CHAPTER 1

Information Theory and Rate Distortion Theory

Please see below examples of equations using `\begin{equation} ...\end{equation}` and `align`.

The M&C macros support various environments like: constraint, construction, convention, conventions, corollary, definition, dictionary, example, lemma, note, notation, observation, property, proposition, remark, rules, theorem, etc.

What if you need an environment that the macros do not provide, like "guesswork?" You create it with `newMCtheorem`:

```
newMCtheorem}{guesswork}{GuessWork}
```

and use then use it:

GuessWork 1.1 This is pure guess work.

1.1 INTRODUCTION

Thus far in the book, the term *information* has been used sparingly and when it has been used, we have purposely been imprecise as to its meaning.

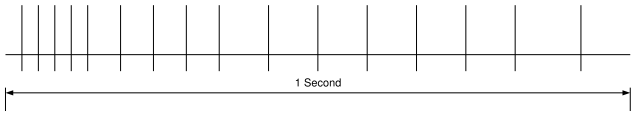

Figure 1.1: Communication system block diagram.

Examples of side-by-side figures.

1.2 ENTROPY AND AVERAGE MUTUAL INFORMATION

Consider a discrete random variable U that takes on the values $\{u_1, u_2, \ldots, u_M\}$, where the set of possible values of U is often called the *alphabet* and the elements of the set

Figure 1.2: Communication system block diagram.

(a) Annotated visualization of the structure of a biological neuron, reconstructed from electron microscope images of 30nm-thick slices of a mouse brain **?**.

(b) The shape of an action potential. A small external voltage stimulus (blue) triggers a cascade of charge build-up inside a neuron (red) via voltage-gated ion channels. The activation threshold is shown as a dotted line. Simulated using a Hodgkin-Huxley model of a neuron **?**.

Figure 1.3: The structure and operation of biological neurons.

are called *letters* of the alphabet. Let $P_U(u)$ denote the probability assignment over the alphabet, then we can define the *self-information* of the event $u = u_j$ by

$$I_U(u_j) = \log \frac{1}{P_U(u_j)} = -\log P_U(u_j). \tag{1.1}$$

Example 1.2 Given a random variable U with four equally likely letters in its alphabet, we wish to find $H(U)$. Clearly, $M = 4$ and $P_U(u_i) = \frac{1}{4}$ for $i = 1, 2, 3, 4$.

$$\begin{aligned} I_{W;X}(w_j; x_k) &= \log \frac{P_{WX}(w_j, x_k)}{P_W(w_j) P_X(x_k)} \\ &= \log \frac{P_{X|W}(x_k|w_j)}{P_X(x_k)} = I_{X;W}(x_k; w_j). \end{aligned} \tag{1.2}$$

Property 1.3 Let U be a random variable with possible values $\{u_1, u_2, \ldots, u_M\}$.

Example ?? illustrates Property ??.

Property 1.4 Let W and X be jointly distributed random variables.

Example 1.5 Here we wish to calculate the mutual information and the average mutual information for the probability assignments (with $M = 2$ and $N = 2$)

$$P_W(w_1) = P_W(w_2) = \tfrac{1}{2} \tag{1.3}$$

Example 1.6 ? The source output is a ternary-valued random variable that takes on the values $\{u_1, u_2, u_3\}$ with probabilities $P(u_1) = 0.7, P(u_2) = 0.15 = P(u_3)$.

Theorem 1.7 (Source Coding Theorem).. *For a DMS with entropy $H(U)$, the minimum average codeword length per source letter $(\bar{n})$ for any code is lower bounded by $H(U)$, that is, $\bar{n} \geq H(U)$, and further, $\bar{n}$ can be made as close to $H(U)$ as desired for some suitably chosen code.*

Theorem 1.8 (Channel Coding Theorem ?).. *Given a DMS with entropy H bits/source letter and a DMC with capacity C bits/source letter, if $H \leq C$, the source output can be encoded for transmission over the channel with an arbitrarily small bit error probability. Further, if $H > C$, the bit error probability is bounded away from 0.*

(a) Points in $\mathbb{R}^2$, subdivided by a single linear classifier. One simple way of understanding linear classifiers is as a line (or hyper-plane, in higher dimensions) which splits space into two regions. In this example, points above the line are mapped to class 1 (red); those below, to class 0 (blue).

(b) Points in $\mathbb{R}^2$, subdivided by a combination of four linear classifiers. Each classifier maps *all* points to class 0 or 1, and an additional linear classifier is used to combine the four. This hierarchical model is strictly more expressive than any linear classifier by itself.

Figure 1.4: Simple elements can be combined to express more complex relationships. This is one basic tenet of deep neural networks.

Proof. This result can be proved in several ways, including calculus of variations **?** inequality; however, an alternative method is used here. □

Table 1.1: Timer0 Compare Output Mode, non-PWM Mode

COM0x1-0	Description
00	Normal port operation
01	Toggle on Compare Match
10	Clear on Compare Match
11	Set on Compare Match

COM0x1-0	Description
00	Normal port operation
01	Toggle on Compare Match
10	Clear on Compare Match
11	Set on Compare Match

SUMMARY

In this chapter we have discussed very briefly some of the salient results from information theory and rate distortion theory and have indicated how these results can be used to bound communication system performance.

PROBLEMS

1.1. A random variable U has a sample space consisting of the set of all possible binary sequences of length N, denoted $\{u_j, j = 1, 2, \ldots, 2^N\}$.

1.2. Given a random variable U with the alphabet $\{u_1, u_2, u_3, u_4\}$ and probability assignments $P(u_1) = 0.8, P(u_2) = 0.1$, $P(u_3) = 0.05, P(u_4) = 0.05$, calculate the entropy of U. Compare your result to a random variable with equally likely values.