Thailand Statistician
April 2020; 18(2): 235-242
http//:statassoc.or.th
Contributed paper

Maximum Entropy for Determining the Transition Probability Matrix of a Markov Chain with a Specified Stationary Distribution

Zohre Nikooravesh*
Department of Basic Sciences, Birjand University of Technology, Birjand, Iran.
*Corresponding author; e-mail: nikooravesh@birjandut.ac.ir

Received: 31 October 2018
Revised: 2 January 2019
Accepted: 3 July 2019

Abstract

In this paper, we try to find the unknown transition probability matrix of a Markov chain that has a specific stationary distribution. Numerous Markov chains can be found with this property, but among these Markov chains, we want to choose one with the highest entropy rate. In this case, we claim that the Markov chain where the rows of its transition probability matrix are of identical distribution to the stationary distribution, has the maximum entropy rate.

Keywords: Entropy rate, optimization problem, infinite state space.

1. Introduction

In probability theory, entropy was introduced by Shannon (1948). The entropy of a random variable X by distribution P_X taking values from a finite set $S = \{x_1, x_2, ..., x_n\}$, with corresponding probability p_i, is defined by Shannon as

$$H(X) = -\sum_i p_i \log p_i, \tag{1}$$

with the convention $0 \log 0 = 0$. The principle of maximum entropy is used in cases where the purpose is the estimation of the probability distribution that has maximum entropy under specific conditions.

Based on this definition, mathematicians were looking for the unknown probability distribution function with the highest degree of uncertainty when some conditions of the distribution were given. This problem known as the maximum entropy problem, is as follows

$$\begin{aligned} \max \quad & H(X) \\ s.t. \quad & \sum_{i=1}^{n} p_i = 1, \\ & \sum_{i=1}^{n} p_i g_j(x_i) = \beta_j, j = 1, 2, ..., n, \\ & p_i \geq 0, i = 1, 2, ..., n, \end{aligned} \tag{2}$$

where β_js are real numbers and g_js are real functions.

The entropy of a random variable is a multivariable function of the space $\mathbb{R}^n$ to $\mathbb{R}$. To maximize it, by differentiating the entropy function with respect to its variables, namely p_i, and by setting them equal to zero, we can find p_is. Since entropy is a convex function , the answer will be the maximum. In addition, p_is must satisfy the conditions (2). Solving these equations is challenging, because of their nonlinearity.

When the number of unknowns in an equation system is greater than the number of equations, we seek the optimal answer instead of solving the equation system. In this paper the optimal answer, is the answer with the highest amount of entropy. The maximum entropy method is a known method for determining the density function of random variables. Here, this method is generalized for Markov chains. Consider a Markov chain $\{X_n\}$ with a state space S, an unknown transition probability matrix $P = \left[p_{ij} \right]$ and an initial distribution π_0. We have $N(N-1)+(N-1) = N^2 - 1$ unknown variables for a finite state space S with N states. Now suppose that, the stationary distribution μ of the Markov chain is known; therefore, we have N equations from the relation $\mu = \mu P$. In this case, there will be an infinite number of answers for unknowns. Now the goal is selecting the answer with the most entropy. In the case of an infinite state space, it is more difficult to find values for unknowns however the maximum entropy method can be used.

Jaynes (1957) was the first to obtain a probability distribution with minimal probability under certain conditions using Shannon's maximum entropy definition.

Ang et al. (1992) proposed a kernel method to approximate the optimal importance sampling density. The method required an initial Monte Carlo run to generate the samples in the failure region. Thus, it was not efficient for problems with small failure probabilities. Au and Beck (1999) proposed an importance sampling scheme to improve the efficiency of the method by using the Markov chain simulation to generate the samples that populate in the regions of most interest.

Zografos (2008) studied the properties, applications, and generalization of the maximum entropy method. In the case of the maximum entropy of discrete probability functions, in stochastic processes, especially in Markov chains, Erik (2009) expresses the concept of maximum entropy, and later, Chliamovitch et al. (2015) complemented his work. Also, Burda et al. (2009) defined a new class of random walk processes that maximize entropy. They proved this maximal entropy random walk is equivalent to generic random walk if it takes place on a regular lattice, but not if the underlying lattice is irregular.

Basset (2015) did the same. He worked with timed region graphs that are to timed automata what finite directed graphs are to finite state automata, that is, automata without labeling on transitions and without initial and final states. He defined stochastic processes over runs of timed region graphs and their (continuous) entropy.

In the discussion of reliability, Dai et al. (2016) introduced the concept of maximum entropy. They introduced a new maximum entropy-based importance sampling scheme. The proposed methodology involves the generation of samples that populate the important region by Markov chain simulation, and the construction of importance sampling density by the maximum entropy density estimation method.

Henter and Kleijn (2016) proposed minimum entropy rate simplification (MERS), an information-theoretic, parameterization-independent framework for simplifying generative models of stochastic processes.

Our study concentrates on the calculation of the unknown transition probability matrix of Markov chains with infinite state space, by help of the maximum entropy method.

This paper consists of four sections. In Section 2, we define the problem of maximum entropy for Markov chains. Section 3 solves the maximum entropy problem for infinite Markov chains in general, by expressing a theorem and proving it. In Section 4, the problem of maximum entropy for Markov chains with triple state space, is investigated and solved directly.

2. Definition of the Problem

Consider a Markov chain $\{X_n\}$ with state space $S=\{1,2,3,...\}$, a stationary distribution $\mu=(\mu_1,\mu_2,\mu_3,...)$ and the unknown transition probability $p_{ij}=\{X_{n+1}=j \mid X_n=i\}$. We intend to calculate the probability values p_{ij} by help of the maximum entropy method. According to the definition of the Shannon entropy rate of the Markov chain, we have

$$H(\chi)=\lim_{n\to\infty}\frac{H(X_1,X_2,...,X_n)}{n}, \tag{3}$$

where X_t is a random variable demonstrating the state at time t, and $H(X_1,X_2,...,X_n)$ is the joint entropy of $(X_1,X_2,...,X_n)$ with the joint distribution $P(x_1,x_2,...,x_n)$ where

$$\begin{aligned} H(X_1,X_2,...,X_n) &= -\sum_{x_1^n\in S^n} p(x_1,x_2,...,x_n)\log p(x_1,x_2,...,x_n) \\ &= -E_{X_1,X_2,...,X_n}\log p(x_1,x_2,...,x_n). \end{aligned} \tag{4}$$

Note that for any sequence $\{X_k\}_{k\geq 1}$ we denote a finite subsequence $x_i,x_{i+1},...,x_j$, $j>i$ by x_i^j.

Shannon (1948) proved the convergence in probability by $-\frac{1}{n}\log P(x_1,x_2,...,x_n)$ to $H(\chi)$. For an ergodic stochastic process X, Cover and Thomas (2006) proved that

$$H(\chi)=\lim_{n\to\infty}H(X_n \mid X_{n-1},...,X_1), \tag{5}$$

and for the homogeneous ergodic Markov chain one can show that

$$H(\chi)=H(X_2 \mid X_1). \tag{6}$$

Thus, we can conclude easily

$$H(\chi)=-\sum_{i-1}^{\infty}\sum_{j-1}^{\infty}\mu_i p_{ij}\log p_{ij}, \tag{7}$$

where μ_i is the unique stationary distribution of state $i\in S$.

Note that one can easily prove $\lim_{n\to\infty}P(X_n=j)=\mu_j$ where μ_j is the stationary distribution of the Markov chain. Now, the problem is to maximize $H(\chi)$ by satisfying the following conditions,

$$\sum_{j-1}^{\infty}p_{ij}=1, i\in S, \tag{8}$$

$$\mu_j=\sum_{i-1}^{\infty}\mu_i p_{ij}, j\in S. \tag{9}$$

The usual solution to solve this problem is by obtaining relationships between p_{ij}s using the relations (8) and (9), and putting the results in relation (7). Then, by differentiating of $H(\chi)$ with respect to its variables and by setting them equal to zero, we get to the desired answer.

In Section 3, this problem will be solved for an infinite countable state space. In this section we have a Theorem that is proved by considering the contradiction in answers. Section 4 solves this problem directly by a usual method for a specific example with a triple state space.

3. Solving the Problem Generally

In this section, we study the maximum entropy method for Markov chains with an infinite countable state space and a specified stationary distribution. This is done in Theorem 1.

Theorem 1 *Consider a Markov chain with a state space* $S = \{1,2,3,...\}$, *and a stationary distribution* $\mu = (\mu_1, \mu_2, ...)$. *The unique Markov chain* $\{X_n\}$, *with the transition probability* $p_{ij} = \mu_j, i, j = 1,2,3,...,$ *has the highest entropy rate with state space* S *and the stationary distribution of* μ.

Proof: $p_{ij} = \mu_j$ is satisfied in relations (8) and (9). We claim that the Markov chain $\{X_n\}$ with this transition matrix has the highest entropy rate. Otherwise, consider the Markov chain $\{X_n'\}$ with $P' \neq P$. Then for P', there is at least one value i and a value l that $p_{il} \neq \mu_l$. In other words, ε can be found such that $p_{il} = \mu_l - \varepsilon$. On the other hand, relations (8) and (9) must be established. Therefore, $k \neq l$, $r \neq i$ and the value δ was established such that $p_{ik} = \mu_k + \varepsilon$, $p_{rl} = \mu_l + \delta$ and $p_{rk} = \mu_k - \delta$. Without loss of generality, we put $i = l = 1$ and $r = k = 2$ for convenience. From the relation (9), we conclude that

$$\begin{cases} \mu_1(\mu_1 - \varepsilon) + \mu_2(\mu_1 + \delta) + \mu_3\mu_1 + \cdots + \mu_n\mu_1 = \mu_1 \\ \mu_1(\mu_2 + \varepsilon) + \mu_2(\mu_2 - \delta) + \mu_3\mu_2 + \cdots + \mu_n\mu_2 = \mu_2 \\ \mu_1\mu_3 + \mu_2\mu_3 + \mu_3\mu_3 + \cdots + \mu_n\mu_3 = \mu_3 \\ \qquad\qquad \vdots \\ \mu_1\mu_n + \mu_2\mu_n + \mu_3\mu_n + \cdots + \mu_n\mu_n = \mu_n \\ \qquad\qquad \vdots \end{cases} \tag{10}$$

Given the relation $\sum_{j=1}^{\infty} \mu_j = 1,$ we have

$$\begin{cases} -\varepsilon\mu_1 + \delta\mu_{2.} = 0 \\ \varepsilon\mu_1 - \delta\mu_{2.} = 0 \end{cases} \Rightarrow \varepsilon\mu_1 = \delta\mu_{2.}. \tag{11}$$

Now, we can write

$$\begin{aligned} -H(\chi) + H(\chi') = & \sum_{i=1}^{\infty}\mu_i \sum_{j=1}^{\infty}\mu_j \log\mu_j - \sum_{i=3}^{\infty}\mu_i \sum_{j=1}^{\infty}\mu_j \log\mu_j \\ & + \mu_1\left[\mu_1\log\mu_1 - (\mu_1 - \varepsilon)\log(\mu_1 - \varepsilon) + \mu_2\log\mu_2 - (\mu_2 + \varepsilon)\log(\mu_2 + \varepsilon)\right] \\ & + \mu_2\left[\mu_1\log\mu_1 - (\mu_1 + \delta)\log(\mu_1 + \delta) + \mu_2\log\mu_2 - (\mu_2 - \delta)\log(\mu_2 - \delta)\right]. \end{aligned} \tag{12}$$

So

$$\begin{aligned} -H(\chi) + H(\chi') = & \mu_1^2\log\mu_1 - \left(\mu_1^2 - \varepsilon\mu_1\right)\left(\log\mu_1 + \log\left(1 - \frac{\varepsilon}{\mu_1}\right)\right) \\ & - \mu_1\left[(\mu_1 - \varepsilon)\log(\mu_1 - \varepsilon) + (\mu_2 + \varepsilon)\log(\mu_2 + \varepsilon) + \sum_{j=3}^{\infty}\mu_j\log\mu_j\right] \\ & - \mu_2\left[(\mu_1 + \delta)\log(\mu_1 + \delta) + (\mu_2 - \delta)\log(\mu_2 - \delta) + \sum_{j=3}^{\infty}\mu_j\log\mu_j\right] \end{aligned} \tag{13}$$

$$+\mu_1\mu_2\log\mu_2-(\mu_1\mu_2+\varepsilon\mu_1)\left(\log\mu_2+\log\left(1+\frac{\varepsilon}{\mu_2}\right)\right)$$
$$+\mu_1\mu_2\log\mu_1-(\mu_1\mu_2+\delta\mu_2)\left(\log\mu_1+\log\left(1+\frac{\delta}{\mu_1}\right)\right)$$
$$+\mu_2^2\log\mu_2-(\mu_2^2-\delta\mu_2)\left(\log\mu_2+\log\left(1-\frac{\delta}{\mu_2}\right)\right).$$

We know that for small x, $\log(1+x)=x-\frac{x^2}{2}+o(x^3)$ and $\log(1-x)=-x-\frac{x^2}{2}+o(x^3)$, so we have

$$\begin{aligned}-H(\chi)+H(\chi') &= \varepsilon\mu_1\log\mu_1-(\mu_1^2-\varepsilon\mu_1)\left(-\frac{\varepsilon}{\mu_1}-\frac{\varepsilon^2}{2\mu_1^2}+o(\xi^3)\right)\\ &-\varepsilon\mu_1\log\mu_2-(\mu_1\mu_2-\varepsilon\mu_1)\left(\frac{\varepsilon}{\mu_2}-\frac{\varepsilon^2}{2\mu_2^2}+o(\xi^3)\right)\\ &-\delta\mu_2\log\mu_1-(\mu_1\mu_2-\delta\mu_2)\left(\frac{\delta}{\mu_1}-\frac{\delta^2}{2\mu_1^2}+o(\xi^3)\right)\\ &+\delta\mu_2\log\mu_2-(\mu_2^2-\delta\mu_2)\left(-\frac{\delta}{\mu_2}-\frac{\delta^2}{2\mu_2^2}+o(\xi^3)\right)\\ &=-\frac{1}{2}(\varepsilon+\delta)^2+o(\xi^3)\le 0,\end{aligned}\tag{14}$$

where $\xi=\max\left\{\frac{\varepsilon}{\mu_1},\frac{\varepsilon}{\mu_2},\frac{\delta}{\mu_1},\frac{\delta}{\mu_2}\right\}$. Equality in the relation (14) is valid only if $\varepsilon=\delta=0$; this completes the proof of theorem.

4. The Triple State Space Markov Chains

In this section, the problem of maximum entropy for Markov chains with a triple state space, is investigated and solved directly. Suppose that we have a Markov chain $\{X_n\}$ with a state space $S=\{1,2,3\}$ and with a known stationary distribution $\mu=(\mu_1,\mu_2,\mu_3)$. By using relation (9) we have

$$\begin{cases}\mu_1=\mu_1p_{11}+\mu_2p_{21}+\mu_3p_{31}\\ \mu_2=\mu_1p_{12}+\mu_2p_{22}+\mu_3p_{32}\\ \mu_1+\mu_2+\mu_3=1\end{cases}\Rightarrow\begin{cases}p_{31}=\frac{1}{\mu_3}\left((1-p_{11})\mu_1-p_{21}\mu_2\right)\\ p_{32}=\frac{1}{\mu_3}\left(-p_{12}\mu_1+(1-p_{22})\mu_2\right)\\ p_{33}=\frac{1}{\mu_3}\left(\mu_3-(1-p_{11}-p_{12})\mu_1\right.\\ \qquad\left.-(1-p_{21}-p_{22})\mu_2\right).\end{cases}\tag{15}$$

Note that we know $p_{i3}=1-p_{i1}-p_{i2}$ for $i=1,2,3$. Now, with respect to the definition of the entropy rate of a Markov chain in relation (7),

$$
\begin{aligned}
-H(\chi) &= \sum_{i=1}^{3}\sum_{j=1}^{3}\mu_i p_{ij}\log p_{ij} \\
&= \mu_1\left(p_{11}\log p_{11} + p_{12}\log p_{12} + (1-p_{11}-p_{12})\log(1-p_{11}-p_{12})\right) \\
&+\mu_2\left(p_{21}\log p_{21} + p_{22}\log p_{22} + (1-p_{21}-p_{22})\log(1-p_{21}-p_{22})\right) \\
&+\left((1-p_{11})\mu_1 - p_{21}\mu_2\right)\log\frac{\left((1-p_{11})\mu_1 - p_{21}\mu_2\right)}{\mu_3} \\
&+\left(-p_{12}\mu_1 + (1-p_{22})\mu_2\right)\log\frac{\left(-p_{12}\mu_1 + (1-p_{22})\mu_2\right)}{\mu_3} \\
&+\left(\mu_3 - (1-p_{11}-p_{12})\mu_1\right. \\
&\left.-(1-p_{21}-p_{22})\mu_2\right)\log\frac{\left(\mu_3 - (1-p_{11}-p_{12})\mu_1 - (1-p_{21}-p_{22})\mu_2\right)}{\mu_3}.
\end{aligned} \tag{16}
$$

Now we obtain some equations for finding p_{ij} s by differentiating of $-H(\chi)$ with respect to p_{11}, p_{12}, p_{21} and p_{22} and by setting them equal to zero.

$$
\begin{aligned}
-\frac{\partial H(\chi)}{\partial p_{11}} &= \mu_1\left(\log p_{11} + 1 - \log(1-p_{11}-p_{12}) - 1\right) \\
&-\mu_1\log\frac{\left((1-p_{11})\mu_1 - p_{21}\mu_2\right)}{\mu_3} - \mu_1 \\
&+\mu_1\log\frac{\left(\mu_3 - (1-p_{11}-p_{12})\mu_1 - (1-p_{21}-p_{22})\mu_2\right)}{\mu_3} + \mu_1.
\end{aligned} \tag{17}
$$

Now,

$$
-\frac{\partial H(\chi)}{\partial p_{11}} = 0 \Rightarrow p_{11}\mu_3 - p_{11}\mu_2 + p_{11}\mu_1 + p_{11}p_{22}\mu_2 + p_{21}\mu_2 + p_{12}\mu_1 - p_{12}p_{21}\mu_2 = \mu_1. \tag{18}
$$

We have the same,

$$
-\frac{\partial H(\chi)}{\partial p_{12}} = 0 \Rightarrow p_{12}\mu_3 + p_{12}p_{21}\mu_2 + p_{22}\mu_2 + p_{11}\mu_2 - p_{11}p_{22}\mu_2 = \mu_2. \tag{19}
$$

$$
-\frac{\partial H(\chi)}{\partial p_{21}} = 0 \Rightarrow p_{21}\mu_3 + p_{12}p_{21}\mu_1 + p_{22}\mu_1 + p_{11}\mu_2 - p_{11}p_{22}\mu_1 = \mu_1. \tag{20}
$$

$$
-\frac{\partial H(\chi)}{\partial p_{22}} = 0 \Rightarrow p_{22}\mu_3 - p_{22}\mu_1 + p_{22}\mu_2 + p_{11}p_{22}\mu_1 + p_{12}\mu_1 + p_{21}\mu_2 - p_{12}p_{21}\mu_1 = \mu_2. \tag{21}
$$

By solving the nonlinear relations (18) to (21), we can obtain the values of p_{ij}. One possible answer is equal to

$$
p_{ij} = \mu_j, j = 1,2,3. \tag{22}
$$

So we have,

$$\begin{aligned} H(\chi) &= -\sum_{i=1}^{3}\sum_{j=1}^{3}\mu_i p_{ij}\log p_{ij} \\ &= -\sum_{i=1}^{3}\sum_{j=1}^{3}\mu_i\mu_j\log\mu_j \\ &= -\sum_{j=1}^{3}\mu_j\log\mu_j \\ &= H(\mu). \end{aligned} \tag{23}$$

$H(\mu)$ is the maximum entropy value for this chain. Therefore, according to the relation (22), it can be concluded that the distribution transition probability matrix rows of Markov chain is the same as the distribution of μ. So

$$\begin{aligned} H(\chi) &= \lim_{n\to\infty} H(X_n \mid X_{n-1}) \\ &\le \lim_{n\to\infty} H(X_n) \\ &= -\lim_{n\to\infty}\sum_{i=1}^{3} p(X_n = j)\log p(X_n = j) \\ &= -\sum_{i=1}^{3}\lim_{n\to\infty} p(X_n = j)\log\lim_{n\to\infty} p(X_n = j) \\ &= -\sum_{j=1}^{3}\mu_j\log\mu_j = H(\mu). \end{aligned} \tag{24}$$

Note that the summation is on the finite number of sentences and logarithm is a continuous function.

5. Conclusions

The maximum entropy method, in cases where some information of the distribution is available, is used to find the distribution among the corresponding distributions with available information and the highest entropy. In this article, this result was generalized for Markov chains. As we know, the uniform distribution among distributions with finite state space has the highest amount of entropy. A generalization of this problem occurs in the Markov chain when there is only stationary distribution of the Markov chain. In this case, a Markov chain has the highest entropy rate when all of its transition probability matrix rows have the same distribution as the stationary distribution.

Acknowledgements

The author acknowledges editor of this journal for the constant encouragement to finalize the paper and would like to thank the reviewers for their critical review and comments on the earlier version of the manuscript.

References

Ang GL, Ang AH, Tang WH. Optimal importance-sampling density estimator. J Eng Mech ASCE. 1992; 118(6): 46-63.

Au SK, Beck JL. A new adaptive importance sampling scheme for reliability calculations. J Struct Saf. 1999; 21(1): 35-58.

Basset N. A maximum entropy stochastic process for a timed automaton. J Inform Comput. 2015; 243: 50-74.

Burda Z, Duda J, Luck JM, Waclaw B. Localization of maximal entropy random walk. Phys Rev Lett. 2009, PRL 102, 160602.

Chliamovitch G, Dupuis A, Chopard B. Maximum entropy rate reconstruction of Markov dynamics. J Entropy. 2015; 17: 3738-3751.

Cover TM, Thomas JA. Elements of information theory. New York: John Wiley & Sons. 2006.

Dai H, Zhang H, Wang W. A new maximum entropy-based importance sampling for reliability anaiysis. J Struct Saf. 2016; 63: 71-80.

Erik VS. Maximum entropy estimation of transition probabilities of reversible Markov chains. J Entropy. 2009; 11: 867-887.

Henter GE, Kleijn WB. Minimum entropy rate simplification of stochastic processes. IEEE Trans Pattern Anal Mach Intell. 2016; 38: 1-14.

Jaynes ET. Information theory and statistical mechanics. J Phys Rev. 1957; 106: 620-630.

Shannon CE. A mathematical theory of communication. J Bell Syst Tech 1948; 27: 379-423.

Zografos K. On some entropy and divergence type measures of variability and dependence for mixed continuous and discrete variables. J Stat Plan Infer. 2008; 138: 3899-3914.