Chapter 2

ENTROPY AND CODING TECHNIQUES

2.1 INFORMATION AND ENTROPY

A binary digit, or “bit,” b, takes one of the values $b = 0$ or $b = 1$. A single bit has the ability to convey a certain amount of information – the information corresponding to the outcome of a binary decision, or “event,” such as a coin toss. If we have N bits, then we can identify the outcomes of N binary decisions.

Intuitively, the average amount of information associated with a binary decision depends upon prior knowledge which we have concerning the likelihoods of the possible outcomes. For example, there is little informative value to including snow conditions in the weather report during summer – in common parlance, the result is a foregone conclusion. By contrast, the binary events which convey most information on average are those which are equally likely. Similarly, the N-bit sequences which convey most information are those for which each bit has equally likely outcomes, regardless of the outcomes of the other bits in the sequence – loosely speaking, these are “entirely random” sequences of bits.

Source coding is the art of mapping each possible output from a given information source to a sequence of binary digits called “code bits.” Ideally, the mapping has the property that the code bits are “entirely random,” i.e., statistically independent, taking values of 0 and 1 with equal probability. In this way, the code bits convey the maximum possible amount of information. Then, provided the mapping is invertible, we can identify the number of code bits with the amount of information in the original source output.

The above concepts were formalized in the pioneering work of Claude Shannon [130]. A quantity known as "entropy" is defined in terms of the statistical properties of the information source. The entropy represents a lower bound on the average number of bits required to represent the source output. Moreover, it is possible to approach this lower bound arbitrarily closely. In fact, practical coding algorithms can achieve average bit rates which are extremely close to the entropy in many applications and when they do so the code bits must be entirely random.

2.1.1 MATHEMATICAL PRELIMINARIES

RANDOM VARIABLES AND VECTORS

Let X denote a random variable. Associated with the random variable is a set of possible outcomes, known as the aphabet, $\mathcal{A}_X$. The outcome of the random variable is denoted x, and is one of the elements of $\mathcal{A}_X$. A random variable is said to be discrete if its alphabet is finite or at most countably infinite. That is, we can enumerate the elements of the alphabet,

$$\mathcal{A}_X = \{\alpha_0, \alpha_1, \alpha_2, \ldots\}$$

In this case, the statistical properties of the random variable are described by its probability mass function (PMF)

$$f_X(x) \triangleq P(X = x) \quad \text{for each } x \in \mathcal{A}_X$$

In words, $f_X(x)$ is the probability of the outcome $X = x$. By contrast, a continuous random variable has uncountably many outcomes, e.g. $\mathcal{A}_X = \mathbb{R}$, the set of all real numbers. In this chapter we will be concerned exclusively with discrete alphabets. As an example, we model binary decisions as random variables whose alphabets have only two entries, usually written $\mathcal{A}_X = \{0, 1\}$. Binary random variables play a special role in coding.

The notion of a random variable is trivially extended to random vectors, $\mathbf{X}$, with alphabet, $\mathcal{A}_{\mathbf{X}}$ and PMF, $f_{\mathbf{X}}(\mathbf{x})$, for each vector, $\mathbf{x} \in \mathcal{A}_{\mathbf{X}}$. An m-dimensional random vector is a collection of m random variables, usually taken as a column vector,

$$\mathbf{X} = \begin{pmatrix} X_0 \\ X_1 \\ \vdots \\ X_{m-1} \end{pmatrix}$$

The PMF, $f_{\mathbf{X}}(\mathbf{x})$, is sometimes written longhand as

It denotes the probability that $X_0 = x_0$, $X_1 = x_1$, ..., and $X_{m-1} = x_{m-1}$ simultaneously. For this reason, it is often called the joint PMF, or joint distribution, for the m random variables.

From the joint distribution of a collection of m random variables, we can obtain the "marginal" distribution of any one of the random variables, X_i, as

$$f_{X_i}(x) = \sum_{\mathbf{x} \ni x_i = x} f_{\mathbf{X}}(\mathbf{x})$$

INDEPENDENCE AND CONDITIONAL PMF'S

We say that two random variables are statistically independent, or simply independent, if their joint distribution is separable; i.e.,

$$f_{X_0,X_1}(x_0, x_1) = f_{X_0}(x_0)\, f_{X_1}(x_1)$$

That is, the probability that both $X_0 = x_0$ and $X_1 = x_1$ is the product of the two marginal probabilities. As suggested by the introductory comments above, the notion of statistical independence plays an important role in coding.

We define the conditional distribution of X_1, given X_0, by

$$f_{X_1|X_0}(x_1, x_0) \triangleq \frac{f_{X_1,X_0}(x_1, x_0)}{f_{X_0}(x_0)} = \frac{f_{X_1,X_0}(x_1, x_0)}{\sum_x f_{X_1,X_0}(x, x_0)}$$

The function, $f_{X_1|X_0}(\cdot, x_0)$, is interpreted as a modified PMF for X_1, where the modification is to reflect the fact that the outcome $X_0 = x_0$ is already known. If the two random variables are statistically independent, we expect that the outcome of X_0 has no bearing on the distribution of X_1 and indeed we find that

$$f_{X_1|X_0}(x_1, x_0) = f_{X_1}(x_1) \text{ if and only if } X_1, X_0 \text{ are independent}$$

We note that the marginal distribution of X_0 and the conditional distribution of X_1, given X_0, together are equivalent to the joint distribution of X_1 and X_0. More generally, we write $f_{X_n|X_{n-1},\ldots,X_0}(x_n, \ldots, x_0)$ for the conditional distribution of X_n, given X_0 through X_{n-1}. The joint distribution of all m random variables of an m-dimensional random vector, X, may be recovered from

$$f_{\mathbf{X}}(\mathbf{x}) = f_{X_0}(x_0)\, f_{X_1|X_0}(x_1, x_0) \cdots f_{X_{m-1}|X_{m-2},\ldots,X_0}(x_{m-1}, \ldots, x_0) \tag{2.1}$$

and the random variables are said to be mutually independent if

EXPECTATION

The expectation of a random variable, X, is denoted $E[X]$ and defined by

$$E[X] \triangleq \sum_{x \in \mathcal{A}_X} x f_X(x)$$

It represents the statistical average or mean of the random variable X. Here, for the first time, we are concerned with the algebraic properties of random variables. More generally, let $g()$ be any function. We may define $Y = g(X)$ to be the random variable whose outcomes are $y = g(x)$ whenever the outcome of X is x. Consequently, the distribution of Y may be found from

$$f_Y(y) = \sum_{x \ni g(x) = y} f_X(x)$$

It is readily shown that the expectation of the new random variable, Y, satisfies

$$E[Y] = E[g(X)] = \sum_{y \in \mathcal{A}_Y} y f_Y(y) = \sum_{x \in \mathcal{A}_X} g(x) f_X(x) \qquad (2.2)$$

Given two random variables, X_0 and X_1, we may define conditional expectations in the most obvious way as

$$E[X_1 \mid X_0 = x_0] \triangleq \sum_{x \in \mathcal{A}_{X_1}} x f_{X_1|X_0}(x, x_0)$$

and for any function, $g()$, we have

$$E[g(X_1) \mid X_0 = x_0] = \sum_{x \in \mathcal{A}_{X_1}} g(x) f_{X_1|X_0}(x, x_0)$$

RANDOM PROCESSES

We conclude this section by introducing the concept of a discrete random process, denoted $\{X_n\}$. A random process is nothing but a sequence of individual random variables, X_n, $n \in \mathbb{Z}$, all having a common alphabet, $\mathcal{A}_X$. The key distinction from a random vector is that there are infinitely many random variables. The statistics are summarized by the vector PMF's, $f_{\mathbf{X}_{i:j}}()$, for all $i < j \in \mathbb{Z}$, where we use the notation, $\mathbf{X}_{i:j}$, to refer to the $(j - i)$-dimensional random vector formed from the elements, X_k, $i \le k < j$, of the random process.

The random process, $\{X_n\}$, is said to be stationary if the vector PMF's satisfy

That is, all collections of m consecutive random variables from the process have exactly the same joint distribution. Thus, a stationary random process is characterized by the PMF's, $f_{\mathbf{X}_{0:m}}$ for each $m = 1, 2, \ldots$. Alternatively, from equation (2.1) we see that stationary random processes are characterized by the marginal distribution, $f_{X_0} \equiv f_X$, together with the sequence of conditional distributions, $f_{X_m|\mathbf{X}_{0:m}}$, for $m = 1, 2, \ldots$.

In most applications we find that the conditional distributions satisfy

$$f_{X_m|\mathbf{X}_{0:m}} = f_{X_m|X_{m-p:m}} \tag{2.3}$$

for a sufficiently large value of the parameter, p. That is, the conditional distribution of X_m given X_0 through X_{m-1}, is actually a function of only the p most recent random variables, X_{m-p} through X_{m-1}. We say that X_m is "conditionally independent" of X_0 through X_{m-p-1}. Conditional independence is a phenomenon which we usually expect to encounter in the information sources which we model using random processes. Indeed statistical dependencies among samples taken from natural physical phenomena such as images and audio are generally of a local nature. For stationary processes, conditional independence means that the entire process is described by a finite number of conditional PMF's

$$f_{X_0},\ f_{X_1|X_0},\ f_{X_2|\mathbf{X}_{0:2}},\ \ldots,\ f_{X_p|\mathbf{X}_{0:p}}$$

These are called Markov random processes with parameter p. A Markov-1 random process is entirely described by f_X and $f_{X_1|X_0}$. If $p = 0$, all elements of the random process are statistically independent with identical distribution, f_X. Such a random process is said to be IID (Independent and Identically Distributed). It is also said to be "memoryless."

Stationary random processes with conditional independence properties (i.e. Markov processes) play an extremely important role in coding, precisely because they are described by a finite number of conditional PMF's. By observing the outcomes of the random process over a finite period of time, we can hope to estimate these conditional PMF's and use these estimates to code future outcomes of the random process. In this way, we need not necessarily have any a priori knowledge concerning the statistics in order to effectively code the source output. Adaptive coders are based on this principle.

The technical condition required to enable estimation of the relevant PMF's from a finite number of outcomes is "ergodicity." To be more precise, suppose we observe the outcomes of random variables X_0

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS

Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS

Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS

Sync your system to PACER to automate legal marketing.

WHAT WILL YOU BUILD? | sales@docketalarm.com | 1-866-77-FASTCASE