Probability in the Engineering and Informational Sciences, **27**, 2013, 53–55.
doi:10.1017/S0269964812000290

A MARKOV CHAIN CHOICE PROBLEM

SHELDON M. ROSS
Epstein Department of Industrial and Systems Engineering
University of Southern California
Los Angeles, CA 90089, USA
E-mail: smross@usc.edu

Consider two independent Markov chains having states 0, 1, and identical transition probabilities. At each stage one of the chains is observed, and a reward equal to the observed state is earned. Assuming prior probabilities on the initial states of the chains it is shown that the myopic policy that always chooses to observe the chain most likely to be in state 1 stochastically maximizes the sequence of rewards earned in each period.

1. INTRODUCTION

Suppose that at each stage one is to choose to observe the current state of one of n independent and identically distributed two-state Markov chains with states 0 and 1. If the observed state is 1, a reward 1 is earned; if it is 0, a reward 0 is earned. The states of the unchosen chains are not learned. Further, suppose that we know the probabilities that each of the chains is initially in state 1.

The problem of determining the policy that maximizes the total expected infinite discounted return was introduced in [1] and also studied in [3] and [2], with the latter two papers not being aware of [1]. All of these papers proved that the myopic policy, defined as the policy that always chooses to observe that chain most likely to be in state 1, is optimal when $n = 2$. Assuming $P_{1,1} > P_{0,1}$, [1] gave a sufficient condition for the myopic policy to be optimal for general n, [3] conjectured that the myopic policy was optimal (without any conditions beyond the assumption that $P_{1,1} > P_{0,1}$) and [2] proved the conjectured optimality. In addition, when $P_{1,1} < P_{0,1}$, [2] proved that the myopic policy is also optimal if $n = 3$ but need not be when $n = 4$. (See these papers for other literature as well as for applications of this model.)

In this paper we present a simple argument that when $n = 2$ the myopic policy is optimal in a very strong sense. Namely, with R_i being the reward earned during period i, we show when $n = 2$ that the myopic policy stochastically maximizes the vector $(R_1, \ldots, R_m)$ for every m. That is, for any m and any increasing function $f(x_1, \ldots, x_m)$, the quantity $E[f(R_1, \ldots, R_m)]$ is maximized when the myopic policy is employed.

2. OPTIMALITY RESULT WHEN $n = 2$

When $n = 2$ the myopic policy can be expressed as follows. Start by choosing the chain that has the highest probability of being in state 1. If $P_{1,1} > P_{0,1}$, then let each subsequent

© Cambridge University Press 2013 0269-9648/13 $25.00

https://doi.org/10.1017/S0269964812000290 Published online by Cambridge University Press

choice be the same as the preceding one if and only if the preceding choice resulted in a reward of 1. If $P_{1,1} < P_{0,1}$, let each subsequent choice be the same as the preceding one if and only if the preceding choice yielded a reward 0. That is, when $P_{1,1} > P_{0,1}$ the myopic policy changes chains when it receives the reward 0, whereas when $P_{1,1} < P_{0,1}$ the myopic policy changes when it receives 1.

PROPOSITION 1: *If $n = 2$, then the myopic policy stochastically maximizes $(R_1, \ldots, R_m)$ for every m.*

PROOF: Arbitrarily number the Markov chains and let $S = X_1 + X_2$, where X_i is the state of Markov chain i in period 1. Let $p_i = P(X_i = 1)$, and suppose, without loss of generality, that $p_1 \geq p_2$. The proof of the proposition is by induction on m. It is immediate for $m = 1$, so assume that for that $E[f(R_1, \ldots, R_m)]$ is maximized by the myopic policy, for any increasing function $f(x_1, \ldots, x_m)$. Now, let $g(x_1, \ldots, x_{m+1})$ be an increasing function defined on $m + 1$ variables. Note, by the induction hypothesis, that the best policy that starts with choosing chain i is the one that then continues with the myopic policy. With π_i denoting the policy that starts by observing chain i and then continues with the myopic policy, observe that

$$E_{\pi_1}[g(R_1, \ldots, R_{m+1})|S = 0] = E_{\pi_2}[g(R_1, \ldots, R_{m+1})|S = 0],$$

where the preceding follows because, conditional on $X_1 = X_2 = 0$, the reward vectors $(R_1, \ldots, R_{m+1})$ yielded by the policies π_1 and π_2 are identically distributed. Similarly,

$$E_{\pi_1}[g(R_1, \ldots, R_{m+1})|S = 2] = E_{\pi_2}[g(R_1, \ldots, R_{m+1})|S = 2].$$

Now, conditional on $S = 1$, it follows that independent of which chain was initially observed the next choice will be the chain that was previously in state 1 as opposed to the one that was previously in state 0 (either because the chain in state 1 was originally chosen and so will be chosen the next period, or because the chain in state 0 was originally chosen and so the other chain will be chosen the next period). Consequently, conditional on $S = 1$, no matter which of the policies $\pi_i, i = 1, 2$ is being used, the rewards $R_2, \ldots, R_{m+1}$ will have the same joint distribution and will be independent of R_1. Because, with $q_i = 1 - p_i$,

$$P_{\pi_1}(R_1 = 1|S = 1) = \frac{p_1 q_2}{p_1 q_2 + p_2 q_1} \geq \frac{p_2 q_1}{p_1 q_2 + p_2 q_1} = P_{\pi_2}(R_1 = 1|S = 1),$$

we can conclude that

$$E_{\pi_1}[g(R_1, \ldots, R_{m+1})|S = 1] \geq E_{\pi_2}[g(R_1, \ldots, R_{m+1})|S = 1].$$

It now follows from the preceding that

$$E_{\pi_1}[g(R_1, \ldots, R_{m+1})] \geq E_{\pi_2}[g(R_1, \ldots, R_{m+1})]$$

and the result is proved. ■

Remarks: (a) The optimality requirement that the policy stochastically maximizes $R_1, \ldots, R_m$ for every m is a very strong requirement. For instance, consider a problem with a two-state Markov with given probabilities for the initial state and with given transition probabilities $P_{i,j}, i, j = 0, 1$. Suppose that at each stage one is to guess the current state of the chain, with a reward of 1 earned if the guess is correct. Further suppose that one immediately learns whether the guess is correct. Then, because one's choices do not affect

https://doi.org/10.1017/S0269964812000290 Published online by Cambridge University Press

subsequent states, it follows that the myopic policy, which always guesses the state that the chain is most likely to be in, will maximize the expected reward earned in any period. However, it need not stochastically maximize the vector of rewards. For instance, suppose there is a 0.6 probability that the chain is in state 0 in period 1, and that $P_{0,0} = 0.6$ and $P_{1,1} = 1$. Suppose $f(R_1, \ldots, R_{10}) = I\left\{\sum_{i=1}^{10} R_i = 10\right\}$. That is, we want to maximize the probability that the first 10 guesses are correct. Because the myopic policy, call it π, will always guess state 0 until it makes a wrong guess, it follows that

$$E_\pi[f(R_1, \ldots, R_{10})] = 0.6^{10}$$

However, with π^* equal to the policy that always guesses state 1

$$E_{\pi^*}[f(R_1, \ldots, R_{10})] = 0.4$$

(b) If in our problem the states of both chains are learned after a guess, then the policy that always chooses the chain most likely to be in state 1 is exactly the myopic policy of our original problem, which immediately shows that it maximizes the total expected discounted cost when only the state of the chosen chain is learned.

Acknowledgments

This material is based upon work supported by the U.S. Army Research Laboratory and the U.S. Army Research Office under grant number W911NF-11-1-0115.

References

1. Koole, G., Liu, Z., & Righter, R. (2001). Optimal transmission policies for noisy channels. *Operations Research* 49(6): 892–899.
2. Ahmad, S.H.A., Liu, M., Javidi, T., Zhao, Q., & Krishnamachari, B. (2009). Optimality of myopic sensing in multichannel opportunistic access. *IEEE Transactions on Information Theory* 55(9): 4040–4050.
3. Zhao, Q., Krishnamachari, B., & Liu, K. (2008). On myopic sensing in multichannel opportunistic access: structure, optimality, and performance. *IEEE Transactions Wireless Communications* 7(12): 5431–5440.

https://doi.org/10.1017/S0269964812000290 Published online by Cambridge University Press