A general internal regret-free strategy*

Ehud Lehrer† and Eilon Solan‡

August 14, 2014

Abstract:

We study sequential decision problems where the decision maker does not observe the states of nature, but rather receives a noisy signal, whose distribution depends on the current state and on the action that she plays. We do not assume that the decision maker considers the worst-case scenario, but rather has a response correspondence, which maps distributions over signals to subjective best responses. We extend the concept of internal regret-free strategy to this setup and provide an algorithm that generates such a strategy.

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

Keywords: internal no regret, no regret, imperfect monitoring, approachability, response correspondence.

*A previous version of this paper, dated July 2007, was titled "Learning to play partially-specified equilibrium". We thank David Lagziel and Yishay Mansour, Marc Teboule, two anonymous referee, and the Associate Editor for helpful comments that improved the presentation of the results. 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 #538/11. Solan acknowledges the support of the Israel Science Foundation, Grant #212/09.

†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.

‡School of Mathematical Sciences, Tel Aviv University, Tel Aviv 69978, Israel. e-mail: eilons@post.tau.ac.il.

1 Introduction

A decision maker (DM) is facing a sequential decision problem. At each stage a state of nature is realized and the DM, who is ignorant of the state of nature, has to choose an action. The DM's stage payoff, which is not told to the DM, depends both on her action and on the state of nature (referred to as *outcomes* at the machine learning literature) at that stage. In a Bayesian setting the DM would have a prior distribution over all possible sequences of states and she could then take decisions that would maximize her expected payoff. In this paper the DM is not Bayesian: she would like to achieve a certain goal whatever be the realized sequences of states, namely, to minimize regret.

Hannan (1957) showed that the DM has a randomized strategy that, comparing average payoffs, could asymptotically perform at least as well as any *stationary* strategy, regardless of the sequence of states. Namely, he proved that the DM has a strategy under which, asymptotically, the actual average payoff is as high as the average payoff that any stationary strategy would obtain, assuming that the states are not affected by the DM's choices. Strategies of this type are called *external regret-free strategies*.

The stronger notion of internal regret-free strategies has been introduced subsequently (see, Foster and Vohra (1997, 1999), Fudenberg and Levine (1999), Hart and Mas-Colell (2000), Stoltz and Lugosi (2005), Cesa-Bianchi and Lugosi (2006), or Blum and Mansour (2007)). A strategy is internal regret-free if for any given sequence of states, and for every action a, the DM cannot gain asymptotically from replacing a by another action a', that is, by playing a' in all stages in which she played a. Hart and Mas-Colell (2000), Foster and Vohra (1998), and Fudenberg and Levine (1999) proved the existence of an internal regret-free strategy when monitoring is perfect.

In this paper we study sequential decision problems with imperfect monitoring. Here, the DM does not observe the state, but rather a noisy signal that may depend also on her action.

To motivate the study, consider the following example, adapted from Mannor and Shimkin (2003). A doctor treating patients who may be suffering from several illnesses has several treatments at her disposal. Suppose that the probability that every given treatment will be effective for every given disease is known, and moreover that even if a treatment is not effective, it still provides the doctor with some information about the true nature of the disease. For each patient the doctor chooses a mixture of treatments and observes the effects of the treatment, thereby increasing the probability of correctly identifying the patient's illness. In this setup, the doctor's strategy is internal regret-free if, in the long run, she never regrets choosing a specific mixture of treatments for those patients who were prescribed that mixture.

The issue of having no regret in decision problems with imperfect monitoring has been mainly treated in a restricted sense of external regret. Rustichini (1999) proved that external regret-free strategies exist, Mannor and Shimkin (2003) provided an approachability-based algorithms that generate such a strategy when the DM's choices do not affect the signals, and Piccolboni and Schindelhauer (2001), Cesa-Bianchi, Lugosi, and Stoltz (2006), and Cesa-Bianchi and Lugosi (2006, Chapter 6) showed regret-free procedures under the condition whereby the payoff matrix of the DM is obtained through a linear transformation from the signalling structure. Lugosi, Mannor and Stoltz (2008) proved the general perfect monitoring case. In another strand of the literature, Foster and Vohra (1997, 1999), Hart and Mas-Colell (2000, 2001), Stoltz and Lugosi (2005), and Blum and Mansour (2007) investigated internal regret-free strategies in the perfect monitoring case.

As mentioned before, a strategy is internal regret-free under perfect monitoring, if for any given sequence of states and for every action a, the DM cannot profit from replacing a with another action. The internal regret test compares at any stage T the actual total payoff over the stages where action a was played up to stage T divided by T, with the corresponding payoff had a been replaced by another action.

In a general sequential decision problem with imperfect monitoring, the empirical frequency of the signals provides only partial information about the frequency of the actual states chosen by nature. There might be many distributions over states that are consistent with past signals. When computing her payoff had she played a different mixed action, the DM could consider any one of these distributions. For instance, among all these distributions the DM could consider the one that entails the worst payoff possible. To maximize her payoff in the worst-case scenario, the DM's optimal action could be mixed (see Example 1 below). In this case, a strategy is regret-free if the DM cannot profit by replacing any mixed action that she played infinitely often by another mixed strategy, where the gain is measured w.r.t. the worst behavior of nature that is consistent with the obtained signals.

We prove that when the DM is pessimistic, namely considers the worst-case scenario, an internal regret-free strategy exists. For this purpose we use Blackwell's approachability theory to construct such a strategy. In a follow-up to our paper, Perchet (2009) showed that one can use calibration methods instead of Blackwell's approachability theory to construct regret-free strategies. In Remark 9 below we elaborate on the differences between Perchet (2009) and our result.

Decision makers are not necessarily pessimistic. An optimistic DM, for instance, believes that nature is benevolent and is trying to help her. In this case she would compare her payoff had nature been cooperative (and played the best distribution over states) with the performance

of an alternative action had nature been playing the best possible distribution against the latter. Our definition and results are general enough to encompass this optimistic approach, as well as other methods to measure the performance of the DM's own strategy.

It is well known that in the setup of multiplayer repeated games, when each player plays an internal regret-free strategy (and each player considers all other players as 'nature'), the play converges to the set of correlated equilibria (see Hart and Mas-Colell, 2000). It turns out that in multiplayer repeated games with imperfect monitoring, when all players play an internal regret-free strategy, the play does not converge to the set of correlated equilibria, but rather to the larger set of partially-specified correlated equilibria (see Lehrer and Solan (2007) for more details).

In recent years the literature on internal regret has expanded in a fast pace. Perchet (2011, 2013) summarizes the relations between the notions of approachability, no regret, and calibration, and Hazan and Kakade (2012) relate the computational complexity of calibration to that of approximate Nash equilibria. The study of the possible optimal rates of internal regret, that was initiated by Cesa-Bianchi, Lugosi, and Stoltz (2006), was recently completed by Foster and Rakhlin (2012).

The paper is arranged as follows. In Section 2 we present a model of sequential decision problems and describe our main results. The proofs are presented in Section 3. The two appendices contain extensions of known results to more general setups, which we need here.

2 The Model and the Main Results

2.1 The Model

A decision maker (DM) faces a sequential decision problem. At every stage the DM takes an action in a finite action set A, and nature chooses a state in a finite set of states Ω. The payoff function of the DM is given by a function $u : A \times \Omega \to \mathbb{R}$, which is assumed (without loss of generality) to be bounded between 0 and 1. The actual action and the state at stage t are denoted by a^t and ω^t, respectively.

We denote[1] by $X = \Delta(A)$ the set of the DM's mixed actions, and by $Y = \Delta(\Omega)$ the set of probability distributions over the set of states, Ω.

The DM does not observe the realized state; rather, she observes a noisy signal that depends on it and on her action. The monitoring structure is given by a function $\mathfrak{s} : A \times \Omega \to \Delta(S)$, where S is a finite set of signals. Thus, at every stage $t \in \mathbb{N}$ the DM observes the signal s^t, chosen according to the probability distribution $\mathfrak{s}(a^t, \omega^t)$. We assume that the DM has

[1] For a finite set Z, the set of probability distributions over Z is denoted by $\Delta(Z)$.

perfect recall, meaning that she recalls all her past actions and the signals she received. For technical reasons it is convenient to assume that the signal includes the action that the DM chose; formally this means that if $\mathfrak{s}(a,\omega)(s') > 0$ and $\mathfrak{s}(a',\omega')(s') > 0$ then $a = a'$. We denote by $a(s)$ the action corresponding to the signal s.

The domains of the payoff function u and the signal function s are extended to $X \times Y$ in a multilinear fashion. Thus, for every $a \in A$ and $y = (y_\omega)_{\omega\in\Omega} \in Y$, $\mathfrak{s}(a,y) = \sum_{\omega\in\Omega} y_\omega \mathfrak{s}(a,\omega) \in \Delta(S)$ is the mixed signal induced by a and y. For every $y \in Y$ the *footprint* of y is $\mu(y) = (\mathfrak{s}(a,y))_{a\in A} \in (\Delta(S))^A$. If, for instance, at every stage the state is chosen independently according to the probability distribution y, and the DM plays the action $a \in A$ infinitely often, then by the strong law of large numbers for martingales, the empirical frequency of signals that the DM observes during the stages in which she played the action a converges to $\mathfrak{s}(a,y)$ with probability 1. Because the DM recalls all past signals, if she plays all actions infinitely often the information she would collect about y is the collection of distributions $(\mathfrak{s}(a,y))_{a\in A}$, which is the footprint that y leaves. Let $\mathcal{M} := \{\mu(y) \in (\Delta(S))^A \colon y \in Y\}$ be the set of all footprints.

We say that two distributions over states $y, y' \in Y$ are equivalent, denoted $y \sim y'$, if they have the same footprint: $\mu(y) = \mu(y')$. For every $\mu' \in \mathcal{M}$ denote by $Y(\mu') := \{y \in Y \colon \mu(y) = \mu'\}$ the set of all distributions over states whose footprint coincides with μ'. The set $Y(\mu)$ is defined by linear equalities, and therefore the set-valued functions $y \mapsto Y(\mu(y))$ and $\mu \mapsto Y(\mu)$ are upper-semi-continuous.[2] By Rockafeller and Wets (2009, Chapter 9, Section E) it follows that the set-valued function $y \mapsto Y(\mu(y))$ is Lipschitz continuous.

2.2 External regret-free Strategies

Definition 1. *(i) Let $\widetilde{X} \subseteq X$ be a set of mixed actions. A* strategy *σ with range $\widetilde{X}$ for the DM is a function that assigns a probability distribution over $\widetilde{X}$ to every finite history in $H := \bigcup_{t=1}^{\infty} (\widetilde{X} \times S)^{t-1}$.*
(ii) A strategy σ with range $\widetilde{X}$ is stationary *if $\sigma(h)$ is independent of h; that is, there is a distribution $p \in \Delta(\widetilde{X})$ such that $\sigma(h) = p$ for every $h \in H$.*

Throughout the paper, $\widetilde{X}$ denotes the range of the strategy σ. When $\widetilde{X}$ is finite we say that σ is a strategy with finite range. In case $\widetilde{X}$ coincides with the set A, a strategy with range $\widetilde{X}$ is a standard behavior strategy.

The concept of internal regret-free strategy that we are about to introduce takes into account the actual mixed actions played at past stages, and not the realized actions. In the definition

[2]Unless indicated otherwise, we use the supremum norm: for every two vectors $x, x' \in \mathbb{R}^d$, the distance between x and x' is $d(x,x') = \max\{|x_i - x'_i|, 1 \le i \le d\}$, and for every subset $D \subseteq \mathbb{R}^d$, the distance between x and D is $d(x,D) := \inf_{x'\in D} d(x,x')$. Likewise, the distance between two subsets Y_1 and Y_2 of Y is given by $d(Y_1,Y_2) := \max_{y_1\in Y_1} \min_{y_2\in Y_2} d(y_1,y_2)$.

of a strategy we therefore allow the strategy to choose randomly a mixed action from $\widetilde{X}$ (and then an action is selected according to the chosen mixed action). This feature of a strategy is essential to the construction of an internal regret-free strategy when monitoring is imperfect.

Given a fixed sequence of states $\vec{\omega} = (\omega^t)_{t\in\mathbb{N}}$, every strategy σ with range $\widetilde{X}$ induces a probability distribution $\mathbb{P}_{\vec{\omega},\sigma}$ over the set of infinite plays $(\widetilde{X} \times S)^\infty$.

Let $\varepsilon \geq 0$. A strategy is external ε-regret-free if the actual (unobserved) long-run average payoff is at least, up to ε, the long-run average payoff that can be guaranteed against the worst distribution over states consistent with the empirical frequency of states. Formally, denote by $\widehat{\omega}^T := \frac{1}{T}\sum_{t=1}^{T}[1(\omega^t)]$ the empirical distribution of states up to stage T.

Definition 2. *[Rustichini, 1999] The strategy σ is* external regret-free *if for every sequence of states $\vec{\omega} = (\omega^t)_{t\in\mathbb{N}}$ and for every $x \in X$,*

$$\limsup_{T\to\infty}\left(\min_{y\sim\widehat{\omega}^T} u(x,y) - \frac{1}{T}\sum_{t=1}^{T} u(a^t,\omega^t)\right) \leq 0 \tag{1}$$

with $\mathbb{P}_{\vec{\omega},\sigma}$-probability 1.

In words, a strategy is external regret-free if the actual long-run average payoff is at least the payoff that the DM would obtain by choosing a fixed alternative mixed action had nature chosen the states according to the worst distribution over state that is equivalent to the empirical distribution of states.

Remark 1. The definition of external no regret, as well as the extensions that we will provide later, implicitly assumes that the sequence of states $\vec{\omega}$ is fixed; that is, nature is oblivious to the DM's choices. The reason is that if nature responds to DM's choices, then Definition 2 cannot be interpreted as having no regret. Indeed, when nature is responsive, the sequence of states is generated as a response to the DM's actions. If the DM deviates from the sequence of actions $(a^1,\ldots,a^T)$ and chooses her actions according to the stationary strategy x, then the realized sequence of actions changes, and, since nature is responsive, the sequence of states might change as well. In particular, the quantity $\min_{y\sim\widehat{\omega}^T} u(x,y)$ in Definition 2 no longer measures the hypothetical worst-case payoff to the DM, and therefore the left-hand side of Eq. (1) no longer measures the DM's regret.

To illustrate this point, consider the Prisoner's Dilemma game that appears in Figure 1, where the DM is Player 1, nature is Player 2, and only the payoffs of the DM are depicted.

	C	D
C	3	0
D	4	1

Figure 1: The Prisoner's Dilemma as a sequential decision problem.

Suppose that nature is strategic and plays the *Grim-Trigger strategy* τ: nature begins with playing C, and at any other stage it plays C if and only if the DM played C in all previous stages. Suppose also that the DM uses the strategy σ that always plays C. Then the long-run average payoff is 3. The strategy σ is *not* regret-free according to Definition 2. Indeed, under (σ, τ) the sequence of realized states is $\vec{\omega} = (C, C, C, \cdots)$, and the left-hand side of Eq. (1) is equal to 1 for $x = [1(D)]$. However, since nature is strategic, if the DM uses the strategy σ' that always plays D, her long-run average payoff is 1, and therefore the quantity $\min_{y \sim \widehat{\omega}^T} u(x, y)$ does not measure the hypothetical worst average payoff to the DM had she played the stationary strategy x.

It should be emphasized that Definition 2 is valid also when nature is strategic. However, its interpretation in this case is questionable.

Remark 2. Standard continuity and compactness arguments show that the convergence in Eq. (1) is uniform over $x \in X$, so that Definition 2 is indeed equivalent to the definition given by Rustichini (1999) for external regret-free strategies.

The main objective of this paper is to introduce the concept of internal regret-free strategies when monitoring is imperfect and to extend the following theorem.

Theorem 1 (Rustichini (1999), Lugosi, Mannor, Stoltz (2008))**.** *There exists an external regret-free strategy with range* $\widetilde{X} = A$.

2.3 Internal Regret-Free Strategies

2.3.1 The definition

Let $x \in \widetilde{X}$, let $a \in A$, and let $T \in \mathbb{N}$. Denote by I_x^T the set of all stages $t \leq T$ in which the DM played the mixed action x, and by $\widehat{\omega}_x^T$ the empirical frequency of states over I_x^T. That is, $\widehat{\omega}_x^T := \frac{1}{|I_x^T|} \sum_{t \in I_x^T} [1(\omega^t)]$, provided that $I_x^T \neq \emptyset$.

Definition 3. *Let* $\varepsilon \geq 0$. *A strategy* σ *with finite range* $\widetilde{X}$ *is* internal ε-regret-free *if for every sequence of states* $\vec{\omega} = (\omega^t)_{t \in \mathbb{N}}$, *every* $x \in \widetilde{X}$, *and every* $x' \in X$,

$$\limsup_{T \to \infty} \left(\min_{y \sim \widehat{\omega}_x^T} u(x', y) - \frac{1}{|I_x^T|} \sum_{t \in I_x^T} u(a^t, \omega^t) \right) \leq \varepsilon \tag{2}$$

holds[3] *on the event* $\{\lim_{T \to \infty} |I_x^T| = \infty\}$, *with respect to* $\mathbb{P}_{\vec{\omega}, \sigma}$.

[3] We say that event B holds on event A if $\mathbb{P}(A \cap B) = \mathbb{P}(A)$. Thus, a certain inequality holds on event A if the probability of all points in A that do not satisfy it is 0.

A strategy is internal regret-free if for every mixed action x that the DM chooses infinitely often, and for every alternative mixed action $x' \in X$ (not necessarily in $\widetilde{X}$), the actual long-run average payoff over the set of stages I_x^T is, up to ε, at least the minimal payoff that the DM would obtain if in those stages she played the stationary strategy x', and nature chose states in a way equivalent to the actual empirical frequency, provided T is sufficiently large.

When monitoring is perfect, the only mixed state equivalent to $\widehat{\omega}_x^T$ is $\widehat{\omega}_x^T$. In this case, Definition 3 with $\widetilde{X} = A$ is identical to the definition of internal no regret introduced by Hart and Mas-Colell (2000).

2.3.2 Remarks

Remark 3. The strategy σ in Definition 3 is required to have a finite range. At first glance it seems that the larger the range of σ, the easier it is to pass the no-regret test, because the DM has a richer set of choices. This is not the case because the number of inequalities that should be met grows as the range of σ grows. Note that whether or not a mixed action x is ε-optimal depends solely on its performance over the set of stages I_x^T. This means that the existence of other mixed actions in the range of σ does not help x to be ε-optimal against $\mu(\widehat{\omega}_x^T)$.

Remark 4. As in the case of perfect monitoring, any internal ε-regret-free strategy is, in particular *external ε-regret-free*, with ε replacing the 0 on the right-hand side of Eq. (1). Indeed, suppose that σ is an internal ε-regret-free strategy and fix $x' \in X$. For every $T \in \mathbb{N}$ denote by $y_T \in Y$ a mixed action where the minimum $\min_{y \sim \widehat{\omega}^T} u(x', y)$ is attained. Then for every $T \in \mathbb{N}$,

$$\begin{aligned}
\min_{y \sim \widehat{\omega}^T} u(x', y) - \frac{1}{T}\sum_{t=1}^{T} u(a^t, \omega^t) &= u(x', y_T) - \frac{1}{T}\sum_{t=1}^{T} u(a^t, \omega^t) \\
&= u(x', y_T) - \sum_{x \in \widetilde{X}} \frac{|I_x^T|}{T}\left(\frac{1}{|I_x^T|}\sum_{t \in I_x^T} u(a^t, \omega^t)\right) \\
&\leq \sum_{x \in \widetilde{X}} \frac{|I_x^T|}{T} \min_{y \sim \widehat{\omega}_x^T} u(x', y) - \sum_{x \in \widetilde{X}} \frac{|I_x^T|}{T}\left(\frac{1}{|I_x^T|}\sum_{t \in I_x^T} u(a^t, \omega^t)\right) \\
&= \sum_{x \in \widetilde{X}} \frac{|I_x^T|}{T}\left(\min_{y \sim \widehat{\omega}_x^T} u(x', y) - \frac{1}{|I_x^T|}\sum_{t \in I_x^T} u(a^t, \omega^t)\right),
\end{aligned}$$

where the inequality holds because $\sum_{x \in \widetilde{X}} \frac{|I_x^T|}{T} y_x \sim \widehat{\omega}^T$ whenever $y_x \sim \widehat{\omega}_x^T$ for every $x \in \widetilde{X}$. Because σ is an internal ε-regret-free strategy with a finite range $\widetilde{X}$, it follows by Definition 3 that the limit superior of the right-hand side is at most ε, as claimed.

Remark 5. In Remark 1 we mentioned that the interpretation of Definitions 2 (and also of Definition 3) as regret is questionable when nature is not oblivious, yet the results hold also in a strategic setup. In fact, as shown in Hart and Mas-Colell (2001), the concept of regret-free strategies can be useful also in strategic setups; in this paper it is proven that if all players use regret-free strategies (and ignore the fact that other players do the same), then the long-run average play converges to the set of correlated equilibria of the one-shot game.

Remark 6. The range of the strategy σ is $\widetilde{X}$ and not X. Yet, in order to be internal ε-regret-free, the performance of a strategy must be better, up to ε, than any alternative mixed action regardless of whether it is in $\widetilde{X}$. For this reason the mixed action x' in Definition 3 is not restricted to $\widetilde{X}$.

Remark 7. In Definition 3 the quantity $\min_{y\sim\widehat{\omega}_x^T} u(x',y)$ represents the hypothetical worst-case expected payoff to the DM if she played the mixed action x' instead of x. This hypothetical payoff is compared to her actual payoff $\frac{1}{|I_x^T|}\sum_{t\in I_x^T} u(a^t,\omega^t)$. Alternatively, one could define internal regret-free strategy by comparing the hypothetical worst-case expected payoff to the expected average payoff of the DM, $\frac{1}{|I_x^T|}\sum_{t\in I_x^T} u(x,\omega^t)$. By the strong law of large numbers for martingales, the difference $\frac{1}{|I_x^T|}\sum_{t\in I_x^T}(u(a^t,\omega^t)-u(x,\omega^t))$ converges to 0 as $|I_x^T|$ goes to infinity, and therefore the current definition and the alternative one coincide.

Remark 8. [Internal no regret with and without perfect monitoring] Definition 3 labels a strategy internal regret-free if the performance of any mixed action played infinitely often is the best possible, over the stage in which it has been played. In the case of perfect monitoring, in contrast, a strategy is internal regret-free if the performance of any pure action played is the best possible, over the stages in which it has been played. The question arises as to why in the imperfect monitoring case only the performances of the mixed actions used are examined, and not the performance of the actions that have actually been used. The reason is explained in the discussion after Example 1 below.

Remark 9. Another variation on the theme of internal no regret could be the definition used in an early version of this paper (Solan and Lehrer, 2007) and adopted by Perchet (2009): a strategy σ is internal ε-regret-free if for every sequence of states $\vec{\omega}=(\omega^t)_{t\in\mathbb{N}}$, every $x\in\widetilde{X}$, and every $x'\in X$,

$$\begin{aligned}&\limsup_{T\to\infty}\min_{y\sim\widehat{\omega}_x^T}\frac{1}{T}\sum_{t\in I_x^T}\big(u(x',y)-u(a^t,\omega^t)-\varepsilon\big)\\&=\limsup_{T\to\infty}\min_{y\sim\widehat{\omega}_x^T}\frac{|I_x^T|}{T}\left(u(x',y)-\frac{\sum_{t\in I_x^T}u(a^t,\omega^t)}{|I_x^T|}-\varepsilon\right)\le 0\end{aligned}\tag{3}$$

with $\mathbb{P}_{\vec{\omega},\sigma}$-probability 1. Clearly, Eq. (2) implies Eq. (3). The opposite implication is incorrect, implying that Definition 3 is strictly stronger than this variation. The reason is the following. As long as $\liminf_{T\to\infty} \frac{|I_x^T|}{T} > 0$, Eq. (3) indeed implies Eq. (2), possibly with a smaller ε. However, if $\liminf_{T\to\infty} \frac{|I_x^T|}{T} = 0$, a condition that depends both on the sequence $\vec{\omega}$ and the strategy σ, then at those stages in which $\frac{|I_x^T|}{T}$ is close to zero, Eq. (3) is mute regarding the actual difference between $\frac{1}{|I_x^T|}\sum_{t\in I_x^T} u(a^t,\omega^t)$ and $\min_{y\sim\widehat{\omega}_x^T} u(x',y)$. In other words, whenever there is no guarantee that actions are asymptotically played with a positive frequency, the definition that uses Eq. (3) does not imply Definition 3.

Our first result is the following.

Theorem 2. *For every $\varepsilon > 0$ there exists a strategy with a finite range that is internal ε-regret-free.*

Theorem 2 is a special case of Theorem 3 below. The proof of Theorem 3 provides an algorithm that constructs an internal ε-regret-free strategy.

2.4 Response Correspondences

2.4.1 Examples

The notion of internal regret-free strategies introduced above refers to a pessimistic DM who considers the worst-case scenario. That is, the DM has no regret if the actual average payoff is at least the payoff that could have been obtained against the worst distribution over states consistent with past signals. However, the DM might have different ways to assess her performance, as illustrated by the following example. This example and the following one serve as a motivation to the concept of response correspondence, which captures the attitude of the DM towards uncertainty.

Example 1. *Consider a decision problem with three states, $\Omega = \{L, M, R\}$, and two actions, $A = \{T, B\}$. The payoffs[4] are given in Figure 2.*

	L	M	R
T	3	0	0
B	0	2	2

Figure 2: The payoffs in Example 1.

Assume that the DM has no information about nature's choices; that is, the set of signals is identical with the set of actions.

[4]For convenience, in the examples we allow payoffs to be greater than 1.

A pessimistic DM who considers the worst-case scenario will play at every stage the mixed action $[\frac{2}{5}(T), \frac{3}{5}(B)]$*, which guarantees an expected payoff of* $\frac{6}{5}$ *per stage. This is the DM's optimal strategy in the zero-sum game defined by the payoff function in Figure 2.*

Example 1 also illustrates why in the definition of internal regret-free strategies we consider the performance of the mixed actions chosen by the DM, rather than the performance of the realized pure actions. In this example, the DM obtains no information about the state of nature. A pessimistic DM has only one strategy which is internal regret-free. This is the stationary strategy that prescribes playing the mixed action $[\frac{2}{5}(T), \frac{3}{5}(B)]$ at every stage. If the performance of the realized pure actions would have been examined against alternatives, every strategy would have generated internal regret. Indeed, in Example 1 both T and B are inferior to $[\frac{2}{5}(T), \frac{3}{5}(B)]$, regardless of the sequence nature chooses.

2.4.2 The definition

The following definition captures the idea that different DMs might have different approaches to uncertainty. A DM is characterized by what we call a *response correspondence*.

Definition 4. *A* response correspondence *is a set-valued function* $\mathrm{R} : \mathcal{M} \hookrightarrow X$ *that satisfies* $\mathrm{R}(\mu) = \mathrm{R}(\mu')$ *whenever* $Y(\mu) = Y(\mu')$.

The interpretation of a response correspondence is as follows. Suppose that the response correspondence of the DM is R. When the DM observes that in a certain subset of stages the empirical footprint coincides with μ, she regards the mixed actions in $\mathrm{R}(\mu)$ as optimal responses to the set of distributions over states that could induce μ.

The condition that $\mathrm{R}(\mu) = \mathrm{R}(\mu')$ whenever $Y(\mu) = Y(\mu')$ reflects the following property. The set of mixed actions that the DM views as optimal depends only on the set of distributions over states that may induce the observed average signal (and not, for instance, on the mixed actions chosen by the DM or on the order by which signals where received). In particular it implies that when the DM receives no information, R is constant.

2.4.3 Examples of response correspondences

Three response correspondences that were mentioned earlier are the following.

Maximum entropy: Denote $y_\mu^{ENT} := \mathrm{argmax}_{y \in Y(\mu)} \left\{ -\sum_{\omega \in \Omega} y_\omega \ln(y_\omega) \right\}$. The response correspondence[5] of a DM who believes in maximal entropy is

$$\mathrm{R}^{ENT}(\mu) := \mathrm{argmax}_{x \in X} \left\{ u(x, y_\mu^{ENT}) \right\}, \quad \forall \mu \in \mathcal{M}.$$

[5]The compactness of $Y(\mu)$ and the concavity of the entropy function ensure that y_μ^{ENT} is well defined.

Worst case: The response correspondence of a pessimistic DM is

$$\mathrm{R}^{WC}(\mu) := \mathrm{argmax}_{x\in X}\left\{\min_{y\in Y(\mu)} u(x,y)\right\}, \quad \forall \mu \in \mathcal{M}.$$

Best case: The response correspondence of an optimistic DM is

$$\mathrm{R}^{BC}(\mu) := \mathrm{argmax}_{x\in X}\left\{\max_{y\in Y(\mu)} u(x,y)\right\}, \quad \forall \mu \in \mathcal{M}.$$

Example 1 continued. *In Example 1 if the DM's attitude towards uncertainty were not pessimistic, she may have different regret-free strategies. Suppose, for example, that the DM were optimistic, considering the best-case scenario. If she played T at every stage, her best-case stage payoff was 3, while if she played B at every stage, it was only 2. On the other hand, if she played a mixed action $[p(T), (1-p)(B)]$ at every stage, her best-case expected stage payoff was $\max\{3p, 2(1-p)\}$. A regret-free strategy for an optimistic DM would then be playing T in all stages.*

A DM who believes in maximum entropy thinks that nature chooses a state with the distribution $[\frac{1}{3}(L), \frac{1}{3}(M), \frac{1}{3}(R)]$. In this case her best response, and her regret-free strategy, is the stationary strategy B.

When the DM does not get any signal, as in Example 1, she takes into account all possible distributions over states, because no distribution is excluded. However, when she receives signals that depend both on her action as well as on the actual state, some distributions over states are ruled out as being inconsistent with the empirical signals. This is illustrated by the following example.

Example 2. *Consider a decision problem with three states, $\Omega = \{L, M, R\}$, two actions, $A = \{T, B\}$, and four signals, $S = \{s_1, s_2, s_3, s_4\}$. The payoffs (at the center) and the signals that the DM observes (at the upper-right corner) are given in Figure 3.*

	L	M	R
T	$3^{\ s_1}$	$3^{\ s_2}$	$0^{\ s_1}$
B	$0^{\ s_3}$	$2^{\ s_4}$	$5^{\ s_3}$

Figure 3: The payoffs and the signals in Example 2.

Suppose that the DM uses the stationary strategy T. At every stage the DM receives the signal s_1 if nature chooses L or R, and the signal s_2 if nature chooses M. Suppose that the long-run frequency of stages in which the DM observed the signal s_1 is 50%. The DM can

deduce from this information that the long-run frequency of the state M is 50%. Thus, during half of the stages the DM's payoff was 3, and during the other half her payoff was either 0 or 3.

If the DM played the stationary strategy B, her payoff would be 2 whenever the signal was s_2, and either 0 or 5 whenever the signal was s_1. To determine whether the DM regrets playing T and not B, the DM should therefore compare the following scenarios:

- *Stationary strategy T: A payoff of 3 during half the stages, and a payoff in $\{3, 0\}$ during the other half.*
- *Stationary strategy B: A payoff of 2 during half the stages, and a payoff in $\{5, 0\}$ during the other half.*
- *Mixed stationary strategy $[p(T), (1-p)(B)]$: An expected payoff $2 + p$ during half the stages, and an expected payoff in $\{3p, 5(1-p)\}$ during the other half.*

The result of this comparison depends on the characteristics of the DM. The reader can verify that the stationary strategy $[\frac{5}{8}(T), \frac{3}{8}(B)]$ maximizes the DM's payoff in the worst-case scenario, hence it is a regret-free strategy in this case. On the other hand, a regret-free strategy of an optimistic DM is to always play B, while a DM who believes in maximal entropy is indifferent between T and B.

2.5 Internal No Regret with Response Correspondences

For every subset $D \subseteq \mathbb{R}^d$ and every $\varepsilon > 0$, denote the set of all points which are ε-close (in the supremum norm) to D by $B(D, \varepsilon)$. That is, $B(D, \varepsilon) := \{x \in \mathbb{R}^d \colon d(x, D) \leq \varepsilon\}$. Thus, for every $\mu \in \mathcal{M}$, $B(\mu, \varepsilon) := B(\{\mu\}, \varepsilon)$ is the set of all footprints that are ε-close to μ, and $\mathrm{R}(B(\mu, \varepsilon))$ stands for the union of all $\mathrm{R}(\mu')$, where $\mu' \in B(\mu, \varepsilon)$. The set $B(\mathrm{R}(B(\mu, \varepsilon)), \varepsilon)$ is then the set of all mixed actions of the DM that are ε-close to optimal strategies against at least one footprint that is ε-close to μ.

Recall that the set-valued function $\mu \mapsto Y(\mu)$ is upper-semi-continuous. It follows that if μ' is close to μ, then for every $y' \in Y(\mu')$ there is $y \in Y(\mu)$ that is close to y'. Thus, when R is continuous a mixed action that is optimal (according to R) against y', is close to an optimal response against y. We conclude that all the mixed actions in $B(\mathrm{R}(B(\mu, \varepsilon)), \varepsilon)$ are close-to-optimal responses against μ. In the following definition we do not distinguish between continuous and noncontinuous response correspondences. In both cases, actions in $B(\mathrm{R}(B(\mu, \varepsilon)), \varepsilon)$ will be considered approximately optimal.

Definition 5. *i. Let $\varepsilon \geq 0$. The mixed action $x' \in \Delta(A)$ is* ε-optimal against *the footprint $\mu \in \mathcal{M}$ if $x' \in B(\mathrm{R}(B(\mu,\varepsilon)),\varepsilon)$. ii. Let σ be a strategy with a finite range $\widetilde{X}$ and let $\varepsilon > 0$. The strategy σ is* internal ε-regret-free w.r.t. the response correspondence R *if for every sequence of states of nature $\vec{\omega}$, every $x \in \widetilde{X}$ for which $\mathbb{P}_{\vec{\omega},\sigma}\left(\lim_{T\to\infty}|I_x^T| = \infty\right) > 0$, and every $\delta > 0$ there exists*[6] *a stage $T' = T'(\vec{\omega},\varepsilon,\delta) \in \mathbb{N}$ such that*

$$\mathbb{P}_{\vec{\omega},\sigma}\left(x \text{ is } \varepsilon-\text{optimal against } \mu(\widehat{\omega}_x^T),\ \forall T \geq T' \mid \{\lim_{T\to\infty}|I_x^T| = \infty\}\right) > 1-\delta.$$

According to Definition 5, the mixed action x is ε-optimal against $\mu(\widehat{\omega}_x^T)$ if it is in the ε-neighborhood (i.e., in $B(\cdot,\varepsilon)$) of a best response (i.e, in $\mathrm{R}(\cdot)$) to a mixed strategy equivalent to $\widehat{\omega}_x^T$. A strategy σ with a finite range is internal ε-regret-free, if for every $\vec{\omega}$ and for every mixed action x that is played infinitely often, the probability that x is ε-optimal against $\mu(\widehat{\omega}_x^T)$ for every T sufficiently large is arbitrarily close to 1.

In the sequel (Proposition 1) we show that when a strategy σ is internal ε-regret-free w.r.t. the response correspondence R^{WC}, it is in particular internal ε-regret-free (see Definition 3).

Example 2 continued. *Suppose that the empirical distribution of signals up to stage T is $[\frac{1}{2}(s_1),\frac{1}{2}(s_2)]$. A DM who believes in maximal entropy assumes that states are chosen stationarily according to the probability distribution $[\frac{1}{4}(L),\frac{1}{2}(M),\frac{1}{4}(R)]$. One can verify that if the state is chosen at every stage using this distribution, the average payoff of the DM is $\frac{9}{4}$ whatever she plays, and therefore any strategy of the DM is internal regret-free w.r.t. R^{ENT}.*

An optimist DM would rather play B: she believes that if she played T her expected average payoff would be 3, whereas if she played B her expected average payoff would be 3.5. Therefore the stationary strategy B is internal regret-free w.r.t. R^{BC}.

The next proposition links the notion of internal ε-regret-free strategies w.r.t. a response correspondence with the notion of internal ε-regret-free strategies presented in Definition 3.

Proposition 1. *Let $\varepsilon > 0$. If the strategy σ with finite range is internal ε-regret-free w.r.t. the response correspondence R^{WC}, then it is internal $L\varepsilon$-regret-free where $L > 0$ depends on the payoff function u and the monitoring structure s, and not, e.g., on σ or on ε.*

Remark 10. [About external no regret with imperfect monitoring] For the sake of completeness we briefly discuss the weaker notion of external regret. Definition 5 requires that, conditional on the event $\{\lim_{T\to\infty}|I_x^T| = \infty\}$, with high probability, in the majority of stages up to stage T, the action played is ε-optimal against the average empirical frequency of states. In contrast,

[6] Because the range of σ is finite, we can omit the dependency of T' on x.

external regret refers to the entire history and does not classify stages according to the mixed action that has been played by the DM. One could define a strategy with a finite range σ to be *external ε-regret-free w.r.t. the response correspondence* R if for any sequence of states of nature $\vec{\omega}$, and for every $\delta > 0$ there exists a stage $T' = T'(\vec{\omega}, \varepsilon, \delta) \in \mathbb{N}$ such that

$$\mathbb{P}_{\vec{\omega},\sigma}\left(\overline{a}^T \text{ is } \varepsilon\text{−optimal against } \mu(\widehat{\omega}^T),\ \forall T \geq T'\right) > 1 - \delta,$$

where $\overline{a}^T$ is the empirical frequency of actions up to stage T. One has to note that as far as R^{WC} is concerned, Proposition 1 claims that internal no regret implies external no regret. This implication does not hold for a general response function.

The following theorem refers to response correspondences and, in light of Proposition 1, generalizes Theorem 2.

Theorem 3. *For every response correspondence* R *and every $\varepsilon > 0$ there exists a strategy with a finite range that is internal ε-regret-free w.r.t.* R*.*

3 Proofs

3.1 Proof of Proposition 1

Because Y is a Lipschitz continuous correspondence, there exists $K > 0$ such that for every $\mu, \mu' \in \mathcal{M}$ and every $y \in Y(\mu)$, there is $y' \in Y(\mu')$ that satisfies $d(y, y') \leq Kd(\mu, \mu')$. In particular,

$$d(\mu, \mu') < \varepsilon \quad \Longrightarrow \quad d(Y(\mu), Y(\mu')) \leq K\varepsilon. \tag{4}$$

Let σ be an internal ε-regret-free strategy with range $\widetilde{X}$ w.r.t. the response correspondence R^{WC} (according to Definition 5). Fix a sequence $\vec{\omega}$ of states, a mixed action $x \in \widetilde{X}$ for which $\mathbb{P}_{\vec{\omega},\sigma}(\lim_{T\to\infty} |I_x^T| = \infty) > 0$, a mixed action $x' \in X$, and $\delta > 0$.

Let $T' = T'(\vec{\omega}, \varepsilon, \delta)$ be the stage given in Definition 5 and let $T \geq T'$ be arbitrary. Because σ is internal ε-regret-free w.r.t. R^{WC}, with $\mathbb{P}_{\vec{\omega},\sigma}$-probability at least $1 - \delta$, x is ε-optimal against $\mu(\widehat{\omega}_x^T)$, so that $x \in B(\mathrm{R}^{WC}(B(\mu(\widehat{\omega}_x^T), \varepsilon)), \varepsilon)$. That is, the distance between x and some $\widehat{x} \in \mathrm{R}^{WC}(B(\mu(\widehat{\omega}_x^T), \varepsilon))$ is at most ε. It follows that there exists $\widehat{\mu}$ such that $\widehat{\mu} \in B(\mu(\widehat{\omega}_x^T), \varepsilon)$ and $\widehat{x} \in \mathrm{R}^{WC}(\widehat{\mu})$. Because (i) $d(\widehat{\mu}, \mu(\widehat{\omega}_x^T)) \leq \varepsilon$, (ii) $d(x, \widehat{x}) \leq \varepsilon$, and (iii) $\widehat{x} \in \mathrm{R}^{WC}(\widehat{\mu})$, it follows

that with probability at least $1-2\delta$,

$$\begin{aligned}
\min_{y\in Y(\mu(\widehat{\omega}_x^T))} u(x,y) &\geq \min_{y\in Y(\widehat{\mu})} u(x,y) - K\varepsilon \quad \text{(follows from (i) and (4))} \\
&\geq \min_{y\in Y(\widehat{\mu})} u(\widehat{x},y) - (K+1)\varepsilon \quad \text{(follows from (ii))} \\
&\geq \min_{y\in Y(\widehat{\mu})} u(x',y) - (K+1)\varepsilon \quad \text{(follows from (iii))} \\
&\geq \min_{y\in Y(\mu(\widehat{\omega}_x^T))} u(x',y) - (2K+1)\varepsilon. \quad \text{(follows from (i) and (4))}
\end{aligned}$$

The result follows by Remark 7. ∎

3.2 Proof of Theorem 3

3.2.1 Random vector-payoff games – background

Blackwell's approachability theory (Blackwell, 1956) is a useful tool in the study of regret-free strategies. Luce and Raifa (1958) cite Blackwell's[7] proof of Hannan's (1957) no regret theorem, which uses this theory. In fact, the approachability technique suggests an algorithm that plays a regret-free strategy. To make the presentation complete, we briefly review the definitions and results that we need for the proof.

A *two-player repeated game with vector payoffs* is given by an $n \times m$ matrix Z, whose entries are $\mathbb{R}^d$-valued random variables,[8] such that for every $i \in I := \{1,\ldots,n\}$ and every $j \in J := \{1,\ldots,m\}$, the random variable $Z(i,j)$ that corresponds to the entry (i,j) has mean $z(i,j) \in \mathbb{R}^d$. At every stage $t \in \mathbb{N}$ the two players, independently and simultaneously, choose actions $i^t \in I$ and $j^t \in J$, and Player 1 obtains an $\mathbb{R}^d$-dimensional payoff $Z^t = (Z_k^t)_{k=1}^d$, which is randomly chosen according to the distribution of $Z(i^t,j^t)$. Player 1 is not informed of the action j^t that Player 2 chose, but only of the realization Z^t. Strategies of the players 1 and 2 are denoted by σ and τ, respectively. Each pair (σ,τ) of strategies naturally defines a probability distribution $\mathbb{P}_{\sigma,\tau}$ over the set of plays, endowed with the σ-algebra spanned by all cylinder sets.

We say that a subset $C \subseteq \mathbb{R}^d$ is *approachable* by Player 1 if he has a strategy σ such that, with $\mathbb{P}_{\sigma,\tau}$-probability 1, the distance between C and the average payoff $\frac{1}{T}\sum_{t=1}^T Z^t$ up to stage T goes to 0 as T goes to infinity, regardless of the strategy τ of Player 2.

To prove the existence of a regret-free strategy, one usually defines a proper two-player repeated game with vector payoffs, where Player 1 is associated with the DM and Player 2 with nature. The payoffs in the game measure the discrepancy between the actual payoff of

[7] Luce and Raifa (1958) refer to Blackwell's invited address to the Institute of Mathematical Statistics, Seattle, August 1956, entitled "Controlled random walks".

[8] In most applications of Blackwell's theory in game theory, the payoffs are uniquely determined by the actions of the players. In our proof, as in Blackwell's original paper, the payoffs are random variables.

the DM and her payoff had she played alternative actions. One then shows that a properly defined set C is approachable by Player 1, and that any strategy of Player 1 that approaches C is regret-free.

In the study of internal regret-free strategies, for each mixed action x that the DM may play, we need to consider the stages in which x was played separately from other stages. Therefore, the vector that should converge to C is not the regular average payoff, but a skewed average payoff, that takes into account the stages in which each mixed action was played and the number of these stages. We thus use Lehrer's (2002) generalization of Blackwell's (1956) approachability theory, which studies repeated games with vector payoffs that incorporate activeness functions. The activeness functions determine the way Player 1 views the average payoff vector at any stage. A *two-player repeated game with vector payoffs and activeness functions* is a two-player repeated game with vector payoffs, supplemented with an infinite sequence of functions $(\chi^t)_{t\in\mathbb{N}}$, where for every $t \in \mathbb{N}$, the function $\chi^t = (\chi^t_k)_{k=1}^d$ maps histories of length t to vectors in $\{0,1\}^d$; for every history $h^t = (i^1, j^1, i^2, j^2, ..., i^t, j^t)$ of length t, $\chi^t(h^t)$ determines which coordinates of the payoff vector at stage t contribute to the average payoff vector. In our construction, $\chi^t(h^t)$ depends only on i^t.

Denote $\overline{\chi}^T = \sum_{t=1}^T \chi^t$. This is the total number of times up to stage T in which each coordinate was active and it depends on the play up to, and including, stage T. Define the skewed average payoff[9] vector $\widehat{Z}^T$ by $\widehat{Z}^T := \frac{1}{\overline{\chi}^T} \cdot \sum_{t=1}^T \chi^t \cdot Z^t$. That is,

$$\widehat{Z}^T_k = \frac{\sum_{\{t\leq T\colon \chi^t_k(h^t)=1\}} Z^t_k}{|\{t \leq T\colon \chi^t_k(h^t) = 1\}|}$$

for each $k = 1, 2, \ldots, d$. Thus, a coordinate is taken into account only in stages in which it contributes to the average payoff.

Lehrer (2002) studies repeated games with vector payoffs and activeness functions in which χ^t depends on past play and *not* on the play at stage t. As mentioned before, for our purposes we need χ^t to depend on i^t and j^t. The generalization of Lehrer's result to this setup is presented in Appendix A.

3.2.2 The proof

The outline of the proof is as follows. After some preparations (Step 1) we define an auxiliary game with vector payoffs and activeness functions (Step 2). We define a target set C, which is close in spirit to the one defined by Blackwell in his alternative proof of Hannan's (1957)

[9]For any two vectors $a, b \in \mathbb{R}^d$ we denote by $a \cdot b$ the coordinate-wise product: $(a \cdot b)_k := a_k b_k$ for every k. Similarly, $\frac{a}{b}$ denotes the coordinate-wise quotient of a and b: $(\frac{a}{b})_k := \frac{a_k}{b_k}$ for every k.

result. We then study some properties of this game (Step 3), and state that the target set C is approachable by Player 1 in the game with activeness functions. The proof of this latter claim appears in Appendix A. We conclude by showing that every strategy that approaches C in the auxiliary game is internal ε-regret-free in the original decision problem (Step 5). We then point out that this strategy is computable.

Step 1: Preparations.

Let R be a response correspondence. Fix $\varepsilon > 0$, set $\eta_0 := \frac{\varepsilon^2}{3}$, and choose an η_0-discretization $Y^* = \{y_1, \ldots, y_L\}$ of the compact set $Y = \Delta(\Omega)$; for every $y \in Y$ there is $\ell \in \{1, 2, \ldots, L\}$ such that $d(y, y_\ell) \leq \eta_0$. For each $\ell \in \{1, 2, \ldots, L\}$ denote $Y_\ell := B(y_\ell, \eta_0) \cap Y$. The set Y_ℓ is a convex polytope, and moreover, $\cup_{\ell=1}^{L} Y_\ell = Y$. The collection $\{Y_\ell, 1 \leq \ell \leq L\}$ is called an η_0-covering in the learning literature.

For each $\ell \in \{1, 2, \ldots, L\}$ denote by $\mu_\ell := \mu(y_\ell) = (\mathfrak{s}(a, y_\ell))_{a \in A}$ the footprint induced by y_ℓ, and choose $\widetilde{x}_\ell \in \mathrm{R}(\mu_\ell)$. This is an optimal response of the DM (with respect to R) when the signals she observes are indistinguishable from the signals generated when nature uses the stationary strategy y_ℓ.

Finally, choose distinct mixed actions $\widetilde{X} = \{x_1, x_2, \ldots, x_L\}$ in X that satisfy the following two properties for each $\ell \in \{1, 2, \ldots, L\}$:

- x_ℓ has a full support and $x_\ell(a) \geq \varepsilon$ for every action $a \in A$; and
- $d(x_\ell, \widetilde{x}_\ell) \leq \varepsilon$.

Step 2: Defining an auxiliary game with random vector payoffs and activeness functions.

Consider an auxiliary two-player repeated game with vector payoffs and activeness functions, where at every stage Player 1 (who represents the DM) chooses a mixed action $x_\ell \in \widetilde{X}$ and Player 2 (who represents nature) chooses a state from Ω. The payoff is $L|S|$-dimensional, where each coordinate corresponds to a pair (mixed action in $\widetilde{X}$, signal in S), and it measures the discrepancy between the actual signal and the theoretical distribution over signals.

The activeness function χ^t is deterministic and depends only on x^t, Player 1's choice at stage t. For every $t \in \mathbb{N}$, $\chi^t_{\ell,s'}(h^t) := 1$ if $x^t = x_\ell$ and is equal to 0 otherwise.

We end the definition of the auxiliary game by defining the payoff function. For every mixed action $x_\ell \in \widetilde{X}$ and every state $\omega \in \Omega$ we thus define the random payoff $Z(x_\ell, \omega) = (Z_{\ell',s'}(x_\ell, \omega))^{s' \in S}_{\ell'=1,\ldots,L} \in \mathbb{R}^{L|S|}$. Because the payoff in the auxiliary game is observed by Player 1, and the auxiliary game should reflect the original decision problem, the coordinate $Z_{\ell',s'}(x_\ell, \omega)$ should depend on the information available to the DM in the original decision problem, namely, on his mixed action x_ℓ and on the observed signal s, and it should be independent of the state

ω. Let

$$Z_{\ell',s'}(x_\ell,\omega)=\begin{cases}0 & \ell'\neq\ell,\\ \mathbf{1}_{\{s'=s\}}-\mathfrak{s}(a(s),y_\ell)(s') & \ell'=\ell.\end{cases}$$

Note that the probability to observe the signal s when Player 1 plays x_ℓ is $\mathfrak{s}(x_\ell,\omega)(s)$.

In other words, the vector payoff is $L|S|$-dimensional, and as in Blackwell's original setup, it is random. The coordinates corresponding to mixed actions different than the one that was actually chosen by the DM are irrelevant and set to 0. This vector measures the difference between the realized signal (when interpreted as a probability distribution in $\Delta(S)$, namely, the Dirac measure concentrated on (ℓ,s)) and the distribution of signals. We emphasize that (a) the mixed action x_ℓ played by Player 1 determines the activeness function (all coordinates $\{(\ell,s'),s'\in S\}$, are active); (b) the observed signal s (and the action of Player 1, which is included in the signal) determines the stage payoff; and (c) both x_ℓ and s determine the distribution of the random payoff. Note that because each x_ℓ has full support, if the players were playing repeatedly the pair (x_ℓ,y), the footprint that y leaves could be reconstructed from the long-run average of the stage payoff vector.

The motivation behind the definition of the payoff function Z is the following. Recall that x_ℓ is almost optimal against y_ℓ, so that when the DM plays x_ℓ, from the DM's perspective y_ℓ induces a "good" distribution over states. Because $\mathfrak{s}(a(s),y_\ell)(s')\neq 0$ if and only if $a(s')=a(s)$, the mean of $Z_{\ell,s'}(x_\ell,\omega)$, denoted $z_{\ell,s'}(x_\ell,\omega)$, satisfies

$$\begin{aligned}z_{\ell,s'}(x_\ell,\omega) &= \mathfrak{s}(x_\ell,\omega)(s')-\sum_{s\in S}\mathfrak{s}(x_\ell,\omega)(s)\cdot\mathfrak{s}(a(s),y_\ell)(s')\\ &= \mathfrak{s}(x_\ell,\omega)(s')-\mathfrak{s}(x_\ell,y_\ell)(s').\end{aligned}\tag{5}$$

For $\ell'\neq\ell$, the mean $z_{\ell',s'}(x_\ell,\omega)$ of $Z_{\ell',s'}(x_\ell,\omega)$ is 0. If $y\in\Delta(\Omega)$ is a distribution over states that satisfies $\sum_{\omega\in\Omega}y(\omega)z_{\ell,s'}(x_\ell,\omega)=0$ for every $s'\in S$, then

$$\mathfrak{s}(x_\ell,y)(s'):=\sum_{\omega\in\Omega}y(\omega)\mathfrak{s}(x_\ell,\omega)(s')=\mathfrak{s}(x_\ell,y_\ell)(s').$$

Therefore, y and y_ℓ are indistinguishable. In particular, x_ℓ is almost optimal against y. Similarly, because $x_\ell(a)\geq\varepsilon$ for every $a\in A$, when $y\in\Delta(\Omega)$ is a distribution that satisfies $\|z_{\ell,s'}(x_\ell,y)\|_\infty\leq\eta$, the mixed actions y and y_ℓ of Player 2 have close footprints: $d(\mu(y),\mu(y_\ell))\leq\frac{\eta}{\varepsilon}$. As we will see below, a strategy is internal ε-regret-free once it guarantees that the average payoff remains near $\vec{0}$.

The auxiliary game we have defined is equivalent to the original sequential decision problem in the following sense. Player 1 (resp. Player 2) in the auxiliary repeated game corresponds to the DM (resp. nature) in the original sequential decision problem, and the random payoff in

the auxiliary repeated game stands in 1-1 relation with the signal that the DM observes in the original sequential decision problem. Every strategy σ of Player 1 in the auxiliary game is a strategy for the DM with range $\widetilde{X}$ in the original sequential decision problem and vice versa. This is so because the information that Player 1 has in the auxiliary repeated game at every stage t is the sequence of his past choices, as well as the sequence of past vector payoffs, which is determined also by the signals that he observes in the original sequential decision problem.

Step 3: Properties of the average payoff in the auxiliary game with vector payoffs.

Recall that Z^t is the payoff in stage t of the auxiliary game, $I^T_{x_\ell}$ is the set of all stages up to stage T in which Player 1 plays the mixed action x_ℓ, and $\widehat{\omega}^T_{x_\ell} = \frac{1}{|I^T_{x_\ell}|}\sum_{t\in I^T_{x_\ell}}[1(\omega^t)]$ is the average play of Player 2 in those stages. Throughout this step we fix a strategy σ of Player 1 and a sequence of states $\vec{\omega} = (\omega^t)_{t\in\mathbb{N}}$ (a pure Markovian strategy of Player 2), and we consider the probability distribution $\mathbb{P}_{\vec{\omega},\sigma}$ on the space of plays. Denote by $I^*_{x_\ell} := \{\lim_{T\to\infty}|I^T_{x_\ell}| = \infty\}$ the event that x_ℓ is played infinitely often and define the random variable L' by $L' := \{\ell \in L\colon x_\ell \text{ is played infinitely often}\}$. Let $\widehat{Z}^T_\ell$ be the average vector payoff in the auxiliary game in all stages in $I^T_{x_\ell}$:

$$\widehat{Z}^T_\ell := \frac{1}{|I^T_{x_\ell}|}\sum_{t\in I^T_{x_\ell}} Z^t.$$

Note that $\mathbb{E}[Z^t \mid h^t] = z(x^t,\omega^t)$. By Lemma 4 in Appendix B, for every $\ell \in L$, on the event $I^*_{x_\ell}$ we have $\lim_{T\to\infty}|\widehat{Z}^T_\ell - z(x_\ell,\widehat{\omega}^T_{x_\ell}))| = 0$ almost surely. Because L is finite, this implies that there is a real-valued function $(\vec{\omega},\eta,T)\mapsto\delta_1(\vec{\omega},\eta,T)$ satisfying $\lim_{T\to\infty}\delta_1(\vec{\omega},\eta,T) = 0$ for every $\eta > 0$, such that for every $T\in\mathbb{N}$, every $\eta > 0$, and every $\ell \in L$,

$$\mathbb{P}_{\vec{\omega},\sigma}\big(\{d(\widehat{Z}^t_\ell, z(x_\ell,\widehat{\omega}^t_{x_\ell})) > \eta, \text{ for some } t \geq T\}\cap I^*_{x_\ell}\big) \leq \delta_1(\vec{\omega},\eta,T). \tag{6}$$

Let $C := B(\vec{0},\eta_0)$ be the ℓ_∞ ball with radius η_0 around $\vec{0}$, and let $\Pi(\widehat{Z}^T)$ denote the closest point to $\widehat{Z}^T$ in C. In some well-known applications of the approachability theorem, such as Aumann and Maschler (1995), Foster (1999), and Hart and Mas-Colell (2000), the set C is defined to be a whole (either the non-negative or the negative) orthant or a translation of it. In our game the sum of the coordinates of the skewed average payoff $\overline{Z}^t$ is 0, and the only point in the non-negative orthant whose coordinates sum up to 0 is $\vec{0}$. We therefore define C to be a ball around $\vec{0}$.

Claim 1. *The set C is approachable by Player 1 in the auxiliary game. That is, there is a strategy σ of Player 1 that guarantees that with $\mathbb{P}_{\sigma,\tau}$-probability 1 one has $\lim_{T\to\infty} d(\widehat{Z}^T_{\ell,s}, \Pi(\widehat{Z}^T)_{\ell,s}) = 0$ on the event $\{l \in L'\} = \{\lim_{T\to\infty}\overline{\chi}^T_{\ell,s}(h^T) = \infty\}$, for every $(\ell,s)\in\{1,\cdots,L\}\times S$ and every strategy τ of Player 2.*

The proof of this claim appears in Appendix A.

Step 4: σ is an ε-internal regret-free strategy w.r.t. R.

Fix a sequence of states of nature $\vec{\omega}$. Recall that L' is the random set of indices in L that satisfy that x_ℓ is played infinitely often. Denote by $\widehat{d}(\widehat{Z}^t, C)$ the distance between $\widehat{Z}^t$ and C, restricted to the coordinates (ℓ, s) where $\ell \in L'$. Because the strategy σ approaches C, we obtain that for every $\eta > 0$ there is a real-valued function $\delta_2(\vec{\omega}, \eta, T)$ that decreases to 0 as T increases and satisfies

$$\mathbb{P}_{\vec{\omega},\sigma}\left(\{\widehat{d}(\widehat{Z}^t, C) > \eta, \text{ for some } t \geq T\} \cap I^*_{x_\ell}\right) \leq \delta_2(\vec{\omega}, \eta, T). \tag{7}$$

Fix $\ell \in L'$. If $\widehat{d}(\widehat{Z}^t, C) < \eta_0$, then by the definition of C it follows that $\|\widehat{Z}^t_{\ell,s}\|_\infty < 2\eta_0$. Eqs. (6) and (7) with $\eta = \eta_0$ imply that for every $T \in \mathbb{N}$,

$$\mathbb{P}_{\vec{\omega},\sigma}\left(\{\|z(x_\ell, \widehat{\omega}^t_{x_\ell})\|_\infty > 3\eta_0 \text{ for some } t \geq T\} \cap \ I^*_{x_\ell}\right) \leq \delta_1(\vec{\omega}, \eta_0, T) + \delta_2(\vec{\omega}, \eta_0, T).$$

Let T be a stage at which Player 1 choose the mixed action x_ℓ and the realized action is a. Then for every $t \geq T$ and every signal s' such that $a(s') = a$ we have $z_{\ell,s'}(x_\ell, \widehat{\omega}^t_{x_\ell}) = x_\ell(a(s'))\Big(\mathfrak{s}(a(s'), \widehat{\omega}^t_{x_\ell}) - \mathfrak{s}(a(s'), y_\ell)\Big)$. It follows that if for each action a of the DM there is a stage before T at which the DM chose the mixed action x_ℓ and the realized action was a, then $\|z(x_\ell, \widehat{\omega}^t_{x_\ell})\|_\infty \leq 3\eta_0$ implies $d(\mu(\widehat{\omega}^t_{x_\ell}), \mu_\ell) \leq \frac{3\eta_0}{\varepsilon} = \varepsilon$, provided that $t \geq T$. Because $d(x_\ell, \widetilde{x}_\ell) < \varepsilon$ implies that $x_\ell \in B(\mathrm{R}(B(\mu_\ell, 0)), \varepsilon)$, in such a case we have $x_\ell \in B(\mathrm{R}(B(\mu(\widehat{\omega}^t_{x_\ell}), \varepsilon)), \varepsilon)$.

Because x_ℓ has full support, at every stage each action of the DM is chosen with a positive probability, bounded away from zero. Hence for every $\eta > 0$ there is $T \in \mathbb{N}$ such that

$$\mathbb{P}_{\vec{\omega},\sigma}\left(\{x_\ell \in B(\mathrm{R}(B(\mu(\widehat{\omega}^t_{x_\ell}), \varepsilon)), \varepsilon) \text{ for some } t \geq T\} \cap I^*_{x_\ell}\right) \leq \delta_1(\vec{\omega}, \eta_0, T) + \delta_2(\vec{\omega}, \eta_0, T) + \eta. \tag{8}$$

Because this is true for any $\ell \in L'$, it follows that σ is indeed ε-internal regret-free w.r.t. R. ∎

Remark 11. In Step 4 of the proof of Theorem 3 we fixed a sequence $\vec{\omega}$ and showed that a properly defined strategy of the DM is ε-regret-free. The proof hinged on the fact that for each $t \in \mathbb{N}$, the state ω^t at stage t is conditionally independent of the DM's choice at that stage given past play. Thus, the sequence $\vec{\omega}$ could be chosen ahead of time or according to a certain stochastic process as the game unfolds. Moreover, if the sequence of states is randomly chosen, the proof shows that for *any* realized sequence $\vec{\omega}$ (rather than almost surely) the strategy of Player 1 is ε-regret-free.

References

[1] Aumann, R. J. and M. Maschler (1995) "Repeated Games with Incomplete Information," MIT Press.

[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] Cesa-Bianchi, N. and G. Lugosi (2006) *Prediction, Learning, and Games*, Cambridge University Press.

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

[6] Foster, D.P. (1999) "A proof of calibration via Blackwell's approachability theorem," *Games and economic behavior*, **29**, 73-78.

[7] Foster, D.P. and A. Rakhlin (2012) "No Internal Regret via Neighborhood Watch," *Journal of Machine Learning Research - Proceedings Track (AISTATS)*, **22**, 382-390.

[8] Foster, D.P. and R.V. Vohra (1997) "Calibrated Learning and Correlated Equilibrium," *Games and Economics Behavior*, **21**, 40-55.

[9] Foster, D.P. and R.V. Vohra (1998) "Asymptotic Calibration," *Biometrika*, **85**, 379-390.

[10] Foster, D.P. and R.V. Vohra (1999) "Regret in the On-Line Desicion Problem," *Games and Economics Behavior*, **29**, 7-36.

[11] Fudenberg D. and D.K. Levine (1999) "Conditional Universal Consistency," *Games and Economics Behavior*, **29**, 104-130.

[12] Hannan, J. (1957) "Approximation to Bayes Risk in Repeated Play," *Contributions to The Theory of Games*, **3**, 97-139.

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

[14] Hart, S. and A. Mas-Colell (2001) "A General Class of Adaptive Strategies," *Journal of Economic Theory*, **98**, 26-54.

[15] Hazan, E. and S.M. Kakade (2012) "(weak) Calibration is Computationally Hard," in Conference on Learning Theory (COLT) 2012.

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

[17] Lehrer, E. and E. Solan (2007) "Learning to Play Partially Specified Equilibrium," mimeo.

[18] Luce, D.R. and H. Raiffa (1958) *Games and Decisions*, John Wiley, N.Y.

[19] Lugosi, G., S. Mannor, and G. Stoltz (2008) "Strategies for Prediction Under Imperfect Monitoring," *Mathematics of Operations Research*, **33**, 513-528.

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

[21] Perchet, V. (2009) "Calibration and Internal No-Regret with Random Signals," ALT2009, 68-82.

[22] Perchet, V. (2011) "Internal Regret with Partial Monitoring Calibration-Based Optimal Algorithms," *Journal of Machine Learning Research*, **12**, 1893-1921.

[23] Perchet, V. (2013) "Approachability, Regret and Calibration; Implications and Equivalences," http://arxiv.org/abs/1301.2663.

[24] Piccolboni, A. and C. Schindelhauer (2001) "Discrete Prediction Games with Arbitrary Feedback and Loss," in COLT 2001, Annual Conference on Computational Learning Theory #14, *Lecture notes in computer science*, **2111**, 208-223.

[25] Rockafeller, R.T. and R.J.B. Wets (2009) *Variational Analysis*. Springer.

[26] Rustichini, A. (1999) "Minimizing Regret: the General Case," *Games and Economic Behavior*, **29**, 224-243.

[27] Stoltz G. and G. Lugosi (2005) "Internal Regret in On-Line Portfolio Selection," *Machine Learning*, **59**, 125-159.

Appendices

A Approachability with Activeness Functions

Lehrer (2002) studies repeated games with vector payoffs and activeness functions that do not depend on the current action of Player 1 (the DM). In the auxiliary game that is defined in the proof of Theorem 3 the activeness functions depend on Player 1's current action. In this section we extend Lehrer's result to this more general setup and prove Claim 1. Thus, we construct a strategy σ for Player 1 in the auxiliary repeated game with vector payoffs that approaches the set C. To simplify the section we do not prove the result in the most general setup, but rather in the formulation discussed here.

Recall that in the auxiliary game, the activeness function for stage t is given by $\chi^t_{\ell,s}(h^t) = 1$ if at stage t Player 1 played x_ℓ and $\chi^t_{\ell,s}(h^t) = 0$, otherwise.

Denote $\overline{\chi}^T = \sum_{t=1}^T \chi^t$. This is the total number of times up to stage T in which each coordinate was active, and it depends on the play up to, and including, stage T. Denote the total skewed payoff up to stage T by $\overline{Z}^T = \sum_{t=1}^T \chi^t Z^t$ (recall that this is a coordinate-wise product and not the inner product), and the average payoff by $\widehat{Z}^T = \frac{\overline{Z}^T}{\overline{\chi}^T}$.

Recall that $C \subset \mathbb{R}^{L\times S}$ is the ℓ_∞ ball with radius η_0 around $\vec{0}$, and $\Pi(x)$ denotes the closest point in C to x, for every $x \in \mathbb{R}^{L\times|S|}$. The (ℓ, s)-coordinate of the vector $\Pi(x)$ will be denoted by $\Pi(x)_{\ell,s}$. Our goal is to show that Player 1 has a strategy σ that guarantees that $\limsup_{T\to\infty} |\widehat{Z}^T_{\ell,s}| \leq \eta_0$, $\mathbb{P}_{\sigma,\tau}$ a.s. for every $1 \leq \ell \leq L$ and $s \in S$ such that $\lim_{T\to\infty} \overline{\chi}^T_{\ell,s} = \infty$, and every strategy τ of Player 2. This result will follow from the following lemma:

Lemma 1 (Lehrer, 2002, Proposition 1)**.** *Let $(h^j)_{j\in\mathbb{N}}$ be a sequence of uniformly bounded functions and let (σ, τ) be a pair of strategies. Define $f^n := \frac{1}{n}\sum_{j=1}^n h^j$, for every $n \in \mathbb{N}$. If $\sum_{n=1}^\infty \frac{\|f^n\|^2}{n} < \infty$, $\mathbb{P}_{\sigma,\tau}$-a.s., then the sequence $(f^n)_{n\in\mathbb{N}}$ converges to 0, $\mathbb{P}_{\sigma,\tau}$-a.s.*

Let τ^n_ℓ be the n-th time in which Player 1 played the mixed action x_ℓ. If this event occurs less than n times, we set $\tau^n_{\ell,s} = 0$. In our setup,

$$f^n_{\ell,s} = \begin{cases} \widehat{Z}^{\tau^n_\ell}_{\ell,s} - \Pi(\widehat{Z}^{\tau^n_\ell})_{\ell,s}, & \tau^n_{\ell,s} > 0, \\ \frac{n-1}{n} f^{n-1}_{\ell,s}, & \tau^n_{\ell,s} = 0. \end{cases}$$

In particular, $\overline{\chi}^{\tau^n_\ell}_{\ell,s} = n$ for every $n \in \mathbb{N}$, $\ell \in L$ and $s \in S$, provided that $\tau^n_{\ell,s} > 0$. To prove Claim 1 we will construct a strategy σ for Player 1 such that (a) $\sum_{n=1}^\infty \frac{\|f^n_{\ell,s}\|^2}{n} < \infty$ with $\mathbb{P}_{\sigma,\tau}$-probability 1 for every strategy τ, every $\ell \in L$, and every $s \in S$, and (b) there is a sequence of uniformly bounded functions $(h^j_{\ell,s})_{j\in\mathbb{N}}$ that satisfies $f^n_{\ell,s} = \frac{1}{n}\sum_{j=1}^n h^j_{\ell,s}$. We start

by constructing the strategy σ (Step 1), then prove that (a) is satisfied (Step 2), and finally prove that (b) holds (Step 3).

Step 1: Constructing a strategy σ for Player 1.

Denote $g^{T-1} := \left(\frac{1}{\overline{\chi}^{T-1}+1}\right)\left(\widehat{Z}^{T-1} - \Pi(\widehat{Z}^{T-1})\right)$. This is a random variable that depends on the history up to (and including) stage $T-1$.

Lemma 2. *For every $T \in \mathbb{N}$ there is an $|L \times S|$-dimensional vector b^{T-1} such that $\Pi(b^{T-1}) = \Pi(\widehat{Z}^{T-1})$ and $b^{T-1} - \Pi(b^{T-1}) = g^{T-1}$.*

Proof. Note that

$$\begin{aligned} g^{T-1}_{\ell,s} = 0 \quad &\iff \quad \Pi(\widehat{Z}^{T-1})_{\ell,s} = \widehat{Z}^{T-1}_{\ell,s}, \\ g^{T-1}_{\ell,s} = \left(\frac{1}{\overline{\chi}^{T-1}+1}\right)_{\ell,s}\left(\widehat{Z}^{T-1}_{\ell,s} - \eta_0\right) > 0 \quad &\iff \quad \widehat{Z}^{T-1}_{\ell,s} > \eta_0, \\ g^{T-1}_{\ell,s} = \left(\frac{1}{\overline{\chi}^{T-1}+1}\right)_{\ell,s}\left(\widehat{Z}^{T-1}_{\ell,s} + \eta_0\right) < 0 \quad &\iff \quad \widehat{Z}^{T-1}_{\ell,s} < -\eta_0. \end{aligned}$$

Define a vector $b^{T-1} \in \mathbb{R}^{L\times S}$ as follows:

$$b^{T-1}_{\ell,s} := \begin{cases} \Pi(\widehat{Z}^{T-1})_{\ell,s} & g^{T-1}_{\ell,s} = 0, \\ g^{T-1}_{\ell,s} + \eta_0 & g^{T-1}_{\ell,s} > 0, \\ g^{T-1}_{\ell,s} - \eta_0 & g^{T-1}_{\ell,s} < 0. \end{cases}$$

Then

$$\Pi(b^{T-1})_{\ell,s} := \begin{cases} \Pi(\widehat{Z}^{T-1})_{\ell,s} & g^{T-1}_{\ell,s} = 0, \\ \eta_0 & g^{T-1}_{\ell,s} > 0, \\ -\eta_0 & g^{T-1}_{\ell,s} < 0. \end{cases}$$

The reader can verify that b^{T-1} satisfies the desired properties. ∎

For every $y \in Y$ there is $\ell(y)$ such that $y \in Y_{\ell(y)}$. Thus, the $\|\cdot\|_\infty$ distance between $\mathfrak{s}(x_{\ell(y)}, y)$ and $\mathfrak{s}(x_{\ell(y)}, y_{\ell(y)})$ is smaller than η_0. By (5), we have $z(x_{\ell(y)}, y) \in C$. This implies that for every supporting hyperplane of C and every mixed action $y \in Y$, the set C and the point $z(x_{\ell(y)}, y)$ lie on the same side of the hyperplane.

Suppose that $b^{T-1} \notin C$. Consider the hyperplane perpendicular to $b^{T-1} - \Pi(b^{T-1})$ and containing the point $\Pi(b^{T-1})$, that is, the hyperplane tangent to C at the point $\Pi(b^{T-1})$. The minmax theorem guarantees that there is $p^T = (p^T_\ell)_{1\le\ell\le L} \in \Delta(\widetilde{X})$, such that for every $\omega \in \Omega$, the vector $z(x_\ell, \omega)$ and C are on the same side of this hyperplane and b^{T-1} on the other.

The strategy σ of Player 1 in the auxiliary game is to play the (history dependent) mixed action p^T when $b^{T-1} \notin C$, and to play arbitrarily when $b^{T-1} \in C$.

The random variable p_ℓ^T represents the probability that the mixed action x_ℓ is chosen by the DM at stage t, given past play. Denote by $p_{\ell,s}^T := p_\ell^T x_\ell(a(s))\mathfrak{s}(a(s),\omega^T)(s)$ the probability that x_ℓ is the mixed action being played at stage T and the realized signal is s. The following expresses the expectation of the inner product of $\langle z(x_\ell,\omega) - \Pi(b^{T-1})$ and $b^{T-1} - \Pi(b^{T-1})$, conditional on the information available to Player 1 at stage T and on the state at stage T being ω, when DM uses p^T:

$$\begin{aligned}&\mathbb{E}_{p^T}^T\Big(\Big\langle z(x_\ell,\omega) - \Pi(b^{T-1}), b^{T-1} - \Pi(b^{T-1})\Big\rangle\Big|\omega^T = \omega\Big)\\ &\quad= \sum_{\ell,s} p_{\ell,s}^T\Big(z(x_\ell,\omega) - \Pi(b^{T-1})\Big)_{\ell,s}\Big(b^{T-1} - \Pi(b^{T-1})\Big)_{\ell,s} \le 0, \quad \forall\omega\in\Omega. \end{aligned} \tag{9}$$

The inequality is due to the choice of p^T. Note that p^T depends only on the play up to (and including) stage $T-1$. In particular, it can be calculated by Player 1 at stage T. Eq. (9) states that the average (across all ℓ and s) of $\Big(z(x_\ell,\omega) - \Pi(b^{T-1})\Big)_{\ell,s}\Big(b^{T-1} - \Pi(b^{T-1})\Big)_{\ell,s}$ is nonpositive, which means that the vectors $z(x_\ell,\omega)$ and b^{T-1} are on different sides of the hyperplane tangent to C at the point $\Pi(b^{T-1})$.

Step 2: Condition (a) holds.

Note that if at stage T Player 1 plays x_ℓ, then

$$(\frac{\overline{\chi}^T - \overline{\chi}^{T-1}}{\overline{\chi}^T})_{\ell,s} = (\frac{1}{\overline{\chi}^{T-1}+1})_{\ell,s}.$$

From now on we fix a strategy τ of Player 2. For every $T\in\mathbb{N}$ let $\mathbb{E}_{\sigma,\tau}^T(\cdot)$ be the conditional expectation under (σ,τ) given the play up to stage $T-1$. The choice of p^T is made so that the following inequality, which will be used in the sequel, holds:

$$\mathbb{E}_{\sigma,\tau}^T\left(\left\langle \frac{\overline{Z}^{T-1}}{\overline{\chi}^{T-1}} - \Pi(\frac{\overline{Z}^{T-1}}{\overline{\chi}^{T-1}}), \frac{Z^T - \Pi(\frac{\overline{Z}^{T-1}}{\overline{\chi}^{T-1}})(\overline{\chi}^T - \overline{\chi}^{T-1})}{\overline{\chi}^T}\right\rangle\right) \le 0. \tag{10}$$

Indeed, denoting by $\tau^T(\omega)$ the conditional probability that $\omega^T = \omega$ given the history up to

stage $T-1$, we have

$$\mathbb{E}^T_{\sigma,\tau}\left(\left\langle \frac{\overline{Z}^{T-1}}{\overline{\chi}^{T-1}} - \Pi(\frac{\overline{Z}^{T-1}}{\overline{\chi}^{T-1}}), \frac{Z^T - \Pi(\frac{\overline{Z}^{T-1}}{\overline{\chi}^{T-1}})(\overline{\chi}^T - \overline{\chi}^{T-1})}{\overline{\chi}^T} \right\rangle\right)$$

$$= \mathbb{E}^T_{\sigma,\tau}\left(\left\langle \frac{Z^T - \Pi(\frac{\overline{Z}^{T-1}}{\overline{\chi}^{T-1}})(\overline{\chi}^T - \overline{\chi}^{T-1})}{\overline{\chi}^T}, \frac{\overline{Z}^{T-1}}{\overline{\chi}^{T-1}} - \Pi(\frac{\overline{Z}^{T-1}}{\overline{\chi}^{T-1}}) \right\rangle\right)$$

$$= \mathbb{E}^T_{\sigma,\tau}\left(\left\langle Z^T - \Pi(\frac{\overline{Z}^{T-1}}{\overline{\chi}^{T-1}}), \frac{(\overline{\chi}^T - \overline{\chi}^{T-1})}{\overline{\chi}^T}\Big(\frac{\overline{Z}^{T-1}}{\overline{\chi}^{T-1}} - \Pi(\frac{\overline{Z}^{T-1}}{\overline{\chi}^{T-1}})\Big) \right\rangle\right) \tag{11}$$

$$= \sum_{\omega\in\Omega}\tau^T(\omega)\mathbb{E}^T_{p^T}\left(\left\langle Z^T - \Pi(\frac{\overline{Z}^{T-1}}{\overline{\chi}^{T-1}}), \frac{(\overline{\chi}^T - \overline{\chi}^{T-1})}{\overline{\chi}^T}\Big(\frac{\overline{Z}^{T-1}}{\overline{\chi}^{T-1}} - \Pi(\frac{\overline{Z}^{T-1}}{\overline{\chi}^{T-1}})\Big) \right\rangle \Big| \omega^T = \omega\right) \tag{12}$$

$$= \sum_{\omega\in\Omega}\tau^T(\omega)\sum_{\ell,s}p^T_{\ell,s}\Big(z(x_\ell,\omega) - \Pi(\frac{\overline{Z}^{T-1}}{\overline{\chi}^{T-1}})\Big)_{\ell,s}\frac{1}{\overline{\chi}^{T-1}_{\ell,s}+1}\Big(\frac{\overline{Z}^{T-1}}{\overline{\chi}^{T-1}} - \Pi(\frac{\overline{Z}^{T-1}}{\overline{\chi}^{T-1}})\Big)_{\ell,s} \tag{13}$$

$$= \sum_{\omega\in\Omega}\tau^T(\omega)\sum_{\ell,s}p^T_{\ell,s}\Big(z(x_\ell,\omega) - \Pi(b^{T-1})\Big)_{\ell,s}\Big(b^{T-1} - \Pi(b^{T-1})\Big)_{\ell,s} \le 0. \tag{14}$$

Eq. (13) holds for two reasons. First, all the terms in Eq. (11) except of Z^T and $\overline{\chi}^T$ are measurable with respect to the history up to stage $T-1$. Second, $(\overline{\chi}^T - \overline{\chi}^{T-1})_{\ell,s} = 1$ only when x_ℓ is played at stage T, an event that occurs with probability p^T_ℓ. Eq. (14) is due to Lemma 2 and the inequality is due to Eq. (9).

Define

$$e^T := \frac{\overline{Z}^T}{\overline{\chi}^T} - \frac{\overline{Z}^{T-1}}{\overline{\chi}^{T-1}} \tag{15}$$

$$= \frac{\overline{Z}^{T-1}}{\overline{\chi}^T} + \frac{Z^T}{\overline{\chi}^T} - \frac{\overline{Z}^{T-1}}{\overline{\chi}^{T-1}}$$

$$= \overline{Z}^{T-1}\left(\frac{1}{\overline{\chi}^T} - \frac{1}{\overline{\chi}^{T-1}}\right) + \frac{Z^T}{\overline{\chi}^T}. \tag{16}$$

Note that either $\overline{\chi}^T_{\ell,s} = \overline{\chi}^{T-1}_{\ell,s}$ and $Z^T_{\ell,s} = 0$, or $\overline{\chi}^T_{\ell,s} = \overline{\chi}^{T-1}_{\ell,s} + 1$ and $|Z^T_{\ell,s}| \le 1$. In particular, if $\overline{\chi}^T_{\ell,s} = \overline{\chi}^{T-1}_{\ell,s}$ then $e^T_{\ell,S} = 0$, while if $\overline{\chi}^T_{\ell,s} = \overline{\chi}^{T-1}_{\ell,s} + 1$ then

$$|e^T_{\ell,s}| = \left|\overline{Z}^{T-1}_{\ell,s}\left(\frac{1}{\overline{\chi}^T_{\ell,s}} - \frac{1}{\overline{\chi}^{T-1}_{\ell,s}}\right) + \frac{Z^T_{\ell,s}}{\overline{\chi}^T_{\ell,s}}\right| \le \frac{|\overline{Z}^{T-1}_{\ell,s}|}{\overline{\chi}^{T-1}_{\ell,s}\overline{\chi}^T_{\ell,s}} + \frac{|Z^T_{\ell,s}|}{\overline{\chi}^T_{\ell,s}} \le \frac{2}{\overline{\chi}^T_{\ell,s}}. \tag{17}$$

In addition,

$$
\begin{aligned}
&\mathbb{E}^T_{\sigma,\tau}(\|f^T\|^2) \\
&= \mathbb{E}^T_{\sigma,\tau}\left(\left\|\frac{\overline{Z}^T}{\overline{\chi}^T} - \Pi(\frac{\overline{Z}^T}{\overline{\chi}^T})\right\|^2\right) \\
&\leq \mathbb{E}^T_{\sigma,\tau}\left(\left\|\frac{\overline{Z}^T}{\overline{\chi}^T} - \Pi(\frac{Z^{T-1}}{\overline{\chi}^{T-1}})\right\|^2\right) && (18) \\
&= \mathbb{E}^T_{\sigma,\tau}\left(\left\|\frac{\overline{Z}^{T-1}}{\overline{\chi}^{T-1}} + e^T - \Pi(\frac{\overline{Z}^{T-1}}{\overline{\chi}^{T-1}})\right\|^2\right) && (19) \\
&= \mathbb{E}^T_{\sigma,\tau}\left(\left\|\frac{\overline{Z}^{T-1}}{\overline{\chi}^{T-1}} - \Pi(\frac{\overline{Z}^{T-1}}{\overline{\chi}^{T-1}})\right\|^2\right) + \mathbb{E}^T_{\sigma,\tau}(\|e^T\|^2) + 2\mathbb{E}^T_{\sigma,\tau}\left(\left\langle\frac{\overline{Z}^{T-1}}{\overline{\chi}^{T-1}} - \Pi(\frac{\overline{Z}^{T-1}}{\overline{\chi}^{T-1}}), e^T\right\rangle\right) \\
&= \mathbb{E}^T_{\sigma,\tau}\left(\left\|\frac{\overline{Z}^{T-1}}{\overline{\chi}^{T-1}} - \Pi(\frac{\overline{Z}^{T-1}}{\overline{\chi}^{T-1}})\right\|^2\right) + \mathbb{E}^T_{\sigma,\tau}(\|e^T\|^2) + 2\mathbb{E}^T_{\sigma,\tau}\left(\left\langle\frac{\overline{Z}^{T-1}}{\overline{\chi}^{T-1}} - \Pi(\frac{\overline{Z}^{T-1}}{\overline{\chi}^{T-1}}), \frac{Z^T}{\overline{\chi}^T}\right\rangle\right) \\
&\quad + 2\mathbb{E}^T_{\sigma,\tau}\left(\left\langle\frac{\overline{Z}^{T-1}}{\overline{\chi}^{T-1}} - \Pi(\frac{\overline{Z}^{T-1}}{\overline{\chi}^{T-1}}), \overline{Z}^{T-1}(\frac{1}{\overline{\chi}^T} - \frac{1}{\overline{\chi}^{T-1}})\right\rangle\right) && (20) \\
&= \mathbb{E}^T_{\sigma,\tau}\left(\|f^{T-1}\|^2\right) + \mathbb{E}^T_{\sigma,\tau}(\|e^T\|^2) + 2\mathbb{E}^T_{\sigma,\tau}\left(\left\langle\frac{\overline{Z}^{T-1}}{\overline{\chi}^{T-1}} - \Pi(\frac{\overline{Z}^{T-1}}{\overline{\chi}^{T-1}}), \frac{Z^T - \Pi(\frac{\overline{Z}^{T-1}}{\overline{\chi}^{T-1}})(\overline{\chi}^T - \overline{\chi}^{T-1})}{\overline{\chi}^T}\right\rangle\right) \\
&\quad + 2\mathbb{E}^T_{\sigma,\tau}\left(\left\langle\frac{\overline{Z}^{T-1}}{\overline{\chi}^{T-1}} - \Pi(\frac{\overline{Z}^{T-1}}{\overline{\chi}^{T-1}}), \overline{Z}^{T-1}(\frac{1}{\overline{\chi}^T} - \frac{1}{\overline{\chi}^{T-1}}) + \frac{\Pi(\frac{\overline{Z}^{T-1}}{\overline{\chi}^{T-1}})(\overline{\chi}^T - \overline{\chi}^{T-1})}{\overline{\chi}^T}\right\rangle\right) && (21) \\
&\leq \mathbb{E}^T_{\sigma,\tau}\left(\|f^{T-1}\|^2\right) + \mathbb{E}^T_{\sigma,\tau}(\|e^T\|^2) \\
&\quad + 2\mathbb{E}^T_{\sigma,\tau}\left(\left\langle\frac{\overline{Z}^{T-1}}{\overline{\chi}^{T-1}} - \Pi(\frac{\overline{Z}^{T-1}}{\overline{\chi}^{T-1}}), \overline{Z}^{T-1}(\frac{1}{\overline{\chi}^T} - \frac{1}{\overline{\chi}^{T-1}}) + \frac{\Pi(\frac{\overline{Z}^{T-1}}{\overline{\chi}^{T-1}})(\overline{\chi}^T - \overline{\chi}^{T-1})}{\overline{\chi}^T}\right\rangle\right). && (22)
\end{aligned}
$$

Indeed, Eq. (18) holds because $\Pi(\frac{\overline{Z}^T}{\overline{\chi}^T})$ is the closest point in C to $\frac{\overline{Z}^T}{\overline{\chi}^T}$; Eq. (19) holds by (15); Eq. (20) holds by (16); Eq. (21) holds because we added and subtracted the same term; and Eq. (22) holds by (10).

Iterating this derivation one obtains for every stage T

$$\begin{aligned} 0 &\leq \mathbb{E}\left(\|f^T\|^2\right) && (23)\\ &\leq \mathbb{E}\left(\|f^1\|^2\right) + \sum_{t=2}^{T}\mathbb{E}(\|e^T\|^2) && (24)\\ &\quad +2\sum_{t=2}^{T}\mathbb{E}\left(\mathbb{E}^t\left(\left\langle \frac{\overline{Z}^{t-1}}{\overline{\chi}^{t-1}} - \Pi(\frac{\overline{Z}^{t-1}}{\overline{\chi}^{t-1}}), \overline{Z}^{t-1}(\frac{1}{\overline{\chi}^t} - \frac{1}{\overline{\chi}^{t-1}}) + \frac{\Pi(\frac{\overline{Z}^{t-1}}{\overline{\chi}^{t-1}})(\overline{\chi}^t - \overline{\chi}^{t-1})}{\overline{\chi}^T}\right\rangle\right)\right). \end{aligned}$$

Consider the second summand in (24). Eq. (17) and the discussion that precedes it imply that

$$\begin{aligned} \sum_{t=2}^{T}\mathbb{E}(\|e^T\|^2) &\leq \sum_{t=2}^{\infty}\mathbb{E}\left(\|e^T\|^2\right) && (25)\\ &= \sum_{t=2}^{\infty}\mathbb{E}(\sum_{\ell,s}(e^T_{\ell,s})^2)\\ &= \sum_{\ell,s}\sum_{t=2}^{\infty}\mathbb{E}((e^T_{\ell,s})^2)\\ &\leq \sum_{\ell,s}\sum_{t=2}^{\infty}\frac{4}{t^2} \leq 8L|S|. \end{aligned}$$

Eqs. (23), (25) imply that the series in the third summand of (24) is greater than $-1-8L|S|$. That is,

$$\begin{aligned} -1-8L|S| &\leq \sum_{t=2}^{T}\mathbb{E}\left(\left\langle \frac{\overline{Z}^{t-1}}{\overline{\chi}^{t-1}} - \Pi(\frac{\overline{Z}^{t-1}}{\overline{\chi}^{t-1}}), \overline{Z}^{t-1}(\frac{1}{\overline{\chi}^t} - \frac{1}{\overline{\chi}^{t-1}}) + \frac{\Pi(\frac{\overline{Z}^{t-1}}{\overline{\chi}^{t-1}})(\overline{\chi}_t - \overline{\chi}^{t-1})}{\overline{\chi}^t}\right\rangle\right)\\ &= \sum_{t=2}^{T}\mathbb{E}\left(\left\langle \frac{\overline{Z}^{t-1}}{\overline{\chi}^{t-1}} - \Pi(\frac{\overline{Z}^{t-1}}{\overline{\chi}^{t-1}}), \frac{\Pi(\frac{\overline{Z}^{t-1}}{\overline{\chi}^{t-1}})(\overline{\chi}^t - \overline{\chi}^{t-1})}{\overline{\chi}^t} - \overline{Z}^{t-1}(\frac{1}{\overline{\chi}^{t-1}} - \frac{1}{\overline{\chi}^t})\right\rangle\right)\\ &= \sum_{t=2}^{T}\mathbb{E}\left(\left\langle \frac{\overline{Z}^{t-1}}{\overline{\chi}^{t-1}} - \Pi(\frac{\overline{Z}^{t-1}}{\overline{\chi}^{t-1}}), \frac{\Pi(\frac{\overline{Z}^{t-1}}{\overline{\chi}^{t-1}})(\overline{\chi}^t - \overline{\chi}^{t-1})}{\overline{\chi}^t} - \overline{Z}^{t-1}(\frac{\overline{\chi}^t - \overline{\chi}^{t-1}}{\overline{\chi}^{t-1}\overline{\chi}^t})\right\rangle\right)\\ &= \sum_{t=2}^{T}\mathbb{E}\left(\left\langle \frac{\overline{Z}^{t-1}}{\overline{\chi}^{t-1}} - \Pi(\frac{\overline{Z}^{t-1}}{\overline{\chi}^{t-1}}), \left(\Pi(\frac{\overline{Z}^{t-1}}{\overline{\chi}^{t-1}}) - \frac{\overline{Z}^{t-1}}{\overline{\chi}^{t-1}}\right)\frac{\overline{\chi}^t - \overline{\chi}^{t-1}}{\overline{\chi}^t}\right\rangle\right). \end{aligned}$$

By letting T go to infinity we deduce that

$$\sum_{t=2}^{\infty}\mathbb{E}\left(\left\langle \frac{\overline{Z}^{t-1}}{\overline{\chi}^{t-1}}-\Pi(\frac{\overline{Z}^{t-1}}{\overline{\chi}^{t-1}}),\left(\Pi(\frac{\overline{Z}^{t-1}}{\overline{\chi}^{t-1}})-\frac{\overline{Z}^{t-1}}{\overline{\chi}^{t-1}}\right)\frac{\overline{\chi}^{t}-\overline{\chi}^{t-1}}{\overline{\chi}^{t}}\right\rangle\right)$$

is bounded, and therefore

$$\sum_{t=2}^{\infty}\left(\frac{\overline{Z}^{t-1}}{\overline{\chi}^{t-1}}-\Pi(\frac{\overline{Z}^{t-1}}{\overline{\chi}^{t-1}})\right)^{2}(\frac{\overline{\chi}^{t}-\overline{\chi}^{t-1}}{\overline{\chi}^{t}})<\infty,\quad \mathbb{P}_{\sigma,\tau}-a.s.$$

This implies that for every $\ell\in L$ and $s\in S$,

$$\sum_{t=2}^{\infty}\left(\frac{\overline{Z}^{t-1}}{\overline{\chi}^{t-1}}-\Pi(\frac{\overline{Z}^{t-1}}{\overline{\chi}^{t-1}})\right)^{2}_{\ell,s}(\frac{\overline{\chi}^{t}-\overline{\chi}^{t-1}}{\overline{\chi}^{t}})_{\ell,s}<\infty,\quad \mathbb{P}_{\sigma,\tau}-a.s. \tag{26}$$

Recall that $f^n_{\ell,s}:=\widehat{Z}^{\tau^n_\ell}_{\ell,s}-\Pi(\widehat{Z}^{\tau^n_\ell})_{\ell,s}=\frac{\overline{Z}^{\tau^n_\ell}}{\overline{\chi}^{\tau^n_\ell}}-\Pi(\frac{\overline{Z}^{\tau^n_\ell}}{\overline{\chi}^{\tau^n_\ell}})_{\ell,s}$ when $\tau^n_\ell>0$. In this case we obtain by (26)

$$\sum_{n=1}^{\infty}\frac{(f^n_{\ell,s})^2}{n}=\sum_{t=1}^{\infty}\left(\frac{\overline{Z}^{t-1}}{\overline{\chi}^{t-1}}-\Pi(\frac{\overline{Z}^{t-1}}{\overline{\chi}^{t-1}})\right)^{2}_{\ell,s}(\frac{\overline{\chi}^{t}-\overline{\chi}^{t-1}}{\overline{\chi}^{t}})_{\ell,s}<\infty,$$

and condition (a) above holds for every $\ell\in L$ and $s\in S$ for which $\lim_{n\to\infty}\tau^n_{\ell,s}=\infty$. If $\lim_{n\to\infty}\tau^n_{\ell,s}<\infty$ then there is $N\in\mathbb{N}$ that satisfies $\tau^n_{\ell,s}=0$ for every $n\geq N$, so that

$$\sum_{n=1}^{\infty}\frac{(f^n_{\ell,s})^2}{n}=\sum_{n=1}^{N-1}\frac{(f^n_{\ell,s})^2}{n}+f^{N-1}_{\ell,s}\sum_{n=N}^{\infty}\frac{(N-1)^2}{n^3}<\infty,$$

and condition (a) holds as well.

Step 3: Condition (b) holds.

By the definition of f^n, for every coordinate $1\leq\ell\leq L$ and every $s\in S$,

$$f^n_{\ell,s}=\begin{cases}(\frac{\overline{Z}^{\tau^n_\ell}}{\overline{\chi}^{\tau^n_\ell}})_{\ell,s}-\eta_0, & \text{if } (\frac{\overline{Z}^{\tau^n_\ell}}{\overline{\chi}^{\tau^n_\ell}})_{\ell,s}>\eta_0\\ (\frac{\overline{Z}^{\tau^n_\ell}}{\overline{\chi}^{\tau^n_\ell}})_{\ell,s}+\eta_0, & \text{if } (\frac{\overline{Z}^{\tau^n_\ell}}{\overline{\chi}^{\tau^n_\ell}})_{\ell,s}<-\eta_0\\ 0, & \text{otherwise.}\end{cases}$$

Note that in the summation (26), if $\overline{\chi}^t=\overline{\chi}^{t-1}$ then the t-th term vanishes. We now present $f^n_{\ell,s}$ as a sum of n functions. Set,

$$\begin{aligned}h^1_{\ell,s} &:= f^1_{\ell,s},\\ h^n_{\ell,s} &:= nf^n_{\ell,s}-(n-1)f^{n-1}_{\ell,s}=f^n_{\ell,s}+(n-1)(f^n_{\ell,s}-f^{n-1}_{\ell,s}),\quad \forall n>1.\end{aligned}\tag{27}$$

Then for every $n \in \mathbb{N}$ we have

$$f^n_{\ell,s} = \frac{h^1_{\ell,s} + h^2_{\ell,s} + \cdots + h^n_{\ell,s}}{n}.$$

It remains to show that the sequence $(h^n)_{n\in\mathbb{N}}$ is uniformly bounded. Fix $n \in \mathbb{N}$, and note that when $\tau^n_\ell > 0$,

$$\left|\frac{\overline{Z}^{\tau^n_\ell}_{\ell,s}}{n} - \frac{\overline{Z}^{\tau^{n-1}_\ell}_{\ell,s}}{n-1}\right| = \left|\frac{Z^{\tau^n_\ell}_{\ell,s} + \overline{Z}^{\tau^{n-1}_\ell}_{\ell,s}}{n} - \frac{\overline{Z}^{\tau^{n-1}_\ell}_{\ell,s}}{n-1}\right| \le \frac{|Z^{\tau^n_\ell}_{\ell,s}|}{n} + \frac{|\overline{Z}^{\tau^{n-1}_\ell}_{\ell,s}|}{n(n-1)}. \tag{28}$$

Case A: Both $|f^n_{\ell,s}|$ and $|f^{n-1}_{\ell,s}|$ are positive. Eqs. (27) and (28) imply that

$$\begin{aligned} |h^n_{\ell,s}| &= |nf^n_{\ell,s} - (n-1)f^{n-1}_{\ell,s}| \\ &\le |f^n_{\ell,s}| + (n-1)|f^n_{\ell,s} - f^{n-1}_{\ell,s}| \\ &\le 1 + |Z^{\tau^n_\ell}_{\ell,s}| + \frac{|\overline{Z}^{\tau^n_\ell}_{\ell,s}|}{n} + 2\eta_0 \le 3 + 2\eta_0, \end{aligned}$$

Case B: $|f^n_{\ell,s}| > 0$ and $|f^{n-1}_{\ell,s}| = 0$. In this case $|\frac{\overline{Z}^{\tau^n_\ell}_{\ell,s}}{n}| > \eta_0$ while $|\frac{\overline{Z}^{\tau^n_\ell}_{\ell,s}}{n-1}| \le \eta_0$. Therefore,

$$|h^n_{\ell,s}| = |nf^n_{\ell,s}| = |\overline{Z}^{\tau^n_\ell}_{\ell,s}| - n\eta_0 \le |\overline{Z}^{\tau^n_\ell}_{\ell,s}| - |\overline{Z}^{\tau^{n-1}_\ell}_{\ell,s}| - \eta_0 \le |\overline{Z}^{\tau^n_\ell}_{\ell,s} - \overline{Z}^{\tau^{n-1}_\ell}_{\ell,s}| - \eta_0 = |Z^{\tau^n_\ell}_{\ell,s}| \le 1.$$

Case C: $|f^n_{\ell,s}| = 0$ and $|f^{n-1}_{\ell,s}| > 0$. Similarly to the previous case,

$$|h^n_{\ell,s}| = |(n-1)f^{n-1}_{\ell,s}| = |\overline{Z}^{\tau^{n-1}_\ell}_{\ell,s}| - (n-1)\eta_0 \le |\overline{Z}^{\tau^{n-1}_\ell}_{\ell,s} - \overline{Z}^{\tau^n_\ell}_{\ell,s}| + \eta_0 \le |Z^{\tau^n_\ell}_{\ell,s}| + \eta_0 \le 1 + \eta_0.$$

Case D: $|f^{\tau^n_\ell}_{\ell,s}| = 0$ and $|f^{\tau^{n-1}_\ell}_{\ell,s}| = 0$. In this case $h^n_{\ell,s} = 0$.

We deduce that if $\tau^n_{\ell,s} > 0$ then $h^n_{\ell,s} \le 3 + 2\eta_0$, while if $\tau^n_{\ell,s} = 0$ then

$$h^n_{\ell,s} = f^n_{\ell,s} + \frac{n-1}{n} f^{n-1}_{\ell,s} \le 2.$$

Therefore condition (b) is satisfied.

Remark 12. The proof can be easily transformed into an algorithm. Indeed, the calculation of $\Pi(\overline{Z}^T)$, the closest point in C to $\overline{Z}^T$ in the Euclidean norm, requires to check whether the absolute value of each coordinate of $\overline{Z}^T$ exceeds η_0, and if so, to round it down to η_0 or up to $-\eta_0$. The distribution p^* defined in Step 4 can be obtained by a linear program. The rest of the construction of σ involves only arithmetic operations. The computational time at each stage is therefore polynomial in L. Because the size of $\widetilde{X}$ is of the order of $L \approx \frac{1}{\varepsilon^{2|\Omega|}}$, it follows that the computational time at each stage is polynomial in $\frac{1}{\varepsilon}$.

B A Variation of the Law of Large Numbers

Let L be a finite set of indices. For every $\ell \in L$ let $\widetilde{V}_\ell$ be a discrete random variable with mean 0 that takes only finitely many values with positive probability. Consider the following stochastic process $(V_T)_{T\in\mathbb{N}}$: An index $C_1 \in L$ is given and V_1 has the same distribution as $\widetilde{V}_{C_1}$. For every stage $T = 2, 3, \cdots$ there is a choice function $C_T = C_T(V_1, \cdots, V_{T-1}) \in L$ that chooses an index in L as a function of past realizations of $V_1, \cdots, V_{T-1}$. Conditional on $V_1, \cdots, V_{T-1}$, the random variable V_T is distributed like $\widetilde{V}_{C_T}$. Formally, for every possible realization of $\widetilde{V}_{C_T}$,

$$\mathbb{P}(V_T = x | V_1 = x_1, \cdots, V_{T-1} = x_{T-1}) = \mathbb{P}(\widetilde{V}_{C_T(x_1,\cdots,x_{T-1})} = x),$$

whenever the left-hand side is well defined.

Let $I_T(\ell) := |\{t \leq T \colon C_t = \ell\}|$ be the number of stages up to stage T in which the chosen index is ℓ. Define $A_\ell := \{C_T = \ell \text{ infinitely often}\}$.

Lemma 3. *For every $\ell \in L$,*

$$\mathbb{P}\left(A_\ell \cap \left\{\lim_{T\to\infty} \frac{\sum_{\{t\leq T;\ C_t=\ell\}} V_t}{I_T(\ell)} = 0\right\}\right) = \mathbb{P}\left(A_\ell\right).$$

Proof. Fix $\ell \in L$. Denote by $\tau(k)$ the k'th time in which $X^T = \widetilde{V}_\ell$. We set $\tau(k) = \infty$ on the event that X^T is equal to $\widetilde{V}_\ell$ less than k times. Define a new stochastic process (V_k) as follows: $V_k := V_{\tau(k)}$ when $\tau(k) < \infty$, and $V_k := 0$ when $\tau_k = \infty$. Then, V_k is uncorrelated with each of $V_1, \cdots, V_{k-1}$. By the law of large numbers, with probability 1 the long-run average of $(V_k)_{k\in\mathbb{N}}$ is 0. Because the long-run averages of $(V_k)_{k\in\mathbb{N}}$ and $\frac{\sum_{\{t\leq T:\ C_t=\ell\}} V_t}{I_T(\ell)}$ coincide on A_ℓ, the desired result follows. ∎

We need a slight variation of Lemma 3, in which the random variables take values in a Euclidean spaces and the index selection function depends also on another stochastic process.

Suppose that the random variables $(\widehat{V}_\ell)_{\ell\in L}$ take values in a finite set Q_1 of a Euclidean space. Let $(W_T)_{T\in\mathbb{N}}$ be a stochastic process that takes values in a finite set Q_2 of a Euclidean space. Consider the following stochastic process, $(V_T)_{T\in\mathbb{N}}$. An index $C_1 \in L$ is given and V_1 has the same distribution as $\widetilde{V}_{C_1}$. For every $T = 2, 3, \cdots$ there is a choice function $C_T = C_T(V_1, \cdots, V_{T-1}, W_1, \cdots, W_{T-1}) \in L$ that chooses an index in L as a function of past realizations of $V_1, \cdots, V_{T-1}, W_1, \cdots, W_{T-1}$. Similar to what we had before, the random variable V_T satisfies,

$$\mathbb{P}(V_T = x | V_1 = x_1, \cdots, V_{T-1} = x_{T-1}, W_1 = w_1, \cdots, W_T = w_T) = \mathbb{P}(\widetilde{V}_{C_T(x_1,\cdots,x_{T-1},w_1,\cdots,w_{T-1})} = x)$$

whenever the left-hand side is well defined.

Note that there is no further assumption of independence between $W_1, W_2, ..$ and $V_1, V_2, \cdots$ beyond this one. In particular, W_T might be dependent of $V_1, \cdots, V_{T-1}$. The function $I_T(\ell)$ plays the same role as before.

Finally, let s be a bi-linear function that assigns to each $(q_1, q_2) \in Q_1 \times Q_2$ a distribution $\mathfrak{s}(q_1, q_2)$ over a finite set in a Euclidian space. Define the stochastic process $U_1, U_2, \cdots$ as follows. For every stage T, conditional on $V_1, \cdots, V_{T-1}, W_1, \cdots, W_{T-1}$, U_T has the distribution $\mathfrak{s}(V_T, W_T)$. Let $A_\ell = \{C_T = \ell \text{ infinitely often}\}$.

Lemma 4. *For every $\ell \in L$,*

$$\begin{aligned}
&\mathbb{P}\left(A_\ell \cap \left\{\lim_{T\to\infty} \frac{\sum_{\{t\leq T:\, C_t=\ell\}} \Big(U_t - \mathbb{E}_t(U_t)\Big)}{I_T(\ell)} = 0\right\}\right) \\
&= \mathbb{P}\left(A_\ell \cap \left\{\lim_{T\to\infty} \left(\frac{\sum_{\{t\leq T:\, C_t=\ell\}} U_t}{I_T(\ell)} - \mathbb{E}(\overline{U}_T) = 0\right)\right\}\right) \\
&= \mathbb{P}\left(A_\ell\right),
\end{aligned}$$

where $\overline{U}_T$ has the distribution $s\Big(\mathbb{E}(\widetilde{V}_\ell), \frac{\sum_{\{t\leq T:\, C_t=\ell\}} \mathbb{E}_t(Wt)}{I_T(\ell)}\Big)$.

The proof is a slight variation of the proof of Lemma 3 and is therefore omitted.