Asymptotically Efficient Adaptive Strategies in Repeated Games Part II: Asymptotic Optimality
Author(s): Nahum Shimkin and Adam Shwartz
Source: *Mathematics of Operations Research*, Vol. 21, No. 2 (May, 1996), pp. 487-512
Published by: INFORMS
Stable URL: http://www.jstor.org/stable/3690245
Accessed: 09/02/2011 08:02

MATHEMATICS OF OPERATIONS RESEARCH
Vol. 21, No. 2, May 1996
Printed in U.S.A.

ASYMPTOTICALLY EFFICIENT ADAPTIVE STRATEGIES IN REPEATED GAMES PART II: ASYMPTOTIC OPTIMALITY

NAHUM SHIMKIN AND ADAM SHWARTZ

This paper continues the analysis of a dynamic decision problem modeled as a two-person repeated game with random rewards, perfect observations, and incomplete information on one side. The emphasis is on strategies of player 1 (the uninformed player) which maximize his worst-case total reward in a strong non-Bayesian sense, namely, for all possible states of nature. An asymptotic bound on performance is first established, followed by the construction of strategies which achieve this bound. The analysis highlights the efficient acquisition of (statistical) information under conflict conditions, and especially the relations between information and payoff which are inherent in this problem.

1. Introduction. This paper continues the study of asymptotically efficient strategies for the model considered in Shimkin and Schwartz (1995). For completeness we summarize briefly the model and relevant notation. Further details and background may be found in Shimkin and Schwartz (1995).

The game model involves two decision makers, player 1 (the maximizer) and player 2, which repeatedly play a matrix game $G(\theta_0)$, known to be a member of a finite set $\{G(\theta),\ \theta \in \Theta\}$. Each $G(\theta)$ is a zero-sum matrix game with random rewards, and with finite action sets $\mathscr{I}$ for player 1 and $\mathscr{J}$ for player 2. The reward structure is thus specified by the probability distributions $\{p_{\theta,i,j}(\cdot)\colon i \in \mathscr{I},\ j \in \mathscr{J}\}$ on a finite reward set $\mathscr{A}$. Perfect observations are assumed, so that after each stage t both decision makers observe and remember the actions (i_t, j_t) and the reward a_t. Rewards accumulate to form the total n-stage reward $\sum_{t=1}^{n} a_t$.

A strategy σ for player 1 is defined as a sequence $\{\sigma_t\}$, which specifies for each history sequence $h_t = \{i_s, j_s, a_s\}_{s=1}^{t-1}$ a "randomized action" $x_t = \sigma_t(h_t)$. Here $x_t \in \mathscr{P}(\mathscr{I})$ is a probability vector over $\mathscr{I}$, used to select the pure action i_t. Randomizations at different stages are performed independently; thus, we consider only behavioral strategies throughout this paper. A strategy τ for player 2 is defined similarly, with $y_t = \tau_t(h_t) \in \mathscr{P}(\mathscr{J})$. The sets of strategies for players 1 and 2 are denoted Σ and $\mathscr{T}$ respectively. Player 1 does not know the value of the true parameter θ_0 (except that it belongs to Θ), so that his strategies cannot depend on θ_0. Such dependence is allowed for player 2. For each triplet (θ_0, σ, τ) in $\Theta \times \Sigma \times \mathscr{T}$, let $P_{\theta_0}^{\sigma,\tau}$ and $E_{\theta_0}^{\sigma,\tau}$ denote the induced probability measure and expectation on the actions-rewards process. Further notations include $\mathscr{P}(\mathscr{I})$ and $\mathscr{P}(\mathscr{J})$—the sets of probability vectors over $\mathscr{I}$ and $\mathscr{J}$; $x_t \in \mathscr{P}(\mathscr{I})$ and $y_t \in \mathscr{P}(\mathscr{J})$—the randomized actions of player 1 and player 2 at stage t; $v(\theta)$—the minimax value of the matrix game $G(\theta)$; and $A_\theta(i,j) = \sum_{a \in \mathscr{A}} a \cdot p_{\theta,i,j}(a)$—the expected reward in $G(\theta)$ given actions (i,j).

The performance measure for player 1 will be defined in terms of the (relative) loss. For fixed σ, τ, θ_0 and $n \geq 1$ define the *loss* $L_n^{\sigma,\tau}(\theta_0)$ and the *worst-case loss*

Received March 18, 1993; revised October 20, 1994.
AMS 1991 subject classification. Primary: 90D20; secondary: 93E05.
OR/MS Index 1978 subject classification. Primary: 239 Games/Stochastic.
Key words. Repeated matrix games, incomplete information, adaptive control, total reward criterion.

0364-765X/96/2102/0487/$01.25

$L_n^\sigma(\theta_0)$ by

$$(1.1)\qquad L_n^\sigma(\theta_0) \triangleq \max_\tau L_n^{\sigma,\tau}(\theta_0) \triangleq \max_\tau E_{\theta_0}^{\sigma,\tau}\left(nv(\theta_0) - \sum_{t=1}^{n} a_t\right).$$

Note that $nv(\theta_0)$ equals the value of the n-stage game *under complete information*, and serves as a reference level for the reward in the incomplete-information game. For each strategy σ, the worst-case loss represents the deficiency in the worst-case (over all strategies of player 2) expected total reward with respect to this level. It is important to note that L_n^σ depends on the parameter θ_0, which is unknown to player 1. It therefore presents, for given n and σ, a *vector* of performance measures, whose entries correspond to the possible values of the true parameter. Ideally, player 1 would like to minimize (reduce to zero) simultaneously all entries of this vector.

While the previous paper (Shimkin and Schwartz 1995) focused on performance of strategies of relatively simple structure, the present paper is concerned with asymptotic (long-term) optimality. In defining a meaningful sense of optimality, we shall follow the asymptotic theory introduced by Lai and Robbins (1985) in relation with the statistical multiarmed bandit problem, and its extension in Agrawal et al. (1989) to controlled i.i.d. processes. The idea is that the *rate of increase* of $L_n^\sigma(\theta_0)$ may be simultaneously minimized for all θ_0. First, a lower bound on the rate of increase of $L_n^\sigma(\theta_0)$ will be established, which is logarithmic in n. More precisely, the bound holds for any strategy σ which is *uniformly good*, i.e., achieves a "satisfactory" rate of increase for *every* possible value of the true parameter θ_0 (see Definition 3.1). It establishes that the rate of increase of $L_n^\sigma(\theta_0)$ is at least $b(\theta_0)\log n$, with $b(\theta_0)$ a nonnegative constant which is explicitly specified. We then proceed to construct a strategy which is asymptotically optimal, in the sense that it satisfies the lower bound. This strategy is essentially based on the (much simpler) value-biased Certainty Equivalence strategy that was analyzed in Shimkin and Schwartz (1995), but some delicate modifications will be required to achieve asymptotic optimality.

In the adaptive control problem (Agrawal et al. 1989), which corresponds to the present model without player 2, it was possible to achieve asymptotic optimality by using a standard parameter estimation scheme, modified by adding a special "probing" phase. Probing is performed whenever it seems that insufficient statistical information was obtained, and is done by choosing actions which are solely dedicated to the efficient acquisition of information. In the present game model, such clear separation of an information acquisition phase is no longer possible. Indeed, depending on the model data, player 2 may be able to hide the value of the true parameter by using "nonrevealing" actions, under which the reward statistics under different parameters coincide (for *all* actions of player 1). Instead, a delicate balance must be maintained here between information acquisition and immediate rewards. Certain (sub-) strategies which are related to Blackwell's approachability theory (Blackwell 1954) will be constructed for that purpose.

The remainder of the paper is organized as follows. In §2 we introduce some simplifying assumptions regarding the model, as well as additional notation. Section 3 contains the asymptotic bound on the loss, and the definition of an asymptotically optimal strategy. Construction of such a strategy commences in §4, where two classes of sub-strategies are developed. These form the basis for the complete strategy, which is presented in §5.

2. Assumptions and further notation. In this paper we apply the following assumptions to the basic model:

ASSUMPTION A1. For each $i \in \mathscr{I}$ and $j \in \mathscr{J}$, the distributions $\{p_{\theta,i,j}(\cdot)\}_{\theta \in \Theta}$ are mutually absolutely continuous. That is, for every θ, θ' and a, $p_{\theta,i,j}(a) = 0$ if and only if $p_{\theta',i,j}(a) = 0$.

ASSUMPTION A2. In each matrix game $G(\theta)$, the optimal strategies of player 1 and player 2, denoted x_θ^* and y_θ^*, are unique.

ASSUMPTION A3. The values of the matrix games $\{G(\theta)\}$ are distinct, namely $v(\theta) \neq v(\theta')$ for $\theta \neq \theta'$.

These assumptions are technical in nature, but without them the construction and analysis of the optimal strategy would be much more complicated. Some of the associated complications (especially with respect to the omission of A1 and A3) may be perceived in the analysis of Shimkin and Schwartz (1995). As for the lower bound, A1 and A3 do not seem crucial; however, relaxing A2 would require a modification of the bound to account for nonuniqueness of y_θ^*.

Some additional definitions and basic relations from Shimkin and Schwartz (1995) are next recalled. $\mathscr{P}(\mathscr{J})$ denotes the set of probability vectors over the (finite) set $\mathscr{J}$. For any matrix $M = \{M(i,j)\}$, the following notation denotes averaging over rows or columns: $M(x,j) \triangleq \Sigma_i x_i M(i,j)$, $M(x,y) \triangleq \Sigma_{i,j} x_i y_j M(i,j)$, and similarly for $M(i,y)$.

The *one-stage loss* is defined by $d_\theta(i,j) = v(\theta) - A_\theta(i,j)$. In this notation the (total) loss may be written as

$$L_n^{\sigma,\tau}(\theta_0) = E_{\theta_0}^{\sigma,\tau} \sum_{t=1}^{n} d_{\theta_0}(i_t, j_t), \tag{2.1}$$

where $d_{\theta_0}(i_t, j_t)$ may be replaced (by appropriate conditioning) by $d_{\theta_0}(x_t, j_t)$, $d_{\theta_0}(i_t, y_t)$ or $d_{\theta_0}(x_t, y_t)$.

Define the *likelihood function* $\lambda_n(\theta) = \Pi_{t=1}^n p_{\theta,i_t,j_t}(a_t)$, and the *log-likelihood ratio*

$$\Lambda_n(\theta, \theta') = \sum_{t=1}^{n} \log \frac{p_{\theta,i_t,j_t}(a_t)}{p_{\theta',i_t,j_t}(a_t)}. \tag{2.2}$$

The corresponding *information divergence* (or *Kullback Leibler information*) is given by

$$I_{\theta,\theta'}(i,j) = \sum_{a \in \mathscr{A}} p_{\theta,i,j}(a) \log \frac{p_{\theta,i,j}(a)}{p_{\theta',i,j}(a)}. \tag{2.3}$$

It is always true that $I_{\theta,\theta'} \geq 0$, and, under Assumption A1, $I_{\theta,\theta'}$ is finite. Thus we need not introduce a truncated version as done in Shimkin and Schwartz (1995).

The parameters in Θ are assumed ordered according to the values $v(\theta)$, so that $\theta > \theta'$ stands for $v(\theta) > v(\theta')$, and $\theta \geq \theta'$ for $v(\theta) \geq v(\theta')$. Finally, $\|\cdot\|$ denotes the Euclidean norm, and $\|\cdot\|_\infty$ the sup-norm.

3. A lower bound on the loss.

In this section we derive an asymptotic lower bound on the worst-case loss. This will be used to define a meaningful non-Bayesian sense of optimal performance for player 1.

The stated objective of player 1 is to minimize (the rate of increase of) the worst-case loss. However, in general the worst-case loss cannot be minimized simultaneously for every possible θ_0. For example, if player 1 plays at every stage his optimal (maximin) strategy in $G(\theta)$ for some fixed $\theta \in \Theta$, then he guarantees zero loss if θ happens to be the true parameter. But if the true parameter is different, his loss may grow *linearly* in n.

To exclude such nonadaptive strategies, we shall restrict attention to strategies which perform "reasonably well" for every parameter, as specified in the following definition (compare Lai and Robbins (1985), Agrawal et al. (1989)).

DEFINITION 3.1. A strategy σ of player 1 is said to be uniformly good if for every $\theta_0 \in \Theta$:

$$L_n^\sigma(\theta_0) = o(n^\alpha) \quad \text{for every } \alpha > 0. \tag{3.1}$$

From Shimkin and Schwartz (1995) we know that the set of uniformly good strategies is nonempty, and in fact there exist strategies which guarantee that the loss rate is $O(\log n)$ at most. Thus, strategies outside this set need not be considered.

For each parameter θ, define an associated set of "bad" parameters (see the end of the section for interpretation of this set and discussion of the lower bound; note that θ is not included in $B(\theta)$):

$$B(\theta) = \{\theta' \in \Theta \colon v(\theta') > v(\theta), I_{\theta,\theta'}(x_\theta^*, y_\theta^*) = 0\}. \tag{3.2}$$

Since $I_{\theta,\theta'}$ is nonnegative, the requirement $I_{\theta,\theta'}(x_\theta^*, y_\theta^*) = 0$ in the last definition is equivalent to: $I_{\theta,\theta'}(i, j) = 0$ for every pair of *relevant actions* in $G(\theta)$, namely $i \in \mathscr{I}_\theta^* \triangleq \{i \colon (x_\theta^*)_i > 0\}$ and $j \in \mathscr{J}_\theta^* \triangleq \{j \colon (y_\theta^*)_j > 0\}$.

THEOREM 3.1. *Let $\theta \in \Theta$ be such that $B(\theta) \neq \varnothing$. Then, for every uniformly good strategy σ of player* 1,

$$\liminf_{n\to\infty} \frac{L_n^\sigma(\theta)}{\log n} \geq b(\theta), \tag{3.3}$$

where (*defining* $0/0 \triangleq \infty$),

$$b(\theta) = \min_{x \in \mathscr{P}(\mathscr{I})} \frac{d_\theta(x, y_\theta^*)}{\min_{\theta' \in B(\theta)} I_{\theta,\theta'}(x, y_\theta^*)} > 0. \tag{3.4}$$

PROOF. Consider a fixed $\theta \in \Theta$ such that $B(\theta) \neq \varnothing$. Let $\tau^\theta = \{y_\theta^*\}$ denote the stationary strategy of player 2, in which $y_n = y_\theta^*$ at each stage (note that this strategy does not depend on the true parameter of the game). It will be proved below that for any uniformly good strategy σ,

$$\liminf_{n\to\infty} \frac{L_n^{\sigma,\tau^\theta}(\theta)}{\log n} \geq b(\theta), \tag{3.5}$$

which clearly establishes the required bound.

Given that player 2 uses τ^θ, player 1 in effect is facing a "controlled i.i.d. process" of the type considered in Agrawal et al. (1989) with "state" $X_n = (a_n, j_n)$. However, the lower bound from Agrawal et al. (1989) does not apply here. The reason is that in the present case the single-stage loss may be either positive or negative, since it is defined with respect to the minimax value $v(\theta_0)$, whereas in the single controller case the single-stage loss is naturally defined with respect to the best achievable reward, and hence is always positive. (See also the remark below (3.11).) Still, it will be possible to follow the proof of Agrawal et al. (1989) after establishing the next two lemmas. It is important to note that the proof of the next lemma requires consideration of other (nonstationary) strategies of player 2 besides τ^θ. In fact, the results to follow do not necessarily hold if player 2 is limited a priori to stationary strategies.

LEMMA 3.1 (COMPARE AGRAWAL ET AL. (1989, LEMMA 3.2)). *Assume that σ is a uniformly good strategy of player* 1. *Then, for any $\theta' \in B(\theta)$,*

$$(3.6)\quad P_{\theta'}^{\sigma,\tau^{\theta}}\left\{\sum_{t=1}^{n} d_{\theta}(i_t, y_{\theta}^{*}) < K \log n\right\} = o(n^{\alpha-1}) \quad \textit{for every } \alpha > 0 \textit{ and } K > 0.$$

PROOF. Fix $\theta' \in B(\theta)$. It is first shown that a small loss under θ' implies a large loss under θ (see (3.10) below). Let $\mathcal{I}_{\theta}^{*} = \{i \in \mathcal{I}\colon (x_{\theta}^{*})_i > 0\}$ be the set of relevant actions of player 1 in $G(\theta)$. It is well known that this is exactly the set of actions which maximize the (expected) reward against y_{θ}^{*} (Parthasarthy and Ragahavan 1971, Theorems 3.1.2 and 3.1.16); that is, $A_{\theta}(i, y_{\theta}^{*}) = v(\theta)$ for $i \in \mathcal{I}_{\theta}^{*}$, and $A_{\theta}(i, y_{\theta}^{*}) < v(\theta)$ for $i \notin \mathcal{I}_{\theta}^{*}$. Consequently,

$$(3.7)\qquad d_{\theta}(i, y_{\theta}^{*}) \equiv v(\theta) - A_{\theta}(i, y_{\theta}^{*}) = 0 \quad \text{for } i \in \mathcal{I}_{\theta}^{*},$$

and, since the action set is finite, there exists a positive constant δ_1 such that

$$(3.8)\qquad d_{\theta}(i, y_{\theta}^{*}) \ge \delta_1 > 0 \quad \text{for } i \notin \mathcal{I}_{\theta}^{*}.$$

Consider now the game $G(\theta')$. By definition, $\theta' \in B(\theta)$ implies $I_{\theta,\theta'}(x_{\theta}^{*}, y_{\theta}^{*}) = 0$ (hence $I_{\theta,\theta'}(i, y_{\theta}^{*}) = 0$ for every $i \in \mathcal{I}_{\theta}^{*}$) and $v(\theta') > v(\theta)$. Noting (3.7),

$$(3.9)\qquad \begin{aligned} d_{\theta'}(i, y_{\theta}^{*}) &= v(\theta') - A_{\theta'}(i, y_{\theta}^{*}) = v(\theta') - A_{\theta}(i, y_{\theta}^{*}) \\ &= v(\theta') - v(\theta) \triangleq \delta_o > 0, \quad i \in \mathcal{I}_{\theta}^{*}. \end{aligned}$$

Denoting $D = \max_i d_{\theta'}(i, y_{\theta}^{*})$, it follows from (3.7)–(3.9) that

$$(3.10)\qquad d_{\theta'}(i, y_{\theta}^{*}) \ge \delta_o - (D + \delta_0)1\{i \notin \mathcal{I}_{\theta}^{*}\} \ge \delta_o - \delta_2 d_{\theta}(i, y_{\theta}^{*}),$$

where $\delta_2 \triangleq (D + \delta_o)/\delta_1 > 0$. Thus, using again the fact that $d_{\theta}(i, y_{\theta}^{*}) \ge 0$ by optimality of y_{θ}^{*},

$$(3.11)\qquad \begin{aligned} &P_{\theta'}^{\sigma,\tau^{\theta}}\left\{\sum_{t=1}^{n} d_{\theta}(i_t, y_{\theta}^{*}) < K \log n\right\} \\ &\quad = P_{\theta'}^{\sigma,\tau^{\theta}}\left\{\sum_{t=1}^{m} d_{\theta}(i_t, y_{\theta}^{*}) < K \log n, \forall m \le n\right\} \\ &\quad \le P_{\theta'}^{\sigma,\tau^{\theta}}\left\{\sum_{t=1}^{m} d_{\theta'}(i_t, y_{\theta}^{*}) \ge \delta_o m - \delta_2 K \log n, \forall m \le n\right\} \\ &\quad \triangleq P_{\theta'}^{\sigma,\tau^{\theta}}\{F_n\}, \end{aligned}$$

where the event F_n is defined accordingly. To establish the lemma, it remains to show that the last probability decays as $o(n^{\alpha-1})$. (Note that application of Chebycheff's inequality, as in the proof of Lemma 3.2 in Agrawal et al. (1989), is impossible here since the loss $\Sigma_1^n d_{\theta'}(i_t, y_{\theta}^{*})$ may be negative.) Fix $n \ge 1$, and consider the following strategy τ' of player 2. First define a stopping time T by

$$T = \min\left\{1 \le m \le n \colon \sum_{t=1}^{m} d_{\theta'}(i_t, y_\theta^*) < \delta_o m - \delta_2 K \log n\right\},$$

and $T = n + 1$ if the minimized set is empty (this is exactly the event F_n). Define τ' as the strategy which chooses $y_t = y_\theta^*$ for $t < T$, and $y_t = y_{\theta'}^*$ thereafter. Since τ^θ and τ' coincide on F_n,

$$P_{\theta'}^{\sigma, \tau^\theta}\{F_n\} = P_{\theta'}^{\sigma, \tau'}\{F_n\}. \tag{3.12}$$

Also, noting that $d_{\theta'}(i, y_{\theta'}^*) \ge 0$ (by optimality of $y_{\theta'}^*$ in $G(\theta')$), we obtain under τ':

$$\begin{aligned} \sum_{t=1}^{n} d_{\theta'}(i_t, y_t) &= \sum_{t=1}^{T-1} d_{\theta'}(i_t, y_\theta^*) + \sum_{t=T}^{n} d_{\theta'}(i_t, y_{\theta'}^*) \\ &\ge \sum_{t=1}^{T-1} d_{\theta'}(i_t, y_\theta^*) \\ &\ge -\delta_2 K \log n + \delta_o (T-1) \\ &\ge -\delta_2 K \log n + \delta_o n 1\{F_n\}, \qquad P_{\theta'}^{\sigma,\tau'}\text{-a.s.}, \end{aligned} \tag{3.13}$$

where the last inequality holds since $T = n + 1$ on F_n. Now, since σ is uniformly good, it follows by (3.1), (1.1), (2.1) and (3.13) that for every $\alpha > 0$:

$$\begin{aligned} o(n^\alpha) = L_n^\sigma(\theta') \ge L_n^{\sigma,\tau'}(\theta') &= E_{\theta'}^{\sigma,\tau'}\left(\sum_{t=1}^{n} d_{\theta'}(i_t, y_t)\right) \\ &\ge -\delta_2 K \log n + \delta_o n P_{\theta'}^{\sigma,\tau'}\{F_n\}, \end{aligned} \tag{3.14}$$

so that:

$$P_{\theta'}^{\sigma,\tau'}\{F_n\} \le o(n^\alpha)/\delta_o n = o(n^{\alpha-1}). \tag{3.15}$$

Together with (3.11) and (3.12), this implies the lemma. □

LEMMA 3.2. *Assume $B(\theta) \ne \varnothing$. Then*
(i) $0 < b(\theta) < \infty$.
(ii) *The minimization in the definition* (3.4) *of $b(\theta)$ can alternatively be taken over $X(\theta) \triangleq \{x \in \mathscr{P}(\mathscr{I})\colon x_i = 0$ if $(x_\theta^*)_i > 0\}$: the set of randomized actions supported on nonrelevant actions.*

PROOF. The inequalities $b(\theta) > 0$ and $b(\theta) < \infty$ follow from the following facts (a) and (b), respectively:

(a) For every $x \in \mathscr{P}(\mathscr{I})$, $d_\theta(x, y_\theta^*) = 0$ implies that $I_{\theta,\theta'}(x, y_\theta^*) = 0$ for every $\theta' \in B(\theta)$. This follows from the definition of $B(\theta)$ by noting that $d_\theta(x, y_\theta^*) = 0$ implies that x is supported on the set $\mathscr{I}_\theta^*$ of relevant actions (cf. (3.7), (3.8)).

(b) $\min_{\theta' \in B(\theta)} I_{\theta,\theta'}(x, y_\theta^*) > 0$ for some $x \in \mathscr{P}(\mathscr{I})$. To see that, note that for every $\theta' \in B(\theta)$,

$$A_{\theta'}(x_{\theta'}^*, y_\theta^*) \ge v(\theta') > v(\theta) \ge A_\theta(x_{\theta'}^*, y_\theta^*),$$

i.e., rewards under θ and θ' are not identical, which implies that $I_{\theta,\theta'}(x^*_{\theta'}, y^*_\theta) \neq 0$. Thus fact (b) is satisfied by choosing x as a convex combination of $\{x^*_{\theta'}: \theta' \in B(\theta)\}$.

Item (ii) of the lemma follows since $i \in \mathscr{I}^*_\theta$ implies that $d_\theta(i, y^*_\theta) = 0$ (see (3.7)), and that $I_{\theta,\theta'}(i, y^*_\theta) = 0$ for all $\theta' \in B(\theta)$ (by definition of $B(\theta)$). □

Based on these lemmas, the proof of Theorem 3.1 may be concluded exactly as the proof of the lower bound in Agrawal et al. (1989). For the reader's convenience we outline the main steps. Given a uniformly good strategy σ, our objective is to establish (3.5). Fix $\rho > 0$, and for each $n \geq 1$ define the event

$$F_n = \left\{ \sum_{t=1}^n d_\theta(i_t, y^*_\theta) < \frac{b(\theta)}{1+2\rho} \log n \right\}.$$

Recalling that $d_\theta(i, y^*_\theta) \geq 0$, we obtain

$$L_n^{\sigma,\tau^\theta}(\theta) = E_\theta^{\sigma,\tau^\theta} \sum_{t=1}^n d_\theta(i_t, y^*_\theta) \geq \left(1 - P_\theta^{\sigma,\tau^\theta}\{F_n\}\right) \frac{b(\theta)}{1+2\rho} \log n.$$

Since $\rho > 0$ is arbitrary, to establish (3.5) it is now sufficient to show that $P_\theta^{\sigma,\tau^\theta}\{F_n\} \to 0$. Let B denote the number of elements in $B(\theta)$. Denote $P_{\theta'} = P_{\theta'}^{\sigma,\tau^\theta}$, and $P_B = B^{-1}\Sigma_{\theta' \in B(\theta)} P_{\theta'}$. Consider the following change of measure, for any event D_n measurable on the sigma algebra generated by $\{i_t, j_t, a_t\}_{t=1}^n$:

$$P_\theta\{D_n\} = \int_{D_n} \frac{dP_\theta}{dP_B} dP_B \leq \int_{D_n} B \min_{\theta' \in B(\theta)} \frac{dP_\theta}{dP_{\theta'}} dP_B = B \int_{D_n} \min_{\theta' \in B(\theta)} \exp\{\Lambda_n(\theta, \theta')\} dP_B,$$

where Λ_n is the log-likelihood ratio (2.2). Note that $\Lambda_n(\theta, \theta')$ is the sum of the (controlled i.i.d.) random variables $X_t \triangleq \log\{p_{\theta, i_t, j_t}(a_t)/p_{\theta', i_t, j_t}(a_t)\}$, with conditional expectation $E_\theta^{\sigma,\tau^\theta}(X_t | i_t = i) = I_{\theta,\theta'}(i, y^*_\theta)$. It follows from, e.g., Lemma 3.1 in Agrawal et al. (1989) that for every $\epsilon > 0$ and $\rho > 0$ there exist a constant $K(\epsilon, \rho)$ and an event $F(\epsilon, \rho)$ with $P_\theta\{F(\epsilon, \rho)\} > 1 - \epsilon$, such that on that event

$$\Lambda_n(\theta, \theta') \leq (1+\rho) n I_{\theta,\theta'}(\hat{x}_n, y^*_\theta) + K(\epsilon, \rho), \qquad \forall n \geq 1,\ \theta' \in B(\theta),$$

where $\hat{x}_n$ denotes the empirical distribution of player 1's actions up to time n, namely $(\hat{x}_n)_i = n^{-1}\Sigma_{t=1}^n \mathbf{1}\{i_t = i\}$. Since $\hat{x}_n \in \mathscr{P}(\mathscr{I})$, it follows from the definition of $b(\theta)$ in (3.4) that

$$\min_{\theta' \in B(\theta)} I_{\theta,\theta'}(\hat{x}_n, y^*_\theta) = d_\theta(\hat{x}_n, y^*_\theta) \frac{\min_{\theta' \in B(\theta)} I_{\theta,\theta'}(\hat{x}_n, y^*_\theta)}{d_\theta(\hat{x}_n, y^*_\theta)}$$

$$\leq d_\theta(\hat{x}_n, y^*_\theta) b(\theta)^{-1} = \frac{1}{n} \sum_{t=1}^n d_\theta(i_t, y^*_\theta) b(\theta)^{-1}.$$

Thus, noting the definitions of F_n and $F(\rho, \epsilon)$,

$$P_\theta\{F_n \cap F(\rho, \epsilon)\} \leq B \exp\left\{(1+\rho)\frac{\log n}{1+2\rho} + K(\epsilon, \rho)\right\} P_B\{F_n\}$$

$$= B\, e^{K(\epsilon, \rho)} n^{(1+\rho)/(1+2\rho)} P_B\{F_n\},$$

which converges to 0 as a consequence of Lemma 3.1. Finally, letting $\epsilon \to 0$ establishes $P_\theta\{F_n\} \to 0$. The proof of Theorem 3.1 is thus complete. □

Theorem 3.1 provides a lower bound on the asymptotic worst-case loss for those parameters which satisfy $B(\theta) \neq \emptyset$. Therefore, the best performance that player 1 can hope for (in terms of the asymptotic rate of the worst-case loss) is to achieve the lower bound for these parameters, while keeping the loss *finite* for the rest. This leads to the following definition of asymptotic optimality.

DEFINITION 3.2. A strategy σ of player 1 is said to be *asymptotically optimal* if

(i) $\limsup_{n\to\infty} L_n^\sigma(\theta_0) < \infty$ for every $\theta_0 \in \Theta$ s.t. $B(\theta_0) = \emptyset$,

(ii) $\limsup_{n\to\infty} L_n^\sigma(\theta_0)/\log n = b(\theta_0)$ for every $\theta_0 \in \Theta$ s.t. $B(\theta_0) \neq \emptyset$.

DISCUSSION. The lower bound of Theorem 3.1 can be rendered a simplified but useful heuristic interpretation, in accordance with Lai and Robbins (1985). Suppose that player 2 uses the strategy $\tau^\theta = \{y_\theta^*\}$ for some θ with $B(\theta) \neq \emptyset$. Suppose that player 1 has (statistical) indications that θ is the true parameter. If this is indeed the case, to achieve zero loss he must choose his actions in the relevant set $\mathscr{I}_\theta^*$. Unfortunately, this may lead to undesired consequences if in fact some $\theta' \in B(\theta)$ is the true parameter. Since $I_{\theta,\theta'}(i, y_\theta^*) = 0$ for every $i \in \mathscr{I}_\theta^*$ and $\theta' \in B(\theta)$ (by definition of the latter), these actions do not yield any statistical information for discriminating θ from θ'. Furthermore, under θ' a positive loss will be incurred at each stage (cf. (3.9)), leading to $O(n)$ loss.

Therefore, in a uniformly good strategy (against τ^θ), player 1 must "probe" the system by playing outside $\mathscr{I}_\theta^*$. To minimize the associated loss, he should choose a probing action which gives the best "loss to information ratio." This is the essence of the constant $b(\theta)$, where information is quantified by the Kullback-Leibler information.

The lower bound (3.3) may now be interpreted as follows. For a strategy of player 1 to be uniformly good (against τ^θ), if θ is the true parameter he must maintain his total information (i.e., a measure of statistical value of the data for discriminating θ and $B(\theta)$, related to the Kullback-Leibler information) at a level of $\log n$ at least (cf. Lai and Robbins (1985)). By performing the required probing optimally, he can keep the probing loss down to $b(\theta)\log n$.

4. Optimal strategies: Preliminary results

4.1. *Discussion and results.* This section is an intermediate step in the construction of an asymptotically optimal strategy. The latter will essentially be based on the *certainty equivalence strategy with biased MLE* which was introduced in Shimkin and Schwartz (1995). To indicate the required modifications in this basic strategy, we start by recalling its definition and performance. This will expose its deficiencies as compared with the required optimal performance. Two families of (sub-) strategies will be introduced to overcome these deficiencies. These strategies are not in themselves adaptive, i.e., each is designed with a specific parameter θ in mind. They will however be used as building blocks for the overall (adaptive) optimal strategy, to be presented in the next section.

Recall the following definitions from Shimkin and Schwartz (1995). The maximum likelihood estimator (MLE) $\hat{\theta}_t$ is the maximizer of the likelihood function $\lambda_{t-1}(\theta) = \prod_{s=1}^{t-1} p_{\theta,i_s,j_s}(a_s)$. For some fixed $Q > 1$, define the sequence

$$K_n = n(\log n)^Q + 1 \tag{4.1}$$

(this is the "smallest" sequence which satisfies requirements (5.1) in Shimkin and Schwartz (1995). Further define the *likely parameters set*:

$$\hat{\Theta}_t = \left\{\theta \in \Theta: \Lambda_{t-1}\left(\hat{\theta}_t, \theta\right) \le \log K_t\right\}, \tag{4.2}$$

and the *value-biased MLE*:

$$\bar{\theta}_t = \arg\max\left\{v(\theta): \theta \in \hat{\Theta}_t\right\}. \tag{4.3}$$

The certainty equivalence strategy with biased MLE, denoted $\bar{\sigma}$, is specified by $x_t = x^*(\bar{\theta}_t)$. The following results have been established for this strategy (Shimkin and Schwartz (1995, Theorems 5.1 and 5.2)):

THEOREM 4.1. *For every $\theta_0 \in \Theta$,*

(i) $L_n^{\bar{\sigma}}(\theta_0) \le O(\log n)$.

(ii) *Assume that $B_2(\theta_0) = \varnothing$, where $B_2(\theta_0) \triangleq \{\theta' \in \Theta$: $I_{\theta_0, \theta'}(x^*_{\theta_0}, j) = 0$ for some $j \in \mathcal{J}^*_{\theta_0}\}$. Then $L_n^{\bar{\sigma}}(\theta_0)$ is bounded.*

Note that the requirement $B_2(\theta_0) = \varnothing$ is equivalent, under Assumption A3, to condition $C_2(\theta_0)$ of Shimkin and Schwartz (1995). It may be readily verified that $B_2(\theta_0) \supset B(\theta_0)$, hence $B_2(\theta_0) = \varnothing$ implies $B(\theta_0) = \varnothing$, so that the last result is compatible with the lower bound of §3.

Compared with the definition of asymptotic optimality, the performance guaranteed by $\bar{\sigma}$ falls short in the following two cases:

(I) $B(\theta_0) = \varnothing$, but $B_2(\theta_0) \neq \varnothing$. Asymptotic optimality requires the loss to be bounded. However, $\bar{\sigma}$ only guarantees a loss of order $O(\log n)$ in this case.

(II) $B(\theta_0) \neq \varnothing$. Then $B_2(\theta_0) \neq \varnothing$, and again $\bar{\sigma}$ guarantees an $O(\log n)$ loss. However, it does not guarantee that the optimal coefficient $b(\theta_0)$ of the lower bound is achieved.

Consider case (I). We shall identify the key properties which enabled us to bound the loss under strategy $\bar{\sigma}$ when $B_2(\theta_0) = \varnothing$, and then attempt to obtain similar properties (by appropriate strategies) under the weaker condition $B(\theta_0) = \varnothing$. For any parameter θ, consider those times when the estimator $\bar{\theta}_n$ equals θ. According to $\bar{\sigma}$, x^*_θ is played at these times. Now the key properties which were used in the proof of Theorem 4.1(ii) are the following relations between loss (positive or negative) and information: For some M, δ and all j,

(a) $d_\theta(x^*_\theta, j) \le 0$.

(b) $B_2(\theta) = \varnothing$, which is equivalent to: $d_\theta(x^*_\theta, j) \le -\delta + M \min_{\theta' > \theta} I_{\theta, \theta'}(x^*_\theta, j)$ for all j.

(c) $d_{\theta'}(x^*_\theta, j) \le -\delta + M I_{\theta', \theta}(x^*_\theta, j)$ for every $\theta' < \theta$.

(The first property is of course a consequence of optimality of x^*_θ in $G(\theta)$, and the other two were established in Shimkin and Schwartz (1995 Lemma 6.1). In (b) we use the convention $\min \varnothing = \infty$.) The interpretation in the context of $\bar{\sigma}$ is as follows. Assume that x^*_θ is played (which occurs at the times when $\bar{\theta}_t = \theta$). If θ happens to be the true parameter, then (a) guarantees nonpositive loss. Moreover, by (b), if for some $\theta' > \theta$ no information is attained (i.e., $I_{\theta, \theta'}$ is low), this will be compensated by strictly negative loss. Also, if some $\theta' < \theta$ happens to be the true parameter, then (c) low $I_{\theta', \theta}$-information is compensated by strictly negative loss.

Unfortunately, property (b) does not hold if $B_2(\theta) \neq \varnothing$. Nonetheless, as long as the (smaller) set $B(\theta)$ is empty, a generalized version of these properties may still be achieved. This requires us to deviate from playing x^*_θ whenever θ is the estimated parameter, and instead use a modified (nonstationary, history-dependent) strategy over these times. The precise formulation follows.

PROPOSITION 4.1. *There exist strategies* $\{\sigma^*(\theta) \in \Sigma\colon \theta \in \Theta\}$ *and positive constants* M_1 *and* δ_1 *such that, for every strategy* τ *of player* 2 *and every* $n \geq 1$, *the following hold*:

(i) $\Sigma_{t=m}^{n} d_\theta(x_t, j_t) \leq M_1$, $m = 1, 2, \dots, n$.

(ii) $\Sigma_{t=1}^{n} d_\theta(x_t, j_t) \leq -\delta_1 n + M_1 + M_1 \min_{\theta' \in G_o(\theta)} \Sigma_{t=1}^{n} I_{\theta, \theta'}(x_t, j_t)$, *where* $G_o(\theta) = \{\theta'\colon \theta' > \theta\} \setminus B(\theta)$.

(iii) $d_{\theta'}(x_t, j_t) \leq -\delta_1 + M_1 I_{\theta', \theta}(x_t, j_t)$ *for every* $\theta' < \theta$, $t \geq 1$.

REMARK 4.1. The relations in Proposition 4.1, as well as in the rest of this section, hold in a sample-path sense, and for every $\theta_0 \in \Theta$.

REMARK 4.2. All the strategies σ of player 1 which appear in the last proposition, as well as in the rest of this section, have the special form $x_t = \sigma_t(j_1, \dots, j_{t-1})$. Thus, x_t depends on the history $h_t = \{i_s, j_s, a_s\}_{s<t}$ only through the actions of player 2.

REMARK 4.3. Properties (i)–(iii) are a generalization of properties (a)–(c) listed above. In fact, when $B_2(\theta) = \varnothing$ then $\sigma^*(\theta)$ may simply be taken as the stationary strategy $x_t \equiv x_\theta^*$.

REMARK 4.4. Note that (i) bounds the loss over any time interval $[m, \dots, n]$, and not just on $[1, \dots, n]$. This will be essential for the results of the next section (cf. the proof of Lemma 5.1).

The proof, as well as the definition of $\sigma^*(\theta)$, are presented in the second part of this section. The main idea in constructing this strategy is as follows. The negation of property (b) above (or of $B_2(\theta) = \varnothing$) may be written as:

$$(4.4)\quad I_{\theta,\theta'}(x_\theta^*, y) = 0 \quad \text{and} \quad d_\theta(x_\theta^*, y) = 0 \quad \text{for some } \theta' > \theta \text{ and } y \in \mathscr{P}(\mathscr{J}).$$

Now, when $B(\theta) = \varnothing$, (4.4) cannot hold for $y = y_\theta^*$. In other words, (4.4) is then satisfied only if player 2 uses an action y which deviates from his optimal one. (Note however that any such randomized action y must be supported on the relevant action set $\mathscr{J}_\theta^*$, as follows from $d_\theta(x_\theta^*, y) = 0$.) Thus, *if* that y was known in advance to player 1, he could achieve an expected reward greater than $v(\theta)$ (i.e., strictly negative loss) in the matrix game $G(\theta)$. When the game is repeated, a similar effect can be achieved (in the long run) by player 1 even if the y_t's are unknown in advance; see Proposition 4.3 below.

Let us turn to the second deficiency noted above, namely case (II). As discussed at the end of the last section, to achieve the optimal coefficient $b(\theta_0)$, player 1's strategy should include an "optimal probing" phase. This phase is intended to accumulate statistical information (quantified by $I_{\theta_0,\theta}$ for $\theta \in B(\theta_0)$) at a minimal loss-per-information ratio. Also, some safeguards should be activated if (due to player 2's actions) insufficient information is obtained.

If player 2 played $y_t = y_{\theta_0}^*$ at every stage, then such "optimal probing" could be achieved on a single-stage basis by any $x^o \in \mathscr{P}(\mathscr{I})$ which is a minimizer in (3.4). (This is trivially satisfied in the single-controller case; cf. Agrawal (1989).) However, since player 2 may play differently, then a stationary strategy $x_t \equiv x^o$ might not yield the desired result: The loss-per-information ratio may then be larger than $b(\theta_0)$, or possibly no information will be obtained.

Again, the problem will be resolved by "punishing" player 2 for playing off $y_{\theta_0}^*$. This can in principle be accomplished by superimposing the one-stage probing strategy x^o on a strategy similar to $\sigma^*(\theta_0)$ of Proposition 4.1. The following result may be thus obtained:

PROPOSITION 4.2. *Let* $\theta \in \Theta$ *be such that* $B(\theta) \neq \varnothing$. *Then there exists a strategy* $\sigma^o(\theta)$ *of player* 1 *and positive constants* M_2, δ_2 *such that, for every* $\tau \in \mathscr{T}$ *and* $n \geq 1$,

(i) *For every* $\epsilon > 0$ *and* $1 \le m \le n$,

$$\sum_{t=m}^{n} d_\theta(x_t, j_t) \le (1+\epsilon) b(\theta) \min_{\theta' \in B(\theta)} \sum_{t=1}^{n} I_{\theta,\theta'}(x_t, j_t) + M(\epsilon),$$

where $b(\theta)$ *is defined in* (3.4), *and* $M(\epsilon) > 0$ *is a constant which depends only on* ϵ.

(ii) $\Sigma_{t=1}^{n} d_\theta(x_t, j_t) \le -\delta_2 n + M_2 + M_2 \min_{\theta' > \theta} \Sigma_{t=1}^{n} I_{\theta,\theta'}(x_t, j_t)$.

(iii) $d_{\theta'}(x_t, j_t) \le -\delta_2 + M_2 I_{\theta',\theta}(x_t, j_t)$ *for every* $\theta' < \theta$, $t \ge 1$.

The bounds in Proposition 4.2 may be roughly interpreted as follows. (i) implies that the information-per-loss ratio (with respect to $B(\theta)$) is close to optimal, provided that information is indeed accumulated (say, at an $O(n)$ rate). Item (ii) implies that if the information rate (with respect to any $\theta' > \theta$, and in particular for $\theta' \in B(\theta)$) is smaller than some critical linear rate, then a strictly negative loss results; compare with (ii) of Proposition 4.1. Finally, (iii) is analogous to Proposition 4.1(iii) or property (c) above.

4.2. *Proofs.* The proofs of Propositions 4.1 and 4.2 depend on a basic result for repeated matrix games, established by different methods in Hannan (1957) and Blackwell (1954), which essentially states the following. In a (complete information) repeated matrix game, each player can asymptotically guarantee for himself an average reward which is no less than what he could guarantee if he knew in advance the empirical frequencies of his opponent's actions. The following (somewhat nonstandard) version of this result will be required here:

PROPOSITION 4.3. (i) *For every* $\theta \in \Theta$, *there exists a strategy* $\bar{\sigma}(\theta)$ *of player* 1 *such that*

$$(4.5) \qquad \frac{1}{n}\sum_{t=1}^{n} A_\theta(x_t, j_t) \ge \max_{x \in \mathscr{P}(\mathscr{J})} A_\theta(x, \bar{y}_n) - \frac{B}{\sqrt{n}}, \qquad \forall \tau \in \mathscr{T}, n \ge 1,$$

where B *is a positive constant, and* $\bar{y}_n$ *is the empirical distribution of player* 2*'s actions up to stage* n.

(ii) *The strategy* $\bar{\sigma}(\theta)$ *may be defined as follows. Let*

$$Q = \left\{(a, y) \in \mathbb{R} \times \mathscr{P}(\mathscr{J}) : a \ge \max_{x \in \mathscr{P}(\mathscr{J})} A_\theta(x, y)\right\},$$

and consider a point $(a, y) \in \mathbb{R} \times \mathscr{P}(\mathscr{J})$ *such that* $(a, y) \notin Q$. *Let* c *denote the closest point in* Q *to* (a, y), *and* $(\alpha, \xi) = c - (a, y)$. *Finally, let* $x^*(a, y)$ *be an optimal (maximin) strategy of player* 1 *in the matrix game with augmented payoff matrix*: $A^{(\alpha,\xi)} \triangleq \alpha A_\theta + \underline{1}'\xi \equiv (\alpha A_\theta(i,j) + \xi_j)$. *Then*

$$\bar{\sigma}(\theta)_n(h_n) = \begin{cases} x^*(\bar{a}_{n-1}, \bar{y}_{n-1}) & \text{if } (\bar{a}_{n-1}, \bar{y}_{n-1}) \notin Q, \\ \text{arbitrary} & \text{otherwise,} \end{cases}$$

where $\bar{a}_n = n^{-1}\Sigma_{t=1}^{n} A_\theta(x_t, j_t)$.

PROOF. As observed in Blackwell (1954), the proof follows by applying general approachability results (Blackwell 1956) to the set Q. Although the approachability result required here is not standard (in that the one-stage payoff depends directly on x_t instead of i_t, and a.s. relations are required), it may be easily inferred from the version which appears, e.g., in Sorin (1980). A direct proof is provided in Shimkin and Shwartz (1993). □

In accordance with Remark 4.2, the strategy $\bar{\sigma}(\theta)$ as defined in (ii) depends on the history only through player 2's actions. Indeed, A_θ is deterministic, and x_t may be recursively eliminated from the equations.

We shall also require the following lemma, which lower-bounds the maximal penalty that a player in a matrix game pays for deviating from his optimal strategy.

Lemma 4.1. *Let A be an $\mathscr{I} \times \mathscr{J}$ zero-sum game matrix with value $v(A)$. Assume that y^* is a unique optimal (minimax) strategy for player 2. Then for some $\delta > 0$ and every $y \in \mathscr{P}(\mathscr{J})$:*

$$\max_{x \in \mathscr{P}(\mathscr{I})} A(x, y) \geq v(A) + \delta \|y - y^*\|. \tag{4.6}$$

Proof. Consider $f(y) \triangleq \max_x A(x, y)$, $y \in \mathscr{P}(\mathscr{J})$. Since y^* is a unique optimal strategy, it follows that $f(y^*) = v(A)$ and $f(y) > v(A)$ for $y \neq y^*$, so that y^* is the unique minimizer of f. Note further that $f(y) = \max_i A(i, y)$, so that f is the maximum of a finite number of linear functions. The inequality (4.6) is an easy consequence of these facts. □

The next lemma will be useful in establishing property (iii) in Propositions 4.1 and 4.2:

Lemma 4.2. *There exist a (small enough) constant $0 < \mu \leq 1/2$ and positive constants δ_3, M_3 such that $\|x - x_\theta^*\|_\infty \leq \mu$ implies*

$$d_{\theta'}(x, j) \leq -\delta_3 + M_3 I_{\theta', \theta}(x, j), \qquad \forall j, \theta, \theta' < \theta.$$

Proof. By Lemma 6.1(i) in Shimkin and Schwartz (1995), there exist positive δ and M such that for every j and $\theta' < \theta$: $d_{\theta'}(x_\theta^*, j) \leq -\delta + M I_{\theta', \theta}(x_\theta^*, j)$. (This is exactly the property (c) discussed at the beginning of this section.) The lemma follows by continuity of $d_{\theta'}$ and $I_{\theta, \theta'}$ in x. □

We proceed now to the proof of Proposition 4.1. It will be convenient to use in the remainder of this section the abbreviated notation:

$$d_\theta\{m:n\} \triangleq \sum_{t=m}^{n} d_\theta(x_t, j_t), \qquad I_{\theta, \theta'}\{m:n\} \triangleq \sum_{t=m}^{n} I_{\theta, \theta'}(x_t, j_t).$$

Also, recall that $G_o(\theta) = \{\theta' > \theta\} \setminus B(\theta)$. As an intermediate step, the following type of strategy is required:

Lemma 4.3. *For each $\theta \in \Theta$, there exists a strategy $\sigma^1(\theta)$ for player 1 and positive constants M_4, δ_4, ϵ_4 such that the following hold for every $\tau \in \mathscr{T}$ and $n \geq 1$.*

(i) $d_\theta\{1:n\} \leq M_4$.

(ii) $\min_{\theta' \in G_o(\theta)} I_{\theta, \theta'}\{1:n\} \leq \epsilon_4 n$ *implies* $d_\theta\{1:n\} \leq -\delta_4 \sqrt[4]{n} + M_4$.

(iii) $\|x_n - x_\theta^*\|_\infty \leq \mu$, *where μ is as in Lemma* 4.2.

Remark. The strategy $\sigma^1(\theta)$ will be based on the strategy $\bar{\sigma}(\theta)$ of Proposition 4.3. However, an essential improvement is property (i), i.e., bounded loss. In contrast, under $\bar{\sigma}(\theta)$ the total loss may be as high as $B\sqrt{n}$ if $\bar{y}_n = y_\theta^*$, since then (4.5) is equivalent to: $d_\theta\{1:n\} \leq B\sqrt{n}$.

PROOF. Let θ be fixed, and let $\bar{\sigma}(\theta)$ be the strategy of Proposition 4.3. For each $0 < \xi < 1$, define the following strategy:

$$(4.7) \qquad \sigma(\xi): \sigma(\xi)_t = \xi\bar{\sigma}(\theta)_t + (1-\xi)x_\theta^*.$$

(This corresponds to independent mixing at each stage of actions $\bar{\sigma}(\theta)_t(h_t)$ and x_θ^*.) The required strategy $\sigma^1(\theta)$ will be defined by restarting $\sigma(\xi)$ at prespecified times, with ξ diminishing to zero. This scheme makes it possible to guarantee property (i), namely bounded loss.

Let $0 < \mu \leq 1/2$ be as defined in Lemma 4.2. Choose a real sequence $\{\xi_k\}_{k\geq 0}$ and a sequence of integers $0 = T_o < T_1 < \cdots$, such that: $0 < \xi_k \leq \mu$, $\xi_k \downarrow 0$, and for some finite constants C_1, C_2:

$$(4.8) \qquad \sum_{k=0}^{\infty} \xi_k\sqrt{T_{k+1} - T_k} \leq C_1,$$

$$(4.9) \qquad \xi_k T_k \geq C_2\sqrt[4]{T_{k+1}} \qquad \forall k \geq 1.$$

Two specific examples are (with $0 < \epsilon < 1/4$): (a) $T_k = 2^k - 1$, $\xi_k = \mu 2^{-k(1/2+\epsilon)}$. (b) $T_k = k^3$, $\xi_k = \mu(1+k)^{-(2+\epsilon)}$.

Finally, define $\sigma^1(\theta)$ as follows:

Strategy $\sigma^1(\theta)$. At stages $t = 1, 2, \ldots, T_1$, play according to $\sigma(\xi_o)$. Next, at stage T_k, $k \geq 1$, reset the history counter to 0, and then play over $t = T_k + 1, \ldots, T_{k+1}$ according to $\sigma(\xi_k)$. More precisely, for $T_k < t \leq T_{k+1}$, $x_t = \xi_k\bar{\sigma}(\theta)_{t-T_k}(j_{T_k+1}, \ldots, j_{t-1}) + (1-\xi_k)x_\theta^*$.

We proceed to upper-bound the loss and lower-bound the information under $\sigma^1(\theta)$. Both bounds will be in terms of $\|\bar{y}_n - y_\theta^*\|$, player 2's average deviation from his optimal strategy in $G(\theta)$. It is assumed in the following that player 2 is using any strategy $\tau \in \mathcal{T}$.

Consider first the strategy $\sigma(\xi)$ defined above, with ξ fixed. Suppose for the moment that this strategy is used throughout by player 1. Then $x_t = \xi\tilde{x}_t + (1-\xi)x_\theta^*$, where $\tilde{x}_t = \bar{\sigma}(\theta)_t(j_1, \ldots, j_{t-1})$. Therefore, by optimality of x_θ^*, Proposition 4.3 and Lemma 4.1:

$$(4.10) \quad \begin{aligned} \frac{1}{n}\sum_{t=1}^{n} A_\theta(x_t, j_t) &= \frac{1}{n}\sum_{t=1}^{n}\{\xi A_\theta(\tilde{x}_t, j_t) + (1-\xi)A_\theta(x_\theta^*, j_t)\} \\ &\geq \xi\frac{1}{n}\sum_{t=1}^{n} A_\theta(\tilde{x}_t, j_t) + (1-\xi)v(\theta) \\ &\geq \xi\left(\max_x A_\theta(x, \bar{y}_n) - B/\sqrt{n}\right) + (1-\xi)v(\theta) \\ &\geq \xi\left(v(\theta) + \delta_\theta\|\bar{y}_n - y_\theta^*\| - B/\sqrt{n}\right) + (1-\xi)v(\theta) \\ &= v(\theta) + \xi\delta_\theta\|\bar{y}_n - y_\theta^*\| - \xi B/\sqrt{n}, \end{aligned}$$

where B and δ_θ are positive constants. (The second inequality follows from Proposition 4.3, even though x_t and not $\tilde{x}_t$ is actually played; to verify that, recall that $\tilde{x}_t$ is a function of the sequence $\{j_t\}$ only, as are all the other variables in inequality (4.5).

Therefore, (4.5) may be interpreted equivalently as a deterministic inequality that holds for all possible values of this sequence.) This can be written equivalently as:

$$(4.11) \qquad d_\theta\{1:n\} \equiv nv(\theta) - \sum_{t=1}^{n} A_\theta(x_t, j_t) \leq -\xi\delta_\theta n\|\bar{y}_n - y_\theta^*\| + \xi B\sqrt{n}\,.$$

Returning to the strategy $\sigma^1(\theta)$, assume henceforth that this strategy is used by player 1. Let $T_k < m \leq T_{k+1}$ for some $k \geq 0$. Observe that $\sigma(\xi_k)$ is started at $t = T_k + 1$. Therefore, (4.11) implies:

$$d_\theta\{T_k + 1 : m\} \leq -\xi_k\delta_\theta(m - T_k)\|\Delta y(m, T_k)\| + \xi_k B\sqrt{m - T_k}\,,$$

where

$$\Delta y(m, T_k) = \frac{1}{m - T_k}\sum_{t=T_k+1}^{m} e_{j_t} - y_\theta^*.$$

Therefore, for any $T_K < n \leq T_{K+1}$, $K \geq 0$:

$$d_\theta\{1:n\} \leq \sum_{k=0}^{K-1}\left(-\xi_k\delta_\theta\Delta T_k\|\Delta y(T_{k+1}, T_k)\| + \xi_k B\sqrt{\Delta T_k}\right)$$
$$+\left(-\xi_K\delta_\theta(n - T_K)\|\Delta y(n, T_K)\| + \xi_K B\sqrt{n - T_K}\right),$$

where $\Delta T_k = T_{k+1} - T_k$. Now, using the fact that $\{\xi_k\}$ is decreasing, the triangle inequality, and (4.8):

$$d_\theta\{1:n\} \leq -\xi_K\delta_\theta\left(\sum_{k=0}^{K-1}\Delta T_k\|\Delta y(T_{k+1}, T_k)\| + (n - T_K)\|\Delta y(n, T_K)\|\right) + B\sum_{k=0}^{\infty}\xi_k\sqrt{\Delta T_k}$$
$$\leq -\xi_K\delta_\theta n\|\bar{y}_n - y_\theta^*\| + BC_1.$$

Moreover, since $T_K < n \leq T_{K+1}$, it follows by (4.9) that:

$$\xi_K n \geq \xi_K T_K \geq C_2\sqrt[4]{T_{K+1}} \geq C_2\sqrt[4]{n}\,,$$

so that, finally, we obtain the upper bound

$$(4.12) \qquad d_\theta\{1:n\} \leq -C_2\delta_\theta\sqrt[4]{n}\|\bar{y}_n - y_\theta^*\| + BC_1.$$

Next, the information will be bounded. Let $\theta' \in G_o(\theta)$. Since $x_t = (1 - \xi_k)x_\theta^* + (\ldots)$ with $1 - \xi_k \geq 1 - \mu \geq 1/2$, and noting that $I_{\theta,\theta'} \geq 0$,

$$(4.13) \qquad \sum_{t=1}^{n} I_{\theta,\theta'}(x_t, j_t) \geq \frac{1}{2}\sum_{t=1}^{n} I_{\theta,\theta'}(x_\theta^*, j_t) = \frac{1}{2}nI_{\theta,\theta'}(x_\theta^*, \bar{y}_n)$$
$$= \frac{1}{2}nI_{\theta,\theta'}(x_\theta^*, y_\theta^*) + \frac{1}{2}nI_{\theta,\theta'}(x_\theta^*, \bar{y}_n - y_\theta^*)$$
$$\geq \frac{1}{2}\beta_1 n - \frac{1}{2}\beta_2 n\|\bar{y}_n - y_\theta^*\|, \qquad \theta' \in G_o(\theta)$$

where

$$\beta_1 = \min_{\theta' \in G_o(\theta)} I_{\theta,\theta'}(x_\theta^*, y_\theta^*) > 0, \qquad \beta_2 = \max_{\theta' \in G_o(\theta)} \max_j I_{\theta,\theta'}(x_\theta^*, j).$$

Note that β_1 is positive by definition of $G_o(\theta)$ and $B(\theta)$ in Proposition 4.1 and (3.2).

We may now proceed to establish (i)–(iii) of the lemma.

(i) Follows immediately from (4.12) (since $C_2\delta_\theta > 0$), for any $M_4 \geq BC_1$.

(ii) Assume that for some $\theta' \in G_o(\theta), \Sigma_{t=1}^n I_{\theta,\theta'}(x_t, j_t) \leq \epsilon n$ where $\epsilon > 0$ will be specified shortly. Then, from (4.13), $\|\bar{y}_n - y_\theta^*\| \geq (\beta_1 - 2\epsilon)/\beta_2$. Therefore, for $\epsilon = \beta_1/4$, (4.12) gives:

$$d_\theta\{1:n\} \leq -C_2\delta_\theta \frac{\beta_1}{2\beta_2}\sqrt[4]{n} + BC_1$$

which clearly implies (ii) for any $\delta_4 \leq C_2\delta_\theta\beta_1/2\beta_2$ and $M_4 \geq BC_1$.

(iii) Recall that, for $T_K < n \leq T_{K+1}, x_n = \xi_K\tilde{x}_n + (1 - \xi_K)x_\theta^*$ for some $\tilde{x}_n \in \mathscr{P}(\mathscr{I})$, and $\xi_K \leq \mu$. Therefore, $\|x_n - x_\theta^*\|_\infty \leq \mu\|\tilde{x}_n - x_\theta^*\|_\infty \leq \mu$. □

Proof of Proposition 4.1. To motivate the definition of $\sigma^*(\theta)$ below, note that property (i) in the proposition requires the loss to be bounded on any time interval $[m, n]$. However, the strategy $\sigma^1(\theta)$ of the previous lemma guarantees that only on $[1, n]$. Thus, if the loss is negative on $[0, m-1]$, say, it might be large on $[m, n]$.

To rectify this problem, we define $\sigma^*(\theta)$ as the strategy which follows $\sigma^1(\theta)$ as long as the loss is above a certain (negative) threshold. However, as soon as it goes below this threshold, the clock is reset and $\sigma^1(\theta)$ is restarted with a new history. The precise definition follows.

Strategy $\sigma^*(\theta)$. Let C_1 be a positive constant. Let $\{m_k\}_{k\geq 0}$ be the sequence of stopping times (possibly infinite) defined recursively by: $m_o = 0, m_{k+1} = \inf\{m \geq m_k + 1\colon\ d_\theta\{m_k + 1\colon\ m\} \leq -C_1\}$. Then, for $m_k + 1 \leq t \leq m_{k+1}$, $\sigma^*(\theta)_t(h_t) \triangleq \sigma^1(\theta)_{t-m_k}(h_t^{(k)})$, where $h_t^{(k)} \triangleq (j_{m_k+1}, \ldots, j_{t-1})$.

Assume that player 1 uses this strategy $\sigma^*(\theta)$. Let $\tau \in \mathscr{T}$ and $n \geq 1$ be fixed, and let $K \geq 0$ be such that $m_K < n \leq m_{K+1}$. Define:

$V_k = \{m_k + 1, \ldots, m_{k+1}\}, 0 \leq k \leq K-1$: the kth (terminated) interval.

$V_K = \{m_K + 1, \ldots, n\}$: the last ($K$th) interval.

By definition of $\{m_k\}$, it follows that on each interval:

$$d_\theta\{m_k + 1 : m\} \geq -C_1 - \hat{D} \qquad \forall m \in V_k, 0 \leq k \leq K, \tag{4.14}$$

where $\hat{D} = \max_{i,j}|d_\theta(i,j)|$. On the other hand, on each terminated interval:

$$d_\theta\{V_k\} \triangleq \sum_{t\in V_k} d_\theta(x_t, j_t) \leq -C_1, \qquad 0 \leq k \leq K-1. \tag{4.15}$$

Finally, since $\sigma^1(\theta)$ is used on each interval, and in particular on the last interval, it follows from Lemma 4.3(i) that:

$$d_\theta\{V_K\} \leq M_4. \tag{4.16}$$

We proceed now to prove assertions (i)–(iii) of the proposition. Note that it is enough to prove each assertion with different constants (M_1, δ_1), since then the maximal M_1 and minimal δ_1 satisfy the assertions simultaneously.

(i) Fix $1 < m \leq n$. Then $m - 1 \in V_k$ for some $0 \leq k \leq K$, so that by (4.14),

$$d_\theta\{m_k + 1: m - 1\} \geq -C_1 - \hat{D}. \tag{4.17}$$

Also, by (4.15) and (4.16) it follows that $d_\theta\{m_k + 1: n\} \leq M_4$. Subtracting the last two inequalities gives $d_\theta\{m : n\} \leq M_4 + C_1 + \hat{D}$, so that (i) holds for $M_1 = M_4 + C_1 + \hat{D}$.

(ii) Since $\sigma^1(\theta)$ is used on each interval, it follows by Lemma 4.3(ii) that, for some positive constants ϵ_4, δ_4 and every $0 \leq k \leq K$, $d_\theta\{V_k\} > -\delta_4\sqrt[4]{|V_k|} + M_4$ implies

$$\min_{\theta' \in G_o(\theta)} I_{\theta,\theta'}\{V_k\} \geq \epsilon_4|V_k|. \tag{4.18}$$

Let $L > 0$ be some large enough constant so that $-\delta_4\sqrt[4]{L} + M_4 < -C_1 - \hat{D}$. It follows then from (4.15) that on each interval V_k, $0 \leq k \leq K$, for which $|V_k| \geq L$:

$$d_\theta\{V_k\} \geq -C_1 - \hat{D} > -\delta_4\sqrt[4]{L} + M_4 \geq -\delta_4\sqrt[4]{|V_k|} + M_4,$$

so that (4.18) holds on that interval. Therefore,

$$\begin{aligned} I_{\theta,\theta'}\{1:n\} &\geq \epsilon_4 \sum_{k=0}^{K} |V_k| 1\{|V_k| \geq L\} \\ &= \epsilon_4\left(n - \sum_{k=0}^{K} |V_k| 1\{|V_k| < L\}\right) \\ &\geq \epsilon_4[n - L(K+1)], \qquad \forall \theta' \in G_o(\theta). \end{aligned} \tag{4.19}$$

On the other hand, by (4.15) and (4.16) it follows that $d_\theta\{1:n\} \leq -C_1K + M_4$. Using (4.19) to eliminate K from this equation, we finally obtain

$$d_\theta\{1:n\} \leq -\frac{C_1}{L}n + (M_4 + C_1) + \frac{C_1}{L\epsilon_4} I_{\theta,\theta'}\{1:n\} \qquad \forall \theta' \in G_o(\theta),$$

which implies (ii) with $M_1 = \max\{M_4 + C_1, C_1/L\epsilon_4\}$ and $\delta_1 = C_1/L$.

(iii) Recall from Lemma 4.3(iii) that, under $\sigma^1(\theta)$, $\|x_n - x_\theta^*\|_\infty \leq \mu$ for every $n \geq 1$. By definition of $\sigma^*(\theta)$, this is valid under $\sigma^*(\theta)$ as well. Therefore, by Lemma 4.2,

$$d_{\theta'}(x_n, j_n) \leq -\delta_3 + M_3 I_{\theta',\theta}(x_n, j_n), \qquad \forall \theta' < \theta,$$

which implies (iii) for any $\delta_1 \leq \delta_3$, $M_1 \geq M_3$. Thus, the proof of Proposition 4.1 is complete. □

We turn now to the proof of Proposition 4.2, which proceeds through the following lemmas.

LEMMA 4.4. *Let $\theta \in \Theta$ be such that $B(\theta) \neq \emptyset$. Then there exists a strategy $\sigma^2(\theta)$ for player* 1 *and positive constants B_5, M_5, δ_5 such that, for every $\tau \in \mathcal{T}$ and $n \geq 1$, the following hold*:

(i) $d_\theta\{1:n\} \leq b(\theta)\min_{\theta' \in B(\theta)} I_{\theta,\theta'}\{1:n\} + B_5\sqrt{n}$.

(ii) $d_\theta\{1:n\} \leq -\delta_5 n + M_5 + M_5 \min_{\theta' > \theta} I_{\theta,\theta'}\{1:n\}$.

(iii) $\|x_n - x_\theta^*\|_\infty \leq \mu$, *with μ as in Lemma* 4.2.

PROOF. Consider some fixed θ such that $B(\theta) \neq \varnothing$. Let $x^o = x^o(\theta) \in \mathscr{P}(\mathscr{I})$ be a minimizer in (3.4), and let $\bar{\sigma}(\theta)$ be the strategy of Proposition 4.3. For any $0 < \epsilon < 1$, define the following strategy σ^ϵ:

$$\sigma^\epsilon\colon\ \sigma_t^\epsilon = (1-\mu)x_\theta^* + \mu\left[\epsilon x^o + (1-\epsilon)\bar{\sigma}(\theta)_t\right].$$

It will be proved that, for ϵ small enough, σ^ϵ satisfies the lemma.

Denote $d_o = d_\theta(x^o, y_\theta^*)$. Note that, by definition of x^o:

$$d_o = b(\theta) \min_{\theta' \in B(\theta)} I_{\theta,\theta'}(x^o, y_\theta^*) > 0. \tag{4.20}$$

Assume now that player 1 uses the strategy σ^ϵ, and player 2 any strategy $\tau \in \mathscr{T}$. Preceeding similarly to (4.10), (4.11), the loss may then be bounded by:

$$d_\theta\{1:n\} \leq \mu\epsilon n d_\theta(x^o, \bar{y}_n) - \mu(1-\epsilon)\left[\delta_\theta n\|\bar{y}_n - y_\theta^*\| - B\sqrt{n}\right],$$

where B and δ_θ are positive constants. Furthermore, note that

$$d_\theta(x^o, \bar{y}_n) \leq d_\theta(x^o, y_\theta^*) + \beta_1\|\bar{y}_n - y_\theta^*\|,$$

where $\beta_1 \triangleq \max_j |d_\theta(x^o, j)|$. Therefore,

$$d_\theta\{1:n\} \leq \mu\epsilon d_o n - \mu\left[\delta_\theta - \epsilon(\delta_\theta + \beta_1)\right]n\|\bar{y}_n - y_\theta^*\| + \mu B\sqrt{n}. \tag{4.21}$$

Next, the information will be lower-bounded. Consider first any $\theta' \in B(\theta)$. Then, by definition of σ^ϵ and (4.20):

$$\begin{aligned} I_{\theta,\theta'}\{1:n\} &\equiv \sum_{t=1}^{n} I_{\theta,\theta'}(x_t, j_t) \geq \sum_{t=1}^{n} \mu\epsilon I_{\theta,\theta'}(x^o, j_t) = \mu\epsilon n I_{\theta,\theta'}(x^o, \bar{y}_n) \\ &= \mu\epsilon n\left(I_{\theta,\theta'}(x^o, y_\theta^*) + I_{\theta,\theta'}(x^o, \bar{y}_n - y_\theta^*)\right) \\ &\geq \mu\epsilon n\left(b(\theta)^{-1}d_o - \beta_2\|\bar{y}_n - y_\theta^*\|\right), \qquad \theta' \in B(\theta), \end{aligned} \tag{4.22}$$

where $\beta_2 \triangleq \max_{j,\theta' > \theta} I_{\theta,\theta'}(x^o, j)$.

Consider now $\theta' \in G_o(\theta) = \{\theta' > \theta\} - B(\theta)$. Then, similarly,

$$\begin{aligned} I_{\theta,\theta'}\{1:n\} &\geq (1-\mu)nI_{\theta,\theta'}(x_\theta^*, \bar{y}_n) \\ &\geq (1-\mu)n\left(I_{\theta,\theta'}(x_\theta^*, y_\theta^*) - \beta_3\|\bar{y}_n - y_\theta^*\|\right) \\ &\geq (1-\mu)n\left(\beta_4 - \beta_3\|\bar{y}_n - y_\theta^*\|\right), \qquad \theta' \in G_o(\theta), \end{aligned} \tag{4.23}$$

where $\beta_3 \triangleq \max_{j,\theta' \in G_o(\theta)} I_{\theta,\theta'}(x_\theta^*, j)$, $\beta_4 \triangleq \min_{\theta' \in G_o(\theta)} I_{\theta,\theta'}(x_\theta^*, y_\theta^*)$. Note that $\beta_4 > 0$ by definition of $G_o(\theta)$ and $B(\theta)$.

Choose $\epsilon > 0$ small enough so that:

$$\epsilon(\delta_\theta + \beta_1 + b(\theta)\beta_2) \leq \tfrac{1}{2}\delta_\theta, \qquad \epsilon(\delta_\theta + \beta_1 + d_o\beta_3/\beta_4) \leq \tfrac{1}{2}\delta_\theta. \tag{4.24}$$

Assertions (i)–(iii) of the lemma now follow from the bounds derived above by simple algebra:

(i) Multiplying (4.22) by $b(\theta)$, subtracting from (4.21) and rearranging yields:

$$d_\theta\{1:n\} \leq b(\theta)I_{\theta,\theta'}\{1:n\} - \beta_5 n\|\bar{y}_n - y_\theta^*\| + \mu B\sqrt{n}, \qquad \theta' \in B(\theta),$$

where $\beta_5 = \mu[\delta_\theta - \epsilon(\delta_\theta + \beta_1 + b(\theta)\beta_2)]$. Since $\beta_5 \geq 0$ by the choice (4.24) of ϵ, (i) follows with $B_5 = \mu B$.

(ii) Using (4.22) to eliminate $\|\bar{y}_n - y_\theta^*\|$ from (4.21) gives:

$$(4.25) \qquad d_\theta\{1:n\} \leq -\beta_7 n + \beta_6 I_{\theta,\theta'}\{1:n\} + \mu B\sqrt{n}, \qquad \theta' \in B(\theta),$$

where $\beta_6 = \delta_\theta/\epsilon\beta_2$, $\beta_7 = \mu d_o[\delta_\theta - \epsilon(\delta_\theta + \beta_1 + b(\theta)\beta_2)]/b(\theta)\beta_2$. Note that $\beta_7 > 0$ by the choice (4.24) of ϵ. A similar calculation with (4.23) used instead of (4.22) gives:

$$(4.26) \qquad d_\theta\{1:n\} \leq -\beta_9 n + \beta_8 I_{\theta,\theta'}\{1:n\} + \mu B\sqrt{n}, \qquad \theta' \in G_o(\theta),$$

where $\beta_8 = \delta_\theta/(1-\mu)\beta_3$, $\beta_9 = \mu\beta_4\beta_3^{-1}[\delta_\theta - \epsilon(\delta_\theta + \beta_1 + d_0\beta_3/\beta_4)]$; note that $\beta_9 > 0$ by choice of ϵ. Combining (4.25) and (4.26) establishes (ii) with, e.g., $\delta_5 = (1/2)\min\{\beta_7, \beta_9\}$ and $M_5 = \max\{\beta_8, \beta_9, \max_{n \geq 1}(\mu B\sqrt{n} - \delta_5 n)\}$.

(iii) Follows directly from the definition of σ^ϵ. □

LEMMA 4.5. *For the strategy $\sigma^2(\theta)$ of the previous lemma, and under the same conditions,*

(i) *For every $\epsilon > 0$ there exists $M'(\epsilon) > 0$ such that:*

$$d_\theta\{1:n\} \leq (1+\epsilon)b(\theta)I_{\theta,\theta'}\{1:n\} + M'(\epsilon), \qquad \forall n \geq 1,\ \theta' \in B(\theta).$$

(ii) *Let $\epsilon_5 = \delta_5/(2M_5)$, $\delta_6 = \delta_5/2$. Then $\min_{\theta' > \theta} I_{\theta,\theta'}\{1:n\} \leq \epsilon_5 n$ implies $d_\theta\{1:n\} \leq -\delta_6 n + M_5$.*

PROOF. Since (ii) follows directly from Lemma 4.4(ii), it remains to prove (i). Assume first that $S_n \triangleq \min_{\theta' \in B(\theta)} I_{\theta,\theta'}\{1:n\} \leq \epsilon_5 n$. Noting that $B(\theta) \subset \{\theta' : \theta' > \theta\}$, it follows by item (ii) of the present lemma that $d_\theta\{1:n\} \leq M_5$. Assume next that $S_n > \epsilon_5 n$. Then, by Lemma 4.4(i):

$$(4.27) \quad d_\theta\{1:n\} \leq b(\theta)S_n + B_5\sqrt{n} \leq (1+\epsilon)b(\theta)S_n + \left(B_5\sqrt{n} - \epsilon b(\theta)\epsilon_5 n\right).$$

Let $M'(\epsilon) \triangleq \max\{M_5, \max_{n \geq 1}(B_5\sqrt{n} - \epsilon b(\theta)\epsilon_5 n)\}$; then (i) follows from the last two bounds. □

PROOF OF PROPOSITION 4.2. Similarly to the proof of Proposition 4.1, it is required to modify the strategy $\sigma^2(\theta)$ so that item (i) will hold on any interval $[m, \ldots, n]$. Thus, define

Strategy $\sigma^o(\theta)$. Defined similarly to $\sigma^*(\theta)$ in the proof of Proposition 4.1, except that $\sigma^1(\theta)$ in that definition is replaced by $\sigma^2(\theta)$ of Lemma 4.4.

Let $\tau \in \mathcal{T}$ and $n \geq 1$ be fixed. Retain the notations in the proof of Proposition 4.1 (i.e., C_1, m_k, V_k and K). We proceed to prove (i)–(iii) of Proposition 4.2.

(i) Consider a fixed $1 \leq m \leq n$. Let $0 \leq k \leq K$ be such that $m - 1 \in V_k$, and note that (4.15) and (4.17) hold true. Moreover, since $\sigma^2(\theta)$ is restarted at $t = m_K + 1$, it follows by Lemma 4.5(i) that for every $\epsilon > 0$,

$$(4.28)\qquad \begin{aligned} d_\theta\{m_k + 1 : n\} &= \sum_{k'=k}^{K} d_\theta\{V_{k'}\} \le d_\theta\{V_K\} \\ &\le (1+\epsilon) b(\theta) \min_{\theta' \in B(\theta)} I_{\theta,\theta'}\{V_K\} + M'(\epsilon) \\ &\le (1+\epsilon) b(\theta) \min_{\theta' \in B(\theta)} I_{\theta,\theta'}\{1:n\} + M'(\epsilon). \end{aligned}$$

Item (i) follows by subtracting (4.17) from (4.28), with $M(\epsilon) = M'(\epsilon) + C_1 + \hat{D}$.

(ii) By Lemma 4.5(ii), it follows that for every $0 \le k \le K$, $d_\theta\{V_k\} > -\delta_6 |V_k| + M_5$ implies

$$(4.29)\qquad \min_{\theta' > \theta} I_{\theta,\theta'}\{V_k\} > \epsilon_5 |V_k|.$$

Let $L > 0$ be a large enough constant so that $-\delta_6 L + M_5 < -C_1 - \hat{D}$. Then on each interval V_k such that $|V_k| \ge L$,

$$d_\theta\{V_k\} \ge -C_1 - \hat{D} > -\delta_6 L + M_5 \ge -\delta_6 |V_k| + M_5,$$

so that (4.29) holds on that interval. Therefore, for every $\theta' > \theta$,

$$(4.30)\qquad \sum_{k=0}^{K-1} I_{\theta,\theta'}\{V_k\} \ge \epsilon_5 \sum_{k=0}^{K-1} |V_k| 1\{|V_k| \ge L\} \ge \epsilon_5 (n - LK - |V_K|).$$

On the other hand, by (4.15) and Lemma 4.4(ii) (applied to the last interval),

$$(4.31)\qquad d_\theta\{1:n\} \le -C_1 K - \delta_5 |V_K| + M_5 + M_5 \min_{\theta' > \theta} I_{\theta,\theta'}\{V_K\}.$$

Using (4.30) to eliminate K from (4.31) and noting that $C_1/L < \delta_5$ by choice of L, it follows that for every $\theta' > \theta$,

$$\begin{aligned} d_\theta\{1:n\} &\le -\frac{C_1}{L}\left(-\frac{1}{\epsilon_5} \sum_{k=0}^{K-1} I_{\theta,\theta'}\{V_k\} + n - |V_K| \right) - \delta_5 |V_K| + M_5 + M_5 I_{\theta,\theta'}\{V_K\} \\ &\le -\frac{C_1}{L} n + M_2 I_{\theta,\theta'}\{1:n\} + M_2, \end{aligned}$$

where $M_2 \triangleq \max\{M_5, C_1/L\epsilon_5\}$. Thus, defining $\delta_2 \triangleq C_1/L > 0$, (ii) is established.

(iii) Follows by Lemma 4.4(iii) and Lemma 4.2, exactly as in the proof of Proposition 4.1(iii). □

5. The optimal strategy. We are now in a position to present a strategy σ^* which is asymptotically optimal. The following definitions will be required.

DEFINITION 5.1. Let $\{m_k\}_{k \ge 1}$ be a strictly increasing sequence of stopping times with respect to the history σ-algebras $\{H_n\}_{n \ge 0}$. Let σ be a given (behavioral) strategy of player 1. By the strategy σ restricted to the times $\{m_k\}$ we refer to the following selection rule at the times m_k, $k \ge 1$: $x_{m_k} = \sigma_k(\tilde{h}_k)$, where $\tilde{h}_k \triangleq \{i_{m_l}, j_{m_l}, a_{m_l}\}_{l=1}^{k-1}$.

Let $\hat{\theta}_t$, $\hat{\Theta}_t$, $\bar{\theta}_t$ and K_t be defined as in §4.1. For every $\theta \in \Theta$ and $t \geq 1$, define the following conditions $C_t^*(\theta)$ and $C_t^o(\theta)$:

$C_t^*(\theta)$: $\bar{\theta}_t = \theta$, and either $B(\theta) = \varnothing$ or else

$$\min_{\theta' \in B(\theta)} \sum_{s=1}^{t-1} I_{\theta,\theta'}(x_s, j_s)\mathbf{1}\{\bar{\theta}_s = \theta\} > \log K_t. \tag{5.1}$$

$C_t^o(\theta)$: $\bar{\theta}_t = \theta$, $B(\theta) \neq \varnothing$, and

$$\min_{\theta' \in B(\theta)} \sum_{s=1}^{t-1} I_{\theta,\theta'}(x_s, j_s)\mathbf{1}\{\bar{\theta}_s = \theta\} \leq \log K_t. \tag{5.2}$$

Note that exactly one of $C_t^*(\theta)$ and $C_t^o(\theta)$ is satisfied when $\bar{\theta}_t = \theta$. To introduce the optimal strategy, observe that for each fixed θ, the times t at which condition $C_t^*(\theta)$ [or $C_t^o(\theta)$] is satisfied form a sequence of increasing stopping times. We shall consider each such sequence separately, and apply on it a restricted version (according to Definition 5.1) of an appropriate substrategy.

Strategy σ^*. For $t = 1, 2, \ldots$. Denote $\bar{\theta} = \bar{\theta}_t$. If $C_t^*(\bar{\theta})$ is satisfied, then play according to the strategy $\sigma^*(\bar{\theta})$ of Proposition 4.1, restricted to the times t' when $C_{t'}^*(\bar{\theta})$ is satisfied. If $C_t^o(\bar{\theta})$ is satisfied, then play according to the strategy $\sigma^o(\bar{\theta})$ of Proposition 4.2, restricted to the times t' when $C_{t'}^o(\bar{\theta})$ is satisfied.

The strategy σ^* may be interpreted as follows. At each stage t, the value-biased MLE $\bar{\theta} = \bar{\theta}_t$ is computed. Then the level of information for discriminating $\bar{\theta}$ from $B(\bar{\theta})$ (quantified as in (5.1) or (5.2)) is evaluated, and compared with the critical level $\log K_t$ (which is slightly larger than $\log t$). If below that level, then the probing strategy $\sigma^o(\bar{\theta})$ is followed. The latter ensures that, if indeed $\bar{\theta} = \theta_0$, additional information will be obtained at a loss-to-information ratio close to $b(\theta_0)$, or else a *negative* loss will accumulate.

If the information level is above the critical level (or if $B(\bar{\theta})$ is empty), then the strategy $\sigma^*(\bar{\theta})$ is used. As discussed in §4.1, this strategy replaces the stationary strategy $\{x_{\bar{\theta}}\}$, and its stronger properties guarantee that the loss associated with the parameters in $\{B_2(\theta_0) \setminus B(\theta_0)\}$ is finite.

Observe that the "information level" in (5.1) and (5.2) is evaluated only over the times when the estimator was identical to the current one. This turns out to be important for the proof of the following theorem, which is the main result of this paper.

THEOREM 5.1. *The strategy σ^* is asymptotically optimal, in the sense of Definition* 3.2.

The rest of this section will be devoted to the proof of this result. The proof strategy is basically similar (and extends) that of Theorem 5.2 in Shimkin and Schwartz (1995) (quoted above as Theorem 4.1(ii)), and some of the results established there will be used here as well.

Assume henceforth that player 1 uses the strategy σ^*, and let $\theta_0 \in \Theta$, $\tau \in \mathcal{T}$, $n \geq 1$ be fixed. In what follows, all relations between random variables hold $P_{\theta_0}^{\sigma^*,\tau}$-a.s. Also, all constants (M, δ, Q, n_o etc.) are independent of τ and n, unless otherwise stated.

It will be convenient to use the abbreviated notation: $d_t = d_{\theta_0}(x_t, j_t)$, $(d_t)^+ = \max\{d_t, 0\}$, $\hat{D} = \max_{i,j} d_{\theta_0}(i, j)$, $I_{\theta_0,\theta}(t) = I_{\theta_0,\theta}(x_t, j_t)$, $E = E_{\theta_0}^{\sigma^*,\tau}$, and finally $l_n = \sum_{t=1}^n d_t \mathbf{1}\{\bar{\theta}_t \geq \theta_0\}$.

By (2.1),

$$(5.3)\qquad L_n^{\sigma^*,\tau}(\theta_o) = E\sum_{t=1}^{n} d_t = E\sum_{t=1}^{n} d_t\mathbf{1}\{\bar{\theta}_t < \theta_o\} + E\sum_{t=1}^{n} d_t\mathbf{1}\{\bar{\theta}_t \geq \theta_0\}$$

$$\leq \hat{D}E\sum_{t=1}^{n}\mathbf{1}\{\bar{\theta}_t < \theta_0\} + El_n \leq \hat{D}Q_1 + El_n,$$

where the last inequality is a basic property of the value-biased estimator $\bar{\theta}_n$, as established (for some $Q_1 < \infty$) (in Shimkin and Schwartz (1995, Lemma 5.1(ii)). We proceed then to bound El_n.

LEMMA 5.1. *For every $\epsilon > 0$, there exists a constant $M_o(\epsilon)$ such that,*

$$(5.4)\qquad l_n \leq (1+\epsilon)b(\theta_0)\log K_n + M_o(\epsilon) + \sum_{t=1}^{n}(d_t)^+\mathbf{1}\{l_t > 0, \bar{\theta}_t > \theta_0\},$$

where $b(\theta_0)$ is defined by (3.4) *if $B(\theta_0) \neq \varnothing$, and $b(\theta_0) \triangleq 0$ otherwise.*

PROOF. Noting Assumption A3, one has $l_n = l_n^a + l_n^b$, where

$$l_n^a = \sum_{t=1}^{n} d_t\mathbf{1}\{\bar{\theta}_t = \theta_0\},\qquad l_n^b = \sum_{t=1}^{n} d_t\mathbf{1}\{\bar{\theta}_t > \theta_0\}.$$

Define the stopping time $m = \max\{0 \leq t \leq n\colon l_t \leq 0\}$, where $l_o \triangleq 0$. Then

$$(5.5)\qquad l_n \leq l_n - l_m = (l_n^a - l_m^a) + (l_n^b - l_m^b).$$

Now,

$$(5.6)\qquad l_n^b - l_m^b \leq \sum_{t=m+1}^{n}(d_t)^+\mathbf{1}\{\bar{\theta}_t > \theta_0\} = \sum_{t=m+1}^{n}(d_t)^+\mathbf{1}\{l_t > 0, \bar{\theta}_t > \theta_0\}$$

$$\leq \sum_{t=1}^{n}(d_t)^+\mathbf{1}\{l_t > 0, \bar{\theta}_t > \theta_o\},$$

where the last equality follows by definition of m.

It remains to upper-bound the term:

$$(5.7)\qquad l_n^a - l_m^a = \sum_{t=m+1}^{n} d_t\mathbf{1}\{\bar{\theta}_t = \theta_0\}$$

$$= \sum_{t=m+1}^{n} d_t\mathbf{1}\{C_t^*(\theta_0)\} + \sum_{t=m+1}^{n} d_t\mathbf{1}\{C_t^o(\theta_0)\},$$

where $C_t^*(\theta_0)$ stands for "$C_t^*(\theta_0)$ is satisfied," and similarly for $C_t^o(\theta_0)$. Note that, by definition of σ^*, player 1's strategy on the times in which $C_t^*(\theta_0)$ is satisfied is the restriction of $\sigma^*(\theta_0)$ to these times. Therefore, by Proposition 4.1(i),

$$(5.8)\qquad \sum_{t=m+1}^{n} d_t\mathbf{1}\{C_t^*(\theta_0)\} \leq M_1.$$

To bound the term involving $\{C_t^o(\theta_0)\}$, note first that if $C_t^o(\theta_0)$ is not satisfied for any $m+1 \le t \le n$, then that term vanishes (this is trivially the case if $B(\theta_0)=\varnothing$). Otherwise, note that player 1's strategy on the times when $C_t^o(\theta_0)$ is satisfied is the restriction of $\sigma^o(\theta_0)$ to these times. Thus, by Proposition 4.2(i) it follows that for every $\epsilon > 0$,

$$\sum_{t=m+1}^{n} d_t \mathbf{1}\{C_t^o(\theta_0)\} \le (1+\epsilon) b(\theta_0) \min_{\theta \in B(\theta_0)} \sum_{t=m+1}^{n} I_{\theta_0,\theta}(t) \mathbf{1}\{C_t^o(\theta_0)\} + M(\epsilon).$$

Define $m' = \max\{m+1 \le t \le n \colon C_t^o(\theta_0) \text{ is satisfied}\}$. Then,

$$\min_{\theta \in B(\theta_0)} \sum_{t=m+1}^{n} I_{\theta_0,\theta}(t) \mathbf{1}\{C_t^o(\theta_0)\} \le \min_{\theta \in B(\theta_0)} \sum_{t=1}^{m'} I_{\theta_0,\theta}(t) \mathbf{1}\{C_t^o(\theta_0)\}$$
$$\le \hat{I} + \log K_{m'} \le \hat{I} + \log K_n,$$

where $\hat{I} = \max_{i,j,\theta} I_{\theta_0,\theta}(i,j)$, and the next to last inequality follows by definition of condition $C_t^o(\theta_0)$ (which is satisfied at $t=m'$). Thus,

$$(5.9)\quad \sum_{t=m+1}^{n} d_t \mathbf{1}\{C_t^*(\theta_0)\} \le (1+\epsilon) b(\theta_0) \log K_n + \left[(1+\epsilon) b(\theta_0) \hat{I} + M(\epsilon)\right],$$

(which holds trivially if $B(\theta_0)=\varnothing$, with $b(\theta_0)=0$).

The lemma now follows from (5.5)–(5.9), with $M_o(\epsilon) = M_1 + (1+\epsilon) b(\theta_0) \hat{I} + M(\epsilon)$. □

To upper-bound the (expected value of the) last term in (5.4), the following lemmas will be required.

LEMMA 5.2. *There exist positive constants* η *and* n_o *such that, for any* $n \ge n_o$, $l_n > 0$ *implies that at least one of the following conditions* $\Omega_1(n) - \Omega_4(n)$ *is satisfied*:

$\Omega_1(n)$: $\sum_{t=1}^n \mathbf{1}\{\bar{\theta}_t < \theta_0\} \ge \eta n$.

$\Omega_2(n)$: $\sum_{t=1}^n I_{\theta_0,\theta}(t)\mathbf{1}\{\bar{\theta}_t = \theta\} \ge \eta n$ *for some* $\theta > \theta_0$.

$\Omega_3(n)$: $\min_{\theta > \theta_0} \sum_{t=1}^n I_{\theta_0,\theta}(t) \ge \eta n$.

$\Omega_4(n)$: *both* $\Omega_{4a}(n)$ *and* $\Omega_{4b}(n)$ *below are satisfied*;

$\Omega_{4a}(n)$: $\min_{\theta \in G_o(\theta_0)} \sum_{t=1}^n I_{\theta_0,\theta}(t) \ge \eta n$, *where* $G_o(\theta_0) = \{\theta > \theta_0\} \setminus B(\theta_0)$.

$\Omega_{4b}(n)$: $N_n^*(\theta_0) \triangleq \sum_{t=1}^n \mathbf{1}\{C_t^*(\theta_0)\} \ge (1/2)n$.

PROOF. Let us first translate the relevant relations in Propositions 4.1 and 4.2 to the setting of the present strategy σ^*. For that purpose, define for each $\theta \in \Theta$:

$$N_n^*(\theta) = \sum_{t=1}^{n} \mathbf{1}\{C_t^*(\theta)\}, \qquad N_n^o(\theta) = \sum_{t=1}^{n} \mathbf{1}\{C_t^o(\theta)\},$$

$$N_n(\theta) = N_n^*(\theta) + N_n^o(\theta) = \sum_{t=1}^{n} \mathbf{1}\{\bar{\theta}_t = \theta\}.$$

Let $M_1, M_2, \delta_1, \delta_2$ be the constants for which Propositions 4.1 and 4.2 hold, and define $M = \max\{M_1, M_2\}$, $\delta = \min\{\delta_1, \delta_2\} > 0$. It then follows from items (i) and (ii) of Proposition 4.1 (upon substituting $\theta \leftarrow \theta_0$ and $\theta' \leftarrow \theta$) that

$$\sum_{t=1}^{n} d_t \mathbf{1}\{C_t^*(\theta_0)\} \le M, \tag{5.10}$$

$$\sum_{t=1}^{n} d_t \mathbf{1}\{C_t^*(\theta_0)\} \le -\delta N_n^*(\theta_0) + M + M \min_{\theta \in G_o(\theta_0)} \sum_{t=1}^{n} I_{\theta_0,\theta}(t)\mathbf{1}\{C_t^*(\theta_0)\}. \tag{5.11}$$

Similarly, by Proposition 4.2(ii),

$$\sum_{t=1}^{n} d_t \mathbf{1}\{C_t^o(\theta_0)\} \le -\delta N_n^o(\theta_0) + M + M \min_{\theta > \theta_0} \sum_{t=1}^{n} I_{\theta_0,\theta}(t)\mathbf{1}\{C_t^o(\theta_0)\}. \tag{5.12}$$

Finally, combining Propositions 4.1(iii) and 4.2(iii) (with $\theta' \leftarrow \theta_0$) gives

$$\sum_{t=1}^{n} d_t \mathbf{1}\{\bar{\theta}_t = \theta\} \le -\delta N_n(\theta) + M \sum_{t=1}^{n} I_{\theta_0,\theta}(t)\mathbf{1}\{\bar{\theta}_t = \theta\}, \qquad \forall \theta > \theta_0. \tag{5.13}$$

Assume now, in contradiction, that $\Omega_1(n) - \Omega_4(n)$ are false. It is required to show that $l_n \le 0$. Write $\overline{\Omega}_i(n)$ for '$\Omega_i(n)$ is false,' and let η be an arbitrary positive constant. Then by $\overline{\Omega}_3(n)$ and (5.12):

$$\sum_{t=1}^{n} d_t \mathbf{1}\{C_t^o(\theta_0)\} \le -\delta N_n^o(\theta_0) + M + M\eta n. \tag{5.14}$$

By (5.13) and $\overline{\Omega}_2(n)$,

$$\sum_{t=1}^{n} d_t \mathbf{1}\{\bar{\theta} = \theta\} \le -\delta N_n(\theta) + M\eta n, \qquad \forall \theta > \theta_0. \tag{5.15}$$

Note also that $\overline{\Omega}_4(n)$ implies that at least one of the following holds:
(a) $\overline{\Omega}_{4b}(n)$, i.e., $N_n^*(\theta_0) < (1/2)n$.
(b) $\Omega_{4b}(n)$ (i.e., $N_n^*(\theta_0) \ge (1/2)n$), and $\overline{\Omega}_{4a}(n)$.
We consider these two cases separately:
(a) By (5.10), (5.14) and (5.15),

$$\begin{aligned} l_n &= \sum_{t=1}^{n} d_t \mathbf{1}\{C_t^*(\theta_0)\} + \sum_{t=1}^{n} d_t \mathbf{1}\{C_t^o(\theta_0)\} + \sum_{\theta > \theta_0} \sum_{t=1}^{n} d_t \mathbf{1}\{\bar{\theta}_t = \theta\} \\ &\le -\delta \left[N_n^o(\theta_0) + \sum_{\theta > \theta_0} N_n(\theta) \right] + 2M + M|\Theta|\eta n. \end{aligned}$$

However

$$N_n^o(\theta_0) + \sum_{\theta > \theta_0} N_n(\theta) = n - N_n^*(\theta_0) - \sum_{\theta < \theta_0} N_n(\theta) \ge \tfrac{1}{2}n - \eta n,$$

which is implied by $\overline{\Omega}_{4b}(n)$ and $\overline{\Omega}_1(n)$, so that in case (a):

$$l_n \le -\tfrac{1}{2}\delta n + \eta(\delta + M|\Theta|)n + 2M. \tag{5.16}$$

(b) Since $\overline{\Omega}_{4a}(n)$ is assumed, it follows from (5.11) that

$$\sum_{t=1}^{n} d_t \mathbf{1}\{C_t^*(\theta_0)\} < -\delta N_n^*(\theta_0) + M + M\eta n. \tag{5.17}$$

Proceeding as in case (a), with (5.17) used in place of (5.10), we get:

$$l_n \leq -\delta\left[N_n(\theta_0) + \sum_{\theta > \theta_0} N_n(\theta)\right] + 2M + (M|\Theta| + M)\eta n. \tag{5.18}$$

Noting that $\overline{\Omega}_1(n)$ implies

$$N_n(\theta_0) + \sum_{\theta > \theta_0} N_n(\theta) = n - \sum_{\theta < \theta_0} N_n(\theta) \leq n - \eta n,$$

it follows that in case (b):

$$l_n \leq -\delta n + \eta(\delta + M|\Theta| + M)n + 2M. \tag{5.19}$$

It is obvious that for η small enough, both (5.16) and (5.19) imply that $l_n \leq 0$ for all n large enough. □

LEMMA 5.3. *Let $\Omega_4(n)$ be as defined in the previous lemma. Then*

$$E \sum_{n=1}^{\infty} (d_n)^+ \mathbf{1}\{\Omega_4(n), \overline{\theta}_n > \theta_0\} \leq Q_3 < \infty. \tag{5.20}$$

PROOF. By definition of $\Omega_4(n)$ and $G_o(\theta_0)$,

$$\begin{aligned}\mathbf{1}\{\Omega_4(n), \overline{\theta}_n > \theta_0\} &= \mathbf{1}\{\Omega_4(n), \overline{\theta}_n \in G_o(\theta_0)\} + \mathbf{1}\{\Omega_4(n), \overline{\theta}_n \in B(\theta_0)\} \\ &\leq \mathbf{1}\{\Omega_{4a}(n), \overline{\theta}_n \in G_o(\theta_0)\} + \mathbf{1}\{\Omega_{4b}(n), \overline{\theta}_n \in B(\theta_0)\}.\end{aligned}$$

We first claim that, for some $Q' < \infty$,

$$E \sum_{n=1}^{\infty} (d_n)^+ \mathbf{1}\{\Omega_{4a}(n), \overline{\theta}_n \in G_o(\theta_0)\} \leq \hat{D} \sum_{n=1}^{\infty} P\{\Omega_{4a}(n), \overline{\theta}_n \in G_o(\theta_0)\} \leq Q'.$$

The proof of the last bound is the same as that of Lemma 6.4(ii) in Shimkin and Schwartz (1995). Namely, by the union bound

$$P\{\Omega_{4a}(n), \overline{\theta}_n \in G_o(\theta_0)\} \leq \sum_{\theta \in G_o(\theta_0)} P\left\{\sum_{t=1}^{n} I_{\theta_0,\theta}(t) > \eta n, \overline{\theta}_t = \theta\right\},$$

and the rest is identical to the above-mentioned proof.

Thus, it remains to bound the term

$$J_3 = J_3(\tau) \triangleq E \sum_{n=1}^{\infty} (d_n)^+ \mathbf{1}\{\Omega_{4b}(n), \overline{\theta}_n \in B(\theta_0)\}.$$

As established in Shimkin and Schwartz (1995, Lemma 5.2), there exists a constant $M < \infty$ such that $d_{\theta_0}(x_\theta^*, j) \le MI_{\theta_0,\theta}(x_\theta^*, j)$ for every j and $\theta > \theta_0$ (hence, in particular, for $\theta \in B(\theta_0)$). Replacing Ω_{4b} by its definition, it follows that

$$J_3 \le \sum_{\theta \in B(\theta_0)} E \sum_{n=1}^{\infty} MI_{\theta_0,\theta}(n)\mathbf{1}\left\{N_n^*(\theta_0) \ge \frac{n}{2}, \bar{\theta}_n = \theta\right\}.$$

Now, $N_n^*(\theta_0) \ge (1/2)n$ implies that $C_m^*(\theta_0)$ is satisfied for some $(1/2)n \le m \le n$, which in turn implies that

$$U_n(\theta) \triangleq \sum_{t=1}^{n-1} I_{\theta_0,\theta}(t)\mathbf{1}\{\bar{\theta}_t = \theta_o\} \ge \log K_{[n/2]} \ge \log K_n - \alpha,$$

where the last inequality follows from the definition of $\{K_n\}$ in (4.1) for some finite constant α (independent of n). Noting further that $\bar{\theta}_n = \theta > \theta_o$ implies $\Lambda_{n-1}(\theta_o, \theta) \le \log K_n$, we finally get

$$\begin{aligned} J_3 &\le M \sum_{\theta \in B(\theta_0)} E \sum_{n=1}^{\infty} I_{\theta_0,\theta}(t)\mathbf{1}\{U_n(\theta) \ge \log K_n - \alpha, \Lambda_{n-1}(\theta_0, \theta) \le \log K_n\} \\ &\le M \sum_{\theta \in B(\theta_0)} Q(\theta) \triangleq Q'', \end{aligned}$$

where the last bound follows for finite constants $\{Q(\theta)\}$ by applying Lemma 3.3(v) of Shimkin and Schwartz (1995) (and the standard translation procedure as described there following equation (4.12)) to each $\theta \in B(\theta_0)$ separately. Thus, letting $Q_3 = Q' + Q''$, (5.20) is established. □

LEMMA 5.4. *The following bound holds*:

$$J_4 = J_4(\tau) \triangleq E \sum_{n=1}^{\infty} (d_n)^+ \mathbf{1}\{l_n > 0, \bar{\theta}_n > \theta_0\} \le Q_4 < \infty.$$

PROOF. By Lemma 5.2,

$$\begin{aligned} J_4 &\le \sum_{i=1}^{4} E \sum_{n=n_o}^{\infty} (d_n)^+ \mathbf{1}\{\Omega_i(n), \bar{\theta}_n > \theta_0\} + \hat{D}n_o \\ &\le \hat{D} \sum_{i=1}^{3} \sum_{n=1}^{\infty} P\{\Omega_i(n), \bar{\theta}_n > \theta_0\} + E \sum_{n=1}^{\infty} (d_n)^+ \mathbf{1}\{\Omega_4(n), \bar{\theta}_n > \theta_0\} + \hat{D}n_o. \end{aligned}$$

The three terms corresponding to $\Omega_1(n) - \Omega_3(n)$ have been bounded in Shimkin and Schwartz (1995, Lemma 6.4) (for any strategy σ of player 1). Therefore, the assertion follows by Lemma 5.3. □

The proof of Theorem 5.1 may now be concluded. By (5.3), Lemma 5.1 and Lemma 5.4, it follows that for every $\epsilon > 0$,

$$L_n^{\sigma^*,\tau}(\theta_0) \le \hat{D}Q_1 + (1 + \epsilon)b(\theta_0)\log K_n + M_o(\epsilon) + Q_4;$$

where $b(\theta_0) = 0$ if $B(\theta_0) = \varnothing$. Thus:

If $B(\theta_0) = \emptyset$, then

$$L_n^{\sigma^*}(\theta_0) = \sup_{\tau} L_n^{\sigma^*,\tau}(\theta_0) \le \hat{D}Q_1 + M_o(1) + Q_4 < \infty, \qquad \forall n \ge 1.$$

If $B(\theta_0) \neq \emptyset$, then

$$\limsup_{n\to\infty} \frac{L_n^{\sigma^*}(\theta_0)}{\log n} \le (1+\epsilon)b(\theta_0),$$

and the required result follows by letting $\epsilon \to 0$. □

Acknowledgement. This work was performed in part while Nahum Shimkin was with the Institute for Mathematics and Its Applications at the University of Minnesota.

References

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

Blackwell, D. (1954). Controlled random walks. *Proc. Internat. Congress Math.* **3**, 336–338.

____ (1956). An analogue for the minimax theorem for vector payoffs. *Pacific J. Math.* **6** 1–8.

Hannan, J. F. (1957). Approximation to Bayes risk in repeated play. *Contributions to the Theory of games*, Vol. 3, Princeton Univ. Press, Princeton, New Jersey, 97–139.

Lai, T. L., H. Robbins (1985). Asymptotically efficient adaptive allocation rules. *Adv. in Appl. Math.* **6** 4–22.

Parthasarathy, T., T. E. S. Ragahavan (1971). *Some Topics in Two-Person Games*, American Elsevier Publishing Co., New York.

Shimkin, N., A. Shwartz (1993). *Asymptotically efficient adaptive strategies in repeated games. Part 2: Asymptotic optimality.* Technical Report, University of Minnesota, IMA preprint series No. 1121.

____, A. Schwartz (1995). Asymptotically efficient adaptive strategies in repeated games. Part 1: Certainty equivalence strategies. *Math. Oper. Res.* **20** 743–767.

Sorin, S. (1980). *An introduction to two-person zero-sum repeated games with incomplete information.* IMSS-Economics, Technical Report No. 312, Stanford University.

N. Shimkin: Department of Electrical Engineering, Technion-Israel Institute of Technology, Haifa 32000, Israel

A. Shwartz: Department of Electrical Engineering, Technion-Israel Institute of Technology, Haifa 32000, Israel