

ELEMENTS OF DISCRETE PROBABILITY

4. Markov Chains. Tails of Distributions.

1 Markov chains

We have already dealt with the situation of repeated trials, specifically of repeated independent binary trials (Bernoulli trials), which led us to the binomial distribution. The analysis can be extended to nonbinary outcomes with moderate complications, but what we wish to analyze here is what happens if we relax the condition of “independent trials”.

In general, suppose we have a set I of possible outcomes and a sequence a_1, a_2, \dots, a_n of trials, with each a_i belonging to some alphabet \mathcal{Z} . The novelty is that outcome a_i is *not* independent of the preceding outcomes a_1, a_2, \dots, a_{i-1} .

A simple example occurs when \mathcal{Z} is the English alphabet and the sequence is the initial segment of a word of the English language, suppose $n = 3$ and $a_1, a_2 = TH$. Then a_3 can be selected within a very small subset of \mathcal{Z} (such as $\{A, E, I, O, U, W, R\}$, each of these symbols has its own distinct probability.

In such cases, the probability of the sequence is not the product of the individual probabilities of its symbols (as it would be in the case of independent symbols), but it is given as

$$\Pr(a_i) \Pr(a_2|a_1) \Pr(a_3|a_1 a_2) \dots \Pr(a_n|a_1 \dots a_{n-1})$$

However, in most situations the dependence of a_n upon the preceding symbols is “local”, i.e., it extends only a few position in the “past”. A most interesting case, of great practical significance and of analytical tractability, is when this dependence is limited to the *immediately preceding* symbol: in such case, the sequence is referred to as a **Markov chain** and

$$\Pr(a_1 a_2 \dots a_n) = \Pr(a_1) \Pr(a_2|a_1) \Pr(a_3|a_2) \dots \Pr(a_n|a_{n-1})$$

It is convenient to view the sequence a_1, a_2, \dots, a_n as generated by an automatic device, which is a special case of automaton. Recall that a finite-state automaton (or machine) is a quintuple $(S, \mathcal{I}, \mathcal{Z}, \delta, \lambda)$, where S is the set set of “states”, \mathcal{I} the input alphabet, \mathcal{Z} the output alphabet, $\delta : S \times \mathcal{I} \rightarrow S$ the next-state function, and

$\lambda : S \times \mathcal{I} \rightarrow \mathcal{Z}$ the output function. We transition from an FSA to a Markov-chain generator as follows

$$\begin{array}{ccccc} (S, & \mathcal{I}, & \mathcal{Z}, & \delta, & \lambda) \\ & & \Downarrow & & \\ (S, & \emptyset, & \mathcal{Z}, & P, & S) \end{array}$$

that is, the generator has no input (autonomous), the output is identical to the present state, and the state-transition function is replaced by the *transition matrix* P .

The transition matrix P is a function:

$$P : S \times S \rightarrow [0, 1]$$

with the additional condition that (letting $S = \{1, 2, \dots, n\}$)

$$\sum_{j=1}^n P_{ij} = 1 \quad \text{for } i = 1, 2, \dots, n$$

Matrix P , called a “stochastic matrix”, is interpreted as follows: P_{ij} is the probability of a transition from state i to state j , or, equivalently, that symbol i is followed by symbol j . Note that each row of this matrix describes a probability density function.

Suppose now that we initially place our Markov-chain generator in a state i . If we wish to express this initial assignment as a distribution over the set of states, we can represent it as vector $\mathbf{w} = (w_1, w_2, \dots, w_n)$ the components of which are 0 except for $w_i = 1$ (the certainty that the state is i). After one transition (if you wish, one time-step) the probability of the states (again, a vector \mathbf{w}' of n component will be given by

$$\mathbf{w}'^T = \mathbf{w}^T P$$

where the subscript ”T” denotes a “row” vector. In general, after n transitions the distribution will be

$$\mathbf{w}^T P^n$$

The matrix P completely characterizes a Markov chain generators. The study of Markov chain is a fascinating subject of probability theory, with fundamental relevance to a large variety of systems, particularly those involving traffic with random arrivals and departures.

Example. Consider the situation where we have n stations aligned, numbered from 1 to n (see the figure below). At each step a person moves from the station where it resides to one of the two adjacent stations, with probability p to the left and probability $q = 1 - p$ to the right. In station 1, the person remains in 1 with probability p ; in station n , he remains in n with probability p . This situation is referred to as “random walk with reflecting (or elastic) barriers”

The transition matrix is illustrated below

$$\begin{bmatrix} p & q & 0 & 0 & \dots & 0 & 0 \\ p & 0 & q & 0 & \dots & 0 & 0 \\ 0 & p & 0 & q & \dots & 0 & 0 \\ & & & \dots & & & \\ 0 & 0 & 0 & 0 & \dots & 0 & q \\ 0 & 0 & 0 & 0 & \dots & p & q \end{bmatrix}$$

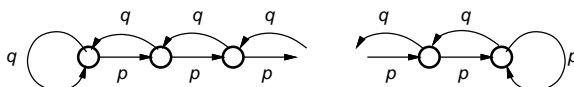


Figure 1: State diagram of Markov chain generator for random-walk with reflecting barriers.

2 The significance of the tails of distributions

As previous examples suggest, very frequently we are interested in evaluating the probability Q that a random variable X we are analyzing lies within a given range of values $[x_1, x_2]$, that is

$$Q = \sum_{x=x_1}^{x_2} \Pr(X = x)$$

It is immediate that

$$Q = 1 - \sum_{x < x_1} \Pr(X = x) - \sum_{x > x_2} \Pr(X = x)$$

where the first and second sum are appropriately denoted *left* and *right tails* of the distribution, respectively (refer to Figure1). Also very frequently we have the task of evaluating the probability that the random variable does not exceed some threshold, in which case the object of our analysis is the tail itself.

If the distribution (density function) is known analytically, then any of the above tasks can be carried out exactly either by deriving an analytical expression or by numerical computation to the desired accuracy. However, frequently the analytical expression is cumbersome or the numerical computation is tedious and time-consuming. Moreover, frequently what is desired is a reliable approximation rather than an accurate result.

More significant is the fact that in most cases a detailed analytical knowledge of the distribution is either not available or, in the spirit of settling for a reliable approximation, not worth deriving. We may have to content ourselves with significant

parameters of the distribution (such as the first two moments) and provide reliable answers using such limited knowledge. This motivates for the analysis of distribution tails that follows.

3 Markov's inequality

The most feeble amount of knowledge we have about a random variable X is its mean value $E[X]$ and the fact that X is nonnegative. It is remarkable that even in this case some statement can be made about the tail of its otherwise unknown distribution. This is expressed by the following theorem.

Theorem 1 (*Markov's inequality*) *Let X be a nonnegative random variable of known mean value $E[X]$. Then, for any positive real k ,*

$$\Pr(X \geq k) \leq \frac{E[X]}{k}.$$

Proof: We assume that X is discrete, with unknown density (g_0, g_1, \dots) (the argument is trivially extensible to non-discrete variables). Starting from the definition of expectation

$$\begin{aligned} E[X] &= \sum_j jg_j \\ &= \sum_{j < k} jg_j + \sum_{j \geq k} jg_j \\ &\geq \sum_{j < k} j \cdot g_j + k \sum_{j \geq k} g_j \\ &\geq 0 + k \Pr(X \geq k) \end{aligned}$$

where, in the third line we have replaced each value j in the right sum with their minimum value k , and in the fourth line we have replaced the nonnegative quantity $\sum_{j < k} j \cdot g_j$ with 0, thereby maintaining the inequality. \square

Notice that, by replacing k with $hE[X]$, Markov's inequality can be placed in the following alternative form:

$$\Pr(X \geq hE[X]) \leq \frac{1}{h}.$$

Example 1. Consider a bottling plant which fills 20,000 cans a day on the average. We ask for an upper bound to the probability that it can fill at least 25,000 cans on a specific day.

Denoting C the random variable "number of cans", Markov's inequality provides the answer:

$$\Pr(C \geq 25000) = \Pr(C \geq 1.25 \cdot 20000) \leq \frac{1}{1.25} = 0.8$$

The weakness of Markov's inequality is essentially the fact that we can only obtain "upper-bounds" on the probability of nonnegative variables. Certainly more interesting would be the availability of lower bounds, since they would provide some estimate of guaranteed performance. The answer to such question is provided by the Chebyshev's inequality, discussed next.

4 Chebyshev's inequality

As is to be expected, a stronger inequality requires some additional information about the distribution of the random variables. Knowledge of expectation and variance enables the formulation of the following bound. Note that the random variables under consideration are no longer restricted to being nonnegative.

Theorem 2 (Chebyshev's inequality) *Let Z be a random variable with expectation $E[Z]$ and variance $\text{var}[Z]$. If r is a positive real number then*

$$\Pr(|Z - E[Z]| \geq r) \leq \frac{\text{var}[Z]}{r^2}$$

Proof: Notice that the event $|Z - E[Z]| \geq r$ is equivalent to the event $(Z - E[Z])^2 \geq r^2$. Next we observe that $(Z - E[Z])^2$ is a nonnegative random variable Y , whose expectation $E[Y]$ is, by definition, the variance $\text{var}[Z]$ of Z . To random variable Y we now apply Markov's inequality with parameter r , i.e.,

$$\Pr(|Z - E[Z]| \geq r) = \Pr(Y \geq r^2) \leq \frac{E[Y]}{r^2} = \frac{\text{var}[Z]}{r^2}$$

□

Example 2. Again, consider a bottling plant which fills 20,000 cans a day on the average; however, this time we also know that the variance of their productivity is 2,000. We ask for a lower bound to the probability that it can fill between 19,000 and 21,000 cans on a specific day.

The fact that the interval being examined is symmetric about the average, allows us to use directly the Chebyshev's inequality. Denoting Z the plant's output, and since $1000/\sqrt{2000} = 22.36\dots$, we have

$$\Pr(|Z - 20000| \geq 1000) \leq \frac{1}{(22.36)^2} = 0.002$$

Due to its generality, we cannot expect the Chebyshev's inequality to be very strong. However, it is very useful, as illustrated by the following corollary:

Corollary 1 *Let Z be a random variable with expectation $E[Z]$ and variance $\text{var}[Z]$. The probability that Z does not deviate from $E[Z]$ by more than 3 standard deviations is at least $8/9$.*

Proof: Indeed, the probability of the event $|Z - E[Z]| \geq 3\sigma$ by Chebyshev's inequality is

$$\Pr(|Z - E[Z]| \geq 3\sigma) \leq \frac{\sigma^2}{(3\sigma)^2} = \frac{1}{9}$$

so that $\Pr(|Z - E[Z]| < 3\sigma) \geq 1 - 1/9 = 8/9 = 0.88888\dots$ □