Asymptotically Efficient Adaptive Strategies in Repeated Games, Part I: Certainty Equivalence Strategies

Nahum Shimkin* and Adam Shwartz†

Revised, March 1994

Abstract

This paper addresses the problem of dynamic decision making in an uncertain and competitive environment. A decision maker (player 1) faces a system about which he has some (parametric) uncertainty, and which is affected also by the actions of other agents. We focus on a worst-case analysis from the viewpoint of player 1, using the simplified model of a repeated matrix game with lack of information on one side, where single-stage rewards are random but announced, and perfect observations are assumed. Certain ideas from the field of stochastic adaptive control are used to formulate performance criteria in a non-Bayesian setting, and to devise appropriate control strategies. The basic performance measure is the *total* reward accumulated by player 1 over all stages played; the purpose of player 1 is to guarantee that his expected total reward will be "close" to what he could guarantee under complete information. The present paper considers adaptive decision strategies of the Certainty Equivalence type, based on a (modified) Maximum Likelihood estimator, and studies their asymptotic (long-term) performance. A sequel paper will be devoted to "asymptotically optimal" strategies.

Key words: repeated matrix games, incomplete information, adaptive control, total reward criterion.

*Rafael Institute, P.O.Box 2250 (#39), Haifa 31021, Israel. E-mail: shimkin@ee.technion.ac.il. This work was performed while this author was with the Department of Electrical Engineering at the Technion – Israel Institute of Technology, and with the Institute for Mathematics and its Applications at the University of Minnesota.

†Department of Electrical Engineering, Technion – Israel Institute of Technology, Haifa 32000, Israel. E-mail: adam@ee.technion.ac.il

1 Introduction

This paper studies certain aspects of dynamic decision making under uncertainty and competition, and extends some ideas from stochastic adaptive control to this setting. Basically, we place ourselves in the position of a decision maker (player 1), facing a dynamic system on which he has incomplete information. The system is influenced also by the actions of other agents (collectively represented here as a single agent, player 2). Taking a worst-case view, we assume that player 2 is doing his best to obstruct us. We also assume that player 1 is non-Bayesian, in the sense that no prior probabilities are prescribed on his initial uncertainty. At the focus of our interest stands the issue of learning, namely the temporal reduction of initial uncertainty, which is made possible by observing the system's response to the players' actions.

We consider the simplest system dynamics, where a fixed single-stage decision problem is repeated in time. The single-stage problem is a parameter-dependent zero-sum matrix game with random rewards, specified by the following elements:

(a) finite action sets $\mathcal{I}$ and $\mathcal{J}$ for player 1 and 2, respectively;
(b) a finite set of parameters Θ;
(c) a finite set $\mathcal{A} \subset \mathbb{R}$ of possible rewards;
(d) for each $(\theta, i, j) \in \Theta \times \mathcal{I} \times \mathcal{J}$, a probability distribution $p_{\theta,i,j}$ over $\mathcal{A}$.

A parameter $\theta_0 \in \Theta$ is first selected (by Nature), followed by the repeated play of the matrix game which corresponds to θ_0. Thus, at each stage $t \geq 1$, player 1 and player 2 choose actions i_t and j_t respectively, and the reward (payoff) a_t to player 1 is determined according to the probability distribution p_{θ_0,i_t,j_t}. Player 1 does not know the true parameter θ_0, but assumes (a worst-case assumption) that player 2 does know it. Perfect observations of actions and rewards are assumed, so that the action of each player at time t may depend on the entire history sequence $\{i_s, j_s, a_s;\ s < t\}$.

The model just described belongs to the general class of repeated games with incomplete information. These games offer a convenient framework for the isolation and examination of various aspects of learning and information in dynamic conflict situations. They have been extensively studied within the classical Bayesian framework for incomplete-information games, typically under the limiting expected average payoff criterion. See, e.g., [20, 24, 5] for surveys of this field.

In this paper we focus on the (long-run) *total* reward, which a refinement of the average criterion. This, together with the non-Bayesian approach, distinguishes the present work from the mainstream of existing research on repeated games with incomplete information. Furthermore, we are concerned only with the case of perfect observations. The approach taken here is closer in spirit to that of non-Bayesian adaptive control ([12, 13, 15]), and is indeed based on certain ideas and methods from that field.

Let us next introduce the performance criterion for player 1. Generally speaking, player 1 wishes to secure a large total reward over all stages played. Assume for the moment that the game proceeds for n stages, with players 1 and 2 using strategies σ and τ, respectively. One may then compute the total (n-stage) expected reward $R_n^{\sigma,\tau}(\theta_0)$, and a reasonable objective for player 1 is to maximize that quantity. Unfortunately, he knows in advance neither player 2's strategy τ nor the true parameter θ_0, so that further specifications are required to make this goal meaningful.

As already mentioned, with respect to player 2 we adopt a worst-case approach, namely we assume that his strategy would be the least favorable one to player 1. (Note however that any strategy of player 2 is subject to the information structure described above; in particular, a player's action cannot depend on the other's action at the same stage, since actions are chosen simultaneously.) This leaves the issue of the unknown parameter. Here a direct worst-case approach is inappropriate, since it fails to bring out fully the learning potential of player 1. The basic approach taken here is to use the "complete information performance" (i.e., the expected total reward that player 1 could guarantee if he knew θ_0) as a reference point for the actual performance, and require that the difference between the two should be small for *all* possible values of θ_0. This approach is well suited to problems where learning and adaptation are key issues, and seems most natural when the complete information performance may indeed be closely approached.

To specify the complete information performance, note that had θ_0 been known, then player 1 could secure at each stage an expected total reward which equals $v(\theta_0)$, the (maximin) value of the matrix game with parameter θ_0. Thus, in n stages player 1 could secure $nv(\theta_0)$. Obviously, in the incomplete information game he can only come close to this level. Define then the *relative loss* (also known as *regret*) for player 1 as the difference between this and the actual expected total reward, namely $L_n^{\sigma,\tau}(\theta_0) := nv(\theta_0) - R_n^{\sigma,\tau}(\theta_0)$. We shall henceforth refer to this quantity simply as *the loss*. Maximization of the loss over all strategies τ of player 2 yields the *worst-case loss*, denoted $L_n^{\sigma}(\theta_0)$. The long run performance of player 1 may now be measured in terms of the asymptotic rate of increase of the worst-case loss (e.g. linear, sub-linear, logarithmic, bounded, etc.). An efficient strategy for player 1 must guarantee a low rate of increase for *all* possible values of the true game parameter.

Implicit in this performance measure are certain assumptions concerning the time horizon n. Formally, this measure is defined for an *infinitely* repeated game. Indeed, the rate of increase of the loss is computed for a *fixed* strategy σ, which must be pre-specified for all stages $t \geq 1$. Practically, it is relevant to the case where the time horizon is very long, and not precisely known to player 1. However, in accordance with our worst-case approach, we do not preclude player 2 from having this information, so that his worst-case strategy is allowed to depend on the time horizon n (cf. eq. (2.3) below).

The total loss criterion is a refinement of the expected *average* reward criterion, and in fact supplies rates of convergence for the latter. Indeed, an "ideal performance" with respect to the average criterion requires only that the worst-case loss would be $o(n)$; a much refined result, namely $O(\log n)$, will be obtained here. We note that if rewards are not observed, then a convergence rate of $O(n^{-1/2})$ of the average reward to the value of the game (in a Bayesian setting) cannot be improved upon in general ([26]); our results clearly imply that with perfect observations this rate is $O(\log n/n)$ at most. Other relevant results concerning the average criterion may be found in [6] and [19], which consider a completely uninformed non-Bayesian player and non-random rewards, and also in [17]. For a related problem of statistical games against nature see [25] and references therein.

What are the strategic problems that confront player 1? Obviously, he may use the observed game history to estimate, in a statistical sense, the true game parameter. We are then confronted with the dual role of control: actions which are good for (statistical) information acquisition may be inefficient in terms of rewards, and vice-versa. On top of this, the effect of player 2 on both aspects should be considered. In particular, the information content of the observations depends also on player 2's actions, and player 1 may not be able to guarantee consistent estimation of the true game parameter. This implies that player 1 cannot isolate the problem of information acquisition (probing), as is possible in comparable adaptive control problems ([13, 15, 16, 1]). Instead, he should rely on inter-relations between information and reward, in trying to guarantee that, no matter what player 2 may do, low information will be compensated by high enough reward as to make this information immaterial. This theme will be dominant in the following analysis.

Our formulation of the performance criterion is closely related to that of [16], where a theory of asymptotic total reward optimality is presented for the statistical Bandit Problem. These results have been extended to various adaptive control models, e.g. [1, 2, 3]. In particular, our game model reduces to that of [1] if player 2 is removed. Asymptotic optimality results in that vein will be presented for the game model in a sequel paper [23].

In the present paper we shall focus on some relatively simple strategies, which are intuitively appealing and easy to implement, study their performance, and identify conditions under which they perform well. These strategies are based on the Certainty Equivalence principle, where a point estimator of the true parameter is substituted in a complete-information optimal control law. First, we consider a strategy which is based on the Maximum Likelihood Estimator (MLE). It turns out that performance is poor in general, in that the loss might increase linearly in time. However, certain conditions are provided which ensure a *bounded* loss (which is of course the best possible "rate of increase"). These results are related to the "closed loop identification problem" which arises in the adaptive control of Markov chains ([13, 15]). Motivated by an idea of [14], we then consider a Certainty Equivalence strategy with a modified estimator, the value-biased MLE. Considerable improvement in performance is obtained: here the loss is $O(\log n)$ at the most and, moreover, a bounded loss is guaranteed under weaker conditions than before.

The paper is organized as follows. The next section describes the model. Section 3 develops some preliminary results required in the analysis. Sections 4 and 5 consider Certainty Equivalence strategies which are based on the MLE and value-biased MLE, respectively, and Section 6 is devoted to the proof of Theorem 5.2. The paper ends with some concluding remarks in Section 7.

Notation: $\mathcal{P}(\mathcal{I})$ denotes the set of probability vectors over the finite action set $\mathcal{I}$, and $\mathcal{P}(\mathcal{J})$ denotes the set of probability vectors over $\mathcal{J}$. An element $x = (x_i)$ of $\mathcal{P}(\mathcal{I})$ is referred to as a "mixed action", and similarly for $y \in \mathcal{P}(\mathcal{J})$. For any $\mathcal{I} \times \mathcal{J}$ matrix $M = \{M(i,j)\}$, let $M(x,y)$ denote the averaged (expected) value of M with respect to the mixed actions x and y, namely

$$M(x,y) = \sum_{i,j} x_i M(i,j) y_j \,. \tag{1.1}$$

The mixed notations $M(i,y)$ and $M(x,j)$ will also be used, with similar interpretation. For positive sequences $\{f_n\}$ and $\{g_n\}$, we write $f_n = o(g_n)$ if $\limsup_n (f_n/g_n) = 0$, and $f_n = O(g_n)$ if $\limsup_n (f_n/g_n) < \infty$. Finally, $\mathbf{1}\{\,\cdot\,\}$ denotes the indicator function.

2 The Model

Let $G(\theta)$ denote the matrix game corresponding to the parameter $\theta \in \Theta$, as described in (a) to (d) above. The repeated game Γ_∞ proceeds as follows. At stage 0, Nature chooses an element $\theta_0 \in \Theta$ ("the true parameter"). This choice is told to player 2, but not to player 1. Then, at each stage $t = 1, 2, \cdots$: player 1 and player 2 simultaneously choose actions $i_t \in \mathcal{I}$ and $j_t \in \mathcal{J}$, respectively. Consequently, the reward $a_t \in \mathcal{A}$ to player 1 is chosen according to the probability distribution $p_{\theta_0, i_t, j_t}(\cdot)$. At the end of each stage, both players observe the actions (i_t, j_t) and the reward a_t.

Perfect recall of past information is assumed. Rewards accumulate to form the total n-stage reward $s_n = \sum_{t=1}^{n} a_t$.

A strategy for each player is a (possibly randomized) rule for choosing his actions at each stage. Since perfect recall is assumed, it follows by the Kuhn-Aumann theorem ([4]) that one can restrict attention (at least as far as the reward sequence distribution is concerned) to *behavioral* strategies, where randomizations are performed independently at each stage. Formally, a (behavioral) strategy for player 1 is a collection of maps $\sigma_t : H_t \to \mathcal{P}(\mathcal{I}), \quad t = 1, 2, \ldots$, where H_t denotes the set of all possible observed histories $h_t = \{i_s, j_s, a_s\}_{s=1}^{t-1}$ up to stage t. Thus $x_t \triangleq \sigma_t(h_t)$ is the randomized action of player 1 at stage t. Strategies of player 2 are defined similarly except that they are allowed to depend explicitly on the true parameter θ_0. A strategy of player 2 is denoted by τ, his strategy set by $\mathcal{T}$, and his randomized action at stage t is denoted y_t. Let $P_{\theta_0}^{\sigma,\tau}$ and $E_{\theta_0}^{\sigma,\tau}$ denote the probability measure and the expectation induced by the triplet (θ_0, σ, τ) on the sample space H_∞.

Let $A_\theta(i, j) = \sum_{a\in\mathcal{A}} a\, p_{\theta,i,j}(a)$ denote the expected reward in the matrix game $G(\theta)$ given actions i and j. The *value* of $G(\theta)$ is given by

$$v(\theta) \triangleq \max_{x\in\mathcal{P}(\mathcal{I})} \min_{y\in\mathcal{P}(\mathcal{J})} A_\theta(x, y) = \min_{y\in\mathcal{P}(\mathcal{J})} \max_{x\in\mathcal{P}(\mathcal{I})} A_\theta(x, y)\,, \tag{2.1}$$

where the notation (1.1) is used.

We now turn to the performance measure for player 1. For each strategy pair (σ, τ), $\theta_0 \in \Theta$ and $n \geq 1$ define the *loss*:

$$L_n^{\sigma,\tau}(\theta_0) = E_{\theta_0}^{\sigma,\tau}\left(nv(\theta_0) - \sum_{t=1}^{n} a_t\right). \tag{2.2}$$

Maximizing over all strategies of player 2 yields the *worst-case loss*:

$$L_n^{\sigma}(\theta_0) = \max_{\tau} L_n^{\sigma,\tau}(\theta_0)\,. \tag{2.3}$$

As motivated in the introduction, the asymptotic rate of the loss will be used as a performance measure for player 1. Thus, we are interested in a strategy σ of player 1 which guarantees a "low" rate (in n) for every possible θ_o.

We define now some additional quantities that will be used in the sequel. The *one-stage loss* is defined as

$$d_\theta(i, j) = v(\theta) - A_\theta(i, j)\,. \tag{2.4}$$

Note that $d_\theta(i, j)$ may be negative. Denote $\hat{D} = \max_{\theta,i,j} d_\theta(i, j)$. The loss may be expressed in terms of the one-stage loss as follows:

$$L_n^{\sigma,\tau}(\theta_0) \;=\; E_{\theta_0}^{\sigma,\tau} \sum_{t=1}^{n} d_{\theta_0}(i_t, j_t) \;=\; E_{\theta_0}^{\sigma,\tau} \sum_{t=1}^{n} d_{\theta_0}(x_t, j_t)\,. \tag{2.5}$$

These relations follow by using the definition (2.4) of d_{θ_0} and applying appropriate conditioning to each term separately.

Let X_θ^* denote the (closed convex) set of optimal (maximin) randomized actions of player 1 in the matrix game $G(\theta)$, and let Y_θ^* denote the set of optimal (minimax) randomized actions of player 2. Let $\mathcal{I}_\theta^* \subset \mathcal{I}$ denote the set of *relevant actions* for player 1 in $G(\theta)$, namely the set of actions which are given positive probability by some optimal randomized action $x^* \in X^*_\theta$. Similarly define $\mathcal{J}^*_\theta$ as the set of relevant actions for player 2. Let $x^*(\theta)$ be an arbitrary point in the relative interior of X^*_θ, which we fix for the rest of the paper. Note that $i \in \mathcal{I}^*_\theta$ if and only if $x^*(\theta)_i > 0$. It follows from Theorems 3.1.2 and 3.1.16 in [21] that $\mathcal{J}^*_\theta$ is exactly the set of actions which minimize the expected reward against $x^*(\theta)$, that is

$$d_\theta(x^*(\theta), j) \stackrel{\triangle}{=} v(\theta) - A_\theta(x^*(\theta), j) = 0 \qquad \text{for } \ j \in \mathcal{J}^*_\theta \, , \tag{2.6}$$

$$d_\theta(x^*(\theta), j) < 0 \qquad \text{for } \ j \notin \mathcal{J}^*_\theta \, . \tag{2.7}$$

Another important quantity is the *information divergence* between $p_{\theta,i,j}$ and $p_{\theta',i,j}$, defined as

$$\hat{I}_{\theta,\theta'}(i,j) = \sum_{a\in\mathcal{A}} p_{\theta,i,j}(a) \log \frac{p_{\theta,i,j}(a)}{p_{\theta',i,j}(a)} \tag{2.8}$$

(where $0 \log 0 \stackrel{\triangle}{=} 0$). The information divergence, also known as the cross-entropy or Kullback-Leibler distance, is a well known measure of statistical distinguishability between probability distributions, and arises naturally as the expected log-likelihood ratio. It is easily verified ([10]) that $\hat{I}_{\theta,\theta'}(i,j) \geq 0$, and $\hat{I}_{\theta,\theta'}(i,j) = 0$ if and only if $p_{\theta,i,j} = p_{\theta',i,j}$. Since $\hat{I}$ may be infinite, it will be convenient to define a "truncated" version:

$$I_{\theta,\theta'}(i,j) = \sum_{a\in A} p_{\theta,i,j}(a) \min\{M_0, \, \log \frac{p_{\theta,i,j}(a)}{p_{\theta',i,j}(a)}\} \, . \tag{2.9}$$

Here $M_o > 0$ is a large enough constant so that $\hat{I}_{\theta,\theta'}(i,j) > 0$ implies $I_{\theta,\theta'}(i,j) > 0$ (such a constant obviously exists since the sets $\Theta, \mathcal{I}, \mathcal{J}, \mathcal{A}$ are finite). It follows that $0 \leq I_{\theta,\theta'}(i,j) \leq M_o$, and

$$I_{\theta,\theta'}(i,j) = 0 \ \text{ if and only if } \ p_{\theta,i,j} = p_{\theta',i,j} \, . \tag{2.10}$$

3 Preliminaries: Controlled I.I.D. Processes

This section develops some results which in essence will be used to bound the deviation of the log-likelihood ratio from its (conditional) mean. These results are derived within a general "controlled i.i.d. process", whose exact relation to the repeated game model is indicated in Lemma 3.1 below.

We consider the following controlled process, which is similar to a one-player version of the repeated game. Let U denote the action space, and Z the state space. For each $u \in U$, let $q(\cdot|u) \in \mathcal{P}(Z)$ be a probability distribution over Z, and let $r : Z \to \mathrm{I\!R}$ be the reward function. At each stage $t \geq 1$, the controller chooses an action $u_t \in U$, the state $z_t \in U$ is randomly chosen according to $q(\cdot|u_t)$, and a one-stage reward $r_t = r(z_t)$ is collected. A control policy π is a sequence of mappings $\pi_t : \, H_t \to \mathcal{P}(U)$, which associate with each observed history sequence $h_t = (u_1, z_1, \cdots, u_{t-1}, z_{t-1})$ a probability distribution over U. (Since the sets U and Z are not necessarily finite, some measurability conditions are required for the above description to induce a well defined stochastic process. Thus, assume that U, Z are measurable sets, $q : U \to \mathcal{P}(Z)$ and $\pi_t : H_n \to \mathcal{P}(U)$ are transition probabilities, where H_t is endowed with the product σ-algebra, and r is a Borel-measurable function.) Let Π denote the set of control policies.

Let R_u denote a random variable distributed as the one-stage reward given action $u \in U$ (i.e. $R_u = r(Z_u)$ where $Z_u \sim q(\cdot|u)$), and let $\overline{R}_u$ denote its expected value: $\overline{R}_u = E(R_u) = \int_Z r(z)q(dz|u)$. Define also the total n-stage reward $S_n = \sum_{t=1}^n r_t$. The following conditions are imposed throughout:

Assumption 3.1

(*i*) The reward function r is bounded: $|r(z)| \leq \hat{R},\ z \in Z$.

(*ii*) $\overline{R}_u \geq 0,\ u \in U$.

(*iii*) There exists a constant $C_o > 0$ such that $E(R_u^2) \leq C_o \overline{R}_u$ for every $u \in U$.

Item (ii) requires the expected reward to be non-negative. The essence of (iii) is that if $\overline{R}_u$ is small, then so is (the second moment of) R_u.

The next lemma and remark summarize the required correspondence between this model and the repeated game model.

Lemma 3.1 *Let $\theta_0, \theta \in \Theta$ be fixed parameters. Consider the following definition:*

$$u_t = (x_t, j_t),\ z_t = (i_t, j_t, a_t),$$
$$r(z) = \min\{M_o,\ \log \frac{p_{\theta_0,i,j}(a)}{p_{\theta,i,j}(a)}\} \quad \forall z \equiv (i, j, a)\,,$$

where M_o is the same as in (2.9), and we arbitrarily define $r(z) = 0$ if $p_{\theta_0,i,j}(a) = 0$. Then Assumption 3.1 is satisfied.

Remark: Under this definition, we have

$$S_n = \sum_{t=1}^n \min\{M_o,\ \log \frac{p_{\theta_0,i,j}(a)}{p_{\theta,i,j}(a)}\} \triangleq \tilde{\Lambda}_n(\theta_0, \theta)\,,$$

(the "truncated log-likelihood ratio"), and $\overline{R}_u = \sum_i x_i I_{\theta_0,\theta}(i,j) = I_{\theta_0,\theta}(x,j)$ for $u = (x,j)$ [cf. (2.9)]. Note that we identify the action u with (x,j), instead of just (i,j). This will enable to apply the results of this section to bound certain expressions that contain $I_{\theta_0,\theta}(x,j)$.

Proof: The reward function $r(z)$ is obviously bounded since $Z = \mathcal{I} \times \mathcal{J} \times \mathcal{A}$ is a finite set. Also, $\overline{R}_u \equiv I_{\theta_0,\theta}(x,j) \geq 0$. To establish the remaining part (iii) of Assumption 3.1, consider first the finite set $U_o = \{(i,j) : i \in \mathcal{I}, j \in \mathcal{J}\}$, taken as a subset of U by embedding pure actions in randomized ones. For every $u = (i,j)$ there exists a positive constant C_u such that $E(R_u^2) \leq C_u \overline{R}_u$. This follows since $E(R_u^2) \leq \hat{R}^2 < \infty$, $\overline{R}_u = I_{\theta_0,\theta}(i,j) \geq 0$, and, by (2.10), $\overline{R}_u = 0$ implies $R_u \equiv 0$. Since U_o is finite, it follows that $E(R_u^2) \leq C_o \overline{R}_u$ for some constant C_o and all $u \in U_o$. But then, for every $u = (x,j) \in U$,

$$E(R_u^2) = \sum_i x_i E(R_{(i,j)}^2) \leq C_o \sum_i x_i \overline{R}_{(i,j)} = C_o \overline{R}_u\,. \qquad \square$$

The next lemma is required in the proof of Lemma 3.3, the latter being the main result of this section.

Lemma 3.2 *There exist constants $\lambda > 0$, $\mu > 0$ such that:*

$$(i) \qquad E(e^{-\lambda R_u}) \leq 1 - \mu \overline{R}_u\,, \quad u \in U \tag{3.1}$$

$$(ii) \qquad \sup_{\pi \in \Pi} P_\pi\{S_n \leq -K\} \leq e^{-\lambda K}\,, \quad \forall K \geq 0\,. \tag{3.2}$$

Proof:

(i) For every $u \in U$ we have $E(e^{-\lambda R_u})\,|_{\lambda=0} = 1$,

$$\frac{d}{d\lambda}\,E(e^{-\lambda R_u})\,|_{\lambda=0} = -E(R_u) = -\overline{R}_u \leq 0 \tag{3.3}$$

and, by Assumptions 3.1(i) and 3.1(iii),

$$\begin{aligned}\frac{d^2}{d\lambda^2}\,E(e^{-\lambda R_u}) &= E(R_u^2\,e^{-\lambda R_u}) \leq E(R_u^2 e^{|R_u|}) \\ &\leq e^{\hat{R}}\,E(R_u^2) \leq (C_o\,e^{\hat{R}})\overline{R}_u, \qquad \forall\,\lambda \in [0,1]\,.\end{aligned} \tag{3.4}$$

Therefore, a second order power-series expansion of $E(e^{-\lambda R_u})$ around $\lambda = 0$ yields

$$E(e^{-\lambda R_u}) \leq 1 - \lambda\overline{R}_u + \frac{\lambda^2}{2}\,(C_o\,e^{\hat{R}})\overline{R}_u, \quad 0 \leq \lambda \leq 1\,, \tag{3.5}$$

and the result follows with $\lambda = (C_o\,e^{\hat{R}})^{-1}$, $\mu = \lambda/2$.

(ii) Let $\lambda > 0$ be the constant for which (3.1) holds. Since $\mathbf{1}\{\beta \leq 0\} \leq e^{-\lambda\beta}$ for every $\beta \in \mathbb{R}$, then

$$P_\pi\{S_n \leq -K\} \leq E_\pi\,e^{-\lambda(S_n+K)} = e^{-\lambda K}\,E_\pi(\prod_{t=1}^{n} e^{-\lambda r_t})\,. \tag{3.6}$$

Now, by standard Dynamic Programming arguments applied to multiplicative cost functionals (e.g. [7, p.66]), it follows that

$$E_\pi(\prod_{t=1}^{n}\,e^{-\lambda r_t}) \leq [\sup_u\,E(e^{-\lambda R_u})]^n \leq 1 \tag{3.7}$$

where the last inequality follows from (3.1). □

Lemma 3.3 *Let $\{\beta_n\}$ be an $o(n)$ positive non-decreasing sequence such that $\beta_n \to \infty$. Let $\overline{S}_n = \sum_{t=1}^n \overline{R}_{u_t}$. Then*

(i) There exists a constant $Q < \infty$ such that: $\sup_\pi E_\pi\left(\sum_{t=1}^{\infty} \overline{R}_{u_t}\;\mathbf{1}\{S_{t-1} \leq 0\}\right) \leq Q\,.$

(ii) $\limsup_{n\to\infty} \frac{1}{\beta_n}\,\sup_\pi E_\pi\left(\sum_{t=1}^{n} \overline{R}_{u_t}\;\mathbf{1}\{S_{t-1} \leq \beta_n\}\right) \leq 1\,.$

(iii) $\sum_{n=1}^{\infty}\,\sup_\pi\,P_\pi\{\overline{S}_n \geq \eta n,\; S_n \leq \beta_n\} < \infty\,, \quad \forall\eta > 0\,.$

(iv) $\sum_{n=1}^{\infty}\,\sup_\pi\,P_\pi\left\{\sum_{t=1}^{n} \overline{R}_{u_t}\;\mathbf{1}\{S_{t-1} \leq \beta_n\} \geq \eta n\right\} < \infty\,, \quad \forall\eta > 0\,.$

(v) *For any $\alpha > 0$, $\epsilon > 0$, there exists a constant $Q = Q(\alpha, \epsilon) < \infty$ such that*

$$E_\pi \sum_{t=1}^{\infty} \overline{R}_{u_t} \, \mathbf{1}\{\overline{S}_{t-1} \geq (1+\epsilon)\beta_t - \alpha, \; S_{t-1} \leq \beta_t\} \leq Q, \quad \forall \pi \in \Pi\,.$$

Proof:

(i) Let $\lambda > 0$, $\mu > 0$ be as in Lemma 3.2(i). Note that $\mathbf{1}\{S_{t-1} \leq 0\} \leq e^{-\lambda S_{t-1}}$, so it suffices to prove that the bound

$$J_1^n \triangleq \sup_\pi E_\pi \left\{ \sum_{t=1}^{n} \overline{R}_{u_t} \, e^{-\lambda S_{t-1}} \right\} \leq Q \tag{3.8}$$

holds for some $Q < \infty$ and every n. Fixing $n > 1$, define for every $1 < m \leq n$:

$$J_m^n = \sup_\pi E_\pi \left\{ \sum_{t=m}^{n} \overline{R}_{u_t} \exp(-\lambda S_m^{t-1}) \right\} \tag{3.9}$$

where $S_m^{t-1} = \sum_{k=m}^{t-1} r_k$ for $t > m$, and $S_m^{m-1} = 0$. Further define $J_{n+1}^n = 0$. Then, for every $1 \leq m \leq n$,

$$\begin{aligned} J_m^n &= \sup_\pi E_\pi \left\{ \overline{R}_{u_m} + e^{-\lambda r_m} \sum_{t=m+1}^{n} \overline{R}_{u_t} \exp(-\lambda S_{m+1}^{t-1}) \right\} \\ &= \sup_{\pi, u} E_\pi \left\{ \overline{R}_u + e^{-\lambda R_u} \sum_{t=m+1}^{n} \overline{R}_{u_t} \exp(-\lambda S_{m+1}^{t-1}) \right\} \\ &= \sup_u \{ \overline{R}_u + E(e^{-\lambda R_u}) J_{m+1}^n \} \end{aligned} \tag{3.10}$$

(which is in essence just the optimality principle of Dynamic Programming). Hence, by Lemma 3.2(i),

$$J_m^n \;\leq\; \sup_u \{\overline{R}_u + (1 - \mu \overline{R}_u) \, J_{m+1}^n\} \;=\; J_{m+1}^n + \sup_u \{\overline{R}_u (1 - \mu \, J_{m+1}^n)\}\,. \tag{3.11}$$

Since $0 \leq \overline{R}_u \leq \hat{R}$, it follows that $J_m^n \leq J_{m+1}^n + \hat{R}$, and also that $J_m^n \leq J_{m+1}^n$ if $J_{m+1}^n \geq \mu^{-1}$. Since $J_{n+1}^n = 0$, this clearly implies that $J_m^n \leq \hat{R} + \mu^{-1}$, and in particular $J_1^n \leq \hat{R} + \mu^{-1}$.

(ii) Let $0 < \epsilon < 1$ be fixed. Then, recalling that $0 \leq \overline{R}_u \leq \hat{R}$,

$$\begin{aligned} J_n &\triangleq \sum_{t=1}^{n} \overline{R}_{u_t} \, \mathbf{1}\{S_{t-1} \leq \beta_n\} \\ &= \sum_{t=1}^{n} \overline{R}_{u_t} \, \mathbf{1}\{S_{t-1} \leq \beta_n, \, S_{t-1} > (1-\epsilon)\overline{S}_{t-1}\} \\ &\quad + \sum_{t=1}^{n} \overline{R}_{u_t} \, \mathbf{1}\{S_{t-1} \leq \beta_n, \, S_{t-1} \leq (1-\epsilon)\overline{S}_{t-1}\} \\ &\leq \sum_{t=1}^{n} \overline{R}_{u_t} \, \mathbf{1}\{\overline{S}_{t-1} < \frac{\beta_n}{1-\epsilon}\} \; + \; \sum_{t=1}^{n} \overline{R}_{u_t} \, \mathbf{1}\{S_{t-1} \leq (1-\epsilon)\overline{S}_{t-1}\} \\ &\leq \frac{\beta_n}{1-\epsilon} + \hat{R} + \sum_{t=1}^{n} \overline{R}_{u_t} \, \mathbf{1}\{S_{t-1} \leq (1-\epsilon)\overline{S}_{t-1}\}\,. \end{aligned} \tag{3.12}$$

(The last inequality follows from the definition of $\overline{S}_t$.) In order to bound the last term, define a modified reward function: $r'(z) = r(z) - (1-\epsilon)\overline{R}_u$. Denoting all quantities related to r' by a prime, it follows that $\overline{R}'_u = \epsilon\overline{R}_u$, and $S'_t = S_t - (1-\epsilon)\overline{S}_t$. Note further that the model assumptions are satisfied for this modified reward function: in particular, r' is bounded, and

$$E(R'_u)^2 \le E(R_u^2) \le C_o\,\overline{R}_u = (\epsilon^{-1}C_o)\overline{R}'_u\,. \tag{3.13}$$

Therefore, by item (i) of the present lemma,

$$E_\pi \sum_{t=1}^n \overline{R}_{u_t}\,\mathbf{1}\{S_{t-1} \le (1-\epsilon)\overline{S}_{t-1}\} \;=\; \epsilon^{-1}E_\pi \sum_{t=1}^n \overline{R}'_{u_t}\,\mathbf{1}\{S'_{t-1} \le 0\} \;\le\; \epsilon^{-1}Q(\epsilon)\,. \tag{3.14}$$

Combining (3.12) and (3.14) gives

$$\limsup_{n\to\infty} \frac{1}{\beta_n}\sup_\pi E_\pi(J_n) \;\le\; \limsup_{n\to\infty}\frac{1}{\beta_n}\left(\frac{\beta_n}{1-\epsilon} + \hat{R} + \epsilon^{-1}Q(\epsilon)\right) \;=\; \frac{1}{1-\epsilon}\,. \tag{3.15}$$

The required result then follows by letting $\epsilon \to 0$.

(iii) Define, as in the proof of (ii), a modified reward function $r'(z) = r(z) - \frac{1}{2}\overline{R}_u$. Then, for every $\pi \in \Pi$ and $\eta > 0$,

$$\begin{aligned}
P_\pi\left\{\overline{S}_n \ge \eta n,\ S_n \le \beta_n\right\} &\le P_\pi\left\{S_n - \frac{1}{2}\,\overline{S}_n \le \beta_n - \frac{1}{2}\,\eta n\right\} \\
&\le P_\pi\left\{S'_n \le \left(\beta_n - \frac{1}{2}\,\eta n\right)\right\} \\
&\le \exp\left(-\lambda'\left(\frac{1}{2}\,\eta n - \beta_n\right)\right) \;\triangleq\; \alpha_n
\end{aligned} \tag{3.16}$$

holds for some $\lambda' > 0$, where the last step follows by Lemma 3.2(ii) applied with the modified reward function. Since $\beta_n = o(n)$, then $\{\alpha_n\}$ is summable and (iii) follows.

(iv) For every $\pi \in \Pi$ and $\eta > 0$,

$$\begin{aligned}
P_\pi\left\{\sum_{t=1}^n \overline{R}_{u_t}\,\mathbf{1}\{S_{t-1} \le \beta_n\} \ge \eta n\right\} &\le P_\pi\left\{\exists m,\ 1 \le m \le n,\ s.t.\ \overline{S}_m \ge \eta n,\ S_{m-1} \le \beta_n\right\} \\
&\le \sum_{m=1}^n P_\pi\left\{\overline{S}_{m-1} \ge \eta n - \hat{R},\ S_{m-1} \le \beta_n\right\} \\
&\le n\alpha_n \exp(\lambda'\hat{R}/2)\,,
\end{aligned} \tag{3.17}$$

where the last inequality follows exactly as in (3.16), with α_n and λ' as defined there. Since the sequence $\{n\alpha_n\}$ is summable, the result follows.

(v) Define, as in the proof of (ii), a modified reward function: $r'(z) = (1+\epsilon)\,r(z) - \overline{R}_u$. Then

$$\begin{aligned}
J_n &\triangleq \sum_{t=1}^n \overline{R}_{u_t}\,\mathbf{1}\{\overline{S}_{t-1} \ge (1+\epsilon)\beta_t - \alpha,\ S_{t-1} \le \beta_t\} \\
&\le \sum_{t=1}^n \overline{R}_{u_t}\,\mathbf{1}\{(1+\epsilon)\,S_{t-1} - \overline{S}_{t-1} \le \alpha\} \\
&= \epsilon^{-1}\sum_{t=1}^n \overline{R}'_{u_t}\,\mathbf{1}\{S'_{t-1} \le \alpha\}\,.
\end{aligned} \tag{3.18}$$

Now $E_\pi J_n$ can be bounded by applying the proof of (i) to the process with this modified reward function. Indeed note that for any $\lambda > 0$,

$$E_\pi J_n \le \epsilon^{-1} e^{\lambda\alpha} E_\pi \sum_{t=1}^{n} \overline{R}'_{u_t} e^{-\lambda S'_{t-1}} .$$

Comparing this expression with (3.8), it follows that for some finite λ' and Q',

$$E_\pi J_n \le \epsilon^{-1} e^{\lambda'\alpha} Q' \triangleq Q < \infty \quad \forall \pi, n .$$

□

4 Certainty Equivalence with the MLE

We consider in this section a simple strategy that is based on the Certainty Equivalence principle and the Maximum Likelihood Estimator (MLE). This means that player 1 computes at each stage the MLE of the unknown parameter θ_o, and then plays the optimal action in the matrix game which corresponds to this estimate. This strategy may give rise to poor performance in general, as indicated at the end of this section. However, we shall provide sufficient conditions, related to the interplay of information and reward in our model, which guarantee a bounded loss even for this simple strategy.

The MLE of θ_0 just prior to stage $t \ge 1$ is given by:

$$\hat{\theta}_t = \arg\max \{\lambda_{t-1}(\theta) : \ \theta \in \Theta\} , \tag{4.1}$$

where

$$\lambda_{t-1}(\theta) = \prod_{s=1}^{t-1} p_{\theta,i_s,j_s}(a_s) \tag{4.2}$$

is the likelihood function. To define $\hat{\theta}_t$ uniquely, we assume that ties in (4.1) are decided according to some fixed ordering of Θ; also, let $\lambda_0(\theta) \triangleq 1$. For every $\theta_0, \theta \in \Theta$, define the log-likelihood ratio:

$$\Lambda_t(\theta_0, \theta) = \log \frac{\lambda_t(\theta_0)}{\lambda_t(\theta)} = \sum_{s=1}^{t} \log \frac{p_{\theta_0,i_s,j_s}(a_s)}{p_{\theta,i_s,j_s}(a_s)} , \tag{4.3}$$

and the "truncated" log-likelihood ratio:

$$\tilde{\Lambda}_t(\theta_0, \theta) = \sum_{s=1}^{t} \min\{M_0, \log \frac{p_{\theta_0,i_s,j_s}(a_s)}{p_{\theta,i_s,j_s}(a_s)}\} , \tag{4.4}$$

where $M_0 > 0$ is the same constant as in (2.9). Note that, by definition of the MLE,

$$\hat{\theta}_t = \theta \quad \Longrightarrow \quad \Lambda_t(\theta_0, \theta) \le 0 . \tag{4.5}$$

The following strategy $\hat{\sigma}$ of player 1 will be considered in this section:

Strategy $\hat{\sigma}$: $x_t = x^*(\hat{\theta}_t)$, where $\hat{\theta}_t$ is the MLE defined in (4.1).

Control policies of this type, namely Certainty Equivalence with the MLE, have been well studied in the context of stochastic adaptive control (e.g. [18, 8, 9, 13, 15]). Performance of these schemes is often hampered by the *closed-loop identification problem*: the prescribed control signals may be inadequate for efficient identification, and poor performance might result. These observations have led to two research directions. The first is to specify appropriate identifiability conditions on the system which ensure "optimal" performance, (see the above-mentioned references on adaptive control). The second is to consider modifications of the basic policy which alleviate the need for such conditions; this will be further discussed in the next section.

We now proceed to formulate an identifiability–type condition which guarantees *bounded* loss for the strategy $\hat{\sigma}$. For each $\theta_0 \in \Theta$, define the following conditions (recall that $I_{\theta_0,\theta}$ is the information divergence defined in (2.9)):

Condition $C_1(\theta_0)$: For every θ and j, $A_{\theta_0}(x^*(\theta), j) < v(\theta_0)$ implies $I_{\theta_0,\theta}(x^*(\theta), j) > 0$.

Condition C_1: $C_1(\theta_0)$ is satisfied for every $\theta_0 \in \Theta$.

Condition C_1 essentially requires low rewards to be "compensated" by the information content of the observed signals. This should hold for all (optimal) actions of the form $x^*(\theta)$, which is just the set of actions which player 1 might employ under the stratey $\hat{\sigma}$. Further disussion of this condition is deferred to the end of the section.

It will be useful to express this condition in terms of the one-stage loss. Recalling the definition of d_θ in (2.4), $C_1(\theta_0)$ reads: For all θ and j, $d_{\theta_0}(x^*(\theta), j) > 0$ implies $I_{\theta_0,\theta}(x^*(\theta), j) > 0$. Since $I_{\theta_0,\theta}$ is non-negative and both Θ and $\mathcal{J}$ are finite sets, this implies (actually equivalent to) that for some $M < \infty$:

$$d_{\theta_0}(x^*(\theta), j) \le M\, I_{\theta_0,\theta}(x^*(\theta), j), \quad \forall\, \theta, j\,. \tag{4.6}$$

We then have the following result.

Theorem 4.1 *Assume that player 1 is using the strategy $\hat{\sigma}$.*

(i) If $C_1(\theta_0)$ is satisfied, then $\limsup_{n\to\infty} L_n^{\hat{\sigma}}(\theta_0) < \infty$.

(ii) Thus, if C_1 is satisfied, then the loss is bounded for every $\theta_0 \in \Theta$.

Proof: The idea of the proof is to upper-bound the loss by the information content of the data (quantified by the information divergence) over the times when the MLE is different from the true parameter (see (4.6), (4.8) below). To bound the latter, we rely on the observation that large information steers the estimator towards the true parameter.

Recalling (2.5), and noting that $x_t = x^*(\hat{\theta}_t)$ under $\hat{\sigma}$, it follows that for every θ_0 and every strategy τ of player 2:

$$L_n^{\hat{\sigma},\tau}(\theta_0) \;=\; E_{\theta_0}^{\hat{\sigma},\tau} \sum_{t=1}^{n} d_{\theta_0}(x^*(\hat{\theta}_t), j_t). \tag{4.7}$$

Consider a fixed θ_0 such that $C_1(\theta_0)$ is satisfied. By optimality of $x^*(\theta_0)$ in $G(\theta_0)$ we have $d_{\theta_0}(x^*(\theta_0), j) \le 0$ for every j. Using this and (4.6) in (4.7) gives (with E standing for $E_{\theta_0}^{\hat{\sigma},\tau}$):

$$L_n^{\hat{\sigma},\tau}(\theta_0) \;=\; E \sum_{t=1}^{n} \sum_{\theta\in\Theta} d_{\theta_0}(x^*(\theta), j_t)\, \mathbf{1}\{\hat{\theta}_t = \theta\}$$

$$
\begin{aligned}
&\le\; E\sum_{t=1}^{n}\sum_{\theta\neq\theta_0} d_{\theta_0}(x^*(\theta),j_t)\,\mathbf{1}\{\hat\theta_t=\theta\}\\
&\le\; M\sum_{\theta\neq\theta_0} E\sum_{t=1}^{n} I_{\theta_0,\theta}(x^*(\theta),j_t)\,\mathbf{1}\{\hat\theta_t=\theta\}\\
&=\; M\sum_{\theta\neq\theta_0} E\sum_{t=1}^{n} I_{\theta_0,\theta}(x_t,j_t)\,\mathbf{1}\{\hat\theta_t=\theta\}\,. \qquad (4.8)
\end{aligned}
$$

Recall that $\Lambda_n(\theta_0,\theta)$ is the log-likelihood ratio (4.3). By (4.5) we have

$$
\mathbf{1}\{\hat\theta_t=\theta\}\le\;\mathbf{1}\{\Lambda_{t-1}(\theta_0,\theta)\le 0\}\,, \qquad (4.9)
$$

so that, noting that $I_{\theta_0,\theta}\ge 0$,

$$
L_n^{\hat\sigma,\tau}(\theta_0)\le M\sum_{\theta\neq\theta_0} E\sum_{t=1}^{\infty} I_{\theta_0,\theta}(x_t,j_t)\,\mathbf{1}\{\Lambda_{t-1}(\theta_0,\theta)\le 0\}\,. \qquad (4.10)
$$

We now proceed to upper-bound the expressions

$$
F^\tau(\theta)\triangleq E_{\theta_0}^{\hat\sigma,\tau}\sum_{t=1}^{\infty} I_{\theta_0,\theta}(x_t,j_t)\,\mathbf{1}\{\Lambda_{t-1}(\theta_0,\theta)\le 0\} \qquad (4.11)
$$

for each $\theta\neq\theta_0$, by employing Lemma 3.3(i) and the correspondence indicated in Lemma 3.1. For that purpose two adjustments are required. First, we have to "replace" $\Lambda_{t-1}(\theta_0,\theta)$ with its truncated version $\tilde\Lambda_{t-1}(\theta_0,\theta)$ (see (4.5)), in order to comply with the required correspondence. Second, we shall have to extend the strategy set of player 2 in order to comply with the "controlled i.i.d. model" of Section 3.

Note first that $\tilde\Lambda_{t-1}(\theta_0,\theta)\le\Lambda_{t-1}(\theta_0,\theta)$ by its definition, so that

$$
F^\tau(\theta)\le\tilde F^\tau(\theta)\triangleq E_{\theta_0}^{\hat\sigma,\tau}\sum_{t=1}^{\infty} I_{\theta_0,\theta}(x_t,j_t)\,\mathbf{1}\{\tilde\Lambda_{t-1}(\theta_0,\theta)\le 0\}\,, \qquad (4.12)
$$

and it suffices to upper-bound $\tilde F^\tau(\theta)$. Now, since the strategy of player 1 is fixed, player 2 can be regarded as a single controller (maximizer) in (4.12). However, since $x_t=x^*(\hat\theta_t)$ depends on the process history, then player 2 is *not* facing a "controlled i.i.d. process". Let us therefore extend the original set $\mathcal{T}$ of strategies available to player 2 by letting him choose at each stage t both j_t, as before, and also $x_t\in\mathcal{P}(\mathcal{I})$. Denote this extended strategy set by Π. Lemma 3.3(i) can now be applied, which gives

$$
F^\tau(\theta)\le\tilde F^\tau(\theta)\le\sup_{\tau\in\mathcal{T}}\tilde F^\tau(\theta)\le\sup_{\pi\in\Pi}\tilde F^\pi(\theta)\le Q(\theta)<\infty,\quad \forall\,\tau\in\mathcal{T}\,. \qquad (4.13)
$$

Combining (4.10), (4.11) and (4.13) yields

$$
L_n^{\hat\sigma,\tau}(\theta_0)\le M\sum_{\theta\neq\theta_0} Q(\theta)<\infty,\quad \forall\,\tau,\,n\,,
$$

and (i) is proved. Since (ii) follows immediately from (i), the proof is complete. $\square$

Discussion: We conclude with a few remarks concerning the results of this section and their implications.

Condition C_1 is not strictly an identifiability condition, since it does *not* guarantee for player 1 the ability to identify (i.e., consistently estimate) the true parameter. For example, player 2 may still have a "non-revealing" action j_0 under which all games are indistinguishable, i.e. $I_{\theta_0,\theta}(x, j_0) = 0$ for all x and θ. Condition C_1 does not preclude such "information hiding", but guarantees that it will be compensated by large enough rewards.

The proposed strategy $\hat{\sigma}$ might perform poorly when condition C_1 is not satisfied. Let θ be a parameter which violates $C_1(\theta_0)$, which means that for some action j' we have

$$(i) \quad A_{\theta_0}(x^*(\theta), j') < v(\theta_0), \quad \text{and} \quad (ii) \quad I_{\theta_0,\theta}(x^*(\theta), j') = 0\,. \tag{4.14}$$

Assume that at some stage t the MLE $\hat{\theta}_t$ equals θ, so that $x_t = x^*(\theta)$ by definition of $\hat{\sigma}$, and that player 2 chooses action j' thereafter. Then the MLE estimator may get "stuck" on θ, since by (ii) the likelihood ratio between θ_0 and θ will not change. At the same time, (i) implies that the reward at each stage will be lower than the value $v(\theta_0)$. This situation may thus persist, leading to an average reward lower than $v(\theta_0)$ (equivalently, to a loss which increases as $O(n)$).

To remedy this problem, the following fact will be crucial: any parameter θ which violates $C_1(\theta_0)$ must have a value lower than that of θ_0, i.e. $v(\theta) < v(\theta_0)$. Indeed, let θ be such that (4.14) is satisfied. Then,

$$v(\theta) \leq A_\theta(x^*(\theta), j') = A_{\theta_0}(x^*(\theta), j') < v(\theta_0)\,, \tag{4.15}$$

where the the equality follows from (4.14)(ii).

To summarize: if condition C_1 is not satisfied, then the MLE may get "stuck" on a wrong parameter θ, while the loss increases linearly. However, this is possible only if θ satisfies $v(\theta) < v(\theta_0)$. This observation holds the key to the improved strategy of the next section.

5 Certainty Equivalence: Value-Biased MLE

In this section we introduce a class of strategies that guarantee a loss of $O(\log n)$ at most. Moreover, bounded loss is guaranteed here under weaker conditions than those of the previous section. These strategies are based on a modified estimator, which takes into account the reward structure of the model. The simple Certainty Equivalence structure is however maintained.

As indicated at the end of the previous section, a basic problem of the MLE-based strategy is that the estimator may adhere to parameters with a lower value than that of the true parameter. To prevent that, a certain bias will be introduced in the estimator in favor of parameters with high value. Naturally, this bias has to be delicate enough so that the identification capability of the estimator will not be destroyed.

The biasing method proposed here relies on the introduction of confidence levels in the estimator. Instead of just the MLE, which is the single parameter that maximizes the likelihood function, consider the set of parameters which *nearly* maximize the likelihood function (to within a prescribed time-varying threshold, or confidence level). We shall refer to this set as the *likely parameters set*. The estimator is then chosen as the member of this set which has the highest value.

The value-biased scheme is closely related to the cost-biased MLE algorithm, introduced in [14] in the context of adaptive control of Markov chains with the average cost criterion. There the bias is introduced by adding a cost-dependent term to the likelihood function; we shall comment on this biasing method at the end of the section. Several adaptive control algorithms have been proposed which incorporate confidence levels in the estimation scheme (e.g. [11, 9, 16, 1]), but not for the purpose of cost related biasing which is crucial here.

Let $\{K_n, n \geq 1\}$ be a sequence of positive numbers, such that

$$\begin{array}{ll} (i) & K_n \uparrow \infty, \quad K_n \geq 1 \\ (ii) & \log K_n = o(n) \\ (iii) & \sum_{n=1}^{\infty} K_n^{-1} < \infty\,. \end{array} \tag{5.1}$$

A specific example, which gives the "lowest" rate in Theorem 5.1 below, is $K_n = n^{1+\epsilon}$ with $\epsilon > 0$.

Let $\hat{\theta}_t$ be the MLE (4.1), and further define the *likely parameters set*:

$$\hat{\Theta}_t = \left\{ \theta \in \Theta : \Lambda_{t-1}(\hat{\theta}_t, \theta) \triangleq \log \frac{\lambda_{t-1}(\hat{\theta}_t)}{\lambda_{t-1}(\theta)} \leq \log K_t \right\}, \tag{5.2}$$

which is the set of parameters which bring the log-likelihood function to within $\log K_t$ of its maximum. The *value-biased maximum likelihood estimator* is given by:

$$\overline{\theta}_t \;=\; \arg\max \left\{ v(\theta) : \theta \in \hat{\Theta}_t \right\}\;; \tag{5.3}$$

if there are several parameters with maximal value, we select among them one with maximal likelihood $\lambda_{t-1}(\theta)$. This is an important rule, which leads to the relation (5.6) below.

The following strategy will be considered in this section.

Strategy $\bar{\sigma}$: $x_t = x^*(\overline{\theta}_t)$, where $\overline{\theta}_t$ is the value-biased MLE (5.3).

Before presenting the main results, we state some basic properties of the proposed estimator. By definition of $\overline{\theta}_t$, the following implications hold (for every $\theta_0 \in \Theta$):

$$v(\overline{\theta}_t) < v(\theta_0) \Longrightarrow \theta_0 \notin \hat{\Theta}_t \;\Longrightarrow\; \frac{\lambda_{t-1}(\hat{\theta}_t)}{\lambda_{t-1}(\theta_0)} > K_t\,. \tag{5.4}$$

Furthermore, since $\overline{\theta}_t \in \hat{\Theta}_t$ and $\Lambda_t(\theta_0, \overline{\theta}_t) \leq \Lambda_t(\hat{\theta}_t, \overline{\theta}_t)$,

$$\Lambda_{t-1}(\theta_0, \overline{\theta}_t) \leq \log K_t\,. \tag{5.5}$$

Finally, by the tie-breaking rule for parameters with equal values,

$$v(\overline{\theta}_t) = v(\theta_0) \Longrightarrow \Lambda_{t-1}(\theta_0, \overline{\theta}_t) \leq 0\,. \tag{5.6}$$

The following lemma, a consequence of (5.4), indicates that the biasing scheme indeed achieves its purpose.

Lemma 5.1

(i) $E_{\theta_0}^{\sigma,\tau} \sum_{t=1}^{\infty} \mathbf{1}\{\theta_0 \notin \hat{\Theta}_t\} \leq Q_1$ *for some finite constant* Q_1 *and all* σ, τ.

(ii) Consequently, $E_{\theta_0}^{\sigma,\tau} \sum_{t=1}^{\infty} \mathbf{1}\left\{v(\overline{\theta}_t) < v(\theta_0)\right\} \leq Q_1$ *for all* σ, τ.

Proof:

(i) Let σ, τ be arbitrary strategies. Then

$$\begin{aligned}
E_{\theta_0}^{\sigma,\tau} \sum_{t=1}^{\infty} \mathbf{1}\{\theta_0 \notin \hat{\Theta}_t\} &= \sum_{t=1}^{\infty} P\{\theta_0 \notin \hat{\Theta}_t\} \\
&= \sum_{t=1}^{\infty} P\{\lambda_{t-1}(\hat{\theta}_t)/\lambda_{t-1}(\theta_0) > K_t\} \\
&\leq \sum_{\theta \neq \theta_0} \sum_{t=1}^{\infty} P\{\lambda_{t-1}(\theta)/\lambda_{t-1}(\theta_0) > K_t\} .
\end{aligned}$$

Now, as is well known the likelihood ratio $\{\lambda_t(\theta)/\lambda_t(\theta_0)\}$ is a positive martingale with expected value 1. It then follows by Markov's inequality that:

$$E_{\theta_0}^{\sigma,\tau} \sum_{t=1}^{\infty} \mathbf{1}\{\theta_0 \notin \hat{\Theta}_t\} \leq \sum_{\theta \neq \theta_0} \sum_{t=1}^{\infty} \frac{1}{K_t} , \tag{5.7}$$

which is finite by the choice (5.1) of $\{K_t\}$.

(ii) From (5.4) it follows that $\mathbf{1}\left\{v(\overline{\theta}_t) < v(\theta_0)\right\} \leq \mathbf{1}\{\theta_0 \in \hat{\Theta}_t\}$, and (ii) is therefore implied by (i). $\square$

Lemma 5.1 implies roughly that the biased estimator $\overline{\theta}_n$ will equal a parameter with a value lower than that of the true parameter not more than a finite number of times. Thus, the effect of such parameters of lower value is no longer significant. Note however that the biasing scheme introduces a new potential problem: Since the estimator is biased towards parameters with higher value, it may favor those over the true parameter, even when the unbiased likelihood function is maximized by the true parameter. The main issue in following analysis will be to bound the loss associated with this effect.

We are now in a position to present the first main result of this section.

Theorem 5.1 *For every* $\theta_0 \in \Theta$ *there exists a constant* $\beta(\theta_0)$ *such that*

$$\limsup_{n\to\infty} \frac{1}{\log K_n} L_n^{\bar{\sigma}}(\theta_0) \leq \beta(\theta_0) . \tag{5.8}$$

Thus, under strategy $\bar{\sigma}$ *the worst-case loss is* $O(\log K_n)$ *at most. In particular, if we choose* $K_n = n^{1+\epsilon}$, *with* $\epsilon > 0$, *then the loss is* $O(\log n)$ *at most.*

The proof of Theorem 5.1 proceeds through some lemmas. We assume in the following that θ_0 is fixed. Define for future reference the the following three sets which are, respectively, the set of parameters with value higher than θ_0, same value as θ_0, and the union of the first two:

$$\begin{aligned} H(\theta_0) &= \{\theta \in \Theta : \ v(\theta) > v(\theta_0)\} && (5.9)\\ S(\theta_0) &= \{\theta \in \Theta : \ \theta \neq \theta_0, \ v(\theta) = v(\theta_0)\} && (5.10)\\ H_0(\theta_0) &= \{\theta \in \Theta : \ \theta \neq \theta_0, \ v(\theta) \geq v(\theta_0)\} \ . && (5.11) \end{aligned}$$

Lemma 5.2 *There exists a constant $M < \infty$ such that $d_{\theta_0}(x^*(\theta), j) \leq M\, I_{\theta_0,\theta}(x^*(\theta), j)$ holds for every j and every $\theta \in \Theta$ which satisfies $v(\theta) \geq v(\theta_0)$.*

Proof: Since Θ and $\mathcal{J}$ are finite sets and $I_{\theta_0,\theta}$ is non-negative, it is enough to show that $v(\theta) \geq v(\theta_0)$ and $I_{\theta_0,\theta}(x^*(\theta), j) = 0$ together imply $d_{\theta_0}(x^*(\theta), j) \leq 0$. Indeed, $I_{\theta_0,\theta}(x^*(\theta), j) = 0$ implies that $A_{\theta_0}(x^*(\theta), j) = A_\theta(x^*(\theta), j)$ (cf. (2.10)), and since $x^*(\theta)$ is optimal in the game matrix A_θ, we get

$$d_{\theta_0}(x^*(\theta), j) \triangleq v(\theta_0) - A_{\theta_0}(x^*(\theta), j) = v(\theta_0) - A_\theta(x^*(\theta), j) \leq v(\theta_0) - v(\theta) \leq 0 \, . \tag{5.12}$$

$\square$

Lemma 5.3 *For every θ with $v(\theta) > v(\theta_0)$:*

$$\limsup_{n\to\infty} \frac{1}{\log K_n} \max_{\tau \in \mathcal{T}} E_{\theta_0}^{\bar{\sigma},\tau} \sum_{t=1}^{n} I_{\theta_0,\theta}(x_t, j_t)\, \mathbf{1}\left\{\Lambda_{t-1}(\theta_0, \theta) \leq \log\, K_n\right\} \leq 1 \, .$$

Proof: Follows from Lemma 3.3(ii), by using exactly the same considerations that were used to bound (4.11). $\square$

Proof of Theorem 5.1: By (2.5) and the definition of $\bar{\sigma}$:

$$L_n^{\bar{\sigma},\tau}(\theta_0) = E_{\theta_0}^{\bar{\sigma},\tau} \sum_{t=1}^{n} d_{\theta_0}(x_t, j_t) = E_{\theta_0}^{\bar{\sigma},\tau} \sum_{t=1}^{n} d_{\theta_0}(x^*(\overline{\theta}_t), j_t) \, . \tag{5.13}$$

Now,

$$\begin{aligned} \sum_{t=1}^{n} d_{\theta_0}(x^*(\overline{\theta}_t), j_t) &= \sum_{t=1}^{n} d_{\theta_0}(x^*(\overline{\theta}_t), j_t) \left[\mathbf{1}\{v(\overline{\theta}_t) < v(\theta_0)\} + \mathbf{1}\{v(\overline{\theta}_t) \geq v(\theta_0)\} \right] \\ &\leq \hat{D} \sum_{t=1}^{n} \mathbf{1}\{v(\overline{\theta}_t) < v(\theta_0)\} + \sum_{\theta \in H_0(\theta_0)} \sum_{t=1}^{n} d_{\theta_0}(x^*(\theta), j_t)\, \mathbf{1}\{\overline{\theta}_t = \theta\} \, , \end{aligned} \tag{5.14}$$

where $\hat{D}$ is an upper-bound on d_{θ_0}, $H_0(\theta_0)$ is defined in (5.11), and $d_{\theta_0}(x^*(\theta_0), j) \leq 0$ was used. We next bound the last term in (5.14). From Lemma 5.2, the control law $x_t = x^*(\overline{\theta}_t)$ and (5.5), it follows that for every $\theta \in H_0(\theta_0)$,

$$\begin{aligned} d_{\theta_0}(x^*(\theta), j_t)\ \mathbf{1}\{\overline{\theta}_t = \theta\} &\leq M\, I_{\theta_0,\theta}(x^*(\theta), j_t)\ \mathbf{1}\{\overline{\theta}_t = \theta\} \\ &= M\, I_{\theta_0,\theta}(x_t, j_t)\ \mathbf{1}\{\overline{\theta}_t = \theta\} \\ &\leq M\, I_{\theta_0,\theta}(x_t, j_t)\ \mathbf{1}\{\Lambda_{t-1}(\theta_0, \theta) \leq \log\, K_t\} \, . \end{aligned} \tag{5.15}$$

Therefore

$$\sum_{t=1}^{n} d_{\theta_0}(x^*(\overline{\theta}_t), j_t) \;\leq\; \hat{D}\sum_{t=1}^{n} \mathbf{1}\{v(\overline{\theta}_t) < v(\theta_0)\} + M \sum_{\theta\in H_0(\theta_0)} \sum_{t=1}^{n} I_{\theta_0,\theta}(x_t, j_t)\, \mathbf{1}\{\Lambda_{t-1}(\theta_0,\theta) \leq \log K_t\}\,. \qquad (5.16)$$

The proof now follows by taking the expectation and applying Lemma 5.1(ii) and Lemma 5.3. □

The proof supports the following heuristic explanation for Theorem 5.1. As already noted, Lemma 5.1 implies that the effect of parameters with lower value than $v(\theta_0)$ on the loss may be ignored. Consider then the effect of parameters with a higher value than $v(\theta_0)$. The biased estimator would prefer such parameters over θ_0 – unless the observations provide sufficient statistical information to overcome the bias. Let θ be a parameter with $v(\theta) > v(\theta_0)$, and assume that $\overline{\theta}_t = \theta$ at some stage, so that $x_t = x^*(\theta)$. Now, the basic information–loss relation of Lemma 5.2 implies that any (positive) loss incurred at that stage is accompanied by a proportional information for discriminating between θ_0 and θ. Thus, if the loss over the times $\{t \leq n : \overline{\theta}_t = \theta\}$ builds up to $O(\log K_n)$, so does the information for discriminating θ_0 and θ, and this information is just sufficient to overcome the bias so that θ is ruled out by the estimator. Consequently, this loss cannot exceed $O(\log K_n)$.

We turn now to formulate conditions under which the strategy $\bar{\sigma}$ guarantees a *bounded* loss. Essentially, an additional information–loss relation will be required to hold under the optimal action $x^*(\theta_0)$; this provides an additional "source of information" for discriminating between θ_0 and parameters with higher value.

Recall that $H(\theta) \triangleq \{\theta' : v(\theta') > v(\theta)\}$. For each $\theta_0 \in \Theta$, define

Condition $C_2(\theta_0)$: For every $\theta \in S(\theta_0)$ (i.e. for every parameter with the same value as θ_o), the following condition $C_3(\theta)$ holds:

$C_3(\theta)$: For each $j \in \mathcal{J}$, either (i) $A_\theta(x^*(\theta), j) > v(\theta)$, or (ii) $\min_{\theta'\in H(\theta)} I_{\theta,\theta'}(x^*(\theta), j) > 0$.

Condition C_2: Condition $C_2(\theta_0)$ holds for every $\theta_0 \in \Theta$.

Note that condition C_2 is equivalent to: $C_3(\theta)$ holds for all θ.

Theorem 5.2 *Assume that player 1 uses strategy $\bar{\sigma}$.*

(i) If $C_2(\theta_0)$ is satisfied, then

$$\limsup_{n\to\infty} L_n^{\bar{\sigma}}(\theta_0) < \infty\,. \qquad (5.17)$$

(ii) Consequently, if C_2 is satisfied, then the worst-case loss is bounded for every $\theta_0 \in \Theta$.

The proof of this result is presented in the next section. Here we compare condition C_2 with condition C_1 of the previous section. It should first be noted that for a given θ_0, conditions $C_1(\theta_0)$ and $C_2(\theta_0)$ are not comparable, since the first pertains to parameters with lower value than that of θ_0, while the second to parameters with higher value. However, it will next be established that the *global* condition C_2 is weaker than C_1, thus implying that strategy $\bar{\sigma}$ guarantees a bounded loss under weaker conditions than those required for $\hat{\sigma}$.

Lemma 5.4 *C_1 implies C_2.*

Proof: We shall prove that if C_2 is not satisfied, then C_1 is not satisfied; more specifically, if $C_2(\theta_0)$ is not satisfied for some θ_0, then $C_1(\theta')$ is not satisfied for some $\theta' \in H(\theta_0)$. Assume then that $C_2(\theta_0)$ is not satisfied. This means that for some $\theta \in S(\theta_0)$, $j \in \mathcal{J}$ and $\theta' \in H(\theta) = H(\theta_0)$, we have (i') $A_\theta(x^*(\theta), j) = v(\theta)$, and (ii') $I_{\theta,\theta'}(x^*(\theta), j) = 0$, which imply that

$$A_{\theta'}(x^*(\theta), j) \;=\; A_\theta(x^*(\theta), j) \;=\; v(\theta) \;<\; v(\theta')\,. \tag{5.18}$$

But (5.18) and (ii') together imply that $C_1(\theta')$ is not satisfied. □

An alternative biasing scheme:

We now consider briefly an alternative biasing method, where the estimator is the maximizer of a biased likelihood function. A similar method was employed in [14].

Let $\{w(\theta) : \theta \in \Theta\}$ be a set of real numbers which are increasing in $v(\theta)$, i.e. $w(\theta) > w(\theta')$ if and only if $v(\theta) > v(\theta')$. Let $\delta = \min\{|w(\theta) - w(\theta')| : w(\theta) \neq w(\theta')\}$ denote the minimal separation between these numbers, and let $\{k_n\}$ be a positive sequence such that (compare (5.1)): *(i)* $k_n \uparrow \infty$, *(ii)* $\log k_n = o(n)$, *(iii)* $\sum_{n=1}^{\infty} k_n^{-1} < \infty$. The value-biased MLE is now defined as

$$\hat{\theta}_t^b = \arg\max_{\theta\in\Theta} \, (k_t)^{w(\theta)} \lambda_{t-1}(\theta) \tag{5.19}$$

where $\lambda_{t-1}(\theta)$ is the likelihood function (4.2). The Certainty Equivalence strategy based on this estimator is $x_t = x^*(\hat{\theta}_t^b)$.

The main results of this section, viz. Theorems 5.1 and 5.2, remain valid under this strategy (with $\log K_n$ replaced by $\log k_n$). This is easily verified by noting that this estimator satisfies properties similar to (5.4)–(5.5), which are the key properties of the estimator $\overline{\theta}_t$.

We note, however, that the two biasing schemes are not completely equivalent. The estimator $\overline{\theta}_t$ effectively provides a uniform bias to all the parameters in $H(\theta_0)$, while in the estimator $\hat{\theta}_t^b$ the bias necessarily increases with $v(\theta)$. The uniform biasing property will prove to be of critical importance in [23], where an "asymptotically optimal" strategy is constructed based on the estimator $\overline{\theta}_t$.

6 Proof of Theorem 5.2

We now present the proof of Theorem 5.2. First we summarize the basic information–loss relations required here, in addition to those of Lemma 5.2.

Lemma 6.1 *There exist positive constants M, δ such that, for every $\theta_0 \in \Theta$ and $j \in \mathcal{J}$:*

(i) $d_{\theta_0}(x^*(\theta), j) \leq -\delta + M I_{\theta_0,\theta}(x^*(\theta), j)$ *for every* $\theta \in H(\theta_0) := \{\theta : v(\theta) > v(\theta_0)\}$.

(ii) If $C_2(\theta_0)$ *holds then* $d_{\theta_0}(x^*(\theta_0), j) \leq -\delta + M \min_{\theta'\in H(\theta_0)} I_{\theta_0,\theta'}(x^*(\theta_0), j)$.

(iii) For every θ, if $v(\theta)$ equals $v(\theta_0)$ and $C_2(\theta)$ is satisfied, then

$$d_{\theta_0}(x^*(\theta), j) \;\leq\; -\delta + M I_{\theta_0,\theta}(x^*(\theta), j) + M \min_{\theta'\in H(\theta_0)} I_{\theta_0,\theta'}(x^*(\theta), j)\,.$$

Proof: Since $I_{\theta_0,\theta_0}(x,j) \equiv 0$, then (ii) is a special case of (iii) for $\theta = \theta_0$. It is therefore only required to prove (i) and (iii). Since Θ and $\mathcal{J}$ are finite sets and $I_{\theta_0,\theta}(x,j) \geq 0$ for all θ, x, j, it is sufficient to establish the following claims (i') and (iii'):
(i') For every $\theta \in H(\theta_0)$, $I_{\theta_0,\theta}(x^*(\theta),j) = 0$ implies $d_{\theta_0}(x^*((\theta),j) < 0$.
(iii') If $v(\theta) = v(\theta_0)$ and $C_2(\theta)$ holds, then $I_{\theta_0,\theta}(x^*(\theta),j) = \min_{\theta' \in H(\theta_0)} I_{\theta_0,\theta'}(x^*(\theta),j) = 0$ implies $d_{\theta_0}(x^*(\theta),j) < 0$.

Claim (i') follows exactly as in the proof of Lemma 5.2, where in the last line the strict inequality $v(\theta_0) - v(\theta) < 0$ may now be used. To establish (iii'), we assume that the assertions there are satisfied, and show that $d_{\theta_0}(x^*(\theta),j) < 0$. Noting (2.10), $I_{\theta_0,\theta}(x^*(\theta),j) = 0$ implies that $I_{\theta_0,\theta'}(x^*(\theta),j) = I_{\theta,\theta'}(x^*(\theta),j)$ for all θ', and in particular for all $\theta' \in H(\theta_0)$. Note that $H(\theta) = H(\theta_0)$ since $v(\theta) = v(\theta_0)$. Therefore

$$\min_{\theta' \in H(\theta)} I_{\theta,\theta'}(x^*(\theta),j) \;=\; \min_{\theta' \in H(\theta_0)} I_{\theta_0,\theta'}(x^*(\theta),j) \;=\; 0\,. \tag{6.1}$$

By $C_2(\theta)$ this implies that $A_\theta(x^*(\theta),j) > v(\theta)$, so that

$$d_{\theta_0}(x^*(\theta),j) \stackrel{\triangle}{=} v(\theta_0) - A_{\theta_0}(x^*(\theta),j) = v(\theta) - A_\theta(x^*(\theta),j) < 0\,. \tag{6.2}$$

□

Remark: Lemma 6.1 reflects the following relations between one-stage loss and information. Item (i) implies that for $x_n = x^*(\theta)$ with $\theta \in H(\theta_0)$, player 1 obtains either positive $I_{\theta_0,\theta}$-information (i.e., information for discriminating between θ_0 and θ), or a negative loss (i.e. expected reward higher than $v(\theta_0)$). Item (ii) means that for $x_n = x^*(\theta_0)$, player 1 obtains either positive $I_{\theta_0,\theta}$-information for *every* $\theta \in H(\theta_0)$, or a negative loss. Item (iii) may be interpreted similarly.

Consider henceforth a fixed $\theta_0 \in \Theta$ and a fixed strategy τ of player 2. Let $H(\theta_0)$, $H_0(\theta_0)$ be defined as in (5.9), (5.11). Recall from (5.13) and (5.14) that

$$L_n^{\bar\sigma,\tau}(\theta_0) \;\leq\; \hat{D} E_{\theta_0}^{\bar\sigma,\tau} \sum_{t=1}^{n} \mathbf{1}\{v(\overline{\theta}_t) < v(\theta_0)\} \;+\; E_{\theta_0}^{\bar\sigma,\tau} \sum_{t=1}^{n} d_{\theta_0}(x^*(\overline{\theta}_t), j_t)\, \mathbf{1}\{v(\overline{\theta}_t) \geq v(\theta_0)\}\,. \tag{6.3}$$

Since the first term on the right-hand-side is bounded by Lemma 5.1(ii), it remains to bound the last term. To this end, define for every $n \geq 1$

$$\ell_n \;=\; \sum_{t=1}^{n} d_{\theta_0}(x^*(\overline{\theta}_t), j_t)\, \mathbf{1}\{v(\overline{\theta}_t) \geq v(\theta_0)\}\,, \tag{6.4}$$

$$\Delta\ell_n \;=\; \ell_n - \ell_{n-1} \;=\; d_{\theta_0}(x^*(\overline{\theta}_n), j_n)\, \mathbf{1}\{v(\overline{\theta}_n) \geq v(\theta_0)\}\,, \tag{6.5}$$

with $\ell_0 \stackrel{\triangle}{=} 0$. The required upper-bound is established in the following lemmas, where the basic idea is that (the expected value of) ℓ_n cannot increase "too much" over those time instants when ℓ_n is positive.

Lemma 6.2
(i) $\ell_n \leq \hat{D} + \sum_{t=1}^{n-1} (\Delta\ell_{t+1})^+\, \mathbf{1}\{\ell_t \geq 0\}$ *for every* $n \geq 1$.
(ii) Consequently, there exists a finite constant Q_3 *such that*

$$E_{\theta_0}^{\bar\sigma,\tau} \ell_n \;\leq\; Q_3 + \hat{D} \sum_{t=1}^{\infty} P_{\theta_0}^{\bar\sigma,\tau}\{\ell_t \geq 0, \overline{\theta}_{t+1} \in H(\theta_0)\}\,.$$

Proof:
(i) Fix $n \geq 1$, and let $m_0 = \max\{0 \leq t \leq n-1 : \ell_t \leq 0\}$. Note that $\Delta \ell_t \leq \hat{D}$. Then

$$\begin{aligned}
\ell_n \ &\leq \ \ell_n - \ell_{m_0} \ \leq \ \sum_{t=m_0}^{n-1} (\Delta \ell_{t+1})^+ \\
&\leq \ \hat{D} + \sum_{t=m_0+1}^{n-1} (\Delta \ell_{t+1})^+ \ = \ \hat{D} + \sum_{t=m_0+1}^{n-1} (\Delta \ell_{t+1})^+ \, \mathbf{1}\{\ell_t > 0\} \\
&\leq \ \hat{D} + \sum_{t=1}^{n-1} (\Delta \ell_{t+1})^+ \, \mathbf{1}\{\ell_t \geq 0\} \, . \qquad (6.6)
\end{aligned}$$

(ii) Recalling the definitions in (5.9) and (5.11) of $H(\theta_0)$ and $S(\theta_0)$, (6.5) may be rewritten as

$$\Delta \ell_t = d_{\theta_0}(x^*(\overline{\theta}_t), j_t) \, \mathbf{1}\{\overline{\theta}_t \in \{\theta_0\} \cup H(\theta_0) \cup S(\theta_0)\} \, . \qquad (6.7)$$

Noting that $d_{\theta_0}(x^*(\theta_0), j) \leq 0$ and $d_{\theta_0}(x^*(\theta), j) \leq \hat{D}$ for $\theta \in H(\theta_0)$, and using Lemma 5.2 for $\theta \in S(\theta_0)$, we get

$$(\Delta \ell_t)^+ \ \leq \ \hat{D} \, \mathbf{1}\{\overline{\theta}_t \in H(\theta_0)\} + M \sum_{\theta \in S(\theta_0)} I_{\theta_0,\theta}(x^*(\theta), j_t) \, \mathbf{1}\{\overline{\theta}_t = \theta\} \, . \qquad (6.8)$$

Therefore, by (i),

$$\ell_n \ \leq \ \hat{D} + \hat{D} \sum_{t=1}^{n-1} \mathbf{1}\{\ell_t \geq 0, \, \overline{\theta}_{t+1} \in H(\theta_0)\} \ + \ M \sum_{\theta \in S(\theta_0)} \sum_{t=1}^{n-1} I_{\theta_0,\theta}(x^*(\theta), j_{t+1}) \, \mathbf{1}\{\overline{\theta}_{t+1} = \theta\} \, . \qquad (6.9)$$

Extending the summations to $+\infty$ and taking expectation gives

$$\begin{aligned}
E_{\theta_0}^{\bar{\sigma},\tau} \ell_n \ &\leq \ \hat{D} + \hat{D} \sum_{t=1}^{\infty} P_{\theta_0}^{\bar{\sigma},\tau}\{\ell_t \geq 0, \, \overline{\theta}_{t+1} \in H(\theta_0)\} \\
&+ \ M \sum_{\theta \in S(\theta_0)} E_{\theta_0}^{\bar{\sigma},\tau} \sum_{t=1}^{\infty} I_{\theta_0,\theta}(x^*(\theta), j_{t+1}) \, \mathbf{1}\{\overline{\theta}_{t+1} = \theta\} \, . \qquad (6.10)
\end{aligned}$$

It remains to bound the last term. Let $\theta \in S(\theta_0)$. Since $v(\theta) = v(\theta_0)$, it follows from (5.6) that

$$E_{\theta_0}^{\bar{\sigma},\tau} \sum_{t=1}^{\infty} I_{\theta_0,\theta}(x^*(\theta), j_t) \, \mathbf{1}\{\overline{\theta}_t = \theta\} \leq E_{\theta_0}^{\bar{\sigma},\tau} \sum_{t=1}^{\infty} I_{\theta_0,\theta}(x_t, j_t) \, \mathbf{1}\{\Lambda_{t-1}(\theta_0, \theta) \leq 0\} \, . \qquad (6.11)$$

The latter term can now be bounded exactly in the same way that $F^\tau(\theta)$ of (4.11) was bounded. □

Lemma 6.3 *Assume that $C_2(\theta)$ is satisfied for every $\theta \in \Theta$ such that $v(\theta) = v(\theta_0)$. Assume that player 1 employs strategy $\bar{\sigma}$, and player 2 is using any strategy τ. Then there exists a constant $\eta > 0$ such that, for every $n \geq 1$: if $\ell_n > 0$, then at least one of the following events $\Omega_1(n)$–$\Omega_3(n)$ holds:*

$$\Omega_1(n) : \qquad \sum_{t=1}^{n} \mathbf{1}\{v(\overline{\theta}_t) < v(\theta_0)\} \geq \eta n \, .$$

$$\Omega_2(n) : \qquad \sum_{t=1}^{n} I_{\theta_0,\theta}(x_t, j_t) \, \mathbf{1}\{\overline{\theta}_t = \theta\} \geq \eta n \ \textit{for some} \ \theta \in H_0(\theta_0) \, .$$

$$\Omega_3(n) : \qquad \min_{\theta \in H(\theta_0)} \sum_{t=1}^{n} I_{\theta_0,\theta}(x_t, j_t) \geq \eta n \, .$$

Proof: Similarly to (6.7), we have

$$\ell_n = \sum_{t=1}^{n} \sum_{\overline{\theta}\in\Theta} d_{\theta_0}(x^*(\overline{\theta}), j_t)\, \mathbf{1}\left\{\overline{\theta}_t = \overline{\theta} \in \{\theta_0\} \cup S(\theta_0) \cup H(\theta_0)\right\} . \tag{6.12}$$

Now, using the appropriate bounds from Lemma 6.1 for each $\overline{\theta}$ in the above sum gives

$$\begin{aligned}
\ell_n \;\; &\leq \;\; \sum_{t=1}^{n} \{ \, [-\delta + M \min_{\theta\in H(\theta_0)} I_{\theta_0,\theta}(x^*(\theta_0), j_t)\,]\; \mathbf{1}\{\overline{\theta}_t = \theta_0\} \\
&\quad + \sum_{\theta\in S(\theta_0)} [-\delta + M I_{\theta_0,\theta}(x^*(\theta), j_t) + M \min_{\theta'\in H(\theta_0)} I_{\theta_0,\theta'}(x^*(\theta), j_t)\,]\; \mathbf{1}\{\overline{\theta}_t = \theta\} \\
&\quad + \sum_{\theta'\in H(\theta_0)} [-\delta + M I_{\theta_0,\theta'}(x^*(\theta'), j_t)\,]\; \mathbf{1}\{\overline{\theta}_t = \theta'\} \; \} \\
&= \;\; -\delta \sum_{t=1}^{n} \mathbf{1}\{v(\overline{\theta}_t) \geq v(\theta_0)\} \; + \; M \sum_{t=1}^{n} \sum_{\theta\in H_0(\theta_0)} I_{\theta_0,\theta}(x_t, j_t)\, \mathbf{1}\{\overline{\theta}_t = \theta\} \\
&\quad + M \sum_{t=1}^{n} \min_{\theta\in H_0(\theta_0)} I_{\theta_0,\theta}(x_t, j_t)\, \mathbf{1}\{\overline{\theta}_t \in \{\theta_0\} \cup S(\theta_0)\} \\
&\leq \;\; -\delta n + \delta \sum_{t=1}^{n} \mathbf{1}\{v(\overline{\theta}_t) < v(\theta_0)\} \; + \; M \sum_{\theta\in H_0(\theta_0)} \sum_{t=1}^{n} I_{\theta_0,\theta}(x_t, j_t)\, \mathbf{1}\{\overline{\theta}_t = \theta\} \\
&\quad + M \min_{\theta\in H(\theta_0)} \sum_{t=1}^{n} I_{\theta_0,\theta}(x_t, j_t)
\end{aligned} \tag{6.13}$$

where in the last steps we used the facts that $x_t = x^*(\overline{\theta}_t)$ under $\bar{\sigma}$, and that $I_{\theta_0,\theta} \geq 0$.

Defining $\eta = \delta/(\delta + M|\Theta|)$, it follows from the last inequality that ℓ_n will be negative unless one of $\Omega_1(n)$–$\Omega_3(n)$ is satisfied. $\square$

Lemma 6.4 *Let $\Omega_1(n)$–$\Omega_3(n)$ be defined as in the previous lemma. Then, for some $Q_2 < \infty$ and every $\tau \in \mathcal{T}$,*

(i) $\displaystyle\sum_{n=1}^{\infty} P_{\theta_0}^{\bar{\sigma},\tau}\{\Omega_1(n)\} \leq Q_2\,,$

(ii) $\displaystyle\sum_{n=1}^{\infty} P_{\theta_0}^{\bar{\sigma},\tau}\{\Omega_2(n)\} \leq Q_2\,,$

(iii) $\displaystyle\sum_{n=1}^{\infty} P_{\theta_0}^{\bar{\sigma},\tau}\{\Omega_3(n),\, \overline{\theta}_n \in H(\theta_0)\} \leq Q_2\,.$

Proof:

(i) Recall from (5.4) that $v(\overline{\theta}_t) < v(\theta_0)$ implies $\lambda_{t-1}(\hat{\theta}_t, \theta_0) := \lambda_{t-1}(\hat{\theta}_t)/\lambda_{t-1}(\theta_0) > K_t$. Therefore (denoting $P := P_{\theta_0}^{\bar{\sigma},\tau}$),

$$\begin{aligned} P\{\Omega_1(n)\} &\leq P\left\{\sum_{t=1}^{n} \mathbf{1}\{\lambda_{t-1}(\hat{\theta}_t, \theta_0) > K_t\} \geq \eta n\right\} \\ &\leq P\{\lambda_{t-1}(\hat{\theta}_t, \theta_0) > K_t \text{ for some } t \geq \eta n\} \\ &\leq P\left\{\sup_{t\geq 1} \lambda_{t-1}(\hat{\theta}_t, \theta_0) > K_{[\eta n]}\right\} \\ &\leq \sum_{\theta\in\Theta} P\left\{\sup_{t\geq 1} \lambda_{t-1}(\theta, \theta_0) > K_{[\eta n]}\right\}, \end{aligned} \tag{6.14}$$

where $[\eta n]$ is the integer part of ηn. Now, since the likelihood ratio $\{\lambda_t(\theta, \theta_0)\}$ is a positive Martingale with expected value 1, it follows by Doob's inequality that $P\{\Omega_1(n)\} \leq |\Theta|\ (K_{[\eta n]})^{-1}$. Since $\{K_n^{-1}\}$ is summable by (5.1), this implies

$$\sum_{n=1}^{\infty} P\{\Omega_1(n)\} \leq |\Theta| \sum_{n=1}^{\infty} K_{[\eta n]}^{-1} < \infty\,. \tag{6.15}$$

(ii) Using the union bound and (5.5),

$$\begin{aligned} P\{\Omega_2(n)\} &\leq \sum_{\theta\in H_0(\theta_0)} P\left\{\sum_{t=1}^{n} I_{\theta_0,\theta}(x_t,\, j_t)\ \mathbf{1}\{\overline{\theta}_t = \theta\} \geq \eta n\right\} \\ &\leq \sum_{\theta\in H_0(\theta_0)} P\left\{\sum_{t=1}^{n} I_{\theta_0,\theta}(x_t,\, j_t)\ \mathbf{1}\{\Lambda_{t-1}(\theta_0, \theta) \leq \log K_t\} \geq \eta n\right\}. \end{aligned} \tag{6.16}$$

Now, using the same procedure as in the proof of Theorem 4.1 (i.e., player 2 is allowed to choose (x_t, j_t), and Λ_{t-1} is replaced by its truncated version $\tilde{\Lambda}_{t-1}$), it follows by Lemmas 3.1 and 3.3(iv) (with $\beta_n = \log K_n$) that, for every $\tau \in \mathcal{T}$,

$$\sum_{n=1}^{\infty} P\{\Omega_2(n)\} \leq \sum_{\theta\in H_0(\theta_0)} Q(\theta) < \infty\,. \tag{6.17}$$

(iii) Similarly to the proof of (ii), it follows that

$$\begin{aligned} P\{\Omega_3(n), \overline{\theta}_{n+1} \in H(\theta_0)\} &\leq \sum_{\theta\in H(\theta_0)} P\left\{\sum_{t=1}^{n} I_{\theta_0,\theta}(x_t, j_t) \geq \eta n,\ \overline{\theta}_{n+1} = \theta\right\} \\ &\leq \sum_{\theta\in H(\theta_0)} P\left\{\sum_{t=1}^{n} I_{\theta_0,\theta}(x_t, j_t) \geq \eta n,\ \Lambda_n(\theta_0, \theta) \leq \log K_{n+1}\right\}. \end{aligned} \tag{6.18}$$

The bound now follows exactly as in (ii), except that Lemma 3.3(iii) is used in place of Lemma 3.3(iv). $\square$

We are now ready to conclude the proof of Theorem 5.2. By (6.3), (6.4), Lemma 5.1(ii) and Lemma 6.2(ii), it follows that for every $\tau \in \mathcal{T}$ and $n \geq 1$:

$$L_n^{\bar{\sigma},\tau}(\theta_0) \;\leq\; \hat{D}Q_1 + E_{\theta_0}^{\bar{\sigma},\tau}\ell_n \;\leq\; \hat{D}Q_1 + Q_3 + \hat{D}\sum_{t=1}^{\infty} P\{\ell_t \geq 0,\, \overline{\theta}_{t+1} \in H(\theta_0)\}\,. \tag{6.19}$$

Moreover, by Lemmas 6.3 and 6.4,

$$\begin{aligned}\sum_{t=1}^{\infty} P\{\ell_t \geq 0,\, \overline{\theta}_{t+1} \in H(\theta_0)\} \;&\leq\; \sum_{t=1}^{\infty} P\{\Omega_1(t)\} + \sum_{t=1}^{\infty} P\{\Omega_2(t)\} + \sum_{t=1}^{\infty} P\{\Omega_3(t),\, \overline{\theta}_{t+1} > \theta_0\} \\ &\leq\; 3Q_2 \;<\; \infty\,, \end{aligned} \tag{6.20}$$

so that $L_n^{\bar{\sigma}}(\theta_0) \leq \hat{D}Q_1 + Q_3 + 3Q_2\hat{D} < \infty$ for every $n \geq 1$. □

7 Concluding Remarks

This paper examined the long-term performance of Certainty Equivalence strategies in an uncertain dynamic game situation. It was shown that these strategies potentially suffer from closed-loop identification problems, similar to those found in comparable adaptive control models, and that these problems can be essentially eliminated by properly modifying the estimator. In particular, Theorem 5.1 established that the worst-case loss can be kept down to $O(\log n)$ by using the value-biased Maximum Likelihood Estimator.

While the latter result seems quite satisfactory, it is still natural to ask whether this is the best that can be attained in general. In the sequel paper [23] it will be established that an increase rate of $O(\log n)$ is in fact the best that can be guaranteed by any strategy; furthermore, the optimal coefficient associated with this increase rate (i.e., the smallest possible coefficient $\beta(\theta_0)$ in Theorem 5.1) will be characterized, and a strategy which attains this "asymptotically optimal" performance will be constructed.

The basic model of this paper may be extended in several directions. Here we studied the case of finite parameter and actions sets; more general sets may be of interest. It should also be of interest to consider systems with non-trivial dynamics, e.g. controlled Markov processes (leading to stochastic game models). More ideas and methods from the field of adaptive control may prove applicable to such models.

References

[1] R. Agrawal, D. Teneketzis, and V. Anantharam, "Asymptotically efficient adaptive allocation schemes for controlled i.i.d. processes: Finite parameter space," *IEEE Trans. Automat. Control* **34**, pp. 258–267, 1989.

[2] R. Agrawal, D. Teneketzis, and V. Anantharam, "Asymptotically efficient adaptive allocation schemes for controlled Markov chains: Finite parameter space," *IEEE Trans. Automat. Control* **34**, pp. 1249–1259, 1989.

[3] V. Anantharam, P. Varaiya, and J. Walrand, "Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays. Part I: iid rewards, Part II: Markovian rewards," *IEEE Trans. Automat. Control* **32**, pp. 968–982, 1987.

[4] R. J. Aumann, "Mixed and behavior strategies in infinite extensive games," in: *Advances in Game Theory: Ann. of Math. Studies* **52**, M. Dresher et al. (Eds.), Princeton University Press, NJ, pp. 627–650, 1964.

[5] R. J. Aumann and S. Hart (Eds.), *Handbook of Game Theory with Economic Applications,* Chapters 5 and 6. North Holland, Amsterdam, 1992.

[6] A. Baños, "On pseudo-games," *Ann. Math. Statist.* **39**, pp. 1932–1945, 1968.

[7] D.P. Bertsekas, *Dynamic Programming and Stochastic Control,* Academic Press, NY, 1976.

[8] V. Borkar and P. Varaiya, "Adaptive control of Markov chains, I: Finite parameter set," *IEEE Trans. Automat. Control* **24**, pp. 953-958, 1979.

[9] V. Borkar and R. Varaiya, "Identification and adaptive control of Markov chains," *SIAM J. Control Optim.* **20**, pp. 470–489, 1982.

[10] T.M. Cover and J.A. Thomas, *Elements of Information Theory,* Wiley, NY, 1991.

[11] B. Doshi and S.E. Shreve, "Strong consistency of a modified maximum likelihood estimator for controlled Markov chains," *J. Appl. Prob.* **17**, pp. 726–734, 1980.

[12] G.C. Goodwin and K.S. Sin, *Adaptive Filtering, Prediction and Control,* Prentice-Hall, NJ, 1984.

[13] P. R. Kumar, "A survey of some results in stochastic adaptive control," *SIAM J. Control Optim.* **23**, pp. 329–380, 1985.

[14] P.R. Kumar and A. Becker, "A new family of optimal adaptive controllers for Markov chains," *IEEE Trans. Automat. Control* **27**, pp. 137-146, 1982.

[15] P.R. Kumar and P. Varaiya, *Stochastic Systems: Estimation, Identification and Adaptive Control,* Prentice-Hall, NJ, 1986.

[16] T.L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Adv. in Appl. Math.* **6**, pp. 4–22, 1985.

[17] S. Lakshmivarahan, *Learning Algorithms Theory and Applications,* Springer-Verlag, NY, 1981.

[18] P. Mandl, "Estimation and control in Markov chains," *Advances in Appl. Prob.* **6**, pp. 40–60, 1974.

[19] N. Megiddo, "On repeated games with incomplete information played by non-Bayesian players", *Internat. J. Game Theory* **9**, pp. 157–167, 1980.

[20] J.-F. Mertens, "Repeated Games," *Proceedings of the International Congress of Mathematicians (Berkeley 1986)*, pp. 1528-1577. American Mathematical Society, 1987.

[21] T. Parthasarathy and T.E.S. Ragahavan, *Some Topics in Two-Person Games,* American Elsevier Publishing Co., NY, 1971.

[22] J. Rosenfeld, "Adaptive competitive decision," in: *Advances in Game Theory: Ann. of Math. Studies* **52**, M. Dresher et al. (eds.), Princeton University Press, NJ, 1964, pp. 69–83.

[23] N. Shimkin and A. Shwartz, "Asymptotically efficient adaptive strategies in repeated games. Part 2: asymptotic optimality," Technical Report, University of Minnesota, IMA preprint series No. 1121, March 1993.

[24] S. Sorin, "An introduction to two-person zero-sum repeated games with incomplete information," IMSS-Economics, Technical Report No. 312, Stanford University, 1980.

[25] J. Van Ryzin, "Repetitive play in finite statistical games with unknown distributions," *Ann. Math. Statist.* **37**, pp. 976–974, 1966.

[26] S. Zamir, "On the relation between finitely and infinitely repeated games with incomplete information," *Internat. J. Game Theory* **1**, pp. 215–229, 1972.