No regret with delayed information*

David Lagziel[†] and Ehud Lehrer[‡]

December 4, 2012

Abstract:

We consider a sequential decision problem where the decision maker is informed of the actual payoff with delay. We introduce deterministic and stochastic delayed information structures under which the decision maker has regret-free strategies. We show how the convergence rate to no-regret is sensitive to changes in the information structure and how, even with a linear time delay, the decision maker has a bounded-regret strategy. We then apply our regret-free strategy to cases with transaction cost and show that the strategy remains regret-free.

Journal of Economic Literature classification numbers: C61, C72, D23, D81, D82, D83

Keywords: Regret-free, delayed information, approachability, transaction cost.

*The authors wish to thank Eilon Solan for his valuable comments. This research was supported in part by the Google Inter-university center for Electronic Markets and Auctions. Lehrer acknowledges the support of the Israel Science Foundation, Grant #762/045. Lagziel acknowledges the support of the Israel Science Foundation, Grant #538/11.

[†] School of Mathematical Sciences, Tel Aviv University, Tel Aviv 69978, Israel. e-mail: davidlag@post.tau.ac.il.

[‡] School of Mathematical Sciences, Tel Aviv University, Tel Aviv 69978, Israel and INSEAD, Bd. de Constance, 77305 Fontainebleau Cedex, France. e-mail: lehrer@post.tau.ac.il.

1 Introduction

1.1 Regret-free strategies

Both in real life and in game theory, the term "regret" relates to the amount of dissatisfaction a person might have about his past actions or behavior. This kind of assessment in retrospect is significant when the decision making is done over and over again.

In a sequential decision problem the decision maker (DM) ought to take an action every period. Upon taking an action he receives a stage-payoff that depends also on a stochastic state of nature. Consider, for instance, a trader interacting with a market. The trader is unaware of the exact way the market evolves. Instead, he assess his past performance in comparison with the best he could do against the market's empirical distribution of states. Having no-regret in this situation formally means that the decision maker's past performance is not inferior to the performance of the best among a list of pre-specified strategies. For instance, having no external regret means that past performance is at least as good as that of the best stationary strategy.

Hannan [14] was the first to discuss regret-free strategies. He proved the existence of an external regret-free strategy. Hannan's results implies[1] that the trader has a strategy that could perform as well as any stationary strategy. In particular, the DM is doing as well as the constant strategy that plays the best response against the empirical distribution of states.

This paper deals with a situation where the information about past outcomes reaches the DM with delay. It applies, for instance, to a case where market information, as it often occurs in real life, does not reach the trader instantaneously but rather with delay. We show that delayed information might still allow regret-free strategies to exist. Exact bounds on the time delay that ensures the existence of a regret-free strategy are provided.

1.2 Main results

The first part of the paper assumes that the DM is eventually informed of the outcome of every stage. That is, any information related to the decision problem that might have occurred in the past, will eventually reach the DM. It turns out that if the time delay grows at a moderated rate,[2] for instance if the information about the state of nature at time n is delayed for 100 periods, then a regret-free strategy exists. A simple example shows that this bound is tight, meaning that if the delay time increases faster than $o(n)$, then a no-regret strategy might not exist.

The second part studies regret-free strategies with stochastic delayed information. We prove

[1] Under the assumption that there are finitely many states of nature.

[2] Formally, if the information about time n is obtained before $o(n)$ periods after n (the notation $o(n)$ stands for any function $f(n)$ such that $f(n)/n \to 0$).

that a regret-free strategy exists even when the DM is not fully informed of past states. Specifically, we show that whenever the DM is getting a fair amount of informative data with sufficiently high probability[3] he has a regret-free strategy. We exemplify this result through a specific model where the information the DM receives about previous outcomes is not only delayed, but also inaccurate, and yet allows a regret-free strategy to exist.

The third part is devoted to studying the rate of convergence of past performance to being regret-free. Blackwell's condition and Hannan's no-regret result prove that the rate of convergence cannot be better than $n^{-1/2}$. In other words, at time n, the best the DM could hope for is that past performance falls short of the best achievable by $cn^{-1/2}$, where c is a constant. We give a new bound on the rate of convergence based on the time delay. We also show that when time delay increases with time in a linear fashion,[4] the strategy discussed becomes a bounded-regret strategy, rather than regret-free.

Analyzing the case of time delay has a further advantage, namely, it enables one to generalize no-regret strategies to cases that involve transaction cost. We show that a regret-free strategy exists also when the DM incurs an extra cost any time he decides to change his mode of action.

1.3 No-regret and approachability

The main tool we use is an extension of Blackwell's Approachability Theorem [2]. Blackwell considered a sequential decision problem where the DM's payoff at each stage is a vector in a finite-dimensional space. A pre-specified target set C in the payoff space is said to be approachable by the DM if he has a strategy that ensures that the difference between the average payoff and its closest point in C converges to zero with probability 1 as time grows unboundedly. Blackwell characterized the sets that are approachable when the DM fully monitors the outcomes of the game and the information about time n is available immediately.

Hart and Mas-Colell [15] were the first to note that no-regret theorems can be proven using Blackwell's Approachability Theorem. Foster and Vohra [10, 11] used games with vector payoffs to show a process that converges to correlated equilibrium. DeMarzo, Kremer and Mansour [9] used regret minimization for option pricing and online algo trading. They showed that using a strategy minimizing his regret, the trader can derive a fairly accurate option price compared to the Black-Scholes model.

In recent years, Approachability Theory has also been used to find processes that converge to correlated equilibrium (see, for instance, Foster and Vohra [10] and Hart and Mas-Colell [15]).

[3] More precisely, information about at least n past outcomes until time $n+o(n)$ with a probability higher than $1-n^{-\alpha}$ (where $\alpha > 1$).

[4] That is, when information about time n arrives at time $n+cn$.

Rustichini [27], using Blackwell's Approachability Theorem, obtained a no-regret theorem when the DM has imperfect monitoring. These applications deal with a finite number of constraints. Lehrer [19] proved an Approachability Theorem when the payoffs are infinite-dimensional; this result has been employed by Lehrer [18] and by Sandroni, Smorodinsky, and Vohra [28] to construct manipulating strategies. Aumann and Maschler [1] utilized Blackwell's Theorem to show the existence of a value in infinitely repeated games with incomplete information. Lehrer and Solan [21] examined simple strategies and their power to guarantee approachability.

All these results were based on the assumption that the DM, the person who tries to approach a target set, is perfectly informed about the outcome of each stage and this information reached him immediately after the stage is over. In this paper, we consider a situation where the information the DM receives is delayed and involves a certain amount of noise: rather than getting instantaneous information about the actual outcome, the DM receives a delayed and noisy signal that might reveal only partial information about past outcomes. While the issue of noisy signal obtained with no delay was partially treated by Lehrer and Solan [22], the subject of no-regret with information lag (using the terminology of Levy [23]) was not discussed so far.

Time delay in dynamic games has been discussed by Shmaya [29] in the context of perfect information infinite zero-sum game. Shmaya analyzed the case where the information is delayed but eventually becomes perfect, i.e., when the DM becomes aware of the history with an arbitrarily high accuracy after a sufficiently long time. Levy [23] analyzed zero-sum stochastic games when one or both players (nature can be considered as the second player) observe past actions of their opponent with a time-dependent delay – information lag. Recently Fudenberg et al. [12] examined equilibrium in repeated games when players' signals about the actions of others arrive with a certain lag.

1.4 Structure of the paper

In Section 2 we present the model, definitions, and main results. Blackwell's Approachability Theorem along with problems that arise when information is delayed are discussed in Section 3. In Section 4 we prove the existence of a regret-free strategy under time delay and partial information. Our proof is based on two specific deterministic and stochastic signaling functions. Section 5 discusses rates of convergence and bounded-regret strategies. We conclude this paper by presenting an additional application of the time-delay regret-free strategy in models that include transactions costs (Section 6).

2 Regret-free strategies

2.1 The model

A sequential decision problem with one DM is a triplet (I,J,U), where I is the finite set of actions, J is the finite set of states of nature, and $U : I \times J \to \mathbb{R}$ is the payoff function. At every stage n, the DM chooses an action $i_n \in I$, simultaneously nature chooses a state $j_n \in J$, and the DM receives a payoff of $U(i_n, j_n) \in \mathbb{R}$.

A history of length n is a sequence of n actions and n states, $h_n = (i_1, j_1, i_2, j_2, ..., i_n, j_n)$. For every n, define H_n to be the set of all histories of length n and $H = \bigcup_n (I \times J)^n$ to be the set of all histories. For any $h_t, h_n \in H$ we say that $h_t \preceq h_n$ (resp. $h_t \prec h_n$) if h_t is a prefix of h_n (resp. $h_t \preceq h_n$ and $t < n$). Denote $\mathcal{H} = (I \times J)^\infty$, where each element $h_\infty \in \mathcal{H}$ will be referred to as a *play.* The n^{th} prefix of the play $h_\infty \in \mathcal{H}$ is denoted by $h_n \in H$.

For any play $h_\infty = (i_1, j_1, i_2, j_2, \dots)$, we denote by $\overline{U}_n(h_\infty) = \overline{U}_n(h_n) = \frac{1}{n} \sum_{t=1}^{n} U(i_t, j_t)$ the average payoff up to stage n. The range of U is extended to $\Delta(I) \times \Delta(J)$ in a multi-linear fashion. For every play h_∞ and every action $i \in I$, let $h_{n,i}$ be the history $(i, j_1, i, j_2, ..., i, j_n)$, where the DM plays constantly the action i ($i_t = i$ for every $t \in \{1, \dots, n\}$), while the states of nature remain unchanged (i.e., coincide with those in h_∞).

A sequential decision problem with time delay and imperfect monitoring is a sequential decision problem (I,J,U) with a set of signals $S = \bigcup_n S_n$ and two functions: the *delay function* $\phi : H \to \mathbb{N}$ and the *signaling function*[5] $\psi : H_n \to \Delta(S_n)$, where S_n is a finite set of signals relevant to histories of length n and $\Delta(S_n)$ is the set of distributions on S_n.

Let $h_n \in H$ be a history of length n. The DM receives at stage $\phi(h_n) \geq n$ a noisy signal $\psi(h_n)$ that stochastically depends on h_n. For every play $h_\infty \in \mathcal{H}$ and every n, denote $Q_n(h_\infty) = Q_n(h_n) = \times_{k;\ \phi(h_k) \leq n} S_k$. The set $Q_n(h_\infty)$ consists of all possible histories of signals that the DM might have received (through the signalling function) up to time n and after h_n. We assume that the DM recalls his past actions and signals. Thus, a strategy σ of the DM is a function $\sigma : \bigcup_n (I^n \times Q_n) \to \Delta(I)$. At any period the DM may condition his mixed action on his past actions as well as on the signals he already received.

A strategy of nature is a function $\tau : H \to \Delta(J)$. Any pair (σ, τ) of the DM's and nature's strategies induces a probability distribution $\Pr_{\sigma,\tau}$ on the set $\mathcal{H}$ of plays.

[5] For the sake of simplicity references to measurability requirements will be henceforth omitted.

2.2 Internal and external regret

Hannan defined a regret-free strategy by comparing the DM's actual payoff to his best hypothetical payoff had he played consistently the same action.

Definition 1. *Let $\varepsilon \geq 0$.*

(i) *A strategy σ is ε-regret-free if for every strategy τ of nature and every action $i \in I$,*

$$\Pr{}_{\sigma,\tau}\left(h_\infty;\ \liminf_{n\to\infty}\left(\overline{U}_n(h_n) - \overline{U}_n(h_{n,i})\right) \geq -\varepsilon\right) = 1.$$

(ii) *A strategy σ is regret-free if it is 0-regret free.*

In other words, assuming that the evolution of the states of nature is independent of the DM's actions, σ is regret-free if it guarantees him no less than what he could get had he played the best response to the empirical distribution of the states of nature. Hannan's no regret notion is usually referred to as *external no regret.*

Remark 1. *Since external regret is based on comparing the actual average payoff with the hypothetical average payoff had the* DM *played constantly some fixed action, we can consider an equivalent decision problem with vector payoffs. Each coordinate i of the vector payoff is the difference between the actual stage payoff and the payoff had the* DM *played action i repeatedly throughout the game. That is, the payoff at time n is $x(i_n, j_n) = (U(i_n, j_n) - U(i, j_n))_{i\in I}$. This is the "regret vector" of the* DM *for deciding to play action i_n instead of any other action i. The regret vector is what links between sequential decision problems and vector-payoff dynamic games.*

External regret is not the only way to evaluate one's strategy. Hart and Mas-Colell [15] and Fudenberg and Levine [13] introduced a stronger no-regret notion than external regret. This notion, commonly referred to as *internal regret*, compares the DM's actual average payoff with what he might have gained had he insistently replaced one action with another.

Formally, for every pair of actions $i, i' \in I$ and every play $h_\infty = (i_1, j_1, i_2, j_2, \dots)$, define the average payoff up to time n by

$$R_n(i, i', h_n) = \frac{1}{n}\sum_{t=1}^{n}\left(\mathbb{I}_{i_t = i} U(i', j_t) + \mathbb{I}_{i_t \neq i} U(i_t, j_t)\right),$$

where $h_n \prec h_\infty$. Given the play h_∞, $R_n(i, i', h_n)$ is the average payoff had the DM played the action i' instead of i every time the latter was played.

Definition 2. *Let $\varepsilon \geq 0$.*

(i) *A strategy σ is ε-internal regret free if for every strategy τ of nature and every two actions $i, i' \in I$,*

$$\Pr{}_{\sigma,\tau}\left(h_\infty;\ \liminf_{n\to\infty}(\overline{U}_n(h_n) - R_n(i,i',h_n)) \geq -\varepsilon\right) = 1.$$

(ii) *A strategy σ is internal regret free if it is* 0*-internal regret free.*

When σ is an internal-regret-free strategy, the DM cannot gain by replacing one action with another throughout the game.

2.3 The main result

The main result shows the balance between the uncertainty embedded in ψ and the delay expressed by ϕ that ensures the existence of a regret-free strategy. Plainly, if the delay is too large (meaning that ϕ increases rapidly) and, in addition, the signaling function ψ is not sufficiently informative, then a regret-free strategy does not exist.

Fix a play h_∞. By time n the DM received a sequence of signals $s \in Q_n(h_\infty)$. In addition he knows his own actions, $i_1, \ldots, i_n$. Had the DM known the strategy τ of nature, he could calculate the probability that $j_{n_1}, \ldots, j_{n_m}$ were the realized states at times $n_1, \ldots, n_m$, resp. This probability is denoted as $\Pr_{\sigma,\tau}(j_{n_1}, \ldots, j_{n_m}|i_1, \ldots, i_n, s)$.

Definition 3. *Let $p \in [0,1]$ and $m \leq n$ be natural numbers. A play h_∞ is* (m,n,r)-revealing *if for $h_n \prec h_\infty$, for every $s \in Q_n(h_\infty)$, and for every τ and σ, there are m periods $n_1, \ldots, n_m \leq n$ such that*

$$\Pr{}_{\sigma,\tau}(j_{n_1}, \ldots, j_{n_m}|i_1, \ldots, i_n, s) \geq r.$$

That is, a play h_∞ is (m,n,r)-revealing if no matter what τ is, for every sequence of signals $s \in Q_n(h_\infty)$, there are m periods $n_1, \ldots, n_m \leq n$ such that the DM can identify the true states, $j_{n_1}, \ldots, j_{n_m}$, with probability of at least r. For example, if $\psi(h_m) = (i_m, j_m)$ and $\phi(h_m) \leq m+n$ then every play h_∞ is $(m,n,1)$-revealing.

Definition 4. *The functions ψ, ϕ are* sufficiently informative *if for every σ and τ, there is $\alpha > 1$ such that the plays h_∞ that are $(n, n+o(n), 1-n^{-\alpha})$-revealing for sufficiently large n have probability 1.*

When the functions ψ, ϕ are sufficiently informative, then from some stage onwards there are at least n stages such that the DM is aware of the realized states of nature during these stages until time $n + o(n)$ with probability of at least $1 - n^{-\alpha}$.

Theorem 1. (Main Theorem) *If ψ and ϕ are sufficiently informative, then the* DM *has a regret-free strategy.*

Remark 2. *Regret-free theorems under imperfect monitoring usually compare the realized payoff to the minimal payoff possible that is still consistent with the signals the* DM *received (e.g., Rustichini [27]). In our model, however, since the* DM *is informed of the states of nature with high probability, we define the no-regret notion as it is defined in perfect monitoring models.*

The bounds stated in Theorem 1 are tight. The following example shows that in case the conditions do not hold, the DM does not necessarily have a regret-free strategy.

2.4 A long delay might prevent the existence of regret-free strategy

Example 1.

Consider the matching pennies game:

	H	T
H	–1	1
T	1	–1

where at each stage n, the DM and nature need to decide between 'Heads' and 'Tails', denoted by H and T, respectively. After each stage the DM receives the payoff indicated in the matrix above, depending on both the realized action and the state.

Suppose that the DM is informed of the realized state j_n only at time $2n$. In terms of the signalling function ψ and the delay function ϕ, this means that $\psi(h_n) = (i_n, j_n)$ and $\phi(h_n) = 2n$ for every $h_n = (i_1, j_1, ..., i_n, j_n)$. We show that the DM does not have a regret-free strategy.

Lemma 1. *The* DM *does not have a regret-free strategy.*

Proof. Consider the possibility that for every $n \geq 0$, the state of nature j_{2^n} is chosen randomly (with equal probabilities) and the following $2^n - 1$ states are identical to j_{2^n}. Since $\phi(h_n) = 2n$, during the periods between 2^n and $2^{n+1} - 1$ the DM is informed only of the states that had been realized prior to stage 2^n.

Fix a stage n, a history h_{2^n}, and a strategy σ of the DM. Denote the average number of times the DM played 'T' (resp., 'H') from stage 2^n until stage $2^{n+1} - 1$ by t_n (resp., b_n) and denote the constant state during these stages by G_n. Denote the average payoff of the DM from stage 2^n until stage $2^{n+1} - 1$ by $\widetilde{U}_{2^n}(h_{2^n})$. Then

$$\begin{aligned}\Pr\left(\exists i : \widetilde{U}_{2^n}(h_{2^n}) \leq 0,\ \widetilde{U}_{2^n}(h_{2^n,i}) = 1\right) &= \Pr\left(G_n = T\right) \cdot \Pr\left(b_n \leq \frac{1}{2},\ i = H \middle| G_n = T\right) \\ &+ \Pr\left(G_n = H\right) \cdot \Pr\left(t_n \leq \frac{1}{2},\ i = T \middle| G_n = H\right) \\ &= \frac{1}{2}\left[\Pr\left(b_n \leq \frac{1}{2}\right) + \Pr\left(t_n \leq \frac{1}{2}\right)\right] \geq \frac{1}{2},\end{aligned}$$

where the last equality is due to the fact that the DM is informed of G_n only after stage $2^{n+1}-1$. Since $\widetilde{U}_t(h_{t,i}) = 1$ for $t = 2^{n-1}, 2^{n-2}$ yields that $\overline{U}_{2^n}(h_{2^n,i}) \geq \frac{1}{2}$, one can show in a similar way that

$$\Pr\left(\exists i : \widetilde{U}_{2^n}(h_{2^n}) \leq 0,\ \widetilde{U}_{2^n}(h_{2^n,i}) = 1,\ \overline{U}_{2^n}(h_{2^n,i}) \geq \frac{1}{2}\right) \geq \frac{1}{8},$$

thus

$$\Pr\left(\exists i : \overline{U}_k(h_k) \leq \overline{U}_k(h_{k,i}) - \frac{1}{4}\right) \geq \Pr\left(\exists i : \overline{U}_k(h_k) \leq \frac{1}{2},\ \overline{U}_k(h_{k,i}) \geq \frac{3}{4}\right) \geq \frac{1}{8},$$

for $k = 2^{n+1}$. Note that the last inequality is independent of the history prior to stage 2^{n-2} and the result follows for the last inequality and the Borel-Cantelli Lemma. ■

3 Blackwell's Approachability Theory with delayed information

In this section we present Blackwell's Approachability Theorem and condition. We show that Blackwell's condition, although sufficient for approachability under perfect monitoring conditions, fails when information is delayed for too long.

3.1 The classical Approachability Theory

Let C be a convex and closed set in $\mathbb{R}^{|I|}$. We will refer to C as the *target set.* For every $x \in \mathbb{R}^{|I|}$ denote by $\Pi_C(x)$ the closest point to x in C and for every sequence $\{x_t\}_{t=1}^{\infty} \subset \mathbb{R}^{|I|}$ denote by $\overline{x}_n = \frac{1}{n}\sum_{t=1}^{n} x_t$ the arithmetic average of its first n elements.

Definition 5. *The sequence $\{x_t\}_{t=1}^{\infty}$* approaches *C if $\|\overline{x}_n - \Pi_c(\overline{x}_n)\|$ converges to zero.*

Recall that the vector payoff of every stage $t \in \mathbb{N}$ in the vector payoff decision problem was defined by $x(i_t, j_t) = (U(i_t, j_t) - U(i, j_t))_{i \in I} \in \mathbb{R}^{|I|}$.

Definition 6. *The set C is approachable by the DM if there exists a strategy σ such that for every strategy τ of nature, $\{x(i_t, j_t)\}_{t=1}^{\infty}$ approaches C almost surely (with respect to the distribution induced by σ and τ).*

The central theme of Approachability Theory is to find conditions under which C is approachable by the DM. The following Blackwell's condition guarantees approachability.

Definition 7. *The set C and the payoff function x satisfy Blackwell's condition if for every $y \in \mathbb{R}^{|I|}$ there is a mixed action $p \in \Delta(I)$ of the DM such that*

$$\langle y - \Pi_C(y), x(p,q) - \Pi_C(y) \rangle \leq 0, \tag{1}$$

for every $q \in \Delta(J)$.

In words, Blackwell condition states that for every vector y in $\mathbb{R}^{|I|}$, there is a mixed action p of the DM that guarantees that, against any mixed state q of nature, the expected payoff corrects the error of y relative to C in the sense that the expected payoff $x(p,q)$ and y sit on different sides of the hyperplane that supports C and is perpendicular to the line $y - \Pi_C(y)$.

The case where $\psi(h_n) = (i_n, j_n)$ and $\phi(h_n) = n$ is the case of *perfect monitoring.* At any stage n the DM is fully informed of the exact previously realized state, j_{n-1}. Blackwell's Approachability Theorem [2] states that under perfect monitoring, if C and x satisfy Blackwell's condition, then C is approachable by the DM. By setting the target set C to be the non-negative orthant, $C = \{x \in \mathbb{R}^{|I|} : x_i \geq 0,\ i = 1, \ldots, |I|\}$, any strategy that approaches C is a regret-free strategy in the original sequential decision problem.

Blackwell [2] provided a sufficient condition for a set to be approachable, and Hou [17] and Spinat [30] fully characterized the family of approachable sets.

3.2 Blackwell's condition does not guarantee approachability when information is delayed

Blackwell's condition does not guarantee approachability when the information received by the DM is delayed for too long. The following example illustrates this point.

Example 2.

Consider the following game:

	L	R
T	(–1,1)	(0,0)
B	(0,0)	(1,–1)

The DM has two actions, T and B, and nature has two states, L and R. Let $C = \{(y_1, y_2);\ y_1, y_2 \geq 0\}$ be the positive orthant of $\mathbb{R}^2$. A straightforward examination shows that x and C satisfy Blackwell's condition. Suppose that the time delay, signaling function, and strategy of nature τ are similar to the ones described in Example 1.

Lemma 2. *With probability 1, there are infinitely many n's such that the average payoff at time 2^n is bounded away from C.*

Proof. For simplicity, let us consider a reduction of the game mentioned above to the 1-dimensional case.

	L	R
T	–1	0
B	0	1

Both the DM and nature maintain the same actions and states defined in the 2-dimensional game, however C will be the singleton $C = \{0\}$. This reduction is obvious due to the fact that the payoffs (in the vector payoff game mentioned above) lie symmetrically with respect to the point $(0, 0)$ along the line $\{(y, -y); -1 \leq y \leq 1\}$ and the possibilities of both the DM and nature preserve their relative features.

Denote by $(Y_n, Z_n) \in \Delta(I) \times J$ the average actions of the DM and state of nature between periods 2^n and $2^{n+1} - 1$. That is, $Y_n = (t_n, b_n)$ and $Z_n \in \{L, R\}$ when Y_n is independent of Z_n. Fix $\epsilon > 0$ and for every n define the event $D_n = \{\|x(Y_n, Z_n)\| < \epsilon\}$. Let $B_m = \bigcap_{n \geq m}^{\infty} D_n$. B_m is an increasing sequence of events. It is sufficient to show that[6] $\Pr(D_n^c \; i.o.) = 1$ or, equivalently, $\Pr(\bigcup_{m=1}^{\infty} B_m) = 0$.

Assume that C is approachable by the DM and $\Pr\left(\bigcup_{m=1}^{\infty} B_m\right) = a > 0$. Note that $\Pr\left(\bigcup_{m=1}^{\infty} B_m\right) = \lim\limits_{m \to \infty} \Pr(B_m)$, therefore there exists $N_0 \in \mathbb{N}$ such that $\Pr(B_m) \geq \frac{a}{2}$ for all $m \geq N_0$.

Note that

$$\begin{aligned} \Pr(D_n) &= \frac{1}{2}\left[\Pr(t_n < \epsilon | Z_n = L) + \Pr(b_n < \epsilon | Z_n = R)\right] \\ &= \frac{1}{2}\left[\Pr(t_n < \epsilon | Z_n = L) + 1 - \Pr(1 - t_n \geq \epsilon | Z_n = R)\right] \\ &= \frac{1}{2}\left[\Pr(t_n < \epsilon | Z_n = L) + 1 - \Pr(t_n \leq 1 - \epsilon | Z_n = R)\right] \leq \frac{1}{2}. \end{aligned}$$

Furthermore,

$$\begin{aligned} \Pr(D_n \cap D_{n+1}) &\leq \frac{1}{2}\left[\Pr(D_{n+1} | D_n)\right] \\ &\leq \frac{1}{2^2}\left[\Pr(t_{n+1} < \epsilon | D_n, Z_{n+1} = L) + \Pr(b_{n+1} < \epsilon | D_n, Z_{n+1} = R)\right] \leq \frac{1}{2^2}. \end{aligned}$$

Continuing inductively we obtain that $\Pr\left(\bigcap_{n \geq m}^{m+k} D_n\right) \leq 1/2^k$, which suggests that for large enough $k \in \mathbb{N}$ and $m \geq N_0$,

$$\Pr(B_m) \leq \Pr\left(\bigcap_{n \geq m}^{m+k} D_n\right) \leq \frac{1}{2^k} < \frac{a}{2},$$

[6] We denote the term "infinitely often" by *i.o.*

in contradiction with $\Pr(B_m) \geq \frac{a}{2} > 0 \; \forall m \geq N_0$. ■

Note that without any time delay the DM can guarantee that the average payoff will converge to C.

4 Regret-free strategies with time delay

In order to prove Theorem 1, we need two preliminary propositions, which are based on specific deterministic and stochastic signaling functions.

Since the validity of the Approachability Theorem under general signaling functions yields the existence of a regret-free strategy, the proofs of the propositions and theorems will be given in terms of approachability instead of no-regret.

4.1 Deterministic information with time delay

In the model of imperfect monitoring (see, e.g., Rustichini [27], Mannor and Shimkin [25], Cesa-Bianchi, Lugosi, and Stoltz [5], Lugosi, Mannor and Stoltz [24], Blum and Mansour [3], Lehrer and Solan [21] and Perchet [32]) the signal that the DM receives does not depend on the entire history h_n. Rather, $\psi(h_n)$ is Markovian and depends only on the last state and action played, that is, $\psi(h_n) = \psi(i_n, j_n)$. In this paper, the signal that the DM receives may depend, and it typically does, on the entire history h_n.

Recall that after time $\phi(h_n)$ the DM receives a signal $\psi(h_n)$. We say that the signal $\psi(h_n)$ *depends deterministically on the history* h_n if, for every n, $\psi(h_n) = \psi(h'_n)$ implies that $h_n = h'_n$. In other words, without loss of generality, $\psi(h_n) = h_n = (i_1, j_1, i_2, j_2, ..., i_n, j_n)$. In Proposition 1, we assume that $\psi(h_n)$ depends deterministically on the history h_n and show that when $\phi(h_n)$ is bounded by $n + o(n)$, a regret-free strategy exists.

Proposition 1. *If $\phi(h_n) = n + o(n)$ and $\psi(h_n)$ depends deterministically on h_n, the* DM *has a regret-free strategy.*

In order to prove Proposition 1, we first need to present our generalization of the Approachability Theorem and more specifically, the generalization of the geometric principle behind approachability.

Approachability's Geometric Principle. The original Blackwell condition is based on a rather simple geometric principle: Let $z_1, z_2, \ldots$ be a bounded sequence of points in $\mathbb{R}^{|I|}$. If for every k

$$\langle z_{k+1} - \Pi_C(\overline{z}_k), \overline{z}_k - \Pi_C(\overline{z}_k) \rangle \leq 0, \tag{2}$$

then the sequence $\{\overline{z}_k\}_{k=1}^{\infty}$ approaches C. That is, $\|\overline{z}_k - \Pi_C(\overline{z}_k)\| \to 0$ as $k \to \infty$.

Lemmas 3 and 4 extend this principle to contexts that are relevant to delayed information. Their proofs are postponed to the Appendix.

Lemma 3. [First extension of approachability′s geometric principle] *Let $z_1, z_2, \ldots$ be a bounded sequence of points in $\mathbb{R}^{|I|}$ and let $\alpha_1, \alpha_2, \ldots$ be a sequence of non-negative real numbers, with $\alpha_1 > 0$. Denote $A_k = \sum_{l=1}^{k} \alpha_l$ and let $\overline{z}_k = \frac{1}{A_k}\sum_{l=1}^{k} \alpha_l z_l$ be the weighted average of $z_1, z_2, \ldots, z_k$. Suppose that $\frac{\alpha_k}{A_k} \to 0$ and $A_k \to \infty$ as $k \to \infty$. If for every k*

$$\langle z_{k+1} - \Pi_C(\overline{z}_k), \overline{z}_k - \Pi_C(\overline{z}_k)\rangle \leq 0, \tag{3}$$

then $\{z_k\}_{k=1}^{\infty}$ approaches C.

Lemma 4. [Second extension of approachability′s geometric principle] *Let $z_1, z_2, \ldots$ be a bounded sequence of points in $\mathbb{R}^{|I|}$ and let $\alpha_1, \alpha_2, \ldots$ be a sequence of non-negative real numbers, with $\alpha_1 > 0$. Denote $A_k = \sum_{l=1}^{k} \alpha_l$ and let $\overline{z}_k = \frac{1}{A_k}\sum_{l=1}^{k} \alpha_l z_l$ be the weighted average of $z_1, z_2, \ldots, z_k$. Suppose that $\frac{\alpha_k}{A_k} \to 0$ and $A_k \to \infty$ as $k \to \infty$. If for every k*

$$\langle z_{k+1} - \Pi_C(\overline{z}_{k-1}), \overline{z}_{k-1} - \Pi_C(\overline{z}_{k-1})\rangle \leq 0, \tag{4}$$

then $\{z_k\}_{k=1}^{\infty}$ approaches C.

The difference between Lemmas 3 and 4 is that in (3) the inner product involves $\overline{z}_k$ while in (4), it involves $\overline{z}_{k-1}$.

The generalization presented in Lemmas 3 and 4 has a simple interpretation. The DM receives new information about past states with a certain delay. The time that passes between two periods where the DM receives new information is the reason for introducing the weights $(\alpha_1, \alpha_2, \ldots)$. The weight α_k is the number stages from the time the DM received the k^{th} signal until he received the $(k+1)^{\text{st}}$ signal. After receiving the new information, the DM can play an action that corrects the error of the average payoff, relative to C (in the sense of Ineq. (4)), until he obtains another piece of information.

Proof of Proposition 1. Define the strategy σ of the DM as follows. Fix a history $h_n \in H$. After history h_n, the DM is familiar with every average payoff, $\overline{x}_t$, such that $h_t \prec h_n$ and $\phi(h_t) \leq n$. Define $t(h_n)$ to be largest t that satisfies $h_t \prec h_n$ and $\phi(h_t) \leq n$. The strategy σ prescribes to play independently the same mixed action at all histories h_m that are strict continuations of h_n (i.e., $h_n \prec h_m$) and satisfy $t(h_n) = t(h_m)$. That is, the DM plays the same mixed action until he receives new information. The mixed action $p^* \in \Delta(I)$ to be played is the one that satisfies

$$\left\langle \overline{x}_{t(h_n)} - \Pi_C(\overline{x}_{t(h_n)}), x(p^*, q) - \Pi_C(\overline{x}_{t(h_n)}) \right\rangle \leq 0, \tag{5}$$

for every mixed state q of nature.

In the first stages (until time $\phi(h_1) + 1$) the DM plays arbitrarily. Since $\phi(h_n)$ is finite for every n and history h_n, σ is well defined.

Fix an infinite sequence of states $\tau = \{j_l\}_{l=1}^{\infty}$. For every play h_∞, let α_k denote the number of stages from the k^{th} signal till the $(k+1)^{\text{st}}$ signal with respect to h_∞ and ϕ. Let $A_k = \sum_{l=1}^{k} \alpha_l$ and assume that the k^{th} signal reached the DM at time n. Since $\phi(h_n) = n + o(n)$, it follows that

$$\alpha_k \leq \phi(h_n) - n = o(n). \tag{6}$$

Note that α_k and A_k are random variables that depend on the play and, by definition,

$$A_k = A_{k-1} + \alpha_k = n + \alpha_k.$$

Hence, $A_k \to \infty$ and $\frac{\alpha_k}{A_k} \to 0$ as $n \to \infty$ (i.e., for every h_∞, see Lemma 8 in the Appendix). In addition, $A_k \to \infty$ if and only if $n \to \infty$.

Consider the rounds played from the k^{th} signal till the $(k+1)^{\text{st}}$ signal as a single round game, denoted by G_k, in which the payoff is the average payoff received during these α_k periods, which we denote by z_k. The action played during these stages, namely in G_k, is the mixed action p that satisfies Eq. (5). This action was played independently.

Note that at stage A_k the average payoff, which was denoted by $\overline{x}_{A_k}$, equals $\frac{1}{A_k}\sum_{l=1}^{k} \alpha_l z_l$. The independence of the action played in every G_k according to the strategy σ (satisfying Eq. (5)) and the properties of α_k and A_k ensure that the conditions for Lemma 4 are satisfied, therefore the set C is approachable by the DM, in the sense that $\|\overline{x}_m - \Pi_C(\overline{x}_m)\|$ convergence to 0. ■

The next proposition weakens the conditions of Theorem 1 by allowing general delay and signaling functions during a finite number of stages.

Proposition 2. *Fix $N_0 \in \mathbb{N}$. If $\phi(h_n) = n + o(n)$ and $\psi(h_n)$ depends deterministically on h_n for all $n \geq N_0$, then the DM has a regret-free strategy.*

Proof. Since

$$\begin{aligned}
\lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} x(i_t, j_t) &= \lim_{n\to\infty} \frac{1}{n} \left[\sum_{t=1}^{N_0} x(i_t, j_t) + \sum_{t=N_0+1}^{n} x(i_t, j_t) \right] \\
&= \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{N_0} x(i_t, j_t) + \lim_{n\to\infty} \frac{1}{n} \sum_{t=N_0+1}^{\infty} x(i_t, j_t),
\end{aligned}$$

it follows that $\lim\limits_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} x(i_t, j_t) = \lim\limits_{n\to\infty} \frac{1}{n}\sum_{t=N_0+1}^{\infty} x(i_t, j_t)$. Thus, every strategy that follows the strategy σ specified in the proof of Proposition 1 only from stage N_0 onwards, is regret-free. ■

4.2 Stochastic information with time delay

This subsection deals with a specific stochastic signaling function. Since this signaling function does not depend deterministically on histories h_n, we will introduce a new regret-free strategy which will be later used in the proof of Theorem 1.

The values of the new signaling function ψ are stochastic and defined as follows. For every possible state of nature there is a specific distribution on the set $\{0, 1\}$. At time $\phi(h_n)$, the DM receives a signal consisting of n bits. Each 0 or 1 bit is chosen with respect to the identifying distribution of the relevant past state of nature. The DM needs to collect enough signals in order to gain a good understanding of past states, but he will never have absolute certainty about them. At each stage there is a positive probability that the DM's understanding of past occurrences is wrong.

Formally, for every $z \in [0, 1]$ denote by $B(z)$ the Bernoulli distribution with parameter z: if $X \sim B(z)$, then $\Pr(X = 1) = z = 1 - \Pr(x = 0)$. Let $\eta : J \to P$ be a bijection from J to $P = \{p_i \ ; \ i = 1, \dots, m, \ 0 \leq p_1 < \dots < p_m \leq 1\}$. The signal $\psi(h_n)$ the DM receives about history h_n is a word of n bits,

$$\psi(h_n) = (X_n^1, \dots, X_n^n),$$

where $X_n^1, \dots, X_n^n$ are independent of each other, independent of $\{X_k^l\}_{k<n, 1\leq l\leq k}$, and $X_n^t \sim B(\eta(j_t))$ for every $n \in \mathbb{N}$. That is, the bit X_n^t which was received at time $\phi(h_n)$ (along with $n-1$ other bits) and relates to the realized state j_t is chosen according to the identifying distribution $B(\eta(j_t))$.

Since the parameters of the Bernoulli distributions for the different states of nature are different from each other, this signaling function gradually enables the DM to have a good assessment of past states. We will now find the number of stages T_n that the DM needs in order to identify n states of nature with high probability.

For every t, denote the moving average of the X_n^t's by

$$F(t, T) = \frac{1}{T - t} \sum_{n=t+1}^{T} X_n^t.$$

Intuitively, one can think of $F(t, T)$ as the average of the signals from time $t + 1$ until time T concerning the state of nature j_t. Fix $0 < \varepsilon < \frac{1}{4} \min\limits_{j,j'\in J} |p_j - p_{j'}|$ and define the event $C_{t,T}$ by

$C_{t,T} = \{|F(t,T) - \eta(j_t)| \leq \varepsilon\}$. Since X_n^t are i.i.d. with distribution $B(\eta(j_t))$, by the strong law of large numbers, $F(t,T)$ converges to $\eta(j_t)$ with probability 1.

We say that the DM *can identify* the states of nature $j_1, \ldots, j_n$ with probability higher than p at time T if $\Pr\left(\bigcap_{t=1}^{n} C_{t,T}\right) \geq p$. Therefore, the average payoff $\overline{x}_n$ is identified with probability higher than p at time T if $j_1, \ldots, j_n$ can be identified with probability higher than p at time T and the average payoff, according to $(i_1, j_1, \ldots, i_n, j_n)$, is $\overline{x}_n$. Obviously, as the game progresses the DM receives more and more signals about j_t and, as Proposition 3 shows, the DM has a regret-free strategy.

Lemma 5. *For every decision problem there is a fixed $\delta > 0$ such that for every sufficiently large n and history h_n, the* DM *can identify with probability higher than $1 - \frac{1}{n^2}$ at time $T_n = n + \lceil \frac{4}{\delta} \ln(n) \rceil$ the states $j_1, \ldots, j_n$.*

It follows from Lemma 5 that for every n and history h_n, the time until the DM can identify h_n with probability higher than $1 - \frac{1}{n^2}$ is bounded from above by $n + o(n)$. The proof of Lemma 5 is given in the Appendix.

Proposition 3. *If for every n the information about histories h_n is obtained through the signaling function ψ defined in Subsection 4.2 and $\phi(h_n) = n + o(n)$, then the* DM *has a regret-free strategy.*

Proof. Let $N_0 \gg 1$ be a positive natural number such that the approximation in Eq. (17) holds for all $n \geq N_0$ (meaning that for every $n \geq N_0$ the DM can identify with probability higher than $1 - \frac{1}{n^2}$ at time $T_n = n + \lceil \frac{4}{\delta} \ln(n) \rceil$ the realized states $j_1, \ldots, j_n$ of nature).

Define the strategy σ of the DM as follows. Until stage T_{N_0}, the strategy σ prescribes playing arbitrarily. For every $n \geq N_0$, the strategy prescribes playing the same mixed action repeatedly from stage T_n until stage T_{n+1}. The mixed action $p^* \in \Delta(I)$ to be played is the one that satisfies

$$\langle \overline{x}_n - \Pi_C(\overline{x}_n), x(p^*, q) - \Pi_C(\overline{x}_n) \rangle \leq 0 \quad \forall q \in \Delta(J).$$

It follows from Lemma 5 that for every $n \geq N_0$ the average payoff $\overline{x}_n$ is identified with probability higher than $1 - \frac{1}{n^2}$ from stage T_n. That is, the DM plays the mixed action $p^* \in \Delta(I)$ according to $\overline{x}_n$ only from the stage he can identify $j_1, \ldots, j_n$ with probability higher than $1 - \frac{1}{n^2}$.

Let V_n be the event that the DM played the mixed action p^* according to $\overline{x}_n$ when $\overline{x}_n$ is not the actual average payoff until time n. This means that for some $t < n$, $|F(t, T_n) - \eta(j_t)| > \varepsilon$. For every $n \geq N_0$, $\Pr(V_n) = \Pr\left(\bigcup_{t=1}^{n} \{C_{t,T_n}\}^c\right) \leq \frac{1}{n^2}$, which implies that $\sum_{n=1}^{\infty} \Pr[V_n] \leq \sum_{n=1}^{N_0} \Pr[V_n] + \sum_{n=N_0+1}^{\infty} \frac{1}{n^2} < \infty$. By the Borrel-Cantelli Lemma, $\Pr[V_n \;\; i.o.] = 0$. It follows that with probability 1 there is $M_0 \in \mathbb{N}$ such that, for all $n \geq M_0$, the DM played the mixed action p^* according to $\overline{x}_n$ when $\overline{x}_n$ is the actual realized average payoff. Since the time delay, including T_n, is bounded by $n + o(n)$, the conditions of Proposition 2 hold and the result follows. ■

4.3 Proof of Theorem 1

In this section we prove Theorem 1. This theorem generalizes both the deterministic and the stochastic cases since it does not depend on a specific signaling function.

Since the functions ϕ and ψ are sufficiently informative, there is an $\alpha > 1$ such that the plays h_∞ that are $\big(n, n+o(n), 1-\frac{1}{n^\alpha}\big)$-revealing for sufficiently large n have probability 1. Thus, for every τ, σ, and $s \in Q_{n+o(n)}(h_\infty)$ there are n stages $t_1, \ldots, t_n$ such that

$$\Pr{}_{\sigma,\tau}(j_{t_1}, \ldots, j_{t_n} | i_1, \ldots, i_{n+o(n)}, s) \geq r.$$

That is, the DM can identify with probability higher than $1 - \frac{1}{n^\alpha}$ at least n previous states of nature by time $n + o(n)$, for some $\alpha > 1$ and every n large enough. Let the DM use the same strategy previously defined in the proof of Proposition 3 and let W_n be the event that the DM played the mixed action p^* according to $\overline{x}_n$ when $\overline{x}_n$ is not the actual average payoff (with a slight abuse of notation, we now use $\overline{x}_n$ to denote the average payoff of the states $t_1, \ldots, t_n$).

By the Borrel-Cantelli Lemma, $\Pr[W_n \quad i.o.] = 0$. Therefore, almost surely there is some $N_1 \in \mathbb{N}$ such that for all $n \geq N_1$, the DM plays the mixed action p^* when $\overline{x}_n$ is the actual average payoff and, by the bound on ϕ, the result follows from Proposition 2.

Note that the DM does not necessarily know h_n, but some n previous plays. However, since the strategy is based only on knowing only the average outcome of previous plays, the result still holds.

5 Bounded regret and rates of convergence

In the previous sections we provided conditions that ensure the existence of a regret-free strategy in the presence of delay. In this section we focus on cases where regret-free strategies need not exist and on the rates of convergence of regret-free strategies to the target set in the vector payoff game[7].

5.1 Bounded regret

We will now show that our previous extensions of Blackwell's Approachability Theorem in Lemmas 3 and 4 are useful in situations with time delay larger than $o(n)$. Example 1 shows that a regret-free strategy need not exist once the time delay is linear. However, bounded-regret strategies – strategies which are ε-regret free – might still exist.

The following theorem shows that the strategy presented in Proposition 1 is still useful in cases where the time delay is linear.

[7] We will sometimes refer to this as *the rates of convergence of regret-free strategies.*

Theorem 2. *If $\beta > 0$ is such that for every play h_∞ and $n \in \mathbb{N}$, $\phi(h_n) \leq n + \beta n$, then the* DM *has a $2M_G\sqrt{7\beta}$-regret free strategy.*

This theorem shows that, given a linear delay function with a slope of β which is small relative to $\frac{1}{(M_G)^2}$, the DM still possesses a small regret-free strategy.

Proof. It follows from the proof of Proposition 1 and Lemmas 3 and 4 that a strategy is ε-regret free if the distance between $\overline{z}_k$ and C is asymptotically bounded by ε.

Eq. (16) and Lemma 7 in the Appendix show that

$$\|\overline{z}_k - \Pi_C(\overline{z}_k)\|^2 \leq 7M^2 \max\left\{\sum_{l=2}^{k} \frac{\alpha_l \alpha_{l-1}}{A_k^2}, \sum_{l=1}^{k} \frac{\alpha_l^2}{A_k^2}\right\} \leq 28M_G^2 \frac{\alpha_{l_k}}{A_{l_k}},$$

where $\alpha_{l_k} = \max\limits_{l=1,\ldots,k} \alpha_l$ and $M \leq 2M_G$.

Since $\phi(h_n) \leq n + \beta n$, the term $\frac{\alpha_{l_k}}{A_{l_k}}$ is bounded from above by β (as $\frac{\alpha_k}{A_k} \leq \frac{n\beta}{n} = \beta$) and the result follows. ■

5.2 Rates of convergence

The original Blackwell condition with perfect monitoring guarantees the DM a rate of convergence of $\frac{1}{\sqrt{n}}$. A natural question is how time delay changes the rate of convergence to the target set, or, equivalently, the rate of convergence of the regret-free strategy. Knowing the differences between the rates of convergence gives a good assessment of the importance of constant monitoring and strategy updating. In cases of limited time to react and other constraints on resources, a DM might prefer not to update his strategy on each round, although he can, if the rate of convergence is sufficient for his needs.

The rate of convergence under time delay is mainly affected by the fact that the DM does not update his action unless he receives new information. Recall that after receiving the k^{th} signal, the DM plays the same action for α_k stages (using the notation of Proposition 1) and this action will be updated only after receiving the $(k+1)^{\text{st}}$ signal.

Theorem 3. *If for every play h_∞ and $n \in \mathbb{N}$, $\phi(h_n) \leq n + o(n)$ and $\psi(h_n)$ depends deterministically*[8] *on h_n, then the rate of convergence of the regret-free strategy is $o(1)$.*[9]

Proof. We have to consider two cases: the α_k's are bounded by some constant $\theta \in \mathbb{R}$ and the α_k's are unbounded.

[8] This condition concerns the deterministic signaling function. The stochastic function requires the condition of Proposition 3.

[9] $o(1)$ is a function that converges to 0 as n goes to infinity.

Using Ineq. (18) from Lemma 7 (in the Appendix), we obtain

$$\frac{1}{A_k^2}\sum_{l=1}^{k}\alpha_l^2 \leq \frac{\alpha_{l_k}}{A_k}, \tag{7}$$

where $\alpha_{l_k} = \max\limits_{l=1,\ldots,k} \alpha_l$. Since α_k is bounded by some large enough θ for every k, the bound on the left-hand side of Ineq. (7) becomes

$$\frac{1}{A_k^2}\sum_{l=1}^{k}\alpha_l^2 \leq \frac{\theta}{A_k} \leq \frac{\theta}{n}.$$

The same holds for the second sum in Ineq. (16), proving that in case of bounded weights the rate of convergence is in fact $n^{-\frac{1}{2}}$, as in the original Approachability Theorem.

On the other hand, the rate of convergence in the case the α_k's are unbounded might be weaker, up to $o(1)$, depending on how the weights behave asymptotically. Ineq. (7) still holds under the current conditions, however α_k is unbounded, therefore α_{l_k} goes to infinity as $n \to \infty$.

From Ineq. (6), the upper and lower bounds on α_{l_k} are $0 \leq \alpha_{l_k} \leq o(n)$. Thus,

$$\frac{1}{A_k^2}\sum_{l=1}^{k}\alpha_l^2 \leq \frac{\alpha_{l_k}}{A_k} \leq \frac{\alpha_{l_k}}{n+\alpha_k} \leq o(1),$$

where the last passages follow also from Ineq. (19) and Eq. (20) in the Appendix, and the statement holds. ■

6 Application: Regret-free strategy with transaction costs

The regret-free strategy with time delay is applicable in scenarios other than time delay, where one of which is the case of trading with transaction costs.

The problem of sequential portfolio selection has been broadly discussed by Cover [6, 7], Cover and Ordentlich [8] and Helmbold, Schapire, Singer and Warmuth [16]. Stoltz and Gabor [31] combined the concept of regret minimization with on-line portfolio selection, but without transaction cost. Blum and Kalai [4] examined Cover's universal algorithm with transaction costs and showed that an updated algorithm can preform almost as well as the constant re-balanced portfolio.

However, a regret-free portfolio selection strategy with transaction cost is yet to be found. This section shows the application of our regret-free strategy with time delay when considering transaction costs.

Consider a model with no time delay where the DM is a trader working with an online learning algorithm. At every stage n the trader chooses a mixed action, $p_n \in \Delta(I)$, which indicates the structure of the portfolio of the trader at time n. Simultaneously, the market moves to a state $q_n \in \Delta(J)$ and the trader receives a stage payoff of $U(p_n, q_n) - c\|p_n - p_{n-1}\|$ for some fixed positive real number c that represents the transaction cost. The payoff at each stage depends on the portfolio of the trader, the market's behavior, and the transaction cost c.

On the one hand, the trader wishes to update his portfolio in order to adjust to the market's behavior. On the other hand, frequent updating might be too costly. It turns out that regret-free strategies with time delay might be of help in this problem.

A regret-free strategy with transaction cost. For every strategy σ of the trader, define the function $l_\sigma : H \to \mathbb{N}$ such that $l_\sigma(h_n)$ is the last stage the trader updated his portfolio with respect to h_n, time n, and σ. For example, if $l_\sigma(h_n) = \frac{n}{2}$ then from stage $\frac{n}{2}$ until stage n the trader constantly chose the same mixed action $p \in \Delta(I)$ according to σ and therefore the transaction cost during these stages was 0.

Fix $0 < \beta < 1$ and $N \in \mathbb{N}$. The regret-free strategy σ_1 of the trader allows him to choose a portfolio arbitrarily until stage N. By definition, $l_{\sigma_1}(h_N)$ is the last stage that the trader updated his portfolio. The strategy prescribes that the trader next updates the portfolio, according to the condition of Eq. (5), only on stage $l_{\sigma_1}(h_N) + l_{\sigma_1}(h_N)^\beta$. In other words, the trader needs to wait from stage $l_{\sigma_1}(h_N)$ another $l_{\sigma_1}(h_N)^\beta$ stages until he can update his portfolio, where $l_{\sigma_1}(h_N)$ is the time of the last update.

Continuing inductively, for every $n > N$ the strategy σ requires the trader to update his portfolio only on stage $l_{\sigma_1}(h_n) + l_{\sigma_1}(h_n)^\beta$. Note that as long as the trader does not update his portfolio from stage n until stage $n + k$ for some $k \in \mathbb{N}$, it holds that $l_{\sigma_1}(h_n) = l_{\sigma_1}(h_{n+m})$ for every $m \leq k$.

This decision to update the portfolio on specific delayed stages relates to the time-delay model, since the mixed actions (according to our proposed strategy in Proposition 1) are updated only after a new signal reaches the DM. It is easy to verify that Proposition 1 and Proposition 2 guarantee that the strategy σ_1 is in fact regret free and the average transaction cost diminishes to 0. That is, the strategy σ_1 guarantees the trader no less than what he could get had he repeatedly chosen the best response portfolio, with respect to the empirical distribution of the states of the market.

However, this strategy can be very costly in terms of rate of convergence. Therefore, we will now prove that an even better regret-free strategy exists.

A regret-free strategy with transaction cost and improved rate of convergence. The improved strategy σ_2 is defined as follows. Let the first update of the portfolio be after the first stage. For every $k \in \mathbb{N}$ the strategy σ_2 prescribes that the k^{th} update of the portfolio, according to the condition of Eq. (5), will be after exactly k rounds from the last update.

Formally, fix a history h_n. The last update of the portfolio of the trader occurred on stage $l_{\sigma_2}(h_n)$. Assume that until stage $l_{\sigma_2}(h_n)$ (not including) the portfolio was updated $k-1$ times for some $k \in \mathbb{N}$. The strategy of the trader states that the next update (the $(k+1)^{\text{st}}$ update), according to the condition of Eq. (5), will be on stage $l_\sigma(h_n) + k + 1$.

Lemma 6. *The strategy σ_2 is regret free with a rate of convergence of $o(n^{-\frac{1}{2}})$.*

Proof. The trader's strategy requires that he delay the updating of the portfolio although he constantly receives new information. The delay defined in strategy σ_2 satisfies the same condition stated in Proposition 1 for the delay function in the time-delay model. Hence, it follows from Proposition 1 and Proposition 2 that σ_2 is regret free.

Fix $k \in \mathbb{N}$. Recall that the α_k's in Proposition 1 denoted the number of stages from the $(k-1)^{\text{st}}$ until the k^{th} signal in the time-delay model. In the transaction cost model α_k translates to the number of stages from the $(k-1)^{\text{st}}$ until the k^{th} update of the trader's portfolio.

Assume that the $(k-1)^{\text{st}}$ update of the portfolio occurred on stage n. According to the strategy σ_2, $\alpha_k = k$ and the next update should be on stage $n+k$. Note that

$$A_k = \sum_{l=1}^{k} \alpha_k = n + \alpha_k = n + k = \sum_{l=1}^{k} l = \frac{k(k+1)}{2},$$

and therefore

$$n = \frac{k(k+1)}{2} - k = \frac{k(k-1)}{2} \leq \frac{(k+1)^2}{2}.$$

Consequently,

$$\frac{\alpha_k}{A_k} = \frac{2k}{k(k+1)} \leq \sqrt{\frac{2}{n}}. \tag{8}$$

Also, note that the average transaction cost until time $n+k$ is $Ck/(n+k)$. As $n = k(k-1)/2 \geq (k-1)^2/2$,

$$\frac{Ck}{n+k} \leq \frac{Ck}{n} \leq \frac{C(\sqrt{2n}+1)}{n}. \tag{9}$$

Since the rate of convergence of the regret-free strategy depends on the average tranaction cost and on the convergence of α_k/A_k to 0 (see Lemma 7 in the Appendix), the result follows from Eqs. (8) and (9). ■

Remark 3. *Our proposed model and strategies differ from previous work in several ways. The natural portfolio selection model considers a list of assets. The utility of the* DM *is the value of his portfolio relative to the closing prices of the assets and his initial investment, and not on a stage payoff. In our model the trader can stay with the same action throughout the stages without* *paying transaction fees, whereas the portfolio selection model requires the trader to sell and buy assets in order to re-balance his portfolio at every stage. Another difference is that our our cumulative payoff is additive and not multiplicative. A full examination of a regret-free portfolio selection with transaction cost is left for future research.*

References

[1] Aumann, R.J. and M. Maschler (1995), *Repeated Games of Incomplete Information*, MIT Press, Cambridge, MA.

[2] Blackwell, D. (1956), *An Analog of the MinMax Theorem for Vector Payoffs*, Pacific Journal of Mathematics, 6, 1–8.

[3] Blum, A. and Y. Mansour (2007), *From External to Internal Regret*, Journal of Machine Learning Research, 8, 1307–1324.

[4] Blum, A. and A. Kalai (1999), *Universal Portfolios with and without Transaction Costs*, Machine Learning, 35, 193–205.

[5] Cesa-Bianchi, N., G. Lugosi, and G. Stoltz (2006), *Regret Minimization under Partial Monitoring*, Mathematics of Operations Research, 31, 562–580.

[6] Cover, T.M. (1991), *Universal portfolios*, Mathematical Finance, 1, 1–29.

[7] Cover, T.M. (1996), *Universal Data Compression and Portfolio Selection*, Proceedings of the 37th IEEE Symposium on Foundations of Computer Science, 534–538.

[8] Cover, T.M. and E. Ordentlich (1996), *Universal Portfolio with Side Information*, IEEE Transactions on Information Theory, 42, 348–363.

[9] DeMarzo, P., I. Kremer, and Y. Mansour (2006), *On-Line Trading Algorithms and Robust Option Pricing*, Proceedings of the 38th Annual ACM Symposium on Theory of Computing, 477-486.

[10] Foster, D. and R. Vohra (1997), *Calibrated Learning and Correlated Equilibrium*, Games and Economic Behavior, 21, 1/2 40–55.

[11] Foster, D. and R.V. Vohra (1999), *Regret in the On-Line Decision Problem*, Games and Economic Behavior, 29, 7–35.

[12] Fudenberg, D., Y. Ishii, and S.D. Kominers (2012), *Delayed-Response Strategies in Repeated Games with Observation Lags*, mimeo.

[13] Fudenberg, D. and D. Levine (1999), *Conditional Universal Consistency*, Games and Economic Behavior, 29, 104–130.

[14] Hannan, J. (1957), *Approximation to Bayes Risk in Repeated Plays.* In M. Dresher, A.W. Tucker, and P. Wolfe, editors, Contributions to the Theory of Games, 3, 97–139, Princeton University Press.

[15] Hart, S. and A. Mas-Colell (2000), *A Simple Adaptive Procedure Leading to Correlated Equilibrium*, Econometrica, 68, 1127–1150.

[16] Helmbold, D.P., R.E. Schapire, Y. Singer and M.K. Warmuth (1998), *On-Line Portfolio Selection using Multiplicative Updates*, Mathematical Finance, 8, 325–344.

[17] Hou T.F. (1971), *Approachability in a Two-Person Game*, The Annals of Mathematical Statistics, 42, 735–744.

[18] Lehrer, E. (2001), *Any Inspection is Manipulable*, Econometrica, 69, 5, 1333–1347.

[19] Lehrer, E. (2002), *Approachability in Infinitely Dimensional Spaces*, International Journal of Game Theory, 31, 255–270.

[20] Lehrer, E. (2003), *A Wide Range No-Regret Theorem*, Games and Economic Behavior, 42, 101–115.

[21] Lehrer, E. and E. Solan (2009), *Approachability with Bounded Memory*, Games and Economic Behavior, 66, 995–1004.

[22] Lehrer, E. and E. Solan (2007), *A General Internal Regret-Free Strategy*, manuscript.

[23] Levy, J. (2012), *Stochastic Games with Information Lag*, Games and Economic Behavior, 74, 243–256.

[24] Lugosi, G., S. Mannor, and G. Stoltz (2006), *Strategies for Prediction under Imperfect Monitoring*, Mathematics of Operations Research, 31, 562–580.

[25] Mannor, S. and N. Shimkin (2003), *On-Line Learning with Imperfect Monitoring*, Proceedings of the 16th Annual Conference on Learning Theory, 552–567. Springer.

[26] Perchet, V. (2011), *Internal Regret with Partial Monitoring: Calibration-Based Optimal Algorithm*, Journal of Machine Learning Research, 12, 1893–1921.

[27] Rustichini, A. (1999), *Minimizing Regret: The General Case*, Games and Economic Behavior, 29, 224–243.

[28] Sandroni, A., R. Smorodinsky, and R.V. Vohra (2003), *Calibration with Many Checking Rules*, Mathematics of Operations Research, 28, 141-153.

[29] Shmaya, E. (2008), *The Determinacy of Infinite Games with Eventual Perfect Monitoring*, Theoretical Economics, 3, 376–382.

[30] Spinat, X. (2002), *A Necessary and Sufficient Condition for Approachability*, Mathematics of Operations Research, 27, 31–44.

[31] Stoltz, G. and G. Lugosi (2005), *Internal Regret in On-Line Portfolio Selection*, Machine Learning, 59, 215–159.

[32] Vieille, N. (1992), *Weak Approachability*, Mathematics of Operations Research, 17, 781–791.

7 Appendix

Proof of Lemma 3.

$$\begin{aligned}
\|\overline{z}_k - \Pi_C(\overline{z}_k)\|^2 &\leq \|\overline{z}_k - \Pi_C(\overline{z}_{k-1})\|^2 && (10)\\
&\leq \left[\frac{A_{k-1}}{A_k}\right]^2 \|\overline{z}_{k-1} - \Pi_C(\overline{z}_{k-1})\|^2 && (11)\\
&+ \left[\frac{\alpha_k}{A_k}\right]^2 \|z_k - \Pi_C(\overline{z}_{k-1})\|^2 \\
&+ 2\frac{\alpha_k}{A_k}\frac{A_{k-1}}{A_k}\langle z_k - \Pi_C(\overline{z}_{k-1}), \overline{z}_{k-1} - \Pi_C(\overline{z}_{k-1})\rangle \\
&\leq \left[\frac{A_{k-1}}{A_k}\right]^2 \|\overline{z}_{k-1} - \Pi_C(\overline{z}_{k-1})\|^2 && (12)\\
&+ \left[\frac{\alpha_k}{A_k}\right]^2 \|z_k - \Pi_C(\overline{z}_{k-1})\|^2,
\end{aligned}$$

where Ineq. (10) and Ineq. (11) follow from the fact that $\Pi_C(\overline{z}_k)$ is the closest point to $\overline{z}_k$ in C and that $\overline{z}_k = \frac{A_{k-1}}{A_k}\overline{z}_{k-1} + \frac{\alpha_k}{A_k}z_k$. Ineq. (12) follows from (3). Continuing inductively we obtain,

$$\begin{aligned}
\|\overline{z}_k - \Pi_C(\overline{z}_k)\|^2 &\leq \left[\frac{\alpha_1}{A_k}\right]^2 \|z_1 - \Pi_C(\overline{z}_1)\|^2 + \sum_{l=2}^{k}\left[\frac{\alpha_l}{A_k}\right]^2 \|z_l - \Pi_C(\overline{z}_{l-1})\|^2 \\
&\leq M\sum_{l=1}^{k}\frac{\alpha_l^2}{A_k^2}, && (13)
\end{aligned}$$

for some large constant M, since $\{z_k\}_{k=1}^{\infty}$ is bounded and C is compact.

Since $\frac{\alpha_k}{A_k} \to 0$ and $A_k \to \infty$, it follows that $\lim\limits_{k\to\infty} \sum\limits_{l=1}^{k} \frac{\alpha_l^2}{A_k^2} = 0$ (see Lemma 7 in the Appendix), and the result follows. ■

Proof of Lemma 4. Starting in the same way as in the proof of Lemma 3, Ineq. (10) and Ineq. (11) still hold and the last term before (12) is

$$\gamma := 2\frac{\alpha_k}{A_k}\frac{A_{k-1}}{A_k}\left\langle z_k - \Pi_C(\overline{z}_{k-1}), \overline{z}_{k-1} - \Pi_C(\overline{z}_{k-1})\right\rangle .$$

The assumption of Lemma 4 does not require γ to be non-positive, as in Lemma 3. Nonetheless,

$$\gamma \quad = \quad 2\frac{\alpha_k}{A_k}\frac{A_{k-1}}{A_k}\left\langle z_k - \Pi_C(\overline{z}_{k-1}) + \Pi_C(\overline{z}_{k-2}) - \Pi_C(\overline{z}_{k-2}), \overline{z}_{k-1} - \Pi_C(\overline{z}_{k-1})\right\rangle .$$

By the bilinearity of the inner product,

$$\begin{aligned}
\gamma \quad &= \quad 2\frac{\alpha_k}{A_k}\frac{A_{k-1}}{A_k}\left\langle z_k - \Pi_C(\overline{z}_{k-2}), \overline{z}_{k-2} - \Pi_C(\overline{z}_{k-2})\right\rangle \\
&+ \quad 2\frac{\alpha_k}{A_k}\frac{A_{k-1}}{A_k}\left\langle z_k - \Pi_C(\overline{z}_{k-2}), \overline{z}_{k-1} - \overline{z}_{k-2} + \Pi_C(\overline{z}_{k-2}) - \Pi_C(\overline{z}_{k-1})\right\rangle \\
&+ \quad 2\frac{\alpha_k}{A_k}\frac{A_{k-1}}{A_k}\left\langle \Pi_C(\overline{z}_{k-2}) - \Pi_C(\overline{z}_{k-1}), \overline{z}_{k-1} - \Pi_C(\overline{z}_{k-1})\right\rangle \\
&\leq \quad 2\frac{\alpha_k}{A_k}\frac{A_{k-1}}{A_k}\left|\left\langle z_k - \Pi_C(\overline{z}_{k-2}), \overline{z}_{k-1} - \overline{z}_{k-2}\right\rangle\right| \qquad (14) \\
&+ \quad 2\frac{\alpha_k}{A_k}\frac{A_{k-1}}{A_k}\left|\left\langle z_k - \Pi_C(\overline{z}_{k-2}), \Pi_C(\overline{z}_{k-2}) - \Pi_C(\overline{z}_{k-1})\right\rangle\right| \\
&+ \quad 2\frac{\alpha_k}{A_k}\frac{A_{k-1}}{A_k}\left|\left\langle \Pi_C(\overline{z}_{k-2}) - \Pi_C(\overline{z}_{k-1}), \overline{z}_{k-1} - \Pi_C(\overline{z}_{k-1})\right\rangle\right| ,
\end{aligned}$$

where (4) yields Ineq. (14). Note that

$$\begin{aligned}
\|\overline{z}_{k-1} - \overline{z}_{k-2}\| \quad &= \quad \left\| \frac{1}{A_{k-1}}\sum_{l=1}^{k-1}\alpha_l z_l - \frac{1}{A_{k-2}}\sum_{l=1}^{k-2}\alpha_l z_l \right\| \\
&= \quad \left\| \frac{\alpha_{k-1}z_{k-1}}{A_{k-1}} - \left(\frac{1}{A_{k-2}} - \frac{1}{A_{k-1}}\right)\sum_{l=1}^{k-2}\alpha_l z_l \right\| \\
&= \quad \left\| \frac{\alpha_{k-1}z_{k-1}}{A_{k-1}} - \frac{\alpha_{k-1}\overline{z}_{k-2}}{A_{k-1}} \right\| \leq M\frac{\alpha_{k-1}}{A_{k-1}}
\end{aligned}$$

for some constant M, as in the proof of Lemma 3. Also note that the mapping $\Pi_C(\cdot)$ is non-expansive. From the Cauchy-Schwartz Inequality one concludes that

$$\gamma \leq 2\frac{\alpha_k}{A_k}\frac{A_{k-1}}{A_k}\left(M^2 + M^2 + M^2\right)\frac{\alpha_{k-1}^2}{A_{k-1}^2} \leq 6\frac{\alpha_k\alpha_{k-1}}{{A_k}^2}M^2. \qquad (15)$$

Combining Ineq. (15) with Ineq. (11) and Ineq. (13) yields

$$\begin{aligned}
\|\overline{z}_k - \Pi_C(\overline{z}_k)\|^2 \quad &\leq \quad 6M^2\sum_{l=2}^{k}\frac{\alpha_l\alpha_{l-1}}{A_k^2} + M^2\sum_{l=1}^{k}\frac{\alpha_l^2}{A_k^2} \\
&\leq \quad 7M^2 \max\left\{\sum_{l=2}^{k}\frac{\alpha_l\alpha_{l-1}}{A_k^2}, \sum_{l=1}^{k}\frac{\alpha_l^2}{A_k^2}\right\}. \qquad (16)
\end{aligned}$$

The assumptions $\frac{\alpha_k}{A_k} \to 0$ and $A_k \to \infty$ imply that $\lim\limits_{k\to\infty} \sum\limits_{l=2}^{k} \frac{\alpha_{l-1}^2 \alpha_l^2}{A_k^2} = 0$ and $\lim\limits_{k\to\infty} \sum\limits_{l=1}^{k} \frac{\alpha_l^2}{A_k^2} = 0$ (see Lemma 7 in the Appendix) and the result follows as in the proof of Lemma 3. ■

Proof of Lemma 5. Let $T \in \mathbb{N}$ be a positive natural number. From time $t+1$ until time T the DM receives $T-t$ signals concerning the realized state j_t, $\{X_{t+1}^t, X_{t+2}^t, \ldots, X_T^t\}$. Since $\mathbf{E}[X_n^t] = \eta(j_t) < \infty$, $\operatorname{Var}[X_n^t] < \infty$, then by the law of large numbers

$$\|F(t,T) - \eta(j_t)\| \to 0 \quad as\ T \to \infty \quad \text{a.s.}$$

Define

$$\delta = \min_{\alpha = \pm 1, \eta(j_t)} \{D\left(\eta(j_t) + \varepsilon\alpha, \eta(j_t)\right)\},$$

where $D(x,y) = x\log(\frac{x}{y}) + (1-x)\log(\frac{1-x}{1-y})$. Using Chernoff's Inequality,

$$\Pr(C_{t,T}) \geq 1 - 2\exp(-\delta(T-t)).$$

Hence,

$$\begin{aligned} \Pr\left(\bigcap_{t=1}^{n} C_{t,T}\right) &\geq \prod_{t=1}^{n} (1 - 2\exp(-\delta(T-t))) \geq (1 - 2\exp(-\delta(T-n)))^n \\ &\approx \exp\left(-\frac{2n}{\exp(\delta(T-n))}\right) \text{ for } n \gg 1. \end{aligned} \tag{17}$$

Taking $T = T_n = n + \lceil \frac{4}{\delta}\ln(n) \rceil$,

$$\Pr\left(\bigcap_{t=1}^{n} C_{t,T}\right) \gtrsim \exp(-2n^{-3}) \geq 1 - \frac{1}{n^2},$$

as desired. ■

Lemma 7. *Let $\alpha_1, \alpha_2, \ldots$ be a sequence of non-negative real numbers and denote $A_k = \sum_{l=1}^{k} \alpha_l$. If $\lim\limits_{k\to\infty} A_k = \infty$ and $\lim\limits_{k\to\infty} \frac{\alpha_k}{A_k} = 0$, then*

(i) $\lim\limits_{k\to\infty} \frac{\sum\limits_{l=1}^{k} \alpha_l^2}{A_k^2} = 0$,

(ii) $\lim\limits_{k\to\infty} \frac{\sum\limits_{l=2}^{k} \alpha_l \cdot \alpha_{l-1}}{A_k^2} = 0$.

Proof. We will start by proving (i). For every k, choose $l_k \in \{1, \ldots, k\}$ that satisfies $\alpha_{l_k} = \max\limits_{l=1,\ldots,k} \alpha_l$. Since all the terms in A_k are non-negative,

$$\frac{\sum_{l=1}^{k}\alpha_l^2}{A_k^2} \leq \frac{\left(\max_{l=1,\ldots,k}\alpha_l\right)\cdot\sum_{l=1}^{k}\alpha_l}{A_k^2} = \frac{\alpha_{l_k}\cdot A_k}{A_k^2} = \frac{\alpha_{l_k}}{A_k} = \frac{\alpha_{l_k}}{A_{l_k}}\cdot\frac{A_{l_k}}{A_k}. \tag{18}$$

The index l_k can either tend to infinity or be finite. If $\lim_{k\to\infty} l_k = \infty$, then $\lim_{k\to\infty}\frac{\alpha_{l_k}}{A_{l_k}} = 0$ and $\frac{A_{l_k}}{A_k} \leq 1$, hence $\lim_{k\to\infty}\frac{\sum_{l=1}^{k}\alpha_l^2}{A_k^2} = 0$. On the other hand, if $\lim_{k\to\infty} l_k < \infty$, then $\lim_{k\to\infty}\frac{A_{l_k}}{A_k} = 0$ and $\frac{\alpha_{l_k}}{A_{l_k}} \leq 1$ and the result follows.

For (ii), note that $\frac{\sum_{l=2}^{k}\alpha_l\cdot\alpha_{l-1}}{A_k^2} \leq \frac{\left(\max_{l=1,\ldots,k-1}\alpha_l\right)\cdot\sum_{l=1}^{k}\alpha_l}{A_k^2} \leq \frac{\left(\max_{l=1,\ldots,k}\alpha_l\right)\cdot\sum_{l=1}^{k}\alpha_l}{A_k^2}$ and continue as in the proof of Ineq. (18). ■

Lemma 8. *Using the terminology of Proposition 1, assume that $\phi(h_n) \leq n + o(n)$ and α_k denotes the number of stages from the k^{th} signal until the $(k+1)^{\text{st}}$ signal for a given h_∞. Then $A_k = \sum_{l=1}^{k}\alpha_l \to \infty$ and $\frac{\alpha_k}{A_k} \to 0$ as $n \to \infty$.*

Proof. The delay function ϕ is non-decreasing with respect to $\prec$ and bounded by $n \leq \phi(h_n) \leq n + o(n)$. Assume that the k^{th} signal reached the DM at time n. From Eq. (6) we conclude that

$$A_k = \sum_{l=1}^{k}\alpha_l = n + \alpha_k \leq \phi(h_n).$$

Therefore,

$$\frac{\alpha_k}{A_k} = \frac{\alpha_k}{n+\alpha_k} \leq \frac{o(n)}{n} \tag{19}$$

and

$$A_k = \alpha_k + n \to \infty \quad \text{as } n \to \infty, \tag{20}$$

as claimed. ■