

Weak Mean-Based Algorithms

Rigel Galgana

Operations Research Center, Massachusetts Institute of Technology
galganarigel@gmail.com

October 5, 2026

Abstract

Various paradigms of algorithmic learning have been proposed and analyzed over the years, the most famous of which is the celebrated *no-regret* learning framework. A more recent and closely related notion is *mean-based* learning, where highly sub-optimal actions are selected with low probability. Mean-based algorithms (MBAs) have two limitations: 1) they restrict behavior on *every* round and 2) computing their long-run behavior is hard. To resolve these issues, we introduce the relaxed class of *weak mean-based algorithms* (WMBAs), which only require highly sub-optimal actions to have vanishing frequency. These WMBAs preserve the main advantage of MBA dynamics in guaranteeing convergence to the *correlated rationalizable set* (CRS) whilst including a wider class of natural learning algorithms as well as those with explicit exploration. Furthermore, we show that their long-run behavior can be computed efficiently. We illustrate the differences between MBAs and WMBAs through multi-unit auctions and price competition games.

1 Introduction

As the world increasingly adopts automated, data-driven decision making—often interacting repeatedly in the same environments over time—it is crucial to understand the possible outcomes of such multi-agent systems. This long-run behavior crucially depends on how they learn. Rather than analyze specific algorithms, we ask what follows from properties that many natural learning algorithms possess. The most famous such property is no regret [2]. A more recent one defines *mean-based algorithms* (MBAs) [13]: *highly sub-optimal* actions, those that fall far behind the best action in hindsight, are played with low probability. Many standard no-regret algorithms are MBAs. Under *MBA dynamics*, in which every player uses an MBA, we obtain concentration in the last iterate on the *correlated rationalizable set* (CRS), the actions that survive iterated elimination of strictly dominated strategies (IESDS) [10]. These MBAs, however, have two primary limitations:

- They restrict behavior on every round,
- Their long-run behavior is hard to characterize.

As such, the class of MBAs excludes many common learning algorithms, such as some mirror descent (MD) algorithms (Propositions 2 and 3) and those with *explicit exploration*—recurring rounds on which the learning algorithm plays a chosen action outright, however poorly that action has performed (Table 1). Furthermore, MBA outcomes are harder to fully characterize, since their long-run behavior depends on the entire history of strategies. As such, we ask our following main research question:

Can we design learning algorithms which 1) preserve MBA rationalizability, 2) include descent-based and explorative algorithms, and 3) enable efficient characterization of outcomes?

1.1 Contributions

To address both limitations by replacing the per-round condition with a time-average condition, we introduce the class of *weak mean-based algorithms* (WMBAs), which only require highly sub-optimal actions to have vanishing empirical frequency. This relaxation includes the preceding learning algorithms and preserves time-average concentration on the CRS. It also yields an exact characterization and polynomial-time computation of the WMBA *extrema*—the worst and best long-run welfare over all WMBAs.

Rationalizability. We show that WMBA dynamics still concentrate strategies on the CRS in time average ([Theorem 1](#)). The guarantee is robust to any exploration that, on average, changes strategies by a vanishing amount ([Theorem 2](#) and [Corollary 1](#)) and to learning from sufficiently accurate reward estimates ([Theorem 3](#)).

Outcomes. Beyond rationalizability, we show that MBA and WMBA dynamics can drive empirical strategies to exactly the same joint distributions. These form the *support-optimal set* [[25](#)], in which each player uses only best responses to the others’ joint strategy, and in two-player games, their marginals form a Nash equilibrium ([Proposition 5](#)). Additional WMBA outcomes therefore arise when empirical frequencies do not converge. An MBA must nearly best respond to the history on every round ([Proposition 4](#)), whereas a WMBA may play an action heavily while it lags, as long as the action later catches up. This freedom, rather than exploration, can widen the range of long-run welfare in both directions ([Theorem 4](#)), since adding explicit exploration to MBAs leaves their range unchanged in our examples ([Proposition 7](#)). We characterize the WMBA extrema exactly and compute them in polynomial time in any finite game ([Propositions 6](#) and [9](#)), whereas the MBA extrema depend on the long-run dynamics of fictitious play (FP), which we can resolve only in special two-player games. Without no regret, the WMBA worst extremum is simply the welfare of the worst rationalizable profile ([Corollary 2](#)). In our examples, however, it is reached only while welfare oscillates, and requiring welfare to converge raises it without closing the gap ([Proposition 8](#)).

Consequences for markets. We illustrate these differences in multi-unit pay-as-bid auctions, where the four classes—MBA and WMBA, each with and without no regret—have pairwise distinct worst extrema for welfare and revenue, and WMBA alone can raise the best extrema ([Theorem 4](#)). Weakening MBA to WMBA can halve the welfare guaranteed at all large times, and it lowers the seller’s revenue guarantee by a fifth even when revenue must converge. It can also raise welfare, revenue, or bidders’ surplus by 16–42% beyond anything MBAs sustain, although in our auction examples this requires some bidder to have positive average regret. In price competition, even a simple WMBA, batched FP (BFP), which updates only after rounds $1, 3, 9, \dots$, guarantees the firms 14% more profit than any MBA can, while the consumer surplus it guarantees is 76% lower ([Section 5](#)).

2 Related Work

Mean-based learning. MBAs were introduced to study selling to a no-regret buyer [[13](#)] and have since been used to analyze bidding and equilibrium selection in repeated auctions [[11](#), [23](#), [35](#)]. Under MBA, actions that are never best responses are played only $o(T)$ times [[14](#)]. Moreover, strategies converge to the CRS both in time average and in the last iterate [[10](#)], and we extend the time-average result to WMBAs. In first-price auctions, bidders using MBAs converge to Nash

Table 1: Learning algorithms classified on arbitrary reward histories. MD and AMD are defined in Section 4.2 and BFP in Section 5, and the two learning algorithms that have no regret but are not WMBA are classified in Section A.4. The last three rows add exploration to FTRL (Section 4.3), either by mixing or through explicit exploration.

Algorithm	Main feature	Examples	No regret	MBA	WMBA
Follow-the-regularized-leader (FTRL) [22]	cumulative scores	entropy (Hedge), Tsallis, Euclidean	✓	✓	✓
Follow-the-leader [16]	cumulative leader	fictitious play (FP)	✗	✓	✓
Mirror descent (MD)	2 actions, $\eta_t = ct^{-\alpha}$, $0 < \alpha \leq 1$	entropy, Tsallis, log-barrier	✗	✗	✓
Averaged mirror descent (AMD)	average of proposals	entropy, Tsallis, log-barrier	✗	✗	✓
Batched fictitious play (BFP)	updates after rounds 3^k	batched bandits [44]	✗	✗	✓
Strongly adaptive algorithms	low regret on every interval	SAOL [18]	✓	✗	✗
Periodic schedule (Lemma 6)	falls back to FTRL		✓	✗	✗
FTRL, mixing	probability $\varepsilon_t \rightarrow 0$	ε -greedy [1], EXP3 [2, 13]	✓	✓	✓
FTRL, explicit exploration	on $o(T)$ rounds	explore-then-commit [27, 37], epoch-greedy [36], forced sampling [3, 30], bidding and pricing algorithms [7, 31, 32]	✓	✗	✓
FTRL, explicit exploration	on $\Theta(T)$ rounds		✗	✗	✗

equilibrium when the highest valuation is shared, and time-average and last-iterate convergence can differ [19]. Rates of MBAs and their relation to no regret have been studied [20], and best-response conditions on limits of empirical strategies, which define the support-optimal set, have been derived from cumulative action frequencies [25].

Rationalizability and equilibrium. Rationalizability [6, 43] and correlated rationalizability [12] were introduced in the 1980s. Learning dynamics that eliminate iteratively dominated actions have been studied [41], and learning algorithms that avoid dominated actions, including all MBAs, cannot collude in a pricing market in which Q-learning agents do [15, 42]. Algorithms that learn rationalizable equilibria under bandit feedback have also been proposed [50]. No regret places empirical strategies in the set of coarse correlated equilibria [16]. To identify the long-run behavior of MBA dynamics in our examples, we use convergence results for FP, in which players best respond to empirical frequencies [5], including a version with errors that vanish [38].

Exploration. Exploration is ubiquitous in online learning: explore-then-commit [27, 37], ε -greedy [1], the uniform mixing in EXP3 [2], successive elimination [21], epoch-greedy [36], deterministic exploration sequences [48], forced sampling in contextual and linear bandits [3, 30, 45], and the explicit exploration behind best-arm identification [26] and change-point detection [8]. Algorithms for repeated auctions and pricing also use explicit exploration to learn competing bids [32], reserve prices [31], click-through rates [28], and demand curves [7]. In Table 1, we classify these modifications of FTRL alongside MD, averaged mirror descent (AMD), and BFP (Section 5). While MBAs are

preserved under exploration with vanishing probabilities, they exclude certain WMBAAs that use explicit exploration. Beyond exploration, we build on MD with decreasing learning rates [22] and averaging [47] to identify the additional WMBAAs in Section 4.2.

3 Model

Game. We consider players $n \in [N]$, each with a finite action set $\mathcal{A}_n$ and a payoff function $u_n : \mathcal{A} \rightarrow [0, 1]$ on the joint action space $\mathcal{A} \doteq \prod_n \mathcal{A}_n$, and we write all such joint objects in bold. Payoffs extend linearly to mixed strategies $p \in \Delta(\mathcal{A}_n)$, to beliefs $\mathbf{q} \in \Delta(\mathcal{A}_{-n})$ over the opponents' profiles, and to joint distributions $\boldsymbol{\mu} \in \Delta(\mathcal{A})$, giving $u_n(p, \mathbf{q})$ and $u_n(\boldsymbol{\mu})$. Identifying each action with its point mass, we write the best responses to $\mathbf{q}$ as

$$\text{BR}_n(\mathbf{q}) \doteq \arg \max_{a \in \mathcal{A}_n} u_n(a, \mathbf{q}).$$

Learning. In each round t , every player n independently chooses $p_{n,t} \in \Delta(\mathcal{A}_n)$, so the joint strategy is $\mathbf{p}_t \doteq \prod_n p_{n,t}$. Player n then observes $(u_n(a, \mathbf{p}_{-n,t}))_{a \in \mathcal{A}_n}$, the payoff each of its actions would have earned against the others' strategies. We call this *expected full-information feedback*, and the sequence of observed vectors the *reward history*. A *learning algorithm* maps each reward history to the next strategy, and every condition below must hold on every reward history. Section 4.4 extends the rationalizability guarantee to estimated rewards.

Empirical strategies. To describe the history of strategies, we use the *empirical distribution* $\bar{\mathbf{p}}_T \doteq T^{-1} \sum_{t=1}^T \mathbf{p}_t$. Its marginal $\bar{p}_{n,T}$ records player n 's *empirical frequencies*, and its marginal $\bar{\mathbf{p}}_{-n,T}$ on $\mathcal{A}_{-n}$ the opponents' joint empirical strategy, which can be correlated across rounds. By linearity, the average payoff that action a would have earned is $u_n(a, \bar{\mathbf{p}}_{-n,T})$, and the average payoff player n did earn is $u_n(\bar{\mathbf{p}}_T)$. The *gap* of a is its shortfall from a best response to the opponents' empirical strategy,

$$\delta_{n,T}(a) \doteq \max_{a' \in \mathcal{A}_n} u_n(a', \bar{\mathbf{p}}_{-n,T}) - u_n(a, \bar{\mathbf{p}}_{-n,T}),$$

and a is *highly sub-optimal* if it falls permanently behind: $\liminf_{T \rightarrow \infty} \delta_{n,T}(a) > 0$. We first recall mean-based algorithms [13], which make actions with large gaps unlikely on every round.

Definition 1 (Mean-based algorithm). A learning algorithm for player n is *mean-based* (an MBA) if, for some nonincreasing $\gamma_T \rightarrow 0$ and all T and $a \in \mathcal{A}_n$,

$$\delta_{n,T}(a) > \gamma_T \implies p_{n,T+1}(a) \leq \gamma_T.$$

We verify this condition for follow-the-regularized-leader (FTRL) with standard regularizers and learning rates, including Hedge, and for regret matching (Sections A.1 and A.2). We now relax the restriction on the next choice to a restriction on empirical frequency.

Definition 2 (Weak mean-based algorithm). A learning algorithm for player n is *weak mean-based* (a WMBA) if every highly sub-optimal action has vanishing empirical frequency: for all $a \in \mathcal{A}_n$,

$$\liminf_{T \rightarrow \infty} \delta_{n,T}(a) > 0 \implies \bar{p}_{n,T}(a) \longrightarrow 0. \quad (1)$$

Definition 2 weakens Definition 1 in two ways. It is harder to activate: an MBA must respond as soon as an action's gap exceeds γ_T , whereas a WMBA responds only once the action has fallen permanently behind. And it asks less once activated: an MBA must assign the action small

probability on the very next round, whereas a WMBA need only play it on a vanishing fraction of rounds. The second weakening is what makes room for explicit exploration. The first is less visible, but it is the one behind the welfare results of [Section 4.6](#) ([Proposition 7](#)): a WMBA may play an action heavily while it trails, provided the action recovers infinitely often. As a result, WMBAs form a larger class.

Proposition 1. $\text{MBA} \subset \text{WMBA}$.

We can see that the inclusion is strict by modifying follow-the-leader to play a fixed action on rounds 2^k , since the result is still a WMBA but violates [Definition 1](#) whenever that action is highly sub-optimal ([Section A.3](#)).

No regret. To control earned payoffs separately from action frequencies, we impose no regret: the average payoff must eventually be at least that of the best fixed action in hindsight:

$$\limsup_{T \rightarrow \infty} \left[\max_{a \in \mathcal{A}_n} u_n(a, \bar{\mathbf{p}}_{-n,T}) - u_n(\bar{\mathbf{p}}_T) \right] \leq 0. \quad (2)$$

Adding (2) to either definition yields the classes MBA_{NR} and WMBA_{NR} of MBAs and WMBAs with no regret. No regret and WMBA do not imply each other ([Table 1](#)).

4 Guarantees

We first prove concentration on the CRS ([Section 4.1](#)), then identify MD families that retain this guarantee ([Section 4.2](#)). We next show how to preserve it under exploration and estimated rewards ([Sections 4.3](#) and [4.4](#)), before characterizing the limits of empirical strategies and long-run welfare ([Sections 4.5](#) and [4.6](#)).

4.1 Rationalizability

We first recall how IESDS determines the CRS. Since an action is a best response to no belief exactly when a mixed strategy strictly dominates it [\[43\]](#), each stage of IESDS removes the actions that are not best responses to any belief over the opponents' remaining actions. Starting from $\mathcal{A}_0 \doteq \mathcal{A}$, let $\mathcal{A}_{k+1} \doteq \prod_n \mathcal{A}_{n,k+1}$, where

$$\mathcal{A}_{n,k+1} \doteq \mathcal{A}_{n,k} \cap \bigcup_{\mathbf{q} \in \Delta(\mathcal{A}_{-n,k})} \text{BR}_n(\mathbf{q}). \quad (3)$$

Following this deletion process, we reach the CRS $\tilde{\mathcal{A}} \doteq \prod_n \tilde{\mathcal{A}}_n$ after finitely many steps [\[12\]](#). We characterize it as the largest product set in which every action is a best response to some belief supported on the set ([Section B.1](#)). We now show that WMBA dynamics concentrate on this set in time average, although not necessarily in the last iterate.

Theorem 1. *WMBA dynamics guarantee $\bar{\mathbf{p}}_T(\tilde{\mathcal{A}}) \rightarrow 1$, although $\mathbf{p}_T(\tilde{\mathcal{A}})$ need not converge to 1.*

We prove [Theorem 1](#) by following the stages of IESDS in (3). Suppose that the actions removed before stage k are played with vanishing frequency. Then the opponents' empirical strategy concentrates on $\mathcal{A}_{-n,k}$, against which every action removed at stage k trails some alternative by a fixed margin ([Lemma 3](#)). Each such action is therefore highly sub-optimal, so a WMBA plays it with vanishing frequency as well ([Section B.2](#)). The conclusion also holds almost surely (a.s.) for realized actions ([Section B.3](#)).

In the second part of [Theorem 1](#), we show that WMBA strategies can return to an action outside the CRS infinitely often, as in [Proposition 1](#). We next ask which natural learning algorithms lie in this larger class.

4.2 Learning Algorithms

To identify natural learning algorithms covered by this guarantee, we consider two MD families. We fix one player, write $\mathcal{A}$ for its action set, and let u_t be its observed reward vector. With MD, we balance the latest rewards against a penalty for moving from the current distribution:

$$p_{t+1} = \arg \max_{p \in \Delta(\mathcal{A})} \{ \eta_t \langle u_t, p \rangle - D_\Psi(p, p_t) \}. \quad (4)$$

We measure this movement by $D_\Psi(p, p') \doteq \Psi(p) - \Psi(p') - \langle \nabla \Psi(p'), p - p' \rangle$, the *Bregman divergence* of a differentiable, strictly convex regularizer Ψ , and we start from a distribution p_1 that gives every action positive probability.

Two actions. We begin with two actions, a_1 and a_2 . Writing $x \doteq p(a_1)$, we call Ψ *steep* when the derivative of $\Psi(x, 1 - x)$ tends to $-\infty$ at 0 and $+\infty$ at 1. We use $\eta_t \doteq ct^{-\alpha}$, with $c > 0$ and $0 < \alpha \leq 1$. Unlike FTRL, which applies the regularizer to the cumulative scores and is an MBA ([Lemma 2](#)), MD weighs each reward by the learning rate of its round, so early rewards carry more weight when learning rates decrease. As a result, an action that led early can remain likely for a while after it falls behind, but not once it has fallen permanently behind ([Section B.4](#)).

Proposition 2. *MD (2 actions) with a steep regularizer is WMBA but not MBA.*

Averaging. To extend to multiple actions, AMD applies (4) to proposals $\hat{p}_t$ with $\eta_t \doteq \eta > 0$ and plays their running average:

$$p_{t+1} = \frac{t}{t+1} p_t + \frac{1}{t+1} \hat{p}_{t+1}, \quad p_1 = \hat{p}_1. \quad (5)$$

We update each proposal from the rewards observed after playing p_t . For this family, we assume at least two actions and take $\Psi(p) \doteq \sum_a \psi_a(p(a))$, with each ψ_a differentiable, strictly convex, and $\psi'_a(x) \rightarrow -\infty$ as $x \rightarrow 0$. Since the average responds slowly, an action that has just fallen behind can again remain likely, while a highly sub-optimal one still fades ([Section B.5](#)).

Proposition 3. *AMD is WMBA but not MBA.*

4.3 Exploration

Beyond changing the update itself, we can add exploration to any WMBA, including the learning algorithms just introduced. We run a base algorithm $\mathcal{L}$ on the full reward history but play p_t in place of its proposal $\hat{p}_t$, with sublinear total deviation on every reward history:

$$\sum_{t=1}^T \sum_a |p_t(a) - \hat{p}_t(a)| = o(T). \quad (6)$$

We now show that WMBA absorbs any such deviation, but no deviation of linear size.

Theorem 2. *If $\mathcal{L}$ is a WMBA, so is the modified learning algorithm. The bound $o(T)$ in (6) cannot be replaced by εT for any $\varepsilon > 0$.*

To prove [Theorem 2](#), we bound the change in any action's frequency by $o(1)$. Since the base learning algorithm uses highly sub-optimal actions with vanishing frequency, the modified one does too ([Section B.6](#)). In particular, we may replace its choices on a sparse set of rounds.

Corollary 1. *If $p_t = \hat{p}_t$ except on $o(T)$ of the first T rounds, the modified learning algorithm is a WMBA.*

With [Corollary 1](#), we can add explicit exploration on $o(T)$ rounds to any WMBA, as in several algorithms listed in [Table 1](#). [Theorem 2](#) also allows exploration mixed into every round,

$$p_t = (1 - \varepsilon_t) \hat{p}_t + \varepsilon_t \nu_t,$$

with arbitrary ν_t , provided $T^{-1} \sum_{t \leq T} \varepsilon_t \rightarrow 0$. If ε_t decreases to zero, this mixing also keeps an MBA an MBA, with threshold $\gamma_T + \varepsilon_{T+1}$, and any deviation of total size $o(T)$ preserves no regret.

4.4 Estimated Rewards

Having allowed exploration, we now consider learning algorithms that use it to estimate payoffs. Suppose each player observes only an estimate $\hat{u}_{n,T}(a)$ of the average payoff $u_n(a, \bar{\mathbf{p}}_{-n,T})$ and is a WMBA with respect to the estimated gaps

$$\hat{\delta}_{n,T}(a) \doteq \max_{a'} \hat{u}_{n,T}(a') - \hat{u}_{n,T}(a),$$

that is,

$$\liminf_{T \rightarrow \infty} \hat{\delta}_{n,T}(a) > 0 \implies \bar{p}_{n,T}(a) \rightarrow 0, \quad (7)$$

and that the estimation errors vanish uniformly, a.s.:

$$\max_{n \in [N]} \max_{a \in \mathcal{A}_n} |\hat{u}_{n,T}(a) - u_n(a, \bar{\mathbf{p}}_{-n,T})| \rightarrow 0. \quad (8)$$

Together, these two conditions preserve rationalizability.

Theorem 3. *Under (7) and (8), $\bar{\mathbf{p}}_T(\tilde{\mathcal{A}}) \rightarrow 1$ a.s.*

We obtain [Theorem 3](#) by noting that vanishing errors change gaps by vanishing amounts. An action that is highly sub-optimal under the true payoffs is then also highly sub-optimal under the estimates, and the proof of [Theorem 1](#) applies on almost every sample path ([Section B.7](#)). Under bandit feedback, we read (8) as a joint requirement on the estimator and the exploration schedule.

4.5 Limits of Empirical Strategies

Having established which actions can persist, we now study how they can be used together. Recall that the gap $\delta_{n,T}(a)$ measures how far a is from a best response to the opponents' joint empirical strategy $\bar{\mathbf{p}}_{-n,T}$. An MBA must therefore nearly best respond to that strategy on every round.

Proposition 4. *If player n uses an MBA, then*

$$\max_{a \in \mathcal{A}_n} u_n(a, \bar{\mathbf{p}}_{-n,T}) - u_n(p_{n,T+1}, \bar{\mathbf{p}}_{-n,T}) \leq (|\mathcal{A}_n| + 1)\gamma_T.$$

For the two-player games studied below, we use [Proposition 4](#) to view the pair of empirical marginals as an FP process with errors that vanish. Such processes need not converge, as in Shapley’s game [46], but we establish convergence to Nash equilibria in our examples [5, 38] ([Section C.4](#)). A WMBA is not bound in this way, since its gaps may grow large and later shrink again. As we show next, however, this freedom does not change where empirical strategies can converge. To describe these limits, we call

$$\mathcal{P}^* \doteq \{\boldsymbol{\mu} \in \Delta(\mathcal{A}) : \text{supp}(\mu_n) \subseteq \text{BR}_n(\boldsymbol{\mu}_{-n}), n \in [N]\}$$

the *support-optimal set* [25], since under each of its joint distributions, every action that a player uses is a best response to the others’ joint strategy. We also call $\boldsymbol{\mu}$ a *coarse correlated equilibrium* (CCE) if

$$u_n(\boldsymbol{\mu}) \geq u_n(a, \boldsymbol{\mu}_{-n}) \quad (n \in [N], a \in \mathcal{A}_n). \quad (9)$$

We now show that, for both classes, $\mathcal{P}^*$ consists of exactly the joint distributions $\boldsymbol{\mu}$ to which the dynamics can converge, that is, with $\bar{\mathbf{p}}_T \rightarrow \boldsymbol{\mu}$ for some profile of learning algorithms from the class.

Proposition 5. *MBA and WMBA dynamics can converge to the same joint distributions, which are those in $\mathcal{P}^*$. Under MBA_{NR} and WMBA_{NR} dynamics, they are the CCEs in $\mathcal{P}^*$.*

In [Proposition 5](#), we show that the two classes agree on every outcome at which empirical strategies can settle. With two players, $\boldsymbol{\mu}_{-n}$ is the other player’s marginal, so $\boldsymbol{\mu} \in \mathcal{P}^*$ exactly when the marginals of $\boldsymbol{\mu}$ form a Nash equilibrium. With more players, however, the opponents’ joint strategy can be correlated, and the marginals need not form one. Within each round the players randomize independently, so any correlation in $\boldsymbol{\mu}$ arises across rounds. Checking whether a given distribution lies in $\mathcal{P}^*$ is easy, but [Proposition 5](#) does not say which distributions in $\mathcal{P}^*$ MBA dynamics actually reach, or whether they converge at all. The idea of the proof is that a *persistent action*—one whose empirical frequency does not vanish—must be a best response in the limit, since otherwise it would be highly sub-optimal. Conversely, a deterministic schedule that falls back to an MBA if another player deviates attains any $\boldsymbol{\mu} \in \mathcal{P}^*$ ([Sections C.1](#) and [C.2](#)).

4.6 Welfare Extrema

To compare the classes beyond the support-optimal set, we measure welfare relative to its maximum,

$$\text{Wel}(\boldsymbol{\mu}) \doteq \frac{\sum_n u_n(\boldsymbol{\mu})}{\max_{\mathbf{a} \in \mathcal{A}} \sum_n u_n(\mathbf{a})} \in [0, 1], \quad (10)$$

assuming the maximum is positive. For a class $\mathcal{C}$, the *worst and best extrema* of long-run welfare are

$$\begin{aligned} \underline{\text{Wel}}(\mathcal{C}) &\doteq \inf_{\mathcal{L}_1, \dots, \mathcal{L}_N \in \mathcal{C}} \liminf_{T \rightarrow \infty} \text{Wel}(\bar{\mathbf{p}}_T), \\ \overline{\text{Wel}}(\mathcal{C}) &\doteq \sup_{\mathcal{L}_1, \dots, \mathcal{L}_N \in \mathcal{C}} \liminf_{T \rightarrow \infty} \text{Wel}(\bar{\mathbf{p}}_T), \end{aligned} \quad (11)$$

where $\mathcal{L}_n$ is player n ’s algorithm. The first is the welfare the class guarantees at all large times, and the second the most it can sustain. Enlarging a class can only widen this range. Both extrema are bounds for the entire class, and they are attained by algorithms built for the purpose, which follow a common deterministic schedule and fall back to Euclidean FTRL if another player deviates ([Section C.1](#)), rather than by standard algorithms such as Hedge. They measure what can be predicted from membership in a class alone.

By [Proposition 5](#), WMBA can leave the MBA range only when the empirical distribution does not converge. In that case, a WMBA needs each persistent action to be an approximate best response only along a subsequence of rounds, so strategies can alternate between increasingly longer blocks at a low-welfare profile and blocks that restore each persistent action to a best response. We now show that this is all the room there is, and describe it exactly. For a product $\mathbf{B} \doteq \prod_n B_n$ of action sets, let $\mathcal{D}(\mathbf{B})$ be the joint distributions supported on $\mathbf{B}$, restricted to CCEs under no regret, and for $a \in B_n$ let

$$Q_a(\mathbf{B}) \doteq \{\boldsymbol{\mu} \in \mathcal{D}(\mathbf{B}) : a \in \text{BR}_n(\boldsymbol{\mu}_{-n})\}$$

be those under which a is a best response. Call $\mathbf{B}$ *admissible* if $Q_a(\mathbf{B}) \neq \emptyset$ for every n and $a \in B_n$, that is, if every action in $\mathbf{B}$ is justified by some joint strategy on $\mathbf{B}$. We now show that the WMBA extrema optimize welfare over admissible products.

Proposition 6. *Let $\mathcal{C}$ be WMBA or WMBA_{NR} , with $\mathcal{D}$ defined accordingly:*

$$\begin{aligned} \underline{\text{Wel}}(\mathcal{C}) &= \min_{\mathbf{B}} \min_{\boldsymbol{\mu} \in \mathcal{D}(\mathbf{B})} \text{Wel}(\boldsymbol{\mu}), \\ \overline{\text{Wel}}(\mathcal{C}) &= \max_{\mathbf{B}} \min_{n, a \in B_n} \max_{\boldsymbol{\mu} \in Q_a(\mathbf{B})} \text{Wel}(\boldsymbol{\mu}), \end{aligned} \tag{12}$$

where $\mathbf{B}$ ranges over admissible products, and both values are attained.

We can read [Proposition 6](#) as follows. The worst extremum is simply the worst joint strategy on an admissible product. The best extremum is subtler, since each persistent action must sometimes be a best response and long-run welfare can be no higher than at those times, so the action with the lowest such ceiling sets the best extremum. In our examples, the best extrema are moreover attained with convergent average welfare ([Section C.3](#)). Without no regret, the worst extremum takes a particularly simple form.

Corollary 2. $\underline{\text{Wel}}(\text{WMBA}) = \min_{\mathbf{a} \in \tilde{\mathcal{A}}} \text{Wel}(\mathbf{a})$, that is, without no regret, WMBA guarantee only the welfare of the worst rationalizable profile.

To obtain [Corollary 2](#) from [Proposition 6](#), we note that, without no regret, $Q_a(\mathbf{B})$ is nonempty exactly when a is a best response to some belief over the other players' actions in $\mathbf{B}$. The admissible products are then those in which every action is a best response to some belief supported on the product, the largest of which is the CRS ([Section 4.1](#)), and a linear objective over $\Delta(\tilde{\mathcal{A}})$ is minimized at a pure profile. The same argument applies to any objective linear in $\boldsymbol{\mu}$, including the market outcomes of [Section 5](#). We will see in [Example 1](#) ([Section D.1](#)) that no regret can raise the guarantee.

Exploration and oscillation. We next ask whether exploration accounts for these outcomes. We call a learning algorithm an *MBA with exploration* if it modifies an MBA as in [\(6\)](#), which makes it a WMBA by [Theorem 2](#). Our comparison concerns the *rationalizable game*—the game restricted to the CRS—which we call *nondegenerate* if no strategy that mixes k actions has more than k best responses in it. We now show that, when this game is a nondegenerate $2 \times n$ game, exploration alone does not widen the MBA range.

Proposition 7. *If the rationalizable game is a nondegenerate $2 \times n$ game, MBAs with exploration have the same welfare extrema as MBAs, with/without no regret.*

We prove [Proposition 7](#) by showing that exploration perturbs FP by only a vanishing amount, so the empirical marginals still approach the set of Nash equilibria ([Section C.4](#)). Since all our

Table 2: Normalized auction outcomes for the instances of [Section D.1](#). With no regret, both classes match MBA’s best extrema after rounding.

Algorithm class $\mathcal{C}$	Worst extremum		Best extremum		
	Welfare (Ex. 1)	Revenue (Ex. 2)	Welfare (Ex. 3)	Revenue (Ex. 4)	Surplus (Ex. 5)
MBA	0.946	0.408	0.837	0.133	0.450
WMBA	0.483	0.165	0.973	0.189	0.543
MBA _{NR}	0.953	0.409	<i>same as MBA</i>		
WMBA _{NR}	0.914	0.272			

examples satisfy this condition, we conclude that in each of them, the wider WMBA range comes entirely from actions that trail for long stretches, not from exploration.

Having separated exploration from the additional outcomes, we turn to oscillation, since in our examples, these outcomes are reached only along trajectories whose average welfare returns infinitely often to higher levels. Requiring welfare to converge defines the *convergent worst extremum* $\underline{\text{Wel}}_c(\mathcal{C})$, the lowest limit of $\text{Wel}(\bar{\mathbf{p}}_T)$ over profiles in $\mathcal{C}$ along which it converges. For an admissible $\mathbf{B}$, let

$$\underline{h}_{\mathbf{B}} \doteq \max_a \min_{Q_a(\mathbf{B})} \text{Wel}, \quad \bar{h}_{\mathbf{B}} \doteq \min_a \max_{Q_a(\mathbf{B})} \text{Wel},$$

with a ranging over every B_n . These are the lowest and highest welfare at which every action in $\mathbf{B}$ can be justified, and $\bar{\text{Wel}}(\mathcal{C}) = \max_{\mathbf{B}} \bar{h}_{\mathbf{B}}$ by [Proposition 6](#). When welfare must converge, a single level must justify every action, which leads to the following characterization.

Proposition 8. *For $\mathcal{C} \in \{\text{WMBA}, \text{WMBA}_{\text{NR}}\}$,*

$$\underline{\text{Wel}}_c(\mathcal{C}) = \min\{\underline{h}_{\mathbf{B}} : \mathbf{B} \text{ admissible, } \underline{h}_{\mathbf{B}} \leq \bar{h}_{\mathbf{B}}\}.$$

The idea behind [Proposition 8](#) is that if welfare converges to h , then every persistent action must be justified at welfare exactly h , which forces $\underline{h}_{\mathbf{B}} \leq h \leq \bar{h}_{\mathbf{B}}$. Conversely, a block schedule over targets at level $\underline{h}_{\mathbf{B}}$ attains this value ([Lemma 5](#)). In our examples, MBA’s worst extrema are already attained with convergent welfare, so requiring convergence changes only the WMBA values, which rise but stay below MBA’s ([Section D](#)).

Computation. We finally show that we can evaluate the extrema in [\(12\)](#) without enumerating products.

Proposition 9. *For $\mathcal{C} \in \{\text{WMBA}, \text{WMBA}_{\text{NR}}\}$, $\underline{\text{Wel}}(\mathcal{C})$ and $\bar{\text{Wel}}(\mathcal{C})$ can be computed exactly in time polynomial in the bit length of the payoff table.*

Our algorithm for [Proposition 9](#) first deletes actions whose $Q_a(\mathbf{B})$ is empty, then repeatedly removes the actions with the lowest values in [\(12\)](#). It solves $O((\sum_n |\mathcal{A}_n|)^2)$ linear programs with at most $|\mathcal{A}|$ variables each, and it applies to any linear objective ([Section C.6](#)).

We have no counterpart of [Proposition 9](#) for MBA that applies across all games, and we know of no other general method, since ours requires MBA strategies to converge, which fails in some games. MBA strategies must follow the FP dynamics, so their long-run behavior depends on whether those

dynamics converge or cycle, and in Shapley’s game they cycle [46]. Even where they converge, in two-player games the extrema range over *Nash couplings*—joint distributions whose marginals form a Nash equilibrium—and related problems, such as maximizing welfare over Nash equilibria, are NP-hard [29]. WMBA avoids both obstacles because each action needs to be justified only at its own moments, against its own distribution. We therefore compute MBA extrema only when the rationalizable game is a nondegenerate $2 \times n$ game, where FP converges, the extrema optimize Wel over Nash couplings (restricted to CCEs under no regret), and the equilibria are few enough to enumerate (Section C.4), as in all our examples.

5 Examples

Many markets now rely on algorithms to set bids and prices. Pricing algorithms compete in Bertrand markets [15], matching platforms serve participants who learn their preferences as they go [39], language-model agents negotiate on behalf of users [9], and AI agents such as Meta’s Muse, released in September 2026 [40], now search for and buy products for consumers. When these algorithms learn from repeated interaction, the outcomes they can sustain depend on the class of algorithms they belong to, so our analysis of WMBA applies directly to such markets. We illustrate this with two classic market games, multi-unit auctions and price competition. Multi-unit auctions sell many identical units at once. They are used in wholesale electricity markets, now strained by the rapidly growing demand of data centers [34], and to sell treasury securities [33], pollution permits [24], and spectrum licenses [17], as well as in procurement auctions.

Multi-unit auctions. In a pay-as-bid (PAB) auction, two bidders with nonincreasing valuations bid for M identical units. The M highest bids win, and each bidder pays its winning bids. We measure welfare, the seller’s revenue, and the bidders’ surplus as fractions of the maximum total value (Section D.1). In these auctions, the four classes separate.

Theorem 4. *For welfare/revenue, there is a PAB instance in which the four classes have pairwise distinct worst extrema. For each of welfare/revenue/bidders’ surplus, there is a PAB instance in which the WMBA best extremum exceeds its MBA counterpart.*

We prove Theorem 4 with five auctions and report their extrema in Table 2 (Section D.1).

Price competition. To complement the auction extrema with a simple learning algorithm, we next consider two firms selling a good that consumers value at 1. Some consumers are loyal to one firm, and the rest buy from the cheaper one, as in Varian’s model of sales [49]. Here even a very simple WMBA reaches outcomes that no MBA can. In BFP, each firm best responds to its rival’s past prices, as in FP, but updates its price only after rounds 1, 3, 9, 27, . . . and keeps it in between, much like a pricing algorithm that is retrained whenever its data triples. Since a price that falls permanently behind is never chosen again, BFP is a WMBA. Between updates, however, it keeps prices that have fallen behind, so it is not an MBA (Section D.2). In an instance with three prices, two such firms spend half of their batches at a high price pair that MBA strategies abandon, and Table 3 compares the resulting guarantees with those of the two classes.

Across both games, we find that WMBA has a noticeably wider range of possible long-run behavior than MBAs, and the extra range can potentially encode stronger algorithmic collusion. In the auctions, WMBA can sustain bidders’ surplus 20% above any MBA outcome, partly at the seller’s expense. In the pricing game, even BFP guarantees the firms 14% more profit than any MBA can, while the consumer surplus it guarantees is 76% lower than under any MBA. While the

Table 3: Lowest long-run average profit and consumer surplus in price competition (Example 6), as fractions of total value. For the classes, the range runs from the worst to the best extremum, and BFP updates after rounds 3^k .

	Profit	Consumer surplus
Any MBA	[0.556, 0.582]	[0.418, 0.444]
BFP	0.663	0.101
Any WMBA	[0.500, 0.696]	[0.067, 0.447]

class extrema rely on constructions that coordinate the players on a common schedule (Section C.1), BFP needs no such design, since each firm only responds to its rival’s past prices.

6 Discussion

We studied which guarantees of MBAs survive when highly sub-optimal actions need only be infrequent, and we found that rationalizability survives but welfare predictions do not. We retain rationalizability for a broader collection of learning algorithms: MD with decreasing learning rates, AMD, periodic updates, and explicit exploration. The two MD families are WMBAs but not MBAs (Propositions 2 and 3), even without explicit exploration. Predictions about welfare and revenue depend instead on strategies responding to the running averages on every round. WMBA dynamics can converge only to the support-optimal distributions that MBA dynamics also reach, but they can sustain other outcomes by never settling on one. In our examples, these outcomes come not from exploration alone but from continuing to play actions that are temporarily behind. The resulting losses are largest in what is guaranteed at all large times, but some, such as the loss in auction revenue, survive even when outcomes must converge.

What we find most striking is that the freedom behind these losses can also produce gains. WMBA strategies can allocate units more efficiently, raise revenue, or raise bidders’ surplus at the seller’s expense, beyond anything MBAs sustain, and who would gain depends on who could steer it: a planner, a platform supplying the bidding agents, or bidders sharing an algorithm. Steering is not always needed, however, since in our pricing game, firms that simply run BFP already guarantee themselves more profit than any MBA can. In our auctions, the gains require some bidder to have positive average regret, so an audit for no regret would exclude them, while one that checks only that each bid is sometimes a best response would not. In the pricing game, by contrast, part of the gain survives no regret. We see this as complementing the finding that learning algorithms that avoid dominated actions, including all MBAs, cannot collude in a pricing market where Q-learning agents do [15, 42]. Since WMBAs also avoid non-rationalizable actions (Theorem 1), our examples suggest that what excludes these outcomes is responding to the running averages on every round, not avoiding dominated actions.

Several questions remain open. The MBA extrema are governed by the long-run dynamics of FP, so we could compute them only in two-player games where those dynamics converge. Characterizing them where FP cycles, as in Shapley’s game, or with more than two players remains open. The contrast with the WMBA extrema, which we can compute in polynomial time in any finite game, suggests to us that the weaker class is also the more tractable one for predicting outcomes. Finally, we would like to know which further learning algorithms are WMBAs on arbitrary reward histories,

whether the convergent worst extremum can be computed efficiently, how close simple uncoordinated learning algorithms, such as BFP, come to the extrema in general, and whether the extrema are attainable under bandit feedback.

AI Use Statement

OpenAI’s GPT-5.6 and GPT-6 and Anthropic’s Claude Fable 5.1 and Claude Opus 5.5 contributed to the writing of this paper and to the formalization of its proofs, and all code was written with these models. We have reviewed all AI-assisted work and take full responsibility for the content of the paper.

References

- [1] Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2–3):235–256, 2002.
- [2] Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.
- [3] Hamsa Bastani and Mohsen Bayati. Online decision making with high-dimensional covariates. *Operations Research*, 68(1):276–294, 2020.
- [4] Michel Benaïm, Josef Hofbauer, and Sylvain Sorin. Stochastic approximations and differential inclusions. *SIAM Journal on Control and Optimization*, 44(1):328–348, 2005.
- [5] Ulrich Berger. Fictitious play in $2 \times n$ games. *Journal of Economic Theory*, 120(2):139–154, 2005.
- [6] B. Douglas Bernheim. Rationalizable strategic behavior. *Econometrica*, 52(4):1007–1028, 1984.
- [7] Omar Besbes and Assaf Zeevi. Dynamic pricing without knowing the demand function: Risk bounds and near-optimal algorithms. *Operations Research*, 57(6):1407–1420, 2009.
- [8] Lilian Besson, Emilie Kaufmann, Odalric-Ambrym Maillard, and Julien Seznec. Efficient change-point detection for tackling piecewise-stationary bandits. *Journal of Machine Learning Research*, 23(77):1–40, 2022.
- [9] Federico Bianchi, Patrick John Chia, Mert Yuksekgonul, Jacopo Tagliabue, Dan Jurafsky, and James Zou. How well can LLMs negotiate? NegotiationArena platform and analysis. In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 3935–3951, 2024.
- [10] Martin Bichler, Julius Durmann, and Matthias Oberlechner. Online optimization algorithms in repeated price competition: Equilibrium learning and algorithmic collusion. *arXiv preprint arXiv:2412.15707*, 2024.
- [11] Martin Bichler, Stephan B. Lunowa, Matthias Oberlechner, Fabian R. Pieroth, and Barbara Wohlmuth. Beyond monotonicity: On the convergence of learning algorithms in standard auction games. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(13):13649–13657, 2025.

- [12] Adam Brandenburger and Eddie Dekel. Rationalizability and correlated equilibria. *Econometrica*, 55(6):1391–1402, 1987.
- [13] Mark Braverman, Jieming Mao, Jon Schneider, and S. Matthew Weinberg. Selling to a no-regret buyer. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 523–538, 2018.
- [14] William Brown, Jon Schneider, and Kiran Vodrahalli. Is learning in games good for the learners? In *Advances in Neural Information Processing Systems*, volume 36, 2023.
- [15] Emilio Calvano, Giacomo Calzolari, Vincenzo Denicolò, and Sergio Pastorello. Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review*, 110(10):3267–3297, 2020.
- [16] Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.
- [17] Peter Cramton. Spectrum auction design. *Review of Industrial Organization*, 42(2):161–190, 2013.
- [18] Amit Daniely, Alon Gonen, and Shai Shalev-Shwartz. Strongly adaptive online learning. In *International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1405–1411, 2015.
- [19] Xiaotie Deng, Xinyan Hu, Tao Lin, and Weiqiang Zheng. Nash convergence of mean-based learning algorithms in first price auctions. In *Proceedings of the ACM Web Conference 2022*, pages 141–150, 2022.
- [20] Julius Durmann and Amelie Kleber. Mean-based algorithms: A lower bound and regret. *arXiv preprint arXiv:2606.04931*, 2026.
- [21] Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research*, 7:1079–1105, 2006.
- [22] Huang Fang, Nicholas J. A. Harvey, Victor S. Portella, and Michael P. Friedlander. Online mirror descent and dual averaging: Keeping pace in the dynamic case. *Journal of Machine Learning Research*, 23(121):1–38, 2022.
- [23] Zhe Feng, Guru Guruganesh, Christopher Liaw, Aranyak Mehta, and Abhishek Sethi. Convergence analysis of no-regret bidding algorithms in repeated auctions. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(6):5399–5406, 2021.
- [24] Rigel Galgana and Negin Golrezaei. Learning in repeated multi-unit pay-as-bid auctions. *Manufacturing & Service Operations Management*, 2024.
- [25] Rigel Galgana and Negin Golrezaei. Gap based learning in markets. *SSRN working paper*, 2026. doi: 10.2139/ssrn.6222981. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6222981.
- [26] Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 998–1027, 2016.

- [27] Aurélien Garivier, Tor Lattimore, and Emilie Kaufmann. On explore-then-commit strategies. In *Advances in Neural Information Processing Systems*, volume 29, pages 784–792, 2016.
- [28] Nicola Gatti, Alessandro Lazaric, Marco Rocco, and Francesco Trovò. Truthful learning mechanisms for multi-slot sponsored search auctions with externalities. *Artificial Intelligence*, 227:93–139, 2015.
- [29] Itzhak Gilboa and Eitan Zemel. Nash and correlated equilibria: Some complexity considerations. *Games and Economic Behavior*, 1(1):80–93, 1989.
- [30] Alexander Goldenshluger and Assaf Zeevi. A linear response bandit problem. *Stochastic Systems*, 3(1):230–261, 2013.
- [31] Negin Golrezaei, Adel Javanmard, and Vahab Mirrokni. Dynamic incentive-aware learning: Robust pricing in contextual auctions. *Operations Research*, 69(1):297–314, 2021.
- [32] Yanjun Han, Tsachy Weissman, and Zhengyuan Zhou. Optimal no-regret learning in repeated first-price auctions. *Operations Research*, 73(1):209–238, 2025.
- [33] Ali Hortaçsu and David McAdams. Mechanism choice and strategic bidding in divisible good auctions: An empirical analysis of the Turkish treasury auction market. *Journal of Political Economy*, 118(5):833–865, 2010.
- [34] International Energy Agency. Energy and AI. World Energy Outlook Special Report, International Energy Agency, Paris, 2025.
- [35] Yoav Kolumbus and Noam Nisan. Auctions between regret-minimizing agents. In *Proceedings of the ACM Web Conference 2022*, pages 100–111, 2022.
- [36] John Langford and Tong Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in Neural Information Processing Systems 20*, pages 817–824, 2007.
- [37] Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.
- [38] David S. Leslie and E. J. Collins. Generalised weakened fictitious play. *Games and Economic Behavior*, 56(2):285–298, 2006.
- [39] Lydia T. Liu, Horia Mania, and Michael Jordan. Competing bandits in matching markets. In *International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 1618–1628, 2020.
- [40] Meta. Introducing Muse: The world’s first personal AI agent built for everyone. Meta Newsroom, September 2026. URL <https://about.fb.com/news/2026/09/introducing-muse-personal-ai-agent/>.
- [41] Paul Milgrom and John Roberts. Adaptive and sophisticated learning in normal form games. *Games and Economic Behavior*, 3(1):82–100, 1991.
- [42] Noam Nisan and Emmanuel Zerah. Domination-avoiding learning agents cannot collude. *arXiv preprint arXiv:2606.01275*, 2026.
- [43] David G. Pearce. Rationalizable strategic behavior and the problem of perfection. *Econometrica*, 52(4):1029–1050, 1984.

- [44] Vianney Perchet, Philippe Rigollet, Sylvain Chassang, and Erik Snowberg. Batched bandit problems. *The Annals of Statistics*, 44(2):660–681, 2016.
- [45] Paat Rusmevichientong and John N. Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.
- [46] Lloyd S. Shapley. Some topics in two-person games. In Melvin Dresher, Lloyd S. Shapley, and Albert W. Tucker, editors, *Advances in Game Theory*, volume 52 of *Annals of Mathematics Studies*. Princeton University Press, 1964.
- [47] Wei Tao, Zhisong Pan, Gaowei Wu, and Qing Tao. Primal averaging: A new gradient evaluation step to attain the optimal individual convergence. *IEEE Transactions on Cybernetics*, 50(2):835–845, 2020.
- [48] Sattar Vakili, Kebin Liu, and Qing Zhao. Deterministic sequencing of exploration and exploitation for multi-armed bandit problems. *IEEE Journal of Selected Topics in Signal Processing*, 7(5):759–767, 2013.
- [49] Hal R. Varian. A model of sales. *American Economic Review*, 70(4):651–659, 1980.
- [50] Yuanhao Wang, Dingwen Kong, Yu Bai, and Chi Jin. Learning rationalizable equilibria in multiplayer games. In *International Conference on Learning Representations*, 2023.

Appendix

We follow the order of the main results. In [Section A](#), we verify the standard algorithms, prove class inclusion, and show that no regret does not imply WMBA. In [Section B](#), we prove rationalizability, the MD results, and robustness to exploration and estimated rewards. We then characterize the limits of empirical strategies and the extrema in [Section C](#), and verify the market instances in [Section D](#). We restate each main result before its proof.

Notation. We use $\|\cdot\|$ for the ℓ_1 norm and $[z]_+ \doteq \max\{z, 0\}$. When discussing one learning algorithm, we omit its player index and write $\mathcal{A}$ for its own action set. We use the *score* of an action, the cumulative payoff it would have earned, and the *regret* of a joint distribution,

$$U_{n,T}(a) \doteq T u_n(a, \bar{\mathbf{p}}_{-n,T}), \quad \text{Reg}_n(\boldsymbol{\mu}) \doteq \max_{a \in \mathcal{A}_n} u_n(a, \boldsymbol{\mu}_{-n}) - u_n(\boldsymbol{\mu}).$$

Thus $T\delta_{n,T}(a) = \max_{a'} U_{n,T}(a') - U_{n,T}(a)$, player n 's cumulative regret is $\text{Reg}_{n,T} \doteq T \text{Reg}_n(\bar{\mathbf{p}}_T)$, (2) reads $\limsup_T \text{Reg}_n(\bar{\mathbf{p}}_T) \leq 0$, and $\boldsymbol{\mu}$ is a CCE exactly when $\text{Reg}_n(\boldsymbol{\mu}) \leq 0$ for every n . Since payoffs lie in $[0, 1]$, they are 1-Lipschitz in $\|\cdot\|$ in each argument. Hence Reg_n is convex and 2-Lipschitz, and so is the gap $\delta_{n,T}(a)$ as a function of $\bar{\mathbf{p}}_T$.

A Algorithms

Euclidean FTRL is the fallback in every construction of [Section C](#), so we record its properties first. We then show that FTRL is mean-based for a general class of regularizers, check two further algorithms, prove [Proposition 1](#), and show that no regret does not imply WMBA.

A.1 Euclidean FTRL

For the fallback used in our constructions, we fix one player and use *full-history Euclidean FTRL* from round $s + 1$. For every $t \geq s$, we play

$$p_{t+1} = \arg \max_{p \in \Delta(\mathcal{A})} \left\{ \langle U_t, p \rangle - \frac{\sqrt{t}}{2} \|p\|_2^2 \right\}. \quad (13)$$

The maximizer is unique and, for the λ_t that makes it a probability vector,

$$p_{t+1}(a) = \frac{[U_t(a) - \lambda_t]_+}{\sqrt{t}}. \quad (14)$$

Lemma 1. *Let a learning algorithm play arbitrarily through round $s \geq 1$ and (13) afterwards. Then, for all $t \geq s$ and $T \geq s$,*

$$\delta_t(a) > t^{-1/2} \implies p_{t+1}(a) = 0, \quad \text{and} \quad \text{Reg}_T \leq \text{Reg}_s + (|\mathcal{A}| + \tfrac{1}{2})\sqrt{T}.$$

Proof. If $p_{t+1}(a) > 0$ and a' maximizes U_t , then (14) yields

$$t \delta_t(a) = U_t(a') - U_t(a) = \sqrt{t} [p_{t+1}(a') - p_{t+1}(a)] \leq \sqrt{t}.$$

For the regret bound, let

$$V_t(U) \doteq \max_{p \in \Delta(\mathcal{A})} \left\{ \langle U, p \rangle - \frac{\sqrt{t}}{2} \|p\|_2^2 \right\}.$$

For $t \geq 2$, $V_t \leq V_{t-1}$, $\nabla V_{t-1}(U_{t-1}) = p_t$ whenever p_t is given by (13), and V_{t-1} has $(t-1)^{-1/2}$ -Lipschitz gradient in the Euclidean norm, by strong convexity of the regularizer [16]. With $u_t \doteq U_t - U_{t-1} \in [0, 1]^{|\mathcal{A}|}$, every $t \geq s+1$ therefore satisfies

$$V_t(U_t) - V_{t-1}(U_{t-1}) \leq V_{t-1}(U_t) - V_{t-1}(U_{t-1}) \leq \langle p_t, u_t \rangle + \frac{\|u_t\|_2^2}{2\sqrt{t-1}}.$$

We sum over $s < t \leq T$ and use

$$\sum_{t=s+1}^T \frac{1}{\sqrt{t-1}} \leq 2\sqrt{T}, \quad \max_a U_T(a) \leq V_T(U_T) + \frac{\sqrt{T}}{2}, \quad V_s(U_s) \leq \max_a U_s(a)$$

to obtain

$$\text{Reg}_T = \max_a U_T(a) - \sum_{t \leq T} \langle p_t, u_t \rangle \leq \frac{\sqrt{T}}{2} + V_s(U_s) - \sum_{t \leq s} \langle p_t, u_t \rangle + |\mathcal{A}| \sqrt{T} \leq \text{Reg}_s + (|\mathcal{A}| + \frac{1}{2}) \sqrt{T}. \quad \square$$

With $s = 1$, Euclidean FTRL is an MBA with $\gamma_T = T^{-1/2}$, and its regret is at most $1 + (|\mathcal{A}| + \frac{1}{2}) \sqrt{T}$ on every reward history. The restart form is what the constructions need: a learning algorithm can follow any plan and still recover both guarantees by switching to (13) at any time.

A.2 General FTRL and Other Algorithms

Having verified Euclidean FTRL, we now show that FTRL is mean-based for a general class of regularizers. With learning rates $\eta_T > 0$, FTRL plays

$$p_{T+1} = \arg \max_{p \in \Delta(\mathcal{A})} \left\{ \langle U_T, p \rangle - \frac{1}{\eta_T} \sum_a \psi(p(a)) \right\}, \quad (15)$$

where $\psi : [0, 1] \rightarrow (-\infty, +\infty]$ is continuous, strictly convex, and differentiable on $(0, 1]$. This covers entropy, Tsallis, log-barrier, and Euclidean regularizers, and strict convexity makes the maximizer unique. For instance, Hedge plays $p_{T+1}(a) \propto \exp(\eta_T U_T(a))$, which is (15) with $\psi(x) = x \log x$.

Lemma 2. *If $\eta_T T$ is nondecreasing and unbounded, then (15) is an MBA.*

Proof. Fix T , write $p \doteq p_{T+1}$, and let a' maximize U_T . First, $p(a') > 0$, since otherwise moving a little mass to a' from an action with positive probability would increase the objective, by strict convexity of ψ . Now let $p(a) > 0$. Since moving a little mass between a and a' in either direction cannot increase the objective, the first-order condition and $p(a') \leq 1$ yield

$$\psi'(p(a)) = \psi'(p(a')) - \eta_T (U_T(a') - U_T(a)) \leq \psi'(1) - \eta_T T \delta_T(a).$$

For $z \geq 0$, let $g(z) \doteq \sup\{x \in (0, 1] : \psi'(x) \leq \psi'(1) - z\}$, with $\sup \emptyset \doteq 0$. Then g is nonincreasing, $g(z) \rightarrow 0$ as $z \rightarrow \infty$, and $p(a) \leq g(\eta_T T \delta_T(a))$ for every action. Hence

$$\gamma_T \doteq \max\{(\eta_T T)^{-1/2}, g((\eta_T T)^{1/2})\}$$

is nonincreasing and tends to zero. Finally, $\delta_T(a) > \gamma_T$ implies $\eta_T T \delta_T(a) > (\eta_T T)^{1/2}$, and hence

$$p(a) \leq g(\eta_T T \delta_T(a)) \leq g((\eta_T T)^{1/2}) \leq \gamma_T. \quad \square$$

With $\eta_T \propto T^{-1/2}$, the entropy, Tsallis, and Euclidean regularizers, which are strongly convex with bounded range, also give no regret [22].

To include a learning algorithm outside FTRL, we consider *regret matching*, which plays $p_{T+1}(a) \propto [U_T(a) - T u(\bar{p}_T)]_+$, or an action maximizing U_T when all these are zero, and has $\text{Reg}_T \leq C\sqrt{T}$ on every reward history for a constant C depending only on $|\mathcal{A}|$ [16]. Since

$$U_T(a) - T u(\bar{p}_T) = \text{Reg}_T - T\delta_T(a),$$

every action with $\delta_T(a) > C/\sqrt{T}$ gets probability zero, so $\gamma_T = C/\sqrt{T}$ works.

MBA alone does not imply no regret, however. *Follow-the-leader*, which is FP in games, puts all its probability on actions with zero gap, so it is an MBA for any nonincreasing $\gamma_T \rightarrow 0$. With two actions, let the payoff vector be $(\frac{1}{2}, 0)$ in the first round and then alternate between $(0, 1)$ and $(1, 0)$, so that the leader changes every round. After the first round, follow-the-leader earns 0 per round, while each fixed action earns about $\frac{1}{2}$ per round on average, so its regret grows linearly in T [16].

A.3 Proof of Proposition 1

Proposition 1. $\text{MBA} \subset \text{WMBA}$.

Proof. If $\liminf_T \delta_T(a) > 0$, then eventually $\delta_T(a) > \gamma_T$, so $p_{T+1}(a) \leq \gamma_T \rightarrow 0$ and $\bar{p}_T(a) \rightarrow 0$ by Cesàro averaging. For strictness, let a learning algorithm follow the leader except on rounds 2^k , where it plays a fixed action a_2 . On every reward history, a highly sub-optimal action eventually stops leading and is then played only on the $O(\log T)$ rounds 2^k , so the learning algorithm is a WMBA. Against the one-player payoffs $u(a_1) = 1$ and $u(a_2) = 0$, however, $\delta_{2^k-1}(a_2) = 1$ while $p_{2^k}(a_2) = 1$ for every k , which violates Definition 1. $\square$

A.4 No Regret Does Not Imply WMBA

Having shown that MBAs are WMBAs, we now classify the two learning algorithms in Table 1 that have no regret but are not WMBAs, and hence not MBAs. Since regret compares the payoff with that of the best fixed action over the whole history, a learning algorithm can play an action that is behind on average, as long as it plays it when it pays off. Both examples use two actions and only the reward vectors $(1, 0)$, which favors a_1 by 1, and $(\frac{1}{2}, 1)$, which favors a_2 by $\frac{1}{2}$.

Strongly adaptive algorithms. These learning algorithms keep the regret against every fixed action small on every interval of rounds [18]. For prediction with expert advice, for instance, their regret on every interval $[q, s]$ is at most

$$O((\sqrt{\log |\mathcal{A}|} + \log(s+1))\sqrt{s-q+1}).$$

We use only that, on every reward history, this regret is at most some $\rho_T = o(T)$ on every interval contained in $[1, T]$. Taking the interval $[1, T]$ itself, we see that they have no regret. For $k \geq 0$, we now supply $(1, 0)$ in rounds $3^k \leq t < 2 \cdot 3^k$ and $(\frac{1}{2}, 1)$ in rounds $2 \cdot 3^k \leq t < 3^{k+1}$. Then $U_T(a_1) - U_T(a_2)$ rises by 1 in each round of the first kind and falls by $\frac{1}{2}$ in each round of the second kind, so it is at least $T/4$ for every T , with equality at $T = 3^{k+1} - 1$, and $\delta_T(a_2) \geq \frac{1}{4}$. Let $T \doteq 3^{k+1} - 1$ be the last round of $I_k \doteq [2 \cdot 3^k, 3^{k+1})$. On I_k , the learning algorithm earns $\sum_{t \in I_k} (1 - p_t(a_1)/2)$, against $|I_k| = 3^k$ for a_2 , so $\sum_{t \in I_k} p_t(a_1) \leq 2\rho_T$. Hence

$$\bar{p}_T(a_2) \geq \frac{3^k - 2\rho_T}{3^{k+1} - 1} \rightarrow \frac{1}{3},$$

so $\bar{p}_T(a_2)$ does not tend to zero, and the learning algorithm is not a WMBA.

Periodic schedule. On the history that supplies $(1, 0)$ in odd rounds and $(\frac{1}{2}, 1)$ in even rounds, $\delta_T(a_2) \rightarrow \frac{1}{4}$, yet playing a_1 in odd rounds and a_2 in even rounds earns 1 per round. As in the proof of [Lemma 6](#), we play this sequence as long as the observed reward vectors agree with this history, and switch permanently to Euclidean FTRL after the first round s on which they differ. Since no action earns more than 1 per round, the regret is nonpositive before round s and $\text{Reg}_s \leq 1$, so [Lemma 1](#) yields $\text{Reg}_T \leq 1 + \frac{5}{2}\sqrt{T}$ on every reward history. On the history itself, however, $\bar{p}_T(a_2) \rightarrow \frac{1}{2}$, so the learning algorithm is not a WMBA.

B Guarantees

We follow [Sections 4.1 to 4.4](#): rationalizability, the MD families, exploration, and estimated rewards. We begin by showing that every action removed by IESDS trails by a uniform margin ([Lemma 3](#)).

B.1 The CRS

We first identify the terminal set of the deletion process. Since $\mathcal{A}_{k+1} \subseteq \mathcal{A}_k$, IESDS [\(3\)](#) stops after finitely many stages at the CRS, and every action in $\tilde{\mathcal{A}}_n$ is a best response to some belief in $\Delta(\tilde{\mathcal{A}}_n)$. Conversely, if a product B has this property, then $B \subseteq \mathcal{A}_k$ implies $B \subseteq \mathcal{A}_{k+1}$, so $B \subseteq \tilde{\mathcal{A}}$: the CRS is the largest such product. By the minimax theorem, an action lacks the property exactly when some mixed strategy strictly dominates it against the opponent profiles in the set, which is why the CRS consists of the recursively undominated actions.

Lemma 3. *If $a \in \mathcal{A}_{n,k} \setminus \mathcal{A}_{n,k+1}$, there is $\kappa > 0$ such that*

$$\max_{a'} u_n(a', \mathbf{q}) - u_n(a, \mathbf{q}) \geq \kappa \quad \text{for all } \mathbf{q} \in \Delta(\mathcal{A}_{-n,k}).$$

Proof. The left side is continuous and, by [\(3\)](#), positive on the compact set $\Delta(\mathcal{A}_{-n,k})$. $\square$

B.2 Proof of [Theorem 1](#)

Theorem 1. *WMBA dynamics guarantee $\bar{p}_T(\tilde{\mathcal{A}}) \rightarrow 1$, although $\mathbf{p}_T(\tilde{\mathcal{A}})$ need not converge to 1.*

Proof. We follow the stages of IESDS in [\(3\)](#) one at a time, showing by induction on k that, for every player n , $\bar{p}_{n,T}(\mathcal{A}_n \setminus \mathcal{A}_{n,k}) \rightarrow 0$ and every action outside $\mathcal{A}_{n,k}$ is highly sub-optimal. The case $k = 0$ holds trivially. Assume the claim for k , and fix n and $a \in \mathcal{A}_{n,k} \setminus \mathcal{A}_{n,k+1}$. By the union bound, the opponents' empirical strategy puts vanishing mass outside $\mathcal{A}_{-n,k}$:

$$\varepsilon_T \doteq 1 - \bar{\mathbf{p}}_{-n,T}(\mathcal{A}_{-n,k}) \leq \sum_{n' \neq n} \bar{p}_{n',T}(\mathcal{A}_{n'} \setminus \mathcal{A}_{n',k}) \rightarrow 0.$$

For large T , decompose

$$\bar{\mathbf{p}}_{-n,T} = (1 - \varepsilon_T) \mathbf{q}_T + \varepsilon_T \mathbf{q}'_T, \quad \mathbf{q}_T \in \Delta(\mathcal{A}_{-n,k}), \quad \mathbf{q}'_T \in \Delta(\mathcal{A}_{-n}).$$

For $a' \in \text{BR}_n(\mathbf{q}_T)$, we use [Lemma 3](#) and $u_n \in [0, 1]$ to obtain, for large T ,

$$\delta_{n,T}(a) \geq u_n(a', \bar{\mathbf{p}}_{-n,T}) - u_n(a, \bar{\mathbf{p}}_{-n,T}) \geq (1 - \varepsilon_T)\kappa - \varepsilon_T \geq \kappa/2.$$

So a is highly sub-optimal, and $\bar{p}_{n,T}(a) \rightarrow 0$ by [\(1\)](#). Summing over the finitely many actions removed at stage k proves the claim for $k + 1$. At the final stage,

$$1 - \bar{\mathbf{p}}_T(\tilde{\mathcal{A}}) \leq \sum_n \bar{p}_{n,T}(\mathcal{A}_n \setminus \tilde{\mathcal{A}}_n) \rightarrow 0.$$

For the last iterate, run the learning algorithm from the proof of [Proposition 1](#) on the one-player payoffs $u(a_1) = 1$ and $u(a_2) = 0$. The CRS is $\{a_1\}$, yet $p_{2^k}(a_2) = 1$ for every k , so $\mathbf{p}_T(\tilde{\mathcal{A}})$ does not converge to 1. $\square$

B.3 Realized Actions

We next transfer the time-average guarantee to the sampled actions. If each $a_{n,t}$ is sampled from $p_{n,t}$, independently across players given the history, then $\mathbf{1}\{a_{n,t} = a\} - p_{n,t}(a)$ is a bounded martingale-difference sequence, so realized frequencies differ from $\bar{p}_{n,T}$ by $o(1)$ a.s. Writing $\mathbf{a}_t \doteq (a_{1,t}, \dots, a_{N,t})$, we conclude that realized actions outside $\tilde{\mathcal{A}}_n$ have vanishing frequency for every n , and so do realized profiles outside $\tilde{\mathcal{A}}$, since

$$\mathbf{1}\{\mathbf{a}_t \notin \tilde{\mathcal{A}}\} \leq \sum_n \mathbf{1}\{a_{n,t} \notin \tilde{\mathcal{A}}_n\}.$$

B.4 Mirror Descent

Having proved rationalizability, we verify the algorithm families introduced in [Section 4.2](#). We begin with [Proposition 2](#), using one player's reward vectors u_t and scores $U_T = \sum_{t \leq T} u_t$. In [\(4\)](#), we update from the current distribution and latest reward, rather than apply a regularizer directly to U_T as in FTRL [\(15\)](#). With decreasing learning rates, we must distinguish these updates [\[22\]](#).

Regularizers. To make the MD update precise, we take Ψ to be continuously differentiable and strictly convex on the interior of the simplex, and interpret its values at the boundary by limits, allowing $+\infty$. For two actions, let $\phi(x) \doteq \Psi(x, 1-x)$. Our steepness assumption is

$$\lim_{x \rightarrow 0} \phi'(x) = -\infty, \quad \lim_{x \rightarrow 1} \phi'(x) = +\infty. \quad (16)$$

We start at $0 < p_1(a_1) < 1$. We then obtain a unique solution in $(0, 1)$ from the first-order condition at every update. We use the following regularizers for both results:

$$\begin{aligned} \text{entropy:} \quad & \Psi(p) \doteq \sum_a p(a) \log p(a), \\ \text{Tsallis:} \quad & \Psi(p) \doteq -\frac{1}{1-\beta} \sum_a p(a)^\beta, \quad 0 < \beta < 1, \\ \text{log-barrier:} \quad & \Psi(p) \doteq -\sum_a \log p(a). \end{aligned}$$

We compute their coordinate derivatives as $1 + \log x$, $-\beta x^{\beta-1}/(1-\beta)$, and $-1/x$, respectively. Each tends to $-\infty$ at zero, and we verify [\(16\)](#) for its two-action restriction. We can also add a quadratic term $\lambda \|p\|_2^2/2$, $\lambda \geq 0$, without changing these boundary limits.

Proposition 2. *MD (2 actions) with a steep regularizer is WMBA but not MBA.*

Proof. We first prove the WMBA guarantee under a rate condition that includes $\eta_t = ct^{-\alpha}$, $0 < \alpha \leq 1$:

$$\eta_t > 0, \quad \eta_{t+1} \leq \eta_t, \quad \sum_{t=1}^{\infty} \eta_t = \infty.$$

Writing $x_t \doteq p_t(a_1)$, we obtain from the first-order condition

$$\phi'(x_{T+1}) = \phi'(x_1) + \sum_{t=1}^T \eta_t (u_t(a_1) - u_t(a_2)). \quad (17)$$

Suppose a_1 is highly sub-optimal. With two actions, $d_T \doteq U_T(a_2) - U_T(a_1) \geq \varepsilon T$ eventually for some $\varepsilon > 0$. By summation by parts, we obtain

$$\begin{aligned} \sum_{t=1}^T \eta_t (u_t(a_2) - u_t(a_1)) &= \eta_T d_T + \sum_{t=1}^{T-1} (\eta_t - \eta_{t+1}) d_t \\ &\geq \varepsilon \left[T\eta_T + \sum_{t=1}^{T-1} t(\eta_t - \eta_{t+1}) \right] - C = \varepsilon \sum_{t=1}^T \eta_t - C \rightarrow \infty, \end{aligned}$$

where C accounts for the finite prefix before $d_t \geq \varepsilon t$. Using (17), we obtain $\phi'(x_{T+1}) \rightarrow -\infty$ and hence $p_{T+1}(a_1) \rightarrow 0$ by (16). We argue identically for a_2 . We obtain (1) by averaging, on every reward history.

To distinguish these learning algorithms from MBAs, we now return to $\eta_t = ct^{-\alpha}$, $0 < \alpha \leq 1$, and construct a reward history violating the next-round bound. In cycle $k \geq 1$, we supply $m_k \doteq 2^{k!}$ reward vectors $(0, 1)$, followed by $\lfloor (1 + \alpha)m_k \rfloor$ reward vectors $(1, 0)$. Let s_k and T_k be the times before and after this cycle. We have $s_k = o(m_k)$ and $\log(1 + s_k) = o(\log m_k)$. At the cycle endpoints,

$$\delta_{T_k}(a_2) \rightarrow \frac{\alpha}{2 + \alpha} > 0. \quad (18)$$

For $0 < \alpha < 1$, we sum the learning rates in the two phases of (17) to obtain

$$\phi'(x_{T_k+1}) = \frac{c}{1 - \alpha} [(2 + \alpha)^{1-\alpha} - 2] m_k^{1-\alpha} + o(m_k^{1-\alpha}) \rightarrow -\infty.$$

We obtain a negative coefficient because $(1 - \alpha) \log(2 + \alpha)$ strictly decreases from $\log 2$, and we bound the preceding cycles by $O(1 + s_k^{1-\alpha})$. For $\alpha = 1$, the first phase contributes $-c \log m_k + O(\log(1 + s_k))$, the second contributes $c \log 3 + o(1)$, and the preceding history contributes $O(\log(1 + s_k))$. Again $\phi'(x_{T_k+1}) \rightarrow -\infty$, so

$$p_{T_k+1}(a_2) \rightarrow 1.$$

Together with (18), we obtain a violation of Definition 1 for every threshold $\gamma_T \rightarrow 0$. Each action becomes a leader during its own rewarding phases, so neither is highly sub-optimal on this history. $\square$

No regret. To check the classification in Table 1, we use the same history to show that this family does not guarantee no regret. During the second phase, $\phi'(x_t)$ increases but its endpoint still tends to $-\infty$, so $p_t(a_1) = o(1)$ uniformly throughout that phase. We earn at most $m_k + o(m_k)$ during cycle k , whereas always choosing a_1 earns $(1 + \alpha)m_k + O(1)$. The preceding history is negligible, and therefore

$$\liminf_{k \rightarrow \infty} \frac{\text{Reg}_{T_k}}{T_k} \geq \frac{\alpha}{2 + \alpha} > 0.$$

B.5 Averaged Mirror Descent

To extend to multiple actions, we average the MD proposals as in (5). We recall $\Psi(p) = \sum_{a \in \mathcal{A}} \psi_a(p(a))$ from Section 4.2, where each ψ_a is continuously differentiable and strictly convex on $(0, 1)$ with $\psi'_a(0+) = -\infty$. We assume $|\mathcal{A}| \geq 2$, use $\eta > 0$, and start with positive probabilities. The proposals and played distributions satisfy

$$\hat{p}_{t+1} = \arg \max_{p \in \Delta(\mathcal{A})} \{ \eta \langle u_t, p \rangle - D_\Psi(p, \hat{p}_t) \}, \quad p_t = \frac{1}{t} \sum_{s=1}^t \hat{p}_s. \quad (19)$$

Unrolling (5) yields the second identity. We update proposals from the reward vectors observed after playing p_t , so the averaging determines the choices that generate our feedback [47].

Proposition 3. *AMD is WMBA but not MBA.*

Proof. We first check that each proposal exists and has positive coordinates. With the boundary limits specified above, the objective in (19) attains a maximum on the compact simplex. A point with infinite penalty cannot maximize it, and at any other boundary point, moving a small mass to a zero coordinate decreases the regularizer by more than the linear term can offset. Strict convexity then yields uniqueness. We subtract the first-order conditions for actions a, a' and sum over time to obtain

$$\psi'_{a'}(\widehat{p}_{T+1}(a')) - \psi'_a(\widehat{p}_{T+1}(a)) = \psi'_{a'}(\widehat{p}_1(a')) - \psi'_a(\widehat{p}_1(a)) + \eta(U_T(a') - U_T(a)). \quad (20)$$

To verify WMBA, suppose a is highly sub-optimal and choose a'_T maximizing U_T . The right-hand side of (20) with $a' = a'_T$ tends to $+\infty$, since the initial derivative differences are fixed and there are finitely many comparators. If $\widehat{p}_{T+1}(a) \geq \varepsilon$ infinitely often for some $\varepsilon \in (0, 1)$, then $a'_T \neq a$ eventually and $\widehat{p}_{T+1}(a'_T) \leq 1 - \varepsilon$. The left-hand side is then at most

$$\max_{a' \neq a} \psi'_{a'}(1 - \varepsilon) - \psi'_a(\varepsilon),$$

a finite constant. From this contradiction, we obtain $\widehat{p}_t(a) \rightarrow 0$. By averaging, we obtain $p_t(a) \rightarrow 0$ and then $\bar{p}_T(a) \rightarrow 0$, proving WMBA. We will also use the same argument in the following form: whenever one action's score exceeds every other score by an amount tending to infinity, its proposal probability tends to one. From the bound above, we also obtain uniform convergence whenever the score differences diverge uniformly.

To show why averaging need not preserve MBA, we now use two actions a_1, a_2 and set $m_k \doteq 2^{k!}$. In each cycle, we reward a_2 by one for m_k rounds, then a_1 by one for $2m_k$ rounds, and we set all other rewards to zero. Write s_k for the starting time and $T_k \doteq s_k + 3m_k$ for the endpoint, so $s_k = o(m_k)$. For each fixed $0 < \tau < 3$ with $\tau \neq 2$, the leading score at $s_k + \lfloor \tau m_k \rfloor$ exceeds every other score by an amount proportional to m_k . Applying the preceding bound to the next proposal, we obtain

$$\widehat{p}_{s_k + \lfloor \tau m_k \rfloor + 1}(a_2) \longrightarrow \begin{cases} 1, & 0 < \tau < 2, \\ 0, & 2 < \tau < 3. \end{cases}$$

We obtain uniform convergence on closed subintervals of $(0, 2)$ and $(2, 3)$. Since probabilities are bounded by one and $s_k = o(m_k)$, summing the proposals yields

$$p_{T_k+1}(a_2) \longrightarrow \frac{2}{3}, \quad \delta_{T_k}(a_2) \longrightarrow \frac{1}{3}. \quad (21)$$

We have therefore violated Definition 1. □

Three active actions. To see that the distinction is not confined to two rewarding actions, we next use a history in which all three actions become leaders. In cycle $k \geq 0$, let $m \doteq 8^k$ and reward a_1 for m rounds, a_2 for $2m$ rounds, then a_3 for $4m$ rounds, with the other rewards zero. At the endpoint $T_k = 8^{k+1} - 1$, the cumulative scores have ratio $1 : 2 : 4$, so $\delta_{T_k}(a_2) = 2/7$. Before the cycle they are $(s, 2s, 4s)$, where $s \doteq (m - 1)/7$. During the last m rounds rewarding a_2 , its score exceeds that of a_1 by at least $s - 1$ and that of a_3 by at least $m - 2s - 1$. Using the score bound above, we conclude that its proposal probability tends uniformly to one on these rounds. Since $T_k + 1 = 8m$, we have

$$p_{T_k+1}(a_2) \geq \frac{1}{8} - o(1), \quad \delta_{T_k}(a_2) = \frac{2}{7}.$$

All three actions become leaders during their respective phases. We thus establish the same strict distinction using each of the three actions.

No regret. To check no regret, we return to the two-action history used for (21). Arithmetic averaging yields $p_{s_k + \lfloor \tau m_k \rfloor + 1}(a_2) \rightarrow 1$ for $0 < \tau \leq 2$ and $2/\tau$ for $2 < \tau \leq 3$. The probability of a_1 has the complementary limit, and the other probabilities vanish. Since the probabilities are bounded, we integrate these limits over the rewarding phases: our payoff in cycle k , divided by m_k , tends to

$$1 + \int_2^3 \left(1 - \frac{2}{\tau}\right) d\tau = 2 - 2\log(3/2).$$

The fixed action a_1 earns $2m_k$, so $\text{Reg}_{T_k}/T_k \rightarrow \frac{2}{3}\log(3/2) > 0$. We have proved the table's no-regret classification for the whole family.

Further averaging choices. The inclusion argument also applies beyond MD. Any WMBA proposal sequence on the supplied history has $T^{-1} \sum_{t \leq T} \hat{p}_t(a) \rightarrow 0$ for every highly sub-optimal action. By playing that average, we obtain a WMBA. When proposals themselves satisfy $\hat{p}_t(a) \rightarrow 0$, we may also use $p_{t+1} = (1 - \rho_t)p_t + \rho_t \hat{p}_{t+1}$ with $0 < \rho_t \leq 1$ and $\sum_t \rho_t = \infty$. After $\hat{p}_t(a) \leq \varepsilon$, we have

$$[p_{t+1}(a) - \varepsilon]_+ \leq (1 - \rho_t)[p_t(a) - \varepsilon]_+,$$

so $p_t(a) \rightarrow 0$. The non-MBA assertion in Proposition 3 uses the arithmetic average in (5).

B.6 Exploration

After verifying the MD families, we show how to preserve WMBA when adding exploration. We bound the change in each action's frequency by the total deviation.

Theorem 2. *If $\mathcal{L}$ is a WMBA, so is the modified learning algorithm. The bound $o(T)$ in (6) cannot be replaced by εT for any $\varepsilon > 0$.*

Proof. For every action,

$$\bar{p}_T(a) \leq \frac{1}{T} \sum_{t \leq T} \hat{p}_t(a) + \frac{1}{T} \sum_{t \leq T} \|p_t - \hat{p}_t\|.$$

If $\liminf_T \delta_T(a) > 0$, the first term vanishes because $\mathcal{L}$ is a WMBA on the realized reward history, and the second by (6). For the second claim, fix $\rho \in (0, 1)$ and one player with $u(a_1) = 1$ and $u(a_2) = 0$, let $\mathcal{L}$ follow the leader with ties toward a_1 , and play $p_t(a_1) = 1 - \rho$ and $p_t(a_2) = \rho$. The total deviation is $2\rho T$, yet $\delta_T(a_2) = 1$ and $\bar{p}_T(a_2) = \rho$. $\square$

Corollary 1. *If $p_t = \hat{p}_t$ except on $o(T)$ of the first T rounds, the modified learning algorithm is a WMBA.*

Proof. $\|p_t - \hat{p}_t\| \leq 2$ on the exceptional rounds and vanishes otherwise, so the total deviation through round T is at most twice the number of exceptional rounds, which is $o(T)$, and Theorem 2 applies. $\square$

The same bound covers the mixed exploration of Section 4.3, since $p_t = (1 - \varepsilon_t)\hat{p}_t + \varepsilon_t \nu_t$ yields $\|p_t - \hat{p}_t\| \leq 2\varepsilon_t$.

B.7 Proof of Theorem 3

To extend the guarantee to estimated rewards, we show that vanishing estimation errors cannot conceal a highly sub-optimal action.

Theorem 3. *Under (7) and (8), $\bar{\mathbf{p}}_T(\tilde{\mathcal{A}}) \rightarrow 1$ a.s.*

Proof. Fix a sample path on which (8) holds. For all n and a ,

$$|\hat{\delta}_{n,T}(a) - \delta_{n,T}(a)| \leq 2 \max_{a'} |\hat{u}_{n,T}(a') - u_n(a', \bar{\mathbf{p}}_{-n,T})| \rightarrow 0,$$

so $\liminf_T \delta_{n,T}(a) > 0$ implies $\liminf_T \hat{\delta}_{n,T}(a) > 0$ and hence $\bar{p}_{n,T}(a) \rightarrow 0$ by (7). Every player thus satisfies (1) along the path, and the proof of Theorem 1 yields $\bar{\mathbf{p}}_T(\tilde{\mathcal{A}}) \rightarrow 1$. As (8) holds a.s., so does the conclusion. $\square$

C Limits of Empirical Strategies and Welfare

We now turn from algorithm membership and rationalizability to limits of empirical strategies and welfare. For necessity, we use one observation: a persistent action must be a best response along some subsequence of rounds, or it would be highly sub-optimal. For sufficiency, we use one construction: players follow a deterministic *schedule* $\sigma \doteq (\sigma_t)_{t \geq 1}$ of pure profiles unless another player departs from it. Playing σ means $\mathbf{p}_t \doteq \sigma_t$, viewed as a point mass, so $\bar{\mathbf{p}}_T$, $\delta_{n,T}$, and $\text{Reg}_{n,T}$ become the empirical quantities of the sequence. We develop the constructions first.

C.1 Constructions

Three lemmas cover the constructions: one target can be tracked with bounded error (Lemma 4), several can be visited so that the worst decides the $\liminf$ (Lemma 5), and any schedule can be played by algorithms that keep, on every reward history, the properties it has along its own path (Lemma 6).

Lemma 4. *For every $\mu \in \Delta(\mathcal{A})$, some schedule plays only profiles in the support of μ and has $|T\bar{\mathbf{p}}_T(\mathbf{a}) - T\mu(\mathbf{a})| < |\mathcal{A}|$ for all T and $\mathbf{a}$.*

Proof. Let $d_t(\mathbf{a}) \doteq t\mu(\mathbf{a}) - N_t(\mathbf{a})$, where $N_t(\mathbf{a})$ counts the plays of $\mathbf{a}$ through round t , so that $\sum_{\mathbf{a}} d_t(\mathbf{a}) = 0$. In round $t + 1$, play a profile maximizing $d_t(\mathbf{a}) + \mu(\mathbf{a})$. These values sum to one, so their maximum is at least $1/|\mathcal{A}|$: the chosen profile keeps $d_{t+1}(\mathbf{a}) \geq 1/|\mathcal{A}| - 1 > -1$, and every other deficit weakly increases. By induction all deficits exceed -1 , and since they sum to zero, each is below $|\mathcal{A}| - 1$. A profile with $\mu(\mathbf{a}) = 0$ has $d_t(\mathbf{a}) + \mu(\mathbf{a}) \leq 0$ and is never chosen. $\square$

Lemma 5. *For any $\mu^{(1)}, \dots, \mu^{(J)} \in \Delta(\mathcal{A})$, some schedule plays only profiles in their supports, keeps $\bar{\mathbf{p}}_T$ within $o(1)$ of their convex hull, and has $\bar{\mathbf{p}}_T \rightarrow \mu^{(j)}$ along a subsequence for every j . Along it,*

- (i) $\liminf_T f(\bar{\mathbf{p}}_T) = \min_j f(\mu^{(j)})$ for every linear f ;
- (ii) every player has no regret if every $\mu^{(j)}$ is a CCE;
- (iii) $\liminf_T \delta_{n,T}(a) = 0$ whenever $a \in \text{BR}_n(\mu_{-n}^{(j)})$ for some j .

Proof. To visit all targets, we play blocks of lengths 2^{2^k} , $k = 1, 2, \dots$, cycling through the targets and tracking the current target from the start of its block (Lemma 4). Each block contributes a count error below $|\mathcal{A}|^2$, and only $O(\log \log T)$ blocks have started by round T , so $\bar{\mathbf{p}}_T$ is within $o(1)$ of a convex combination of the targets. Each block is longer than all earlier blocks together by a factor tending to infinity, so at the end of each block, $\bar{\mathbf{p}}_T$ is within $o(1)$ of that block's target. Claim (i) follows because f is linear and bounded; (ii) because Reg_n is Lipschitz, convex, and nonpositive at every target; and (iii) because $\delta_{n,T}(a)$ is Lipschitz in $\bar{\mathbf{p}}_T$ and vanishes at every target under which a is a best response. $\square$

Lemma 6. *Every schedule σ is played in self-play by a profile of algorithms in which player n 's algorithm*

- (i) *is an MBA if, along σ , $\delta_{n,t}(\sigma_{n,t+1}) \leq C/\sqrt{t}$ for all t and some constant C ;*
- (ii) *is a WMBA if (1) holds for player n along σ ;*
- (iii) *has no regret if (2) holds for player n along σ .*

Proof. To implement the schedule on every reward history, we have player n play $\sigma_{n,t+1}$ as long as the reward vectors observed so far are those σ generates, $u_n(\cdot, \sigma_{-n,\tau})$ for $\tau \leq t$, and switch permanently to full-history Euclidean FTRL after the first round s on which they differ. In self-play no difference arises, so the profile plays σ , and on every history that agrees with σ forever each property holds by assumption. On any other history, Lemma 1 applies with this s . (i) In rounds $t+1 \leq s$ the only action played has gap at most $C/\sqrt{t}$, and in later rounds actions with gap above $t^{-1/2}$ get probability zero, so Definition 1 holds with $\gamma_t = \min\{1, \max\{C, 1\}/\sqrt{t}\}$ on every history. (ii) An action with $\liminf_t \delta_{n,t}(a) > 0$ eventually has gap above $t^{-1/2}$ at every round, so it is played on only finitely many rounds and $\bar{p}_{n,T}(a) \rightarrow 0$. (iii) With $\text{Reg}_{n,s-1}$ the regret along σ through round $s-1$,

$$\text{Reg}_{n,T} \leq \text{Reg}_{n,s} + (|\mathcal{A}_n| + \frac{1}{2})\sqrt{T} \leq \text{Reg}_{n,s-1} + 1 + (|\mathcal{A}_n| + \frac{1}{2})\sqrt{T},$$

and as s is fixed, $\limsup_T \text{Reg}_{n,T}/T \leq 0$. $\square$

C.2 Limits of Empirical Strategies

With the constructions in place, we first relate each MBA decision to the observed empirical strategies.

Proposition 4. *If player n uses an MBA, then*

$$\max_{a \in \mathcal{A}_n} u_n(a, \bar{\mathbf{p}}_{-n,T}) - u_n(p_{n,T+1}, \bar{\mathbf{p}}_{-n,T}) \leq (|\mathcal{A}_n| + 1)\gamma_T.$$

Proof. By definition of the gaps,

$$u_n(p_{n,T+1}, \bar{\mathbf{p}}_{-n,T}) = \max_{a'} u_n(a', \bar{\mathbf{p}}_{-n,T}) - \sum_a p_{n,T+1}(a) \delta_{n,T}(a),$$

so it suffices to bound the sum. Actions with $\delta_{n,T}(a) \leq \gamma_T$ contribute at most γ_T in total, and each of the others has $p_{n,T+1}(a) \leq \gamma_T$ and $\delta_{n,T}(a) \leq 1$. $\square$

We next identify the joint distributions to which empirical strategies can converge under either class.

Proposition 5. *MBA and WMBA dynamics can converge to the same joint distributions, which are those in $\mathcal{P}^*$. Under MBA_{NR} and WMBA_{NR} dynamics, they are the CCEs in $\mathcal{P}^*$.*

Proof. Necessity. Suppose $\bar{\mathbf{p}}_T \rightarrow \boldsymbol{\mu}$ under WMBA dynamics. Then, for every a ,

$$\delta_{n,T}(a) \rightarrow \max_{a'} u_n(a', \boldsymbol{\mu}_{-n}) - u_n(a, \boldsymbol{\mu}_{-n}).$$

If $\mu_n(a) > 0$, then $\bar{p}_{n,T}(a) \not\rightarrow 0$, so by (1) this limit is zero, that is, $a \in \text{BR}_n(\boldsymbol{\mu}_{-n})$. Hence $\boldsymbol{\mu} \in \mathcal{P}^*$. Under no regret, $\text{Reg}_n(\boldsymbol{\mu}) = \lim_T \text{Reg}_n(\bar{\mathbf{p}}_T) \leq 0$ by (2), so $\boldsymbol{\mu}$ is a CCE. Since MBAs are WMBAs, the same holds for them.

Sufficiency. Let $\boldsymbol{\mu} \in \mathcal{P}^*$, and let $\boldsymbol{\sigma}$ track $\boldsymbol{\mu}$ (Lemma 4), so that $\|\bar{\mathbf{p}}_T - \boldsymbol{\mu}\| < |\mathcal{A}|^2/T$. Every scheduled action of player n lies in the support of μ_n , hence in $\text{BR}_n(\boldsymbol{\mu}_{-n})$, so its gap along $\boldsymbol{\sigma}$ is below $2|\mathcal{A}|^2/T$. By Lemma 6(i) the algorithms are MBAs, and in self-play $\bar{\mathbf{p}}_T \rightarrow \boldsymbol{\mu}$. If $\boldsymbol{\mu}$ is also a CCE, then along $\boldsymbol{\sigma}$

$$\text{Reg}_{n,T} = T \text{Reg}_n(\bar{\mathbf{p}}_T) \leq T \text{Reg}_n(\boldsymbol{\mu}) + 2|\mathcal{A}|^2 \leq 2|\mathcal{A}|^2,$$

so by Lemma 6(iii) the algorithms also have no regret. $\square$

Along this schedule, we have $\text{Wel}(\bar{\mathbf{p}}_T) \rightarrow \text{Wel}(\boldsymbol{\mu})$, so we attain the MBA extrema of Section C.4 with convergent welfare. Membership in $\mathcal{P}^*$ is also easy to test, since we only compute the marginals of $\boldsymbol{\mu}$, check that every action in the support of μ_n best responds to $\boldsymbol{\mu}_{-n}$, and add (9) under no regret. This requires $O(N|\mathcal{A}|)$ arithmetic operations.

C.3 WMBA Extrema

We now allow empirical strategies not to converge and prove the exact characterization in Proposition 6. Its necessity part rests on the observation at the start of Section C: a persistent action must be a best response along some subsequence of rounds. Its sufficiency part uses the schedule of Lemma 5, which visits several target distributions in increasingly longer blocks.

Proposition 6. *Let $\mathcal{C}$ be WMBA or WMBA_{NR} , with $\mathcal{D}$ defined accordingly:*

$$\begin{aligned} \underline{\text{Wel}}(\mathcal{C}) &= \min_{\mathbf{B}} \min_{\boldsymbol{\mu} \in \mathcal{D}(\mathbf{B})} \text{Wel}(\boldsymbol{\mu}), \\ \overline{\text{Wel}}(\mathcal{C}) &= \max_{\mathbf{B}} \min_{n, a \in B_n} \max_{\boldsymbol{\mu} \in Q_a(\mathbf{B})} \text{Wel}(\boldsymbol{\mu}), \end{aligned} \tag{12}$$

where $\mathbf{B}$ ranges over admissible products, and both values are attained.

Proof. Necessity. Fix a profile in $\mathcal{C}$, and let B_n be player n 's set of persistent actions, those with $\bar{p}_{n,T}(a) \not\rightarrow 0$. Every subsequential limit of $\bar{\mathbf{p}}_T$ is supported on $\mathbf{B}$ and, under no regret, is a CCE, so it lies in $\mathcal{D}(\mathbf{B})$; hence $\liminf_T \text{Wel}(\bar{\mathbf{p}}_T) \geq \min_{\mathcal{D}(\mathbf{B})} \text{Wel}$. For $a \in B_n$, (1) forces $\liminf_T \delta_{n,T}(a) = 0$, so along some subsequence $\bar{\mathbf{p}}_T \rightarrow \boldsymbol{\mu}^{(a)}$ with $a \in \text{BR}_n(\boldsymbol{\mu}_{-n}^{(a)})$, that is, $\boldsymbol{\mu}^{(a)} \in Q_a(\mathbf{B})$. Thus $\mathbf{B}$ is admissible, and

$$\liminf_T \text{Wel}(\bar{\mathbf{p}}_T) \leq \text{Wel}(\boldsymbol{\mu}^{(a)}) \leq \max_{Q_a(\mathbf{B})} \text{Wel} \quad \text{for every } n \text{ and } a \in B_n.$$

Sufficiency. Let $\mathbf{B}$ be admissible. For the lower value, take a minimizer $\boldsymbol{\mu}^{(0)}$ of Wel over $\mathcal{D}(\mathbf{B})$ and any $\boldsymbol{\mu}^{(a)} \in Q_a(\mathbf{B})$ for each n and $a \in B_n$; for the upper value, take maximizers $\boldsymbol{\mu}^{(a)}$ of Wel over $Q_a(\mathbf{B})$. Using the block schedule over these targets (Lemma 5), we play only profiles in $\mathbf{B}$

and obtain, for every action in $\mathbf{B}$, a subsequence along which its gap vanishes. Thus we satisfy (1) for every player, and with no regret, each target is a CCE, so we also satisfy (2). Its lim inf welfare is the smallest target welfare: $\text{Wel}(\boldsymbol{\mu}^{(0)})$ in the first case, since every $\boldsymbol{\mu}^{(a)}$ lies in $\mathcal{D}(\mathbf{B})$, and $\min_a \text{Wel}(\boldsymbol{\mu}^{(a)})$ in the second. Lemma 6(ii) and (iii) turn the schedule into a profile in $\mathcal{C}$. $\square$

Corollary 2. $\underline{\text{Wel}}(\text{WMBA}) = \min_{\mathbf{a} \in \tilde{\mathcal{A}}} \text{Wel}(\mathbf{a})$, that is, without no regret, WMBA's guarantee only the welfare of the worst rationalizable profile.

Proof. Without no regret, $\mathcal{D}(\mathbf{B}) = \Delta(\mathbf{B})$, and $Q_a(\mathbf{B}) \neq \emptyset$ exactly when $a \in \text{BR}_n(\mathbf{q})$ for some $\mathbf{q} \in \Delta(\mathbf{B}_{-n})$, since we can obtain such a joint distribution by fixing player n 's action at a and drawing the opponents' actions from $\mathbf{q}$. Admissibility therefore means that every action best responds to some belief on the product. Every admissible product lies in the CRS, which is itself admissible (Section B.1). The lower value in (12) is therefore the minimum of the linear function Wel over $\Delta(\tilde{\mathcal{A}})$, which is attained at a pure profile. $\square$

Convergent best extrema. To determine whether the attained best extremum also has convergent welfare, we test whether the product $\mathbf{B}$ attaining $\overline{\text{Wel}}(\mathcal{C}) = h$ admits

$$\boldsymbol{\mu}^{(a)} \in Q_a(\mathbf{B}) \quad \text{with} \quad \text{Wel}(\boldsymbol{\mu}^{(a)}) = h \quad \text{for every } n \text{ and } a \in B_n. \quad (22)$$

When this test succeeds, the block schedule has $\text{Wel}(\bar{\mathbf{p}}_T) \rightarrow h$ by Lemma 5. We verify (22) for every best extremum in Tables 2 and 4.

C.4 MBA Extrema and Exploration

Throughout this subsection, we consider two players. We have no general formula for MBA extrema comparable to (12). Each empirical marginal satisfies $\bar{p}_{n,t+1} = \bar{p}_{n,t} + (p_{n,t+1} - \bar{p}_{n,t})/(t+1)$, so on the time scale $\log t$ the pair of marginals moves toward the current strategies. Under WMBA, we can construct paths between any joint distributions meeting the conditions of Proposition 6. Under MBA, we instead use Proposition 4 to identify the pair of empirical marginals as a generalized weakened FP process. Its limit sets are internally chain-transitive sets of the corresponding dynamics [4]. Under the convergence conditions used below, these sets lie in the set of Nash equilibria. Otherwise, as in Shapley's game [46], we may obtain cycles. The joint empirical distribution additionally records correlation across rounds, which we retain when computing welfare.

Even when the empirical strategies converge, a second difficulty remains. Every persistent action must be a best response to the same limit, a fixed-point condition that couples the actions, so the extrema range over Nash couplings (Lemma 7), and related problems, such as maximizing welfare over Nash equilibria, are NP-hard [29]. In Proposition 6, by contrast, each action is justified separately. We therefore compute the MBA extrema only under the hypothesis of Lemma 8, which holds in all our examples and removes both difficulties: the empirical marginals converge to the set of Nash equilibria, and the equilibria have supports of size at most two.

Lemma 7. *If the empirical marginals under MBA dynamics always approach the set of Nash equilibria, then $\underline{\text{Wel}}(\text{MBA})$ and $\overline{\text{Wel}}(\text{MBA})$ are the minimum and maximum of Wel over Nash couplings, and those of MBA_{NR} are the minimum and maximum over the CCEs among them.*

Proof. Every subsequential limit of $\bar{\mathbf{p}}_T$ is a Nash coupling and, under no regret, a CCE, so the class extrema lie between the stated optima, which exist by compactness. Each optimum lies in $\mathcal{P}^*$, so MBA dynamics can converge to it by Proposition 5, with convergent welfare (Section C.2). $\square$

For each Nash equilibrium (x, y) , these optima are linear programs over the couplings of x and y ,

$$\mu \in \Delta(\mathcal{A}), \quad \mu_1 = x, \quad \mu_2 = y,$$

with (9) added under no regret. With finitely many equilibria, the extrema are the extreme values over all of them. We optimize over couplings rather than the product $x \otimes y$ because empirical strategies can be correlated across rounds. The same reasoning shows why WMBAs dynamics can leave the MBA range only if their empirical marginals do not converge. If they converge, the lim inf of welfare is attained by a subsequential limit that is a Nash coupling and, under no regret, a CCE, so an MBA profile attains that welfare by [Proposition 5](#).

Lemma 8. *If the rationalizable game is a nondegenerate $2 \times n$ game [5], then under the dynamics of MBAs with exploration, and in particular under MBA dynamics, the empirical marginals approach the set of Nash equilibria.*

Proof. We show that the pair of empirical marginals, conditioned on the CRS, is a generalized weakened FP process. We write $\hat{p}_{n,t}$ for player n 's base MBA proposal and $\mathbf{e}_t \doteq (p_{1,t} - \hat{p}_{1,t}, p_{2,t} - \hat{p}_{2,t})$, so $\sum_{t \leq T} \|\mathbf{e}_t\| = o(T)$, with $\mathbf{e}_t = 0$ for MBAs. The learning algorithms are WMBAs ([Theorem 2](#)), so by [Section B.2](#) every action outside $\tilde{\mathcal{A}}_n$ is highly sub-optimal and $\bar{p}_{n,t}(\tilde{\mathcal{A}}_n) \rightarrow 1$. Since the base learning algorithm is an MBA, also $\hat{p}_{n,t+1}(\mathcal{A}_n \setminus \tilde{\mathcal{A}}_n) \leq |\mathcal{A}_n| \gamma_t$ eventually. We condition $\bar{p}_{n,t}$ and $\hat{p}_{n,t+1}$ on $\tilde{\mathcal{A}}_n$ to obtain $\pi_{n,t}$ and $\hat{\rho}_{n,t+1}$, and write $\boldsymbol{\pi}_t \doteq (\pi_{1,t}, \pi_{2,t})$ and $\hat{\boldsymbol{\rho}}_{t+1} \doteq (\hat{\rho}_{1,t+1}, \hat{\rho}_{2,t+1})$. With $\theta_t \doteq 1/(t+1)$, we rewrite the marginal updates $\bar{p}_{n,t+1} = (1 - \theta_t)\bar{p}_{n,t} + \theta_t p_{n,t+1}$ as

$$\boldsymbol{\pi}_{t+1} = (1 - \theta_t) \boldsymbol{\pi}_t + \theta_t (\hat{\boldsymbol{\rho}}_{t+1} + \mathbf{M}_{t+1}), \quad \|\mathbf{M}_{t+1}\| \leq C(\|\mathbf{e}_{t+1}\| + \zeta_t),$$

for a constant C and some $\zeta_t \rightarrow 0$, because the mass outside $\tilde{\mathcal{A}}_n$ vanishes in both $\bar{p}_{n,t}$ and $\hat{p}_{n,t+1}$. Over any window $s \leq t < k$ with $\sum_t \theta_t \leq \tau$, we have $k \leq e^\tau(s+1)$, so

$$\sum_{t=s}^{k-1} \theta_t \|\mathbf{M}_{t+1}\| \leq C \left(\frac{1}{s+1} \sum_{t \leq k} \|\mathbf{e}_t\| + \tau \sup_{t \geq s} \zeta_t \right) \rightarrow 0 \quad \text{as } s \rightarrow \infty,$$

which is the perturbation condition of generalized weakened FP [38]. By [Proposition 4](#), applied to the base learning algorithm, and the Lipschitz property of payoffs, $\hat{\rho}_{n,t+1}$ is a ξ_t -best response to $\pi_{-n,t}$ in the rationalizable game, with $\xi_t \rightarrow 0$. Hence $(\boldsymbol{\pi}_t)$ is a generalized weakened FP process in the rationalizable game, and such processes converge to the set of Nash equilibria of nondegenerate $2 \times n$ games [5, 38]. Finally, the rationalizable and full games have the same Nash equilibria, because equilibria are supported on the CRS and every best response to a belief on the CRS lies in it, and $\sum_{n=1}^2 \|\bar{p}_{n,t} - \pi_{n,t}\| \rightarrow 0$. $\square$

For each example, our code computes the CRS, checks the nondegeneracy conditions of the rationalizable game (every pure strategy has a unique best response, and no mixed strategy of the two-action player has more than two), and enumerates the Nash equilibria by solving the support equations for pure and two-action supports and testing every deviation in the full game.

[Lemma 8](#) also shows that exploration alone does not widen the MBA range.

Proposition 7. *If the rationalizable game is a nondegenerate $2 \times n$ game, MBAs with exploration have the same welfare extrema as MBAs, with/without no regret.*

Proof. By [Lemma 8](#), every subsequential limit of $\bar{\mathbf{p}}_T$ under the dynamics of MBAs with exploration is a Nash coupling and, under no regret, a CCE. The extrema of MBAs with exploration therefore lie between the minimum and maximum of Wel over these distributions, which are the MBA extrema by [Lemma 7](#). Since every MBA is an MBA with exploration, the extrema coincide. $\square$

C.5 Convergent Worst Extremum

We next prove [Proposition 8](#), which characterizes the lowest limit we can attain when welfare converges. We use the quantities $\underline{h}_B$ and $\bar{h}_B$ defined in [Section 4.6](#).

Proposition 8. *For $\mathcal{C} \in \{\text{WMBA}, \text{WMBA}_{\text{NR}}\}$,*

$$\underline{\text{Wel}}_{\mathcal{C}}(\mathcal{C}) = \min\{\underline{h}_B : \mathbf{B} \text{ admissible, } \underline{h}_B \leq \bar{h}_B\}.$$

Proof. Necessity. Suppose $\text{Wel}(\bar{\mathbf{p}}_T) \rightarrow h$ along a profile in $\mathcal{C}$, and let $\mathbf{B}$ be the product of persistent actions, which is admissible by the proof of [Proposition 6](#). For each n and $a \in B_n$, some subsequential limit $\mu^{(a)}$ of $\bar{\mathbf{p}}_T$ lies in $Q_a(\mathbf{B})$, and $\text{Wel}(\mu^{(a)}) = h$ because the welfare converges. Hence

$$\min_{Q_a(\mathbf{B})} \text{Wel} \leq h \leq \max_{Q_a(\mathbf{B})} \text{Wel} \quad \text{for every } n \text{ and } a \in B_n,$$

that is, $\underline{h}_B \leq h \leq \bar{h}_B$.

Sufficiency. Let $\mathbf{B}$ be admissible with $\underline{h}_B \leq \bar{h}_B$. Each $Q_a(\mathbf{B})$ is convex and contains distributions with welfare at most $\underline{h}_B$ and at least $\bar{h}_B \geq \underline{h}_B$, so it contains some $\mu^{(a)}$ with $\text{Wel}(\mu^{(a)}) = \underline{h}_B$. The block schedule over these targets has $\text{Wel}(\bar{\mathbf{p}}_T) \rightarrow \underline{h}_B$ ([Lemma 5](#)) and satisfies (1) and, under no regret, (2), and [Lemma 6](#) turns it into a profile in $\mathcal{C}$. $\square$

In our examples the MBA worst extremum is attained with convergent welfare ([Lemma 7](#)), so only the WMBA values change. We do not have a polynomial-time method for the convergent worst extremum. We compute it by enumerating the products inside the CRS, which contains every admissible product ([Section B.1](#)), and this is feasible because the CRS is small in our examples. Every linear program is certified exactly, and the minimizing product comes with the targets $\mu^{(a)}$ at the common level $\underline{h}_B$.

C.6 Efficient Computation

Having characterized the outcomes, we finish by proving [Proposition 9](#): the feasible sets shrink with the candidate product, allowing us to evaluate (12) without enumerating products.

Proposition 9. *For $\mathcal{C} \in \{\text{WMBA}, \text{WMBA}_{\text{NR}}\}$, $\underline{\text{Wel}}(\mathcal{C})$ and $\overline{\text{Wel}}(\mathcal{C})$ can be computed exactly in time polynomial in the bit length of the payoff table.*

Proof. For $\mathbf{B}' \subseteq \mathbf{B}$, extending distributions by zero yields $\mathcal{D}(\mathbf{B}') \subseteq \mathcal{D}(\mathbf{B})$ and $Q_a(\mathbf{B}') \subseteq Q_a(\mathbf{B})$, since the deviation constraints do not depend on the support. We obtain a polytope with at most $|\mathcal{A}|$ variables and $O(K)$ payoff inequalities, where $K \doteq \sum_n |\mathcal{A}_n|$, together with the probability constraints. We test feasibility and optimize Wel by linear programming.

Lower value. Starting from $\mathbf{B} = \mathcal{A}$, repeatedly delete every action a with $Q_a(\mathbf{B}) = \emptyset$. By monotonicity, a deleted action belongs to no admissible product, so every admissible product survives, and the terminal product $\mathbf{B}_\star$ is admissible and contains them all. Hence $\underline{\text{Wel}}(\mathcal{C}) = \min_{\mathcal{D}(\mathbf{B}_\star)} \text{Wel}$.

Upper value. Starting from $\mathbf{B} = \mathbf{B}_\star$, compute $v(a) \doteq \max_{Q_a(\mathbf{B})} \text{Wel}$ for every n and $a \in B_n$, deleting actions with empty $Q_a(\mathbf{B})$ as before. When none is empty, $\mathbf{B}$ is admissible with value $h \doteq \min_a v(a)$, and we record h and delete every action with $v(a) = h$. By monotonicity, an admissible product containing a deleted action has value at most h , so every admissible product with a larger value survives the pass. Stop when a player's set is empty. An admissible product with value above every record would have survived every pass, which is impossible, so the largest record is $\overline{\text{Wel}}(\mathcal{C})$.

Each pass deletes an action or stops, so at most K passes solve at most K linear programs each. Their coefficients have bit length polynomial in the input, so they can be solved and compared exactly in polynomial time, as can the normalization in (10). Nothing in the argument uses more about Wel than linearity. The bound is polynomial in the size of the explicit payoff table, which grows exponentially with the number of players. $\square$

D Instances and Certificates

Every class extremum in a table comes from applying [Proposition 6](#) and [Lemmas 7](#) and [8](#) to an explicit game, with every linear program certified exactly, and the entries for BFP come from the certified cycle of [Section D.2](#) ([Section D.3](#)).

D.1 The Auctions

We begin with the auction model underlying [Theorem 4](#). Two bidders compete for M identical units in a pay-as-bid (PAB) auction. Bidder n 's valuation for its j th unit is $v_{n,j}$, nonincreasing in j , and it submits a nonincreasing vector of integer bids with $0 \leq b_{n,j} < v_{n,j}$. The M highest bids win, with any remaining units going to a uniformly random subset of the bids tied at the cutoff, zero bids included, and each bidder pays its winning bids. A bidder that wins w_n units and pays P_n has utility $\sum_{j \leq w_n} v_{n,j} - P_n$. Relative to the maximum total value OPT of an allocation, we measure the value created (welfare) and the seller's revenue,

$$\begin{aligned} \text{Wel}(\mu) &\doteq \mathbb{E}_\mu \left[\sum_n \sum_{j \leq w_n} v_{n,j} \right] / \text{OPT}, \\ \text{Rev}(\mu) &\doteq \mathbb{E}_\mu [P_1 + P_2] / \text{OPT}, \end{aligned} \tag{23}$$

as well as the bidders' surplus, their total utility, which equals $\text{Wel} - \text{Rev}$. Class extrema are defined as in (11).

We prove [Theorem 4](#) with five auctions. In [Examples 1](#) and [2](#), we separate the worst extrema of the four classes, and in [Examples 3](#) to [5](#), we show that WMBA raises the best extrema ([Table 2](#)).

Example 1. $M = 5$ units, with valuations

$$\begin{aligned} v_1 &= (3.25, 2.5, 1.75, 1.75, 1.5), \\ v_2 &= (7.5, 6.75, 2.25, 1, 0.75). \end{aligned}$$

Example 2. $M = 4$ units, with valuations

$$\begin{aligned} v_1 &= (4.5, 3.75, 2, 2), \\ v_2 &= (5.9, 4, 3, 1). \end{aligned}$$

Example 3. $M = 3$ units, with valuations

$$\begin{aligned} v_1 &= (3.6, 3.6, 1), \\ v_2 &= (2.01, 1.65, 1.5). \end{aligned}$$

Example 4. $M = 2$ units, with valuations

$$\begin{aligned} v_1 &= (2.1, 2), \\ v_2 &= (9, 0.5). \end{aligned}$$

Example 5. $M = 5$ units, with valuations

$$\begin{aligned} v_1 &= (2.16, 2.16, 2.16, 2.16, 2.16), \\ v_2 &= (3.26, 3.26, 3.26, 3.26, 1). \end{aligned}$$

We now discuss what [Theorem 4](#) means for each outcome.

Allocation efficiency. In [Example 1](#), we see that the welfare guarantee falls from 0.946 under MBA to 0.483 under WMBA. By [Corollary 2](#), 0.483 is the welfare of the worst rationalizable bid profile, in which bidder 2, who values the first units most, bids zero and wins nothing. That value is reached only while welfare oscillates. If average welfare must converge, the WMBA guarantee is instead 0.941, barely below MBA’s ([Proposition 8](#)). With no regret, the WMBA guarantee rises to 0.914, still short of 0.953 under MBA. In [Example 3](#), WMBA instead raises the best extremum by 16%, from 0.837 to 0.973.

Seller revenue. Turning to the seller, in [Example 2](#) we see that WMBA lowers the revenue guarantee from 0.408 to 0.165, and from 0.409 to 0.272 with no regret. Part of this loss survives when revenue must converge, as the guarantees become 0.325 and 0.382, about 20% and 7% below MBA’s. In [Example 4](#), WMBA instead raises the best extremum by 42%, from 0.133 to 0.189.

Bidders’ surplus. For bidders, in [Example 5](#) we obtain a 20% increase in the best extremum of their surplus under WMBA, from 0.450 to 0.543, partly through lower payments, that is, at the seller’s expense.

Coordination and regret. In [Examples 3 to 5](#), the gains vanish under no regret: with it, each objective takes a single value, the same in both classes. Exceeding that value therefore requires some bidder to have positive average regret along a subsequence of rounds, and our constructions attain the gains by coordinating the bidders on a common schedule ([Section C.1](#)). No regret does not remove the losses in the same way, since in [Examples 1 and 2](#), WMBA_{NR} still has lower worst extrema than MBA.

Theorem 4. *For welfare/revenue, there is a PAB instance in which the four classes have pairwise distinct worst extrema. For each of welfare/revenue/bidders’ surplus, there is a PAB instance in which the WMBA best extremum exceeds its MBA counterpart.*

Proof. For each of [Examples 1 to 5](#), we use the code to enumerate every feasible bid vector, average allocations and payments over tie-breaking, and normalize utilities by OPT so that they lie in $[0, 1]$. We then apply the same programs to each objective in [\(23\)](#), keeping deviations to every original bid, verify the hypotheses of [Lemma 8](#), and check [\(22\)](#) for every best extremum. This yields every entry of [Table 2](#), and each value is attained. In each column of worst extrema, the four values are pairwise distinct. In each column of best extrema, the WMBA value exceeds the other three, which agree after rounding, and the two no-regret values agree exactly. $\square$

D.2 Price Competition

We next specify the pricing game used to compare the class extrema with BFP. Two firms with zero costs sell a good that every consumer values at 1. Each loyal consumer buys from its own firm at any price up to 1, and each shopper buys from the cheaper firm, with shoppers split evenly when the prices tie. Since every consumer buys, total surplus equals the number of consumers, and we measure industry profit and consumer surplus as fractions of it, so the two sum to 1 in every round.

Example 6. One consumer is loyal to firm 1, five are loyal to firm 2, and nine are shoppers. Each firm charges 0.5, 0.9, or 1.

To compute the class extrema, we first observe that firm 1’s price 1 is never a best response, although no single price dominates it, so the CRS is $\{0.5, 0.9\} \times \{0.5, 0.9, 1\}$. This rationalizable game is a nondegenerate 2×3 game, so [Lemma 8](#) applies. Its unique Nash equilibrium is mixed: firm 1 charges 0.5 with probability $8/9$ and 0.9 otherwise, and firm 2 charges 0.5 with probability $80/117$ and 1 otherwise, the price dispersion of Varian’s model. Since firm 2 never charges 0.9 in this equilibrium, MBA dynamics play the pair $(0.9, 0.9)$ with vanishing frequency, by [Lemma 8](#). [Table 4](#) reports the class extrema, computed as for the auctions.

Table 4: Class extrema in price competition ([Example 6](#)), from the worst to the best extremum. A single value means that both extrema coincide.

Algorithm class $\mathcal{C}$	Profit	Consumer surplus
MBA	[0.556, 0.582]	[0.418, 0.444]
WMBA	[0.500, 0.696]	[0.067, 0.447]
MBA _{NR}	0.564	0.436
WMBA _{NR}	[0.525, 0.651]	[0.140, 0.436]

Profit. For industry profit, we obtain a WMBA best extremum of 0.696, 20% above MBA’s best extremum of 0.582. With no regret, we can sustain 0.651 under WMBA, still 15% above MBA’s 0.564. Unlike in the auctions, the gain survives no regret.

Consumer surplus. For consumers, we obtain a lower guarantee, as consumer surplus falls from 0.418 under MBA to 0.067 under WMBA. By [Corollary 2](#), 0.067 is the consumer surplus of the worst rationalizable price pair, in which firm 1 charges 0.9 and firm 2 charges 1. With no regret, the guarantee falls from 0.436 to 0.140. Part of the loss survives when consumer surplus must converge, as the WMBA guarantee becomes 0.304, 27% below MBA’s. The worst extremum of consumer surplus is not one minus the best extremum of profit, since both are $\liminf$ guarantees and WMBA strategies can return to high prices infinitely often without staying there.

BFP. To obtain an outcome beyond the MBA range with a simple learning algorithm, we let each firm play an arbitrary price in round 1. After rounds 1, 3, 9, 27, $\dots$, both firms switch to a best response to the rival’s empirical distribution of prices so far, breaking ties toward the lower price, and keep that price until their next update. BFP is a WMBA, since a price that trails some alternative by a fixed margin from some time on is not a best response at any later update, so it is never chosen again. The same argument shows that any MBA updated only at geometrically spaced times is a WMBA. BFP is not an MBA, however, since a held price can fall behind within a batch. Indeed, on the cycle below, some firm keeps a price that trails another by a fixed amount at the end of every period. From each of the nine starting price pairs, and with ties broken either way, the firms’ prices enter the same cycle of eight batches and stay on it: two batches at $(0.5, 0.9)$, then $(0.5, 1)$, $(0.9, 1)$, and four batches at $(0.9, 0.9)$. Within a batch, long-run averages move monotonically, so their extremes occur at the ends of batches. Long-run average profit stays between 0.663 and 0.899,

so even its lowest value is 14% above MBA’s best extremum of 0.582, and consumer surplus stays between 0.101 and 0.337, below MBA’s worst extremum of 0.418.

We also verify its no-regret classification in [Table 1](#). Since BFP chooses pure actions, we can assign reward zero to its chosen action and one to every other action on each round. BFP then earns zero, while some fixed action earns at least $T/2$ when there are at least two actions. Thus it does not guarantee no regret on every reward history.

D.3 Exact Verification

To verify the computations behind these instances, we obtain candidate solutions by floating-point optimization and check every claim in exact arithmetic. For each optimum, the certificate is a feasible solution and a feasible dual solution with equal objective values. For each empty $Q_a(\mathbf{B})$, it is a nonnegative combination of the program’s homogeneous inequalities that is strictly positive in every coordinate, which no probability vector can satisfy. A separate verifier rebuilds the payoff tables and programs from the game definitions and replays all certificates, deletion steps, and equilibrium supports without an optimization solver, and checks that the targets in [\(22\)](#) share one objective value. The reported extrema are never supplied to the computation, and every strict comparison uses exact values, not their roundings.

The convergent worst extrema of [Proposition 8](#) come from a second script, which enumerates the products inside the CRS, certifies each of its linear programs in the same way, and records, for the minimizing product, targets as in [\(22\)](#) that share one objective value. As a consistency check, it confirms that every convergent worst extremum lies between the worst extremum of its class and that of the corresponding MBA class.

For BFP, we use a third script that follows its updates exactly from every starting price pair. It computes the cycle’s periodic orbit in rational arithmetic and checks that every step is a strict best response. Once the empirical frequencies are within the distance that these margins allow, each batch shrinks the distance by a factor of three, so the strategies stay on the cycle, and the script certifies that the empirical frequencies reach this distance from every start.