# Robust approachability and regret minimization in games with partial monitoring

**Shie Mannor**
Technion, Haifa
Israel
shie@ee.technion.ac.il

**Vianney Perchet**
Ecole normale supérieure, Cachan
France
vianney.perchet@normalesup.org

**Gilles Stoltz**
Ecole normale supérieure,* Paris
& HEC Paris
France
gilles.stoltz@ens.fr

## Abstract

Approachability has become a standard tool in analyzing learning algorithms in the adversarial online learning setup. We develop a variant of approachability for games where there is ambiguity in the obtained reward that belongs to a set, rather than being a single vector. Using this variant we tackle the problem of approachability in games with partial monitoring and develop simple and efficient algorithms (i.e., with constant per-step complexity) for this setup. We finally consider external and internal regret in repeated games with partial monitoring, for which we derive regret-minimizing strategies based on approachability theory.

## 1 Introduction

Blackwell's approachability theory and its variants has become a standard and useful tool in analyzing online learning algorithms (Cesa-Bianchi and Lugosi, 2006) and algorithms for learning in games (Hart and Mas-Colell, 2000, 2001). The first application of Blackwell's approachability to learning in the online setup is due to Blackwell himself in Blackwell (1956b). Numerous other contributions are summarized in Cesa-Bianchi and Lugosi (2006). Blackwell's approachability theory enjoys a clear geometric interpretation that allows it to be used in situations where online convex optimization or exponential weights do not seem to be easily applicable and, in some sense, to go beyond the minimization of the regret and/or to control quantities of a different flavor; e.g., in Mannor et al. (2009), to minimize the regret together with path constraints, and in Mannor and Shimkin (2008), to minimize the regret in games whose stage duration is not fixed. Recently, it has been shown that approachability and low regret learning are equivalent in the sense that efficient reductions exist from one to the other (Abernethy et al., 2011). Another recent paper (Rakhlin et al., 2011) showed that approachability can be analyzed from the perspective of learnability using tools from learning theory.

In this paper we consider approachability and online learning with partial monitoring in games against Nature. In partial monitoring the decision maker does not know how much reward was obtained and only gets a (random) signal whose distribution depends on the action of the decision maker and the action of Nature. There are two extremes of this setup that are well studied. On the one extreme we have the case where the signal includes the reward itself (or a signal that can be used to unbiasedly estimate the reward), which is essentially the celebrated bandits setup. The other extreme is the case where the signal is not informative (i.e., it tells the decision maker nothing about the actual reward obtained); this setting then essentially consists of repeating the same situation over and over again, as no information is gained over time. We consider a setup encompassing these situations and more general ones, in which the signal is indicative of the actual reward, but is not necessarily a sufficient statistics thereof. The difficulty is that the decision maker cannot compute the actual reward he obtained nor the actions of Nature.

Regret minimization with partial monitoring has been studied in several papers in the learning theory community. Piccolboni and Schindelhauer (2001), Mannor and Shimkin (2003), Cesa-Bianchi et al. (2006) study special cases where an accurate estimation of the rewards (or worst-case rewards) of the decision maker is possible thanks to some extra structure. A general policy with vanishing regret is presented in Lugosi et al. (2008). This policy is based on exponential weights and a specific estimation procedure for the (worst-case) obtained rewards. In contrast, we provide approachability-based results for the problem of regret minimization. On route, we define a new type of approachability setup, with enables to re-derive the extension of approachability to the partial monitoring vector-valued setting proposed by Perchet (2011a).

*CNRS – Ecole normale supérieure, Paris – INRIA, within the project-team CLASSIC

More importantly, we provide algorithms for this approachability problem that are more efficient in the sense that, unlike previous works in the domain, their complexity is constant over all steps. Moreover, their rates of convergence are, as in Blackwell (1956b) but for the first time in this general framework, independent of the game at hand.

The paper is organized as follows. In Section 2 we recall some basic facts from approachability theory. In Section 3 we propose a novel setup for approachability, termed "robust approachability," where instead of obtaining a vector-valued reward, the decision maker obtains a set, that represents the ambiguity concerning his reward. We provide a simple characterization of approachable convex sets and an algorithm for the set-valued reward setup. In Section 4 we show how to apply the robust approachability framework to the repeated vector-valued games with partial monitoring. We provide a simple and constructive algorithm for this setup. Previous results for approachability in this setup were either non-constructive (Rustichini, 1999) or were highly inefficient as they relied on some sort of lifting to the space of probability measures on mixed actions (Perchet, 2011a) and typically required a grid that is progressively refined (leading to a step complexity that is exponential in the number $T$ of past steps). In Section 5 we apply our results for both external and internal regret minimization with partial monitoring. In both cases our proofs are simple, lead to algorithms with constant complexity at each step, and are accompanied with rates. Our results for external regret have rates similar to Lugosi et al. (2008), but our proof is direct and simpler. For internal regret minimization we present the first algorithm not relying on a grid being refined over time and the first convergence rates.

## 2 Some basic facts from approachability theory

In this section we recall the most basic versions of Blackwell's approachability theorem for vector-valued payoff functions.

We consider a vector-valued game between two players, a decision maker (first player) and Nature (second player), with respective finite action sets $\mathcal{A}$ and $\mathcal{B}$, whose cardinalities are referred to as $N_{\mathcal{A}}$ and $N_{\mathcal{B}}$. We denote by $d$ the dimension of the reward vector and equip $\mathbb{R}^d$ with the $\ell^2$–norm $\|\cdot\|_2$. The payoff function of the first player is given by a mapping $m : \mathcal{A} \times \mathcal{B} \to \mathbb{R}^d$, which is multi-linearly extended to $\Delta(\mathcal{A}) \times \Delta(\mathcal{B})$, the set of product-distributions over $\mathcal{A} \times \mathcal{B}$.

We consider two frameworks, depending on whether pure or mixed actions are taken.

**Pure actions taken and observed.** We denote by $A_1, A_2, \ldots$ and $B_1, B_2, \ldots$ the actions in $\mathcal{A}$ and $\mathcal{B}$ sequentially taken by each player; they are possibly given by randomized strategies, i.e., the actions $A_t$ and $B_t$ were obtained by random draws according to respective probability distributions denoted by $\boldsymbol{x}_t \in \Delta(\mathcal{A})$ and $\boldsymbol{y}_t \in \Delta(\mathcal{B})$. For now, we assume that the first player has a full monitoring of the pure actions taken by the opponent player: at the end of round $t$, when receiving the payoff $m(A_t, B_t)$, the pure action $B_t$ is revealed to him.

**Definition 1** *A set $\mathcal{C} \subseteq \mathbb{R}^d$ is* $m$–approachable with pure actions *if there exists a strategy*[1] *of the first player such that for all strategies of the second player,*

$$\limsup_{T\to\infty} \inf_{c\in\mathcal{C}} \left\| c - \frac{1}{T}\sum_{t=1}^{T} m\big(A_t, B_t\big) \right\|_2 = 0 \qquad \textit{a.s.}$$

*That is, the first player has a strategy that ensures that the average of his vector-valued payoffs converges to the set $\mathcal{C}$.*

**Mixed actions taken and observed.** In this case, we denote by $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$ and $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots$ the actions in $\Delta(\mathcal{A})$ and $\Delta(\mathcal{B})$ sequentially taken by each player. We also assume a full monitoring for the first player: at the end of round $t$, when receiving the payoff $m(\boldsymbol{x}_t, \boldsymbol{y}_t)$, the mixed action $\boldsymbol{y}_t$ is revealed to him.

**Definition 2** *In this context, a set $\mathcal{C} \subseteq \mathbb{R}^d$ is* $m$–approachable with mixed actions *if there exists a strategy of the first player such that for all strategies of the second player,*

$$\limsup_{T\to\infty} \inf_{c\in\mathcal{C}} \left\| c - \frac{1}{T}\sum_{t=1}^{T} m\big(\boldsymbol{x}_t, \boldsymbol{y}_t\big) \right\|_2 = 0 \qquad \textit{a.s.}$$

[1]The original definition given by Blackwell requires uniformity w.r.t. the strategy set of the opponent. We ignore the uniformity to avoid excessive nomenclature.

**Necessary and sufficient condition for approachability.** For closed convex sets there is a simple characterization of approachability that is a direct consequence of the minimax theorem; the condition is the same for the two settings, whether pure or mixed actions are taken and observed.

**Theorem 3 (Blackwell 1956a, Theorem 3)** *A closed convex set $\mathcal{C} \subseteq \mathbb{R}^d$ is approachable (with pure or mixed actions) if and only if*

$$\forall\, \boldsymbol{y} \in \Delta(\mathcal{B}),\ \ \exists\, \boldsymbol{x} \in \Delta(\mathcal{A}), \qquad\qquad m(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{C}\,.$$

*In the latter case, an explicit strategy achieves the following convergence rates. We denote by $M$ a bound in norm over $m$, i.e.,*

$$\max_{(a,b)\in\mathcal{A}\times\mathcal{B}} \big\| m(a,b) \big\|_2 \leqslant M\,.$$

*With mixed actions taken and observed, for all strategies of the second player, with probability 1,*

$$\inf_{c\in\mathcal{C}} \left\| c - \frac{1}{T}\sum_{t=1}^{T} m\big(\boldsymbol{x}_t, \boldsymbol{y}_t\big) \right\|_2 \leqslant \frac{2M}{\sqrt{T}}\,.$$

*With pure actions taken and observed, for all $\delta \in (0,1)$ and for all strategies of the second player, with probability at least $1-\delta$,*

$$\inf_{c\in\mathcal{C}} \left\| c - \frac{1}{T}\sum_{t=1}^{T} m\big(A_t, B_t\big) \right\|_2 \leqslant \frac{2M}{\sqrt{T}}\Big(1 + 2\sqrt{\ln(2/\delta)}\Big).$$

The proof is provided in Appendix B.1 for completeness.

**An associated strategy (that is efficient depending on the geometry of $\mathcal{C}$).** Blackwell suggested a simple strategy with a geometric flavor. (This strategy can be extracted from the proof of Theorem 3 provided in appendix.)

Play an arbitrary $\boldsymbol{x}_1$. For $t \geqslant 1$, given the vector-valued quantities

$$\widehat{m}_t = \frac{1}{t}\sum_{\tau=1}^{t} m(\boldsymbol{x}_\tau, B_\tau) \qquad \text{or} \qquad \widehat{m}_t = \frac{1}{t}\sum_{\tau=1}^{t} m(\boldsymbol{x}_\tau, \boldsymbol{y}_\tau)\,,$$

depending on whether pure or mixed actions are taken and observed, compute the projection $c_t$ (in $\ell^2$–norm) of $\widehat{m}_t$ on $\mathcal{C}$. Find a mixed action $\boldsymbol{x}_{t+1}$ that solves the minimax equation

$$\min_{\boldsymbol{x}\in\Delta(\mathcal{A})}\ \max_{\boldsymbol{y}\in\Delta(\mathcal{B})} \big\langle \widehat{m}_t - c_t, m(\boldsymbol{x}, \boldsymbol{y}) \big\rangle\,, \tag{1}$$

where $\langle\,\cdot\,,\,\cdot\,\rangle$ is the Euclidian inner product in $\mathbb{R}^d$. The minimax problem above is easily seen to be a (scalar) zero-sum game and is therefore efficiently solvable using, e.g., linear programming: the associated complexity is polynomial in $N_{\mathcal{A}}$ and $N_{\mathcal{B}}$. All in all, this strategy is efficient as soon as the computations of the required projections onto $\mathcal{C}$ in $\ell^2$–norm can be performed efficiently.

In the case when pure actions are taken and observed, it only remains to draw $A_{t+1}$ at random according to $\boldsymbol{x}_{t+1}$.

## 3 Robust approachability

In this section we extend the results of the previous section to set-valued payoff functions. To this end, we denote by $\mathcal{S}\big(\mathbb{R}^d\big)$ the set of all subsets of $\mathbb{R}^d$ and consider a set-valued payoff function $\overline{m} : \mathcal{A}\times\mathcal{B} \to \mathcal{S}\big(\mathbb{R}^d\big)$.

**Pure actions taken and observed.** At each round $t$, the players choose simultaneously respective actions $A_t \in \mathcal{A}$ and $B_t \in \mathcal{B}$, possibly at random according to mixed distributions $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$. Full monitoring still takes place for the first player: he observes $B_t$ at the end of round $t$. However, as a result, the first player gets the *subset* $\overline{m}(A_t, B_t)$ as a payoff. This models the ambiguity or uncertainty associated with some true underlying payoff gained.

We extend $\overline{m}$ multi-linearly to $\Delta(\mathcal{A})\times\Delta(\mathcal{B})$ and even to $\Delta(\mathcal{A}\times\mathcal{B})$, the set of joint probability distributions on $\mathcal{A}\times\mathcal{B}$, as follows. Let

$$\mu = \big(\mu_{a,b}\big)_{(a,b)\in\mathcal{A}\times\mathcal{B}}$$

be such a joint probability distribution; then $\overline{m}(\mu)$ is defined as a finite convex combination[2] of subsets of $\mathbb{R}^d$,

$$\overline{m}(\mu) = \sum_{a\in\mathcal{A}} \sum_{b\in\mathcal{B}} \mu_{a,b}\, \overline{m}(a,b)\,.$$

When $\mu$ is the product-distribution of some $\boldsymbol{x} \in \Delta(\mathcal{A})$ and $\boldsymbol{y} \in \Delta(\mathcal{B})$, we use the notation $\overline{m}(\mu) = \overline{m}(\boldsymbol{x},\boldsymbol{y})$.

We denote by

$$\pi_T = \frac{1}{T}\sum_{t=1}^{T} \delta_{(A_t,B_t)}$$

the empirical distribution of the pairs $(A_t, B_t)$ of actions taken during the first $T$ rounds and will be interested in the behavior of

$$\frac{1}{T}\sum_{t=1}^{T} \overline{m}(A_t,B_t)\,,$$

which can also be rewritten here in a compact way as $\overline{m}(\pi_T)$, by linearity of the extension of $\overline{m}$.

**Definition 4** *Let $\mathcal{C} \subseteq \mathbb{R}^d$ be some set; $\mathcal{C}$ is $\overline{m}$–approachable with pure actions if there exists a strategy of the first player such that for all strategies of the second player,*

$$\limsup_{T\to\infty} \quad \sup_{d\in\overline{m}(\pi_T)} \inf_{c\in\mathcal{C}} \|c-d\|_2 \;= 0 \qquad \textit{a.s.}$$

That is, when $\mathcal{C}$ is $\overline{m}$–approachable with pure actions, the first player has a strategy that ensures that the average of the sets of payoffs converges to the set $\mathcal{C}$: the sets $\overline{m}(\pi_T)$ are included in $\varepsilon_T$–neighborhoods of $\mathcal{C}$, where the sequence of $\varepsilon_T$ tends almost-surely to $0$.

**Mixed actions taken and observed.** At each round $t$, the players choose simultaneously respective mixed actions $\boldsymbol{x}_t \in \Delta(\mathcal{A})$ and $\boldsymbol{y}_t \in \Delta(\mathcal{B})$. Full monitoring still takes place for the first player: he observes $\boldsymbol{y}_t$ at the end of round $t$; he however gets the subset $\overline{m}(\boldsymbol{x}_t,\boldsymbol{y}_t)$ as a payoff (which, again, accounts for the uncertainty).

The product-distribution of two elements $\boldsymbol{x} = (x_a)_{a\in\mathcal{A}} \in \Delta(\mathcal{A})$ and $\boldsymbol{y} = (y_b)_{b\in\mathcal{B}} \in \Delta(\mathcal{B})$ will be denoted by $\boldsymbol{x}\otimes\boldsymbol{y}$; it gives a probability mass of $x_a y_b$ to each pair $(a,b) \in \mathcal{A}\times\mathcal{B}$. We consider the empirical joint distribution of mixed actions taken during the first $T$ rounds,

$$\nu_T = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}_t \otimes \boldsymbol{y}_t\,,$$

and will be interested in the behavior of

$$\frac{1}{T}\sum_{t=1}^{T} \overline{m}(\boldsymbol{x}_t,\boldsymbol{y}_t)\,,$$

which can also be rewritten here in a compact way as $\overline{m}(\nu_T)$, by linearity of the extension of $\overline{m}$.

**Definition 5** *Let $\mathcal{C} \subseteq \mathbb{R}^d$ be some set; $\mathcal{C}$ is $\overline{m}$–approachable with mixed actions if there exists a strategy of the first player such that for all strategies of the second player,*

$$\limsup_{T\to\infty} \quad \sup_{d\in\overline{m}(\nu_T)} \inf_{c\in\mathcal{C}} \|c-d\|_2 \;= 0 \qquad \textit{a.s.}$$

**A useful continuity lemma.** Before proceeding we provide a continuity lemma. It can be reformulated as indicating that for all joint distributions $\mu$ and $\nu$ over $\mathcal{A}\times\mathcal{B}$, the set $\overline{m}(\mu)$ is contained in a $M\,\|\mu-\nu\|_1$–neighborhood of $\overline{m}(\nu)$, where $M$ is a bound in $\ell^2$–norm on $\overline{m}$; this is a fact that we will use repeatedly below.

**Lemma 6** *Let $\mu$ and $\nu$ be two probability distributions over $\mathcal{A}\times\mathcal{B}$. We assume that the set-valued function $\overline{m}$ is bounded in norm by $M$, i.e., that there exists a real number $M > 0$ such that*

$$\forall (a,b) \in \mathcal{A}\times\mathcal{B}, \qquad \sup_{d\in\overline{m}(a,b)} \|d\|_2 \leqslant M\,.$$

*Then*

$$\sup_{d\in\overline{m}(\mu)} \inf_{c\in\overline{m}(\nu)} \|d-c\|_2 \;\leqslant M\,\|\mu-\nu\|_1 \leqslant M\sqrt{N_{\mathcal{A}}N_{\mathcal{B}}}\,\|\mu-\nu\|_2\,,$$

*where the norms in the right-hand side are respectively the $\ell^1$ and $\ell^2$–norms between probability distributions.*

[2] For two sets $S,\ T$ and $\alpha \in [0,1]$, the convex combination $\alpha S + (1-\alpha)T$ is defined as

$$\big\{\alpha s + (1-\alpha)t,\ \ s\in S \text{ and } t\in T\big\}\,.$$

**Proof:** Let $d$ be an element of $\overline{m}(\mu)$; it can be written as

$$d = \sum_{a\in\mathcal{A}} \sum_{b\in\mathcal{B}} \mu_{a,b}\, \theta_{a,b}$$

for some elements $\theta_{a,b} \in \overline{m}(a,b)$. We consider

$$c = \sum_{a\in\mathcal{A}} \sum_{b\in\mathcal{B}} \nu_{a,b}\, \theta_{a,b}\,,$$

which is an element of $\overline{m}(\nu)$. Then by the triangle inequality,

$$\|d-c\|_2 = \left\| \sum_{a\in\mathcal{A}} \sum_{b\in\mathcal{B}} \big(\mu_{a,b}-\nu_{a,b}\big)\theta_{a,b} \right\|_2 \leqslant \sum_{a\in\mathcal{A}} \sum_{b\in\mathcal{B}} \big|\mu_{a,b}-\nu_{a,b}\big|\, \|\theta_{a,b}\|_2 \leqslant M \sum_{a\in\mathcal{A}} \sum_{b\in\mathcal{B}} \big|\mu_{a,b}-\nu_{a,b}\big|\,.$$

This entails the first claimed inequality. The second one follows from an application of the Cauchy-Schwarz inequality. ∎

**Necessary and sufficient condition for approachability.** We state the condition in the theorem below, as well as the associated convergence rates. Explicit strategies can be deduced from the proof, which is based on Theorem 3; these strategies are efficient as soon as projections in $\ell^2$–norm onto the set $\widetilde{\mathcal{C}}$ defined in (3) can be computed efficiently. The latter fact depends on the respective geometries of $\overline{m}$ and $\mathcal{C}$.

**Theorem 7** *Suppose that the set-valued function $\overline{m}$ is bounded in norm by $M$. A closed convex set $\mathcal{C} \subseteq \mathbb{R}^d$ is approachable (with pure or mixed actions) if and only if the following robust approachability condition is satisfied,*

$$\forall\, \boldsymbol{y} \in \Delta(\mathcal{B}),\ \ \exists\, \boldsymbol{x} \in \Delta(\mathcal{A}), \qquad \overline{m}(\boldsymbol{x},\boldsymbol{y}) \subseteq \mathcal{C}\,. \tag{RAC}$$

*In the latter case, the following convergence rates are achieved by a strategy constructed in the proof. With mixed actions taken and observed, for all strategies of the second player, with probability 1,*

$$\sup_{d\in\overline{m}(\nu_T)} \inf_{c\in\mathcal{C}} \|c-d\|_2 \leqslant \frac{2M}{\sqrt{T}} \sqrt{N_{\mathcal{A}} N_{\mathcal{B}}}\,.$$

*With pure actions taken and observed, for all $\delta \in (0,1)$ and for all strategies of the second player, with probability at least $1-\delta$,*

$$\sup_{d\in\overline{m}(\pi_T)} \inf_{c\in\mathcal{C}} \|c-d\|_2 \leqslant \frac{2M}{\sqrt{T}} \sqrt{N_{\mathcal{A}} N_{\mathcal{B}}} \Big(1 + 2\sqrt{\ln(2/\delta)}\Big).$$

**Proof: Condition** (RAC) **is necessary.** If the condition does not hold, then there exists $\boldsymbol{y}_0 \in \Delta(\mathcal{B})$ such that for every $\boldsymbol{x} \in \mathcal{A}$, the set $\overline{m}(\boldsymbol{x},\boldsymbol{y}_0)$ is not included in $\mathcal{C}$, i.e., it contains at least one point not in $\mathcal{C}$. We then define a mapping $D : \Delta(A) \to \mathbb{R}$ by

$$\forall\, \boldsymbol{x} \in \Delta(\mathcal{A}), \qquad D(\boldsymbol{x}) = \sup_{d\in\overline{m}(\boldsymbol{x},\boldsymbol{y}_0)} \inf_{c\in\mathcal{C}} \|c-d\|_2\ .$$

Since $\mathcal{C}$ is closed, distances of given individual points to $\mathcal{C}$ are achieved; therefore, by the choice of $\boldsymbol{y}_0$, we get that $D(\boldsymbol{x}) > 0$ for all $\boldsymbol{x} \in \Delta(\mathcal{A})$.

We now show that $D$ is continuous on the compact set $\Delta(\mathcal{A})$; it thus attains its minimum, whose value we denote by $D_{\min} > 0$. More precisely, it suffices to show that for all $\boldsymbol{x}, \boldsymbol{x}' \in \Delta(\mathcal{A})$, the condition $\|\boldsymbol{x}'-\boldsymbol{x}\|_1 \leqslant \varepsilon$ implies that $D(\boldsymbol{x}) - D(\boldsymbol{x}') \leqslant M\varepsilon$. Indeed, fix $\delta > 0$ and let $d_{\delta,\boldsymbol{x}} \in \overline{m}(\boldsymbol{x},\boldsymbol{y}_0)$ be such that

$$D(\boldsymbol{x}) \leqslant \inf_{c\in\mathcal{C}} \big\| c - d_{\delta,\boldsymbol{x}} \big\|_2 + \delta\,. \tag{2}$$

By Lemma 6 (with the choices $\mu = \boldsymbol{x}\otimes\boldsymbol{y}_0$ and $\nu = \boldsymbol{x}'\otimes\boldsymbol{y}_0$) there exists $d_{\delta,\boldsymbol{x}'} \in \overline{m}(\boldsymbol{x}',\boldsymbol{y}_0)$ such that $\big\| d_{\delta,\boldsymbol{x}} - d_{\delta,\boldsymbol{x}'} \big\|_2 \leqslant M\varepsilon + \delta$. The triangle inequality entails that

$$\inf_{c\in\mathcal{C}} \big\| c - d_{\delta,\boldsymbol{x}} \big\|_2 \leqslant \inf_{c\in\mathcal{C}} \big\| c - d_{\delta,\boldsymbol{x}'} \big\|_2 + M\varepsilon + \delta\,.$$

Substituting in (2), we get that

$$D(\boldsymbol{x}) \leqslant M\varepsilon + 2\delta + \inf_{c\in\mathcal{C}} \big\| c - d_{\delta,\boldsymbol{x}'} \big\|_2 \leqslant M\varepsilon + 2\delta + D(\boldsymbol{x}')\,,$$

which, letting $\delta \to 0$, proves our continuity claim.

Assume now that the second player chooses at each round $\boldsymbol{y}_t = \boldsymbol{y}_0$ as his mixed action. In the case of mixed actions taken and observed, denoting

$$\overline{\boldsymbol{x}}_T = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}_t \,,$$

we get that $\nu_t = \overline{\boldsymbol{x}}_T \otimes \boldsymbol{y}_0$, and hence, for all strategies of the first player and for all $T \geqslant 1$,

$$\sup_{d \in \overline{m}(\nu_T)} \inf_{c \in \mathcal{C}} \|c-d\|_2 = D(\overline{\boldsymbol{x}}_T) \geqslant D_{\min} > 0 \,,$$

which shows that $\mathcal{C}$ is not approachable. The case of pure actions taken and observed is treated similarly, with the sole addition of a concentration argument. By repeated uses of the Hoeffding-Azuma inequality together with an application of the Borel-Cantelli lemma, $\delta_T = \|\pi_T - \nu_T\|_1 \to 0$ almost surely as $T \to \infty$. By applying Lemma 6 as above, we get

$$\sup_{d \in \overline{m}(\pi_T)} \inf_{c \in \mathcal{C}} \|c-d\|_2 \geqslant \sup_{d \in \overline{m}(\nu_T)} \inf_{c \in \mathcal{C}} \|c-d\|_2 - M\delta_T \geqslant D_{\min} - M\delta_T \,;$$

we simply take the $\liminf$ in the above inequalities to conclude the argument. ■

**Proof: Condition** (RAC) **is sufficient.** We first show that there exists a strategy of the first player such that, for all strategies of the opponent player, the sequences $(\pi_T)$ or $(\nu_T)$ of the empirical distributions of actions converge to the set

$$\widetilde{\mathcal{C}} = \big\{ \mu \in \Delta(\mathcal{A} \times \mathcal{B}) : \ \overline{m}(\mu) \subseteq \mathcal{C} \big\} \tag{3}$$

in $\ell^2$–norm, at the rates prescribed by Theorem 3.

To do so, we identify probability distributions over $\mathcal{A} \times \mathcal{B}$ with vectors in $\mathbb{R}^{\mathcal{A}\times\mathcal{B}}$ and consider the vector-valued payoff function

$$m : (a,b) \in \mathcal{A} \times \mathcal{B} \longmapsto \delta_{(a,b)} \in \mathbb{R}^{\mathcal{A}\times\mathcal{B}} \,,$$

which we extend multi-linearly to $\Delta(\mathcal{A}) \times \Delta(\mathcal{B})$. We have that

$$\pi_T = \frac{1}{T}\sum_{t=1}^{T} m(A_t, B_t) \qquad \text{and} \qquad \nu_T = \frac{1}{T}\sum_{t=1}^{T} m(\boldsymbol{x}_t, \boldsymbol{y}_t)$$

and we therefore only need to show that $\widetilde{C}$ is $m$–approachable (with pure or mixed actions).

Since $\overline{m}$ is a linear function on $\Delta(\mathcal{A} \times \mathcal{B})$ and $\mathcal{C}$ is convex, the set $\widetilde{\mathcal{C}}$ is convex as well. In addition, since $\mathcal{C}$ is closed, $\widetilde{\mathcal{C}}$ is also closed. We can therefore apply the original version of the approachability theorem (stated in Theorem 3). The desired existence result follows therefore from the fact that by assumption, for all $\boldsymbol{y} \in \Delta(\mathcal{B})$, there exists some $\boldsymbol{x} \in \Delta(\mathcal{A})$ such that $\mu = m(\boldsymbol{x}, \boldsymbol{y})$, the product-distribution between $\boldsymbol{x}$ and $\boldsymbol{y}$, belongs to $\widetilde{\mathcal{C}}$, as it satisfies $\overline{m}(\mu) = \overline{m}(\boldsymbol{x}, \boldsymbol{y}) \subseteq \mathcal{C}$.

Let $P_{\widetilde{\mathcal{C}}}$ denote the projection operator onto $\widetilde{\mathcal{C}}$. We therefore have proved the existence of explicit (and possibly efficient) strategies—along the lines of the ones presented around (1)—such that, for all strategies of the second player, with probability $1-\delta$,

$$\varepsilon_T := \Big\| \pi_T - P_{\widetilde{\mathcal{C}}}(\pi_T) \Big\|_2 = \inf_{\mu \in \widetilde{\mathcal{C}}} \|\pi_T - \mu\|_2 \leqslant \frac{2}{\sqrt{T}} \Big( 1 + \sqrt{2\ln(2/\delta)} \Big),$$

and with probability 1, $\varepsilon'_T := \Big\| \nu_T - P_{\widetilde{\mathcal{C}}}(\nu_T) \Big\|_2 = \inf_{\mu \in \widetilde{\mathcal{C}}} \|\nu_T - \mu\|_2 \leqslant \dfrac{2}{\sqrt{T}} \,.$

Lemma 6 entails that the sets $\overline{m}(\pi_T)$ are included in $M\sqrt{N_{\mathcal{A}}N_{\mathcal{B}}}\,\varepsilon_T$–neighborhoods of $\overline{m}\big(P_{\widetilde{\mathcal{C}}}(\pi_T)\big)$, and thus, by definition of $\widetilde{\mathcal{C}}$, in $M\sqrt{N_{\mathcal{A}}N_{\mathcal{B}}}\,\varepsilon_T$–neighborhoods of $\mathcal{C}$. A similar statement holds for the sets the sets $\overline{m}(\nu_T)$ and this completes the proof. ■

## 4 Application to games with partial monitoring

A repeated vector-valued game with partial monitoring is described as follows (see, e.g., Mertens et al., 1994, Rustichini, 1999 and the references therein). The players have respective finite action sets $\mathcal{I}$ and $\mathcal{J}$. We denote by $r : \mathcal{I} \times \mathcal{J} \to \mathbb{R}^d$ the vector-valued payoff function of the first player and extend it multi-linearly to $\Delta(\mathcal{I}) \times \Delta(\mathcal{J})$. At each round, players simultaneously choose their actions $I_t \in \mathcal{I}$ and $J_t \in \mathcal{J}$, possibly at random according to probability distributions denoted by $\boldsymbol{p}_t \in \Delta(\mathcal{I})$ and $\boldsymbol{q}_t \in \Delta(\mathcal{J})$. At the end of a round, the first player does not observe $J_t$ or $r(I_t, J_t)$ but only a signal. There is a finite set $\mathcal{H}$ of possible signals; the feedback $S_t$ that is given to the first player is drawn at random according to the distribution $H(I_t, J_t)$, where the mapping $H : \mathcal{I} \times \mathcal{J} \to \Delta(\mathcal{H})$ is known by the first player.

**Example 1** Examples of such partial monitoring games are provided by, e.g., Cesa-Bianchi et al. (2006, Section 2), among which we can cite the apple tasting problem, the label-efficient prediction constraint, and the multi-armed bandit settings.

Some additional notation will be useful. We denote by $R$ the norm of (the linear extension of) $r$,

$$R = \max_{(i,j)\in\mathcal{I}\times\mathcal{J}} \big\| r(i,j) \big\|_2 .$$

The cardinalities of the finite sets $\mathcal{I}$, $\mathcal{J}$, and $\mathcal{H}$ will be referred to as $N_{\mathcal{I}}$, $N_{\mathcal{J}}$, and $N_{\mathcal{H}}$.

Definition 1 can be extended as follows in this setting; the only new ingredient is the signaling structure, the aim is unchanged.

**Definition 8** *Let $\mathcal{C} \subseteq \mathbb{R}^d$ be some set; $\mathcal{C}$ is $r$–approachable for the signaling structure $H$ if there exists a strategy of the first player such that for all strategies of the second player,*

$$\limsup_{T\to\infty} \ \inf_{c\in\mathcal{C}} \left\| c - \frac{1}{T}\sum_{t=1}^{T} r(I_t, J_t) \right\|_2 = 0 \qquad \text{a.s.}$$

*That is, the first player has a strategy that ensures that the sequence of his average vector-valued payoffs converges to the set $\mathcal{C}$, even if he only observes the random signals $S_t$ as a feedback.*

A necessary and sufficient condition for $r$–approachability with the signaling structure $H$ was stated and proved by Perchet (2011a); we therefore need to indicate where our contribution lies. First, both proofs are constructive but our strategy can be efficient (as soon as some projection operator can be efficiently implemented) whereas the one of Perchet (2011a) relies on auxiliary strategies that are calibrated and that require a grid that is progressively refined to be so (leading to a step complexity that is exponential in the number $T$ of past steps). Second, we are able to exhibit convergence rates. Third, as far as elegancy is concerned, our proof is short, compact, and more direct than the one of Perchet (2011a), which relied on several layers of complicated notions (internal regret in partial monitoring, calibration of auxiliary strategies, etc.).

To recall the mentioned approachability condition of Perchet (2011a) we need some additional notation: for all $\boldsymbol{q} \in \Delta(\mathcal{J})$, we denote by $\widetilde{H}(\boldsymbol{q})$ the element in $\Delta(\mathcal{H})^{\mathcal{I}}$ defined as follows. For all $i \in \mathcal{I}$, its $i$–th component is given by the following convex combination of probability distributions over $\mathcal{H}$,

$$\widetilde{H}(\boldsymbol{q})_i = H(i,\boldsymbol{q}) = \sum_{j\in\mathcal{J}} q_j H(i,j) .$$

Finally, we denote by $\mathcal{F}$ the set of feasible vectors of probability distributions over $\mathcal{H}$:

$$\mathcal{F} = \Big\{ \widetilde{H}(\boldsymbol{q}) : \ \ \boldsymbol{q} \in \Delta(\mathcal{J}) \Big\}.$$

A generic element of $\mathcal{F}$ will be denoted by $\sigma \in \mathcal{F}$. The necessary and sufficient condition exhibited by Perchet (2011a) for the $r$–approachability of $\mathcal{C}$ for the signaling structure $H$ can now be recalled.

**Condition 1** *The signaling structure $H$, the vector-payoff function $r$, and the set $\mathcal{C}$ satisfy*

$$\forall\, \boldsymbol{q} \in \Delta(\mathcal{J}),\ \ \exists\, \boldsymbol{p} \in \Delta(\mathcal{I}),\ \ \forall\, \boldsymbol{q}' \in \Delta(\mathcal{J}), \qquad \widetilde{H}(\boldsymbol{q}) = \widetilde{H}(\boldsymbol{q}') \ \ \Rightarrow \ \ r(\boldsymbol{p},\boldsymbol{q}') \in \mathcal{C} .$$

*Defining the set-valued function $\overline{m}$, for all $\boldsymbol{p} \in \Delta(\mathcal{I})$ and $\sigma \in \mathcal{F}$, by*

$$\overline{m}(\boldsymbol{p},\sigma) = \big\{ r(\boldsymbol{p},\boldsymbol{q}') : \ \ \boldsymbol{q}' \in \Delta(\mathcal{J}) \text{ such that } \widetilde{H}(\boldsymbol{q}') = \sigma \big\} ,$$

*the condition can be equivalently reformulated as*

$$\forall\, \sigma \in \mathcal{F},\ \ \exists\, \boldsymbol{p} \in \Delta(\mathcal{I}), \qquad \overline{m}(\boldsymbol{p},\sigma) \subseteq \mathcal{C} .$$

**This condition is necessary.** The next two sections show (in a constructive way and by constructing strategies) that Condition 1 is sufficient for $r$–approachability of closed convex sets $\mathcal{C}$ given the signaling structure $H$. That this condition is necessary was already proved in Perchet (2011a); a slightly simpler argument can however be found in Appendix A.1.

### 4.1 Approachability for deterministic feedback signals only depending on outcome

In this section, we assume that $H$ is of the following form: it only contains Dirac masses, and these Dirac masses $H(i,j)$ only depend on $j$. Put differently, the signals $S_t$ are deterministic functions of the actions $J_t$; we thus denote by $h : \mathcal{J} \to \mathcal{H}$ the function such that $S_t = h(J_t)$ for all $t$ and extend it linearly to $\Delta(\mathcal{J})$. The condition stated above takes the following simpler form (we assume with no loss of generality that all elements in $\mathcal{H}$ are associated with at least one action $j \in \mathcal{J}$, so that $\mathcal{F}$ can be identified with $\mathcal{H}$):

$$\forall\, \sigma \in \Delta(\mathcal{H}),\ \ \exists\, \boldsymbol{p} \in \Delta(\mathcal{I}), \qquad \overline{m}(\boldsymbol{p}, \sigma) \subseteq \mathcal{C}\,, \tag{4}$$

where

$$\overline{m}(\boldsymbol{p}, \sigma) = \big\{ r(\boldsymbol{p}, \boldsymbol{q}') :\ \ \boldsymbol{q}' \in \Delta(\mathcal{J}) \text{ such that } h(\boldsymbol{q}') = \sigma \big\}\,.$$

The fact that $\boldsymbol{p}$ and $\sigma$ are unrelated in the definition above entails that $\overline{m}$ is linear, i.e., that for all $\boldsymbol{p} \in \Delta(\mathcal{I})$ and $\sigma \in \Delta(\mathcal{H})$,

$$\overline{m}(\boldsymbol{p}, \sigma) = \sum_{i \in \mathcal{I}} \sum_{s \in \mathcal{H}} p_i\, \sigma_s\, \overline{m}(i, s)\,.$$

In addition, $\overline{m}$ is also bounded in norm by $R$. Therefore, we are exactly in the setting of Section 3.

**Theorem 9** *A closed convex $\mathcal{C}$ is $r$–approachable for the signaling structure $h$ if and only if (4) holds. In this case, there exists an explicit strategy to do so, at the following rate: for all $T$, with probability at least $1-\delta$,*

$$\inf_{c \in \mathcal{C}} \left\| c - \frac{1}{T} \sum_{t=1}^{T} r(I_t, J_t) \right\|_2 \leqslant \frac{2R}{\sqrt{T}} \sqrt{N_{\mathcal{I}} N_{\mathcal{H}}} \Big( 1 + 2\sqrt{\ln(2/\delta)} \Big).$$

**Proof:** We need only to show that the stated condition entails approachability. Since by definition of $\overline{m}$ and because of the particular signaling structure $h$,

$$\frac{1}{T} \sum_{t=1}^{T} r(I_t, J_t) \ \ \in \ \ \frac{1}{T} \sum_{t=1}^{T} \overline{m}(I_t, S_t)\,,$$

it is enough to show that $\mathcal{C}$ is $\overline{m}$–approachable (when the signals $S_t$ are observed, which is the case). But Theorem 7 indicates that this is the case when (4) holds. ■

The efficiency of the obtained strategy depends on the respective geometries of $\overline{m}$ and $\mathcal{C}$, as was indicated before the statement of Theorem 7.

### 4.2 Approachability with general signaling structures

In this section we consider the case where the signal structure is general. We start from a technical lemma that is needed to show that $\overline{m}(\boldsymbol{p}, \sigma)$ can be written as a *finite* convex combination of sets of the form $\overline{m}(i, b)$. We then describe a (possibly) efficient strategy for approachability followed by convergence rate guarantees.

#### 4.2.1 A preliminary technical result.

With general signaling structures, $\overline{m}$ is not linear, it only satisfies that for all $\boldsymbol{p} \in \Delta(\mathcal{I})$, all pairs $\sigma,\ \sigma' \in \mathcal{F}$, and all $\alpha \in [0,1]$,

$$\alpha\, \overline{m}(\boldsymbol{p}, \sigma) + (1-\alpha)\, \overline{m}(\boldsymbol{p}, \sigma') \ \subseteq\ \overline{m}(\boldsymbol{p},\, \alpha\sigma + (1-\alpha)\sigma')\,,$$

with a strict inclusion in general. (Specific examples can be provided.) Therefore, a direct appeal to Theorem 7 is not possible anymore.

However, a similar linearity property on a lifted finite set is given by the geometric lemma stated below. It follows from an application of Rambau and Ziegler (1996, Proposition 2.4), which entails that since $\widetilde{H}$ is linear on the polytope $\Delta(\mathcal{J})$, its inverse application $\widetilde{H}^{-1}$ is a piecewise linear mapping of $\mathcal{F}$ into the subsets of $\Delta(\mathcal{J})$; the detailed proof can be found in Appendix A.2.

**Lemma 10** *There exist a finite subset $\mathcal{B} \subseteq \mathcal{F}$ and a mapping $\Phi : \mathcal{F} \to \Delta(\mathcal{B})$ such that*

$$\forall\, \sigma \in \mathcal{F},\ \ \forall\, \boldsymbol{p} \in \Delta(\mathcal{I}), \qquad \overline{m}(\boldsymbol{p}, \sigma) = \sum_{i \in \mathcal{I}} \sum_{b \in \mathcal{B}} p_i\, \Phi_b(\sigma)\, \overline{m}(i, b)\,,$$

*where we denoted the convex weight vector $\Phi(\sigma) \in \Delta(\mathcal{B})$ by $\big(\Phi_b(\sigma)\big)_{b \in \mathcal{B}}$.*

**Definition 11** *We denote by $\overline{\overline{m}}$ the linear extension to $\Delta(\mathcal{I} \times \mathcal{B})$ of the restriction of $\overline{m}$ to $\mathcal{I} \times \mathcal{B}$, so that for all $\boldsymbol{p} \in \Delta(\mathcal{I})$ and $\sigma \in \mathcal{F}$,*

$$\overline{m}(\boldsymbol{p}, \sigma) = \overline{\overline{m}}\big(\boldsymbol{p},\ \Phi(\sigma)\big)\,.$$

**Remark 1** The proof shows that $\Phi$ is piecewise linear on a finite decomposition of $\mathcal{F}$; it is therefore Lipschitz on $\mathcal{F}$. We denote by $\kappa_\Phi$ its Lipschitz constant with respect to the $\ell^2$–norm.

*Parameters*: an integer block length $L \geqslant 1$, an exploration parameter $\gamma \in [0,1]$, a strategy $\Psi$ for $\overline{\overline{m}}$–approachability of $\mathcal{C}$

*Notation*: $\boldsymbol{u} \in \Delta(\mathcal{I})$ is the uniform distribution over $\mathcal{I}$, $P_{\mathcal{F}}$ denotes the projection operator in $\ell^2$–norm of $\mathbb{R}^{\mathcal{H}\times\mathcal{I}}$ onto $\mathcal{F}$

*Initialization*: compute the finite set $\mathcal{B}$ and the mapping $\Phi : \mathcal{F} \to \Delta(\mathcal{B})$ of Lemma 10, pick an arbitrary $\boldsymbol{x}_1 \in \Delta(\mathcal{I})$

*For all blocks* $n = 1, 2, \ldots,$

1. define $\boldsymbol{p}_n = (1-\gamma)\,\boldsymbol{x}_n + \gamma\,\boldsymbol{u}$;
2. for rounds $t = (n-1)L+1,\ \ldots,\ nL$,
   2.1 drawn an action $I_t \in \mathcal{I}$ at random according to $\boldsymbol{p}_n$;
   2.2 get the signal $S_t$;
3. form the estimated vector of probability distributions over signals,
$$\widetilde{\sigma}_n = \left( \frac{1}{L} \sum_{t=(n-1)L+1}^{nL} \frac{\mathbb{I}_{\{S_t=s\}}\mathbb{I}_{\{I_t=i\}}}{p_{I_t,n}} \right)_{(i,s)\in\mathcal{I}\times\mathcal{H}};$$
4. compute the projection $\widehat{\sigma}_n = P_{\mathcal{F}}\big(\widetilde{\sigma}_n\big)$;
5. choose $\boldsymbol{x}_{n+1} = \Psi\Big(\boldsymbol{x}_1,\, \Phi\big(\widehat{\sigma}_1\big),\, \ldots,\, \boldsymbol{x}_n,\, \Phi\big(\widehat{\sigma}_n\big)\Big)$.

Figure 1: The proposed strategy, which plays in blocks.

#### 4.2.2 Construction of a strategy to approach $\mathcal{C}$.

The approaching strategy for the original problem is based on a strategy $\Psi$ for $\overline{\overline{m}}$–approachability of $\mathcal{C}$, provided by Theorem 7 and thus solving repeatedly minimax problems of the form (1). We therefore first need to prove the existence of such a $\Psi$.

**Lemma 12** *Under Condition 1, the closed convex set $\mathcal{C}$ is approachable.*

**Proof:** We show that Condition (RAC) in Theorem 7 is satisfied, that is, that for all $\boldsymbol{y} \in \Delta(\mathcal{B})$, there exists a $\boldsymbol{p} \in \Delta(\mathcal{I})$ such that $\overline{\overline{m}}(\boldsymbol{p},\boldsymbol{y}) \subseteq \mathcal{C}$. By linearity of $\overline{\overline{m}}$ (for the following equality) and by definition of $\overline{m}$ (for the following inclusion),

$$\overline{\overline{m}}(\boldsymbol{p},\boldsymbol{y}) = \sum_{b\in\mathcal{B}} y_b\, \overline{m}(\boldsymbol{p},b) \ \subseteq\ \overline{m}\left(\boldsymbol{p},\ \sum_{b\in\mathcal{B}} y_b\, b\right).$$

The existence of the desired $\boldsymbol{p}$ is therefore ensured by Condition 1, applied with $\sigma = \sum_{b\in\mathcal{B}} y_b\, b$. ■

We consider the strategy described in Figure 1. It forces exploration at a $\gamma$ rate, as is usual in situations with partial monitoring. One of its key ingredient, that conditionally unbiased estimators are available, is extracted from Lugosi et al. (2008, Section 6): in block $n$ we consider

$$\widehat{H}_t = \frac{\mathbb{I}_{\{S_t=s\}}\mathbb{I}_{\{I_t=i\}}}{p_{I_t,n}} \in \mathbb{R}^{\mathcal{H}\times\mathcal{I}};$$

averaging over the respective random draws of $I_t$ and $S_t$ according to $\boldsymbol{p}_n$ and $H(I_t, J_t)$, i.e., taking the conditional expectation $\mathbb{E}_t$ with respect to $\boldsymbol{p}_n$ and $J_t$, we get

$$\mathbb{E}_t\big[\widehat{H}_t\big] = \widetilde{H}\big(\delta_{J_t}\big). \tag{5}$$

This is why, by concentration-of-the-measure argument, we will be able to show that for $L$ large enough, $\widetilde{\sigma}_n$ is close to $\widetilde{H}\big(\widehat{\boldsymbol{q}}_n\big)$, where

$$\widehat{\boldsymbol{q}}_n = \frac{1}{L} \sum_{t=(n-1)L+1}^{nL} \delta_{J_t}. \tag{6}$$

Actually, since $\mathcal{F} \subseteq \Delta(\mathcal{H})^{\mathcal{I}}$, we have a natural embedding of $\mathcal{F}$ into $\mathbb{R}^{\mathcal{H}\times\mathcal{I}}$ and we can define $P_{\mathcal{F}}$, the convex projection operator onto $\mathcal{F}$ (in $\ell^2$–norm). Instead of using directly $\widetilde{\sigma}_n$, we consider in our strategy $\widehat{\sigma}_n = P_{\mathcal{F}}\big(\widetilde{\sigma}_n\big)$, which is even closer to $H\big(\widehat{\boldsymbol{q}}_n\big)$.

#### 4.2.3 Performance guarantee.

We provide a performance bound for fixed parameters $\gamma$ and $L$ tuned as functions of $T$. The proof is provided in Appendix A.3. Adaptation to $T \to \infty$ can be performed either by resorting to a standard doubling trick (see, e.g., Cesa-Bianchi and Lugosi 2006, page 17) or by taking time-varying parameters $\gamma_t$ and $L_t$.

**Theorem 13** *Under the assumptions of Lemma 12, consider the strategy of Figure 1, run with parameters $\gamma \in [0,1]$ and $L \geqslant 1$ and fed with a strategy $\Psi$ for $\overline{\overline{m}}$–approachability of $\mathcal{C}$, provided by the indicated lemma. Then, for all rounds $T \geqslant L+1$ and with probability at least $1-\delta$,*

$$\inf_{c\in\mathcal{C}} \left\| c - \frac{1}{T}\sum_{t=1}^{T} r(I_t,J_t) \right\|_2 \leqslant \frac{2L}{T}R + 4R\sqrt{\frac{\ln\bigl((2T)/(L\delta)\bigr)}{T}} + 2\gamma R + \frac{2R}{\sqrt{T/L-1}}\sqrt{N_{\mathcal{I}}N_{\mathcal{B}}} + R\kappa_\Phi N_{\mathcal{H}}\sqrt{N_{\mathcal{I}}}\left(\sqrt{\frac{2N_{\mathcal{I}}}{\gamma L}\ln\frac{2N_{\mathcal{I}}N_{\mathcal{H}}T}{L\delta}} + \frac{1}{3}\frac{N_{\mathcal{I}}}{\gamma L}\ln\frac{2N_{\mathcal{I}}N_{\mathcal{H}}T}{L\delta}\right).$$

*In particular, for all $T \geqslant 1$, the choices of $L = \lceil T^{3/5} \rceil$ and $\gamma = T^{-1/5}$ imply that with probability at least $1-\delta$,*

$$\inf_{c\in\mathcal{C}} \left\| c - \frac{1}{T}\sum_{t=1}^{T} r(I_t,J_t) \right\|_2 \leqslant \square\left(T^{-1/5}\sqrt{\ln\frac{T}{\delta}} + T^{-2/5}\ln\frac{T}{\delta}\right)$$

*for some constant $\square$ depending only on $\mathcal{C}$ and on the game $(r,\, H)$ at hand.*

The efficiency of the strategy of Figure 1 depends on whether it can be fed with an efficient approachability strategy $\Psi$, which in turn depends on the respective geometries of $\overline{m}$ and $\mathcal{C}$, as was indicated before the statement of Theorem 7. (Note that the projection onto $\mathcal{F}$ can be performed in polynomial time, as the latter closed convex set is defined by finitely many linear constraints, and that the computation of $\overline{\overline{m}}$ can be performed beforehand.)

## 5 Application to regret minimization

In this section we analyze external and internal regret minimization in repeated games with partial monitoring from the approachability perspective. Using the results developed for vector-valued games with partial monitoring, we show how to—in particular—minimize regret in both setups.

### 5.1 External regret

We consider in this section the framework and aim introduced by Rustichini (1999) and studied, sometimes in special cases, by Piccolboni and Schindelhauer (2001), Mannor and Shimkin (2003), Cesa-Bianchi et al. (2006), Lugosi et al. (2008). We show that our general strategy can be used for regret minimization.

Scalar payoffs are obtained (but not observed) by the first player: the payoff function $r$ is a mapping $\mathcal{I}\times\mathcal{J}\to\mathbb{R}$; we still denote by $R$ a bound on $|r|$. We define in this section

$$\widehat{\boldsymbol{q}}_T = \frac{1}{T}\sum_{t=1}^{T}\delta_{J_T}$$

as the empirical distribution of the actions taken by the second player. The external regret of the first player at round $T$ equals by definition

$$R_T^{\text{ext}} = \max_{\boldsymbol{p}\in\Delta(\mathcal{I})}\ \rho\Bigl(\boldsymbol{p}, \widetilde{H}\bigl(\widehat{\boldsymbol{q}}_T\bigr)\Bigr) - \frac{1}{T}\sum_{t=1}^{T} r(I_t,J_t)\,,$$

where $\rho : \Delta(\mathcal{I})\times\mathcal{F}$ is defined as follows: for all $\boldsymbol{p}\in\Delta(\mathcal{I})$ and $\sigma\in\mathcal{F}$,

$$\rho(\boldsymbol{p},\sigma) = \min\Bigl\{ r(\boldsymbol{p},\boldsymbol{q}) \,:\, \boldsymbol{q} \text{ such that } \widetilde{H}(\boldsymbol{q}) = \sigma \Bigr\}\,.$$

The function $\rho$ is continuous in its first argument and therefore the supremum in the defining expression of $R_T^{\text{ext}}$ is a maximum.

We recall briefly why, intuitively, this is the natural notion of external regret to consider in this case. Indeed, the first term in the definition of $R_T^{\text{ext}}$ is (close to) the worst-case average payoff obtained by the first player when playing consistently a mixed action $\boldsymbol{p}$ against a sequence of mixed actions inducing the same laws on the signals.

The following result is an easy consequence of Theorem 13, as is explained below; it corresponds to the main result of Lugosi et al. (2008), with the same convergence rate but with a different strategy. (However, Perchet 2011b, Section 2.3 exhibited an efficient strategy achieving a convergence rate of order $T^{-1/3}$, which is optimal; a question is thus whether the rates exhibited in Theorem 13 could be improved.)

**Corollary 14** *For all $T$, the first player has a strategy such that, for all strategies of the second player and with probability at least $1-\delta$,*

$$R_T^{ext} \leqslant \square \left( T^{-1/5} \sqrt{\ln \frac{T}{\delta}} + T^{-2/5} \ln \frac{T}{\delta} \right)$$

*for some constant $\square$ depending only on the game $(r,\ H)$ at hand.*

The proof below is an extension to the setting of partial monitoring of the original proof and strategy of Blackwell (1956b) for the case of external regret under full monitoring: in the case of full monitoring the vector-payoff function $\underline{r}$ and the set $\mathcal{C}$ considered in our proof are equal to the ones considered by Blackwell.

**Proof:** As usual, we embed $\Delta(\mathcal{J})$ into $\mathbb{R}^{\mathcal{J}}$ so that in this proof we will be working in the vector space $\mathbb{R} \times \mathbb{R}^{\mathcal{J}}$. We consider the convex set $\mathcal{C}$ and the vector-valued payoff function $\underline{r}$ respectively defined by

$$\mathcal{C} = \left\{ (z, \boldsymbol{q}) \in \mathbb{R} \times \Delta(\mathcal{J}) : \ z \geqslant \max_{\boldsymbol{p} \in \Delta(\mathcal{I})} \rho\big(\boldsymbol{p},\, \widetilde{H}(\boldsymbol{q})\big) \right\} \qquad \text{and} \qquad \underline{r}(i,j) = \left[ \begin{array}{c} r(i,j) \\ \delta_j \end{array} \right],$$

for all $(i,j) \in \mathcal{I} \times \mathcal{J}$. We now show that $\mathcal{C}$ is $\underline{r}$–approachable for $H$, i.e., by the results of Section 4, that Condition 1 is satisfied. To do so, we associate with each $\boldsymbol{q} \in \Delta(\mathcal{J})$ an element $\phi(\boldsymbol{q}) \in \Delta(\mathcal{I})$ such that

$$\phi(\boldsymbol{q}) \in \operatorname*{argmax}_{\boldsymbol{p} \in \Delta(\mathcal{I})} \rho\big(\boldsymbol{p},\, \widetilde{H}(\boldsymbol{q})\big)\,.$$

Then, given any $\boldsymbol{q} \in \Delta(\mathcal{J})$, we note that for all $\boldsymbol{q}'$ satisfying $\widetilde{H}(\boldsymbol{q}') = \widetilde{H}(\boldsymbol{q})$, we have, by definition of $\rho$,

$$r\big(\phi(\boldsymbol{q}),\, \boldsymbol{q}'\big) \geqslant \rho\big(\phi(\boldsymbol{q}),\, \widetilde{H}(\boldsymbol{q}')\big) = \max_{\boldsymbol{p} \in \Delta(\mathcal{I})} \rho\big(\boldsymbol{p},\, \widetilde{H}(\boldsymbol{q}')\big)\,,$$

which shows that $\underline{r}\big(\phi(\boldsymbol{q}),\, \boldsymbol{q}'\big) \in \mathcal{C}$. The required condition is thus satisfied.

To exhibit the convergence rates, we use the fact that the mapping

$$\boldsymbol{q} \in \Delta(\mathcal{J}) \longmapsto \max_{\boldsymbol{p} \in \Delta(\mathcal{I})} \rho\big(\boldsymbol{p}, \widetilde{H}(\boldsymbol{q})\big)$$

is Lipschitz, with Lipschitz constant in $\ell^2$–norm denoted by $L_\rho$; this fact is proved below. Now, the regret is non positive as soons as $\sum_{t=1}^T \underline{r}(I_t, J_t)/T$ belongs to $\mathcal{C}$; we therefore only need to consider the case when this average is not in $\mathcal{C}$. In the latter case, we denote by $(\widetilde{r}_T, \widetilde{\boldsymbol{q}}_T)$ its projection in $\ell^2$–norm onto $\mathcal{C}$. We have first that the defining inequality of $\mathcal{C}$ is an equality on its border, so that

$$\widetilde{r}_T = \max_{\boldsymbol{p} \in \Delta(\mathcal{I})} \ \rho\Big(\boldsymbol{p}, \widetilde{H}\big(\widetilde{\boldsymbol{q}}_T\big)\Big)\,;$$

and second, that

$$\begin{aligned}
R_T^{\text{ext}} &= \max_{\boldsymbol{p} \in \Delta(\mathcal{I})} \ \rho\Big(\boldsymbol{p}, \widetilde{H}\big(\widehat{\boldsymbol{q}}_T\big)\Big) - \frac{1}{T} \sum_{t=1}^T r(I_t, J_t) \\
&\leqslant \left| \max_{\boldsymbol{p} \in \Delta(\mathcal{I})} \ \rho\Big(\boldsymbol{p}, \widetilde{H}\big(\widehat{\boldsymbol{q}}_T\big)\Big) - \max_{\boldsymbol{p} \in \Delta(\mathcal{I})} \ \rho\Big(\boldsymbol{p}, \widetilde{H}\big(\widetilde{\boldsymbol{q}}_T\big)\Big) \right| + \left| \widetilde{r}_T - \frac{1}{T} \sum_{t=1}^T r(I_t, J_t) \right| \\
&\leqslant L_\rho \left\| \widehat{\boldsymbol{q}}_T - \widetilde{\boldsymbol{q}}_T \right\|_2 + \left| \widetilde{r}_T - \frac{1}{T} \sum_{t=1}^T r(I_t, J_t) \right| \\
&\leqslant \sqrt{2}\, \max\big\{L_\rho, 1\big\} \left\| \left[ \begin{array}{c} \widetilde{r}_T \\ \widetilde{\boldsymbol{q}}_T \end{array} \right] - \frac{1}{T} \sum_{t=1}^T \underline{r}(I_t, J_t) \right\|_2 \\
&= \sqrt{2}\, \max\big\{L_\rho, 1\big\} \inf_{c \in \mathcal{C}} \left\| c - \frac{1}{T} \sum_{t=1}^T \underline{r}(I_t, J_t) \right\|_2 .
\end{aligned}$$

The rates follow from the ones indicated in Theorem 13.

It only remains to prove the indicated Lipschitzness. (All Lipschitzness statements that follow will be with respect to the $\ell^2$–norm.) We prove that $L_\rho = R\kappa_\Phi\sqrt{N_\mathcal{I}N_\mathcal{J}N_\mathcal{B}}$. On the one hand, for every $i \in \mathcal{I}$, the mappings $\boldsymbol{q} \in \Delta(\mathcal{J}) \mapsto H(i,\boldsymbol{q})$ are $\sqrt{N_\mathcal{J}}$–Lipschitz as the $\|H(i,j)\|_2$ are bounded by 1 for all $j$. Thus, the mapping $\boldsymbol{q} \in \Delta(\mathcal{J}) \mapsto \widetilde{H}(\boldsymbol{q})$ is $\sqrt{N_\mathcal{I}N_\mathcal{J}}$–Lipschitz. On the other hand, we have by definition that for all $\boldsymbol{p} \in \Delta(\mathcal{I})$ and $\sigma \in \mathcal{F}$,

$$\rho(\boldsymbol{p},\sigma) = \min\ \overline{\overline{m}}\big(\boldsymbol{p},\Phi(\sigma)\big)\,,$$

and that (by Remark 1) the mapping $\sigma \in \mathcal{F} \mapsto \Phi(\sigma)$ is $\kappa_\Phi$–Lipschitz; this entails, by Lemma 6, that for all $\boldsymbol{p} \in \Delta(\mathcal{I})$, the mapping $\sigma \in \mathcal{F} \mapsto \rho(\boldsymbol{p},\sigma)$ is $R\sqrt{N_\mathcal{B}}\ \kappa_\Phi$–Lipschitz. In particular, since the latter Lipschitz constant is independent of $\boldsymbol{p}$, the mapping $\sigma \in \mathcal{F} \mapsto \max_{\boldsymbol{p}\in\Delta(\mathcal{I})}\rho(\boldsymbol{p},\sigma)$ is $R\sqrt{N_\mathcal{B}}\ \kappa_\Phi$–Lipschitz as well. Combining the two Lipschitz mappings yields yet another Lipschitz mapping, whose Lipschitz constant is the product of the Lipschitz constants of the base two mappings. ■

### 5.2 Internal / swap regret

Foster and Vohra (1999) defined internal regret with full monitoring as follows. A player has no internal regret if, for every action $i \in \mathcal{I}$, he has no external regret on the stages when this specific action $i$ was played. In other words, $i$ is the best response to the empirical distribution of action of the other player on these stages.

With partial monitoring, the first player evaluates his payoffs in some pessimistic way through the function $\rho$ defined above. This function is not linear over $\Delta(\mathcal{I})$ in general (it is concave), so that the best responses are not necessarily pure actions $i \in \mathcal{I}$ but mixed actions, i.e., elements of $\Delta(\mathcal{I})$. Following Lehrer and Solan (2007) we therefore should partition the stages not depending on the pure actions actually played but on the mixed actions $\boldsymbol{p}_t \in \Delta(\mathcal{I})$ used to draw them. To this end, it is convenient to assume that the strategies of the first player need to pick these mixed actions in a finite (but possibly thin) grid of $\Delta(\mathcal{I})$, which we denote by $\big\{\boldsymbol{p}_g,\ g \in \mathcal{G}\big\}$, where $\mathcal{G}$ is a finite set. At each round, the first player picks an index $G_t \in \mathcal{G}$ and uses the distribution $\boldsymbol{p}_{G_t}$ to draw his action $I_t$. Up to a standard concentration-of-the-measure argument, we will measure the payoff at round $t$ with $r\big(\boldsymbol{p}_{G_t},J_t\big)$ rather than with $r(I_t,J_t)$.

For each $g \in \mathcal{G}$, we denote by $N_T(g)$ the number of stages in $\{1,\dots,T\}$ for which we had $G_t = g$ and, whenever $N_T(g) > 0$,

$$\widehat{\boldsymbol{q}}_{T,g} = \frac{1}{N_T(g)}\sum_{t:G_t=g}\delta_{J_t}\,.$$

We define $\widehat{\boldsymbol{q}}_{T,g}$ is an arbitrary way when $N_T(g) = 0$. The internal regret of the first player at round $T$ is measured as

$$R_T^{\text{int}} = \max_{g,g'\in\mathcal{G}}\frac{N_T(g)}{T}\left(\rho\Big(\boldsymbol{p}_{g'},\widetilde{H}\big(\widehat{\boldsymbol{q}}_{T,g}\big)\Big) - r\big(\boldsymbol{p}_g,\widehat{\boldsymbol{q}}_{T,g}\big)\right).$$

Actually, our proof technique rather leads to the minimization of some swap regret (see Blum and Mansour, 2007 for the definition of swap regret in full monitoring):

$$R_T^{\text{swap}} = \sum_{g\in\mathcal{G}}\frac{N_T(g)}{T}\left(\max_{g'\in\mathcal{G}}\rho\Big(\boldsymbol{p}_{g'},\widetilde{H}\big(\widehat{\boldsymbol{q}}_{T,g}\big)\Big) - r\big(\boldsymbol{p}_g,\widehat{\boldsymbol{q}}_{T,g}\big)\right).$$

Again, the following bound on the swap regret easily follows from Theorem 13; the latter constructs a simple and direct strategy to control the swap regret, thus also the internal regret. It therefore improves on the results of Lehrer and Solan (2007), Perchet (2009), two articles which presented complicated strategies to do so (strategies based on auxiliary strategies using a grid that needs to be refined over time and whose complexities is exponential in the size of these grids). Moreover, we provide convergence rates.

**Corollary 15** *For all $T$, the first player has an explicit strategy such that, for all strategies of the second player and with probability at least $1-\delta$,*

$$R_T^{\text{swap}} \ \leqslant\ \square\left(T^{-1/5}\sqrt{\ln\frac{T}{\delta}} + T^{-2/5}\ln\frac{T}{\delta}\right)$$

*for some constant $\square$ depending only on the game $(r,\ H)$ at hand and on the size of the finite grid $\mathcal{G}$.*

The proof of this corollary is based on ideas similar to the ones used in the proof of Corollary 14; it is deferred to Appendix A.4.

**Acknowledgements**

Shie Mannor was partially supported by the ISF under contract 890015. Vianney Perchet benefited from the support of the ANR under grant ANR-10-BLAN 0112. Gilles Stoltz acknowledges support from the French National Research Agency (ANR) under grant EXPLO/RA ("Exploration–exploitation for efficient resource allocation") and by the PASCAL2 Network of Excellence under EC grant no. 506778.

## References

J. Abernethy, P. L. Bartlett, and E. Hazan. Blackwell approachability and low-regret learning are equivalent. In *Proceedings of the Twenty-Fourth Annual Conference on Learning Theory (COLT'11)*. Omnipress, 2011.

D. Blackwell. An analog of the minimax theorem for vector payoffs. *Pacific Journal of Mathematics*, 6:1–8, 1956a.

D. Blackwell. Controlled random walks. In *Proceedings of the International Congress of Mathematicians, 1954, Amsterdam, vol. III*, pages 336–338, 1956b.

A. Blum and Y. Mansour. From external to internal regret. *Journal of Machine Learning Research*, 8:1307–1324, 2007.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

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

N. Cesa-Bianchi, G. Lugosi, and G. Stoltz. Regret minimization under partial monitoring. *Mathematics of Operations Research*, 31:562–580, 2006.

X. Chen and H. White. Laws of large numbers for Hilbert space-valued mixingales with applications. *Econometric Theory*, 12:284–304, 1996.

D. Foster and R. Vohra. Regret in the on-line decision problem. *Games and Economic Behavior*, 29:7–36, 1999.

D.A. Freedman. On tail probabilities for martingales. *Annals of Probability*, 3:100–118, 1975.

J.E. Goodman and J. O'Rourke, editors. *Handbook of Discrete and Computational Geometry*. Discrete Mathematics and its Applications. Chapman & Hall/CRC, Boca Raton, FL, second edition, 2004.

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

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

E. Lehrer and E. Solan. Learning to play partially-specified equilibrium. Mimeo, 2007.

G. Lugosi, S. Mannor, and G. Stoltz. Strategies for prediction under imperfect monitoring. *Mathematics of Operations Research*, 33:513–528, 2008. An extended abstract was presented at COLT'07.

S. Mannor and N. Shimkin. On-line learning with imperfect monitoring. In *Proceedings of the Sixteenth Annual Conference on Learning Theory (COLT'03)*, pages 552–567. Springer, 2003.

S. Mannor and N. Shimkin. Regret minimization in repeated matrix games with variable stage duration. *Games and Economic Behavior*, 63(1):227–258, 2008.

S. Mannor, J. Tsitsiklis, and J. Y. Yu. Online learning with sample path constraints. *Journal of Machine Learning Research*, 10(Mar):569–590, 2009.

J.-F. Mertens, S. Sorin, and S. Zamir. Repeated games. Technical Report no. 9420, 9421, 9422, Université de Louvain-la-Neuve, 1994.

V. Perchet. Calibration and internal no-regret with random signals. In *Proceedings of the Twentieth International Conference on Algorithmic Learning Theory (ALT'09)*, pages 68–82, 2009.

V. Perchet. Approachability of convex sets in games with partial monitoring. *Journal of Optimization Theory and Applications*, 149:665–677, 2011a.

V. Perchet. Internal regret with partial monitoring calibration-based optimal algorithms. *Journal of Machine Learning Research*, 2011b. In press.

A. Piccolboni and C. Schindelhauer. Discrete prediction games with arbitrary feedback and loss. In *Proceedings of the Fourteenth Annual Conference on Computational Learning Theory (COLT'01)*, pages 208–223, 2001.

A. Rakhlin, K. Sridharan, and A. Tewari. Online learning: Beyond regret. In *Proceedings of the Twenty-Fourth Annual Conference on Learning Theory (COLT'11)*. Omnipress, 2011.

J. Rambau and G. Ziegler. Projections of polytopes and the generalized Baues conjecture. *Discrete and Computational Geometry*, 16:215–237, 1996.

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

# A Appendix beyond the COLT page limit

An conference version of this paper was published in the *Proceedings of the Twenty-Fourth Annual Conference on Learning Theory* (COLT'11); this appendix details some material which was alluded at in this conference version but could not be published therein because of the page limit.

### A.1 Proof of Condition 1 being necessary for approachability

When Condition 1 is not satisfied, there exists a vector $\sigma_0 \in \mathcal{F}$ such that for all $\boldsymbol{p} \in \Delta(\mathcal{I})$, there exists a $\phi(\boldsymbol{p}) \in \Delta(\mathcal{J})$ such that $\widetilde{H}\big(\phi(\boldsymbol{p})\big) = \sigma_0$ and $r\big(\boldsymbol{p}, \phi(\boldsymbol{p})\big) \notin \mathcal{C}$. We denote by $\widetilde{H}^{-1}(\sigma_0)$ the set of all $\boldsymbol{q}$ such that $\widetilde{H}(\boldsymbol{q}) = \sigma_0$. We will consider a small subclass of the possible strategies of the second player: only those which prescribe him to play at each round the same element of $\widetilde{H}^{-1}(\sigma_0)$. We will show that for all strategies of the first player, there exists a strategy of the second player of the form mentioned above such that, with some *positive* probability,

$$\limsup_{T\to\infty} \quad \inf_{c\in\mathcal{C}} \left\| c - \frac{1}{T}\sum_{t=1}^{T} r(I_t, J_t) \right\|_2 > 0\,. \tag{7}$$

We first note that by concentration of the measure (by the Hoeffding-Azuma inequality and the Borel-Cantelli lemma), if $\boldsymbol{q} \in \widetilde{H}^{-1}(\sigma_0)$ is the element repeatedly played by the second player,

$$\lim_{T\to\infty} \quad \left\| \frac{1}{T}\sum_{t=1}^{T} r(I_t, \boldsymbol{q}) - \frac{1}{T}\sum_{t=1}^{T} r(I_t, J_t) \right\|_2 = 0 \qquad \text{a.s.} \tag{8}$$

Now, all considered strategies of the second player are indistinguishable to the first player, since they all induce the same vector $\sigma_0$ of probability distributions over signals. Therefore, the law of

$$\widehat{\boldsymbol{p}}_T = \frac{1}{T}\sum_{t=1}^{T} \delta_{I_t}$$

only depends on $T$ and $\sigma_0$ (and on the strategy of the first player). We denote by $\overline{\boldsymbol{p}}_T$ the common expectation of the $\widehat{\boldsymbol{p}}_T$ as the $\boldsymbol{q}$ vary in $\widetilde{H}^{-1}(\sigma_0)$; the expectation has to be understood with respect to the auxiliary randomizations taken (to draw the pure actions from the mixed actions $\boldsymbol{p}_t$ and $\boldsymbol{q}$ and to draw the signals).

We denote by $d_{\mathcal{C}}$ the Euclidian distance to the closed convex set $\mathcal{C}$; it is a continuous and convex function (see Boyd and Vandenberghe 2004, Example 3.16). In particular, it is bounded on the set of all feasible payoff vectors $r(\boldsymbol{p}, \boldsymbol{q})$, as $\boldsymbol{p}$ and $\boldsymbol{q}$ vary. By the dominated convergence theorem and in view of (8), to prove (7) it thus suffices to show that for all strategies of the first player, there exists a strategy of the second player in $\widetilde{H}^{-1}(\sigma_0)$ such that

$$\limsup_{T\to\infty} \quad \mathbb{E}\left[ \inf_{c\in\mathcal{C}} \left\| c - \frac{1}{T}\sum_{t=1}^{T} r(I_t, \boldsymbol{q}) \right\|_2 \right] = \limsup_{T\to\infty} \quad \mathbb{E}\Big[ d_{\mathcal{C}}\Big( r\big(\widehat{\boldsymbol{p}}_T, \boldsymbol{q}\big)\Big)\Big] > 0\,.$$

By Jensen's inequality,

$$\mathbb{E}\Big[ d_{\mathcal{C}}\Big( r\big(\widehat{\boldsymbol{p}}_T, \boldsymbol{q}\big)\Big)\Big] \geqslant d_{\mathcal{C}}\Big( \mathbb{E}\Big[ r\big(\widehat{\boldsymbol{p}}_T, \boldsymbol{q}\big)\Big]\Big) = d_{\mathcal{C}}\Big( r\big(\overline{\boldsymbol{p}}_T, \boldsymbol{q}\big)\Big)\,.$$

By the Bolzano-Weierstrass property, for all strategies of the first player, the sequence of the $\overline{\boldsymbol{p}}_T$ has values in the compact space $\Delta(\mathcal{I})$; thus, it admits a converging subsequence, which we denote by $\overline{\boldsymbol{p}}_{\varphi(T)}$ and whose limit point we denote by $\overline{\boldsymbol{p}}_\infty$. (This limit point depends solely on the strategy of the first player and on $\sigma_0$.) By considering $\boldsymbol{q} = \phi\big(\overline{\boldsymbol{p}}_\infty\big)$ and putting the pieces together, we get that

$$\limsup_{T\to\infty} \quad \mathbb{E}\Big[ d_{\mathcal{C}}\Big( r\big(\widehat{\boldsymbol{p}}_T, \boldsymbol{q}\big)\Big)\Big] \geqslant \limsup_{T\to\infty} \quad d_{\mathcal{C}}\Big( r\big(\overline{\boldsymbol{p}}_{\varphi(T)}, \boldsymbol{q}\big)\Big) = d_{\mathcal{C}}\left( r\Big(\overline{\boldsymbol{p}}_\infty, \phi\big(\overline{\boldsymbol{p}}_\infty\big)\Big)\right) > 0\,,$$

since by definition of $\phi$, the vector $r\big(\overline{\boldsymbol{p}}_\infty, \phi(\overline{\boldsymbol{p}}_\infty)\big)$ is not in the closed convex set $\mathcal{C}$.

### A.2 Proof of Lemma 10

**Proof:** Rambau and Ziegler (1996, Proposition 2.4) state that since $\widetilde{H}$ is linear on the polytope $\Delta(\mathcal{J})$, its inverse application $\widetilde{H}^{-1}$ is a piecewise linear mapping of $\mathcal{F}$ into the subsets of $\Delta(\mathcal{J})$, which means that there exists a finite decomposition of $\mathcal{F}$ into polytopes $\{P_1, \ldots, P_K\}$ each on which $\widetilde{H}^{-1}$ is linear. Up to a

triangulation (see, e.g. Goodman and O'Rourke 2004, Chapter 14), we can assume that each $P_k$ is a simplex. Denote by $\mathcal{B}_k \subseteq \mathcal{F}$ the set of vertices of $P_k$; then, the finite subset stated in the lemma is

$$\mathcal{B} = \bigcup_{k=1}^{K} \mathcal{B}_k \,,$$

the set of all vertices of all the simplices.

Fix any $\sigma \in \mathcal{F}$. It belongs to some simplex $P_k$, so that there exists a convex decomposition $\sigma = \sum_{b \in \mathcal{B}_k} \lambda_b \, b$; this decomposition is unique within the simplex $P_k$. If $\sigma$ belongs to two different simplices, then it actually belongs to their common face and the two possible decompositions coincide (some coefficients $\lambda_b$ in the above decomposition are null). All in all, with each $\sigma \in \mathcal{F}$, we can associate a unique decomposition in $\mathcal{B}$,

$$\sigma = \sum_{b \in \mathcal{B}} \Phi_b(\sigma) \, b \,,$$

where the coefficients $\big(\Phi_b(\sigma)\big)_{b \in \mathcal{B}}$ form a convex weight vector over $\mathcal{B}$, i.e., belong to $\Delta(\mathcal{B})$; in addition, $\Phi_b(\sigma) > 0$ only if $b \in \mathcal{B}_k$, where $k$ is such that $\sigma \in P_k$.

Since $\widetilde{H}^{-1}$ is linear on each simplex $P_1, \, \ldots, \, P_K$, we therefore get

$$\widetilde{H}^{-1}(\sigma) = \sum_{b \in \mathcal{B}} \Phi_b(\sigma) \, \widetilde{H}^{-1}(b) \,.$$

Finally, the result is a consequence of the fact that

$$\overline{m}(\boldsymbol{p}, \sigma) = r\Big(\boldsymbol{p}, \, \widetilde{H}^{-1}(\sigma)\Big) = r\left(\boldsymbol{p}, \sum_{b \in \mathcal{B}} \Phi_b(\sigma) \, \widetilde{H}^{-1}(b)\right) \,,$$

which implies, by linearity of $r$, that

$$\overline{m}(\boldsymbol{p}, \sigma) = \sum_{b \in \mathcal{B}} \Phi_b(\sigma) \, r\Big(\boldsymbol{p}, \widetilde{H}^{-1}(b)\Big) = \sum_{b \in \mathcal{B}} \Phi_b(\sigma) \, \overline{m}(\boldsymbol{p}, b) \,.$$

The proof is concluded by noting that by definition, for all $\sigma \in \mathcal{F}$, the applications $\boldsymbol{p} \in \Delta(\mathcal{I}) \mapsto \overline{m}(\boldsymbol{p}, \sigma)$ are linear. ■

### A.3 Proof of Theorem 13

**Proof:** We write $T$ as $T = NL + k$ where $N$ is an integer and $0 \leqslant k \leqslant L - 1$ and will show successively that (possibly with overwhelming probability only) the following statements hold.

$$\frac{1}{T} \sum_{t=1}^{T} r(I_t, J_t) \qquad \text{is close to} \qquad \frac{1}{NL} \sum_{t=1}^{NL} r(I_t, J_t) \,; \tag{9}$$

$$\frac{1}{NL} \sum_{t=1}^{NL} r(I_t, J_t) \qquad \text{is close to} \qquad \frac{1}{N} \sum_{n=1}^{N} r\big(\boldsymbol{p}_n, \, \widehat{\boldsymbol{q}}_n\big) \,; \tag{10}$$

$$\frac{1}{N} \sum_{n=1}^{N} r\big(\boldsymbol{p}_n, \, \widehat{\boldsymbol{q}}_n\big) \qquad \text{is close to} \qquad \frac{1}{N} \sum_{n=1}^{N} r\big(\boldsymbol{x}_n, \, \widehat{\boldsymbol{q}}_n\big) \,; \tag{11}$$

$$\frac{1}{N} \sum_{n=1}^{N} r\big(\boldsymbol{x}_n, \, \widehat{\boldsymbol{q}}_n\big) \qquad \text{belongs to the set} \qquad \frac{1}{N} \sum_{n=1}^{N} \overline{m}\Big(\boldsymbol{x}_n, \, \widetilde{H}\big(\widehat{\boldsymbol{q}}_n\big)\Big) \,;$$

$$\frac{1}{N} \sum_{n=1}^{N} \overline{m}\Big(\boldsymbol{x}_n, \, \widetilde{H}\big(\widehat{\boldsymbol{q}}_n\big)\Big) \qquad \text{is equal to the set} \qquad \frac{1}{N} \sum_{n=1}^{N} \overline{\overline{m}}\Big(\boldsymbol{x}_n, \, \Phi\Big(\widetilde{H}\big(\widehat{\boldsymbol{q}}_n\big)\Big)\Big) \,;$$

$$\frac{1}{N} \sum_{n=1}^{N} \overline{\overline{m}}\Big(\boldsymbol{x}_n, \, \Phi\Big(\widetilde{H}\big(\widehat{\boldsymbol{q}}_n\big)\Big)\Big) \qquad \text{is close to the set} \qquad \frac{1}{N} \sum_{n=1}^{N} \overline{\overline{m}}\Big(\boldsymbol{x}_n, \, \Phi\big(\widehat{\sigma}_n\big)\Big) \,; \tag{12}$$

$$\frac{1}{N} \sum_{n=1}^{N} \overline{\overline{m}}\Big(\boldsymbol{x}_n, \, \Phi\big(\widehat{\sigma}_n\big)\Big) \qquad \text{is close to the set} \qquad \mathcal{C} \,; \tag{13}$$

where we recall that the notation $\widehat{\boldsymbol{q}}_n$ was defined in (6). Actually, we will show below the numbered statements only; the first unnumbered statement is immediate by the very definition of $\overline{m}$ and the second one follows from Definition 11.

**Step 1: the term** (9)**.** A direct calculation decomposing the sum over $T$ elements into a sum over the $NL$ first elements and the $k$ remaining ones shows that

$$\left\| \frac{1}{T}\sum_{t=1}^{T} r(I_t, J_t) - \frac{1}{NL}\sum_{t=1}^{NL} r(I_t, J_t) \right\|_2 \leqslant R\left(\frac{k}{T} + \left(\frac{1}{NL} - \frac{1}{T}\right)NL\right) = \frac{2k}{T}R \leqslant \frac{2L}{T}R\,.$$

**Step 2: the term** (10)**.** We note that by defining $\mathbb{E}_t$ the conditional expectation with respect to $(I_1, S_1, J_1)$, …, $(I_{t-1}, S_{t-1}, J_{t-1})$ and $J_t$, which fixes the values of the law $\boldsymbol{p}'_t$ of $I_t$ and the value of $J_t$, we have

$$\mathbb{E}_t\big[r(I_t, J_t)\big] = r(\boldsymbol{p}'_t, J_t)\,.$$

We note that by definition of the forecaster, $\boldsymbol{p}'_t = \boldsymbol{p}_n$ if $t$ belongs to the $n$–th block. By a version of the Hoeffding-Azuma inequality for sums of Hilbert space-valued martingale differences proved in[3] Chen and White (1996, Lemma 3.2), we therefore get that with probability at least $1-\delta$,

$$\left\| \frac{1}{NL}\sum_{t=1}^{NL} r(I_t, J_t) - \frac{1}{N}\sum_{n=1}^{N} r\big(\boldsymbol{p}_n, \widehat{\boldsymbol{q}}_n\big) \right\|_2 \leqslant 4R\sqrt{\frac{\ln(2/\delta)}{T}}\,.$$

**Step 3: the term** (11)**.** Since by definition $\boldsymbol{p}_n = (1-\gamma)\,\boldsymbol{x}_n + \gamma\,\boldsymbol{u}$, we get

$$\left\| \frac{1}{N}\sum_{n=1}^{N} r\big(\boldsymbol{p}_n, \widehat{\boldsymbol{q}}_n\big) - \frac{1}{N}\sum_{n=1}^{N} r\big(\boldsymbol{x}_n, \widehat{\boldsymbol{q}}_n\big) \right\|_2 \leqslant 2\gamma R\,.$$

**Step 4: the term** (12)**.** We fix a given block $n$. It can be extracted from Lugosi et al. (2008, proof of Theorem 6.1) that with probability $1-\delta$,

$$\left\| \widehat{\sigma}_n - \widetilde{H}\big(\widehat{\boldsymbol{q}}_n\big) \right\|_2 \leqslant \sqrt{N_{\mathcal{I}} N_{\mathcal{H}}}\left(\sqrt{\frac{2N_{\mathcal{I}}}{\gamma L}\ln\frac{2N_{\mathcal{I}}N_{\mathcal{H}}}{\delta}} + \frac{1}{3}\frac{N_{\mathcal{I}}}{\gamma L}\ln\frac{2N_{\mathcal{I}}N_{\mathcal{H}}}{\delta}\right). \tag{14}$$

(For the sake of completeness this extracted statement is however proved again in Appendix B.2 below.) Since $\Phi$ is Lipschitz (see Remark 1), with Lipschitz constant in $\ell^2$–norm denoted by $\kappa_\Phi$, we get that with probability $1-\delta$,

$$\left\| \Phi\big(\widehat{\sigma}_n\big) - \Phi\Big(\widetilde{H}\big(\widehat{\boldsymbol{q}}_n\big)\Big) \right\|_2 \leqslant \kappa_\Phi\sqrt{N_{\mathcal{I}} N_{\mathcal{H}}}\left(\sqrt{\frac{2N_{\mathcal{I}}}{\gamma L}\ln\frac{2N_{\mathcal{I}}N_{\mathcal{H}}}{\delta}} + \frac{1}{3}\frac{N_{\mathcal{I}}}{\gamma L}\ln\frac{2N_{\mathcal{I}}N_{\mathcal{H}}}{\delta}\right).$$

By a union bound, the above bound holds for all blocks $n = 1, \ldots, N$ with probability at least $1 - N\delta$. Finally, an application of Lemma 6 shows that

$$\frac{1}{N}\sum_{n=1}^{N}\overline{\overline{m}}\Big(\boldsymbol{x}_n,\, \Phi\Big(\widetilde{H}\big(\widehat{\boldsymbol{q}}_n\big)\Big)\Big) \qquad \text{is in a } \varepsilon_T\text{–neighborhood (in } \ell^2\text{–norm) of} \qquad \frac{1}{N}\sum_{n=1}^{N}\overline{\overline{m}}\Big(\boldsymbol{x}_n,\, \Phi\big(\widehat{\sigma}_n\big)\Big)\,,$$

where

$$\varepsilon_T = R\sqrt{N_{\mathcal{H}}} \;\times\; \kappa_\Phi\sqrt{N_{\mathcal{I}} N_{\mathcal{H}}}\left(\sqrt{\frac{2N_{\mathcal{I}}}{\gamma L}\ln\frac{2N_{\mathcal{I}}N_{\mathcal{H}}}{\delta}} + \frac{1}{3}\frac{N_{\mathcal{I}}}{\gamma L}\ln\frac{2N_{\mathcal{I}}N_{\mathcal{H}}}{\delta}\right).$$

**Step 5: the term** (13)**.** Since $\mathcal{C}$ is $\overline{\overline{m}}$–approachable and by definition of the choices of the $\boldsymbol{x}_n$ in Figure 1, we get by Theorem 7, with probability 1,

$$\inf_{c\in\mathcal{C}} \left\| c - \frac{1}{N}\sum_{n=1}^{N}\overline{\overline{m}}\Big(\boldsymbol{x}_n,\, \Phi\big(\widehat{\sigma}_n\big)\Big) \right\|_2 \leqslant \frac{2R}{\sqrt{N}}\sqrt{N_{\mathcal{I}} N_{\mathcal{B}}} \leqslant \frac{2R}{\sqrt{T/L - 1}}\sqrt{N_{\mathcal{I}} N_{\mathcal{B}}}\,,$$

since $T/L \leqslant N + k/L \leqslant N+1$.

The proof is concluded by putting the pieces together, thanks to a triangle inequality, by noting that $T/L \leqslant N+1$, and by considering $L\delta/T \leqslant \delta/(N+1)$ instead of $\delta$. ■

[3] We use the fact that $\sqrt{u}\,e^{-u} \leqslant e^{-u/2}$ for all $u \geqslant 0$.

### A.4 Proof of Corollary 15

We provide here the existence proof of strategies minimizing the swap regret. The proof follows the same lines as the one of Corollary 14.

**Proof:** In this proof we will be working in the vector space $\big(\mathbb{R} \times \mathbb{R}^{\mathcal{J}}\big)^{\mathcal{G}}$. We first extend linearly $r$ from $\Delta(\mathcal{I}) \times \Delta(\mathcal{J})$ to $\Delta(\mathcal{I}) \times \mathbb{R}_+^{\mathcal{J}}$ and extend also continuously (but not linearly) $\rho : \mathcal{G} \times \Delta(\mathcal{J}) \to \mathbb{R}$ into a mapping $\zeta : \mathcal{G} \times \mathbb{R}_+^{\mathcal{J}} \to \mathbb{R}$ as follows: for all $g \in \mathcal{G}$ and $\boldsymbol{v} \in \mathbb{R}_+^{\mathcal{J}}$,

$$\zeta(\boldsymbol{p}_g, \boldsymbol{v}) = \begin{cases} 0 & \text{if } \|\boldsymbol{v}\|_1 = 0, \\ \|\boldsymbol{v}\|_1 \; \rho\Big(\boldsymbol{p}_g, \, \widetilde{H}\Big(\dfrac{\boldsymbol{v}}{\|\boldsymbol{v}\|_1}\Big)\Big) & \text{if } \|\boldsymbol{v}\|_1 > 0. \end{cases}$$

The convex set $\mathcal{C}$ and the vector-valued payoff function $\underline{r}$ are then respectively defined by

$$\mathcal{C} = \Big\{(z_g, \boldsymbol{v}_g) \in \big(\mathbb{R} \times \mathbb{R}_+^{\mathcal{J}}\big)^{\mathcal{G}} : \;\; \forall\, g \in \mathcal{G}, \;\; z_g \geqslant \max_{g' \in \mathcal{G}} \zeta\big(\boldsymbol{p}_{g'}, \boldsymbol{v}_g\big)\Big\}$$

and, for all $(g, j) \in \mathcal{G} \times \mathcal{J}$,

$$\underline{r}(g, j) = \left[\begin{array}{c} r\big(\boldsymbol{p}_g, j\big)\, \mathbb{I}_{\{g'=g\}} \\ \delta_j\, \mathbb{I}_{\{g'=g\}} \end{array}\right]_{g' \in \mathcal{G}}.$$

To show that $\mathcal{C}$ is $\underline{r}$–approachable, we associate with each $\boldsymbol{q} \in \Delta(\mathcal{J})$ an element $g^\star(\boldsymbol{q}) \in \mathcal{G}$ such that

$$g^\star(\boldsymbol{q}) \in \operatorname*{argmax}_{g \in \mathcal{G}} \rho\big(\boldsymbol{p}_g, \, \widetilde{H}(\boldsymbol{q})\big)\,.$$

Then, given any $\boldsymbol{q} \in \Delta(\mathcal{J})$, we note that for all $\boldsymbol{q}'$ satisfying $\widetilde{H}(\boldsymbol{q}') = \widetilde{H}(\boldsymbol{q})$, the components of the vector $\underline{r}\big(g^\star(\boldsymbol{q}), \boldsymbol{q}'\big)$ are all null but the ones corresponding to $g^\star(\boldsymbol{q})$, for which we have

$$\max_{g' \in \mathcal{G}} \zeta\big(\boldsymbol{p}_{g'}, \boldsymbol{q}'\big) = \max_{g' \in \mathcal{G}} \rho\Big(\boldsymbol{p}_{g'}, \widetilde{H}\big(\boldsymbol{q}'\big)\Big) = \max_{g' \in \mathcal{G}} \rho\Big(\boldsymbol{p}_{g'}, \widetilde{H}\big(\boldsymbol{q}\big)\Big) = \rho\Big(\boldsymbol{p}_{g^\star(\boldsymbol{q})}, \widetilde{H}\big(\boldsymbol{q}\big)\Big) \leqslant r\big(\boldsymbol{p}_{g^\star(\boldsymbol{q})}, \boldsymbol{q}'\big)\,,$$

where the last line is by definition of $\rho$. Therefore, $\underline{r}\big(g^\star(\boldsymbol{q}), \boldsymbol{q}'\big) \in \mathcal{C}$. The required condition in Lemma 12 and Theorem 13 is thus satisfied, hence the desired approachability.

We now exhibit the convergence rates. As in the proof of Corollary 14, we need only to consider the case where $\sum_{t=1}^T \underline{r}(I_t, J_t)/T$ is not in $\mathcal{C}$, for otherwise, the swap regret is non positive. We denote by $(\widetilde{r}_{T,g}, \widetilde{\boldsymbol{v}}_{T,g})_{g \in G}$ the projection in $\ell^2$–norm of $\sum_{t=1}^T \underline{r}(I_t, J_t)/T$ onto $\mathcal{C}$, and by $\widehat{\boldsymbol{v}}_{T,g} = \big(N_T(g)/T\big)\, \widehat{\boldsymbol{q}}_{T,g}$ the realized frequency of playing each $g \in \mathcal{G}$. Since the projection lies on the border of $\mathcal{C}$, we have that for all $g \in \mathcal{G}$,

$$\widetilde{r}_{T,g} = \max_{g' \in \mathcal{G}} \zeta\big(\boldsymbol{p}_{g'}, \widetilde{\boldsymbol{v}}_{T,g}\big)\,.$$

We will prove below that

$$\boldsymbol{v} \in \mathbb{R}_+^{\mathcal{J}} \;\longmapsto\; \max_{g \in \mathcal{G}} \zeta\big(\boldsymbol{p}_g, \boldsymbol{v}\big)$$

is $L_\zeta$–Lipschitz, for some constant $L_\zeta > 0$. Then, as for the external regret,

$$\begin{aligned}
R_T^{\text{swap}} &= \sum_{g \in \mathcal{G}} \frac{N_T(g)}{T} \Big(\max_{g' \in \mathcal{G}} \rho\Big(\boldsymbol{p}_{g'}, \widetilde{H}\big(\widehat{\boldsymbol{q}}_{T,g}\big)\Big) - r\big(\boldsymbol{p}_g, \widehat{\boldsymbol{q}}_{T,g}\big)\Big) \\
&= \sum_{g \in \mathcal{G}} \Big(\max_{g' \in \mathcal{G}} \zeta\big(\boldsymbol{p}_{g'}, \widehat{\boldsymbol{v}}_{T,g}\big) - r\big(\boldsymbol{p}_g, \widehat{\boldsymbol{v}}_{T,g}\big)\Big) \\
&\leqslant \sum_{g \in \mathcal{G}} \Bigg(\Big|\max_{g' \in \mathcal{G}} \zeta\big(\boldsymbol{p}_{g'}, \widehat{\boldsymbol{v}}_{T,g}\big) - \max_{g' \in \mathcal{G}} \zeta\big(\boldsymbol{p}_{g'}, \widetilde{\boldsymbol{v}}_{T,g}\big)\Big| + \big|\, \widetilde{r}_{T,g} - r\big(\boldsymbol{p}_g, \widehat{\boldsymbol{v}}_{T,g}\big)\big|\Bigg) \\
&\leqslant \sum_{g \in \mathcal{G}} \Big(L_\zeta \big\|\, \widehat{\boldsymbol{v}}_{T,g} - \widetilde{\boldsymbol{v}}_{T,g} \,\big\|_2 + \big|\, \widetilde{r}_{T,g} - r\big(\boldsymbol{p}_g, \widehat{\boldsymbol{v}}_{T,g}\big)\big|\Big) \\
&\leqslant \sqrt{2N_{\mathcal{G}}}\, \max\big\{L_\zeta, 1\big\} \left\| \left[\begin{array}{c} \widetilde{r}_{T,g} \\ \widetilde{\boldsymbol{v}}_{T,g} \end{array}\right]_{g \in \mathcal{G}} - \; \frac{1}{T} \sum_{t=1}^T \underline{r}(I_t, J_t) \right\|_2 \\
&= \sqrt{2N_{\mathcal{G}}}\, \max\big\{L_\zeta, 1\big\} \; \inf_{c \in \mathcal{C}} \left\| c - \frac{1}{T} \sum_{t=1}^T \underline{r}(I_t, J_t) \right\|_2,
\end{aligned}$$

where we denoted by $N_{\mathcal{G}}$ the cardinality of $\mathcal{G}$. Resorting to the convergence rate stated in Theorem 13 concludes the proof, up to the stated Lipschitzness of $\zeta$, which we now prove.

It suffices to show that for all fixed elements $\boldsymbol{p} \in \Delta(\mathcal{I})$, the functions $\boldsymbol{v} \in \mathbb{R}_+^{\mathcal{J}} \mapsto \zeta(\boldsymbol{p}, \boldsymbol{v})$ are Lipschitz, with a Lipschitz constant that is independent of $\boldsymbol{p}$.

Consider two elements $\boldsymbol{v},\, \boldsymbol{v}' \in \mathbb{R}_+^{\mathcal{J}}$. If $\|\boldsymbol{v}'\| = 0$ and $\|v\|_1 > 0$, then

$$\left|\zeta(\boldsymbol{p}, \boldsymbol{v}) - \zeta(\boldsymbol{p}, \boldsymbol{v}')\right| = \left|\zeta(\boldsymbol{p}, \boldsymbol{v})\right| = \|\boldsymbol{v}\|_1 \left| \rho\left(\boldsymbol{p},\, \widetilde{H}\left(\frac{\boldsymbol{v}}{\|\boldsymbol{v}\|_1}\right)\right)\right| \leqslant R\,\|\boldsymbol{v}\|_1 = R\,\|\boldsymbol{v} - \boldsymbol{v}'\|_1 \ .$$

In the case where both $\boldsymbol{v}$ and $\boldsymbol{v}'$ are non zero,

$$\begin{aligned}
&\zeta(\boldsymbol{p}, \boldsymbol{v}) - \zeta(\boldsymbol{p}, \boldsymbol{v}') \\
&= \|\boldsymbol{v}\|_1\, \rho\left(\boldsymbol{p},\, \widetilde{H}\left(\frac{\boldsymbol{v}}{\|\boldsymbol{v}\|_1}\right)\right) - \|\boldsymbol{v}'\|_1\, \rho\left(\boldsymbol{p},\, \widetilde{H}\left(\frac{\boldsymbol{v}'}{\|\boldsymbol{v}'\|_1}\right)\right) \\
&= \|\boldsymbol{v}\|_1 \left[\rho\left(\boldsymbol{p},\, \widetilde{H}\left(\frac{\boldsymbol{v}}{\|\boldsymbol{v}\|_1}\right)\right) - \rho\left(\boldsymbol{p},\, \widetilde{H}\left(\frac{\boldsymbol{v}'}{\|\boldsymbol{v}'\|_1}\right)\right)\right] + \left(\|\boldsymbol{v}\|_1 - \|\boldsymbol{v}'\|_1\right) \rho\left(\boldsymbol{p},\, \widetilde{H}\left(\frac{\boldsymbol{v}'}{\|\boldsymbol{v}'\|_1}\right)\right).
\end{aligned}$$

Therefore, by using the Lipschitzness proved at the end of the proof of Corollary 14, by two applications of the triangle inequality, and by noting that $\|\cdot\|_2 \leqslant \|\cdot\|_1 \leqslant \sqrt{N_{\mathcal{J}}}\ \|\cdot\|_2$, we get

$$\begin{aligned}
\left|\zeta(\boldsymbol{p}, \boldsymbol{v}) - \zeta(\boldsymbol{p}, \boldsymbol{v}')\right| &\leqslant \|\boldsymbol{v}\|_1\, L_\rho \left\| \frac{\boldsymbol{v}}{\|\boldsymbol{v}\|_1} - \frac{\boldsymbol{v}'}{\|\boldsymbol{v}'\|_1} \right\|_2 + \|\boldsymbol{v} - \boldsymbol{v}'\|_1\, R \\
&\leqslant L_\rho \left\| \boldsymbol{v} - \boldsymbol{v}' + \left(1 - \frac{\|\boldsymbol{v}\|_1}{\|\boldsymbol{v}'\|_1}\right) \boldsymbol{v}' \right\|_2 + R\sqrt{N_{\mathcal{J}}}\ \|\boldsymbol{v} - \boldsymbol{v}'\|_2 \\
&\leqslant \left(L_\rho + R\sqrt{N_{\mathcal{J}}}\right) \|\boldsymbol{v} - \boldsymbol{v}'\|_2 + L_\rho \left|1 - \frac{\|\boldsymbol{v}\|_1}{\|\boldsymbol{v}'\|_1}\right| \|\boldsymbol{v}'\|_2 \\
&\leqslant \left(L_\rho + R\sqrt{N_{\mathcal{J}}}\right) \|\boldsymbol{v} - \boldsymbol{v}'\|_2 + L_\rho \left|\|\boldsymbol{v}'\|_1 - \|\boldsymbol{v}\|_1\right| \frac{\|\boldsymbol{v}'\|_2}{\|\boldsymbol{v}'\|_1} \\
&\leqslant \left(L_\rho + R\sqrt{N_{\mathcal{J}}}\right) \|\boldsymbol{v} - \boldsymbol{v}'\|_2 + L_\rho\, \|\boldsymbol{v} - \boldsymbol{v}'\|_1 \\
&\leqslant \left(L_\rho + (R + L_\rho)\sqrt{N_{\mathcal{J}}}\right) \|\boldsymbol{v} - \boldsymbol{v}'\|_2 \ .
\end{aligned}$$

We therefore proved the required Lipschitzness, with constant $L_\zeta = L_\rho + (R + L_\rho)\sqrt{N_{\mathcal{J}}}$. ■

## B Proofs of results extracted from other works

The proofs below reproduce arguments that were published elsewhere; we rewrite them with our notation only for the convenience of the readers and to make this paper fully self-contained.

### B.1 Proof of the basic approachability results

This material is standard and can be found, e.g., in Cesa-Bianchi and Lugosi (2006, Section 7.7 and Exercise 7.23).

**Proof: (of Theorem 3)** Since $c_t$ is the projection of $\widehat{m}_t$ on the closed convex set $\mathcal{C}$ with respect to the $\ell^2$–norm, the following geometric property is satisfied:

$$\forall\, c \in \mathcal{C}, \qquad \langle \widehat{m}_t - c_t,\, c - c_t \rangle \leqslant 0\,.$$

By assumption, for every $\boldsymbol{y} \in \Delta(\mathcal{B})$, there exists $\boldsymbol{x} \in \Delta(\mathcal{A})$ such that $m(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{C}$; therefore, the above-stated geometric property implies that

$$\max_{\boldsymbol{y}\in\Delta(\mathcal{B})} \min_{\boldsymbol{x}\in\Delta(\mathcal{A})} \langle \widehat{m}_t - c_t,\, m(\boldsymbol{x}, \boldsymbol{y}) - c_t \rangle \leqslant 0\,.$$

By von Neumann's minimax theorem,

$$\max_{\boldsymbol{y}\in\Delta(\mathcal{B})} \min_{\boldsymbol{x}\in\Delta(\mathcal{A})} \langle \widehat{m}_t - c_t,\, m(\boldsymbol{x}, \boldsymbol{y}) - c_t \rangle = \min_{\boldsymbol{x}\in\Delta(\mathcal{A})} \max_{\boldsymbol{y}\in\Delta(\mathcal{B})} \langle \widehat{m}_t - c_t,\, m(\boldsymbol{x}, \boldsymbol{y}) - c_t \rangle \leqslant 0\,.$$

In view of the defining minimax choice of $\boldsymbol{x}_{t+1} \in \Delta(\mathcal{A})$, the above inequality yields that for all $\boldsymbol{z}_{t+1} \in \Delta(\mathcal{B})$,

$$\langle \widehat{m}_t - c_t,\, m(\boldsymbol{x}_{t+1}, \boldsymbol{z}_{t+1}) - c_t \rangle \leqslant 0\,. \tag{15}$$

In the rest of the proof, we choose $\boldsymbol{z}_{t+1}$ to be either $\boldsymbol{y}_{t+1}$ or $\delta_{B_{t+1}}$, depending on whether mixed or pure actions are taken and observed; in particular, we have the rewriting

$$\widehat{m}_t = \frac{1}{t} \sum_{\tau=1}^{t} m(\boldsymbol{x}_\tau, \boldsymbol{z}_\tau)\,.$$

Straightforward calculation show that

$$\begin{aligned}
\widehat{m}_{t+1} &= \frac{1}{t+1} \sum_{\tau=1}^{t+1} m(\boldsymbol{x}_\tau, \boldsymbol{z}_\tau) \\
&= \frac{1}{t+1} m(\boldsymbol{x}_{t+1}, \boldsymbol{z}_{t+1}) + \frac{1}{t} \sum_{\tau=1}^{t} m(\boldsymbol{x}_\tau, \boldsymbol{z}_\tau) - \frac{1}{t(t+1)} \sum_{\tau=1}^{t} m(\boldsymbol{x}_\tau, \boldsymbol{z}_\tau) \\
&= \widehat{m}_t + \frac{1}{t+1} \big( m(\boldsymbol{x}_{t+1}, \boldsymbol{z}_{t+1}) - \widehat{m}_t \big)\,.
\end{aligned}$$

Denote by

$$d_t = \inf_{c\in\mathcal{C}} \|c - \widehat{m}_t\|_2 = \|c_t - \widehat{m}_t\|_2$$

the $\ell^2$–distance of $\widehat{m}_t$ to $\mathcal{C}$. Now, for all $t \geqslant 1$,

$$\begin{aligned}
d_{t+1}^2 \leqslant \|c_t - \widehat{m}_{t+1}\|_2^2 &= \left\| \big(c_t - \widehat{m}_t\big) + \frac{1}{t+1} \big( \widehat{m}_t - m(\boldsymbol{x}_{t+1}, \boldsymbol{z}_{t+1}) \big) \right\|_2^2 \\
&= \|c_t - \widehat{m}_t\|_2^2 + \frac{2}{t+1} \langle \widehat{m}_t - c_t, m(\boldsymbol{x}_{t+1}, \boldsymbol{z}_{t+1}) - \widehat{m}_t \rangle \\
&\quad + \frac{\big\| \widehat{m}_t - m(\boldsymbol{x}_{t+1}, \boldsymbol{z}_{t+1}) \big\|_2^2}{(t+1)^2} \\
&= \left( 1 - \frac{2}{t+1} \right) \underbrace{\|c_t - \widehat{m}_t\|_2^2}_{= d_t^2} + \frac{2}{t+1} \underbrace{\langle \widehat{m}_t - c_t, m(\boldsymbol{x}_{t+1}, \boldsymbol{z}_{t+1}) - c_t \rangle}_{\leqslant 0 \text{ by } (15)} \\
&\quad + \frac{\big\| \widehat{m}_t - m(\boldsymbol{x}_{t+1}, \boldsymbol{z}_{t+1}) \big\|_2^2}{(t+1)^2}\,.
\end{aligned}$$

By the triangle inequality,

$$\left\| \widehat{m}_t - m(\boldsymbol{x}_{t+1}, \boldsymbol{z}_{t+1}) \right\|_2^2 \leqslant 4M^2 \, ;$$

thus, we proved that

$$d_{t+1}^2 \leqslant \left(1 - \frac{2}{t+1}\right) d_t^2 + \frac{4M^2}{t+1} \, .$$

Since

$$\left(1 - \frac{2}{t+1}\right) \frac{1}{t} + \frac{1}{(t+1)^2} = \frac{(t+1)^2 - 2(t+1) + t}{t(t+1)^2} = \frac{t^2 - 1 + t}{t(t+1)^2} \leqslant \frac{1}{t+1} \, ,$$

a simple induction argument yields that $d_T^2 \leqslant 4M^2/T$ for all $T \geqslant 1$; which concludes the proof in the case of mixed actions taken and observed.

In the case of pure actions taken and observed, we need an additional concentration argument. We denote by $\mathbb{E}_t$ the conditional expectation at round $t$ with respect to $B_t$ and the $\boldsymbol{x}_s$, $A_s$, $B_s$, where $1 \leqslant s \leqslant t-1$; we have

$$\mathbb{E}_t \big[ m(A_t, B_t) \big] = m(\boldsymbol{x}_t, B_t) = m(\boldsymbol{x}_t, \boldsymbol{z}_t) \, .$$

In addition, the quantities $m(\boldsymbol{x}_t, \boldsymbol{z}_t) - m(A_t, B_t)$ are bounded in norm by $2M$. By the version of the Hoeffding-Azuma inequality for sums of Hilbert space-valued martingale differences already used in the proof of Theorem 13, we therefore have that for all $T \geqslant 1$, with probability at least $1 - \delta$,

$$\left\| \frac{1}{T} \sum_{t=1}^{T} m(A_t, B_t) - \frac{1}{T} \sum_{t=1}^{T} m(\boldsymbol{x}_t, \boldsymbol{z}_t) \right\|_2 \leqslant 4M \sqrt{\frac{\ln(2/\delta)}{T}} \, ,$$

which, combined with the deterministic bound on $d_T$, entails, still with probability at least $1 - \delta$,

$$\inf_{c \in \mathcal{C}} \left\| c - \frac{1}{T} \sum_{t=1}^{T} m(A_t, B_t) \right\|_2 \leqslant \frac{2M}{\sqrt{T}} \Big( 1 + 2\sqrt{\ln(2/\delta)} \Big) .$$

This concludes the proof. ■

### B.2 Proof of a concentration argument

We re-prove here the inequality (14), that is directly extracted from Lugosi et al. (2008, Section 6). Again, this is only for the sake of self-containment.

**Proof:** For all $(i, j) \in \mathcal{I} \times \mathcal{J}$, the quantity $H(i, j)$ is a probability distribution over $\mathcal{H}$; we denote by $H_s(i, j)$ the probability mass that it puts on some element $s \in \mathcal{H}$.

We consider a fixed block $n$. Equation (5) indicates that for each pair $(i, s) \in \mathcal{I} \times \mathcal{H}$,

$$\sum_{t=(n-1)L+1}^{nL} \left( \frac{\mathbb{I}_{\{S_t = s\}} \mathbb{I}_{\{I_t = i\}}}{p_{I_t,n}} - H_s(i, J_t) \right)$$

is a sum of $L$ elements of a martingale difference sequence. The conditional variances of the increments are bounded by

$$\mathbb{E}_t \left[ \left( \frac{\mathbb{I}_{\{S_t = s\}} \mathbb{I}_{\{I_t = i\}}}{p_{I_t,n}} \right)^2 \right] \leqslant \frac{1}{p_{i,n}^2} \, \mathbb{E}_t \big[ \mathbb{I}_{\{I_t = i\}} \big] = \frac{1}{p_{i,n}} \, ;$$

since by definition of the strategy, $\boldsymbol{p}_n = (1 - \gamma) \, \boldsymbol{x}_n + \gamma \, \boldsymbol{u}$, we have that $p_{i,n} \geqslant \gamma / N_{\mathcal{I}}$, which shows that the sum of the conditional variances is bounded by

$$\sum_{t=(n-1)L+1}^{nL} \mathrm{Var}_t \left( \frac{\mathbb{I}_{\{S_t = s\}} \mathbb{I}_{\{I_t = i\}}}{p_{I_t,n}} \right) \leqslant \frac{L N_{\mathcal{I}}}{\gamma} \, .$$

The Bernstein-Freedman inequality (see Freedman 1975 or Cesa-Bianchi et al. 2006, Lemma A.1) therefore indicates that with probability at least $1 - \delta$,

$$\left| \frac{1}{L} \sum_{t=(n-1)L+1}^{nL} \frac{\mathbb{I}_{\{S_t = s\}} \mathbb{I}_{\{I_t = i\}}}{p_{I_t,n}} - \underbrace{\frac{1}{L} \sum_{t=(n-1)L+1}^{nL} H_s(i, J_t)}_{= \, H_s(i, \widehat{\boldsymbol{q}}_n) \text{ by (6)}} \right| \leqslant \sqrt{2 \frac{N_{\mathcal{I}}}{\gamma L} \ln \frac{2}{\delta}} + \frac{1}{3} \frac{N_{\mathcal{I}}}{\gamma L} \ln \frac{2}{\delta} \, .$$

Therefore, by summing the above inequalities over $i \in \mathcal{I}$ and $s \in \mathcal{H}$, we get (after a union bound) that with probability at least $1 - N_{\mathcal{I}} N_{\mathcal{H}} \delta$,

$$\left\| \widetilde{\sigma}_n - \widetilde{H}(\widehat{\boldsymbol{q}}_n) \right\|_2 \leqslant \sqrt{N_{\mathcal{I}} N_{\mathcal{H}}} \left( \sqrt{\frac{2N_{\mathcal{I}}}{\gamma L} \ln \frac{2}{\delta}} + \frac{1}{3} \frac{N_{\mathcal{I}}}{\gamma L} \ln \frac{2}{\delta} \right).$$

Finally, since $\widehat{\sigma}_n$ is the projection in the $\ell^2$–norm of $\widetilde{\sigma}_n$ onto the convex set $\mathcal{F}$, to which $\widetilde{H}(\widehat{\boldsymbol{q}}_n)$ belongs, we have that

$$\left\| \widehat{\sigma}_n - \widetilde{H}(\widehat{\boldsymbol{q}}_n) \right\|_2 \leqslant \left\| \widetilde{\sigma}_n - \widetilde{H}(\widehat{\boldsymbol{q}}_n) \right\|_2 ,$$

and this concludes the proof. ■