Probability in the Engineering and Informational Sciences, **25**, 2011, 307–342.
doi:10.1017/S0269964811000039

OPTIMAL MIXING OF MARKOV DECISION RULES FOR MDP CONTROL

DINARD VAN DER LAAN
Tinbergen Institute and
Department of Econometrics and Operations Research
VU University
De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands
E-mail: dalaan@feweb.vu.nl

In this article we study Markov decision process (MDP) problems with the restriction that at decision epochs, only a finite number of given Markov decision rules are admissible. For example, the set of admissible Markov decision rules $\mathcal{D}$ could consist of some easy-implementable decision rules. Additionally, many open-loop control problems can be modeled as an MDP with such a restriction on the admissible decision rules. Within the class of available policies, optimal policies are generally nonstationary and it is difficult to prove that some policy is optimal. We give an example with two admissible decision rules—$\mathcal{D} = \{d^1, d^2\}$—for which we conjecture that the nonstationary periodic Markov policy determined by its period cycle $(d^1, d^1, d^2, d^1, d^2, d^1, d^2, d^1, d^2)$ is optimal. This conjecture is supported by results that we obtain on the structure of optimal $\mathcal{D}$ Markov policies in general. We also present some numerical results that give additional confirmation for the conjecture for the particular example we consider.

1. INTRODUCTION

Markov decision processes (MDP) are a well-established tool for optimizing the control of stochastic systems. A complex system as MDP is applied in, for example, telecommunication, manufacturing systems, and call centers. Classically solving the MDP results in an (optimal) policy that for every system state yields a corresponding (optimal) control action. To implement this policy at any decision event, the current system state has to be known (or determined) before the corresponding control action is chosen. In practice, such implementation might not be convenient. Moreover, for complex systems with a large (multidimensional) state space, it is hard and practically impossible to find the optimal state-dependent policy.

0269-9648/11 $25.00

In this article we consider MDP for which decisions should be taken at an infinite discrete set T of consecutive decision epochs. For such MDP, a general (possibly nonstationary) Markovian policy specifies for each decision epoch $t \in T$ a decision rule to be applied at t where a decision rule can be represented by a mapping from the state space to a corresponding action space. For purposes in this article all of these spaces are assumed to be finite. Still, in general, it soon becomes intractable to determine the optimal policy if the state and/or action space(s) get larger. Moreover, optimal decision rules might have a complicated structure and be hard to implement in practice. Therefore, a basic idea in this article is to optimize over a (much) smaller set of Markov policies by restricting the set of admissible decision rules and, thus, the corresponding mappings from state space to action space are also restricted and, for example, have a specific structure.

We will refer to such a problem of MDP optimization over Markov policies with decision rules restricted to some given finite set $\mathcal{D}$ as $\mathcal{D}$ restricted MDP. Policies that are applicable to $\mathcal{D}$ restricted MDP will be referred to as $\mathcal{D}$ mixing policies. These $\mathcal{D}$ mixing policies correspond to infinite sequences $(d_1, d_2, \ldots)$ of Markov decision rules restricted by $d_t \in \mathcal{D}$ for $t = 1, 2, \ldots$. We note that optimization for $\mathcal{D}$ restricted MDP investigated in the present article is quite different from classical constrained MDP as investigated in, for example, [23] and [5]. In these articles on classical constrained MDP, the existence of particular optimal stationary randomized policies (respectively optimal state-action frequencies) is shown under some conditions. For the $\mathcal{D}$ restricted MDP investigated in the present article we show that even in the case of simple finite state and action spaces, there do not exist stationary randomized $\mathcal{D}$ mixing policies that are optimal with respect to maximizing the long-run average reward. In fact, we obtain nonstationary $\mathcal{D}$ mixing policies with better performance than all stationary randomized $\mathcal{D}$ mixing policies and we have results on optimality within a certain class of such nonstationary $\mathcal{D}$ mixing policies.

To clarify the problem and the type of questions we are investigating in this article, we now give a key example in which the problem of optimization for a particular $\mathcal{D}$ restricted MDP is described. Despite that this particular problem has a simple description with very small state space and action space, it turns out that the questions related to optimization over $\mathcal{D}$ mixing policies are intriguing and not easy to answer. Example 1 will be used throughout the article to illustrate the general results we obtain on optimization for $\mathcal{D}$ restricted MDP.

Example 1: A machine is operated that can be in two states, state—space $S = \{1, 2\}$—where state 1 is referred to as the bad state and state 2 as the good state. At every decision epoch t, $t = 1, 2, \ldots$, the operator has to decide whether the machine goes in working or repair mode for one time unit until the next decision epoch. Thus, there is a common action space $\mathcal{A} = \{1, 2\}$ for both states where action 1 refers to working mode and action 2 refers to repair mode. If action 1 is chosen, then there is a probability of .2 that the machine will be in a bad state at the next decision epoch if the machine is currently in a good state. Moreover, for action 1, the machine will certainly be in a bad state at the next decision epoch if the current state is bad. For action 2, there is

https://doi.org/10.1017/S0269964811000039 Published online by Cambridge University Press

a probability of .3 that the machine will be in a good state at the next decision epoch if the machine is currently in a bad state. Moreover, for action 2, the machine will certainly be in a good state at the next decision epoch if the current state is good. The only case in which a positive immediate reward of 1 is obtained if action 1 is chosen and the machine is currently in a good state; in every other case, we assume that the immediate reward is zero. Without restriction on the decision rules, the optimal policy is very easily obtained for this MDP with the average reward criterion. Of course, if the machine is in a good state, then action 1 will be optimal, and if the machine is in a bad state, then action 2 is optimal. However, in this article we investigate $\mathcal{D}$ restricted MDP where $\mathcal{D}$ is a (given) finite set of decision rules. In particular for this example consider optimization over $\mathcal{D} = \{d^1, d^2\}$ mixing policies where decision d^1 is choosing action 1 (work) for both states and d^2 is choosing action 2 (repair) for both states. Note that d^1 and d^2 are deterministic open-loop decision rules and restriction to $\{d^1, d^2\}$ mixing policies could be considered, for example, if observing the (current) state of the machine has some cost.

Now that we have described this particular $\mathcal{D}$ restricted MDP in detail, we focus on the problem of maximizing the long-run average reward over all $\mathcal{D}$ mixing policies and which questions arise. It soon turns out that optimizing over $\{d^1, d^2\}$ mixing policies is far from easy even for this problem with a very small state space and action space. First, we note that the problem of maximizing the long-run average reward over all $\{d^1, d^2\}$ mixing policies is completely specified by the transition matrix P_1 and expected immediate reward vector $r(d^1)$ induced by decision rule d^1 (respectively the transition matrix P_2 and expected immediate reward vector $r(d^2)$ induced by decision rule d^2). From the model description above, it follows for this example that

$$P_1 = \begin{pmatrix} 1 & 0 \\ 0.2 & 0.8 \end{pmatrix}, \qquad r(d^1) = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \qquad P_2 = \begin{pmatrix} 0.7 & 0.3 \\ 0 & 1 \end{pmatrix}, \qquad r(d^2) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \tag{1}$$

It is easily seen that the most straightforward (being both deterministic and stationary) $\{d^1, d^2\}$-mixing policies $(d^1, d^1, \ldots)$ and $(d^2, d^2, \ldots)$ give a long-run average reward of zero, which certainly can be improved. To improve this, one could consider policies that are still stationary but might be randomized. These are the class of so-called Bernoulli $\{d^1, d^2\}$-mixing policies, which at every decision epoch t apply decision rule d^1 with probability $\theta \in [0, 1]$ and apply d^2 with probability $1 - \theta$. Applying such a Bernoulli policy of rate $\theta \in [0, 1]$ induces a homogeneous Markov chain on state space $S = \{1, 2\}$ with transition matrix

$$B_\theta = \theta P_1 + (1 - \theta) P_2 = \begin{pmatrix} 0.7 + 0.3\theta & 0.3 - 0.3\theta \\ 0.2\theta & 1 - 0.2\theta \end{pmatrix}. \tag{2}$$

It is easily obtained that $b_\theta^T = (2\theta/(3 - \theta), (3 - 3\theta)/(3 - \theta))$ is the row vector corresponding to the stationary distribution of B_θ for any $\theta \in [0, 1]$. Using this, we can compute the performance function $g(\theta)$ of Bernoulli mixing policies as function

https://doi.org/10.1017/S0269964811000039 Published online by Cambridge University Press

of θ and it follows that

$$g(\theta) = \theta \left(\frac{2\theta}{3-\theta}, \frac{3-3\theta}{3-\theta}\right)\begin{pmatrix}0\\1\end{pmatrix} + (1-\theta)\left(\frac{2\theta}{3-\theta}, \frac{3-3\theta}{3-\theta}\right)\begin{pmatrix}0\\0\end{pmatrix} = \frac{3\theta-3\theta^2}{3-\theta}.$$

Obviously, $g(0) = g(1) = 0$ and $g(\theta) > 0$ for $\theta \in (0, 1)$. Moreover, $g(\theta)$ is continuously differentiable for $\theta \in [0, 1]$ and, thus, it follows that $g'(\theta^*) = 0$ for any optimal $\theta^* \in [0, 1]$. Since $g'(\theta) = 3(\theta^2 - 6\theta + 3)/(\theta - 3)^2$, it follows that $\theta^* = 3 - \sqrt{6} \sim 0.551$ is the unique value for $\theta \in [0, 1]$ that maximizes the performance over all Bernoulli mixing policies with varying rate θ. The performance of this optimal Bernoulli policy equals $g(3 - \sqrt{6}) = 15 - 6\sqrt{6} \sim 0.303$.

Being able to optimize over all Bernoulli mixing policies as in Example 1 is not the end of story with respect to optimizing the long-run average reward over all $\mathcal{D}$ mixing policies. One should also consider mixing policies that are nonstationary, which complicate matters because such policies induce inhomogeneous Markov chains. For Example 1, would the simple deterministic but nonstationary, round-robin policy $(d^1, d^2, d^1, d^2, \ldots)$ not improve the performance 0.303 of the best Bernoulli mixing policy? Moreover, could we improve even more maybe by applying decision rule d^1 slightly more often than d^2, as is the case for the optimal Bernoulli policy having rate $\theta^* = 3 - \sqrt{6} \sim 0.551$. For example, could a periodic policy with period cycle $(d^1, d^1, d^2, d^1, d^2, d^1, d^2, d^1, d^2)$ using decision rule d^1 with a proportion of $\frac{5}{9}$ be even better? Indeed, we will show that it gives a better performance and, moreover, we present strong evidence that this latter periodic policy is the overall optimal mixing policy for the problem described earlier (see Conjecture 34). The conjecture will be based both on theoretical results we obtain having implications on the structure of optimal mixing policies and numerical results that indicate that this policy is the best among all $\mathcal{D}$ mixing policies having the appropriate structure. On the other hand, we give arguments that the problem of optimizing over all $\mathcal{D}$ mixing policies is, in general, not an easy problem. It turns out that even for problems with small state and action spaces, as in Example 1, it might be hard to prove for some given policy that it is optimal within the class of $\mathcal{D}$ mixing policies.

Keeping Example 1 as key example in mind, we proceed in this article as follows. Section 2 introduces some basic notation and $\mathcal{D}$ restricted MDP concepts. In Section 2.1 the link between these concepts and open-loop control problems is emphasized. In Section 3.1 some basic properties of Bernoulli policies are given and methods to optimize within this particular class of $\mathcal{D}$ mixing policies are discussed. Subsequently, in Section 3.2 deterministic $\mathcal{D}$ mixing policies are introduced. Comparing with Bernoulli policies, the advantages and disadvantages of applying and optimizing within the class of deterministic $\mathcal{D}$ mixing policies are discussed. The problem of computing the performance of a given deterministic $\mathcal{D}$ mixing is considered, and for so-called periodic policies, a method is given and illustrated with an example.

In Section 4, for any given $\mathcal{D}$ restricted MDP, an associated (unrestricted) MDP is defined such that there is equivalence with the $\mathcal{D}$ restricted MDP. This equivalence gives some useful results on optimal policies and associated sample paths. Then some conditions for $\mathcal{D}$ restricted MDP are given that are shown to be sufficient for the existence of optimal stationary deterministic Markovian policies for the associated full observation MDP. Moreover, it is shown that these conditions are sufficient for performances of deterministic $\mathcal{D}$ mixing policies to be independent of the initial state distribution from which additional results are deduced. In Section 5 some subclasses of deterministic $\mathcal{D}$ mixing policies are introduced, algorithms are given to optimize the performance over such a subclass, and the efficiency of this approach is considered. A subclass of deterministic $\mathcal{D}$ mixing policies that is considered in particular are the so-called policies with regular structure for which the corresponding sequence of decision rules is a so-called regular sequence. Because this notion of regularity is often used in this article we now give a formal definition. More details and useful properties are given in Section 5.

Let $U = (u_1, u_2, \ldots)$ be an infinite sequence of zeros and ones. Denote by $s_k(n) := \sum_{j=k}^{k+n-1} u_j$ the number of ones in the subsequence of length n beginning at the kth element of U and put $s(n) := s_1(n)$. U is said to have a density of $\theta \in [0, 1]$ if $\lim_{n\to\infty}(s(n)/n) = \theta$. Thus, if an infinite sequence U of zeros and ones has a density θ, then θ is the asymptotic frequency of the ones in U. In that case, it might intuitively be clear that the positions of ones in the sequence are more regularly distributed if for all $k, n \in \mathbb{N}$, absolute deviations between $s_k(n)$ and $n\theta$ are small. The following fundamental definition is based on this intuition and defines exactly when an infinite sequence of zeros and ones is (most) regular. We also define when a sequence is so-called eventually regular.

DEFINITION 2: *Let $U = (u_1, u_2, \ldots)$ be an infinite sequence of zeros and ones. Then U is called regular of density θ if for $s_k(n)$, the number of ones in the corresponding subsequence of length n, it holds that*

$$|s_k(n) - n\theta| < 1 \text{ for every } k, n \in \mathbb{N}. \tag{3}$$

An infinite sequence of zeros and ones is called eventually regular if it has a suffix that is regular of some density.

One of the main results in this article will be the existence of optimal deterministic $\mathcal{D}$ mixing policies corresponding to a regular sequence if some conditions are satisfied. For open-loop control problems, the existence of optimal policies with a regular structure has been investigated previously and [1] gives an overview on this. The obtained results were for queuing networks assumed to have particular topological properties. Moreover, in [1] the main condition for this optimality of a regular policy is multimodularity of the performance function. For many problems, this condition of multimodularity is hard to check. In the present article the optimality of regularity is investigated for an MDP setting, which is applicable to many cases that are not in the framework of [1]. We will obtain sufficient conditions for the existence of a regular policy that is optimal and are entirely different from [1] in both formulation and possible applications. Indeed, the conditions obtained in the present article are

https://doi.org/10.1017/S0269964811000039 Published online by Cambridge University Press

formulated in terms of $\mathcal{D}$ restricted MDP and the associated full observation MDP. This is applicable for other types of problems than open-loop control in the particular queuing networks considered in [1].

In Section 5.2 we apply regular $\mathcal{D}$ mixing policies to the problem described in Example 1 and compare it with Bernoulli mixing policies. Thereafter, in Section 6 it is proved that for $\mathcal{D}$ restricted MDP with $\mathcal{D}$ consisting of (only) two different decision rules, some generally applicable conditions are sufficient for the existence of an optimal $\mathcal{D}$ mixing policies within the subclass of deterministic $\mathcal{D}$ mixing policies with regular structure. The associated full observation MDP is considered to formulate the main condition for this result. It is shown that for the full observation MDP, the existence of an optimal stationary deterministic Markovian policy having some type of threshold structure is, together with some easy checkable minor conditions, sufficient to obtain the result on optimality within the subclass of policies with regular structure. The application of this result is illustrated with an example. Some concluding remarks are made about possible generalizations of the main result and possible connections with comparable MDP or optimal control problems.

2. MIXING OF MDP DECISION RULES

In this article we consider an infinite-horizon discrete-time ($T = \{1, 2, \ldots\}$ is the set of decision epochs) MDP with a finite state space S and, for every $s \in S$, a finite action space A_s. Then $A = \cup_{s \in S} A_s$ is the common finite action space. Most readers will be familiar with the standard setup for MDP (as explained in textbooks such as [22,24]). We need only be clear about assumptions and notations. A Markov decision rule d maps the set of states into the set of probability distributions on the action space; that is, $d : S \to \mathcal{P}(A)$. A Markov policy, π is denoted $\pi = (d_1, d_2, \ldots)$, with the meaning that if at time t the process is in state s, then the probability of choosing any action $a \in A_s$ follow from the the action space distribution $d_t(s)$. A stationary Markov policy is one for which $\pi = (d, d, \ldots)$. A Markov deterministic (MD) decision rule is equivalent to a mapping from the state space to the action space $d : S \to A$ and an MD policy π applies an MD decision rule for all $t \in T$. The expected immediate reward upon taking action a in state s is denoted by $r(s, a)$. Our objective is to maximize expected average reward, which we may write as

$$g^{\pi}(x) := \liminf_{N \to \infty} \frac{1}{N} \mathbb{E}_x^{\pi} \left[\sum_{t=1}^{N} r(x_t, a_t) \right], \tag{4}$$

where $r(x_t, a_t)$ denotes the reward obtained at time t and $\mathbb{E}_x^{\pi}$ denotes the expectation conditional on the initial state distribution being x and policy π being employed.

In general terms, the problem we investigate in this article is optimization over policies for which all of the decision rules $d_t, t = 1, 2, \ldots,$ are restricted to be elements of some finite set of particular Markov decision rules. Thus, we would like to maximize

https://doi.org/10.1017/S0269964811000039 Published online by Cambridge University Press

for any initial state distribution x, the performance $g^{\pi}(x)$ over policies $\pi = (d_1, d_2, \ldots)$ restricted to $d_t \in \mathcal{D}$ for every $t \in T$, where $\mathcal{D}$ is a given finite set of MD decision rules.

Such policies will be called $\mathcal{D}$ mixing policies. In practical applications, the given set $\mathcal{D}$ of admissible decision rules typically consist of easy implementable deterministic decision rules determined by some straightforward heuristic. For example, in a routing problem such heuristic MD decision rules could be d^1 "route arriving jobs to the shortest queue" or d^2 "route arriving jobs to the queue being served by the fastest server." In general, such heuristic MD decision rules are suboptimal with respect to performance optimization.

2.1. Open-Loop Control and Corresponding $\mathcal{D}$ Mixing Policies

For many applications, performance optimization yields an MDP for which it is desirable to use an open-loop control mechanism. In this case, the choice of an action should not depend on the (current) system state. For example, if, at decision epochs, observing the current system state is relatively expensive, time-consuming, or not possible at all, then open-loop (state-independent) control should be considered. We would like to point out that the restriction $d_t \in \mathcal{D}$ we always have in this article determining the class of policies within, we seek an optimal policy can be seen as open-loop although d_t in itself is allowed to be closed-loop since the action determined by decision rule d_t may be state dependent. Vice versa, an important class of open-loop problems might be modeled as $\mathcal{D}$ restricted MDP, on which we focus in this article.

Indeed, for an open-loop control problem, assume that there is a common action space A for all states. The simplest case is $A = \{a, b\}$; that is, in every state, the same two actions a and b are available. For example, in a queueing problem with admission control with decision epochs corresponding to arrivals of jobs, action a could be to accept the new arriving job and action b to decline it. If $A = \{a, b\}$, then the only two decision rules that obey the rules of open-loop control are d^1 that chooses action a in every state $s \in S$ and d^2 that chooses action b in every state $s \in S$. Then d^1 induces a stationary Markov chain with some corresponding transition matrix P_1 and d^2 induces a stationary Markov chain with some corresponding transition matrix P_2. Moreover, any Markov open-loop control policy π is of the form $\pi = (d_1, d_2, \ldots)$ with $d_t \in \{d^1, d^2\}$ for every $t \in T$ and it follows that optimizing the performance over all open-loop control policies with two available actions in every state can be considered as a special case of optimization over $\mathcal{D} = \{d^1, d^2\}$ mixing policies. In general, optimizing open-loop control with any finite common action space A corresponds to optimization over $\mathcal{D}$ mixing policies, where $\mathcal{D}$ has the same cardinality as the action space A.

3. PERFORMANCE COMPUTATION AND OPTIMIZATION OF MIXING POLICIES

In the examples we give in this article, the set of admissible decision rules $\mathcal{D}$ consists of two MD decision rules, say d^1 and d^2. Then $\pi^1 = (d^1, d^1, \ldots)$ and $\pi^2 = (d^2, d^2, \ldots)$

https://doi.org/10.1017/S0269964811000039 Published online by Cambridge University Press

are the only stationary deterministic policies that are feasible for a $\mathcal{D} = \{d^1, d^2\}$ restricted MDP. Both π^1 and π^2 are not optimal if d^1 (respectively d^2) are suboptimal decision rules. Then some Markov policies $\pi = (d_1, d_2, \ldots)$ with $d_t \in \{d^1, d^2\}$ for all $t \in T$ could improve the performance of both π_1 and π_2 if π is not restricted to be both stationary and deterministic. In that case, our goal is to maximize the performance g^π over the set of admissible policies for $\mathcal{D}$ restricted MDP resulting in a performance that is strictly larger than $\max(g^{\pi_1}, g^{\pi_2})$. To obtain an admissible policy π with $g^\pi > \max(g^{\pi_1}, g^{\pi_2})$, the decision rules d^1 and d^2 have to be *mixed* in some way.

Since π_1 and π_2 are stationary policies, they induce stationary discrete-time Markov chains on S with corresponding transition matrixes (say, P_1 and P_2, respectively). We assume that both Markov chains are unichain and aperiodic. In other words, both Markov chains have exactly one recurrent class that is aperiodic and let p_1 and p_2 be the corresponding unique stationary distributions satisfying $p_1^T = p_1^T P_1$, $\sum_{s\in S} p_1(s) = 1$ (respectively $p_2^T = p_2^T P_2$, $\sum_{s\in S} p_2(s) = 1$), where p_1^T and p_2^T are the row vectors representing the stationary distributions p_1 and p_2, respectively. The finiteness of S guarantees the existence of p_1 and p_2 and the performances g^{π_1} and g^{π_2} of both policies might be directly computed from p_1 and p_2, respectively. From the existence of such unique stationary distributions p_1 and p_2, it follows that the performances g^{π_1} and g^{π_2} of the two stationary policies are independent of the initial state distribution. Indeed, for all initial state distributions on S, the performance of policy π_1 is given by $g^{\pi_1} = \sum_{s\in S} p_1(s) r(s, d^1(s))$. Similarly, $g^{\pi_2} = \sum_{s\in S} p_2(s) r(s, d^2(s))$ gives the performance of policy π_2.

We can generalize these formulas to compute the performance of any (randomized) stationary policy $\pi = (d, d, \ldots) \in \Pi^{MR}$, where d is some randomized decision rule inducing a stationary unichain aperiodic Markov chain. Indeed, let p be the unique stationary distribution on state space S of the induced Markov chain and let $r(d)_s := \sum_{a\in A_s} r(s,a)\mathrm{P}(a|s,d)$ be the expected immediate reward in state $s \in S$ given that MR decision rule d is applied. Then $r(d)$ is the vector containing the expected rewards for all states $s \in S$ if decision rule d is applied, and the expected performance of π is given by

$$g^\pi = \sum_{s\in S} p_s r(d)_s = p \cdot r(d), \text{ the inner product of } p \text{ and } r(d). \qquad \mathbf{(5)}$$

3.1. Bernoulli Policies

Recall that in Example 1 we obtained the optimal Bernoulli policy for the $\mathcal{D}$ restricted MDP introduced in that example. In this subsection we introduce Bernoulli policies for general $\mathcal{D}$ restricted MDP and discuss some main properties and methods to optimize over Bernoulli policies.

A Bernoulli policy can be implemented as follows. Given some MDP and $\mathcal{D} = \{d^1, d^2\}$, a set of two admissible MD decision rules for controlling the MDP. Consider the following randomized algorithm to generate $\mathcal{D}$ mixing policies $\pi = (d_1, d_2, \ldots)$.

https://doi.org/10.1017/S0269964811000039 Published online by Cambridge University Press

Let $\theta \in [0,1]$ be given. For $t = 1, 2, \ldots$, generate independent random numbers u_t uniformly distributed on $[0,1]$ and put

$$d_t = \begin{cases} d^1 & \text{if } u_t \in [0,\theta] \\ d^2 & \text{if } u_t \in (\theta, 1]. \end{cases}$$

In other words, for every decision epoch $t \in T$, an independent θ-coin is flipped; its outcome determines the decision rule that is applied at t. For all $t \in T$ with probability θ, the first decision rule is applied, and with probability $1 - \theta$, the second decision rule is applied. Policies generated by this randomized algorithm are called Bernoulli policies of rate θ. Note that this implementation of a Bernoulli policy can easily be generalized for the case that $\mathcal{D}$ consists of more than two decision rules, but this generalization is not explored in this article.

The randomization of the policy in the Bernoulli algorithm makes actual implementation of such policies in practice somewhat awkward, but a nice property is that the performance of Bernoulli policies is relatively easy to compute or approximate. This makes it tractable to optimize the performance over all Bernoulli policies and, in particular, analytic methods are available for optimizing the Bernoulli parameter θ. The following property of the Bernoulli policy is useful to analyze and compute or approximate performances.

LEMMA 3: *Assume an MDP with finite state space S where decisions rules d^1 and d^2 induce stationary and aperiodic unichain Markov chains with corresponding transition matrices P_1 (respectively P_2). Then any Bernoulli policy mixing d^1 and d^2 with rate $\theta \in [0,1]$ induces a stationary aperiodic unichain Markov chain on S with transition matrix*

$$B_\theta = \theta P_1 + (1-\theta) P_2, \tag{6}$$

which has an unique stationary distribution b_θ satisfying $b_\theta^T B_\theta = b_\theta^T$ and $\sum_{s \in S} b_\theta(s) = 1$, where b_θ^T is the row vector representing b_θ.

In other words, the Bernoulli policy mixing two decision rules induces a stationary Markov chain with unique stationary distribution b_θ depending on the Bernoulli parameter θ. From this, it follows that given the MDP and decision rules d^1 and d^2, the expected performance of the Bernoulli policy does not depend on the initial state distribution x and is a function $g(\theta)$ of the Bernoulli parameter θ. By (5) we have

$$\begin{aligned} g(\theta) &= \sum_{s \in S} (b_\theta)_s [\theta r(s, d^1(s)) + (1-\theta) r(s, d^2(s))] \\ &= \theta(b_\theta \cdot r(d^1)) + (1-\theta)(b_\theta \cdot r(d^2)). \end{aligned} \tag{7}$$

Optimizing the expected performance $g(\theta)$ of Bernoulli policies over $\theta \in [0,1]$ is relatively easy if $g(\theta)$ is a smooth function of $\theta \in [0,1]$. Indeed (see, e.g., [7] or [12]), it follows for a family of Bernoulli policies as given by (6) and any $\theta \in [0,1]$ that $g(\theta)$

https://doi.org/10.1017/S0269964811000039 Published online by Cambridge University Press

is n-differentiable at θ for any $n \in \mathbb{N}$ and, thus, smooth at θ. Combining this with (7) gives the following result.

Proposition 4: *The performance function $g(\theta)$ is smooth on the interval $[0, 1]$ and there exists some $\theta^* \in [0, 1]$ maximizing the performance $g(\theta)$ of Bernoulli policies. For any θ^* maximizing $g(\theta)$, it holds that $g'(\theta^*) = 0$ if $\theta^* \notin \{0, 1\}$.*

Remark 5: In the case of complex systems for which the MDP has a very large state space, it may not be tractable to obtain exact expressions for the stationary distribution b_θ and performance function $g(\theta)$. However, then an approximation of the stationary distribution b_θ can be obtained by methods like Markov chain Monte Carlo and then the expected performance $g(\theta)$ of the Bernoulli policy could also be approximated by plugging in the approximation of b_θ in (7). In this way, the optimal θ^* maximizing $g(\theta)$ can be approximated in such cases. Alternatively, gradient estimation by measure-valued differentiation could be applied to approximate some (optimal) value $\theta^* \in [0, 1]$ for which $g'(\theta^*) = 0$. In [6] this simulation technique is applied to a call center operation problem with two types of jobs having different service requirements for which in various ways, two reasonable applicable decision rules are obtained that are mixed to improve the system performance. The technique is relatively fast to approximate an optimal value for θ. For more theoretical results and background on this, see, for example, [12–14].

3.2. Deterministic Nonstationary Policies

To maximize the performance of some $\mathcal{D}$ mixing policy π, it is desirable that $g^\pi(x)$ as defined by (4) does not depend on the initial state distribution x. However, for nonstationary $\mathcal{D}$ mixing policies $\pi = (d_1, d_2, \ldots)$ with $d_t \in \mathcal{D}$ for $t = 1, 2, \ldots$, the performance might depend on the initial state distribution even if all transition matrixes corresponding to decision rules in $\mathcal{D}$ are unichain and aperiodic. Example 7 will illustrate this. Therefore, in Section 4 we will provide some additional (in addition to being unichain and aperiodic) sufficient condition on the transition matrixes associated with $\mathcal{D}$ such that also for nonstationary $\mathcal{D}$ mixing policies π, the performance will not depend on the initial state distribution. Under this assumption, the performance of some $\mathcal{D}$ mixing policy π might simply be denoted by g^π as in (5) or (7) for stationary (randomized) $\mathcal{D}$ mixing policies.

In this subsection we consider deterministic $\mathcal{D}$ mixing policies that are represented as an infinite deterministic sequence describing for every decision epoch for which the MD decision rule in $\mathcal{D}$ is applied. For $\mathcal{D} = \{d^1, d^2\}$, if we let symbol 1 correspond to decision rule d^1 and symbol 0 correspond to decision rule d^2, then we have a one-to-one correspondence between deterministic $\mathcal{D}$ mixing policies and one-sided infinite sequences $U = (u_1, u_2, \ldots)$ of zeros and ones. Therefore, an infinite sequence $(u_1, u_2, \ldots)$ is identified with a deterministic $\mathcal{D}$ mixing policy where u_t determines the decision rule that is applied at decision epoch t for $t = 1, 2, \ldots$. Thus, if $\mathcal{D} = \{d^1, d^2\}$, then optimizing the performance over all deterministic $\mathcal{D}$ mixing policies corresponds

https://doi.org/10.1017/S0269964811000039 Published online by Cambridge University Press

to optimization over the set $\{0,1\}^{\mathbb{N}}$ of all one-sided infinite sequences of zeros and ones. More generally, if $\mathcal{D} = \{d^1, d^2, \ldots, d^n\}$, then it follows analogously that optimizing the performance over all deterministic $\mathcal{D}$ mixing policies corresponds to optimization over a corresponding set W, where $W = \{\mathcal{A}\}^{\mathbb{N}}$ are all one-sided infinite words over some finite alphabet $\mathcal{A}$. A word is, by definition, a sequence of symbols from a finite alphabet, and for $\mathcal{D} = \{d^1, d^2, \ldots, d^n\}$, the corresponding alphabet $\mathcal{A}$ consists of n (different) symbols.

One of the positive aspects of applying a deterministic $\mathcal{D}$ mixing policy such as represented above as infinite decision sequence $U = (u_1, u_2, \ldots)$ is that the implementation is more straightforward than for randomized policies like the Bernoulli policies. Indeed, at decision epoch t, only the (easy implementable) MD decision rule determined by u_t has to be implemented and it is not necessary to "flip a coin" (randomization) at every decision epoch. Moreover, arguably the most important advantage of deterministic mixing policies over Bernoulli policies is that, in general, good (not necessarily optimal) deterministic mixing policies easily outperform the best (optimized) Bernoulli policies.

We have seen (Lemma 3) that an advantage of applying a stationary (Bernoulli) mixing policy is that it induces a stationary Markov chain on the state space S. However, deterministic mixing policies given by some infinite decision sequence $U = (u_1, u_2, \ldots)$ as described earlier do not induce a stationary Markov chain except for degenerate policies for which $u_t = u_1$ for every decision epoch t. Therefore, it is also not possible to obtain the performance of deterministic mixing policies by computing a unique stationary distribution. The fact that computing the performance is harder than for Bernoulli policies is one of the reasons that optimizing over deterministic mixing policies is much harder than for Bernoulli policies.

For periodic deterministic mixing policies, there exists an algorithm to compute the performance, but computation times will increase with the period. A deterministic mixing policy is periodic with period k if for the corresponding decision sequence $U = (u_1, u_2, \ldots)$, it holds that $u_t = u_{t+k}$ for $t = 1, 2, \ldots$. Theorem 6 yields a formula for computing the performance of periodic deterministic mixing policies under some assumptions.

THEOREM 6: *Let π be a deterministic $\mathcal{D}$ mixing policy with corresponding decision sequence $U = (u_1, u_2, \ldots)$ that is periodic with period k. Let X_t be the state at decision epoch t when policy π is applied and let $d_t \in \mathcal{D}$ be the decision rule corresponding to u_t to be applied at decision epoch t. For $m = 1, 2, \ldots, k$, assume that the stationary Markov chain $\{X_t, t = m, m+k, m+2k, \ldots\}$ has unique stationary distribution b_m. Then for the long-run average reward g^π, we have that*

$$g^\pi = \frac{1}{k} \sum_{m=1}^{k} b_m \cdot r(d^m). \tag{8}$$

In Theorem 6 the assumption is made that for all m, the stationary Markov chain $\{X_t, t = m, m+k, m+2k, \ldots\}$ has an unique stationary distribution. Note that, according to (8), this implies that the performance g^{π} of a such a periodic policy π does not depend on the initial state distribution. This is a necessary condition and it should be realized that for this condition to hold, it is not sufficient that all the transition matrixes associated to the decision rules in $\mathcal{D}$ are unichain and aperiodic. Recall that unichain and aperiodic is sufficient (Lemma 3) for Bernoulli policies, but for periodic deterministic mixing policies, we have the following counterexample.

Example 7: Let $\mathcal{D} = \{d^1, d^2\}$, and the periodic deterministic $\mathcal{D}$ mixing policy π corresponding to decision sequence $U = (1,0,1,0,\ldots) = (1,0)^{\infty}$ is applied. Let

$$P_1 = \begin{pmatrix} 0 & 0.5 & 0.5 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} \quad \text{and} \quad P_2 = \begin{pmatrix} 0 & 0 & 1 \\ 0.5 & 0 & 0.5 \\ 0 & 1 & 0 \end{pmatrix}$$

be the transition matrixes corresponding to decision rules d^1 (respectively d^2). It is easily seen that P_1 and P_2 are unichain and aperiodic. However,

$$A_1 := P_1 P_2 = \begin{pmatrix} 0.25 & 0.5 & 0.25 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

is obviously not unichain and, thus, the Markov chain $\{X_1, X_3, X_5, \ldots\}$ does not have an unique stationary distribution. Similarly,

$$A_2 := P_2 P_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0.5 & 0.25 & 0.25 \\ 0 & 0 & 1 \end{pmatrix}$$

is not unichain and, thus, also the Markov chain $\{X_2, X_4, X_6, \ldots\}$ does not have an unique stationary distribution. In fact, in this example the performance of policy π in general depends on the initial state distribution and cannot be computed by (8).

In Section 4 we will give some conditions on the transition matrixes corresponding to the decision sequences in $\mathcal{D}$ that are sufficient for the independence of the performance of $\mathcal{D}$ mixing policies on the initial state distribution and that under such conditions, (8) is certainly valid.

We also note that (8) could be seen as a generalization of (5). Indeed, the latter formula is then for the special case $k = 1$. Additionally, it follows that if (8) is applied to compute the performance, the computational effort increases with the period k of the decision sequence U. In fact, it is easily seen that for a fixed set $\mathcal{D}$ of admissible decision rules, the computation time of computing the performance of a periodic deterministic $\mathcal{D}$ mixing policy by applying (8) increases linearly with the period k of the decision sequence.

In contrast, in the case of stationary (Bernoulli) policies, often a closed formula for the performance $g(\theta)$ can be given. In fact, for Bernoulli policies, the computational

https://doi.org/10.1017/S0269964811000039 Published online by Cambridge University Press

effort to obtain the performance hardly depends on which Bernoulli policy is applied, as all Bernoulli policies induce stationary Markov chains. Thus, in comparison with Bernoulli policies, computing the performance is harder for deterministic mixing policies because, in general, such policies do not induce stationary Markov chains. Moreover, for periodic policies, the computational effort increases with the period. In addition, for deterministic mixing policies that are not periodic, we do not have an algorithm (even if the assumptions are satisfied that the performance is independent of the initial state distribution) to compute the exact performance in finite time and we think it is, in general, only possible to approximate the performance of such a policy. Therefore, the optimization over deterministic mixing policies is harder than optimization over all Bernoulli policies. Another issue is that deterministic $\mathcal{D}$ mixing policies should be optimized over the infinite discrete set W of all possible decision sequences $U = (u_1, u_2, \ldots)$ with $u_n \in \mathcal{D}$ for $n = 1, 2, \ldots$, the structure of which is more complicated than for Bernoulli policies for which the optimization can be done over a bounded convex set, as we have seen in Example 1.

4. THE ASSOCIATE MDP

In this section we define an associated MDP that is equivalent to optimizing over $\mathcal{D}$ mixing policies for $\mathcal{D} = \{d^1, d^2\}$. The advantage of considering the associated MDP is that decision rules are no longer restricted to $\mathcal{D}$. Therefore, in contrast to the class of $\mathcal{D}$ mixing policies, the existence of optimal Markov policies that are both stationary and deterministic holds if certain conditions are satisfied. Then the existence of an optimal stationary deterministic Markov policy for the associated MDP can be used to obtain structural properties of some optimal policy within the class of $\mathcal{D}$ mixing policies. In this way, we will obtain results about optimality within certain subclasses of $\mathcal{D}$ mixing policies if some conditions are satisfied. In addition to such benefits, the associated MDP formulation also gives new issues to consider. For example, in the associated MDP, the state space is not finite anymore, but continuous. Therefore despite the equivalence of the associated MDP formulation, obtaining an optimal $\mathcal{D}$ mixing policy remains, in general, a difficult problem. However, from the equivalence, we will obtain some structural results.

Before we give the associated continuous state space MDP that is equivalent to optimizing within the class of $\mathcal{D} = \{d^1, d^2\}$ mixing policies, we recall some definitions and notations. Let S be the finite state space and $r(d^1)$ ($r(d^2)$) be the immediate reward vector for decision rule d^1 (respectively d^2). Moreover, let P_1 be the transition matrix associated to d^1 and P_2 be the transition matrix associated to d^2. Then the associated MDP with continuous state space is defined as follows:

- The state space X is the set of all probability distributions on S.
- The action space $\widetilde{A} := \{d^1, d^2\}$ for all $x \in X$.
- For all $x \in X$, the immediate rewards $r(x, d^1)$ and $r(x, d^2)$ are given by the inner products $r(x, d^1) := x \cdot r(d^1)$ (respectively $r(x, d^2) := x \cdot r(d^2)$).

https://doi.org/10.1017/S0269964811000039 Published online by Cambridge University Press

- For action d^1, state transitions are given by the state space mapping $x \to xP_1$ for all $x \in X$, where x is represented as an $|S|$-dimensional row vector. Analogously, for d^2, state transitions are given by the state space mapping $x \to xP_2$ for all $x \in X$ being represented by a row vector.
- Let $\widetilde{\Omega} = \{X \times \widetilde{A}\}^\infty$ be the sample space for the stochastic process generated by the MDP when some admissible policy $\widetilde{\pi}$ is applied. A sample path $\widetilde{\omega} \in \widetilde{\Omega}$ is an alternating sequence $\widetilde{\omega} = (x_1, a_1, x_2, a_2, \ldots)$ of states and actions. For $t = 1, 2, \ldots$, let variables $\widetilde{X}_t$ and $\widetilde{Y}_t$ be defined by $\widetilde{X}_t(\omega) = x_t$ and $\widetilde{Y}_t(\omega) = a_t$, respectively. The optimality criterion is again the lim inf average reward criterion. In other words, for initial state distribution $\widetilde{X}_1 = x \in X$, the performance of policy $\widetilde{\pi}$ is given by

$$g^{\widetilde{\pi}}(x) := \liminf_{N\to\infty} \frac{1}{N} \mathbb{E}_x^{\widetilde{\pi}} \left\{ \sum_{t=1}^{N} r(\widetilde{X}_t, \widetilde{Y}_t) \right\}. \qquad \textbf{(9)}$$

The equivalence between $\mathcal{D}$ restricted MDP with finite state space S and the above-defined MDP with a state space X of all probability distributions on S follows from the well-known equivalence between a partial observation MDP and an associated full observation MDP because we also have equivalence between $\mathcal{D}$ restricted MDP and a partial observation MDP, as explained at the end of Section 2. Equivalence between a partial observation MDP and a corresponding full observation MDP is applied in [17] for a particular problem, whereas in [8] the equivalence is described and explored in a more general setup. Here, we do not go in details about the equivalence between both models, but for this article the following properties are most important.

Let $\widetilde{\pi}$ be a Markov policy to be applied for the full observation MDP and $\widetilde{\omega} = (x_1, a_1, x_2, a_2, \ldots) \in \widetilde{\Omega}$ be an associated sample path. Define π as the $\mathcal{D}$ mixing policy defined by the infinite sequence of decision rules $(a_1, a_2, \ldots)$ corresponding to sample path $\widetilde{\omega}$. Then for the performances $g^{\widetilde{\pi}}(x_1)$ and $g^{\pi}(x_1)$ as defined by (9) and (4), respectively, it almost surely holds that $g^{\widetilde{\pi}}(x_1) = g^{\pi}(x_1)$. Similarly, it follows that if there exists an optimal stationary deterministic Markov policy $\widetilde{\pi}$ for the full observation MDP, then the deterministic $\mathcal{D}$ mixing policy π obtained from the sample path $\widetilde{\omega}$associated to $\widetilde{\pi}$ is an optimal $\mathcal{D}$ mixing policy for the initial state distribution x_1.

Next, two conditions for $\mathcal{D}$ restricted MDP are given. If they are satisfied, then some useful results (Theorem 10 and Corollary 11) follow immediately from the above explained equivalence between $\mathcal{D}$ restricted MDP and the associated full observation MDP.

Condition 8: *For all deterministic $\mathcal{D}$ mixing policies π and initial state distributions $x, y \in X$, it holds that $g^{\pi}(x) = g^{\pi}(y)$. In other words, performances of deterministic $\mathcal{D}$ mixing policies do not depend on the initial state distribution and the performance of such a policy π can be denoted by g^{π}.*

Condition 9: *There exists some optimal stationary deterministic Markov policy for the full observation MDP associated to the $\mathcal{D}$ restricted MDP.*

https://doi.org/10.1017/S0269964811000039 Published online by Cambridge University Press

Theorem 10: *Suppose Condition* 8 *is satisfied for a $\mathcal{D}$ restricted MDP. Let $\widetilde{\pi}$ be a stationary deterministic Markov policy for the associated full observation MDP and $\widetilde{\omega} = (x_1, a_1, x_2, a_2, \ldots) \in \widetilde{\Omega}$ be an associated sample path. For $t = 1, 2, \ldots,$ let π_t be the deterministic $\mathcal{D}$ mixing policy given by the infinite sequence of decision rules $(a_t, a_{t+1}, \ldots)$. Then the performances $g^{\widetilde{\pi}}(x_1)$ and $g^{\pi_t}(x_1)$ for $t = 1, 2, \ldots$ are independent of the initial state distribution x_1 and, moreover, we have that*

$$g^{\widetilde{\pi}} = g^{\pi_t} \quad \textit{for } t = 1, 2, \ldots. \tag{10}$$

Corollary 11: *Suppose Conditions* 8 *and* 9 *are satisfied for some $\mathcal{D}$ restricted MDP. Let $\widetilde{\pi}$ be an optimal stationary deterministic Markov policy for the associated full observation MDP, and for $t = 1, 2, \ldots,$ let π_t be the deterministic $\mathcal{D}$ mixing policy defined as in Theorem* 10. *Then for all $t = 1, 2, \ldots,$ policy π_t is an optimal $\mathcal{D}$ mixing policy.*

Note that Corollary 11 implies that for $\mathcal{D}$ restricted MDP, there exists an optimal policy that is not randomized if Conditions 8 and 9 are satisfied. In Section 6 we obtain under some additional conditions, the existence of an optimal $\mathcal{D}$ mixing policy within the class of deterministic $\mathcal{D}$ mixing policies corresponding to so-called regular sequences.

4.1. Stationary Optimal Policies for the Associated MDP

Next, we wish to apply Corollary 11 to obtain structural results on optimal $\mathcal{D}$ mixing policies. To apply Corollary 11, it is sufficient that Conditions 8 and 9 are satisfied. In Example 7 we have seen that for Condition 8 to be satisfied it is not sufficient for the relevant transition matrixes to be unichain and aperiodic. Moreover, because an associated full observation MDP has an uncountable state space X, it is, in general, a priori not clear whether Condition 9 is satisfied. Therefore, we wish to apply a sufficient condition according to Corollary 4.1 in [9] for Condition 9 to be satisfied. This result basically states that for an MDP with finite action set equivalent with a partially observable MDP with finite state space, a uniformly boundedness condition is sufficient for the existence of an appropriate solution of the corresponding average cost optimality equation (ACOE) implying the existence of optimal stationary deterministic Markov policies. Then Condition 9 is satisfied. Moreover, in the case that the uniformly boundedness condition is applicable, which implies that Condition 9 is satisfied, it will follow that Condition 8 is satisfied as well. The uniformly boundedness condition as given in [9] is that the difference in optimal discounted costs is uniformly bounded over the state space X. First, we reformulate the condition given in [9] for rewards instead of costs.

Let $x \in X$ be an initial state, and for $t = 1, 2, \ldots,$ let variables $\widetilde{X}_t$ and $\widetilde{Y}_t$ be defined as in (9) for policy $\widetilde{\pi}$. Since the number of components $|S|$ of both reward vectors $r(d^1)$ and $r(d^2)$ is finite with no loss of generality, we assume for the obtained rewards $r(\widetilde{X}_t, \widetilde{Y}_t)$ of the reward process that $0 \leq r(\widetilde{X}_t, \widetilde{Y}_t) \leq B$ for $t = 1, 2, \ldots.$ For discount

https://doi.org/10.1017/S0269964811000039 Published online by Cambridge University Press

factor $0 < \beta < 1$, initial state $x \in X$, and policy $\widetilde{\pi}$, the discounted reward $R_\beta(x, \widetilde{\pi})$ (DR) is defined by

$$R_\beta(x, \widetilde{\pi}) := \lim_{N \to \infty} \mathbb{E}_x^{\widetilde{\pi}} \left\{ \sum_{t=1}^{N} \beta^{t-1} r(\widetilde{X}_t, \widetilde{Y}_t) \right\}. \quad \textbf{(11)}$$

By the assumption on the rewards, the limit in (11) exists and is nonnegative. The optimal β-discounted reward for initial state $x \in X$ is given by $R_\beta^*(x) := \sup_{\widetilde{\pi}} R_\beta(x, \widetilde{\pi})$, the supremum being taken over all admissible policies. With these definitions, the uniformly boundedness condition for optimal discounted rewards is the following.

Condition 12: *There exists some $M \in \mathbb{R}$ such that for all $x, y \in X$ and $0 < \beta < 1$, it holds that*

$$|R_\beta^*(x) - R_\beta^*(y)| \leq M. \quad \textbf{(12)}$$

After some additional notation and definitions, we are able to present conditions on the existence of stationary optimal policies for full observation MDP associated with some $\mathcal{D}$ restricted MDP. These conditions are easy to check for the problems considered in this article. It follows that if these conditions are satisfied, then Condition 12 is satisfied and, therefore, Conditions 8 and 9 as well. It turns out for Condition 12 to be satisfied that some sufficient conditions on transition matrixes induced by decision rules in $\mathcal{D}$ can be formulated in terms of Dobrushin's coefficient of ergodicity of a transition matrix.

Definition 13: *Let $P = (p_{ij})$ be a transition matrix on some finite state space S. Dobrushin's coefficient of ergodicity of P is defined as*

$$\rho_0(P) = \frac{1}{2} \max_{i,j} \sum_{k=1}^{|S|} |p_{ik} - p_{jk}|. \quad \textbf{(13)}$$

Lemma 14 states some well-known (see, e.g., [21]) useful properties of Dobrushin's coefficient.

Lemma 14:

1. $0 \leq \rho_0(P) \leq 1$.
2. $\rho_0(P) = 0$ *if and only if P has identical rows.*
3. $\rho_0(P_1 \cdot P_2) \leq \rho_0(P_1) \cdot \rho_0(P_2)$.
4. *There exists some positive integer N with $\rho_0(P^N) < 1$ if and only if P is unichain and aperiodic.*

A useful property of Dobrushin's coefficient has to do with the l_1-distance between probability distributions on the finite state space S. For $x, y \in X$, denote by

$$||x - y||_1 := \sum_{i=1}^{|S|} |x_i - y_i|$$

https://doi.org/10.1017/S0269964811000039 Published online by Cambridge University Press

the l_1-distance between probability distributions x and y on S. Then the following lemma (see [21]) holds.

LEMMA 15: $||x-y||_1$ *is a metric on the set of state space probability distributions* X *with the property that* $||x-y||_1 \leq 2$ *for all* $x,y \in X$. *Moreover, for any* $x,y \in X$ *and transition matrix* P *on* S, *we have that*

$$||xP - yP||_1 \leq \rho_0(P)||x-y||_1. \tag{14}$$

In other words, if $\rho_0(P) < 1$, then P induces a contraction mapping on X. In the following results, Lemmas 14 and 15 will be applied to show that Condition 12 is satisfied under several specific assumptions on the transition matrices.

THEOREM 16: *Let* $\mathcal{D} = \{d^1, d^2\}$ *and let* P_1 *and* P_2 *be the transition matrixes induced by decision rule* d^1 (*respectively* d^2). *Assume that (at least) one of the two transition matrixes has Dobrushin's coefficient smaller than* 1 *and both transition matrices are unichain and aperiodic. Then Condition* 12 *is satisfied.*

PROOF: Let $x,y \in X$ be arbitrarily chosen probability distributions. We should show that for all discount factors $0 < \beta < 1$, it holds that $|R^*_\beta(x) - R^*_\beta(y)| \leq M$ for some $M \in \mathbb{R}$. Let $0 < \beta < 1$ be any given discount factor. Without loss of generality, we assume that $R^*_\beta(x) \geq R^*_\beta(y)$. Moreover, according to Theorem 4.2.3 in [15], there exists some optimal stationary deterministic Markov policy $\widetilde{\pi}$ for the given β. Let $\widetilde{\omega} = (x_1, a_1, x_2, a_2, \ldots)$ be the sample path with initial state $x_1 = x$ for this optimal policy $\widetilde{\pi}$. Tracking sample path $\widetilde{\omega}$ for $t = 1, 2, \ldots$, let $r(a_t)$ be the reward vector for decision rule $a_t \in \mathcal{D}$, $A_t \in \{P_1, P_2\}$ be the transition matrix corresponding to a_t, and B_t be the matrix product given by $B_t := \prod_{k=1}^{t-1} A_t$ with the convention that B_1 is the identity matrix. Then $x_t = x_1 B_t$ for $t = 1, 2, \ldots$, and by (11), we have that

$$R^*_\beta(x) = R_\beta(x, \widetilde{\pi}) = \lim_{k\to\infty} \sum_{t=1}^{k} \beta^{t-1} r(x_t, a_t) = \lim_{k\to\infty} \sum_{t=1}^{k} \beta^{t-1} (xB_t) \cdot r(a_t). \tag{15}$$

The infinite sequence of decision rules $(a_1, a_2, \ldots)$ defines a policy $\widetilde{\pi}'$ for which the sample path $\widetilde{\omega}'$ for initial state $y_1 = y$ is given by $\widetilde{\omega}' = (y_1, a_1, y_2, a_2, \ldots)$ with $y_t = yB_t$ for $t = 1, 2, \ldots$. Hence,

$$R^*_\beta(y) \geq R_\beta(y, \widetilde{\pi}') = \lim_{k\to\infty} \sum_{t=1}^{k} \beta^{t-1} r(y_t, a_t) = \lim_{k\to\infty} \sum_{t=1}^{k} \beta^{t-1} (yB_t) \cdot r(a_t). \tag{16}$$

Recall that we could assume that all components of the reward vectors $r(a_t)$ are nonnegative and bounded from above by some $B > 0$. Thus, by (15), (16), and

https://doi.org/10.1017/S0269964811000039 Published online by Cambridge University Press

Lemma 15, we have

$$R_\beta^*(x) - R_\beta^*(y) \leq \lim_{k\to\infty} \sum_{t=1}^{k} \beta^{t-1}((xB_t)\cdot r(a_t) - (yB_t)\cdot r(a_t))$$

$$= \lim_{k\to\infty} \sum_{t=1}^{k} \beta^{t-1}(xB_t - yB_t)\cdot r(a_t) \leq \lim_{k\to\infty} \sum_{t=1}^{k} ||xB_t - yB_t||_1 B$$

$$\leq B \lim_{k\to\infty} \sum_{t=1}^{k} \rho_0(B_t)||x-y||_1 \leq 2B \lim_{k\to\infty} \sum_{t=1}^{k} \rho_0(B_t). \quad \textbf{(17)}$$

Without loss of generality, we may assume that $\rho_0(P_1) = \gamma_1 < 1$. Moreover, since P_2 is unichain and aperiodic, there exists by property 4 of Lemma 14, some $N \in \mathbb{N}$ such that $\rho_0(P_2^N) = \gamma_2 < 1$. Put $\gamma = \max(\gamma_1, \gamma_2)$. Then it follows by properties 1 and 3 of Lemma 14 that $\rho_0(B_{N+1}) \leq \gamma < 1$, as the matrix product B_{N+1} contains at least one P_1 or $B_{N+1} = P_2^N$. Similarly, it follows that $\rho_0(B_{t+N}) \leq \gamma\rho_0(B_t)$ for $t = 1, 2, \ldots$. Combining this with $0 \leq \rho_0(B_t) \leq 1$ for $t = 1, 2, \ldots$, it follows that

$$\lim_{k\to\infty} \sum_{t=1}^{k} \rho_0(B_t) \leq N + \lim_{k\to\infty} \sum_{t=N+1}^{k} \rho_0(B_t)$$

$$= N + \lim_{k\to\infty} \sum_{t=1}^{k} \rho_0(B_{t+N}) \leq N + \gamma \lim_{k\to\infty} \sum_{t=1}^{k} \rho_0(B_t).$$

Hence, $(1-\gamma)\lim_{k\to\infty}\sum_{t=1}^{k}\rho_0(B_t) \leq N$ and, thus, $\lim_{k\to\infty}\sum_{t=1}^{k}\rho_0(B_t) \leq N/(1-\gamma)$. Combining this with (17), we obtain $R_\beta^*(x) - R_\beta^*(y) \leq 2BN/(1-\gamma)$. Thus, we have shown that for all $x, y \in X$ and $0 < \beta < 1$, it holds that

$$|R_\beta^*(x) - R_\beta^*(y)| \leq \frac{2BN}{1-\gamma} \quad \textbf{(18)}$$

and, thus, Condition 12 is satisfied with $M = 2BN/(1-\gamma)$. ■

We have just shown that Condition 12 is satisfied if both transition matrixes P_1 and P_2 are unichain and aperiodic and at least one of them has Dobrushin's coefficient smaller than 1. Because Condition 12 is satisfied, it also follows that Condition 9 is satisfied. Moreover, in the proof of Theorem 16, we have shown something additional that is also useful—namely it also follows that for any deterministic $\mathcal{D}$ mixing policy $\pi = (a_1, a_2, \ldots)$ and any time t, the difference in expected accumulated total (undiscounted) rewards up to time t for any two initial state distributions $x, y \in X$ is uniformly bounded by $M = 2BN/(1-\gamma)$. From this, it immediately follows that the expected long-run average reward g^π of such a policy π does not depend on the initial state distribution. Thus, Condition 8 is also satisfied for a $\mathcal{D}$ restricted MDP

https://doi.org/10.1017/S0269964811000039 Published online by Cambridge University Press

as in Theorem 16. Thus, Theorem 10 and Corollary 11 are applicable for such a $\mathcal{D}$ restricted MDP.

Example 17: Consider the $\mathcal{D} = \{d^1, d^2\}$ restricted MDP considered in Example 1. Recall (1) describing P_1, P_2, $r(d^1)$, and $r(d^2)$. It follows that $\rho_0(P_1) = 0.8$ and $\rho_0(P_2) = 0.7$. Hence, (18) holds for $B = 1, N = 1$, and $\gamma = 0.8$ and, thus, Condition 12 is satisfied for $M = 10$. Then, as explained, also Conditions 9 and 8 are satisfied. Hence, Theorem 10 and Corollary 11 are applicable to obtain structural results on optimal $\mathcal{D}$ mixing policies. Later we consider this example again to optimize over $\mathcal{D}$ mixing policies.

Theorem 16 can its consequences can easily be generalized to be applicable for more $\mathcal{D}$ restricted MDP problems. Indeed, from the proof, it is easily seen that the conditions on the transition matrixes given in Theorem 16 are a special case of the following more general result.

THEOREM 18: *Consider a $\mathcal{D}$ restricted MDP with $\mathcal{D} = \{d^1, d^2, \ldots, d^n\}$ and let $\mathcal{A} = \{P_1, P_2, \ldots, P_n\}$ be the set of n corresponding transition matrixes. Suppose there exists some $\gamma < 1$ and positive integer N such that for all n^N matrix products A of the form $A = \prod_{k=1}^{N} A_k$ with $A_k \in \mathcal{A}$ for $k = 1, 2, \ldots, N$, it holds that $\rho_0(A) \leq \gamma$, then for the associated (full observation) MDP, Condition* 12 *is satisfied for $M = 2BN/(1-\gamma)$. Moreover, Conditions* 9 *and* 8 *are also satisfied.*

PROOF: Similar to the proof of Theorem 16, it follows that (18) holds and, thus, Condition 12 is satisfied. Then, as explained earlier, it also follows that Conditions 9 and 8 are satisfied. ■

To conclude this section, the following example provides for $\mathcal{D} = \{d^1, d^2\}$, a case where Theorem 18 is applicable and Theorem 16 is not. Thus, also in the case of $\mathcal{D} = \{d^1, d^2\}$, Theorem 18 is a useful extension on Theorem 16.

Example 19: Consider a $\mathcal{D} = \{d^1, d^2\}$ restricted MDP with state space $S = \{1, 2, 3\}$. For decision rule d^1, the transition matrix P and reward vector $r(d^1)$ are as follows:

$$P = \begin{pmatrix} 0 & 0.5 & 0.5 \\ 1 & 0 & 0 \\ 0.5 & 0.5 & 0 \end{pmatrix}, \qquad r(d^1) = \begin{pmatrix} 2 \\ 0 \\ 3 \end{pmatrix}.$$

For the other decision rule d^2, the transition matrix P_2 and reward vector $r(d^2)$ are as follows:

$$Q = \begin{pmatrix} 1 & 0 & 0 \\ 0.5 & 0.5 & 0 \\ 0 & 0.5 & 0.5 \end{pmatrix}, \qquad r(d^2) = \begin{pmatrix} 0 \\ 2 \\ 0 \end{pmatrix}.$$

Then $\rho_0(P_1) = \rho_0(P_2) = 1$ and, thus, Theorem 16 is not applicable in this case. However, it is easy to check that $\rho_0(P_1^2) = 0.75$, $\rho_0(P_2^2) = 0.75$, $\rho_0(P_1P_2) = 0.75$, and

https://doi.org/10.1017/S0269964811000039 Published online by Cambridge University Press

$\rho_0(P_2P_1) = 0.75$. Thus, Theorem 18 is applicable with $N = 2$, $\gamma = 0.75$, and $B = 3$. Hence, Condition 12 is satisfied for $M = 48$ and Conditions 9 and 8 are satisfied. Thus, Theorem 10 and Corollary 11 are applicable in this case.

5. OPTIMIZING DETERMINISTIC MIXING POLICIES

To simplify notation and definitions, we continue to investigate the case that $\mathcal{D} = \{d^1, d^2\}$, which implies that deterministic mixing policies correspond to infinite sequences $U = (u_1, u_2, \ldots)$ of zeros and ones as explained earlier. However, for many aspects, a straightforward generalization to the case $\mathcal{D} = \{d_1, d_2, \ldots, d_n\}$ is possible. The main issue for deterministic mixing policies is that optimization of the performance seems very hard, because the set W of all infinite sequences of zeros and ones is infinite and discrete. Let $W^p \subseteq W$ be the subset of all periodic sequences of zeros and ones. If Condition 8 is satisfied for some deterministic mixing policy π corresponding to $U \in W^p$, then, as in the following example, the performance g^π is computable by (8).

Example 20: Consider again Example 1, whose characteristics were summarized by (1). Instead of the performance of Bernoulli policies, we now compute the performance of the deterministic mixing policy π with corresponding decision sequence $U = (1, 0, 1, 0, \ldots) = (1, 0)^\infty$, which obviously is periodic with period 2. For $t = 1, 2, \ldots$, let $X_t \in \{1, 2\}$ be the state at decision epoch t when policy π is applied. Then $\{X_t, t = 1, 2, \ldots\}$ is a Markov chain that is not stationary. However, $\{X_t, t = 1, 3, 5, \ldots\}$ is a stationary Markov chain with transition matrix

$$A_1 = P_1P_2 = \begin{pmatrix} 0.7 & 0.3 \\ 0.14 & 0.86 \end{pmatrix}.$$

It is easily verified that this Markov chain has unique stationary distribution $b_1^T = \left(\frac{7}{22}, \frac{15}{22}\right)$. Analogously, $\{X_t, t = 2, 4, 6, \ldots\}$ is a stationary Markov chain with transition matrix

$$A_2 = P_2P_1 = \begin{pmatrix} 0.76 & 0.24 \\ 0.20 & 0.80 \end{pmatrix}$$

and unique stationary distribution $b_2^T = \left(\frac{5}{11}, \frac{6}{11}\right)$. It follows that for $t = 1, 3, \ldots$, the long-run average reward is given by the inner product $b_1 \cdot r(d^1) = \frac{15}{22}$, and for $t = 2, 4, \ldots$, the long-run average reward is given by the inner product $b_2 \cdot r(d^2) = 0$. This implies that for the performance g^π, we have that $g^\pi = \frac{1}{2}\left(\frac{15}{22} + 0\right) = \frac{15}{44} \sim 0.341$.

Thus, for periodic policies we may compute the performance, but a problem is that the set W^p of all periodic sequences of zeros and ones remains infinite and discrete. However, optimizing over specific relatively small subsets of W^p is tractable by enumeration of all performances. If the optimal performance within such a subset

https://doi.org/10.1017/S0269964811000039 Published online by Cambridge University Press

is close (or even better equal) to the optimal performance within W^p (and possibly also W), then we may obtain (almost) optimal mixing policies in such a way.

One might choose a positive integer n and optimize over the finite subset of periodic sequences with period smaller or equal than n. We denote such subset by $W^p(n) \subseteq W^p$. Optimization over such a subset could in practice give good results. For example, consider the problem investigated in Examples 1 and 20. Optimization over the small set $W^p(2)$ results in the deterministic mixing policy corresponding to the periodic sequence with period cycle $(1, 0)$, which, according to Example 20, yields a performance of 0.341, which improves the performance of the optimal stationary Bernoulli mixing policy, which equals 0.303, as was shown in Example 1.

However, optimization over sets $W^p(n)$ has some disadvantages. First, the cardinality of $W^p(n)$ increases exponentially in n and, therefore, it is only tractable for rather small n. Additionally, if n gets smaller, then the optimal performance within the subset is likely to decrease. Thus, there is a trade-off between computation time and performance and a priori it is unknown what would be a good choice for the maximal period n. For the problem of Example 1, we have seen that, for period $n = 2$, already a policy exists that improves on the optimal Bernoulli policy, but for a larger state space, it is likely that a larger period n is necessary to improve on the optimal Bernoulli policy. In general, for fixed period n, we cannot say a priori whether the optimal performance over $Wp(n)$ is better than for the optimal Bernoulli mixing policy. Of course, the optimal performance over $Wp(n)$ is not better than the optimal performance over W, but nothing is known about the difference. This lack of guarantees for the optimal performance over $Wp(n)$ motivates one to investigate optimization over other subsets of W. In this article we consider, in particular, the subset of so-called regular sequences, which are introduced in the next subsection.

5.1. Regular Sequences and Corresponding Policies

Consider again the $\mathcal{D} = \{d^1, d^2\}$ restricted MDP from Example 1 characterized by (1). The best performance of a $\mathcal{D}$ mixing policy we have obtained so far for this example is 0.341, which is obtained by the deterministic mixing policy that corresponds to the periodic decision sequence with period cycle $(1, 0) \in Wp(2)$. This performance of 0.341 might be improved by optimizing over $Wp(n)$ for some larger value of n. Indeed, optimizing over $Wp(10)$ by applying (8) numerous times, it follows that the periodic decision sequence with period cycle $(1, 1, 0, 1, 0, 1, 0, 1, 0)$ yields an expected long-run average reward of $0.3435 > 0.341$ and that this performance is optimal over $Wp(10)$. The question is whether such improving sequences can only be found by such an exhaustive search over the set Wp or if it is possible to characterize some subset of W that certainly contain improving decision sequences, provided they exist. Such characterization is useful if searching over the subset is considerably faster than searching over W (Wp) and then it would be especially nice if the considered subset contains a decision sequence corresponding to a policy that is optimal over all $\mathcal{D}$ mixing policies. Indeed, we can characterize some subset of W having all of these desired properties if some conditions are satisfied. This is the subset $\mathcal{R} \subseteq W$

https://doi.org/10.1017/S0269964811000039 Published online by Cambridge University Press

of so-called regular sequences of zeros and ones. In the sequel, we define this subset and give some of the most useful properties and characterizations. We show how an effective optimization over this subset $\mathcal{R}$ can be performed and we will apply this to the $\mathcal{D} = \{d^1, d^2\}$ restricted MDP from Example 1. In Section 6 we give some conditions that are shown to be sufficient to that some optimal $\mathcal{D}$ mixing policy corresponds to a decision sequence that is a regular sequence. Thus, in that case, the optimal performance can indeed be found within this set of regular sequences.

Definition 21: *Let $U = (u_1, u_2, \ldots)$ be an infinite sequence of certain symbols. A suffix of U is an infinite sequence of the form $(u_n, u_{n+1}, \ldots)$ for some $n \in \mathbb{N}$. A finite subsequence of U is a finite sequence of the form $(u_k, u_{k+1}, \ldots, u_l)$ for some $k, l \in \mathbb{N}$ with $k \leq l$.*

In the sequel, $U = (u_1, u_2, \ldots)$ is assumed to be an infinite sequence of zeros and ones and recall Definition 2 defining when such a sequence is (eventually) regular. The subset $\mathcal{R} \subseteq W$ of regular sequences is defined as the set of infinite sequences of zeros and ones that are regular for some density $\theta \in [0, 1]$. It is obvious that if some sequence U is regular and thus element of $\mathcal{R}$ that (3) holds for some unique $\theta \in [0, 1]$. For infinite sequences of zeros and ones, very closely related to this notion of being (eventually) regular but possibly more convenient to apply is the notion of being (eventually) balanced. This notion is defined as follows.

Definition 22: *Let $U = (u_1, u_2, \ldots)$ be an infinite sequence of zeros and ones. Then U is called balanced if*

$$|s_k(n) - s_l(n)| \leq 1 \quad \text{for every } k, l, n \in \mathbb{N}. \qquad \textbf{(19)}$$

In other words, U is balanced if for any two finite subsequences of the same length, the number of ones contained in these subsequences differs by at most 1.

An infinite sequence of zeros and ones is called eventually balanced if it has a suffix that is balanced.

A complete classification of balanced sequences was given in [20]. Next we enumerate in Propositions 23 and 24 for regular (balanced) sequences the most important properties and connections that are useful for the present article. These results are obvious or might be retrieved from results in [20], [25], or [18]. The terminology in these references somewhat differs from each other and the present article, but Propositions 23 and 24 summarize the results on regular sequences, which will be applied in the remaining of this article.

Proposition 23: *For infinite sequences of zeros and ones, the following properties hold.*

1. *All regular sequences are balanced.*
2. *All balanced sequences are eventually regular.*

https://doi.org/10.1017/S0269964811000039 Published online by Cambridge University Press

3. *A sequence is eventually regular if and only if it is eventually balanced.*
4. *Let $U = (u_1, u_2, \dots)$ be an infinite sequence of zeros and ones and $V = (v_1, v_2, \dots)$ be defined by $v_n = 1 - u_n$ for all $n \in \mathbb{N}$. Then U is balanced if and only if V is balanced. Moreover, U is regular of density θ if and only if V is regular of density $1 - \theta$.*
5. *For every $\theta \in [0, 1]$, there exist some regular sequence(s) of density θ. Indeed, for given $\theta \in [0, 1]$, a regular sequence $U = (u_1, u_2, \dots)$ of density θ can be obtained as follows. Choose some arbitrary $\phi \in \mathbb{R}$ and let U be determined either by*

$$u_n = \lfloor n\theta + \phi \rfloor - \lfloor (n-1)\theta + \phi \rfloor \quad \text{for all } n \in \mathbb{N} \tag{20}$$

or by

$$u_n = \lceil n\theta + \phi \rceil - \lceil (n-1)\theta + \phi \rceil \quad \text{for all } n \in \mathbb{N}. \tag{21}$$

Then U is regular of density θ. Moreover, an infinite sequence of zeros and ones $(u_1, u_2, \dots)$ can be determined for some $\phi \in \mathbb{R}$ by either (20) or (21) if and only if the sequence is regular of density θ.
6. *A regular sequence of density θ is periodic if and only if θ is rational. If $\theta = p/q$ with $p, q \in \mathbb{N}$, p and q coprime, then the regular sequence has a period cycle of length q containing exactly p ones and $q - p$ zeros.*

The following result will be useful for an efficient maximization of the performance over the set $\mathcal{R}$ of regular sequences, as it implies that for regular sequences the performance is uniquely determined by the density θ of the sequence.

Proposition 24: *Let U and V be regular sequences and suppose they both have density θ. Then the set of all finite subsequences of U equals the set of all finite subsequences of V. Moreover, either V is a suffix of U or U is a suffix of V. If θ is rational, then the period cycles of U and V are cyclic shifts of each other.*

For example, $U = (1, 0, 1, 0, 0)^\infty$ and $V = (0, 1, 0, 1, 0)^\infty$ are regular sequences of the same density $\frac{2}{5}$ and indeed the period cycles $(1, 0, 1, 0, 0)$ and $(0, 1, 0, 1, 0)$ are cyclic shifts of each other.

5.2. Optimization over Regular Sequences

We have defined the subset $\mathcal{R}$ of regular sequences and described some important properties of regular (balanced) sequences. In this subsection our objective is to apply this and optimize the performance over $\mathcal{R}$ in an efficient manner. Regular and/or balanced sequences have been applied in open-loop control of particular queuing systems. In [11] it was proved for some specific admission control problem that the optimal control sequence is a regular sequence. Subsequently, regular sequences have been applied (see, e.g., [2–4,10,16]) to several admission, routing, and polling problems. In such applications to queuing and discrete-event systems, the optimality of regular

https://doi.org/10.1017/S0269964811000039 Published online by Cambridge University Press

sequences for open-loop control follows from multimodularity of an appropriate performance criterion such as expected workload in a queue or expected waiting times. Multimodularity is a property of functions defined on a discrete set that is comparable to convexity for functions defined on a continuous set. The concept of multimodularity and its applications are discussed in detail in [1] and an overview of control problems is given for which optimality of regular sequences can be established by multimodularity. In [1] several assumptions such as specifications on the topology of the queuing system are used to obtain multimodularity.

In the present article the objective is to apply regular sequences for general $\mathcal{D}$ restricted MDP optimizing the long-run average reward instead of some specific open-loop queuing control problem with a specific performance criterion, as in the above-mentioned references. A consequence of this generality is that specific properties of performance functions yielding, for example, multimodularity cannot be used. Therefore, any results on the optimality of some policy corresponding to a regular decision sequence have to be obtained in another way. We are able to do this since in Section 6, where we show that if for the associated full observation MDP, an optimal stationary and deterministic policy exists satisfying some specific properties, the existence of an optimal $\mathcal{D}$ mixing policy corresponding to some regular sequence follows. Thus, this is a new approach to establish the optimality of regular sequences for some (restricted) MDP problems without being dependent on multimodularity of the performance function. Next, we discuss and illustrate with an example the issue of optimizing the performance over $\mathcal{R}$—the set of all regular sequences of zeros and ones.

To optimize over $\mathcal{R}$ for some $\mathcal{D} = \{d^1, d^2\}$ restricted MDP problem, we assume that Conditions 8 and 9 are satisfied such that Theorem 10 and Corollary 11 are applicable. In Section 4 we have given some sufficient conditions for this that are easy to check. Then, by Theorem 10 and Proposition 24, it follows that all deterministic $\mathcal{D}$ mixing policies corresponding to regular sequences of the same density $\theta \in [0, 1]$ have the same performance. Thus, we may denote by $h(\theta)$ the long-run average reward of a deterministic $\mathcal{D}$ mixing policy corresponding to a regular decision sequence of density θ. Then we have that maximizing the performance over $\mathcal{R}$ is nothing more than maximizing the function $h(\theta)$ over $\theta \in [0, 1]$. Recall from Section 3.1 that this problem is rather similar to finding the optimal Bernoulli policy for which a performance function $g(\theta)$ should be maximized over $\theta \in [0, 1]$. We also note that, previously, in the admission, routing, and polling problems in which regular sequences have been applied, the optimization in most cases was reduced to a maximization or minimization over the density θ of the regular sequence. In these cases, $h(\theta) \geq g(\theta)$ for all θ in case of maximization or $h(\theta) \leq g(\theta)$ in case of minimization. Hence, the optimal value of $h(\theta)$ improves the optimal performance over all Bernoulli policies. We expect and would like to show that this property also holds for $\mathcal{D}$ restricted MDP problems such as the one we consider in the present article.

Recall that, for Bernoulli policies, it is not difficult to maximize $g(\theta)$ because for any $\theta \in [0, 1]$, the value $g(\theta)$ can be computed quickly and possibly a closed formula for $g(\theta)$ can be obtained as in Example 1. However, maximizing $h(\theta)$ is more

https://doi.org/10.1017/S0269964811000039 Published online by Cambridge University Press

difficult. First, it seems, in general, impossible to obtain a closed formula for $h(\theta)$, and for irrational θ, we do not even have a finite algorithm to compute $h(\theta)$. On the positive side, if θ is rational, the value $h(\theta)$ is computable by the following finite algorithm.

Algorithm 25: If Condition 8 is satisfied, this algorithm computes the performance $h(\theta)$ of any $\mathcal{D} = \{d^1, d^2\}$ mixing policy corresponding to a regular decision sequence of rational density $\theta \in [0, 1]$.

1. Determine coprime integers p and q with $p \geq 0$ and $q > 0$ such that $\theta = p/q$.
2. Choose some default value for ϕ, say $\phi = 0$, and then for $n = 1, 2, \ldots, q$, compute u_n by (20). The obtained sequence $(u_1, u_2, \ldots, u_q)$ is a period cycle of a regular sequence of density $\theta = p/q$.
3. Apply (8) to compute the long-run average reward g^π of the periodic policy π with period cycle $(u_1, u_2, \ldots, u_q)$. The value $h(\theta)$ is obtained by putting $h(\theta) = g^\pi$.

The running time of Algorithm 25 increases in the denominator q of θ because the period cycle of the regular sequence of density θ is of length q. Thus, for given $\theta = p/q$, the computation time is of order $\Omega(q)$, and to obtain or approximate the maximal value of $h(\theta)$, it seems most efficient to apply Algorithm 25 to a set of densities θ with bounded denominator q. For example, Algorithm 25 can be applied to obtain a maximum of $h(\theta)$ over the set $\mathcal{R} \cap W_p(n)$ for some $n \in \mathbb{N}$. For such maximization, the algorithm has to be applied only $O(n^2)$ times, and for each run, the period cycle of the decision sequence is at most n. Therefore, the total computation time is polynomial in n and the algorithm terminates relatively quickly if neither n nor the state space is very large.

For the $\mathcal{D}$ restricted MDP from Example 1, the algorithm quickly maximizes the performance over $\mathcal{R} \cap W_p(n)$ for a maximal period of, for example, $n = 200$. Applying the algorithm, it follows that the regular sequence with period cycle $(1, 1, 0, 1, 0, 1, 0, 1, 0)$ and density $\theta = \frac{5}{9}$ maximizes the performance over this set. Recall from the previous subsection that this particular decision sequence yields an expected long-run average reward of 0.3435. Applying Algorithm 25 for larger values of n does not give another improvement. Results in the sequel of this article support the optimality of this regular sequence of density $\frac{5}{9}$ over all feasible $\mathcal{D}$ mixing policies for the $\mathcal{D}$ restricted MDP from Example 1.

We note that $\frac{5}{9}$ is close but not equal to $\theta^* = 3 - \sqrt{6} \sim 0.551$, which maximizes (recall Example 1) the performance over Bernoulli policies over rate θ. Figure 1, in which, for $\theta \in [0, 1]$, the performance of Bernoulli policies and deterministic $\mathcal{D}$ mixing policies given by a regular sequence of density θ are plotted, illustrates this. Recall from Example 1 that for Bernoulli policies of rate θ, the performance $g(\theta)$ is according to the function $g(\theta) = (3\theta - 3\theta^2)/(3 - \theta)$. For regular sequences of density θ, we do not have a closed formula, but the performance $h(\theta)$ (computed by Algorithm 25) is plotted for all $\theta = k/100$ for $k = 0, 1, \ldots, 100$. Thus, $g(\theta)$ is the

https://doi.org/10.1017/S0269964811000039 Published online by Cambridge University Press

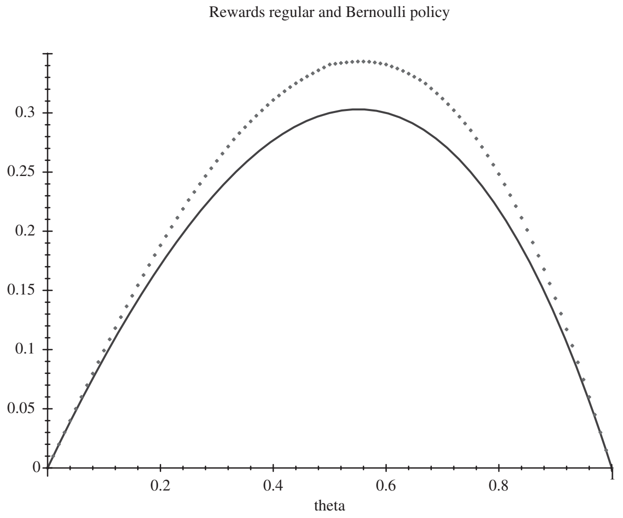

FIGURE 1. The performance of regular sequences versus Bernoulli policies.

solid smooth curve in Figure 1, and the isolated points $(\theta, h(\theta))$ for $\theta = k/100$ also seem to be situated on some smooth curve. This suggests that the performance $h(\theta)$ for regular sequences is continuous for $\theta \in [0, 1]$ just as $g(\theta)$, which is known to be continuous. It is also interesting to note that Figure 1 visually confirms that $h(\theta)$ is never smaller than $g(\theta)$ and that the difference in performance $h(\theta) - g(\theta)$ appears

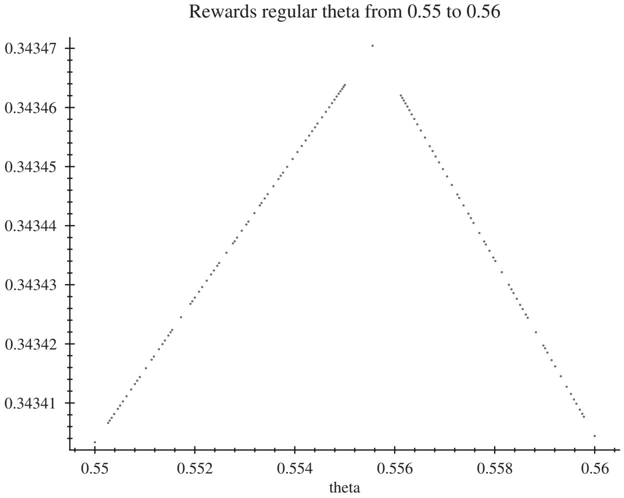

FIGURE 2. The performance of regular sequences for rational densities in the interval [0.55, 0.56] and denominator at most 200.

https://doi.org/10.1017/S0269964811000039 Published online by Cambridge University Press

to be maximal around the value where $h(\theta)$ is maximal. Moreover, Figure 1 confirms that the value of θ that maximizes $h(\theta)$ could well be $\theta = \frac{5}{9}$ and that the maximizing value for $h(\theta)$ is close to the value that maximizes $g(\theta)$. Figure 2 visually confirms that for $\theta = \frac{5}{9}$, the value of $h(\theta)$ is maximal. In Figure 2, the value of θ is varying over the small interval $[0.55, 0.56]$, and in this interval, all points $(\theta, h(\theta))$ are plotted for all rational $\theta = m/n$ with denominator $n \leq 200$. In Figure 2, the point $(\frac{5}{9}, h(\frac{5}{9}))$ is obviously the top one. Moreover, from the triangular shape that is recognizable in Figure 2, it may be concluded that for θ in this small interval around $\frac{5}{9}$, the value of $h(\theta)$ increases approximately linearly if θ approximates $\frac{5}{9}$.

6. SUFFICIENT CONDITIONS FOR OPTIMALITY OF A REGULAR SEQUENCE

In this section we show that certain conditions for $\mathcal{D} = \{d^1, d^2\}$ restricted MDP are sufficient for the existence of an optimal $\mathcal{D}$ mixing policy that is deterministic corresponding to a regular zero–one decision sequence. This is a main result. After that, we also discuss the applicability of the results to $\mathcal{D}$ restricted MDP problems and, in particular, the problem introduced in Example 1.

First, we formulate and prove a key result that states that some infinite sequence of zeros and ones generated by iterating some function on the interval [0, 1] is eventually regular if the functions satisfies certain conditions. In the sequel, we denote by I the interval [0, 1].

Iteration 26: Let $x_1, x^* \in I$ be given. Let $f_1, f_2 : I \to I$ be given functions and $f : I \to I$ be defined by

$$f(x) = \begin{cases} f_1(x) & \text{if } x \leq x^* \\ f_2(x) & \text{if } x > x^*. \end{cases} \tag{22}$$

Consecutively, for $n = 1, 2, \ldots$, determine u_n and x_{n+1} iteratively by

$$u_n := \begin{cases} 0 & \text{if } x_n \leq x^* \\ 1 & \text{if } x_n > x^* \end{cases} \qquad \text{and} \qquad x_{n+1} := f(x_n). \tag{23}$$

THEOREM 27: *Let $U = (u_1, u_2, \ldots)$ be an infinite sequence of zeros and ones generated by Iteration* 26 *with $f_1, f_2 : I \to I$ both monotonically increasing and, moreover,*

$$f_1(f_2(x)) \geq f_2(f_1(x)) \quad \textit{for all } x \in I. \tag{24}$$

Then U is an eventually regular sequence.

To prove Theorem 27, we apply Lemma 28, which follows immediately from Proposition 2.1.3 in [18]. As in [18], for sequences (or so-called words) $a = (a_1, a_2, \ldots, a_n)$ and $b = (b_1, b_2, \ldots, b_m)$ the concatenation $(a_1, a_2, \ldots, a_n, b_1, b_2, \ldots, b_m)$ is denoted by ab.

https://doi.org/10.1017/S0269964811000039 Published online by Cambridge University Press

LEMMA 28: *Let $U = (u_1, u_2, \ldots)$ be an infinite sequence of zeros and ones. Then U is balanced if and only if there does not exist some (possibly empty) finite sequence w of zeros and ones such that both $0w0$ and $1w1$ are subsequences of U.*

Proof of Theorem 27: We distinguish a few cases. In the first case, suppose $f_1(x^*) \leq x^*$. Then $f_1(x) \leq x^*$ for all $0 \leq x \leq x^*$ because f_1 is monotonically increasing. Thus, if $x_n \leq x^*$ for some $N \in \mathbb{N}$, then $x_n \leq x^*$ for $n = N, N+1, \ldots$ and, thus, $(u_N, u_{N+1}, \ldots) = (0, 0, \ldots)$ is regular of density 0. Hence, U is an eventually regular sequence. If, on the other hand, $x_n > x^*$ for $n = 1, 2, \ldots$, then $U = (1, 1, \ldots)$ is regular of density 1.

In the second case, suppose $f_2(x^*) > x^*$. Then it follows analogously to the first case that there exists some $N \in \mathbb{N}$ such that $x_n > x^*$ for all $n \geq N$ or $x_n \leq x^*$ for $n = 0, 1, \ldots$. Hence, either U is eventually regular of density 1 or U is regular of density 0.

In the third and most important case, we suppose that $f_2(x^*) \leq x^* < f_1(x^*)$ and let J denote the interval $[f_2(x^*), f_1(x^*)]$. Note that if $x_n < f_2(x^*)$, then $x_{n+1} = f_1(x_n) \leq f_1(x^*)$, and if $x_n > f_1(x^*)$, then $x_{n+1} = f_2(x_n) \geq f_2(x^*)$. Thus, either $x_n < f_2(x^*) \leq x^*$ for $n = 1, 2, \ldots$, or $x_n > f_1(x^*) > x^*$ for $n = 1, 2, \ldots$, or $x_N \in J$ for some $N \in \mathbb{N}$. Thus, either U is regular of density 0 or U is regular of density 1 or $x_N \in J$ for some $N \in \mathbb{N}$. Suppose $x_N \in J$ for some $N \in \mathbb{N}$. If $f_2(x^*) \leq x_N \leq x^*$, then $x_{N+1} = f_1(x_N) \leq f_1(x^*)$, and by (24) we also have that

$$x_{N+1} = f_1(x_N) \geq f_1(f_2(x^*)) \geq f_2(f_1(x^*)) \geq f_2(x^*)$$

and, thus, $x_{N+1} \in J$. Similarly, if $x^* < x_N \leq f_1(x^*)$, then $x_{N+1} = f_2(x_N) \geq f_2(x^*)$, and by (24) we also have that

$$x_{N+1} = f_2(x_N) \leq f_2(f_1(x^*)) \leq f_1(f_2(x^*)) \leq f_1(x^*)$$

and, thus, $x_{N+1} \in J$. Hence, if $x_N \in J$, then it follows by induction that $x_n \in J$ for $n = N, N+1, \ldots$. Thus, in this third case we may assume that there exists some $N \in \mathbb{N}$ such that $x_n \in J$ for all $n \geq N$.

Consider the suffix $U' := (u_N, u_{N+1}, \ldots)$ of U. Suppose $u_m = 0, u_n = 1$ for some $m, n \in N$. Assume there exists some $k \in \mathbb{N}$ for which $u_{m+k} \neq u_{n+k}$ and let k_0 be the minimal positive integer satisfying $u_{m+k_0} \neq u_{n+k_0}$. We claim that it then follows that

$$u_{m+k_0} = 1 \quad \text{and} \quad u_{n+k_0} = 0.$$

To verify this claim, note that by (24) and $x_n, x_m \in J$, we have that

$$x_{m+1} = f_1(x_m) \geq f_1(f_2(x^*)) \geq f_2(f_1(x^*)) \geq f_2(x_n) = x_{n+1}.$$

Thus, if $k_0 = 1$, then $x_{m+k_0} \geq x_{n+k_0}$, implying $1 \geq u_{m+k_0} \geq u_{n+k_0} \geq 0$. Hence, $u_{m+k_0} = 1$ and $u_{n+k_0} = 0$ follows from the fact that $u_{m+k_0} \neq u_{n+k_0}$. If $k_0 \geq 2$, then we have that $u_{m+1} = u_{n+1}$. Because both f_1, f_2 are monotonically increasing and thus order-preserving, it follows that $x_{m+2} \geq x_{n+2}$ by either $x_{m+2} = f_1(x_{m+1}) \geq f_1(x_{n+1}) = x_{n+2}$

or $x_{m+2} = f_2(x_{m+1}) \geq f_2(x_{n+1}) = x_{n+2}$. By applying this order-preserving property of both f_1 and f_2 repetitively, it follows again that $x_{m+k_0} \geq x_{n+k_0}$ and, thus, $u_{m+k_0} = 1$ and $u_{n+k_0} = 0$, as above. Thus, the claim holds, but then it follows that there does not exist some (possibly empty) finite sequence w of zeros and ones such that both $0w0$ and $1w1$ are subsequences of U'. Thus, by Lemma 28, it follows that U' is balanced. Thus, by definition, U is eventually balanced because U' is a suffix of U, and by Proposition 23, it follows that U is an eventually regular sequence. ■

Our next aim is to apply Theorem 27 to $\mathcal{D} = \{d^1, d^2\}$ restricted MDP problems with finite state space S satisfying some specific properties. For this, we consider again the associated full observation MDP with continuous state space X of probability distributions on S as we introduced in Section 4. First, we restrict to problems for which Conditions 8 and 9 are satisfied. Recall that in Section 4.1, we have investigated when these two conditions are satisfied and we have seen that they are satisfied for a considerable class of problems. Now, we define an extra condition that should hold in particular for the applicability of Theorem 27. In the sequel, this new condition will be called the threshold condition, as basically it says that for the associated full observation MDP, some optimal stationary deterministic Markov policy (which exists according to Condition 9) has some "threshold structure." In Definition 29 we define this notion of "threshold structure" for such policies, which is followed by Condition 30 stating our threshold condition for $\mathcal{D} = \{d^1, d^2\}$ restricted MDP.

Definition 29: *Let $h : X \to \mathcal{A} = \{d^1, d^2\}$ be the mapping corresponding to a stationary deterministic Markov policy $\tilde{\pi}$. Then we say that mapping h and policy $\tilde{\pi}$ have threshold structure if there exists some $i \in S$ and $x^0 \in I$ such that for all $x = (x_1, x_2, \ldots, x_{|S|}) \in X$, we either have that $h(x) = d^1$ if and only if $x_i \leq x^0$ ($x_i < x^0$) or $h(x) = d^2$ if and only if $x_i \leq x^0$ ($x_i < x^0$).*

Condition 30: *For the associated full observation MDP that is equivalent to the considered $\mathcal{D} = \{d^1, d^2\}$ restricted MDP, there exist some optimal stationary deterministic Markov policy $\tilde{\pi}$ having a threshold structure as defined in Definition* 29.

Proposition 31 connects Condition 30 with Iteration 26 in the case of a two-state state space as, for example, in the $\mathcal{D} = \{d^1, d^2\}$ restricted MDP of Example 1. Then Theorem 27 will be applicable if the appropriate functions f_1 and f_2 have the properties stated in Theorem 27. Additionally, from Proposition 31, it follows for such two-state cases, the appropriate f_1 and f_2 are linear, which in the sequel will be useful to check the properties to apply Theorem 27.

Proposition 31: *Consider a $\mathcal{D} = \{d^1, d^2\}$ restricted MDP with state space $S = \{1, 2\}$. Suppose that Condition* 30 *is satisfied and let $\tilde{\pi}$ be a stationary deterministic Markov policy for the associated full observation MDP having a threshold structure. Let $\tilde{\omega} = (y_1, a_1, y_2, a_2, \ldots) \in \tilde{\Omega}$ be an associated sample path. For $n = 1, 2, \ldots$, let $v_n, w_n \in I$ satisfying $v_n + w_n = 1$ be such that $y_n = (v_n, w_n)$. Then there exist $x_1, x^* \in I$ and linear*

https://doi.org/10.1017/S0269964811000039 Published online by Cambridge University Press

functions $f_1, f_2 \colon I \to I$ *such that the sequences* $(u_1, u_2, \ldots)$ *and* $(x_1, x_2, \ldots)$ *generated by Iteration* 26 *satisfy the following properties.*

1. *Either* $x_n = v_n$ *for* $n = 1, 2, \ldots$ *or* $x_n = w_n$ *for* $n = 1, 2, \ldots$.
2. *Either* (25) *or* (26) *holds:*

$$u_n = \begin{cases} 0 & \textit{if } a_n = d^1 \\ 1 & \textit{if } a_n = d^2 \end{cases} \qquad \textit{for } n = 1, 2, \ldots, \tag{25}$$

$$u_n = \begin{cases} 0 & \textit{if } a_n = d^2 \\ 1 & \textit{if } a_n = d^1 \end{cases} \qquad \textit{for } n = 1, 2, \ldots. \tag{26}$$

Proof: Let P_1 be the transition matrix corresponding to d^1 and P_2 be the transition matrix corresponding to d^2. Let $a, b, c, d \in I$ be such that

$$P = \begin{pmatrix} a & 1-a \\ 1-b & b \end{pmatrix}, \qquad P_2 = \begin{pmatrix} c & 1-c \\ 1-d & d \end{pmatrix}. \tag{27}$$

Let $h \colon X \to \{d^1, d^2\}$ be the mapping corresponding to $\widetilde{\pi}$. Then h has a threshold structure (see Definition 29) and assume h has this property for state $i = 1$. Now, we distinguish several cases.

In the first case, suppose that there exists some $x^0 \in I$ such that for any $\hat{x} = (\hat{x}_1, \hat{x}_2) \in X$, it holds that $h(\hat{x}) = d^1$ if and only if $\hat{x}_1 \leq x^0$. Then we claim that by putting $x_1 = v_1, x^* = x^0, f_1(x) = (a + b - 1)x + 1 - b$ for all $x \in I$, and $f_2(x) = (c + d - 1)x + 1 - d$ for all $x \in I$, the sequences $(u_1, u_2, \ldots)$ and $(x_1, x_2, \ldots)$ generated by Iteration 26 satisfy $x_n = v_n$ for $n = 1, 2, \ldots$ and, moreover, $u_n = 0$ if and only if $a_n = d^1$. We prove this claim by induction to n. For $n = 1$ we already have $x_1 = v_1$. If $v_1 \leq x^0$, then $a_1 = h(y_1) = d^1$, $x_1 \leq x^*$ and, thus, $u_1 = 0$. On the other hand, if $v_1 > x^0$, then $a_1 = h(y_1) = d^2$, $x_1 > x^*$ and, thus, $u_1 = 1$. Thus, the claim holds for $n = 1$. Suppose the claim holds for $n = k$ and, thus, $x_k = v_k$. Distinguish the cases $v_k \leq x^0$ and $v_k > x^0$. Suppose $v_k \leq x^0$. Then $a_k = h(y_k) = d^1$ and, thus, $u_k = 0$ by the induction claim. Then by Iteration 26, it follows that $x_{k+1} = f_1(x_k) = f_1(v_k) = (a + b - 1)v_k + 1 - b$. Additionally we have

$$\begin{aligned} y_{k+1} &= (v_k, w_k)P_1 = (v_k, w_k)\begin{pmatrix} a & 1-a \\ 1-b & b \end{pmatrix} \\ &= (av_k + (1-b)w_k, (1-a)v_k + bw_k). \end{aligned}$$

Hence, $v_{k+1} = av_k + (1-b)w_k = av_k + (1-b)(1-v_k) = (a + b - 1)v_k + 1 - b$ and, thus, $x_{k+1} = v_{k+1}$ if $v_k \leq x^0$. Suppose $v_k > x^0$. Then $a_k = h(y_k) = d^2$ and, thus, $u_k = 1$ by the induction claim. Then, by Iteration 26, it follows that $x_{k+1} = f_2(x_k) =$

$f_2(v_k) = (c+d-1)v_k + 1 - d$. Additionally, we have

$$y_{k+1} = (v_k, w_k)P_2 = (v_k, w_k)\begin{pmatrix} c & 1-c \\ 1-d & d \end{pmatrix}$$
$$= (cv_k + (1-d)w_k, (1-c)v_k + dw_k).$$

Hence, $v_{k+1} = cv_k + (1-d)w_k = cv_k + (1-d)(1-v_k) = (c+d-1)v_k + 1 - d$ and, thus, $x_{k+1} = v_{k+1}$ if $v_k > x^0$. Thus, we have proved that $x_{k+1} = v_{k+1}$. Then, for $n = k+1$, it follows that $u_n = 0$ if and only if $a_n = d^1$ similarly as for $n = 1$ and the induction proof is finished.

In the second case, suppose that there exists some $x^0 \in I$ such that for any $\hat{x} = (\hat{x}_1, \hat{x}_2) \in X$, it holds that $h(\hat{x}) = d^1$ if and only if $\hat{x}_1 < x^0$. Then we claim that by putting $x_1 = w_1, x^* = 1 - x^0, f_1(x) = (c+d-1)x + 1 - c$ for all $x \in I$, and $f_2(x) = (a+b-1)x + 1 - a$ for all $x \in I$, the sequences $(u_1, u_2, \ldots)$ and $(x_1, x_2, \ldots)$ generated by Iteration 26 satisfy $x_n = w_n$ for $n = 1, 2, \ldots$ and, moreover, $u_n = 0$ if and only if $a_n = d^2$. This claim follows also by induction analogously to the above first case.

In the third case, suppose that there exists some $x^0 \in I$ such that for any $\hat{x} = (\hat{x}_1, \hat{x}_2) \in X$, it holds that $h(\hat{x}) = d^2$ if and only if $\hat{x}_1 \leq x^0$. Then we claim that by putting $x_1 = v_1, x^* = x^0, f_1(x) = (c+d-1)x + 1 - d$ for all $x \in I$, and $f_2(x) = (a+b-1)x + 1 - b$ for all $x \in I$, the sequences $(u_1, u_2, \ldots)$ and $(x_1, x_2, \ldots)$ generated by Iteration 26 satisfy $x_n = v_n$ for $n = 1, 2, \ldots$ and, moreover, $u_n = 0$ if and only if $a_n = d^2$. This claim follows also by induction analogously to the above first case.

In the fourth and last case, suppose that there exists some $x^0 \in I$ such that for any $\hat{x} = (\hat{x}_1, \hat{x}_2) \in X$, it holds that $h(\hat{x}) = d^2$ if and only if $\hat{x}_1 < x^0$. Then we claim that by putting $x_1 = w_1, x^* = 1 - x^0, f_1(x) = (a+b-1)x + 1 - a$ for all $x \in I$, and $f_2(x) = (c+d-1)x + 1 - c$ for all $x \in I$, the sequences $(u_1, u_2, \ldots)$ and $(x_1, x_2, \ldots)$ generated by Iteration 26 satisfy $x_n = w_n$ for $n = 1, 2, \ldots$ and, moreover, $u_n = 0$ if and only if $a_n = d^1$. This claim follows also by induction analogously to the above first case.

This finishes the proof for the case that h has a threshold structure for state $i = 1$. For the case that h has a threshold structure for state $i = 2$, it could be proved similar as for $i = 1$ by distinguishing four different cases and obtaining the appropriate x_1, x^*, f_1, and f_2 for all of these cases. However, it follows more elegantly by noting that if $S = \{1, 2\}$, any threshold structure for state $i = 2$ is equivalent to a threshold structure for state $i = 1$ and vice versa. For example, suppose there exists some $x^0 \in I$ such that for any $\hat{x} = (\hat{x}_1, \hat{x}_2) \in X$, it holds that $h(\hat{x}) = d^1$ if and only if $\hat{x}_2 \leq x^0$. This is a threshold structure according to Definition 29 for state $i = 2$. Obviously, it is equivalent to $h(\hat{x}) = d^2$ if and only if $\hat{x}_1 < 1 - x^0$, which gives a threshold structure according to Definition 29 for state $i = 1$. ■

Theorem 32 is a main result in this article that is based on combining Theorem 18, Theorem 10, Corollary 11, Proposition 23, Proposition 31, and Theorem 27.

https://doi.org/10.1017/S0269964811000039 Published online by Cambridge University Press

THEOREM 32: *Consider a $\mathcal{D} = \{d^1, d^2\}$ restricted MDP with state space $S = \{1, 2\}$. Let $a, b, c, d \in I$ be such that (27) holds, where P_1 is the transition matrix corresponding to d^1 and P_2 is the transition matrix corresponding to d^2. Suppose that $1 \le a + b < 2$, $1 \le c + d < 2$, and Condition* 30 *is satisfied. Then for the asociated full observation MDP, there exists some optimal stationary deterministic Markov policy $\widetilde{\pi}$ having a threshold structure as in Definition* 29.

Moreover, let $\widetilde{\omega} = (y_1, a_1, y_2, a_2, \ldots) \in \widetilde{\Omega}$ be an associated sample path for $\widetilde{\pi}$ and let $f_1, f_2 : I \to I$ be linear functions as obtained in the proof of Proposition 31. *If there exists some $x \in I$ for which $f_1(f_2(x)) \ge f_2(f_1(x))$, then for the $\mathcal{D} = \{d^1, d^2\}$ restricted MDP, there exist an optimal $\mathcal{D}$ mixing policy that is deterministic and the corresponding decision sequence of zeros and ones is a regular sequence. In particular, there exist some $n \in \mathbb{N}$ such that for all positive integers $t \ge n$, the infinite sequence of decision rules $(a_t, a_{t+1}, \ldots)$ determines an optimal $\mathcal{D}$ mixing policy for which the corresponding sequence of zeros and ones is a regular sequence.*

PROOF: We have $P_1 = \begin{pmatrix} a & 1-a \\ 1-b & b \end{pmatrix}$ and, thus, by (13) it is easily seen that $\rho_0(P_1) = |a + b - 1|$. Since $1 \le a + b < 2$, it follows that $\rho_0(P_1) < 1$ and, similarly, we also have that $\rho_0(P_2) < 1$. Thus, Theorem 18 is applicable for $N = 1$ and, thus, it follows for the associated full observation MDP that Conditions 12, 9, and 8 are satisfied. Thus, there exist optimal stationary deterministic Markov policies for the associated MDP, and because Condition 30 is also satisfied, it follows that there exists some optimal stationary deterministic Markov policy $\widetilde{\pi}$ having a threshold structure as in Definition 29.

Let $\widetilde{\omega} = (y_1, a_1, y_2, a_2, \ldots) \in \widetilde{\Omega}$ be an associated sample path for $\widetilde{\pi}$ as in Proposition 31. For the $\mathcal{D} = \{d^1, d^2\}$ restricted MDP, we have by Theorem 10 and Corollary 11 that all deterministic $\mathcal{D}$ mixing policies π_t, $t = 1, 2, \ldots,$ given by the infinite sequence of decision rules $(a_t, a_{t+1}, \ldots)$ have the same performance. Moreover, this performance is optimal with respect to maximizing the long-run average reward for the $\mathcal{D}$ restricted MDP, as $(y_1, a_1, y_2, a_2, \ldots)$ is a sample path of policy $\widetilde{\pi}$ that is optimal for the associated full observation MDP. Thus, for all $t = 1, 2, \ldots,$ policy π_t is an optimal $\mathcal{D}$ mixing policy.

Let $U := (u_1, u_2, \ldots)$ and $(x_1, x_2, \ldots)$ be the infinite sequences generated by Iteration 26 for linear functions f_1 *and* f_2 and appropriate $x_1, x^* \in I$ as in Proposition 31. Then U is an infinite sequence of zeros and ones, and by Proposition 31 we have that either (25) or (26) holds. Moreover, according to the proof of Proposition 31, we can assume that either the slope of f_1 is $a + b - 1$ and the slope of f_2 is $c + d - 1$ or the slope of f_1 is $c + d - 1$ and the slope of f_2 is $a + b - 1$. Either way, it follows that f_1 and f_2 are monotonically increasing functions because $a + b \ge 1$ and $c + d \ge 1$. Moreover, it follows that the composite functions $f_1 \circ f_2$ and $f_2 \circ f_1$ are both linear functions with slope $(a + b - 1)(c + d - 1)$ mapping I to I. Hence, $f_1(f_2(x)) \ge f_2(f_1(x))$ for some $x \in I$ implies that $f_1(f_2(x)) \ge f_2(f_1(x))$ for all $x \in I$. Thus, if $f_1(f_2(x)) \ge f_2(f_1(x))$ for some $x \in I$, then the properties demanded in Theorem 27 for the functions f_1 and f_2

https://doi.org/10.1017/S0269964811000039 Published online by Cambridge University Press

are satisfied and, thus, the sequence U generated by Iteration 26 is an eventually regular sequence. Thus, there exists some $n \in \mathbb{N}$ such that for every positive integer $t \geq n$, the infinite sequence $U_t := (u_t, u_{t+1}, \ldots)$ is a regular sequence of zeros and ones.

Recall from Section 3.2 that, by convention, symbol 1 corresponds to action d^1 and symbol 0 corresponds to action d^2. Following this convention, let $U' := (u'_1, u'_2, \ldots)$ be the infinite sequence of zeros and ones corresponding to $(a_1, a_2, \ldots)$. Then, for $t = 1, 2, \ldots$, we have that $U'_t = (u'_t, u'_{t+1}, \ldots)$ is the infinite sequence of zeros and ones corresponding to optimal $\mathcal{D}$ mixing policy $\pi_t = (a_t, a_{t+1}, \ldots)$. In the case that (25) holds, then it follows that $u'_n = 1 - u_n$ for $n = 1, 2, \ldots$. Thus, by property 4 of Proposition 23, it follows for $t = 1, 2, \ldots$ that U'_t is regular of density $1 - \theta$ if and only if U_t is regular of density θ. Thus, for every positive integer $t \geq n$, the sequence U'_t corresponding to optimal $\mathcal{D}$ mixing policy π_t is regular, as U_t is regular for $t = n, n+1, \ldots$. In the case that (26) holds, then it follows that $u'_n = u_n$ for $n = 1, 2, \ldots$. Thus, it follows for $t = 1, 2, \ldots$ that sequence U'_t is exactly the same as sequence U_t. Thus, for every positive integer $t \geq n$, the sequence U'_t corresponding to optimal $\mathcal{D}$ mixing policy π_t is regular of some density θ because U_t is regular of some density θ for $t = n, n+1, \ldots$. ■

In Example 33 we apply Theorem 32 to the $\mathcal{D} = \{d^1, d^2\}$ restricted MDP of Example 1.

Example 33: Consider again the $\mathcal{D} = \{d^1, d^2\}$ restricted MDP of Example 1. Let $a, b, c, d \in I$ be defined as in Theorem 32. For this example, we have that $a = 1$, $b = 0.8$, $c = 0.7$, and $d = 1$. Thus, the conditions $1 \leq a + b < 2$ and $1 \leq c + d < 2$ are satisfied. Moreover, recall from Example 17 that Conditions 12, 9, and 8 are satisfied. Thus, for the associated MDP, there exist some optimal stationary deterministic Markov policy $\widetilde{\pi}$. We will not prove that also Condition 30 is satisfied, but we note that it seems plausible. Indeed, let p be the probability that the machine is in state 1 (the bad state) at a decision epoch. Indeed, it seems plausible that there exists some threshold probability p^* such that if p is smaller than p^* then it is optimal to choose action 1 (work), whereas if p is larger than p^*, then it is optimal to choose action 2 (repair). Thus, assume Condition 30 is satisfied and that policy $\widetilde{\pi}$ has indeed a threshold structure. Let $\widetilde{\omega} = (y_1, a_1, y_2, a_2, \ldots) \in \widetilde{\Omega}$ be an associated sample path for $\widetilde{\pi}$, and for $n = 1, 2, \ldots$, let $v_n, w_n \in I$ be such that $y_n = (v_n, w_n)$ for $n = 1, 2, \ldots$ as in Theorem 32.

Now, we distinguish two cases of plausible threshold structures that optimal policy $\widetilde{\pi}$ could have in this example. In the first case, suppose there exists some $p^* \in I$ such that policy $\widetilde{\pi}$ chooses decision rule d^1 and, thus, action 1 (work) if and only if $p \leq p^*$ and it chooses decision rule d^2 and, thus, action 2 (repair) if and only if $p > p^*$. Then following the proof of Proposition 31, we put $x_1 = v_1, x^* = p^*, f_1(x) = (a + b - 1)x + 1 - b = 0.8x + 0.2$, and $f_2(x) = (c + d - 1)x + 1 - d = 0.7x$. Then $f_1(f_2(x)) = 0.56x + 0.2$ and $f_2(f_1(x)) = 0.56x + 0.14$. Thus for this threshold structure (24) is indeed satisfied.

In the second case, suppose there exists some $p^* \in I$ such that policy $\widetilde{\pi}$ chooses, at a decision epoch, decision rule d^1 and, thus, action 1 (work) if and only if $p < p^*$ and it chooses decision rule d^2 and, thus, action 2 (repair) if and only if $p \geq p^*$.

https://doi.org/10.1017/S0269964811000039 Published online by Cambridge University Press

Then following the proof of Proposition 31, we put $x_1 = w_1$, $x^* = 1 - p^*$, $f_1(x) = (c + d - 1)x + 1 - c = 0.7x + 0.3$, and $f_2(x) = (a + b - 1)x + 1 - a = 0.8x$. Then $f_1(f_2(x)) = 0.56x + 0.3$, and $f_2(f_1(x)) = 0.56x + 0.24$. Thus, also for this threshold structure, (24) is indeed satisfied.

Thus, we may conclude that if Condition 30 is satisfied for this example, then (24) holds for all plausible threshold structures. Then it follows by Theorem 32 that there exists some $n \in \mathbb{N}$ such that for all positive integers $t \geq n$, the infinite sequence of decision rules $(a_t, a_{t+1}, \ldots)$ determines an optimal $\mathcal{D}$ mixing policy for which the corresponding sequence of zeros and ones is a regular sequence. In other words, under the assumption that Condition 30 holds, it follows for this example that an optimal $\mathcal{D}$ mixing policy is among the deterministic Markov $\mathcal{D}$ mixing policies for which the corresponding decision sequence is in the set $\mathcal{R}$ of regular sequences of zeros and ones and the maximal performance is obtained by maximizing performances over $\mathcal{R}$.

Recall from Section 5.2 that for the $\mathcal{D} = \{d^1, d^2\}$ restricted MDP of Example 1, the maximal performance over $\mathcal{R} \cap W_p(200)$ equals 0.3435 (rounded to four decimals) and is obtained by the regular sequence with period cycle $(1, 1, 0, 1, 0, 1, 0, 1, 0)$ of density $\theta = \frac{5}{9}$. Moreover, Figures 1 and 2 did give additional visual support for density $\frac{5}{9}$ maximizing the performance over $\mathcal{R}$. If $\theta = \frac{5}{9}$ indeed maximizes $h(\theta)$ and Condition 30 holds, then it follows from Theorem 32 that for this $\mathcal{D} = \{d^1, d^2\}$ restricted MDP, the $\mathcal{D} = \{d^1, d^2\}$ mixing policy corresponding to period cycle $(1, 1, 0, 1, 0, 1, 0, 1, 0)$ (with symbol 1 corresponding to choosing d^1 and symbol 0 corresponding to choosing d^2) is optimal and the maximal long-run average reward is 0.3435, which is obtained by this policy.

Vice versa, this would imply that there should exist some p^* and corresponding threshold property as described above such that for the associated full observation MDP and the corresponding policy $\widetilde{\pi}$ induces, for any initial state distribution $x \in X$, an infinite sequence of decision rules that is eventually periodic according to the period cycle $(1, 1, 0, 1, 0, 1, 0, 1, 0)$. Indeed, this is the case for $p^* = 0.47$ (in fact for some interval containing 0.47). The reader can check that by putting $x^* = 1 - p^* = 0.53$, $f_1(x) = 0.7x + 0.3$, and $f_2(x) = 0.8x$ as in the second distinguished case above; then, for any $x_1 \in [0, 1]$, the infinite sequence $(u_1, u_2, \ldots)$ of zeros and ones obtained according to Iteration 26 eventually becomes periodic with period cycle $(1, 1, 0, 1, 0, 1, 0, 1, 0)$. Moreover, for initial state distribution $y_1 = (1 - x_1, x_1) \in X$, the sample path $\widetilde{\omega} = (y_1, a_1, y_2, a_2, \ldots) \in \widetilde{\Omega}$ obtained by applying the threshold property $a_n = d^1$ if and only if $y_n \cdot (1, 0) < 0.47$ satisfies (26). Thus for this $\mathcal{D}$ restricted MDP, we have established some additional confirmation for the optimality of the $\mathcal{D} = \{d^1, d^2\}$ mixing policy corresponding to a regular decision sequence with period cycle of density $\frac{5}{9}$ yielding a performance of 0.3435.

Resuming, our results all support, but do not completely prove, the following conjecture.

Conjecture 34: For the problem introduced in Example 1 the periodic Markov policy with period cycle $(d^1, d^1, d^2, d^1, d^2, d^1, d^2, d^1, d^2)$ is optimal within the class of $\mathcal{D} = \{d^1, d^2\}$ mixing policies.

https://doi.org/10.1017/S0269964811000039 Published online by Cambridge University Press

7. CONCLUDING REMARKS

We have shown that for a class of $\mathcal{D}$ restricted MDP, the optimality of a deterministic policy corresponding to a regular sequence is assured if some threshold condition is satisfied for the corresponding full observation MDP. In the present article we have not investigated whether the threshold condition actually holds for the corresponding full observation MDP. However, for many comparable MDP, such a threshold structure of optimal stationary policies has been investigated and established. For example, in Section 5.3 of [24] for the so-called searching for a moving-target problem, it was conjectured that the optimal policy has a simple threshold structure—namely search location 1 if and only if at the decision epoch the probability p that the target is at location 1 is larger (or equal) than a certain threshold probability p^*. In [19] the existence of such optimal threshold probability p^* and corresponding policy is proved for many cases of such searching for moving-target MDP. Condition 30 for MDP associated with $\mathcal{D} = \{d^1, d^2\}$ restricted MDP is similar and possibly for some problem classes, it can be established by similar methods as in [19].

If Condition 30 indeed holds, then the (desired) optimality within the class of policies corresponding to regular sequences follows if some additional (and easy checkable) technical conditions (see Theorem 32) are satisfied for the transition matrices induced by the applicable decision rules in $\mathcal{D}$. Note that we have proved that these additional conditions stated in Theorem 32 are sufficient, but possibly these technical conditions can be weakened. Moreover, it is interesting whether Theorem 32 can be generalized to $\mathcal{D} = \{d^1, d^2\}$ restricted MDP with S consisting of more than two states. Another issue is whether the results on optimality of regular sequences can be extended from the relatively simple threshold structure given by Condition 30 to more involved cases where, for example, an optimal stationary policy is determined by multiple thresholds.

References

1. Altman, B., Gaujal, E., & Hordijk, A. (2003). *Discrete-event control of stochastic networks: Multimodularity and regularity*. Lecture Notes in Mathematics. New York: Springer Verlag.
2. Altman, E., Gaujal, B., & Hordijk, A. (2000). Balanced sequences and optimal routing. *Journal of the ACM* 47: 752–775.
3. Altman, E., Gaujal, B., & Hordijk, A. (2000). Multimodularity, convexity and optimization properties. *Mathematics of Operations Research* 25: 324–347.
4. Altman, E., Gaujal, B., Hordijk, A., & Koole, G. (1998) Optimal admission, routing and service assignment control: the case of single buffer queues. In *the 37th IEEE Conference on Decision and Control*, Tampa, FL, Vol. 2, pp. 2119–2124.
5. Altman, E. & Shwartz, A. (1991). Markov decision problems and state-action frequencies. *SIAM Journal on Control and Optimization* 29: 786–809.
6. Bhulai, S., Farenhorst-Yuan, T., Heidergott, B., & van der Laan, D.A. (2010). Optimal balanced control for call centers. Technical report, Tinbergen Institute.
7. Cao, X.R. (1998). The MacLaurin series for performance functions of Markov chains. *Advances in Applied Probability* 30: 676–692.
8. Fernández-Gaucherand, E., Araposthathis, A., & Marcus, S.I. (1991). On the average cost optimality equation and the structure of optimal policies for partially observable Markov decision processes. *Annals of Operations Research* 29: 439–470.

https://doi.org/10.1017/S0269964811000039 Published online by Cambridge University Press

9. Fernández-Gaucherand, E., Araposthathis, A., & Marcus, S.I. (1991). Remarks on the existence of solutions to the average cost optimality equation in Markov decision processes. *Systems and Control Letters* 15: 425–432.
10. Gaujal, B., Hordijk, A., & van der Laan, D.A. (2007). On the optimal policy for deterministic and exponential polling systems. *Probability in the Engineering and Informational Sciences* 21: 157–187.
11. Hajek, B. (1985). Extremal splittings of point processes. *Mathematics of Operations Research* 10(4): 543–556.
12. Heidergott, B. & Hordijk, A. (2003). Taylor series expansions for stationary Markov chains. *Advances in Applied Probability* 35: 1046–1070.
13. Heidergott, B. & Vázquez-Abad, F. (2008). Measure valued differentiation for Markov chains. *Journal of Optimization and Applications* 136: 187–209.
14. Heidergott, B., Vázquez-Abad, F.J., Pflug, G., & Farenhorst-Yuan, T. (2010). Gradient estimation for discrete-event systems by measure-valued differentiation. *ACM Transactions on Modeling and Computer Simulation (TOMACS)* 20: 1–28.
15. Hernández-Lerma, O. & Lasserre, J.B. (1996). *Discrete-time Markov control processes: Basic optimality criteria*. New York: Springer.
16. Hordijk, A. & van der Laan, D.A. (2005). On the average waiting time for regular routing to deterministic queues. *Mathematics of Operations Research* 30: 521–544.
17. Koole, G. (1999). On the static assignment to parallel servers. *IEEE Transactions on Automatic Control* 44: 1588–1592.
18. Lothaire, M. (2002). *Algebraic combinatorics on words*. Cambridge: Cambridge University Press.
19. MacPhee, I.M. & Jordan, B.P. (1995). Optimal search for a moving target. *Probability in the Engineering and Informational Sciences* 9: 159–182.
20. Morse, M. & Hedlund, G.A. (1940). Symbolic dynamics II — sturmian trajectories. *American Journal of Mathematics* 62: 1–42.
21. Pflug, G.C. (1996). *Optimization of stochastic models*. Amsterdam: Kluwer Academic.
22. Puterman, M. (1994). *Markov decision processes: Discrete stochastic dynamic programming*. New York: John Wiley and Sons.
23. Ross, K.W. (1989). Randomized and past-dependent policies for Markov decision processes with multiple constraints. *Operations Research* 37: 474–477.
24. Ross, S.M. (1983). *Introduction to stochastic dynamic programming*. New York: Academic Press.
25. Tijdeman, R. (2000). Fraenkel's conjecture for six sequences. *Discrete Mathematics* 222: 223–234.

https://doi.org/10.1017/S0269964811000039 Published online by Cambridge University Press