

Dynamic Non-Bayesian Decision Making in Multi-Agent Systems

Dov Monderer and Moshe Tennenholtz
Faculty of Industrial Engineering and Management
Technion – Israel Institute of Technology
Haifa 32000
Israel

August, 1998

Abstract

We consider a group of several Non-Bayesian agents that can fully coordinate their activities and share their past experience in order to obtain a joint goal in face of uncertainty. The reward obtained by each agent is a function of the environment state but not of the action taken by other agents at the group. The environment state (controlled by Nature) may change arbitrarily, and the reward function is initially unknown. Two basic feedback structures are considered. In one of them – the perfect monitoring case – the agents are able to observe the previous environment state as part of their feedback, while in the other – the imperfect monitoring case – all that is available to the agents are the rewards obtained. Both of these settings refer to partially observable processes, where the current environment state is unknown. Our study refers to the competitive ratio criterion. It is shown that, for the imperfect monitoring case, there exists an efficient stochastic policy that ensures that the competitive ratio is obtained for all agents at almost all stages with an arbitrarily high probability, where efficiency is measured in terms of rate of convergence. It is also shown that if the agents are restricted only to deterministic policies then such a policy does not exist, even in the perfect monitoring case.

1 Introduction

Work in Economics has devoted much attention to coordination and cooperation in multi-agent systems [11]. Agents may compete for resources, and may cooperate in order to obtain desired goals. We will be concerned with fully cooperative agents that have similar goals, captured by similar utility functions, and that are able to fully coordinate their actions and to share their past information. We do not handle the general case of multi-agent learning, in which there are limited communications and partial observability of one agent about the others' observations, actions and rewards¹. Our agents can be viewed as components of a single master-agent [16]. This master-agent is an entity (decision maker) that can perform several actions simultaneously and that gets appropriate feedback for these actions based on the current environment state. For example, if Alice and Bob decide on joint activities in the stock-market then they can coordinate their activities and share their information, acting as a single-entity. The feedback each of them will get depends on her/his actions and on the state of the stock-market (where activities of other interested parties are taken to be part of this state).

Decision making is a central task of artificial agents [18, 24, 23]. 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 simple 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 (problem) 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. Such a problem is best represented by a matrix with columns indexed by the states, rows indexed by the actions and the rewards as entries. When the reward function is not known to the agent we say that the agent has *payoff uncertainty* and we refer to the problem as a *problem with incomplete information*[7]. When modeling a problem with incomplete information one must also describe the underlying assumptions on the knowledge of the agent about the reward function. For example, the agent may know bounds on his rewards,

¹See [20] for a discussion of various approaches to multi-agent learning

or he may know (or partially know) an underlying probabilistic structure. In a dynamic (multistage) decision-making setup the agent faces static decision problems over stages. At each stage the agent selects an action to be performed and the environment selects a state. When we have several agents, then at each stage they need to select a joint action to be performed. The history of actions and states determine the immediate rewards obtained and may effect the next state of the environment.

Most of the work on reinforcement learning in artificial intelligence [9] has adopted the view of an agent operating in a probabilistic Bayesian setting, where the agent's last action and the last state determine the next environment state based on a given probability distribution. Naturally, the learner may not be a-priori familiar with this probability distribution, but the existence of the underlying probabilistic model is a key issue in the system's modeling. However, this assumption is not an ultimate one. In particular, much work in other areas in AI and in economics have dealt with non-probabilistic settings in which the environment changes in an unpredictable manner². When the agent does not know the influence of his choices on the selection of the next state (i.e., he is not certain about the environment strategy), we say that the agent has *strategic uncertainty*. In this paper we use a general model for the representation of agents-environment interactions in which the agents have both payoff and strategic uncertainty. We deal with non-Bayesian agents who face a repeated game with incomplete information against Nature.

In a repeated game against Nature an agent faces the same static decision problem at each stage while the environment state is taken to be an action chosen by external opponents. The decision problem is called a game to stress the fact that the agent's action and the state are independently chosen. 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. As we said, we consider an agent that has both payoff uncertainty and strategic uncertainty. That is, he is a-priori ignorant about the utility function (i.e., the game is of incomplete information) as well as about the state selection strategy of Nature. The agent is non-Bayesian in the sense that he does not assume any probabilistic model concerning nature's strategy

²There are many intermediate cases where it is assumed that the changes are probabilistic with a non-Markovian structure.

and in the sense that he does not assume any probabilistic model concerning the reward function, though he may assume lower and upper bounds³. In a multi-agent dynamic decision problem there are several agents that interact with the environment. Given the assumption that an agent's immediate reward is not affected by the other agents' current actions, the multi-agent dynamic decision problem is in fact a tuple of (single-agent) dynamic decision problems, where the agents can coordinate their actions at each point. Hence, we can model the multi-agent setup using several identical repeated games with incomplete information, that are related to each other due to the fact that Nature can choose only one state at each stage. That is, at each stage Nature chooses the same state for all games. The key feature is that we allow the agents in these games to coordinate their activities and share their information. As it turns out, this leads to powerful techniques and useful results.

We consider an example to illustrate the above-mentioned notions and model. Consider a slot machine with 4 handles. A player can control two handles; he may pull one of them, both of them, or none of them. If there are two players then each of them would control a different pair of handles. The payoff of each agent depends on whether he pulls one, two, or no handle, and in the state of the machine, but not in the action taken by the other player. The way the payoff refers to the state of the environment is initially unknown to the agents. The payment for each player depends on its action and on the state of the environment in a way that is initially unknown to him. The state of the machine changes in an unpredictable way, that may depend on the actions taken by the agents. As we can see the agents are playing in this case in a multi-agent dynamic decision problem with both payoff uncertainty and strategic uncertainty. This situation fits nicely into our general setting.

Given the above general setting we are interested in what the good actions for the agents are. Consider the case of a one stage game against Nature, in which the utility function is known, but the agents cannot observe the

³Repeated 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: [21, 3, 12], and more recently [7, 13], and the evolving literature on learning (e.g., [6]). The incomplete information setup in which the player is ignorant about the game being played was inspired by [8]. See e.g., [2] for a comprehensive survey. Most of the above literature deals with (partially) Bayesian agents.

current environment state when selecting their actions. How should the agents choose their actions? Work on decision making under uncertainty has suggested several approaches⁴. The question of which decision criterion (e.g., expected utility maximization, maximin, competitive ratio, etc.), is the natural one for a particular context is beyond the scope of this paper. A formal method for the analysis of different decision criteria, is using an axiomatic approach [19, 14, 4].

In this paper we deal with the competitive ratio decision criterion (or its additive variant, termed the minmax regret decision criterion [14])⁵. According to this approach an agent would choose an action that minimizes the worst case ratio between the payoff he could have obtained had he known the environment state to the payoff he would actually obtain⁶.

Given a repeated game with incomplete information against Nature, the agents would not be able to choose their one stage optimal action. 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 is obtained in *most* stages. We are interested in particular in efficient policies, where efficiency is measured in terms of rate of convergence. Hence, we are interested in a (joint) policy that if adopted by the agents would guarantee each of them on almost any stage along time, with high probability, at least the payoff guaranteed by an action leading to the competitive ratio. Moreover, they will not have to wait much before they will start getting this type of satisfactory behavior.

We distinguish between two information structures. In one of them, the *imperfect monitoring* case, all the information available to the agents is the rewards they obtain for their actions. In the *perfect monitoring* case, they are also able to observe the environment state at the previous stage. The former type of information structure is appropriate for bandit problem such as the one presented in our slot machine example, while the latter structure is appropriate to situations such as those discussed in the stock-market setting (where the agents are able to see the market activities on previous days). In both settings the current environment state is not observable.

⁴See e.g., [19, 14, 12], and [10].

⁵By [15], the main result in this paper holds for the single agent setup, with the maximin criterion and does not hold for one agent who uses the competitive ratio criterion.

⁶The competitive ratio decision criterion has been found to be most useful in settings such as on-line algorithms (e.g., [17]).

In Section 2 we define the above mentioned setting, model, and concepts. In Sections 3 we discuss the case of perfect monitoring. We give an example in which no deterministic policy (i.e., pure strategy) of the agents guarantees our optimality criterion. This shows that the existence of several agents may not improve upon the behavior of the system in the perfect monitoring case, when compared to work on non-Bayesian decision making in single-agent contexts. In Section 4 we consider the imperfect monitoring case. We show that there exists a stochastic (joint) policy that guarantees our optimality criterion. This is complementary to a negative result obtained in this regard in the single-agent case (see [15]; it shows that the existence of several agents in the system enable to obtain efficient behavior that could not be obtained by only one agent.

2 The Basic Framework and Model

In this section we define our basic setting and terminology. We follow the definitions of the single-agent case (see [15]), while adjusting the setting to the multi-agent context as well.

A (one-shot) *decision problem* (with payoff certainty and strategic uncertainty) is a 3-tuple $D = \langle A, S, u \rangle$, 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 numerical values $u(a, s)$ is context-dependent. 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 [12]. In this paper we deal with a multi-agent dynamic setup. In a dynamic setup an agent faces the decision problem D , without knowing the utility function u , over an infinite number of stages, $t = 1, 2, \dots$. 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 unpredictable 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 .⁷ A *dynamic decision problem* (with payoff uncertainty and strategic uncertainty) is therefore represented for the agent by a pair $DD = \langle A, S \rangle$ of finite sets.

A *multi-agent dynamic decision problem* is a pair $MDD = \langle N, DD \rangle$ where N is a finite set of agents, that are simultaneously facing the dynamic decision problem DD . In this paper we deal with the case $|N| = 2$, and we will refer to the agents as 1 and 2 respectively.⁸ The pair of agents are assumed to act in a coordinated fashion; they coordinate the actions to be taken by them at each point, and they share the feedback they get. The immediate payoff obtained by each agent, however, is independent of the action taken by the other agent and depends on on its action and on the state of Nature.

At each stage t , Nature chooses a state $s_t \in S$. The agents, who do not know the chosen state, choose a joint action $(a_t, b_t) \in A^2$, and receive the rewards $(u(a_t, s_t), u(b_t, s_t))$ respectively. We distinguish between two informational structures. In the *perfect monitoring* case, the state s_t is revealed to the agents alongside the payoff $(u(a_t, s_t), u(b_t, s_t))$. In the *imperfect monitoring* case, the states are not revealed to the agents. We assume that the agents can coordinate their activities, and therefore they have a shared history. A generic history available to the agents at stage $t + 1$ is denoted by h_t . In the imperfect monitoring case, $h_t \in H_t^{imp} = (A \times A \times R_+ \times R_+)^t$, where R_+ denotes the set of positive real numbers. In the perfect monitoring case, $h_t \in H_t^p = (A \times A \times S \times R_+ \times R_+)^t$. 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 $H^{imp} = \cup_{t=0}^{\infty} H_t^{imp}$, $H^p = \cup_{t=0}^{\infty} H_t^p$, and let H stand for either H^{imp} or H^p . A (joint) *strategy*⁹ for the agents in a multi-agent dynamic decision problem is a function $F : H \rightarrow \Delta(A^2)$, where $\Delta(A^2)$ denotes the set of probability measures over A^2 (the joint strategies of the agents). That is, for every $h_t \in H$, $F(h_t) : A^2 \rightarrow [0, 1]$ and $\sum_{(a,b) \in A^2} F(h_t)(a, b) = 1$. In other words, if the agents observe the history h_t then they choose (a_{t+1}, b_{t+1}) by randomizing amongst their joint actions, with the probability $F(h_t)(a, b)$ assigned to

⁷All the results of this paper remain unchanged if the agent does not initially know the set S , but rather an upper bound on n_S .

⁸All of our results remain valid to any fixed number of agent, that is greater than 1.

⁹Strategy 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.

the joint action (a, b) . A strategy F is called *pure* if $F(h_t)$ is a probability measure concentrated on a singleton for every $t \geq 0$.

The strategy recommended to the agents is chosen according to a “long-run” decision criterion which is induced by the *competitive ratio* one-stage decision criterion. The competitive ratio decision criterion, that is described below, 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 [12, 14].

For every $s \in S$ let $M(s)$ be the maximal payoff an 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 $D = \langle A, S, u \rangle$. 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 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 $G : Q \rightarrow \Delta(S)$, and $Q = \cup_{t=0}^{\infty} Q_t = \cup_{t=0}^{\infty} (A^2 \times S)^t$. Nature can be viewed as an additional player that knows the reward function. Its strategy may of course refer to the whole history of actions and states until a given point and may depend on the payoff function.

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 the set of infinite histories $Q_{\infty} = (A^2 \times S)_{\infty}$ endowed with the natural measurable

structure. For an event $B \subseteq Q_\infty$ we will denote the probability of B according to μ by $\mu(B)$ or by $Prob_\mu(B)$. More precisely, the probability measure μ is uniquely defined by its values for finite cylinder sets: Let $A_t : Q_\infty \rightarrow A^2$ and $S_t : Q_\infty \rightarrow S$ be the coordinate random variables which contain the values of the actions and states selected by the agents and the environment in stage t (respectively). That is, $A_t(h) = (a_t, b_t)$ and $S_t(h) = s_t$ for every $h = ((a_1, b_1, s_1), (a_2, b_2, s_2), \dots)$ in Q_∞ . Similarly, define A_t^i to be the projection of A_t on the actions of agent i . Then for every $T \geq 1$ and for every $((a_1, b_1, s_1), \dots, (a_T, b_T, s_T)) \in Q_T$,

$$Prob_\mu((A_t, S_t) = ((a_t, b_t), s_t) \text{ for all } 1 \leq t \leq T) = \prod_{t=1}^T F(\varphi_{t-1})(a_t)G(\psi_{t-1})(s_t),$$

where ψ_0 and φ_0 are the empty histories, and for $2 \leq t \leq T$ we have

$$\psi_{t-1} = ((a_1, b_1, s_1), \dots, (a_{t-1}, b_{t-1}, s_{t-1})),$$

while the definition of φ_{t-1} depends on the monitoring structure. In the perfect monitoring case,

$$\varphi_{t-1} = ((a_1, b_1, s_1, u(a_1, s_1), u(b_1, s_1)), \dots, (a_{t-1}, b_{t-1}, s_{t-1}, u(a_{t-1}, s_{t-1}), u(b_{t-1}, s_{t-1}))).$$

and in the imperfect monitoring case

$$\varphi_{t-1} = ((a_1, b_1, u(a_1, s_1), u(b_1, s_1)), \dots, (a_{t-1}, b_{t-1}, u(a_{t-1}, s_{t-1}), u(b_{t-1}, s_{t-1}))).$$

We now define some auxiliary additional random variables on Q_∞ .

Let $X_t = 1$ if $c(A_t^i, S_t) \leq CR$ for $i = 1$ and $i = 2$, and $X_t = 0$ otherwise, and let $N_T = \sum_{t=1}^T X_t$. 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

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

where $\mu = \mu_{F,G,u}$. A strategy F is *optimal* if it is δ -optimal for all $\delta > 0$.

Roughly speaking, N_T measures the number of stages in which the competitive ratio (or a better value) has been obtained by both agents in the first T iterations. In an δ -optimal strategy there exists a number K , such that if the system runs for $T \geq K$ iterations we can get with high probability

that N_T is close to 1 (i.e., almost all iterations are good ones). In an optimal strategy we guarantee that we can get as close as we wish to the situation where all iterations are good ones, with a probability that is as high as we wish. Notice that we require that the above-mentioned useful property will hold for every payoff function and for every strategy of Nature. This strong requirement is a consequence of the non-Bayesian setup; since we do not have any clue about the reward function or about the strategy selected by Nature (and this strategy may yield arbitrary sequences of states to be reached), the best policy would be to insist on good behavior against any behavior adopted by Nature. Notice however that two other relaxations are introduced here; we require successful behavior in most stages, and that the whole process would be successful only with some (very high) probability.¹⁰

The major objective is to find a policy that will enable (1) to hold for every multi-agent dynamic decision problem and every Nature strategy. Moreover, we wish (1) to hold for small enough K . If K is small then our agents can benefit from obtaining the desired behavior already in an early stage. This will be the subject of the following sections. We complete this section with a useful technical observation. This observation has been made in the case of a single agent [15], and it is true in the multi-agent case as well. We show that a strategy F is δ -optimal if it satisfies the optimality criterion (1) for every reward function and for every stationary strategy of nature, where a stationary strategy is defined by a sequence of states $z = (s_t)_{t=1}^{\infty}$. In this strategy Nature chooses s_t at stage t , independent of the history. Indeed, assume that F is a strategy for which (1) holds for every reward function and for every stationary strategy of Nature, then we show that F is δ -optimal.

Given any payoff function u and any strategy G , δ -optimality with respect to stationary strategies 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

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

The above captures the fact that in our non-Bayesian setting we need to present a joint strategy that will be good for both agents for any sequence of states chosen by Nature, regardless of the way in which this sequence of states has been chosen.

¹⁰These relaxations are in the spirit of PAC-learning [22].

3 Perfect Monitoring

Previous work on single-agent non-Bayesian decision making[15] has shown that optimal behavior can not be obtained using a deterministic policy (i.e., a pure strategy can not obtain our optimality criterion). In this section we show that the introduction of several agents into the system may not enable to overcome the need to consider stochastic policies in order to obtain optimal behavior in our non-Bayesian setup. In the next example we show a dynamic decision problem such that, even in the perfect monitoring case two coordinating agents cannot meet our optimality criterion (for small δ).

Proposition 3.1: *There exists a multi-agent dynamic decision problem and $\delta > 0$, for which there is no δ -optimal strategy.*

Proof:

Let $A = \{a^1, a^2, a^3\}$ and $S = \{s^1, s^2, s^3\}$ and let $0 < \delta < 0.1$. Assume in negation that in the two-agent dynamic decision problem defined by A and S the agents have a δ optimal pure strategy f .

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

$$\begin{array}{c} \underline{D_1} \\ \left(\begin{array}{ccc} 1 & 1 & 20 \\ 1 & 2 & 1 \\ 2 & 1 & 1 \end{array} \right) \end{array}$$

$$\begin{array}{c} \underline{D_2} \\ \left(\begin{array}{ccc} 1 & 1 & 2 \\ 1 & 20 & 1 \\ 2 & 1 & 1 \end{array} \right) \end{array}$$

$$\begin{array}{c} \underline{D_3} \\ \left(\begin{array}{ccc} 1 & 1 & 2 \\ 1 & 2 & 1 \\ 20 & 1 & 1 \end{array} \right) \end{array}$$

with the corresponding ratio matrices:

$$\underline{C_1}$$

$$\begin{pmatrix} 2 & 2 & 1 \\ 2 & 1 & 20 \\ 1 & 2 & 20 \end{pmatrix}$$

$$\underline{C_2}$$

$$\begin{pmatrix} 2 & 20 & 1 \\ 2 & 1 & 2 \\ 1 & 20 & 2 \end{pmatrix}$$

$$\underline{C_3}$$

$$\begin{pmatrix} 20 & 2 & 1 \\ 20 & 1 & 2 \\ 1 & 2 & 2 \end{pmatrix}$$

Assume in addition that in all three cases Nature uses the strategy g , defined as follows:

1. if $f(h_t) = (a^1, a^2)$ then $g(h_t) = s_1$,
2. if $f(h_t) = (a^1, a^3)$ then $g(h_t) = s_2$,
3. if $f(h_t) = (a^2, a^3)$ then $g(h_t) = s_3$,
4. if $f(h_t) = (a^1, a^1)$ then $g(h_t) = s_1$,
5. if $f(h_t) = (a^2, a^2)$ then $g(h_t) = s_3$,
6. if $f(h_t) = (a^3, a^3)$ then $g(h_t) = s_2$.

Nature's behavior will be similar in h_t when the agents select (a_i, a_j) and when they select (a_j, a_i) . Let $\delta < 0.1$. Let N_T^i denote N_T for decision problem i . Since f is δ -optimal, there exists K such that for every $T \geq K$, $N_T^1 \geq (1 - \delta)T$, $N_T^2 \geq (1 - \delta)T$, and $N_T^3 \geq (1 - \delta)T$. Note also that the same sequence $((a_t, b_t, z_t))_{t \geq 1}$ of joint actions and states is generated in all three cases.

Given our decision problems and the competitive ratio matrices, we have:

1. If (a_1, a_2) (or, similarly, (a_2, a_1)) appears in the generated history as a joint action in more than $\frac{1}{9}$ of the joint actions, then in D_3 the optimality criteria will not hold.
2. If (a_1, a_3) (or, similarly, (a_3, a_1)) appears in the generated history as a joint action in more than $\frac{1}{9}$ of the joint actions, then in D_2 the optimality criteria will not hold.
3. If (a_2, a_3) (or, similarly, (a_3, a_2)) appears in the generated history as a joint action in more than $\frac{1}{9}$ of the joint actions, then in D_1 the optimality criteria will not hold.
4. If (a_1, a_1) appears in the generated history as a joint action in more than $\frac{1}{9}$ of the joint actions, then in D_3 our optimality criteria will not hold.
5. If (a_2, a_2) appears in the generated history as a joint action in more than $\frac{1}{9}$ of the joint actions, then in D_1 our optimality criteria will not hold.
6. If (a_3, a_3) appears in the generated history as a joint action in more than $\frac{1}{9}$ of the joint actions, then in D_2 our optimality criteria will not hold.

Combining the above, with the fact that at least one pair of the form (a_i, a_j) should appear at least $\frac{1}{9}$ of the times in the agents' history, we get the desired result.

■

4 Imperfect Monitoring

In this section we show that by having more than one agent we can guarantee our long-range optimality criterion also in the case of imperfect monitoring. This shows that the existence of several agents improves upon the behavior of a single agent in our non-Bayesian setup. In the single-agent setup there is no optimal strategy for the imperfect monitoring case [15]. This will enable to improve upon agents' behavior in setups such as bandit problems.

Our result shows the existence of a δ -optimal strategy. It also guarantees that the desired behavior will be obtained after polynomially many stages. Our result is constructive. We first present the rough idea of the strategy employed in our proof. If the utility function was known to the agents then they could use the competitive ratio action. Since the utility function is initially unknown then the agents will use a greedy strategy, where they select an action that is optimal as far as the competitive ratio is concerned, according to the agents' knowledge at the given point. However, the agents will try from time to time to sample a random action. A major problem is that the agents can not associate payoffs with states (because they do not observe the states) and therefore they can not compute ratios using observations about the utility of a particular action in a particular state. In order to overcome this problem, only one agent will sample non-optimal actions. This will enable to check whether the current candidate for an optimal action is indeed such, without observing what the actual state is. Our strategy chooses a right tradeoff between exploration and exploitation phases in order to yield the desired result. We now introduce our main theorem.

Theorem 4.1: *Let $MDD = \langle N, DD \rangle$ be a multi-agent dynamic decision problem, where $DD = \langle A, S \rangle$ is a dynamic decision problem. Then, for every $\delta > 0$ there exists a δ -optimal strategy. Moreover, the δ -optimal strategy can be chosen to be efficient in the sense that K (in (1)) can be taken to be polynomial in $\max(n, \frac{1}{\delta})$.*

Proof:

Without lose of generality assume $|N| = 2$. 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 $M = \frac{8}{\delta}$. That is,

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

At each stage $T \geq 1$ we will construct a vector $C_T^F \in R^A$, and a subset of the actions W_T in the following way: $C_0^F(a) = 1$ for every action $a \in A$. At each stage $T > 1$, if (a_{T-1}, b_{T-1}) has been performed in stage $T - 1$, and

(q_{T-1}, r_{T-1}) have been observed, then $C_T^F(a_{T-1}) = \max(C_{T-1}^F(a_{T-1}), \frac{r_{T-1}}{q_{T-1}})$, and $C_T^F(b_{T-1}) = \max(C_{T-1}^F(b_{T-1}), \frac{q_{T-1}}{r_{T-1}})$. Finally W_T is the set of all $a \in A$ at which $\min_{b \in A} C_T^F(b)$ is obtained. We refer to elements in W_T as the *temporarily good* actions at stage T . Let $(Z_t)_{t \geq 1}$ be a sequence of i.i.d. $\{0, 1\}$ random variables with $\text{Prob}(Z_t = 1) = 1 - \frac{1}{M}$. This sequence is generated as part of the strategy, independent of the observed history. That is at each stage, before choosing their actions, the agents (or one of them) flip a coin, independently of their past observations. At each stage t the agents observe Z_t . If $Z_t = 1$, both agents choose an action from W_t by randomizing with equal probabilities (where both agents take a similar action). If $Z_t = 0$ then agent 1 behave as before, but agent 2 randomizes with equal probabilities amongst the actions in A . This complete the description of the strategy F . 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 (1) holds with K being the upper integer value of $\alpha = \max(\alpha_1 + 2, \alpha_2 + 2)$, where

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

Recall that $X_t = 1$ if $c(A_t^i, s_t) \leq CR$ for both agents ($i = 1, 2$) and $X_t = 0$ otherwise, and that $N_T = \sum_{t=1}^T X_t$. By a slight change of notation, we denote by $P_\mu = \text{Prob}_\mu$ the probability measure induced by F , u and the sequence of states on $(A^2 \times S \times \{0, 1\})_\infty$ (where $\{0, 1\}$ corresponds to the Z_t values).

Let $\varepsilon = \frac{\delta}{8}$. Define

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

Roughly speaking, B_K captures the cases where temporarily good actions are selected by both agents in most stages. Notice that agent 1 will always select a temporarily good action. We wish first to show that actions that are not temporarily good actions will not be selected too frequently.

Claim 5.1.1:

$$P_\mu(\overline{B_K}) < \frac{\delta}{2}.$$

Proof of Claim 5.1.1:

By [5] (see also [1]), for every T ,

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

Recall that given a set S , \overline{S} denotes the complement of S . Hence,

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

Therefore

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

Since $K > \alpha_1$,

$$P_\mu(\overline{B_K}) < \frac{\delta}{2}. \quad (2)$$

■

Let

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

Roughly speaking, L_K captures the case where competitive ratio actions (or better actions in this regard) are selected by both agents in most stages.

In order to prove that F is δ -optimal (i.e., that (1) is satisfied), we have to prove that

$$P_\mu(\overline{L_K}) < \delta. \quad (3)$$

By (2) it suffices to prove the following claim:

Claim 5.1.2:

$$P_\mu(\overline{L_K} | B_K) \leq \frac{\delta}{2}. \quad (4)$$

Proof of Claim 5.1.2:

We 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 let $Y_t = 0$ 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 let $Y_t^s = 0$ 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 $A_t^1 = a$, and let $Y_t^{s,a} = 0$ otherwise. Notice that $Z_t = 1$ and $A_t^1 = a$ imply that $A_t^2 = a$ in the sequences constructed by our strategy. Let

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

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

$$P_\mu(\overline{L}_K | B_K) \leq P_\mu(\exists T \geq K, R_T \geq g | B_K). \quad (5)$$

In order to prove (5) we show that

$$\overline{L}_K \cap B_K \subseteq \{\exists T \geq K, R_T \geq g\} \cap B_K.$$

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

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

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

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

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

(5) implies that it suffices to prove that

$$P_\mu(\exists T \geq K, R_T \geq g | B_K) \leq \frac{\delta}{2}. \quad (6)$$

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

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

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

$$\gamma = P_\mu \left(\exists T \geq K, \quad R_T^{s,a} \geq \frac{g}{n^2} | B_K \right) \leq \frac{\delta}{2n^2}. \quad (7)$$

In order to prove (7), note that if the inequality $R_T^{s,a} \geq \frac{g}{n^2}$ 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 $1 \leq t \leq T$ (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 by agent 2 in a stage \bar{t} with $s_{\bar{t}} = s$, while agent 1 performs a , then $a \notin W_t$ for all $t \geq \bar{t}$. Therefore

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

Hence

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

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

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

■

■

Theorem 4.1 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 $\max(n, \frac{1}{\delta})$.

For analytical completeness, we end this section by mentioning the existence of an optimal strategy (and not merely a δ -optimal strategy). Such an optimal strategy is obtained by utilizing δ_m -optimal strategies (whose existence was proved in Theorem 5.1) for intervals of stages with sizes that converge to infinity, when $\delta_m \rightarrow 0$. A similar Theorem has been proven for the single-agent case, and therefore we only state the related corollary.

Corollary 4.2: *In every multi-agent dynamic decision problem with perfect monitoring there exists an optimal strategy.*

5 Conclusion and Discussion

This work extends previous work on dynamic non-Bayesian decision-making to the framework of multi-agent systems. We considered the case of agents with identical goals that coordinate their activities and share their information in order to obtain these goals in face of uncertainty. The uncertainty includes both uncertainty about the strategy of Nature and about the utility function. Our aim is to find strategies that optimize a long-range competitive ratio criteria for both agents.

We have shown that the fact we have a multi-agent system plays a significant role in devising optimal behavior. The existence of several agents in the system enables to get optimal behavior in the case of imperfect monitoring. This has been shown to be impossible in the single-agent case. Given the importance of settings with imperfect monitoring in modelling basic problems, such as bandit problems, our results expose a non-trivial use of the existence of several agents in a system.

The above suggests that one may wish to consider the use and potential of several coordinated agents in the context of perfect monitoring. In this case, a single-agent can obtain the related optimal behavior, but a stochastic (non-pure) policy is essential for that. We have shown that the existence of several agents may not change this result, although it has changed the achievements guaranteed in the imperfect monitoring case.

We proved our results for two agents. In the imperfect monitoring case, our result, Theorem 5.1 implies the analogous result for more than two agents. That is, the case $n = 2$ is the hardest case. However, our counter example is not easily generalized to the case $n > 2$. We conjecture that it does hold for every n . That is, given $n > 2$, there exists a decision problem in which n agents do not have pure (non-probabilistic) δ -optimal strategies. Regarding many agents it is also important to note that for a fixed size of a decision problem, if the number of agents exceed (or it is equal to) the number of actions (i.e., rows), then the agents have pure optimal strategies in the imperfect monitoring case. This can be proved similarly to the proof concerning the maxmin criteria in Proposition 5.1 of [15].

There are various limitations to our study and various assumptions that have to be taken into account before applying our ideas. First, the reader should notice that our study is especially useful where the amount of information available to the agents is minimal, or when we wish to consider very

complicated (and in particular non-Markovian) environment behaviors. In addition, the description of a simple decision problem as a one-shot game in strategic form might also be quite problematic; for example, it might be problematic to represent in this case agents' interaction where at each point in time the agent can select among a huge number of actions. On the other hand, our results enable to have a full solution to the related exploration-exploitation problem, while obtaining a polynomial convergence to the desired behavior. This fundamental property is missing from most of the related work in reinforcement learning.

We have already mentioned that our results refer in fact to the case where an agent can have several representatives, rather than to the general multi-agent case. Indeed, we have shown that the ability of a master-agent to perform several actions simultaneously enables it to learn and adapt its behavior efficiently, without observing the exact environment behaviors. This observation is the basic idea behind the main theorem of this paper.

References

- [1] N. Alon, J.H. Spencer, and P. Erdos. *The Probabilistic Method*. Wiley-Interscience, 1992.
- [2] R.J. Aumann and M.B. Maschler. *Repeated Games with Incomplete Information*. The MIT Press, 1995.
- [3] D. Blackwell. An analog of the minimax theorem for vector payoffs. *Pacific Journal of Mathematics*, 6:1–8, 1956.
- [4] R. Brafman and M. Tennenholtz. Axiom Systems for Qualitative Decision Criteria. In *Proceedings of AAAI-97*, 1997.
- [5] H. Chernoff. A measure of the asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Annals of Mathematical Statistics*, 23:493–509, 1952.
- [6] D. Fudenberg and D. Levine. Theory of learning in games. miemo, 1997.
- [7] D. Fudenberg and J. Tirole. *Game Theory*. MIT Press, 1991.

- [8] J.C. Harsanyi. Games with incomplete information played by bayesian players, parts i, ii, iii. *Management Science*, 14:159–182, 1967.
- [9] L.P. Kaelbling, M.L. Littman, and A.W. Moore. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 4:237–258, 1996.
- [10] D. Kreps. *Notes on the Theory of Choice*. Westview press, 1988.
- [11] D. Kreps. *A Course in Microeconomic Theory*. Princeton University Press, 1990.
- [12] R. D. Luce and H. Raiffa. *Games and Decisions- Introduction and Critical Survey*. John Wiley and Sons, 1957.
- [13] J-F. Mertens, S. Sorin, and S. Zamir. Repeated games, part a. CORE, DP-9420, 1995.
- [14] J. Milnor. Games Against Nature. In R. M. Thrall, C.H. Coombs, and R.L. Davis, editors, *Decision Processes*. John Wiley & Sons, 1954.
- [15] D. Monderer and M. Tennenholtz. Dynamic Non-Bayesian Decision-Making. *Journal of Artificial Intelligence Research*, 7:231–248, 1997.
- [16] Y. Moses and M. Tennenholtz. Multi-Entity Models. *Machine Intelligence*, 14:63–88, 1995.
- [17] C.H. Papadimitriou and M. Yannakakis. Shortest Paths Without a Map. In *Automata, Languages and Programming. 16th International Colloquium Proceedings*, pages 610–620, 1989.
- [18] S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 1995.
- [19] L.J. Savage. *The Foundations of Statistics*. Dover Publications, New York, 1972.
- [20] S. Sen. Adaptation and Learning in Multiagent Systems. IJCAI-95 Workshop Program, Working Notes, 1995.
- [21] L.S. Shapley. Stochastic games. *Proceeding of the National Academic of Sciences (USA)*, 39:1095–1100, 1953.

- [22] L. G. Valiant. A theory of the learnable. *Comm. ACM*, 27(11):1134–1142, 1984.
- [23] M. Wellman and J. Doyle. Modular utility representation for decision-theoretic planning. In *Proceedings of the first international conference on AI planning systems*. Morgan Kaufmann, 1992.
- [24] M.P. Wellman. Reasoning about preference models. Technical Report MIT/LCS/TR-340, Laboratory for Computer Science, MIT, 1985.