ARTICLE IN PRESS

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Economic Theory ▮ (▮▮▮▮) ▮▮▮–▮▮▮

JOURNAL OF Economic Theory

http://www.elsevier.com/locate/jet

Randomization and simplification in dynamic decision-making

Ehud Kalai[a,b] and Eilon Solan[a,c,*]

[a] *MEDS Department, Kellogg School of Management, Northwestern University, 2001 Sheridan Road, Evanston, IL 60208, USA*

[b] *Department of Mathematics, College of Arts and Sciences, Northwestern University, Evanston, IL 60208-2001, USA*

[c] *School of Mathematical Sciences, Tel Aviv University, Tel Aviv 69978, Israel*

Received 22 June 2000; final version received 24 April 2002

Abstract

Randomization adds beneficial flexibility to the construction of optimal simple plans in dynamic environments. A decision-maker, restricted to the use of simple plans, may find a stochastic plan that strictly outperforms all deterministic ones. This is true even in noninteractive deterministic environments where the decision-maker's choices have no influence on his signals nor on the future evolution of the system. We describe a natural decomposition of simple plans into two components: an *action selection* rule and a *behavior modification* rule. In noninteractive environments optimal simple plans do not require randomization in the action selection rule. Only randomization in the behavior modification rule may be necessary.

JEL classification: C61; C44

Keywords: Dynamic decision-making; Randomization; Simplification

1. Introduction

A celebrated result of Blackwell [2] states that in every Markov decision problem (MDP) with finite state and action spaces, the decision-maker has an optimal

[*] Corresponding author. MEDS Department, Kellogg School of Management, Northwestern University, 2001 Sheridan Road, Evanston, IL 60208, USA.

E-mail addresses: kalai@kellogg.northwestern.edu (E. Kalai), e-solan@kellogg.northwestern.edu (E. Solan).

0022-0531/03/$ - see front matter © 2003 Elsevier Science (USA). All rights reserved.
doi:10.1016/S0022-0531(03)00072-3

deterministic stationary plan. What happens to this result if we restrict the decision-maker to *simple* plans?

The main conclusions to be taken from this paper are the following. First, and somewhat surprising, is the observation that Blackwell's conclusion no longer holds (see Example 2). Even if restricted to noninteractive environments, where the decision-maker actions have no influence on the information he receives nor on the future evolution of the environment, a simple plan that uses randomization may outperform all deterministic simple plans.

Given the above observation, one would like to know more about the nature of such beneficial randomization. The paper describes a natural decomposition of simple plans into two components: an *action selection* rule and a *behavior modification* rule. It goes on to show (Theorem 1) that Blackwell's conclusion is partly retained. Namely, in noninteractive environments, optimal simple plans do not require randomization in the action selection rule. Only randomization in the behavior modification rule may be necessary.

Before proceeding with the main results, the following is a brief elaboration of Blackwell's result and its extensions to more general MDPs.

The plans obtained by Blackwell [2] are optimal in very strong senses. One can find an optimal deterministic stationary plan for every discount parameter that the decision-maker uses for the purpose of evaluating his payoffs. Moreover, one can find a single plan that is optimal for all MDPs with discount parameter below some critical level. Such a plan is also optimal under the average cost criterion; that is, when the decision-maker tries to maximize the limsup of his average daily payoffs (see, e.g., [1, Theorem 4.3]).

Moreover, one can generalize Blackwell's result to MDPs unrestricted to finitely many actions or states, provided that continuity and compactness assumptions are satisfied: for every discount parameter there is an optimal deterministic stationary plan (which may depend on the discount parameter), and, under an appropriate ergodicity assumption, there is an optimal deterministic stationary plan that is optimal under the average cost criterion (see, e.g., [1, Theorems 6.4 and 6.5]).

The situation becomes more involved when we move to MDPs with partial observation (see, e.g., [1] or [3]). These are MDPs where the decision-maker does not observe the actual state of the world, but, rather, observes a stochastic signal, that depends on the actual state and on his action. The standard approach to deal with this model is to define an auxiliary problem, where the state variable is the space of probability distributions over states of the world—the state of the auxiliary problem at stage n is the conditional probability over states given the information of the decision-maker up to that stage. One can then apply tools from the theory of Markov decision problems with general state space.

Unfortunately, the auxiliary problem does not necessarily satisfy an ergodicity condition, hence the existence of an optimal plan is not assured. Moreover, it is well known that in MDPs with partial observation there need not exist an optimal deterministic plan (see [10, Example 7.1.3]) nor an optimal stationary plan in the auxiliary problem (see [10, Example 7.1.5]). This means that when partial observation is present, the conditional distribution over states is not a sufficient statistic for an optimal plan. Nevertheless, Rosenberg et al. [9] proved that if the state and action spaces are finite, as is the space of signals the decision-maker can observe, then there exist epsilon uniformly optimal plans for patient decision-makers. That is, for every $\varepsilon > 0$ there exist $\lambda > 0$ and a plan σ, such that σ is ε-optimal for every discount parameter smaller than λ: under the discounted evaluation, no plan outperforms σ by more than ε.

In general, the epsilon uniformly optimal plan is neither deterministic nor stationary.[1] However, if the model is deterministic, that is, both the new state of the world and the signal depend deterministically on the current state and on the action chosen by the decision-maker, then this plan can be chosen to be deterministic and stationary.

In this paper, an MDP with partial observation is described by the following entities. (1) Finite sets of *states* (of the world), of *actions* and of *signals*. One of the states is designated as the *initial state*. (2) A *payoff function* that assigns to every state and action a real number. (3) A stochastic *information function*, assigning to every state and action a probability distribution over the set of signals. (4) A stochastic *transition function*, assigning to every state and action a probability distribution over the set of states.

The MDP evolves as follows. At the initial state the decision-maker chooses an action (with the possible aid of a randomization device). According to the state and the selected action, the decision-maker is awarded the associated payoff, and is told the signal that is generated by the stochastic information function.[2] A new state is selected by the stochastic transition function and the process repeats itself with the newly selected state playing the role of the initial state.

The present note studies the structure of optimal *simple* plans for MDPs with partial observation. Simplicity of plans may be imposed by exogenous considerations, for example if the plan needs to be communicated to a less than fully able executor, or as a cost savings device, especially if the loss to payoff is not significant. Example 1 below illustrates the point.

Formally, a decision-maker's plan is a function that assigns to every sequence of past signals a probability distribution over the set of actions.

Every such plan induces a subplan in each stage of the evolution of the MDP; that is, the plan's action selection as a function of the finite sequences of observed future signals from the current stage on. One measure of the complexity of a plan is the number of *different* subplans it induces. A stationary plan, for example, induces the same subplan in each stage, while a periodic plan with period p induces p different subplans. As Kalai and Standford [5] observed, this is also the size of the smallest automaton that can implement the plan. Indeed, one can represent each induced subplan by a state of the automaton, and, for every state (= subplan) and every

[1] Here stationary means stationary in the auxiliary problem.

[2] Notice that this formulation allows both, situations where the decision-maker is or is not informed of his realized payoffs. The modeler has the freedom to include the realized payoff in the signal that the decision-maker receives. But even when the decision-maker is not told his payoff, he may be able to infer it from his signal, as is the case in some of the examples presented here.

signal, the new state is the one that corresponds to the induced subplan conditioned on the signal.

This representation gives rise to a natural decomposition of a plan; a plan is given by an *action selection* rule and a *behavior modification* rule. The former indicates which action should be taken at the current stage, and the latter indicates what continuation plan to use from now on.

The definitions follow, and we refer the reader to [4] for general discussion, and to the survey by Kalai [6] for elaboration on uses in decision theory and game theory.

Example 1 (Long seasons). Consider the following deterministic MDP with partial observation with 365 states $\{Winter_1, Winter_2, \ldots, Winter_{165}, Summer_1, Summer_2, \ldots, Summer_{200}\}$, two actions $\{Take\ Umbrella,\ Don't\ take\ Umbrella\}$, and two signals $\{Rainy, Shiny\}$.[3]

Transition is deterministic and independent of the action chosen by the decision-maker: W_1 is the initial state, each state W_i leads to state W_{i+1}, except for state W_{165} that leads to state S_1, and each state S_i leads to state S_{i+1}, except for state S_{200} that leads to state W_1. Payoff is given by:

$$r(W_i, T) = 1, \quad r(W_i, D) = 0, \quad i = 1, \ldots, 165,$$

$$r(S_i, T) = 0, \quad r(S_i, D) = 1, \quad i = 1, \ldots, 200.$$

Signals depend deterministically on the state: the signal in state W_i, $i = 1, \ldots, 165$, is *Rainy*, and the signal in state S_i, $i = 1, \ldots, 200$, is *Shiny*. The MDP is depicted in Fig. 1.

Clearly, by using a plan that counts to 165, and then to 200, a decision-maker can achieve an average payoff 1, and cannot do better than that.

Let us now consider only plans that can be described by two-state automata. An *automaton C* is given by the following.

- A finite state space with one designated as the initial state.
- A finite action set with an action selection rule that assigns to every state of the automaton a probability distribution over actions.
- A finite set of input signals.
- A transition rule that assigns to every state and every input signal a probability distribution over the next state.

Notice that the word state in this paper has a double use, since states of the world are different from states of the automaton. When it is not clear from the context, we refer to states as either states of the automaton or states of the world.

The automaton "plays" as follows. (i) Starting at the two initial states, of the world and the automaton, the decision-maker takes the random action generated by the automaton's initial state, and is paid off accordingly. (ii) The automaton is given

[3] Below, when referring to states and actions, we usually abbreviate $Winter_i$ by W_i, *Take Umbrella* by T, etc.

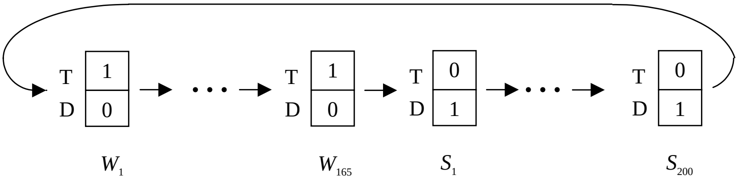

Fig. 1. The MDP with long seasons.

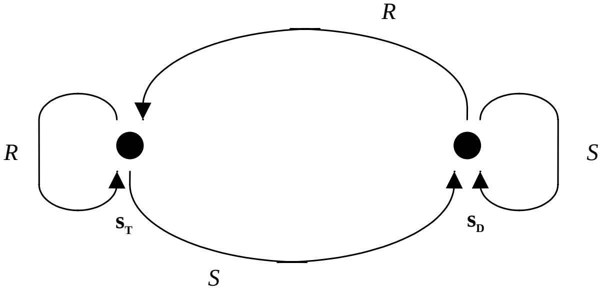

Fig. 2. The optimal two state automaton.

the random input signal generated stochastically at the state of the world and the action just taken. (iii) Based on the initial state of the world and the action chosen by the automaton, a new state of the world is selected according to the stochastic transition function of the MDP. (iv) According to its transition rule, the automaton chooses its next state stochastically, according to the probability distribution at the initial state and the input signal. (v) The process repeats itself with the current states of the world and the automaton playing the roles of the initial states.

If a plan can be described by an automaton, then the expected average payoff of this plan is well defined. Moreover, as the discount parameter goes to 0, the discounted evaluation of the stage payoffs converges to the expected average payoff.

One can verify that the optimal two state automaton in this example is the one that appears in Fig. 2, where the action *Take umbrella* is taken in state $\mathbf{s}_T$ (the initial state), and the action *Don't take umbrella* is taken in state $\mathbf{s}_D$.

The arrows that leave each state denote the (deterministic) transition rule of the automaton, and the labels next to the arrows describe the associated (deterministic) input signals about the state just visited. This automaton "misses" twice in every year: in states W_1 and S_1. Its expected payoff is, therefore, 363/365, which is pretty close to the optimal value 1.

In the present note we consider MDPs with partial observation, and with the added restriction of the decision-maker to use plans of complexity n, i.e., ones that can be implemented by automata with n or fewer states. We assume that n is fixed throughout the paper. In the sequel, we identify each plan with the automaton that implements it.

Since we bound the size of the automata the decision-maker can use, unless there is a sufficient statistic with a small state space, the epsilon uniformly optimal plan that exists by Rosenberg et al. [9] may not be feasible.

Let the set of complexity n plans, $C(n)$, consist of all plans that can be described by automata with n states. The main interest of this note is the optimal or nearly optimal automata in this set.[4] Two subsets of plans are of interest: deterministic-transition automata and deterministic-action automata. *Deterministic-transition automata* are automata where the transition rule is deterministic, but the action selection rule may be nondeterministic, whereas *deterministic-action automata* are automata where the action selection rule is deterministic, but the transition rule may be nondeterministic. An automaton is *deterministic* if both transition rule and action selection rule are deterministic. The automaton that appears in Fig. 2 is deterministic.

A MDP with partial observation is called *noninteractive*, if its transition function and information function are both independent of the action chosen by the decision-maker. This means that in every stage of the process, the action chosen by the decision-maker only affects his stage payoff. It has no effect on his next signal or on the future evolution of the environment.

Theorem 1. *Consider a noninteractive MDP. For every complexity n automaton $C \in C(n)$ there exists a deterministic-action automaton in $C(n)$ that performs at least as good as C.*

As already stated, one may view the above theorem as conveying both a positive and a negative result. The positive one is the fact that as in the intuition developed from Blackwell and the follow up literature, despite the restriction to simple rules we can still choose actions deterministically.

But perhaps more surprising is the negative result that is illustrated by the fact that Theorem 1 is sharp. More specifically, optimization among simple plans does require randomization, even if only in the automaton's transition rules. As explained in the sequel, randomization introduces a certain amount of flexibility that is lacking in rigidly deterministic simple plans. Such flexibility may lead to strictly higher expected payoffs.

Example 2 below illustrates this point. There, the optimal automaton in $C(2)$ requires the use of randomization in the transition rule. To make the point as sharp

[4] In general, as pointed out by Abraham Neyman, there need not be optimal plans in $C(n)$ for any n. As an example, take a MDP with 4 states $\{Bad, Good, t_0, t_1\}$, *Bad* being the initial state, 2 actions $\{Invest, Collect\}$, and no signals. States t_0 and t_1 are absorbing, with payoff 0 and 1 respectively. In both states *Bad* and *Good*, payoff is 0 whichever action is chosen. Transition is as follows:

$$q(t_0 \mid Bad,\ Collect) = q(t_1 \mid Good,\ Collect) = 1$$

$$q(Good \mid Bad,\ Invest) = 1 - q(Bad \mid Bad,\ Invest) = 1/2,\ q(Good \mid Good,\ Invest) = 1.$$

In this MDP, no plan guarantees an expected average payoff 1. The automaton in $C(1)$ that plays repeatedly the action *Collect* with probability p and the action *Invest* with probability $1-p$ guarantees an expected payoff $1 - p/(1-(1-p)/2)$, which goes to 1 as p goes to 0.

as possible the example uses a deterministic MDP, so the use of randomization by the decision-maker has nothing to do with randomness in the environment.

Additional examples that follow illustrate that Theorem 1 is tight in other aspects.

2. Proof of Theorem 1 and Examples

Proof of Theorem 1. Let a noninteractive MDP with partial observation be given. Let Ω be the set of states of the world, A the actions of the decision-maker, and $r(\omega, a)$ the daily payoff when the action a is chosen in state ω.

Let an automaton C in $C(n)$ be given, and denote its state space by S. Let $p_C(s, \omega)$ be the conditional probability that the MDP is in state ω provided the decision-maker uses C and that the automaton is in state s. Formally, we let $p_C(s) = \lim_{n\to\infty} \mathbf{E}_C[\#\{k \leqslant n \mid \text{at stage } k, \text{ the automaton is in state } s\}]/n$, which is well defined. Whenever $p_C(s) > 0$, $p_C(s, \omega) = \lim_{n\to\infty} \mathbf{E}_C[\#\{k \leqslant n \mid \text{at stage } k, \text{ the automaton is in state } s \text{ and the state of the world is } \omega\}]/(n \times p_C(s))$.

Since the actions of the decision-maker influence neither the transition function of the MDP nor the distribution of signals, we have for every automaton D in $C(n)$ that has the same transitions as C

$$p_C(s, \omega) = p_D(s, \omega) \quad \text{for every } s \in S \text{ and every } \omega \in \Omega.$$

Since the payoff is linear in the actions of the decision-maker, one can choose, for every state s, an action a_s of the decision-maker that maximizes $\sum_{\omega \in \Omega} p_C(s, \omega) r(\omega, a)$. a_s is defined arbitrarily when $p_C(s) = 0$.

Define a deterministic-action automaton D in $C(n)$ as follows. It has the same state space and transition rule as C, and in every state $s \in S$ the action a_s is taken. The automaton D achieves at least as high a payoff as C. □

We now show that Theorem 1 is tight in the following senses. In Example 2 we show that even for a deterministic noninteractive MDP there need not be an optimal deterministic simple automaton. In Example 3 we show that if the actions influence the distribution of the signals, but not the transitions of the MDP, Theorem 1 no longer holds. In Example 4 we show that if the actions influence the transitions of the MDP, but not the distribution of the signals, Theorem 1 need not hold as well. In Example 5 we show that Theorem 1 does not generalize to a model with two decision-makers.

In all these examples the MDP is deterministic and we take $n = 2$: the decision-maker can use automata of size two. Analogous examples can be constructed for any finite n.

Example 2 (Short seasons, deterministic noninteractive MDP requiring automaton with stochastic transitions). Take a MDP with 3 states $\{Winter_1, Winter_2, Summer\}$, two actions $\{$*Take umbrella*, *Don't take umbrella*$\}$, and two signals $\{Rainy, Shiny\}$.

Transition is deterministic and periodic: W_1 is the initial state, it leads to state W_2, which leads to state S, which leads to state W_1.

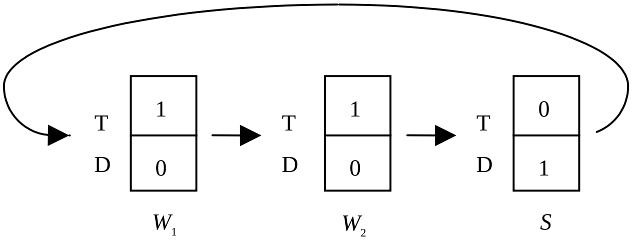

Fig. 3. The MDP with short seasons.

Payoff is given by:

$$r(W_i, T) = 1, \quad r(W_i, D) = 0, \quad i = 1, 2,$$
$$r(S, T) = 0, \quad r(S, D) = 1.$$

The signal is *Rainy* in states W_1 and W_2, and is *Shiny* in state S. The MDP is depicted in Fig. 3.

The best the decision-maker can do using deterministic automata in $C(2)$ is 2/3, which is obtained by always taking an umbrella. Indeed, if the automaton uses only a single action, the best it yields is 2/3. If it uses both actions, denote the two states of the automaton by $\mathbf{s}_T$ (action T is taken) and $\mathbf{s}_D$ (action D is taken). If the transition from $\mathbf{s}_T$ given the signal *Rainy* is to $\mathbf{s}_D$, the automaton misses in state W_2, while if it is to stay at $\mathbf{s}_T$, it misses in state S.

We now show that there is a deterministic action automaton that yields expected average payoff 3/4. One can show that this automaton is optimal in $C(2)$.

Consider the deterministic action automaton $K(p)$, that depends on a single parameter $p \in [0, 1]$, and has the following transition rule.

In $K(p)$, p is the probability to move to state $\mathbf{s}_D$ if the current state is $\mathbf{s}_T$ and the observed signal is *Rainy*. The initial state is $\mathbf{s}_T$ (see Fig. 4).

Under $K(p)$, after a shiny day the automaton will be in state $\mathbf{s}_T$, hence the probability of success in state W_1 is 1, the probability of success in state W_2 is $1 - p$, and the probability of success in state S is $1 - (1-p)^2 = 2p - p^2$. In particular, the expected average payoff is $(2 + p - p^2)/3$, which is maximized at $p = 1/2$, and gives an expected average payoff 3/4.

Example 3. The present example shows that even in partially noninteractive MDPs, where the choice of an action influences the next signal but not the transition to the next state of the world, the conclusion of Theorem 1 no longer holds.

Take a MDP with $2r$, states $\{W_1, W_2, \ldots, W_r, S_1, S_2, \ldots, S_r\}$, where $r > 1$ is even, two actions $\{Home, Work\}$ and three signals $\{0, 1, -1\}$.

Transition is deterministic and independent of the actions of the decision-maker: W_1 is the initial state, each state W_i leads to state W_{i+1}, except for state W_r that leads to state S_1, each state S_i leads to state S_{i+1}, except for state S_r that leads to W_1.

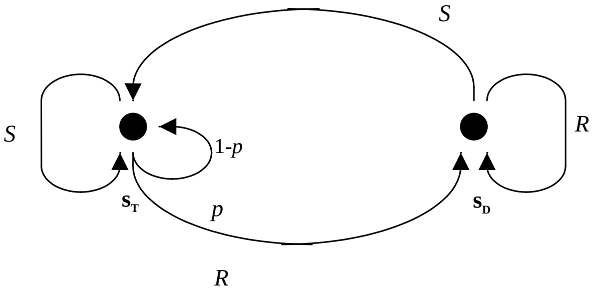

Fig. 4. The transition of the automaton $K(p)$.

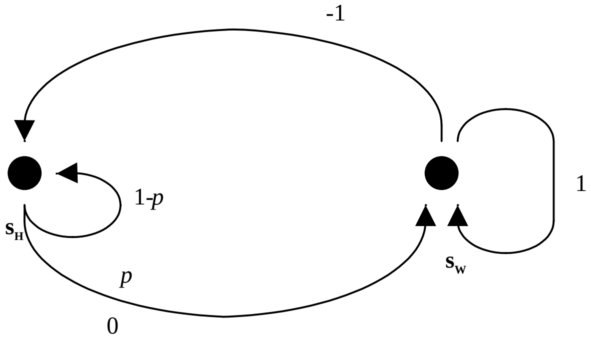

Fig. 5. The automaton $H(p)$.

The payoff is given by:

$$r(W_i, H) = r(S_i, H) = 0, \quad i = 1, \ldots, r,$$

$$r(S_i, W) = 1, \quad r(W_i, W) = -1, \quad i = 1, \ldots, r.$$

The signal the decision-maker receives coincides with his payoff. Thus, as long as the decision-makers remains at home, he does not learn anything about the weather, but once he goes to work, he observes it.

Consider the two-state automaton $H(p)$ that appears in Fig. 5.

In state $\mathbf{s}_H$ (the initial state), the action H is taken; in state $\mathbf{s}_W$, the action W is taken. Transition is as follows: if the automaton is in state $\mathbf{s}_W$ and the signal is 1, the automaton remains in the same state, whereas if the signal is -1, it moves to state $\mathbf{s}_H$. If the automaton is in state $\mathbf{s}_H$, the signal is 0, and then the automaton remains in state $\mathbf{s}_H$ with probability $1-p$, and moves to state $\mathbf{s}_W$ with probability p.

One can verify that the optimal deterministic automaton in $C(2)$ is $H(1)$, which achieves an expected average payoff 1/4: this automaton goes to work all summer, and every other day during winter. Consider now the automaton $H(p)$ for $0<p<1$. During winter, the automaton goes to work in each day with probability p. Once summer starts, it takes the automaton on average $1/p$ stages to observe that, and

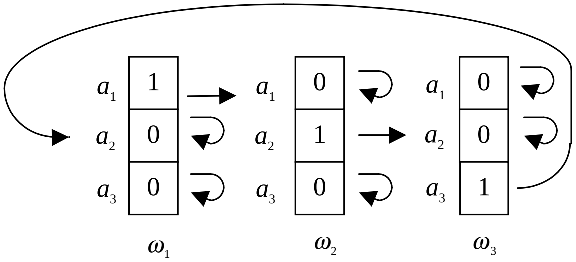

Fig. 6. The MDP.

	b_1	b_2	b_3
a_1	1	0	0
a_2	0	1	0
a_3	0	0	1

Fig. 7. The unique state of the MDP.

thereafter it goes to work for the remaining of the season. When p is small, and r is much larger than $1/p$, the expected average payoff of this automaton is close to 1.

Example 4. The present example shows that if the selected action influences the transition of the MDP but not the signals, Theorem 1 no longer holds.

Consider an MDP with three states $\{\omega_1, \omega_2, \omega_3\}$ (ω_1 being the initial state), three actions $\{a_1, a_2, a_3\}$ and no signals. The payoff is 1 if action a_i is chosen in state ω_i, and 0 otherwise:

$$r(\omega_i, a_j) = 1 \quad \text{if } i = j, \qquad r(\omega_i, a_j) = 0 \quad \text{if } i \neq j.$$

Transition is as appears in Fig. 6.

Any deterministic action automaton in $C(2)$ can use only two actions, and therefore its expected average payoff is 0. However, there is an automaton in $C(1)$ that achieves on average 1/3 (choose all three actions with equal probabilities).

The same point can be made by properly adapting the example studied by Piccioni and Rubinstein [8].

Example 5. The present example shows that Theorem 1 cannot be generalized to the setup of several decision-makers.

We consider an MDP with a single state and *two* decision-makers DM1 and DM2, or a repeated game. Each of the two decision-makers has three actions, $\{a_1, a_2, a_3\}$ and $\{b_1, b_2, b_3\}$, respectively (see Fig. 7).

There is one state of the world, so the transition is trivial, and there are no signals. The MDP is *zero-sum*, so the sum of the payoffs of the two decision-makers is 0. The payoff of DM1 is:

$$r(a_i, b_j) = 1 \quad \text{if } i = j, \qquad r(a_i, b_j) = 0 \quad \text{if } i \neq j.$$

By using an automaton with a single state, that chooses each action with probability 1/3, DM1 can guarantee that his expected average payoff is at least 1/3, and DM2 can guarantee that his expected average payoff does not fall below $-1/3$. However, for any deterministic action automaton of DM1 in $C(2)$, DM2 has a deterministic automaton in $C(1)$ that guarantees him an expected average payoff 0. Indeed, any deterministic action automaton of DM1 in $C(2)$ uses at most two actions. If DM2 uses the automaton that deterministically chooses the action that corresponds to the one not used by DM2, the expected average payoff is 0.

3. Interpretation and discussion

1. *Example 1: On randomization, flexibility and bounded recall.* At first glance, it seems surprising that a decision-maker would choose to randomize in a noninteractive one-person decision problem. Under the conditions of Kuhn's [7] theorem (restricted to the case of one player), any nondeterministic plan is a convex combination of deterministic plans, with payoffs being linear in the convex combinations. Thus, no nondeterministic plan could do better than all deterministic plans. Fig. 8 helps clear the situation.

For one cycle (W_1, W_2, S), the graph describes the probability tree of the optimal plan in Example 2. It gives the paths that can occur in the cycle.

The state of the automaton at the beginning of the cycle is $\mathbf{s}_T$. Then, a signal *Rainy* is observed: with probability 0.5 the new state is $\mathbf{s}_T$, and with probability 0.5 it is $\mathbf{s}_D$. Again a signal *Rainy* is received, and a new state is chosen. A bold circle means that at that stage the action chosen by the automaton is "correct", and a thin circle means it is "incorrect".

Note that any path yields an average payoff at least 2/3. The path $\mathbf{s}_T$–$\mathbf{s}_T$–$\mathbf{s}_T$ can be represented by an automaton that prescribes always taking an umbrella, and the path $\mathbf{s}_T$–$\mathbf{s}_D$–$\mathbf{s}_D$ can be represented by an automaton that prescribes taking an umbrella after a shiny day, and not taking an umbrella after a rainy day. These two automata are deterministic, and yield average payoff 2/3. The middle path, $\mathbf{s}_T$–$\mathbf{s}_T$–

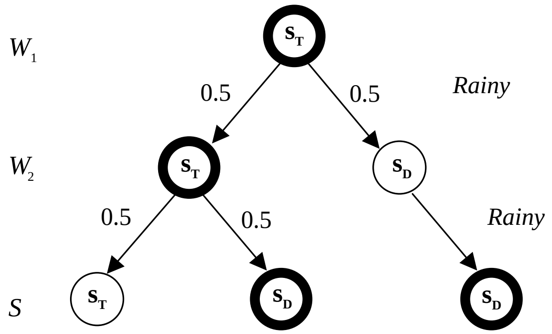

Fig. 8.

$\mathbf{s}_D$, yields average payoff 1, but alas, cannot be generated by a deterministic automaton of size 2.

Thus, with probability 0.25, the randomizing automaton adds to the decision-maker a behavior pattern not possible with deterministic automata of size 2. Or, in other words, flexibility is not possible otherwise. This is exactly where our gain came from.

It is also easy to see why the conclusion of Kuhn's theorem does not hold. Kuhn's decomposition of the optimal plan in that example involves two deterministic plans (automata) of two states, and one deterministic plan of three states, which is not permissible.

In terms of Kuhn's assumptions, requiring a decision-maker to use plans describable by automata with bounded number of states, forces him to have imperfect recall—the automaton only "knows" what state it is in but not how it got there.

Example 2 shows that simplicity, flexibility and randomization remain tied together even in the *most elementary* environments.

2. *Example 3: On randomization and learning.* In Example 3 randomization is used for experimenting and learning. This example illustrates a subtle, yet important, point. Here, the selected action, while not affecting the transitions, does have an effect on the future state of the system, namely the informational state. This is not the case in Example 2 where the decision-maker observes the same signal regardless of his action. Thus, the randomization illustrated in Example 2 is more revealing since it takes place in environments that are in a certain sense more truly time independent of the decision-maker actions.

The overall logic of the automaton $H(p)$ depicted in Fig. 5 is clear. At the beginning of each "season", the state of the world is unknown. Randomization allows the decision-maker to experiment, and find out the true state of the world. The optimal rate of such experiments is carefully balanced. It is not too high, so that he does not experiment too often once the information is already revealed, and it is not too low, so the information is not revealed too late.

One could argue, however, that even in Example 2 there is informational time dependency which is not due to the physical rules of the environment, but to the constrains (self-imposed or not) on the decision-maker. Even if given information, a limited decision-maker may not have the ability to "remember it" due to a limitation on the number of states. His optimal automaton is constructed to do learning of information that is available from the environment, but not in his limited memory. In this sense we may want to differentiate and think of Example 3 as one involving exogenous dependencies between environment changes and action, whereas in Examples 1 and 2 this dependency is endogenously chosen.

Viewed in this way, the optimal design of a simple plan should address the issue of what information should be forgotten, or ignored, to be generated when needed by experimentation.

3. *On the type of simplification device.* There is no universal agreement on the proper way of measuring complexity, or simplicity, of plans. Two-state automata enable us to represent clearly and concisely simple rules of thumb. Note that

automata are more suitable to dynamic environments with signals than Markovian plans, since the latter force the decision-maker to ignore the observed signal. Moreover, the number of states of the automaton can serve as a fine measure of the complexity of the plan. This is unlike Turing machines where all computable decision rules can be described by the same universal Turing machine, that requires only a small number of states (see, e.g., [4]).

But under most reasonable measures of complexity two-state automata would be considered simple. Thus, regardless of the debate on complexity, at a minimum, this note establishes a connection between randomization and a strong version of simplicity.

4. *The complexity of randomization*. The discussion above suggested that the performance of simple plans may be improved through randomization. It ignored however, the cost and complexity of the randomization process itself. This may be the case if the randomization is done in one's mind, but not if the randomization is done by the use of some costly device. Does randomization improve performance even when its cost is taken into consideration seems like an interesting open question.

5. *Open problems*. The above examples raise a large number of general open questions. For example,

(a) When is optimality obtained by a random (rather than a deterministic) automaton?
(b) Can one bound the performance of the optimal automaton (with an exogenous bound on the number of states)?
(c) Can one bound the improvement by which the optimal random automaton will outperform the optimal deterministic automaton?

Acknowledgments

The authors thank Adam Kalai, Ehud Lehrer and Abraham Neyman for helpful conversations, and an anonymous referee for very helpful suggestions that drastically improved the paper. Kalai acknowledges financial support from the National Science foundation, Grant No. SES 9904931.

References

[1] A. Arapostathis, V.S. Borkar, E. Fernádez-Gaucherand, M.K. Ghosh, S.I. Marcus, Discrete-time controlled Markov processes with average cost criterion: a survey, SIAM J. Control Optim. 31 (1993) 282–344.

[2] D. Blackwell, Discrete dynamic programming, Ann. Math. Statist. 33 (1962) 719–726.

[3] K. Hinderer, Foundations of Non-stationary Dynamic Programming with Discrete Time Parameter, Springer, Berlin, 1970.

[4] J.E. Hopcroft, J.D. Ullman, Introduction to Automata Theory, Languages, and Computation, Addison–Wesley Series in Computer Science, Addison–Wesley Publishing Co., Reading, MA, 1979.

[5] E. Kalai, W. Stanford, Finite rationality and interpersonal complexity in repeated games, Econometrica 56 (2) (1988) 387–410.

[6] E. Kalai, Bounded rationality and strategic complexity in repeated games, in: T. Ichiishi, A. Neyman, Y. Tauman (Eds.), Game Theory and Applications, Academic Press, New York, 1990, pp. 131–157.

[7] H.W. Kuhn, Extensive games and the problem of information, in: H.W. Kuhn, A.W. Tucker (Eds.), Contributions to the Theory of Games I, Princeton University Press, Princeton, 1953, pp. 193–216.

[8] M. Piccioni, A. Rubinstein, On the interpretation of decision problems with imperfect recall, Games Econom. Behav. 20 (1997) 3–25.

[9] D. Rosenberg, E. Solan, N. Vieille, Blackwell optimality in Markov decision processes with partial observation, Ann. Statist. 30 (2002) 1178–1193.

[10] L.I. Sennott, Stochastic Dynamic Programming and the Control of Queueing Systems, Wiley, New York, 1999.