ECE 1502 — Information Theory

Discrete Probability Refresher

F. R. Kschischang
Dept. of Electrical and Computer Engineering
University of Toronto

January 13, 1999 — revised January 11, 2006

Probability theory plays a central role in information theory. Information sources and communications channels are modelled probabilistically, and the key measures of information theory (like entropy and channel capacity) are defined in terms of the underlying random variables. The student of information theory is expected to have some familiarity with probability and the theory of random variables. In some cases, however, these ideas may not be as fresh in the student's memory as they could be. This set of notes is intended as an informal refresher of the basic notions of discrete probability, with an emphasis on those ideas that are needed in the study of information theory. Of course, a more formal and complete development can be found in most undergraduate or graduate texts on probability and random variables (e.g., [1, 2]).

1 Discrete Random Variables

A discrete random variable is used to model a "random experiment" with a finite or countable number of possible outcomes. For example, the outcome resulting from the toss of a coin, the roll of a die, or a count of the number of the telephone call attempts made during a given hour can all be modelled as discrete random variables.

The set of all possible outcomes is called the *sample space*, or *range*, or *alphabet* of the random variable in question. Here, "discrete" means that the sample space $\mathcal{S}$ is finite or countable, i.e., $\mathcal{S}$ can be placed into a one-to-one correspondence with a subset of the integers. For example, a coin toss has sample space $\{\text{heads}, \text{tails}\}$; a regular die roll has sample space {⚀, ⚁, ⚂, ⚃, ⚄, ⚅}; a count of telephone call attempts has sample space $\{0, 1, 2, \ldots\}$. This latter sample space is infinite, but still countable. It contains some very large numbers (like 10^{99}). While one may argue that the occurrence of such a large number of telephone call attempts is absurd, we find it convenient to include such outcomes. Later we can assign a vanishingly small (or even zero) probability to the absurd outcomes.

Let X be a random variable with sample space $\mathcal{S}_X$. A *probability mass function* (pmf) for X is a mapping

$$p_X : \mathcal{S}_X \to [0,1]$$

from $\mathcal{S}_X$ to the closed unit interval $[0,1]$ satisfying

$$\sum_{x \in \mathcal{S}} p_X(x) = 1.$$

The number $p_X(x)$ is the *probability* that the outcome of the given random experiment is x, i.e.,

$$p_X(x) := P[X = x].$$

Example. A <u>Bernoulli</u> random variable X has sample space $\mathcal{S}_X = \{0, 1\}$. The pmf is

$$\left\{ \begin{array}{l} p_X(0) = 1 - p, \\ p_X(1) = p \end{array} \right. , \; 0 \leq p \leq 1.$$

The sum of N independent[1] Bernoulli random variables, $Y = \sum_{i=1}^{N} X_i$ has $\mathcal{S}_Y = \{0, 1, \ldots, N\}$. The pmf for Y is

$$p_Y(k) = \binom{N}{k} p^k (1-p)^{N-k}, \quad k \in \mathcal{S}_Y.$$

This represents the probability of having exactly k heads in N independent coin tosses, where $P[\text{heads}] = p$.

Some Notation:

To avoid excessive use of subscripts, we will identify the a random variable by the letter used in the argument of its probability mass function, i.e., we will use the convention

$$\begin{aligned} p_X(x) &\equiv p(x) \\ p_Y(y) &\equiv p(y). \end{aligned}$$

Strictly speaking this is ambiguous, since the same symbol 'p' is used to identify two different probability mass functions; however, no confusion should arise with this notation, and we can always make use of subscripts to avoid ambiguity if necessary.

2 Vector Random Variables

Often the elements of the sample space $\mathcal{S}_X$ of a random variable X are real numbers, in which case X is a (real) scalar random variable. If the elements of $\mathcal{S}_X$ are vectors of real numbers, then X is a (real) vector random variable.

Suppose Z is a vector random variable with a sample space in which each element has has two components (X, Y), i.e.,

$$\begin{aligned} \mathcal{S}_Z &= \{z_1, z_2, \ldots\} \\ &= \{(x_1, y_1), (x_2, y_2), \ldots\}. \end{aligned}$$

[1] Independence is defined formally later.

The *projection* of $\mathcal{S}_Z$ on its first coordinate is

$$\mathcal{S}_X = \{x : \text{for some } y, (x, y) \in S_Z\}.$$

Similarly, the projection of $\mathcal{S}_Z$ on its second coordinate is

$$\mathcal{S}_Y = \{y : \text{for some } x, (x, y) \in S_Z\}.$$

Example. If $Z = (X, Y)$ and $\mathcal{S}_Z = \{(0, 0), (1, 0), (1, 1)\}$, then $\mathcal{S}_X = \mathcal{S}_Y = \{0, 1\}$.

In general, if $Z = (X, Y)$, then

$$\mathcal{S}_Z \subseteq \mathcal{S}_X \times \mathcal{S}_Y, \tag{1}$$

where

$$\mathcal{S}_X \times \mathcal{S}_Y = \{(x, y) : x \in \mathcal{S}_X, y \in \mathcal{S}_Y\}$$

is the Cartesian product of $\mathcal{S}_X$ and $\mathcal{S}_Y$. In general the containment relation (1) is strict, i.e., $\mathcal{S}_Z \neq \mathcal{S}_X \times \mathcal{S}_Y$. However, we can always define a new random variable Z' having the sample space $\mathcal{S}_{Z'} = \mathcal{S}_X \times \mathcal{S}_Y$. The sample space $\mathcal{S}_{Z'}$ is said to be in *product form.* The pmf of Z can be extended to a pmf for Z' by assigning zero probability to any events in $\mathcal{S}_{Z'}$ that do not appear in $\mathcal{S}_Z$. The random variable Z' will be indistinguishable from the random variable Z. Thus we can always assume that a vector random variable $Z = (X, Y)$ has a sample space in product form. This argument is easily extended to vector random variables having more than two components.

A vector random variable $Z = (X, Y)$ can be thought of as a combination of two random variables X and Y. The pmf for Z is also called the *joint* pmf for X and Y, and is denoted

$$\begin{aligned} p_Z(x, y) &= p_{X,Y}(x, y) \\ &= P[Z = (x, y)] \\ &= P[X = x, Y = y] \end{aligned}$$

where the comma in the last equation denotes a logical 'AND' operation.

From $p_{X,Y}(x, y)$ we can find $p_X(x)$:

$$p_X(x) \equiv p(x) = \sum_{y \in \mathcal{S}_Y} p_{X,Y}(x, y);$$

similarly,

$$p_Y(y) \equiv p(y) = \sum_{x \in \mathcal{S}_X} p_{X,Y}(x, y).$$

These probability mass functions are usually referred to as the *marginal* pmfs associated with vector random variable (X, Y).

Some More Notation:

Again, to avoid the excessive use of subscripts, we will use the convention

$$p_{X,Y}(x, y) \equiv p(x, y).$$

3 Events

An *event* A is a subset of the discrete sample space $\mathcal{S}$. The probability of the event A is

$$\begin{aligned} P[A] &= P[\text{some outcome contained in } A \text{ occurs}] \\ &= \sum_{x \in A} p(x). \end{aligned}$$

In particular,

$$\begin{aligned} P[\mathcal{S}] &= \sum_{x \in \mathcal{S}} p(x) = 1 \\ P[\phi] &= \sum_{x \in \phi} p(x) = 0, \end{aligned}$$

where ϕ is the empty (or null) event.

Example. A fair coin is tossed N times, and A is the event that an even number of heads occurs. Then

$$\begin{aligned} P[A] &= \sum_{\substack{k=0 \\ k \text{ even}}}^{N} P[\text{exactly } k \text{ heads occurs}] \\ &= \sum_{\substack{k=0 \\ k \text{ even}}}^{N} \binom{N}{k} (\frac{1}{2})^k (\frac{1}{2})^{N-k} \\ &= (\frac{1}{2})^N \sum_{\substack{k=0 \\ k \text{ even}}}^{N} \binom{N}{k} \\ &= \frac{2^{N-1}}{2^N} = \frac{1}{2}. \end{aligned}$$

4 Conditional Probability

Let A and B be events, with $P[A] > 0$. The *conditional probability* of B, given that A occurred, is

$$P[B|A] = \frac{P[A \cap B]}{P[A]}.$$

Thus, $P[A|A] = 1$, and $P[B|A] = 0$ if $A \cap B = \phi$.

Also, if $Z = (X, Y)$ and $p_X(x_k) > 0$, then

$$\begin{aligned} p_{Y|X}(y_j|x_k) &= P[Y = y_j | X = x_k] \\ &= \frac{P[X = x_k, Y = y_j]}{P[X = x_k]} \\ &= \frac{p_{X,Y}(x_k, y_j)}{p_X(x_k)} \end{aligned}$$

The random variables X and Y are *independent* if

$$\forall (x,y) \in \mathcal{S}_{X,Y} \left(p_{X,Y}(x,y) = p_X(x) p_Y(y) \right).$$

If X and Y are independent, then

$$p_{X|Y}(x|y) = \frac{p_{X,Y}(x,y)}{p_Y(y)} = \frac{p_X(x)p_Y(y)}{p_Y(Y)} = p_X(x),$$

and

$$p_{Y|X}(y|x) = \frac{p_{X,Y}(x,y)}{p_X(x)} = \frac{p_X(x)p_Y(y)}{p_X(X)} = p_Y(y),$$

i.e., knowledge of X does not affect the statistics of Y, and vice versa. As we will see later in the course, if X and Y are independent, then X provides no *information* about Y and vice-versa.

More generally, n random variables $X_1, \ldots, X_n$ are independent if their joint probability mass function factors as a product of marginals, i.e., if

$$p_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} p_{X_i}(x_i)$$

for all possible values $x_1, x_2, \ldots, x_n$. A collection $X_1, \ldots, X_n$ of random variables is said to be i.i.d. (independent, identically distributed) if they are independent and if the marginal pmfs are all the same, i.e., if $p_{X_i} = p_{X_j}$ for i and j.

Still More Notation:

Again, we'll avoid subscripts, and use the notation

$$p_{Y|X}(y|x) \equiv p(y|x).$$

In the simplified notation, $p(y|x) = p(x,y)/p(x)$ and $p(x|y) = p(x,y)/p(y)$. Similarly, in this notation, if $X_1, \ldots, X_n$ is a collection of independent random variables, the joint probability mass function $p(x_1, \ldots, x_n)$ factors as

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i).$$

5 Expected Value

If X is a random variable, the *expected value* (or mean) of X, denoted $E[X]$, is

$$E[X] = \sum_{x \in \mathcal{S}_X} x p_X(x).$$

The expected value of the random variable $g(X)$ is

$$E[g(X)] = \sum_{x \in \mathcal{S}_X} g(x) p_X(x).$$

In particular, $E[X^n]$, for n a positive integer, is the nth moment of X. Thus the expected value of X is the first moment of X. The *variance* of X, defined as the second moment of $X - E[X]$, can be computed as $\mathrm{VAR}[X] = E[X^2] - E[X]^2$. The variance is a measure of the "spread" of a random variable about its mean. Note that for any constant a, $E[aX] = aE[X]$ and $\mathrm{VAR}[aX] = a^2\mathrm{VAR}[X]$.

The *correlation* between two random variables X and Y is the expected value of their product, i.e., $E[XY]$. If $E[XY] = E[X]E[Y]$, then X and Y are said to be uncorrelated. Clearly if X and Y are independent, then they are uncorrelated, but the converse is not necessarily true.

If $X_1, X_2, \ldots, X_n$ is any sequence of random variables, then

$$E[X_1 + X_2 + \ldots + X_n] = E[X_1] + E[X_2] + \cdots + E[X_n],$$

i.e., the expected value of a sum of random variables is the sum of their expected values. If, in addition, $X_1, X_2, \ldots, X_n$ are pairwise uncorrelated, then the additive property holds also for the variance, i.e.,

$$\mathrm{VAR}[X_1 + X_2 + \cdots + X_n] = \mathrm{VAR}[X_1] + \mathrm{VAR}[X_2] + \cdots + \mathrm{VAR}[X_n].$$

6 The Markov and Chebyshev Inequalities

If X is a random-variable taking on *non-negative values only* and having expected value $E[X]$, then, for every value $a > 0$,

$$P[X \geq a] \leq \frac{E[X]}{a},$$

a result known as *Markov's Inequality.* This result can be derived from the following chain of inequalities. We have

$$\begin{aligned} E[X] &= \sum_{x \geq 0} xp(x) = \sum_{0 \leq x < a} xp(x) + \sum_{x \geq a} xp(x) \\ &\geq \sum_{x \geq a} xp(x) \\ &\geq \sum_{x \geq a} ap(x) \\ &= aP[X \geq a] \end{aligned}$$

Now if X is any random variable, then $Y = (X - E[X])^2$ is a random variable taking on non-negative values only, and hence Markov's Inequality applies. Take $a = k^2$ for some positive value k, we find

$$P[Y \geq k^2] = P[(X - E[X])^2 \geq k^2] = P[|X - E[X]| \geq k] \leq \frac{\mathrm{VAR}[X]}{k^2},$$

a result known as *Chebyshev's Inequality.*

7 The Weak Law of Large Numbers

Let $X_1, X_2, \ldots,$ be an i.i.d. sequence of random variables with mean m and finite variance σ^2. Suppose we observe the first n of these variables. An estimator for the mean m is then

$$M_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

As the following theorem shows, if n is sufficiently large, then with high probability M_n is close to the mean m.

Theorem 1 (The Weak Law of Large Numbers) *For all $\epsilon > 0$ and all $\delta > 0$ there exists a positive integer n_0 such that for all $n \geq n_0$,*

$$P[|M_n - m| \geq \epsilon] \leq \delta.$$

Proof: Note that M_n is a random variable with mean m and variance σ^2/n. It follows from Chebyshev's Inequality that

$$P[|M_n - m| \geq \epsilon] \leq \frac{\sigma^2}{n\epsilon^2}.$$

Take $n_0 = \lceil \sigma^2/(\epsilon^2\delta) \rceil$. Then for every $n \geq n_0$, we have $P[|M_n - m| \geq \epsilon] \leq \delta$. ∎

A more complicated argument would allow us to omit the requirement that the random variables have finite variance.

We sometimes write that $M_n \xrightarrow{p} m$ (read "M_n converges in probability to m"), meaning that $P[|M_n - m| \geq \epsilon] \to 0$ as $n \to \infty$.

References

[1] A. Leon-Garcia, *Probability and Random Processes for Electrical Engineering*, 2nd Edition. Don Mills, Ontario: Addison-Wesley, 1994.

[2] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, 2nd Edition. Toronto: McGraw-Hill, 1984.