Information Theory

Alec Zabel-Mena

February 27, 2022

Contents

Chapter 1

Preliminaries.

1.1 Probability Theory.

Definition. We define a **probability space** to be a triple $(S, \mathcal{B}, P)$ where S is a nonempty set called the **sample space**, $\mathcal{B}$ is a Borel set of subsets of S, and P is a nonnegative function on $\mathcal{B}$ with the properties

(1) $P(S) = 1$.

(2) $P(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(A_n)$.

where each $A_m \cap A_n =$ whenever $m \neq n$. We call P the **probability measure**. We call the value of $P(A)$ the **probaility** of A.

Example 1.1. If $S = \{s_1, s_2, \dots\}$ is finite, or countable, and $\mathcal{B} = 2^S$, and $\{p_n\}$ is a sequence of nonegative numbers with $\sum p_n = 1$, then the function defined by $P(A) = \sum \{p_n : s_n \in A\}$ is a probability measure, and makes $(S, \mathcal{B}, P)$ a prpbability space.

Definition. Let S be the sample space of some probability space with measure P. We define a **random variable** to be a mapping $X : S \to R$, where we call R the **range** of X. If R is a subset of an n-dimensional Euclidean space, then we call X a **random vector**, whose components are random variables with 1-dimensional range.

Definition. Let $(S, \mathcal{B}, P)$ be a probability space. We define two random variables X and Y to be **equal almost everywhere** if the probability of the set $\{s : X(s) \neq Y(s)\} = 0$, i.e. $P(X) \neq P(Y)$. We write $X = Y$.

Definition. We define the **expectation**, or **average** of a random variable X, with range $\mathbb{R}$ to be the function $E(X)$, defined by:

$$E(X) = \int_S X(s) \ \mathrm{d}P \tag{1.1}$$

where the integral take is the Lebesgue integral.

Definition. Let S be a sample space with probability measure P. We define the **distribution** of a random variable X with range R to be the probability measure P_X induced by P on R such that $P_X(A) = \{s : X(s) \in A\}$, for any Borel set of $\mathbb{R}$. We define the function:

$$F_X(x) = P(\{s : X(s \leq x)\}) \tag{1.2}$$

to be the **distribution** function of X.

We can see that the expectation of X via the above definitions is given by:

$$E(X) = \int_R x \ \mathrm{d}P_x = \int_{-\infty}^{\infty} x \ \mathrm{d}F_X(x)$$

Where the integral taken with respect to P_X is Lebeusue, and the integral taken with respect to $F_X(x)$ is Riemann-Stieltjes.

In the case where S is discrete, and hence so is R, for any random vairable X, we can then define the expectations of X to be the sums:

$$E(X) = \sum_x p(x)x \tag{1.3}$$

$$E(f(X)) = \sum_x p(x)f(x) \tag{1.4}$$

where $f : S \to \mathbb{R}$ and $f(X)$ is a random variable. It may be that f is not well behaved, or undefined for certain values of x.

Definition. Let X be a random variable. We say that X has **density** if the distribution function of X has the form:

$$F_X(x) = \int_{-\infty}^{x} up(u) \ \mathrm{d}u \tag{1.5}$$

for some nonegative function p. We then define the expectations of X to be

$$E(X) = \int_{-\infty}^{\infty} up(u) \ \mathrm{d}u \tag{1.6}$$

$$E(f(X)) = \int_{-\infty}^{\infty} f(u)p(u) \ \mathrm{d}u \tag{1.7}$$

$$\tag{1.8}$$

Definition. Let $(S, \mathcal{B}, P)$ be a probability space, and let $A_1, \ldots, A_n \in \mathcal{B}$. We say that $A_1, \ldots, a_n$ are **independent** if for every subset of them, $A_{i_1}, \ldots, A_{i_m}$,

$$P(A_{i_1} \cap \ldots P(A_{i_m})) = P(A_{i_1}) \ldots P(A_{i_m}) \tag{1.9}$$

Moreover, a collection $X_1, \ldots X_n$ of random variables are indepenedent if the evets $A_i = \{s : X_i(s) \in S_i\}$ are indepenedent for any $S_1, \ldots S_n \in \mathcal{B}$.

Definition. If X and Y are real random variables defined on a sample space S, then the mapping $s \to (X(s), Y(s))$ induces a probability measure P_{XY} on the field of 2-dimensional Borel sets, called the **joint distribution** which is defined by the function

$$F_{XY}(x, y) = P(\{x : X(s) \leq x, Y(s) \leq y\}) \tag{1.10}$$

Lemma 1.1.1. *The random variables X and Y are independent if, and only if $F_{XY} = F_X(x)F_Y(y)$. With F_X and F_Y the distributions of X and Y, respectively.*

Theorem 1.1.2 (The Weak Law of Large Numbers.)**.** *Let $X_1, \dots X_n$ be independent random variables, each with finite expectation μ, and with the same distribution. Then for each $\epsilon > 0$,*

$$\lim_{n\to\infty} P(|\frac{\sum X_i}{n} - \mu| \geq \epsilon) = 0 \tag{1.11}$$

1.2 Convexity and Jensen's Inequality.

Definition. A subset $K \subseteq E^n$, of a Euclidean space E is said to be **convex** if for any two points in K, the line segment adjoining them is also in K. That is for any $x, y \in K$, the line $l(x, y) = tx + (1 - t)y \in K$, for any $t \in [0, 1]$.

Definition. Let $K \subseteq E^n$ be a subset of a Euclidean space E. We call a point $x \in K$ a **convex combination** of the points $x_1, \dots, x_n \in K$ if there exist scalars $a_1, \dots, a_n$ with $\sum a_i = 1$ such that $\sum a_i x_i = x$. We call the set of all convex combinations of K a **convex hull**.

Lemma 1.2.1. *A set K is convex if, and only if every comvex combination of points of K is also in K.*

Proof. Suppose K is convex. Then for any $x, y \in K$, let $l(x, y) = xt + (1 - t)y$ be the line adjoining x and y. Then $l(x, y) \in K$ is a point of K which is a convex combination of x and y.

Now suppose that for any points $x_1, \dots, x_n \in K$, that their convex combination is also in K, suppose the same is true for the points $y_1, \dots y_n$; not necessarily the same convex combination as those for each x_i. Let $x = \sum a_i x_i$ and $y = \sum b_i y_i$ for scalars a_i, b_i where $1 \leq i \leq n$. Now consider the line $l(x, y) = tx + (1 - t)y$ for some $t \in [0, 1]$. Then $l(x, y) = \sum ta_i x_i + \sum (1 - t)b_i y_i = \sum ta_i x_i + (1 - t)b_i y_i$. So $l(x, y)$ is a sum of convex combinations of the points $a_i x_i, b_i y_i \in K$ for each i; by hypothesis, this makes $l(x, y) \in K$. ■

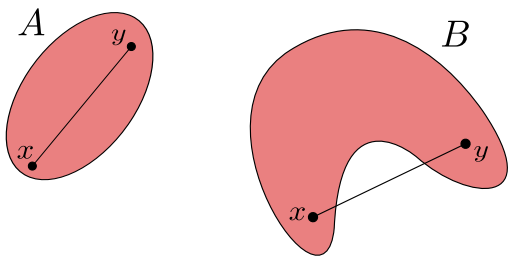

Figure 1.1: A convex set, A and a non-convex set B.

Definition. Let f be a real-valued function, and let $K \subseteq \operatorname{dom} f$ be convex. We call f **convex up** if, for every $x, y \in K$ and $t \in [0, 1]$,

$$f(tx + (1 - t)y) \leq tf(x) + (1 - t)f(y) \tag{1.12}$$

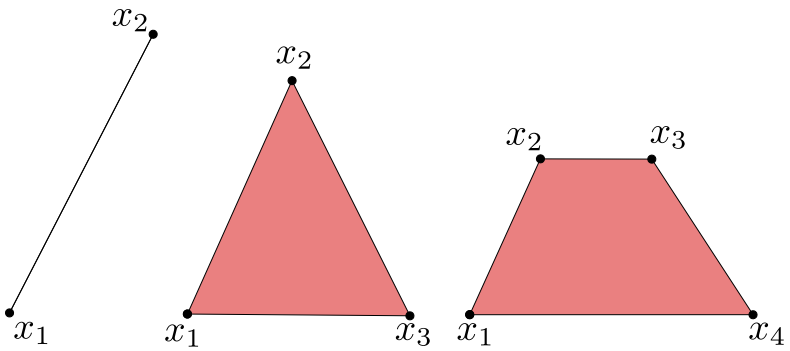

Figure 1.2: 2-dimensional convex hulls.

We call f **convex down** if, for every $x, y \in K$ and $t \in [0,1]$,

$$f(tx + (1-t)y) \geq tf(x) + (1-t)f(y) \tag{1.13}$$

If either equality is strict, i.e. $f(tx + (1-t)y) \neq tf(x) + (1-t)f(y)$ for some x and y, then we call f **strictly convex** (up or down).

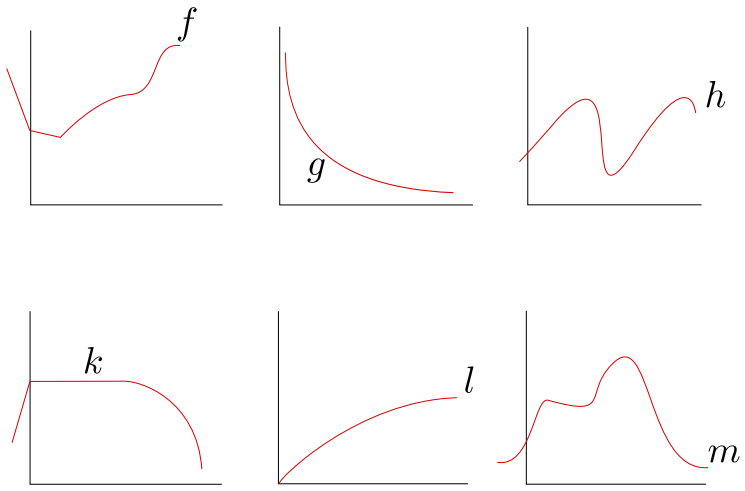

Figure 1.3: The following are real-valued functions. The function f is convex up, the function h is strictly convex up, while g is not convex. Similarly, k is convex down, l is strictly convex down, while m is not convex.

Lemma 1.2.2. *If f is a real-valued function, and $K \subseteq \operatorname{dom} f$ is a convex open set, then if f is convex (up or down), then f is continuous.*

Example 1.2. The above lemma does not hold for closed sets. If $K = [0,1]$, and $f(x) = x$ for all $x \in (0,1]$ and $f(0) = 1$, then f is discontinuous despite being convex up.

Lemma 1.2.3. *If f is a twice differentiable real-valued function with first derivative $f'(x)$ for every $x \in K$, then f is convex up if, and only if f' is nondecreasing. Similarly, f is convex down if, and only if f' is nonincreasing.*

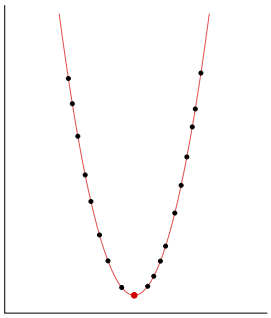

Figure 1.4: Jensen's Inequality says that if a mass distribution is placed on the graph of the given function, then the center of mass lies above, or on the graph.

Theorem 1.2.4 (Jensen's Inequality)**.** *Let E be a Euclidean space, and let K be an interval in E^1. Let $F(x)$ be a probaility distribution concentrated on K, and X be the associated random variable with the prpbality $P(X \leq x) = F(x)$. Then if the expectation $E(X)$ exists, and if f is a function that is convex up, then:*

$$E(f(X)) \geq f(E(X)) \tag{1.14}$$

If f is convex down, then:

$$E(f(X)) \leq f(E(X)) \tag{1.15}$$

If f is strictly convex, then (1.14) *and* (1.15) *are strict, except possibly when $P(X = x_0) = 1$ for some $x_0 \in K$.*

We now finish the section by introdicing examples of Jensen's inequality with probability distributions. There are two accompying theorems, and one is the continuous analog of the other.

Example 1.3. Let $S = \{s_1, s_2, \dots\} \subseteq \mathbb{R}$ be a discrete sample space of real numbers. Let $p(s_i)$ be a nonnegative function such that $\sum p(s_i) = 1$. Let $1(s_i)$ be any other nonnegative function defined on S, and define the random variable X by:

$$X(s) = \frac{q(s)}{p(s)}$$

for any $s \in S$. THen by Jensen's inequality we have $E(\log(X)) \leq \log E(X)$, which gives the following inequality:

$$\sum p(s_i) \log \frac{1}{p(s_i)} \leq \sum p(s) \log \frac{1}{q(s)} + \log \alpha$$

where $\alpha = \sum q(s_i)$.

Furthermore, $\log x$ is strictly convex down, so equality holds if, and only if $X = \beta$ for some constant β. Since $\sum p(s_i) = 1$, we get $\beta = \alpha$, and so equality holds if, and only if $q(s_i) = (s_i)$ for all i such that $p(s_i) \neq 0$.

Theorem 1.2.5. *Let $\{p_i\}_{i\in\mathbb{Z}}$ be a sequence of positive real numbers such that $\sum p_i = 1$. If $\{q_i\}$ is any other sequence of nonegative reals with $\sum q_i = \alpha$, then:*

$$\sum p_i \log p_i^{-1} \leq \sum p_i \log q_i^{-1} + \log \alpha \tag{1.16}$$

with equality hilding if, and only if $q_i = \alpha p_i$ for all i.

Example 1.4. Take $S = \mathbb{R}$, and let $p(x)$ be a nonnegative density function such that $\int_{-\infty}^{\infty} p(x)\ \mathrm{d}x = 1$. Then p induces a probability measure on S. Now, let $q(x)$ be any other nonnegative function defined on S and define the random variable X as was in example 1.3. Then, again by Jensen's inequalty, we get:

$$\int_{-\infty}^{\infty} p(x) \log \frac{q(x)}{p(x)}\ \mathrm{d}x \leq$$

$\log \int_I q(x)\ \mathrm{d}x$

where I is a measurable subset of $\mathbb{R}$.

Theorem 1.2.6. *Let I be a measurable subset of $\mathbb{R}$ and let $p(x)$ be a positive function defined on I, with $\int_I p(x)\ \mathrm{d}x = 1$. If $q(x)$ is a nonnegative positive function defined on I with $\int q(x)\ \mathrm{d}x = \alpha$, then:*

$$\int_I p(x) \log p(x)^{-1}\ \mathrm{d}x \leq \int_I p(x) \log q(x)^{-1} + \log \alpha \tag{1.17}$$

with equality holding if, and only if $q(x) = \alpha p(x)$.

Chapter 2

Entropy, and Mutual Information.

2.1 Discrete Random Variables.

Definition. Let X be a discrete random variable with finite, or countable range R, $R = \{x_1, x_2, \dots\}$. Let $p_i = P(X = x_i)$. We define the **entropy** base b of X to be:

$$H_b(X) = \sum_{i \geq 1} p_i \log_b p_i^{-1} \tag{2.1}$$

If $b = 2$, we write $\log_2 = \log$ and if $b = e$ we write $\log_e = \ln$. Additionally, we define $p_i \log_b p_i^{-1} = 0$ whenever $p_i = 0$. We define $H(X) = \infty$ in the case that R is infinite and H diverges.

Example 2.1. (1) If X is the outcome of the roll of a fair 6-sided die, then $R = \{1, 2, 3, 4, 5, 6\}$ and $p_i = \frac{1}{6}$. Thus, $H_b(X) = \log_b 6$.

(2) Let $R = \mathbb{Z}/2\mathbb{Z}$, and define X by the prbabilities $P(X = 0) = p$ and $P(X = 1) = 1 - p$. Then $H_2(X) = -p \log p - (1-p) \log (1-p)$. We call H_2 the **binary entropy function** and denote it simply as H. The majority of the work done in these notes will concern this function in particular.

(3) If $p = (p_1, \dots, p_r)$ is any probability vector (where $p_i \geq 0$ and $\sum p_i = 1$), we define the **entropy function** base b of p to be $H(p) = H(p_1, \dots, p_r) = \sum p_i \log p_i^{-1}$.

(4) If $A = \sum_{n=2}^{\infty} n \log^2 n^{-1}$, and if X is a random variable defined by $P(X = n) = An \log^2 n^{-1}$, for $n \geq 2$, then $H(X) = \infty$.

Definition. Let X be a discrete random variable with range R. We define the **ammount of information** in base b, provided by the event $X = x$ to be the function $I_b(x) = -\log_b P(X = x)$, where $x \in R$. When $b = 2$, we write $I_b = I$.

Lemma 2.1.1. *For any discrete random variable X, $H_b(X) = E(I_b(X))$.*

Proof. This follows from the definitions of $I_b(X)$, $H_b(X)$, and the expectation $E(I_b(X))$. ■

Example 2.2. Define X by the probabilities $P(X = 0) = P(X = 1) = \frac{1}{2}$. Then $I(0) = I(1) = H(X) = \log 2$.

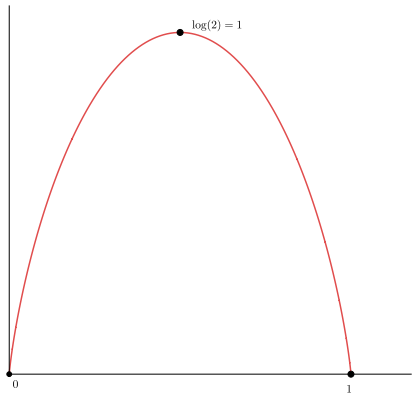

Figure 2.1: The graph of the binary entropy function $H_2(X) = H(X)$. Notice that $H(X) = 0$ at 0 and 1.

Theorem 2.1.2. *Let X be a discrete random variable with range $R = \{x_1, \ldots, x_r\}$. Then:*

$$0 \leq H_b(X) \leq \log r \tag{2.2}$$

Furthermore, $H_b(X) = 0$ if, and only if $p_i = 1$ for some i, and $H_b(X) = \log r$ if, and only if $p_i = \frac{1}{r}$ for all i.

Proof. Since each $0 \leq p_i \leq 1$, $p_i^{-1} \geq 0$, thus $\log p_i^{-1} \geq 0$. This implies that $H_b(X) \geq 0$. Now, $p \log p^{-1} = 0$ if, and only if $p = 0$ or $p = 1$. Thus, $H_b(X) = 0$ if and only if $p = 0$ or $p = 1$, that is, if each $p_i = 0$ or each $p_i = 1$.

Notice now, that the function $\log x$ is convex down. Thus, by theorem 1.2.5, we get $H(X) = \sum p_i \log p_i^{-1} \leq \sum p_i \frac{1}{p_i} = \log r$, so $H_b(X) = \log r$ if, and only if p_i is constant for each i, i.e. $p_i = \frac{1}{r}$ for all i. ■

Corollary. *If $p = (p_1, \ldots, p_r)$ is a probability vector, then $H_b(P)$ attains its maximum value at uniquely $p = (\frac{1}{r}, \ldots, \frac{1}{r})$.*

Definition. Let X and Y be random variables. We define the **conditional entropy** of X given Y to be the entropy $H(X|Y)$ defined by:

$$H(X|Y) = E(\log p(x|y)^{-1}) \tag{2.3}$$

Where $P(X|Y) = P(X = x, Y = y)$. We again define $0 \log 0 = 0$ and $H(X|Y) = \infty$ whenever the expectation diverges.

Remark. The notion of conditional entropy can also be motivated by a model of a communications channel called the **discrete memoryless channel** (DMC). DMC accept a r input symbols and give s output symbols. The channel is memoryless as the current output only depends on the current input and not any previous inputs. DMCs can be visualized in the following figures 2.2 and 2.3. We can describe the behaviour of a DCM with an $r \times s$ matrix of **transition proabilities**, $(p(x|y))$, where the ij-th entry is the probability $p(i|j)$.

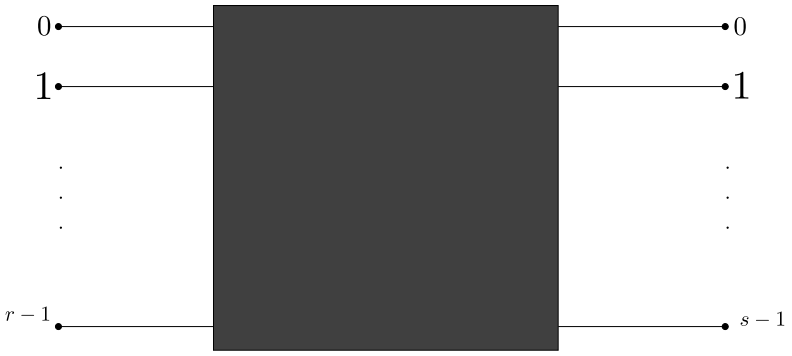

Figure 2.2: A discrete memoryless channel viewed as a black box.

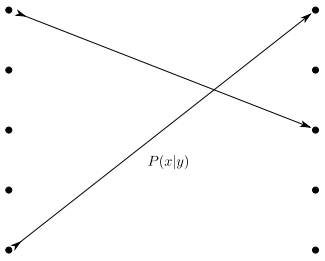

Figure 2.3: A discrete memoryless channel viewd in terms of the current inputs and outputs.

Example 2.3. (1) We define a **binary symmetric source** to be an object which transmits the bits 0 and 1 at a given rate R. We define a **binary symmetric channel** (BSC) to be an object capable of transimitting bits generated by a source one bit per unit time. It is entirely possible that the BSC is not reliable with its trasnmission, and can have a probaility $0 \leq p \leq \frac{1}{2}$ called the **raw bit error probability** which is the chance a transmitted bit is not the same as was generated by the source.

We can see that the BSC is a DMC. Now, let $r = s = 2$ for the BSC. Then the input-output graph of the BSC, with its error proabilitiy is as described in figure 2.4

(2) If we take a DMC with $r = 2$ and $s = 3$, then we have inputs 0 and 1, and outputs 0, 1, and 2, where we take 2 to be an **erasure** of the input bit. We call this cahnnel a **binary erasure channel**. Such channels may arise if the inputs into such a (physical)

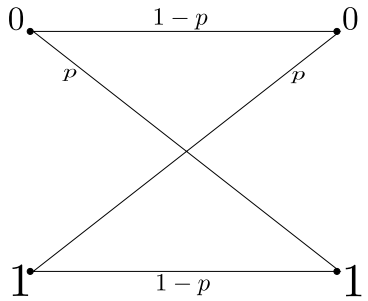

Figure 2.4: Input-Output graph of the BSC.

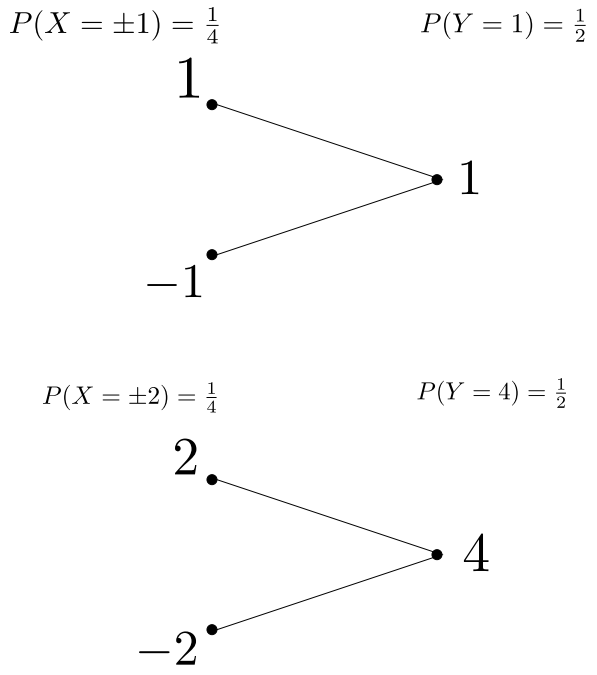

Figure 2.5:

channel were two sqare waves signifying voltage. Such waves are rarely perfect, and are often recieved in analog form with noise. Thus it is entriely possible for a bit going through this channel is erased.

Lemma 2.1.3. *For any discrete memoryless channel, there is a pair of discrete random variables X and Y such that the channel takes $X \to Y$. Conversely, for any pair of discrete random variables, (X, Y), there is a discrete memoryless channel taking $X \to Y$.*

Proof. Select the inputs, X, according to the probaility distribution $p(x)$ on $\mathbb{Z}/r\mathbb{Z}$. Now, define the random variable Y, then $p(x, y) = p(x)p(y|x)$, then $py) = \sum_x p(y|x)p(x)$. We also get $p(x|y) = \frac{p(x,y)}{p(y)}$ which g ves us the result. ■

Example 2.4. (1) Let X be defined by the probabilities $P(X = 1) = P(X = -1) = P(X = 2) = P(X = -2) = \frac{1}{4}$ and $Y = X^2$. We get the DMC found in figure **??** from the pair (X, X^2)

(2) Consider the Binary Erasure channeldescribed in example 2.3(2), where 2 is the erasure of a single bit. We induce the DMC to be the graph: $0 \to 0$, $0 \to 2$, $1 \to 2$, and $1 \to 1$ with the following probabilities: $\frac{3}{4}$, $\frac{1}{4}$, $\frac{1}{2}$, and $\frac{1}{2}$, respectively (see figure **??**). Let $P(X = 0) = \frac{2}{3}$ and $P(X = 1) = \frac{1}{3}$. Then we get $H(X) = \frac{2}{3}\log\frac{3}{2} + \frac{1}{3}\log 3$ bits. Additionally, $H(X|Y = 0) = 0$, $H(X|Y = 1) = 0$ and $H(X|Y = 2) = 1$. So if Y is the erasure of a bit, there must be uncertainty about X. Fortunately, on average we have certainty, so we can be confident the correct bit was transmitted.

Theorem 2.1.4. *Let X, Y, and Z be discrete random variables. Define for each z, $A(z) = \sum_{x,y} p(y)p(x|x, y)$. Then:*

$$H(X|Y) \leq H(Z) + E(\log A) \tag{2.4}$$

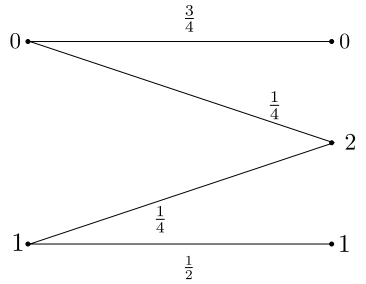

Figure 2.6:

Proof. By definition,

$$H(X|Y) = E(\log px|y^{-1}) = \sum_{x,y,z} p(x,y,z) \log p(x|y)^{-1} = \sum p(z) \sum \frac{p(x,y,z)}{p(z)} \log p(x|y)^{-1}$$

Now, fixing z we get $\frac{p(x,y,z)}{p(z)} = p(x,y|z)$. Now, by Jensen's inequality to the inner sum, we get

$$H(X|Y) \leq \sum p(z) \log (\frac{1}{p(z)} \sum \frac{p(x,y,z)}{p(x|y)}) = \sum p(z) \log p(z)^{-1} + \sum p(z) \log \sum \frac{p(x,y,z)}{p(x|y)}$$

Now, $\frac{p(x,y,z)}{p(x|y)} = \frac{p(x,y,z)p(y)}{p(x,y)} = p(y)p(z|x,y)$; which establishes the result. ■

Corollary (Fano's Inequality)**.** *If X and Y each take values in the sequence $\{x_i\}_{i=1}^r$, and $P_e = P(X \neq Y)$, then*

$$H(X|Y) \leq H(P_e) + P_e \log r - 1 \tag{2.5}$$

Proof. Take $Z = 0$ if $X = Y$ and $Z = 1$ if $X \neq Y$. Then $A(0) = 1$ and $A(1) = r - 1$. ■

Example 2.5. COnsider the DMC of example 2.4. Taking $r = 3$, $P(X = Y) = \frac{2}{3}$, and $P_e = \frac{1}{3}$, we get $H(X|Y) \leq H(\frac{1}{3}) + \frac{1}{3} \log 2 = \frac{1}{3} \log 3 + \frac{2}{3} \log \frac{3}{2} + \frac{1}{3} \log 2 = \log 3 - \frac{1}{3}$ bits.

Definition. Let X and Y be random variables. Let $H(X)$ be the entropy of X independent of Y, and $H(X|Y)$ the entropy of X geven Y. We define the **mutual information** between X and Y to be:

$$I(X,Y) = H(X) - H(X|Y) \tag{2.6}$$

We define the **mutual information** of the random variables X, Y, and Z to be:

$$I(X,Y,Z) = E(\log \frac{p(z|x,y)}{p(z)}) \tag{2.7}$$

Lemma 2.1.5. *Given random variables X and Y, we have*

$$I(X,Y) = \sum_{x,y} p(x,y) \log p(x|y)p(x)^{-1} = \sum p(x,y) \log \frac{p(x,y)}{p(x)p(y)} = \sum p(x,y) \log p(y|x)p(y)^{-1} \tag{2.8}$$

Proof. $I(X,Y) = H(X) - H(X|Y) = \sum p(x) \log p(x)^{-1} - \sum p(x,y) \log p(x,y)^{-1}$. ■

Definition. We define the **joint entropy** between the random variables X and Y to be:

$$H(X,Y) = \sum p(x,y) \log p(x,y)^{-1} \tag{2.9}$$

Lemma 2.1.6. *The following hold:*

(1) $I(X,Y) = I(Y,X)$.

(2) $I(X,Y) = H(Y) - H(Y|X)$.

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

Proof. By the above lemma, $I(X,Y) = \sum p(x,y) \log p(y|x)p(y)^{-1} = I(Y,X)$. Now, from this we get $I(X,Y) = I(Y,X) = H(Y) - H(Y|X)$, by definition.

Now notice that $H(X,Y) = H(X) + H(Y|X)$. So $I(X,Y) = I(Y,X) = H(X) + H(Y) - H(Y|X)$. ■

Lemma 2.1.7. *For any discrete random variables X, and Y, $I(X,Y) \geq 0$. Moreover, $I(X,Y) = 0$ if, and only if X and Y are independent of each other.*

Proof. That $I(X,Y) \geq 0$ follows from definition. Now, if $I(X,Y) = 0$, then $H(X)+H(Y) = H(X,Y)$ which implies independence. Conversely, we get that same result. ■

Theorem 2.1.8. *Given discrete random variables X, Y, and Z, $I(X,Y,Z) \geq I(Y,Z)$, with equality holding if, and only if $p(z|x,y) = p(z|y)$ for all triples (x,y,z) and $p(x,y,z) > 0$.*

Proof. $I(Y,Z) - I(X,Y,Z) = E(\log \frac{p(z|y)}{p(z)} - \log \frac{p(z|x,y)}{p(z)}) = E(\log \frac{p(z|y)}{p(z|x,y)}) = \sum p(x,y,z) \log \frac{p(x|y)}{p(z|x,y)}$. Now, by Jensen's inequality, $I(Y,Z) - I(X,Y,Z) \leq \log \sum p(x,y,z) \frac{p(x|y)}{p(z|x,y)} = \log 1 = 0$. ■

We can state the above results in terms of Markov chains.

Theorem 2.1.9. *If (X,Y,Z) is a Markov chain, then*

$$I(X,Z) = \begin{cases} I(X,Z) \\ I(Y,Z) \end{cases} \tag{2.10}$$

Example 2.6. (1) If X_1, X_2, X_3 are independent random variables, then the triple $(X_1, X_1 + X_1, X_1 + X_2 + X_3)$ forms a Markov chain. So $I(X_1, X_1 + X_2 + X_3) \leq I(X_1, X_1 + X_2)$.

(2) Define X by $P(X = 0) = P(X = 1) = \frac{1}{2}$, on two DMCs both with error proabilities p. Then $I(X,Y) = 1 - H(p)$ and $I(X,Z) = 1 - H(2p(1-p))$.

Theorem 2.1.10. *Given discrete random variables X and Y, $I(X,Y)$ is a convex down function of the probabilities $p(x)$.*

Proof. Fix the transformation probabilities $p(y|x)$, and let X_1 and X_2 be two random variables with probability distributions $p_1(x)$ and $p_2(x)$. Define the probability distribution of X to be the convex combination $p(x) = \alpha p_1(x)_\beta p_2(x)$.

Now, consider Y_1, Y_1, and Y such that $X_1 \to Y_1$, $X_2 \to Y_2$, and $X \to Y$ through some DMC. Then we must have:

$$\alpha I(X_1, Y_1) + \beta I(X_2, Y_2) - I(X, Y) = \sum p_1(x) \log \frac{p(y)}{p_1(y)} + \sum p_2(x) \log \frac{p(y)}{p_2(y)}$$

By Jensen's inequality, we get:

$$\sum p_1(x) \log \frac{p(y)}{p_1(y)} \leq \log \sum p_1(x, y) \frac{p(y)}{p_1(y)} = \log \sum \frac{p(y)}{p_1(y)} p_1(y) = 1$$

Hence, both sums are less than 0, which makes $I(X, Y)$ convex down. ■

Corollary. *The entropy of the proabilitie vector $p = (p_1, \ldots, p_r)$, $H(p)$ is convex down.*

Proof. Define X by the distribution $P(X = x_i) = p_i$. Then $I(X, X) = H(X) = H(p)$ where $p = (p_1, \ldots, p_r)$. ■

Theorem 2.1.11. *$I(X, Y)$ is convex up in the transition probabilities $p(y|x)$.*

Proof. The proof is analogous to that of theorem 2.1.10. ■

2.2 Discrete Random Vectors.

Definition. Let $X = (X_1, \ldots, X_n)$ be a discrete random vector, with probability $p(x) = P(X_1 = x_1 \ldots, X_n = x_n)$. We define the **entropy** of X in base b to be:

$$H_b(X) = \sum_x p(x) \log_b p(x)^{-1} \tag{2.11}$$

Remark. We can similarly extend the previous definitions to discrete random vectors.

Now, consider the random vector $U = (U_1, \ldots, U_k)$. We describe a communications model where U is taken to $X = (X_1, \ldots, X_k)$ via an encoder. X is then sent through a (possibly noisy) channel as the output Y, which is a (possibly) noisy version of X. We then take Y to V through a decoder and send V to the destination. Ideally, we want $U = V$, however, since Y is the reesult of sending X through a (possibly) noisy channel, it may not be the case. In fact, we see that the sequence $U \to X \to Y \to V$ forms a Markov chain.

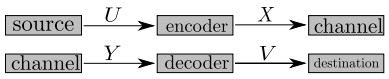

Figure 2.7: The Sequence $U \to X \to Y \to V$ seen as sending U to a recpient.

Now, this implies that $p(y|x) = p(v, y, x) = p(v|y)$. Notice also that $I(U, V) \leq I(X, V) \leq I(X, Y)$. We then come to a central theorem called the *data processing theorem.*

Theorem 2.2.1 (The Data Processing Theorem)**.** *Given a Markov chain of discrete random vectors $U \to X \to Y \to V$, we have:*

$$I(X,Y) \leq I(X,Y) \tag{2.12}$$

Theorem 2.2.2. *If $X = (X_1, \ldots, X_n)$ is a random vector, with X_i and X_j independent for each $i \neq j$, then:*

$$I(X,Y) \geq \sum_{i=1}^{n} I(X_i, Y_i) \tag{2.13}$$

Proof. We have:

$$I(X,Y) = E(\log \frac{p(x|y)}{p(x)}) = E(\log \frac{p(x|y)}{p(x_1)\ldots p(x_n)})$$

On the other hand:

$$\sum I(X_i,Y_i) = \sum E(\log \frac{p(x_i|y_i)}{p(x_i)}) = E(\log \frac{p(x_1|y_1)\ldots p(x_n|y_n)}{p(x_1)\ldots p(x_n)})$$

So, by Jensen's inequality, we get:

$$\sum I(X_i,Y_i) - I(X,Y) \leq \log E(\frac{p(x_1|y_1)\ldots p(x_n|y_n)}{p(x|y)}) = 0$$

■

Example 2.7. LEt $X_1, \ldots, X_n$ be independent identically distributed random variables with common entropy H.Let π be a permutation on the set $\{1, \ldots, n\}$ and let $Y_i = X_{\pi(i)}$. Then $I(X,Y) = nH$, but $\sum I(X_i,Y_i) = kH$ with k the number of fixed points ($\pi(x) = x$) of π. If $\pi(i) \equiv i + 1 \mod n$, then $k = 0$.

Theorem 2.2.3. *If $X = (X_1, \ldots, X_n)$, and $Y = (Y_1, \ldots, Y_n)$ are random vectors over a discrete memoryless channel, then $I(X,Y) \leq \sum I(X_i,Y_i)$.*

Corollary. $H(X) \leq \sum H(X_i)$

Example 2.8. Let X be a random variable with entropy H and let $X_1 = \cdots = X_n = Y_1 = \cdots = Y_n = X$. Then $I(X,Y) = H$ and $\sum I(X_i,Y_i) = nH$.

2.3 Continuous Random Variables and Vectors.

Definition. If X is a random variable with distribution $F(x) = P(X \leq x)$, and if $P = \{S_i\}_{i \in \mathbb{Z}^+}$ is a partition of $\mathbb{R}$ into finite or countable Lebesgue measurable subsets, we define the **quantization** of X by P to be the discrete random variable $[X]_P$ with probability distribution

$$P([X]_P = i) = P(X \in \S_i) = \int_{S_i} \mathrm{d}F(x_i) \tag{2.14}$$

Definition. We define the **mutual information** between two random variables X and Y to be

$$I(X,Y) = \sup_{P,Q} I([X]_P, [Y]_Q) \tag{2.15}$$

where P and Q are partitions of $\mathbb{R}$ into finite or countable Lebesgue measurable subsets. If $X = (X_1, \ldots, X_n)$ and $Y = (Y_1, \ldots, Y_m)$, then:

$$I(X,Y) = \sup I(X,Y) \tag{2.16}$$

Where the supremum is taken over all partitions of $[X] = ([X_1], \ldots, [X_n])$ and $[Y] = ([Y_1], \ldots, [Y_m])$.

Lemma 2.3.1. *If P_1 and P_2 are partitions of $\mathbb{R}$, and P_2 is a refinement of P_1, then $I([X]_{P_1}, Y) \leq I([X]_{P_2}, Y)$.*

Corollary. *If $\{P_i\}_{i=1}^n$ and $\{Q_i\}_{i=1}^n$ are collections of partitions of $\mathbb{R}$, associated with the random vectors X and Y, and if P_{i+1} is a refinement of P_i, and Q_{i+1} a refinement of Q_i, then:*

$$I([X],[Y]) \geq I([X]_i, [Y]_i) \tag{2.17}$$

For $1 \leq i \leq n$.

Now, for the following results, assume X and Y have continuous joint density $p(x,y)$ defined by

$$P(X \in A, Y \in Y) = \int_A \int_B p(x,y) \ \mathrm{d}x \, \mathrm{d}y$$

for intervals A and B of $\mathbb{R}$ Let the densities of X and Y be defined by

$$p(x) = \int_{-\infty}^{\infty} p(x,y) \ \mathrm{d}y$$
$$q(y) = \int_{-\infty}^{\infty} p(x,y) \ \mathrm{d}x$$

respectively, as well as having conditional probabilities:

$$p(x|y) = \frac{p(x,y)}{q(y)}$$
$$p(y|x) = \frac{p(x,y)}{p(x)}$$

The, assuming that

$$h(X) = \int_{-\infty}^{\infty} p(x) \log p(x)^{-1} \ \mathrm{d}x$$
$$h(X|Y) = \int_{-\infty}^{\infty} p(x,y) \log p(x,y)^{-1} \ \mathrm{d}x$$

exist, we get:

Theorem 2.3.2. $I(X,Y) = h(X) - h(X|Y)$.

Proof. Choose $0 < \epsilon_1 < \epsilon_2$ arbitrarily small and let $\{x_i\}_{i\in\mathbb{Z}}$ be a strictly increasing sequence of points such that $\Delta x_i = x_i - x_{i-1}$ satisfies $\epsilon_1 < \Delta x_i < \epsilon_2$. Similarly, let $\{y_j\}_{j\in\mathbb{Z}}$ be a strictly increasing sequence of points such that $\Delta y_j = y_i - y_{j-1}$ satisfies $\epsilon_1 < \Delta y_j < \epsilon_2$.

Now let $[X]$ be the quantization of the random variable X from the partiotion $P = \{[x_{i-1}, x_i) : i \in \mathbb{Z}\}$ of $\mathbb{R}$ and let $[Y]$ be the quantization of the random variable Y from the partiotion $Q = \{[y_{j-1}, y_j) : j \in \mathbb{Z}\}$ of $\mathbb{R}$. Let

$$p(i) = P([X] = i) = \int_{x_{i-1}}^{x_i} p(x)\ \mathrm{d}x$$
$$q(j) = P([Y] = j) = \int_{y_{j-1}}^{y_j} q(y)\ \mathrm{d}y$$
$$p(x,y) = P([X] = i, [Y] = j) = \int_{y_{j-1}}^{y_j}\int_{x_{i-1}}^{x_i} p(x,y)\ \mathrm{d}y\,\mathrm{d}x$$

and

$$p(i|j) = P([X] = i|[Y] = j) = \frac{p(i,j)}{1(j)}$$

Then, by the mean value theorems for integrals, we have $p(i) = \Delta x_i p(s_i)$ and $q(j) = \Delta y_j p(t_j)$ for $s_i \in [x_{i-1}, x_i)$ and $t_j \in [y_{j-1}, y_j)$. Additionally, we have $p(i,j) = p(s_{ij}|t_{ij})\Delta x_i \Delta y_j q(t_j)$ for $(s_{ij}, t_{ij}) \in [x_{i-1}, x_i) \times [y_{j-1}, y_j)$. Thus, $p(i|j) = \Delta x_i p(s_{ij}|t_{ij})$.

Now, we have $I([X],[Y]) = H([X]) - H([X]|[Y]) = \sum p(i)\log p(i)^{-1} - \sum p(i,j)\log p(i,j)^{-1}$. Then:

$$H([X]) = \sum \Delta x_i p(s_i)\log p(s_i)^{-1} + \sum \Delta x_i p(s_i)\log {\Delta x_i}^{-1}$$
$$H([X]|[Y]) = \sum \Delta x_i \Delta y_j p(s_{ij}|t_{ij})q(t_j)\log p(s_{ij}|t_{ij})^{-1} + \sum \Delta x_i \log {\Delta x_i}^{-1} \sum \Delta y_i p(s_{ij}|t_{ij})q(y_j)$$

Now, as $\epsilon_2 \to 0$, we see that $H([X]) \to h(X) + \log \epsilon_1^{-1}$ and $H([X]|[Y]) \to h(X|Y) + p(s_i)$. Thus, we see that as $\epsilon_2 \to 0$, then $I([X],[Y]) \to h(X) - h(X|Y)$. ∎

Definition. Assume that X and Y have joint density $p(x,y)$, with X having density $p(x)$. We define the **differential entropy** of X to be:

$$h(X) = \int_{-\infty}^{\infty} p(x)\log p(x)^{-1}\ \mathrm{d}x \tag{2.18}$$

We define the **conditional differential entropy** of X given Y to be:

$$h(X|Y) = \int_{-\infty}^{\infty} p(x,y)\log p(x,y)^{-1}\ \mathrm{d}x \tag{2.19}$$

Example 2.9. Let $X = (X_1, \ldots, X_n)$ with X_i idnependent Gaussian random variables with mean μ and variance $\frac{2}{i}$. Then the density for X is

$$g(x) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma_i^2}\right) \tag{2.20}$$

which is called the **Gaussian density**. Now, computing the differential entropy of X we get:

$$\begin{aligned} h(X) &= \int g(x) \log \frac{1}{g(x)} \\ &= \sum_{i=1}^{n} \frac{1}{2} \log 2\pi\sigma_i^2 + \frac{1}{2} \\ &= \frac{n}{2} \log 2\pi e \sqrt[n]{\sigma_1^2 \dots \sigma_n^2} \end{aligned}$$

Theorem 2.3.3. *IF $X = (X_1, \dots, X_n)$ has density $p(x)$, and if $E((x_i - \mu_i)^2) = \sigma_i^2$ for $1 \leq i \leq n$, then $h(X) \leq \frac{n}{2} \log 2\pi e \sqrt[n]{\sigma_1^2 \dots \sigma_n^2}$; with equality holding if, and only if $p(x) = g(x)$.*

Proof. By hypothesis, if $p_i(x)$ is the marginal distribution of X_i, we have $\int p_i(x) \ \mathrm{d}x = 1$ and $\int p_i(x)(x_i - \mu_i)^2 \ \mathrm{d}x = \sigma_i^2$. Thus

$$\int p(x) \log \frac{1}{g(x)} \ \mathrm{d}x = \frac{n}{2} \log 2\pi e \sqrt[n]{\sigma_1^2 \dots \sigma_n^2}$$

So if Y is an n-dimensional random Gaussian (i.e. it has Gaussian distribution) vector, then

$$h(X) - h(X|Y) = \int p(x) \log \frac{g(x)}{p(x)} \ \mathrm{d}x \leq \log \int g(x) \ \mathrm{d}x = 0$$

by Jensen's inequality. ■

We now state, but don't prove the theorems from previous sections which are true for continuous random variables.

Theorem 2.3.4. $I(X, Y) \geq 0$.

Theorem 2.3.5. $I([X], [Y]) \geq \sum I([X_i], [Y_i])$.

Theorem 2.3.6. $p(y|x) = \prod p(x_i|y_i)$.

Corollary. *For any discrete memoryless channel,*

$$I([X], [Y]) \leq \sum I([X_i], [Y_i])$$

Chapter 3

Discrete Memoryless Channels and Capacity-Cost Functions.

3.1 Capacity and Cost.

We define now in a more precise manner the discrete memoryless channel.

Definition. A **discrete memoryless channel** is a triple $D = (A_X, A_Y, Q)$, where A_X and A_Y are finite sets of size $|A_X| = r$ and $|A_Y| = s$, called the **input alphabet**, and **output alphabet**, respectively; and Q is an $r \times s$ matrix called the **transitional probability matrix** whose entries are $(p(y|x))$. We also define a map $b : A_X \to \mathbb{R}$ called the **cost function** of D, and we call the value $b(x)$ the **cost** of x.

Example 3.1. (1) Let $A_X = A_Y = \mathbb{F}_2$ with $Q = \begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix}$, where $0 \leq p \leq \frac{1}{2}$, and $b(0) = 0$, $b(1) = 1$. This describes a DMC called the **binary symmetric channel**.

(2) Let $A_X = \{0, \frac{1}{2}, 1\}$ and $A_Y = \mathbb{F}_2$, with $Q = \begin{pmatrix} 1 & 0 \\ \frac{1}{2} & \frac{1}{2} \\ 0 & 1 \end{pmatrix}$ with $b(0) = b(1) = 1$ and $b(\frac{1}{2}) = 0$.

(3) Let $A_X = A_Y = \mathbb{F}_3$, with $Q = I_{3\times 3}$ and $b(0) = b(1) = 1$ and $b(2) = 4$.

Definition. Let $D = (A_X, A_Y, Q)$ be a DMC with cost $b(x)$. Assume that D is used n consecutive times with input and output sequences $x = \{x_i\}_{i=1}^n$ and $y = \{y_i\}_{i=1}^n$. The **memoryless assumption** that y_i is a function only of x_i is defined to be the product $\prod_{i=1}^n p(x_i, x_i)$. We define the **cost** of sending x over the channel to be:

$$b(x) = \sum_{i=1}^{n} b(x_i) \tag{3.1}$$

We define the **average cost** of sending x to be:

$$\overline{b}(x) = E(b(x)) = \sum_{i=1}^{n} p(x_i) b(x_i) \tag{3.2}$$

Definition. For a DMC $D = (A_X, A_Y, Q)$ with cost $b(x_i)$ and $n \in \mathbb{Z}^+$; we define the **n-th capacity-cost function** of D to be:

$$C_n(\beta) = \max_{x,y \in A_X^n \times A_Y^n} \{I(X,Y) : \overline{b}(x) \leq n\beta\} \tag{3.3}$$

Where X and Y are the random variables associated with x and y. We call x the **test source**, and say it is **β-admissable** if $\overline{b}(x) \leq n\beta$.

Lemma 3.1.1. *For a test source $x = \{x_i\}$, if $\beta_{\min} = \min_{x_i \in A_X} b(x_i)$, then $n\beta_{\min} \leq \overline{b}(x)$.*

Proof. Notice that $\beta_{\min} \leq \frac{\sum p(x_i)b(x_i)}{n} = \frac{\overline{b}(x)}{n}$. ■

Corollary. *$C_n(\beta)$ is defined for all $\beta \geq \beta_{\min}$.*

Corollary. *If $\beta_{\min} \leq \beta_1 \leq \beta_2$, then $C_n(\beta_1) \leq C_n(\beta_2)$.*

Proof. Let $C_i = \{I(X,Y) : \overline{b}(x) \leq n\beta_i\}$ for $1 \leq i \leq 2$. Then if x achieves $C_n(\beta_1)$, that is, $I(X,Y) = C_n(\beta_1) \in C_1$ and $\overline{b}(x) \leq n\beta_1$, then $\overline{b}(x) \leq n\beta_2$. This makes $I(X,Y) \in C_2$. Thus $C_1 \subseteq C_2$ implying $C_n(\beta_1) \leq C_n(\beta_2)$. ■

Definition. Let $D = (A_X, A_Y, Q)$ be a DMC. We define the **capacity-cost function** of D to be:

$$C(\beta) = \sup \frac{C_n(\beta)}{n} \tag{3.4}$$

Theorem 3.1.2. *For any DMC, the n-th capacity-cost, $C_n(\beta)$ is convex down for all $\beta \geq \beta_{\min}$.*

Proof. Let $\alpha_1, \alpha_2 \geq 0$ with $\alpha_1 + \alpha_2 = 1$ and let $\beta_1, \beta_2 \geq \beta_{\min}$. Let x_1 and x_2 be test sources with probailities $p_1(x)$ and $p_2(x)$, both achieving $C_n(\beta_1)$ abd $C_n(\beta_2)$, respectively. Define x to be the test source with probability $p(x) = \alpha_1 p_1(x) + \alpha_2 p_2(x)$. Then $\overline{b}(x) = \sum_x p(x)b(x) = \alpha_1 \sum p_1(x)b(x) + \alpha_2 \sum p_2(x)b(x) = \alpha_1 \overline{b}(x_1) + \alpha_2 \overline{b}(x_2) \leq n(\alpha_1\beta_1 + \alpha_2\beta_2)$. Additionally, since $I(X,Y)$ is convex down, we have:

$$C_n(\alpha_1\beta_1 + \alpha_2\beta_2) \geq I(X,Y) \geq \alpha_1 I(X_1,Y_1) + \alpha_2 I(X_2,Y_2) = \alpha_1 C_n(\beta_1) + \alpha_2 C_n(\beta_2)$$

■

Theorem 3.1.3. *For any DMC, $C_n(\beta) = nC_1(\beta)$, for $n \in \mathbb{Z}^+$ and all $\beta \geq \beta_{\min}$.*

Proof. Let $x = \{x_i\}$ be β-admissable, achieving $C_n(\beta)$. By theorem 2.2.2, $I(X,Y) \leq \sum I(X_i,Y_i)$. Defining $\beta_i = \overline{b}(x_i)$, we get $\sum \beta_i = \sum\sum p(x_i)b(x_i) = \overline{b}(x) \leq n\beta$. We also have $I(X_i,Y) \leq C_1(\beta_i)$. Then, by Jensen's inequality:

$$\frac{1}{n}\sum C_1(\beta_i) \leq C_1(\frac{1}{n}\sum \beta_i) = C_1(\frac{\overline{b}(x)}{n}) \leq C_1(\beta)$$

So C_1 is increasing, and $\sum C_1(\beta_i) \leq nC_1(\beta)$. Then we get from the above results, that $C_n(\beta) \leq nC_1(\beta)$.

Now, let $x = \{x_i\}$ be a β-admissable test source achieving $C_1(\beta)$; i.e. $\overline{b}(x) \leq \beta$, and $I(X,Y) = C_1(\beta)$. Now, consider the random variables $X_1, \ldots, X_n$, corresponding to $x_1, \ldots, x_n$ and assume they are independent and identically distributed with distributions that of X, associated with x. Then $\overline{b}(x) = \sum \overline{b}(x_i) \leq n\beta$ and $I(X,Y) = \sum I(X_i,Y_i) = nC_1(\beta)$. This makes $C_n(\beta) \geq nC_1(\beta)$. ■

Corollary. *For a memoryless channel, $C(\beta) = C_1(\beta)$.*

Lemma 3.1.4. *For all $\beta \geq \beta_{\min}$ $C(\beta)$ is constant for β sufficiently large.*

Proof. Define $C_{\max} = \max\{C(\beta) : \beta \geq \beta_{\min}\} = \max_{x \in A_X}\{I(X,Y)\}$. Define:

$$\beta_{\max} = \min\{\bar{b}(x) : I(X,Y) = C_{\max}\}$$

Then $C(\beta) = C_{\max}$ for $\beta \geq \beta_{max}$, and $C(\beta) < C_{\max}$ otherwise. ■

Definition. For any *DMC* D, we definie the **channel capacity** of D to be:

$$C_{\max} = \max\{C(\beta) : \beta \geq \beta_{\min}\} \tag{3.5}$$

Lemma 3.1.5. *A test source x is $\beta_{\min}$-admissable if, and only if $p(x) = 0$ whenever $b(x) > \beta_{\min}$.*

Example 3.2. Consider the BSC, that is, the triple $(\mathbb{F}_2, \mathbb{F}_2, Q)$ where $Q = \begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix}$, with $0 \leq p \leq \frac{1}{2}$ and costs $b(0) = 0$ and $b(1) = 1$. Then $\beta_{\min} = 0$, and the reduced channel has only one input, so $C_{\min} = C(0) = 0$. Now, let x be a test source achieving $C(\beta)$ for $0 \leq \beta \leq \beta_{\max}$. Then $p(1) = \beta$ and $p(0) = 1 - \beta$, and $C(\beta) = H(Y) - H(X|Y) = H((1-\beta)(1-p) + \beta p) - H(p)$. Then $H((1-\beta)(1-p) + \beta p)$ attains maximum at $\beta = \frac{1}{2}$, so $\beta_{\max} = \frac{1}{2}$. So, $C(\beta)$ gives the curve:

$$C(\beta) = \begin{cases} H((1-\beta)(1-p) + \beta p) - H(p), 0 \leq \beta < \frac{1}{2} \\ \log 2 - H(p), \beta \geq \frac{1}{2} \end{cases}$$

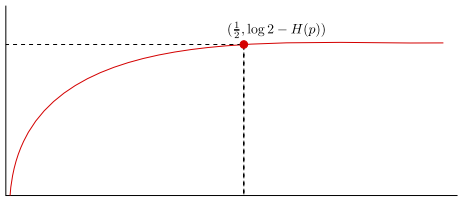

Figure 3.1: The Capacity-Cost function of the BSC, $C(\beta)$.

Definition. We call a *DMC* D with transition matirx Q symmetric if Q is a symmetric matrix.

Theorem 3.1.6. *If D is a DMC with r inputs and s outputs, then the channel capacity for D is a acheived with equiprobable inputs at:*

$$C_{\max} = \log s - H(q_0, \ldots, q_{s-1}) \tag{3.6}$$

where $(q_1 \ \ldots \ q_{s-1})$ is any row of the transition matrix Q of D.

Proof. We have $I(X,Y) = H(Y) - H(Y|X) = H(Y) - \sum_x p(x)H(Y|x)$. Now,

$$H(Y|x) = \sum_y p(y|x)\log\left(\frac{1}{p(y|x)}\right) = H(q_0, \dots q_{s-1})$$

is independent of X, so by theorem 2.1.2, $H(Y) \leq \log s$, with equality holding if, and only if $p(y) = \frac{1}{s}$ for all $y \in A_Y$. This is implied by the transition matrix, and so gives us the result. ■

Example 3.3. (1) Consider a DMC with transition matrix $Q = \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{6} & \frac{1}{6} \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} \end{pmatrix}$ then, the DMC has channel capacity $C_{\max} = \log 4 - H(\frac{1}{3}, \frac{1}{3}, \frac{1}{6}, \frac{1}{6}) = \log \frac{\sqrt[3]{2^5}}{3}$ bits.

(2) Consider the **r-ary symmetric channel** with transition $r \times r$ matrix (q_{xy}) where $q_{xy} = \epsilon$ if $x \neq y$ and $q_{xy} = 1 - (r-1)\epsilon$ otherwise. For $r = 2$, we get the BSC, and for $r = 4$, we get

$$Q = \begin{pmatrix} 1-3\epsilon & \epsilon & \epsilon & \epsilon \\ \epsilon & 1-3\epsilon & \epsilon & \epsilon \\ \epsilon & \epsilon & 1-3\epsilon & \epsilon \\ \epsilon & \epsilon & \epsilon & 1-3\epsilon \end{pmatrix}$$

The channel capacity for this channel is $\log r - H(1-(r-1)\epsilon, \epsilon, \epsilon, \epsilon) = \log r + (1-r)\epsilon \log \epsilon + (1 - r\epsilon + \epsilon)\log(1 - r\epsilon + \epsilon)$.

3.2 The Channel Coding Theorem.

We begin this section with two results.

Lemma 3.2.1. *For any discrete memoryless channel, with capacity $C_{\max}$, the DMC can transmit at most $C_{\max}$ bits of information per unit time.*

Corollary. *If X is a test source achieving $C(\beta_{\max}) = C_{\max}$, then the DMC can transmit at leeast $C_{\max}$ bits of information per unit time.*

Remark. So, if we choose our test source right, we can transmit, with the DMC, exactly $C_{\max}$ bits of information per unit time; i.e. we transmit all communcations optimally at channel capacity.

Definition. We define the **rate** of a channel to be the ratio

$$\frac{k}{n} \tag{3.7}$$

of number of symbols transmitted per channel use.

Lemma 3.2.2. *Given a channel with rate $\frac{k}{n}$, and $\epsilon > 0$ small enough, such that for input and output sequences $\{U_i\}$ and $\{\hat{U}_i\}$, respectively, $P(\hat{U}_i \neq U_i) < \epsilon$, then:*

$$\frac{k}{n} \leq \frac{C(\beta)}{1 - H(\epsilon)} \tag{3.8}$$

Proof. Let $U = \{U_i\}_{i=1}^k$ be a sequence of independent random variable with distribution $P(U = 0) = P(U = 1) = \frac{1}{2}$. Let $X = (X_1, \ldots, X_n)$ and $Y = (Y_1, \ldots, Y_n)$ be channel inputs and outputs respectively, with X an encoding of U. Now, from Y decode the recieved message as $\hat{U} = (\hat{U}_1, \ldots, \hat{U}_k)$.

Suppose $P(\hat{U}_i) < \epsilon$ for $1 \leq i \leq k$ and $\epsilon > 0$ small enough. Then $I(U, \hat{U}) \geq \sum_{i=1}^k I(U_i, \hat{U}_i) = \sum H(U_i) - H(U_i|\hat{U}_i) = \sum \log 2 - H(U_i|\hat{U}_i) \geq \log 2 - H(\epsilon)$ by theorem 2.2.2 and Fano's inequality. So $I(U, \hat{U}) \geq k - kH(\epsilon)$ Then, by the data processing theorem, we have:

$$I(U, \hat{U}) \leq I(X, Y) \leq C_n(\beta) = nC(\beta)$$

thus we get,

$$\frac{k}{n} \leq \frac{C(\beta)}{1 - H(\epsilon)}$$

■

Remark. What this says is that we cannot cimmunicate above the channel capacity $C_{\max}$.

We come now to the next big definition of information theory.

Definition. Let $n \in \mathbb{Z}^+$ and let $D = (A_X, A_Y, Q)$ be a discrete memoryless cahnnel. We define a **channel code** of **length** n to be a subset $\mathcal{C} \subseteq A_X^n$ with $\mathcal{C} = \{x_1, \ldots x_M\}$ for some $M \in \mathbb{Z}^+$. We call the elements of $\mathcal{C}$ **codewords**.

We define the **rate** of the code $\mathcal{C}$ to be :

$$R = \frac{1}{n} \log M \tag{3.9}$$

We say that $\mathcal{C}$ is β**-admissible** of $b(x_i) = \sum_j b(x_{ij}) \leq n\beta$ for all i and $\beta > \beta_0$.

We define a **decoding rule** for the code $\mathcal{C}$ to be a map $f : A_Y^n \to \mathcal{C} \cup \{e\}$, where e is called the **error**. We define an **encoding rule** to be a $1 - 1$ map from all possible source sequences into $\mathcal{C}$.

We define the **error probability** of transmitting the codeword x_i of $\mathcal{C}$ to be:

$$P_e^{(i)} = P(f(y) \neq x_i) = \sum_{f(y) \neq x_i} p(y|x_i) \tag{3.10}$$

where $p(y|x_i) = \prod_{j=1}^n p(y_i|x_{ij})$ is the product of all transitional probabilities of D, i.e. the product of all the entries of Q.

Remark. Here, we take the codeword x to be an encoding of the message u which we regard as already having been transmitted.

Example 3.4. (1) Consider the binary symmetric channe (BSC) wiht $b(0) = b(1) = 0$. For $n = 3$, $M = 2$, define the code $\mathcal{C} \subseteq \mathbb{F}_2^3$ by $\mathcal{C} = \{(000), (111)\}$. The rate for the code is $R = \frac{1}{3} \log 2 = \frac{1}{3}$.

Now, define the decoding rule $f : y_1y_2y_3 \to xxx$ such that $x = y_1$ if $y_1 = y_2$ and $y_1 = y_3$, $x = y_2$ if $y_2 = y_1$ or $y_2 = y_3$ and $x = y_3$ if $y_3 = y_1$ or $y_3 = y_2$; i.e. we take the majority vote on the three bits. Then $P_e^{(1)} = P_e^{(2)} = 3p^2 - 2p^3 < p < \frac{1}{2}$.

(2) Consider the DMC defined in example 3.1(2), with $b(0) = b(1) = 1$ and $b(\frac{1}{2}) = 0$. Take $e = \frac{1}{2}$ to be the error. Define the code $\mathcal{C}A_X^n$ by $\mathcal{C} = \{(x_1, \ldots, x_k, \frac{1}{2}, \ldots, \frac{1}{2}) : x_i \in \mathbb{F}_2 \text{ and } 1 \leq i \leq k\}$, for $k \leq n$. Take $M = 2^k$; then the rate is $R = \frac{1}{n} \log 2^k = \frac{k}{n}$.

Notice for $\beta > R$, that $b(x_i) \leq \beta n$, so $\mathcal{C}$ is β-admissible. Take the decoding rule for $\mathcal{C}$, f defined by the rule $(y_1, \ldots, y_n) \to (y_1, \ldots, y_k, \frac{1}{2}, \ldots, \frac{1}{2})$. Then $P_e^{(i)} = 0$.

(3) Let $A_X = \mathbb{F}_2$ and $A_Y = \mathbb{Z}/4\mathbb{Z}$, and let the transition matrix be that of example 3.3(1), with $b(0) = b(1) = 0$. Let $n = 2$ and $M = 2$ and consider the code $\mathcal{C} = \{(00), (11)\}$. The rate for the code is $R = \frac{1}{2}$.

Now, take the decoding rule $f : (y_1 y_2) \to a_{ij}$ where (a_{ij}) is the matrix:

$$\begin{pmatrix} 00 & 00 & 00 & e \\ 00 & 00 & e & 11 \\ 00 & e & 11 & 11 \\ e & 11 & 11 & 11 \end{pmatrix}$$

whose rows are all possible values for $y_1 \in \mathbb{Z}/4\mathbb{Z}$ and whose columns are all possible values for $y_2 \in \mathbb{Z}/4\mathbb{Z}$. Then $P_e^{(i)} = \frac{4}{9}$ for $1 \leq i \leq 2$.

We now come to the channel coding theorem.

Theorem 3.2.3 (The Channel Coding Theorem)**.** *Let D be a DMC with capacity-cost function $C(\beta)$. Then for any $\beta_0 \geq \beta_{\min}$, and real numbers $\beta > \beta_0$, $R < C(\beta_0)$ and $\epsilon > 0$, and for n sufficiently large, there exists a code $\mathcal{C}$ of length n such that:*

(1) Every $x \in \mathcal{C}$ is β-admissible.

(2) $M \geq 2^{\lceil Rn \rceil}$.

(3) $P_e^{(i)} < \epsilon$ for all $1 \leq i \leq M$.

Proof. Let n be sufficiently large, and consider the set $\mathcal{O} = A_X^n \times A_Y^n$ of all pairs (x, y) of input and output vectors. Define $p : \mathcal{O} \to \mathbb{R}$ by

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

where $p(x) = \prod_{i=1}^n p(x_i)$ is the probability distribution on A_X^n, achieving $C(\beta_0)$ and $p(y|x) = \prod_{i=1}^n p(y_i|x_i)$ is the product of all transitional probabilities of the discrete memoryless channel D.

Now, choose R' such that $R < R' < C(\beta_0)$ and define $T \subseteq \mathcal{O}$ by:

$$T = \{(x, y) : I(x, y) \geq nR'\}$$

where $I(x, y) = \log \frac{p(x|x)}{p(y)}$. Now, define the set of all β-admissible codewords $B \subseteq A_X^n$ by:

$$B = \{x : b(x) \leq \beta n\}$$

That is, B is the collection of all balls of radius lessthan or equal to βn. Then define $T' \subseteq T$ by:

$$T' = \{(x, y) : (x, y) \in T \text{ and } x \in B\}$$

Now, let $\mathcal{C} = \{x_1, \ldots, x_M\}$ be any code of length n, and consider the ball about y of radius less than or equal to βn:

$$S(y) = \{x : (x, y) \in T'\} \subseteq B$$

Define then, the decoding rule f of $\mathcal{C}$ as follows: For any $y \in A_Y^n$, if $S(y)$ contains exactly

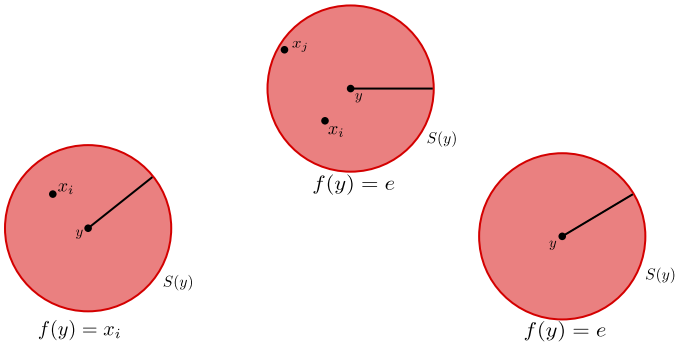

Figure 3.2: The Decoding spheres of the code $\mathcal{C}$.

one such codeword of $\mathcal{C}$, x_i, then $f(y) = x_i$. Otherwise, $f(y) = e$, an error. That is, if there are no codewords, or more than one (disctinct) codeword in any given sphere, we introduce an error.

Now, transmitting x_i and recieving y, error can occure if, and only if either $x_i \notin S(y)$, or $x_i, x_j \in S(y)$ for $i \neq j$ and x_i and x_j distinct. Thus, we get:

$$P_e^{(i)} = P(x_i \in S(y)) + \sum_{j=1, j\neq i}^{M} P(x_j \in S(y))$$

Define then the indicator functions Δ and Δ' of T' as:

$$\Delta(x, y) = \begin{cases} 1, & (x, y) \in T' \\ 0, & (x, y) \notin T' \end{cases}$$

and

$$\Delta'(x, y) = \begin{cases} 0, & (x, y) \in T' \\ 1, & (x, y) \notin T' \end{cases}$$

Then $P_e^{(i)} \leq \sum \Delta'(x_i, y) p(y|x_i) + \sum_{j\neq i} \sum_y \Delta(x_j, y) p(y|x_i) = Q_i(x_1, \ldots, x_M)$ (note, Q_i is not the transition matrix Q).

Now, define the probability distribution on A_X^n b y:

$$P(x_1, \ldots, x_M) = \prod_{i=1}^{M} p(x_i)$$

This distribution corresponds to picking a code $\mathcal{C}$ at randomby picking its codewords independently according to the probability distribution $p(x)$ achieving $C(\beta_0)$. Then viewing $Q_i(x_1, \ldots, x_M)$ as a random variable on A_X^n, we find :

$$E(Q_i(x_1, \ldots, x_M)) = E_1 + \sum_{j=1}^{M} E_2^{(j)}$$

Now, $E_1 = \sum \prod_{i=1}^{M} p(x_i) \sum \Delta'(x_i, y) p(y|x_i) = \sum p(x_i) p(y|x_i) \Delta'(x_i, y) = \sum p(x, y) \Delta'(x, y) = P((x, y) \notin T') \leq P((x, y) \notin T) + P(x \notin B)$. Thus:

$$E_1 \leq P(I(x, y) < nR') + P(b(x) > \beta n)$$

Notice, however that $I(x, y) = \sum I(x_k, y_k)$ which has mean $C(\beta_0)$; and since $R' < C(\beta_0)$, by the weak law of large numbers:

$$\lim_{n \to \infty} P(I(x, y) < nR') = 0$$

So, $E_1 \to_0$ as $n \to \infty$. Similarly, $P(b(x) < \beta n) \to 0$ as $n \to \infty$.

On the otherhand, $E_2^{(j)} = \sum_{i=1}^{M} \prod p(x_i) \sum \Delta(x_j, y) p(y|x_j) = \sum p(x_j) \Delta(x_j, y) \sum p(x_i) p(y|x_i) = \sum p(x_j) \Delta(x_j, y) p(y)$. So

$$E_2^{(j)} \leq \sum_{(x,y) \in T} p(x) p(y)$$

Now, $p(x)p(y) \leq p(x, y) 2^{-R'n}$, thus $E_2^{(j)} \leq 2^{-R'n}$, so we get:

$$E(Q_i) \leq P(I(x, y) < nR') + P(b(x) > \beta n) + M_2^{-R'n}$$

So $E(Q_i) \leq M_2^{-R'n}$ as $n \to \infty$. If $M = 2^{\lceil Rn \rceil + 1}$, then $E(Q_i) \leq 2^{-n(R'-R)+2}$; and since $R' > R$, we get $E(Q_i) \to 0$ as $n \to \infty$. Therefore, for n large enough, and $M = 2^{\lceil Rn \rceil + 1}$, we get:

$$E(Q_i) < \frac{\epsilon}{2}$$

For $\epsilon > 0$ small enough, and we have established condition (2) of the theorem.

Now define:

$$P_e(x_1, \ldots, x_M) = \frac{1}{M} \sum_{i}^{M} P_e^{(i)}(x_1, \ldots, x_M)$$

to be the overall error probability assuming each codeword of the code $\mathcal{C}$ is transmitted with probability $\frac{1}{M}$. Regardign $P_e(x_1, \ldots, x_M)$ as a random variable over A_X^n, we have for n large, and $M = 2^{\lceil Rn \rceil + 1}$. that $E(P_e) < \frac{\epsilon}{2}$. Thus there is a code for which $P_e(x_1, \ldots, x_M) < \frac{\epsilon}{2}$, and consequently for which:

$$P_e^{(i)} < \frac{\epsilon}{2} < \epsilon$$

For $\epsilon > 0$ small enough.

Now, it may be that the code $\mathcal{C}$ contains codewords x_i for which $b(x_i) > \beta n$, or $P_E^{(i)} \geq \epsilon$, or both; furthermore, if more than half the codewords have $P_e^{(i)} \geq \epsilon$, then $P_e > \frac{\epsilon}{2}$, which

cannot happen. So deleting all the codewords x_i for which $P_e^{(i)} \geq \epsilon$ from $\mathcal{C}$, we get a code $\mathcal{C}'$ with $|\mathcal{C}'| = 2^{\lceil Rn \rceil}$ codewords for which $P_e^{(i)} < \epsilon$. Notice, then that if $b(x_i) > \beta n$ for some $x_i \in \mathcal{C}'$, then $x_i \notin S(y)$, making $P_e^{(i)} = 1 > \epsilon$. So $\mathcal{C}'$ cannot contain non β-admissible codewords, and we complete the proof. ■

Remark. The proof of the channel coding theorem is very involved and warrants closer study of the arguments to better understand.

Corollary (The Channel Coding Theorem for Discrete Memoryless Channels.)**.** *For any $R < C_{\max}$ and $\epsilon > 0$, there exists a code $\mathcal{C} = \{x_1, \ldots, x_M\}$ of length n and decoding rule such that:*

(1) $M \geq 2^{\lceil Rn \rceil}$....

(2) $P_e^{(i)} < \epsilon$ for all $1 \leq i \leq M$.

Proof. Let $\beta_0 = \beta_{\max}$. ■

Bibliography

[1] R. McEliece, *The theory of information and coding*. Cambridge: Cambridge University Press, 2001.