THE TEACHING OF MATHEMATICS

2006, Vol. IX, 1, pp. 1–12

A CLASSROOM NOTE: ENTROPY, INFORMATION, AND MARKOV PROPERTY

Zoran R. Pop-Stojanović

Abstract. How to introduce the concept of the Markov Property in an elementary Probability Theory course? From this author's teaching experience, it appears that the best way that gives a natural intuitive flavor and preserves the mathematical rigor, is to use concepts of *entropy* and *information* from the classical Shannon Information Theory, as suggested in the brilliant monograph of A. Rényi [5]. Following this path, the connection between Entropy and Markov Property is presented.

ZDM Subject Classification: K65; *AMS Subject Classification*: Primary: 97D40, 60J10. Secondary: 94A15.

Key words and phrases: Random variable, Independence, Entropy, Information, Conditional Probability, Sufficient Function, Markov Chain.

1. Introduction

Anyone who has taught a Probability Theory course at any level encountered a quandary as how to introduce the concept of *Markov property* in such a way as to convey to students its intuitive meaning, while preserving fully the mathematical rigor that defines this concept. While teaching a course on Markov Processes on a graduate level, it appears that introduction of the Markov property in its most rigorous form is paradoxically also the easiest way to do the job! Of course, at that level, students have a sufficient background in the sophisticated machinery of the Theory of Measure, that allows this to be done. But, while teaching a Probability Theory course on Elementary and even Intermediate level, where a teacher has no a luxury of assuming the knowledge of concepts of the Measure Theory, this job is in a way, much more difficult. Although on this levels students are familiar with the concept of conditional probabilities, presented usually in discrete cases, that is, in cases of chains, the definition is too formal, and it is not quite clear as how from it one concludes that this property is a probabilistic analogue of a familiar property of dynamical systems, namely, if one has a system of particles and the position and velocities of all particles are known at the *present*, then the future evolution of such a system is fully determined. From the author's experience the best way for presenting the Markov property to students in elementary and even intermediate Probability Theory courses, is to follow the path outlined in A. Rényi's brilliant monograph *Foundations of Probability*, [5]. The main idea used there is, to express the Markov property via concepts of *Entropy and Information!* [2], [3]. Such a presentation is perfectly suited for our information age, where students are at ease in operating computers and cell phones. The mentioned monograph appeared in

1970—just before the dawn of the new computer age. By using elements of the Shannon Information Theory, the essence of the Markov property is easily understood. Its natural connection with conditional probabilities is easily established. Finally, following this development, that is also mentioned in [5], one arrives at *rooted random trees and forests*—one of the latest research topics in contemporary Probability Theory [4].

2. Setting

Throughout this exposition, all random variables are real-valued, and assuming a finite number of different values. They are all given on a fixed probability space $(\Omega, \mathcal{F}, \mathrm{P})$, where the sample space Ω is often finite, and σ-algebra $\mathcal{F}$ of random events, is the superset of Ω. This is a typical set-up that one finds in an elementary Probability Theory course. Now, following Shannon's ideas, one has the following

Definition 1. Let X be a random variable taking different values $x_1, x_2, \ldots, x_N$ such that $\mathrm{P}[X = x_k] = p_k,\ k = 1, 2, \ldots, N$. The *entropy* $H(X)$ of X is defined by

$$H(X) = \sum_{k=1}^{N} p_k \log_2 \frac{1}{p_k}. \tag{1}$$

Remark. Obviously, $H(X) \geq 0$, and $H(X) = 0$ if and only if $N = 1$, i.e., if X is a constant random variable. The entropy $H(X)$ of X could be interpreted as a measure of the amount of uncertainty carried by the value of random variable X, *but before observing the actual value of* X. Yet, the another interpretation of $H(X)$ is that it represents a measure of the amount of information received from X, but after the actual value of X has been observed. Equation (1) is known as *Shannon formula.* In short, the entropy measures the amount of uncertainty carried by a random variable.

As $H(X)$ depends only on the distribution $\mathcal{P} = \{p_1, p_2, \ldots, p_N\}$ of X, the following notation will be used as well.

$$H[\mathcal{P}] = \sum_{k=1}^{N} p_k \log_2 \frac{1}{p_k}. \tag{1'}$$

Now, let f be any function whose domain is the set $\{x_1, x_2, \ldots, x_N\}$ such that $f(x_i) \neq f(x_j)$, if $i \neq j$. Then, $\mathrm{P}[f(X) = f(x_k)] = \mathrm{P}[X = x_k] = p_k$, and

$$H(f(X)) = H(X). \tag{2}$$

More generally,

Lemma 1. *For any function f, one has*

$$H(f(X)) \leq H(X), \tag{3}$$

where the equality holds in (3) *if and only if,* $f(x_i) \neq f(x_j)$*, for* $i \neq j$*. (*X *is a random variable taking finite number of different values.)*

Proof. Let $\{y_1, y_2, \ldots, y_r\}$ denote the set of different values of $f(X)$, and put $D_j \equiv \{k;\ \text{such that } f(x_k) = y_j,\ j = 1, 2, \ldots, r\}$. Let

$$\forall j,\ q_j \equiv \mathrm{P}[f(X) = y_j] = \sum_{k \in D_j} p_k.$$

Then, one has:

$$H(X) - H(f(X)) = \sum_{j=1}^{r} q_j H[\mathcal{P}_j], \tag{4}$$

where $\mathcal{P}_j$ denotes the probability distribution $\{\frac{p_k}{q_j};\ k \in D_j\}$. Since $\forall j,\ H[\mathcal{P}_j] \geq 0$, with equality sign holding if and only if D_j is a singleton, the inequality (3) follows. ■

Definition 2. Let $\mathcal{P} = \{p_1, p_2, \ldots, p_N\}$ and $\mathcal{Q} = \{q_1, q_2, \ldots, q_N\}$ be two finite probability distributions containing the equal number N, $(N \geq 2)$, of positive, and different terms. The *divergence* $D(\mathcal{P}, \mathcal{Q})$ of distribution $\mathcal{P}$ from distribution $\mathcal{Q}$ is defined by the formula:

$$D(\mathcal{P}, \mathcal{Q}) = \sum_{k=1}^{N} p_k \log_2 \frac{p_k}{q_k}. \tag{5}$$

Observe that the divergence $D(\mathcal{P}, \mathcal{Q})$ may be interpreted as a measure of the discrepancy of the two distributions $\mathcal{P}$ and $\mathcal{Q}$. It is asymmetric, i.e., in general, $D(\mathcal{P}, \mathcal{Q}) \neq D(\mathcal{Q}, \mathcal{P})$. Also, notice that definition of $D(\mathcal{P}, \mathcal{Q})$ depends on the labelling of terms of distributions $\mathcal{P}$ and $\mathcal{Q}$.

Lemma 1. *Let* $\mathcal{P}$ *and* $\mathcal{Q}$ *be any pair of distributions for which* $D(\mathcal{P}, \mathcal{Q})$ *is defined. Then, one has:*

$$D(\mathcal{P}, \mathcal{Q}) \geq 0. \tag{6}$$

The equality sign holds if and only if $p_k = q_k,\ k = 1, 2, \ldots, N$.

Proof. Using the inequality $\ln(1 + x) \leq x,$ for $x > -1$, where the equality sign holds only for $x = 0$, one gets:

$$-D(\mathcal{P}, \mathcal{Q}) \ln 2 = \sum_{k=1}^{N} p_k \ln \left(1 + \frac{q_k - p_k}{p_k}\right) \leq \sum_{k=1}^{N} (q_k - p_k) = 0,$$

thus showing (6). ■

This lemma has several important consequences.

Corollary 1. *Given a discrete random variable* X *taking on* N *different values. Then, its entropy is maximal if and only if* $p_k = \frac{1}{N},\ k = 1, 2, \ldots, N$*. In this case,* $H(X) = \log_2 N$.

Proof. Denote by $\mathcal{P}_N$ the probability distribution $\{p_1, p_2, \ldots, p_N\}$ and let $\mathcal{U}_N$ denote the uniform distribution $\{\frac{1}{N}, \ldots, \frac{1}{N}\}$. Then, using the previous lemma, one gets:

$$\log_2 N - H(X) = D(\mathcal{P}_N, \mathcal{U}_N) \geq 0. \quad \blacksquare \tag{7}$$

Here is an important consequence of this corollary: the uncertainty concerning the value of a discrete random variable X that takes N different values is maximal, if all these values are *equiprobable.*

One of the basic inequalities used in Information Theory is given by the following

COROLLARY 2. *If $\mathcal{P} = \{p_1, p_2, \ldots, p_n\}$ is a probability distribution and $\mathcal{Q} = \{q_1, q_2, \ldots, q_N\}$ is a set of N positive numbers such that $\sum_{k=1}^{N} q_k \leq 1$, then,*

$$\sum_{k=1}^{N} p_k \log_2 \frac{p_k}{q_k} \geq 0. \quad \blacksquare$$

Proof. Set $\lambda \equiv \sum_{k=1}^{N} q_k$ and ${q_k}' = \frac{q_k}{\lambda}$, $k = 1, 2, \ldots, N$. Then, $\mathcal{Q}' = \{{q_1}', \ldots, {q_N}'\}$ is a probability distribution, and by Lemma 1, $D(\mathcal{P}, \mathcal{Q}') \geq 0$. Thus,

$$\sum_{k=1}^{N} p_k \log_2 \frac{p_k}{q_k} = \log_2 \frac{1}{\lambda} + D(\mathcal{P}, \mathcal{Q}') \geq 0. \quad \blacksquare$$

Now one can introduce the basic concept that will be used in characterization of the Markov Property.

DEFINITION 3. The joint information $I(X, Y)$ of random variables X and Y where each one of them takes only a finite number of different values, is defined by:

$$I(X, Y) \equiv H(X) + H(Y) - H((X, Y)), \tag{8}$$

where $H((X, Y))$ denotes the entropy of the random vector (X, Y).

Observe that if different values of X and Y are $x_1, \ldots, x_N$ and $y_1, \ldots, y_M$, respectively, and if the joint distribution of X and Y is $r_{jk} = \mathrm{P}[X = x_j, Y = y_k]$, $j = 1, 2, \ldots, N$, $k = 1, 2, \ldots, M$, then the entropy $H((X, Y))$ of the random vector (X, Y), is given by:

$$H((X, Y)) = \sum_{j=1}^{N} \sum_{k=1}^{M} r_{jk} \log_2 \frac{1}{r_{jk}}.$$

REMARK. Put $p_j = \mathrm{P}[X = x_j]$, $j = 1, 2, \ldots, N$, and $q_k = \mathrm{P}[Y = y_k]$, $k = 1, 2, \ldots, M$. Then one has:

$$\sum_{k=1}^{M} r_{jk} = p_j, \ j = 1, 2, \ldots, N, \text{ and}$$

$$\sum_{j=1}^{N} r_{jk} = q_k, \ k = 1, 2, \ldots, M.$$

Hence,

$$H(X) = \sum_{j=1}^{N} p_j \log_2 \frac{1}{p_j} = \sum_{j=1}^{N}\sum_{k=1}^{M} r_{jk} \log_2 \frac{1}{p_j},$$
$$H(Y) = \sum_{k=1}^{M} q_k \log_2 \frac{1}{q_k} = \sum_{j=1}^{N}\sum_{k=1}^{M} r_{jk} \log_2 \frac{1}{q_k}.$$

Thus,

$$I(X,Y) = \sum_{j=1}^{N}\sum_{k=1}^{M} r_{jk} \log_2 \frac{r_{jk}}{p_j q_k}.$$

REMARK. Let $\mathcal{R}$ denote the joint probability distribution of random variables X and Y, and let $\mathcal{PQ}$ be probability distribution $\{p_j q_k;\ j = 1, 2, \ldots, N,\ k = 1, 2, \ldots, M\}$, that is, $\mathcal{PQ}$ would be the joint distribution of random variables X and Y, *were they independent*, as *they are not assumed to be here*. Then

$$I(X,Y) = D(\mathcal{R}, \mathcal{PQ}). \tag{9}$$

From (9) and Lemma 2, one gets the following

THEOREM 1. *For any two random variables X and Y, one has:*

$$I(X,Y) \geq 0. \tag{10}$$

The equality sign holds in (10) if and only if X and Y are independent.

A very important ramification of this theorem: in classical Shannon Information Theory, $I(X,Y)$ is interpreted as *the amount of information gained about the random variable X based on observations of random variable Y*. From the definition of $I(X,Y)$, it follows that $I(X,Y) = I(Y,X)$, that is, amounts of information gained about one of the random variables based on observations of the other, are equal. This amount of information is always nonnegative and equal to zero *if and only if X and Y are independent random variables*. Or, two paraphrase this, *two random variables are independent if and only if by observing one of them, one gets no information about the other*. Hence, Theorem 1 sheds a new light on the meaning of independence—one of the basic concepts that is introduced in an elementary Probability Theory course. Actually, this theorem could be used as the definition of independence of two random variables. By defining independence this ways, students will get a clear intuitive meaning of this concept. To elaborate further on this, one recalls the definition of conditional probabilities:

$$p_{j|k} \equiv \mathrm{P}[X = x_j | Y = y_k] = \frac{r_{jk}}{q_k},\ q_k > 0;$$
$$q_{j|k} \equiv \mathrm{P}[Y = y_k | X = x_j] = \frac{r_{jk}}{p_j},\ p_j > 0.$$

Denote by $\mathcal{P}_k$ the distribution $\{p_{j|k};\ j = 1, 2, \ldots, N\}$ for $k = 1, 2, \ldots, M$, and by $\mathcal{Q}_j$ the distribution $\{q_{k|j};\ k = 1, 2, \ldots, M, \}$ for $j = 1, 2, \ldots, N$. Thus, $\mathcal{P}_k$ is

the conditional distribution of X given $Y = y_k,\ k = 1, 2, \ldots, M$, and $\mathcal{Q}_j$ is the conditional distribution of Y given $X = x_j,\ j = 1, 2, \ldots, N$. Put

$$H(X|Y) = \sum_{k=1}^{M} q_k H[\mathcal{P}_k], \text{ and}$$
$$H(Y|X) = \sum_{j=1}^{N} \mathrm{P}_j H[\mathcal{Q}_j].$$

The quantity $H(X|Y)$, (resp. $H(Y|X)$), is interpreted as the average conditional entropy of X given Y. (Resp. of Y given X). It is clear that

$$H(X|Y) = H((X, Y)) - H(Y), \tag{11}$$

and from Lemma 1 it follows that

$$H(X|Y) \geq 0, \tag{12}$$

since Y is a function of (X, Y), and thus

$$H(Y) \leq H((X, Y)).$$

Furthermore, it follows from (11) that

$$I(X, Y) = H(X) - H(X|Y) = H(Y) - H(Y|X). \tag{13}$$

Formula (13) has the following, and very important interpretation: the information gained about X by observing Y is equal to the decrease of uncertainty concerning X obtained by observing Y (and conversely). It follows from (12) that

$$H(X|Y) \leq H(X). \tag{14}$$

Thus, one has the following corollary to Theorem 1: by observing Y, the uncertainty concerning X is nonincreasing; it remains unchanged if and only if random variables X and Y are independent. Another way of paraphrasing Theorem 1 is to state that for any pair of random variables X and Y, one has

$$H((X, Y)) \leq H(X) + H(Y),$$

where the equality sign holds if and only if X and Y are independent. From (13) one gets that

$$I(X, Y) \leq H(X). \tag{15}$$

Observe that in (15) equality sign holds if and only if X is w.p. 1 constant for a given value of random variable Y. That is, if X is a function of Y. In particular,

$$I(X, X) = H(X). \tag{16}$$

Hence, the entropy of a random variable X measures the amount of information concerning itself that is contained in its value. Since $I(X, Y)$ is symmetric, (15) implies that

$$I(X, Y) \leq \min(H(X), H(Y)).$$

Theorem 1 demonstrated that a close relationship between the notion of independence of two random variables, and that of information. This relationship can now be extended to a case of more than two independent random variables, and it is described by the following theorem.

THEOREM 2. *Given a finite set of jointly distributed random variables* $\{X_i;\ i = 1, 2, \ldots, n,\ n > 2\}$ *on probability space* $(\Omega, \mathcal{F}, \mathcal{P})$. *Assume that for every* $i,\ i = 1, 2, \ldots, n,\ X_i$ *takes only a finite number of different values. Denote by* $H((X_1, X_2, \ldots, X_n))$ *the entropy of the joint distribution of* $X_1, X_2, \ldots, X_n$. *Then,*

$$H((X_1, X_2, \ldots, X_n)) \leq \sum_{k=1}^{n} H(X_k), \tag{17}$$

where the equality sign holds if and only if $X_1, X_2, \ldots, X_n$ *are independent.*

Proof. Let $x_{kj},\ j = 1, 2, \ldots, N_k$, denote the set of different values taken by random variables $X_k,\ k = 1, 2, \ldots, n$. Put $p_{kj} = \mathrm{P}[X_k = x_{kj}],\ k = 1, 2, \ldots, n,\ j = 1, 2, \ldots, N_k$. Denote by $\mathcal{P}_k$ the distribution $\{p_{k1}, p_{k2}, \ldots, p_{kN_k}\}$ and by $\mathcal{P}_1\mathcal{P}_2 \cdots \mathcal{P}_n$ the probability distribution whose terms are $p_{1j_1}p_{2j_2} \cdots p_{nj_n}$. Let $\mathcal{R}$ denote the joint distribution of the random variables $X_1, X_2, \ldots, X_n$, that is, $\mathcal{R} = \{r(j_1, \ldots, j_n)\}$, where

$$r(j_1, \ldots, j_n) = \mathrm{P}[X_1 = x_{1j_1}, \ldots, X_n = x_{nj_n}].$$

Then, one has:

$$H((X_1, X_2, \ldots, X_n)) - \sum_{k=1}^{n} H(X_k) = \sum_{j_1=1}^{N_1} \cdots \sum_{j_n=1}^{N_n} r(j_1, \ldots, j_n) \log_2 \frac{r(j_1, \ldots, j_n)}{p_{1j_1} \cdots p_{nj_n}}.$$

Thus,

$$H((X_1, X_2, \ldots, X_n)) - \sum_{k=1}^{n} H(X_k) = D(\mathcal{R}, \mathcal{P}_1\mathcal{P}_2 \cdots \mathcal{P}_n) \geq 0, \tag{18}$$

where the equality sign holds in (18) if and only if $\mathcal{R} = \mathcal{P}_1\mathcal{P}_2 \cdots \mathcal{P}_n$, i.e., if $X_1, X_2, \ldots, X_n$ are independent. ■

DEFINITION 4. Put

$$I(X_1, X_2, \ldots, X_n) \equiv H((X_1, X_2, \ldots, X_n)) - \sum_{k=1}^{n} H(X_k). \tag{19}$$

$I(X_1, X_2, \ldots, X_n)$ is called the joint information provided by random variable $X_1, X_2, \ldots, X_n$.

Observe that (19) generalizes Definition 1. The immediate consequence of Theorem 2 is the following corollary.

COROLLARY 3. *Random variables* $\{X_k,\ k = 1, 2, \ldots, n\}$ *are independent if and only if their joint information is equal to zero.*

Another re-statement of Theorem 2 is the following corollary.

COROLLARY 4. *Random variables* $\{X_k;\ k = 1, 2, \ldots, n\}$ *are independent if and only if for each* $k,\ k = 1, 2, \ldots, n-1$, *one has:*

$$I((X_1, X_2, \ldots, X_k), X_{k+1}) = 0. \tag{20}$$

REMARK. What does (20) say? It states, that when X_ks are independent, by observing $X_1, X_2, \ldots, X_k,\ k = 1, 2, \ldots, n-1$, *one does not get information about* X_{k+1}.

Proof. For $k \geq 2$ one has that

$$I(X_1, \ldots, X_{k+1}) - I(X_1, \ldots, X_k) = I((X_1, \ldots, X_k), X_{k+1}). \tag{21}$$

After summing (21) for $k = 2, 3, \ldots, n-1$, and after adding $I(X_1, X_2)$ to both sides of the equality, one gets:

$$I(X_1, X_2, \ldots, X_n) = \sum_{k=1}^{n-1} I((X_1, \ldots, X_k), X_{k+1}). \tag{22}$$

Using (22), one gets the statement of the corollary. ■

Observe that although $I(X_1, X_2, \ldots, X_n)$ is invariant with respect to any permutation of random variables $X_1, X_2, \ldots, X_n$, the right-hand side of (22) is not invariant with respect to permutations of X_ks. It can be shown that there are $(2n-3)!!$ decompositions of the right-hand side of (22). More precisely, this number represents the number of decomposition of $I(X_1, X_2, \ldots, X_n)$ into the sum of $n-1$ joint informations. (This number is obtained by considering oriented rooted trees). However, to each of these decompositions for joint information corresponds a criterion of independence of random elements in question. This leads to the following theorem.

THEOREM 3. *Let* $X_1, X_2, \ldots, X_n$ *be a set of independent random variables given on probability space* $(\Omega, \mathcal{F}, \mathcal{P})$, *each of them taking on a finite number of different values, and let* Z *be an arbitrary random variable given on the same probability space and also taking on a finite number of different values. Then,*

$$\sum_{k=1}^{n} I(X_k, Z) \leq I((X_1, X_2, \ldots, X_n), Z). \tag{23}$$

Proof. It suffices to prove (23) for $n = 2$, and then use the induction argument to show its validity for $n > 2$. Assume X, Y are independent random variables satisfying the assumptions of the theorem, and whose values are denoted by x, y, respectively. Likewise, let Z be an arbitrary random variable satisfying the assumptions of the theorem, and whose values are denoted by z. Then, one has to show

$$I(X, Z) + I(Y, Z) \leq I((X, Y), Z). \tag{24}$$

From the definition of information, it follows that

$$I((X,Y),Z) - I(X,Z) - I(Y,Z) = D(\mathcal{P},\mathcal{Q}),$$

where $\mathcal{P}$ is the joint distribution of X, Y and Z, that is, $\mathcal{P} = \{\mathrm{P}[X = x, Y = y, Z = z]\}$, and $\mathcal{Q}$ is the probability distribution whose terms are

$$Q(x,y,z) = \frac{\mathrm{P}[X = x, Z = z]\mathrm{P}[Y = y, Z = z]}{\mathrm{P}[Z = z]}.$$

Then, with all this in place, (24) follows from Lemma 2. The induction argument completes the proof. ■

Remark. The notion of the joint information can be extended for arbitrary random variables. However, we are not going to do it here. For example, if X, Y are continuous random variables having the joint probability density function $f_{X,Y}$ and whose marginal probability densities are f_X, f_Y, respectively, then their joint information is given by:

$$I(X,Y) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_{X,Y}(x,y)\log_2\frac{f_{X,Y}(x,y)}{f_X(x)f_Y(y)}\,\mathrm{d}x\,\mathrm{d}y,$$

provided that the integral on the right-hand-side of this equation converges.

3. Sufficient Function

The main ingredient for formulating the Markov property in terms of Entropy and Information, is the concept of a *sufficient function*. It is introduced by the following definition.

Definition 5. Let X, Y be random variables given on a probability space $(\Omega, \mathcal{F}, \mathcal{P})$ such that $I(X,Y)$ is finite. Given a real-valued, Borel measurable function g. Then, the random variable $g \circ X \equiv g(X)$ is called *a sufficient function of random variable X for random variable Y*, if one has

$$I(g(X),Y) = I(X,Y).$$

In other words, *$g(X)$ is sufficient for Y if $g(X)$ contains all information on random variable Y provided by random variable X*. The main characterization of a sufficient function, that connects concepts of information and conditional probabilities, and that will be used in formulation of the Markov property, is given by the following theorem.

Theorem 4. *Given random variables X and Y each of them taking on a finite number of values $x_1, \ldots, x_N$ and $y_1, \ldots, y_M$, respectively. Let g be a real-valued, Borel function. Then, $g(X)$ is a sufficient function for Y if and only if the conditional probability distribution of Y given a value of X, depends on the value of $g(X)$ only, that is,*

$$\mathcal{P}[Y = y_k | X = x_i] = \mathcal{P}[Y = y_k | X = x_j] \tag{25}$$

whenever $g(x_i) = g(x_j)$, $i = 1, 2, \ldots, N$, $j = 1, 2, \ldots, N$.

Observe that relation (25) could be expressed differently, namely, X and Y are independent random variables when the value of $g(X)$ is fixed, that is, for any z such that $\mathrm{P}[g(X) = z] > 0$, one has:

$$\mathrm{P}[X = x_j, Y = y_k | g(X) = z] = \mathrm{P}[X = x_j | g(X) = z]\mathrm{P}[Y = y_k | g(X) = z]. \tag{26}$$

Proof. Denote by $\mathcal{P} = \{p_j;\ j = 1, 2, \ldots, N\}$, where $p_j = \mathrm{P}[X = x_j]$ the distribution of X, and by $\mathcal{Q} = \{q_k;\ k = 1, 2, \ldots, M\}$, where $q_k = \mathrm{P}[Y = y_k]$, the distribution of Y. Denote by $\mathcal{R} = \{r_{jk};\ j = 1, 2, \ldots, N,\ k = 1, 2, \ldots, M\}$, where $r_{jk} = \mathrm{P}[X = x_j, Y = y_k]$ the joint distribution of X, Y. Let $\{z_1, z_2, \ldots, z_s\}$ be the set of different values of function g taken on the set $\{x_1, x_2, \ldots, x_N\}$. Put $g(x_j) = z_l,\ j \in D_l$, where $D_1, D_2, \ldots, D_s$ is a partition of the set $\{1, 2, \ldots, N\}$. Furthermore, define $v(x_j) = l$ if $j \in D_l$, $t_{lk} = \mathrm{P}[g(X) = z_l, Y = y_k]$ and $t_l = \mathrm{P}[g(X) = z_l]$. The numbers

$$u_{jk} = \frac{t_{v(x_j)k} \cdot p_j}{t_{v(x_j)}},\ j = 1, 2, \ldots, N,\ k = 1, 2, \ldots, M, \tag{27}$$

form a probability distribution $\mathcal{U} = \{u_{jk}\}$, and one has:

$$I(X, Y) - I(g(X), Y) = D(\mathcal{R}, \mathcal{U}), \tag{28}$$

thus, $I(X, Y) = I(g(X), Y)$ if and only if $\mathcal{R} = \mathcal{U}$; that is, if

$$\frac{t_{v(x_j)k} \cdot p_j}{t_{v(x_j)}} = r_{jk}, \tag{29}$$

i.e., if and only if

$$\frac{r_{ik}}{p_i} = \frac{r_{jk}}{p_j} \text{ whenever } g(x_i) = g(x_j). \tag{30}$$

Hence, (25) is the necessary and sufficient condition for

$$I(g(X), Y) = I(X, Y).$$

Furthermore, if $g(x_j) = z_l$, then

$$\begin{aligned} \mathrm{P}[X = x_j, Y = y_k | g(X) = z_l] &= \frac{\mathrm{P}[X = x_j, Y = y_k]}{\mathrm{P}[g(X) = z_l]} \\ &= \mathrm{P}[X = x_j | g(X) = z_l]\mathrm{P}[Y = y_k | X = x_j]. \end{aligned}$$

Hence, (26) is equivalent to

$$\mathrm{P}[Y = y_k | X = x_j] = \mathrm{P}[Y = y_k | g(X) = z_l]$$

if $g(x_j) = z_l$. Thus, (26) is equivalent to (25). ∎

4. Markov Chains

Using concepts of *entropy* and *information* with their basic characterizations just introduced, now one can introduce a Markov Chain—a concept close to that of *independence*. Here, the simplest case of a Markov Chain will be introduced: that one depending on a discrete parameter, and with a finite number of states.

DEFINITION 6. A sequence $\{X_n;\ n = 1, 2, \dots\}$ of real-valued random variables given on a probability space $(\Omega, \mathcal{F}, \mathcal{P})$, each taking on a finite number of values, is called a *discrete-parameter Markov Chain* if

$$I((X_1, X_2, \dots, X_n), X_{n+1}) = I(X_n, X_{n+1}),\ n = 1, 2, \dots. \tag{31}$$

The relation (31) expresses the *Markov Property* in a natural way that is easy to comprehend intuitively, while at the same time, preserves all the necessary mathematical rigor. Definition 6 tells us that a sequence of random variables $\{X_n;\ n = 1, 2, \dots\}$ is a Markov Chain if X_n *contains all information about* X_{n+1} *that is present in the random vector* $(X_1, X_2, \dots, X_n)$. To put it yet another way, this sequence of random variables is a Markov Chain if X_n, *now considered as a function on* $X_1, X_2, \dots, X_n$, is a *sufficient function* for X_{n+1} for $n = 1, 2, \dots$.

Recall the traditional definition of a discrete-parameter Markov Chain with a finite number if states:

DEFINITION 7. A sequence $\{X_n;\ n = 1, 2, \dots\}$ of real-valued random variables given on a probability space $(\Omega, \mathcal{F}, \mathcal{P})$, each taking on a finite number of values, is called a *Markov Chain* if

$$\mathcal{P}[X_{n+1} = x_{n+1} | X_1, X_2, \dots, X_n] = \mathcal{P}[X_{n+1} = x_{n+1} | X_n]. \tag{32}$$

The requirement (32), although very deep, appears to students who are taking an elementary Probability Theory course, as purely formalistic—just made in order *to make our life easy!*

EXERCISE 1. Show that Definitions 6 and 7 are equivalent. Hint: use Theorem 4 and the basic property of a sufficient function.

EXERCISE 2. A person is walking on a straight line who at each point of time $n,\ n = 1, 2, \dots$, either takes one step to the right with probability p, $p > 0$, or one step to the left with probability $q = 1 - p$. Let X_n denotes his/her position on the line at instant n. Using both Definitions 6 and 7, show that X_n is a Markov Chain. (This is the well-known *Random Walk Model.*)

EXERCISE 3. Given a sequence of independent, identically distributed random variables $\{X_n;\ n = 1, 2, \dots\}$ such that $\mathrm{P}[X_n = -1] = p > 0,\ \mathrm{P}[X_n = +1] = q = 1 - p$. Put $Y_n = X_n X_{n+1},\ n = 1, 2, \dots$. Show that sequence $Y_n;\ n = 1, 2, \dots$, *is not a Markov Chain* if $p \neq q$.

5. A History Note

Today, Theory of Markov Processes is an essential part of the Probability Theory, and it has wide applications in many areas of sciences. This theory was named after *A. A. Markov,* the celebrated Russian mathematician, who laid the foundations of the theory in a series of papers starting in 1907. One of his first application of this theory was in investigation of the way the vowels and consonants alternate in literary works in Russian literature. Markov carried out such a study on Pushkin's *Eugene Onegin* and on Aksakov's *The Childhood Years of Bagrov's Grandson.* Although the modern linguists have shown that a natural language *is not* a Markov Chain, nevertheless, the work on a study of Russian language, a natural language, contributed to the creation of what is known today, as the Theory of Markov Processes!

REFERENCES

1. K. L. Chung, *Markov Chains with Stationary Transition Probabilities*, 2nd edition, Springer, Berlin, 1967.
2. A. Feinstein, *Foundations of Information Theory*, McGraw-Hill, New York, 1958.
3. S. Kullbach, *Information Theory and Statistics*, Willey, New York, 1959.
4. J-F. Le Gall, *Random Trees and Applications*, Cornell University Summer School in Probability, 2005.
5. A. Rényi, *Foundations of Probability* Holden-Day, Inc., San Francisco, 1970.

Department of Mathematics, University of Florida, Gainesville, Florida 32611, U.S.A.
E-mail: zps@math.ufl.edu