INFORMATION THEORY

Contents

1
Introduction

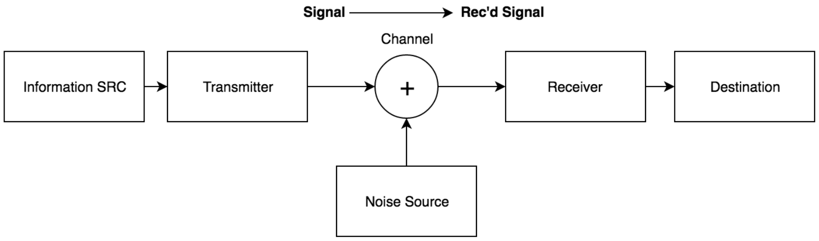
Two other sets of notes for this course are available at https://github.com/mananshah99/infotheory and http://tiny.cc/infotheory1 .

1.1 Why study Information Theory?

Information Theory is a mathematical framework for thinking about what it means to have efficient communication. Examples of information include:

- Email
- Telegraph
- Images
- Speech
- Video

Much of today's digital world revolves around transmitting information: we zip files, email them across the internet, download MP3s, etc. The ideas of information theory give us a rigorous way of characterizing streams of information. The cornerstone of our model will start with the following pipeline of sorts



This model gives us a general way of abstracting the transmission of information. On the left we have a source of information (where a signal originates) that is sent to a transmitter. A channel then allows that signal to flow to a receiver, although noise may be added at this stage. Finally, the receiver sends the signal to the destination.

Part I

Theory

2
Stochastic Variables

Definition 2.1 (Stochastic Variable)**.** A **stochastic variable** is a real-valued function of an outcome of an experiment.

Remark. Stochastic variables are also called random variables, but this term seems to imply all possibilities are equally likely or cannot be determined. The word stochastic generally means something like "depending upon probabilities."

Example 2.1. Toss a coin 7 times. The number of heads in the sequence could be a stochastic variable.

Remark. The 7-long sequence itself is not a stochastic variable, since it is not a real value. We should always have a clear way of assigning a real number to the outcome.

Example 2.2. Sum two rolls of a die. This value could be a stochastic variable. The number of 5's rolled could also be a stochastic variable.

A stochastic variable can either be discrete, taking on values from a countable set or continuous, taking on values on an interval from the real number line $\mathbb{R}$.

2.1 *Probability Measure*

Definition 2.2 (Sample Space)**.** The **sample space** is the set of all possible outcomes for an experiment.

The sample space is typically represented with Ω.

Definition 2.3 (Event)**.** An **event** is a subset of the sample space Ω.

Remark. Any event belongs to the power set of Ω.[1] Thus, events range from the empty set to a singleton set (i.e. a set with size unity) to Ω itself and anything in between.

$$\begin{tikzcd}
& X \arrow[dr, "\text{can take on values from}"] & \\
\Pr \arrow[ur, "\text{function of}"] \arrow[rr, "\text{of specific value}"] & & x
\end{tikzcd}$$

Definition 2.4 (Probability Measure)**.** The **probability measure** $\Pr$ is a real-valued function that assigns events probabilities obeying the **Kolmogorov axioms**.

Remark. Our notions of probability are fairly intuitive, so a careful treatment of the Kolmogorov axioms is not needed. They are available in the appendix, however.

[1] The Wikipedia article on events has good examples.

If each element of an event A is equally likely, then

$$\Pr(A) = \frac{|A|}{|\Omega|}$$

Notation. In general, if X is a stochastic variable, the probability mass function can be written in two equivalent ways:

$$p_X(x) = \Pr(X = x) \tag{2.1}$$

2.2 *Bayes' Theorem*

Events A and B are said to be independent if the occurence of one does not affect the probabilities of the other. When events A and B are not independent, knowing what A is gives us information about what B might be (and vice versa). The notation $\Pr(A|B)$ denotes the probability of A given that or conditioned upon B occuring. Consider that for A and B to happen that B must first happen and A must happen under those circumstances.

$$\Pr(A \cap B) = \Pr(B) \cdot \Pr\left(A|B\right) \tag{2.2}$$

From this definition, Thomas Bayes determined how to compute the support B provides for A given *priors* and *posteriors*.

Theorem 2.1 (Bayes' Theorem)**.**

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)}$$

Proof. Equation (2.2) is symmetric since A and B is the same as B and A. Thus, $\Pr(A|B)\Pr(B) = \Pr(B|A)\Pr(B)$. □

2.3 *Probability Mass Functions*

A **probability mass function** (pmf) characterizes a discrete stochastic variable by returning the probability measure of some x in Ω occuring.

Example 2.3. Consider 2 tosses of a fair coin. What is the probability mass function (pmf) of the number of heads given this experiment?

Solution. It is useful to construct a table of all the possibilities.

	Heads	Tails
Heads	2	1
Tails	1	0

From the table, we can conclude that

$$\Pr(X = x) = \begin{cases} 1/4 & x = 0 \\ 1/2 & x = 1 \\ 1/4 & x = 2 \\ 0 & \text{otherwise} \end{cases}$$

Example 2.4. Consider a 4-sided die rolled twice. What is the probability mass function for the maximum value of 2 rolls?

Solution. As before, let us think about the various possibilities. There are $4 \times 4 = 16$ total possibilities, and the maximum value can take on values 1, 2, 3, and 4. To take on a value of 1, both rolls must have been a 1; the probability of this happening is 1/16. To take on a value of 2, one of the rolls must have been a 2 and the other one must have been a 1 or a 2. The possibilites are enumerated as (2,1), (2,2), (1,2). Thus, its chance is 3/16. For a max value of 3, the possibilities are (3,1), (3,2), (3,3), (1,3), (2,3). Thus,

$$p_X(x) = \begin{cases} 1/16 & x = 1 \\ 3/16 & x = 2 \\ 5/16 & x = 3 \\ 7/16 & x = 4 \\ 0 & \text{otherwise} \end{cases}$$

There are a few common discrete stochastic variables that are worth discussing separately. The chapter on Discrete Stochastic Variables covers 4 important ones. The following sections will assume familiarity with these variables.

2.4 *Expectation and Variance*

As is often the case in mathematics, we like to define general operators on objects to understand their properties. Firstly, note that we can use functions of stochastic variables to build other stochastic variables.

Example 2.5 (Simple Functions of a Stochastic Variable)**.** Consider the following pmf

$$p_X(x) = \begin{cases} 1/9 & x \in \{-4, -3, -2, -1, 0, 1, 2, 3, 4\} \\ 0 & \text{otherwise} \end{cases}$$

Here are two functions of X and their probability mass functions:

a) $Y = |X| \implies p_Y(y) = \begin{cases} 2/9 & y \in \{1, 2, 3, 4\} \\ 1/9 & y = 0 \end{cases}$

b) $Z = X^2 \implies p_Z(z) = \begin{cases} 2/9 & z \in \{1,4,9,16\} \\ 1/9 & z = 0 \end{cases}$

A special class of functions, known as moments, include two operators, **expectation** and **variance**.

Definition 2.5 (Expectation)**.** The expectation of a function g of a stochastic variable X is

$$\mathbb{E}[g(X)] \equiv \sum_{x\in\Omega} g(x) \cdot \Pr(X = x)$$

Expectation returns the mean of the function g, that is, what we expect the average value of $g(X)$ to be if we run the experiment an infinite number of times.

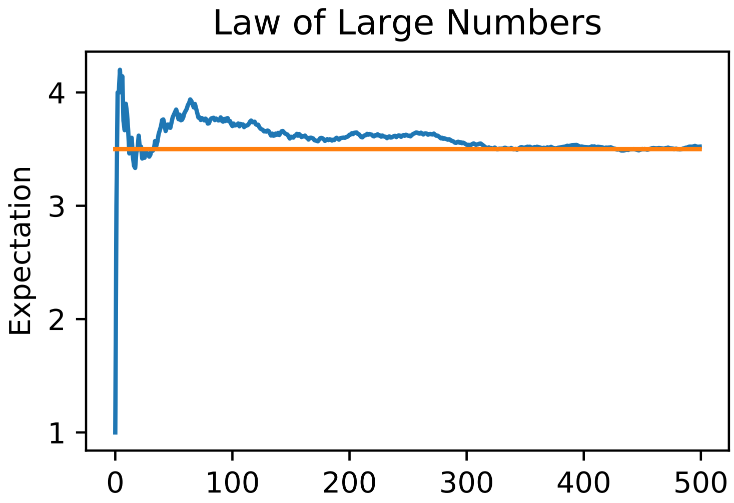
Exercise 2.1 (Expected Value of a Die)**.** Find $\mathbb{E}[X]$ where X is the value of a roll of a die.

Solution. Each value has a $1/6$ chance of appearing. Therefore,

$$\mathbb{E}[X] = \sum_{i=1}^{6} \frac{1}{6} i = \frac{21}{6} = 3.5$$

We include a numerical solution showing convergence:

```
N  = 500; plt.plot(np.cumsum(np.random.randint(1, 7, size=N)) / np.arange(1, N + 1))
```



Theorem 2.2 (Linearity of Expectation)**.**

$$\mathbb{E}\left[aX + bY\right] = a\mathbb{E}[X] + b\mathbb{E}[Y]$$

The nth moment about x_0 is defined as $\mathbb{E}[(x - x_0)^n]$.

Definition 2.6 (Variance)**.** The variance is the 2nd moment about the mean.

$$\mathrm{Var}[X] \equiv \mathbb{E}[(X - \mathbb{E}[X])^2] \tag{2.3}$$

Unfortunately, the definition given above in Equation (2.3) tends to be difficult to compute by hand. A little algebra results in a computationally simpler alternative.

Lemma 2.1 (Determinism of Expectation of Expectation)**.**

$$\mathbb{E}\Big[\mathbb{E}[X]\Big] = \mathbb{E}[X]$$

Theorem 2.3 (Computationally Simpler Alternative for Variance)**.**

$$\mathrm{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}\,[X]^2 \tag{2.4}$$

Proof.

$$\begin{aligned}\mathrm{Var}[X] &= \mathbb{E}\Big[X^2 + \mathbb{E}[X]^2 - 2X\mathbb{E}[X]\Big] \\ &= \mathbb{E}[X^2] + \mathbb{E}[X]^2 - 2\mathbb{E}[X]\mathbb{E}[X] \\ &= \mathbb{E}[X^2] - \mathbb{E}[X]^2.\end{aligned}$$

□

As an aside, the indicator function $\mathbb{1}_A$ indicates whether certain events are in A or not.

$$\mathbb{1}_A(\omega) = \begin{cases} 1 & \omega \in A \\ 0 & \omega \notin A \end{cases}$$

As a result,

$$\begin{aligned}\mathbb{E}[\mathbb{1}_A(\omega)] &= \Pr(A) \\ \mathrm{Var}[\mathbb{1}_A(\omega)] &= P(A)(1 - P(A))\end{aligned}$$

Lemma 2.2. *If X and Y are independent,*

$$\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y]$$

Proof.

$$\begin{aligned}\mathrm{Var}[X+Y] &= \mathbb{E}[(X+Y)^2] - \mathbb{E}[X+Y]^2 \\ &= \mathbb{E}[X^2 + 2XY + Y^2] - \mathbb{E}[X+Y]^2 \\ &= \mathbb{E}[X^2] + 2\mathbb{E}[XY] + \mathbb{E}[Y^2] - (\mathbb{E}[X+Y])^2 \\ &= \mathbb{E}[X^2] + 2\mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[Y^2] - \Big(\mathbb{E}[X]^2 + 2\mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[Y]^2\Big) \\ &= \mathbb{E}[X^2] + 2\mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[Y^2] - \mathbb{E}[X]^2 - 2\mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[Y]^2 \\ &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 + \mathbb{E}[Y^2] - \mathbb{E}[Y]^2 \\ &= \mathrm{Var}[X] + \mathrm{Var}[Y]\end{aligned}$$

□

Table 2.2: Expectation and Variance of Common Distributions

Distribution	Expectation	Variance
Binomial	np	$np(1-p)$
Geometric	$1/p$	$(1-p)/p^2$
Poisson	λ	λ

2.5 *Probability Density Functions*

Now we turn to stochastic variables that return values on an interval of the real number line. We can use our knowledge of discrete stochastic variables to find analagous versions for continuous stochastic variables. Instead of a probability mass function, we call the function for a continuous stochastic variables **probability density function** (pdf).

Notation. Probability density functions are typically denoted with lowercase English letters, most commonly f.

The requirement that the total probability must add up to unity is still in place:

$$\int_{\mathbb{R}} f_X(x)\,\mathrm{d}x = 1$$

However, asking whether $X = x$ is no longer a well-formed question, since X has a continuum. The probability that X takes on the exact value x is nil. Therefore, probabilities of a continuous stochastic variable may only be queried in the following form.

$$\Pr(a \leq X \leq b) = \int\limits_a^b f_X(x)\,\mathrm{d}x$$

By extension, the expectation of a pdf can be computed as

$$\mathbb{E}[X] = \int_{\mathbb{R}} x f_X(x)\,\mathrm{d}x$$

The chapter on Continuous Stochastic Variables discusses 3 important stochastic variables: uniform, exponential, and Gaussian. These variables will come up later, but for the sake of cleanliness, the proofs and computations have been moved into the appendix.

3
Entropy

Why did we spend the time reviewing probability for information? The answer lies in Proposition 3.1 below:

Proposition 3.1 (Stochasticity of Information)**.** *Information can be modelled as samples from a stochastic source.*

Consider the fact that many emails you send will have a typical structure: a greeting, a body of text, a conclusion. Or that the changes in the frames of a video tend to be rather small. Intuitively, we have a sense that the repetitive elements of data are not information-dense, and therefore, when we transmit this information, we should really only focus on what is novel about each message.

Shannon's original example was to show that text can modelled probabilistically in this way. Let's imagine that we want to generate some text that looks like English. We can first start by creating a sample space Ω that includes letters and spaces. Then, we sample randomly. This is callled a zero-order approximation. Next, we can refine the approximation by making characters more or less frequent. Adapting the probability based on the frequency with which that character appears in a corpus (e.g. make *e* show up about 12% of the time) is called a first-order approximation. An even more refined approach would consider digram (2-character sequences) and their frequencies.

3.1 Surprisal

Now that you're convinced that information can be modelled stochastically, we can consider what information means.

Example 3.1 (The Q-U Question)**.** Consider a game where you predict the next letter of a piece of English text given all the previous letters. You are given the phrase `elephants are q`. You know that the letter `q` is nearly always followed by a `u`. You would then predict that the next letter is a `u`, and you would be confident in your guess. If, by some odd reason, that next letter is not a u, you would be "surprised".

What we've captured in this example is a working conception of information.

Definition 3.1 (Information)**.** New information (which is really the only kind of information we care about) **is** the "surprisal" of an information source. The more surprised you are, the more information you gain, since you didn't expect to see that result.

Shannon's formula for entropy actually has a similarity to thermodynamic entropy:

$$S = -k_B \sum p_i \ln p_i$$

We now need a way of quantifying surprisal/information. The core of our theory is that surprisal should be inversely correlated with the probability of occurence. So naturally, we gravitate towards picking something like $1/p$ as our information function. Consider the limit cases, however, of literally using $1/p$. When $p = 0$, we have infinite surprisal, and when $p = 1$, we have 1 unit of surprisal. If something is guaranteed to happen, 1 unit is an odd baseline to use. As a result, we pick $\log(1/p)$, which is more attractive for a few reasons.

1. Continuity
2. Monotonically decreasing in p
3. Never negative
4. With $p = 1$, information becomes 0
5. Information due to independent events is additive

Typically, we will pick a base of 2 although any other sensible base works. We will usually omit the subscript/base unless it needs to be made explicit.

To each event, we now attach a surprisal value. To characterize a stochastic variable as a whole, we now define entropy.

Definition 3.2 (Entropy)**.** The entropy H of a stochastic variable X is the expectation of surprisal of X.

$$\begin{aligned} H(X) &\equiv \mathbb{E}\left[\log \frac{1}{\Pr(X = x)}\right] \\ &= \sum_i p_i \log(1/p_i) \\ &= -\sum_i p_i \log p_i \end{aligned}$$

Lemma 3.1 (Entropy is Nonnegative)**.**

$$\forall X, H(X) \geq 0$$

Proof.

$$\begin{aligned} H(X) &\geq 0 \\ -\sum_{x\in\Omega} \Pr(X=x)\log\Pr(X=x) &\geq 0 \\ \sum_{x\in\Omega} \Pr(X=x)\log\Pr(X=x) &\leq 0 \end{aligned}$$

Note that

$$\forall x \in \Omega, 0 \leq \Pr(X=x) \leq 1 \implies \forall x \in \Omega, \log\Pr(X=x) \leq 0.$$

Since a weighted sum of negative numbers with nonnegative numbers can never be positive, entropy can never be negative. □

Exercise 3.1 (Entropy of Bernoulli Distribution)**.** Find the entropy of a general Bernoulli distribution and plot the entropy against the probability of heads.

Solution.

$$H(X) = p\log\frac{1}{p} + (1-p)\log\frac{1}{1-p}$$

Plotting for every possible value for p, we yield a nice graph:

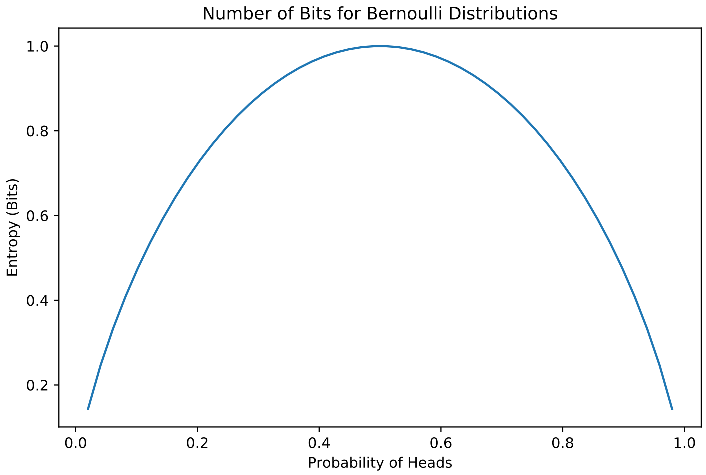


Figure 3.1: Bernoulli Entropy

When p is 0 or 1, we need 0 bits of information, which makes sense because the result was guaranteed. As we go more towards complete randomness (which colloquially, we might also call "entropy" from a physics standpoint), we need more bits to represent the possibilites (a maximum of 1 in this case).

Remark. But what does it mean to have 0.47 bits, which we might have if $p(\text{heads}) = 0.9$? Imagine that we had a 100-long sequence of coin flips and we transmitted the information. For the purely random case (i.e. using a fair coin), we would need 100 bits. However, for this extremely unfair case, we could get away with 47 bits without losing any information (on average).

3.2 *Bit Representations*

While we will be overloading the word bit in different contexts in this book, it is useful to understand what it represents. As noted before, bit is an abbreviation of "binary digit." When we talk about a bit in computer science, we typically mean 0 or 1, low voltage or high voltage, etc. Here, we take a bit to mean something like the answer to a single yes or no question with yes and no equally likely. In other words, a coin toss. That is, one bit captures the information of a Bernoulli distribution with $p = 0.5$. From there, we can meaningfully interpret values of entropy as telling us roughly how many of these yes/no questions or coin flips or sequence of binary digits are needed to transmit the information on average.

Exercise 3.2 (Entropy of a Fair Dice Roll)**.** Find the entropy of a fair dice roll.

Solution. Note that entropy does not care about the actual values of X. Therefore, the entropy is computed as

$$\begin{aligned} H(x) &= \sum_x p(x) \log(1/p(x)) \\ &= \sum_{x=1}^{6} \frac{1}{6} \log(6) \\ &= 6 \cdot \frac{1}{6} \log(6) \\ &= \log 6 \approx 2.585 \end{aligned}$$

Exercise 3.3 (Double the Possibilities)**.** Repeat Exercise 3.2 except that there are double the number of possible values (i.e. a 12-sided die).

Solution. Intuitively, we just need to add one more bit to flip between the first 6 and last 6 values. Mathematically, we consider $\log(12)$, which by log properties is $\log(2) + \log(6) = 1 + \log(6)$.

3.3 *Jensen's Inequality*

Definition 3.3 (Convex Function)**.** A function $f(x)$ is **convex** on the interval (a, b) if it is concave up on that interval (i.e. second derivative is positive).

Alternatively, it obeys the property that for all (x_1, x_2) within the interval (a, b) and for all λ normalized between 0 and 1,

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

Theorem 3.1 (Jensen's Inequality)**.** *For a stochastic variable X and a convex function f,*

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

Intuitively, if we think of expected value as an average, it makes sense that the average value of a convex function would never exceed the function's value at its average (i.e. its midpoint).

Theorem 3.2. *If X assumes real values $\{x_1, \ldots, x_n\}$ and $0 \leq H(X) \leq \log r$. Then,*

$$\forall 1 \leq i \leq n, p_i = \frac{1}{r} \iff H(X) = \log r$$

Theorem 3.2 tells us that an equiprobable distribution maximizes entropy. We now see that the intuition from Exercise 3.1 can be generalized.

3.4 *Joint & Conditional Entropy*

Definition 3.4 (Joint Entropy)**.**

$$H(X,Y) \equiv -\sum_x \sum_y \Pr(x,y) \log(\Pr(x,y))$$

Definition 3.5 (Conditional Entropy)**.**

$$H(Y|X) \equiv \sum_x \Pr(X=x) H(Y|X=x)$$

Theorem 3.3.

$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y) \tag{3.1}$$

Intuitively, we can think of Equation (3.1) as saying that a complete characterization of correlated variables X and Y can be described using as many bits needed to fully describe X and then however many additional bits it takes to describe Y given X (and vice versa).

Given that we have different kinds of entropy, it is helpful to write down one equation that makes the definitions easier to digest:

$$H(\cdot) = \mathbb{E}\left[\log \frac{1}{\Pr(\cdot)}\right] \tag{3.2}$$

Equation (3.2) generates $H(X)$, $H(X,Y)$ and $H(X|Y)$.

Theorem 3.4. *Let X be a stochastic variable and g a function of X. Then,*

$$H(g(X)) \leq H(X)$$

Proof. Using Theorem 3.3, we can expand

$$H(X, g(X)) = H(X) + H(g(X)|X)$$

However, $g(X)$ is completely determined by X. In other words, no additional bits are needed to describe $g(X)$ if one has the bits describing X. Thus, $H(g(X)|X) = 0$, and

$$H(X, g(X)) = H(X). \tag{3.3}$$

Additionally, we can use the alternate symmetric expansion

$$H(X, g(X)) = H(g(X)) + H(X|g(X))$$

If g happens to be an injective function, then $H(X|g(X)) = 0$ since knowing $g(x)$ allows us to trace back the specific x which generated it. However, injectivity is not guaranteed. In general, g could take different x inputs and produce the same output with them. This conflation means that we would require additional bits. Whether injective or not, we can guarantee that $H(g(X)) + H(X|g(X))$ will not be lower than $H(g(X))$. Or, more to the point,

$$H(X, g(X)) \geq H(g(X)). \tag{3.4}$$

We now combine Equations (3.3) and (3.4) to assert that $H(g(X)) \leq H(X)$. □

Interestingly, we can use conditional entropy to define a **metric**.

Definition 3.6 (Metric)**.** A metric on a set X is a function $d : X \times X \to [0, \infty]$ that defines distance with the following constraints:

1. $\forall x, y \in X : d(x, y) \geq 0$
2. $\forall x, y \in X : d(x, y) = 0 \iff x = y$
3. $\forall x, y \in X : d(x, y) = d(y, x)$
4. $\forall x, y, z \in X : d(x, z) \leq d(x, y) + d(y, z)$

Our goal is to show that $d(X, Y) = H(X|Y) + H(Y|X)$ works. From Lemma 3.1, property (1) should clearly be true. Our intuition should match up with property (2): the only way to be 0 is if the number of bits that describes X completely describes Y. Property (3) should also be clear by symmetry. Property (4), intuitively, says that it takes less bits to describe X and Z together than to describe X and Y and Y and Z (effectively, all 3).

3.5 *Differential Entropy*

We would now like to generalize the concept of entropy to continuous stochastic variables. We call this version **differential entropy**. However, we cannot simply switch the sum for an integral:

Limit of Discrete Entropy. Imagine a discrete stochastic variable X. The probability that X^Δ is in the ith bin is $p(x_i)\Delta x$. Then, we can imagine limiting Δx to an infinitesimal dx.

$$\begin{aligned} H(X^\Delta) &= \sum_x \Pr(X = x) \log \frac{1}{\Pr(X = x)} \\ &= \sum_i p(x_i)\Delta x \log \left(\frac{1}{p(x_i)\Delta x} \right) \\ &= \sum_i \left[p(x_i)\Delta x \log \frac{1}{p(x_i)} + p(x_i)\Delta x \log \frac{1}{\Delta x} \right] \end{aligned}$$

If we allow the bins to become infinitesimal,

$$H(X) = \int_{\mathbb{R}} f_X(x) \log \frac{1}{f_X(x)} \, \mathrm{d}x + \sum_i p(x_i)\Delta x \log \frac{1}{\Delta x}$$

Unfortunately, that second term, the sum, cannot be turned into an integral because it grows to infinity. □

In any case, we drop this problematic term and label it differential entropy.

Definition 3.7 (Differential Entropy)**.** The differential entropy h of a continuous stochastic variable X is

$$h(\mathrm{X}) \equiv - \int_{\mathbb{R}} f_X(x) \log f_X(x) \, \mathrm{d}x$$

Intuitively, this infinity term represents information regarding precision. That is, if Δx were not infinitesimal, $H(X^\Delta)$ could be computed since the bins would have an actual size. As the bins get smaller and smaller, we need more information to pinpoint the exact numbers. If you actually knew a number to all its decimal places, you would have infinite information. It may be helpful to think of floating-point numbers here. If we actually wanted computers to represent all real numbers, it would take an infinite amount of information.

Exercise 3.4 (Entropy of Uniform)**.** Find $h(X)$ for X a uniform distribution.

Solution. On the interval $[a, b]$,

$$h(X) = \int_a^b \frac{1}{b-a} \log(b-a)\,\mathrm{d}x = \log(b-a) \tag{3.5}$$

Notice that if $a = 0$ and $b \leq 1$, we have 0 and negative values of entropy, so differential entropy really isn't like discrete entropy.

Exercise 3.5 (Entropy of Exponential)**.** Find $h(X)$ for X an exponential distribution.

Solution.

$$\begin{aligned} h(X) &= -\int_0^\infty f_X(x) \log f_X(x)\,\mathrm{d}x \\ &= -\int_0^\infty \lambda e^{-\lambda x} \log\left(\lambda e^{-\lambda x}\right)\,\mathrm{d}x \\ &= -\int_0^\infty \lambda e^{-\lambda x} \left(\log \lambda - \lambda x \log e\right)\,\mathrm{d}x \\ &= -\log \lambda + \lambda \log e \int_0^\infty x \lambda e^{-\lambda x}\,\mathrm{d}x \end{aligned}$$

Recall that the expectation of the exponential is $1/\lambda$. Thus,

$$h(X) = -\log \lambda + \lambda \frac{1}{\lambda} \log e = \log(e/\lambda) \tag{3.6}$$

Exercise 3.6 (Entropy of Gaussian)**.** Find $h(X)$ for X a Gaussian distribution.

Solution. The Gaussian distribution is a function with a lot of components, so we'll simplify the proof by packaging some of the parts

$$\begin{aligned} f_X(x) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-x^2/(2\sigma^2)) \\ &= c \exp g(x) \end{aligned}$$

We can drop the μ because the integral expands over the real line anyways. A u-substitution of $y = x - \mu$ would not change the integral.

The trick to this question is to expand only the log side of the differential entropy:

$$
\begin{aligned}
h(X) &= -\int_{\mathbb{R}} f_X(x) \log f_X(x)\, \mathrm{d}x && \text{definition} \\
&= -\int_{\mathbb{R}} f_X(x) \log(c \exp(g(x)))\, \mathrm{d}x && \text{substitution} \\
&= -\int_{\mathbb{R}} \left(f_X(x) \log c + f(x) \log \exp(g(x)) \right)\, \mathrm{d}x && \text{distributive property} \\
&= -\left(\log c \cancelto{1}{\int_{\mathbb{R}} f_X(x)\, \mathrm{d}x} + \int_{\mathbb{R}} f(x) \frac{\ln \exp(g(x))}{\ln 2}\, \mathrm{d}x \right) && \text{change of base} \\
&= -\left(\log c + \frac{-1}{2\sigma^2 \ln 2} \int_{\mathbb{R}} f(x) x^2\, \mathrm{d}x \right) && \text{substitution} \\
&= -\left(\log c + \frac{-1}{2\sigma^2 \ln 2} \mathbb{E}[X^2] \right) && \text{definition} \\
&= -\left(\log c + \frac{-\sigma^2}{2\sigma^2 \ln 2} \right) && \text{Gaussian properties}
\end{aligned}
$$

The expression simplifies to

$$h(X) = \log(2\pi e \sigma^2)/2 \tag{3.7}$$

Having done all the work of finding the differential entropy for three continuous stochastic variables, we make a few claims with important ramifications:

Claim: Given a fixed upper and lower bound, no pdf can have a larger entropy than the uniform stochastic variable, the entropy given in Equation (3.5).

Claim: Given a positive stochastic variable with a mean μ with no other constraints, no pdf can have a larger entropy than the exponential stochastic variable, the entropy given in Equation (3.6). [1]

Claim: Given a stochastic variable of variance σ^2 with no other constraints, no pdf can have a larger entropy than the Gaussian, the entropy given in Equation (3.7).

[1] In other words, the exponential random variable is the most equiprobable way of creating a convergent series.

4
Source Coding

4.1 Encoding the English Alphabet

Here's a practical problem we would like to solve: Given the 26 letters of the English alphabet and assuming letters are coming independently, design an encoder (a schema that converts text into a binary message) to minimize the expected number of bits used per letter. Essentially, can we find an encoding of the English alphabet using a zero-order approximation?

We'll start with a simple solution:

1. Compute how many bits it would take if each letter had the same number of bits. The number of bits needed is given by $\lceil \log_2(26) \rceil$[1]
2. Then, A becomes `00000`, B becomes `00001`, and so on and so forth.
3. Store the characters and their numbers in a matrix. The matrix serves as both the encoding and decoding scheme.

Solution #1 is actually the best approach if each character had an equal probability (i.e. $1/26$) of appearing. However, certain characters tend to appear more than others. Therefore, using the same number of bits to represent a commonly occuring character like *e* and a infrequent character like *z* is not making the best use of each bit.

Our next solution will take into account frequency. The three most common letters in the English alphabet are E, T, and A. Assign E the value `0`, T the value `1`, and A the value `10`. However, it now becomes impossible to determine whether `10` is a "TA" message or an "E" message. We need to avoid such prefix-collisions to interpret messages without ambiguity.

4.2 Huffman Coding

In 1952, David Huffman, a student at MIT, developed an algorithm for efficiently encoding symbols without any loss of information. The algorithm follows:

Notation. Let Σ denote an alphabet of symbols to be encoded.

[1] The upper brackets denote the ceiling function or greatest integer function. The number of bits must be a natural number.

```
Data: A map of symbols and their weights
Result: A binary tree representing an encoding scheme
while The map has more than 1 key do
    Remove the two symbols $a$ and $b$ with the least weights ;
    Insert a new symbol, a tree with leaves $a$ and $b$, with a
      weight equal to the sum of their individual weights ;
end
```

Algorithm 1: Building a Huffman Tree

Theorem 4.1 (Huffman Coding Optimality)**.** *If X is a random variable, and L is the expected number of bits per letter using Huffman coding,*

$$H(X) \leq L(X) \leq H(X) + 1$$

An intuitive video explaining Huffman coding can be found here: https://www.youtube.com/watch?v=JsTptu56GM8 . A Python program is available here: https://github.com/ashwinreddy/huffman .

Example 4.1. Let $X \in \Sigma$ where $\Sigma = \{1, 2, 3, 4, 5\}$ with probabilities .25, .25, .2, .15, .15, respectively.

a) Draw the Huffman Tree
b) Compute $\mathbb{E}[L(X)]$ where $L(X)$ is the length of the encoding for X.
c) Find $H(X)$
d) Repeat for a ternary tree

Solution. Drawing the tree, we find

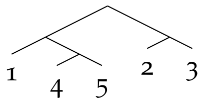


Figure 4.1: Huffman Tree with 2 branches

The expected word length is between 2 and 3 since each node is either at the second or third level. 30% of the time, we expect to see a 4 or a 5, which will require 3 bits. The other 70% of the time, we expect to use two bits.

$$\mathbb{E}[L(X)] = 0.3 \cdot 3 + 0.7 \cdot 2 = 2.3$$

Also, $H(X) \approx 2.28$. Huffman coding is within the band of optimality.

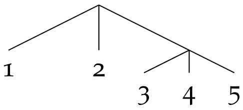


Figure 4.2: Huffman Tree with 3 Branches

The expected word length is 1.5 bits. We adjust $H(X)$ using $\log_3$ accordingly to find $H(X) \approx 1.3$.

To remain efficient, one must insert an appropriate number of dummy symbols into the list with weights 0. Add enough symbols so that the total number of symbols can be written as $1 + k(D-1)$ where D is the base being used for the smallest possible $k \in \mathbb{N}$.

In the special case that $D = 2$, the formula simplifies to $1 + k$. In other words, in base 2, dummy symbols do not need to be added if there is already at least one symbol.

4.3 *Lagrange Multipliers*

To prove Huffman optimality, we will use Lagrange multipliers. Recall that Lagrange multipliers solve constraint/constrained optimization problems. For this family of problems, we attempt to minimize or maximize a function that is constrained by some relation of the function's variables. Lagrange Multipliers provides the following algorithm to solving this problem.

1. Let $f(x_1, \ldots, x_n)$ be a function whose extrema we would like to find. Let $g(x_1, \ldots, x_n) = k$ be a constraint imposed on the $x_1, \ldots, x_n$.
2. Solve the equation $\vec{\nabla} f = \lambda \vec{\nabla} g$ while still imposing the constraint that $g(x_1, \ldots, x_n) = k$. This system should provide enough equations for the variables involved.[2]
3. Collect the solutions. These are possible extrema.

Alternatively, we can construct a function

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

Then, solving the constrained optimization problem reduces to finding solutions to $\vec{\nabla}\mathcal{L} = \vec{0}$.

Example 4.2. Find extrema of $f(x, y) = x^2 - \ln x$ subject to $8x - 3y = 0$.

Solution. Let $g(x) = 8x + 3y$. We solve two equations simultaneously:

$$\vec{\nabla} f(x, y) = \lambda \vec{\nabla} g(x, y) \tag{4.1}$$

$$8x + 3y = 0 \tag{4.2}$$

We then expand Equation (4.1), a vector equation, into its two component equations. To do this, we must first calculate the gradients.

[2] We introduced a λ not involved in the original system, which adds an additional equation because we must also solve for λ.

$$\begin{aligned}\frac{\partial f}{\partial x} = \lambda\frac{\partial g}{\partial x} &\implies 2xy - \frac{1}{x} = 8\lambda \\ \frac{\partial f}{\partial y} = \lambda\frac{\partial g}{\partial y} &\implies x^2 = 3\lambda \\ 8x + 3y = 0\end{aligned}$$

Solving all three equations, we find that $x = -1/2$, $y = 4/3$, and $\lambda = 1/12$.

Solution (Alternative Solution). The constraint $8x + 3y = 0$ can be rewritten as $y(x) = -\frac{8}{3}x$. Then $f(x, y)$ is really only a function of x:

$$\begin{aligned}f(x, y(x)) = f(x) &= x^2\left(-\frac{8}{3}x\right) - \ln x \\ &= -\frac{8}{3}x^3 - \ln x\end{aligned}$$

We can find minima of f w.r.t x from here:

$$\frac{\mathrm{d}f}{\mathrm{d}x} = -8x^2 - \frac{1}{x}$$

$$\begin{aligned}-8x^2 - \frac{1}{x} &= 0 \\ -8x^2 &= \frac{1}{x} \\ -8x^3 &= 1 \\ x^3 &= -\frac{1}{8} \\ x &= -1/2\end{aligned}$$

The value of y can be recovered by simply evaluating $y(-1/2) = 4/3$.

4.4 *Coding Classifications*

Instantaneous is a synonym for prefix-free

- All codes
- Nonsingular codes (nonsingular codes are when the mapping from source symbols to bit string is injective.)
- Uniquely decodable codes
- Instantaneous codes

4.5 *Huffman Optimality*

Theorem 4.2 (Kraft Inequality)**.** *All instaneous codes with d symbols with code word lengths $\ell_1, \ldots, \ell_m$ must satisfy*

$$\sum_{i=1}^{m} 2^{-\ell_i} \leq 1$$

Proof. Note that an instaneous entails that no code word is the ancestor of another code word on the tree.

1. Let ℓ_{max} be the length of the longest code word.
2. For the sake of argument, grow the tree out to length ℓ_{max} (create and expand nodes out to this depth)
3. Code words with length ℓ_i have $2^{\ell_{max}-\ell_i}$ descendants at the level ℓ_{max}.
4. Since instantaneous codes do not grow to this level, it must be the case that

$$\sum_i 2^{\ell_{max}-\ell_i} \leq \sum_i 2^{\ell_{max}}$$

Since $2^{\ell_{max}}$ is constant given the ℓ_i's,

$$\sum_{i=1}^{m} 2^{-\ell_i} \leq 1$$

□

We can now write down a constraint-optimization problem. We would like to minimize the expected length given the constraint defined by Theorem 4.2.

$$\min L = \mathbb{E}[\ell] \qquad \sum_{i=1}^{m} 2^{-\ell_i} \leq 1$$

We formulate a Lagrangian

$$\mathcal{L}(\ell_1, \ell_2, \ldots, \ell_n, \lambda) = \sum_i p_i \ell_i - \lambda \left(\sum 2^{-\ell_i} - 1 \right)$$

$$\frac{\partial \mathcal{L}}{\partial \ell_i} = p_i + \lambda \left(2^{-\ell_i} \ln 2 \right)$$

$$L^* = \sum p_i \ell_i^* = -\sum p_i \log_2 p_i = H(X)$$

Since $\log \frac{1}{p_i}$ may not be an integer, we use $\ell_i = \lceil \log \frac{1}{p_i} \rceil$. We show that it still satisfies Theorem 4.2.

$$\sum_i 2^{-\ell_i} \leq \sum_i 2^{-\log \frac{1}{p_i}} = \sum p_i = 1 \tag{4.3}$$

$$\log \frac{1}{p_i} \leq l_i \leq \log \frac{1}{p_i} + 1 \tag{4.4}$$

$$\sum p_i \log \frac{1}{p_i} \leq \sum p_i l_i \leq \sum p_i \log \frac{1}{p_i} + 1 \tag{4.5}$$

$$H(X) \leq L(X) \leq H(X) + 1 \tag{4.6}$$

The 1 bit is split over the block

Huffman produces an optimal code but does not have to be the only way.

Lemma 4.1. *For any distribution, there exists an instantaneous, optimal code minimizing the expected length that satisfies*

a) $p_j > p_k \implies \ell_j \leq \ell_k$.
b) The two longest code words have the same length
c) The two longest code words differ only in the last bit, and correspond to the two least likely symbols.

Proof. We prove each part in order.

a) Consider C'_m with code words j and k swapped. Assume C_m optimal.

$$\begin{aligned} L(C'_m) - L(C_m) &= \sum_i p_i \ell'_i - \sum_i p_i \ell_i \\ &= \sum_i p_i \ell'_i - \sum_i p_i \ell_i \\ &= p_j \ell_k + p_k \ell_j - p_j \ell_j - p_k \ell_k \\ &= \underbrace{(p_j - p_k)}_{>0}(\ell_k - \ell_j) \end{aligned}$$

b)

c) Suppose there is a max length code word without any siblings. THen, we can delete the last bit of the code word and still satisfy the prefix-free property. This leads to a contradiction. Therefore, every maximal lnegth code word in any optimal code must have siblings. Then, exchange longest code words such that the lowest probability code word are associated with 2 siblings (doesn't change $\mathbb{E}[L]$)

Given optimal code, it can be shuffled into Huffman without changing expected word length, implying optimality.

Let p be a tuple $(p_1, p_2, \ldots, p_m)$ from greatest to least. Let $C^\star_m(p)$ be the optimal distribution. If $p' = (p_1, p_2, \ldots, p_{m-2}, p_{m-1} + p_m)$ and $C^\star_{m-1}(p')$ optimal code for p'

1) Take $C^\star_{m-1}(p')$ and extend $(p_{m-1} + p_m)$ node by adding 0 to form a code word for p_{m-1} and 1 to form a code word for p_m.

$$L(p) = L^\star(p') + p_{m-1} + p_m \tag{4.7}$$

2) Take $C^\star_m(p)$ and merge code word for 2 lowest symbols

$$L(p') = L^\star(p) - p_{m-1} - p_m \tag{4.8}$$

Add Equations (4.7) and (4.8):

$$L(p) + L(p') = L^\star(p') + L^\star(p)$$

Use induction to show this is true for all levels of the tree. □

Theorem 4.3 (Huffman Optimality)**.** *Huffman coding is optimal.*

5
Coupled Stochastic Variables

5.1 Fun with Expectations

$$\mathbb{E}[X|Y=y] = \sum_x P(X=x|Y=y) = \int_{\mathbb{R}} x f_{x|y}(x|y)dx$$

Exercise 5.1. A miner is lost in a cave. Facing north, he finds three caverns which seem to continue on. If he follows the trail through the first cavern, he will find his way back home in 3 hours. Through the second cavern, he will follow a circuitous route such that he will enter the cave facing north again. Through the third cavern, the same will occur but in 7 hours. The miner is very scared and cold and forgets which path he took every time he returns to the cave. He also never turns around to see the hole(s) that took him back to the cave. Find the expected time it takes for him to go to safety.

Solution. Imagine a Markov model with loops of 5 and 7 back to themselves with probability 1/3 each and an edge with 3 out to the final state.

$$\mathbb{E}[X] = \sum_x xp(x)$$

There's a 1/3 probability straight off the bat that he escapes the cave and doesn't have to deal with any of this nonsense. In this case, $xp(x) = 3 \cdot \frac{1}{3} = 1$. However, there's a 2/3 chance that he takes 6 hours to come back (since the 5 and 7 hour paths have equal probabilities of being selected, given that the first path was not). From there, the process repeats. In essence, the expression looks like

$$\frac{1}{3}(3) + \frac{2}{3}\left[6 + \frac{1}{3}(3) + \frac{2}{3}\left[6 + \frac{1}{3}(3) + \dots\right.\right.$$

Given that the expression continues on infinitely, we cannot evaluate it as a closed-form expression. Let us instead consider generically the sum

$$k(j + k(j + k(j + \dots$$

If we write

$$f(n) = k\left[j + f(n)\right]$$

Then

$$\begin{aligned} f(n) &= kj + kf(n) \\ f(n) - kf(n) &= kj \\ f(n)(1-k) &= kj \\ f(n) &= \frac{kj}{1-k} \end{aligned}$$

Using our values $k = 2/3$ and $j = 7$, we evaluate the entire expression from earlier as 15 hours.

Exercise 5.2. What is the expected number of cards one must flip to obtain 2 aces.

5.2 *Conditional Probability*

Given the joint probability function $p(X = x, Y = y)$ and a function $Z = g(X, Y)$ for random variables X and Y, can we find the probability function for Z?

Definition 5.1 (Cumulative Distribution Function)**.**

$$F_X(x) \equiv \Pr(X \leq x)$$

Mathematically,

$$F_X(x) = \int\limits_{-\infty}^{x} f_X(u)\,\mathrm{d}u$$

Therefore,

$$f_X(x) = \frac{\mathrm{d}F_X(x)}{\mathrm{d}x}$$

So if we can find $F_X(x)$, we can get $f_X(x)$ pretty easily. Let D be a region in the XY plane where $g(x, y) \leq z$ is satisfied. Then,

$$F_Z(z) = \int_D f(x,y)\,\mathrm{d}A \implies f_Z(z) = \frac{\mathrm{d}}{\mathrm{d}z}\int_D f(x,y)\,\mathrm{d}A$$

Example 5.1. Find the probability function for $Z = X + Y$.

Solution.

$$\begin{aligned} f_Z(z) &= \tfrac{\mathrm{d}}{\mathrm{d}z} \int_{\mathbb{R}} \int_{-\infty}^{z-x} f(x,y)\,\mathrm{d}y\,\mathrm{d}x \\ &= \int_{\mathbb{R}} \left[\tfrac{\partial}{\partial z} \int_{-\infty}^{z-x} f(x,y)\,\mathrm{d}y \right] \mathrm{d}x \\ &= \int_{\mathbb{R}} f(x, z-x)\,\mathrm{d}x \end{aligned}$$

If we further assume X and Y are independent, then

$$f_Z(z) = \int_{\mathbb{R}} f_X(x) f_Y(z-x)\,\mathrm{d}x = (f_Y * f_X)(z) = (f_X * f_Y)(z)$$

5.3 *Relative Entropy & Mututal Information*

Definition 5.2 (Relative Entropy / Kullback-Leibler Divergence)**.**

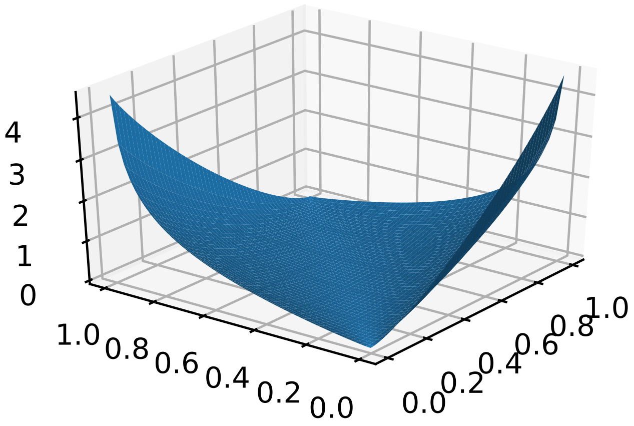
$$D_{KL}(P \parallel Q) \equiv \sum_x p(x) \log \frac{p(x)}{q(x)} = \mathbb{E}_p \left[\log \frac{p(x)}{q(x)} \right]$$

The KL divergence is the minimum possible inefficiency of using

Example 5.2. Let X be Bern(r) and Y be Bern(s). Find $D_{KL}(X \parallel Y)$

Solution.

$$D(X \parallel Y) = (1-r) \log \frac{1-r}{1-s} + r \log \frac{r}{s}$$



Definition 5.3 (Mutual Information)**.**

$$I(X;Y) \equiv D_{KL}(p(x,y) \parallel p(x)p(y))$$

Lemma 5.1.

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \tag{5.1}$$

Corollary 5.1.

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

Example 5.3. Find $I(X;Y)$ for the table below.

Table 5.1: Joint Entropy for X (columns) and Y (rows)

	1	2	3	4
1	1/8	1/16	1/32	1/32
2	1/16	1/8	1/32	1/32
3	1/16	1/16	1/16	1/16
4	1/4	0	0	0

Solution.

$$H(X) = \sum p(x) \log \frac{1}{p(x)} = 14/8$$

$$H(X|Y) = 11/8$$

Then, $I(X;Y) = 3/8$

Theorem 5.1 (Chain Rule for Entropy)**.** *Let $X_1, \ldots, X_n$ be drawn according to $p(x_1, \ldots, x_n)$. Then,*

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

Proof. First, recall that

$$H(X_1, X_2) = H(X_1) + H(X_2|X_1)$$

$$H(X_1, X_2, X_3) = H(X_1) + H(X_2, X_3|X_1)$$

□

Theorem 5.2 (Information Inequality)**.**

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

Proof.

$$\begin{aligned} -D_{KL}(p \parallel q) &= -\sum p(x) \log \frac{p(x)}{q(x)} \\ &= \sum p(x) \log \frac{q(x)}{p(x)} \\ &\leq \log \sum p(x) \frac{q(x)}{p(x)} \\ &= \log \sum q(x) \\ &= \log 1 \\ &= 0 \end{aligned}$$

$$-D_{KL}(p \parallel q) \leq 0 \implies D_{KL}(p \parallel q) \geq 0$$

□

Corollary 5.2.

$$I(X;Y) \geq 0$$

Exercise 5.3. Show that $H(X) \leq \log|\mathscr{H}|$. $|\mathscr{H}|$ denotes the number of elements in the range of X of equality iff X has a uniform distribution.

Proof.

$$\begin{aligned} D_{KL}(p \parallel \mathscr{U}) &= \mathbb{E}_p \left[\log \frac{p(x)}{q(x)} \right] \\ &= \mathbb{E}_p \left[\log \left(|\mathscr{H}| p(x) \right) \right] \\ &= \sum p(x) \log \left(|\mathscr{H}| p(x) \right) \\ &= \sum p(x) \left(\log|\mathscr{H}| + \log p(x) \right) \\ &= \sum p(x) \log|\mathscr{H}| + \sum p(x) \log p(x) \\ &= \log|\mathscr{H}| - H(X) \end{aligned}$$

□

Lemma 5.2. *The following statement holds with equality iff X and Y are independent.*

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

Intuition. Knowing another R.V. value cannot, on average, increase the surprisal of a R.V.

Exercise 5.4. Do something with the table below.

X	1	2
Y		
1	0	3/4
2	1/8	1/8

Solution.

$$H(X) = \sum p(x) \log \frac{1}{p(x)} = \frac{1}{8} \log 8 + \frac{7}{8} \log \frac{8}{7} \approx .544$$

$$H(X|Y = 1) = \frac{3}{4} \log \frac{4}{3} \approx .311$$

$$H(X|Y = 2) = 1$$

$$H(X|Y) \approx 0.483$$

6
Channel Coding

We now move to discussing channels.

Examples of channels include:

Example 6.1. Fiber optical / light-based transmissions

Example 6.2. Air or free space transmissions

Example 6.3. Coppper

Each medium will have two characteristics relating to how it affects the signal crossing it: attentuation (decrease in signal strength) and time delay or latency. These actually are not as problematic as one might think: given the characteristics, one can easily decode the message on the other end. The real problem is noise.

We might imagine a black box that takes an input waveform and produces an output. We might also consider the same with noise added.

We now have a pdf over waveforms, conditioned on input: $P(Y|X = x)$.

Assumption. Noise is independent of input.

6.1 *Channel Capacity*

1) Send something; before it comes out of the pipe, what is the surprisal of getting input exactly?

$$\log \frac{1}{p(x)}$$

2) Now, Y pops out. Surprisal of X being sent?

$$\log \frac{1}{p(x|y)}$$

3) What is the value (i.e. information or decrease in surprise) of seeing Y pop out?

$$\log \frac{1}{p(x)} - \log \frac{1}{p(x|y)}$$

4) What is the expectation of the information value of seeing Y pop out?

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

Remark. Simply put, mutual information is expected information value.

Definition 6.1 (Channel Capacity)**.** The channel capacity is the maximum mutual information over all sources.

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

Example 6.4. Find the channel capacity of a binary channel. Assume the probability of sending a 0 is q and sending a 1 is $1-q$.

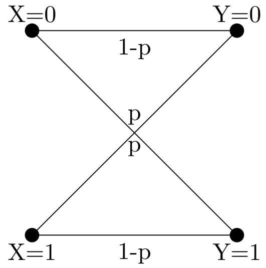


Figure 6.1: Binary Channel

Solution.

$$I(X;Y) = H(Y) - H(p)$$

$$\max I(X;Y) = 1 - H(p)$$

Notice that the channel capacity for a noiseless binary channel is just 1.

Example 6.5. Find the channel capacity of a Gaussian $Z \sim \mathcal{N}(0, \sigma_z^2)$. $Y = X + Z$.

For formulas previously where we used discrete entropy H, the differential entropy h can be replaced without modification.

Solution.

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

The question is what is $h(Y)$. We have to make some assumptions about the structure of X. We assume it is normal. The variance of the sum of two normal distributions is the sum of the variances of the two normal distributions. Thus,

$$\sigma_y^2 = \sigma_x^2 + \sigma_z^2$$

Example 6.6. Find the channel capacity of a

Appendix

A
Discrete Stochastic Variables

A.1 Bernoulli Distribution

A Bernoulli distribution represents the number of heads in tossing a potentially unfair coin once. The unfairness is characterized by a probability p that the coin lands heads. Therefore, the probability that the coin lands tails is $1 - p$.

Definition A.1 (Bernoulli Distribution)**.**

$$\text{Bern}(x) \equiv \begin{cases} p & x = 1 \\ 1 - p & x = 0 \end{cases} \tag{A.1}$$

A.2 Binomial Distribution

Next, the binomial distribution can be imagined as tossing the same unfair coin N times and counting the number of heads. The probability that the number of heads is nil is $(1 - p)^N$ since the implication is that the coin came up tails every single time. The probability that the number of heads is exactly one is $Np(1 - p)^{N-1}$. This is because it must have come up heads once with probability p and tails $N - 1$ times with probability $1 - p$. Additionally, the one heads could come up at any point in the sequence, which introduces a factor of $\binom{N}{k}$.

Definition A.2 (Binomial Distribution)**.**

$$\mathbb{P}(X = k; n, p) \equiv \binom{n}{k} p^k (1 - p)^{n-k} \tag{A.2}$$

You can check that $\mathbb{P}(X = 1; 1, p) = \text{Bern}(p)$

A.3 Geometric Distribution

In a geometric distribution, we keep tossing the coin until there is one heads.

Definition A.3 (Geometric Distribution)**.**

$$p_X(x) = (1-p)^{x-1}p \tag{A.3}$$

A.4 Poisson Distribution

Definition A.4 (Poisson Distribution)**.**

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

Similar to a binomial distribution, the Poisson distribution can be thought of as the number of replacements needed for a biased lightbulb (rather than a coin) in a given amount of time. In this case, we are typically dealing with a large n and a small p, which leads to a "moderate" np.[1] If $\lambda \equiv np$ (can be thought of as the expected number of times the bulb will burn out),

Proof. Start with binomial distribution, using λ instead of p:

$$\begin{aligned} p_X(x=k) &= \lim_{n\to\infty}\lim_{p\to 0}\frac{n!}{(n-k)!k!}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k} \\ &= \lim_{n\to\infty}\lim_{p\to 0}\frac{n(n-1)\dots(n-k+1)}{n^k}\cdot\frac{\lambda^k}{k!}\frac{(1-\lambda/n)^n}{(1-\lambda/n)^k)} \\ &= e^{-\lambda}\lambda^k/k! \end{aligned}$$

□

What can be modelled with a Poisson distribution?

- Number of customers entering a bank in a given period of time
- Number of misprints on a page
- Number of alpha particles discharged from a radioactive substance.

[1] I have no idea what this means.

B
Continuous Stochastic Variables

B.1 Uniform

Definition B.1 (Uniform Stochastic Variable)**.** For the interval $[a, b]$, the uniform stochastic variable assigns all x in the interval the same probability, so that

$$f_X(x) = \frac{1}{b-a}$$

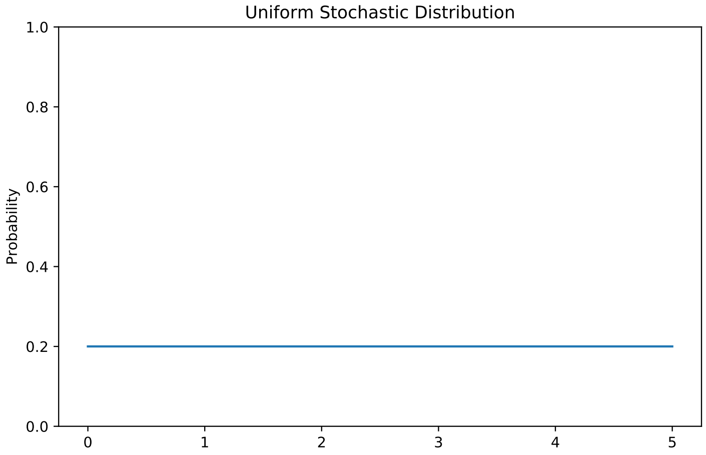


Figure B.1: A Uniform Distribution is Constant

Intuitively, the expectation for the uniform stochastic variable on the interval $[a, b]$ is $(a+b)/2$.

Proof.

$$\begin{aligned}
\mathbb{E}[X] &= \int_{\mathbb{R}} x f_X(x)\,\mathrm{d}x \\
&= \int_{\mathbb{R}} x \frac{1}{b-a}\,\mathrm{d}x \\
&= \frac{1}{b-a}\int_a^b x\,\mathrm{d}x \\
&= \frac{1}{b-a}\left[\frac{b^2-a^2}{2}\right] \\
&= \frac{(b+a)(b-a)}{2(b-a)} \\
&= \frac{b+a}{2}
\end{aligned}$$

□

The variance may also be computed using the formula in Equation (2.4).

$$\begin{aligned}
\mathrm{Var}[X] &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 \\
&= \mathbb{E}[X^2] - \left(\frac{a+b}{2}\right)^2 \\
&= \int_a^b x^2 f_X(x)\,\mathrm{d}x - \left(\frac{b+a}{2}\right)^2 \\
&= \frac{1}{b-a}\left[\frac{b^3-a^3}{3}\right] - \left(\frac{b+a}{2}\right)^2 \\
&= \frac{1}{12}(a-b)^2
\end{aligned}$$

B.2 Exponential

Definition B.2 (Exponential Stochastic Variable)**.**

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

Expectation of Exponential Stochastic Variable.

$$\begin{aligned}
\mathbb{E}[X] &= \int_0^\infty \lambda e^{-\lambda x}\,\mathrm{d}x \\
&= \lambda \int_0^\infty e^{-\lambda x}\,\mathrm{d}x
\end{aligned}$$

The integral can be evaluated by using integration by parts with $u = x$ and $\mathrm{d}v = e^{-\lambda x}\,\mathrm{d}x$. We record $\mathrm{d}u = \mathrm{d}x$ and $v = -\frac{e^{-\lambda x}}{\lambda}$.

$$\begin{aligned}\int_0^\infty e^{-\lambda x}\,\mathrm{d}x &= \left[-\frac{xe^{-\lambda x}}{\lambda}\right]_0^\infty + \frac{1}{\lambda}\int_0^\infty e^{-\lambda x}\,\mathrm{d}x \\ &= \frac{1}{\lambda}\left[-\frac{e^{-\lambda x}}{\lambda}\right]_0^\infty \\ &= \frac{1}{\lambda^2}\end{aligned}$$

Finally, multiply the λ from the original expression to obtain $1/\lambda$.

□

Definition B.3 (Memoryless Property)**.** A stochastic variable is memoryless iff

$$p(x > s + t \,|\, x > t) = p(x > s) \qquad s, t \geq 0$$

The exponential stochastic variable is memoryless.

B.3 Gaussian Distribution

We now give a separate treatment for the very common Gaussian distribution (aka normal distribution). A Gaussian or normal distribution is essentially what people imagine when they are talking about a "bell curve." The base function is $\exp(-x^2/2)$, whose graph is given in Figure B.2.

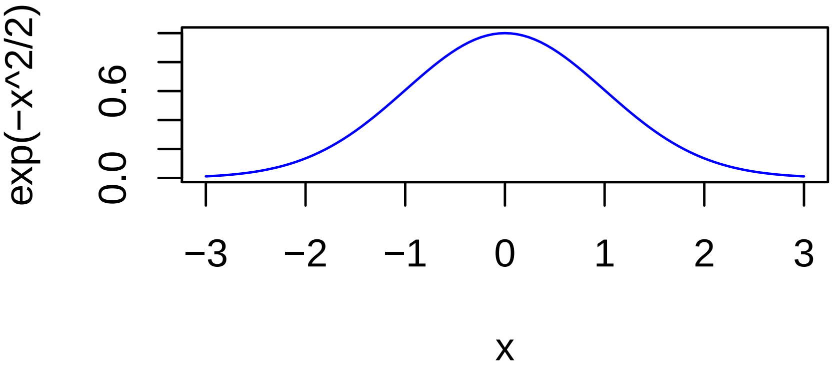


Figure B.2: Simple Gaussian

This function is known as the **standard normal**. However, we have not yet checked that it integrates to unity.

Integral of Gaussian. We want to find $I = \int_{\mathbb{R}} f(x)\,\mathrm{d}x$. However, a simplified I is not possible as is. The solution is a bit tricky.

$$
\begin{aligned}
I &= \sqrt{\left(\int_{\mathbb{R}} f(x)\,\mathrm{d}x\right)\left(\int_{\mathbb{R}} f(x)\,\mathrm{d}x\right)} \\
&= \sqrt{\left(\int_{\mathbb{R}} f(x)\,\mathrm{d}x\right)\left(\int_{\mathbb{R}} f(y)\,\mathrm{d}y\right)} \\
&= \sqrt{\iint_{\mathbb{R}\times\mathbb{R}} \left[f(x)f(y)\right]\,\mathrm{d}x\,\mathrm{d}y} \\
&= \sqrt{\iint_{\mathbb{R}\times\mathbb{R}} e^{-(x^2+y^2)/2}\,\mathrm{d}x\,\mathrm{d}y} \\
&= \sqrt{\int\limits_0^\infty \int\limits_0^{2\pi} e^{-r^2/2} r\,\mathrm{d}\theta\,\mathrm{d}r} \\
&= \sqrt{\left(\int\limits_0^{2\pi} \mathrm{d}\theta\right)\left(\int\limits_0^\infty \left[e^{-r^2/r} r\right]\mathrm{d}r\right)} \\
&= \sqrt{2\pi\left(e^{-r}\left(-1-r\right)\right)}\,\Bigg|_0^\infty \\
&= \sqrt{2\pi}
\end{aligned}
$$

□

The general form of the Gaussian includes a factor of $1/\sqrt{2\pi}$ for normalization:

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

Properties:

1. $\mathbb{E}[X] = \mu$
2. $\mathrm{Var}[X] = \sigma^2$

Mean of Gaussian Stochastic Variable. To show that the mean is μ, we first construct a new stochastic variable that is a function of X.

$$Z = \frac{X-\mu}{\sigma}$$

$$\mathbb{E}[Z] = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{\infty} z e^{-z^2/2}\,\mathrm{d}z$$

If we look at a plot of the function $\exp(-x^2/2)$ as in Figure B.2, it becomes apparent that this function is even. Multiplied by an odd function z, the integrand is odd with endpoints $[-a, a]$ (in this case, $a \to \infty$). Thus, $\mathbb{E}[Z] = 0$. Performing the appropriate shift, we find that this implies that $\mathbb{E}[X] = \mu$. □

Similarly, we can compute the variance of Z.

$$\mathrm{Var}[Z] = \sigma^2$$

C
Extra Formalism

C.1 Kolmogorov Axioms

There are three Kolmogorov axioms:

1. $\forall A \in \Omega, \mathbb{P}(A) \geq 0$ (There are no events with a negative probability of happening)

2. $\mathbb{P}(\Omega) = 1$ (The probability of *something* in the sample space happening must be 100%)

3. For a sequence of disjoint[1] sets $A_1, A_2, \ldots, \mathbb{P}\left(\bigcup_i A_i\right) = \sum_i \mathbb{P}(A_i)$ (The probability of mutually exclusive events happening is the total probability of any one happening)

[1] Disjoint sets have no elements in common. If A and B are disjoint, $A \cap B = \varnothing$