Information Theory

Manan Shah
manan.shah.777@gmail.com
The Harker School

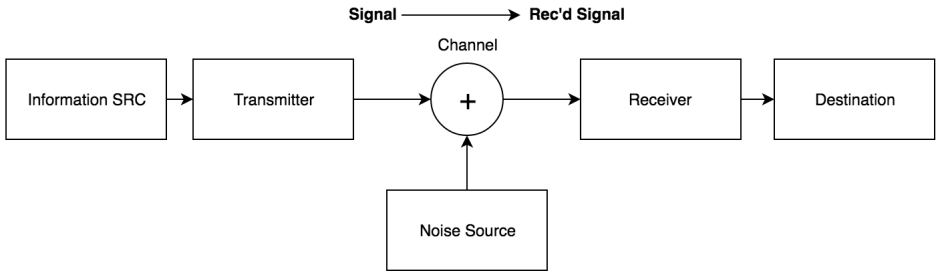
October 7, 2016

This document contains lecture notes from Harker's Advanced Topics in Mathematics class in Information Theory, Parts I and II. These notes were taken using TeXShop and LaTeX2ϵ and will be updated for each class. The reader is advised to note any errata at the source control repository https://github.com/mananshah99/infotheory.

Contents

1 August 22, 2016

1.1 Class Overview



Information may be conveyed in multiple ways; for example, by a person talking, a video, or pictures. Our goal is to model this kind of information as a system with one output and no input. The information is therefore a box that emits output probabilistically, and we will therefore be spending 1-2 weeks on probability, specifically regarding distribution functions and expectation. This is illustrated in the "information source" box—the job of the information source is to **provide the signal**.

We will next define the compression problem; having obtained numbers as a stream, the problem is to remove the redundancy in a stream. In particular, we will discuss both lossless and lossy compression. One of the most critical ideas in this course is that of entropy, and we will define the concept (in general, a notion of how much information is originating from the source). We will also discuss the sensibility of such a definition and potentially expand upon it. An overarching concept here will be our definition of the performance of a compression algorithm via the source compression theorem and the boundedness of compression algorithms. This is illustrated in the "transmitter" box—the job of the transmitter is to **remove all redundancy**.

We will subsequently discuss the channel circle, which we will model with conditional probability distribution functions (PDFs). The output of this box, the receiver signal, is input to the receiver box, which has the job of **reconstruction of the signal and the noise**; this is known as the coding problem. Much like the compression problem of the source, the channel has the capacity problem (how much information can be transmitted?). And similar to the compression theorem, we will discuss the channel coding theorem, which will help us determine which parts of a code will require more redundancy for transmission.

We will finally discuss Shannon's theorem, in which he states that a communication with no loss and arbitrary cost (due to a constraint on the amount of bandwith, power, etc.) is possible if the entropy of the source H is less than the capacity of the channel C. In our course, we will first discuss the pipeline for lossless compression and then move to lossy compression.

2 August 24, 2016

Knowing what a random variable is and what it means is important for the source to transmitter relationship.

2.1 Random Variables

Definition 1. A random variable is a real-valued function of the outcome of an experiment. By definition, it is not deterministic.

Notation 1. We will represent random variables with capital letters (for example, X, Y, Z).

Example 1 (What are Random Variables?)**.** (a) A coin is tossed ten times. Determine the number of heads obtained from the coin toss. The number of heads is the random variable in this case. (b) Given two rolls of die, identify their sum (a random variable) and the second roll raised to the fifth power (another random variable). (c) The time to transmit a message, the delay with which a message is received, and the number of errors in a message are both themselves random variables.

Two types of random variables exist; discrete variables and continuous variables.

Definition 2 (Discrete RV)**.** A discrete random variable is one that has a finite set of outcomes (or is countably infinite), and has a probability associated with each outcome.

Definition 3 (Continuous RV)**.** A continuous random variable is one that has a non-finite, uncountable set of outcomes.

Example 2. Define the function

$$sgn(a) = \begin{cases} 1 & a > 0 \\ 0 & a = 0 \\ -1 & a < 0 \end{cases}$$

The variable a is a continuous random variable, while $sgn(a)$ is a discrete function (with outputs $\in \{1, 0, -1\}$)

2.1.1 Probability Mass Function (pmf)

Probability mass functions are functions of a discrete random variable, describing all instances of a random variable occurring and its associated probabilities. Define X as a random variable, and write $P(X = x)$ as $p_X(x)$. Then $\sum_x p_X(x) = 1$. Consider two tosses of a coin, letting X equal the number of heads obtained. We therefore have

$$p_X(x) = \begin{cases} 1/4 & x = 0 \\ 1/2 & x = 1 \\ 1/4 & x = 2 \end{cases}$$

note that $p_X(x)$ is clearly a discrete function that takes on values $x \in \{0, 1, 2\}$ with respective output probabilities.

2.1.2 Probability Density Function (pdf)

Probability density functions are functions of a continuous random variable. Define X as a random variable. In this case, the expression $P(X = x)$ is meaningless (in fact, it is defined as 0). The probability of a range of values from $x = a$ to $x = b$ is more relevant; in particular, $\int_{-\infty}^{\infty} p_X(x) = 1$.

Example 3. Edgar's driving time to school is between 15 and 20 minutes if it is sunny and between 20 and 25 minutes if it is a rainy day. Assume a day is sunny with 2/3 probability and rainy with 1/3 probability. Construct the respective PDF.

The PDF construction is quite simple; remember that $\int_{-\infty}^{\infty} p_X(x) = 1$.

$$p_X(x) = \begin{cases} 2/15 & 15 \leq x \leq 20 \\ 1/15 & 20 \leq x \leq 25 \\ 0 & x \geq 25 \end{cases}$$

2.1.3 Types of Random Variables

We will define these distributions in terms of a coin toss.

Bernoulli (Discrete). Toss a coin, and it define the probability of heads as p. The respective PMF is

$$p_X(x) = \begin{cases} p & x = 1 \\ (1-p) & x = 0 \end{cases}$$

Binomial (Discrete). Toss a coin N times, and define x as the number of heads obtained in N tosses. Let $N = 10$, and define the probability of heads as p. Then the respective PMF is

$$p_X(x) = \begin{cases} (1-p)^{10} & x = 0 \\ \binom{10}{2} p^2 (1-p)^8 & x = 2 \\ \dots & \text{other values of } x \in [0, 10] \cap \mathbb{N} \end{cases}$$

Geometric (Discrete). Repeatedly toss a coin until the first "success." This means that we may theoretically have countably infinite $P(X = 1), P(X = 2), P(X = 3), \dots$ output values. Define success to be a head, and the probability of heads as p. Then, we have

$$p_X(x) = \begin{cases} p & x = 1 \\ (1-p)p & x = 2 \\ (1-p)^2 p & x = 3 \\ \dots & \text{other values of } x \end{cases}$$

Poisson (Discrete, $\sim$ Binomial). Instead of tossing a biased coin and counting the number of heads (a binomial distribution), count the number of times one must replace a biased lightbulb in a time t (with the lightbulb being on or off). This makes the poisson distribution a discrete one. Our distribution function is

$$p_X(x) = \frac{e^{-\lambda} \lambda^x}{x!}$$

where λ is an arbitrary constant. The number of decay events in some radioactive element, for example, may be modeled using a Poisson distribution. Anything that involves units of time and something happening within a period of time t may include a Poisson distribution.

Exponential (Continuous, $\sim$ Geometric). Given a biased lightbulb, identify the amount of time until the bulb burns out. Examples of the use of this function include rates (as opposed to

numbers defined with the Poisson distribution). The PDF is defined as

$$p_X(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

where λ is an arbitrary constant.

Gaussian (Continuous). If one does enough experiments of a binomial random variable, the distribution will end up looking like a Gaussian. The PDF is

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

where μ and σ are parameters that can be treated as constants.

3 August 26, 2016

3.1 Properties of Random Variables

We will be defining expectation and variance, and deriving equations regarding these properties of random variables.

3.1.1 Expectation (Mean) of a Random Variable

Discrete. The notation is $E[X]$, where X is the random variable. We have

$$E[X] = \mu = \sum_x x_i p(x_i)$$

For example, the expected value of rolling a die is 3.5.

Continuous. The notation is again $E[X]$, where X is the random variable. We have

$$E[X] = \mu = \int_{-\infty}^{\infty} x p(x) dx$$

In fact, it is possible to find the expected value of any function of a random variable. For example,

$$E[X^2] = \sum_x x_i^2 p(x_i)$$

$E[X]$ is called the "first moment" of a variable, and $E[X^2]$ is called the "second moment."

3.1.2 Variance of a Random Variable

We will denote the variance of a random variable as $V[X]$ (in class, this is denoted $Var[X]$). The variance depends on the expectation, and is written as

$$V[X] = E[(X - E[X])^2] = E[X^2] - E[X]^2$$

It is possible to expand this condensed form in both discrete and continuous forms.

Discrete. $V[X] = \sum_x (x_i - \mu)^2 p(x_i)$ **Continuous.** $V[X] = \int_{-\infty}^{\infty} (x_i - \mu)^2 p(x_i) dx$

The variance is always a positive quantity, and the standard deviation $\sigma = \sqrt{V[X]}$.

3.1.3 The Mean and Variance of a Bernoulli Distribution

Recall the definition of a Bernoulli random variable,

$$p_X(x) = \begin{cases} p & x = 1 \\ (1-p) & x = 0 \end{cases}$$

We can write

$$\begin{aligned} E[X] &= \sum_x x_i p(x_i) \\ &= 1(p) + 0(1-p) \\ &= p \end{aligned}$$

The variance is defined as $E[(x-\mu)^2]$, so we have

$$\begin{aligned} V[X] &= \sum_x (x_i - p)^2 p(x_i) \\ &= (1-p)^2 p + (0-p)^2(1-p) \\ &= p(1-p) \end{aligned}$$

3.1.4 The Mean and Variance of a Geometric Distribution

Recall the definition of a geometric random variable,

$$p_X(x) = \begin{cases} p & x = 1 \\ (1-p)p & x = 2 \\ (1-p)^2 p & x = 3 \\ \dots & \text{other values of } x \end{cases}$$

We can write

$$\begin{aligned} E[X] &= p(1-p)^0 + 2p(1-p)^1 + \dots \\ &= p(1-p)^0 + p(1-p)^1 + p(1-p)^2 + \dots = 1 \\ &\quad + p(1-p)^1 + p(1-p)^2 + \dots = 1-p \\ &\quad + p(1-p)^2 + \dots = (1-p)^2 \end{aligned}$$

The critical insight here is that $2p(1-p)$ can be written as $p(1-p) + p(1-p)$, and $3p(1-p)^2$ can be written the same way, with each sub-series forming a geometric series. We then sum the resulting geometric series $1 + (1-p) + (1-p)^2 + \dots$, yielding $E[X] = \frac{1}{p}$.

The variance is defined as $V[X] = E[X^2] - E[X]^2$, which simplifies (after much algebra) to

$$V[X] = \frac{1-p}{p^2}$$

3.1.5 The Mean and Variance of a Binomial Distribution

Recall the definition of a binomial random variable,

$$p_X(x) = \begin{cases} (1-p)^{10} & x = 0 \\ \binom{10}{2} p^2 (1-p)^8 & x = 2 \\ \dots & \text{other values of } x \in [0, 10] \cap \mathbb{N} \end{cases}$$

We can write

$$\begin{aligned} E[X] &= \sum_x x_i p(x_i) \\ &= 0p(0) + 1p(1) + 2p(2) + \dots \\ &= 1\left(p(1-p)^{n-1}\binom{n}{1}\right) + 2\left(p^2(1-p)^{n-2}\binom{n}{2}\right) + \dots \\ &= np\left(\binom{n-1}{0}(1-p)^{n-1} + \binom{n-1}{1}p(1-p)^{n-2} + \binom{n-1}{n-1}p^{n-1}\right) \\ &= np((p+1)-p)^{n-1} \\ &= np \end{aligned}$$

Note that a binomial random variable is really just a bunch of independent binomial random variables added up. So we can sum the variances to obtain $V[X] = np(1-p)$. In retrospect, we could have done the same for $E[X]$.

4 August 30, 2016

4.1 Wrapping up Means and Variances

Recall the following facts from last time. Bernoulli: $E[X] = p$, $V[X] = p(1-p)$. Binomial: $E[X] = np$, $V[X] = np(1-p)$. Geometric: $E[X] = 1/p$, $V[X] = (1-p)/p^2$.

4.1.1 The Mean and Variance of a Poisson Distribution

We may go over this in the future. For now, note that $E[X] = V[X] = \lambda$.

4.1.2 The Mean and Variance of a Exponential Distribution

Here, we have $E[X] = 1/\lambda$ and $V[X] = 1/\lambda^2$. The derivation for $E[X]$ is as follows, noting that $p(x) = \lambda e^{-\lambda x}$. So,

$$\begin{aligned} E[X] &= \int_0^{+\infty} x\lambda e^{-\lambda x} dx \\ &= \lambda\left[\frac{-x}{\lambda}e^{-\lambda x}\Big|_0^{\infty} + \frac{1}{\lambda}\int_0^{+\infty} e^{-\lambda x} dx\right] \\ &= 0 + \left(\frac{-1}{\lambda}\right)e^{-\lambda x}\Big|_0^{\infty} \\ &= \frac{1}{\lambda} \end{aligned}$$

4.1.3 The Mean and Variance of a Gaussian Distribution

Remember that the PDF is

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

where μ and σ are parameters that can be treated as constants. The integration involves the error function, and yields $E[X] = \mu$ and $V[X] = \sigma^2$.

4.2 Linearity of Expectation

The central idea is that, if one has properties of random variables X and Y, it would be nice to also know properties about the sum $X + Y$.

Theorem 1 (Linearity of Expectation)**.** $E[X + Y] = E[X] + E[Y]$.

Proof. This statement is the same as

$$\sum_x \sum_y (x + y) p(X = x, Y = y)$$

where the probability function is certainly not conditional, but may not be represented as the product of both individual probabilities as no assumptions are made of independence. This may be rewritten as

$$\sum_x \sum_y x p(X = x, Y = y) + \sum_x \sum_y y p(X = x, Y = y)$$

which is equivalent to

$$\sum_x x \sum_y p(X = x, Y = y) + \sum_y \sum_x y p(X = x, Y = y)$$

Note that, in the left sum indexed by y, the probabilities of y sum to 1 so the resulting value is simply $p(X = x)$. The same is true for the right sum, yielding

$$\sum_x x p(X = x) + \sum_y y p(Y = y)$$

which completes the proof. □

Note that this result is not true for the variances; i.e. $V[X + Y] = V[X] + V[Y]$. The linearity of variances is only true when X and Y are independent (really, uncorrelated).

Example 4. Let $Y = aX + b$. We have $E[Y] = aE[X] + b$. The linear shift does not affect variance, and so our resulting variance is $V[Y] = a^2 V[X]$.

Example 5. You have an urn with three types of balls: 8 blue, 4 white, and 2 orange. Choose two balls at random. You win two dollars for every blue ball, lose one dollar for every white ball, and are not changed by an orange ball. Determine all values of X, your winnings, the probability of X, and $E[X]$.

A listing of possible winnings is as follows: 4, 2, 1, 0, -1, -2. $P(4) = \binom{8}{2}/\binom{14}{2}$. $P(2) = \binom{8}{1}\binom{2}{1}/\binom{14}{2}$. $P(1) = \binom{8}{1}\binom{4}{1}/\binom{14}{2}$. $P(0) = \binom{2}{2}/\binom{14}{2}$. $P(-1) = \binom{2}{1}\binom{4}{1}/\binom{14}{2}$. $P(-2) = \binom{4}{2}/\binom{14}{2}$. Multiply each value with its probability to obtain $E[X] = 12/7 \approx 1.71$.

Example 6. One of the numbers $(1-10)$ is randomly chosen. You are to try to guess the number by asking yes or no questions. Compute the expected number of questions if (a) your ith question is "is it i?", and (b) with each question, you try to eliminate $1/2$ of the remaining as nearly as possible.

(a) The expected value is

$$E[X] = \frac{1}{10} \cdot 1 + \frac{1}{10} \cdot 2 + \frac{1}{10} \cdot 3 + \cdots = 5.5$$

(b) Draw a decision tree (for example, first split has left branch $1, 2, 3, 4, 5$ and left branch has $6, 7, 8, 9, 10$). Count the probability for each number of moves and obtain

$$E[X] = 4 \cdot \frac{4}{10} + 3 \cdot \frac{6}{10} = 3.4$$

Example 7. Suppose that the average numbers of cars abandoned in a week on a certain highway is 2.2. Approximate the probability that there will be (a) no abandoned cars this week and (b) at least two abandoned cars in the next week.

(a) Use a Poisson distribution. $E[X] = \lambda = e^{-2.2}$.

(b) This is expressed as $1 - p_X(x = 1) - p_X(x = 0)$ which equals $1 - 2.2e^{-2.2} - e^{-2.2}$.

5 September 1, 2016

This class was primarily reviewing Problemset 2.

Theorem 2 (Bayes)**.** $P(A|B) \cdot P(B) = P(A\&B) = P(A \cap B) = P(B|A) \cdot P(A)$.

6 September 6, 2016

We will first go over some problems, and then talk more about information sources and whether it makes sense to model them as random variables.

6.1 Review Problems

Example 8. X is a random variable with probability with probability distribution function. Let

$$f(x) = \begin{cases} c(1-x^2) & -1 < x < 1 \\ 0 & x = \text{anything else} \end{cases}$$

(a) What is c? $\int f(x)dx = 1$, so we have $(cx - \frac{cx^3}{3}|_{-1}^{1} = 1$, and $c = 3/4$.
(b) Recall the definition of $E[X] = \int xf(x)dx$. So, we have $E[X] = \int_{-1}^{1} x(3/4)(1-x^2)dx = 0$.

Example 9. The PDF of lifetime of a certain type of electronic device is

$$f(x) = \begin{cases} 10/x^2 & x > 10 \\ 0 & x \leq 10 \end{cases}$$

(a) What is $P(X > 20)$? Integrate from 20 to infinity and solve to obtain $1/2$.
(b) What is the probability that, of 6 such types of devices, at least three will function at least 15

hours? State assumptions. Assume that each case is independent. We have $P(X > 15) = 2/3$. So our expression for exactly 3 is

$$1 - P(0) - P(1) - P(2)$$

which equals

$$1 - \binom{6}{0}(1/3)^6 - \binom{6}{1}(1/3)^5(2/3) - \binom{6}{2}(1/3)^4(2/3)^2$$

Example 10. Find $E[X]$ of the following
(a)

$$f(x) = \begin{cases} 5/x^2 & x > 5 \\ 0 & x \leq 5 \end{cases}$$

(b)

$$f(x) = \begin{cases} (x/4)e^{-x/2} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

(a) is 0, (b) is 4 (integrate by parts twice)

6.2 Information Sources

Can these sources be modeled as a system with no input and one output and output emits symbols probabilistically?

Type	Source (Black Box)	Data	Characteristics of PDF	Can you send less?
e-mail	computer	text	frequency of letters, greetings, formatting	Possible
telephone	receiver	sound	decibel range, tone, background noise, pauses	Possible (similar to email)
images	camera	pixels	background, same pixel values for large areas	Possible
video	camera	series of images	stationary objects in video	Possible

7 September 8, 2016 (Andrew)

7.1 Modeling English Probabilistically

Today, we will go through the same thought experiment that Claude Shannon performed. Looking at English text, Shannon first analyzed whether it could be modeled probabilistically.

Zeroth Order. A simplistic zeroth order approximation involves assuming uniform distribution across all letters. This may be conducted by rolling a 27 sided die. A sample output might look like *XFOML RXKHRJFFJUJ ZLPWCFWKCYJ FFJEYVKCDSGHYD QPAAMKBZAACCIB-ZLHJQD.*

First Order. A first order approximation leverages the probabilities of certain letters. For example, $P_X(E) \approx 0.13$ and $P_X(W) \approx 0.02$. A sample output might look like *OCRO HLI RGWR NMIELWIS EU LL NBNESBEYA THEEI.*

Second Order. A second order approximation utilizes the probability of a letter appearing given the preceding letter. A sample output might look like *ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY ACAIW D ILONONASIVE.*

***n*th Order.** We can extend this out to an nth order approximation, which uses information about the preceding $n-1$ letters to predict the value of the nth character. The larger the value of n, the more like English the model's output becomes.

Shannon assumed that all sources are **ergodic**. Ergodic sources are those whose distributions do not change with time. While this may not hold true universally, it is a reasonable approximation for most practical applications of his work.

Example 11. The time (in hours) required to repair a machine is an exponentially distributed random variable with $\lambda = \frac{1}{2}$.
(a) What is P(repair time > 2hrs)? $P(X > 2) = 1 - \int_0^2 \frac{1}{2} e^{-\frac{1}{2}x} dx$
(b) What is P(repair time > 10 | repair time > 9)? $\mathrm{P(X > 10 \mid X > 9)} = \frac{P(X>10)}{P(X>9)} = \frac{e^{-5}}{e^{-4.5}}$

7.2 Memoryless Property

A random variable is said to be **memoryless** if information about the past does not have an impact on future results. For instance, in the case of Example 17 (b), the probability of waiting 10 hours given that you've waited 9 hours is equivalent to the probability of waiting for 1 hour. The exponential distribution is the only continuous random variable and geometric is the only discrete one. Formally stated:

$$P(X > m + n | X > n) = P(X > m)$$

Example 12. Say that two people, Ms. Jones and Mr. Brown, are being serviced by two mail attendants whose time to completion is an exponential random variable. Mr. Smith is standing behind both. What is the probability that he will be the last person to leave the post office? 0.5 because the distribution is memoryless.

Example 13. The speed of a particle can be modeled with

$$f(X) = \begin{cases} ax^2 e^{-bx^2} & x >= 0 \\ 0 & x < 0 \end{cases}$$

Where $b = \frac{m}{2kT}$. What value of a satisfies this pdf?
$\int_0^\infty ax^2 e^{-bx^2} dx = \frac{\sqrt{\pi} a}{4b^{3/2}}$
$\frac{\sqrt{\pi} a}{4b^{3/2}} = 1$
$a = \frac{4b^{3/2}}{\sqrt{\pi}}$

Example 14. Given that $Z \sim N(0, 1)$, and $g(x)$ is differentiable, prove that

$$E[g'(Z)] = E[Z * g(Z)]$$

We'll finish this example next time.

8 September 13, 2016

Quick review of Bayes' theorem—we always have

$$P(A|B) = \frac{P(A\&B)}{P(B)}$$
$$P(B|A) = \frac{P(B\&A)}{P(A)}$$

If A and B are independent, $P(A\&B) = P(A)P(B)$, as $P(B|A) = P(B)$ and $P(A|B) = P(A)$.

8.1 Definition of Information

Our goal is to motivate the definition of a quantity called entropy. Consider a large stream of output. We are already convinced that this stream can be modeled as a random variable. Now, we will determine how much information is in the stream.

Example 15 (Motivation)**.** Consider the letter "q". We are asked to guess the next letter; an obvious guess is "u". If we are told the letter is indeed "u", there is not much information gain (we are not very surprised). If, however, the letter is "w", significantly more information is revealed (and we are surprised as a result). We would therefore like to identify a function that can represent this idea.

In line with this idea, let's try different functions. $f(p) = 1/p$ and $f(p) = 1 - p$ are both monotonically decreasing functions of p that represent an increasing value with a decreasing probability. For a stream of letters, however, this becomes a problem. Additionally, we would like to think about this idea in terms of information (is there more or less information?). To resolve this issue, we will consider the concept of entropy.

8.2 Entropy

Definition 4 (Entropy)**.** Let X be a random variable.

$$H(X) = \sum_x p(x) \log\left(\frac{1}{p(x)}\right)$$

The entropy $H(X)$ represents the average amount of information of variable X. From this definition, it is easy to see that we have $f(p) = \log(1/p(x))$ which is actually quite similar to our initial discussion. Note that this is equivalent to

$$E\left[\log \frac{1}{p(x)}\right]$$

which is why we refer to entropy as "average information". In both cases, the logarithm is base 2 as we refer to information in bits.

Example 16. What is the entropy of a fair coin?

$$\begin{aligned} H(X) &= \sum_x p(x) \log\left(\frac{1}{p(x)}\right) \\ &= \frac{1}{2}\log\frac{1}{1/2} + \frac{1}{2}\log\frac{1}{1/2} \\ &= 1 \end{aligned}$$

Example 17. What is the entropy of a biased coin? Note that we expect $H(X)$ to be smaller in this case as per our previous discussion. Let the coin have probability of heads 0.9.

$$\begin{aligned} H(X) &= \sum_x p(x) \log\left(\frac{1}{p(x)}\right) \\ &= \frac{1}{10}\log\frac{1}{1/10} + \frac{9}{10}\log\frac{1}{9/10} \\ &= 0.47 \end{aligned}$$

Example 18. What is the entropy of a biased coin with probability p? Note that we expect $H(X)$ to be smaller in this case as per our previous discussion. Let the coin have probability of heads p.

$$\begin{aligned} H(X) &= \sum_x p(x) \log\left(\frac{1}{p(x)}\right) \\ &= p\log\frac{1}{p} + (1-p)\log\frac{1}{1-p} \end{aligned}$$

As expected, this represents a parabola with $\frac{dH(X)}{dx} = 0$ at $x = 0.5$. One may plot this distribution on Mathematica with the command `Plot[-p Log[2, p] - (1 - p) Log[2, 1 - p], p, 0, 1]`.

Remark 1. If all symbols occur with equal probability, $H(X)$ will simply be $\log m$ bits where m is the number of possible values. Equivalently, $2^{H(X)} = m$ equally probable values. So, transmitting information only requires m bits for each value X.

8.3 Quiz Review

Example 19. X is a binomial random variable with expected value 6 and variance 2.4. Find $P\{X = 5\}$.

Example 20. Suppose two 4-sided dice have sides numbered 1-4. Let X denote the sum of the dice. Find all values of X, the PMF of X, and $E[X]$.

Example 21. X is a random variable having possible values of 0 and 1 such that $P(X = 0) = 2P(X = 1)$. Find the PMF, expectation, and variance of X.

Example 22. A local soccer team has 5 more games to play. If it wins, they will play its final 4 games in upper bracket and win each game in this bracket is 0.4. If it plays in lower bracket, the probability of winning is 0.7. Given that the probability that it wins the game this weekend is 0.5, what is the probability that it wins at least 3 of its final 4 games?

Example 23. Let pdf of X be

$$f(X) = \begin{cases} a + bx^2 & 0 \le x \le 1 \\ 0 & o/w \end{cases}$$

If $E[X] = \frac{3}{5}$, find a and b. All solutions are left to the reader.

9 September 14, 2016

Quiz today; no notes.

10 September 16, 2016

Problemset for Thursday is seven problems on entropy. This is a strictly individual assignment, but collaboration is allowed so long as each student has a unique solution.

10.1 Review of Entropy

We have so far defined the term "entropy" and done problems on fair and biased coin. With m equally likely symbols, $H(X) = \log_2 m$ and $m = 2^{H(X)}$.

Example 24. The entropy of a fair six-sided die is simply $\log_2 6$. Know how to solve this problem with a biased die (different faces have different probabilities).

10.2 Properties of $H(X)$

From our previous discussion, we have the following properties for $H(X)$

- $H(X) \geq 0$
- $H_b(X) = (\log_b a)H_a(X)$, where $H_q(X)$ is $H(X)$ using a logarithm of base q
- $H(X)$ does not depend on outcome, only on p's
- $E[X] \geq H(X)$

Note that "nat" (using the natural logarithm) is equivalent to $1/\ln 2$ bits.

Example 25. Let X be a geometric random variable with $p = 1/2$. It is useful to know that

$$\sum_{n=1}^{\infty} r^n = \frac{r}{1-r}$$
$$\sum_{n=1}^{\infty} nr^n = \frac{r}{(1-r)^2}$$

The solution is straight from the definition of entropy; the reader can confirm $H(X) = 2$.

Example 26. Ask "efficient" questions of the form "Is X in the set $\mathcal{S}$" to identify what the element X is (think binary search), where X is a geometric random variable with $p = 1/2$. Determine what your questions will look like, the expected number of questions, and the expected number of questions of $H(X)$.

By the memoryless property, it is equally likely that your guess will be correct or incorrect (if you say "is X 1", the chance is $1/2$, and the chance of the rest is $1/2$. Similarly, if you say "is X 2?", the chance is again $1/2$–there is no way to optimize this method). The expected number of questions $E[\text{Questions}] = \sum(\text{question \#})P(\text{question \#})$. So, we have $1(1/2) + 2(1/4) + 3(1/8) + \cdots$. We can follow the "sum of sums" approach delineated previously in the proof of the mean and variance of a geometric distribution to obtain 2. As 2 is also the entropy of such a geometric random variable, we are sure that this is the best method; this follows from $E[X] \geq H(X)$.

Theorem 3 (Jensen's Inequality)**.** If f is a concave up function ($f'' > 0$) on a given region and $a + b = 1$, we have

$$af(x_1) + bf(x_2) \geq f(ax_1 + bx_2)$$

Flip the inequality if f is concave down ($f'' < 0$). A more general form of this is

$$\sum c_i f(x_i) \geq f(\sum c_i x_i)$$

where $\sum c_i = 1$. Again, flip the inequality if f is concave down.

Theorem 4. Let X assume values $\{x_1, x_2, x_3, x_4, \ldots, x_r\}$. We may also denote this as "let the PMF of X be an r-dimensional vector $< p_1, p_2, \ldots, p_r >$". We have that (a) $0 \leq H(X) \leq \log r$, (b) $H(X) = 0$ if and only if $p_i = 1$ for some i, and (c) $H(X) = \log r$ if and only if $p_i = 1/r$ for all i.

Proof. (a) Jensen's inequality allows us to write

$$H(X) = \sum_x p(x) \log\left(\frac{1}{p(x)}\right) \leq \log \sum p_i(1/p_i) = \log r$$

(b) If $p_i = 1$, all of the other probabilities must be zero. Therefore, $H(X) = \log 1 = 0$. To prove the inverted statement, use the idea that $p(x) \log \frac{1}{p}$ can never be negative. In fact, $p(x) \log \frac{1}{p}$ is only ever zero when $p = 1$. Since every other p would produce a positive value for entropy, the only vector p that will satisfy $H(X) = 0$ is where one p_i is 1 and the rest are 0.

(c) For the first part of the if and only if, $p_i = 1/r$, the operand of the sum can be taken out of the sum, simply leaving $\log r$. We can also write $2^{H(X)} = r$ which implies that there are r equally likely outcomes. This necessitates that each outcome has probability $p_i = 1/r$. □

11 September 20, 2016

11.1 Differential Entropy

We will discuss how to define entropy for continuous random variables.

Theorem 5. The differential entropy $h(x) = H_{\text{diff}}(X) = h(f) = \int f(x) \log(1/f(x)) dx$.

Proof. Define a random variable X^Δ with a pmf that has bins of arbitrary size Δx. If we define x_i to be the event that X is in bin i, we could write the entropy of X^Δ as

$$H(X^\Delta) = \sum_i P(x_i)\Delta x \log \frac{1}{P(x_i)\Delta x}$$
$$H(X^\Delta) = \sum_i P(x_i)\Delta x \log \frac{1}{P(x_i)} + P(x_i)\Delta x \log \frac{1}{\Delta x}$$
$$H(X^\Delta) = \sum_i P(x_i) \log \frac{1}{P(x_i)} \Delta x + \sum_i P(x_i)\Delta x \log \frac{1}{\Delta x}$$

Let $\Delta x \to 0$:

$$H(X^\Delta) = \boxed{\int f(x) \log \frac{1}{f(x)} dx} + \infty$$

□

We define the quantity $\int f(x) \log \frac{1}{f(x)} dx$ to be the differential entropy of the pmf f. The quantity ∞ is referred to as the precision factor. From this, we determine that the total entropy of the variable increases as the measurements of the variable become more precise.

Notation 2. We write the differential entropy as either $h(X)$, $H_{\text{diff}}(X)$, or $h(f)$.

11.2 Properties of Differential Entropy

Theorem 6. If $Y = X + c$, $h(Y) = h(X)$

Proof. Left to the reader.

Theorem 7. If $Y = cX$, $h(Y) = h(X) + \log|c|$

Proof. Let f be the distribution of X and g be the distribution of Y. Then

$$h(Y) = \int g(y) \log \frac{1}{g(y)} dy$$

With $y = cx$ and $dy = dx$, we can write the integral as

$$h(Y) = \int c\, g(x) \log \frac{1}{g(cx)} dx$$

But, since we have $g(cx) = f(x)/c$, this resolves to

$$h(Y) = \int f(x) \log \frac{c}{f(x)} dx$$

which is simply $h(X) + \log|c|$. □

Remark 2. Note that, although Theorem 7 seems to indicate that more information is encoded with a simple multiplication by a constant factor, it turns out that the multiplication increases the number of bins, resulting in precision error.

11.2.1 Entropy of a Uniform Distribution

Theorem 8. If $X \sim \mathcal{U}(0, a)$, then the differential entropy $h(X) = \log a$

Proof. We have

$$h(X) = \int f(x) \log\left(\frac{1}{f(x)}\right) dx$$

which can be simplified to

$$\int_0^a \frac{1}{a} \log(a) dx = \log a$$

□

Remark 3. In the discrete case (and with $a \to 2a$ in the uniform distribution), if the number of outcomes is doubled, the entropy H goes up by one bit. Note also that the entropy can be 0 or negative.

- if $a = 1, h(x) = 0$
- if $a = 2, h(x) = 1$ (it takes one more bit than $\mathcal{U}(0, 1)$)
- if $a = 1/2, h(x) = -1$ (it takes one less bit than $\mathcal{U}(0, 1)$)

Theorem 9. Given a fixed lower bound and upper bound, no PDF can have a larger entropy than a uniform distribution.

Proof. Heuristically, the probability density function on $\{x_1, x_2, .., .x_n\}$ with maximum entropy turns out to be the one that corresponds to the least amount of knowledge of $\{x_1, x_2, .., .x_n\}$, in other words the uniform distribution. For a more formal proof consider the following:

A probability density function on $\{x_1, x_2, .., .x_n\}$ is a set of nonnegative real numbers $p_1, ..., p_n$ that add up to 1. Entropy is a continuous function of the n-tuples $(p_1, ..., p_n)$, and these points lie in a compact subset of $\mathbb{R}^n$, so there is an n-tuple where entropy is maximized. We want to show this occurs at $(1/n, ..., 1/n)$ and nowhere else.

Suppose the p_j are not all equal, say $p_1 < p_2$. (Clearly $n \neq 1$.) We will find a new probability density with higher entropy. It then follows, since entropy is maximized at some n-tuple, that entropy is uniquely maximized at the n-tuple with $p_i = 1/n$ for all i. Since $p_1 < p_2$, for small positive ε we have $p_1 + \varepsilon < p_2 - \varepsilon$. The entropy of $\{p_1 + \varepsilon, p_2 - \varepsilon, p_3, ..., p_n\}$ minus the entropy of $\{p_1, p_2, p_3, ..., p_n\}$ equals

$$-p_1 \log\left(\frac{p_1 + \varepsilon}{p_1}\right) - \varepsilon \log(p_1 + \varepsilon) - p_2 \log\left(\frac{p_2 - \varepsilon}{p_2}\right) + \varepsilon \log(p_2 - \varepsilon)$$

To complete the proof, we want to show this is positive for small enough ε. Rewrite the above equation as

$$-p_1 \log\left(1 + \frac{\varepsilon}{p_1}\right) - \varepsilon\left(\log p_1 + \log\left(1 + \frac{\varepsilon}{p_1}\right)\right) - p_2 \log\left(1 - \frac{\varepsilon}{p_2}\right) + \varepsilon\left(\log p_2 + \log\left(1 - \frac{\varepsilon}{p_2}\right)\right)$$

Recalling that $\log(1 + x) = x + O(x^2)$ for small x, the above equation is

$$-\varepsilon - \varepsilon \log p_1 + \varepsilon + \varepsilon \log p_2 + O(\varepsilon^2) = \varepsilon \log(p_2/p_1) + O(\varepsilon^2)$$

which is positive when ε is small enough since $p_1 < p_2$.

□

11.2.2 Entropy of an Exponential Distribution

Example 27. Find $h(X)$ for the function

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

It is easier to replace the log with a ln in the definition of entropy. Integrating by parts, we obtain

$$h(X) = 1 + \ln\frac{1}{\lambda} = \ln\frac{e}{\lambda} = \log\frac{e}{\lambda} \text{ bits}$$

Theorem 10. Given a random variable with mean μ and unconstrained otherwise, the exponential distribution achieves maximum entropy. Recall that the expected value for an exponential distribution is $1/\lambda$.

12 September 22, 2016

As a review, we defined differential entropy as

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

and we discussed two distributions: the uniform distribution $\mathcal{U}(0, a)$ which has a differential entropy of $\log a$ bits and the exponential distribution which has a differential entropy $h(X) = \log(\frac{e}{\lambda})$.

12.1 Entropy of a Gaussian Distribution

Theorem 11. $h(X)$ for the function

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

where μ and σ are parameters that can be treated as constants is $\frac{1}{2}\log 2\pi e\sigma^2$.

Proof. We will start with the definition of entropy

$$\begin{aligned} h(X) &= \frac{1}{\sigma\sqrt{2\pi}} \int e^{\frac{-(x-\mu)^2}{2\sigma^2}} \log\left(\frac{1}{\frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}}\right) dx \\ &= \frac{-1}{\sigma\sqrt{2\pi}} \int e^{\frac{-(x-\mu)^2}{2\sigma^2}} \left[\log \frac{1}{\sigma\sqrt{2\pi}} + \log e^{\frac{-(x-\mu)^2}{2\sigma^2}}\right] dx \end{aligned}$$

The first additive term of the integral resolves to $\log \sigma\sqrt{2\pi}$ as it is simply the PDF (which integrates to 1). The second term is written as

$$\frac{1}{\sigma\sqrt{2\pi}} \int e^{\frac{-(x-\mu)^2}{2\sigma^2}} \cdot \frac{(x-\mu)^2}{2\sigma^2} dx$$

which can be shown to resolve to the variance of $(x-\mu)$. Some algebra resolves this to $1/2$ in nats or $1/2\ln 2$ in bits. The overall entropy is therefore $\frac{1}{2}\log 2\pi e\sigma^2$. □

13 September 26, 2016

13.1 Source Coding

We will be discussing source coding. Our goal is to remove redundancy from the source and code it "efficiently" and "losslessly". For the purposes of our discussion today, we will be using the English alphabet.

Example 28. Given the 26 letters of the English alphabet and their probabilities (assume letters are independent), find a way to encode the distribution (convert it into 0s and 1s) and decode the distribution (return it to the original state). If we decide to use 5 bits for each letter, it is easy to cover the entire alphabet (there are 32 combinations).

Example 29. Although the first encoding is acceptable, we would next like to associate the length of the code with the probability of the symbol. Given that the most probable letter is an "E", we can send a 0, and the second probable letter ("A") can be encoded with a 1. The next letter, say "T", we can send 01, and so on. There is a problem here, however: we don't have a stop coding, so it is impossible to differentiate the end of a string and the start of another one.

Our general issue is that *no code can be a prefix of another code.* Our two conditions are as follows:

- *Variable Length Codes.* Assign smaller codes to larger probability symbols
- *Prefix "Free" Codes.* No letter can be a prefix of another letter's code (instantaneous codes)

Remark 4. A code is a prefix code if and only if it can be visualized with a tree.

Example 30. Consider the code mapping E $\to$ 0, T $\to$ 10, A $\to$ 110, and O $\to$ 111. This can certainly represented by a tree with binary splits, and therefore works with these four letters.

14 September 28, 2016

14.1 Huffman Coding

As we discussed previously, our coding mechanism relies on two principles: frequent symbols have shorter code words and that our compression achieved is as close to the maximum entropy close to the entropy of the source (within 1 bit).

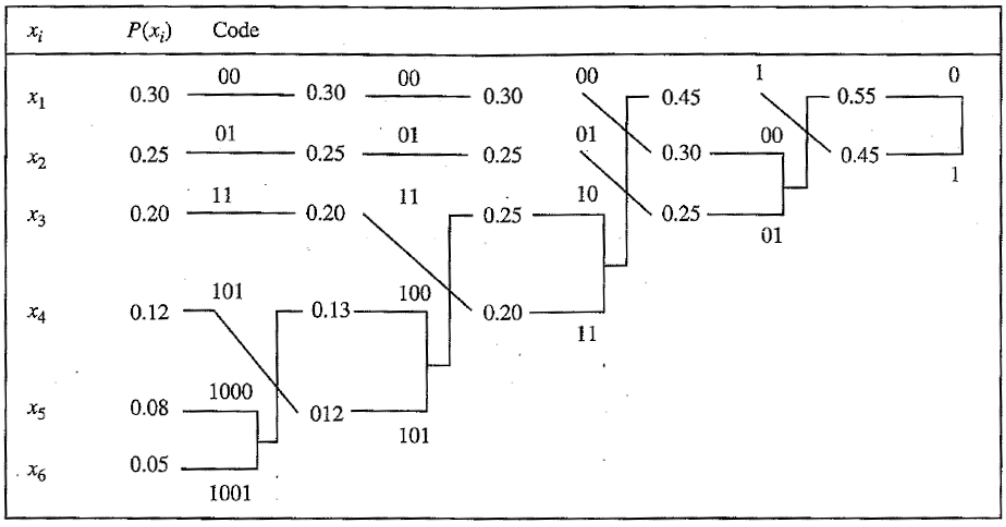
The Huffman coding algorithm can be written as follows.

- Find the two least probable symbols
- Combine these symbols to make an imaginary composite symbol
- Replace the two least probable symbols with the probability as the sum of the components
- Repeat the process until you have 1 symbol left

The decoding process is written as

- The code for each symbol is obtained by concatenating binary numbers that lead to a symbol
- Each time two symbols are combined, two paths are labelled with a 1 (upper) and 0 (lower)
- Wash, rinse, repeat.

Example 31. The following details a sample Huffman encoding for sample $x_1 \dots x_6$.



Example 32. Given a list $X \in \{1, 2, 3, 4, 5\}$ with probabilities $\{0.25, 0.25, 0.2, 0.15, 0.15\}$. Identify the expected value of the length and the entropy $H(X)$. With the optimal Huffman code, $E[\text{Length}] = 2.3$ bits and $H(X) = 2.29$ bits.

Example 33. The Huffman code works for n-ary trees. In particular, the same example above with a ternary division yields $E[\text{Length}] = 1.5$ ternary digits. Ternary trees may require an "additional" value with probability 0 to balance the tree–insert this at the beginning to include it in overall computation.

Example 34. Which of these cannot be Huffman codes for any probability assignment? (a) $\{0, 10, 11\}$, (b) $\{00, 01, 10, 110\}$, (c) $\{01, 10\}$. (b) and (c) cannot be codes.

14.2 Huffman and 20 Questions

Let $X_i = 1$ or 0 if i-th object is good or defective, respectively. Let $X_i, X_2, \ldots, X_n$ be independent with $\Pr\{X_i = 1\} = p_i$. Also note that $p_1 > p_2 > \cdots > p_n > 1/2$. The goal is to determine the set of all defective objects. (a) Give a good lower bound on the minimum average number of questions.

15 October 3, 2016

15.1 Lagrange Optimization

Given a $f(x, y)$ or $f(x, y, z)$, we would like to determine the maximum and minimum of f subject to a constraint $g(x, y)$ or $g(x, y, z)$. We can write

$$\mathcal{L}(x, y, z) = f - \lambda g$$

of which we can find the critical points of $\mathcal{L}$ and determine whether those points are minima or maxima. (We'll discuss the intuition behind this later.) These critical points would be found by taking $\partial\mathcal{L}/\partial x, \partial\mathcal{L}/\partial y, \partial\mathcal{L}/\partial z, \partial\mathcal{L}/\partial\lambda$ and setting them equal to 0.

Example 35 (Lagrange)**.** Find the min and max values of $f(x, y, z) = x^2y^2z^2$ constrained by $x^2 + y^2 + z^2 = 1$. We have

$$f_x = 2xy^2z^2$$
$$f_y = 2yx^2z^2$$
$$f_z = 2zx^2y^2$$

and similarly $g_x = 2x$, $g_y = 2y$, and $g_z = 2z$. So, we have $f_x = \lambda g_x \implies 2xy^2z^2 = \lambda 2z$. We also have $f_y = \lambda g_y \implies 2yx^2z^2 = \lambda 2y$ and $f_z = \lambda 2z \implies 2zx^2y^2 = \lambda 2z$. Solving for lambda in each equation yields $\lambda = y^2z^2, \lambda = x^2z^2$, and $\lambda = y^2x^2$. These equations may be combined into the constraint (noting that $x^2 = y^2 = z^2$) so we have $3z^2 = 1$. So $z = \pm 1/\sqrt{3} = x = y$ for the maximum (where $f(x, y, z) = 1/27$). The minimum is obtained by setting one of x, y, or z to 0 and the other two to fit the constraint (in which case $f(x, y, z) = 0$).

Example 36. $f(x, y) = xy^2$ on the circle $x^2+y^2 = 1$. The solution is $x = \pm 1/\sqrt{3}$ and $y = \pm\sqrt{2}/\sqrt{3}$. These coordinates must be evaluated in f to obtain minima and maxima.

Example 37. $f(x, y) = 3x + y$ constrained on $x^2 + y^2 = 10$. The solution is $x = \pm 3$ and $y = \pm 1$. Again, these coordinates must be evaluated to in f to determine minima and maxima.

15.2 Review of Huffman Coding

Consider X with PMF $\{1/21, 2/21, 3/21, 4/21, 5/21, 6/21\}$. Find the (a) binary and (b) ternary Huffman coding, and determine the lengths of each. (a) and (b) are left to the reader; the binary expected length is 2.42 and the ternary expected length is 1.62.

16 October 5, 2016

16.1 Proof of Source Coding Theorem: Part 1 (LB: $H[X]$, UB: $H[X]+1$)

We will show that the expected length of an optimal prefix encoding n-ary tree is between $H[X]$ and $H[X]+1$ (it has a lower bound of the entropy and the upper bound of (entropy + 1)).[1]

Lemma 1 (Kraft Inequality)**.** For any prefix code, the code word lengths $l_1, l_2, \ldots, l_m$ must satisfy

$$\sum_i 2^{-l_i} < 1$$

Conversely, given a set of code word lengths with this property, then there is a code with these lengths.

Proof. For the first half of the Lemma, imagine a unit square. One may partition this square an arbitrary number of times, yet the area will remain constant (2^{-l_i} is the area of a sub-square in the larger square). For the converse: given a set of code word lengths which satisfy Kraft, label the first node (lexicographically such that any '0' comes before any '1')[2] of depth l_1 as code word 1. Remove all descendants from this node (so that no code is a prefix of another code) and label the first remaining node of depth l_2 as code word 2. Continuing this process will generate a prefix code with $l_1, l_2, \ldots, l_m$. This is essentially a description of how to construct a prefix tree. □

We have shown that any prefix code satisfies the Kraft inequality, and any code that satisfies the Kraft inequality is a prefix code. Our problem is now to find the prefix code with the minimum expected length. This is equivalent to finding a set of lengths that satisfy the Kraft inequality and whose expected length $L = \sum p_i l_i$ is less than any other prefix code. We must therefore minimize $L = \sum p_i l_i$ subject to the constraint $\sum 2^{-l_i} \leq 1$. We will neglect the integer constraint on l_i and assume equality in the constraint. Following our method of Lagrange optimization,

$$J = \sum p_i l_i + \lambda(\sum 2^{-l_i} - 1)$$

Taking the partial with respect to l_i, we have

$$\frac{\partial J}{\partial l_i} = p_i - \lambda 2^{-l_i} \ln 2$$

because $\sum p_i l_i = p_1 l_1 + p_2 l_2 + \ldots$ Setting the derivative equal to 0 yields

$$2^{-l_i} = \frac{p_i}{\lambda \ln 2}$$

which we may substitute into the constraint to find λ[3].

$$\sum \frac{p_i}{\lambda \ln 2} = 1$$
$$\lambda = \frac{1}{\ln 2}$$

[1] Our proof will discuss binary trees, but it can be shown that it generalizes to an n-ary tree.
[2] Here, '00' would come before '001'
[3] The optimal code is achieved with equality.

We may then substitute this back into our original equation to obtain $p_i = 2^{-l_i}$, so $l_i^* = -\log_2 p_i$. The optimal code (denoted with a $*$) is therefore

$$L^* = \sum p_i l_i^* = -\sum p_i \log_2 p_i = H[X]$$

In reality, $L^* \geq H[X]$. We have therefore proved the expected length L^* is bounded below. Since $\log_2 \frac{1}{p_i}$ may not be an integer, let $l_i = \lceil \log_2 \frac{1}{p_i} \rceil$. These lengths will still satsify Kraft since

$$\sum 2^{-\lceil \log_2 \frac{1}{p_i} \rceil} \leq 2^{-\log \frac{1}{p_i}} = \sum p_i = 1$$

This choice of code word lengths satisfies $\log_2 \frac{1}{p_i} \leq l_i \leq \log_2 \frac{1}{p_i} + 1$ due to the property of the ceiling. Multiplying by p_i and summing over i, we get

$$\sum p_i \log_2 \frac{1}{p_i} \leq p_i l_i < \sum \left[p_i \log_2 \frac{1}{p_i} \right] + \sum p_i$$

Analogously, $H[X] \leq L < H[X] + 1$. Since $L \geq L^*$, $H[X] \leq L^* < H[X] + 1$.

17 October 7, 2016

There are certainly many optimal codes, of which we will prove Huffman is one.

17.1 Proof of Source Coding Theorem: Part 2 (Optimality of Huffman)

Before we can show that Huffman is optimal, we must prove some properties of a particular optimal code.

Lemma 2. For any distribution, there exists an optimal prefix code with a minimum expected length that satisfies the following:

- If $p_j > p_k$, then $l_j \leq l_k$ (p denotes probability, l denotes length)
- Two largest code words code words have the same length
- Two longest code words differ only in the last bit and correspond to the least likely symbols

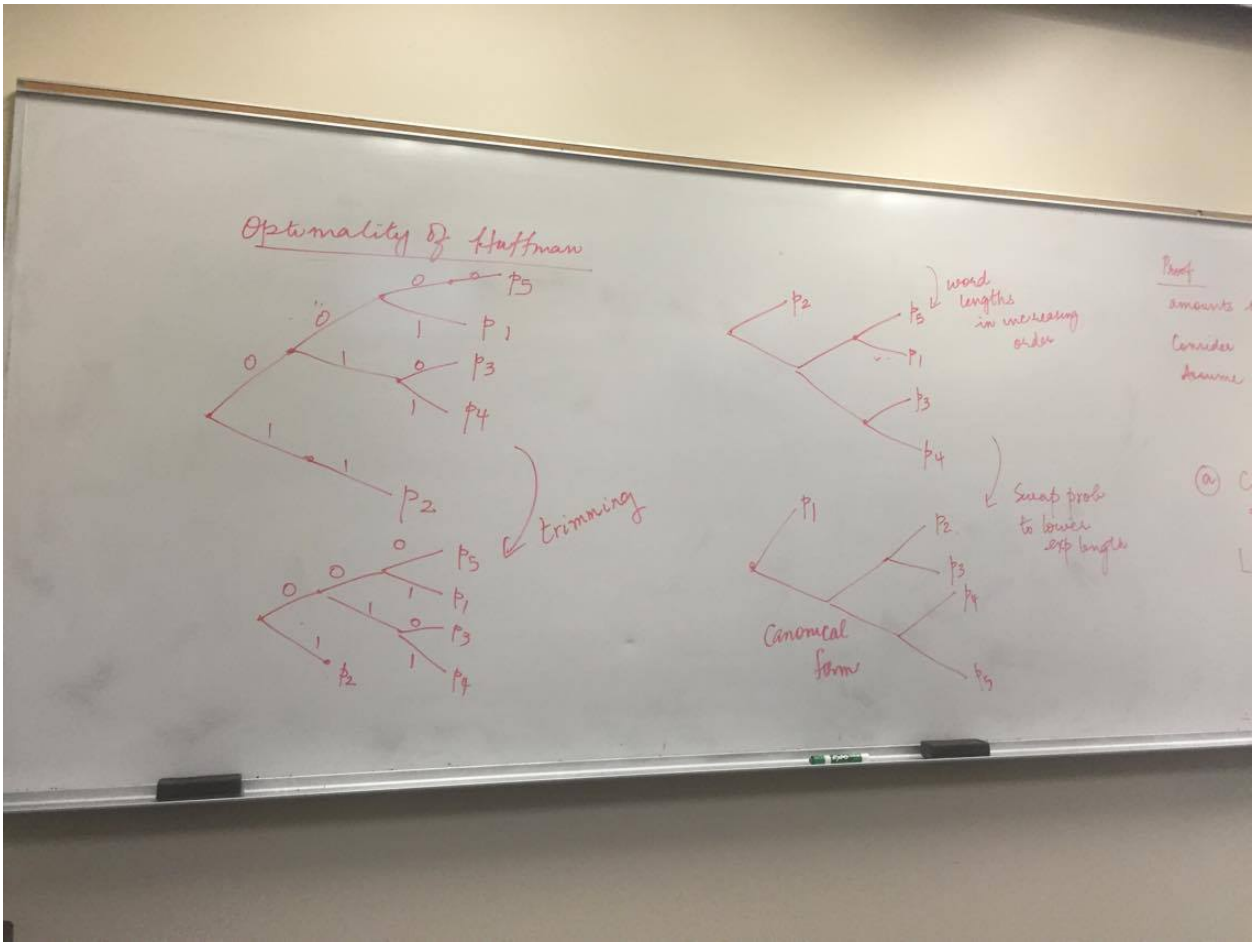
Proof. The proof really amounts to swapping, trimming, and rearranging code words. Consider an optimal code C_m with m symbols. Assume that $p_1 \geq p_2 \geq p_3 \geq \cdots \geq p_m$. As the code is optimal, $\sum_i p_i l_i$ is minimal (as close to $H[X]$ as possible, but certainly within the bounds $H[X]$ and $H[X] + 1$).

Consider a new code C'_m with code words j and k of C_m interchanged. The difference between the expected lengths of both code words $L(C'_m) - L(C_m) = \sum p_i l'_i - \sum p_i l_i$ reduces to $p_j l_k p_k l_j - p_j l_j - p_k l_k$, which equates to $(p_j - p_k)(l_k - l_j)$. By hypothesis, $(p_j - p_k) \geq 0$. Since this code is the "best" optimal code (as defined by the sum $\sum_i p_i l_i$ as minimal), $L(C'_m) - L(C_m) \geq 0$, and $l_k \geq l_i$.

Next, if the 2 longest code words are not of the same length, then one can delete the last bit of the longest one and still satisfy the prefix property.

Finally, if there is a maximum length code word without a sibling (a code that differs in the last bit), then we can delete the last bit of the code word and still satisfy the prefix property (similar to (b)). This reduces the expected length, contradicting the hypothesis that the prefix code has "minimum expected length" ($\sum p_i l_i$ is minimal). So, the two longest code words must differ only in the last bit—every maximum length code word in any optimal code has a sibling. □

Given a code tree, we can (a) trim it, (b) change the order so that the word lengths are in increasing order (the shorter ones are on the top), and (c) swap the probabilities to lower expected length. The final state is known as the "canonical form." Any tree can be converted to a canonical form by following this series of steps.



Theorem 12. The Huffman encoding is a best optimal code (there exist other codes that achieve the same minimum length, but there is no better code).

Proof. We will show by induction that, if there is an optimal code at every level of the tree, then the tree is optimal. Define C_m as the code for $\{p_1, p_2, \ldots, p_m\}$ such that $p_1 \geq p_2 \geq \cdots \geq p_m$. Define a merged code C_{m-1} as a code obtained by taking the two longest code words and alloting them to a symbol with probability $p_{m-1} + p_m$. The original expected length of our tree is

$$L(C_m) = \sum_{i=1}^{m} p_i l_i$$

This is equivalent to

$$L(C_m) = \sum_{i-1}^{m-2} p_i l'_i + p_{m-1}(l'_{m-1} + 1) + p_m(l'_{m-1} + 1)$$

where l' is the length of the tree without the last two elements. By pulling $p_x l_x$ terms into the sum, we can write this as

$$\sum_{i=1}^{m-1} p_i l'_i + p_{m-1} + p_m$$

The expected length is therefore equal to

$$L(C_m) = L(C_{m-1}) + p_m + p_{m-1}$$

Therefore, minimizing $L(C_m)$ is equivalent to minimizing $L(C_{m-1})$. We have reduced the problem to □

Appendix A – Random Variables

X	Pr[X=x]	E[X]	Var[X]	H[X]
Bernoulli	NA	p	$p(1-p)$	$-p\log p-(1-p)\log(1-p)$
Binomial	$\binom{n}{k}p^k(1-p)^{n-k}$	np	$np(1-p)$	NA
Geometric	$p(1-p)^{k-1}$	$\frac{1}{p}$	$\frac{1-p}{p^2}$	$\frac{-p\log p-(1-p)\log(1-p)}{p}$
Poisson	$\frac{\lambda^k e^{-\lambda}}{k!}$	λ	λ	NA
Uniform	$\frac{1}{b-a}$	$\frac{1}{2}(a+b)$	$\frac{1}{12}(b-a)^2$	$\log(b-a)$
Exponential	$\lambda e^{-\lambda x}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$\log\frac{e}{\lambda}$
Gaussian	$\frac{1}{\sqrt{2\sigma^2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2	$\frac{1}{2}\log 2\pi e\sigma^2$

Appendix B – Quotes

- "What's p, entropy?" (Shaya Zarkesh)
- "I can't do the math" (Swapnil Garg)
- "I still don't get it" (Swapnil)
- "Swapnil—remember when, in 8th grade, I made JMO and you didn't?" (Shaya)
  "I made USAMO." (Swapnil)
- "I'm just happy I'm not on the quotes page" (Neymika Jain)
- "How many fingers has a Martian?" (Text)
- "2, 4, 6, 8, who do we appreciate? Not info theory!" (Neymika)
- "I don't want to have as many [quotes] as Shaya..." (Neymika)
- "So the entropy is bounded by $H[X]$ and $H[X]+1$, right?" (Manan Shah)
- "Isn't swapping kind of like gargling? Like you can put the code words in your mouth and gargle them..." (Rose Guan)
- "Couldn't we make a pun with swapping and Swapnil?" (Neymika)
- "This is Cauchy-Schwartz" (Shaya)
  "No, it's just factoring" (Kai Siang-Ang)
  "No, it's Cauchy" (Shaya)
  "No, Shaya—No Cauchy, No Schwartz. Not related." (Dr. Aiyer)
- "I would be confused too if I actually understood what was happening" (Neymika)
- (Phone ringing) "I guess it was mine" (Dr. Aiyer)
  "Anu it was yours!" (Shaya)