INTRODUCTION TO INFORMATION THEORY

Masud Mansuripur
College of Engineering
Boston University

PRENTICE-HALL, INC.
Englewood Cliffs, New Jersey 07632

1

MATHEMATICAL PRELIMINARIES

Introduction The purpose of this chapter is to familiarize the reader with some of the mathematical tools that are needed for the study of information theory. Although the tools are simple and their mastering requires no knowledge of advanced mathematics, the reader must have a certain level of mathematical maturity and sophistication in order to use them effectively. Knowledge of probability theory at an introductory level and familiarity with combinatorial techniques are the only prerequisites for the study of this book; although Section 1.1 is devoted to a review of probability, it is no substitute for formal education in this area. Section 1.2 describes the Chebyshev inequality and the weak law of large numbers. In Section 1.3 we define convexity and prove an important relationship, known as the Jensen inequality, for convex functions. A simple derivation of Stirling's approximation to the factorial function is given in Section 1.4.

1.1 REVIEW OF PROBABILITY

In a probability experiment the outcome can be any member of a given set E. The set E is known as the sample space of the experiment. Certain subsets of E are defined as events. If the outcome of a trial belongs to the event E_i, then E_i is said to have occurred. To each E_i, a probability p_i is assigned. The assignment is not arbitrary, and the collection of p_i's must satisfy the axioms of probability theory. For the most part, in this book we assume either that the set E is countable or that it is divided into a countable number of elementary events. In either case, we have a discrete probability space.

Example 1

In a coin-flipping experiment, $E = \{H, T\}$. If the coin is fair $P\{H\} = P\{T\} = 0.5$.

Example 2

If the experiment is to throw a die, then $E = \{1, 2, 3, 4, 5, 6\}$. If the die is fair, $P\{1\} = P\{2\} = \cdots = P\{6\} = \frac{1}{6}$, $P\{2, 5\} = \frac{1}{3}$, $P\{1, 3, 5\} = \frac{1}{2}$.

Example 3

Let the experiment be the weather forecast for a certain day; then $E = \{\text{sunny, cloudy, rainy, snowy}\}$. A possible probability assignment is $P\{\text{sunny}\} = \frac{1}{2}$, $P\{\text{cloudy}\} = \frac{1}{4}$, $P\{\text{rainy}\} = \frac{1}{8}$, $P\{\text{snowy}\} = \frac{1}{8}$.

Example 4

Take a typical English text and pick a letter at random; then $E = \{a, b, c, \ldots, x, y, z\}$. The probability distribution will be $P\{a\} = 0.0642, \ldots, P\{e\} = 0.103, \ldots, P\{z\} = 0.0005$.

If the outcome of an experiment is always the same, it is said to be a degenerate (or certain) experiment.

Example 5

In the English language, the letter that follows q in a word is always u. The experiment is to open a book at random and to pick the letter following the first q encountered. The sample space is $E = \{a, b, c, \ldots, x, y, z\}$, but $P\{u\} = 1$ and $P\{a\} = \cdots = P\{z\} = 0$.

A discrete random variable $\mathbf{x}$ is defined by assigning a real number x_i to each elementary event E_i in a discrete sample space. The probability of x_i is then denoted by $p(x_i)$. The average (or expected value) and the variance of $\mathbf{x}$ are defined as follows:

$$E(\mathbf{x}) = \bar{x} = \sum_i x_i p(x_i)$$

$$\begin{aligned}\text{Var}(\mathbf{x}) &= E((\mathbf{x} - \bar{x})^2) = \sum_i (x_i - \bar{x})^2 p(x_i) \\ &= \sum_i (x_i^2 + \bar{x}^2 - 2x_i\bar{x})p(x_i) \\ &= E(\mathbf{x}^2) + \bar{x}^2 - 2\bar{x} \cdot \bar{x} = E(\mathbf{x}^2) - \bar{x}^2\end{aligned}$$

where $E(\mathbf{x}^2)$ is the expected value of $\mathbf{x}^2$. In general, the expected value of a function of $\mathbf{x}$, such as $f(\mathbf{x})$, is given by

$$E(f(\mathbf{x})) = \sum_i f(x_i) \cdot p(x_i)$$

A pair of random variables $(\mathbf{x}, \mathbf{y})$ associated with an experiment forms a joint random variable. If $\mathbf{x}, \mathbf{y}$ are discrete, the joint probability distribution is defined as $\{p_{ij}\}$, where $p_{ij} = P\{\mathbf{x} = x_i, \mathbf{y} = y_j\}$.

Example 6

The experiment is to throw a coin and a die. The possible outcomes are

$$(\text{H}, 1), (\text{H}, 2), \ldots, (\text{H}, 6), (\text{T}, 1), (\text{T}, 2), \ldots, (\text{T}, 6)$$

If the coin and the die are fair and independent, $p_{ij} = \frac{1}{12}$ for all i, j. The diagram of Figure 1.1 is often helpful in visualizing the situation. Using the diagram, it is easy to verify that

(i) The sum of the probabilities p_{ij} is equal to 1.
(ii) $P\{\text{Coin} = \text{Head}\} = P\{(\text{H}, 1)\} + P\{(\text{H}, 2)\} + \cdots + P\{(\text{H}, 6)\} = \frac{1}{2}$.
(iii) $P\{\text{Die} = 5\} = P\{(\text{H}, 5)\} + P\{(\text{T}, 5)\} = \frac{1}{6}$.

Now assume that the coin and the die are somehow interacting and as a result their outcomes are not independent. The joint density may have the distribution shown in Figure 1.2. The following statements can be directly verified from the diagram.

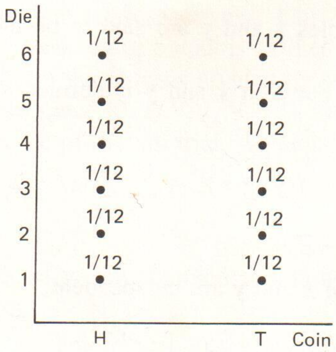


Figure 1.1

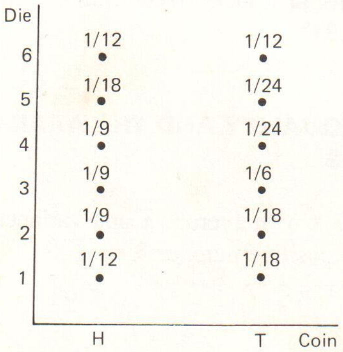


Figure 1.2

(i) $P\{(\mathrm{H}, 5)\} = \frac{1}{18}$
(ii) $P\{\text{Coin} = \mathrm{H}\} = \frac{5}{9}$
(iii) $P\{\text{Coin} = \mathrm{T}\} = \frac{4}{9}$
(iv) $P\{\text{Die} = 5\} = \frac{7}{72}$
(v) $P\{\text{Die} = 5 \text{ given Coin} = \mathrm{H}\} =$

$$\frac{1/18}{(1/12) + (1/18) + (1/9) + (1/9) + (1/9) + (1/12)} = 1/10$$

The conditional density of **y** given **x** is defined as

$$p(y_j \mid x_i) = \frac{p(x_i, y_j)}{p(x_i)}$$

where the marginal density $p(x_i)$ is assumed to be nonzero. The marginal densities are defined as follows:

$$p(x_i) = \sum_j p(x_i, y_j)$$

$$p(y_j) = \sum_i p(x_i, y_j)$$

The random variables **x** and **y** are said to be independent if $p(x_i, y_j) = p(x_i)p(y_j)$ for all i, j.

The correlation C between **x** and **y** is defined as the expected value of $(\mathbf{x} - \bar{x})(\mathbf{y} - \bar{y})$:

$$\begin{aligned} C(\mathbf{x}, \mathbf{y}) &= E((\mathbf{x} - \bar{x})(\mathbf{y} - \bar{y})) = \sum_i \sum_j (x_i - \bar{x})(y_j - \bar{y})p(x_i, y_j) \\ &= E(\mathbf{xy}) - \bar{x}\bar{y} \end{aligned}$$

In the special case where **x** and **y** are independent,

$$E(\mathbf{xy}) = \sum_i \sum_j x_i y_j p(x_i) p(y_j) = \sum_i x_i p(x_i) \sum_j y_j p(y_j) = \bar{x}\bar{y}$$

Independence thus results in a lack of correlation. The opposite, however, is not true (see Problem 1.3).

1.2 CHEBYSHEV INEQUALITY AND THE WEAK LAW OF LARGE NUMBERS

Given a random variable **x** with average $\bar{x}$ and variance σ_x^2, the Chebyshev inequality for an arbitrary positive number δ is

$$P\{|\mathbf{x} - \bar{x}| \geq \delta\} \leq \frac{\sigma_x^2}{\delta^2}$$

A simple proof of this result is given here (for another proof, see Problem 1.4). Define the function $f(x)$ as follows:

$$f(x) = \begin{cases} 1, & \text{if } |x - \bar{x}| \geq \delta \\ 0, & \text{if } |x - \bar{x}| < \delta \end{cases}$$

Then

$$P\{|\mathbf{x} - \bar{x}| \geq \delta\} = \sum f(x_i)p(x_i)$$

Upon inspection of Figure 1.3, it is obvious that

$$f(x) \leq \left[\frac{x - \bar{x}}{\delta}\right]^2$$

Therefore,

$$P\{|\mathbf{x} - \bar{x}| \geq \delta\} \leq \sum_i \left[\frac{x_i - \bar{x}}{\delta}\right]^2 p(x_i) = \frac{\sigma_x^2}{\delta^2}$$

The Chebyshev inequality can now be used to derive the weak law of large numbers. Consider a binary experiment where the outcomes are 0 and 1 with probabilities p_0 and $1 - p_0$, respectively. This experiment is repeated N times independently, and the average output is defined as $\mathbf{y}_N$; that is, $\mathbf{y}_N$ is equal to the total number of 1's in the N trials divided by N. Obviously, $\mathbf{y}_N$ is a random variable with sample space $\{0, 1/N, 2/N, \ldots, 1\}$. Defining $\mathbf{x}^{(n)}$ as the r. v. associated with the outcome of the nth trial, we have

$$\mathbf{y}_N = \frac{1}{N}\sum_{n=1}^{N} \mathbf{x}^{(n)}$$

The average and variance of $\mathbf{y}_N$ are obtained as follows:

$$\bar{y}_N = \frac{1}{N}\sum_{n=1}^{N} E(\mathbf{x}^{(n)}) = \frac{1}{N}\sum_{n=1}^{N} \bar{x} = \bar{x}$$

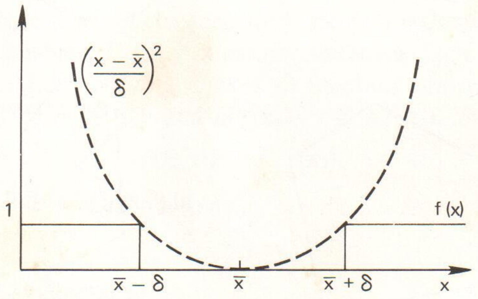


Figure 1.3

$$\sigma_y^2 = E((\mathbf{y}_N - \bar{y}_N)^2) = \frac{1}{N^2}E\left(\left[\sum_{n=1}^{N} (\mathbf{x}^{(n)} - \bar{x})\right]^2\right)$$
$$= \frac{1}{N^2}\sum_{n=1}^{N} E((\mathbf{x}^{(n)} - \bar{x})^2) = \frac{\sigma_x^2}{N}$$

For an arbitrary positive number ε, the Chebyshev inequality

$$P\{|\mathbf{y}_N - \bar{y}_N| \geq \varepsilon\} \leq \frac{\sigma_y^2}{\varepsilon^2}$$

leads to the following statement of the weak law of large numbers:

$$P\left\{\left|\left[\frac{1}{N}\sum_{n=1}^{N}\mathbf{x}^{(n)}\right] - \bar{x}\right| \geqslant \varepsilon\right\} \leqslant \frac{\sigma_x^2}{N\varepsilon^2}$$

Notice that the right side approaches zero with increasing N. The weak law of large numbers thus asserts that the sample average of $\mathbf{x}$ approaches the statistical average $\bar{x}$ with high probability as $N \to \infty$.

1.3 CONVEX SETS AND FUNCTIONS: JENSEN'S INEQUALITY

In the Euclidean space, a set S is convex if, for every pair of points P_1, P_2 in S, the straight line connecting P_1 to P_2 is completely contained in S. Figure 1.4a shows two convex sets in the two-dimensional Euclidean space. The sets of Figure 1.4b are not convex.

If $P_1 = (x_1, \ldots, x_n)$ and $P_2 = (y_1, \ldots, y_n)$ are points in the Euclidean n-space, then the straight line connecting them is represented by the set of points

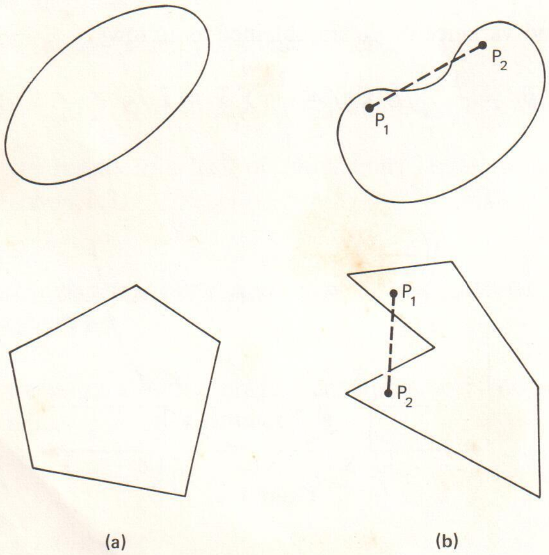


Figure 1.4

P, where $P = \lambda P_1 + (1 - \lambda)P_2 = [\lambda x_1 + (1 - \lambda)y_1, \ldots, \lambda x_n + (1 - \lambda)y_n]$, and λ is a real number in the interval $[0, 1]$. An important example of a convex set is the set of all n-dimensional probability distributions $\{p_1, \ldots, p_n\}$ (see Problem 1.6).

The real function $f(P)$, defined on a convex set S, is convex cap ($\cap$) if for every pair of points P_1, P_2 in the set and for every λ in $[0, 1]$ the following inequality is satisfied:

$$f[\lambda P_1 + (1 - \lambda)P_2] \geq \lambda f(P_1) + (1 - \lambda)f(P_2)$$

This property is shown graphically in Figure 1.5 for a function defined on a one-dimensional space. If the direction of inequality is reversed (for all P_1, P_2, λ), then the function is convex cup ($\cup$).

Theorem. If $\lambda_1, \ldots, \lambda_N$ are nonnegative numbers whose sum is unity, then, for every set of points $P_1, \ldots, P_N$ in the domain of the convex $\cap$ function $f(P)$, the following inequality is valid:

$$f\left(\sum_{n=1}^{N} \lambda_n P_n\right) \geq \sum_{n=1}^{N} \lambda_n f(P_n)$$

Proof. We use induction on N. For $N = 2$, the inequality is valid by definition of convexity. Assuming the validity of the theorem for $N - 1$, we prove it for N.

$$\begin{aligned} f\left(\sum_{n=1}^{N} \lambda_n P_n\right) &= f\left(\lambda_N P_N + (1 - \lambda_N) \sum_{n=1}^{N-1} \frac{\lambda_n}{1 - \lambda_N} P_n\right) \\ &\geq \lambda_N f(P_N) + (1 - \lambda_N) f\left(\sum_{n=1}^{N-1} \frac{\lambda_n}{1 - \lambda_N} P_n\right) \\ &\geq \lambda_N f(P_N) + (1 - \lambda_N) \sum_{n=1}^{N-1} \frac{\lambda_n}{1 - \lambda_N} f(P_n) \\ &= \sum_{n=1}^{N} \lambda_n f(P_n) \end{aligned}$$

The proof is thus complete.

A direct consequence of the preceding theorem is Jensen's inequality for discrete random variables. Let r. v. $\mathbf{x}$ assume the values $x_1, \ldots, x_n$ with probabilities $p_1, \ldots, p_n$. Let $f(x)$ be a convex $\cap$ function whose domain includes $x_1, \ldots, x_n$. Now $E(\mathbf{x}) = \sum_i p_i x_i$ and $E(f(\mathbf{x})) = \sum_i p_i f(x_i)$. Therefore,

$$f(E(\mathbf{x})) \geq E(f(\mathbf{x}))$$

This is known as Jensen's inequality.

1.4 STIRLING'S FORMULA

In this section we derive Stirling's formula, which provides upper and lower bounds to $N!$. Consider the function $\ln(x)$ whose integral from 1 to N is given

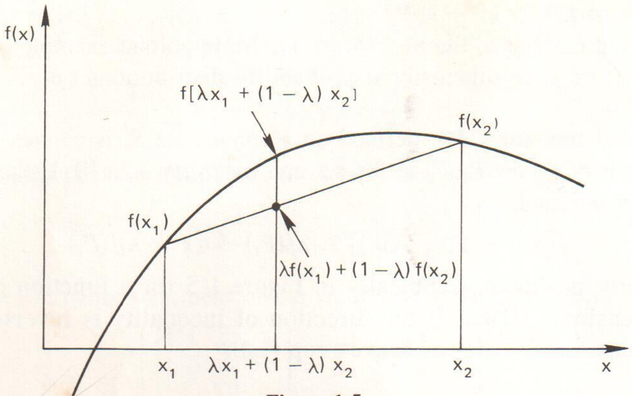


Figure 1.5

by

$$\int_1^N \ln(x)\,dx = [x\ln(x) - x]_1^N = N\ln(N) - N + 1$$

This integral is underapproximated by the trapezoid method of Figure 1.6a and overapproximated by the midpoint method of Figure 1.6b. Consequently,

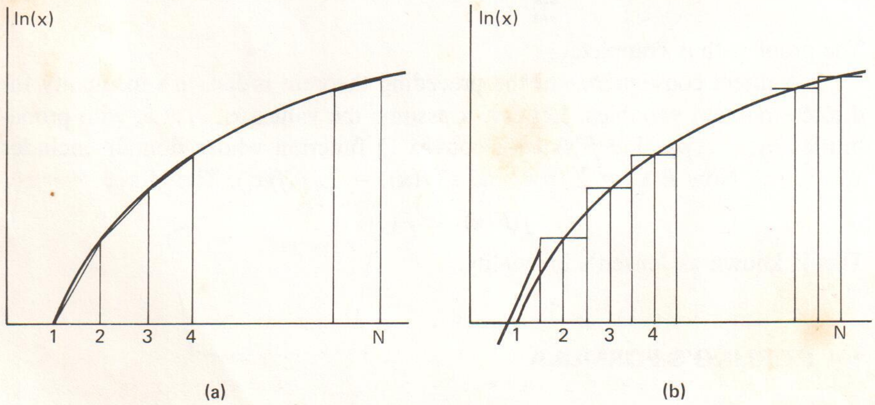
$$\int_1^N \ln(x)\,dx \geq \ln 2 + \ln 3 + \cdots + \ln(N-1) + \tfrac{1}{2}\ln(N) = \ln(N!) - \tfrac{1}{2}\ln(N)$$

$$\int_1^N \ln(x)\,dx \leq (\tfrac{1}{8}) + \ln 2 + \ln 3 + \cdots + \ln(N-1) + \tfrac{1}{2}\ln(N)$$

$$= \ln(N!) + (\tfrac{1}{8}) - \tfrac{1}{2}\ln(N)$$

Therefore,

$$N\ln(N) + \tfrac{1}{2}\ln(N) - N + (\tfrac{7}{8}) \leq \ln(N!) \leq N\ln(N) + \tfrac{1}{2}\ln(N) - N + 1$$



(a)

(b)

Figure 1.6

which yields Stirling's formula as

$$N^N e^{-N}\sqrt{N}e^{7/8} \le N! \le N^N e^{-N}\sqrt{N}e$$

PROBLEMS

1.1. Three events E_1, E_2, and E_3, defined on the same space, have probabilities $P(E_1) = P(E_2) = P(E_3) = \frac{1}{4}$. Let E_0 be the event that one or more of the events E_1, E_2, E_3 occurs.

(a) Find $P(E_0)$ when:

(1) E_1, E_2, E_3 are disjoint.

(2) E_1, E_2, E_3 are statistically independent.

(3) E_1, E_2, E_3 are three names for the same event.

(b) Find the maximum values that $P(E_0)$ can assume when:

(1) Nothing is known about the independence or disjointness of E_1, E_2, E_3.

(2) It is known that E_1, E_2, E_3 are pairwise independent; that is, the probability of realizing both E_i and E_j is $P(E_i)P(E_j)$, but nothing is known about the probability of realizing all three events together.

Hint: Use Venn diagrams.

1.2. A box contains two dice, one fair and the other loaded, so that for the first one $P(1) = P(2) = \cdots P(6) = \frac{1}{6}$ and for the second one $P(6) = \frac{2}{3}$, $P(1) = \cdots = P(5) = \frac{1}{15}$. We choose one die from the box at random and roll it. If the outcome is the number 6, what is the probability that the loaded die has been selected? What if the die is rolled twice and the outcome of both trials is the number 6?

1.3. Let **x** and **y** be discrete random variables,

(a) Prove that $E(\mathbf{x} + \mathbf{y}) = E(\mathbf{x}) + E(\mathbf{y})$

(b) If **x** and **y** are independent, prove that $E(\mathbf{xy}) = E(\mathbf{x})\cdot E(\mathbf{y})$; that is, **x** and **y** are uncorrelated.

(c) Is it possible for **x** and **y** to be dependent but uncorrelated? Give an example.

(d) If **x** and **y** are independent, prove that $\mathrm{Var}(\mathbf{x} + \mathbf{y}) = \mathrm{Var}(\mathbf{x}) + \mathrm{Var}(\mathbf{y})$. Is this relationship valid when **x** and **y** are dependent but uncorrelated?

1.4. **(a)** For any random variable **y** that assumes only nonnegative values, prove the following inequality:

$$P\{\mathbf{y} \ge \delta\} \le \frac{\overline{y}}{\delta}$$

where δ is an arbitrary positive number.

(b) Let **x** be a random variable with average $\overline{x}$ and variance σ_x^2. Define a nonnegative random variable $\mathbf{y} = (\mathbf{x} - \overline{x})^2$ and show that

$$P\{|\mathbf{x} - \overline{x}| \ge \delta\} \le \frac{\sigma_x^2}{\delta^2}$$

This is the Chebyshev inequality derived in Section 1.2.

1.5. A sequence of independent identically distributed random variables $\mathbf{y}_1, \ldots, \mathbf{y}_N$ with average $\overline{y}$ and standard deviation σ_y is given. Define the random variable **x** as follows:

$$\mathbf{x} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{y}_n$$

Show that the Chebyshev inequality for **x** is

$$P\{|\mathbf{x} - \bar{y}| \geq \delta\} \leq \frac{\sigma_y^2}{N\delta^2}$$

and investigate the limit when $N \to \infty$.

1.6. Prove that the set of all n-dimensional probability distributions $\{p_1, \ldots, p_n\}$ is convex. For $n = 3$, show the set graphically.

1.7. The function $f(x)$ is defined on the open interval (a, b) and is convex $\cap$. Prove that $f(x)$ is continuous. Would this conclusion be valid if the domain of the function were a closed interval?

1.8. The function $f(x)$ is defined and has second derivative on the interval (a, b). Prove that the necessary and sufficient condition for $f(x)$ to be convex $\cap$ is

$$\frac{d^2f(x)}{dx^2} \leq 0$$

for all points in (a, b).