The Empirical Bayes Envelope and Regret Minimization in Competitive Markov Decision Processes

Shie Mannor and Nahum Shimkin

Department of Electrical Engineering
Technion, Israel Institute of Technology
Haifa 32000, Israel

Revised, July 2002

Abstract. This paper proposes an extension of the regret minimizing framework from repeated matrix games to stochastic game models, under appropriate recurrence conditions. A decision maker (P1) who wishes to maximize his long-term average reward is facing a Markovian environment, which may also be affected by arbitrary actions of other agents. The latter are collectively modeled as a second player, P2, whose strategy is arbitrary. Both states and actions are fully observed by both players. While P1 may obviously secure the min-max value of the game, he may wish to improve on that when the opponent is not playing a worst-case strategy. For repeated matrix games, an achievable goal is presented by the Bayes envelope, that traces P1's best-response payoff against the observable frequencies of P2's actions. We propose a generalization to the stochastic game framework, under recurrence conditions that amount to fixed-state reachability. The *empirical Bayes envelope* (EBE) is defined as P1's best-response payoff against the stationary strategies of P2 that agree with the observed *state-action* frequencies. As the EBE may not be attainable in general, we consider its lower convex hull, the CBE, which is proved to be achievable by P1. The analysis relies on Blackwell's approachability theory. The CBE is lower bounded by the value of the game, and for irreducible games turns out to be strictly above the value whenever P2's frequencies deviate from a worst-case strategy. In the special case of single-controller games where P2 alone affects the state transitions, the EBE itself is shown to be attainable.

Keywords: Bayes envelope, Controlled Markov processes, Stochastic games, Regret minimization, Approachability.

1 Introduction

We consider a decision maker, or player (P1), who faces an arbitrary opponent (P2) in a dynamic environment. This situation is captured through a stochastic game model, where the opponent's strategy is arbitrary, and in particular is not subject to rationality considerations. From the modelling point of view, P2 may be used to model possible variations in P1's environment that are not captured within the basic Markovian model, such as the effect of other agents, and non-stationary moves of nature. A basic assumption, however, is that P2's actions, as well as the visited states, are observed by P1. Examples of such a scenario include routing in channels with selective availability, admission control in queues with unmodelled server vacations or service rate variations, path planning with occasionally blocked pathways, service scheduling in face of varying arrival rate, and chess playing against a "sub-optimal" opponent.

P1's goal is to maximize his long-term average reward. Obviously, he may secure for himself the maximin value of the stochastic game. However, since P2 does not in general play an optimal (worst-case) strategy, this goal may be too conservative. Indeed, the sequential nature of the game presents the opportunity to monitor the opponent's actions and adapt one's choices accordingly. For repeated matrix games, an elegant formulation of P1's goal has been introduced in Hannan (1957), using the *Bayes envelope* concept. In this paper we propose as extension of this concept to the stochastic game model, under appropriate recurrence conditions. Specifically, we assume that a fixed state is reachable by each player from any state (Assumption 1). This implies, in particular, the existence of ϵ-optimal strategies in the zero-sum game. In addition, finite state and action spaces are assumed throughout.

This framework may be compared to adaptive control schemes for Markov decision processes (MDPs). Had P2 been restricted to stationary strategies, then well established methods of adaptive control (see Kumar and Varaiya (1986), Bertsekas and Tsitsiklis (1995)) may be used to obtain the best-response strategy and reward against the stationary strategy employed by P2. However, such schemes provide no performance guarantees against non-stationary strategies, and may indeed fail to deliver even the value of the game against such strategies. Our goal here is to provide explicit performance guarantees that hold for general strategies of P2.

For repeated matrix games, the Bayes envelope is defined as the maximal reward that P1 could achieve against the observed relative frequencies of the opponent's actions. More precisely, given the sequence $\{b_k\}_{k=1,\ldots,t}$ of P2's actions in a repeated game with reward matrix $r(a,b)$, the relative frequencies are $y_t(b) = \frac{1}{t}\sum_{k=1}^{t} 1\{b_k = b\}$, and the Bayes envelope at $y_t = y$ is given by $r^*(y) \overset{\triangle}{=} \max_a \sum_b r(a,b)y(b)$. It was established by Hannan (1957) that the Bayes envelope may be asymptotically attained in such games, in the sense that $\hat{r}_t - r^*(y_t)$ is asymptotically non-negative, for any strategy of the second player. This result was subsequently proved by Blackwell

(1956b) using the theory of approachability for repeated games with vector payoffs. Strategies that attain the Bayes envelope are often referred to as *regret-minimizing* or *no-regret* strategies. While these classical results rely on perfect monitoring of the opponent's actions, recent extension consider the case where only some related signal may be observed. Auer et al. (1995) and Freund and Schapire (1999) assume that the signal consists of the actual reward at each stage, and show that no-regret strategies exist in this case as well. An extension to a general signal structure has been carried out in Rustichini (1999), where existence of no-regret strategies with respect to an appropriately relaxed Bayes envelope was established.

In extending the regret minimization framework to stochastic games, two distinct approaches may be pursued. The one we consider here is based on state-action frequencies: both the modified Bayes envelope we propose and the corresponding regret-minimizing strategies depend only on the empirical frequencies of the observed state and (opponent's) action. Another possible approach could be based on an analogy between single-stage actions in repeated matrix games and finite-interval strategies in the stochastic game model. This results in a highly complex scheme, which we do not pursue in this paper save for some comments in the concluding section.

The *empirical Bayes envelope* (EBE) for stochastic games is defined in the state-action frequency space, as the best reward that P1 can secure against the set of stationary strategies of P2 which are compatible with the observed state-action frequencies. Two different notions of compatibility are considered, leading to two variants of the EBE. In either case, it turns out that the EBE is *not* attainable in general. We therefore revert to its lower convex hull, the CBE, and show that this relaxed envelope is indeed attainable. The performance level guaranteed by the CBE is, trivially, never below the value of the game. In the case of *irreducible* games, we show that the CBE is strictly larger than the value whenever the empirical frequencies of the opponent's actions deviate from a minimax strategy.

A specific case of interest is the single-controller game, where P2 alone determines the state transitions. In this case the EBE itself turns out to be convex, and therefore attainable. In fact, using direct analysis this result is established without any recurrence conditions.

The paper is organized as follows. Section 2 presents the stochastic game model and recalls the basic approachability results that are needed here. In Section 3 we define the empirical Bayes envelope for stochastic games, prove that its convex hull, the CBE, is attainable, and establish the basic properties and performance guarantees provided by the latter. Section 4 considers the single-controller case along with some applications. In Section 5 we present an example of a game in which EBE is not attainable. Some concluding remarks are offered in Section 6.

2 Model and Preliminaries

2.1 The Model

We consider a two-person stochastic game model (e.g., Filar and Vrieze, 1996) with finite state and action spaces. We refer to the players as P1 (the regret-minimizing player) and P2. The model is defined by the five-tuple $(\mathcal{S}, \mathcal{A}, \mathcal{B}, P, r)$, where:

1. $\mathcal{S}$ is the finite set of states of the Markov process, $\mathcal{S} = \{1, \ldots, S\}$.
2. $\mathcal{A}$ is the set of actions of P1 in each state, $\mathcal{A} = \{1, \ldots, A\}$.
3. $\mathcal{B}$ is the set of actions of P2 in each state, $\mathcal{B} = \{1, \ldots, B\}$.
4. P is the state transition kernel, $P : \mathcal{S} \times \mathcal{A} \times \mathcal{B} \times \mathcal{S} \to [0, 1]$ such that $P(s_2|s_1, a_1, b_1)$ is the probability that the next state is s_2 given that current state is s_1, P1 plays a_1 and P2 plays b_1.
5. r is P1's reward function, $r : \mathcal{S} \times \mathcal{A} \times \mathcal{B} \to \mathrm{I\!R}$. $r(s, a, b)$ is the reward obtained when P1 plays a, P2 plays b and the current state is s.

At each time epoch $t = 1, 2, \ldots$, both players observe the current state s_t, and then P1 and P2 choose actions a_t and b_t, respectively. As a result P1 receives a reward of $r_t = r(s_t, a_t, b_t)$ and the next state is determined according to the transition probability $P(\cdot|s_t, a_t, b_t)$. A behavioral strategy $\sigma_1 \in \Sigma_1$ for P1 is a mapping from all possible histories to the mixed action Δ^A, where Δ^A is the set of probability vectors over $\mathcal{A}$. Similarly, a strategy $\sigma_2 \in \Sigma_2$ for P2 is a mapping from all possible histories to the mixed action Δ^B. A strategy of either players is *stationary* if the mixed action at each time instant t depends only on the state s_t. A stationary strategy g for P2 is thus identified with the vector of conditional probabilities $g = (g(b|s)) \in (\Delta^B)^S$, and similarly for a stationary strategy f for P1. The set of stationary strategies of P1 (resp. P2) is denoted by F (resp. G), and the set of stationary deterministic strategies of P1 is denoted by F_D. Let $P^s_{\sigma_1,\sigma_2}$ denote the probability measure induced on the sequence of states and actions by the strategy pair σ_1 and σ_2 and initial state s, and let $E^s_{\sigma_1,\sigma_2}$ denote the corresponding expectation operator. We are primarily interested in the average reward:

$$\hat{r}_t = \frac{1}{t} \sum_{\tau=1}^{t} r_\tau \,. \tag{2.1}$$

We shall also occasionally consider a related vector-valued stochastic game which has the same transition structure as the original game, but a vector-valued reward function. We denote the corresponding reward vector at time t by $m_t = m(s_t, a_t, b_t) \in \mathrm{I\!R}^k$, where $k > 1$ is a given integer.

More generally, the reward m_t may be stochastic, in which case $m(s_t, a_t, b_t)$ denotes its mean, and finite second moments are assumed. In a similar manner to (2.1) the average reward vector is defined as $\hat{m}_t = \frac{1}{t}\sum_{\tau=1}^{t} m_\tau$. Given the vector-valued stochastic game and a vector $u \in \mathbb{R}^k$, we define the projected game $\Gamma_s(u)$ as the zero-sum stochastic game with the scalar reward given by $r_t = m_t \cdot u$ ($\cdot$ is the inner product in $\mathbb{R}^k$) and initial state s. $\Gamma_s(u)$ has a value $\mathrm{v}\Gamma_s(u)$ if

$$\begin{aligned} \mathrm{v}\Gamma_s(u) &= \sup_{\sigma_1}\inf_{\sigma_2}\liminf_{t\to\infty} E^s_{\sigma_1\sigma_2}(\hat{m}_t \cdot u) \\ &= \inf_{\sigma_2}\sup_{\sigma_1}\limsup_{t\to\infty} E^s_{\sigma_1\sigma_2}(\hat{m}_t \cdot u)\,. \end{aligned} \tag{2.2}$$

As is well known (Mertens and Neyman, 1981) this value exists for any finite stochastic game. Our basic assumption regarding the state transition structure of the model is the following:

Assumption 1 *The game is 0-reachable by each of the players. This means that there exists a particular state ("state 0") and a strategy for each player which guarantees that state 0 is reached with probability 1 for any strategy of the other player and from any initial state.*

The strategy of either player which is used to reach state 0 can always be taken as a stationary one (e.g., Lemma 3.3 in Mannor and Shimkin, 2000b). Assumption 1 implies the existence of ϵ-optimal stationary strategies for both players (see, e.g., Filar and Vrieze, 1996). Hence the infimum and supremum in (2.2) can be taken over stationary strategies. Moreover, the value of the game is independent of the initial state. We note that the existence of optimal stationary strategies requires stronger conditions, e.g., the existence of a state which is recurrent under *all* (stationary) strategies, as assumed in Shimkin and Shwartz (1993).

In principle, our results may well apply under other conditions on the state transition structure which guarantee ϵ-optimal stationary strategies. For other conditions which are somehow less explicit see Filar and Vrieze (1996), Flesch et al. (1998). Observe further that (ϵ-) optimality in stationary strategies is an underlying assumption in our formulation of the EBE, and particularly concerning its safety properties with respect to the value of the game.

For a pair of stationary strategies f and g for P1 and P2, respectively, we denote by $r(f,g)$ the expected average reward of the two stationary strategies from state 0, that is

$$r(f,g) \triangleq \lim_{t\to\infty} E^0_{f,g}(\hat{r}_t)\,.$$

Note that this is well defined regardless of any recurrence assumption. We shall also use the notion of *best response reward.* Given a stationary strategy g for P2, the best response reward is given by

$$r^{BR}(g) \triangleq \max_{f\in F} r(f,g)\,.$$

Note that the max is obtained on F, since that for a fixed stationary g the best response strategy can be taken stationary (and even deterministic).

2.2 Approachability Theory

We next recall some definitions and results concerning approachability in stochastic games with vector rewards. The theory of approachability for repeated matrix games with vector rewards has found various uses in game theory and related applications; see, e.g., Fudenberg and Levine (1995), Hart and Mas-Colell (2000), Lehrer (1998), Spinat (1999) and the recent special issue Vohra et al. (1999). We describe here the extension to stochastic games in Mannor and Shimkin (2000b) which applies under the the recurrence conditions assumed in this paper.

Let $d(\cdot,\cdot)$ denote the Euclidean norm in $\mathbb{R}^k$. Recall that given a set of probability measures $\{P_\gamma\}_{\gamma\in\Gamma}$ defined on a common measurable space, and a sequence X_t of random variables on that space such that $X_t \to 0$ P_γ a.s. (almost surely) for every $\gamma \in \Gamma$, we say that $X_t \to 0$ *at a uniform rate* over Γ if for every $\epsilon > 0$

$$\lim_{T\to\infty} \sup_{\gamma\in\Gamma} P_\gamma \left(\sup_{t\geq T} |X_t| > \epsilon \right) = 0\,.$$

Definition 2.1 *A set $B \subseteq \mathbb{R}^k$ is* approachable *from state s by P1 if there exists a B-*approaching *strategy σ_1^* such that*

$$d(\hat{m}_t, B) \to 0 \quad P^s_{\sigma_1^*\sigma_2}\text{ -a.s.}\quad \text{for every } \sigma_2\,,$$

at a uniform rate over $\sigma_2 \in \Sigma_2$. B is excludable *from s by P2 if there exists a B-*excluding *strategy $\sigma_2^* \in \Sigma_2$ such that for some $\delta > 0$, and every $\sigma_1 \in \Sigma_1$*

$$\liminf_{t\to\infty} d(\hat{m}_t, B) > \delta \quad P^s_{\sigma_1\sigma_2^*}\text{ -a.s.}\quad \text{for every } \sigma_1\,,$$

at a uniform rate over $\sigma_1 \in \Sigma_1$.

A set is said to be approachable if it is approachable from all states, and similarly it is excludable if it is excludable from all states.

Sufficient conditions for approachability in stochastic games have been presented by Shimkin and Shwartz (1993), under a single-state recurrence condition (namely, that a fixed state is recurrent under *all* strategies), and refined by Mannor and Shimkin (2000b) for games which are 0-reachable. These conditions essentially carry over the geometric insight of Blackwell (1956a) to the stochastic game model, with stationary strategies replacing single-stage actions. As for approaching strategies, the basic idea is to switch between stationary strategies at appropriate switching instants, with the next strategy chosen as the optimal one in the u-projected game (cf. $\Gamma(u)$ above), where u is the direction of a shortest distance from the current reward vector to the set B. The main difference between the two recurrence conditions is in the choice of switching times: Under the conditions of Shimkin and Shwartz (1993), switching between stationary strategies can simply be performed whenever the recurrent reference state is reached. Under the weaker 0-reachability condition, the switching scheme becomes more involved and requires a slow increase of the switching intervals. We refer the reader to the respective papers for a detailed description. See also Milman (2000) for a different approach which defines approachability from all states simultaneously, and relies on a uniform convergence result for general stochastic games.

We shall require here the following result on approachability of *convex* sets. For every pair of stationary strategies $f \in F$ of P1 and $g \in G$ of P2, let $m(f,g) \overset{\triangle}{=} \lim_{t\to\infty} E^0_{f,g}\hat{m}_t$ denote the long-run average expected reward. Note that this is well defined (irrespective of any recurrence assumptions) for stationary strategies as the state space is finite and we fix the initial state. Let

$$M(F,g) \overset{\triangle}{=} \overline{\mathrm{co}}(\{m(f,g), f \in F\}) \subseteq \mathrm{I\!R}^k$$

denote the closed convex hull of the set of average rewards that P1 can obtain against g in stationary strategies. The next theorem characterizes convex approachable sets for games which are 0-reachable by P1.

Theorem 2.1 *(Mannor and Shimkin, 2000b, Theorem 4.1) Consider a closed convex set $C \subseteq \mathrm{I\!R}^k$, and assume that the game is 0-reachable by both players. Then*

(i) C is approachable if and only if either one of the following equivalent conditions is satisfied:

a. $M(F,g) \cap C \neq \emptyset$ for every stationary strategy $g \in G$.

b. $\mathrm{v}\Gamma_0(u) \geq \min_{y\in C}(y \cdot u)$ for every $u \in U(C)$. Here $U(C)$ is the set of all unit vectors u in $\mathrm{I\!R}^k$ which are in the direction of the shortest distance from some $x \notin C$ to C.

(ii) If C is not approachable then it is excludable by P2.

Remark 1: Theorem 2.1 does not provide convergence rate guarantees. Under the stronger assumption that state 0 is recurrent for every pair of stationary strategies, the following bound on the convergence rate was obtained by Shimkin and Shwartz (1993):

$$P\left\{\sup_{t\geq T} d(\hat{m}_t, C) > \epsilon\right\} \leq \frac{Q}{\epsilon^2 T},$$

where Q depends only on the parameters of the stochastic game. It can be shown in a similar manner to Mertens et al. (1994) that also $E\left(d(\hat{m}_t, C)^2\right) < \frac{Q'}{t}$, where Q' depends only on the game parameters. Furthermore, it is not difficult to prove that $d(\hat{m}_t, C)$ behaves almost surely like $\Theta(t^{-1/3})$ (that is $\liminf_{t\to\infty} \log(d(\hat{m}_t, C))/log(t) \leq -1/3$), see Mertens et al., 1994).

3 The Empirical Bayes Envelope for Stochastic Games

In this section we define the empirical Bayes envelope (*EBE*) for the stochastic game model, prove our main results regarding the attainability of its convex hull, the *CBE*, and establish the basic properties of these performance envelopes. The empirical Bayes envelope is defined as the best-response reward of P1 (in stationary strategies) to the stationary strategies of P2 that are compatible with the empirical state-action frequencies of P2. A similar line of reasoning leads to the standard Bayes envelope for repeated matrix games (Hannan, 1957), and our definitions reduce to the latter when specialized to a single-state stochastic game.

We suggest in fact two alternative definitions for the *EBE*. In the first, we identify a compatible strategy of P2 by considering the state-action frequencies at each state separately. The mixed action at each state is simply defined as the empirical distribution of P2's actions taken at that state. We refer to the resulting envelope as the *state-based EBE*. The second definition, the *dynamics-based EBE*, does take the state transition structure into account, by requiring that a compatible strategy of P2 can give rise to the observed state-action frequencies when paired with *some* stationary strategy of P1. As such strategies need not exist for every vector of state-action frequencies, the *EBE* is only partially defined in this case. The extension to the entire space is handled by introducing the lower convex hull of that envelope, which is in fact required in both cases to obtain an attainable envelope. Comparing the two proposed envelopes, the first is simpler, but is dominated from above (hence more conservative) by the second.

Let us start by defining formally the concept of reward envelope in the space of state-action frequencies. Denote by $y_t(s,b)$ the t-stage relative frequency of P2's action b in state s, namely

$$y_t(s,b) = \frac{1}{t}\sum_{\tau=1}^{t} 1\{s_\tau = s, b_\tau = b\}\,, \tag{3.1}$$

and let $y_t \in \Delta^{SB}$ be the state-action frequency vector. A reward *envelope* is a function $h : \Delta^{SB} \to \mathbb{R}$. We shall be looking for envelopes that can be attained in the sense of the following definition.

Definition 3.1 *The reward envelope* $h : \Delta^{SB} \to \mathbb{R}$ *is* attainable *by P1 if there exists a strategy* σ_1^* *such that*

$$\liminf_{t\to\infty} (\hat{r}_t - h(y_t)) \geq 0 \quad P^s_{\sigma_1^*\sigma_2}\text{-}a.s.,$$

for every strategy σ_2 *and initial state* s*, at a uniform rate over* $\sigma_2 \in \Sigma_2$

We refer to the strategy σ_1^* in this definition as an h-attaining strategy for P1.

In the sequel we shall use the *continuous lower convex hull* of certain functions. Recall that for a function $h : X \to \mathbb{R}\cup\infty$, with X a convex compact set, the lower convex hull is the largest function $h^c : X \to \mathbb{R}\cup\infty$ which is nowhere larger than h and is convex (e.g., Rockafellar, 1970). The lower convex hull is continuous in the (relative) interior of its domain $D = \{y \in X : h^c(y) < \infty\}$, with possible discontinuities at the boundary. The *closure* of the lower convex hull, h^{cl}, is obtained by extending h^c continuously to the boundary of its domain (see Stoer and Witzgall, 1970), and is also a convex function. Finally, to streamline the definitions we define the continuous lower convex hull of h as a *continuous* function h^{cc} which coincides with h^{cl} on D; the continuous extension outside of D may be arbitrarily chosen, and need not be specified here. Henceforth we shall use simply h^c to denote the continuous lower convex hull of a function h.

3.1 State-based *EBE*

In this subsection we give our first definition of the *EBE*. This will be based on compatible strategies of P2 that are prescribed on a state-by-state basis.

Consider for now a state-action frequency vector y so that $y(s) \stackrel{\triangle}{=} \sum_{b=1}^{B} y(s,b) > 0$ for every state s. The associated stationary strategy $g = g^+(y)$ of P2 is defined as

$$g^+(y)(b|s) = \frac{y(s,b)}{y(s)}\,. \tag{3.2}$$

Obviously, if P2 uses a stationary strategy, the latter provides a consistent estimate of that strategy (at least in those states which are visited infinitely often). We can now define the empirical Bayes envelope at y as:

$$r^+(y) \stackrel{\triangle}{=} r^{BR}(g^+(y))\,. \tag{3.3}$$

This definition reduces to the standard definition of the Bayes envelope for repeated matrix games when the state space is reduced to a singleton.

When $y(s) = 0$ for some state s, a slight generalization is required. Since no information is available to determine P2's choice in such states, we allow for an arbitrary choice of P2's actions there. More precisely, the set of all stationary strategies that are consistent with y is specified by:

$$G^+(y) = \{g \in G : y(s,b) = y(s)g(b|s) \ \ \forall (s,b) \in \mathcal{S} \times \mathcal{B}\}\,. \tag{3.4}$$

As noted, $G^+(y)$ is a singleton when every state has been visited (in particular, in the interior of Δ^{SB}), and in any case it is non empty, convex, and closed. The empirical Bayes envelope can now be defined in general by:

$$r^+(y) \stackrel{\triangle}{=} \inf_{g \in G^+(y)} r^{BR}(g) = \inf_{g \in G^+(y)} \sup_{f \in F} r(f,g)\,. \tag{3.5}$$

Note that $r^+(y)$ is finite for every $y \in \Delta^{SB}$. When $G^+(y)$ is a singleton, namely it equals $g^+(y)$, $r^+(y)$ is actually the optimal value of the Markov decision process which is induced by $g^+(y)$. Otherwise, $r^+(y)$ is the value of the stochastic game which is obtained from the original one by fixing P2's actions at those states for which $y(s) \neq 0$. It can also be shown that r^+ is continuous in the interior of Δ^{SB}. It is obvious that $r^+(y)$ is never below the minimax value of the game, since the infimum in (3.5) is over a *subset* of G, which is P2's set of stationary strategies in the original game (recall that under our recurrence assumption, the game has ϵ-optimal stationary strategies). Moreover, it is strictly above the value when $G^+(y)$ does not include an optimal strategy of P1.

Unfortunately, $r^+(y)$ need not be attainable in general; a counterexample is given Section 5. For that reason we will consider the lower convex hull of $r^+(y)$, denoted by r^{+c}, which is shown to be attainable in Theorem 3.1. In fact, we will show the attainability of a higher reward envelope which is defined next.

3.2 The Dynamics-based EBE and the CBE

The definition of r^+ above did not take into account the state-transition dynamics for the purpose of estimating P2's strategy, but rather considered each state in isolation. We next define an envelope which does take the game dynamics into consideration. For any pair (f, g) of stationary strategies, the limiting state-action occupation measure $y_{f,g}$ is defined as

$$y_{f,g}(s,b) \overset{\triangle}{=} \lim_{t\to\infty} E^0_{f,g}(y_t(s,b)) .$$

The limiting state occupation measure is defined similarly, and obviously coincides with the invariant measure when the latter exists. We shall identify the set of P2's strategies that are compatible with y as those strategies that can give rise to y as the limiting occupation measure (when paired with *some* stationary strategy of P1).

The definition in (3.4) of compatible strategies is thus refined as follows:

$$\begin{aligned} G(y) \quad &= \{g \in G\colon \;\; y(s,b) = y(s)g(b|s) \quad \forall (s,b) \in \mathcal{S} \times \mathcal{B} \\ &\qquad \text{and } \exists f \in F \text{ s.t. } y(s) = y_{f,g}(s) \quad \forall s \in \mathcal{S}\} . \end{aligned} \tag{3.6}$$

Obviously, $G(y) \subseteq G^+(y)$. Note that (3.6) implies that $y(s,b) = y_{f,g}(s,b)$, since $y_{f,g}(s,b) = y_{f,g}(s)g(b|s)$. Note further that (3.6) can be written similarly to (3.4) as

$$G(y) = \{g \in G : \exists f \in F \text{ s.t. } y(s,b) = y_{f,g}(s)g(b|s) \quad \forall (s,b) \in \mathcal{S} \times \mathcal{B}\} ,$$

since the latter definition implies that $y(s) = \sum_b y(s,b) = \sum_b y_{f,g}(s)g(b|s) = y_{f,g}(s)$.

We can now present our first attempt at the definition of the dynamics-based EBE:

$$r^*(y) \overset{\triangle}{=} \inf_{g\in G(y)} r^{BR}(g) = \inf_{g\in G(y)} \sup_{f\in F} r(f,g) . \tag{3.7}$$

If $G(y) = \emptyset$ then we set $r^*(y) = \infty$.

Note that $r^+(y) \leq r^*(y)$, since $G(y) \subseteq G^+(y)$. As in the case of r^+, we can show that the continuous lower convex hull r^c of r^* is attainable. Thus, r^* dominates r^+ in terms of its performance guarantees.

Still, there exists an improved envelope that is both attainable and easier to calculate. For each stationary strategy g of P2, let $f(g)$ be an optimal strategy of P1 in the Markov Decision

Process induced by g. As is well known, $f(g)$ can always be chosen stationary and deterministic. Redefine r^* according to (3.7) for $y \in \{y_{f(g),g} : g \in G\}$, and as ∞ outside of this set. Formally, let

$$G^*(y) \overset{\triangle}{=} \{g \in G : \text{ so that } y(s,b) = y_{f(g),g}(s)g(b|s) \;\; \forall (s,b) \in \mathcal{S} \times \mathcal{B}\} .$$

The dynamic-based Bayes envelop is defined as:

$$r^*(y) = \inf_{g \in G^*(y)} r^{BR}(g) . \tag{3.8}$$

Note that r^* is now assigned a finite value on a smaller set of points than before, as $G^*(y) \subseteq G(y)$. Thus, its lower convex hull will in general be higher. We remind that $G^*(y)$ is at most a singleton when $y(s) > 0$ for all s. Moreover, when $G^*(y)$ is not a singleton, it is not difficult to verify that the same reward r^{BR} is obtained for all $g \in G^*(y)$.

We define r^c, the dynamics-based convex Bayes envelope (CBE):

Definition 3.2 *The dynamics-based convex Bayes envelope, $r^c : \Delta^{SB} \to \mathbb{R}$, is the continuous lower convex hull of r^* as given in (3.8).*

Remark 2: It should be noted that different choices of $f(g)$ may result in different envelopes, see Example 2. Choosing $f(g)$ as a *deterministic* optimal strategy against g, results in an envelope which can be calculated. For each of the deterministic strategies $f \in F_d$, one can calculate P2's possible state-action frequency vectors as the polyhedron which is induced by the Markov Decision Process with f fixed (this is a polyhedron by Theorem 8.9.3 from Puterman, 1994). The reward for each polyhedron is a linear functional. This means that given $g \in G$, one can calculate the finitely many state-action frequency vector $y_{f,g}$ for $f \in F_d$ and for each of them find the reward which is induced by a linear functional. Moreover, r^c itself can be calculated since it is the lower convex hull of points on the polyhedrons. This requires finding $f(g)$ for every g. If, for example, the game is irreducible then finding $f(g)$ amounts to finding the maximum of finitely many rational functions. In any case, finding $f(g)$ can be done numerically by finding the intersection between the state-action polyhedron induced by g with each of the state-action polyhedrons induced by the deterministic strategies $f \in F_D$. This calculation requires solving several systems of linear equations.

Since $r^*(y) \geq r^+(y)$, the same holds for their lower convex hulls: $r^c \geq r^{+c}$. We now provide an example in which the inequality is strict, thus showing that the dynamics-based CBE is indeed a strict improvement of the state-based CBE.

Example 1: Consider a game with three states, $\mathcal{S} = \{s^1, s^2, s^3\}$. P1 and P2 may choose actions only at state s^1, in which $\mathcal{A} = \mathcal{B} = \{1, 2\}$. The transition probabilities are: $p(s^1|s^2) = p(s^1|s^3) = 1$ and $p(s^2|s^1, a, b) = 0$ if $a = b$ and 1 if $a \neq b$. The reward is 0 in states s^2, s^3 and a matching penny game in state s^1 (i.e., $r(s^1,a,b)=1$ if $a=b$ and 0 if $a\neq b$). The state-action frequency vector is $y=(y(s^1,1),y(s^1,2),y(s^2),y(s^3))$. P2's strategy g is defined by the frequency of action 1 in state s^1. It can be easily seen that when $g=1$, $y_1=(1/2,0,\alpha/2,(1-\alpha)/2), 0\leq\alpha\leq 1$; when $g=1/2$, $y_{1/2}=(1/4,1/4,1/4,1/4)$; and that when $g=0$, $y_0=(0,1/2,\alpha/2,(1-\alpha)/2), 0\leq\alpha\leq 1$. The point $y=(1/4,1/4,1/2,0)$ is not feasible, but it is in the convex hull of feasible points (the mean of $(1/2,0,1/2,0)$ and $(0,1/2,1/2,0)$). It so happens that $r^{+c}(y)=r^+(y)=\mathrm{v}=1/4$ but $r^c(y)=\frac{1}{2}r^*(1/2,0,1/2,0)+\frac{1}{2}r^*(0,1/2,1/2,0)=1/2$. Note that $r^c(y)=1/2$ (strict equality) since the only points that take finite values of r^* and give y as their convex combination are $(1/2,0,1/2,0)$ and $(0,1/2,1/2,0)$. ∎

3.3 Properties of EBE and CBE

We start by showing that CBE is attainable. We focus on the dynamics-based CBE as the state-based CBE is lower - its attainability is therefore implied.

Theorem 3.1 *Suppose that the game is 0-reachable (Assumption 1). Then the envelope $r^c : \Delta^{SB}\to \mathrm{I\!R}$ (Definition 3.2) is attainable by P1.*

Proof: In order to use approachability arguments let us introduce the following vector-valued reward function. Define the $1+SB$ dimensional reward vector $m=(m_r,m_y)\in \mathrm{I\!R}\times\Delta^{BS}$, where m_y is indexed by the state-action pairs (s,b). Given that the state is s, P2 played action b and the observed reward was r, the reward vector m is given by

$$m=(r,e_{sb}) \tag{3.9}$$

where e_{sb} is a unit vector with 1 at the entry that corresponds to (s,b). Thus, the average reward vector $\hat{m}_t=\frac{1}{t}\sum_{\tau=1}^{t}m_\tau$ can be written as $\hat{m}_t=(\hat{r}_t,\hat{y}_t)$, and consists of the average scalar reward and the vector of empirical state-action frequencies. Recall that for a pair of stationary strategies f of P1 and g of P2 $m(f,g)=\lim_{t\to\infty}E^0_{f,g}\hat{m}_t$ denotes the long-term average reward under f and g.

Let $C\triangleq\{(r,y):r\geq r^c(y)\}$. We shall show that C is approachable. The difference between attainability and approachability is that approachability requires that the Euclidean distance of $\hat{m}_t=(\hat{r}_t,y_t)$ to C converges to zero, while attainability requires the reward coordinate difference $\hat{r}_t-h(y_t)$ to become non-negative. By definition r^c is continuous on Δ^{SB}. It follows that if C is approachable then r^c is attainable.

Since C is convex it suffices by Theorem 2.1 to prove that for every $g\in G$, $M(F,g)\cap C\neq\emptyset$. Fix $g^0\in G$. Consider the state-action frequency vector y^0 that is the limit occupancy measure corresponding to the pair of strategies $f(g^0)$ and g^0. It suffices to show that the point $m^0\triangleq$

$(r(f(g^0),g^0),y^0) \in M(F,g^0)$ is in C. By definition of r^c, $r^c(y^0) \leq r^*(y^0)$. The reward $r^*(y^0)$ is finite, it thus satisfies

$$r^*(y^0) = \inf_{g:y_{f(g),g}=y} \sup_{f\in F} r(f(g),g) \leq \sup_{f\in F} r(f,g^0) = r(f(g^0),g^0) \tag{3.10}$$

so that $m^0 \in C$ and the result follows. ∎

Since the game is 0-reachable by both players, there exist ϵ-optimal stationary strategies. The definition of r^* implies that $r^*(y) \geq \mathrm{v}$, where v is the value of the zero-sum game. It immediately follows that $r^c(y) \geq \mathrm{v}$. This means that the *CBE* is *safe*, in terms of Fudenberg and Levine (1998). An obvious desirable property is to have a higher reward than v when P2's observed play (as reflected by the empirical distribution of P2's actions at the states visited) is suboptimal. We shall establish that property for the class of *irreducible* stochastic games. Recall that a stochastic game is irreducible if all states are recurrent under every pair of stationary strategies. Let $G^{opt} \subseteq G$ denote the set of stationary minimax-optimal strategies for P2. This set is not empty for irreducible games (Theorem 5.1.5 from Filar and Vrieze, 1996). Let Y_0 denote the set of feasible limit state-action frequencies. For an irreducible game this set satisfies that for some $\epsilon > 0$ every $y \in Y_0$ satisfies $y(s) > \epsilon$ for all s. Let

$$\tilde{v}(g) \triangleq \inf_{y:g\in G^+(y),y\in Y_0} r^{+c}(y)\,, \tag{3.11}$$

denote the guaranteed performance level when P2's strategy seems to coincides with g and P1 plays an r^c-attaining strategy. The proof below is given for *CBE* defined as the lower convex hull of r^+ as defined in (3.3), hence it naturally holds for the higher r^c from Definition 3.2.

Theorem 3.2 *Suppose the game is irreducible. Then:*

(i) $r^c(y) > \mathrm{v}$ *for any* $y \in Y_0$ *such that* $g^+(y) \notin G^{opt}$.

(ii) Furthermore, $\tilde{v}(g) > \mathrm{v}$ *for* $g \notin G^{opt}$.

The proof relies on the structure of the set G^{opt} of minimax optimal strategies for P2 under the irreducibility assumption which is provided in the following lemma.

Lemma 3.3 *For an irreducible game, the set of optimal stationary strategies* G^{opt} *is a Cartesian product of convex sets.*

Proof: From Proposition 5.1 in Patek (1997) we know that there is a unique (up to an additive constant) vector $w^* \in \mathbb{R}^S$ such that for every $s \in \mathcal{S}$

$$\mathrm{val}\left(r(s,a,b) + \sum_{s'} P(s'|s,a,b)w^*(s')\right) = w^*(s) + \mathrm{v}\,, \tag{3.12}$$

where the operator val is the minimax value defined for each matrix game (i.e., $\inf_{g(\cdot|s)}\sup_{f(\cdot|s)}$). We require Lemma 5.3.1 from Filar and Vrieze (1996) which we quote for completeness. Given a pair of stationary strategies f and g of P1 and P2, respectively, if for every $s \in \mathcal{S}$

$$\sum_{a,b} f(a|s)g(b|s)r(s,a,b) + \sum_{a,b,s'} P(s'|s,a,b)w^*(s') \geq w^*(s) + \mathrm{v}\,, \tag{3.13}$$

then $r(f,g) \geq \mathrm{v}$ (and the same holds if all inequalities are reversed). Furthermore, $r(f,g) = \mathrm{v}$ if and only if equality holds for all recurrent states. We claim that $g \in G^{opt}$ if and only if for every s the mixed action $g(\cdot|s)$ is in the set $G^{opt}(s)$ of optimal strategies in the matrix game defined in (3.12). Suppose $g(\cdot|s)$ is optimal in every such game. Then from the above quoted Lemma, for every $f \in F$ we have that $r(f,g) \leq \mathrm{v}$, so that g is optimal. Suppose that $g(\cdot|s)$ is not optimal for a matrix game that is induced by some state s'. There is a strategy for P1, f^* that guarantees a higher reward than the value of the game defined for state s' in (3.13) for P1. Using the quoted Lemma again, $r(f,g) = \mathrm{v}$ if and only if equality holds in (3.13). Since equality does not hold for all states $r(f^*,g) \neq \mathrm{v}$, applying the Lemma again with the reversed side of the inequality shows that $r(f^*,g) > \mathrm{v}$, so that g is not minimax optimal.

As a result the set G^{opt} is a Cartesian product of the optimal strategy sets at each state, namely $G^{opt} = \bigotimes_s G^{opt}(s)$. Furthermore, each set $G^{opt}(s)$ is convex, as follows from the well known fact that for a zero-sum matrix game the set of optimal strategies is convex. ∎

Proof of Theorem 3.2:

Suppose $g \notin G^{opt}$ and let $y \in Y_0$ be a point such that $G^+(y) = \{g\}$. Assume by contradiction that $r^{+c}(y) = \mathrm{v}$. Since $g \notin G^{opt}$, then $r^+(y) > \mathrm{v}$ by definition (3.3). Since r^{+c} is convex, then by Caratheodory's theorem there exist k ($k \leq SB + 1$) points $\{y_i\}_{i=1}^k$ such that for each point y_i there exists a strategy $g^i \in G^+(y_i) \cap G^{opt}$ and for some $0 < \alpha_i < 1, \sum_i \alpha_i = 1$, we have that $y = \sum_{i=1}^k \alpha_i y_i$. It follows that for every state s there exists $\{\beta_i\}_{i=1}^k$ such that $0 \leq \beta_i \leq 1$, $\sum_{i=1}^k \beta_i = 1$ and for every b:

$$g^+(y)(b|s) = \sum_{i=1}^k \beta_i g^i(b|s)\,,$$

specifically $\beta_i = \frac{\alpha_i y_i(s)}{\sum_{i=1}^k \alpha_i y_i(s)}$. Consequently, it holds that for every s, $g(y)(b|s) \in G^{opt}(s)$. Since G^{opt} is a Cartesian product of convex sets product per s, it follows that g belongs to G^{opt} contradicting the assumption, so (i) follows. Part (ii) follows since r^{+c} is continuous, and $\{y : g \in G(y)\}$ is closed (it is actually convex) so that the infimum of r^{+c} is achieved on the set. From (i) we know that $r^{+c}(y) > \mathrm{v}$ for every y in $\{y : g \in G(y)\}$ which implies $\inf_{\{y: g\in G(y)\}} r^{+c}(y) > \mathrm{v}$. ∎

The above proof fails when some states are transient for certain strategies, which is the case when the game is not irreducible. The following example satisfies that $\tilde{v}(g) = \mathrm{v}$ for every strategy g, even though the best response reward to some strategies is larger than v. We note however, that this example is not necessarily typical. For each not irreducible game the performance guarantees of the CBE need to be verified. See also Example 3 in this respect.

Example 2: Let the states space be defined by $\mathcal{S} = \{s^1, s^2, s^3\}$; the actions of P1 at state s^1 and s^3 are $\mathcal{A} = \{1, 2\}$ (P1 does not have actions in state s^2); and the actions of P2 in state s^3 are $\mathcal{B} = \{1, 2\}$ (P2 does not have actions in states s^1 and s^2). The state transition structure is defined so that $p(s^2|s^1, 1, \cdot) = 1$ and $p(s^3|s^1, 2, \cdot) = 1$. The transition probability from states s^2, s^3 satisfy $p(s^1|s^2) = 1$ and $p(s^1|s^3) = 1$. The reward for states s^1 and s^2 is constant, that is: $r(s^1, \cdot, \cdot) = r(s^2, \cdot, \cdot) = 3/4$. The actions of P1 and P2 in state s^3 form a "matching penny" game. So that $r(s^3, 1, 1) = r(s^3, 2, 2) = 1$ and $r(s^3, 2, 1) = r(s^3, 1, 2) = 0$. The state-action frequency are: $(y(s^1), y(s^2), y(s^3, 1), y(s^3, 2))$. A straightforward calculation shows that the point $y_0 = (1/2, 1/2, 0, 0)$ satisfies that $r^*(y_0) = 3/4 = \mathrm{v}$. Since $G^+(y_0) = G$ it follows that the above example $\tilde{v}(g) = \mathrm{v}$ for all g. Note, however, that for some $y \in \Delta^{SB}$ a higher reward is guaranteed. ∎

Remark 3: The EBE concept may be extended to games with *partial* observation of P2's actions. In these games, P1 observes a signal which is a (stochastic) function of both players' actual actions and the current state. We still assume that the state is fully observed. For these games, one can formulate the envelope in the state-signal space in a very similar manner to the development above. If we define $m(f, g)$ as the expectation of the state-signal frequency starting from state 0, then both a state-based EBE and a dynamics-based EBE can be defined and their lower convex hull can be attained.

Remark 4: The EBE does not take into account the empirical frequencies of P1's own actions. It is possible to include P1's actions in the state-action frequency vector, making it a vector in Δ^{SBA}; this may in fact be embedded in the state-signal framework of the previous remark, where the "signal" is now the action pair (a, b) rather than b alone. This would result in an envelope which is at least as high as EBE for any given history (since P1's actions themselves are taken into consideration). An appropriately defined CBE can be proved to be attainable. It can be shown that the projection of the EBE which includes the actions of both players equals EBE which is defined by only P2's action presented before.

Remark 5: A major issue regarding CBE is its maximality. By this we mean that there might be functions which are higher than r^c and are still attainable. When CBE it is not maximal one may consider "optimal" enlargements of CBE. Relevant notions of optimality are the Pareto efficiency of the guaranteed reward – either in the frequency space, compared with $r^c(y)$ for every y, or in the strategy space, compared with $\tilde{v}(g)$. This may possibly be done with the assistance of the necessary and sufficient conditions introduced in Spinat (1999). This topic is under current investigation.

Remark 6: The performance guarantees of the dynamics-based CBE may depend on the choice of $\{f(g)\}$. It is an interesting question how to choose $f(g)$, when the choice exists, so that r^c would be maximal. The following example shows that the right choice of $\{f(g)\}$ may be critical.

Example 3: The example is a modification of Example 2. Replace the reward in states s^1 and s^2 with 1/2. Recall that g in this game is determined by $g(b|s^3)$. Consider the following two optimal strategies against $g_0 = g(1|s^3) = \frac{1}{2}$: $f_1(1|s^1) = 1$ (s^3 is not reached in this case) and $f_2(2|s^1) = 1, f_2(1|s^3) = \frac{1}{2}$. If f_1 is chosen then the point $y_0 = (1/2, 1/2, 0, 0)$ satisfies $r^c(y) = \frac{1}{2}$. Since $G(y_0) = G$, it follows that $\tilde{v}(g) = \mathrm{v} = 1/2$ for every $g \in G$. If f_2 is chosen, then it is easy to show that r^* is finite only on $\{(1/2, 0, \frac{\epsilon}{2}, \frac{1-\epsilon}{2}),\ 0 \leq \epsilon \leq 1\}$. The attainable reward for every vector in this set is the Bayes reward since the stochastic game reduces to a matrix game and the CBE is simply given by the envelope function of the matrix game

$$r^c(1/2, 0, \frac{\epsilon}{2}, \frac{1-\epsilon}{2}) = 1/4 + 1/2 \max\{\epsilon, 1-\epsilon\}\,.$$

It can be further shown that the corresponding $\tilde{v}(g)$ in (3.11) satisfies:

$$\tilde{v}(g) \overset{\triangle}{=} \inf_{y: g \in G(y)} r^c(y) = \max_{f \in F} r(f, g) = r^{BR}(g)\,.$$

■

4 Single Controller Games

In this section we consider the special case in which P1 does not affect the state transitions, that is $P(s'|s,a,b) = P(s'|s,b)$. The resulting model can be viewed as a sequence of matrix games where the next game to be played is determined only by P2's action. These models are often called single controller games (e.g., Filar and Vrieze, 1996) as only one of the players controls the state transitions. As it turns out, from the regret minimization perspective P1 need not concern himself here with the state transitions, and the stochastic game can be partitioned into a sequence of interleaved repeated matrix games. Thus, performing optimally in each matrix game suffices for performing optimally in the overall stochastic game. We first relate this case to our general framework, by showing in Proposition 4.1 that EBE is convex and therefore attainable. In Subsection 4.1 we provide a general result which applies without any structural assumptions on the stochastic game and provides a better convergence rate. In Subsection 4.2 we outline an example for an application of the single controller game - the k-th order Bayes envelope for repeated matrix games.

We begin with a characterization of EBE for games in which P2 controls the state transitions. Note that in this case every feasible point y defines a set of strategies g which agree with y, so that $r^+ = r^*$ for feasible points.

Proposition 4.1 *Suppose that P1 does not affect the state transitions, i.e.,* $P(s'|s,a,b) = P(s'|s,b)$*. Then* EBE *is convex.*

Proof: Under the assumption that $P(s'|s,a,b) = P(s'|s,b)$, the stochastic game dynamics effectively reduces to that of a Markov Decision Process with a single decision maker. Let $Y_0 \subseteq \Delta^{SB}$ be the set of feasible state-action frequencies for P2. It is well known (e.g., Theorem 8.9.4 in Puterman, 1994) that Y_0 is a convex set. Since only P2 affects the transitions, then for a given y and for every $g \in G(y), f \in F$, the state-action frequencies $y_{f,g}$ depends only on g and equals y_g. Thus, the reward $r(f,g)$ can be written as a function of y_g (and not g):

$$r(f,g) = \sum_{s,a,b} y_g(s,b) f(a|s) r(s,a,b)\,, \tag{4.1}$$

It follows that $r^*(y)$ in (3.8) reduces to:

$$\begin{aligned} r^*(y) &= \inf_{g:y_g=y} \sup_{f \in F} r(f,g) \\ &= \sup_{f \in F} \sum_{s,a,b} y(s,b) f(a|s) r(s,a,b)\,. \end{aligned}$$

This implies that EBE is in fact a convex set since r^* is the maximum of linear functions. ∎

Since EBE, r^*, is convex we can apply Theorem 3.1 to deduce the attainability of r^*. Note that as P1 does not affect the state transition, the 0-reachability assumption (for P1) reduces to the requirement that the state 0 is recurrent under all strategies of P2. We therefore have the following corollary:

Corollary 4.2 *Suppose that 0 is recurrent in the single controller game controlled by P2, then the* EBE *is attainable.*

The above result was obtained using a general framework. By using specific analysis, we shall be able to dispense of any recurrence assumptions.

4.1 Generalized Results With Improved Convergence Rate

In this sub-section we show the empirical Bayes envelope for every single controller stochastic game is attainable without resorting to any assumptions on the game dynamics. We offer a simple algorithm for attaining the EBE. The algorithm takes advantage of the fact that the single controller game can be effectively partitioned into a set of repeated matrix games, each corresponding to a single state. Let us re-define $r^*(y)$ so that it will be well defined without any assumptions on the game dynamics and for every $y \in \Delta^{SB}$:

$$r^*(y) \stackrel{\triangle}{=} \max_{f \in F} \sum_{s,a,b} y(s,b) f(a|s) r(s,a,b)\,. \tag{4.2}$$

As shown in the proof of Proposition 4.1, if the game has a recurrent state then the original definition of r^* in equation (3.8) is the same as (4.2) for every feasible state-action frequency vector y.

Theorem 4.3 *Assume that P1 does not affect the state transitions, i.e., $P(s'|s,a,b) = P(s'|s,b)$. Then:*

i. The EBE, r^*, is attainable.

ii. The convergence rate satisfies $\hat{r}_t - r^*(y_t) \geq -Ct^{-1/3}$ almost surely, where C is a constant that depends on the parameters of the game.

Proof: *EBE* is attained if the average reward $\hat{r}_t$ and the state-action frequencies y_t satisfy

$$\liminf_{n\to\infty}(\hat{r}_t - r^*(y_t)) \geq 0\,. \tag{4.3}$$

Starting from equation (4.2), we have that:

$$\begin{aligned} \hat{r}_t - r^*(y_t) &= \hat{r}_t - \sum_s \max_{f(\cdot|s)} \sum_b \sum_a y_t(s,b) f(a|s) r(s,a,b) \\ &= \hat{r}_t - \sum_s \max_a \sum_b y_t(s,b) r(s,a,b)\,, \end{aligned} \tag{4.4}$$

where $y_t(s,b)$ is the relative frequency of state s and action b measured up to time t (see (3.1)). The last equality is justified since f affects the inner sum only through the actions at state s. But $\hat{r}_t = \sum_{s,a,b} y_t(s,b) f_t(a|s) r(s,a,b)$, where $f_t(a|s)$ is the relative frequency of choosing a at state s by time t so

$$\hat{r}_t - r^*(y_t) = \sum_s y_t(s) \left(\sum_b \sum_a g_t(b|s) f_t(a|s) r(s,a,b) - \max_a \sum_b g_t(b|s) r(s,a,b) \right)\,, \tag{4.5}$$

where $g_t(b|s) \triangleq \frac{y_t(s,b)}{y_t(s)}$ and $y_t(s) = \sum_b y_t(s,b)$ (if $y_t(s) = 0$ then $f_t(\cdot|s)$ and $g_t(\cdot|s)$ are arbitrary). Suppose P1's strategy is to play a regret minimizing strategy for every state s separately for the game in which P1's actions are $a \in \mathcal{A}$, P2's actions are $b \in \mathcal{B}$, and the expected reward is $r(s,a,b)$. Regret minimizing strategies for repeated matrix games exist - for example, Hannan (1957), Freund and Schapire (1999), Hart and Mas-Colell (2001). In that case, (4.5) becomes the sum of elements that tend to zero as $t \to \infty$ (if the state is visited often enough). We have that

$$\hat{r}_t - r^*(y_t) = \sum_s y_t(s) R_t(s)\,, \tag{4.6}$$

where $R_t(s)$ is the regret of the repeated game which is played in state s game by time t:

$$R_t(s) \triangleq \sum_b \sum_a g_t(b|s) f_t(a|s) r(s,a,b) - \max_a \sum_b g_t(b|s) r(s,a,b)\,.$$

Now, P1's strategy ensures that $\{R_t(s)\}^- \to 0$ almost surely ($\{x\}^- \stackrel{\triangle}{=} \min\{0,x\}$) for every state s that is visited infinitely often, while $y_t(s) \to 0$ for the other states. It then follows from (4.6) that (4.3) holds.

The second part follows by using a regret minimizing algorithm that satisfies a regret bound of $Ct^{-1/3}$ for each matrix game (t is the number of time the game is played). There are several algorithms with this regret rate, we refer the reader to Hannan (1957), Freund and Schapire (1999). If such an algorithm is used then every element in the sum (4.6) can be bounded by $y_t(s)R_t(s) = \frac{n_s(t)}{t}C_s(n_s(t))^{-1/3}$, where $n_s(t)$ is the number of epochs spent in state s until time t and C_s is a constant which depends only on $r(s,\cdot,\cdot)$. It follows that (4.6) can be bounded by:

$$\begin{aligned} \hat{r}_t - r^*(y_t) &\geq -\sum_s \frac{n_s(t)}{t} C_s n_s(t)^{-1/3} \\ &= -\frac{1}{t}\sum_s C_s n_s(t)^{2/3} \\ &\geq -\frac{1}{t}\sqrt{\sum_s C_s^2}\sqrt{\sum_s n_s(t)^{4/3}} \\ &\geq -\frac{1}{t}\sqrt{\sum_s C_s^2}\sqrt{t^{4/3}} \\ &\geq -\frac{1}{t^{1/3}}\sqrt{\sum_s C_s^2}\,, \end{aligned} \tag{4.7}$$

where the third inequality follows from Cauchy-Schwartz inequality, and the fourth from the convexity of $x^{4/3}$ and the definition of $n_s(t)$. ∎

It is worth noting that a similar scheme will work even for an unknown game. In Freund and Schapire (1999), an algorithm with regret of $Ct^{-1/3}$ for an unknown repeated matrix game was introduced. By using this algorithm as the basic algorithm for each matrix game we get the same results even if the game is initially unknown.

4.2 The k-th order Bayes envelope

As an immediate application of the single controller stochastic game we briefly consider the k-th order extension of the Bayes envelope for repeated matrix games. This has some interesting applications to the classical problem of prediction of Markov sequences.

Recall that a matrix game is defined by the 3-tuple $(\mathcal{A},\mathcal{B},r)$ where $\mathcal{A}$ is P1's action set, $\mathcal{B}$ is P2's action set and $r : \mathcal{A}\times\mathcal{B} \to \mathbb{R}$ is P1's reward. The Bayes envelope of a matrix game is defined as:

$$r^*(y_t) \stackrel{\triangle}{=} \max_a \sum_b y_t(b) r(a,b)\,, \tag{4.8}$$

where $y_t \in \Delta(\mathcal{B})$ is the empirical measure of P2's actions by time t, and as usual is attained when $\hat{r}_t \geq r^*(y_t) + o(1)$. The k-th order Bayes envelope accounts for possible the correlations (up to order k) between P2's actions. To motivate the definition below, assume that P2 is stationary and that the conditional probabilities $p^k(b_t|b_{t-1}, \dots b_{t-k})$ are known by P1 for any k-sequence $c_t^k = (b_{t-1}, \dots b_{t-k})$ of P2's actions. Then the average reward that can be secured by P1 is

$$r^{*k} \overset{\triangle}{=} \sum_{c \in \mathcal{B}^k} y^k(c^k) \max_a \sum_{b \in \mathcal{B}} p^k(b|c^k) r(a,b) = \sum_{c \in \mathcal{B}^k} \max_a \sum_{b \in \mathcal{B}} y^{k+1}(c^k b) r(a,b) \,,$$

where $y^k(c^k)$ is the probability of the k-length sequence c^k of P2's actions, and cb denotes the concatenation of c and b. Accordingly, the k-order Bayes Envelope is defined as

$$\hat{r}_t \geq r^{*k}(y_t) \overset{\triangle}{=} \sum_{c \in \mathcal{B}^k} \max_a \sum_{b \in \mathcal{B}} y_t^{k+1}(cb) r(a,b) \,, \tag{4.9}$$

where y_t^{k+1} is the empirical frequency of $k+1$ length sequences of P2's actions. Obviously, the standard Bayes envelope (4.8) is the 0-order Bayes envelope according to this definition. Theorem 4.3 implies that EBE as defined in (4.2) is attainable. A straightforward interpretation of that envelope in terms of the original matrix game now asserts that the k-order Bayes envelope is attainable.

Another interesting application that fits into the single controller framework is the problem of prediction with expert advice (see Vovk, 1998). The embedding of this problem in single controller games is discussed in Mannor and Shimkin (2000a).

5 *EBE* Is Unattainable

In this section we analyze a specific example of a stochastic game in which the empirical Bayes envelope is not convex and not attainable. This game is a single controller game in which P1 controls the state transitions. The game is also irreducible. The example also sheds light on the structure and character of the *EBE*. In this section we will concentrate on r^+, since it is lower than r^*.

The game is defined by $\mathcal{S} = \{s^1, s^2\}$; $\mathcal{A} = \{0,1\}$ in both states; and $\mathcal{B} = \{0,1\}$ in both states. The transition probabilities are the same in both states and are determined by P1's action: $P(s^1|\cdot, 0, \cdot) = 0.99$, $P(s^2|\cdot, 1, \cdot) = 0.99$. Thus, P1 determines the next state by declaring it (up to a probability of 0.01). The reward of P1 is determined by P2's action only : $r(\cdot,\cdot,0) = 1$, $r(\cdot,\cdot,1) = -1$. The value of the zero-sum game with the same reward is -1, as the minimax strategy of P2 is to always play action 1.

Let $y = (y(s^1,0), y(s^1,1), y(s^2,0), y(s^2,1))$ denote the state-action frequency vector. The game is defined so that P1 determines the state frequencies and P2 determines the reward at each state.

We now calculate the empirical Bayes envelope explicitly. Assume P1 chooses action 0 at state s^1 with probability $g(0|s^1)$ and chooses action 0 at state s^2 with probability $g(0|s^2)$. Let $y(s^1)$ and $y(s^2)$ be the steady state probabilities of state s^1 and s^2. It can be shown that:

$$y(s^1) = \frac{0.98g(0|s^2) + 0.01}{1 + 0.98g(0|s^2) - 0.98g(0|s^1)}$$

Since $(g(0|s^1), g(0|s^2)) \in [0,1] \times [0,1]$ are arbitrary it follows that P1 can set $y(s^1) \in [0.01, 0.99]$ at will. The observed strategy $g = g^+(y)$ is defined by $g^+(b|s) = \frac{y(s,b)}{y(s)}$. Given g, if $g(0|s^1) - g(1|s^1) > g(0|s^2) - g(1|s^2)$ then P1 finds it more rewarding to be in state s^1, while if the inequality is reversed then state s^2 is preferred. As a result the best strategy of P1 is to have state s^1 visited with frequency 0.99 if the above inequality holds, or have state s^2 visited with frequency 0.99 if this inequality is reversed. If $g(0|s^1) - g(1|s^1) = g(0|s^2) - g(1|s^2)$ then any strategy of P1 will produce identical reward. As a result the Bayes envelope function is given by $r^+(y) = r^{BR}(g^+(y))$, where

$$r^{BR}(g) = \begin{cases} 0.99(g(0|s^1) - g(1|s^1)) + 0.01(g(0|s^2) - g(1|s^2)) & \text{if} \quad g(0|s^1) > g(0|s^2) \\ 0.01(g(0|s^1) - g(1|s^1)) + 0.99(g(0|s^2) - g(1|s^2)) & \text{if} \quad g(0|s^1) \leq g(0|s^2) \end{cases}$$

is the best-response reward of P1 against g.

To establish explicitly that EBE is not attainable, it suffices to show that P2 can keep the average reward away from $r^+(y_t)$ at arbitrarily large times; more precisely, there exists $\epsilon > 0$ so that for any for any time T, there exists a strategy of P2 so that (with large probability) the average reward $\hat{r}_t$ for some $t > T$ will be ϵ-lower than $r^+(y_t)$. Given T, define the following strategy for P2:

$$g_t = \begin{cases} (1, 0, \frac{1}{2}, \frac{1}{2}) & 0 < t < T \\ (0, 1, \frac{1}{2}, \frac{1}{2}) & T \leq t < 2T \\ (0, 1, 1, 0) & 2T \leq t < 3T \end{cases}$$

Here the entries of g_t correspond to $(g(0|s^1), g(1|s^1), g(0|s^2), g(1|s^2))$. We shall give here a qualitative argument, complete details can be found in Mannor and Shimkin (2000a).

Assume that T is sufficiently large so that the empirical frequencies are close enough to their expected value with probability close to 1. Note that for such T each state s^1 or s^2 is visited with frequency 0.01 at least. After the first T epochs, the observed strategy $g(y_T)$ equals $(1, 0, \frac{1}{2}, \frac{1}{2})$, and the best-response reward of P1 is 0.99. P1 is obliged to have the empirical frequency of state s^1 near the maximum, or else the average reward will be strictly less than $r^+(y)$. In the next T time epochs P1 should have the occupation frequency of state s^1 near the maximum too. This is the case since otherwise the average reward at time $2T$ will be away from the EBE. Indeed, assuming for simplicity that the empirical frequency of state s^1 by time T was exactly 0.99, an occupation measure of $0.99(1-\alpha)$ over the second T-period can be seen to yield an average reward (by $2T$) of $\hat{r}_{2T} = 0.99\alpha$, while the observed strategy is $g = (\frac{1}{2-\alpha}, \frac{1-\alpha}{2-\alpha}, \frac{1}{2}, \frac{1}{2})$ which gives a best-response

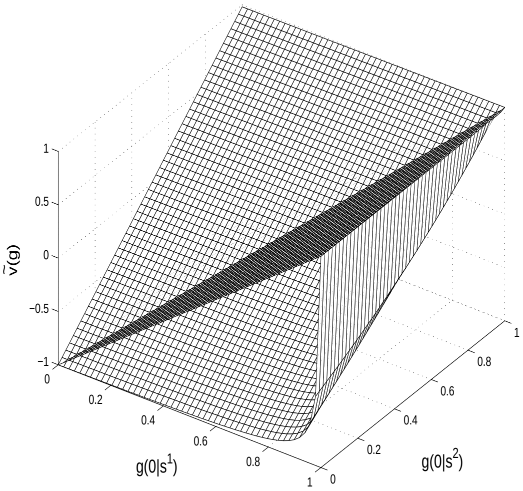

Figure 1: The upper envelope is $r^{BR}(g)$, the lower envelope is $\tilde{v}(g)$.

reward of $r^+(y_{2T}) = 0.99\frac{\alpha}{2-\alpha}$. The two rewards coincide only for $\alpha = 0$. Thus, P1 reaches time $2T$ with average reward (arbitrarily close to) 0, $y(s^1) = 0.99$, and $g(y) = (\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2})$. Finally, at the last T-period, the Bayes reward will surely be denied, since the best-response reward will be determined by the observed strategy at state s^2, which (due to the relatively low occupancy of that state by $2T$) increases more quickly than possible for the average reward. Indeed, if P1 chooses $y(s^1) = \beta$ over that period ($0.01 \leq \beta \leq 0.99$), then we obtain an average reward of $\hat{r}_{3T} = \frac{1-2\beta}{3}$, while $g(y) = (\frac{0.33}{0.66+\beta}, \frac{0.33+\beta}{0.66+\beta}, \frac{1.01-\beta}{1.02-\beta}, \frac{0.01}{1.02-\beta})$ and consequently $r^+(y_{3T}) = 0.99\frac{1-\beta}{1.02-\beta} + 0.01\frac{-\beta}{0.66+\beta}$. A straightforward calculation shows that the difference between the average reward and $r^+(y)$ remains bounded away from 0 (it is actually more than 0.33). Now the whole process can be repeated beyond $3T$ by taking a large enough interval T'. The same analysis applies, as the first $3T$ epochs are negligible for T' large enough. By repeating this construction EBE is denied infinitely often.

In Figure 1, we show the performance envelopes for the game which is described above. The planar coordinates are $g(0|s^1)$ and $g(0|s^2)$, the third dimension is the guaranteed reward. In the figure, there are two surfaces. The upper surface is the best response reward and the lower surface is the guaranteed reward $\tilde{v}(g)$ when attaining CBE (defined in (3.11)). This value was calculated using linear programming for a grid of points in the $g(0|s^1) \times g(0|s^2)$ plane. One can notice that when either $g(0|s^1)$ or $g(0|s^2)$ are extreme, the guaranteed reward when attaining CBE is close to the best response reward, however as they become central, $\tilde{v}(g)$ becomes less tight, but still strictly above the value of the game which is -1.

It is interesting to note that if the state transitions become deterministic (i.e., 0 instead of 0.01 and 1 instead of 0.99) then EBE become attainable. Indeed, it may be seen that

$r^+(y(s^1,0), y(s^1,1), 0, 0) = y(s^1,0) - y(s^1,1)$ since the worst strategy of P2 which is consistent with such state-action frequency vector yields a reward of -1 in s^2. A strategy that attains EBE is therefore: always remain in state s^1. However, if P1's strategies are restricted to strategies with some exploration (i.e., all states must be visited a small fraction of the time) then we are left with the same problem as before and EBE is not attainable.

6 Conclusion and Open Questions

We have proposed in this paper a notion of regret in stochastic games which relies on state-action frequencies. The empirical Bayes envelope was defined as the best-response reward that can be secured against the set of stationary strategies of the opponent which are consistent with the observed state-action frequencies. Two options for this set of the opponent's strategies were proposed and led to distinct definitions of the EBE. The first is the state-based EBE that regarded only the relative frequencies of the opponent's actions at each state. This was refined by the dynamics-based EBE, which accounts also for the feasibility of the observed state frequencies. As the EBE need not be attainable in general, we proposed its lower convex hull, the CBE, as our general solution concept, and established the existence of no-regret strategies with respect to it.

The (unattainable) EBE guarantees, by its very definition, the best response payoff when opponent uses a stationary strategy. The lower convex hull operation inherent in the CBE obviously degrades the performance guarantee. This gap seems to be inherent in the formulation of attainable performance envelope (against general strategies) which relies on the state-action frequencies alone. An exception is the case of single-controller games, where the EBE itself is convex, hence attainable. It was shown that for irreducible games the CBE is higher than the value of the game *whenever* the opponent's strategy is seen to deviate from an optimal one.

It should be emphasized that the approach of this paper is viable only under recurrence assumptions of the type that were imposed here (Assumption 1). In general, optimal maximin strategies may be inherently non-stationary, and a meaningful regret minimizing scheme is not evident.

Some questions and possible extensions remain concerning the proposed framework. The strategy we used for attaining the CBE is based on a geometric viewpoint within the approachability framework; it would be interesting to have such strategies formulated in a more direct way, perhaps in the style of Freund and Schapire (1999). Our definition of the dynamics-based CBE is non-unique, and essentially depends on the particular choice of best-response strategies for P1. It would be interesting to formulate an appropriate notion of minimality (or undominance), and characterize minimal envelopes. It would also be of interest to apply the concepts of regret minimization to a learning situation, where the system parameters are not known in advance, or the problem is too complex for explicit solution. A Reinforcement Learning approach (e.g., Bertsekas and Tsitsiklis, 1995; Kaelbling et al., 1996) may be especially relevant. Further applications of the theory should be found in areas that include communications over arbitrary channels (Lapidoth and Narayan, 1998), universal prediction (Feder and Merhav, 1998), utilization of expert advice (Vovk, 1998), competitive queuing problems (Shimkin and Shwartz, 1993), and others.

Finally, one may consider other concepts of a Bayes envelope which depart from the state-action frequency viewpoint. One possibility would be to work in the space of (empirical frequencies) of the opponent's strategies over large intervals, viewing them as actions in an appropriately defined repeated super-game. This idea is further discussed in Mannor and Shimkin (2000a). The drawback of this approach is the huge space on which the Bayes envelope is defined, and the apparent requirement for an explicit experimentation of every (deterministic) stationary strategy of P1 which renders this approach impractical for none but the smallest models.

Acknowledgement: This research was supported by the Fund for Promotion of Research at the Technion. We are grateful to an anonymous referee for pointing out several significant points and assisting us in making this paper more intelligible. A specific observation by another referee led to the introduction of the dynamics-based *EBE*.

References

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. (1995). Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proc. 36th Annual Symposium on Foundations of Computer Science* (pp. 322–331). IEEE Computer Society Press.

Bertsekas, D., and Tsitsiklis, J. (1995). *Neuro-dynamic programming.* Athena Scientific.

Blackwell, D. (1956a). An analog of the minimax theorem for vector payoffs. *Pacific J. Math.*, *6*(1), 1–8.

Blackwell, D. (1956b). Controlled random walks. In *Proc. International Congress of Mathematicians, 1954* (Vol. III, pp. 336–338). North-Holland.

Feder, M., and Merhav, N. (1998). Universal prediction. *IEEE Transaction on Information Theory*, *44*(6), 2124–2147.

Filar, J., and Vrieze, K. (1996). *Competitive Markov decision processes.* Springer Verlag.

Flesch, J., Thuijsman, F., and Vrieze, O. J. (1998). Simplifying optimal strategies in stochastic games. *SIAM J. Control Optim.*, *36*(4), 1331–1347.

Freund, Y., and Schapire, R. (1999). Adaptive game playing using multiplicative weights. *Games and Economic Behavior*, *29*, 79–103.

Fudenberg, D., and Levine, D. (1995). Universal consistency and cautious fictitious play. *Journal of Economic Dynamic and Control*, *19*, 1065–1990.

Fudenberg, D., and Levine, D. (1998). *The theory of learning in games.* MIT Press.

Hannan, J. (1957). Approximation to Bayes risk in repeated play. In M. Dresher, A. W. Tucker, and P. Wolde (Eds.), *Contribution to the theory of games, III* (pp. 97–139). Princeton University Press.

Hart, S., and Mas-Colell, A. (2000). A simple adaptive procedure leading to correlated equilibrium. *Econometrica*, *68*, 1127-1150.

Hart, S., and Mas-Colell, A. (2001). A general class of adaptive strategies. *Journal of Economic Theory*, *98*, 26-54.

Kaelbling, L., Littman, M., and Moore., A. (1996). Reinforcement learning - a survey. *Journal of Artificial Intelligence Research*, *4*, 237-285.

Kumar, P. R., and Varaiya, P. (1986). *Stochastic systems: Estimation, identification and adaptive control.* Englewood Cliffs, N. J.: Prentice Hall.

Lapidoth, A., and Narayan, P. (1998). Reliable communication under channel uncertainty. *IEEE Transactions on Information Theory*, *44*, 2148-2177.

Lehrer, E. (1998, May). *Approachability in infinite dimensional spaces and an application: A universal algorithm for generating extended normal numbers.* (Preprint)

Mannor, S., and Shimkin, N. (2000a). *The empirical Bayes envelope approach to regret minimization in stochastic games* (Technical report EE- No. 1262). Faculty of Electrical Engineering, Technion, Israel.

Mannor, S., and Shimkin, N. (2000b). *Generalized approachability results for stochastic games with a single communicating state* (Technical report EE- No. 1263). Faculty of Electrical Engineering, Technion, Israel. (Appeared in ORP3, 2001; submitted to EJOR)

Mertens, J., and Neyman, A. (1981). Stochastic games. *International Journal of Game Theory*, *10*(2), 53–66.

Mertens, J., Sorin, S., and Zamir, S. (1994). *Repeated games* (CORE Reprint Nos. Dps 9420, 9421 and 9422). Center for Operation Research and Economics, Universite Catholique De Louvain, Belgium.

Milman, E. (2000). *Uniform properties of stochastic games and approachability.* Unpublished master's thesis, Tel Aviv University.

Patek, S. (1997). *Stochastic shortest path games.* Unpublished doctoral dissertation, LIDS MIT.

Puterman, M. (1994). *Markov decision processes.* Wiley-Interscience.

Rockafellar, R. (1970). *Convex analysis.* Princeton University Press.

Rustichini, A. (1999). Minimizing regret: the general case. *Games and Economic Behavior*, *29*, 224–243.

Shimkin, N., and Shwartz, A. (1993). Guaranteed performance regions in Markovian systems with competing decision makers. *IEEE Trans. on Automatic Control*, *38*(1), 84–95.

Spinat, X. (1999). *An approachability condition for general sets* (Technical Report No. 496). Ecole Polytechnique, Paris.

Stoer, J., and Witzgall, C. (1970). *Convexity and optimization in finite dimensions* (Vol. I). Springer-Verlag.

Vohra, R., Levine, D. K., and Foster, D. (Eds.). (1999). Special issue on learning in games. *Games and Economic Behavior*, *29*(1).

Vovk, V. (1998). A game of prediction with experts advice. *Journal of Computer and Systems Sciences*, *56*(2), 153–173.