DYNAMIC NON BAYESIAN DECISION MAKING

Dov Monderer and Moshe Tennenholtz

Faculty of Industrial Engineering and Management,
Technion–Israel Institute of
Technology, Haifa , Israel
e–mail:
dov@ie.technion.ac.il
moshet@ie.technion.ac.il

Abstract.

The model of a non-Bayesian agent who faces a repeated game with incomplete information against Nature is an appropriate tool for modeling general agent-environment interactions. In such a model the environment state (controlled by Nature) may change arbitrarily, and the feedback/reward function is initially unknown. The agent is not Bayesian, that is he does not form a prior probability neither on the state selection strategy of Nature, nor on his reward function. A policy for the agent is a function which assigns an action to every history of observations and actions. Two basic feedback structures are considered. In one of them – the perfect monitoring case – the agent is able to observe the previous environment state as part of its feedback, while in the other – the imperfect monitoring case – all that is available to the agent is the reward obtained. Both of these settings refer to partially observable processes, where the current environment state is unknown. Our main result refers to the competitive ratio criterion in the perfect monitoring case; We prove the existence of an efficient stochastic policy which ensures that the competitive ratio is obtained at almost all stages with an arbitrarily high probability, where efficiency is measured in terms of rate of convergence. It is further shown that such an optimal policy does not exist in the imperfect monitoring case. Moreover, it is proved that in the perfect monitoring case there does not exist a deterministic policy that satisfies our long run optimality criterion. In addition, we discuss the maxmin criterion and prove that a deterministic efficient optimal strategy does exist in the imperfect monitoring case under this criterion. Finally

This work was supported by the Fund for the Promotion of Research in the Technion.

we show that our approach to long-run optimality can be viewed as qualitative, which distinguishes it from previous work in this area.

1. Introduction.

Decision-making is a central task of artificial agents (Russell and Norvig (1995); Wellman (1985); Wellman and Doyle (1992)). At each point in time, an agent needs to select among several actions. This may be a simple decision, which takes place only once, or a more complicated decision where a series of decisions has to be made. The question of "what should the right actions be" is the basic issue discussed in both of these settings, and is of fundamental importance to the design of artificial agents.

A static decision-making context for an artificial agent consists of a set of actions that the agent may perform, a set of possible environment states, and a utility/reward function which determines the feedback for the agent when it performs a particular action in a particular state. In a dynamic decision-making setup the agent faces static decision problems over stages. At each stage the agent selects an action to be performed. The action selection strategy may depend on the history of actions and observations (e.g., rewards) obtained by the agent in the past. As a result of the agent's choices, the state of the environment may change.

An appropriate model for the representation of the above-mentioned agent-environment interactions is the model of a non-Bayesian agent who faces a repeated game with incomplete information against Nature [1]. In a repeated game against Nature the agent faces the same static decision problem at each stage while the environment state is taken to be an action chosen by its opponent. The fact that the game is repeated refers to the fact that the set of actions, the set of possible states, and the one shot utility function do not vary with time [2]. The incomplete information assumption means that the agent does not initially know the utility function, and the fact that the agent is non-Bayesian means that he does not form a prior probability neither on the values of the utility function, nor on the state selection

[1] Repeated games with complete information, or more generally, multistage games and stochastic games have been extensively studied in Game Theory and Economics. A very partial list includes: Shapley (1953), Blackwell (1956), Luce and Raiffa (1957), and more recently, Fudenberg and Tirole (1993), Mertens, Sorin and Zamir (1995), and the evolving literature on learning (e.g., Fudenberg and Levine (1996)). The incomplete information setup in which the player is ignorant about the game being played was inspired by Harsanyi (1967-8). See e.g., Aumann and Maschler (1995) for a comprehensive survey. Most of the above literature deals with (partially) Bayesian agents. Some of the rare exceptions are cited in Section 6.

[2] In the most general setup, those sets may vary with time. No useful analysis can be done in a model where those changes are completely arbitrary.

strategy of Nature. As an example for the generality of the above-mentioned setting, consider the model of Markov Decision Processes. A Markov Decision Process can be considered as a repeated game against Nature in which the agent's action in a given state determines (in a probabilistic fashion) the next state to be obtained. That is, the agent has a structural assumption on the state selection strategy.

When the agent performs an action a_t in state s_t, part of its feedback would be $u(a_t, s_t)$, where u is the reward function. We distinguish between two basic feedback structures. In one of them – the perfect monitoring case – the agent is able to observe the previous environment state as part of its feedback, while on the other – the imperfect monitoring case – all that is available to the agent is the reward obtained [3]. Notice that in both of these feedback structures, the current state is not observed by the agent when it is called to select an action. Such is also the case in the evolving literature on the problem of controlling partially observable Markov decision processes [4] (see e.g., Lovejoy (1991); Cassandra et. al. (1994)).

Consider the case of a one stage game against Nature, in which the utility function is known, but the agent can not observe the current environment state when selecting its action. How should the agent choose its action? Work on decision-making under uncertainty has suggested several approaches [5]. One of these approaches is the maxmin (safety-level) approach. According to this approach the agent would choose an action which maximizes its worst case payoff. Another approach is the competitive ratio approach (or its additive variant, termed the minmax regret decision criterion). According to this approach an agent would choose an action which minimizes the worst case ratio between the payoff it could have obtained had it known the environment state to the payoff it would actually obtain [6].

Given a repeated game with incomplete information against Nature, the agent would not be able to obtain its one stage optimal action (with respect to the competitive ratio or maxmin value criteria) at each stage, since the utility function is initially unknown. This calls for a precise definition of a long-run optimality criterion. In this paper we are mainly concerned with policies (strategies) guaranteeing that the optimal competitive ratio (or the maxmin value) is obtained in *most* stages.

[3] Notice that the former assumption is very popular in the related game theory literature (see e.g., Aumann and Maschler (1995)).

[4] In contrast, Q-learning theory (see e.g., Watkins (1989); Watkins and Dayan (1992); Sutton (1992)) does assume a knowledge of the current state.

[5] See e.g., Savage (1954), Milnor (1954), Luce and Raiffa (1957), and Kreps (1988).

[6] The competitive ratio decision criterion has been found to be most useful in settings such as on-line algorithms (e.g., Papadimidtriou and Yanakakis (1989))

We are interested in particular in efficient policies, where efficiency is measured in terms of rate of convergence.

In Section 2 we define the above mentioned setting and introduce some preliminaries. In Sections 3 and 4 we discuss the long-run competitive ratio criterion: In Section 3 we show that even in the perfect monitoring case, a deterministic optimal policy does not exist. However, we show that there exist an efficient stochastic policy which ensure that the competitive ratio long-run criterion holds with a high probability. In Section 4 we show that such stochastic policies do not exist in the imperfect monitoring case. In Section 5 we prove for both, the perfect and imperfect monitoring cases, that there exist a deterministic efficient optimal policy for the long-run maxmin criterion. In Section 6 we compare our notions of long-run optimality to other criteria appearing in some of the related literature. In particular, we show that our approach to long-run optimality can be interpreted as qualitative, which distinguishes it from previous work in this area.

2. Preliminaries. A *decision problem* is a 3-tuple $D =< A, S, u >$, where A and S are finite sets and u is a real-valued function defined on $A \times S$ with $u(a, s) > 0$ for every $(a, s) \in A \times S$. Elements of A are called *actions* and those of S are called *states*. u is called the *utility function*. The interpretation of the the numerical values $u(a, s)$ is context-dependent [7]. Let n_A denote the number of actions in A, let n_S denote the number of states in S and let $n = \max(n_A, n_S)$.

The above-mentioned setting is a classical static setting for decision-making, where there is uncertainty about the actual state of nature (Luce and Raiffa (1957)). In this paper we deal with a dynamic setup, in which the agent faces the decision problem D over infinite number of stages, $t = 1, 2, \ldots$. As we have explained in the introduction, this setting enables us to capture general dynamic non-Bayesian decision making contexts, where the environment may change its behavior in an arbitrary and unpredicted fashion. As mentioned in the introduction, this is best captured by means of a repeated game against Nature. The state of the environment at each point plays the role of an action taken by Nature in the corresponding game. The agent knows the sets A and S, but he does not know the payoff function u.[8] A dynamic decision problem is therefore represented for the agent by a pair $DD =< A, S >$ of finite sets. At each stage t, Nature chooses a state $s_t \in S$. The agent, who does not know the chosen state, chooses $a_t \in A$, and receives $u(a_t, s_t)$.

[7] See the discussion at Section 6.

[8] All the results of this paper remain unchanged if the agent does not initially know the set S, but rather an upper bound on n.

We distinguish between two informational structures. In the *perfect monitoring* case, the state s_t is revealed to the agent alongside the payoff $u(a_t, s_t)$. In the *imperfect monitoring* case, the states are not revealed to the agent. A generic history available to the agent at stage $t+1$ is denoted by h_t. In the perfect monitoring case, $h_t \in H_t^p = (A \times S \times R_+)^t$, where R_+ denotes the set of positive real numbers. In the imperfect monitoring case, $h_t \in H_t^{imp} = (A \times R_+)^t$. In the particular case $t = 0$ we assume that $H_0^p = H_0^{imp} = \{e\}$ is a singleton containing the empty history e. Let $H^p = \cup_{t=0}^{\infty} H_t^p$ and let $H^{imp} = \cup_{t=0}^{\infty} H_t^{imp}$. The symbol H will be used for both H^p and H^{imp}. A *strategy*[9] for the agent in a dynamic decision problem is a function $F : H \to \Delta(A)$, where $\Delta(A)$ denotes the set of probability measures over A. A strategy F is called *pure* if $F(h_t)$ is a probability measure concentrated on a singleton for every $t \geq 0$.

In Sections 2–4 the strategy recommended to the agent is chosen according to a "long-run" decision criterion which is induced by the *competitive ratio* one-stage decision criterion. This competitive ratio decision criterion, that is described bellow, may be used by an agent who faces the decision problem only once, and who knows the payoff function u as well as the sets A and S. There are other "reasonable" decision criteria that could be used instead. One of them is the *maxmin* decision criterion to be discussed in Section 5, while another is the minmax regret decision criterion (Luce and Raiffa (1957); Milnor (1954)). The latter is a simple variant of the competitive ratio (and can be treated similarly), and therefore will not be treated explicitly in this paper.

For every $s \in S$ let $M(s)$ be the maximal payoff the agent can get when the state is s. That is

$$M(s) = \max_{a \in A} u(a, s).$$

For every $a \in A$ and $s \in S$ define

$$c(a, s) = \frac{M(s)}{u(a, s)}.$$

Denote $c(a) = \max_{s \in S} c(a, s)$, and let

$$CR = \min_{a \in A} c(a) = \min_{a \in A} \left(\max_{s \in S} c(a, s) \right).$$

CR is called the *competitive ratio* of D. Any action a for which $CR = c(a)$ is called a *competitive ratio action*, or in short a CR action. An agent which chooses a CR

[9] Strategy is a decision-theoretic concept. It coincides with the term *policy* used in the control theory literature, and with the term *protocol* used in the distributed systems literature.

action guarantees receiving at least $\frac{1}{CR}$ fraction from what it could have gotten, had it known the state s. That is,$u(a,s) \geq \frac{1}{CR}M(s)$ for every $s \in S$. This agent cannot guarantee a bigger fraction.

In the long-run decision problem, a non-Bayesian agent does not form a prior probability on the way Nature is choosing the states. Nature may choose a fixed sequence of states or, more generally, use a probabilistic strategy G, where $G : (A \times S)_\infty \to \Delta(S)$, and $(A\times S)_\infty = (A\times S)\times(A\times S)\times\dots$. A payoff function u and a pair of probabilistic strategies F, G, where G can depend on u, generate a probability measure $\mu = \mu_{F,G,u}$ over $(A \times S)_\infty$ endowed with the natural measurable structure. Let (A_t, S_t) denote the coordinate random variables in this probability space. Let $X_t = 1$ if $c(A_t, S_t) \leq CR$ and $X_t = 0$ otherwise, and let $N_T = \sum_{t=1}^{T} X_t$.[10] Let $\delta > 0$. A strategy F is δ-optimal if there exists an integer K such that for every payoff function u and every Nature's strategy G

$$Prob_\mu(N_T \geq (1-\delta)T \quad \text{for every } T \geq K) \geq 1-\delta. \tag{2.1}$$

A strategy F is *optimal* if it is δ-optimal for all $\delta > 0$.

The major objective is to find a policy that will enable (2.1) to hold for every dynamic decision problem. Moreover, we wish (2.1) to hold for small enough K. This will be the subject of the following section.

Every sequence of states $z = (s_t)_{t=1}^{\infty}$ can be regarded as a strategy of Nature. In this strategy Nature chooses s_t at stage t, independent of the history. It is important to notice that if F is a strategy for which (2.1) holds for every such stationary strategy of Nature, then F is δ-optimal. To show this, assume that (2.1) holds for every stationary strategy of Nature . Given any payoff function u and any strategy G, this assumption implies that for $\mu = \mu_{F,G,u}$,

$$Prob_\mu(N_T \geq (1-\delta)T \quad \text{for every } T \geq K)|S_1, S_2, \dots) \geq 1-\delta,$$

with probability one. Therefore

$$Prob_\mu(N_T \geq (1-\delta)T \quad \text{for every } T \geq K) \geq 1-\delta.$$

3. Perfect Monitoring.

We now introduce our main theorem.

[10] Note that the function $c(a,s)$ depends on the payoff function u and therefore so do the random variables X_t and N_t.

Theorem 1. *Let $DD =< A, S >$ be a dynamic decision problem with perfect monitoring. Then for every $\delta > 0$ there exists a δ-optimal strategy. Moreover, a δ optimal strategy can be chosen to be efficient in the sense that K (in (2.1)) can be taken to be polynomial in $max(n, \frac{1}{\delta})$.*

Proof. Recall that n_A and n_S denote the number of actions and states respectively, and that $n = \max(n_A, n_S)$. In this proof we assume for simplicity that $n = n_A = n_S$. Only slight modifications are required for removing this assumption. Without loss of generality, $\delta < 1$. We define a strategy F as follows: Let

$$\frac{1}{M} = \frac{\delta}{8}.$$

At each stage $T \geq 1$ we will construct matrices U_T^F, C_T^F and a subset of actions W_T in the following way: Define $U_1^F(a, s) = *$ for each a, s. At each stage $T > 1$, if a_{T-1} has been performed in stage $T-1$, and s_{T-1} has been observed, then update U by replacing the $*$ in the (a_{T-1}, s_{T-1}) entry with $u(a_{T-1}, s_{T-1})$. At each stage $T \geq 1$, if $U_T^F(a, s) = *$, define $C_T^F(a, s) = 1$. If $U_T^F(a, s) \neq *$, $C_T^F(a, s) = \max_{\{b: U_T^F(b,s) \neq *\}} \frac{U_T^F(b,s)}{U_T^F(a,s)}$. Finally W_T is the set of all $a \in A$ at which $\min_{b \in A} \max_{s \in S} C_T^F(b, s)$ is obtained. We refer to elements in W_T as the *good* actions at stage T. Let $(Z_t)_{t \geq 1}$, be a sequence of i.i.d. $\{0, 1\}$ random variables with $Prob(Z_t = 1) = 1 - \frac{1}{M}$. This sequence is generated as part of the strategy, independent of the observed history. At each stage t the agent observes Z_t. If $Z_t = 1$, the agent chooses an action from W_T by randomizing with equal probabilities. If $Z_T = 0$ the agent randomizes with equal probabilities amongst the actions in A. Let u be a given payoff function, and let $(s_t)_{t=1}^{\infty}$ be a given sequence of states. We proceed to show that (2.1) holds with K being the upper integer value of $\alpha = \max(\alpha_1 + 2, \alpha_2 + 2)$, where

$$\alpha_1 = \frac{128}{\delta^2} \ln\left(\frac{256}{\delta^3}\right) \quad \text{and} \quad \alpha_2 = \frac{n^2(n\frac{8}{\delta} + 1) \ln\left(\frac{2n^2}{\delta}\right) + 1}{1 - \frac{3}{4}\delta}.$$

Recall that $X_t = 1$ if $c(A_t, s_t) \leq CR$ and $X_t = 0$ otherwise, and that $N_T = \sum_{t=1}^{T} X_t$. We denote by P_μ the probability measure induced by F, u and the sequence of states on $(A \times \{0, 1\})_\infty$. Let $\varepsilon = \frac{\delta}{8}$. Define

$$B_K = \left\{ \sum_{t=1}^{T} Z_t \geq \left(1 - \frac{1}{M} - \varepsilon\right) T \quad \text{for all } T \geq K \right\}.$$

By Chernoff(1952) (see also Alon et. al. (1992)), for every T,

$$P_\mu\left(\sum_{t=1}^{T} Z_t < (1-\frac{1}{M}-\varepsilon)T\right) \leq e^{-\frac{\varepsilon^2 T}{2}}.$$

Hence,

$$P_\mu(B_K^c) \leq \sum_{T=k}^{\infty} P_\mu\left(\sum_{t=1}^{T} Z_t < (1-\frac{1}{M}-\varepsilon)T\right) \leq \sum_{T=K}^{\infty} e^{-\frac{\varepsilon^2 T}{2}}.$$

Therefore

$$P_\mu(B_K^c) \leq \int_{k-1}^{\infty} e^{-\frac{\varepsilon^2 T}{2}} dT = \frac{2}{\varepsilon^2} e^{-\frac{\varepsilon^2 (K-1)}{2}}.$$

Since $K > \alpha_1$,

$$P_\mu(B_K^c) < \frac{\delta}{2}. \tag{3.1}$$

Define:

$$L_K = \{N_T \geq (1-\delta)T \quad \text{for every } T \geq K\}.$$

In order to prove that F is δ-optimal (i.e., that (2.1) is satisfied), we have to prove that

$$P_\mu(L_K^c) < \delta. \tag{3.2}$$

By (3.1) it suffices to prove that

$$P_\mu(L_K^c | B_K) \leq \frac{\delta}{2}. \tag{3.3}$$

We now define for every $t \geq 1$, $s \in S$ and $a \in A$ six auxiliary random variables, $Y_t, R_t, Y_t^s, R_t^s, Y_t^{s,a}, R_t^{s,a}$. Let $Y_t = 1$ whenever $Z_t = 1$ and $X_t = 0$, and $Y_t = 0$ otherwise. Let

$$R_T = \sum_{t=1}^{T} Y_t.$$

For every $s \in S$ let $Y_t^s = 1$ whenever $Y_t = 1$ and $s_t = s$, and $Y_t^s = 0$ otherwise. Let

$$R_T^s = \sum_{t=1}^{T} Y_t^s.$$

For every $s \in S$ and for every $a \in A$, let $Y_t^{s,a} = 1$ whenever $Y_t^s = 1$ and $A_t = a$, and $Y_t^{s,a} = 0$ otherwise. Let

$$R_T^{s,a} = \sum_{t=1}^{T} Y_t^{s,a}.$$

Let g be the integer value of $(1-\frac{3}{4}\delta)K$. We now show that

$$P_\mu(L_K^c|B_K) \le P_\mu(\exists T \ge K \quad , R_T \ge g|B_K). \tag{3.4}$$

In order to prove (3.4) we show that

$$L_K^c \cap B_K \subseteq \{\exists T \ge K \quad , R_T \ge g\} \cap B_K.$$

Indeed, if w is a path in B_K such that for every $T \ge K$ $R_T < g$, then, at w, for every $T \ge K$,

$$N_T \ge \sum_{1\le t\le T, Z_t=1} X_t \ge V_T - \sum_{t=1}^{T} Y_t,$$

where V_T denotes the number of stages $1 \le t \le T$ for which $Z_t = 1$. Since $w \in B_K$,

$$N_T \ge (1-\frac{1}{M}-\varepsilon)T - R_T > (1-\frac{1}{M}-\varepsilon)T - g$$

for every $T \ge K$. Since $\frac{1}{M} = \varepsilon = \frac{\delta}{8}$ and $g \le (1-\frac{3}{4}\delta)K$, $N_T \ge (1-\delta)T$ for every $T \ge K$. Hence, $w \in L_K$.

(3.4) implies that it suffices to prove that

$$P_\mu(\exists T \ge K \quad , R_T \ge g|B_K) \le \frac{\delta}{2}. \tag{3.5}$$

Therefore it suffices to prove that for every $s \in S$,

$$P_\mu\left(\exists T \ge K, \quad R_T^s \ge \frac{g}{n}|B_K\right) \le \frac{\delta}{2n}.$$

Hence it suffices to prove that for every $s \in S$ and every $a \in A$,

$$\gamma = P_\mu\left(\exists T \ge K, \quad R_T^{s,a} \ge \frac{g}{n^2}|B_K\right) \le \frac{\delta}{2n^2}. \tag{3.6}$$

In order to prove (3.6) , note that if the inequality $R_T^{s,a} \ge \frac{g}{n^2}$ is satisfied at w, then $c(a,s) > CR$ and a is nevertheless considered to be a good action in at least $\frac{g}{n^2}$ stages $1 \le t \le T$ (w.l.o.g. assume that $\frac{g}{n^2}$ is an integer). Let $b \in A$ satisfy $\frac{u(b,s)}{u(a,s)} > CR$. If b is ever played in a stage $\bar{t}$ with $s_{\bar{t}} = s$, then $a \notin W_t$ for all $t \ge \bar{t}$. Therefore

$$\gamma \le P\mu\left(\exists T \ge K, \quad b \text{ is not played in the first } \frac{g}{n^2} \text{ stages at which } s_t = s|B_K\right).$$

Hence

$$\gamma \le \left(1-\frac{1}{nM}\right)^{\frac{g}{n^2}}.$$

As $(1-\frac{1}{x})^{x+1} \leq e^{-x}$ for $x \geq 1$,

$$\gamma \leq e^{-\frac{g}{n^2(nM+1)}} < \frac{\delta}{2n^2}. \quad \square$$

Theorem 1 shows that efficient dynamic non-Bayesian decisions may be obtained by an appropriate stochastic policy. Moreover, it shows that δ-optimality can be obtained in time which is a (low degree) polynomial in $max(n, \frac{1}{\delta})$. An interesting question is whether similar results can be obtained by a pure/deterministic strategy. As the following example shows, deterministic strategies do not suffice for that job.

We give an example in which the agent does not have a δ optimal *pure* strategy.

Example 1.

Let $A = \{a_1, a_2\}$ and $S = \{s_1, s_2\}$. Assume in negation that the agent has a δ optimal pure strategy f.

Consider the following two decision problems whose rows are indexed by the actions and whose columns are indexed by the states:

$$D_1 = \begin{pmatrix} 1 & 30 \\ 10 & 2 \end{pmatrix} \quad D_2 = \begin{pmatrix} 1 & 10 \\ 30 & 2 \end{pmatrix},$$

with the corresponding ratio matrices:

$$C_1 = \begin{pmatrix} 30 & 1 \\ 1 & 5 \end{pmatrix} \quad C_2 = \begin{pmatrix} 10 & 1 \\ 1 & 15 \end{pmatrix},$$

Assume in addition that in both cases Nature uses the strategy g , defined as follows: $g(h_t) = s_i$ if $f(h_t) = a_i$, $i = 1, 2$. That is, for every t, $(a_t, z_t) = (a_1, s_1)$ or $(a_t, z_t) = (a_2, s_2)$. Let $\delta < 0.25$. Let N_T^i denote N_T for decision problem i. Since f is δ-optimal, there exists K such that for every $T \geq K$, $N_T^1 \geq (1-\delta)T$ and $N_T^2 \geq (1-\delta)T$. Note also that the same sequence (a_t, z_t) is generated in both cases. $N_K^1 \geq (1-\delta)K$ implies that (a_1, s_1) is used at more than half of the stages $1, 2, \ldots, K$. On the other hand, $N_K^2 \geq (1-\delta)K$ implies that (a_2, s_2) is used at more than half of the stages $1, 2, \ldots, K$. A contradiction. $\square$

We complete this section with the following result:

Corollary 1. *In every dynamic decision problem with perfect monitoring there exists an optimal strategy.*

Proof. For $m \geq 1$, let F_m be a $\frac{\delta_m}{2}$-optimal strategy, where $(\delta_m)_{m=1}^{\infty}$ is a decreasing sequence with $\lim_{m\to\infty} \delta_m = 0$. Let $(K_m)_{m=1}^{\infty}$ be an increasing sequence of integers such that for every $m \geq 1$

$$Prob\left(N_T \geq (1-\frac{\delta_m}{2})T \quad \text{for every } T \geq K_m\right) \geq 1 - \frac{\delta_m}{2},$$

and

$$K_{m+1} \geq 2\frac{\sum_{j=1}^{m} K_j}{\delta_m}.$$

Let F be the strategy that for $m \geq 1$ utilizes F_m at the stages $K_0+\cdots+K_{m-1}+1 \leq t \leq K_0+\cdots+K_{m-1}+K_m$, where $K_0 = 0$. It can be easily verified that F is optimal. □

4. Imperfect Monitoring.

We proceed to give an example for the imperfect monitoring case, where for sufficiently small $\delta > 0$, the agent does not have a δ optimal strategy.

Example 2 (Non existence of δ-optimal strategies in the imperfect monitoring case).

Let $A = \{a_1, a_2\}$, and $S = \{s_1, s_2, s_3\}$. Let $\delta < \frac{1}{8}$. Assume in negation that there exists a δ optimal strategy F. Consider the following two decision problems whose rows are indexed by the actions and whose columns are indexed by the states:

$$D_1 = \begin{pmatrix} 2a & 2b & 2c \\ a & b & c \end{pmatrix} \quad D_2 = \begin{pmatrix} 2a & 2b & 2c \\ b & c & a \end{pmatrix},$$

where a, b and c are positive numbers satisfying: $a > 4b > 16c$. For $i = 1, 2$, let $C_i = (c_i(a, s))$, $a \in A$ and $s \in S$ be the ratio matrices. That is:

$$C_1 = \begin{pmatrix} 1 & 1 & 1 \\ 2 & 2 & 2 \end{pmatrix} \quad C_2 = \begin{pmatrix} 1 & 1 & \frac{a}{2c} \\ \frac{2a}{b} & \frac{2b}{c} & 1 \end{pmatrix}.$$

Note that a_1 is the unique CR action in D_1 and a_2 is the unique CR action in D_2. Assume that Nature uses strategy G which randomizes at each stage with equal probabilities (of $\frac{1}{3}$) amongst all 3 states. For $i = 1, 2$ let μ_1 and μ_2 be the probability measures induced by F and G on $(A \times S)^\infty$ in the decision problems D_1 and D_2 respectively. Obviously $\mu_1 \neq \mu_2$, but it can be easily verified that for every $i \in \{1, 2\}$, the distribution of the stochastic process $(N_T^i)_{T=1}^\infty$ with respect to μ_j does not depend on j. More precisely.

$$\mu_1(N_t^i \in M_t \quad \text{for all } t) = \mu_2(N_t^i \in M_t \quad \text{for all } t),$$

for every sequence $(M_t)_{t=1}^\infty$ with $M_t \subseteq \{0, 1, \ldots, t\}$ for all $t \geq 1$. As F is δ-optimal, then there exists an integer K such that with a probability at least $1-\delta$ with respect to μ_1, and hence with respect to μ_2, $N_T^1 \geq (1-\delta)T$ for every $T \geq K$. This implies that with a probability of at least $1-\delta$, a_1 is played at least at $1-\delta$ of

the stages up to time T, for all $T \geq K$. Let CR_2 and c_t^2 denote CR and c_t of decision problem 2, respectively. Note that if $A_t = a_1$, then $C_2(A_t, S_t) \leq CR_2$ if and only if $S_t = s_3$. On the other hand, if T is sufficiently large, then in at least one third of the stages, Nature chooses s_3. Hence, at most 2 thirds of the stages $c_t^2 \leq CR_2 = \max(\frac{2a}{b}, \frac{2b}{c})$. Hence, F cannot be optimal. □

5. Safety level. For the sake of comparison we discuss in this section the safety level (known also as maxmin) criterion. Let $D =< A, S, u >$ be a decision problem. Denote

$$v = \max_a \min_s u(a, s)$$

v is called the *safety level* of the agent (or the maxmin value). Every action a for which $u(a, s) \geq v$ for every s is called a safety level action. We consider now the imperfect monitoring model for the dynamic decision problem $< A, S >$. Every sequence of states $z = (z_t)_{t=1}^{\infty}$ with $z_t \in S$ for every $t \geq 1$ and every pure strategy f of the agent induce a sequence of payoffs $(u_t^{z,f})_{t=1}^{\infty}$. Let $N_T^{z,f} = N_T$ denote the number of stages up to stage T at which the agent's payoff exceeds the safety level v. That is,

$$N_T = \#\{1 \leq t \leq T : \quad u(a_t, z_t) \geq v\}.$$

We say that f is *safety level optimal* if for every decision problem and for every sequence of states,

$$\lim_{T\to\infty} \frac{1}{T} N_T = 1,$$

and the convergence holds uniformly w.r.t. all payoff functions u and all sequences of states in S. That is, for every $\delta > 0$ there exists $K = K(\delta)$ such that $N_T^{z,f} \geq (1 - \delta)T$ for every $T \geq K$ for every decision problem $< A, S, u >$ and for every sequence of states z.

Proposition 3. *Every dynamic decision problem possesses a safety level optimal strategy in the imperfect monitoring case, and consequently in the perfect monitoring case. Moreover, such an optimal strategy can be chosen to be strongly efficient in the sense that for every sequence of states there exist at most $n_A - 1$ stages at which the agent receives a payoff smaller than his safety level, where n_A denotes the number of actions.*

Proof. Let $n = n_A$. Define a strategy f as follows: Play each of the actions in A in the first n stages. For every $T \geq n + 1$, let $v_T(a) = \min u(a_t, s_t)$ where the min ranges over all stages $t \leq T - 1$ for which $a_t = a$. Choose a_T to maximize $v_T(a)$

over $a \in A$. It is obvious that for every sequence of states there are at most $n-1$ stages at which $u_t < v$. Hence $N_T \geq T - n$. Thus for $K(\delta) = \frac{n}{\delta}$, $\frac{1}{T}N_T \geq 1 - \delta$ for every $T \geq K(\delta)$. $\square$

6. Qualitative vs. Quantitative. Note that all the notations established in Section 5, as well as Proposition 3 remain unchanged if we assume that the utility function u takes values in a totally pre-ordered set without any group structure. That is our approach to decision making is qualitative (or ordinal). This distinguishes our work from previous work on non-Bayesian repeated games against Nature, which used the probabilistic safety level criterion as a basic solution concept for the one shot game. These works, that include Blackwell (1956), Hannan (1957), Banos (1968), Megiddo (1980), and more recently Auer,Cesa-Bianchi,Freund and Schapire (1996) and Hart and Mas-Colell (1997), used several versions of long-run solution concepts, all based on some optimization of the average over time of the utility values.

This work is to the best of our knowledge the first to introduce an efficient dynamic optimal policy for the basic competitive ratio context. Our study and results in sections 2-4 can be easily adapted to the case of *qualitative* competitive ratio as well. In this approach the utility function takes values in some totally pre-ordered set and, in addition, we assume a regret function which maps $A \times A \times S$ to some totally pre-ordered set as well. Given a pair of actions a and b and a state s, the regret function will determine the qualitative regret of the agent when action a is performed instead of b in state s. The qualitative regret of action a will be the maximal regret of this action over all states. The optimal qualitative competitive ratio will be obtained by the action for which the qualitative regret is minimal. Notice that no arithmetic calculations are needed (or make any sense) for this qualitative version. All that is required from the agent is that it will be able to compare its regret values and outcomes (once obtained) in a qualitative manner. Our results can be easily adapted to the case of qualitative competitive ratio. For ease of exposition, however, we used the quantitative version of this model.

REFERENCES.

Alon, N, Spencer J.H., and Erdos. P. [1992], The Probabilistic Method, *Wiley-Interscience, New York.*

Auer, P., N. Cesa-Bianchi, Y. Freund and R.E. Schapire [1995], Gambling in a Rigged Casino: The Adversial Multi-Armed Bandit Problem, *Proceedings of the*

36th Annual Symposium on Foundations of Computer Science, 322–331.

Aumann, R.J. and M.B. Maschler, [1995], Repeated Games with Incomplete Information, *The MIT Press, 1995.*

Banos, A. [1968], On Pseudo Games, *The Annals of Mathematical Statistics* 39, 1932-1945.

Blackwell, D. [1956], An Analog of the Minmax Theorem for Vector Payoffs, *Pacific Journal of Mathematics 6, 1–8.*

Cassandra, A.R., Kaelbling, L.P., and Littman, M.L. [1994], Acting optimally in partially observable stochastic domain, In *Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94), page 1023–1028, Seattle, Washington, AAAI Press.*

Chernoff, H. [1952] A measure of the asymptotic efficiency for tests of a hypothesis based on the sum of observations, *Annals of Mathematical Statistics, 23, 493-509*

Fudenberg, D. and Tirole, J. [1993], Game Theory, *The MIT Press,* Cambridge.

Fudenberg, D. and Levine, D. [1997], Theory of Learning in Games, *mimeo.*

Hannan, J. [1957], Approximation to Bayes Risk in Repeated Play, in *Contributions to the Theory of Games, vol. III (Annals of Mathematics Studies 39), M. Dresher, A.W. Tucker and P. Wolfe (eds.), 97-139*, Princeton, Princeton University Press.

Harsanyi, J. C. [1967-68], Games with Incomplete Information Played by Bayesian Players,Parts I, II, III *Management Science 14, 159–182.*

Hart, S. and Mas Colell, A. [1997], A Simple Adaptive Procedure Leading to Correlated Equilibrium, Discussion paper 126, Center for Rationality and Interactive Decision Theory, Hebrew University, January 1997.

Kreps, D. [1988], Notes on the Theory of Choice, *Westview Press, London.*

Lovejoy, W.S. [1991], A survey of algorithmic methods for partially observed Markov decision processes, *Annals of Operations Research, 28 (1-4):47-66.*

Luce. R.D and Raiffa, H. [1957] Games and Decisions- Introduction and Critical Survey, *John Wiley and Sons.*

Megiddo, N. [1980], On Repeated Games with Incomplete Information Played by Non-Bayesian Players, *International Journal of Game Theory 9, 157-167.*

Milnor, J. [1954], Games Against Nature, In *R. M. Thrall, C.H. Coombs and R.L. Davis (eds.), Decision Processes, John Wiley & Sons.*

Mertens, J-F., Sorin, S and Zamir ,S. [1995], Repeated Games, Part A, *CORE, DP-9420.*

Papadimitriou, C.H. and Yanakakis, M. [1989], Shortest Paths Without a Map, *Automata, Languages and Programming. 16th International Colloquium Proceedings, pages 610-620.*

Russell, S.J. and Norvig, P. [1995], Artificial Intelligence - A Modern Approach, *Prentice Hall, New Jersey.*

Savage,L. J. [1954], The Foundations of Statistics, *John Wiley and sons, New York.* Revised and enlarged revision, Dover, New York, 1972.

Shapley, L. S. [1953], Stochastic games, *Proceeding of the National Academic of Sciences (USA) 39, 1095–1100.*

Sutton, R.S. [1992], Special Issue on Reinforcement Learning, *Machine Learning 8, 225–227.*

Watkins, C.J. [1989], Models of Delayed Reinforcement, *PhD Thesis, Psychology Department, Cambridge University, Cambridge, United Kingdom.*

Watkins, C.J. and Dayan, P. [1992], Technical Note: Q-Learning, *Machine Learning, Volume 8, No. 3–4, May 1992.*

Wellman, M.P. [1985], Reasoning About Preference Models, *Technical Report MIT/LCS/TR-340, Laboratory for Computer Science, MIT, Cambridge.*

Wellman, M. and Doyle, J. [1992], Modular Utility Representation for Decision-Theoretic Planning. In *Proceedings of the first international conference on AI planning systems*, College Park, Maryland, Morgan KaufmannL.