On-line Learning with Imperfect Monitoring

Shie Mannor[1] and Nahum Shimkin[2]

[1] Laboratory for Information and Decision Systems
Massachusetts Institute of Technology, Cambridge, MA 02139
shie@mit.edu
[2] Department of Electrical Engineering
Technion, Haifa 32000, Israel
shimkin@ee.technion.ac.il

Abstract. We study on-line play of repeated matrix games in which the observations of past actions of the other player and the obtained reward are partial and stochastic. We define the Partial Observation Bayes Envelope (POBE) as the best reward against the worst-case stationary strategy of the opponent that agrees with past observations. Our goal is to have the (unobserved) average reward above the POBE. For the case where the observations (but not necessarily the rewards) depend on the opponent play alone, an algorithm for attaining the POBE is derived. This algorithm is based on an application of approachability theory combined with a worst-case view over the unobserved rewards. We also suggest a simplified solution concept for general signaling structure. This concept may fall short of the POBE.

1 Introduction

Repeated games provide the opportunity for each player to adjust her play according to the observed past, in particular the observed actions of the other players, or the rewards obtained for different choices of actions. The regret minimization framework allows to exploit this idea in a non-strategic framework, without imposing any restrictions or rationality assumptions on the strategies employed by the other players. The idea in regret minimization is to set a desired goal for the average payoff, depending on the observed moves in the game, and show that this goal may be obtained asymptotically. The most common goal is the Bayes envelope – the maximal average payoff that a player could secure for herself had she known in advance the relative empirical frequencies of the other players' actions. Obviously, with such prior knowledge, this payoff could be secured simply by playing the stationary strategy which repeats at every stage the best response (in the single-shot game) to the given relative empirical frequencies. The difference between the actual average payoff and the current Bayes envelope is termed the average regret. A strategy which asymptotically secures non-positive average regret for all possible strategies of the other player has been termed *regret minimizing*, and more recently *universally consistent* ([1]).

To motivate the following discussion let us consider a doctor that attends many patients. Suppose that each patient may either have disease A or disease

B. The doctor may treat each patient using one of two treatments - 1 or 2. Treatment 1 is effective only against disease A while treatment 2 is effective only against disease B. Suppose that the doctor does not know if treatment 1 is successful or not, however she does know if treatment 2 is successful or not. As many patients arrive, the doctor's overall goal is to have the best success rate as if she knew in advance the patients' disease distribution. This situation can be captured in the following table. Each entry (r, s) in the table represents the doctor's reward, r, and by the observation, s, she receives.

	Disease A	Disease B
Treatment 1	$(1, a)$	$(0, a)$
Treatment 2	$(0, b)$	$(1, c)$

If the doctor was informed the results of each treatment the game reduces to a matching pennies game, and the doctor can obtain the best rate possible. Since this is not the case, a refined machinery is needed. Suppose that by time t the doctor observed signals a, b, and c for π_a, π_b, and π_c fraction of the time (respectively). If it was known in advance that the patients' disease is a stationary process, then (assuming $\pi_b + \pi_c \neq 0$) the best response is $r^*(\pi_a, \pi_b, \pi_c) = \frac{\max(\pi_b, \pi_c)}{\pi_b + \pi_c}$. If $\pi_a = 1$ then no information is gathered and the worst case disease distribution is $1/2 - 1/2$, in which case the doctor cannot hope to gain more than $1/2$. The function r^* is shown in Figure 1(a). We will show that while r^* may not be attainable in general its lower convex hull, r^c, is attainable. r^c is presented in Figure 1(b). The difference between r^* and r^c is plotted in Figure 1(c). We note that by attaining r^c a higher reward than the pessimistic $\frac{1}{2}$ is obtained for most observation frequency vectors.

The case of perfect monitoring is well studied. Regret minimizing strategies were originally provided in [2] and later in [3] (see also [1] for a more modern approach). Recently, the feasibility of regret minimization has been established even when a perfect monitoring of the opponents actions is not available. The adversarial bandit formulation of [4] considers the case where only the reward at each stage is observed (in addition to the player's own action). It was shown

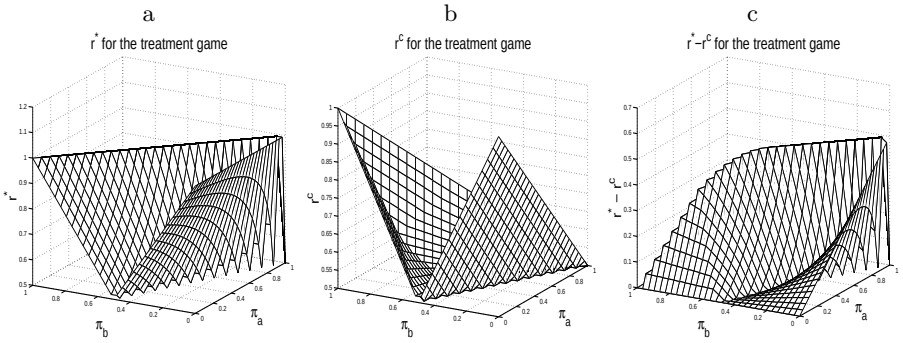

Fig. 1. r^* (a), r^c (b), and the difference $r^* - r^c$ (c) for the treatment game.

that regret minimizing strategies exist in this case as well. In [5] the regret minimization framework was extended to a general signalling structure, where the observed signals are random functions of the two players' actions at each stage. The Bayes envelope in this case must be weakened to take into account the imperfect monitoring. This modified Bayes envelope is still defined in the space of the (now unobserved) empirical actions frequency of the other player, but now the worst-case payoff over all stationary strategies of the other player, which induce the same signal frequencies as the empirically observed signal frequencies, is considered. This envelope is termed Partial Observation Bayes Envelope (POBE). The existence of regret minimizing strategies with respect to the POBE was established in [5]. The proofs there rely on the approachability theory, lifted to the space of measures over mixed actions. Concrete algorithms that attain the POBE were not supplied in [5]. A simplified model was suggested in [6]. In the model analyzed there the average reward can be consistently estimated from the available signal data. Since the model of [6] essentially allows observing the reward, the POBE of that model coincides with the perfect monitoring Bayes envelope and efficient algorithms which are based on multiplicative weights were derived.

In this paper we consider the general signalling model where the reward is not assumed observed. We provide an explicit algorithm for attaining the POBE when P2 alone affects the observation. We also suggest a simplified concept for arbitrary signaling structure. This concept can be easily attained, possess some non-trivial performance guarantees, but may fall short of the POBE. We note in passing that an essential part of our imperfect monitoring model is that the game and signalling are *known* to the player. This is required in order to allow meaningful inference from the observed signals. A note about the terminology is due. We use the words "game" and "opponent" in this paper, however the model does *not* assume that the other participating agents are rational. Consequently, the model may be considered as a game against Nature (as in the treatment game above), where Nature's play is not assumed stationary nor adversarial.

2 The Model

In order to model imperfect monitoring we consider a finite action two-person game that is repeated in time. We refer to the players as P1 (the regret-minimizing player) and P2. The stage game is assumed here to be a finite game in strategic form, namely a matrix game. Let $a \in \mathcal{A}$ and $b \in \mathcal{B}$ denote the (finite) set of actions of P1 and P2 in this game, respectively. The strategies of P1 and P2 in the repeated game will be denoted by σ and ρ, respectively. A strategy is a map from the set of all possible histories to the set of mixed actions of each respective player in the stage game.

When P1 plays an action a and P2 plays an action b, P1 obtains a stochastic reward with expected value $r(a,b)$. This reward is assumed to be a positive bounded random variable. The average reward until time t is $\hat{r}_t \stackrel{\triangle}{=} \frac{1}{t}\sum_{\tau=1}^{t} r_\tau$ (r_τ is the reward at time τ). When P1 plays a mixed strategy $p \in \Delta(\mathcal{A})$ ($\Delta(\mathcal{A})$ is

the set of probability vectors on $\mathcal{A}$) and P2 plays a mixed strategy $q \in \Delta(\mathcal{B})$ the obtained reward has the expected value of $r(p,q) \stackrel{\triangle}{=} \sum_{a\in\mathcal{A}} \sum_{b\in\mathcal{B}} p(a)q(b)r(a,b)$. The value of the associated one shot zero-sum game is denoted by v. Note that since P2's reward is unknown and possibly not relevant, v is not the value of any concrete game.

We assume that there are "signals" that are received by P1 after each stage of the game. These signals carry some information regarding P2's last action. Let $\mathcal{S} = \{1, \ldots, S\}$ denote the (finite) set of possible signals. Given that actions a and b were played at a given stage t, the observed signal s_t is a random variable with a given distribution $P(s|a,b)$. We further denote by $s(p,q)$ the expected signal frequency vector in $\Delta(\mathcal{S})$ which is generated when P1 plays p and P2 plays q. That is, if $\pi = s(p,q)$ then $\pi(s) = \sum_{a,b} P(s|a,b)p(a)q(b)$ for every $s \in \mathcal{S}$.

Let P2's empirical strategy by time t be denoted by q_t. That is, $q_t(b) = \frac{1}{t}\sum_{\tau=1}^{t} 1_{\{b_\tau=b\}}$. The Bayes envelope with respect to $q_t(b)$ is defined as $r^*_{BE}(q_t) \stackrel{\triangle}{=} \max_{p\in\Delta(A)} r(p,q_t)$. If P1 knew q_t in advance, then P1 could have played the maximizing (one-shot game) action against q_t repeatedly, and attain an average reward as high as $r^*_{BE}(q_t)$. Since q_t is not known in advance, an adaptive framework is needed. Even if we suppose that q_t is estimated as the repeated game is played, one must limit the performance comparison to q_t which are distinguishable. For that purpose we define the follows congruence class of q:

$$Q(q) = \{q' \in \Delta(\mathcal{B}) \ : \text{for every } p \in \Delta(\mathcal{A}) \quad s(p,q') = s(p,q)\}\,. \tag{1}$$

It can be shown that it suffices to examine the set of pure actions for P2 in (1). Essentially, P1 cannot distinguish between different q in $Q(q)$, therefore any scheme cannot strive to achieve more than:

$$r^*(q) \stackrel{\triangle}{=} \max_{p\in\Delta(A)} \min_{q\in Q(q)} r(p,q)\,. \tag{2}$$

We call (2) the Partial Observation Bayes Envelope (POBE) and consider it as the target to be achieved. Note that $r^*(q)$ depends on the *actual* unobserved strategy of P2. Formally, our goal is to attain envelopes in the sense of the following definition:

Definition 1. *A function $r : \Delta(\mathcal{B}) \to \mathbb{R}$ is* attainable *by P1 if there exists a strategy σ of P1 such that for every strategy ρ of P2:*

$$\liminf_{t\to\infty} (\hat{r}_t - r(q_t)) \geq 0 \quad P_{\sigma,\rho}\ a.s.$$

where $P_{\sigma,\rho}$ is the probability measure induced by σ and ρ.

3 P2 Determines the Signals

This section discusses the case where P2 alone determines the signal probability that is $P(s|a,b) = P(s|b)$. As a result one can write $s(q)$ for the expected signal

frequency when q is played (rather than $s(p,q)$). We will show that for the sake of consistency it suffices to consider the empirical signal frequency vector $\pi_t \in \Delta(\mathcal{S})$ which is defined as $\pi_t(s) \stackrel{\triangle}{=} \frac{1}{t}\sum_{\tau=1}^{t} 1_{\{s_\tau = s\}}$. With some abuse of notation, let $Q(\pi) = \{q \in \Delta(\mathcal{B}) : \ \pi = s(q)\}$ denote the set of all possible stationary strategies (or, equivalently, mixed single stage strategies) that can possibly result in the observation π. Furthermore, we define the set of "possible" observations as $\Pi \stackrel{\triangle}{=} \{\pi \in \Delta(\mathcal{S}) : Q(\pi) \neq \emptyset\}$. Observe that Π is a convex set. Since every q induces a single π it stands to reason to define a Bayes-like envelope as a function of the signal frequency π:

$$r^*(\pi) \stackrel{\triangle}{=} \min_{q \in Q(J(\pi))} r^*(q) = \max_{p \in \Delta(\mathcal{A})} \min_{q \in Q(J(\pi))} r(p,q), \tag{3}$$

where we define $J(\pi) : \Delta(\mathcal{S}) \to \Pi$ to be the Euclidean projection to of $\Delta(\mathcal{S})$ to Π. The following proposition provides the basic property of the POBE in the special case considered in this section.

Proposition 1. *If P1 does not affect the signals then $r^*(\pi)$ is a convex function of π on Π.*

Proof. Let $\pi_1, \ldots, \pi_k$ be probability vectors in Π. We have to show that for a convex combination $\alpha_1, \ldots, \alpha_k$ we have that $r^*(\sum_{i=1}^k \alpha_i \pi_i) \leq \sum_{i=1}^k \alpha_i r^*(\pi_i)$. Since P1 does not affect the signal probability we can write $Q(\pi) = \{q : Hq = \pi\}$ for the signalling matrix H ($H_{bs} = P(s|b)$). Let q_i denote a minimax mixed action in the stage game that agrees with π_i, that is $r^*(\pi_i) = \max_p r(p, q_i)$ (such a mixed action exists by the minimax theorem since $Q(\pi)$ is convex for every π.) Let $q_\alpha \stackrel{\triangle}{=} \sum_{i=1}^k \alpha_i q_i$, we have that $Hq_\alpha = \sum_{i=1}^k \alpha_i H q_i = \sum_{i=1}^k \alpha_i \pi_i$. Recalling the definition:

$$\begin{aligned} r^*(\sum_{i=1}^k \alpha_i \pi_i) &= \min_{q \in Q(\sum_{i=1}^k \alpha_i \pi_i)} \max_{p \in \Delta(\mathcal{A})} r(p,q) = \max_{p \in \Delta(\mathcal{A})} \min_{q \in Q(\sum_{i=1}^k \alpha_i \pi_i)} r(p,q) \\ &\leq \max_{p \in \Delta(\mathcal{A})} r(p, q_\alpha) = \max_{p \in \Delta(\mathcal{A})} \sum_{i=1}^k \alpha_i r(p, q_i) \\ &\leq \sum_{i=1}^k \max_{p \in \Delta(\mathcal{A})} \alpha_i r(p, q_i) = \sum_{i=1}^k \alpha_i r^*(\pi_i)\,. \end{aligned}$$

The second equality follows from the minimax theorem and the convexity of $Q(\pi)$. The first inequality holds since we fixed a specific q that agrees with the signals $\alpha_i \pi_i$. The third equality is a result of the linearity of the reward r in q. The second inequality is justified by the convexity of the max operator. □

Since P2 alone affects the signals probability, and by the continuity of $r^*(\pi)$ and $r^*(q)$ we next claim that $r^*(q_t) \to r^*(\pi_t)$ almost surely.

Proposition 2. *Suppose P1 does not affect the signals. Then for every pair of strategies σ, ρ*

$$\lim_{t \to \infty} r^*(q_t) - r^*(\pi_t) = 0 \quad \mathbf{P}_{\sigma,\rho}\text{-a.s.}$$

Proof. Since $Q(q)$ is convex and r is bilinear it follows that there exists $\tilde{q}_t \in \Delta(\mathcal{B})$ such that $r^*(q_t) = \max_p r(p, \tilde{q}_t)$. Let $\tilde{\pi}_t = s(\tilde{q}_t)$. It follows by our definitions so far that $r^*(\tilde{\pi}_t) = r^*(q_t)$. If the signals were deterministic then $\tilde{\pi}_t = \pi_t$ and the result follows. Generally, the signals are random so we need a more complicated argument. By Proposition 1, $r^*(\pi)$ is convex and therefore Lipschitz continuous over Π. It is therefore enough to show that $\|\pi_t - \tilde{\pi}_t\| \to 0$ almost surely. Note that the strategy ρ which governs the distribution of q_t and π_t is *not* assumed stationary so standard large deviation bounds cannot be immediately applied. Let $n_t(b)$ be the number of times action b was played by time t (i.e. $tq_t(b)$), and similarly let $n_t(b,s) = \sum_{\tau=1}^{t} 1_{\{b_\tau = b \& s_\tau = s\}}$ count how many times action b was used by P2 and signal s observed. Define the event $E_t(b,s) = \{|n_t(b,s) - n_t(b)P(s|b)| \geq \delta t\}$. By the union bound we have that:

$$\mathbf{P}(\|\pi_t - \tilde{\pi}_t\| \geq \epsilon) \leq \sum_{b,s} \mathbf{P}(E_t(b,s)),$$

with $\delta = \epsilon/SB$. To bound the probability of $E_t(b,s)$ define the events $F_\ell(b,s) = \{|f_\ell(b,s) - \ell P(s|b)| \geq \delta t\}$ where f_ℓ is the sum of ℓ independently distributed Bernoulli random variables with bias $P(s|b)$. Let $F(b,s) = \bigcup_{1 \leq \ell \leq t} F_\ell(b,s)$. We will now reason in terms of a single probability space on which the controlled process can be defined, under any strategy. Formally speaking, we need to define a joint probability space, but as this is standard we omit it for the sake of brevity. It follows that

$$\mathbf{P}(E_t(b,s)) \leq \mathbf{P}(F(b,s)) \leq \sum_{\ell=1}^{t} \mathbf{P}(F_\ell(b,s)) \leq \sum_{\ell=1}^{t} 2e^{-(\frac{\delta t}{\ell})^2 \ell c(b,s)} \leq 2te^{-\delta^2 t c'(b,s)},$$

where the third inequality is due to Hoeffding's inequality, and $c(b,s), c'(b,s)$ are some constants. Applying the union bound again gives $\mathbf{P}(\|\pi_t - \tilde{\pi}_t\| \geq \epsilon) \leq cte^{-t\epsilon^2 c'}$ for some constants c, c'. By the Borel-Cantelli Lemma it follows that $\mathbf{P}\left(\|\pi_t - \tilde{\pi}_t\| \geq \epsilon \text{ i.o}\right) = 0$. □

We can therefore term $r^*(\pi)$ the POBE as well. We start with discussing the case where the reward is observed in Section 3.1. We then discuss the case where the reward is not observed and a refined scheme is required. An example for a game where P2 determines the signals is presented in Section 3.3

3.1 Reward is observed

When the reward is observed, the Bayes envelope itself is attainable, as in, e.g., [4]. We now provide some insight to the case where the reward is not observed, using the proof technique of [3] for the perfect monitoring case. Let us recall the following definition. The setup is a repeated game with vector-valued reward vector m_t which average by time t is denoted by $\hat{m}_t = \frac{1}{t}\sum_{\tau=1}^{t} m_\tau$; see [7].

Definition 2. *A set $B \subseteq \mathbb{R}^k$ is* approachable *by P1 if there exists a B-*approaching *strategy σ^* of P1 such that*

$$d(\hat{m}_t, B) \to 0 \quad P_{\sigma^*,\rho}\text{-a.s., for every } \rho$$

where d is the Euclidean point to set distance.

If the reward itself is observed then a straightforward application of approachability theory would result in attaining $r^*(q_t)$.

Theorem 1. *If the reward is observed then the set*

$$\mathbf{B} = \{(\hat{r}, \pi) : \ \pi \in \mathit{\Pi} \ ; \ \hat{r} \geq r^*(\pi)\} \subseteq \mathbb{R} \times \mathit{\Pi} \, . \tag{4}$$

is approachable. Consequently, every B-approaching strategy attains the POBE.

Proof. We apply approachability theory for repeated matrix games (e.g., [7]). $r^*(\pi)$ is convex on $\mathit{\Pi}$ and therefore continuous (e.g., [8]). In order to apply approachability arguments, we construct the following game with vector-valued payoffs. Define the $1 + S$ dimensional reward vector $m = (\hat{r}, \pi) \in \mathbb{R} \times \mathit{\Delta}(\mathcal{S})$, indexed by $k \in \{0\} \cup \mathcal{S}$. When the observed signal was s and the observed reward was r,

$$m(k) \stackrel{\triangle}{=} \begin{cases} r \text{ if } k = 0 \\ 0 \text{ if } k \in \mathcal{S}, \ k \neq s \\ 1 \text{ if } k \in \mathcal{S}, \ k = s \end{cases} . \tag{5}$$

Thus, the first coordinate of the vector-valued average payoff vector is the average reward, the other coordinates are the relative frequencies of the signals. We note that m is generally a random vector, and is observed by P1 according to our assumptions. For a pair of mixed strategies $p \in \mathit{\Delta}(\mathcal{A})$ and $q \in \mathit{\Delta}(\mathcal{B})$, let

$$m(p, q) \stackrel{\triangle}{=} \sum_{a \in \mathcal{A}} p(a) \sum_{b \in \mathcal{B}} q(b) m(a, b) \, , \tag{6}$$

where $m(a, b)$ is the entry of the vector-valued game defined in Eq. (5). Since $\mathbf{B}$ is convex it suffices to prove that $\forall q \in \mathit{\Delta}(\mathcal{B})$ the set $M(\mathcal{A}, q) \stackrel{\triangle}{=} \text{CO}\big(\{m(a, q)\}_{a \in \mathcal{A}}\big)$ (CO is the convex hull operator) intersects $\mathbf{B}$ ($m(a, q)$ denotes the expected reward vector in the one shot game when P1 plays the action a and P2 plays the strategy q). Fix q, in the original game P1 has an optimal deterministic action $a^* \in \mathcal{A}$ against q. Consider the signal π that satisfies $\pi(s) = \sum_{b \in \mathcal{B}} P(s|a^*, b) q(b)$ for every $s \in \mathcal{S}$. By definition $q \in Q(\pi)$. r^* satisfies $r^*(\pi) = \max_{p \in \mathit{\Delta}(\mathcal{A})} \min_{q \in Q(\pi)} r(p, q)$ but since the pure strategy a^* is optimal against q, we have that $\forall p \in \mathit{\Delta}(\mathcal{A})$, $r(p, q) \leq r(a^*, q)$ so at π we have that $r^*(\pi) \leq r(a^*, q)$. We conclude that $\mathbf{B}$ is approachable since $(r(a^*, q), s(q)) \in \mathbf{B}$ holds for every $q \in \mathit{\Delta}(\mathcal{B})$. Since $\mathbf{B}$ is approachable and since r^* is continuous it follows that by our definitions that by approaching $\mathbf{B}$ the first coordinates difference decreases to 0 so the result follows. □

It follows by [7] that the approaching strategy is to play at time t a minimax mixed action in the one shot matrix game defined by the projection of the vector reward, defined in Eq. (5), on the direction from $(\hat{r}_t, \pi_t)$ to a closest point in $\mathbf{B}$ (defined in Eq. (4)).

3.2 Unobserved reward

In this section we discuss the case where the reward is not observed. Surprisingly, we are still able to attain the POBE when P1 does not affect the signaling structure.

The following strategy attains the POBE in the case where the reward is not observed by approaching the set $\mathbf{B}$. The strategy advances is stages. In each stage, the same mixed action is used repeatedly. A note about the notations is due. We use superscripts for stage indices and subscript for time indices.

Algorithm 1 : An Attaining Strategy when the Reward is Unobserved

1. Start: $t = 0$, $i = 1$, $t^0 = 0$.
2. Choose an arbitrary p^1, Goto 6.
3. If $\tilde{r}_t \geq r^*(\pi_t)$ let p^i be arbitrary, Goto 6.
4. If $\tilde{r}_t < r^*(\pi_t)$ find the direction u^i from the point $(\tilde{r}_t, \pi_t)$ to the closest point in the set $\mathbf{B}$ defined in Eq. (4).
5. Let p^i be a maximizer of $\max_{p \in \Delta(\mathcal{A})} \min_{q \in \Delta(\mathcal{B})} m(p,q) \cdot u^i$, where $m(p,q)$ is given in (6), and $\cdot$ is the standard dot product.
6. Repeat T^i times:
   $t = t + 1$.
   Pick a random action a_t according to p^i, play a_t and observe s_t.
7. $t^i = t$. Set: $\pi^i(s) = \frac{1}{T^i} \sum_{\tau=t^{i-1}+1}^{t^i} 1_{\{s_\tau = s\}}$; $\tilde{r}^i = \min_{q \in Q(J(\pi^i))} r(p^i, q)$;
   $\tilde{r}_{t^i} = \frac{1}{t^i} \sum_{j=1}^{i} T^j \tilde{r}^j$.
8. $i = i + 1$. Goto 3.

Theorem 2. *By playing Algorithm 1 with $T^i = i^2$ the POBE is attained.*

Proof. The proof is deferred to the Appendix.

The basic idea behind Algorithm 1 is that the fictitious reward $\tilde{r}$ replaces the true (unobserved) reward. For large t the average reward and the observed signals frequency change slowly, so that by choosing T^i small enough the same strategy is almost optimal during the i-th interval. By choosing T^i large enough we guarantee that the average reward is asymptotically not lower than the fictitious reward. We also note that any polynomial T^i would lead to a similar convergence result.

3.3 An Example

We now provide an example to the case where the signals depend only on P2's actions. This example is related to the field of Internet Protocols and is motivated by source routing and the ability of modern routers to supply information to the sending machine. In TCP communication protocol a node (e.g., a computer) sends a packet and receives an acknowledgement signal from the destination node. It may happen that a packet gets lost in the way and should be resent. The packet is resent if there is an indication that the packet is lost (i.e., after enough time). Suppose a node sends packets with not only destination address

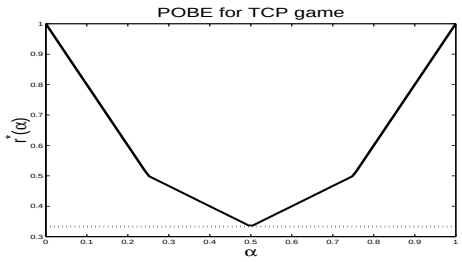

Fig. 2. $r^*(\alpha)$ is the convex line above the value v = 1/3 (dotted).

but also route of the packet (this is called source routing). The router sends the packet in the route required by the node, however since the router has a better picture of the network it may know if the packet has a good chance to get lost. The router may be able to return a small message containing some limited information (this is not an acknowledge - just an indication). See [9] for an example of a scheme where a single bit is used to provide information regarding congestion. The following example demonstrates such a situation. Suppose that there are three links, out of which only one is operative. The node does not know which link is operative. The router may use only a single bit to hint the node which of the three links was working. The performance measure is to get as many packets across. Let P1 be the node that sends packets (the row player) and P2 is the network (the column player), P1's reward-signal matrix is (each entry is denoted by (r, s)):

	Link 1 up	Link 2 up	Link 3 up
Send via link 1	$(1, a)$	$(0, a\|b)$	$(0, b)$
Send via link 2	$(0, a)$	$(1, a\|b)$	$(0, b)$
Send via link 3	$(0, a)$	$(0, a\|b)$	$(1, b)$

The available signals are $\{a, b\}$. The signals generating probability $P(s|b)$ satisfies $P(a|1) = P(b|3) = 1$ and $P(a|2) = P(b|2) = 1/2$. This means that when P2 plays action 1 or 3 P1 observes signals a, b deterministically (respectively), and when action 2 is played a random signal is observed with equal probabilities. Observe that this example cannot be cast into the model of [6], and that the Bayes envelope (rather than the POBE) cannot be attained. To see that consider two strategies of P2, the first is to always play action 2, and the second to always play actions 1 and 3 with equal probability. P1 cannot distinguish between the two strategies. The POBE of either strategies is 1/3, however the best response reward against the first is 1 and against the second is 1/2. It follows that if P2 throws a coin before the game begins and with probability 1/3 plays the first strategy and with probability 2/3 plays the second, then no matter what strategy P1 employs a reward of no more than 1/3 can be guaranteed. Let α be the frequency of signal a. We identify the signal frequency with $0 \leq \alpha \leq 1$. Obviously, $r^*(\alpha)$ is symmetric around 1/2. The possible stage game strategies that agree with α satisfy $Q(\alpha) = \text{CO}\{(\alpha, 0, 1-\alpha), (0, 2\alpha, 1-2\alpha)\}$, which leads

to the conclusion that $r^*(\alpha) = 1 - 2\alpha$ for $\alpha \leq 1/4$. For $1/4 \leq \alpha \leq 1/2$ a straight forward calculation shows that $r^*(\alpha) = \min_{0\leq\beta\leq 1} \max\{\alpha\beta, 2\alpha - 2\alpha\beta, 1 + \alpha\beta - 2\alpha\} = \frac{2}{3} - \frac{2\alpha}{3}$. The graph of $r^*(\alpha)$ is given in Figure 2. The value of the game is $\mathrm{v} = 1/3$ which is the result of the unique single stage game minimax strategy $q^* = (1/3, 1/3, 1/3)$. Note that $r^*(\pi) > \mathrm{v}$ for every π for which $q^* \notin Q(\pi)$. This behavior is typical, as explained in Remark 1.

4 The General Case

In this section we study the general model (i.e., both players affect the signals). We suggest a simplified target which is fairly easy to attain, though it does not promise a reward as high as $r^*(q_t)$.

The idea is to define a function of the signal frequency, which is attainable. Redefine

$$Q(\pi) \stackrel{\triangle}{=} \{q \in \Delta(\mathcal{B}) : \quad \text{there exists } p \in \Delta(\mathcal{A}) \text{ such that } \pi = s(p, q)\}.$$

$Q(\pi)$ is the set of all possible stationary strategies (or, equivalently, mixed actions) that can possibly result in the observation π. As before, let $\Pi = \{\pi : Q(\pi) \neq \emptyset\}$. Note that in this case Π may be not convex. We can now formulate the empirical Bayes envelope in the reward signal space, $r^*_E(\pi) : \Delta(\mathcal{S}) \to \mathbb{R}$, as

$$r^*_E(\pi) \stackrel{\triangle}{=} \max_{p\in\Delta(\mathcal{A})} \min_{q\in Q(J(\pi))} r(p, q),$$

with $J(\pi)$ being the projection onto Π (with ambiguities resolved in arbitrary manner). This definition coincides with Eq. (3) for games where P1 does not affect the signalling structure. It turns out that in general r^*_E is *not* attainable ([10]). This may be the case even if r^*_E is well defined for all π and $Q(\pi) \neq \emptyset$ for all π. An attainable solution concept we suggest is the *convex Bayes envelope* that is defined as the lower convex hull of $r^*_E(\pi)$ on Π and is denoted by r^c (where $r^c(\pi) = r^c(J^c(\pi))$ for $\pi \notin \mathrm{CO}(\Pi)$, and $J^c(\pi)$ is the projection of $\Delta(\mathcal{S})$ onto $\mathrm{CO}(\Pi)$.) Let $\mathbf{C}$ as the set $\{(r, \pi) : r \geq r^c(\pi)\}$.

Theorem 3. *Suppose that Algorithm 1 is used with* $\mathbf{C}$ *replacing* $\mathbf{B}$. *Then for every strategy of P2:*

$$\liminf_{t\to\infty} \hat{r}_t - r^c(\pi_t) \geq 0 \quad a.s.$$

Proof. (Outline) The set $\mathbf{C}$ is convex by definition. Showing that $\mathbf{C}$ is approachable in a game were both signals and reward are observed follows exactly as in Theorem 1. When the reward is not observed, an analogue algorithm to Algorithm 1 may be suggested with replacing $\mathbf{B}$ with $\mathbf{C}$ and the analysis of Theorem 2 still holds. □

Remark 1. By attaining $r^c(\pi_t)$ instead of $r^*(q_t)$ the performance guarantees deteriorate. Obviously $r^c(\pi) \geq \mathrm{v}$ where v is the value of the one shot zero-sum game. It can be shown ([10]) that if P2 has a unique minimax strategy q^* then $r^c(\pi) > \mathrm{v}$ for every π for which $q^* \notin Q(\pi)$.

Recall the treatment game from the introduction. In this game both players affect the signals. As shown in Figure 1, r^* is not convex. Moreover, it is not even continuous near the point $(\pi_a, \pi_b) = (1, 0)$. The difference between the attainable envelope, r^c and the r^* is plotted in Figure 1c. In this game $r^c(\pi) > \mathrm{v}$ for every π that does not agree with the unique minimax strategy $(1/2, 1/2)$ of Nature.

The main question which remains open is how to attain the POBE (from Definition 1) in the general case where both players affect the signalling probabilities. We believe that combining the idea of fictitious reward with a prediction with expert advice framework may be used for attaining the POBE.

Acknowledgements This research was supported by the fund for the promotion of research at the Technion.

References

1. D. Fudenberg and D. Levine. Universal consistency and cautious fictitious play. *Journal of Economic Dynamic and Control*, 19:1065–1190, 1995.
2. J. Hannan. Approximation to Bayes risk in repeated play. In M. Dresher, A. W. Tucker, and P. Wolde, editors, *Contribution to The Theory of Games, III*, pages 97–139. Princeton University Press, 1957.
3. D. Blackwell. Controlled random walks. In *Proc. International Congress of Mathematicians, 1954*, volume III, pages 336–338. North-Holland, 1956.
4. P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The non-stochastic multi-armed bandit problem. To appear in SIAM journal of Computation, 2002.
5. A. Rustichini. Minimizing regret: the general case. *Games and Economic Behavior*, 29:224–243, November 1999.
6. A. Piccolboni and C. Schindelhauer. Discrete prediction games with arbitrary feedback and loss. In D. Helmbold and B. Williamson, editors, *Fourteenth Annual Conference on Computation Learning Theory*, pages 208–223. Springer, 2001.
7. D. Blackwell. An analog of the minimax theorem for vector payoffs. *Pacific J. Math.*, 6(1):1–8, 1956.
8. R.T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.
9. K. Ramakrishnan, S. Floyd, and D. Black. The addition of explicit congestion notification (ECN) to IP. IETF, Tech. Rep., 2001.
10. S. Mannor and N. Shimkin. Regret minimization in signal space for repeated matrix games with partial observations. Technical report EE- 1242, Faculty of Electrical Engineering, Technion, Israel, March 2000. Available from http://web.mit.edu/˜shie/www/pubs.htm.
11. N. Shimkin. Extremal large deviations in controlled I.I.D. processes with applications to hypothesis testing. *Adv. Appl. Prob.*, 25:875–894, 1993.
12. D.P. Bertsekas and J.N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1996.

A Proof of Theorem 2

Let $r^i = \frac{1}{T^i} \sum_{\tau=t^{i-1}+1}^{t^i} r_\tau$ be the actual (unobserved) reward in the i-th interval. The actual (unobserved) average reward can be written as

$$\hat{r}_{t^i} = \frac{1}{t^i} \sum_{j=1}^{i} T^j r^j. \tag{7}$$

First, let us bound the probability that the worst case estimate is too optimistic:

Lemma 1. *There exist constants C and D such that for every i:*

$$\mathbf{P}(\tilde{r}^i \geq r^i + \epsilon) \leq CT^i e^{-D\epsilon^2 T^i}.$$

Proof. Let q^i denote P2's true unobserved empirical frequency by the ith interval. That is, $q^i(b) = \frac{1}{T^i} \sum_{\tau=t^{i-1}+1}^{t^i} 1_{\{b_\tau = b\}}$. We have that:

$$\mathbf{P}(\tilde{r}^i \geq r^i + \epsilon) \leq \mathbf{P}(r(p^i, q^i) - r^i \geq \epsilon/2) + \mathbf{P}(\tilde{r}^i - r(p^i, q^i) \geq \epsilon/2)\,. \qquad (8)$$

Recall that $r(p^i, q^i)$ is the expected reward in the one shot game when P1 plays p^i and P2 plays q^i. We now bound each of the terms in Eq. (8). As in the proof of Proposition 2 we cannot assume stationarity of the strategy of P2, so we will use a similar technique. One can use the results of [11], however, this is at the expense of finite time bounds. Let us reset the time of the beginning of the ith interval to 1. Consider the event

$$E(a,b) = \{|\sum_{\tau=1}^{T^i} 1_{\{a_\tau = a, b_\tau = b\}} r_\tau - n_{T^i}(a,b) r(a,b)| > \delta T^i\},$$

where $r(a,b)$ is the expected reward when P1 plays a and P2 plays b, and $n_{T^i}(a,b) = \sum_{\tau=1}^{T^i} 1_{\{a_\tau = a, b_\tau = b\}}$ counts the number of time P1 played a and P2 played b. It follows by the union bound that

$$\mathbf{P}(r(p^i, q^i) - r^i \geq \epsilon/2) \leq \sum_{a,b} E(a,b),$$

with $\delta = \frac{\epsilon}{2AB}$. Consider the event $F_\ell(a,b) = \{|\sum_{k=1}^{\ell} f_k(a,b) - \ell r(a,b)| > \delta T^i\}$ where $f_k(a,b)$ is an I.I.D. random variable with the same distribution as r when a and b are played by P1 and P2, respectively. Reasoning in terms of a single probability space we get that:

$$\mathbf{P}(E(a,b)) \leq \sum_{\ell=1}^{T^i} \mathbf{P}(F_\ell(a,b)) \leq \sum_{\ell=1}^{T^i} 2e^{-c(a,b)(\frac{\delta T^i}{\ell})^2 \ell} \leq 2T^i e^{-c'(a,b)\epsilon^2 T^i},$$

where the second inequality follows by Hoeffding's inequality, and $c(a,b), c'(a,b)$ are some constants. Summing over a, b and using the union bound we get that $\mathbf{P}(r(p^i, q^i) - r^i \geq \epsilon/2) \leq C_1 T^i e^{-D_1 T^i \epsilon^2}$ for every p^i and q^i. Note that the inequality holds with the same constants for all p^i and q^i since r is bounded for every a and b and therefore for all a and b ($\mathcal{A}$ and $\mathcal{B}$ are finite.)

To bound the second term of Eq. (8) let $\tilde{\pi}^i = s(q^i)$. In Proposition 2 we proved that there are constants c and c' such that: $\mathbf{P}(\|\pi^i - \tilde{\pi}^i\| > \epsilon) \leq cT^i e^{-c' T^i \epsilon^2}$. Consider the function $r_{p^i}(\pi) = \inf_{q \in Q(J(\pi))} r(p^i, q)$. It follows that $r_{p^i}(\pi) : \Delta(\mathcal{S}) \to \mathbb{R}$ is a continuous function. Moreover, it is easily verified that $\{r_{p^i}\}_{p^i \in \Delta(\mathcal{A})}$

are Lipschitz continuous, and that the Lipschitz constant of all $\{r_{p^i}\}_{p^i \in \Delta(\mathcal{A})}$ is bounded by some K. By our definitions, we have that $r_{p^i}(\tilde{\pi}^i) \leq r(p^i, q^i)$. Using the uniform continuity it follows that with probability of at least $1 - cT^i e^{-\frac{c'}{4K^2} T^i \epsilon^2}$ we have that $\|\tilde{\pi} - \pi^i\| \leq \frac{\epsilon}{2K}$ so that $|r_{p^i}(\tilde{\pi}^i) - r_{p^i}(\pi^i)| \leq \epsilon/2$. Recalling that $\tilde{r}^i = r_{p^i}(\pi^i)$, it follows that the probability of the event $\{\tilde{r}^i - r(p^i, q^i) \geq \epsilon/2\}$ is at most $cT^i e^{-\frac{c'}{4K^2} T^i \epsilon^2}$. The lemma follows by combining the bounds for the two terms. □

The asymptotical relation between $\tilde{r}$ and $\hat{r}$ is given in the following lemma.

Lemma 2. *For every strategy of P2* $\liminf_{i \to \infty} \hat{r}_{t^i} - \tilde{r}_{t^i} \geq 0$ *almost surely.*

Proof. By our choice of $T^i = i^2$ we can apply the Borel-Cantelli Lemma to deduce that for every $\epsilon > 0$ we have that $\mathbf{P}(\tilde{r}^i \geq r^i + \epsilon \text{ i.o}) = 0$. Recalling our definitions we have that $\hat{r}_{t^i} - \tilde{r}_{t^i} = \frac{1}{t^i} \sum_{j=1}^{i} T^j (r^j - \tilde{r}^j)$. Since after some random index k we have that $r^j - \tilde{r}^j > -\epsilon/2$ for every $j > k$, and since the reward is bounded until t^k we have that for some random ℓ for all $j > \ell$ we have that $\hat{r}_{t^j} - \tilde{r}_{t^j} \geq -\epsilon$. The lemma follows since this is true for every $\epsilon > 0$. □

We can now imitate the proof of Blackwell's theorem. In order to provide a Blackwell like theorem we need to provide a geometric condition regarding the behavior of the fictitious reward in every interval. Let C^i be the closest point in $\mathbf{B}$ to the point $(\tilde{r}_{t^{i-1}}, \pi_{t^{i-1}})$ in the $S+1$ dimensional space (using Euclidean norm). That is $C^i = \text{argmin}_{c \in \mathbf{B}} d(c, (\tilde{r}_{t^{i-1}}, \pi_{t^{i-1}}))$. Note that C^i is well defined by the convexity of $\mathbf{B}$. When $\tilde{r}_{t^{i-1}} < r^*(\pi_{t^{i-1}})$ (step 4 of the algorithm) the direction u^i is set as $u^i = \frac{C^i - (\tilde{r}_{t^{i-1}}, \pi_{t^{i-1}})}{\|C^i - (\tilde{r}_{t^{i-1}}, \pi_{t^{i-1}})\|}$. Though not critical to the following when $\tilde{r}_{t^{i-1}} \geq r^*(\pi_{t^{i-1}})$ we define $u^i = 0$. Finally, we set $\tilde{m}^i = (\tilde{r}^i, \pi^i) \in \mathbb{R} \times \Delta(\mathcal{S})$ the $S+1$ dimensional vector whose first coordinate is the added fictitious reward, and whose remaining S coordinates are added to the observation frequency vector (normalized by the length of the ith stage.)

Lemma 3. *When Algorithm 1 is used then for every strategy ρ of P2 we have that*

$$\liminf_{i \to \infty} u^i \cdot (\tilde{m}^i - C^i) \geq 0 \quad a.s.$$

Proof. Fix $\epsilon > 0$, we will show that $\liminf_{i \to \infty} u^i \cdot (\tilde{m}^i - C^i) \geq -\epsilon$ almost surely. If $\tilde{r}_{t^{i-1}} \geq r^*(\pi_{t^{i-1}})$ then $u^i = 0$ and equality holds. Assume $\tilde{r}_{t^{i-1}} < r^*(\pi_{t^{i-1}})$. By Proposition 1 the set $\mathbf{B}$ is approachable in the original game when the reward is observed. It follows from Blackwell's theorem that in the original game by choosing p^i we have that for every $q \in \Delta(\mathcal{B})$

$$u^i \cdot (m(p^i, q) - C^i) \geq 0, \tag{9}$$

where $m(p, q)$ is given in Eq. (6). Suppose first that $\pi^i \in \Pi$. Recalling our definitions, there exists $q' \in \Delta(\mathcal{B})$ such that $\tilde{r}^i = r(p^i, q')$ and $\pi^i = s(q')$. Now, Eq. (9) holds for q', so that $u^i \cdot (m(p^i, q') - C^i) \geq 0$, and therefore $u^i \cdot (\tilde{m}^i - C^i) \geq 0$. As in the proof of Lemma 1, we can show that $\mathbf{P}(\|\pi^i - \tilde{\pi}^i\| \geq \epsilon) \leq cT^i e^{-c' T^i \epsilon^2}$,

so that by our choice of T^i and using the Borel-Cantelli lemma $\mathbf{P}(\|\pi^i - \tilde{\pi}^i\| \geq \epsilon \text{ i.o}) = 0$. Let $\tilde{q}^i$ be the minimizer of $\min_{q \in J(\pi^i)} r(p^i, q)$. we have that:

$$\begin{aligned} u^i \cdot (\tilde{m}^i - C^i) &= u^i \cdot ((\tilde{r}^i, J(\pi^i)) - C^i) + u^i \cdot \left((\tilde{r}^i, \pi^i) - (\tilde{r}^i, J(\pi^i))\right) \\ &\geq -\|\pi^i - J(\pi^i)\|, \end{aligned}$$

where we used Eq. (9) for the first term and the Cauchy-Schwartz inequality for the second. The result follows from some random time on since from some time on $\|\pi^i - \tilde{\pi}^i\| \leq \epsilon$. □

We can now use Lemma 3 to show that $(\tilde{r}_{t^i}, \pi_{t^i})$ converge to $\mathbf{B}$.

Lemma 4. *Suppose that Algorithm 1 is used. Then for every strategy of P2:*

$$\lim_{i \to \infty} d\left((\tilde{r}_{t^i}, \pi_{t^i}), \mathbf{B}\right) = 0 \quad a.s.$$

Proof. Let $\mathbf{B}^\eta$ be the η-expansion of $\mathbf{B}$ (i.e. the union of all the points whose distance from $\mathbf{B}$ is η or less). Let $\tilde{m}_{t^i} \stackrel{\triangle}{=} (\tilde{r}_{t^i}, \pi_{t^i})$ be the $S+1$ vector whose first coordinate is the fictitious reward and whose remaining S coordinates are the empirical signal frequency. We will show that $d\left(\tilde{m}_{t^i}, \mathbf{B}^\eta\right) \to 0$.
Fix $\eta > 0$. By our choices of T^i we have that $t^i = \frac{i(i+1)(2i+1)}{6}$. By Lemma 3 we have that after some finite random time either $\tilde{m}_{t^{i-1}} \in \mathbf{B}^\eta$ or that there is an advancement in direction u^i, i.e.:

$$\tilde{m}^i \cdot u^i \geq u^i \cdot C^i - \eta\,,$$

where C^i is the closest point to $\tilde{m}_{t^{i-1}}$ in $\mathbf{B}$. Let y_{t^i} be the closest point to $\tilde{m}_{t^i}$ in $\mathbf{B}^\eta$. It follows that if $\tilde{m}_{t^i} \notin \mathbf{B}^\eta$ then $y_{t^i} = C^{i+1} - \eta u^{i+1}$. We therefore have that after some finite (a.s.) random time

$$\tilde{m}^{i+1} \cdot u^{i+1} \geq u^{i+1} \cdot C^i - \eta u^{i+1} \cdot u^{i+1} = u^{i+1} \cdot y_{t^i}. \tag{10}$$

Let d_i denote the distance from $\tilde{m}_{t^i}$ to $\mathbf{B}^\eta$, i.e., $d_i = d((\tilde{r}_{t^i}, \pi_{t^i}), \mathbf{B}^\eta)$.

Consider the square of the distance, d_i^2. It follows that:

$$\begin{aligned} d_{i+1}^2 &= \|\tilde{m}_{t^{i+1}} - y_{t^{i+1}}\|_2^2 \leq \|\tilde{m}_{t^{i+1}} - y_{t^i}\|_2^2 = \|\tilde{m}_{t^{i+1}} - \tilde{m}_{t^i} + \tilde{m}_{t^i} - y_{t^i}\| \\ &= \|\tilde{m}_{t^i} - y_{t^i}\|_2^2 + \|\tilde{m}_{t^{i+1}} - \tilde{m}_{t^i}\|_2^2 + 2(\tilde{m}_{t^i} - y_{t^i}) \cdot (\tilde{m}_{t^{i+1}} - \tilde{m}_{t^i}). \end{aligned}$$

The first element is simply d_i^2. Since $t_i = O(i^3)$ and the reward is bounded, the second element can be bounded by $\frac{D}{i^2}$. The third element is more tricky. Since

$$\tilde{m}_{t^{i+1}} - \tilde{m}_{t^i} = (\frac{1}{t^{i+1}} - \frac{1}{t^i}) t^i \tilde{m}_{t^i} + \frac{T^{i+1}}{t^{i+1}} \tilde{m}^{i+1} = \frac{t^i - t^{i+1}}{t^{i+1}} \tilde{m}_{t^i} + \frac{T^i}{t^{i+1}} \tilde{m}^{i+1}$$

we have that:

$$\begin{aligned} (\tilde{m}_{t^i} - y_{t^i}) \cdot (\tilde{m}_{t^{i+1}} - \tilde{m}_{t^i}) &= (\tilde{m}_{t^i} - y_{t^i}) \cdot (\tilde{m}_{t^{i+1}} - \frac{t^i - t^{i+1}}{t^{i+1}} y_{t^i} + \frac{t^i - t^{i+1}}{t^{i+1}} y_{t^i} - \tilde{m}_{t^i}) \\ &= \frac{t^i - t^{i+1}}{t^{i+1}} (\tilde{m}_{t^i} - y_{t^i}) \cdot (\tilde{m}_{t^i} - y_{t^i}) \\ &\quad + \frac{t^i - t^{i+1}}{t^{i+1}} (\tilde{m}_{t^i} - y_{t^i}) \cdot \left(\tilde{m}^{i+1} - y_{t^i}\right). \end{aligned} \tag{11}$$

Since $i^3 c \geq t^i$ (for some c) and recalling that $T^i = i^2$, we can bound the fraction $\frac{t^{i+1}-t^i}{t^{i+1}} \geq C/i$. The first term (Eq. (11)) is therefore bounded by $-d_i^2 \frac{C_1}{i}$. We get the following inequality:

$$d_{i+1}^2 \leq (1 - \frac{C_1}{i}) d_i^2 + \frac{C_2}{i^2} + 2C_3 \frac{(\tilde{m}_{t^i} - y_{t^i}) \cdot (\tilde{m}^{i+1} - y_i)}{i+1}\,, \tag{12}$$

where C_1,C_2, and C_3 are positive constants. Now according to Eq. (10) the last term in (12) is negative after some random time. We therefore have that $d_{i+1}^2 \leq (1 - \frac{C_1}{i}) d_i^2 + \frac{C_2}{i^2}$ from some point on. It now follows that $d_{i+1}^2 \to 0$ almost surely using, e.g., [12, Page 117]. □

Using Lemma 4 we have that $d\big((\tilde{r}_{t^i}, \pi_{t^i}), \mathbf{B}\big) \to 0$ almost surely. By Lemma 2 we have that $\liminf_{i\to\infty} \hat{r}_{t^i} - \tilde{r}_{t^i} \geq 0$. Since $\mathbf{B}$ is the epigraph of r^* we therefore have that also $d(\hat{m}_{t^i}, \mathbf{B}) \to 0$ almost surely, where $\hat{m}_t = (\hat{r}_{t^i}, \pi_{t^i})$. To conclude the proof we need to show that the above bound holds for all t and not just for t^i. Let $i(t)$ denote the maximal t^i smaller than t (i.e. $i(t) = \max\{i : t^i \leq t\}$). By the triangle inequality

$$d(\hat{m}_t, \mathbf{B}) \leq d(\hat{m}_{t^{i(t)}}, \mathbf{B}) + \|\hat{m}_{t^{i(t)}} - \hat{m}_t\|_2\,.$$

The first term converges to 0 by the above. To bound the second term we let $m_\tau = (r_\tau, \pi_\tau)$.

$$\begin{aligned}
\|\hat{m}_{t^{i(t)}} - \hat{m}_t\|_2 &= \big\|\frac{1}{t}\sum_{\tau=1}^{t} m_\tau - \frac{1}{t^{i(t)}}\sum_{\tau=1}^{t^{i(t)}} m_\tau\big\|_2 = \big\|\frac{1}{t}\sum_{\tau=t^{i(t)}+1}^{t} m_\tau + \frac{t^{i(t)}-t}{t t^{i(t)}}\sum_{\tau=1}^{t^{i(t)}} m_\tau\big\|_2 \\
&\leq \frac{1}{t}\big\|\sum_{\tau=t^{i(t)}+1}^{t} m_\tau\big\|_2 + \frac{t-t^{i(t)}}{t t^{i(t)}}\big\|\sum_{\tau=1}^{t^{i(t)}} m_\tau\big\|_2 \\
&= \frac{t-t^{i(t)}}{t}\left(\frac{1}{t-t^{i(t)}}\big\|\sum_{\tau=t^{i(t)}+1}^{t} m_\tau\big\|_2 + \frac{1}{t^{i(t)}}\big\|\sum_{\tau=1}^{t^{i(t)}} m_\tau\big\|_2\right),
\end{aligned} \tag{13}$$

where the inequality is due to the triangle inequality. By our construction of t^i it follows that $\lim_{t\to\infty} \frac{t-t^{i(t)}}{t} = 0$. By the boundedness of m_t it follows that both terms inside the parenthesis of Eq. (13) are contained in some ball of finite radius. Consequently, Eq. (13) converges to 0 almost surely.

Since $d(\hat{m}_t, \mathbf{B})$ converges to 0 almost surely it follows using the same continuity argument in Theorem 1 that $\liminf_{t\to\infty} \hat{r}_t - r^*(\pi_t) \geq 0$ almost surely. Finally, using Proposition 2 we have that $\lim_{t\to\infty} r^*(\pi_t) - r^*(q_t) = 0$ almost surely. The result follows by combining the two limits. □