$$

Proof. We can prove via the usual method of a random coding argument for achievability and Fano's inequality for the converse. □

Choosing the inputs to be independent Gaussians is best (to maximize the mutual information), say $X^{(k)} \sim \mathcal{N}(0, P_k)$ (we must have $\sum_{k=1}^{K} P_k \le P$). This leads to the problem

$$\max_{\sum_{k=1}^{K} P_k = P} \sum_{k=1}^{K} \frac{1}{2} \log \left(1 + \frac{P_k}{\sigma_k^2}\right).$$

Use the Lagrange multiplier technique: The Lagrangian is

$$\mathcal{L}(P_1, \dots, P_k, \lambda) = \sum_{k=1}^{K} \frac{1}{2} \log \left(1 + \frac{P_k}{\sigma_K^2}\right) + \lambda \left(\sum_{k=1}^{K} P_k - P\right).$$

Then

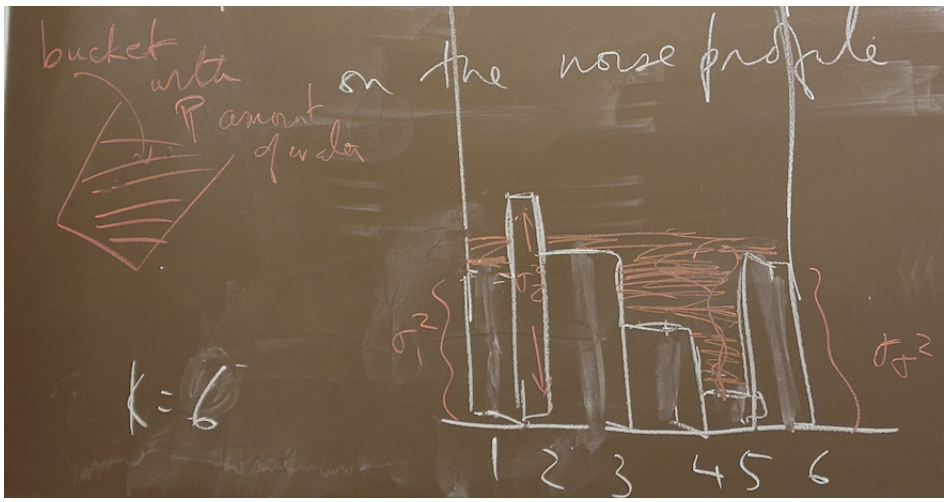
$$\frac{\partial \mathcal{L}}{\partial P_k} = (\log_2 e) \cdot \frac{1/\sigma_k^2}{2(1 + P_k/\sigma_k^2)} + \lambda$$

$$= \frac{\log_2 e}{2} \cdot \frac{1}{\sigma_k^2 + P_k} + \lambda$$

We also need to bring in the non-negativity constraints. With these taken into account, the optimality criterion is that at the optimum, $\frac{\partial \mathcal{L}}{\partial P_k}$ must be ≤ 0 with strict inequality allowed only at $P_k^* = 0$. This leads to

$$\frac{\log_2 e}{2} \frac{1}{\sigma_k^2 + P_k^*} \leq -\lambda^*$$

for all k¡ with strict inequality only if $P_k^* = 0$. That is, $\sigma_k^* + P_k^* =$ constant, except possibly for k such that $P_k^* = 0$. This is waterfilling the available power $P = \sum_{k=1}^{L} P_k^*$ on the noise power. Imagine filling up the following bucket with water:



21.2 Power-constrained waveform channels with colored noise

What does this have to do with waveform channels in colored noise?

Definition 21.1. For a discrete time stationary process $(U_k, k \in \mathbb{Z})$, the **autocorrelation function** is

$$R_{U,U}(m,n) := \mathbb{E}[U_m U_n].$$

This is dependent only on $m - n$, and we may call it $R_{U,U}(m-n)$.

Definition 21.2. We call $(U_n, n \in \mathbb{Z})$ **wide sense stationary (WSS)** if $R_{U,U}(m,n)$ is dependent only on $m - n$ and if $\mathbb{E}[U_n]$ is constant.

Definition 21.3. The **power spectral density** of the process $(U_n, n \in \mathbb{Z})$ (assuming the sampling time is T) is

$$S_{U,U}(f) = \sum_{n=-\infty}^{\infty} R_{U,U}(n) e^{-i2\pi f n T},$$

which is periodic with period $2\pi/T$.

The coefficient of the autocorrelation function can be recovered as

$$\frac{1}{2W}\int_{-W}^{W} e^{-in\frac{\pi f}{W}} S_{U,U}(f)\,df. \quad \text{where} \quad W = \frac{\pi}{T}.$$

If the parallel Gaussian channel model is viewed as coming from quantizing the communication bandwidth into K levels with the noise power roughly flat over those levels, this leads to the capacity formula for power-constrained waveform channels with colored noise:

$$C = \int_{-W}^{W} \frac{1}{2}\log\left(1 + \frac{\max\{\nu - S_{U,U}(f), 0\}}{S_{U,U}(f)}\right) df,$$

where ν is chosen by waterfilling as the unique level with $\int \max\{\nu - S_{U,U}(f), 0\}\,df = P$.

Observe that if you consider the Toeplitz matrix

$$R_{U,U}^{(n)} = \begin{bmatrix} R_{U,U}(0) & R_{U,U}(1) & \cdots & R_{U,U}(n-1) \\ R_{U,U}(1) & \ddots & \ddots & \\ \vdots & & & R_{U,U}(1) \\ R_{U,U}(n-1) & \cdots & R_{U,U}(1) & R_{U,U}(0) \end{bmatrix},$$

then $w^\top R_{U,U}^{(n)} w = \mathbb{E}[(\sum_{\ell=0}^{n1} e_\ell U_\ell)^2]$. This matrix is positive semidefinite, so it has nonnegative, real eigenvalues $\tau_{n,1}, \ldots, \tau_{n,n}$.

Theorem 21.2 (Szegö)**.** *The fraction of these eigenvalues that lie in $(f_0, f_0 + \varepsilon)$ for any f_0 converges to a limit in the sense that for any function $F : \mathbb{R}_+ \to \mathbb{R}$ that is continuous,*

$$\frac{1}{n}\sum_{k=1}^{n} F(\tau_{n,k}) \to \frac{T}{2\pi}\int_{-\pi/T}^{\pi/T} F(S(f))\,df.$$

Where does this theorem come from? Think about this in terms of eigenvalues:

$$R_{U,U}^{(n)} w_k^{(n)} = \tau_{n,k} w_k^{(n)}, \qquad w_k^{(n)} = \begin{bmatrix} w_{k,1}^{(n)} \\ \vdots \\ w_{k,n} \end{bmatrix}$$

normalized to make $\|w_k^{(n)}\|_2 = 1$. Associate to this

$$\psi_k^{(n)}(f) = \sum_{\ell=1}^{n} w_{k,\ell}^{(n)} e^{-i\frac{\ell\pi f}{W}},$$

which is a periodic function of period $2W$. Then

$$\int_{-W}^{W} |\psi_n^{(k)}(f)|^2 \, df = \|w_k^{(n)}\|_2^2 = 1,$$

and

$$\begin{aligned}
\frac{1}{2W} \int_{-W}^{W} |\psi_k^{(n)}(f)|^2 S_W(f) \, df &= \frac{1}{2W} \int_{-W}^{W} \sum_{\ell=1}^{n} \sum_{j=1}^{n} w_{k,\ell}^{(n)} w_{k,j}^{(n)} e^{-i\frac{\pi f}{W}(j-\ell)} S_W(f) \, df \\
&= (w_k^{(n)})^\top R_{U,U}^{(n)} (w_k^{(n)}) \\
&= \tau_{n,k}.
\end{aligned}$$

22 Network Information Theory

In this lecture, we will be discussing **multiuser/network information theory**. There is a recent book on it by Abbas El Gamal and Young-Han Kim.

22.1 Shannon capacity region of a multiuser DMC

In the **multiple access channel** model, there are multiple transmitters and a single receiver. For example, we could think of a cell tower receiving multiple signals. The channel is modeled as in the DMC case and also as in the power constrained Gaussian channel model. We will study the Shannon capacity region in Shannon's block coding formulation. Good news: This is known, unlike many problems in network information theory.

Consider the 2 transmitter case:

- $\mathscr{X}_1$ is the input alphabet of transmitter 1.
- $\mathscr{X}_2$ is the input alphabet of transmitter 2.
- $\mathscr{Y}$ is the output alphabet.
- The channel model in a simple use is $(p(y \mid x_1, x_2) \geq 0, \sum_y p(y \mid x_1, x_2) = 1 \, \forall x_1, x_2)$.
- The encoding map of transmitters 1 and 2 are

$$e_n^{(1)} : [M_n^{(1)}] \mapsto \mathcal{X}_1^n, \qquad e_n^{(2)} : [M_n^{(2)}] \mapsto \mathcal{X}_2^n,$$

  and the decoding map is

$$d_n : \mathcal{Y}^n \mapsto [M_n^{(1)}] \times [M_n^{(2)}].$$

Like a Pavlovian dog, let's turn the Shannon crank.

Definition 22.1. If there exists $((e_n^{(1)}, e_n^{(2)}, d_n), n \geq 1)$ such that

$$\liminf_n \frac{1}{n} \log M_n^{(1)} \geq R_1,$$

$$\liminf_n \frac{1}{n} \log M_n^{(2)} \geq R_2,$$

$$\mathbb{P}(d_n(e_n^{(1)}(W_{1,n}), e_n^{(2)}(W_{2,n})) \neq (W_{1,n}, W_{2,n}) \to 0,$$

where $W_{i,n} \sim \text{Unif}[M_n^{(i)}]$, and $W_{1,n} \amalg W_{2,n}$, we say that the rate pair (R_1, R_2) is **achievable**.

Theorem 22.1. *The closure*[12] *of the set of achievable rate pairs is the closed convex hull of the union of all the sets of rate pairs of the type*

$$\begin{aligned}\mathcal{R}_{p(x_1)p(x_2)} = \{(R_1, R_2) : R_1 &< I(X_1; Y \mid X_2),\\ R_2 &< I(X_2; Y \mid X_1),\\ R_1 + R_2 &< I(X_1, X_2; Y)\}\end{aligned}$$

for some $p(x_1)p(x_2)$, where $p(y \mid x_1, x_2)$ is given by the channel.

In general, each of these regions looks like a polyhedron.

Remark 22.1. A more elegant way to write this region is as

$$\begin{aligned}\{(R_1, R_2) : R_1 &< I(X_1; Y \mid X_2, Q),\\ R_2 &< I(X_2; Y \mid X_1, Q),\\ R_1 + R_2 &< I(X_1, X_2; Y \mid Q)\},\end{aligned}$$

where the joint distribution is

$$p(q)p(x_1 \mid q)p(x_2 \mid q)p(y \mid x_1, x_2),$$

and $Q \in \mathcal{Q}$, a finite set of size ≤ 4.

Proof. Achievability is via a random coding argument. Given $p(x_1)p(x_2)$ and $(R_1, R_2) in \mathcal{R}_{p(x_1)p(x_2)}$ and clock length n, transmitter 1 constructs the random codebook

$$\begin{bmatrix} X_{1,1}(1) & \cdots & X_{1,n}(1) \\ \vdots & & \vdots \\ X_{1,1}(m_1) & \cdots & X_{1,n}(m_1) \\ \vdots & & \vdots \\ X_{1,1}(\lceil 2^{n(R_1-\delta)} \rceil) & \cdots & X_{1,n}(\lceil 2^{n(R_1-\delta)} \rceil) \end{bmatrix},$$

and transmitter 2 constructs the random codebook

$$\begin{bmatrix} X_{2,1}(1) & \cdots & X_{2,n}(1) \\ \vdots & & \vdots \\ X_{2,1}(m_1) & \cdots & X_{2,n}(m_1) \\ \vdots & & \vdots \\ X_{2,1}(\lceil 2^{n(R_2-\delta)} \rceil) & \cdots & X_{2,n}(\lceil 2^{n(R_2-\delta)} \rceil) \end{bmatrix}.$$

[12] We take the closure because this is an engineering class, where we don't want to bother with the boundary.

Let $W_{1,n} = \text{Unif}([M_n^{(1)}])$ and $W_{2,n} = \text{Unif}([M_n^{(2)}])$, where $M_n^{(1)} = 2^{nR_1}$ and $M_n^{(2)} = 2^{nR_2}$. Then decode via

$$d_n(Y^n) = \begin{cases} (m_1, m_2) & \text{if there is a unique } (m_1, m_2) \text{ such that} \\ & \quad ((X_1)_1^n(m_1), (X_2)_1^n(m_2), Y_1^n) \text{ is } \varepsilon\text{-jointly weakly typical} \\ \text{arbitrary} & \text{if there is no such } (m_1, m_2) \text{ or more than 1 such } (m_1, m_2). \end{cases}$$

Then, by symmetry,

$$\begin{aligned} &\mathbb{P}(d_n(e_n^{(1)}(W_{1,n}), e_n^{(2)}(W_{2,n})) \neq (W_{1,n}, W_{2,n})) \\ &\quad = \mathbb{P}(d_n(e_n^{(1)}(1), e_n^{(2)}(1)) \neq (1,1)) \\ &\quad \leq \mathbb{P}(E_{1,1}^c) + \sum_{i \neq 1} \mathbb{P}(E_{i,1}) + \sum_{j \neq 1} \mathbb{P}(E_{1,j}) + \sum_{i \neq 1, j \neq 1} \mathbb{P}(E_{i,j}), \end{aligned}$$

where $E_{i,j}$ is the event that $((X_1)_1^n(i), (X_2)_1^n(j), Y_1^n)$ us ε-jointly weakly typical. Then $\mathbb{P}(E_{1,1}) \to 0$ by the weak law of large numbers, and

$$\begin{aligned} \mathbb{P}(E_{i,1}) &= \sum_{((x_1)_1^n, (x_2)_1^n, y_1^n) \in A_\varepsilon^{(n)}} p((x_1)_1^n) p((x_2)_1^n, y_1^n) \\ &\leq |A_\varepsilon^{(n)}| 2^{-nH(X_1)} 2^{n\varepsilon} \\ &\leq 2^{nH(X_1, X_2, Y)} 2^{n\varepsilon} 2^{-nH(X_2, Y)} 2^{n\varepsilon} 2^{-nH(X_1)} 2^{n\varepsilon} \\ &= 2^{-nI(X_1; X_2; Y)} 2^{3n\varepsilon} \\ &= 2^{-nI(X_1; Y | X_2)} 2^{3n\varepsilon} \end{aligned}$$

because $I(X_1; X_2) = 0$. Hence,

$$\sum_{i \neq 1} \mathbb{P}(E_{i,1}) \leq 2^{n(R_1 - \delta)} 2^{-nI(X_1; X_2 | Y)} 2^{3n\varepsilon},$$

so if $R_1 < I(X_1; X_2 \mid Y) - 3\varepsilon + \delta$, then this goes to 0 as $n \to \infty$. We can apply a similar argument to $\mathbb{P}(E_{1,j})$.

When $i \neq 1$ and $j \neq 1$,

$$\begin{aligned} \mathbb{P}(E_{i,j}) &\leq \sum_{((x_1)_1^n, (x_2)_1^n, y_1^n) \in A_\varepsilon^{(n)}} \frac{\underbrace{p((x_1)_1^n)}}{\prod_{t=1}^n p_{X_1}(x_{1,t})} \frac{\underbrace{p((x_2)_1^n)}}{\prod_{t=1}^n p_{X_2}(x_{2,t})} \; p(y_1^n) \\ &\leq |A_\varepsilon^{(n)}| 2^{-H(X_1)} 2^{-n\varepsilon} 2^{-nH(X_2)} 2^{n\varepsilon} 2^{-nH(Y)} 2^{n\varepsilon} \\ &\leq 2^{n(H(X_1, X_2, Y) - H(X_1) - H(X_2) - H(Y))} 2^{4n\varepsilon}. \end{aligned}$$

This tells us that if $R_1 + R_2 \leq I(X_1, X_2; Y) - 4\varepsilon + 2\delta$, then $\sum_{i \neq 1, j \neq 1} p(E_{i,j}) \to 0$ as $n \to \infty$.

For the converse, we use Fano's inequality 3 times. For any $((e_n^{(1)}, e_n^{(2)}, d_n), n \geq 1$,

$$\begin{aligned} nR_1 + o(n) &= \log\lceil 2^{nR_1} \rceil \\ &= H(W_1) \\ &= I(_1; Y_1^n) + H(W_1 \mid Y_1^n) \\ &\leq I(W_1 \mid Y_1^n) + n\varepsilon_n, \end{aligned}$$

where $\varepsilon_n \to 0$ from Fano's inequality because $H(W_1 \mid Y_1^n) \leq H(W_1, W_2 \mid Y_1^n)$ and $H(W_1, W_2 \mid Y_1^n) \leq h(p_{\text{error}}^{(n)}) + n(R_1 + R_2)p_{\text{error}}^{(n)}$.

$$\begin{aligned} &\leq I((X_1)1^n(W_1); Y_1^n) \\ &= H((X_1)_1^n(W_1)) - H((X_1)_1^n(W_1) \mid Y_1^n) + n\varepsilon_n \\ &= H((X_1)_1^n(W_1) \mid (X_1)_1^n(W_2)) - H((X_1)_1^n(W_1) \mid Y_1^n, (X_2)_1^n(W_2)) + n\varepsilon_n \\ &= I((X_1)_1^n(W_1); Y_1^n \mid (X_2)_1^n(W_2)) + n\varepsilon_n \\ &= H(Y_1^n \mid (X_2)_1^n(W_2)) - H(Y_1^n \mid (X_1)_1^n(W_1), (X_2)_1^n(W_2)) + n\varepsilon_n \\ &\leq \sum_{i=1}^n H(Y_i \mid (X_2)_1^n(W_2)) - \sum_{i=1}^n H(Y_i \mid X_{1,i}(W_1), X_{2,i}(W_2)) + \varepsilon_n \\ &= \sum_{i=1}^n I(Y_i; X_{1,i} \mid X_{2,i}) + n\varepsilon_n. \end{aligned}$$

We get $R_1 \leq \frac{1}{n}\sum_{i=1}^n I(Y_i; X_{1,n} \mid X_{2,i}) + \varepsilon_n$ and similar bounds for R_2 and $R_1 + R_2$. □

22.2 Achievable rate pairs of a multiuser AWGN channel

In the case of Gaussian noise, we have input X_1 with power constraint P_1 and input X_2 with power constraint P_2. WIth $\mathcal{N}(0, \sigma^2)$ noise, the result is more explicit:

Theorem 22.2. *With Gaussian noise, the set of achievable rate pairs is*

$$\left\{(R_1, R_2) : \begin{aligned} R_1 &\leq \frac{1}{2}\log\left(1 + \frac{P_1}{\sigma^2}\right) \\ R_2 &\leq \frac{1}{2}\log\left(1 + \frac{P_2}{\sigma^2}\right) \\ R_1 + R_2 &\leq \frac{1}{2}\log\left(1 + \frac{P_1 + P_2}{\sigma^2}\right) \end{aligned}\right\}.$$

23 Two Receiver Broadcast Channels

23.1 Degraded two receiver broadcast channels

The two receiver broadcast channel (for a discrete memoryless channel) is defined via

- $p(y_1, y_2 \mid x)$, which is nonnegative with $\sum_{y_1,y_2} p(y_1, y_2 \mid x) = 1$ for all x,
- Input alphabet $x \in \mathscr{X}$,
- Output alphabet $\mathscr{Y}_1$ of receiver 1,
- Output alphabet $\mathscr{Y}_2$ of receiver 2,
- Memorylessness of the channel, given by

$$p(y_{1,[1:n]}, y_{2,[1:n]} \mid x_{[1:n]}) = \prod_{i=1}^{n} p(y_{1,i}, y_{2,i} \mid x_i),$$

  where $y_{1,[1:n]}$ is new notation for $(y_1)_1^n$,
- Encoding map $e_n : [M_n^{(1)}] \times [M_n^{(2)}] \to \mathscr{X}^n$ of block length n,
- Decoding map $d_n : \mathscr{Y}^n \to [M_n^{(1)}] \times [M_n^{(2)}]$ of block length n,
- Rate region given by the closure of the set

$$\begin{aligned}\{(R_1, R_2) : \exists((e_n, d_n), n \geq 1) \text{ s.t. } & \liminf_n \tfrac{1}{n} \log M_n^{(1)} \geq R_1, \\ & \liminf_n \tfrac{1}{n} \log M_n^{(2)} \geq R_2, \\ & \lim_{n\to\infty} P(d_n(e_n(W_{1,n}, W_{2,n})) \neq (W_{1,n}, W_{2,n})) = 0, \}\end{aligned}$$

  where $W_{1,n} \sim \text{Unif}([M_n^{(1)}]), W_{2,n} \sim \text{Unif}([M_n^{(2)}])$.

The bad news is that finding the rate region has been an open problem for about 50 years. A special case where the rate region is known is called the *stochastically degraded* case.

Definition 23.1. $p(y_1, y_2 \mid x)$ is called **physically degraded** if

$$p(y_1, y_2 \mid x) = p(y_1 \mid x)p(y_2 \mid y_1).$$

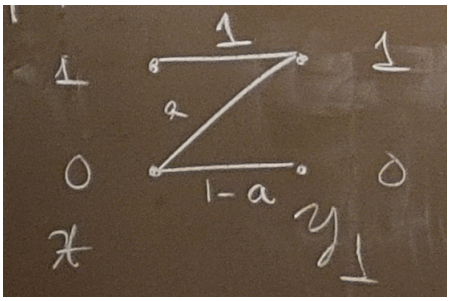
It is called **stochastically degraded** if there exists some distribution $p'(y_2 \mid y_1)$ such that

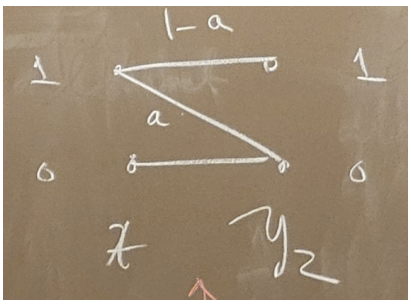
$$p(y_2 \mid x) = \sum_{y_1} p(y_1 \mid x)p'(y_2 \mid y_1).$$

The physical degradation condition means that we have the Markov chain $X - Y_1 - Y_2$. The stochastic degradation condition does not require $X - Y_1 - Y_2$ but is "equivalent" since the rate region only depends on $p(y_1 \mid x)$ and $p(y_2 \mid x)$.

Example 23.1 (Stochastically but not physically degraded channel)**.** Let $\mathscr{X} = \mathscr{Y}_1 = \mathscr{Y}_2 = \{0, 1\}$, and suppose that $Y_1 = X \oplus Z$, where $Z \in \{0, 1\}$, $\mathbb{P}(Z = 1) = a = 1 - \mathbb{P}(Z = 0)$. Here, $0 < a < 1$. Also, let $Y_2 = Z$, where $Z \amalg X$. This is not a physically degraded channel, since $X - Y_1 - Y_2$ is false (e.g. knowing both X and Y_1 determines Y_2). But it is stochastically degraded because we can replace Y_2 by Z', where $Z' \stackrel{d}{=} Z$, $Z' \amalg (X, Z)$.

Example 23.2 (Broadcast channel that is not stochastically degraded)**.** Let $\mathscr{X} = \mathscr{Y}_1 = \mathscr{Y}_2 = \{0, 1\}$ with $p(y_1 \mid x)$ given by a Z-channel and $p(y_2 \mid x)$ given by a different Z-channel.





We claim that there cannot be any $p'(y_2 \mid y_1)$ such that the stochastic degradation condition holds, i.e.

$$p(y_2 \mid x) = sum_{y_1} p(y_1 \mid x) p'(y_2 \mid y_1).$$

If such a p' existed, then

$$\begin{aligned} 0 &= p_{Y_2|X}(1 \mid 0) \\ &= p_{Y_1|X}(1 \mid 0) p'(1 \mid 1) + p_{Y_1|X}(0 \mid 0) p'(1 \mid 0). \end{aligned}$$

That is,

$$0 = p'(1 \mid 1) + (1 - a) p'(1 \mid 0),$$

so

$$p'(1 \mid 1) = p'(1 \mid 0) = 0,$$

which makes

$$p'(0 \mid 1) = p(0 \mid 0) = 1.$$

Then $p(y_2 \mid x) = \sum_{y_1} p(y_1 \mid x) p'(y_2 \mid y_1)$ gives the the wrong channel.

23.2 Capacity region for a stochastically degraded broadcast channel

Theorem 23.1. *The capacity region for independent private messages over a stochastically degraded broadcast channel is the closure of the convex hull of*

$$\{(R_1, R_2) : R_2 \le I(U; Y_2), R_1 \le I(X; Y_1 \mid U)\}$$

for some $p(x)p(x \mid u)p(y_1, y_2 \mid x)$*, where* $U \in \mathscr{U}$ *and* $|\mathscr{U}| \le \max\{|\mathscr{X}|, |\mathscr{Y}_1|, |\mathscr{Y}_2|\}$.

Think of these U variables as information that receiver 1, the stronger receiver, can use to get a better signal.

Proof. We will use a random coding achievability argument. The codebook is going to be comprised of $2^{n(R_1-\delta)}2^{n(R_2-\delta)}$ codewords in $\mathscr{X}^n$, organized as $2^{n(R_2-\delta)}$ clusters, each with $2^{n(R_1-\delta)}$ codewords.

Generate $2^{n(R_2-\delta)}$ independent sequences $(U_1(m_2), \dots, W_m(m_2))$ with $1 \le m_2 2^{n(R_2-\delta)}$, and entries $\overset{\text{iid}}{\sim} p(u)$. For each m_2, generate $2^{n(R_1-\delta)}$ sequences $(X_1(m_1, m_2), \dots, X_n(m_1, m_2))$ with $1 \le m_1 \le 2^{n(R_1-\delta)}$ and, for each m_1, joint law $\prod_{i=1}^n p(x_i \mid U_i(m_2))$ (independently over m_1).

To send (m_1, m_2) the transmitter sends $(X_1(m_1, m_2), \dots, X_n(m_1, m_2))$. Receiver 2, receiving $(Y_{2,1}, \dots, Y_{2,n})$, determines all m_2 such that $(U_{[1:n]}(m_2), Y_{2,[1:m]})$ is ε-jointly weakly typical. If there is only one such message, it decodes as m_2. If there are none or more than one such message, it decodes arbitrarily. Receiver 1, receiving $(Y_{1,1}, \dots, Y_{1,n})$, finds all (m_1, m_2) such that $(U_{[1:n]}(m_2), X_{[1:n]}(m_1, m_2), Y_{1,[1:n]})$ is ε-jointly weakly typical. If there is only one such message, it decodes as m_1. If there are none or more than one such message, it decodes arbitrarily.

If we take the probability over the random codebook, W_1, and W_2, symmetry gives us

$$\mathbb{P}(d_n(e_n(W_{1,n}, W_{2,n})) \ne (W_{1,n}, W_{2,n}))] = \mathbb{P}(d_n(e_n(1,1)) \ne (1,1)),$$

so we can condition on the message pair $(m_1, m_2) = (1, 1)$ being sent.

The error events for receiver 2 are

$$E_n^{(2)} = \{(U_{[1:n]}(1), Y_{2,[1:n]}) \notin A^{(n)}_{\varepsilon,(U,Y_2)}\}, \qquad E_{n,i}^{(2)} = \{(U_{[1:n]}(i), Y_{2,[1:n]}) \in A^{(n)}_{\varepsilon,(U,Y_2)}\}$$

for $i \ne 1$. By the weak law of large numbers,

$$\mathbb{P}(E_n^{(2)}) \xrightarrow{n\to\infty} 0$$

On the other hand,

$$\mathbb{P}(E_{n,i}^{(2)}) \le 2^{-nI(U;Y_2)}2^{3n\varepsilon},$$

so we want $2^{n(R_2-\delta)}2^{-nI(U;Y_2)}2^{n3\varepsilon} \to 0$, i.e. $R_2 < U(U; Y_2) - 3\varepsilon + \delta$.

The error events for receiver 1 are

$$E_n^{(1)} = \{(U_{[1:n]}(1), X_{[1:n]}(1,1), Y_{1,[1:n]}) \notin A^{(n)}_{\varepsilon,(U,X,Y_1)}\}, \qquad E_{n,i}^{(1)} = \{(U_{[1:n]}(i) Y_{1,[1:n]}) \in A^{(n)}_{\varepsilon,(U,Y_2)}\}$$

for $i \neq 1$. By the weak law of large numbers,

$$\mathbb{P}(E_n^{(1)}) \xrightarrow{n\to\infty} 0.$$

On the other hand,

$$\mathbb{P}(E_{n,i}^{(1)}) \leq 2^{-nI(U;Y_1)} 2^{3n\varepsilon}.$$

There are $2^{n(R_2-\delta)}$, and $I(U;Y_1) \geq I(U;Y_2)$, so the earlier condition on R_2 ensures $\sum_{i\neq 1} \mathbb{P}(E_{n,i}^{(1)}) \to 0$.

For $j \neq 1$, we also have the error event

$$E_{n,1,j}^{(1)} = \{(U_{[1:n]}(1), X_{[1:n]}(j,1), Y_{1,[1:n]}) \in A^{(n)}_{\varepsilon,(U,X,Y_1)}\}.$$

Then

$$\mathbb{P}(E_{n,1,j}^{(1)}) = \sum_{u_{[1:n]}, x_{[1:n]}, y_{1,[1:n]} \in A_\varepsilon^{(n)}} 2^{-nH(U,Y_1)} 2^{n\varepsilon} 2^{-nH(X|U)} 2^{n\varepsilon}$$

The size of $A_\varepsilon^{(n)}$ is $\leq 2^{nH(U,X,Y_1)} 2^{n\varepsilon}$.

$$\leq 2^{-nI(X;Y_1|U)} 2^{n4\varepsilon}.$$

The converse part of the proof is homework. □

23.3 Capacity region for a stochastically degraded Gaussian broadcast channel

The Gaussian case (with power constrained to P, receiver 1 noise $\mathcal{N}(0,\sigma_1^2)$, and receiver noise $\mathcal{N}(0,\sigma_2^2)$ with $\sigma_2^2 > \sigma_1^2$) is automatically stochastically degraded.

Theorem 23.2. *The rate region is the union of the sets of the form*

$$\{(R_1, R_2) : R_2 \leq C((1-\alpha)P, \alpha P + \sigma_2^2), R_1 \leq C(\alpha P, \sigma_1^2)\}$$

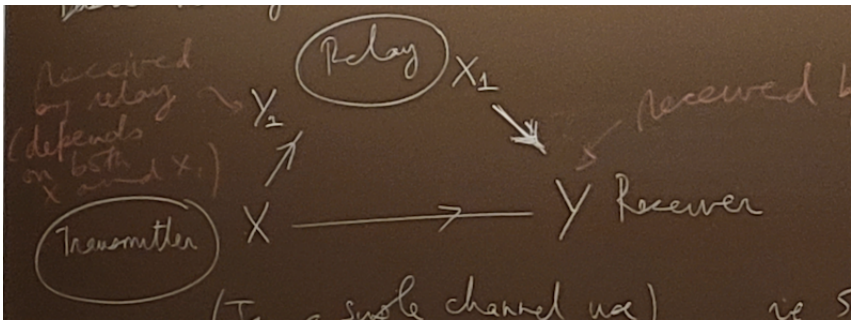
over $0 < \alpha < 1$*, where*

$$C(P, \sigma^2) = \frac{1}{2} \log\left(1 + \frac{P}{\sigma^2}\right).$$

24 The Relay Channel Model, One Shot Information Theory, and Rate Distortion Theory

24.1 The relay channel model

The basic relay channel model (in the discrete memoryless case) has a transmitter, a relay, and a receiver. In a single channel use, the transmitter inputs X. The relay receives Y_1 (which depends on X and X_1) and sends X_1, and the receiver receives Y, which depends both on X and X_1.



The channel is described by $p(y, y_1 \mid x, x_1)$ with $y \in \mathcal{Y}$, $y_1 \in \mathcal{Y}_1$, $x \in \mathscr{X}$, and $x_1 \in \mathscr{X}$.

We use our Shannon persona to study the Shannon capacity asymptotically as block length goes to ∞. The new twist is that in deciding the k-th input with $1 \leq k \leq n$, the relay can use the past $k-1$ observations. The overall probability distribution is

$$p(m)p(x_{[1:n]} \mid m) \prod_{i=1}^{n} p(x_{1,i} \mid \underbrace{y_{1,1}, \ldots, y_{1,i-1}}_{y_{1,[1:i-1]}}) \prod_{i=1}^{n} p(y_i, y_{1,i} \mid x_i x_{1,i})$$

in either a deterministic or random coding scheme (for proof purposes), where $m \in [M_n] = [2^{nR}]$.

In a fixed coding scheme,

$$p(x_{[1:n]} \mid m) = \mathbb{1}_{\{e_n(m)=x_{[1:n]}\}},$$

where $e_n : [M_n] \to \mathscr{X}^n$ is an encoding map, and

$$p(x_{1,i} \mid y_{1,[1:i-1]}) = \mathbb{1}_{\{f_i(y_{1,[1:i-1]})=x_{1,i}\}},$$

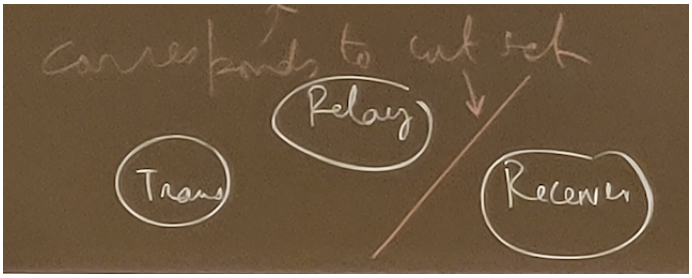
where $f_1, \ldots, f_n$ are the relay's encoding rules. We also need the decoding map $d_n : \mathscr{Y}^n \to [M_n]$.

The Shannon capacity region, defined as usual as the supremum of rates at which the error probability (asymptotically in n) goes to zero, is unknown. Here is a basic theorem in this area.

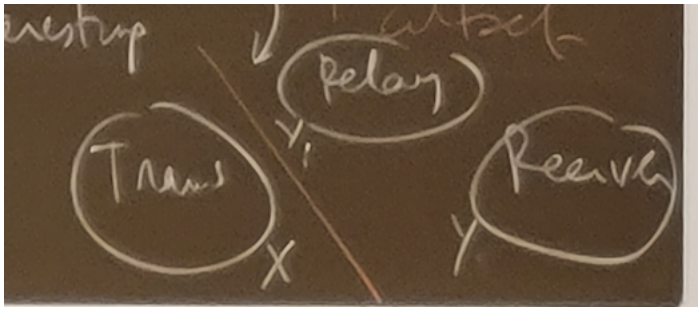
Theorem 24.1 (Cut-set bound)**.**

$$C \leq \sup_{p(x,x_1)} \min\{I(X, X_1; Y), I(X; Y, Y_1 \mid X_1)\}.$$

The bound in terms of $I(X, X_1; Y)$ should be thought of as the case where the transmitter and the relay can communicate freely; here is the following cut-set in a picture:



The bound in terms of $I(X; Y, Y_1 \mid X_1)$ should be thought of as the case where the relay and the receiver can communicate freely. This cut-set looks like



The first term is more straightforward, while the second is more interesting. We'll omit the details of the proof and include them in a handout later.

24.2 One shot information theory

One shot information theory (in the discrete memoryless case) involves a single use of a DMC. A message $m \in \{1, \dots, L\}$ is encoded as $x(m) \in \mathscr{X}$, received as y via $[p(y \mid x)]$ through the channel, and decoded as $d(y) = \widehat{m}$. We want to study $p_e := \mathbb{P}(\widehat{W} \neq W)$ as a function of L, where $W \sim \text{Unif}(\{1, \dots, L\})$.

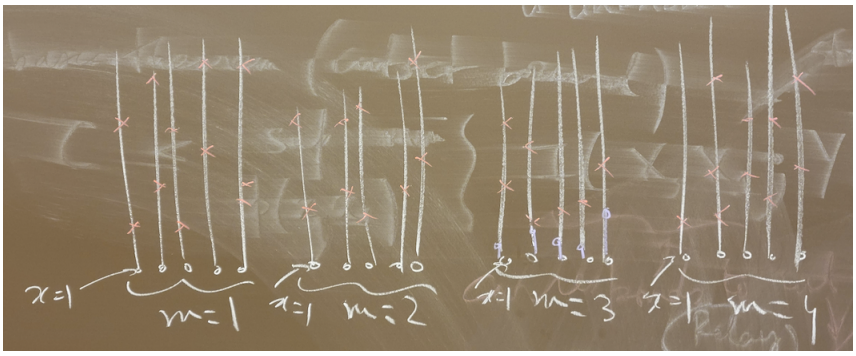
Theorem 24.2 (Poisson matching lemma)**.** *For any input distribution distribution p_X,*

$$p_e \leq \mathbb{E}\left[1 - \frac{1}{1 + L2^{-i_{X;Y}(X;Y)}}\right],$$

where $i_{X,Y}(x; y) := \log \frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)}$.

This uses a Poisson process on $\mathscr{X} \times [L] \times \mathbb{R}_+$. You may think of this as one copy of $\mathbb{R}_+$ for each $x \in X, 1 \le m \le 1$ and an independent rate 1 Poisson process on each line. That is, the points are placed with iid Exp(1) interarrival times.

Example 24.1. Here is what this looks like when $|\mathscr{X}| = 5$ and $L = 4$.



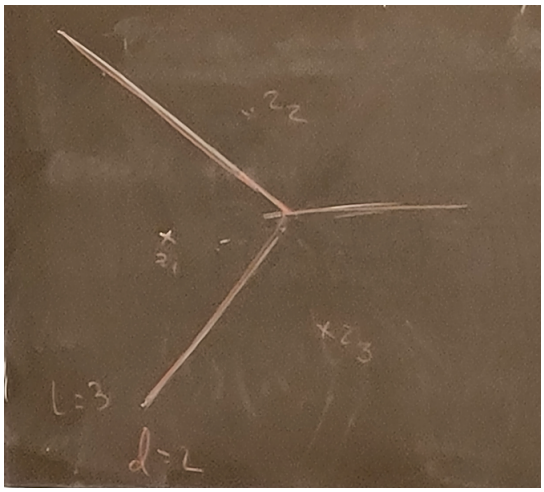
This Poisson structure is shared randomness between transmitter and receiver. The transmitter knows m and scales up the profile $(p_x(x), x \in \mathscr{X})$ in the m-th group until it hits a point in the Poisson process. If that is in line x, input x into the channel. The receiver scales up the distribution $\frac{1}{L} p_{X|Y}(x \mid Y)$ (which is computed from p_X and $p_{Y|X}$ using Bayes' rule) until it heats a point of the Poisson process. Then the receiver returns $\widehat{m}$, which is the block of lines in which the hit occurs.

24.3 Rate distortion theory

Rate distortion theory is a "Shannon mindset theory" which tries to do an asymptotic version of vector quantization. Here is the basic vector quantization problem. Say $Z \in \mathbb{R}^d$ is random, and you are allowed to place L points $z_1, \dots, z_: \in \mathbb{R}^d$. The aim is to minimize $\mathbb{E}[\min_{1 \le \ell \le L} (Z - z_\ell)^2]$.

Given $z_1, \dots, z_L$, $\mathbb{R}^d$ gets decomposed into **Voronoi cells**, which are the points closest

to a given z_ℓ.



But given a region D, the best choice of $z \in \mathbb{R}^d$ to map that region to will be the one that minimizes $\mathbb{E}[(Z-z)^2 \mathbb{1}_{\{Z \in D\}}]$.

In the rate distortion formulation, the block length is n, the alphabet is $\mathscr{X}$, and the finite source sequence $x_{[1:n]} \in \mathscr{X}^n$ can be represented by 2^{nR} points via $f_n : \mathscr{X}^n \to \{1, \ldots, 2^{nR}\}$. The decompressor sees $f_n(x_{[1:n]})$ and reproduces it as $\widehat{x}_{[1:n]} \in \widehat{\mathscr{X}}^n$ via $g_n : [2^{nR}] \to \widehat{\mathscr{X}}^n$. The aim is to minimize

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} d(X_i, \widehat{X}_i)\right].$$

Here, $X_1, \ldots, X_n$ are iid, $d : \mathscr{X} \times \widehat{\mathscr{X}} \to \mathbb{R}$ is some distortion measure, and $\widehat{X}_{[1:n]} = g_n(f_n(X_{[1:n]}))$.

25 Rate Distortion Theory

25.1 Shannon's rate distortion theorem

In rate distortion theory,, we have an iid $\mathscr{X}$-valued source $X_1, \ldots, X_n$. At the compressor, we have

$$f_n : \mathscr{X}^n \to \{1, 2, \ldots, 2^{nR}\}.$$

Assume $f_n(X^n)$ is perfectly received at the decompressor. The decompressor uses a map $g_n : \{1, \ldots, 2^{nR}\} \to \widehat{\mathscr{X}}^n$, where $\widehat{\mathscr{X}}$ could be different from $\mathscr{X}$. Call (f_n, g_n) a $(2^{nR}, n)$ **distortion code**. We are also given a cost metric $d : \mathscr{X} \times \widehat{\mathscr{X}} \to \mathbb{R}_+$ called the **distortion measure**.

Definition 25.1. (R, D) is called **achievable** if there exists $((f_n, g_n), n \geq 1)$ of $(2^{nR}, n)$ distortion codes such that

$$\limsup_n \frac{1}{n} \mathbb{E}[d(X^n, g_n(f_n(X^n)))] \leq D,$$

where $d(x^n, \widehat{x}^n)$ denotes $\sum_{i=1}^n d(x_i, \widehat{x}_i)$.

Theorem 25.1 (Shannon's rate distortion theorem)**.** *Let*

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

where $p(x)$ is the marginal distribution of the source. (R, D) is achievable if $R > R^{(I)}(D)$ and not achievable if $R < R^{(I)}(D)$.

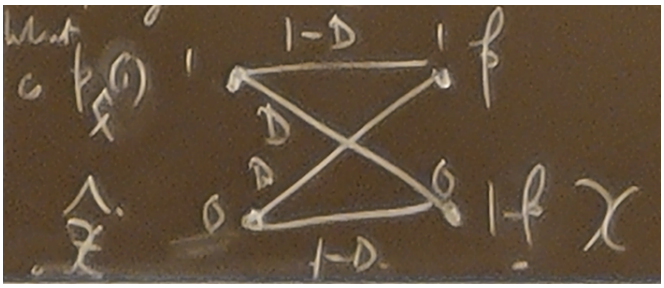
We write $R^{(I)}(D)$ as $R(D)$, the **rate distortion function**.

Example 25.1 (Bernoulli source)**.** Let $\mathscr{X} = \{0, 1\}$ with $p(1) = p$ and $p(0) = 1 - p$, reproduction alphabet $\widehat{\mathscr{X}} = \{0, 1\}$, and distortion measure $d(0, 0) = 0 = d(1, 1)$, $d(0, 1) = 1 = d(1, 0)$. Here,

$$R(D) = \begin{cases} h(p) - h(D)Y & 0 \leq D \leq \min\{p, 1 - p\} \\ 0 & \text{otherwise} \end{cases}$$

To see this, if $D > \min\{p, 1-p\}$ then if $p < 1/2$, represent all binary sequences of length n by 0^n; if $p > 1/2$, represent all binary sequences of length n by 1^n. If $d \leq \min\{p, 1 - p\}$, we can choose the optimizing $p(\widehat{x} \mid x)$ by defining the corresponding $p(x \mid \widehat{x})$ via a binary symmetric channel with crossover probability D ($p(\widehat{x})$ has to be chosen correctly to get the

correct $p(x)$).



We must have

$$p_{\widehat{X}}(1)(1-D) + (1-p_{\widehat{X}}(1))D = p$$

which gives

$$p_{\widehat{X}}(1) = \frac{p-D}{1-2D}.$$

This makes sense because $D \le \min\{p, 1-p\}$ and $D \le 1/2$.

If we made this choice, then

$$\begin{aligned} I(X;\widehat{X}) &= H(X) - H(X \mid \widehat{X}) \\ &= h(p) - h(D). \end{aligned}$$

To show that this is the best choice, we need to show that $I(X;\widehat{X}) \ge h(p) - h(D)$ for all other choices of $p(\widehat{x} \mid x)$. This holds because for any other choice of $p(\widehat{x} \mid x)$,

$$\begin{aligned} I(\widehat{X};X) &= H(X) - H(X \mid \widehat{X}) \\ &= h(p) - H(X \mid \widehat{X}) \\ &= h(p) - H(X \oplus \widehat{X} \mid \widehat{X}) \\ &\ge h(p) - H(X \oplus \widehat{X}) \\ &\ge h(p) - h(D). \end{aligned}$$

25.2 Proof of the rate distortion theorem

Let's prove the theorem.

Proof. Converse: We want to show that if $R < R^{(I)}(D)$¡ then (R, D) is not achievable. First, observe that $R^{(I)}(D)$ is a convex function of D (using $I(X;\widehat{X})$ is convex in $[p(\widehat{x} \mid x)]$ for fixed $(p(x), x \in \mathscr{X})$). Consider any sequence $((f_n, g_n), n \ge 1)$ of $(2^{nR}, n)$ rate distortion codes. Then

$$\begin{aligned} nR &\ge H(\widehat{X}^n) \\ &\ge I(X^n;\widehat{X}^n) \end{aligned}$$

$$= H(X^n) - H(X^n \mid \widehat{X}^n)$$

Use the chain rule

$$= H(X^n) - \sum_{i=1}^{n} H(X_i \mid X^{i-1}, \widehat{X}^n)$$

$$= \sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} H(X_i \mid X^{i-1}, \widehat{X}^n)$$

Conditioning on more decreases the entropy, so

$$\geq \sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} H(X_i \mid \widehat{X}_i)$$

$$= \sum_{i=1}^{n} I(X_i; \widehat{X}_i)$$

If D_i denotes $\mathbb{E}[d(X_i, \widehat{X}_i)]$, then $I(X_i; \widehat{X}_i) \geq R^{(I)}(D_i)$.

$$\geq \sum_{i=1}^{n} R^{(I)}(D_i)$$

By convexity of $R^{(I)}$,

$$= nR^{(I)}(D).$$

Achievability: We use a random coding argument. Given $p(x, \widehat{x})$, define the set

$$A_{d,\varepsilon}^{(n)} := \left\{ (x^n, \widehat{x}^n) : (x^n, \widehat{x}^n) \in A_\varepsilon^{(n)}, \left| \frac{1}{n} \sum_i d(x_i, \widehat{x}_i) - \mathbb{E}[d(X, \widehat{X})] \right| < \varepsilon \right\}.$$

We can show that $\mathbb{P}((X^n, \widehat{X}^n) \in A_{d,\varepsilon}^{(n)}) \to 1$ as $n \to \infty$, where $(X_i, \widehat{X}_i)$ are iid $\sim p(x, \widehat{x})$. We can also show that

$$(1-\varepsilon)2^{nH(X,\widehat{X})}2^{-n\varepsilon} \leq |A_{d,\varepsilon}^{(n)}| \leq 2^{nH(X,\widehat{X})}2^{n\varepsilon},$$

where the lower bound holds for all sufficiently large n. If $X^n \amalg \widetilde{\widehat{X}}^n$, where X^n is iid $\sim p(x)$ and $\widehat{X}^n$ is iid $\sim p(\widehat{x})$, then

$$(1-\varepsilon)2^{-nI(X;\widehat{X})}2^{-n3\varepsilon} \leq \mathbb{P}(X^n, \widetilde{\widehat{X}}^n \in A_{d,\varepsilon}^{(n)}) \leq 2^{-nI(X;\widehat{X})}2^{n3\varepsilon}.$$

So generate 2^{nR} sequences

$$\begin{bmatrix} \widehat{X}_1(1) & \cdots & \widehat{X}_n(1) \\ \vdots & & \vdots \\ \widehat{X}_1(2^{nR}) & \cdots & \widehat{X}_n(2^{nR}) \end{bmatrix}$$

with entries iid over the coordinates and $\sim p(\widehat{x})$. To construct $f_n : \mathscr{X}^n \to \{1, \ldots, 2^{nR}\}$, on seeing x^n, find a row ℓ such that $(x^n, \widehat{X}^n(\ell)) \in A_{d,\varepsilon}^{(n)}$ if such exists. Then define $g_n : \{1, \ldots, 2^{nR}\} \to \widehat{\mathscr{X}}^n$ by $g_n(\ell)$ is row ℓ.

We claim that if $R > I(X; \widehat{X})$, then $\mathbb{P}(\text{row exists for } X^n) \to 1$ as $n \to \infty$. If $R > I(X; \widehat{X}) + 3\varepsilon$, then

$$\left(1 - (1-\varepsilon)2^{-nI(X;\widehat{X})}2^{-3n\varepsilon}\right)^{2^{nR}} \to 0.$$

This completes the proof. □

25.3 The rate distortion function with a Gaussian source

For an iid Gaussian source $X_1, \ldots, X_n$ iid $\sim \mathcal{N}(0, \sigma^2)$, $\mathscr{X} = \widehat{\mathscr{X}} = \mathbb{R}$, and distortion $d(x, \widehat{x}) = (x - \widehat{x})^2$ with the goal of asymptotical per letter distortion at most D (i.e. $\frac{1}{n}\mathbb{E}[\sum_{i=1}^n (X_i - \widehat{X}_i)^2] \le D$ with $\widehat{X}^n = g_n(f_n(X^n))$), we have

$$R^{(I)}(D) = \begin{cases} \frac{1}{2}\log\frac{\sigma^2}{D} & \text{if } D < \sigma^2 \\ 0 & \text{if } D > \sigma^2. \end{cases}$$

The first case is achieved via $Z \sim \mathcal{N}(0, \sigma^2 - D)$, where $\widehat{X} \amalg Z$. Here, $I(X; \widehat{X}) = h(X) - h(X \mid \widehat{X}) = \frac{1}{2}\log\frac{\sigma^2}{D}$.

26 Convex Dual of the Cumulant Generating Function and Sanov's Theorem

26.1 The cumulant generating function and convex duality

Suppose $X \in \mathbb{R}^d$ is a random variable.

Definition 26.1. The map $\theta \mapsto \mathbb{E}[e^{\theta^\top X}]$ with $\theta \in \mathbb{R}^d$ is called the **moment generating function**.

Definition 26.2. The map $\theta \mapsto \log \mathbb{E}[e^{\theta^\top X}]$ with $\theta \in \mathbb{R}^d$ is called the **cumulant generating function**.

If we differentiate the moment generating function with respect to θ and set $\theta = 0$, we get the moments of X. Likewise, doing the same to the cumulant generating function gives us the cumulants of X. One advantage of working with the cumulant generating function is that it is convex.

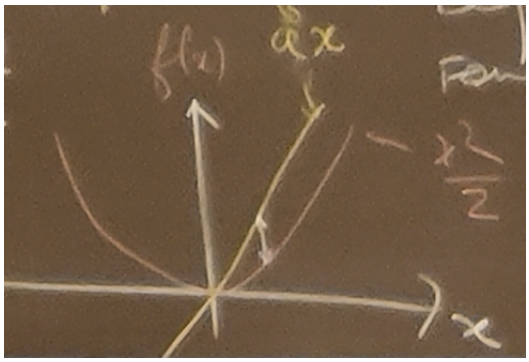
We have dealt with finite (and countable) random variables and some densities. For a finite random variable $X \in \mathscr{X}$ with $|\mathscr{X}| = d$,it is interesting to consider $Z \in \mathbb{R}^d$ where $Z = e_i$ with probability p_i (here, e_i is the i-th basis vector). Then

$$\log \mathbb{E}[e^{\theta^\top Z}] = \log \sum_{i=1}^{d} p_i e^{\theta_i}$$

because $\theta^\top e_i = \theta_i$ for $i = 1, \ldots, d$.

To any (extended real-valued) convex function there is a *dual*[13] convex function on $\mathbb{R}^d$.

Example 26.1. Let $d = 1$ and consider $f(x) = x^2/2$. Consider a line ax of slope ax and look at the height that separates the line from the function. Find the point at which this height is the greatest to calculate the dual $\widehat{f}(a) := \sup_{x \in \mathbb{R}} ax - f(x)$.



Here, we can calculate $\widehat{f}(a) = a^2/2$. In a related sense to how the Gaussian is self-dual for the Fourier transform, this function is self-dual for the Frenchel-Legendre transform.

[13] This is somtimes called Fenchel duality, Legendre duality, or Fenchel-Legendre duality.

Example 26.2. Let $f(x) = e^x$. To find $\widehat{f}(a)$, since $f'(x) = a$ for x, if $a > 0$, this occurs if $x = \ln a$, and if $a \leq$, this is impossible. So we get

$$\begin{aligned}\widehat{f}(a) &= \sup_x (ax - e^x) \\ &= \begin{cases} a \ln a - a & a > 0 \\ 0 & a = 0 \\ \infty & a < 0. \end{cases}\end{aligned}$$

What if $d > 1$?

Definition 26.3. Suppose $\Phi : \mathbb{R}^d \mapsto \mathbb{R} \cup \{\infty\}$ is convex. Its **Fenchel-Lengendre dual** is

$$\widehat{\Phi}(a) := \sup_{x \in \mathbb{R}^d} a^\top x - \Phi(x)$$

for $a \in \mathbb{R}^d$.

Again,

$$\widehat{\Phi}(a) = a^\top x_a - \Phi(x_a),$$

where x_a is defined by $\nabla \Phi(x_a) = a$ (if x_a exists). It can be shown that

$$\Phi(x) = \sup_a x^\top a - \widehat{\Phi}(a).$$

To check this where Φ expresses all derivatives, write

$$\Phi(x) \geq x^\top a - \widehat{\Phi}(a) \quad \forall x, a \iff \widehat{\Phi}(a) \geq a^\top x - \Phi(x) \quad \forall x, a.$$

Proposition 26.1. *Let X take values in $\mathscr{X}$ with $|\mathscr{X}| = d$ and $p_i = \mathbb{P}(X = i)$. Let $Z = e_i$ iff $X = i$ (i.e. $P(Z = e_i) = p_i$ for $1 \leq i \leq d$). Then the Fenchel dual of $\Phi(\theta) = \ln \mathbb{E}[e^{\theta^\top Z}]$ is*

$$\widehat{\Phi}(a) = \begin{cases} D(a \,||\, p) & \text{if } a \text{ is a probability distribution} \\ \infty & \text{otherwise.} \end{cases}$$

Proof. Here,

$$\Phi_Z(\theta) = \ln \sum_{i=1}^d p_i e^{\theta_i},$$

so

$$\nabla \Phi_Z(\theta) = \begin{bmatrix} \frac{p_1 e^{\theta_1}}{\sum_{i=1}^d p_i e^{\theta_i}} \\ \vdots \end{bmatrix}.$$

This expresses only gradients that are probability distributions (means where $p_i \neq 0$). We have

$$\widehat{\Phi}_X(a) = a^\top p_a - \ln \sum_{i=1}^{d} p_i e^{\theta_{ai}},$$

where θ_a is defined in terms of a via $\nabla\Phi(\theta_a) = a$, i.e. $p_i e^{\theta_i}$ is proportional to a_i (i.e. $\theta_i = \ln \frac{a_i}{p_i} + \text{constant}$). The constant is $\log \sum_{i=1}^{d} p_i e^{(\theta_a)_i} = 0$.

$$\begin{aligned} &= \sum_{i=1}^{d} a_i \ln \frac{a_i}{p_i} - \cancelto{0}{\ln\left(\sum_{i=1}^{d} p_i e^{\ln \frac{a_i}{p_i}}\right)} \\ &= D(a \,||\, p). \end{aligned}$$

□

26.2 Large deviations and Sanov's theorem

Roughly speaking, a basic large deviations theory result is of the form: If $Z_1, Z_2, \ldots$ are iid $\mathbb{R}^d$-valued with $\log \mathbb{E}[e^{\theta^\top Z}]$ denoted $\Phi_Z(\theta)$ and $\mathbb{E}[Z_1] = 0 \in \mathbb{R}^d$, then for any open set $A \subseteq \mathbb{R}^d$,

$$\liminf_{n\to\infty} -\frac{1}{n} \log \mathbb{P}\left(\frac{Z_1 + \cdots + Z_n}{n} \in A\right) \leq \inf_{z\in A} \widehat{\Phi}_Z(z).$$

Here is a special case.

If $X_1, X_2, \ldots,$ are i.i.d. $\mathscr{X}$-valued with $\mathscr{X} = \{1, 2, \ldots, d\}$ and $Z_1, Z_2, \ldots$ are i.i.d. $\mathbb{R}^d$-valued creased from $X_1, X_2, \ldots$, then observe that $\frac{Z_1+\cdots+Z_n}{n}$ is equivalent to the empirical distribution of $(X_1, \ldots, X_n)$, i.e. $\frac{Z_1+\cdots+Z_n}{n} = \sum_{i=1}^{d} \frac{N(i|x^n)}{n} e_i$. Let $P_{x^n} := (\frac{N(i|x^n)}{n}, i = 1, \ldots, d)$. So for any open subset $A \subseteq$ simplex in $\mathbb{R}^d$,

$$\liminf_{n} -\frac{1}{n} \log \mathbb{P}(P_{X^n} \in A) \leq \inf_{a\in A} D(a \,||\, p).$$

Recall that if $x^n = (x_1, \ldots, x_n) \in \mathscr{X}^n$ with $\mathscr{X}$ finite and if $\mathcal{P}$ denotes the set of probability distributions on X, then $p_{x^n} \in \mathcal{P}$ denotes $(\frac{N(x|x^n)}{n}, x \in \mathscr{X})$ and $\mathcal{P}_n$ denotes the set of all such P_{x_n}. For an n**-type** $P \in \mathcal{P}_n$, the **typicality set for** P refers to $T(P) := \{x^n \in \mathscr{X}^n : P_{x^n} = P\}$. For $Q \in \mathcal{P}$,

$$\begin{aligned} Q(x^n) &= \prod_{i=1}^{n} q(x_i) \\ &= \prod_{x\in X} q(x)^{N(x|x^n)} \\ &= 2^{-n(H(P_{x^n}) + D(P_{x^n}||Q))}. \end{aligned}$$

We also proved that for $P \in \mathcal{P}_n$,

$$P^n(T(P)) \geq P^n(T(\widetilde{P})) \qquad \forall \widetilde{P} \in \mathcal{P}_n,$$

$|P_n| \leq (n+1)^{|\mathscr{X}|}$, and for $P \in \mathcal{P}_n$,

$$\frac{1}{(n+1)^{|\mathscr{X}|}} 2^{nH(P)} \leq |T(P)| \leq 2^{nH(P)}.$$

Theorem 26.1 (Sanov)**.** *Let $\mathscr{X}$ be finite, $X_1, X_2, \ldots \overset{\text{iid}}{\sim} Q$, and $E \subseteq \mathcal{P}$. Assume that E is the closure of its interior. Then*

$$\lim_{n\to\infty} \frac{1}{n} \log Q^n(P_{X^n} \in E) = -D(P^* \,||\, Q),$$

where

$$P^* = \arg\min_{P \in E} D(P \,||\, Q).$$

Remark 26.1. Since E is closed and $D(\cdot \,||\, Q)$ is continuous, this argmin exists. P^* is called the I**-projection** of Q onto E.

Proof. For the upper bound,

$$\begin{aligned} Q^n(P_{X^n} \in E) &= Q^n(P_{X^n} \in E \cap \mathcal{P}_n) \\ &\leq (n+1)^{|\mathscr{X}|} 2^{-nD(P^*||Q)} \end{aligned}$$

For the lower bound, for any $P^{(n)} \in \mathcal{P}_n \cap E$,

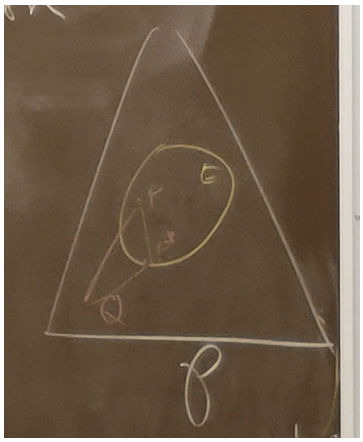
$$\begin{aligned} Q^n(P_{X^n} \in E) &\geq Q^n(T(P^{(n)})) \\ &\geq \frac{1}{(n+1)^{|\mathscr{X}|}} 2^{-nD(P^{(n)}||Q)}. \end{aligned}$$

Choose $P^{(n)} \to P^*$. □

Here is a nice observation about the I-projection of Q onto a *convex* set E.

Proposition 26.2. *For all $P \in E$,*

$$D(P \,||\, Q) \geq D(P \,||\, P^*) + D(P^* \,||\, Q).$$



This tells us that we should think of $D(P \,||\, Q)$ as the *square* of a distance.

Proof. Consider the relative entropy $D(\lambda P + (1-\lambda)P^* \,||\, Q)$ for $\lambda \in [0,1]$. Differentiate in λ. It must be nonnegative. $\square$

27 I-Projection in Sanov's Theorem and Hypothesis Testing

27.1 Properties of I-projection in Sanov's theorem

Last time, we proved Sanov's theorem:

Theorem 27.1 (Sanov)**.** *Let $X_1, X_2, \ldots \overset{\text{iid}}{\sim} Q$ be $\mathscr{X}$-valued random variables, and let P_{x^n} be the type of x^n: $P_{x^n}(x) = \frac{N(x|x^N)}{n}$. Let $\mathcal{P}$ be the set of probability distributions on $\mathcal{X}$, and assume that $E \subseteq \mathcal{P}$ is the closure of its interior. Then*

$$\lim_{n\to\infty} \frac{1}{n} \log Q^n(P_{X^n} \in E) = -D(P^* \,||\, Q),$$

where

$$P^* = \arg\min_{P\in E} D(P \,||\, Q).$$

P^* is called the I**-projection** of Q onto E.

Definition 27.1. Let $\mathscr{X}$ be finite. Given $Q \in \mathcal{P}$ and $h : \mathscr{X} \to \mathbb{R}$, the probability distribution of the form

$$\frac{Q(x)e^{\lambda h(x)}}{\sum_{a\in\mathscr{X}} Q(a)e^{\lambda h(a)}}$$

is called an **exponential transform** of Q.

Proposition 27.1. *Suppose E is defined as*

$$E = \left\{ P : \sum_x g_j(x)P(x) \geq \alpha_j, j = 1, \ldots, k \right\}.$$

Then P^* will be an exponential transform of Q.

Proof. Assume $Q(x) > 0$ for all x. We want

$$\max \sum_x P(x) \log \frac{P(x)}{Q(x)},$$

subject to

$$\begin{cases} \sum_a P(a)g_j(a) \geq \alpha_j, & j = 1, \ldots, k \\ P(x) \geq 0 & x \in \mathscr{X} \\ \sum_x P(x) = 1. \end{cases}$$

where the variables are $(P(x), x \in \mathscr{X})$ and $Q \in \mathcal{P}$ is fixed. The correct Lagrangian is

$$\sum_x P(x) \log \frac{P(x)}{Q(x)} + \sum_{j=1}^k \lambda_j \left(\sum_x P(x)g_j(x) - \alpha_j \right) - \sum_x \mu_x P(x) + \nu \left(\sum_x P(x) - 1 \right).$$

Write the KKT conditions for this:

$$\lambda_{*j} \geq 0,$$
$$\mu_x^* \geq 0,$$
$$\lambda_j^* \left(\alpha_j - \sum_x P^*(x) g_j(x) \right) = 0 \qquad \forall j,$$
$$\mu_x^* P^*(x) = 0 \qquad \forall x.$$

Differentiate this to get

$$\log \frac{P^*(x)}{Q(x)} + 1 + \sum_j \lambda_j g_j(x) - \mu_x^* + \nu^* = 0 \qquad \forall x.$$

Since $P^*(x)$ cannot be 0 for any x, we must have $\mu_x^* = 0$. □

We also can show the following.

Theorem 27.2.

$$\lim_{n\to\infty} Q^n(X_1 = a \mid P_{X^n} \in E) = P^*(a) \qquad \forall a \in \mathscr{X}.$$

Proof. Given $\delta > 0$, let $A = \{P \in E : D(P \mid\mid Q) \leq D(P^* \mid\mid Q) + 2\delta\}$. The Sanov theorem calculation tells us that

$$Q^n(E \setminus A) \leq (n+1)^{|\mathscr{X}|} 2^{-n(D^*(P||Q)+2\delta)}.$$

For large enough n,

$$Q^n(A) \geq \frac{1}{(n+1)^{|\mathscr{X}|}} 2^{-n(D^*(P||Q)+\delta)}.$$

This proves that

$$Q^n(P_{X^n} \in A \mid P_{X^n} \in E) \xrightarrow{n\to\infty} 1.$$

If E is convex, we can use $D(P \mid\mid P^*) + D(P^* \mid\mid Q) \leq D(P \mid\mid Q)$ for all $P \in E$ to show that

$$Q^n(D(P_{X^n} \mid\mid P^*) \leq 2\delta \mid P_{X^n} \in E) \xrightarrow{n\to\infty} 1.$$

Then use Pinsker's inequality:

$$D(P_1 \mid\mid P_2) \geq \frac{1}{2\ln 2} \|P_1 - P_2\|_1^2 \qquad \forall P_1, P_2.$$ □

27.2 The Neyman-Pearson framework of hypothesis testing

Here is the Neyman-Pearson formulation of hypothesis testing with two hypotheses H_1 and H_2. Under H_1, assume that $X_1, X_2, \ldots,$ are iid $\mathscr{X}$-valued with $X_i \sim P_1$. Under H_2, assume that $X_1, X_2, \ldots,$ are iid $\mathscr{X}$-valued with $X_i \sim P_2$. Given a "threshold" T, define

$$A_n(T) = \left\{ x^n : \frac{P_1^n(x^n)}{P_2^n(x^n)} > T \right\}.$$

Definition 27.2. A **hypothesis test** is a function $\mathscr{X}^n \to \{1, 2\}$.

Equivalently, it means we choose a set $B \subseteq \mathscr{X}^n$ on which to decide H_1, and on B^c we decide H_2.

Let $\mathbb{1}_B$ denote the indicator function of B. Observe that

$$(\mathbb{1}_{A_n(T)}(x^n) - \mathbb{1}_B(x^n))(P_1^n(x^n) - TP_2^n(x^n)) \geq 0 \qquad \forall x^n.$$

Summing this up over x^n,

$$\underbrace{\sum_{x^n \in A_n(T)} P_1(x^n)}_{\underbrace{1 - \mathbb{P}_1^n(X^n \notin A_n(T))}_{\alpha^*}} - T \underbrace{\sum_{x^n \in A_n(T)} P_2^n(x^n)}_{\beta^*} - \underbrace{\sum_{x^n \in B} P_1^n(x^n)}_{1-\alpha} + T \underbrace{\sum_{x^n \in B} P_2^n(x^n)}_{\beta} \geq 0.$$

We get

$$T(\beta - \beta^*) - (\alpha^* - \alpha) \geq 0,$$

so if $\alpha \leq \alpha^*$, then $\beta \geq \beta^*$. Hence, if one tries to minimize $\mathbb{P}(\text{error} \mid H_2)$ given a bound on $\mathbb{P}(\text{error} \mid H_1)$, then we use a threshold test.

Theorem 27.3 (Stein's lemma)**.** *For any $\varepsilon > 0$, let*

$$\beta_n^\varepsilon := \min_{B \subseteq \mathscr{X}^n} \{\beta_n : \alpha_n \leq \varepsilon\}.$$

Then

$$\lim_{n \to \infty} \frac{1}{n} \log \beta_n^\varepsilon = -D(P_1 \mid\mid P_2).$$

The intuition is that for all $\delta > 0$, the ball $C_n = \{P \in \mathcal{P} : D(P \mid\mid P_1) \leq \delta\}$ has $P_1^n(C_n) \to 1$ as $n \to \infty$ and

$$\liminf_{n \to \infty} -\frac{1}{n} \log P_2^n(C_n) \geq D(P_1 \mid\mid P_2) - \eta,$$

where $\eta \to 0$ as $\delta \to 0$.

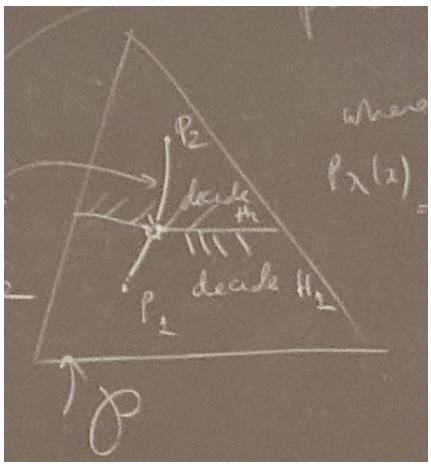
27.3 The Bayesian framework of hypothesis testing

In the Bayesian view of hypothesis testing, π_1 is the prior probability of H_1, and π_2 is the prior of H_2. The optimal test is to decide H_1 if

$$\frac{\pi_1 P_1^n(x^n)}{\pi_2 P_2^n(x^n)} \geq 1.$$

This is related to information geometry, which is about the space of probability distributions with separation defined by relative entropy. If $P_1, P_2 \in \mathcal{P}$, then there is a statistically natural path connecting them, parameterized by $\lambda \in [0,1]$, where

$$P_\lambda(x) = \frac{P_1^\lambda(x) P_2^{1-\lambda}(x)}{\sum_a P_1^\lambda(a) P_2^{1-\lambda}(a)}.$$



P_λ arises by studying the minimum of $D(P \,||\, P_2)$ subject to $D(P \,||\, P_2) - D(P \,||\, P_1) = K$. Why this constraint? This is because

$$\left\{ x^n : \frac{P_1(x^n)}{P_2(x^n)} \geq T \right\} = \left\{ x^n : D(P_{x^n} \,||\, P_2) - D(P_{x^n} \,||\, P_1) \geq \frac{1}{n} \log T \right\}.$$

Theorem 27.4. *Assume that $\pi_1 > 0$ and $\pi_2 > 0$. Let $\alpha_n^* = P_1^n(A_n(\frac{\pi_2}{\pi_1})^c)$, and let $\beta_n^* = P_2^n(A_n(\frac{\pi_2}{\pi_1}))$. Then*

$$\lim_{n\to\infty} \frac{1}{n} \log(\pi_1 \alpha_n^* + \pi_2 \beta_n^*) \to -D(P_{\lambda^*} \,||\, P_2),$$

where $D(P_{\lambda^*} \,||\, P_2) = D(P_{\lambda^*} \,||\, P_1)$.