BEST RESPONSE CONVERGENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Does convergence of one's mixed strategy in multi-agent learning dynamics guarantee convergence of best response? The answer is no: already in a simple matching pennies game, the best responses cycle whilst the strategies converge to a mixed Nash equilibrium. However, when both agents converge to a strict *pure* equilibrium, their best responses also converge. In this paper, we show that the best responses of agents converging to a pure strategy *need not* converge. In particular, we use physics-inspired techniques to show necessary and sufficient conditions for when best responses converge in FTRL dynamics. To do this, we analyze learning near *boundary equilibria*, a class of non-strict Nash equilibria where one player's limiting action yields the same payoff as several other actions, which we call *alternatives*. As a function of the algorithms' learning rates, we identify sharp best-response thresholds for Hedge. For example, with symmetric learning rates $t^{-\alpha}$, the threshold under Hedge is $\alpha = 3/7$. A further consequence of our analysis is that larger learning rates can make the gains from deviation vanish more slowly, despite faster concentration on the limiting action. We illustrate our results through simulating learning dynamics in inspection games.

1 INTRODUCTION

In a game played repeatedly, each player's payoffs depend on its own actions and those of the other players. As the players learn from past interactions, their strategies evolve together. If these strategies converge, must the players' best responses converge as well? We study this question for follow-the-regularized-leader (FTRL), a family of no-regret algorithms that chooses action probabilities using past payoffs and a regularization penalty (Section 3). An action can retain a high probability because of its past payoffs even when another action currently earns more. We call an action that maximizes expected payoff against the opponents' current mixed strategies a *current best response*. We ask whether this best response eventually stops changing when the mixed strategies under FTRL converge.

Convergence in strategies vs. in payoffs. At a Nash equilibrium (NE), no player can gain by changing their own strategy. At a strict pure Nash equilibrium (PNE), each player's chosen action is uniquely optimal, so sufficiently close play has the same best response. By contrast, near a mixed Nash equilibrium (MNE) around which play rotates, payoff differences can change sign repeatedly as play approaches the equilibrium. We focus on boundary equilibria, where one player converges to a pure action, with one or more other actions yielding the same payoff at the limit. We call these other actions *alternatives*. Here best responses can either stabilize or keep changing, depending on how play approaches the equilibrium. Our k-boundary model has k alternatives (Definition 2). The following examples illustrate how changes in play can restore the incentive to choose the equilibrium action:

- **Inspection and compliance.** A firm can comply or violate a rule in one of several activities, while an inspector can inspect at most one activity at no additional cost. A violation can save the firm the cost of complying, but detection exposes it to a penalty. Under compliance, all inspection choices yield the same payoff. As the algorithm gradually weakens the inspector's past preference for particular inspections, some activities may be inspected too rarely, making violations profitable. Violations then make inspection of those activities more profitable and restore the incentive to comply.
- **First-price auctions.** Boundary equilibria in auctions and pricing games are especially natural applications. Here, one agent captures the entire market, leaving competitors indifferent among

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

actions that still lose and earn nothing. Their high bids or low prices can nevertheless keep the winner from bidding less or charging more, helping sustain the equilibrium [37]. If their learned preferences for more competitive bids weaken, the winner may profit by lowering its bid. This can let a competitor win at a profit by bidding more aggressively. Renewed competition then restores the original bidder's incentive to bid higher.

- **Investment and repayment.** An investor can withhold funding or finance a productive project whose recipient chooses how much to repay. Without investment, every repayment policy gives the recipient the same payoff. As a learned preference for retaining the proceeds weakens, the prospect of more generous repayment can make investment attractive again. Investment then restores the incentive to retain the proceeds and can make withholding funds optimal again.

We illustrate the issue with a 1-boundary game, which has two actions per player. Each entry lists the payoffs of players A and B, in that order:

	B_0	B_1
A_0	$(1, 0)$	$(0, 0)$
A_1	$(0, 1)$	$(2, 0)$.

(1)

We write $q_t \doteq \Pr_t(B_1)$ for the probability that B chooses B_1 at time t, and $u_A(a, q_t)$ for A's expected payoff from action a. Since $u_A(A_0, q_t) - u_A(A_1, q_t) = 1 - 3q_t$, A is indifferent at $q_t = 1/3$. When player A assigns no probability to A_1, player B is indifferent between B_0 and B_1. The game therefore has a continuum of NE with $\Pr(A_0) = 1$ and $\Pr(B_1) \leq 1/3$. At the endpoint $q = 1/3$, A_1 yields the same payoff as the equilibrium action A_0. If q_t approaches $1/3$ from below, A_0 remains best. If q_t instead crosses $1/3$ infinitely often, the best response keeps changing. Both behaviors are compatible with player A's strategy converging to A_0. We ask how the learning algorithm determines which behavior occurs.

Contributions. Player n uses FTRL with learning rate $\eta_t^n \doteq t^{-\alpha_n}$, $0 < \alpha_n < 1$. We call t^{α_n} the *inverse learning rate*, or *inverse rate* for short; it weights the penalty in the FTRL objective (4).

A threshold for best-response stability. For Hedge, the entropy-regularized form of FTRL, we identify the sharp threshold $\alpha = 3/7$ for common learning rates $t^{-\alpha}$ (Theorem 3). For any fixed number of alternatives, every finite initialization converges to the boundary strategy profile. Above the threshold, the equilibrium action eventually remains uniquely best; below it, the best response switches infinitely often for almost every initialization. The threshold reflects two vanishing effects: the *drift* along a *reference path*—specified trajectories for $s_i(t)$ and $r_i(t)$, defined in Appendix D—makes the equilibrium action better, while *oscillations* around that path can make an alternative better (Figure 1). We establish the threshold without assuming approach to the reference (Lemma 1). Surprisingly, the same threshold governs any fixed number of interacting alternatives. The familiar square-root schedule $t^{-1/2}$ [27] lies on the stable side.

Larger learning rates can slow payoff convergence. As a secondary result, we quantify the weighted payoff gap $\mathrm{Gap}(t) = g_A(t)/3 + g_B(t)$ in the 1-boundary game, where $g_n(t)$ is the gain from switching to a current best response, defined in Equation 3. For every $0 < \alpha < 3/7$, we obtain sharp rates for the largest recurring values of $\mathrm{Gap}(t)$: these values decay more slowly as α decreases. Above $3/7$, the limiting action remains best, and the remaining probability on the alternative determines $\mathrm{Gap}(t)$. Thus a larger learning rate can speed up concentration on the limiting action while slowing payoff convergence (Proposition 3). We obtain the same decay rates for the average $T^{-1} \int_1^T \mathrm{Gap}(t)\,\mathrm{d}t$ (Proposition 5). We also relate action probabilities to the shortfall $\sigma_t^n(a)$ in Equation 2 (Proposition 1) and derive matching upper and lower convergence rates in the 1-boundary model (Theorems 1 and 2). The rate is determined either by the opponent's inverse rate or by the learner's response to $\eta_t^n \sigma_t^n(a)$, depending on how quickly the learner reduces the probability of an action with a persistent disadvantage. We also derive a threshold for FTRL with Tsallis entropy and a stability condition for FTRL with a quadratic penalty, called *Euclidean FTRL* (Propositions 7 and 6; Table 1). The regularizers are defined in Appendices D and A, respectively.

2 Related Work

Best response and repeated play. Classical models of learning in games differ in the information to which players respond. Fictitious play uses empirical beliefs formed from past play, whereas best-response dynamics respond to current strategies [3, 6, 19, 23, 34, 40, 43, 46]. Replicator and best-reply dynamics provide other descriptions of adjustment to payoff incentives [11, 20–22, 41]. Empirical beliefs may converge to a mixed equilibrium while period-by-period strategies do not [25]. Modified variants of fictitious play make period-by-period play converge as well [45].

FTRL and last-iterate convergence. FTRL uses the accumulated payoffs $U_t^n(a)$ from Equation 2, with a regularizer controlling how strongly $U_t^n(a) - U_t^n(a')$ affects action probabilities. It belongs to the broader literature on no-regret learning [9, 29, 39, 42]. In games, time averages can converge even when the current mixed strategies do not, and regularized learning can exhibit cycling or instability near non-strict and mixed equilibria [10, 13, 31, 47]. Stability and attraction of equilibria under regularized learning are studied in [14, 16]. We focus on trajectories that converge to an equilibrium where other actions yield the same payoff as the limiting action. Last-iterate convergence concerns the current mixed strategy rather than its time average. Other algorithms respond to current or predicted incentives through gradient information, extrapolation, or momentum [12, 26, 32, 33, 35, 36, 38]. Under monotonicity or related stability conditions, optimistic gradient and optimistic mirror descent admit tight last-iterate rates [2, 4, 7, 17, 24]. Optimism, mutation, and changes in the treatment of past payoffs can also improve convergence of FTRL in some settings [1, 8, 15, 50]; extrapolated variants can behave differently from ordinary FTRL in broad classes of finite games [28]. We distinguish actions maximizing $U_t^n(a)$ from those that are currently best along the same FTRL trajectory.

Dynamical-systems methods. Convex duality and energy estimates are used to study regularized learning in games [30]. In optimization, differential-equation, Lyapunov, and variational analyses explain how damping and oscillation affect accelerated methods [44, 48, 49]. Approximating trajectories over bounded time intervals is also central to the method based on ordinary differential equations (ODEs) for stochastic approximation [5]. In zero-sum games, volume preservation yields recurrence results for regularized learning [31]. We use related tools to separate the movement toward a boundary equilibrium from oscillations around the reference path and determine whether the best action eventually stays fixed. After rescaling the deviations from the reference path, our volume argument rules out their convergence to zero for almost every initialization (Section 3; Appendix D).

3 Model and FTRL Response Estimates

Player n has finite action set $\mathcal{A}_n$ and mixed action $\mathbf{p}_t^n \in \Delta(\mathcal{A}_n)$, where $\Delta(\mathcal{A}_n)$ is the set of probability distributions on $\mathcal{A}_n$. We write $\mathbf{p}_t^{-n}$ for the opponents' mixed actions. For a pure action a, $u_n(a, \mathbf{p}_t^{-n})$ denotes player n's expected one-period payoff against the opponents' independently randomized mixed actions. We define

$$U_t^n(a) \doteq U_1^n(a) + \int_1^t u_n(a, \mathbf{p}_{t'}^{-n})\,\mathrm{d}t', \qquad \sigma_t^n(a) \doteq \max_{a' \in \mathcal{A}_n} U_t^n(a') - U_t^n(a). \tag{2}$$

The finite values $U_1^n(a)$ specify the *initialization*. Thus $U_t^n(a)$ is the payoff action a would have accumulated against the observed path, including its initial value, and $\sigma_t^n(a)$ is the shortfall from the largest accumulated payoff. We use *almost every initialization* with respect to Lebesgue measure on these initial utility vectors.

Both $u_n(a, \mathbf{p}_t^{-n}) - u_n(a', \mathbf{p}_t^{-n})$ and $U_t^n(a) - U_t^n(a')$ can be negative. By contrast, the *deviation gain* is how much a player can earn by replacing its current mixed action with a best response:

$$g_n(t) \doteq \max_{a \in \mathcal{A}_n} u_n(a, \mathbf{p}_t^{-n}) - u_n(\mathbf{p}_t^n, \mathbf{p}_t^{-n}). \tag{3}$$

Here we extend payoffs linearly to mixed actions. The maximum in Equation 3 ensures $g_n(t) \geq 0$.

For nonnegative quantities, we write $x_t \lesssim \widetilde{x}_t$ if $x_t \leq C\widetilde{x}_t$ for all large t and some constant $C < \infty$, and $x_t \asymp \widetilde{x}_t$ if both $x_t \lesssim \widetilde{x}_t$ and $\widetilde{x}_t \lesssim x_t$ hold.

FTRL and the response estimate. We define FTRL as follows: with learning rate $\eta_t^n \doteq t^{-\alpha_n}$ and a regularizer Φ_n—a convex penalty on mixed actions—player n chooses

$$\mathbf{p}_t^n \in \arg\max_{\mathbf{p}\in\Delta(\mathcal{A}_n)} \left\{ \sum_{a\in\mathcal{A}_n} \mathbf{p}(a) U_t^n(a) - (\eta_t^n)^{-1}\Phi_n(\mathbf{p}) \right\}. \tag{4}$$

The inverse rate $t^{\alpha_n} = 1/\eta_t^n$ weights the regularization penalty. For $t > 1$, a smaller α_n gives a larger learning rate and a smaller penalty weight. Multiplying the shortfall in Equation 2 by the learning rate gives $\eta_t^n \sigma_t^n(a)$. We scale $U_t^n(a) - U_t^n(a')$ in the same way. A shortfall of order t^{α_n} therefore gives $\eta_t^n \sigma_t^n(a) \asymp 1$ and can change the mixed strategy by a constant amount.

Definition 1 (Separable FTRL regularizer)**.** A separable FTRL regularizer on a finite action set $\mathcal{A}$ is $\Phi(\mathbf{p}) \doteq \sum_{a\in\mathcal{A}} \phi(\mathbf{p}(a))$, with ϕ differentiable and strictly convex on positive probabilities.

Under the conditions in Appendix A, we define the nonincreasing *response function* $f(\widehat{\sigma}) \doteq (\phi')^{-1}(-\widehat{\sigma})$ for sufficiently large $\widehat{\sigma}$. Here $\widehat{\sigma}$ is a generic value of $\eta_t^n \sigma_t^n(a)$, and f gives the order of the probability assigned to that action.

Proposition 1. *For separable FTRL on a fixed finite action set, suppose $\eta_t \doteq t^{-\alpha}$, $\alpha \in (0,1)$, and the conditions in Appendix A hold. For this response, uniformly over actions and all large t,*

$$\mathbf{p}_t(a) \asymp f(\eta_t \sigma_t(a)), \qquad \lim_{\widehat{\sigma}\to\infty} f(\widehat{\sigma}) = 0. \tag{5}$$

For Hedge, $f(\widehat{\sigma}) \asymp e^{-\widehat{\sigma}}$, so increasing $\eta_t\sigma_t(a)$ reduces the action's probability exponentially. Tsallis FTRL gives a polynomial decrease: $f(\widehat{\sigma}) \asymp \widehat{\sigma}^{-\beta}$, where $\beta > 1$ is the parameter in the regularizer (23). Euclidean FTRL instead assigns probability zero once $\eta_t\sigma_t(a)$ is large enough. Appendix A gives the formulas for Hedge (18) and Euclidean FTRL with two actions (19).

To derive convergence rates near boundary equilibria, we consider both FTRL dynamics satisfying Proposition 1 and other dynamics satisfying Equation 5. We write the nonincreasing response and exponent for player n as f_n and α_n.

For decreasing f_n and small $p > 0$, we define its *generalized inverse* by $f_n^{-1}(p) \doteq \inf\{\widehat{\sigma} \geq 0 : f_n(\widehat{\sigma}) \leq p\}$. To turn probability rates into rates for $\sigma_t^n(a)$, we use the following inverse condition in Theorem 1: $f_n^{-1}(cp_t) \asymp f_n^{-1}(p_t)$ for every fixed $c > 0$ and every sequence $p_t \to 0$.

Boundary games. We write α_A and α_B for the learning-rate exponents of players A and B, respectively. To study whether multiple alternatives change best-response stability, we allow k payoff differences to vanish simultaneously (Theorem 3). We call the following family k-boundary games, where k is the number of alternatives.

Definition 2 (k-boundary game)**.** Fix an integer $k \geq 1$, constants $c_A, c_B > 0$, and $\bar{q} \in (0, 1/(k+1))$. Player A has actions $A_0, A_1, \ldots, A_k$, and player B has actions $B_0, B_1, \ldots, B_k$. Writing $\mathbf{1}\{E\}$ for the indicator of an event E, for $i \in \{1,\ldots,k\}$ and $j \in \{0,\ldots,k\}$ we set

$$u_A(A_0, B_j) \doteq c_A\bar{q}, \quad u_A(A_i, B_j) \doteq c_A\mathbf{1}\{i=j\}, \qquad u_B(B_j, A_0) \doteq 0, u_B(B_j, A_i) \doteq c_B\mathbf{1}\{i\neq j\}.$$

We write $\mathbf{q} \doteq (q_0,\ldots,q_k)$ and $\mathbf{y} \doteq (y_0,\ldots,y_k)$ for the two players' mixed actions, with $q_i \doteq \Pr(B_i)$ and $y_i \doteq \Pr(A_i)$. We obtain

$$u_A(A_0,\mathbf{q}) - u_A(A_i,\mathbf{q}) = c_A(\bar{q} - q_i), \qquad u_B(B_0,\mathbf{y}) - u_B(B_i,\mathbf{y}) = c_B y_i. \tag{6}$$

The NE are the profiles with $\Pr(A_0) = 1$ and $q_i \leq \bar{q}$ for every $i \geq 1$. At the *corner*, $q_1 = \cdots = q_k = \bar{q}$ and $q_0 = 1 - k\bar{q}$, so all k alternatives $A_1,\ldots,A_k$ yield the same payoff as A_0.

We call the $k = 1$ case of Definition 2 a 1-boundary game. Each player has two actions, and we write $y \doteq \Pr(A_1)$ and $q \doteq \Pr(B_1)$. The NE have $y = 0$ and $q \leq \bar{q}$: B is indifferent when $y = 0$, and A is indifferent at $q = \bar{q}$.

Example 1 (1-boundary game)**.** The game in Equation 1 has the same payoff differences as the $k = 1$ case of Definition 2, with $\bar{q} = 1/3$, $c_A = 3$, and $c_B = 1$.

For a 1-boundary game, Y_t is the current probability y_t multiplied by time, while H_t accumulates that probability:

$$y_t \doteq \Pr_t(A_1), \qquad Y_t \doteq t y_t, \qquad H_t \doteq H_1 + \int_1^t y_{t'} \, \mathrm{d}t'. \tag{7}$$

Here $H_1 \doteq [U_1^B(B_0) - U_1^B(B_1)]/c_B$, so $c_B H_t = U_t^B(B_0) - U_t^B(B_1)$. We define

$$x_t \doteq c_A(\bar{q} - q_t), \qquad X_t \doteq X_1 + \int_1^t x_{t'} \, \mathrm{d}t' = U_t^A(A_0) - U_t^A(A_1), \tag{8}$$

where $X_1 \doteq U_1^A(A_0) - U_1^A(A_1)$ is finite. The sign of x_t determines player A's current best response, while X_t determines the FTRL probability of A_1 through Equation 5.

To derive the probability rate in Theorem 1, we assume that player B assigns nearly equal probabilities when $\eta_t^B[U_t^B(B_0) - U_t^B(B_1)] \to 0$. Since $c_B H_t = U_t^B(B_0) - U_t^B(B_1)$, this means

$$H_t = o(t^{\alpha_B}) \quad \Longrightarrow \quad q_t = \frac{1}{2} + o(1). \tag{9}$$

This is natural for an algorithm that treats actions symmetrically. The symmetric, strictly convex FTRL regularizers considered here satisfy this assumption: equal inputs give equal probabilities, and continuity gives the limit when the inputs approach equality.

4 Results

We organize the section around two questions: how play approaches equilibrium, and whether the best response settles along the way. We begin with strict pure, mixed, and boundary equilibria to show why these questions differ. We then study how quickly the probability of a single alternative decreases, before asking whether it can become profitable again even as that probability vanishes. Next, we extend the best-response analysis to several alternatives under Hedge, then return to one alternative to quantify how quickly gains from deviation disappear. We close by examining how the choice of regularizer affects best-response stability.

Three Equilibrium Geometries. We compare three ways that play can approach an equilibrium. At strict PNE, best responses stabilize along convergent paths. Near the matching-pennies MNE, play rotates around the equilibrium while converging to it. Boundary NE can exhibit either behavior.

Unique strict PNE and matching pennies. We first compare a game with a unique strict PNE and matching pennies:

	B_0	B_1
A_0	$(0, 1)$	$(0, 0)$
A_1	$(1, 0)$	$(1, 1)$.

(10)

	B_0	B_1
A_0	$(1, -1)$	$(-1, 1)$
A_1	$(-1, 1)$	$(1, -1)$.

(11)

Equation 10 has a unique strict PNE and eventually fixed best responses. Equation 11 is matching pennies, with incentives that cycle around its unique MNE.

In Equation 10, (A_1, B_1) is the unique strict PNE. Player A always prefers A_1, while player B prefers B_0 after A_0 and B_1 after A_1. Each player's equilibrium action earns one more unit than its alternative, so best responses remain fixed near the equilibrium.

In Equation 11, the unique MNE is $\Pr(A_1) = \Pr(B_1) = 1/2$. The players' incentives induce rotation around this equilibrium.

Proposition 2. *For Hedge in matching pennies* (11)*, with $\eta_t \doteq t^{-\alpha}$ and $\alpha \in (0, 1)$, almost every initialization converges in mixed strategy to $\Pr(A_1) = \Pr(B_1) = 1/2$, while both current best responses switch infinitely often.*

The pair $(s(t), r(t))$—defined in the proof of Proposition 2—approaches zero while continuing to rotate around the origin (Appendix B). Thus the mixed strategies converge, but the best responses keep switching.

Boundary NE. In the 1-boundary game of Example 1, player A converges to a pure action, with its alternative yielding the same payoff at the limit. Unlike the two preceding cases, best-response stability depends on the learning-rate exponent. Figure 1 pairs each game's current payoff directions with its Hedge trajectories for two exponents.

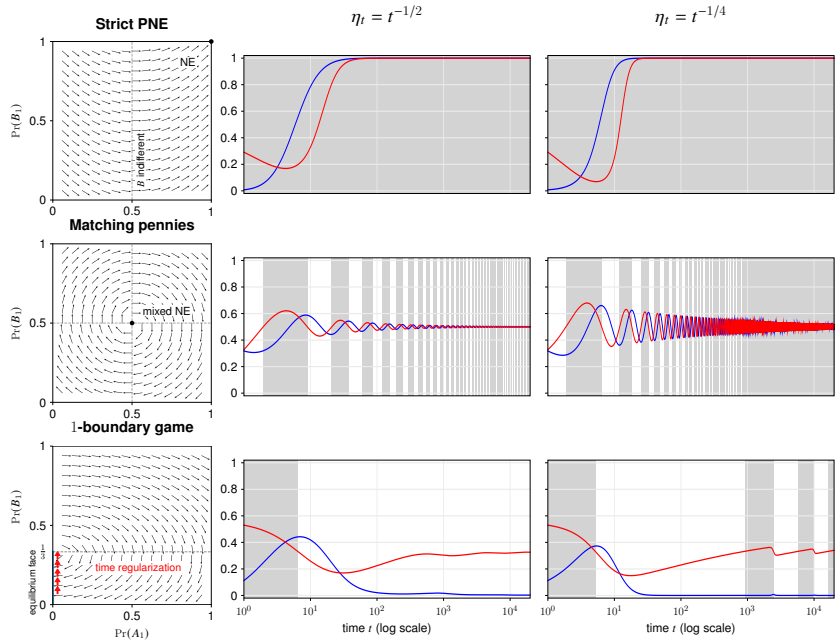


Figure 1: Payoff directions (left) and Hedge trajectories (middle: $\eta_t \doteq t^{-1/2}$; right: $\eta_t \doteq t^{-1/4}$). Rows show the strict-PNE game, matching pennies, and the 1-boundary game. The coordinates in the left-hand plots are $\Pr(A_1)$ and $\Pr(B_1)$. Red arrows along $\Pr(A_1) = 0$ show how the increasing inverse rate changes play. In the trajectories, blue and red curves show $\Pr_t(A_1)$ and $\Pr_t(B_1)$ from random initial cumulative utilities, and solid gray bands indicate when A_1 is best for player A. Time is logarithmic. Only the 1-boundary game changes from eventual best-response stability to persistent switching between these exponents.

Last-Iterate Rates with One Alternative. We now focus on the $k = 1$ case of Definition 2 and ask how quickly player A reduces the probability of A_1. The answer depends on how strongly the algorithm responds to a persistent payoff disadvantage.

For Theorems 1 and 2, we assume Equation 5 and $X_t \geq 0$ eventually, so $X_t = \sigma_t^A(A_1)$ eventually. Thus $y_t \asymp f_A(t^{-\alpha_A} X_t)$ and $Y_t \asymp t f_A(t^{-\alpha_A} X_t)$.

Fast and slow pruning conditions. *Pruning* is the reduction in y_t as X_t grows. If x_t remains bounded below by a positive constant, then $X_t \asymp t$. We call pruning *fast* or *slow* relative to B according to whether the cumulative probability assigned to such an action is smaller or larger in order than B's inverse rate t^{α_B}. The respective conditions are, for every $c > 0$:

$$\int_1^t f_A(ct'^{1-\alpha_A})\,\mathrm{d}t' = o(t^{\alpha_B}), \qquad (12) \qquad\qquad \int_1^t f_A(ct'^{1-\alpha_A})\,\mathrm{d}t' = \omega(t^{\alpha_B}). \qquad (13)$$

Fast pruning holds for exponential responses and responses that vanish past a finite threshold. The *barrier regularizer* $\phi(p) \doteq 1/p$, whose penalty diverges as an action probability approaches zero, gives $f_A(\widehat{\sigma}) \asymp \widehat{\sigma}^{-1/2}$ and satisfies slow pruning for common exponents $0 < \alpha < 1$.

We call Y_t *rate-regular* if exactly one relation holds:

$$Y_t = o(t^{\alpha_B}), \qquad Y_t \asymp t^{\alpha_B}, \qquad Y_t = \omega(t^{\alpha_B}). \qquad (14)$$

Intuitively, the ratio Y_t/t^{α_B} must tend to zero, stay bounded above and away from zero, or grow without bound. This rules out repeated shifts between different orders of growth.

Fast Pruning. Under fast pruning, neither $Y_t = o(t^{\alpha_B})$ nor $Y_t = \omega(t^{\alpha_B})$ can persist. In the first case, player B assigns probability $1/2 + o(1)$ to each action by Equation 9, making A_1 better. In the second, B learns to play B_0, making A_1 worse and forcing its probability below the assumed rate. Rate regularity leaves $Y_t \asymp t^{\alpha_B}$.

Theorem 1. *In a* 1*-boundary game satisfying Equations 5 and 9, eventual* $X_t \geq 0$*, rate regularity* (14)*, and fast pruning* (12)*,*

$$y_t \asymp t^{\alpha_B - 1}, \qquad X_t \asymp t^{\alpha_A} f_A^{-1}(t^{\alpha_B - 1}), \qquad \frac{1}{t}\int_1^t q_{t'}\,\mathrm{d}t' \longrightarrow \bar{q}.$$

Slow Pruning. Under slow pruning, the probability assigned to A_1 remains large enough for player B to learn that B_0 is better, even when $X_t \asymp t$. As B moves toward B_0, X_t continues to grow at the same rate.

Theorem 2. *In a* 1*-boundary game satisfying Equation 5, eventual* $X_t \geq 0$*, slow pruning* (13)*, and* $f_A(c\widehat{\sigma}) \asymp f_A(\widehat{\sigma})$ *for every fixed* $c > 0$*,*

$$y_t \asymp f_A(t^{1-\alpha_A}), \qquad X_t \asymp t, \qquad q_t \longrightarrow 0.$$

Best Responses Along Last Iterates. The rates above describe how play concentrates, not whether the action it favors remains best. We now turn to this second question. In a 1-boundary game, we use $x_t = c_A(\bar{q} - q_t)$ to identify which action is best: A_0 when $x_t > 0$ and A_1 when $x_t < 0$. $X_t = X_1 + \int_1^t x_{t'}\,\mathrm{d}t'$ can grow even while x_t changes sign infinitely often.

Definition 3 (Best-response stability)**.** A learning path is *best-response stable* for player n if one action is a best response to the opponents' last-iterate mixed strategy for all sufficiently large t. It is *strictly best-response stable* if that action is eventually the unique best response.

By Theorem 2, $q_t \to 0$, so $x_t = c_A(\bar{q} - q_t) \to c_A\bar{q} > 0$. Thus A_0 eventually remains the unique best response, giving Corollary 1 directly.

Corollary 1. *Under Theorem 2,* A_0 *is eventually player A's unique current best response.*

Hedge with k alternatives. We next turn to Hedge, where stability depends on the learning-rate exponents, and allow any fixed number of alternatives (Theorem 3). Both players use Hedge in a k-boundary game from Definition 2, with $k \geq 1$, exponents $\alpha_A, \alpha_B \in (0, 1)$, and finite initial cumulative utilities.

Theorem 3. *Every trajectory converges to* $y_0 = 1$ *and* $q_i = \bar{q}$ *for* $1 \leq i \leq k$*. The best-response threshold is independent of* k*:*

(i) If $5\alpha_A + 2\alpha_B > 3$*, then* A_0 *is eventually the unique current best response.*
(ii) If $5\alpha_A + 2\alpha_B < 3$*, the best response switches infinitely often for almost every initialization.*

Remarkably, for every fixed number of alternatives, common exponents give stability above $3/7$ and switching for almost every initialization throughout $0 < \alpha < 3/7$. To see the restoring forces in Example 1, we view $x_t = \dot{X}_t$ and $y_t = \dot{H}_t$ as the horizontal and vertical coordinates: they are growth rates, not cumulative values.

- If $y_t > \alpha_B H_t/t$, new payoffs strengthen B's preference for B_0, so q_t falls and x_t rises.
- If $y_t < \alpha_B H_t/t$, regularization weakens B's preference for B_0, so x_t falls.
- If $x_t > \alpha_A X_t/t$, A's preference for A_0 strengthens and y_t falls.
- If $x_t < \alpha_A X_t/t$, A's preference for A_0 weakens and y_t rises.

Thus more play of A_1 pushes the incentive for A_0 upward, while a stronger incentive for A_0 pushes play of A_1 downward. These are the restoring forces, but they need not stop the oscillations. We prove this using a *limiting oscillator*—a linear system obtained by rescaling coordinates and time near the reference path.

Payoff convergence. Even when the best response keeps changing, the gain from deviating can vanish. We next ask how quickly this happens in the 1-boundary game of Example 1. We combine the deviation gains $g_A(t)$ and $g_B(t)$ from Equation 3 into the weighted *payoff gap* $\mathrm{Gap}(t) \doteq g_A(t)/3 + g_B(t)$. This quantity is zero exactly at NE.

Using $x_t = 1 - 3q_t$ from Example 1, we obtain

$$\mathrm{Gap}(t) = \frac{y_t}{3} + \left(q_t - \frac{1}{3}\right)_+ = \frac{y_t}{3} + \frac{(-x_t)_+}{3}, \tag{15}$$

where $(x)_+ \doteq \max\{x, 0\}$. The second term is zero exactly when A_0 is a current best response.

For Proposition 3, both players use Hedge with a common exponent $\alpha \in (0,1) \setminus \{3/7\}$ and finite initial cumulative utilities. For $0 < \alpha < 3/7$, we write $\delta_\alpha \doteq \max\{(1+\alpha)/3, (1+3\alpha)/4\}$.

Proposition 3. *If $\alpha > 3/7$, then $\mathrm{Gap}(t) \asymp t^{-(1-\alpha)}$. For $0 < \alpha < 3/7$, $\mathrm{Gap}(t) = O(t^{-\delta_\alpha})$ for all finite initializations, and almost every initialization satisfies*

$$0 < \limsup_{t\to\infty} t^{\delta_\alpha}\, \mathrm{Gap}(t) < \infty. \tag{16}$$

Equation 16 makes the rate sharp for recurring payoff gaps: $t^{\delta_\alpha}\,\mathrm{Gap}(t)$ is bounded and stays bounded away from zero along some sequence of times tending to infinity. Below $3/7$, a larger learning rate makes these gaps decay more slowly. The exponent is $(1+\alpha)/3$ for $\alpha \le 1/5$ and $(1+3\alpha)/4$ for $\alpha > 1/5$; Appendix D explains the change. Above $3/7$, the best response is eventually A_0, so the payoff gap is $\mathrm{Gap}(t) = y_t/3$. For example, choosing $t^{-1/4}$ rather than $t^{-1/2}$ accelerates concentration from $t^{-1/2}$ to $t^{-3/4}$, but slows the decay of recurring payoff gaps from $t^{-1/2}$ to $t^{-7/16}$. In Proposition 5, we obtain the same decay exponent for $T^{-1}\int_1^T \mathrm{Gap}(t)\,\mathrm{d}t$.

Other regularizers. Finally, we ask how the stability condition changes when we replace entropy with another regularizer. Table 1 summarizes the conditions for Hedge, Tsallis, and Euclidean FTRL. For Tsallis FTRL, the response satisfies $f(\widehat{\sigma}) \asymp \widehat{\sigma}^{-\beta}$ as $\widehat{\sigma} \to \infty$. Proposition 7 in Appendix D gives the threshold $(3\beta-5)/(7\beta-5)$ for common exponents. Proposition 6 in Appendix D gives the Euclidean condition.

Table 1: Best-response stability guarantees. The Hedge row concerns every fixed k; the other rows concern Example 1. For Tsallis, $\beta > 1$, $\gamma \doteq (1-\alpha_B)/\beta$, and $\alpha_A + \gamma < 1$. The opposite strict inequalities give switching for almost every initialization for Hedge and Tsallis; the Euclidean guarantee assumes $y_t \asymp t^{\alpha_B - 1} > 0$ eventually.

Response rule	Condition for stability	Common exponents
Hedge, Thm. 3	$5\alpha_A + 2\alpha_B > 3$	$\alpha > 3/7$
Tsallis FTRL, Prop. 7	$5\alpha_A + 2\alpha_B + 5\gamma > 3$	$\alpha > \frac{3\beta-5}{7\beta-5}$
Euclidean FTRL, Prop. 6	$5\alpha_A + 3\alpha_B > 4$ and $3\alpha_A + 5\alpha_B > 4$	$\alpha > 1/2$

5 Experiments

In the inspection game, firm A chooses compliance, A_0, or a violation in one of k activities, A_i. Inspector B inspects one activity, B_i, or none, B_0. A violation is detected exactly when the inspector chooses the same activity; inspection carries no separate cost.

Compliance gives both players zero payoff. An undetected violation gives the firm $\bar{q}$ and the inspector -1; a detected violation gives the firm $-(1-\bar{q})$ and the inspector zero. We set $\bar{q} \doteq 2/(2k+1)$. If activity i is inspected with probability $q_i(t)$, compliance earns $q_i(t) - \bar{q}$ more than violating in that activity. Each violation is therefore exactly deterred at inspection probability $\bar{q}$. $u_B(B_0, \mathbf{y}_t) - u_B(B_i, \mathbf{y}_t) = -y_i(t)$. This related game is an empirical test, not a direct application of Theorem 3; Appendix E explains the difference in payoff signs. We simulate Hedge with $\alpha = 1/4$ or $\alpha = 1/2$ for both players. Here we integrate the continuous-time equations; the separate discrete-time experiments use the update in Appendix E.

Two activities with random initial utilities. For $k = 2$, we sample ten initial pairs $(s(1), r(1))$ and use the same pairs for both exponents; Appendix E defines these coordinates and gives the setup.

We define $x_i(t) \doteq u_A(A_0, \mathbf{q}_t) - u_A(A_i, \mathbf{q}_t) = q_i(t) - 2/5$. Figure 2 plots $\min_i x_i(t)$: compliance is uniquely best when it is positive, and some violation is profitable when it is negative.

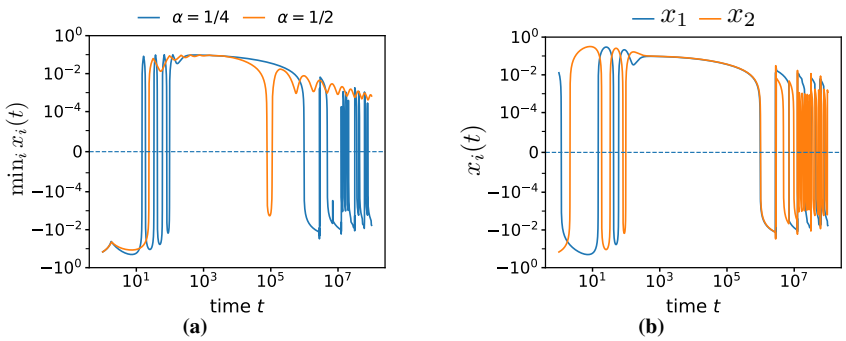


Figure 2: $x_i(t) = q_i(t) - 2/5$ in the two-activity inspection game, from a common initialization. **(a)** The minimum $\min_i x_i(t)$ repeatedly changes sign for $\alpha = 1/4$, but stays positive after the initial changes for $\alpha = 1/2$. **(b)** At $\alpha = 1/4$, $x_1(t)$ and $x_2(t)$ cross zero at different times. Horizontal axes are logarithmic; vertical axes are linear near zero and logarithmic elsewhere.

Across all ten initializations, the firm's best response changes during $[10^6, 10^8]$ for $\alpha = 1/4$, whereas compliance remains best throughout this window for $\alpha = 1/2$.

More activities. For each $k \in \{2, 3, 5\}$ and exponent, we also use twenty perturbed starting states in which compliance and all violations have nearly equal payoffs. During $[10^8, 10^9]$, the firm's best response changes in every $\alpha = 1/4$ run, whereas compliance remains best in every $\alpha = 1/2$ run.

Discrete-time FTRL. In the 1-boundary and two-activity inspection games, round-by-round FTRL gives best-response changes in all ten $\alpha = 1/4$ runs during rounds $10^6, \ldots, 10^7$, while A_0 remains best throughout this window in all ten $\alpha = 1/2$ runs (Table 2; Figure 4). These additional experiments, detailed in Appendix E, suggest that our methodology may extend to discrete-time settings. The main challenges are to control accumulated discretization errors on the scale of the vanishing x_t, and to adapt our energy and volume estimates to the discrete update. Related work obtains discrete-time guarantees by bounding discretization errors in regret estimates [27], interpolating iterates to analyze their limits [5], and constructing discrete Lyapunov functions to prove optimization rates [49].

6 Conclusion

We distinguish convergent play from a fixed best response: a player can increasingly favor one action while its current best response keeps changing. Our results (Theorem 3) identify when time-dependent regularization makes the best response settle and when it keeps switching.

We also show that faster concentration can make the weighted payoff gap $\mathrm{Gap}(t)$ vanish more slowly (Propositions 3 and 5). Our results (Propositions 7 and 6; Table 1) also show how the stability conditions depend on the learning schedule and regularizer. Inspection simulations and discrete-time experiments both show best responses that keep changing at $\alpha = 1/4$ but settle at $\alpha = 1/2$ (Figures 2 and 4; Table 2). These findings motivate studying which actions remain optimal along a learning path, not only where the path converges.

Use of artificial intelligence. In this work, we used the GPT 6 language model to help formalize the proof of the sharp payoff rates for $0 < \alpha < 1/5$ (Propositions 3 and 5; Appendix D). In this range, the oscillations differ qualitatively from those for $1/5 < \alpha < 3/7$, so the proof uses the exact nonlinear response rather than a quadratic approximation. We developed the original formulation in x_t, X_t, y_t, Y_t; our current approach studies the resulting oscillations to determine whether the best action eventually stops changing. GPT 5.6 and GPT 6 helped prepare the experiments and wrote all simulation and plotting code.

Additionally, we used GPT 5.6 and GPT 6 for writing and polishing the manuscript and for drafting and revising the experimental section. This included organizing the results and separating the proof of Theorem 3 into distinct steps, as summarized in Appendix D. We used GPT 5.1, GPT 5.2, GPT 5.6, and GPT 6 to suggest papers to cite and to help draft the related work section.

The revision process included proof audits, simulation reruns, and numerical checks of the implementation. We take responsibility for the final manuscript, mathematical claims, and computational results, including all content prepared with artificial intelligence.

REFERENCES

[1] Kenshi Abe, Mitsuki Sakamoto, and Atsushi Iwasaki. Mutation-driven follow the regularized leader for last-iterate convergence in zero-sum games. In *Proceedings of the 38th Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pp. 1–10, 2022.

[2] Ioannis Anagnostides, Ioannis Panageas, Gabriele Farina, and Tuomas Sandholm. On last-iterate convergence beyond zero-sum games. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 536–581, 2022.

[3] Baruch Awerbuch, Yossi Azar, Amir Epstein, Vahab S. Mirrokni, and Alexander Skopalik. Fast convergence to nearly optimal solutions in potential games. In *Proceedings of the 9th ACM Conference on Electronic Commerce*, pp. 264–273. ACM, 2008.

[4] Waïss Azizian, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. The last-iterate convergence rate of optimistic mirror descent in stochastic variational inequalities. In *Proceedings of the 34th Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 326–358, 2021.

[5] Michel Benaïm. Dynamics of stochastic approximation algorithms. In Jacques Azéma, Michel Émery, Michel Ledoux, and Marc Yor (eds.), *Séminaire de Probabilités XXXIII*, volume 1709 of *Lecture Notes in Mathematics*, pp. 1–68. Springer, Berlin, Heidelberg, 1999. doi: 10.1007/BFb0096509.

[6] George W. Brown. Iterative solution of games by fictitious play. In T. C. Koopmans (ed.), *Activity Analysis of Production and Allocation*, pp. 374–376. Wiley, New York, 1951.

[7] Yang Cai, Argyris Oikonomou, and Weiqiang Zheng. Finite-time last-iterate convergence for learning in multi-player games. In *Advances in Neural Information Processing Systems*, volume 35, pp. 33904–33919, 2022.

[8] Yang Cai, Gabriele Farina, Julien Grand-Clément, Christian Kroer, Chung-Wei Lee, Haipeng Luo, and Weiqiang Zheng. Fast last-iterate convergence of learning in games requires forgetful algorithms. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

[9] Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

[10] Yun Kuen Cheung and Georgios Piliouras. Vortices instead of equilibria in MinMax optimization: Chaos and butterfly effects of online learning in zero-sum games. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pp. 807–834, 2019.

[11] Ross Cressman and Yi Tao. The replicator equation and other game dynamics. *Proceedings of the National Academy of Sciences*, 111(Suppl. 3):10810–10817, 2014. doi: 10.1073/pnas.1400823111.

[12] Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training GANs with optimism. In *International Conference on Learning Representations*, 2018.

[13] Tanner Fiez, Ryann Sim, Stratis Skoulakis, Georgios Piliouras, and Lillian J. Ratliff. Online learning in periodic zero-sum games. In *Advances in Neural Information Processing Systems*, volume 34, pp. 10313–10325, 2021.

[14] Lampros Flokas, Emmanouil-Vasileios Vlatakis-Gkaragkounis, Thanasis Lianeas, Panayotis Mertikopoulos, and Georgios Piliouras. No-regret learning and mixed Nash equilibria: They do not mix. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1380–1391, 2020.

[15] Angeliki Giannou, Emmanouil-Vasileios Vlatakis-Gkaragkounis, and Panayotis Mertikopoulos. On the rate of convergence of regularized learning in games: From bandits and uncertainty to optimism and beyond. In *Advances in Neural Information Processing Systems*, volume 34, pp. 22655–22666, 2021.

[16] Angeliki Giannou, Emmanouil-Vasileios Vlatakis-Gkaragkounis, and Panayotis Mertikopoulos. Survival of the strictest: Stable and unstable equilibria under regularized learning with partial information. In *Proceedings of the 34th Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 2147–2148, 2021.

[17] Noah Golowich, Sarath Pattathil, and Constantinos Daskalakis. Tight last-iterate convergence rates for no-regret learning in multi-player games. In *Advances in Neural Information Processing Systems*, volume 33, pp. 20766–20778, 2020.

[18] Ernst Hairer, Syvert Paul Nørsett, and Gerhard Wanner. *Solving Ordinary Differential Equations I: Nonstiff Problems*, volume 8 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, Heidelberg, second revised edition, 1993. doi: 10.1007/978-3-540-78862-1.

[19] Torsten Heinrich, Yoojin Jang, Luca Mungo, Marco Pangallo, Alex Scott, Bassel Tarbush, and Samuel Wiese. Best-response dynamics, playing sequences, and convergence to equilibrium in random games. *International Journal of Game Theory*, 52(3):703–735, 2023.

[20] Josef Hofbauer and William H. Sandholm. Stable games and their dynamics. *Journal of Economic Theory*, 144(4):1665–1693, 2009.

[21] Josef Hofbauer and Karl Sigmund. Evolutionary game dynamics. *Bulletin of the American Mathematical Society*, 40(4):479–519, 2003.

[22] Josef Hofbauer, Sylvain Sorin, and Yannick Viossat. Time average replicator and best-reply dynamics. *Mathematics of Operations Research*, 34(2):263–269, 2009.

[23] Ed Hopkins. A note on best response dynamics. *Games and Economic Behavior*, 29(1–2): 138–150, 1999.

[24] Yu-Guan Hsieh, Kimon Antonakopoulos, and Panayotis Mertikopoulos. Adaptive learning in continuous games: Optimal regret bounds and convergence to Nash equilibrium. In *Proceedings of the 34th Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 2388–2422, 2021.

[25] James S. Jordan. Three problems in learning mixed-strategy Nash equilibria. *Games and Economic Behavior*, 5(3):368–386, 1993. doi: 10.1006/game.1993.1022.

[26] Anatoli Juditsky, Arkadi Nemirovski, and Claire Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.

[27] Joon Kwon and Panayotis Mertikopoulos. A continuous-time approach to online optimization. *Journal of Dynamics and Games*, 4(2):125–148, 2017. doi: 10.3934/jdg.2017008.

[28] Davide Legacci, Panayotis Mertikopoulos, Christos H. Papadimitriou, Georgios Piliouras, and Bary S. R. Pradelski. No-regret learning in harmonic games: Extrapolation in the face of conflicting interests. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

[29] H. Brendan McMahan. A survey of algorithms and analysis for adaptive online learning. *Journal of Machine Learning Research*, 18(90):1–50, 2017.

[30] Panayotis Mertikopoulos and William H. Sandholm. Learning in games via reinforcement and regularization. *Mathematics of Operations Research*, 41(4):1297–1324, 2016. doi: 10.1287/moor.2016.0778.

[31] Panayotis Mertikopoulos, Christos H. Papadimitriou, and Georgios Piliouras. Cycles in adversarial regularized learning. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 2703–2717, 2018. doi: 10.1137/1.9781611975031.172.

[32] Panayotis Mertikopoulos, Bruno Lecouat, Houssam Zenati, Chuan-Sheng Foo, Vijay Chandrasekhar, and Georgios Piliouras. Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile. In *International Conference on Learning Representations*, 2019.

[33] Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 1497–1507, 2020.

[34] Dov Monderer and Lloyd S. Shapley. Potential games. *Games and Economic Behavior*, 14(1):124–143, 1996.

[35] Arkadi Nemirovski. Prox-method with rate of convergence $O(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

[36] Yurii E. Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Doklady Akademii Nauk SSSR*, 269(3):543–547, 1983.

[37] Martin J. Osborne and Ariel Rubinstein. *Models in Microeconomic Theory*. Open Book Publishers, Cambridge, UK, expanded second edition, 2023. doi: 10.11647/OBP.0362.

[38] Boris T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964. doi: 10.1016/0041-5553(64)90137-5.

[39] Alexander Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

[40] Julia Robinson. An iterative method of solving a game. *Annals of Mathematics*, 54(2):296–301, 1951.

[41] William H. Sandholm. Local stability under evolutionary game dynamics. *Theoretical Economics*, 5(1):27–50, 2010.

[42] Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2012. doi: 10.1561/2200000018.

[43] Lloyd S. Shapley. Some topics in two-person games. In M. Dresher, L. S. Shapley, and A. W. Tucker (eds.), *Advances in Game Theory*, pp. 1–28. Princeton University Press, Princeton, 1964.

[44] Weijie Su, Stephen Boyd, and Emmanuel J. Candès. A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. *Journal of Machine Learning Research*, 17(153):1–43, 2016. URL `https://jmlr.org/papers/v17/15-084.html`.

[45] Brian Swenson, Soummya Kar, and João Xavier. From weak learning to strong learning in fictitious play type algorithms. arXiv preprint arXiv:1504.04920, 2015. URL `https://arxiv.org/abs/1504.04920`.

[46] Brian Swenson, Ryan Murray, and Soummya Kar. On best-response dynamics in potential games. *SIAM Journal on Control and Optimization*, 56(4):2734–2767, 2018.

[47] Emmanouil-Vasileios Vlatakis-Gkaragkounis, Lampros Flokas, and Georgios Piliouras. Poincaré recurrence, cycles and spurious equilibria in gradient-descent-ascent for non-convex non-concave zero-sum games. In *Advances in Neural Information Processing Systems*, volume 32, pp. 10450–10461, 2019.

[48] Andre Wibisono, Ashia C. Wilson, and Michael I. Jordan. A variational perspective on accelerated methods in optimization. *Proceedings of the National Academy of Sciences*, 113 (47):E7351–E7358, 2016. doi: 10.1073/pnas.1614734113.

[49] Ashia C. Wilson, Ben Recht, and Michael I. Jordan. A Lyapunov analysis of accelerated methods in optimization. *Journal of Machine Learning Research*, 22(113):1–34, 2021. URL `https://jmlr.org/papers/v22/20-195.html`.

[50] Yuepeng Yang and Cong Ma. $O(T^{-1})$ convergence of optimistic-follow-the-regularized-leader in two-player zero-sum Markov games. In *International Conference on Learning Representations*, 2023.

[51] Julian Zimmert and Yevgeny Seldin. Tsallis-INF: An optimal algorithm for stochastic and adversarial bandits. *Journal of Machine Learning Research*, 22(28):1–49, 2021. URL `https://jmlr.org/papers/v22/19-753.html`.

A FTRL Response Estimates

We use the FTRL objective, response notation, and asymptotic notation from Section 3. We prove Proposition 1 and give the exact Hedge and Euclidean responses. In the proofs, $c_0, c_1, \dots$ and $C_0, C_1, \dots$ denote positive constants independent of time; their values may change between bounds. Dots denote derivatives with respect to t, and $\|\cdot\|$ denotes the Euclidean norm.

Regularity conditions for the response estimate. Proposition 1 uses the regularizer and $\eta_t\sigma_t(a)$ from Section 3. We use the *lower-semicontinuous convex extension* of ϕ at 0 and 1: its endpoint limits, which may be $+\infty$. The argument requires two conditions on the Karush–Kuhn–Tucker (KKT) equation 17, the first-order optimality condition with the probability constraint. First, its multiplier λ_t, enforcing that probabilities sum to one, is bounded. Second, we choose $\widehat{\sigma}_0 \geq 0$ such that $(\phi')^{-1}(-\widehat{\sigma})$ exists for $\widehat{\sigma} \geq \widehat{\sigma}_0$. We set $f(\widehat{\sigma}) \doteq (\phi')^{-1}(-\widehat{\sigma})$ there and $f(\widehat{\sigma}) \doteq f(\widehat{\sigma}_0)$ on $[0, \widehat{\sigma}_0)$. We require $f(\widehat{\sigma}+\lambda) \asymp f(\widehat{\sigma})$ as $\widehat{\sigma} \to \infty$, uniformly for λ in each bounded interval. Bounded λ_t makes the constants in the probability bounds independent of the action and time.

Proposition 1. *For separable FTRL on a fixed finite action set, suppose $\eta_t \doteq t^{-\alpha}$, $\alpha \in (0,1)$, and the conditions in Appendix A hold. For this response, uniformly over actions and all large t,*

$$\mathbf{p}_t(a) \asymp f(\eta_t\sigma_t(a)), \qquad \lim_{\widehat{\sigma}\to\infty} f(\widehat{\sigma}) = 0. \tag{5}$$

Proof. We drop the player index, scale the FTRL objective (4) by η_t and subtract $\eta_t \max_a U_t(a)$:

$$\mathbf{p}_t \in \arg\max_{\mathbf{p}\in\Delta(\mathcal{A})} \{-\eta_t \sum_{a\in\mathcal{A}} \mathbf{p}(a)\sigma_t(a) - \Phi(\mathbf{p})\}.$$

The inverse defining f exists for arbitrarily large arguments, so $\phi'(0+) = -\infty$, $f(\widehat{\sigma}) \to 0$ as $\widehat{\sigma} \to \infty$, and finite utilities give positive probabilities. For $\Phi(\mathbf{p}) = \sum_a \phi(\mathbf{p}(a))$, we let λ_t enforce $\sum_a \mathbf{p}(a) = 1$. The KKT condition is

$$\phi'(\mathbf{p}_t(a)) = \lambda_t - \eta_t\sigma_t(a). \tag{17}$$

We choose a fixed $\widehat{\sigma}_1$ large enough that the inverse formula for f applies whenever $\eta_t\sigma_t(a) \geq \widehat{\sigma}_1$. We consider two cases:

1. $\eta_t\sigma_t(a) \geq \widehat{\sigma}_1$**.** Equation 17 gives $\mathbf{p}_t(a) = f(\eta_t\sigma_t(a) - \lambda_t) \asymp f(\eta_t\sigma_t(a))$.

2. $\eta_t\sigma_t(a) < \widehat{\sigma}_1$**.** Bounded λ_t keeps $\mathbf{p}_t(a)$ bounded away from zero. The constant extension of f therefore preserves the same two-sided bound.

□

Checking the assumptions. An action maximizing $U_t(a)$ has probability at least $1/|\mathcal{A}|$. When $\phi'(1)$ is finite, the bounds $\phi'(1/|\mathcal{A}|) \le \lambda_t \le \phi'(1)$ verify the multiplier assumption. The constant extension of f specifies its value when $\eta_t\sigma_t(a)$ is bounded. For Hedge, multiplying a small probability by a fixed positive constant changes its inverse response by an additive constant, which gives the inverse-response condition required by Theorem 1.

Hedge and Euclidean responses. For Hedge, $\phi(p) \doteq p\log p$. Equation 17 and a denominator between 1 and $|\mathcal{A}|$ give

$$\mathbf{p}_t(a) = \frac{\exp(-\eta_t\sigma_t(a))}{\sum_{a'\in\mathcal{A}}\exp(-\eta_t\sigma_t(a'))} \asymp e^{-\eta_t\sigma_t(a)}. \tag{18}$$

Euclidean FTRL with two actions. For two actions with $U_t(0) \ge U_t(1)$, we have $\sigma_t(1) = U_t(0) - U_t(1)$. We use $(x)_+ \doteq \max\{x, 0\}$. For Euclidean FTRL we use $\Phi(\mathbf{p}) \doteq \frac{1}{2}\sum_a \mathbf{p}(a)^2$, which gives

$$\mathbf{p}_t(1) = \frac{(1-\eta_t\sigma_t(1))_+}{2}. \tag{19}$$

Thus the response has a finite cutoff: once $\eta_t\sigma_t(1) \ge 1$, action 1 receives probability zero.

Scope of the response estimate. Hedge satisfies the conditions of Proposition 1. Euclidean FTRL has an exact zero-probability cutoff and uses Equation 19 instead.

B Equilibrium Convergence Proofs

We use pure Nash equilibrium (PNE), mixed Nash equilibrium (MNE), and last-iterate convergence as in Section 1 and Section 4. Best-response stability is defined in Definition 3.

Strict PNE. We prove Proposition 4 using strictness and continuity.

Proposition 4. *If a learning path converges in last iterate to a strict PNE, then both players' current best responses eventually stabilize to their equilibrium actions.*

Proof. At a strict PNE, every alternative action has a positive payoff disadvantage. By continuity and finiteness of the action sets, these disadvantages remain positive in a neighborhood of the equilibrium, which any convergent path eventually enters. □

In the strict-equilibrium game (10), every path satisfying Equation 5 with $f_n(\widehat{\sigma}) \to 0$ as $\widehat{\sigma} \to \infty$ converges to (A_1, B_1). Indeed, A_1 always earns one more unit than A_0, so $U_t^A(A_1) - U_t^A(A_0) = t + O(1)$ and $\Pr_t(A_1) \to 1$. Then $u_B(B_1, \mathbf{p}_t^A) - u_B(B_0, \mathbf{p}_t^A) \to 1$, so $U_t^B(B_1) - U_t^B(B_0) \sim t$. Equation 5 gives $\Pr_t(B_1) \to 1$, and Proposition 4 gives eventual best-response stability.

Matching pennies. We prove Proposition 2 by tracking $(s(t), r(t))$, defined below.

Proposition 2. *For Hedge in matching pennies* (11)*, with $\eta_t \doteq t^{-\alpha}$ and $\alpha \in (0,1)$, almost every initialization converges in mixed strategy to $\Pr(A_1) = \Pr(B_1) = 1/2$, while both current best responses switch infinitely often.*

Proof. We multiply $U_t^n(a) - U_t^n(a')$ by the learning rate, as in Section 3, and define

$$s(t) \doteq t^{-\alpha}\big(U_t^A(A_0) - U_t^A(A_1)\big), \qquad r(t) \doteq t^{-\alpha}\big(U_t^B(B_0) - U_t^B(B_1)\big).$$

Hedge gives

$$y_t \doteq \Pr_t(A_1) = \frac{1}{1+e^{s(t)}}, \qquad q_t \doteq \Pr_t(B_1) = \frac{1}{1+e^{r(t)}}. \tag{20}$$

The payoffs in Equation 11 give

$$\dot{s} = -\frac{\alpha}{t}s + 2t^{-\alpha}\frac{e^r - 1}{e^r + 1}, \qquad \dot{r} = -\frac{\alpha}{t}r - 2t^{-\alpha}\frac{e^s - 1}{e^s + 1}. \tag{21}$$

We define the energy $\mathcal{E}(s, r) \doteq 2\log((1 + e^s)(1 + e^r)/4) - s - r$. Differentiating along Equation 21 yields

$$\dot{\mathcal{E}}(t) = -\frac{\alpha}{t}\left(\frac{s(e^s - 1)}{e^s + 1} + \frac{r(e^r - 1)}{e^r + 1}\right).$$

For every real s', $s'(e^{s'} - 1)/(e^{s'} + 1) \geq 2\log(1 + e^{s'}) - s' - 2\log 2 \geq 0$. Hence $\dot{\mathcal{E}} \leq -\alpha\mathcal{E}/t$ and $\mathcal{E} \to 0$, so $s, r \to 0$. Equation 20 gives $y_t, q_t \to 1/2$.

By uniqueness, a path starting away from the equilibrium $(s, r) = (0, 0)$ never reaches it at finite time. We choose a continuous *polar angle* $\theta(t)$ along $(s(t), r(t))$, following its rotation without resetting after each turn. Then

$$\dot{\theta}(t) = \frac{s\dot{r} - r\dot{s}}{s^2 + r^2} = -\frac{2t^{-\alpha}}{s^2 + r^2}\left(\frac{s(e^s - 1)}{e^s + 1} + \frac{r(e^r - 1)}{e^r + 1}\right).$$

Since $(e^{s'} - 1)/(e^{s'} + 1) = s'/2 + O(s'^3)$ near zero, $\dot{\theta} = -t^{-\alpha} + o(t^{-\alpha})$. The integral of $t^{-\alpha}$ diverges because $\alpha < 1$, so $\theta(t) \to -\infty$ and both s and r change sign infinitely often. These signs determine the two players' current best responses. □

C Rate Proofs for One Alternative

We use f_A, α_A, α_B and the inverse condition from Section 3, the variables x_t, X_t, y_t, Y_t, H_t from Section 3, and pruning and rate regularity from Section 4.

Proof of Theorem 1. We exclude the two rates in Equation 14 that are incompatible with its assumptions.

Theorem 1. *In a* 1*-boundary game satisfying Equations 5 and 9, eventual* $X_t \geq 0$*, rate regularity* (14)*, and fast pruning* (12)*,*

$$y_t \asymp t^{\alpha_B - 1}, \qquad X_t \asymp t^{\alpha_A} f_A^{-1}(t^{\alpha_B - 1}), \qquad \frac{1}{t}\int_1^t q_{t'}\,\mathrm{d}t' \longrightarrow \bar{q}.$$

Proof. Since $X_t \geq 0$ eventually, Equation 5 gives $y_t \asymp f_A(t^{-\alpha_A} X_t)$. For every $c > 0$, monotonicity and fast pruning (12) give

$$t f_A(ct^{1-\alpha_A}) \leq 2\int_{t/2}^t f_A(ct'^{1-\alpha_A})\,\mathrm{d}t' = o(t^{\alpha_B}).$$

Under rate regularity (14), there are three possible growth rates. We rule out two of them:

1. $Y_t = \omega(t^{\alpha_B})$**.** We have $H_t = \omega(t^{\alpha_B})$ and hence $c_B t^{-\alpha_B} H_t \to \infty$, so Equation 5 gives $q_t \to 0$. Thus x_t is eventually bounded below by a positive constant and $X_t \geq ct$ for some $c > 0$. The response estimate and the preceding fast-pruning bound give $Y_t \lesssim t f_A(ct^{1-\alpha_A}) = o(t^{\alpha_B})$, again a contradiction.

2. $Y_t = o(t^{\alpha_B})$**.** We have $H_t = o(t^{\alpha_B})$, so Equation 9 gives $q_t \to 1/2$. Since $\bar{q} < 1/2$, $x_t = c_A(\bar{q} - q_t)$ is eventually bounded above by a negative constant. This forces $X_t \to -\infty$, contradicting eventual $X_t \geq 0$.

The only remaining case under rate regularity (14) is $Y_t \asymp t^{\alpha_B}$ and hence $y_t \asymp t^{\alpha_B - 1}$. Applying the inverse condition from Section 3 to $f_A(t^{-\alpha_A} X_t) \asymp t^{\alpha_B - 1}$ yields $X_t \asymp t^{\alpha_A} f_A^{-1}(t^{\alpha_B - 1})$.

Finally, $X_t = o(t)$. Otherwise, $X_{t_n} \geq \varepsilon t_n$ along some sequence $t_n \to \infty$, and fast pruning would contradict $Y_t \asymp t^{\alpha_B}$:

$$Y_{t_n} \lesssim t_n f_A(\varepsilon t_n^{1-\alpha_A}) = o(t_n^{\alpha_B}).$$

Dividing $X_t = X_1 + \int_1^t c_A(\bar{q} - q_{t'})\,\mathrm{d}t'$ by t proves $t^{-1}\int_1^t q_{t'}\,\mathrm{d}t' \to \bar{q}$. □

Proof of Theorem 2. We show that the accumulated probability of A_1 makes player B converge to B_0.

Theorem 2. *In a* 1*-boundary game satisfying Equation 5, eventual* $X_t \geq 0$*, slow pruning* (13)*, and* $f_A(c\widehat{\sigma}) \asymp f_A(\widehat{\sigma})$ *for every fixed* $c > 0$*,*

$$y_t \asymp f_A(t^{1-\alpha_A}), \qquad X_t \asymp t, \qquad q_t \longrightarrow 0.$$

Proof. Bounded one-period payoffs give $X_t \leq Ct$ eventually. Since $X_t \geq 0$ and f_A is nonincreasing, Equation 5 gives $y_t \gtrsim f_A(Ct^{1-\alpha_A})$. Slow pruning (13) then implies $H_t = \omega(t^{\alpha_B})$, so $c_B t^{-\alpha_B} H_t \to \infty$ and $q_t \to 0$. Consequently, $x_t \to c_A \bar{q} > 0$ and $X_t \asymp t$. For some constants $0 < c < C < \infty$,

$$f_A(Ct^{1-\alpha_A}) \lesssim y_t \lesssim f_A(ct^{1-\alpha_A}).$$

The assumption $f_A(c\widehat{\sigma}) \asymp f_A(\widehat{\sigma})$ gives $y_t \asymp f_A(t^{1-\alpha_A})$. □

A slow-pruning response. For the example in Section 4, we take $\phi(p) \doteq 1/p$ on $(0, 1]$ and $\phi(0) \doteq +\infty$. Since $\phi'(p) = -p^{-2}$, Appendix A gives $f_A(\widehat{\sigma}) \asymp \widehat{\sigma}^{-1/2}$ as $\widehat{\sigma} \to \infty$. For $\alpha_A = \alpha_B = \alpha \in (0, 1)$ and every fixed $c > 0$,

$$\int_1^t f_A(ct'^{1-\alpha})\,\mathrm{d}t' \asymp \int_1^t t'^{-(1-\alpha)/2}\,\mathrm{d}t' \asymp t^{(1+\alpha)/2} = \omega(t^{\alpha}).$$

Thus Equation 13 holds, and $f_A(c\widehat{\sigma}) \asymp f_A(\widehat{\sigma})$ for every fixed $c > 0$.

D Best-Response Threshold Proofs

Proof Overview. The Hedge and Tsallis arguments use the same steps. We use s_i, r_i and their reference coordinates $\bar{s}, \bar{r}$, defined in Appendix D. At common coordinates, $u_A(A_0, \mathbf{q}) - u_A(A_i, \mathbf{q}) = F(r)$, with F defined there. We first prove convergence to the boundary strategy profile. For switching, an eventually fixed best response would force approach to the reference, which we exclude for almost every initialization. For stability, we bound the error relative to $F(\bar{r})$.

1. **We define the responses and reference path.** We specify the exact equations and the reference in Appendix D. We estimate $F(\bar{r})$ in Equation 29 and the *residual* $\rho_r(t)$ in Equation 30. The residual is the error left when we substitute the reference into the dynamics.

2. **We prove convergence from every finite initialization.** The convex functions $\mathcal{V}_t, \mathcal{W}_t$ in Equations 31 and 32 are *potentials*: their gradients give the differences between current and reference responses. We combine them into an *energy*, a nonnegative function bounding both deviations. Its derivative satisfies Equation 35, which yields convergence of the mixed strategies. On the stable side, adding a product of the deviations gives the stronger approach estimate (37).

3. **An eventually fixed best response forces approach to the reference.** In Lemma 2, we use eventual optimality of A_0 to show that each $t^{\alpha_A} s_i(t)$ is nondecreasing. A bound on their growth limits how far the opponent can stay from indifference. Averages just before and after t then determine the current action probabilities, giving Equation 38.

4. **A fixed best response forces the rescaled deviations to vanish.** The coordinate and time scales (43) produce the limiting oscillator (46). Lemma 4 approximates the trajectory over bounded intervals. A fixed best response would impose sign restrictions on the coordinates of $C\widehat{r}_\infty$ in Equation 46. Their oscillation and zero mean force the limiting solution to vanish.

5. **We use volume preservation to prove switching.** Volume preservation means that the map from initial states to rescaled states preserves Lebesgue volume. The coordinate scales exactly cancel the contraction in Equation 47. Lemma 3 rules out vanishing rescaled deviations for almost every initialization. Together with Steps 3–4, this proves the switching claims.

6. **We show that the error is $o(F(\bar{r}))$.** One correction of the reference makes the second equation exact. The local energy (50) gives Equation 53. With an integrating factor, we show that the error is $o(F(\bar{r}))$, so A_0 eventually stays uniquely best.

7. **We quantify** $\mathrm{Gap}(t)$ **throughout the switching range.** Here both players use Hedge with $\alpha_A = \alpha_B = \alpha$. For $1/5 < \alpha < 3/7$, Equation 58 bounds the rescaled energy and gives a positive limit for almost every initialization. For $0 < \alpha \le 1/5$, the exact nonlinear energy (64) replaces the quadratic approximation. In both cases, the interval bound (60) and payoff identity (15) give the rates in Propositions 3 and 5. They show that a larger learning rate can slow the disappearance of profitable deviations despite faster concentration.

Proposition 6 proves convergence for Euclidean FTRL and gives its sufficient stability condition using Equation 72. Proposition 7 applies the shared estimates to Tsallis FTRL.

Global Convergence and Approach to the Reference. For Hedge in Definition 2, we multiply each difference by its player's learning rate and define

$$s_i(t) \doteq t^{-\alpha_A}\big(U_t^A(A_0) - U_t^A(A_i)\big), \qquad r_i(t) \doteq t^{-\alpha_B}\big(U_t^B(B_0) - U_t^B(B_i)\big).$$

We write $s = (s_1, \dots, s_k)$ and $r = (r_1, \dots, r_k)$, and let $\mathbf{1} \in \mathbb{R}^k$ be the all-ones vector. Sums over alternatives run from 1 to k. The exact probabilities and equations are

$$y_i = \frac{e^{-s_i}}{1+\sum_j e^{-s_j}}, \qquad q_i = \frac{e^{-r_i}}{1+\sum_j e^{-r_j}}, \qquad y_0 = 1 - \sum_i y_i, \quad q_0 = 1 - \sum_i q_i,$$

$$\dot{s}_i = -\frac{\alpha_A}{t} s_i + c_A t^{-\alpha_A}(\bar{q} - q_i), \qquad \dot{r}_i = -\frac{\alpha_B}{t} r_i + c_B t^{-\alpha_B} y_i. \tag{22}$$

When all coordinates within each player are equal, we write s, r for their common scalar values. We define the payoff terms $F(r) \doteq c_A(\bar{q} - e^{-r}/(1+ke^{-r}))$ and $G(s) \doteq c_B e^{-s}/(1+ke^{-s})$. The value $r_\star \doteq \log((1-k\bar{q})/\bar{q}) > 0$ makes A indifferent, since $F(r_\star) = 0$.

We also treat Tsallis FTRL in Example 1, where $k = 1$ and $c_B = 1$. Both players use

$$\Phi_\beta(\mathbf{p}) \doteq -\frac{\beta}{\beta - 1} \sum_{a \in \mathcal{A}} \mathbf{p}(a)^{1-1/\beta}, \qquad \beta > 1. \tag{23}$$

For each $s \in \mathbb{R}$, its exact response $G(s) \in (0,1)$ is the unique probability satisfying

$$s = G(s)^{-1/\beta} - (1 - G(s))^{-1/\beta}. \tag{24}$$

For this response, we set $F(r) \doteq 1 - 3G(r)$ and $r_\star \doteq 3^{1/\beta} - (3/2)^{1/\beta} > 0$. With $s = t^{-\alpha_A} X_t$ and $r = t^{-\alpha_B} H_t$, we obtain

$$\dot{s} = -\frac{\alpha_A}{t} s + t^{-\alpha_A} F(r), \qquad \dot{r} = -\frac{\alpha_B}{t} r + t^{-\alpha_B} G(s), \qquad y_t = G(s), \quad q_t = G(r). \tag{25}$$

Equation 25 also describes Hedge in Example 1, using $G(s) = (1+e^s)^{-1}$ and $F = 1 - 3G$.

For the shared estimates below, $\alpha_A, \alpha_B \in (0,1)$ and initial cumulative utilities are finite. We set

$$\gamma \doteq \begin{cases} 0, & \text{Hedge}, \\ (1-\alpha_B)/\beta, & \text{Tsallis}, \end{cases} \qquad h \doteq 1 - \alpha_A - \gamma > 0, \qquad \nu \doteq \frac{3\alpha_A + 2\alpha_B + 3\gamma - 1}{2h}. \tag{26}$$

For Tsallis, the assumption $h > 0$ is the fast-pruning condition $\beta(1-\alpha_A) > 1 - \alpha_B$. For large t, the common scalar reference coordinates satisfy

$$G(\bar{s}(t)) \doteq \alpha_B r_\star t^{\alpha_B - 1}, \qquad F(\bar{r}(t)) \doteq t^{\alpha_A}\left(\dot{\bar{s}}(t) + \frac{\alpha_A}{t}\bar{s}(t)\right). \tag{27}$$

Lemma 1. *Every finite initialization satisfies $s_i \to \infty$ and $r_i \to r_\star$ for Hedge, and $s \to \infty$, $r \to r_\star$ for Tsallis. If $\nu > 0$, then*

$$\begin{aligned} &s_i - \bar{s} \to 0, \quad t^{h/2}(r_i - \bar{r}) \to 0, \quad y_i \sim (\alpha_B r_\star / c_B) t^{\alpha_B - 1}, && \textit{Hedge}, \\ &s/\bar{s} \to 1, \quad t^{h/2}(r - \bar{r}) \to 0, \quad y_t \sim \alpha_B r_\star t^{\alpha_B - 1}, && \textit{Tsallis}. \end{aligned}$$

Proof. **Response and reference estimates.** The exact responses are smooth and bounded, so the equations for (s, r) have global solutions. For Tsallis, the right-hand side of $s = y^{-1/\beta} - (1-y)^{-1/\beta}$ is strictly decreasing and onto $\mathbb{R}$. Inverse differentiation gives

$$G'(s) = -\frac{\beta}{y^{-1-1/\beta} + (1-y)^{-1-1/\beta}}, \qquad y = G(s),$$
$$G(s) \sim s^{-\beta},\ G'(s) \sim -\beta s^{-\beta-1},\ G''(s) \sim \beta(\beta+1)s^{-\beta-2},\ G'''(s) \sim -\beta(\beta+1)(\beta+2)s^{-\beta-3}. \tag{28}$$

For the common Hedge coordinates, the jth derivative satisfies $G^{(j)}(s) \sim (-1)^j c_B e^{-s}$ for each fixed integer $j \geq 0$. Differentiating Equation 27 gives

$$\begin{aligned} &\bar{s} = (1-\alpha_B)\log t + \log\frac{c_B}{\alpha_B r_\star} + o(1), \quad \bar{r} - r_\star \sim \frac{\alpha_A(1-\alpha_B)}{c_A \bar{q}(1-k\bar{q})} t^{-h}\log t && \text{(Hedge)}, \\ &\bar{s} \sim (\alpha_B r_\star)^{-1/\beta} t^{\gamma}, && \bar{r} - r_\star \sim \frac{(\alpha_A+\gamma)(\alpha_B r_\star)^{-1/\beta}}{F'(r_\star)} t^{-h} \quad \text{(Tsallis)}. \end{aligned} \tag{29}$$

For Hedge, $t\dot{\bar{s}} \to 1 - \alpha_B$; for Tsallis, $t\dot{\bar{s}}/\bar{s} \to \gamma$. In both cases $\dot{\bar{s}} > 0$ eventually and $t\dot{\bar{r}} \to 0$. The residual in the second equation is

$$\rho_r(t) \doteq -\dot{\bar{r}} - \frac{\alpha_B}{t}(\bar{r} - r_\star) = \begin{cases} O(t^{-1-h}\log t), & \text{Hedge}, \\ O(t^{-1-h}), & \text{Tsallis}. \end{cases} \tag{30}$$

Its derivative has the same bound with one additional factor t^{-1}, by the explicit inverse response and Equation 28.

Convex potentials. We use deviations $\widetilde{s} \doteq s - \bar{s}\mathbf{1}$ and $\widetilde{r} \doteq r - \bar{r}\mathbf{1}$; for Tsallis, $k = 1$ and these are scalars. **Hedge.** We take

$$\begin{aligned} \mathcal{V}_t(\widetilde{r}) &\doteq c_A\left[\log\frac{1 + e^{-\bar{r}}\sum_i e^{-\widetilde{r}_i}}{1 + ke^{-\bar{r}}} + \frac{e^{-\bar{r}}}{1+ke^{-\bar{r}}}\mathbf{1}^\top\widetilde{r}\right], \\ \mathcal{W}_t(\widetilde{s}) &\doteq c_B\left[\log\frac{1 + e^{-\bar{s}}\sum_i e^{-\widetilde{s}_i}}{1 + ke^{-\bar{s}}} + \frac{e^{-\bar{s}}}{1+ke^{-\bar{s}}}\mathbf{1}^\top\widetilde{s}\right]. \end{aligned} \tag{31}$$

Tsallis. We use

$$\begin{aligned} \mathcal{V}_t(\widetilde{r}) &\doteq \int_{\bar{r}}^{\bar{r}+\widetilde{r}} [F(r') - F(\bar{r})]\,\mathrm{d}r', \\ \mathcal{W}_t(\widetilde{s}) &\doteq \int_{\bar{s}}^{\bar{s}+\widetilde{s}} [G(\bar{s}) - G(s')]\,\mathrm{d}s'. \end{aligned} \tag{32}$$

These are nonnegative convex functions, minimized at zero, whose gradients give differences between current and reference responses. Subscripts t index the potentials, and state subscripts denote partial derivatives. In $\partial_{\bar{s}}\mathcal{W}$ and $\partial_{\bar{r}}\mathcal{V}$, the deviations are held fixed. Dots denote derivatives along trajectories in t; we omit the time index on the potentials below. In both cases the exact centered equations are

$$\dot{\widetilde{s}} = -\frac{\alpha_A}{t}\widetilde{s} + t^{-\alpha_A}\nabla\mathcal{V}, \qquad \dot{\widetilde{r}} = -\frac{\alpha_B}{t}\widetilde{r} - t^{-\alpha_B}\nabla\mathcal{W} + \rho_r\mathbf{1}. \tag{33}$$

Convexity gives $\widetilde{r}^\top\nabla\mathcal{V} \geq \mathcal{V}$ and $\widetilde{s}^\top\nabla\mathcal{W} \geq \mathcal{W}$.

Hedge. We set $\bar{y} \doteq e^{-\bar{s}}/(1 + ke^{-\bar{s}})$. Direct differentiation yields the global identity

$$\partial_{\bar{s}}\mathcal{W} - k\bar{y}\mathcal{W} = c_B\sum_i\left(\bar{y} - y_i + \bar{y}\log\frac{y_i}{\bar{y}}\right) \leq 0.$$

The sign follows from $\log(y_i/\bar{y}) \leq y_i/\bar{y} - 1$. Since $\bar{y} \to 0$ and $t\dot{\bar{s}}$ is bounded, this controls the motion of the first reference coordinate globally. Also $|\partial_{\bar{r}}\mathcal{V}| \leq \mathcal{V}$. To verify the latter bound, we write $M \doteq \mathrm{diag}(q_1, \ldots, q_k) - (q_i q_j)_{i,j=1}^k$ for the Hessian of $\log(1 + \sum_i e^{-r_i})$, the logarithm of the Hedge denominator. Under a common shift of its inputs,

$$\partial_{\bar{r}}M = -q_0 M + q_0(q_i q_j)_{i,j=1}^k, \qquad -M \preceq \partial_{\bar{r}}M \preceq M.$$

The matrix inequalities mean that $M - \partial_{\bar{r}} M$ and $M + \partial_{\bar{r}} M$ are positive semidefinite. The inequalities follow from Cauchy–Schwarz and $M \succeq q_0 \operatorname{diag}(q_1, \ldots, q_k)$. We integrate these inequalities in the Hessian representation of $\mathcal{V}_t(\widetilde{r})$ along $\bar{r}\mathbf{1} + \theta\widetilde{r}$, $0 \le \theta \le 1$.

Tsallis. Bounded responses give $|\mathcal{V}_{\widetilde{r}}| \le 3$, and $|\partial_{\bar{r}}\mathcal{V}| \le C_0\mathcal{V}$: Taylor expansion proves this near zero, positivity proves it on compact sets away from zero, and both functions grow at most linearly at infinity, with $\mathcal{V}$ growing at least linearly. Eventual convexity of G gives

$$t\dot{\bar{s}}\,\partial_{\bar{s}}\mathcal{W} \le o(1)\widetilde{s}\mathcal{W}_{\widetilde{s}}. \tag{34}$$

We take s_0 large enough that G is convex on $[s_0, \infty)$. Since $\bar{s} \to \infty$, we have $\bar{s} \ge s_0$ eventually. We distinguish two cases:

1. $s \ge s_0$**.** Convexity gives $\partial_{\bar{s}}\mathcal{W} = \widetilde{s}G'(\bar{s}) + G(\bar{s}) - G(s) \le 0$.
2. $s < s_0$**.** We have $\widetilde{s} < 0$, $\widetilde{s}\mathcal{W}_{\widetilde{s}} \ge |\widetilde{s}|G(s_0)/2$, and $\partial_{\bar{s}}\mathcal{W} \le |\widetilde{s}||G'(\bar{s})|$. Since $t\dot{\bar{s}}|G'(\bar{s})| \to 0$, we obtain Equation 34 in this case as well.

Convergence of the strategy profile. We use the energy

$$\mathcal{E}_0 \doteq (1 + t^{\alpha_B - \alpha_A})\mathcal{V} + (1 + t^{\alpha_A - \alpha_B})\mathcal{W}.$$

Its interaction terms cancel because $(1+t^{\alpha_B-\alpha_A})t^{-\alpha_B} = (1+t^{\alpha_A-\alpha_B})t^{-\alpha_A}$. The preceding bounds give

$$t\dot{\mathcal{E}}_0 \le -\big(\min\{\alpha_A, \alpha_B\} - o(1)\big)\mathcal{E}_0 + C_0 t|\rho_r(t)|(1 + t^{\alpha_B - \alpha_A}). \tag{35}$$

The last term is $O((t^{\alpha_A - 1} + t^{\alpha_B - 1})\log t)$ for Hedge and $O(t^{-h} + t^{\alpha_B + \gamma - 1})$ for Tsallis. Both tend to zero, because $\alpha_B + \gamma - 1 = -(1 - \alpha_B)(1 - 1/\beta) < 0$ for Tsallis. An integrating factor in $\log t$ gives $\mathcal{V}, \mathcal{W} \to 0$. **Hedge.** These potentials are scaled relative entropies from the reference distribution $\bar{\mathbf{p}}$ to the current distribution $\mathbf{p}$, of the form $\sum_a \bar{\mathbf{p}}(a)\log(\bar{\mathbf{p}}(a)/\mathbf{p}(a))$. Their vanishing, compactness of the simplex, and lower semicontinuity of relative entropy yield $q_i \to \bar{q}$ and $y_0 \to 1$. Thus $r_i = \log(q_0/q_i) \to r_\star$ and $s_i = \log(y_0/y_i) \to \infty$. **Tsallis.** From $\mathcal{V} \to 0$, we obtain $r \to r_\star$. If $s \le s_0 < \bar{s}$ along a sequence, then

$$\mathcal{W}_t(s - \bar{s}) \ge \mathcal{W}_t(s_0 - \bar{s}) \longrightarrow \int_{s_0}^{\infty} G(s')\,\mathrm{d}s' > 0,$$

contradicting $\mathcal{W} \to 0$. Here $\bar{s}G(\bar{s}) \to 0$, since $\beta > 1$.

Relative approach when $\nu > 0$. For Hedge, every coordinate between s and $\bar{s}\mathbf{1}$ tends to infinity. The softmax Hessian on this entire segment is therefore its diagonal exponential Hessian up to a relative $o(1)$ error in quadratic forms. For Tsallis, Equation 28 holds uniformly between s and $\bar{s}$. Integrating the respective Hessians gives

$$\mathcal{W} = \begin{cases} c_B \bar{y} \sum_i (\widetilde{s}_i + e^{-\widetilde{s}_i} - 1)(1 + o(1)), & \text{Hedge}, \\ \bar{s}^{1-\beta}\left[\dfrac{\widetilde{s}}{\bar{s}} + \dfrac{(1 + \widetilde{s}/\bar{s})^{1-\beta} - 1}{\beta - 1}\right](1 + o(1)), & \text{Tsallis}. \end{cases} \tag{36}$$

For Hedge, differentiating its Hessian under a common shift gives $\partial_{\bar{s}}\mathcal{W} = (-1 + o(1))\mathcal{W}$ uniformly on that segment. For Tsallis, differentiating the Hessian integral under simultaneous scaling of $\bar{s}, \widetilde{s}$ and using $sG''(s)/G'(s) \to -(\beta + 1)$ gives $\bar{s}\partial_{\bar{s}}\mathcal{W} + \widetilde{s}\mathcal{W}_{\widetilde{s}} = (1 - \beta + o(1))\mathcal{W}$. In both cases,

$$t\dot{\bar{s}}\,\partial_{\bar{s}}\mathcal{W} = -(1 - \alpha_B - \gamma)\mathcal{W} - \gamma\widetilde{s}^{\top}\nabla\mathcal{W} + o(\mathcal{W} + \widetilde{s}^{\top}\nabla\mathcal{W}).$$

Also $\mathcal{V} \asymp \|\widetilde{r}\|^2$ and $\widetilde{r}^{\top}\nabla\mathcal{V} = (2 + o(1))\mathcal{V}$. We set

$$\mathcal{E}_1 \doteq t^h\mathcal{V} + t^{1-\alpha_B-\gamma}\mathcal{W}, \qquad \widehat{\mathcal{E}}_1 \doteq \mathcal{E}_1 + \lambda t^{-\gamma}\widetilde{s}^{\top}\widetilde{r}, \qquad \lambda \doteq \frac{2\alpha_B - \alpha_A - \gamma - 1}{4}.$$

The functions in Equation 36 bound the absolute value of their argument by a constant times their square root plus themselves. Hence

$$\|t^{-\gamma}\widetilde{s}\| \le C_0(\sqrt{\mathcal{E}_1} + \mathcal{E}_1), \qquad |t^{-\gamma}\widetilde{s}^{\top}\widetilde{r}| \le C_0(t^{-h/2} + \|\widetilde{r}\|)\mathcal{E}_1 = o(\mathcal{E}_1).$$

Thus $\widehat{\mathcal{E}}_1 = (1 + o(1))\mathcal{E}_1$ is eventually nonnegative. Differentiation and the exact cancellation in Equation 33 give

$$\begin{aligned} t\dot{\widehat{\mathcal{E}}}_1 = &-(\nu h - o(1))t^h\mathcal{V} - (\nu h/2 - o(1))t^{1-\alpha_B-\gamma}\widetilde{s}^{\top}\nabla\mathcal{W} + o(\mathcal{E}_1) \\ &+ \rho_r\left(t^{h+1}\mathbf{1}^{\top}\nabla\mathcal{V} + \lambda t^{1-\gamma}\mathbf{1}^{\top}\widetilde{s}\right). \end{aligned}$$

Here $h - 2\alpha_B + 2\lambda = -\nu h$ and $\alpha_A + \gamma + \lambda = \nu h/2$. The last terms are bounded by

$$C_0(1+\log t)\left[t^{-h/2}\sqrt{\mathcal{E}_1} + t^{-h}(\sqrt{\mathcal{E}_1} + \mathcal{E}_1)\right].$$

Since $\nu > 0$, Young's inequality and convexity yield

$$t\dot{\widehat{\mathcal{E}}}_1 \le -c_0\widehat{\mathcal{E}}_1 + C_0 t^{-h}(1+\log t)^2$$

for fixed $c_0, C_0 > 0$ at large t. An integrating factor gives $\mathcal{E}_1 \to 0$. Consequently,

$$t^{h/2}\widetilde{r} \to 0, \qquad \begin{cases} s - \bar{s}\mathbf{1} \to 0, & \text{Hedge}, \\ s/\bar{s} \to 1, & \text{Tsallis}. \end{cases} \tag{37}$$

The probability asymptotics follow from the exact responses and Equation 27. □

Approach forced by an eventually fixed best response.

Lemma 2. *For the games, algorithms, and exponents in Lemma 1, an eventually fixed best response must be A_0 and implies*

$$\begin{array}{llll} y_i \sim (\alpha_B r_\star/c_B)t^{\alpha_B-1}, & s_i - \bar{s} \to 0, & r_i - \bar{r} \to 0, & \textit{Hedge}, \\ y_t \sim \alpha_B r_\star t^{\alpha_B-1}, & s/\bar{s} \to 1, & r - \bar{r} \to 0, & \textit{Tsallis}. \end{array} \tag{38}$$

Proof. We first distinguish which action could remain best:

1. A_i **remains best for some** $i \ge 1$**.** Then $t^{\alpha_A}s_i$ is nonincreasing, contradicting $s_i \to \infty$ from Lemma 1.

2. A_0 **remains best.** For Hedge, $q_i \le \bar{q}$ and $q_0 \ge 1 - k\bar{q}$ imply $r_i = \log(q_0/q_i) \ge r_\star$. For Tsallis, $F(r) \ge 0$ implies $r \ge r_\star$.

Thus only A_0 can remain best, and each $t^{\alpha_A}s_i$ is eventually nondecreasing.

A growth bound for the scores. The equation for r_i and its limit give

$$c_B\int_t^{2t} y_i(t')\,\mathrm{d}t' = (2t)^{\alpha_B}r_i(2t) - t^{\alpha_B}r_i(t) \sim r_\star(2^{\alpha_B} - 1)t^{\alpha_B}.$$

For Tsallis, we use $k = 1$ and $c_B = 1$. For $t' \in [t, 2t]$, monotonicity gives $s_i(t') \ge 2^{-\alpha_A}s_i(t)$. The Hedge bound $y_i(t') \le e^{-s_i(t')}$ and the Tsallis bound $G(s) \asymp s^{-\beta}$ therefore imply

$$s_i(t) = O(\log t) \quad \text{for Hedge}, \qquad s(t) = O(t^\gamma) \quad \text{for Tsallis}. \tag{39}$$

We bound the distance from $r_\star$**.** We define $\Delta r(t) \doteq \sum_i (r_i(t) - r_\star) \ge 0$. For Hedge,

$$\Delta r = k\log\frac{q_0}{1-k\bar{q}} - \sum_i \log\frac{q_i}{\bar{q}} \asymp \sum_i(\bar{q} - q_i)$$

near the limiting distribution in the region $q_i \le \bar{q}$. For Tsallis, $F(r) \asymp r - r_\star = \Delta r$ in this region. Thus $\sum_i c_A(\bar{q} - q_i) \ge c_0\Delta r$ for Hedge and $F(r) \ge c_0\Delta r$ for Tsallis, for some fixed $c_0 > 0$ eventually. Also $\frac{\mathrm{d}}{\mathrm{d}t}\Delta r \ge -C_0/t$, since r_i is bounded and $y_i \ge 0$. Whenever $\Delta r(t) > 0$, it remains at least $\Delta r(t)/2$ throughout $[t, t + c_1 t\Delta r(t)]$ for a sufficiently small fixed $c_1 > 0$. We integrate $\frac{\mathrm{d}}{\mathrm{d}t}(t^{\alpha_A}\sum_i s_i)$ and apply Equation 39 to obtain

$$c_2 t\Delta r(t)^2 \le \left[t'^{\alpha_A}\sum_i s_i(t')\right]_t^{t+c_1 t\Delta r(t)} \le \begin{cases} C_1 t^{\alpha_A}\log t, & \text{Hedge}, \\ C_1 t^{\alpha_A+\gamma}, & \text{Tsallis}. \end{cases}$$

In particular,

$$0 \le r_i(t) - r_\star \le \Delta r(t) = O(t^{-h/2}\sqrt{\log t}). \tag{40}$$

We bound the current probability using short averages. We take $\Delta t \doteq t^{1-h/4}$. The equation for r_i and the endpoint estimate (40) yield

$$\frac{1}{\Delta t}\int_t^{t+\Delta t} y_i(t')\,\mathrm{d}t' \sim \frac{\alpha_B r_\star}{c_B} t^{\alpha_B-1}, \qquad \frac{1}{\Delta t}\int_{t-\Delta t}^{t} y_i(t')\,\mathrm{d}t' \sim \frac{\alpha_B r_\star}{c_B} t^{\alpha_B-1}. \tag{41}$$

The relative endpoint error is $O(t^{-h/4}\sqrt{\log t}) = o(1)$. Monotonicity of $t^{\alpha_A} s_i$ and the logarithmic bound for Hedge give $s_i(t') \geq s_i(t) - o(1)$ on the forward interval and $s_i(t') \leq s_i(t) + o(1)$ on the backward interval; its softmax denominators tend to one uniformly there. For Tsallis, monotonicity gives the corresponding multiplicative bounds $s(t') \geq (t/t')^{\alpha_A} s(t)$ forward and $s(t') \leq (t/t')^{\alpha_A} s(t)$ backward. Since $(t/t')^{\alpha_A} = 1 + O(t^{-h/4})$ and $G(s) \sim s^{-\beta}$, both algorithms satisfy

$$y_i(t') \leq (1+o(1))y_i(t) \quad (t \leq t' \leq t+\Delta t), \qquad y_i(t') \geq (1-o(1))y_i(t) \quad (t-\Delta t \leq t' \leq t).$$

Equation 41 now determines $y_i(t)$ to leading order. The exact responses give $s_i - \bar{s} \to 0$ for Hedge and $s/\bar{s} \to 1$ for Tsallis. The limits of r_i and $\bar{r}$ complete the proof. □

Parameter and response conventions. For Hedge, $h > 0$ is automatic. In the common scalar coordinates, $G(s) = c_B y_i$ includes the payoff factor c_B; in Example 1, $c_B = 1$, so it is also the action probability. The Tsallis regularizer (23) is a positive rescaling of negative Tsallis entropy, up to an additive constant [51].

Local Estimates and the Hedge Threshold. We use (s, r) from Equations 22 and 25, with the reference in Equation 27. For Hedge, states and deviations are vectors, while the common reference coordinates are scalars; for Tsallis, $k = 1$. The local estimates below apply whenever

$$\begin{cases} \max_i |s_i - \bar{s}| + \max_i |r_i - \bar{r}| \to 0, & \text{Hedge,} \\ |s/\bar{s} - 1| + |r - \bar{r}| \to 0, & \text{Tsallis.} \end{cases} \tag{42}$$

Lemma 1 establishes this for every initialization when $\nu > 0$; Lemma 2 establishes it whenever a best response is eventually fixed.

Coordinate scales and payoff equations. We use h, ν, γ from Equation 26, with $h > 0$ and $\gamma = 0$ for Hedge. We define the new time variable and coordinate scales:

$$\tau(t) \doteq \frac{2}{h} t^{h/2}, \qquad A_s(t) \doteq t^{\gamma - \nu h/2}, \qquad A_r(t) \doteq t^{-(\nu+1)h/2}. \tag{43}$$

For the switching argument, we use deviations from the original reference: $\widetilde{s} \doteq s - \bar{s}\mathbf{1}$ and $\widetilde{r} \doteq r - \bar{r}\mathbf{1}$, with $\widehat{s} \doteq \widetilde{s}/A_s$ and $\widehat{r} \doteq \widetilde{r}/A_r$. Primes on these rescaled trajectories denote derivatives with respect to τ; response derivatives such as $G'(s)$ remain derivatives with respect to their arguments. The scales obey $A_s A_r = t^{-(\alpha_A+\alpha_B)}$, $A_s/t^\gamma \asymp \tau^{-\nu}$, and $A_r \asymp \tau^{-\nu-1}$. We set

$$\begin{aligned} &C \doteq c_A(\bar{q} I_k - \bar{q}^2 \mathbf{1}\mathbf{1}^\top), \quad d \doteq \alpha_B r_\star && \text{for Hedge,} \\ &C \doteq F'(r_\star), && \text{for Tsallis.} \end{aligned} \qquad d \doteq \beta(\alpha_B r_\star)^{1+1/\beta}$$

Here I_k is the $k \times k$ identity matrix. The Hedge matrix has positive eigenvalues $c_A\bar{q}(1 - k\bar{q})$ and $c_A\bar{q}$; in both cases $d > 0$. We reserve C for this fixed coefficient and use C_0 for constants in bounds. Along a trajectory satisfying Equation 42, first-order expansion of the exact equations gives

$$\widehat{s}' = C\widehat{r} + o(\|\widehat{s}\| + \|\widehat{r}\|), \qquad \widehat{r}' = -d\widehat{s} + o(\|\widehat{s}\| + \|\widehat{r}\|) + O(\tau^{\nu-2}(1+\log\tau))\mathbf{1}. \tag{44}$$

For Hedge, the relative errors are bounded by a constant times $\tau^{-1} + \|s - \bar{s}\mathbf{1}\| + \|r - \bar{r}\mathbf{1}\| + t^{\alpha_B - 1} + |\bar{r} - r_\star|$. For Tsallis, Equation 28 and $s/\bar{s} \to 1$ give the same vanishing relative errors, with $|s - \bar{s}|/\bar{s}$ in place of $\|s - \bar{s}\mathbf{1}\|$. At $s = \bar{s}\mathbf{1}$, the scaled Hedge Jacobian in the second equation is $-\alpha_B r_\star I_k + O(t^{\alpha_B - 1})\mathbf{1}\mathbf{1}^\top$; for Tsallis, $t^{1+\gamma-\alpha_B} G'(\bar{s}) \to -d$. The additive error in Equation 44 is $\rho_r/(A_r\dot{\tau})$.

For $x \doteq (c_A(\bar{q} - q_i))_{i=1}^k$, or $x \doteq F(r)$ for Tsallis, Taylor expansion gives

$$x = F(\bar{r})\mathbf{1} + C(r - \bar{r}\mathbf{1}) + o(\|r - \bar{r}\mathbf{1}\|), \qquad \frac{x}{A_r} = C\widehat{r} + o(\|\widehat{r}\|) + O(\tau^{\nu-1}(1+\log\tau))\mathbf{1}. \tag{45}$$

Eventually fixed best responses. We use two facts about the rescaled trajectories. We write Leb for Lebesgue volume; a map preserves volume if it preserves the volume of every measurable set.

Lemma 3. *Let T_j be injective volume-preserving maps on $\mathbb{R}^{2k}$. For almost every initial state (s, r), the sequence $T_j(s, r)$ does not converge to zero.*

Proof. For any bounded measurable subset E of $\{(s, r) : T_j(s, r) \to 0\}$ and any ball B_ε centered at zero, dominated convergence gives

$$\mathrm{Leb}(E) = \lim_{j\to\infty} \mathrm{Leb}\{(s, r) \in E : T_j(s, r) \in B_\varepsilon\} \le \mathrm{Leb}(B_\varepsilon).$$

We let $\varepsilon \to 0$ and then exhaust the initial space by bounded sets. □

Lemma 4. *Let $C \in \mathbb{R}^{k\times k}$ be symmetric positive definite and $d > 0$. Suppose $\widehat{s}' = C\widehat{r} + o(\|\widehat{s}\| + \|\widehat{r}\|) + o(1)$ and $\widehat{r}' = -d\widehat{s} + o(\|\widehat{s}\| + \|\widehat{r}\|) + o(1)$. If x, divided by a positive scale, equals $C\widehat{r} + o(\|\widehat{s}\| + \|\widehat{r}\|) + o(1)$, an eventually fixed best response implies $(\widehat{s}, \widehat{r}) \to 0$.*

Proof. Otherwise, we choose $\tau_j \to \infty$ with $\|(\widehat{s}(\tau_j), \widehat{r}(\tau_j))\| \ge c_0 > 0$. After taking a subsequence, the normalized initial states converge to a unit vector. On each fixed interval $[0, \tau_0]$, we divide the shifted path by its initial norm. Gronwall's inequality first bounds it and then gives uniform convergence to a nonzero solution of

$$\widehat{s}'_\infty = C\widehat{r}_\infty, \qquad \widehat{r}'_\infty = -d\widehat{s}_\infty. \tag{46}$$

Applying the same rescaling to x gives the limit $C\widehat{r}_\infty$. We distinguish the possible fixed best responses:

1. A_0 **stays best.** Every coordinate of $C\widehat{r}_\infty$ is nonnegative.
2. A_i **stays best for some** $i \ge 1$**.** Coordinate i of $C\widehat{r}_\infty$ is nonpositive, and every other coordinate minus coordinate i is nonnegative.

These inequalities hold for all future times in the limit: division by a positive initial norm leaves their signs unchanged, and limits preserve the non-strict inequalities.

Diagonalizing C, each coordinate and each coordinate difference is a finite sum of sinusoids of nonzero frequency. Such a function has mean zero and positive mean square unless it vanishes identically. A bounded nonnegative function x satisfies $x^2 \le \|x\|_\infty x$, so these two averages exclude a fixed sign unless $x = 0$. The best-response inequalities therefore force every coordinate of $C\widehat{r}_\infty$ to be zero. Equation 46 then gives $\widehat{s}_\infty = \widehat{r}_\infty = 0$, a contradiction. □

The divergence of Equation 22, or of the system (25), is $-k(\alpha_A + \alpha_B)/t$—with $k = 1$ for Tsallis. The *flow* φ_t maps each initial state at time one to its state at time t. Its Jacobian $D\varphi_t$, the derivative with respect to the initial state, satisfies

$$\det D\varphi_t = t^{-k(\alpha_A+\alpha_B)} = (A_s A_r)^k. \tag{47}$$

Translating by a reference common to all initializations and dividing by A_s, A_r therefore preserves volume. By Lemma 3, applied at integer times, the rescaled trajectory does not tend to zero for almost every initial state.

An energy bound for stability. For stability, we replace the first reference coordinate by $\bar{s}_\mathrm{c}$ and retain the function $\bar{r}(t)$:

$$\bar{s}_\mathrm{c} \doteq G^{-1}\left[t^{\alpha_B}\left(\dot{\bar{r}} + \frac{\alpha_B}{t}\bar{r}\right)\right]. \tag{48}$$

The argument of G^{-1} differs from $G(\bar{s})$ by $-t^{\alpha_B}\rho_r$. Using $(G^{-1})'$, which is of order $t^{1-\alpha_B+\gamma}$ at the reference argument, and differentiating with respect to t gives

$$\bar{s}_\mathrm{c} - \bar{s} = O(t^{\gamma-h}(1 + \log t)), \qquad \rho_s \doteq -\dot{\bar{s}}_\mathrm{c} + \dot{\bar{s}} - \frac{\alpha_A}{t}(\bar{s}_\mathrm{c} - \bar{s}) = O(t^{\gamma-h-1}(1 + \log t)). \tag{49}$$

For Hedge these estimates follow from $G^{-1}(c_B y) = \log((1 - ky)/y)$; for Tsallis they follow from Equation 24 and the reference derivatives. The corrected reference satisfies the equation for r in (22) exactly; the same holds in (25) for Tsallis.

We set $\widetilde{s} \doteq s - \bar{s}_c \mathbf{1}$, $\widetilde{r} \doteq r - \bar{r}\mathbf{1}$, $\widehat{s} \doteq \widetilde{s}/A_s$, and $\widehat{r} \doteq \widetilde{r}/A_r$. For Hedge, we use the convex potentials (31), with $\bar{s}$ replaced by $\bar{s}_c$ in $\mathcal{W}_t$. For Tsallis, we use Equation 32 with the same replacement. The local energy is the function

$$\mathcal{E}(\tau, \widehat{s}, \widehat{r}) \doteq \frac{t^{\alpha_B}}{\dot{\tau}} \mathcal{V}_t(A_r \widehat{r}) + \frac{t^{\alpha_A}}{\dot{\tau}} \mathcal{W}_t(A_s \widehat{s}) - \frac{1 + \alpha_A + \gamma - 2\alpha_B}{2h\tau} \widehat{s}^\top \widehat{r}. \tag{50}$$

Here t is determined by $\tau(t)$ in Equation 43. With state derivatives taken at fixed τ, the exact equations are

$$\widehat{s}' = \nabla_{\widehat{r}} \mathcal{E} + \widehat{\rho}_s \mathbf{1}, \qquad \widehat{r}' = -\nabla_{\widehat{s}} \mathcal{E}, \qquad \widehat{\rho}_s \doteq \frac{\rho_s}{A_s \dot{\tau}} = O(\tau^{\nu-3}(1 + \log \tau)). \tag{51}$$

Along approaching trajectories, Taylor expansion yields

$$\mathcal{E} = \tfrac{1}{2} \widehat{r}^\top C \widehat{r} + \tfrac{d}{2} \|\widehat{s}\|^2 + o(\|\widehat{s}\|^2 + \|\widehat{r}\|^2), \qquad \|\nabla \mathcal{E}\| \le C_0 \sqrt{\mathcal{E}}.$$

Holding $\widehat{s}, \widehat{r}$ fixed, we bound the explicit time derivative by

$$|\partial_\tau \mathcal{E}| \le \frac{o(1)}{\tau} \mathcal{E}. \tag{52}$$

At zero, the Hessian with respect to $\widehat{r}$ tends to C, and the Hessian with respect to $\widehat{s}$ tends to dI_k. Their derivatives with respect to τ, multiplied by τ, tend to zero by differentiating the explicit references (27) and (48) and using Equation 28. The higher terms and their explicit time derivatives are bounded relative to the quadratic form by $\|\widetilde{s}\| + \|\widetilde{r}\|$ for Hedge and $|\widetilde{s}|/\bar{s} + |\widetilde{r}|$ for Tsallis, with an additional factor τ^{-1} for the derivatives. These quantities tend to zero by Equations 42 and 49. The derivative of the cross term in (50) is $O(\tau^{-2})(\|\widehat{s}\|^2 + \|\widehat{r}\|^2)$.

For $\sqrt{\mathcal{E}}$, Equations 51–52 give $\sqrt{\mathcal{E}}' \le o(1)\sqrt{\mathcal{E}}/\tau + C_0 |\widehat{\rho}_s|$. At zeros we derive the bound for $\sqrt{\mathcal{E} + \varepsilon}$ and let $\varepsilon \to 0$. If $\nu > 1$, then eventually

$$\tau \left(\frac{\sqrt{\mathcal{E}}}{\tau^{\nu-1}} \right)' \le -\frac{\nu - 1}{2} \frac{\sqrt{\mathcal{E}}}{\tau^{\nu-1}} + C_0 \frac{1 + \log \tau}{\tau}. \tag{53}$$

An integrating factor gives $\sqrt{\mathcal{E}} = o(\tau^{\nu-1})$. Since $A_r \tau^{\nu-1} \asymp t^{-h}$, we have proved

$$\|r - \bar{r}\mathbf{1}\| = o(t^{-h}) \qquad (\nu > 1). \tag{54}$$

We apply the estimates to the Hedge game and exponents specified before Theorem 3.

Theorem 3. *Every trajectory converges to $y_0 = 1$ and $q_i = \bar{q}$ for $1 \le i \le k$. The best-response threshold is independent of k:*

(i) If $5\alpha_A + 2\alpha_B > 3$, then A_0 is eventually the unique current best response.
(ii) If $5\alpha_A + 2\alpha_B < 3$, the best response switches infinitely often for almost every initialization.

Proof. Lemma 1 gives convergence to the stated strategy profile for every finite initialization. We consider each strict inequality separately:

1. $5\alpha_A + 2\alpha_B > 3$**.** We have $\nu > 1$. The same lemma gives Equation 42, so the stability estimate (54) applies. The reference satisfies $F(\bar{r}) \sim \alpha_A(1 - \alpha_B) t^{-h} \log t > 0$ by Equation 29. Thus $F(\bar{r})$ dominates the error in Equation 45, proving Theorem 3(i).

2. $5\alpha_A + 2\alpha_B < 3$**.** We have $\nu < 1$. We show that an eventually fixed best response is impossible for almost every initialization. Lemma 2 gives approach to the original reference, so $\widehat{s} = (s - \bar{s}\mathbf{1})/A_s$ and $\widehat{r} = (r - \bar{r}\mathbf{1})/A_r$ satisfy Equations 44 and 45. Their additive terms tend to zero whenever $\nu < 1$. Lemma 4 therefore forces $(\widehat{s}, \widehat{r}) \to 0$. Equation 47 and Lemma 3 exclude this behavior for almost every initialization, proving Theorem 3(ii).

□

Additional actions that are strictly worse. If $u_n(a', \mathbf{p}_t^{-n}) - u_n(a, \mathbf{p}_t^{-n}) \ge \varepsilon > 0$ for a fixed action a' in a neighborhood eventually containing the path, then $U_t^n(a') - U_t^n(a) \ge \varepsilon t + O(1)$. With learning rate $t^{-\alpha}$, $\alpha \in (0, 1)$, the exact Hedge formula (18) gives

$$\Pr_t(a) = O\left(e^{-\varepsilon t^{1-\alpha}/2} \right).$$

Thus these probabilities are smaller than every power of t.

Scope and coordinate conventions. Adding a common constant to each player's initial utilities leaves their differences unchanged, so Fubini's theorem transfers the statements for almost every initialization from initial differences to initial utility vectors. Example 1 agrees with the $k = 1$ game of Definition 2 up to payoff constants that leave the payoff differences unchanged. Theorem 3 establishes strategy convergence at equality $5\alpha_A + 2\alpha_B = 3$, but makes no best-response claim there.

Payoff Convergence in the 1-Boundary Game. The game, algorithm, exponent, and initial utilities are as specified before Proposition 3. We use $g_A(t), g_B(t)$ from Equation 3 and $\mathrm{Gap}(t)$ from Equation 15. We use s, r and their reference $\bar{s}, \bar{r}$ from Appendix D, and center at that reference: $\widetilde{s} \doteq s - \bar{s}$ and $\widetilde{r} \doteq r - \bar{r}$. Here $h = 1 - \alpha$, $C = 2/3$, $d = \alpha \log 2$, and δ_α is defined before Proposition 3.

Proposition 3. *If $\alpha > 3/7$, then $\mathrm{Gap}(t) \asymp t^{-(1-\alpha)}$. For $0 < \alpha < 3/7$, $\mathrm{Gap}(t) = O(t^{-\delta_\alpha})$ for all finite initializations, and almost every initialization satisfies*

$$0 < \limsup_{t\to\infty} t^{\delta_\alpha}\,\mathrm{Gap}(t) < \infty. \tag{16}$$

Proof. To obtain the sharp payoff rates, we must distinguish $\alpha = 1/5$, where the oscillations themselves change. For $1/5 < \alpha < 3/7$, $\widetilde{s}$ tends to zero, so a quadratic approximation of the energy is valid. At and below $1/5$, $\widetilde{s}$ need not tend to zero, so we keep the exact nonlinear response. We prove matching upper and lower bounds in both ranges; the change in the sharp exponent is therefore not an artefact of the analysis.

The normalization satisfies $\mathrm{Gap}(t) \le g_A(t) + g_B(t) \le 3\,\mathrm{Gap}(t)$. We have

$$g_A(t) = \begin{cases} y_t x_t, & x_t \ge 0, \\ (1 - y_t)(-x_t), & x_t < 0, \end{cases} \qquad g_B(t) = q_t y_t.$$

Thus $g_A = y_t x_t + (-x_t)_+$ and $g_B = q_t y_t$, which prove Equation 15 using $x_t = 1 - 3q_t$.

Claim. For each fixed $0 < \alpha < 3/7$, every finite initialization satisfies $\mathrm{Gap}(t) = O(t^{-\delta_\alpha})$, where

$$\delta_\alpha = \begin{cases} (1+\alpha)/3, & 0 < \alpha \le 1/5, \\ (1+3\alpha)/4, & 1/5 < \alpha < 3/7. \end{cases}$$

For almost every initialization, this rate is sharp in the sense of Equation 16. The two exponents agree at $\alpha = 1/5$, where both equal $2/5$.

We first settle the stable case, then prove the claim above and at or below $1/5$.

1. $\alpha > 3/7$**.** Here $\nu > 1$, so Lemma 1 gives $y_t \sim dt^{-(1-\alpha)}$. Theorem 3 makes $x_t > 0$ eventually, so

$$\mathrm{Gap}(t) = \frac{y_t}{3} \sim \frac{\alpha \log 2}{3} t^{-(1-\alpha)}. \tag{55}$$

2. $1/5 < \alpha < 3/7$**. Bounds on deviations from the reference.** Now $0 < \nu = (5\alpha - 1)/(2(1 - \alpha)) < 1$. We take $\widehat{s} = \widetilde{s}/A_s$, $\widehat{r} = \widetilde{r}/A_r$, and τ from Equation 43, so $A_r = t^{-(1+3\alpha)/4} = t^{-\delta_\alpha}$. Lemma 1 gives $\widetilde{s}, \widetilde{r} \to 0$ and $y_t \sim dt^{-(1-\alpha)}$. We use the local energy from Equation 50 with the first reference coordinate set to $\bar{s}$. Subscripts $\widehat{s}, \widehat{r}$ denote its partial derivatives at fixed τ. The exact equations are

$$\widehat{s}' = \mathcal{E}_{\widehat{r}}, \qquad \widehat{r}' = -\mathcal{E}_{\widehat{s}} + \widehat{\rho}_r, \qquad \widehat{\rho}_r \doteq \frac{\rho_r}{A_r \dot{\tau}} = O(\tau^{\nu-2}(1 + \log \tau)). \tag{56}$$

The inequality $\nu < 1$ makes $\widehat{\rho}_r$ integrable. The quadratic form and its partial derivative with respect to τ satisfy

$$\mathcal{E} = \frac{d}{2}\widehat{s}^2 + \frac{C}{2}\widehat{r}^2 + o(\widehat{s}^2 + \widehat{r}^2), \qquad |\partial_\tau \mathcal{E}| \le C_0 \left(\tau^{-2} + \frac{|\widetilde{s}| + |\widetilde{r}|}{\tau}\right)(\widehat{s}^2 + \widehat{r}^2). \tag{57}$$

For the derivative bound, the coefficients of $\widehat{s}^2$ and $\widehat{r}^2$ differ from their limits by $O(\tau^{-2}(1+\log\tau))$, and their derivatives gain a factor τ^{-1}. The nonlinear terms and their derivatives have the additional factors $|\widetilde{s}|+|\widetilde{r}|$ and $(|\widetilde{s}|+|\widetilde{r}|)/\tau$, respectively, as in the verification of Equation 52; the cross term contributes $O(\tau^{-2})$. For $\sqrt{\mathcal{E}}$, exact cancellation in Equation 56 yields

$$|\sqrt{\mathcal{E}}'| \le C_0\left(\tau^{-2}+\frac{|\widetilde{s}|+|\widetilde{r}|}{\tau}\right)\sqrt{\mathcal{E}}+C_0|\widehat{\rho}_r|. \tag{58}$$

At zeros we derive the bound using $\sqrt{\mathcal{E}+\varepsilon}$ and let $\varepsilon\to 0$. Since $\widetilde{s},\widetilde{r}\to 0$, an integrating factor first gives $\sqrt{\mathcal{E}}=O(\tau^{\varepsilon})$ for every $\varepsilon>0$. We choose $\varepsilon<\nu$. The identities $A_s\asymp\tau^{-\nu}$ and $A_r\asymp\tau^{-\nu-1}$ make the coefficient of $\sqrt{\mathcal{E}}$ in Equation 58 integrable. A second integrating factor bounds $\sqrt{\mathcal{E}}$, and then $\int|\sqrt{\mathcal{E}}'|<\infty$. Hence $\sqrt{\mathcal{E}}$ has a finite limit. If the limit is zero, Equation 57 gives $(\widehat{s},\widehat{r})\to 0$. Equation 47 and Lemma 3 exclude this for almost every initialization. For those initializations,

$$0<c_0\le\sqrt{\widehat{s}(\tau)^2+\widehat{r}(\tau)^2}\le C_0<\infty \quad \text{for all sufficiently large } \tau. \tag{59}$$

A lower bound on $\mathrm{Gap}(t)$ over fixed intervals. Boundedness and Equation 44 give $\widehat{s}'=C\widehat{r}+o(1)$ and $\widehat{r}'=-d\widehat{s}+o(1)$. We take $\tau_0\doteq 2\pi/\sqrt{Cd}$. On each interval of length τ_0, Gronwall's inequality compares the path with the exact oscillator started at its left endpoint, with error tending to zero. The integral of $(-\widehat{r})_+$ over one period of that oscillator is

$$\frac{2}{\sqrt{Cd}}\sqrt{\widehat{r}(\tau)^2+(d/C)\widehat{s}(\tau)^2}.$$

It has a positive lower bound under Equation 59. Since the positive-part function is Lipschitz, there is a fixed $c_1>0$ such that

$$\int_{\tau}^{\tau+\tau_0}(-\widehat{r}(\tau'))_+\,\mathrm{d}\tau'\ge c_1 \quad \text{for every sufficiently large } \tau. \tag{60}$$

Also $\bar{r}-\log 2=o(A_r)$ and $y_t=o(A_r)$. Taylor expansion of $q=(1+e^r)^{-1}$, whose derivative at $r=\log 2$ is $-2/9$, therefore gives

$$\mathrm{Gap}(t)=t^{-\delta_\alpha}\left[\frac{2}{9}(-\widehat{r}(\tau(t)))_+ +o(1)\right]. \tag{61}$$

Boundedness proves $\mathrm{Gap}=O(t^{-\delta_\alpha})$ for every finite initialization. For almost every initialization, Equation 60 gives the positive, finite limit superior in Equation 16.

3. $0<\alpha\le 1/5$.
We keep the nonlinear response. Here $\delta_\alpha=(1+\alpha)/3$. We set

$$\kappa\doteq\frac{1-5\alpha}{3},\quad \widehat{s}\doteq t^{-\kappa}\widetilde{s},\quad \widehat{r}\doteq t^{\delta_\alpha}\widetilde{r},\quad \tau(t)\doteq t^{\delta_\alpha}/\delta_\alpha. \tag{62}$$

Thus $\kappa\ge 0$, $2\delta_\alpha+\kappa=h$, and $\kappa-\delta_\alpha=-2\alpha$. We use the exact scalar potentials $\mathcal{V}_t,\mathcal{W}_t$ from Equation 31 and write $\bar{y}\doteq dt^{-h}$. In particular,

$$\mathcal{W}_t(\widetilde{s})=\log\frac{1-\bar{y}}{1-y_t}+\bar{y}\widetilde{s},\qquad \mathcal{W}_{\widetilde{s}}=\bar{y}-y_t. \tag{63}$$

The global convergence in Lemma 1 gives $s\to\infty$ and $\widetilde{r}\to 0$, even though $\widetilde{s}$ need not tend to zero. The Hessian estimate in Equation 36 uses only these limits, so it still gives

$$\mathcal{V}_t(\widetilde{r})\sim\frac{C}{2}\widetilde{r}^2,\qquad \mathcal{W}_t(\widetilde{s})=dt^{-h}(\widetilde{s}+e^{-\widetilde{s}}-1)(1+o(1)).$$

We define

$$\mathcal{E}\doteq t^{2\delta_\alpha}(\mathcal{V}+\mathcal{W}),\qquad \widehat{\mathcal{E}}\doteq\mathcal{E}+(\alpha-\delta_\alpha)t^{-\kappa}\widetilde{s}\widetilde{r}. \tag{64}$$

The inequality $|\widetilde{s}|\le C_0[\sqrt{\widetilde{s}+e^{-\widetilde{s}}-1}+\widetilde{s}+e^{-\widetilde{s}}-1]$ implies

$$|t^{-\kappa}\widetilde{s}\widetilde{r}|\le C_0(t^{-h/2}+|\widetilde{r}|)\mathcal{E}=o(\mathcal{E}).$$

Hence $\widehat{\mathcal{E}}=(1+o(1))\mathcal{E}$ is eventually nonnegative and controls $\widehat{r}^2$ and $t^{-\kappa}(\widetilde{s}+e^{-\widetilde{s}}-1)$.

The energy has a finite limit. Differentiating Equation 64 with the exact centered equations (33), we obtain

$$t\dot{\widetilde{\mathcal{E}}} = \kappa t^{2\delta_\alpha}(\widetilde{s}\mathcal{W}_{\widetilde{s}} - \mathcal{W}) + \rho_{\mathcal{E}}(t), \tag{65}$$

where

$$\begin{aligned}\rho_{\mathcal{E}}(t) = t^{2\delta_\alpha}\big[\delta_\alpha(2\mathcal{V} - \widetilde{r}\mathcal{V}_{\widetilde{r}}) + t\dot{\widetilde{s}}\,\partial_{\bar{s}}\mathcal{W} + h\mathcal{W} + t\dot{\widetilde{r}}\,\partial_{\bar{r}}\mathcal{V} + t\rho_r\mathcal{V}_{\widetilde{r}}\big]\\ - \delta_\alpha(\alpha - \delta_\alpha)t^{-\kappa}\widetilde{s}\widetilde{r} + (\alpha - \delta_\alpha)t^{1-\kappa}\rho_r\widetilde{s}.\end{aligned}$$

We now bound these terms without approximating $\widetilde{s}$ by zero. Smoothness gives $|2\mathcal{V} - \widetilde{r}\mathcal{V}_{\widetilde{r}}| \le C_0|\widetilde{r}|\mathcal{V}$ and $|\partial_{\bar{r}}\mathcal{V}| \le C_0\mathcal{V}$. The Hessian of $\mathcal{W}$ is $G(s)(1-G(s))$; its derivative under a common shift of s and $\bar{s}$ is $(-1+2G(s))G(s)(1-G(s))$. Integrating along the segment between $\bar{s}$ and s and using $t\dot{\bar{s}} = h/(1-\bar{y})$ therefore gives

$$|t\dot{\bar{s}}\,\partial_{\bar{s}}\mathcal{W} + h\mathcal{W}| \le C_0(y_t + \bar{y})\mathcal{W}.$$

Together with $t\dot{\bar{r}} = O(t^{-h}(1+\log t))$ and $\rho_r = O(t^{-h-1}(1+\log t))$, these estimates yield

$$\begin{aligned}|\rho_{\mathcal{E}}(t)| \le C_0\big(|\widetilde{r}| + y_t + \bar{y} + t^{-h/2} + t^{-h}(1+\log t)\big)\widehat{\mathcal{E}}\\ + C_0\big(t^{-\kappa-\delta_\alpha} + t^{-h-\kappa/2}\big)(1+\log t)\sqrt{\widehat{\mathcal{E}}}.\end{aligned} \tag{66}$$

Convexity gives $\widetilde{s}\mathcal{W}_{\widetilde{s}} - \mathcal{W} \ge 0$. Equation 63 also gives

$$\widetilde{s}\mathcal{W}_{\widetilde{s}} - \mathcal{W} = -\widetilde{s}y_t + \log\frac{1-y_t}{1-\bar{y}} \le C_0\big(\bar{y} + (1+\log t)y_t\big).$$

Indeed, when $\widetilde{s} \ge 0$, we have $y_t \le \bar{y}$ and the expression is at most $C_0\bar{y}$; when $\widetilde{s} < 0$, it is at most $|\widetilde{s}|y_t$, with $|\widetilde{s}| \le \bar{s} = O(1+\log t)$ eventually because $s \to \infty$. Since $H_t = t^\alpha r_t$ and $r_t \to \log 2$, integration by parts gives

$$\int^\infty \frac{y_t}{t}\,\mathrm{d}t < \infty, \qquad \int^\infty t^{2\delta_\alpha - 1}(1+\log t)y_t\,\mathrm{d}t < \infty \quad (\kappa > 0). \tag{67}$$

For the second integral, the boundary term is $O(t^{-\kappa}(1+\log t))$, since $2\delta_\alpha - 1 + \alpha = -\kappa$; its derivative gives an integrable bound. Thus the first term in Equation 65 is integrable in $\mathrm{d}t/t$ when $\kappa > 0$, and it is zero when $\kappa = 0$.

Using $\sqrt{\widehat{\mathcal{E}}} \le 1 + \widehat{\mathcal{E}}$, Equations 66 and 67 leave only $|\widetilde{r}|\widehat{\mathcal{E}}$ without an integrable coefficient in $\mathrm{d}t/t$. Since $\widetilde{r} \to 0$, an integrating factor first gives $\widehat{\mathcal{E}} = O(t^\varepsilon)$ for every $\varepsilon > 0$. The bound $|\widetilde{r}| \le C_0 t^{-\delta_\alpha}\sqrt{\widehat{\mathcal{E}}}$ then makes $|\widetilde{r}|$ integrable in $\mathrm{d}t/t$ as well. A second integrating factor bounds $\widehat{\mathcal{E}}$; Equation 65 now gives $\int|\dot{\widehat{\mathcal{E}}}|\,\mathrm{d}t < \infty$. Hence

$$\widehat{\mathcal{E}}(t) \longrightarrow \mathcal{E}_\infty \in [0,\infty), \qquad |\widehat{s}| + |\widehat{r}| = O(1), \qquad y_t = O(t^{-2\delta_\alpha}). \tag{68}$$

For the probability bound, bounded energy gives $\widetilde{s} + e^{-\widetilde{s}} - 1 = O(t^\kappa)$, hence $e^{-\widetilde{s}} = O(t^\kappa)$ and $y_t \le 2\bar{y}e^{-\widetilde{s}} = O(t^{-2\delta_\alpha})$. If $\mathcal{E}_\infty = 0$, then $\widehat{r} \to 0$ and $|\widehat{s}| \le C_0(t^{-\kappa/2}\sqrt{\widehat{\mathcal{E}}} + \widehat{\mathcal{E}}) \to 0$. The coordinate scales have product $t^{\kappa-\delta_\alpha} = t^{-2\alpha}$, exactly canceling the flow determinant in Equation 47. Lemma 3 therefore gives $\mathcal{E}_\infty > 0$ for almost every initialization.

The negative values of $\widehat{r}$ persist over intervals. We fix an initialization with $\mathcal{E}_\infty > 0$. Primes now denote differentiation with respect to the time in Equation 62. The exact equations and boundedness give

$$\begin{aligned}\widehat{s}' &= t^{\delta_\alpha}[F(\bar{r} + t^{-\delta_\alpha}\widehat{r}) - F(\bar{r})] - (\kappa+\alpha)t^{-\delta_\alpha}\widehat{s} = C\widehat{r} + o(1),\\ \widehat{r}' &= -d + t^h y_t + o(1).\end{aligned} \tag{69}$$

We distinguish the endpoint from the strict inequality:

(a) $\alpha = 1/5$**.** Here $\kappa = 0$ and $\widehat{s} = \widetilde{s}$ is bounded. The equations converge uniformly on bounded sets to

$$\widehat{s}' = C\widehat{r}, \qquad \widehat{r}' = d(e^{-\widehat{s}} - 1),$$

whose conserved energy is $C\widehat{r}^2/2 + d(\widehat{s} + e^{-\widehat{s}} - 1)$. Every positive energy level is a closed curve around zero, because the potential is strictly convex and grows without bound in both directions. The period is finite and varies continuously with the energy on compact intervals

bounded away from zero. Equation 68 confines the path to such an interval. On each bounded time interval, Gronwall's inequality compares it with this limiting system. Choosing τ_0 larger than the largest period on the interval gives Equation 60: on a full period, $\int(-\widehat{r})_+ \,\mathrm{d}\tau$ equals the difference between the largest and smallest values of $\widehat{s}$, divided by C, and has a positive lower bound.

(b) $0 < \alpha < 1/5$**.** Here $\kappa > 0$. We choose $M \geq 1$ with $|\widehat{s}|, |\widehat{r}| \leq M$ eventually, and consider any interval $I = [\tau, \tau + \tau_0]$ of fixed length. Brackets $[\,\cdot\,]_I$ denote the value at $\tau + \tau_0$ minus its value at τ. The second equation in (69) gives

$$\int_I t^h y_t \,\mathrm{d}\tau' \leq 2M + d\tau_0 + o(1), \qquad \int_I t^{2\delta_\alpha} y_t \,\mathrm{d}\tau' = o(1),$$

where t is evaluated at τ'. The second relation follows from $2\delta_\alpha = h - \kappa$. We also have $\int_I |\widehat{s}| t^h y_t \,\mathrm{d}\tau' = o(1)$: for $\widetilde{s} \geq 0$, $\widetilde{s} t^h y_t \leq 2d\widetilde{s}e^{-\widetilde{s}} \leq 2d$, while for $\widetilde{s} < 0$, $|\widetilde{s}| = O(1 + \log t)$; we multiply these bounds by $t^{-\kappa}$ and use the first integral estimate. Expanding the logarithm in Equation 63 gives, uniformly on the interval,

$$\widehat{\mathcal{E}} = \frac{C}{2}\widehat{r}^2 + d\widehat{s} + t^{2\delta_\alpha} y_t + o(1).$$

We integrate this identity and the derivative of $\widehat{sr}$ from Equation 69:

$$\frac{C}{2}\int_I \widehat{r}^2 \,\mathrm{d}\tau' + d\int_I \widehat{s}\,\mathrm{d}\tau' = \mathcal{E}_\infty \tau_0 + o(1),$$
$$C\int_I \widehat{r}^2 \,\mathrm{d}\tau' - d\int_I \widehat{s}\,\mathrm{d}\tau' = [\widehat{sr}]_I + o(1).$$

Their sum yields $(3C/2)\int_I \widehat{r}^2 \,\mathrm{d}\tau' = \mathcal{E}_\infty \tau_0 + [\widehat{sr}]_I + o(1)$. Also $\int_I \widehat{r}\,\mathrm{d}\tau' = [\widehat{s}]_I/C + o(1)$ by the first equation in (69). For a sufficiently large fixed τ_0, the bounds $|[\widehat{sr}]_I| \leq 2M^2$ and $|[\widehat{s}]_I| \leq 2M$ therefore give

$$\int_I (-\widehat{r})_+ \,\mathrm{d}\tau' = \frac{1}{2}\left(\int_I |\widehat{r}|\,\mathrm{d}\tau' - \int_I \widehat{r}\,\mathrm{d}\tau'\right) \geq \frac{1}{2M}\int_I \widehat{r}^2 \,\mathrm{d}\tau' - \frac{M}{C} - o(1) > c_1 > 0.$$

This proves Equation 60 in this case too.

Finally, $y_t = O(t^{-2\delta_\alpha}) = o(t^{-\delta_\alpha})$ and $\bar{r} - \log 2 = O(t^{-h}(1 + \log t)) = o(t^{-\delta_\alpha})$. Taylor expansion of $q = (1 + e^r)^{-1}$ therefore gives Equation 61 with the time in Equation 62. Boundedness proves the upper bound for every finite initialization. The interval bound proves the positive, finite limit superior for almost every initialization.

The last two cases prove the claim. Together with the first case, they establish Proposition 3. □

Figure 3 illustrates the different oscillation shapes in the exact Hedge dynamics. At $\alpha = 0.1$, the plotted cycle has a long positive excursion and a sharp return; at $\alpha = 0.3$, it is more symmetric and closer to a sinusoid.

Proposition 5. *For the trajectories in Proposition 3,*

$$\frac{1}{T}\int_1^T \mathrm{Gap}(t)\,\mathrm{d}t \asymp \begin{cases} T^{-\delta_\alpha}, & 0 < \alpha < 3/7, \\ T^{-(1-\alpha)}, & 3/7 < \alpha < 1. \end{cases}$$

The first line holds for almost every initialization; the second holds for all finite initializations.

Proof. 1. $\alpha > 3/7$**.** Integrating Equation 55 gives $T^{-1}\int_1^T \mathrm{Gap}(t)\,\mathrm{d}t \sim (\log 2/3)T^{-(1-\alpha)}$.

2. $0 < \alpha < 3/7$**.** The pointwise bound gives the upper estimate. For almost every initialization, Equation 60 gives the lower estimate in the time variable used in the corresponding case. Both time changes have the form $\tau(t) = ct^\xi$ with $c, \xi > 0$. We partition $[\tau(T/2), \tau(T)]$ into intervals of length τ_0. There are order T^ξ complete intervals, and on them $t^{-\delta_\alpha} \asymp T^{-\delta_\alpha}$ and $\mathrm{d}t/\,\mathrm{d}\tau \asymp T^{1-\xi}$. Equations 60 and 61, with the uniform vanishing of the remainder at late times, give

$$\int_{T/2}^T \mathrm{Gap}(t)\,\mathrm{d}t \gtrsim T^{-\delta_\alpha} T^{1-\xi} T^\xi = T^{1-\delta_\alpha}.$$

This proves the lower bound and the stated rate.

□

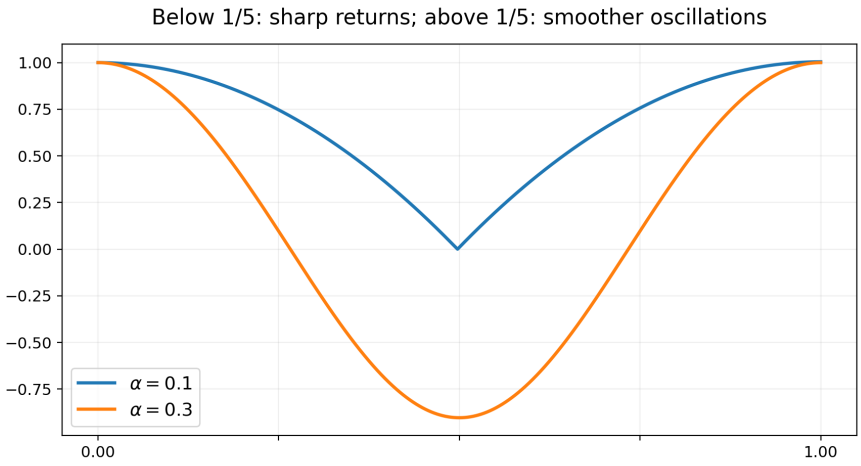


Figure 3: Late oscillation shapes for the exact Hedge dynamics with a common exponent $\alpha = 0.1$ and $\alpha = 0.3$. The horizontal axis is the fraction of one late cycle. The vertical axis is $\widetilde{s}(t) = s(t) - \bar{s}(t)$ divided by its positive peak. Each curve is normalized separately. Below $1/5$, the oscillation has a long positive excursion and a sharp return; above $1/5$, it is more symmetric and closer to a sinusoid.

Interpretation of the sharp rates. The matching upper and lower bounds show that the change at $\alpha = 1/5$ is not an artefact of the analysis. It concerns the payoff-decay exponent, whereas $3/7$ remains the sharp best-response threshold in Theorem 3. Equation 16 gives sharpness along recurring times, not a lower bound at every time. Proposition 3 makes no rate claim at $\alpha = 3/7$. The numerical curves illustrate the proof rather than establish its sharpness.

Euclidean FTRL. We consider Euclidean FTRL with two actions per player in Example 1, with $\alpha_A, \alpha_B \in (0, 1)$ and finite initial cumulative utilities at $t = 1$. We use the quadratic regularizer from Appendix A and the variables X_t, H_t from Section 3. We set $s(t) \doteq t^{-\alpha_A} X_t$ and $r(t) \doteq t^{-\alpha_B} H_t$.

Proposition 6. *Every trajectory satisfies $y_t \to 0$ and $q_t \to 1/3$. If $y_t \asymp t^{\alpha_B - 1} > 0$ eventually, then A_0 is eventually the unique current best response when*

$$5\alpha_A + 3\alpha_B > 4 \qquad \textit{and} \qquad 3\alpha_A + 5\alpha_B > 4.$$

Proof. **Convergence to the boundary profile.** The exact response, including its cutoff, is

$$G(s) \doteq \min\{1, \max\{0, (1 - s)/2\}\}, \qquad F(r) \doteq 1 - 3G(r).$$

It gives Equation 25 with $r_\star = 1/3$. For large t, we take $\bar{s} \doteq 1 - (2\alpha_B/3)t^{\alpha_B - 1}$ and $\bar{r} \doteq 1/3 + (2/3)t^{\alpha_A}(\dot{\bar{s}} + \alpha_A \bar{s}/t)$. Both reference responses are interior, and the residual is $\rho_r \doteq -\dot{\bar{r}} - \alpha_B(\bar{r} - 1/3)/t = O(t^{\alpha_A - 2})$. With $\widetilde{s} = s - \bar{s}$ and $\widetilde{r} = r - \bar{r}$, the convex potentials (32) give the exact centered system (33). Since F, G are continuous and piecewise linear, these potentials are continuously differentiable along the trajectory. For $\bar{r}$ near $1/3$, their explicit piecewise quadratic formulas give $|\mathcal{V}_{\widetilde{r}}| \le 3$ and $|\partial_{\bar{r}} \mathcal{V}| \le C_0 \mathcal{V}$. To bound $\partial_{\bar{s}} \mathcal{W}$, we distinguish two cases:

1. $s \ge -1$. We have $\partial_{\bar{s}} \mathcal{W} = -\widetilde{s}/2 + G(\bar{s}) - G(s) \le 0$.
2. $s < -1$. We have $\widetilde{s}\mathcal{W}_{\widetilde{s}} \ge |\widetilde{s}|/2$ eventually and $\partial_{\bar{s}} \mathcal{W} \le |\widetilde{s}|/2$.

Since $t\dot{\widetilde{s}} = O(t^{\alpha_B - 1}) \to 0$, we obtain Equation 34 in both cases. The same energy $\mathcal{E}_0$ as in Lemma 1 therefore satisfies

$$t\dot{\mathcal{E}}_0 \le -(\min\{\alpha_A, \alpha_B\} - o(1))\mathcal{E}_0 + C_0(t^{\alpha_A - 1} + t^{\alpha_B - 1}).$$

Thus $\mathcal{V}, \mathcal{W} \to 0$ and $r \to 1/3$. If $s \le 1 - \varepsilon$ along a sequence, then $\mathcal{W} \ge \mathcal{W}_t(1 - \varepsilon - \bar{s}) \to \int_{1-\varepsilon}^{1} G(s')\,\mathrm{d}s' > 0$, a contradiction. Hence $\liminf s \ge 1$, so $y_t \to 0$ and $q_t \to 1/3$.

We also distinguish which action could remain best:

1. A_1 **remains best.** Then $\frac{\mathrm{d}}{\mathrm{d}t}(t^{\alpha_A}s) \le 0$, contradicting $\liminf s \ge 1$.
2. A_0 **remains best.** Now $t^{\alpha_A}s$ is nondecreasing. If $s(t_j) \ge 1+\varepsilon$ along a sequence, we choose $1 < c < (1+\varepsilon)^{1/\alpha_A}$. Then $s > 1$ and $y = 0$ on $[t_j, ct_j]$, giving $r(ct_j) = c^{-\alpha_B} r(t_j) \to c^{-\alpha_B}/3 < 1/3$, contrary to $r \to 1/3$.

Thus an eventually fixed best response must be A_0 and also implies $s \to 1$.

Stability under the stated probability rate. We now assume $y_t \asymp t^{\alpha_B - 1} > 0$ eventually. The exact response gives $s = 1 - 2y_t \to 1$. Since $r \to 1/3$, both players eventually assign positive probability to both actions. Equation 19 now gives $y_t = (1-s)/2$, $q_t = (1-r)/2$, and

$$\dot{s} = -\frac{\alpha_A}{t}s + t^{-\alpha_A}\frac{3r-1}{2}, \qquad \dot{r} = -\frac{\alpha_B}{t}r + t^{-\alpha_B}\frac{1-s}{2}.$$

For the reference defined above,

$$\bar{r} - \frac{1}{3} = \frac{2\alpha_A}{3}t^{\alpha_A - 1} + O(t^{\alpha_A + \alpha_B - 2}). \tag{70}$$

The residual ρ_r also satisfies $\dot{\rho}_r = O(t^{\alpha_A - 3})$. We set $\bar{s}_{\mathrm{c}} \doteq \bar{s} + 2t^{\alpha_B}\rho_r$, $\widetilde{s} \doteq s - \bar{s}_{\mathrm{c}}$, and $\widetilde{r} \doteq r - \bar{r}$. This correction cancels the residual in the equation for $\widetilde{r}$, leaving

$$\dot{\widetilde{s}} = -\frac{\alpha_A}{t}\widetilde{s} + \frac{3}{2}t^{-\alpha_A}\widetilde{r} + \rho_s, \qquad \dot{\widetilde{r}} = -\frac{\alpha_B}{t}\widetilde{r} - \frac{1}{2}t^{-\alpha_B}\widetilde{s}, \tag{71}$$

where $\rho_s \doteq -\dot{\bar{s}}_{\mathrm{c}} + \dot{\bar{s}} - \alpha_A(\bar{s}_{\mathrm{c}} - \bar{s})/t = O(t^{\alpha_A + \alpha_B - 3})$. For this linear system, the coordinate scales and rescaled time are

$$A_s \doteq t^{-(3\alpha_A + \alpha_B)/4}, \qquad A_r \doteq t^{-(\alpha_A + 3\alpha_B)/4}, \qquad \tau \doteq \frac{2t^{1-(\alpha_A+\alpha_B)/2}}{2 - \alpha_A - \alpha_B}.$$

We set $\widehat{s} \doteq \widetilde{s}/A_s$ and $\widehat{r} \doteq \widetilde{r}/A_r$; primes on these trajectories denote τ derivatives. With $\kappa \doteq (\alpha_A - \alpha_B)/(2(2 - \alpha_A - \alpha_B))$ and $\widehat{\rho}_s \doteq \rho_s/(A_s\dot{\tau})$, Equation 71 becomes $\widehat{s}' = (3/2)\widehat{r} - \kappa\widehat{s}/\tau + \widehat{\rho}_s$, $\widehat{r}' = -\widehat{s}/2 + \kappa\widehat{r}/\tau$. The energy is the eventually positive quadratic form

$$\mathcal{E}(\tau, \widehat{s}, \widehat{r}) \doteq \frac{1}{4}\widehat{s}^2 + \frac{3}{4}\widehat{r}^2 - \frac{\kappa}{\tau}\widehat{s}\widehat{r} \quad \text{with} \quad \mathcal{E}' = \frac{\kappa}{\tau^2}\widehat{s}\widehat{r} + \widehat{\rho}_s\left(\frac{1}{2}\widehat{s} - \frac{\kappa}{\tau}\widehat{r}\right). \tag{72}$$

Here $\mathcal{E}'$ is the derivative along the trajectory. Since $\mathcal{E} \asymp \widehat{s}^2 + \widehat{r}^2$, an integrating factor for $\sqrt{\mathcal{E}}$ gives, for fixed sufficiently large T,

$$\sqrt{\widehat{s}^2 + \widehat{r}^2} \lesssim 1 + \int_T^t \frac{|\rho_s(t')|}{A_s(t')}\,\mathrm{d}t' \lesssim 1 + \int_T^t t'^{(7\alpha_A + 5\alpha_B - 12)/4}\,\mathrm{d}t'.$$

Multiplying by the scales and including $\bar{s}_{\mathrm{c}} - \bar{s} = O(t^{\alpha_A + \alpha_B - 2})$ yields

$$|s - \bar{s}| \lesssim (1 + \log t)\left[t^{-(3\alpha_A + \alpha_B)/4} + t^{\alpha_A + \alpha_B - 2}\right],$$
$$|r - \bar{r}| \lesssim (1 + \log t)\left[t^{-(\alpha_A + 3\alpha_B)/4} + t^{(3\alpha_A + \alpha_B - 4)/2}\right].$$

The first bound is $o(t^{\alpha_B - 1})$ under $3\alpha_A + 5\alpha_B > 4$. The second is $o(t^{\alpha_A - 1})$ under $5\alpha_A + 3\alpha_B > 4$, since $\alpha_A + \alpha_B < 2$. Equation 70 therefore gives $r - 1/3 \sim (2\alpha_A/3)t^{\alpha_A - 1} > 0$. Hence $q_t < 1/3$ eventually, and A_0 is uniquely best. □

Scope of the stability condition. For common exponents, the two inequalities reduce to $\alpha > 1/2$. This is a sufficient condition under the stated assumption $y_t \asymp t^{\alpha_B - 1} > 0$ eventually, not an unconditional sharp threshold. The convergence assertion in Proposition 6 does not need that additional assumption.

Tsallis FTRL. Both players use the regularizer (23) with $\beta > 1$ in Example 1. For Proposition 7, we take $\alpha_A, \alpha_B \in (0, 1)$, finite initial cumulative utilities, and

$$\gamma \doteq (1 - \alpha_B)/\beta, \qquad \alpha_A + \gamma < 1. \tag{73}$$

The last inequality is the fast-pruning condition for this response. We use s, r, F, G and the reference from Appendix D. The indifference point is $r_\star \doteq 3^{1/\beta} - (3/2)^{1/\beta}$, and h, ν are defined in Equation 26.

Proposition 7. *Every trajectory satisfies $y_t \to 0$ and $q_t \to 1/3$. Moreover:*

(i) If $5\alpha_A + 2\alpha_B + 5\gamma > 3$, then A_0 is eventually the unique current best response.
(ii) If $5\alpha_A + 2\alpha_B + 5\gamma < 3$, the best response switches infinitely often for almost every initialization.

Proof. Equation 73 gives $h > 0$, so Lemma 1 proves the strategy limits for every initialization. We consider each strict inequality separately:

1. $5\alpha_A + 2\alpha_B + 5\gamma > 3$**.** We have $\nu > 1$, and the same lemma gives $s/\bar{s} \to 1$ and $r - \bar{r} \to 0$. Equation 54 now yields $r - \bar{r} = o(t^{-h})$, while Equation 29 gives $\bar{r} - r_\star \asymp t^{-h} > 0$. Thus $F(r) > 0$ eventually, proving (i).

2. $5\alpha_A + 2\alpha_B + 5\gamma < 3$**.** We have $\nu < 1$. An eventually fixed best response would give Equation 42 by Lemma 2. Using the original reference in Equations 44 and 45, both additive terms tend to zero. Lemma 4 implies $(\widehat{s}, \widehat{r}) \to 0$. Equation 47 and Lemma 3 rule this out for almost every initialization, proving (ii).

□

Common exponents and the equality case. For common exponents, Equation 73 holds for every $\alpha \in (0, 1)$. The dividing value is $\alpha = (3\beta - 5)/(7\beta - 5)$: it equals $1/9$ at $\beta = 2$ and tends to $3/7$ as $\beta \to \infty$. For $1 < \beta \leq 5/3$, every $\alpha \in (0, 1)$ satisfies the stability condition. When the dividing value is in $(0, 1)$, Proposition 7 makes no best-response claim at equality.

E Additional Experimental Details

Initialization and observation window for the two-activity game. For the inspection experiments, we use $s_i(t) = t^{-\alpha}[U_t^A(A_0) - U_t^A(A_i)]$ and $r_i(t) = t^{-\alpha}[U_t^B(B_0) - U_t^B(B_i)]$, as in Appendix D. In the two-activity experiment, we start at $t = 1$ and draw $s_i(1), r_i(1)$ for $i = 1, 2$ independently from $\mathcal{N}(0, 0.5^2)$, the normal distribution with mean zero and standard deviation 0.5. We reuse the ten sampled vectors for $\alpha = 1/4$ and $\alpha = 1/2$ and simulate to $t = 10^8$. Late changes are changes in the firm's current best response observed during $[10^6, 10^8]$. Ties are assigned to the lowest-index action.

Figure 2 uses seed 3 for both exponents. Its summary shows 39 observed changes in the firm's best response for $\alpha = 1/4$, including 30 after 10^6 and a last observed change near 7.60×10^7. All three actions A_0, A_1, A_2 occur as best responses after 10^6. For $\alpha = 1/2$, we observe five transient changes, with the last near 1.09×10^5. Across the ten initializations, every $\alpha = 1/4$ run has late changes, whereas compliance remains best throughout the late window in every $\alpha = 1/2$ run.

Initialization with more activities. For each $k \in \{2, 3, 5\}$ and $\alpha \in \{1/4, 1/2\}$, we construct a symmetric starting state with low violation probabilities and nearly equal payoffs for the firm's actions. We sample twenty perturbations of (s, r), start at $t = 10^6$, and simulate to 10^9. At the starting time, each player's perturbation has an independently sampled direction uniform on the unit sphere, with radius $3t^{(1-5\alpha)/4}$ for A and $3t^{-(1+3\alpha)/4}$ for B. Late changes are changes in the firm's best response observed during $[10^8, 10^9]$. All twenty $\alpha = 1/4$ runs have such changes for each tested k; compliance remains best throughout this window in every $\alpha = 1/2$ run.

Numerical integration and observed changes. We integrate the Hedge equations for (s, r) in the time variable $\log t$ using an adaptive Runge–Kutta method [18]. For the two-activity runs from random initial utilities, the maximum step size in $\log t$ is 0.004, the relative tolerance is 10^{-9}, and the absolute tolerance is 10^{-11}. For runs from perturbations of the symmetric starting state, these settings are 0.002, 2×10^{-9}, and 10^{-11}. The payoff vectors are evaluated against current mixed strategies.

We detect best-response changes between consecutive observations on logarithmically spaced grids: $50{,}000$ points for random starts and $25{,}000$ for perturbed starts. Counts and times of the last observed best-response changes use these complete observation grids. Figure 2 displays every fifth observation, while the reported switching counts use the complete grid.

Scope of the inspection experiments. The two payoff differences in the inspection game reverse the signs in Equation 6, and $\bar{q} > 1/(k + 1)$. This is therefore an empirical extension, not a game

covered directly by Theorem 3. Counts on the observation grids need not capture changes between observations, and a finite observation window does not establish eventual stability or infinitely many changes.

Discrete-Time FTRL. We simulate the discrete-time counterpart of the Hedge update (18). At each integer round $t \geq 1$, both players compute their mixed strategies from $U_t^n(a)$ and then add their expected payoff vectors simultaneously:

$$\mathbf{p}_t^n(a) \doteq \frac{\exp(t^{-\alpha} U_t^n(a))}{\sum_{a' \in \mathcal{A}_n} \exp(t^{-\alpha} U_t^n(a'))}, \qquad U_{t+1}^n(a) \doteq U_t^n(a) + u_n(a, \mathbf{p}_t^{-n}). \tag{74}$$

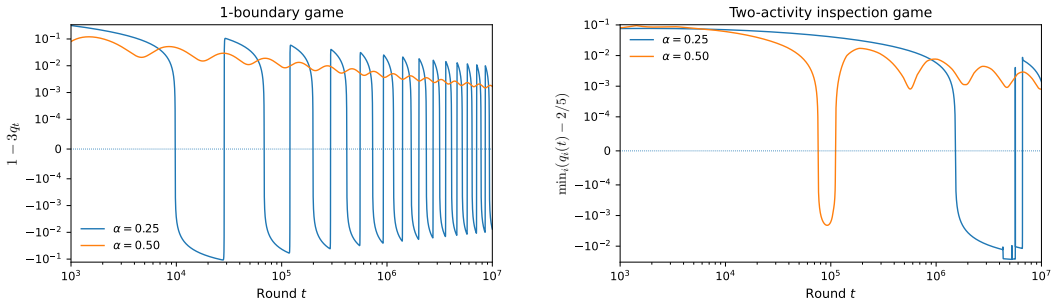
The payoff vector is evaluated against the opponent's mixed strategy. Each exponential uses $t^{-\alpha} U_t^n(a)$. This is the exponential-weights update with a time-varying learning rate studied in [27].

We use the 1-boundary game of Example 1 and the two-activity inspection game in Section 5. In each game, we independently draw ten vectors of initial utility differences from $\mathcal{N}(0, 0.5^2)$ and reuse them for $\alpha = 1/4$ and $\alpha = 1/2$. We run every integer round from 1 through 10^7 and count changes of player A's current best response during rounds $10^6, \ldots, 10^7$. Ties are assigned to the lowest-index action. All forty runs use the same *full-information update* (74): each player receives the expected payoff of every action against the current opponent mixture.

Table 2: Discrete-time FTRL: changes in player A's best response during rounds $10^6, \ldots, 10^7$. Each row contains ten initializations; the last three columns summarize their change counts. Counts use every integer round.

Game	α	Runs with changes	Minimum	Median	Maximum
1-boundary	$1/4$	$10/10$	17	29	51
1-boundary	$1/2$	$0/10$	0	0	0
Two-activity inspection	$1/4$	$10/10$	6	7	8
Two-activity inspection	$1/2$	$0/10$	0	0	0

In both games, player A's best response changes in the late window for every $\alpha = 1/4$ run, while A_0 remains best for every $\alpha = 1/2$ run. In Figure 4, we plot $1 - 3q_t$ and $\min_i(q_i(t) - 2/5)$ for one fixed initialization in each game. As in the inspection simulations (Figure 2), $\alpha = 1/4$ produces late best-response changes while $\alpha = 1/2$ keeps A_0 best throughout the observed late window.



(a) **(b)**

Figure 4: Discrete-time FTRL for seed 3, shown from round 10^3. **(a)** $1 - 3q_t$ in the 1-boundary game. **(b)** $\min_i(q_i(t) - 2/5)$ in the inspection game. Positive values make A_0 uniquely best; negative values make an alternative better. The horizontal axes are logarithmic. The vertical axes are linear within 10^{-4} of zero and logarithmic on either side. In both games, $\alpha = 1/4$ produces late sign changes, while $\alpha = 1/2$ keeps the plotted quantity positive throughout the late window.

Implementation. We update both players' $U_t^n(a)$ simultaneously in double precision and count every change in player A's current best response. Probabilities use a *shifted softmax*: we subtract $\max_{a'} t^{-\alpha} U_t^n(a')$ from each $t^{-\alpha} U_t^n(a)$ before exponentiating, leaving the probabilities unchanged while avoiding overflow. A separate implementation using the full payoff matrices agrees with the

implementation using $s_i(t), r_i(t)$ on the first 200 rounds to within 6×10^{-15}. Repeating all forty runs with *compensated summation*, which carries a correction for rounding error when adding payoffs, preserves every switch count and the last round in which a best-response change was observed; final probabilities differ by less than 7×10^{-12}. The trajectories in Figure 4 are sampled on a logarithmic grid. **Scope of the discrete-time evidence.** The counts cover every round in the stated window, but do not establish an asymptotic threshold for the update (74).