SCALE-FREE FAST CONVERGENCE IN GAMES

Anonymous authors
Paper under double-blind review

ABSTRACT

Scale-freeness in games, that is, invariance of learning dynamics under positive rescaling of utilities, has emerged as a desirable property. Yet, almost all fast convergence guarantees in learning in games require prior knowledge of the utility scale. To address this, we develop scale-free learning dynamics that achieve fast convergence without any prior information about utilities. For two-player zero-sum games, we obtain such dynamics with external regret bounded by $\widetilde{O}(A_{\mathrm{diff}})$, where A_{diff} is the payoff range, which implies an $\widetilde{O}(A_{\mathrm{diff}}/T)$ convergence rate to Nash equilibrium after T rounds. For multiplayer general-sum games, we obtain such dynamics with swap regret bounded by $O(U_{\max}\sqrt{\log T})$, where $U_{\max}$ is the utility scale and the dependence on the number of players and actions is ignored. This bound yields an $O(U_{\max}\sqrt{\log T}/T)$ convergence rate to correlated equilibrium. Our learning dynamics are based on optimistic follow-the-regularized-leader with adaptive learning rates, and our analyses exploit negative terms in regret bounds without scale-dependent tuning. Specifically, for two-player zero-sum games, we introduce a new analysis based on a stopping time, and for general-sum games, we combine a hybrid regularizer and clipped utilities based on past observations with a new analysis of the time-varying negative terms.

1 INTRODUCTION

Learning in games studies repeated strategic interactions in which each player adapts their strategy online to minimize regret (Freund & Schapire, 1999; Hart & Mas-Colell, 2000; Cesa-Bianchi & Lugosi, 2006). Such no-regret learning enables equilibrium computation. For example, in two-player zero-sum games, if each player achieves external regret Reg^T, then the average play after T rounds is an $O(\mathsf{Reg}^T/T)$-approximate Nash equilibrium (Freund & Schapire, 1999). This paradigm underlies practical successes ranging from superhuman game AI (Bowling et al., 2015; Moravčík et al., 2017; Perolat et al., 2022; FAIR et al., 2022) to recent methods for aligning LLMs (Munos et al., 2024; Swamy et al., 2024).

A standard assumption in learning in games is that the underlying utility functions are bounded and that the exact bound is known to the players. In two-player zero-sum games, this corresponds to knowing a scale parameter such as the maximum absolute magnitude $A_{\max} = \max_{i,j}|A_{i,j}|$ for a payoff matrix A. In multiplayer general-sum games, this corresponds to knowing $U_{\max}$ such that all utility values lie in $[-U_{\max}, U_{\max}]$. However, in many applications, it is difficult to know the scale of utilities in advance, making it desirable to design learning dynamics without such knowledge.

Motivated by the connection between no-regret online learning and equilibrium computation, a natural way to remove the assumption that the scale is known in advance is to adopt *scale-free* online learning (Cesa-Bianchi et al., 2007). For example, in two-player zero-sum games, using scale-free online learning, we can obtain an external regret bound of $O(A_{\max}\sqrt{T})$ without knowing $A_{\max}$, which implies an $O(A_{\max}/\sqrt{T})$ convergence rate to Nash equilibrium.

However, this result is not satisfactory, since one can obtain much faster convergence rates when the scale is known (Daskalakis et al., 2011). One representative approach to obtaining such fast rates is to use an optimistic online learning algorithm, such as optimistic follow-the-regularized-leader (OFTRL) or optimistic online mirror descent (Chiang et al., 2013; Rakhlin & Sridharan, 2013). In two-player zero-sum games, if $A_{\max}$ is known, the external regret can be bounded by $\widetilde{O}(A_{\max})$ (Rakhlin & Sridharan, 2013; Syrgkanis et al., 2015), which yields an $\widetilde{O}(A_{\max}/T)$ convergence rate to Nash equilibrium. In multiplayer general-sum games, if $U_{\max}$ is known, the

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.

Table 1: Comparison of individual regret upper bounds after T rounds. **Top:** external regret of the x-player in two-player zero-sum games with a payoff matrix $A \in [-A_{\max}, A_{\max}]^{m_x \times m_y}$, where $m = \max\{m_x, m_y\}$. The parameter $\delta > 0$ is an instance-dependent constant, which can be arbitrarily small (in the bounds below, $\delta = \min\{\|g^1\|, \|\ell^1\|\}$; see Appendix D.1). The value $A_{\mathrm{diff}} = \max_{i,j} A_{i,j} - \min_{i,j} A_{i,j} \le 2A_{\max}$ is the range of the entries of A, which can be much smaller than $A_{\max}$. **Bottom:** swap regret of player i in multiplayer general-sum games with n players and m actions. The value $U_{\max}$ is the scale of the underlying utility function. The bound of "Scale-free OLO" is achieved by simply using scale-free online linear optimization (OLO) algorithms (see Appendix C.4 for details). "$\sqrt{\text{Corrupt}}$?" indicates whether the effect of opponent deviations and corrupted observed utilities can be kept to a square-root dependence (see Appendix B.3).

Reference	Regret bound	$\sqrt{\text{Corrupt}}$?	Scale-free & Scale-agnostic?
▶ **Two-player zero-sum games** (external regret)			
Syrgkanis et al. (2015)	$A_{\max} \log m$	No	No
Tsuchiya et al. (2025)	$A_{\max} \log m$	Yes	No
Scale-free OLO	$A_{\max}\sqrt{T \log m}$	Yes	**Yes**
Tsuchiya et al. (2025)	$A_{\max} \log m/\delta$	Yes	**Yes**
Zhang et al. (2025)	$A_{\max} m^{3/2}/\delta$	Yes	**Yes**
This work (Theorem 1)	$A_{\mathrm{diff}} \log m$	Yes	**Yes**
▶ **Multiplayer general-sum games** (swap regret)			
Chen & Peng (2020)	$U_{\max}\sqrt{n}(m \log m)^{3/4} T^{1/4}$	No	No
Anagnostides et al. (2022b)	$U_{\max} n m^{5/2} \log T$	No	No
Tsuchiya et al. (2025)	$U_{\max} n m^{5/2} \log T$	Yes	No
Tsuchiya (2026b)	$U_{\max} n m^2 \sqrt{\log m \log T}$	No	No
Scale-free OLO (Proposition 2)	$U_{\max}\sqrt{mT \log m}$	Yes	**Yes**
This work (Theorem 2)	$U_{\max} n m^2 \sqrt{\log m \log T}$	No	**Yes**

swap regret is bounded by $O(U_{\max} \log T)$ (Anagnostides et al., 2022b), and more recently by $O(U_{\max}\sqrt{\log T})$ (Tsuchiya, 2026b), where we ignore the dependence on the number of players and actions. The latter bound yields an $O(U_{\max}\sqrt{\log T}/T)$ convergence rate to correlated equilibrium.

These observations lead to a natural question: *can we construct learning dynamics that are invariant under positive rescaling of utilities and agnostic to their scale while still achieving fast convergence?* In particular, can such learning dynamics match, up to constant factors, the best-known fast rates achieved when the scale is known in advance?

Being scale-agnostic and being scale-free are closely related. Informally, learning dynamics are scale-free if rescaling the utilities by any positive constant does not change the sequence of strategies (see Section 2.4 for formal definitions). Scale-freeness is widely regarded as a practically useful property in learning in games; for instance, regret matching, one of the most powerful methods for solving games, is scale-free (Hart & Mas-Colell, 2000). Several studies suggest that such invariance properties may be crucial for strong empirical performance for zero-sum games (Brown & Sandholm, 2019; Grand-Clément & Kroer, 2021; Chakrabarti et al., 2024; Zhang et al., 2025). Accordingly, our goal is to construct scale-free learning dynamics with fast convergence.

1.1 Contributions of this paper

We affirmatively answer the above question by constructing learning dynamics with scale-free and scale-agnostic fast convergence in two-player zero-sum games and multiplayer general-sum games.

Two-player zero-sum games. We start with two-player zero-sum games with a payoff matrix A and provide scale-agnostic and scale-free learning dynamics under which the external regrets of the x- and y-players are $O(A_{\mathrm{diff}} \log m)$ (Theorem 1), where A_{diff} is the range of the entries of A and m is the maximum number of actions among the players. Consequently, the average play after T rounds is an $O(A_{\mathrm{diff}} \log m/T)$-approximate Nash equilibrium. This learning dynamic is robust against opponent deviations and adversarial corruption of observed utilities (see Appendix B.3 for details). This result affirmatively answers the question raised in Zhang et al. (2025) of whether one can retain the scale-free fast convergence while removing the dependence on the $1/\delta$ factor in their regret bound, where $\delta > 0$ can be arbitrarily small (see Appendix D.1 for detailed discussion).

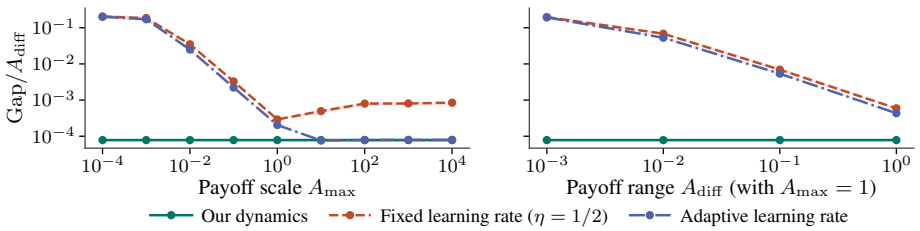

Figure 1: Nash gap after $T = 10^4$ rounds divided by A_{diff} (medians and 25th and 75th percentiles over 30 random games). Optimistic Hedge uses learning rates set for $A_{\max} = 1$ (see Appendix E).

Moreover, our regret bound and convergence rates depend on the payoff range A_{diff}, which can be much smaller than the maximum absolute magnitude $A_{\max}$. See Table 1 for comparison with existing bounds. Figure 1 shows that, in numerical experiments, the Nash gap of our dynamics divided by A_{diff} does not change with $A_{\max}$ or A_{diff}, whereas optimistic Hedge with learning rates set for $A_{\max} = 1$ converges more slowly when $A_{\max}$ or A_{diff} is small (see Appendix E for details).

The learning dynamic that achieves the above convergence rate is based on optimistic Hedge with an adaptive learning rate. Fast convergence of such optimistic algorithms relies on a negative term in their regret upper bounds, and existing analyses exploit this term by bounding the learning rate by a constant that depends on the scale. For example, many existing works use a constant learning rate (Syrgkanis et al., 2015; Anagnostides et al., 2022b). Adaptive learning rates were introduced to learning in games by Rakhlin & Sridharan (2013), and those of Tsuchiya et al. (2025) yield a regret bound independent of T, but they are also upper bounded by a constant that requires prior knowledge of $A_{\max}$. Another closely related work is Hsieh et al. (2021), which employs adaptive learning rates in continuous games, but their learning rate involves an offset parameter $\tau_i > 0$ fixed in advance, and hence their dynamics are not scale-free.

To overcome these issues, we introduce two refinements in the design and analysis of optimistic Hedge. The first is to incorporate the squared path length of the gradients observed by the opponent into the learning rate. The second is a new analysis based on a stopping time that tracks the growth of the squared path length of the gradients. This analysis exploits the negative term without bounding the learning rate by a constant, and thus yields a fast convergence rate without knowing the scale. Note that the A_{diff}-dependent bound comes from centering the gradients at the first round and from an AdaHedge-type analysis (Orabona & Pál, 2018), which depends on the gradients only through their increments. We also provide a scale-agnostic learning dynamic with fast convergence that requires no communication of gradients in Appendix B.5.

Multiplayer general-sum games. We also establish learning dynamics for multiplayer general-sum games. To obtain a *correlated equilibrium* (Aumann, 1974; Foster & Vohra, 1997; Hart & Mas-Colell, 2000) in this setting, we focus on minimizing the *swap regret*. For a known horizon T, we provide scale-agnostic ($U_{\max}$-agnostic) and scale-free learning dynamics under which the swap regret of every player is $O(U_{\max} n m^2 \sqrt{\log m \log T})$ (Theorem 2), where n is the number of players and m is the maximum number of actions among the players. Consequently, the time-averaged joint distribution is an $O(U_{\max} n m^2 \sqrt{\log m \log T / T})$-approximate correlated equilibrium. To our knowledge, ours is the first scale-agnostic and scale-free learning dynamic with a fast swap regret bound, and hence with fast convergence to correlated equilibrium, in multiplayer general-sum games. A comparison with existing swap regret bounds is summarized in Table 1, and a detailed discussion follows Theorem 2.

We employ the well-known reduction of Blum & Mansour (2007), which converts swap regret minimization into several instances of external regret minimization. We instantiate each external regret minimizer with OFTRL using the hybrid regularizer of Tsuchiya (2026b), which combines the negative Shannon entropy and the log barrier. As in the two-player zero-sum case, the adaptive learning rates for swap regret minimization in Tsuchiya et al. (2025) allow us to attain the fast rate, but their tuning requires prior knowledge of $U_{\max}$. To handle the unknown scale, we use clipping based on past observations (Cutkosky, 2019, Section 4) and an adaptive learning rate, which depends on the observed utility norms and the maximum squared path length of the clipped utility vectors.

Compared with the two-player case, there are two additional difficulties. First, the negative terms in the regret bounds of OFTRL control changes in the output distributions of the external regret

minimizers, whereas the players' strategies are the stationary distributions of the Markov chains formed from these distributions. We relate the two by the results of Tsuchiya (2026b, Theorem 3 and Lemma 3), which bound changes in the stationary distribution by the Bregman divergences induced by the log-barrier regularizer, and this makes our algorithm and analysis significantly simpler than those based on the analysis of Anagnostides et al. (2022b). Second, the utility vector of each player changes with the strategies of all other players, so these changes cannot be absorbed by the negative term of a single opponent, and we absorb them jointly by a common learning rate. Moreover, as in the two-player case, the observed scale can stay far below $U_{\max}$, whereas the changes in the utility vectors caused by the players' strategy changes scale with $U_{\max}$. To handle this, we keep the time-varying coefficients of the negative terms at all rounds and use an integral inequality to bound the maximum squared path length without knowing $U_{\max}$. See Section 4 for details. Additional related work is discussed in Appendix D.

2 Preliminaries

Notation. For any natural number $n \in \mathbb{N}$, we write $[n] = \{1, \dots, n\}$. Let $\mathbf{0}$ and $\mathbf{1}$ be the all-zero and all-one vectors, respectively. For a vector x, we write $x(a)$ for its a-th entry and $\|x\|_p$ for its ℓ_p-norm, where $p \in [1, \infty]$. For a matrix A, we write $A(k, \cdot)$ for its k-th row and $A_{i,j}$ for its (i, j) element. We use $\Delta(\mathcal{K})$ for the set of probability distributions over $\mathcal{K}$, and $\Delta_d = \{x \in [0, 1]^d : \|x\|_1 = 1\}$ for the $(d-1)$-dimensional probability simplex. For brevity, we may use $f \lesssim g$ to denote $f = O(g)$. For a sequence $z = (z^1, \dots, z^T)$ and $q \in [1, \infty]$, we define the squared path-length values in terms of the ℓ_q-norm up to round $t-1$ and from round t as $P_q^t(z) = \sum_{s=1}^{t-1} \|z^s - z^{s-1}\|_q^2$ and $Q_q^t(z) = \sum_{s=t}^{T} \|z^s - z^{s-1}\|_q^2$, respectively, where we omit the dependence on T and let $z^0 = \mathbf{0}$ for simplicity. Note that $\sum_{s=1}^{T} \|z^s - z^{s-1}\|_q^2 = P_q^t(z) + Q_q^t(z)$. Throughout this paper, we frequently use i to index players, a to index actions, and d or m to denote the dimension of a feasible set.

2.1 Online linear optimization

Setup. Online linear optimization is a central problem in online learning and has been used to study learning dynamics in games. In this setting, a player is given a convex set $\mathcal{K} \subseteq \mathbb{R}^d$ before the interaction begins. Then, at each round $t = 1, \dots, T$, the player chooses a point $w^t \in \mathcal{K}$ using the information observed so far, and then the environment selects a loss vector $h^t \in \mathbb{R}^d$ without seeing w^t. The player then incurs loss $\langle w^t, h^t \rangle \in \mathbb{R}$ and observes h^t.

External regret and swap regret. The external regret is defined as the difference between the cumulative loss actually incurred by the player and that of the best fixed point in hindsight, that is, $\mathsf{Reg}^T = \max_{u \in \mathcal{K}} \mathsf{Reg}^T(u)$ for $\mathsf{Reg}^T(w^*) = \sum_{t=1}^{T} \langle w^t - w^*, h^t \rangle$. The swap regret compares the player's performance with that of any swapping rule, and we consider it only when the feasible set $\mathcal{K}$ is the probability simplex. Let $\mathcal{M}_d = \{\mathbf{M} \in [0, 1]^{d \times d} \colon \mathbf{M}(k, \cdot) \in \Delta_d \text{ for } k \in [d]\}$ denote the set of $d \times d$ row-stochastic matrices, which we also call transition probability matrices. The swap regret is then defined as $\mathsf{SwapReg}^T = \max_{\mathbf{M} \in \mathcal{M}_d} \mathsf{SwapReg}^T(\mathbf{M})$ for $\mathsf{SwapReg}^T(\mathbf{M}) = \sum_{t=1}^{T} \langle w^t, h^t - \mathbf{M} h^t \rangle$.

Optimistic follow-the-regularized-leader. A standard and widely used algorithmic framework for online linear optimization is follow-the-regularized-leader (FTRL). Here we describe its generalization, optimistic FTRL (OFTRL) (Chiang et al., 2013; Rakhlin & Sridharan, 2013), which is known to be a particularly powerful framework for learning in games. OFTRL selects a point $w^t \in \mathcal{K}$ at round $t \in [T]$ by $w^t \in \arg\min_{w \in \mathcal{K}} \{\langle w, m^t + \sum_{s=1}^{t-1} h^s \rangle + \psi^t(w)\}$, where ψ^t is a convex regularizer over $\mathcal{K}$ and $m^t \in \mathbb{R}^d$ is an optimistic prediction of the true loss vector h^t. When the optimistic prediction m^t is the zero vector for all rounds, OFTRL is called FTRL.

The prediction m^t needs to be computed using only information available up to round $t-1$, and the regret may be smaller when m^t is closer to the true loss vector h^t. This property is useful for obtaining fast convergence in learning in games, since the loss vectors are induced by the strategies of the other players (Rakhlin & Sridharan, 2013; Syrgkanis et al., 2015). For this reason, a common choice is $m^t = h^{t-1}$. OFTRL is known to achieve an RVU (Regret bounded by Variation in Utilities) bound (Syrgkanis et al., 2015) that contains a negative term, which is useful for obtaining fast convergence in games. For a concrete statement, see Lemma 5 in Appendix B.1.

2.2 TWO-PLAYER ZERO-SUM GAMES

Here we describe the problem of learning in two-player zero-sum games. It is characterized by an unknown payoff matrix $A \in [-A_{\max}, A_{\max}]^{m_x \times m_y}$, where $A_{\max} \geq 0$ is a scale parameter and m_x and m_y are the numbers of actions available to the x- and y-players, respectively. Note that most existing formulations of learning in games assume that $A_{\max} = 1$ is known to each player.

The interaction proceeds as follows. At each round $t = 1, \dots, T$, the x-player selects a *mixed strategy* (or simply a strategy) $x^t \in \Delta_{m_x}$, and the y-player simultaneously selects $y^t \in \Delta_{m_y}$. The x-player then observes a gain vector $g^t = Ay^t \in [-A_{\max}, A_{\max}]^{m_x}$ and receives reward $\langle x^t, g^t \rangle$, while the y-player observes a loss vector $\ell^t = A^\top x^t \in [-A_{\max}, A_{\max}]^{m_y}$ and incurs loss $\langle y^t, \ell^t \rangle$.

For each player, this interaction can be viewed as online linear optimization: for the x-player, the loss vector is $(-g^t)$ and the feasible set is $\mathcal{K} = \Delta_{m_x}$, and for the y-player, the loss vector is ℓ^t and $\mathcal{K} = \Delta_{m_y}$. Hence, the external regret of the x-player and that of the y-player can naturally be defined as $\mathsf{Reg}_x^T = \max_{x^* \in \Delta_{m_x}} \mathsf{Reg}_x^T(x^*)$ for $\mathsf{Reg}_x^T(x^*) = \sum_{t=1}^T \langle x^* - x^t, g^t \rangle$ and $\mathsf{Reg}_y^T = \max_{y^* \in \Delta_{m_y}} \mathsf{Reg}_y^T(y^*)$ for $\mathsf{Reg}_y^T(y^*) = \sum_{t=1}^T \langle y^t - y^*, \ell^t \rangle$, respectively.

In a two-player zero-sum game with payoff matrix A, a pair of probability distributions $\sigma = (x^*, y^*)$ over the action sets $[m_x]$ and $[m_y]$ is called an *ε-approximate Nash equilibrium* for $\varepsilon \geq 0$ if, for any $x \in \Delta_{m_x}$ and $y \in \Delta_{m_y}$, $x^\top A y^* - \varepsilon \leq x^{*\top} A y^* \leq x^{*\top} A y + \varepsilon$. When $\varepsilon = 0$, we call σ a *Nash equilibrium*. The product distribution of the average play $(\frac{1}{T}\sum_{t=1}^T x^t, \frac{1}{T}\sum_{t=1}^T y^t)$ is a $((\mathsf{Reg}_x^T + \mathsf{Reg}_y^T)/T)$-approximate Nash equilibrium (Freund & Schapire, 1999), which connects no-external-regret learning dynamics to Nash equilibrium computation.

2.3 MULTIPLAYER GENERAL-SUM GAMES

We now turn to multiplayer general-sum games. Let $n \geq 2$ denote the number of players, with player set $[n] = \{1, \dots, n\}$. Each player $i \in [n]$ has an action set $\mathcal{A}_i$ of size $|\mathcal{A}_i| = m_i$ and an unknown utility function $u_i \colon \mathcal{A}_1 \times \cdots \times \mathcal{A}_n \to [-U_{\max}, U_{\max}]$, where $U_{\max} \geq 0$ is an unknown scale parameter. Let $m = \max_{i \in [n]} m_i$, where we assume $m \geq 2$. The interaction proceeds as follows. At each round $t = 1, \dots, T$, every player $i \in [n]$ selects a mixed strategy $x_i^t \in \Delta_{m_i}$ and then observes an *expected utility* vector $u_i^t \in [-U_{\max}, U_{\max}]^{m_i}$. Here, the a_i-th component of the expected utility vector u_i^t is given by $u_i^t(a_i) = \mathbb{E}_{a_{-i} \sim x_{-i}^t}[u_i(a_i, a_{-i})]$, which is the expected reward obtained by player i when choosing action a_i while the other players act according to $x_{-i}^t = (x_1^t, \dots, x_{i-1}^t, x_{i+1}^t, \dots, x_n^t)$.

Again, for each player, this interaction can be viewed as online linear optimization: for player i, the loss vector is $(-u_i^t)$ and the feasible set is $\mathcal{K} = \Delta_{m_i}$. Thus, the swap regret of each player i can naturally be defined as $\mathsf{SwapReg}_{x_i}^T = \max_{\mathbf{M} \in \mathcal{M}_{m_i}} \mathsf{SwapReg}_{x_i}^T(\mathbf{M})$ for $\mathsf{SwapReg}_{x_i}^T(\mathbf{M}) = \sum_{t=1}^T \langle x_i^t, \mathbf{M} u_i^t - u_i^t \rangle$, where $\mathcal{M}_{m_i}$ is the set of $m_i \times m_i$ row-stochastic matrices defined in Section 2.1.

A probability distribution σ over the joint action space $\times_{i=1}^n \mathcal{A}_i$ is called an *ε-approximate correlated equilibrium* for $\varepsilon \geq 0$ if, for every player $i \in [n]$ and every (swap) function $\varphi_i \colon \mathcal{A}_i \to \mathcal{A}_i$ that replaces an action a_i by $\varphi_i(a_i)$, it holds that $\mathbb{E}_{a \sim \sigma}[u_i(a)] \geq \mathbb{E}_{a \sim \sigma}[u_i(\varphi_i(a_i), a_{-i})] - \varepsilon$. When $\varepsilon = 0$, we simply call σ a *correlated equilibrium*.

Let $\sigma^t = \otimes_{i \in [n]} x_i^t \in \Delta(\times_{i=1}^n \mathcal{A}_i)$ denote the joint distribution at round t, defined by $\sigma^t(a_1, \dots, a_n) = \prod_{i \in [n]} x_i^t(a_i)$ for each $a_i \in \mathcal{A}_i$. The time-averaged distribution $\sigma = \frac{1}{T}\sum_{t=1}^T \sigma^t$ is a $(\max_{i \in [n]} \mathsf{SwapReg}_{x_i}^T / T)$-approximate correlated equilibrium (Foster & Vohra, 1997), and hence no-swap-regret learning dynamics yield an approximate correlated equilibrium.

2.4 SCALE-FREE AND SCALE-AGNOSTIC LEARNING DYNAMICS

Here we give definitions of *scale-free* and *scale-agnostic* online linear optimization algorithms and learning dynamics in games. Note that neither property implies the other, since dividing the utilities by a known scale makes any learning dynamic scale-free and Appendix B.5 provides a learning dynamic that is scale-agnostic but not scale-free.

Definition 1 (Scale-free / scale-agnostic online linear optimization)**.** An online linear optimization algorithm is said to be *scale-free* if, for any sequence of loss vectors $(h^t)_{t=1}^T$ and any constant $c > 0$, the algorithm outputs the same sequence of points $(w^t)_{t=1}^T$ when run on $(h^t)_{t=1}^T$ and when run on $(ch^t)_{t=1}^T$. Moreover, an online linear optimization algorithm is said to be *scale-agnostic* if it does not rely on any prior information about the loss vectors $(h^t)_{t=1}^T$ except for their dimension.

Definition 2 (Scale-free / scale-agnostic learning dynamics in games)**.** A learning dynamic in games is said to be *scale-free* if, for any functions $u_1, \dots, u_n$ and any constant $c > 0$, the dynamic outputs the same sequence of strategies when run on utility functions $u_1, \dots, u_n$ and when run on $cu_1, \dots, cu_n$. A learning dynamic in games is said to be *scale-agnostic* if the dynamic does not rely on any prior information about the underlying utilities except for their dimension.

3 Scale-Free Dynamics for Two-Player Zero-Sum Games

This section investigates scale-free learning dynamics for two-player zero-sum games.

3.1 Proposed learning dynamic

Here we present scale-free and scale-agnostic learning dynamics with a fast convergence rate. Let $M_x = \max\{4, \log m_x / 2^{3/2}\}$ and $M_y = \max\{4, \log m_y / 2^{3/2}\}$. Define $M = \max\{M_x, M_y\}$. For $\mu_g = (\max_{a \in [m_x]} g^1(a) + \min_{a \in [m_x]} g^1(a))/2$ and $\mu_\ell = (\max_{b \in [m_y]} \ell^1(b) + \min_{b \in [m_y]} \ell^1(b))/2$, define the centered gradient sequences[1] $g_\mu^0 = \mathbf{0}$, $g_\mu^t = g^t - \mu_g \mathbf{1}$ for $t \geq 1$, and $\ell_\mu^0 = \mathbf{0}$, $\ell_\mu^t = \ell^t - \mu_\ell \mathbf{1}$ for $t \geq 1$. We then consider the optimistic Hedge algorithm, which determines the strategies of the x- and y-players via $x^1 = \frac{1}{m_x}\mathbf{1}$, $y^1 = \frac{1}{m_y}\mathbf{1}$ and, for $t \geq 2$,

$$\begin{aligned} x^t(a) &\propto \exp\left(\eta_x^t \left(\sum_{s=1}^{t-1} g_\mu^s(a) + g_\mu^{t-1}(a)\right)\right), \quad \eta_x^t = \sqrt{\frac{M_x}{P_\infty^t(g_\mu) + P_\infty^t(\ell_\mu)}}, \\ y^t(b) &\propto \exp\left(-\eta_y^t \left(\sum_{s=1}^{t-1} \ell_\mu^s(b) + \ell_\mu^{t-1}(b)\right)\right), \quad \eta_y^t = \sqrt{\frac{M_y}{P_\infty^t(g_\mu) + P_\infty^t(\ell_\mu)}}, \end{aligned} \tag{1}$$

where we recall that $P_q^t(z) = \sum_{s=1}^{t-1} \|z^s - z^{s-1}\|_q^2$. If the denominators of the learning rates are zero, we set $x^t = \frac{1}{m_x}\mathbf{1}$ and $y^t = \frac{1}{m_y}\mathbf{1}$. The learning rates η_x^t and η_y^t can be computed using only the observations up to round $t-1$. Note also that optimistic Hedge corresponds to OFTRL with the negative Shannon entropy regularizer.

3.2 External regret upper bounds

Let $m = \max\{m_x, m_y\}$. Then the above learning dynamic guarantees the following bounds.

Theorem 1 (Scale-free fast convergence to Nash equilibrium)**.** *Suppose that the x- and y-players use the algorithms in (1). Then, it holds that* $\mathsf{Reg}_x^T \leq 91 A_{\mathrm{diff}} \sqrt{(M+1)M_x} = O(A_{\mathrm{diff}} \log m)$ *and* $\mathsf{Reg}_y^T \leq 91 A_{\mathrm{diff}} \sqrt{(M+1)M_y} = O(A_{\mathrm{diff}} \log m)$*. Consequently, the average play is an $O(A_{\mathrm{diff}} \log m / T)$-approximate Nash equilibrium.*

Unlike existing scale-free learning results, our analysis shows that the convergence rate can be controlled not by the maximum absolute entry $A_{\max}$ of the payoff matrix, but rather by the payoff range A_{diff}, the difference between the maximum and the minimum entries in the payoff matrix. Note that our $O(A_{\mathrm{diff}} \log m)$ regret bound does not contradict the lower bound for optimistic Hedge (Tsuchiya, 2026a), since their lower bound applies to optimistic Hedge with a constant learning rate. Our analysis also extends to the corrupted regime (Tsuchiya et al., 2025), where the opponent may deviate from the prescribed algorithm or the observed gradients may be adversarially corrupted. We defer the details to Appendix B.

[1] The centered sequences are introduced only to control the initial path-length term in the learning rates: since the shifts are action-independent, they affect the update in (1) only through the learning rates, and they ensure that the first increment of the path length is controlled by A_{diff} rather than by $A_{\max}$. In fact, for $t \geq 2$ we have $P_\infty^t(g_\mu) = \|g^1 - \mu_g \mathbf{1}\|_\infty^2 + \sum_{s=2}^{t-1} \|g^s - g^{s-1}\|_\infty^2$.

3.3 Proof technique and sketch of Theorem 1

Here we describe the proof technique and give a proof sketch of Theorem 1; the full proof is given in Appendix B.2, and the extension to the corrupted regime in Appendix B.4.

Since $x^*, x^t \in \Delta_{m_x}$, we have $\langle x^* - x^t, g^t \rangle = \langle x^* - x^t, g_\mu^t \rangle$, and thus Reg_x^T is the regret of optimistic Hedge run on the centered loss sequence $(-g_\mu^t)_t$ with the learning rate η_x^t in (1). Applying the standard RVU bound for the AdaHedge-type update (Lemma 5 in Appendix B.1), we have

$$\mathsf{Reg}_x^T \le \sqrt{32\big(P_\infty^{T+1}(g_\mu) + P_\infty^{T+1}(\ell_\mu)\big) M_x} - \sum_{t=1}^{T} \frac{1}{4\eta_x^t} \|x^{t+1} - x^t\|_1^2 , \tag{2}$$

and the same bound with M_y and η_y^t holds for Reg_y^T. The squared path lengths $P_\infty^{T+1}(g_\mu)$ and $P_\infty^{T+1}(\ell_\mu)$ in (2) are those of the observed gradients, and their increments are bounded by the strategy changes, since $\|g_\mu^t - g_\mu^{t-1}\|_\infty \le \bar{A}_{\mathrm{diff}} \|y^t - y^{t-1}\|_1$ and $\|\ell_\mu^t - \ell_\mu^{t-1}\|_\infty \le \bar{A}_{\mathrm{diff}} \|x^t - x^{t-1}\|_1$ for $t \ge 2$, where $\bar{A}_{\mathrm{diff}} = A_{\mathrm{diff}}/2$ (see Lemma 6).

When the scale is known, the negative term of the RVU bound in (2) can be exploited by a learning rate bounded by a constant of order $1/A_{\max}$ (*e.g.,* Syrgkanis et al. 2015; Tsuchiya et al. 2025). By using this and summing Reg_x^T and Reg_y^T, the path-length term of each player is then absorbed by the negative term in (2) of the opponent, and we obtain $\mathsf{Reg}_x^T + \mathsf{Reg}_y^T \lesssim A_{\max} \log m$.

However, this approach requires prior knowledge of $A_{\max}$. To resolve this, we notice that the learning rate in (1) satisfies $1/\eta_x^t = \sqrt{(P_\infty^t(g_\mu) + P_\infty^t(\ell_\mu))/M_x}$, which is small while the observed path lengths are small, and thus the negative term in (2) can compensate for the increments of the path lengths only after the path lengths have reached order $A_{\mathrm{diff}}^2 M$, and the rounds before that must be handled separately.

The following lemma is useful for exploiting the negative term in the RVU bound to obtain a fast constant regret bound even when the algorithm is agnostic to $A_{\max}$ and A_{diff}.

Lemma 1. *Let $z^1, \dots, z^T \in \Delta_d$, $h^t = Az^t$ for $t \in [T]$, and $c_1, c_2 > 0$. Define $\mu_h = (\max_a h^1(a) + \min_a h^1(a))/2$, $h_\mu^0 = \mathbf{0}$, and $h_\mu^t = h^t - \mu_h \mathbf{1}$ for $t \ge 1$. Define $\tau^\circ(h_\mu; c_1, c_2)$ by*

$$\tau^\circ(h_\mu; c_1, c_2) = \begin{cases} \min\big\{t \in \{2, \dots, T+1\} \colon P_\infty^t(h_\mu) > \frac{c_2}{c_1} + A_{\mathrm{diff}}^2 \big\} & \text{if } P_\infty^{T+1}(h_\mu) > \frac{c_2}{c_1} + A_{\mathrm{diff}}^2, \\ \infty & \text{otherwise}, \end{cases} \tag{3}$$

where we recall $P_\infty^t(h_\mu) = \sum_{s=1}^{t-1} \|h_\mu^s - h_\mu^{s-1}\|_\infty^2$. Then, we have $\tau^\circ \ge 3$, and if $\tau^\circ(h_\mu; c_1, c_2) < \infty$, it holds that

$$P_\infty^{\tau^\circ}(h_\mu) \le \frac{c_2}{c_1} + 2A_{\mathrm{diff}}^2 , \ P_\infty^{\tau^\circ - 1}(h_\mu) \ge \frac{c_2}{c_1} , \ c_1 \sqrt{P_\infty^{\tau^\circ}(h_\mu)} + \frac{c_2}{\sqrt{P_\infty^{\tau^\circ - 1}(h_\mu)}} \le \sqrt{2} c_1 A_{\mathrm{diff}} + 2\sqrt{c_1 c_2} .$$

In the proof of Theorem 1, we fix $\tau_x, \tau_y \in \{2, \dots, T+1\}$ and split each path length in the first term of (2). We show that the increments of $P_\infty^{T+1}(g_\mu)$ over rounds $\tau_y, \dots, T$ sum to at most $\bar{A}_{\mathrm{diff}}^2 Q_1^{\tau_y}(y)$ (see Lemma 6), and similarly for ℓ_μ. These increments can be canceled by the negative terms in (2) up to terms proportional to $\eta_x^{\tau_x - 1}$ and $\eta_y^{\tau_y - 1}$ by considering the worst case with respect to $Q_1^{\tau_y}(y)$ and $Q_1^{\tau_x}(x)$ since the learning rates are nonincreasing. Combining this observation and using $\eta_x^{\tau_x - 1} \le \sqrt{M_x / P_\infty^{\tau_x - 1}(\ell_\mu)}$ and $\eta_y^{\tau_y - 1} \le \sqrt{M_y / P_\infty^{\tau_y - 1}(g_\mu)}$, we obtain

$$\mathsf{Reg}_x^T + \mathsf{Reg}_y^T \le 2^{7/2} \sum_{(h_\mu, \tau)} \left(\sqrt{P_\infty^\tau(h_\mu) M} + \frac{2^{9/2} A_{\mathrm{diff}}^2 M^{3/2}}{\sqrt{P_\infty^{\tau - 1}(h_\mu)}} \right) - \frac{Q_1^{\tau_x}(x)}{8\eta_x^{\tau_x - 1}} - \frac{Q_1^{\tau_y}(y)}{8\eta_y^{\tau_y - 1}} , \tag{4}$$

where the sum is over $(h_\mu, \tau) \in \{(\ell_\mu, \tau_x), (g_\mu, \tau_y)\}$.

Now we choose the stopping times by (3) with $c_1 = \sqrt{M}$ and $c_2 = 2^{9/2} A_{\mathrm{diff}}^2 M^{3/2}$, namely $\tau_x = \min\{\tau^\circ(\ell_\mu; c_1, c_2), T+1\}$ and $\tau_y = \min\{\tau^\circ(g_\mu; c_1, c_2), T+1\}$. (This proof sketch assumes that both stopping times in (3) are finite.) Then, by Lemma 1, we have $P_\infty^\tau(h_\mu) \lesssim c_2/c_1 + 2A_{\mathrm{diff}}^2 \lesssim A_{\mathrm{diff}}^2 M$ and $P_\infty^{\tau - 1}(h_\mu) \ge c_2/c_1$, and thus $\eta_x^{\tau_x - 1} \lesssim 1/A_{\mathrm{diff}}$. The last inequality of Lemma 1 also implies that the sum in (4) is at most $O(A_{\mathrm{diff}} M)$.

Algorithm 1: No-swap-regret algorithm of player i in multiplayer general-sum games

Set $U^0 = 0$.
for $t = 1, 2, \dots, T$ **do**
  Compute $y_{i,a}^t \in \Delta_{m_i}$ by OFTRL in (6) for each expert $a \in \mathcal{A}_i$ (at $t = 1$, $y_{i,a}^1 = \frac{1}{m_i}\mathbf{1}$).
  Let $\mathbf{Q}_i^t \in [0,1]^{m_i \times m_i}$ be a matrix whose a-th row is $y_{i,a}^t$, that is, $\mathbf{Q}_i^t(a,\cdot) = (y_{i,a}^t)^\top$.
  Let x_i^t be a stationary distribution of the Markov chain $\mathbf{Q}_i^t$, that is, $(\mathbf{Q}_i^t)^\top x_i^t = x_i^t$.
  Play strategy $x_i^t \in \Delta_{m_i}$.
  Observe a utility vector $u_i^t \in [-U_{\max}, U_{\max}]^{m_i}$, where $u_i^t(a_i) = \mathbb{E}_{a_{-i} \sim x_{-i}^t}[u_i(a_i, a_{-i})]$.
  Let $u_{i,a}^t = x_i^t(a) u_i^t \in [-x_i^t(a) U_{\max}, x_i^t(a) U_{\max}]^{m_i}$ for each $a \in \mathcal{A}_i$.
  Compute $U^t = \max\{U^{t-1}, \max_{j \in [n]} \|u_j^t\|_\infty\}$.
  Clip the utility vectors to obtain $\bar{u}_i^t$ and $(\bar{u}_{i,a}^t)_{a \in \mathcal{A}_i}$ via (5).

Finally, combining (4) with $\mathsf{Reg}_x^T + \mathsf{Reg}_y^T \geq 0$, and the facts that the sum in (4) is $O(A_{\mathrm{diff}} M)$ and $1/\eta_x^{\tau_x - 1}, 1/\eta_y^{\tau_y - 1} \gtrsim A_{\mathrm{diff}}$, we obtain $Q_1^{\tau_x}(x), Q_1^{\tau_y}(y) \lesssim M$. Substituting these bounds and $P_\infty^{\tau_x}(\ell_\mu), P_\infty^{\tau_y}(g_\mu) \lesssim A_{\mathrm{diff}}^2 M$ into (2) gives $\mathsf{Reg}_x^T \lesssim A_{\mathrm{diff}} \sqrt{M M_x}$, which is the bound in Theorem 1.

Comparison with existing analysis. Instead of bounding the learning rate in advance, we control the contribution of the rounds with a large learning rate through the observed path length, which is bounded by the stopping rule itself. This approach replaces the scale-dependent constant upper bound on the learning rate in Tsuchiya et al. (2025), which requires prior knowledge of $A_{\max}$, and the assumption $\delta = \min\{\|g^1\|, \|\ell^1\|\} > 0$ in Zhang et al. (2025), whose bound scales with $1/\delta$ (see Appendix D.1). Finally, the bound depends on A_{diff} rather than $A_{\max}$ because the centering makes the first increment $\|g_\mu^1 - g_\mu^0\|_\infty$ at most $\bar{A}_{\mathrm{diff}}$ and the AdaHedge-type bound in (2) depends on the gradients only through their increments.

4 Scale-Free Dynamics for Multiplayer General-Sum Games

This section presents scale-free and scale-agnostic dynamics for multiplayer general-sum games.

4.1 Preliminaries: Swap Regret to External Regret

Here we describe the reduction of Blum & Mansour (2007) from swap regret minimization to external regret minimization. We focus on the construction for player i. The reduction runs m_i external regret minimizers, which we refer to as "experts" and index by actions $a \in \mathcal{A}_i$. For each $a \in \mathcal{A}_i$, define the utility vector fed to expert a at round t by $u_{i,a}^t = x_i^t(a) u_i^t \in [-U_{\max} x_i^t(a), U_{\max} x_i^t(a)]^{m_i}$. Let $y_{i,a}^t \in \Delta_{m_i}$ denote the output distribution produced by expert a at round t. Using $(y_{i,a}^t)_{a \in \mathcal{A}_i}$, we form a transition matrix $\mathbf{Q}_i^t \in [0,1]^{m_i \times m_i}$ by setting its a-th row to be $y_{i,a}^t$, *i.e.*, $\mathbf{Q}_i^t(a,\cdot) = (y_{i,a}^t)^\top$. We then choose x_i^t as a stationary distribution of the Markov chain induced by $\mathbf{Q}_i^t$, namely, $(\mathbf{Q}_i^t)^\top x_i^t = x_i^t$ (for the column vector x_i^t), and use this x_i^t as player i's strategy.

For a comparator $y^* \in \Delta_{m_i}$, define $\mathsf{Reg}_{i,a}^T(y^*) = \sum_{t=1}^T \langle y^* - y_{i,a}^t, u_{i,a}^t \rangle$, and let $\mathsf{Reg}_{i,a}^T = \max_{y \in \Delta_{m_i}} \mathsf{Reg}_{i,a}^T(y)$ be the external regret of expert $a \in \mathcal{A}_i$ for player $i \in [n]$. Blum & Mansour (2007) show that, for every $\mathbf{M} \in \mathcal{M}_{m_i}$, this construction satisfies $\mathsf{SwapReg}_{x_i}^T(\mathbf{M}) = \sum_{a \in \mathcal{A}_i} \mathsf{Reg}_{i,a}^T(\mathbf{M}(a,\cdot))$. Since the rows of $\mathbf{M}$ can be chosen independently, maximizing both sides gives $\mathsf{SwapReg}_{x_i}^T = \sum_{a \in \mathcal{A}_i} \mathsf{Reg}_{i,a}^T$.

4.2 Proposed Learning Dynamic

By the reduction above, to minimize $\mathsf{SwapReg}_{x_i}^T(\mathbf{M})$, it suffices to design an external regret minimization algorithm that guarantees small regret $\mathsf{Reg}_{i,a}^T(\mathbf{M}(a,\cdot))$ for each expert $a \in \mathcal{A}_i$. Each expert runs OFTRL on clipped utility vectors with an adaptive learning rate. Our complete algorithm is summarized in Algorithm 1, and we explain its components below.

Clipping. To handle the unknown utility scale, we use clipping based on past observations, following Cutkosky (2019, Section 4). Here we use a common scale across players: let $U^t =$

$\max_{s\in[t]}\max_{j\in[n]}\|u_j^s\|_\infty$, with $U^0 = 0$. In the first round, we set $\bar{u}_i^1 = u_i^1$ and $\bar{u}_{i,a}^1 = u_{i,a}^1$ (if $U^1 = 0$, all players keep playing uniform distributions, as explained in Appendix C). After observing the utility vectors at round $t \geq 2$, we compute the clipped utility vectors by

$$\bar{u}_{i,a}^t = \frac{U^{t-1}}{U^t} u_{i,a}^t\,, \quad \bar{u}_i^t = \frac{U^{t-1}}{U^t} u_i^t\,. \tag{5}$$

Thus, $\bar{u}_{i,a}^t = x_i^t(a)\bar{u}_i^t$, and $\|\bar{u}_i^t\|_\infty \leq U^{t-1}$ for $t \geq 2$. In particular, the clipped vectors are bounded by a scale known before the current round, even when the observed utility norm increases.

Online learner. We run OFTRL on the clipped utility vectors in (5) using the hybrid regularizer of Tsuchiya (2026b), which combines the negative Shannon entropy and the log barrier. Let $H(y) = -\sum_b y(b)\log y(b)$ be the Shannon entropy and $\phi(y) = -\sum_b \log y(b)$ be the log-barrier function. Then, for each $a \in \mathcal{A}_i$ and $t \geq 2$, we compute $y_{i,a}^t \in \Delta_{m_i}$ by

$$y_{i,a}^t = \operatorname*{arg\,max}_{y\in\Delta_{m_i}} \left\{ \left\langle y, \bar{u}_{i,a}^{t-1} + \sum_{s=1}^{t-1} \bar{u}_{i,a}^s \right\rangle - \psi^t(y) \right\}, \quad \eta^t = \frac{\alpha}{\sqrt{\gamma(U^{t-1})^2 + \max_{j\in[n]} P_\infty^t(\bar{u}_j)}}\,, \tag{6}$$

where $\psi^t(y) = (\theta\phi(y) - H(y))/\eta^t$ for $\theta = \log m/(m\log T)$, $\alpha = \sqrt{m\log m}$ and $\gamma = 2^{10}m^3\log T$, and we set $\bar{u}_j^0 = \mathbf{0}$. Here, η^t is the common learning rate of all players and experts at round t. The relative weight θ of the two regularizers is fixed throughout the T rounds, and only their common multiplier $1/\eta^t$ changes with the observations. The learning rate is computed from the observations up to round $t-1$, but this computation requires sharing two statistics across players, the maximum utility norm and the maximum squared path length, and thus the learning dynamic is not uncoupled. The value $\gamma(U^{t-1})^2$ in (6) controls the increase of $\max_j P_\infty^t(\bar{u}_j)$ and ensures that the negative term is large enough to bound the contribution of changes in the player's own strategy (see Lemma 15).

4.3 Swap Regret Upper Bounds

Here we present regret upper bounds for Algorithm 1.

Theorem 2 (Scale-free fast convergence to correlated equilibrium)**.** *Suppose that $T \geq 3$ and every player uses Algorithm 1. Then, $\mathsf{SwapReg}_{x_i}^T \lesssim U_{\max} nm^2\sqrt{\log m\log T}$ for all $i \in [n]$. Consequently, the time-averaged history of joint play after T rounds is an $O(U_{\max}nm^2\sqrt{\log m\log T}/T)$-approximate correlated equilibrium.*

The proof is deferred to Appendix C. The bound matches the $O(U_{\max}nm^2\sqrt{\log m\log T})$ bound obtained by Tsuchiya (2026b, Theorem 2) after normalizing the utilities by a known scale, but does not require this knowledge. Unlike their uncoupled dynamics, our learning dynamic uses the shared statistics described above. Another limitation is that, unlike Tsuchiya et al. (2025, Section 5), we do not establish robustness to adversarial corruption of utilities.

Techniques. As in the two-player case, the proof exploits the negative terms of the RVU bounds to control the squared path lengths of the observed vectors. The swap regret of player i is at most $O(\alpha(\sqrt{\gamma}\,U_{\max} + \sqrt{\mathbb{S}^{T+1}}))$ minus a multiple of $\sum_t \|x_i^{t+1} - x_i^t\|_1^2/\eta^t$ for $\mathbb{S}^t = \max_{j\in[n]} P_\infty^t(\bar{u}_j)$ (Lemma 15). This negative term comes from the log barrier, which controls the changes in stationary distributions (Tsuchiya, 2026b, Theorem 3 and Lemma 3). Unlike in two-player zero-sum games, the utility vector of a player changes with the strategies of all other players, and thus its changes cannot be absorbed by the negative term of a single opponent. To resolve this, we use a common adaptive learning rate in (6), and define it through the maximum of the players' squared path lengths.

Moreover, as in the two-player zero-sum games, the scale the players have observed can stay far below $U_{\max}$, whereas the changes in the utility vectors caused by the strategy changes scale with $U_{\max}$ (Lemma 16). The coefficient of the negative term is proportional to $1/\eta^t = \sqrt{\gamma(U^{t-1})^2 + \mathbb{S}^t}/\alpha$, and thus it can be too small to absorb these changes while U^{t-1} and $\mathbb{S}^t$ are small. To handle this, we exploit the fact that this coefficient increases with the squared path length, namely $1/\eta^t \geq \sqrt{\mathbb{S}^{t+1}}/\alpha$ owing to the term $\gamma(U^{t-1})^2$. Since the increase $\mathbb{S}^{t+1} - \mathbb{S}^t$ is caused by these changes, the negative terms of all players are at least a multiple of $\sum_t \sqrt{\mathbb{S}^t}\,(\mathbb{S}^{t+1} - \mathbb{S}^t)$ in magnitude, and this sum is at least $\frac{2}{3}(\mathbb{S}^{T+1})^{3/2} - 2U_{\max}\mathbb{S}^{T+1}$ (Lemma 17). Since swap regret is nonnegative, the negative terms cannot exceed the positive terms, which grow only as $\sqrt{\mathbb{S}^{T+1}}$. This gives $\mathbb{S}^{T+1} = O(U_{\max}^2 n^2 m^3 \log T)$, and substituting it into the positive terms yields the bound in Theorem 2.

AI USE STATEMENT

The problem formulation, the algorithms, and the analysis in the original version of this paper were developed without generative AI tools, except that we used such tools for proofreading. In revising the original version of the paper, we used generative AI tools (GPT-5.6-Sol, GPT-6-Astra, Fable-5, and Opus 5.5) to check the proofs in the appendix for errors and to proofread and edit the text for readability, and used them to simplify a small part of the proofs. We also used them to design the experiments, implement and test the code, and draft and proofread the experimental section. We have reviewed all AI-assisted work, and we take responsibility for the final content of this work.

ETHICS STATEMENT

This work does not involve human subjects, personal data, or applications that raise ethical concerns.

REPRODUCIBILITY STATEMENT

The assumptions of all theoretical results are stated in the theorems, the learning dynamics are given in Sections 3 and 4, and complete proofs are given in Appendices A to C. The experimental setup and comparison methods are described in Appendix E. The code for the experiments is provided in the supplementary material.

REFERENCES

Ioannis Anagnostides, Constantinos Daskalakis, Gabriele Farina, Maxwell Fishelson, Noah Golowich, and Tuomas Sandholm. Near-optimal no-regret learning for correlated equilibria in multi-player general-sum games. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 736–749. Association for Computing Machinery, 2022a.

Ioannis Anagnostides, Gabriele Farina, Christian Kroer, Chung-Wei Lee, Haipeng Luo, and Tuomas Sandholm. Uncoupled learning dynamics with $O(\log T)$ swap regret in multiplayer games. In *Advances in Neural Information Processing Systems*, volume 35, pp. 3292–3304. Curran Associates, Inc., 2022b.

Robert J. Aumann. Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics*, 1(1):67–96, 1974.

Avrim Blum and Yishay Mansour. From external to internal regret. *Journal of Machine Learning Research*, 8(47):1307–1324, 2007.

Michael Bowling, Neil Burch, Michael Johanson, and Oskari Tammelin. Heads-up limit hold'em poker is solved. *Science*, 347(6218):145–149, 2015.

Noam Brown and Tuomas Sandholm. Solving imperfect-information games via discounted regret minimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 1829–1836, 2019.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

Nicolò Cesa-Bianchi, Yishay Mansour, and Gilles Stoltz. Improved second-order bounds for prediction with expert advice. *Machine Learning*, 66(2):321–352, 2007.

Darshan Chakrabarti, Julien Grand-Clément, and Christian Kroer. Extensive-form game solving via Blackwell approachability on treeplexes. In *Advances in Neural Information Processing Systems*, volume 37, pp. 35257–35287. Curran Associates, Inc., 2024.

Liyu Chen, Haipeng Luo, and Chen-Yu Wei. Impossible tuning made possible: A new expert algorithm and its applications. In *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134, pp. 1216–1259. PMLR, 2021.

Xi Chen and Binghui Peng. Hedging in games: Faster convergence of external and swap regrets. In *Advances in Neural Information Processing Systems*, volume 33, pp. 18990–18999. Curran Associates, Inc., 2020.

Chao-Kai Chiang, Chia-Jung Lee, and Chi-Jen Lu. Beating bandits in gradually evolving worlds. In *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30, pp. 210–227. PMLR, 2013.

Ashok Cutkosky. Artificial constraints and hints for unbounded online learning. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99, pp. 874–894. PMLR, 2019.

Constantinos Daskalakis, Alan Deckelbaum, and Anthony Kim. Near-optimal no-regret algorithms for zero-sum games. In *Proceedings of the Twenty-Second Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 235–254. Society for Industrial and Applied Mathematics, 2011.

Steven de Rooij, Tim van Erven, Peter D. Grünwald, and Wouter M. Koolen. Follow the leader if you can, hedge if you must. *Journal of Machine Learning Research*, 15(37):1281–1316, 2014.

Meta Fundamental AI Research Diplomacy Team FAIR, Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, Athul Paul Jacob, Mojtaba Komeili, Karthik Konath, Minae Kwon, Adam Lerer, Mike Lewis, Alexander H. Miller, Sasha Mitts, Adithya Renduchintala, Stephen Roller, Dirk Rowe, Weiyan Shi, Joe Spisak, Alexander Wei, David Wu, Hugh Zhang, and Markus Zijlstra. Human-level play in the game of Diplomacy by combining language models with strategic reasoning. *Science*, 378 (6624):1067–1074, 2022.

Dean P. Foster and Rakesh V. Vohra. Calibrated learning and correlated equilibrium. *Games and Economic Behavior*, 21(1):40–55, 1997.

Dylan J Foster, Zhiyuan Li, Thodoris Lykouris, Karthik Sridharan, and Eva Tardos. Learning in games: Robustness of fast convergence. In *Advances in Neural Information Processing Systems*, volume 29, pp. 4734–4742. Curran Associates, Inc., 2016.

Yoav Freund and Robert E. Schapire. Adaptive game playing using multiplicative weights. *Games and Economic Behavior*, 29(1):79–103, 1999.

Julien Grand-Clément and Christian Kroer. Conic Blackwell algorithm: Parameter-free convex-concave saddle-point solving. In *Advances in Neural Information Processing Systems*, volume 34, pp. 9587–9599. Curran Associates, Inc., 2021.

Sergiu Hart and Andreu Mas-Colell. A simple adaptive procedure leading to correlated equilibrium. *Econometrica*, 68(5):1127–1150, 2000.

Yu-Guan Hsieh, Kimon Antonakopoulos, and Panayotis Mertikopoulos. Adaptive learning in continuous games: Optimal regret bounds and convergence to Nash equilibrium. In *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134, pp. 2388–2422. PMLR, 2021.

Zakaria Mhammedi, Wouter M Koolen, and Tim Van Erven. Lipschitz adaptivity with multiple learning rates in online learning. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99, pp. 2490–2511. PMLR, 2019.

Matej Moravčík, Martin Schmid, Neil Burch, Viliam Lisý, Dustin Morrill, Nolan Bard, Trevor Davis, Kevin Waugh, Michael Johanson, and Michael Bowling. DeepStack: Expert-level artificial intelligence in heads-up no-limit poker. *Science*, 356(6337):508–513, 2017.

Remi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Côme Fiegel, Andrea Michi, Marco Selvi, Sertan Girgin, Nikola Momchev, Olivier Bachem, Daniel J Mankowitz, Doina Precup, and Bilal Piot. Nash learning from human feedback. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 36743–36768. PMLR, 2024.

Francesco Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.

Francesco Orabona and Dávid Pál. Scale-free online learning. *Theoretical Computer Science*, 716: 50–69, 2018. Special Issue on ALT 2015.

Julien Perolat, Bart De Vylder, Daniel Hennes, Eugene Tarassov, Florian Strub, Vincent de Boer, Paul Muller, Jerome T. Connor, Neil Burch, Thomas Anthony, Stephen McAleer, Romuald Elie, Sarah H. Cen, Zhe Wang, Audrunas Gruslys, Aleksandra Malysheva, Mina Khan, Sherjil Ozair, Finbarr Timbers, Toby Pohlen, Tom Eccles, Mark Rowland, Marc Lanctot, Jean-Baptiste Lespiau, Bilal Piot, Shayegan Omidshafiei, Edward Lockhart, Laurent Sifre, Nathalie Beauguerlange, Remi Munos, David Silver, Satinder Singh, Demis Hassabis, and Karl Tuyls. Mastering the game of Stratego with model-free multiagent reinforcement learning. *Science*, 378(6623):990–996, 2022.

Sasha Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. In *Advances in Neural Information Processing Systems*, volume 26, pp. 3066–3074. Curran Associates, Inc., 2013.

Gokul Swamy, Christoph Dann, Rahul Kidambi, Steven Wu, and Alekh Agarwal. A minimaximalist approach to reinforcement learning from human feedback. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 47345–47377. PMLR, 2024.

Vasilis Syrgkanis, Alekh Agarwal, Haipeng Luo, and Robert E Schapire. Fast convergence of regularized learning in games. In *Advances in Neural Information Processing Systems*, volume 28, pp. 2989–2997. Curran Associates, Inc., 2015.

Taira Tsuchiya. Tight regret upper and lower bounds for optimistic Hedge in two-player zero-sum games. In *Proceedings of The 29th International Conference on Artificial Intelligence and Statistics*, volume 300, pp. 4870–4878. PMLR, 2026a.

Taira Tsuchiya. Sublogarithmic swap regret in multiplayer general-sum games via hybrid regularization. *arXiv preprint arXiv:2608.04149*, 2026b.

Taira Tsuchiya, Shinji Ito, and Haipeng Luo. Corrupted learning dynamics in games. In *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291, pp. 5506–5552. PMLR, 2025.

Chen-Yu Wei, Chung-Wei Lee, Mengxiao Zhang, and Haipeng Luo. Linear last-iterate convergence in constrained saddle-point optimization. In *International Conference on Learning Representations*, 2021.

Brian Hu Zhang, Ioannis Anagnostides, and Tuomas Sandholm. Scale-invariant regret matching and online learning with optimal convergence: Bridging theory and practice in zero-sum games. *arXiv preprint arXiv:2510.04407*, 2025.

A Regret Analysis of Optimistic FTRL

This section provides regret upper bounds of optimistic follow-the-regularized-leader (OFTRL) for online linear optimization. First, we provide a general analysis of OFTRL. Then, we present an RVU bound for OFTRL with the negative Shannon entropy regularizer, which is used in Appendices B and C. The analysis in Appendix C combines the general OFTRL bound with a bound for the hybrid regularizer (Lemma 12). The notation in this section follows that of the online linear optimization setup described in Section 2.1.

Let $D_\psi(x, y)$ denote the Bregman divergence between x and y induced by a differentiable convex function ψ, that is, $D_\psi(x, y) = \psi(x) - \psi(y) - \langle \nabla\psi(y), x - y\rangle$. The following lemma provides a regret bound for OFTRL, which is adapted from Tsuchiya et al. (2025, Lemma 16).

Lemma 2. *Let $\mathcal{K} \subseteq \mathbb{R}^d$ be a nonempty closed and bounded convex set. Let $w^t \in \arg\min_{x\in\mathcal{K}}\{\langle x, m^t + \sum_{s=1}^{t-1} h^s\rangle + \psi^t(x)\}$ be the output of OFTRL at round t. Then, for any $w^* \in \mathcal{K}$,*

$$\sum_{t=1}^{T}\langle w^t - w^*, h^t\rangle \le \psi^{T+1}(w^*) - \psi^1(w^1) + \sum_{t=1}^{T}\left(\psi^t(w^{t+1}) - \psi^{t+1}(w^{t+1})\right) + \sum_{t=1}^{T}\left(\langle w^t - w^{t+1}, h^t - m^t\rangle - D_{\psi^t}(w^{t+1}, w^t)\right) + \langle w^* - w^{T+1}, m^{T+1}\rangle\,. \tag{7}$$

Lemma 2 yields the following regret upper bound for OFTRL with the negative Shannon entropy regularizer, which corresponds to optimistic Hedge.

Lemma 3 (RVU bound for optimistic Hedge)**.** *Let $\psi^t(x) = -\frac{1}{\eta^t}H(x)$ for $H(x) = \sum_{k=1}^{d} x(k)\log(1/x(k))$ be the negative Shannon entropy regularizer with a nonincreasing learning rate η^t. Let $w^t \in \arg\min_{x\in\Delta_d}\{\langle x, m^t + \sum_{s=1}^{t-1} h^s\rangle + \psi^t(x)\}$ be the output of OFTRL at round t, and let $m^{T+1} = \mathbf{0}$. Then, for any $w^* \in \Delta_d$,*

$$\sum_{t=1}^{T}\langle w^t - w^*, h^t\rangle \le \frac{\log d}{\eta^{T+1}} + \sum_{t=1}^{T}\min\left\{2\|h^t - m^t\|_\infty, \eta^t\|h^t - m^t\|_\infty^2\right\} - \sum_{t=1}^{T}\frac{1}{4\eta^t}\|w^t - w^{t+1}\|_1^2\,. \tag{8}$$

Proof. We will upper bound the RHS of (7) in Lemma 2. Since $H(x) \le \log d$ for all $x \in \Delta_d$, we have

$$\psi^{T+1}(w^*) - \psi^1(w^1) + \sum_{t=1}^{T}\left(\psi^t(w^{t+1}) - \psi^{t+1}(w^{t+1})\right) \le \frac{\log d}{\eta^1} + \sum_{t=1}^{T}\left(\frac{1}{\eta^{t+1}} - \frac{1}{\eta^t}\right)\log d = \frac{\log d}{\eta^{T+1}}\,.$$

Since ψ^t is $(1/\eta^t)$-strongly convex with respect to $\|\cdot\|_1$, we also have

$$\begin{aligned}
&\langle w^t - w^{t+1}, h^t - m^t\rangle - D_{\psi^t}(w^{t+1}, w^t)\\
&\le \|w^t - w^{t+1}\|_1\|h^t - m^t\|_\infty - \frac{1}{2\eta^t}\|w^t - w^{t+1}\|_1^2\\
&= \|w^t - w^{t+1}\|_1\|h^t - m^t\|_\infty - \frac{1}{4\eta^t}\|w^t - w^{t+1}\|_1^2 - \frac{1}{4\eta^t}\|w^t - w^{t+1}\|_1^2\\
&\le \min\left\{2\|h^t - m^t\|_\infty, \eta^t\|h^t - m^t\|_\infty^2\right\} - \frac{1}{4\eta^t}\|w^t - w^{t+1}\|_1^2\,,
\end{aligned}$$

where the first inequality follows from Hölder's inequality and the $(1/\eta^t)$-strong convexity of ψ^t with respect to $\|\cdot\|_1$, and the last inequality follows from $b\sqrt{z} - az \le b^2/(4a)$ for $a > 0, b \ge 0$, and $z \ge 0$. Combining Lemma 2 with the above two inequalities and $\langle w^* - w^{T+1}, m^{T+1}\rangle = 0$ (by $m^{T+1} = \mathbf{0}$) completes the proof of Lemma 3. ■

The following lemma, which follows from Lemma 3 by choosing an adaptive learning rate, is a variant of well-known upper bounds in the literature on scale-free online learning (see, *e.g.,* Orabona & Pál, 2018). The difference is that we use OFTRL rather than FTRL here, and we keep the negative term in order to derive an RVU bound.

Lemma 4 (RVU bound for optimistic Hedge with AdaHedge-type learning rate)**.** *Suppose that the same assumptions as in Lemma 3 hold. We also assume that the learning rate is given by* $\eta^t = \sqrt{c^2 / \sum_{s=1}^{t-1}(\|h^s - m^s\|_\infty^2 + \nu^s)}$ *for some constant* $c \geq 2$ *and nonnegative reals* $\nu^1, \dots, \nu^T$*. When the denominator is zero, we set* $\eta^t = \infty$ *and interpret* $1/\eta^t = 0$*. Then, for any* $w^* \in \Delta_d$*,*

$$\sum_{t=1}^{T} \langle w^t - w^*, h^t \rangle \leq \left(\frac{\log d}{c} + 2\sqrt{2}c \right) \sqrt{\sum_{t=1}^{T} (\|h^t - m^t\|_\infty^2 + \nu^t)} - \sum_{t=1}^{T} \frac{1}{4\eta^t} \|w^t - w^{t+1}\|_1^2 .$$

In particular, when $c = c^* := \max\{2, \sqrt{\log d}/2^{3/4}\}$*, for any* $w^* \in \Delta_d$*,*

$$\sum_{t=1}^{T} \langle w^t - w^*, h^t \rangle \leq \sqrt{\sum_{t=1}^{T} (\|h^t - m^t\|_\infty^2 + \nu^t) \log_+(d)} - \sum_{t=1}^{T} \frac{1}{4\eta^t} \|w^t - w^{t+1}\|_1^2 ,$$

where we defined

$$\log_+(d) = \begin{cases} (\log(d)/2 + 4\sqrt{2})^2 & \text{if } \log d \leq 2^{7/2} \\ 2^{7/2} \log d & \text{if } \log d > 2^{7/2} \end{cases} = O(\log d) .$$

Proof. We will upper bound the RHS of (8) in Lemma 3. We denote $\Delta^t = \|h^t - m^t\|_\infty$ for simplicity. Then,

$$\begin{aligned} \min\{2\Delta^t, \eta^t (\Delta^t)^2\} &\leq \min\left\{ 2\Delta^t, \frac{c(\Delta^t)^2}{\sqrt{\sum_{s=1}^{t-1} (\Delta^s)^2}} \right\} = \sqrt{\min\left\{ 4(\Delta^t)^2, \frac{c^2 (\Delta^t)^4}{\sum_{s=1}^{t-1} (\Delta^s)^2} \right\}} \\ &\leq \sqrt{\frac{2}{\frac{1}{4(\Delta^t)^2} + \frac{\sum_{s=1}^{t-1} (\Delta^s)^2}{c^2 (\Delta^t)^4}}} = \frac{\sqrt{2} c (\Delta^t)^2}{\sqrt{(c^2/4)(\Delta^t)^2 + \sum_{s=1}^{t-1} (\Delta^s)^2}} , \end{aligned}$$

where we used the AdaHedge-type analysis in Orabona (2019). From this inequality, the second term in (8) is evaluated as

$$\begin{aligned} \sum_{t=1}^{T} \min\{2\Delta^t, \eta^t (\Delta^t)^2\} &\leq \sum_{t=1}^{T} \frac{\sqrt{2} c (\Delta^t)^2}{\sqrt{(c^2/4)(\Delta^t)^2 + \sum_{s=1}^{t-1} (\Delta^s)^2}} \\ &\leq 2\sqrt{2}c \sqrt{\sum_{t=1}^{T} (\Delta^t)^2} \leq 2\sqrt{2}c \sqrt{\sum_{t=1}^{T} ((\Delta^t)^2 + \nu^t)} , \end{aligned}$$

where the second inequality follows from the assumption that $c \geq 2$ and the fact that $\sum_{t=1}^{T} z^t / \sqrt{\sum_{s=1}^{t} z^s} \leq 2\sqrt{\sum_{t=1}^{T} z^t}$ for $z^1, \dots, z^T \geq 0$, and the last inequality from $\nu^t \geq 0$. Therefore, (8) implies

$$\sum_{t=1}^{T} \langle w^t - w^*, h^t \rangle \leq \left(\frac{\log d}{c} + 2\sqrt{2}c \right) \sqrt{\sum_{t=1}^{T} (\|h^t - m^t\|_\infty^2 + \nu^t)} - \sum_{t=1}^{T} \frac{1}{4\eta^t} \|w^t - w^{t+1}\|_1^2 , \quad (9)$$

which is the first desired inequality.

Now let $f(c) = \frac{\log d}{c} + 2\sqrt{2}c$ for $c \geq 2$. Then, $c^* = \max\{2, \sqrt{\log d}/2^{3/4}\}$ minimizes f, and the optimal value satisfies

$$f(c^*) = \begin{cases} \log(d)/2 + 4\sqrt{2} & \text{if } \log d \leq 2^{7/2} \\ 2^{7/4}\sqrt{\log d} & \text{if } \log d > 2^{7/2} \end{cases} \leq \begin{cases} 8\sqrt{2} & \text{if } \log d \leq 2^{7/2} \\ 2^{7/4}\sqrt{\log d} & \text{if } \log d > 2^{7/2} \end{cases} = O(\sqrt{\log d}) .$$

Since we have $\log_+(d) = (f(c^*))^2$, combining this with (9) gives the second desired upper bound, which completes the proof. ■

B Deferred Proofs for Two-Player Zero-Sum Games from Section 3

This section provides the details and deferred proofs from Section 3.

B.1 Auxiliary lemmas and proof of Lemma 1

The following lemma, used in Section 3, is a minor variant of the well-known regret upper bound for the AdaHedge-type update in (1).

Lemma 5. *Consider online linear optimization over the probability simplex Δ_d. Suppose that for some $\nu^1, \dots, \nu^T \geq 0$, the points are chosen by optimistic Hedge with the adaptive learning rate given by $w^t(a) \propto \exp\bigl(-\eta^t\bigl(\sum_{s=1}^{t-1} h^s(a) + h^{t-1}(a)\bigr)\bigr)$ for $a \in [d]$ and $\eta^t = \sqrt{D/\sum_{s=1}^{t-1}\bigl(\|h^s - m^s\|_\infty^2 + \nu^s\bigr)}$ for $D = \max\{4, \log(d)/2^{3/2}\}$, where $m^s = h^{s-1}$ is the optimistic prediction used in the above update (with $h^0 = \mathbf{0}$), and $m^{T+1} = \mathbf{0}$ for the update that defines w^{T+1}. When the denominator of η^t is zero, we set $\eta^t = \infty$ and interpret $1/\eta^t = 0$, in which case the update outputs the uniform distribution. Then, for any $w^* \in \Delta_d$, it holds that*

$$\sum_{t=1}^{T} \langle w^t - w^*, h^t \rangle \leq \sqrt{32 \sum_{t=1}^{T} (\|h^t - m^t\|_\infty^2 + \nu^t)\, D} - \sum_{t=1}^{T} \frac{1}{4\eta^t} \|w^t - w^{t+1}\|_1^2 .$$

Lemma 4 immediately yields Lemma 5 since $\log_+(d) \leq 32 \max\{4, \log d/2^{3/2}\}$.

The following lemma is useful for obtaining A_{diff}-dependent regret bounds.

Lemma 6. *For $m, n \in \mathbb{N}$ and $A \in \mathbb{R}^{m \times n}$, let $A_{\mathrm{diff}} = \max_{i,j} A_{i,j} - \min_{i,j} A_{i,j}$. Then, for any $x, x' \in \Delta_m$ and $y, y' \in \Delta_n$, it holds that $\|A(y - y')\|_\infty \leq \frac{\max_i(\max_j A_{ij} - \min_j A_{ij})}{2} \|y - y'\|_1 \leq \frac{A_{\mathrm{diff}}}{2} \|y - y'\|_1$ and $\|A^\top(x - x')\|_\infty \leq \frac{\max_j(\max_i A_{ij} - \min_i A_{ij})}{2} \|x - x'\|_1 \leq \frac{A_{\mathrm{diff}}}{2} \|x - x'\|_1$.*

Proof. We show the first inequality, and the second follows by applying the same argument to $A^\top$. Note that for any $\xi \in \mathbb{R}^m$, it holds that $\|A(y - y')\|_\infty = \|(A - \xi \mathbf{1}^\top)(y - y')\|_\infty = \max_{i \in [m]} |\langle A(i, \cdot) - \xi(i)\mathbf{1}, y - y' \rangle| \leq \max_{i \in [m]} \|A(i, \cdot) - \xi(i)\mathbf{1}\|_\infty \|y - y'\|_1$, where the first equality follows from $y, y' \in \Delta_n$ and the last inequality from Hölder's inequality. Choosing $\xi(i) = (\max_j A_{ij} + \min_j A_{ij})/2$ for each $i \in [m]$ completes the proof. ∎

Proof of Lemma 1. For simplicity, we write $\tau^\circ = \tau^\circ(h_\mu; c_1, c_2)$, omitting the dependence on h_μ, c_1, c_2 in this proof. We first bound the increments of the centered sequence $(h_\mu^t)_{t \geq 0}$:

$$\|h_\mu^1 - h_\mu^0\|_\infty \leq \frac{A_{\mathrm{diff}}}{2}, \qquad \|h_\mu^s - h_\mu^{s-1}\|_\infty \leq A_{\mathrm{diff}} \quad \text{for } s \geq 2 .$$

For the first bound, since each entry of $h^1 = Az^1$ is a convex combination of entries of A, we have $\max_a h^1(a) - \min_a h^1(a) \leq A_{\mathrm{diff}}$, and hence $\|h_\mu^1 - h_\mu^0\|_\infty = \|h^1 - \mu_h \mathbf{1}\|_\infty = \bigl(\max_a h^1(a) - \min_a h^1(a)\bigr)/2 \leq A_{\mathrm{diff}}/2$, where we used $h_\mu^0 = \mathbf{0}$ and the definition of μ_h. For $s \geq 2$, we have $h_\mu^s - h_\mu^{s-1} = h^s - h^{s-1} = A(z^s - z^{s-1})$, and hence Lemma 6 and $\|z^s - z^{s-1}\|_1 \leq 2$ for $z^s, z^{s-1} \in \Delta_d$ give $\|h_\mu^s - h_\mu^{s-1}\|_\infty \leq (A_{\mathrm{diff}}/2)\|z^s - z^{s-1}\|_1 \leq A_{\mathrm{diff}}$. In particular, every increment of $t \mapsto P_\infty^t(h_\mu)$ is bounded as

$$P_\infty^{t+1}(h_\mu) - P_\infty^t(h_\mu) = \|h_\mu^t - h_\mu^{t-1}\|_\infty^2 \leq A_{\mathrm{diff}}^2 \qquad \text{for } t \geq 1 . \tag{10}$$

We next show $\tau^\circ \geq 3$. Indeed, $P_\infty^1(h_\mu) = 0$ and $P_\infty^2(h_\mu) = \|h_\mu^1 - h_\mu^0\|_\infty^2 \leq A_{\mathrm{diff}}^2/4 \leq c_2/c_1 + A_{\mathrm{diff}}^2$, so the condition $P_\infty^t(h_\mu) > c_2/c_1 + A_{\mathrm{diff}}^2$ fails at $t \in \{1, 2\}$.

Suppose now that $\tau^\circ < \infty$, so that $\tau^\circ \in \{3, \dots, T+1\}$. By the minimality of τ°, we have

$$P_\infty^{\tau^\circ - 1}(h_\mu) \leq c_2/c_1 + A_{\mathrm{diff}}^2 .$$

Combining this with the increment bound (10), we obtain

$$P_\infty^{\tau^\circ}(h_\mu) = P_\infty^{\tau^\circ - 1}(h_\mu) + \|h_\mu^{\tau^\circ - 1} - h_\mu^{\tau^\circ - 2}\|_\infty^2 \leq \frac{c_2}{c_1} + 2A_{\mathrm{diff}}^2 .$$

Similarly, since the definition of τ° implies $P_\infty^{\tau^\circ}(h_\mu) > c_2/c_1 + A_{\mathrm{diff}}^2$, we have

$$P_\infty^{\tau^\circ - 1}(h_\mu) = P_\infty^{\tau^\circ}(h_\mu) - \|h_\mu^{\tau^\circ - 1} - h_\mu^{\tau^\circ - 2}\|_\infty^2 > \frac{c_2}{c_1} .$$

Combining the last two inequalities, we obtain

$$c_1 \sqrt{P_\infty^{\tau^\circ}(h_\mu)} + \frac{c_2}{\sqrt{P_\infty^{\tau^\circ - 1}(h_\mu)}} \leq c_1 \sqrt{\frac{c_2}{c_1} + 2A_{\mathrm{diff}}^2} + \frac{c_2}{\sqrt{c_2/c_1}} \leq \sqrt{2} c_1 A_{\mathrm{diff}} + 2\sqrt{c_1 c_2} ,$$

which completes the proof. ∎

B.2 Proof of Theorem 1

Here we provide the proof of Theorem 1 deferred from Section 3.3.

Proof of Theorem 1. If $A_{\mathrm{diff}} = 0$, then g^t and ℓ^t are constant vectors for all t, and hence $\mathsf{Reg}_x^T = \mathsf{Reg}_y^T = 0$. We thus assume $A_{\mathrm{diff}} > 0$ in what follows. Since $x^*, x^t \in \Delta_{m_x}$ and $y^*, y^t \in \Delta_{m_y}$, we have $\langle x^* - x^t, g^t \rangle = \langle x^* - x^t, g_\mu^t \rangle$ and $\langle y^t - y^*, \ell^t \rangle = \langle y^t - y^*, \ell_\mu^t \rangle$, and thus the regrets coincide with those of optimistic Hedge run on the loss sequences $(-g_\mu^t)_t$ and $(\ell_\mu^t)_t$ with optimistic predictions $-g_\mu^{t-1}$ and ℓ_μ^{t-1}, respectively. Let $\bar{A}_{\mathrm{diff}} = A_{\mathrm{diff}}/2$ for simplicity. From Lemma 6 and the identities $g_\mu^t - g_\mu^{t-1} = g^t - g^{t-1}$ and $\ell_\mu^t - \ell_\mu^{t-1} = \ell^t - \ell^{t-1}$ for $t \geq 2$, we have

$$\|g_\mu^t - g_\mu^{t-1}\|_\infty \leq \bar{A}_{\mathrm{diff}} \|y^t - y^{t-1}\|_1 \,, \quad \|\ell_\mu^t - \ell_\mu^{t-1}\|_\infty \leq \bar{A}_{\mathrm{diff}} \|x^t - x^{t-1}\|_1 \quad \text{for } t \geq 2 \,, \tag{11}$$

while the initial increments satisfy $\|g_\mu^1 - g_\mu^0\|_\infty \leq \bar{A}_{\mathrm{diff}}$ and $\|\ell_\mu^1 - \ell_\mu^0\|_\infty \leq \bar{A}_{\mathrm{diff}}$ (see the proof of Lemma 1 in Appendix B.1). We fix arbitrary $\tau_x, \tau_y \in \{2, \dots, T+1\}$. Then, applying Lemma 5 with $D = M_x$ and $\nu^t = \|\ell_\mu^t - \ell_\mu^{t-1}\|_\infty^2$, we obtain

$$\begin{aligned}
\mathsf{Reg}_x^T &\leq \sqrt{32(P_\infty^{T+1}(g_\mu) + P_\infty^{T+1}(\ell_\mu))M_x} - \sum_{t=1}^{T} \frac{1}{4\eta_x^t} \|x^{t+1} - x^t\|_1^2 \\
&\leq \sqrt{32\big[P_\infty^{\tau_y}(g_\mu) + \bar{A}_{\mathrm{diff}}^2 Q_1^{\tau_y}(y) + P_\infty^{\tau_x}(\ell_\mu) + \bar{A}_{\mathrm{diff}}^2 Q_1^{\tau_x}(x)\big] M_x} - \frac{1}{4\eta_x^{\tau_x - 1}} Q_1^{\tau_x}(x) \,,
\end{aligned} \tag{12}$$

where the last inequality holds for the following reasons. First, since $\tau_y \geq 2$, every increment of $P_\infty^{T+1}(g_\mu)$ with index $s \geq \tau_y$ satisfies $\|g_\mu^s - g_\mu^{s-1}\|_\infty^2 \leq \bar{A}_{\mathrm{diff}}^2 \|y^s - y^{s-1}\|_1^2$ by (11), and hence $P_\infty^{T+1}(g_\mu) \leq P_\infty^{\tau_y}(g_\mu) + \bar{A}_{\mathrm{diff}}^2 Q_1^{\tau_y}(y)$. Similarly, $P_\infty^{T+1}(\ell_\mu) \leq P_\infty^{\tau_x}(\ell_\mu) + \bar{A}_{\mathrm{diff}}^2 Q_1^{\tau_x}(x)$. Second, since $(\eta_x^t)_t$ is nonincreasing, we have $\sum_{t=1}^T \frac{1}{4\eta_x^t} \|x^{t+1} - x^t\|_1^2 \geq \sum_{s=\tau_x}^T \frac{1}{4\eta_x^{s-1}} \|x^s - x^{s-1}\|_1^2 \geq \frac{1}{4\eta_x^{\tau_x - 1}} Q_1^{\tau_x}(x)$, where $Q_1^{T+1}(x) = 0$ when $\tau_x = T+1$. Similarly,

$$\mathsf{Reg}_y^T \leq \sqrt{32\big[P_\infty^{\tau_y}(g_\mu) + \bar{A}_{\mathrm{diff}}^2 Q_1^{\tau_y}(y) + P_\infty^{\tau_x}(\ell_\mu) + \bar{A}_{\mathrm{diff}}^2 Q_1^{\tau_x}(x)\big] M_y} - \frac{1}{4\eta_y^{\tau_y - 1}} Q_1^{\tau_y}(y) \,. \tag{13}$$

Combining the upper bounds on Reg_x^T in (12) and Reg_y^T in (13), we obtain

$$\begin{aligned}
\mathsf{Reg}_x^T + \mathsf{Reg}_y^T &\leq \sqrt{128(P_\infty^{\tau_y}(g_\mu) + P_\infty^{\tau_x}(\ell_\mu))M} + \sqrt{128\bar{A}_{\mathrm{diff}}^2 Q_1^{\tau_y}(y) M} - \frac{1}{4\eta_y^{\tau_y - 1}} Q_1^{\tau_y}(y) \\
&\qquad + \sqrt{128\bar{A}_{\mathrm{diff}}^2 Q_1^{\tau_x}(x) M} - \frac{1}{4\eta_x^{\tau_x - 1}} Q_1^{\tau_x}(x) \\
&\leq \sqrt{128(P_\infty^{\tau_y}(g_\mu) + P_\infty^{\tau_x}(\ell_\mu))M} + 256\bar{A}_{\mathrm{diff}}^2 M(\eta_x^{\tau_x - 1} + \eta_y^{\tau_y - 1}) - \left(\frac{1}{8\eta_x^{\tau_x - 1}} Q_1^{\tau_x}(x) + \frac{1}{8\eta_y^{\tau_y - 1}} Q_1^{\tau_y}(y)\right) \\
&\leq 2^{7/2} \left[\sqrt{P_\infty^{\tau_x}(\ell_\mu) M} + \frac{2^{9/2} A_{\mathrm{diff}}^2 M^{3/2}}{\sqrt{P_\infty^{\tau_x - 1}(\ell_\mu)}} + \sqrt{P_\infty^{\tau_y}(g_\mu) M} + \frac{2^{9/2} A_{\mathrm{diff}}^2 M^{3/2}}{\sqrt{P_\infty^{\tau_y - 1}(g_\mu)}}\right] \\
&\qquad - \left(\frac{1}{8\eta_x^{\tau_x - 1}} Q_1^{\tau_x}(x) + \frac{1}{8\eta_y^{\tau_y - 1}} Q_1^{\tau_y}(y)\right) ,
\end{aligned} \tag{14}$$

where the second inequality follows by considering the worst cases with respect to $Q_1^{\tau_y}(y)$ and $Q_1^{\tau_x}(x)$ using the inequality $b\sqrt{z} - az \leq b^2/(4a)$ that holds for $a > 0$, $b \geq 0$, and $z \geq 0$, and the last inequality follows from the subadditivity of $\sqrt{\cdot}$, $\bar{A}_{\mathrm{diff}} = A_{\mathrm{diff}}/2$, and the definitions of the learning rates in (1): $\eta_x^{\tau_x - 1} \leq \sqrt{M_x / P_\infty^{\tau_x - 1}(\ell_\mu)}$ and $\eta_y^{\tau_y - 1} \leq \sqrt{M_y / P_\infty^{\tau_y - 1}(g_\mu)}$.

Now using the stopping time τ° in (3) of Lemma 1, we choose $\tau_x = \min\{\tau^\circ(\ell_\mu), T+1\}, \tau_y = \min\{\tau^\circ(g_\mu), T+1\} \in \{2, \dots, T+1\}$ for

$$\tau^\circ(\ell_\mu) \coloneqq \tau^\circ(\ell_\mu; \sqrt{M}, 2^{9/2} A_{\mathrm{diff}}^2 M^{3/2}) \,, \quad \tau^\circ(g_\mu) \coloneqq \tau^\circ(g_\mu; \sqrt{M}, 2^{9/2} A_{\mathrm{diff}}^2 M^{3/2}) \,.$$

We first consider the case when $\tau^\circ(\ell_\mu), \tau^\circ(g_\mu) < \infty$ and treat the remaining cases at the end of the proof. From Lemma 1, the terms in (14) are bounded as

$$\sqrt{P_\infty^{\tau_x}(\ell_\mu) M} + \frac{2^{9/2} A_{\mathrm{diff}}^2 M^{3/2}}{\sqrt{P_\infty^{\tau_x - 1}(\ell_\mu)}} \leq 11 A_{\mathrm{diff}} M \,, \; \sqrt{P_\infty^{\tau_y}(g_\mu) M} + \frac{2^{9/2} A_{\mathrm{diff}}^2 M^{3/2}}{\sqrt{P_\infty^{\tau_y - 1}(g_\mu)}} \leq 11 A_{\mathrm{diff}} M \,, \tag{15}$$

where we used $M \geq 1$. Substituting (15) into (14), we have

$$\mathsf{Reg}_x^T + \mathsf{Reg}_y^T \leq 2^{7/2} \cdot 22 A_{\mathrm{diff}} M - \left(\frac{1}{8\eta_x^{\tau_x - 1}} Q_1^{\tau_x}(x) + \frac{1}{8\eta_y^{\tau_y - 1}} Q_1^{\tau_y}(y) \right). \tag{16}$$

Now from Lemma 1 we have $P_\infty^{\tau_x - 1}(\ell_\mu) \geq 2^{9/2} A_{\mathrm{diff}}^2 M$ and thus

$$\eta_x^{\tau_x - 1} \leq \sqrt{M_x / P_\infty^{\tau_x - 1}(\ell_\mu)} \leq \sqrt{M_x / (2^{9/2} A_{\mathrm{diff}}^2 M)} \leq 1/(4.75 A_{\mathrm{diff}}) .$$

Therefore, combining (16) with the fact that $\mathsf{Reg}_x^T + \mathsf{Reg}_y^T \geq 0$ and the last inequality, we obtain

$$Q_1^{\tau_x}(x) \leq 8 \cdot 2^{7/2} \cdot 22\, \eta_x^{\tau_x - 1} A_{\mathrm{diff}} M \leq \frac{8 \cdot 2^{7/2} \cdot 22}{4.75} M \leq 420 M .$$

By the same argument, we also have $Q_1^{\tau_y}(y) \leq 420M$. From Lemma 1, we also have

$$P_\infty^{\tau_x}(\ell_\mu) \leq 2A_{\mathrm{diff}}^2 (8\sqrt{2}M + 1) , \qquad P_\infty^{\tau_y}(g_\mu) \leq 2A_{\mathrm{diff}}^2 (8\sqrt{2}M + 1) .$$

Finally, by substituting the above upper bounds on the squared path lengths $P_\infty^{\tau_x}(\ell_\mu)$, $P_\infty^{\tau_y}(g_\mu)$, $Q_1^{\tau_x}(x)$, and $Q_1^{\tau_y}(y)$ into (12) and (13), we obtain

$$\begin{aligned} \mathsf{Reg}_x^T &\leq \sqrt{32[4A_{\mathrm{diff}}^2 (8\sqrt{2}M + 1) + 2 \cdot 420 \bar{A}_{\mathrm{diff}}^2 M] M_x} \\ &\leq \sqrt{32 \cdot 256\, A_{\mathrm{diff}}^2 (M+1) M_x} \leq 91 A_{\mathrm{diff}} \sqrt{(M+1) M_x} , \end{aligned}$$

where we used $2 \cdot 420 \bar{A}_{\mathrm{diff}}^2 = 210 A_{\mathrm{diff}}^2$, $(4 \cdot 8\sqrt{2} + 210)M + 4 \leq 256(M+1)$, and $\sqrt{32 \cdot 256} = 2^{13/2} \leq 91$. Similarly, we obtain $\mathsf{Reg}_y^T \leq 91 A_{\mathrm{diff}} \sqrt{(M+1) M_y}$, which is the desired bound in the case $\tau^\circ(\ell_\mu), \tau^\circ(g_\mu) < \infty$.

It remains to consider the cases $\tau^\circ(\ell_\mu) = \infty$ or $\tau^\circ(g_\mu) = \infty$. If $\tau^\circ(\ell_\mu) = \infty$ and $\tau^\circ(g_\mu) < \infty$, then $\tau_x = T + 1$, and hence $Q_1^{\tau_x}(x) = 0$ and $P_\infty^{T+1}(\ell_\mu) \leq 2^{9/2} A_{\mathrm{diff}}^2 M + A_{\mathrm{diff}}^2$ by the definition of τ° in (3). The latter bound gives $\sqrt{128 P_\infty^{T+1}(\ell_\mu) M} \leq 2^{7/2} \cdot 5 A_{\mathrm{diff}} M \leq 2^{7/2} \cdot 11 A_{\mathrm{diff}} M$. Using this bound in the first line of (14), we again obtain (16), and the bounds after it hold with $Q_1^{\tau_x}(x) = 0$. The case $\tau^\circ(\ell_\mu) < \infty$ and $\tau^\circ(g_\mu) = \infty$ follows in the same manner with the roles of the two players swapped. If $\tau^\circ(\ell_\mu) = \tau^\circ(g_\mu) = \infty$, substituting $P_\infty^{T+1}(g_\mu), P_\infty^{T+1}(\ell_\mu) \leq 2^{9/2} A_{\mathrm{diff}}^2 M + A_{\mathrm{diff}}^2$ and $Q_1^{T+1}(x) = Q_1^{T+1}(y) = 0$ into (12) and (13) gives the bounds in Theorem 1, which completes the proof. ■

B.3 Corrupted procedure in two-player zero-sum games

We now introduce the corrupted regime of Tsuchiya et al. (2025), which allows opponent deviations and adversarially corrupted utilities. Let $\widehat{x}^t \in \Delta_{m_x}$ and $\widehat{y}^t \in \Delta_{m_y}$ denote the strategies suggested by the prescribed algorithm of the x-player and y-player, respectively, at round t, and let $x^t \in \Delta_{m_x}$ and $y^t \in \Delta_{m_y}$ denote the strategy actually chosen by each player at round t. We then summarize the corrupted learning procedure for a two-player zero-sum game with payoff matrix A, in which two settings, (I) and (II), differ in the payoff received in Step 4:

At each round $t = 1, \ldots, T$:

1. A prescribed algorithm suggests strategies $\widehat{x}^t \in \Delta_{m_x}$ and $\widehat{y}^t \in \Delta_{m_y}$;
2. The x-player selects a strategy $x^t \leftarrow \widehat{x}^t + \widehat{c}_x^t$ and the y-player selects $y^t \leftarrow \widehat{y}^t + \widehat{c}_y^t$;
3. The x-player observes a corrupted expected reward vector $\widetilde{g}^t = g^t + \widetilde{c}_x^t$ for $g^t = Ay^t$ and the y-player observes a corrupted expected loss vector $\widetilde{\ell}^t = \ell^t + \widetilde{c}_y^t$ for $\ell^t = A^\top x^t$;
4. The x-player gains a payoff of $\langle x^t, g^t \rangle$ in Setting (I) and $\langle x^t, \widetilde{g}^t \rangle$ in Setting (II), and the y-player incurs a loss of $\langle y^t, \ell^t \rangle$ in Setting (I) and $\langle y^t, \widetilde{\ell}^t \rangle$ in Setting (II).

Here, $\widehat{c}_x^t$ and $\widetilde{c}_x^t$ denote the x-player's strategy deviation and utility corruption at round t, respectively. These satisfy $\sum_{t=1}^T \|\widehat{c}_x^t\|_1 = \sum_{t=1}^T \|x^t - \widehat{x}^t\|_1 \leq \widehat{C}_x$ and $\sum_{t=1}^T \|\widetilde{c}_x^t\|_\infty = \sum_{t=1}^T \|g^t - \widetilde{g}^t\|_\infty \leq \widetilde{C}_x$, and we define $C_x = (1+\rho) A_{\max} \widehat{C}_x + 2\widetilde{C}_x$[2]. Similarly, $\widehat{c}_y^t$ and $\widetilde{c}_y^t$ denote the y-player's strategy

[2]The choice of the coefficients here follows from Proposition 1, where ρ is the constant of the bounded-corruption assumption (17) below.

deviation and utility corruption at round t, respectively, with $\sum_{t=1}^T \|\widehat{c}_y^t\|_1 = \sum_{t=1}^T \|y^t - \widehat{y}^t\|_1 \le \widehat{C}_y$ and $\sum_{t=1}^T \|\widetilde{c}_y^t\|_\infty = \sum_{t=1}^T \|\ell^t - \widetilde{\ell}^t\|_\infty \le \widetilde{C}_y$, and we define $C_y = (1+\rho)A_{\max}\widehat{C}_y + 2\widetilde{C}_y$. In addition to these total budgets, we consider the standard *bounded-corruption* model, in which the per-round utility corruptions are bounded: for a universal constant $\rho > 0$,

$$\|\widetilde{c}_x^t\|_\infty \le \rho A_{\max} \qquad \text{and} \qquad \|\widetilde{c}_y^t\|_\infty \le \rho A_{\max} \qquad \text{for all } t \in [T]\,. \tag{17}$$

This corresponds to the normalized setting of Tsuchiya et al. (2025), where $A_{\max} = 1$ and the corrupted observed utilities remain bounded as well. We emphasize that the assumption (17) is used only in the analysis. The algorithm itself requires no knowledge of $A_{\max}$ or ρ. Note that the corrupted regime with $\widehat{C}_x = \widehat{C}_y = \widetilde{C}_x = \widetilde{C}_y = 0$ corresponds to the *honest regime* presented in Section 2.2, in which there is no deviation from the prescribed strategies and no corruption in utilities. See Tsuchiya et al. (2025, Remark 8) for the motivation behind the two settings (I) and (II).

In the corrupted regime, the following four types of external regret can be defined:

$$\mathsf{Reg}_{x,g}^T(x^*) = \sum_{t=1}^T \langle x^* - x^t, g^t\rangle\,, \quad \mathsf{Reg}_{\widehat{x},g}^T(x^*) = \sum_{t=1}^T \langle x^* - \widehat{x}^t, g^t\rangle\,,$$
$$\mathsf{Reg}_{x,\widetilde{g}}^T(x^*) = \sum_{t=1}^T \langle x^* - x^t, \widetilde{g}^t\rangle\,, \quad \mathsf{Reg}_{\widehat{x},\widetilde{g}}^T(x^*) = \sum_{t=1}^T \langle x^* - \widehat{x}^t, \widetilde{g}^t\rangle\,.$$

As discussed in Tsuchiya et al. (2025, Remark 8), depending on the cause of the corruption in the utility vector g^t, it is natural to consider either $\mathsf{Reg}_{x,g}^T(x^*)$ or $\mathsf{Reg}_{x,\widetilde{g}}^T(x^*)$ as the evaluation metric for the player. Specifically, in Setting (I), it is natural to use $\mathsf{Reg}_{x,g}^T(x^*)$, while in Setting (II), it is natural to use $\mathsf{Reg}_{x,\widetilde{g}}^T(x^*)$. The four different types of external regret of the y-player, $\mathsf{Reg}_{y,\ell}^T(y^*)$, $\mathsf{Reg}_{\widehat{y},\ell}^T(y^*)$, $\mathsf{Reg}_{y,\widetilde{\ell}}^T(y^*)$, and $\mathsf{Reg}_{\widehat{y},\widetilde{\ell}}^T(y^*)$, can be similarly defined. Note that we have $\mathsf{Reg}_x^T(x^*) = \mathsf{Reg}_{x,g}^T(x^*)$ and $\mathsf{Reg}_y^T(y^*) = \mathsf{Reg}_{y,\ell}^T(y^*)$.

For these external regrets, the following inequalities hold. They are minor generalizations of Tsuchiya et al. (2025, Proposition 9) to the scale-agnostic setting.

Proposition 1. *Under the bounded-corruption assumption (17), for any $x^* \in \Delta_{m_x}$, $|\mathsf{Reg}_{x,g}^T(x^*) - \mathsf{Reg}_{\widehat{x},g}^T(x^*)| \le A_{\max}\widehat{C}_x$, $|\mathsf{Reg}_{x,\widetilde{g}}^T(x^*) - \mathsf{Reg}_{\widehat{x},\widetilde{g}}^T(x^*)| \le (1+\rho)A_{\max}\widehat{C}_x$, $|\mathsf{Reg}_{x,g}^T(x^*) - \mathsf{Reg}_{x,\widetilde{g}}^T(x^*)| \le 2\widetilde{C}_x$, and $|\mathsf{Reg}_{\widehat{x},g}^T(x^*) - \mathsf{Reg}_{\widehat{x},\widetilde{g}}^T(x^*)| \le 2\widetilde{C}_x$. Similar inequalities hold for $\mathsf{Reg}_{y,\ell}^T(y^*), \mathsf{Reg}_{\widehat{y},\ell}^T(y^*), \mathsf{Reg}_{y,\widetilde{\ell}}^T(y^*), \mathsf{Reg}_{\widehat{y},\widetilde{\ell}}^T(y^*)$.*

Proof of Proposition 1. From the triangle inequality and Hölder's inequality, we have

$$\left|\mathsf{Reg}_{x,g}^T(x^*) - \mathsf{Reg}_{\widehat{x},g}^T(x^*)\right| = \left|\sum_{t=1}^T \langle \widehat{x}^t - x^t, g^t\rangle\right| \le \sum_{t=1}^T \|\widehat{x}^t - x^t\|_1 \|g^t\|_\infty \le A_{\max}\widehat{C}_x\,.$$

Similarly, we also have

$$\left|\mathsf{Reg}_{x,g}^T(x^*) - \mathsf{Reg}_{x,\widetilde{g}}^T(x^*)\right| = \left|\sum_{t=1}^T \langle x^* - x^t, g^t - \widetilde{g}^t\rangle\right| \le \sum_{t=1}^T \|x^* - x^t\|_1 \|g^t - \widetilde{g}^t\|_\infty \le 2\widetilde{C}_x\,.$$

For the comparison under the corrupted observed utilities, the bounded-corruption assumption (17) gives $\|\widetilde{g}^t\|_\infty \le \|g^t\|_\infty + \|\widetilde{c}_x^t\|_\infty \le (1+\rho)A_{\max}$, and hence

$$\left|\mathsf{Reg}_{x,\widetilde{g}}^T(x^*) - \mathsf{Reg}_{\widehat{x},\widetilde{g}}^T(x^*)\right| = \left|\sum_{t=1}^T \langle \widehat{x}^t - x^t, \widetilde{g}^t\rangle\right| \le \sum_{t=1}^T \|\widehat{x}^t - x^t\|_1 \|\widetilde{g}^t\|_\infty \le (1+\rho)A_{\max}\widehat{C}_x\,.$$

The other inequalities can be proven in the same manner, which completes the proof. ■

B.4 Extension of Theorem 1 to the corrupted regime

Here we provide the extension of Theorem 1 to the corrupted regime.

We use the algorithms for the corrupted regime, which simply replace the non-corrupted (unobserved) gradients g^t and ℓ^t in (1) with the corrupted (observable) gradients $\widetilde{g}^t$ and $\widetilde{\ell}^t$. For

$$\mu_{\widetilde{g}} = \frac{\max_{a\in[m_x]} \widetilde{g}^1(a) + \min_{a\in[m_x]} \widetilde{g}^1(a)}{2}\,, \qquad \mu_{\widetilde{\ell}} = \frac{\max_{b\in[m_y]} \widetilde{\ell}^1(b) + \min_{b\in[m_y]} \widetilde{\ell}^1(b)}{2}\,,$$

we define the centered corrupted gradient sequences

$$\widetilde{g}_\mu^0 = \mathbf{0}\,, \quad \widetilde{g}_\mu^t = \widetilde{g}^t - \mu_{\widetilde{g}}\mathbf{1} \quad \text{for } t \geq 1\,, \qquad \widetilde{\ell}_\mu^0 = \mathbf{0}\,, \quad \widetilde{\ell}_\mu^t = \widetilde{\ell}^t - \mu_{\widetilde{\ell}}\mathbf{1} \quad \text{for } t \geq 1\,.$$

Then, the strategies of the x- and y-players, $(\widehat{x}^t)_{t=1}^T$ and $(\widehat{y}^t)_{t=1}^T$, are determined as follows:

$$\begin{aligned}
\widehat{x}^t(a) &\propto \exp\left(\eta_x^t\left(\sum_{s=1}^{t-1}\widetilde{g}_\mu^s(a) + \widetilde{g}_\mu^{t-1}(a)\right)\right), \quad \eta_x^t = \sqrt{\frac{M_x}{P_\infty^t(\widetilde{g}_\mu) + P_\infty^t(\widetilde{\ell}_\mu)}}\,, \\
\widehat{y}^t(b) &\propto \exp\left(-\eta_y^t\left(\sum_{s=1}^{t-1}\widetilde{\ell}_\mu^s(b) + \widetilde{\ell}_\mu^{t-1}(b)\right)\right), \quad \eta_y^t = \sqrt{\frac{M_y}{P_\infty^t(\widetilde{g}_\mu) + P_\infty^t(\widetilde{\ell}_\mu)}}\,,
\end{aligned} \tag{18}$$

where η_x^t, η_y^t are the learning rates of the x- and y-players at round t. As in the honest regime, if the denominators of the learning rates are zero, then all the centered corrupted gradients observed so far are zero vectors, and we set $\widehat{x}^t = \frac{1}{m_x}\mathbf{1}$ and $\widehat{y}^t = \frac{1}{m_y}\mathbf{1}$, which corresponds to running (18) with an arbitrary finite learning rate. Note that $\widehat{x}^1 = \frac{1}{m_x}\mathbf{1}$ and $\widehat{y}^1 = \frac{1}{m_y}\mathbf{1}$, and recall that $\widetilde{g}^t = g^t + \widetilde{c}_x^t$ for $g^t = Ay^t$ and $\widetilde{\ell}^t = \ell^t + \widetilde{c}_y^t$ for $\ell^t = A^\top x^t$.

The learning dynamic specified by (18) guarantees the following bounds.

Theorem 3 (Scale-free fast convergence to Nash equilibrium in the corrupted regime)**.** *Suppose that the x- and y-players use the algorithms in (18) to obtain strategies $(\widehat{x}^t)_{t=1}^T$ and $(\widehat{y}^t)_{t=1}^T$, respectively. Then, in the corrupted regime under the bounded-corruption assumption (17), it holds that*

$$\begin{aligned}
\mathsf{Reg}_{x,\widetilde{g}}^T &= O\left(A_{\mathrm{diff}}\sqrt{\log m \log m_x} + \sqrt{\left[A_{\max}\widetilde{C} + A_{\mathrm{diff}}^2\widehat{C} + A_{\mathrm{diff}}C\right]\log m_x} + A_{\max}\widehat{C}_x\right), \\
\mathsf{Reg}_{y,\widetilde{\ell}}^T &= O\left(A_{\mathrm{diff}}\sqrt{\log m \log m_y} + \sqrt{\left[A_{\max}\widetilde{C} + A_{\mathrm{diff}}^2\widehat{C} + A_{\mathrm{diff}}C\right]\log m_y} + A_{\max}\widehat{C}_y\right),
\end{aligned}$$

and

$$\begin{aligned}
\mathsf{Reg}_{x,g}^T &= O\left(A_{\mathrm{diff}}\sqrt{\log m \log m_x} + \sqrt{\left[A_{\max}\widetilde{C} + A_{\mathrm{diff}}^2\widehat{C} + A_{\mathrm{diff}}C\right]\log m_x} + C_x\right), \\
\mathsf{Reg}_{y,\ell}^T &= O\left(A_{\mathrm{diff}}\sqrt{\log m \log m_y} + \sqrt{\left[A_{\max}\widetilde{C} + A_{\mathrm{diff}}^2\widehat{C} + A_{\mathrm{diff}}C\right]\log m_y} + C_y\right),
\end{aligned}$$

where $\widetilde{C} = \widetilde{C}_x + \widetilde{C}_y$, $\widehat{C} = \widehat{C}_x + \widehat{C}_y$, and $C = C_x + C_y$.

Note that a linear dependence on a player's own corruption level is unavoidable since a matching lower bound is known (Tsuchiya et al., 2025, Theorem 15 (ii)). In contrast, the opponent's strategic deviations enter only through a square root. Under the bounded-corruption assumption (17), the utility-corruption budgets $\widetilde{C}_x, \widetilde{C}_y$ also enter only through a square root in Setting (II). In Setting (I), a player's own utility corruption additionally contributes $2\widetilde{C}_x$ through C_x for the x-player and $2\widetilde{C}_y$ through C_y for the y-player. When $A_{\max} = 1$, the term $A_{\max}(\widetilde{C}_x + \widetilde{C}_y)$ reduces to $\widetilde{C}_x + \widetilde{C}_y$, as in Tsuchiya et al. (2025). The square-root dependence on the opponent's strategic deviations and on utility corruption in Setting (II) is optimal (Tsuchiya et al., 2025, Theorem 15 (i), (iii)). We also remark that our path-length-based analysis uses the assumption (17) to convert the budget $\sum_{t=1}^T\|\widetilde{c}_x^t\|_\infty$ into a bound on $\sum_t\|\widetilde{c}_x^t - \widetilde{c}_x^{t-1}\|_\infty^2$. If one only assumes the ℓ_1 budgets $\sum_{t=1}^T\|\widetilde{c}_x^t\|_\infty \leq \widetilde{C}_x$ without any per-round bound, then a single large corruption of magnitude $\widetilde{C}_x$ makes $\sum_t\|\widetilde{c}_x^t - \widetilde{c}_x^{t-1}\|_\infty^2$ of order $\widetilde{C}_x^2$, and the analysis below yields only a linear dependence on the corruption budgets.

To prove Theorem 3, we recall that from Lemma 6, for $t \geq 2$ we have

$$\|g^t - g^{t-1}\|_\infty \leq (A_{\mathrm{diff}}/2)\|y^t - y^{t-1}\|_1\,, \quad \|\ell^t - \ell^{t-1}\|_\infty \leq (A_{\mathrm{diff}}/2)\|x^t - x^{t-1}\|_1\,. \tag{19}$$

Note also that, since the centering shifts are constant over rounds, the increments for $t \geq 2$ satisfy $\widetilde{g}_\mu^t - \widetilde{g}_\mu^{t-1} = \widetilde{g}^t - \widetilde{g}^{t-1}$ and $\widetilde{\ell}_\mu^t - \widetilde{\ell}_\mu^{t-1} = \widetilde{\ell}^t - \widetilde{\ell}^{t-1}$. In particular, for any $\tau \geq 2$ we have $Q_\infty^\tau(\widetilde{g}_\mu) = Q_\infty^\tau(\widetilde{g})$ and $Q_\infty^\tau(\widetilde{\ell}_\mu) = Q_\infty^\tau(\widetilde{\ell})$.

We next state the following lemma, which gives upper bounds on $Q_\infty^{\tau_y}(\widetilde{g}_\mu)$ and $Q_\infty^{\tau_x}(\widetilde{\ell}_\mu)$.

Lemma 7. *Under the bounded-corruption assumption (17), for any $\tau_x, \tau_y \in \{2, \dots, T+1\}$, it holds that*

$$Q_\infty^{\tau_y}(\widetilde{g}_\mu) \le A_{\text{diff}}^2 \big(Q_1^{\tau_y}(\widehat{y}) + 8\widehat{C}_y\big) + 8\rho A_{\max}\widetilde{C}_x\,,$$
$$Q_\infty^{\tau_x}(\widetilde{\ell}_\mu) \le A_{\text{diff}}^2 \big(Q_1^{\tau_x}(\widehat{x}) + 8\widehat{C}_x\big) + 8\rho A_{\max}\widetilde{C}_y\,.$$

Proof. We prove the bound for $\widetilde{g}_\mu$. The bound for $\widetilde{\ell}_\mu$ follows by the same argument. Since $Q_\infty^{\tau_y}(\widetilde{g}_\mu) = Q_\infty^{\tau_y}(\widetilde{g})$ and $\widetilde{g}^s = g^s + \widetilde{c}_x^s$, using $\|a+b\|_\infty^2 \le 2\|a\|_\infty^2 + 2\|b\|_\infty^2$, we have

$$Q_\infty^{\tau_y}(\widetilde{g}_\mu) = \sum_{s=\tau_y}^{T} \|\widetilde{g}^s - \widetilde{g}^{s-1}\|_\infty^2 \le 2\sum_{s=\tau_y}^{T} \|g^s - g^{s-1}\|_\infty^2 + 2\sum_{s=\tau_y}^{T} \|\widetilde{c}_x^s - \widetilde{c}_x^{s-1}\|_\infty^2\,.$$

The second sum is bounded as

$$2\sum_{s=\tau_y}^{T} \|\widetilde{c}_x^s - \widetilde{c}_x^{s-1}\|_\infty^2 \le 4\sum_{s=\tau_y}^{T} \big(\|\widetilde{c}_x^s\|_\infty^2 + \|\widetilde{c}_x^{s-1}\|_\infty^2\big) \le 8\sum_{t=1}^{T} \|\widetilde{c}_x^t\|_\infty^2 \le 8\rho A_{\max}\sum_{t=1}^{T} \|\widetilde{c}_x^t\|_\infty \le 8\rho A_{\max}\widetilde{C}_x\,,$$

where the third inequality follows from the bound (17). For the first sum, by (19) and the triangle inequality $\|y^s - y^{s-1}\|_1 \le \|\widehat{c}_y^s\|_1 + \|\widehat{y}^s - \widehat{y}^{s-1}\|_1 + \|\widehat{c}_y^{s-1}\|_1$, we have

$$\begin{aligned}
2\sum_{s=\tau_y}^{T} \|g^s - g^{s-1}\|_\infty^2 &\le \frac{A_{\text{diff}}^2}{2}\sum_{s=\tau_y}^{T} \|y^s - y^{s-1}\|_1^2 \\
&\le \frac{A_{\text{diff}}^2}{2}\sum_{s=\tau_y}^{T} \big(2\|\widehat{y}^s - \widehat{y}^{s-1}\|_1^2 + 4\|\widehat{c}_y^s\|_1^2 + 4\|\widehat{c}_y^{s-1}\|_1^2\big) \\
&\le \frac{A_{\text{diff}}^2}{2}\left(2Q_1^{\tau_y}(\widehat{y}) + 8\sum_{t=1}^{T}\|\widehat{c}_y^t\|_1^2\right) \le A_{\text{diff}}^2\big(Q_1^{\tau_y}(\widehat{y}) + 8\widehat{C}_y\big)\,,
\end{aligned}$$

where the second inequality follows from $(a+b+c)^2 \le 2b^2 + 4a^2 + 4c^2$, and the last inequality from $\sum_{t=1}^{T}\|\widehat{c}_y^t\|_1^2 \le 2\sum_{t=1}^{T}\|\widehat{c}_y^t\|_1 \le 2\widehat{C}_y$, which holds since $\|\widehat{c}_y^t\|_1 = \|y^t - \widehat{y}^t\|_1 \le 2$. Combining the last three displays completes the proof. ∎

Lemma 1 does not directly apply to the corrupted sequences $\widetilde{g}_\mu, \widetilde{\ell}_\mu$, since their increments contain the terms $\widetilde{c}_x^t - \widetilde{c}_x^{t-1}$ and $\widetilde{c}_y^t - \widetilde{c}_y^{t-1}$, respectively, and are thus no longer bounded by A_{diff}. We instead use the following variant, in which $\widetilde{D}$ is an upper bound on the increments of the centered corrupted sequence.

Lemma 8. *Let $(\widetilde{h}^t)_{t=1}^T$ be a sequence of the form $\widetilde{h}^t = Az^t + c^t$ for $z^t \in \Delta_d$ and arbitrary $c^t \in \mathbb{R}^d$, and let $c_1, c_2 > 0$. Define $\mu_{\widetilde{h}} = (\max_a \widetilde{h}^1(a) + \min_a \widetilde{h}^1(a))/2$, $\widetilde{h}_\mu^0 = \mathbf{0}$, $\widetilde{h}_\mu^t = \widetilde{h}^t - \mu_{\widetilde{h}}\mathbf{1}$ for $t \ge 1$, and*

$$\widetilde{D} \coloneqq A_{\text{diff}} + 2\max_{t\in[T]}\|c^t\|_\infty\,.$$

Define $\widetilde{\tau}^\circ(\widetilde{h}_\mu; c_1, c_2) \in \{2, \dots, T+1\} \cup \{\infty\}$ as in (3) with A_{diff} replaced by $\widetilde{D}$, that is, $\widetilde{\tau}^\circ = \min\big\{t \in \{2, \dots, T+1\} : P_\infty^t(\widetilde{h}_\mu) > c_2/c_1 + \widetilde{D}^2\big\}$ if $P_\infty^{T+1}(\widetilde{h}_\mu) > c_2/c_1 + \widetilde{D}^2$, and $\widetilde{\tau}^\circ = \infty$ otherwise. Then, we have $\widetilde{\tau}^\circ \ge 3$, and if $\widetilde{\tau}^\circ < \infty$, it holds that

$$P_\infty^{\widetilde{\tau}^\circ}(\widetilde{h}_\mu) \le \frac{c_2}{c_1} + 2\widetilde{D}^2\,,\ P_\infty^{\widetilde{\tau}^\circ - 1}(\widetilde{h}_\mu) \ge \frac{c_2}{c_1}\,,\ c_1\sqrt{P_\infty^{\widetilde{\tau}^\circ}(\widetilde{h}_\mu)} + \frac{c_2}{\sqrt{P_\infty^{\widetilde{\tau}^\circ - 1}(\widetilde{h}_\mu)}} \le \sqrt{2}c_1\widetilde{D} + 2\sqrt{c_1 c_2}\,.$$

Proof. It suffices to show the increment bounds

$$\|\widetilde{h}_\mu^1 - \widetilde{h}_\mu^0\|_\infty \le \widetilde{D}/2\,, \qquad \|\widetilde{h}_\mu^s - \widetilde{h}_\mu^{s-1}\|_\infty \le \widetilde{D} \quad \text{for } s \ge 2\,,$$

since the rest of the argument is then identical to the proof of Lemma 1 with A_{diff} replaced by $\widetilde{D}$. For the first bound, since each entry of Az^1 is a convex combination of entries of A and $\|c^1\|_\infty \le \max_t\|c^t\|_\infty$, we have $\max_a \widetilde{h}^1(a) - \min_a \widetilde{h}^1(a) \le A_{\text{diff}} + 2\|c^1\|_\infty \le \widetilde{D}$, and hence $\|\widetilde{h}_\mu^1 - \widetilde{h}_\mu^0\|_\infty = \big(\max_a \widetilde{h}^1(a) - \min_a \widetilde{h}^1(a)\big)/2 \le \widetilde{D}/2$. For $s \ge 2$, we have $\widetilde{h}_\mu^s - \widetilde{h}_\mu^{s-1} = A(z^s - z^{s-1}) + (c^s - c^{s-1})$, and hence, by Lemma 6 and the triangle inequality, $\|\widetilde{h}_\mu^s - \widetilde{h}_\mu^{s-1}\|_\infty \le A_{\text{diff}} + \|c^s\|_\infty + \|c^{s-1}\|_\infty \le \widetilde{D}$, which completes the proof. ∎

Now we are ready to prove Theorem 3.

Proof of Theorem 3. If $A_{\mathrm{diff}} = 0$, then g^t and ℓ^t are constant vectors, so that $\mathsf{Reg}_{x,g}^T = \mathsf{Reg}_{y,\ell}^T = 0$. Moreover, in this case all the increments of $\widetilde{g}_\mu$ come from the corruptions. Since the entries of $\widetilde{g}^1$ have range at most $2\|\widetilde{c}_x^1\|_\infty$, we have $\|\widetilde{g}_\mu^1 - \widetilde{g}_\mu^0\|_\infty \le \|\widetilde{c}_x^1\|_\infty \le \min\{\rho A_{\max}, \widetilde{C}_x\}$, and hence, together with the argument in the proof of Lemma 7, $P_\infty^{T+1}(\widetilde{g}_\mu) \le \min\{\rho A_{\max}, \widetilde{C}_x\}^2 + 8\rho A_{\max}\widetilde{C}_x \le 9\rho A_{\max}\widetilde{C}_x$, and similarly $P_\infty^{T+1}(\widetilde{\ell}_\mu) \le 9\rho A_{\max}\widetilde{C}_y$. Plugging these bounds into (20) and (22) below, which do not rely on $A_{\mathrm{diff}} > 0$, and using Proposition 1 yields the bounds stated in Theorem 3. We thus assume $A_{\mathrm{diff}} > 0$ in what follows. Since $\widehat{x}^t, x^* \in \Delta_{m_x}$ and $\widehat{y}^t, y^* \in \Delta_{m_y}$, we have $\langle x^* - \widehat{x}^t, \widetilde{g}^t\rangle = \langle x^* - \widehat{x}^t, \widetilde{g}_\mu^t\rangle$ and $\langle \widehat{y}^t - y^*, \widetilde{\ell}^t\rangle = \langle \widehat{y}^t - y^*, \widetilde{\ell}_\mu^t\rangle$. Hence, the algorithm in (18) is optimistic Hedge run on the centered loss sequences $(-\widetilde{g}_\mu^t)_t$ and $(\widetilde{\ell}_\mu^t)_t$ with optimistic predictions $-\widetilde{g}_\mu^{t-1}$ and $\widetilde{\ell}_\mu^{t-1}$. Its regrets $\mathsf{Reg}_{\widehat{x},\widetilde{g}}^T$ and $\mathsf{Reg}_{\widehat{y},\widetilde{\ell}}^T$ can be bounded by Lemma 5 applied to these centered sequences. Now we fix arbitrary $\tau_x, \tau_y \in \{2, \dots, T+1\}$. Then, from Lemma 5, we have

$$\mathsf{Reg}_{\widehat{x},\widetilde{g}}^T \le \sqrt{32(P_\infty^{T+1}(\widetilde{g}_\mu) + P_\infty^{T+1}(\widetilde{\ell}_\mu))M_x} - \sum_{t=1}^T \frac{1}{4\eta_x^t}\|\widehat{x}^{t+1} - \widehat{x}^t\|_1^2 \tag{20}$$

$$= \sqrt{32(P_\infty^{\tau_y}(\widetilde{g}_\mu) + P_\infty^{\tau_x}(\widetilde{\ell}_\mu) + Q_\infty^{\tau_y}(\widetilde{g}_\mu) + Q_\infty^{\tau_x}(\widetilde{\ell}_\mu))M_x} - \sum_{t=1}^T \frac{1}{4\eta_x^t}\|\widehat{x}^{t+1} - \widehat{x}^t\|_1^2$$

$$\le \sqrt{32\Big[P_\infty^{\tau_y}(\widetilde{g}_\mu) + P_\infty^{\tau_x}(\widetilde{\ell}_\mu) + A_{\mathrm{diff}}^2\big(Q_1^{\tau_x}(\widehat{x}) + Q_1^{\tau_y}(\widehat{y}) + 8\widehat{C}_x + 8\widehat{C}_y\big) + 8\rho A_{\max}(\widetilde{C}_x + \widetilde{C}_y)\Big]M_x} - \frac{1}{4\eta_x^{\tau_x - 1}}Q_1^{\tau_x}(\widehat{x})\,, \tag{21}$$

where the equality follows from $P_\infty^{T+1}(z) = P_\infty^\tau(z) + Q_\infty^\tau(z)$, and the last inequality follows from Lemma 7 and the inequality

$$\sum_{t=1}^T \frac{1}{4\eta_x^t}\|\widehat{x}^{t+1} - \widehat{x}^t\|_1^2 \ge \sum_{s=\tau_x}^T \frac{1}{4\eta_x^{s-1}}\|\widehat{x}^s - \widehat{x}^{s-1}\|_1^2 \ge \frac{1}{4\eta_x^{\tau_x - 1}}Q_1^{\tau_x}(\widehat{x})$$

since $(\eta_x^t)_t$ is nonincreasing and $\tau_x \ge 2$. Note that $Q_1^{T+1}(\widehat{x}) = 0$ if $\tau_x = T+1$. Similarly,

$$\mathsf{Reg}_{\widehat{y},\widetilde{\ell}}^T \le \sqrt{32(P_\infty^{T+1}(\widetilde{g}_\mu) + P_\infty^{T+1}(\widetilde{\ell}_\mu))M_y} - \sum_{t=1}^T \frac{1}{4\eta_y^t}\|\widehat{y}^{t+1} - \widehat{y}^t\|_1^2 \tag{22}$$

$$\le \sqrt{32\Big[P_\infty^{\tau_y}(\widetilde{g}_\mu) + P_\infty^{\tau_x}(\widetilde{\ell}_\mu) + A_{\mathrm{diff}}^2\big(Q_1^{\tau_x}(\widehat{x}) + Q_1^{\tau_y}(\widehat{y}) + 8\widehat{C}_x + 8\widehat{C}_y\big) + 8\rho A_{\max}(\widetilde{C}_x + \widetilde{C}_y)\Big]M_y} - \frac{1}{4\eta_y^{\tau_y - 1}}Q_1^{\tau_y}(\widehat{y})\,. \tag{23}$$

In what follows, we provide the regret upper bounds on $\mathsf{Reg}_{x,g}^T$ and $\mathsf{Reg}_{y,\ell}^T$, and the regret bounds for $\mathsf{Reg}_{x,\widetilde{g}}^T$ and $\mathsf{Reg}_{y,\widetilde{\ell}}^T$ can be obtained in a similar manner. Define

$$\kappa = \sqrt{128\big[8A_{\mathrm{diff}}^2(\widehat{C}_x + \widehat{C}_y) + 8\rho A_{\max}(\widetilde{C}_x + \widetilde{C}_y)\big]M} + C_x + C_y\,. \tag{24}$$

From Proposition 1, we have $\mathsf{Reg}_{x,g}^T \le \mathsf{Reg}_{\widehat{x},\widetilde{g}}^T + A_{\max}\widehat{C}_x + 2\widetilde{C}_x \le \mathsf{Reg}_{\widehat{x},\widetilde{g}}^T + C_x$ and $\mathsf{Reg}_{y,\ell}^T \le \mathsf{Reg}_{\widehat{y},\widetilde{\ell}}^T + C_y$. Then, by combining the upper bounds in (21) and (23) and using the subadditivity of

$z \mapsto \sqrt{z}$, we obtain

$$
\begin{aligned}
&\mathsf{Reg}_{x,g}^T + \mathsf{Reg}_{y,\ell}^T \\
&\le \sqrt{128(P_\infty^{\tau_y}(\widetilde{g}_\mu) + P_\infty^{\tau_x}(\widetilde{\ell}_\mu))M} \\
&\quad + \left(\sqrt{128A_{\text{diff}}^2 Q_1^{\tau_x}(\widehat{x})M} - \frac{1}{8\eta_x^{\tau_x - 1}} Q_1^{\tau_x}(\widehat{x})\right) + \left(\sqrt{128A_{\text{diff}}^2 Q_1^{\tau_y}(\widehat{y})M} - \frac{1}{8\eta_y^{\tau_y - 1}} Q_1^{\tau_y}(\widehat{y})\right) \\
&\quad - \left(\frac{1}{8\eta_x^{\tau_x - 1}} Q_1^{\tau_x}(\widehat{x}) + \frac{1}{8\eta_y^{\tau_y - 1}} Q_1^{\tau_y}(\widehat{y})\right) + \kappa \qquad (25) \\
&\le \sqrt{128(P_\infty^{\tau_y}(\widetilde{g}_\mu) + P_\infty^{\tau_x}(\widetilde{\ell}_\mu))M} + 256A_{\text{diff}}^2 M(\eta_x^{\tau_x - 1} + \eta_y^{\tau_y - 1}) \\
&\quad - \left(\frac{1}{8\eta_x^{\tau_x - 1}} Q_1^{\tau_x}(\widehat{x}) + \frac{1}{8\eta_y^{\tau_y - 1}} Q_1^{\tau_y}(\widehat{y})\right) + \kappa \\
&\le 2^{7/2}\left(\sqrt{P_\infty^{\tau_x}(\widetilde{\ell}_\mu)M} + \frac{2^{9/2}A_{\text{diff}}^2 M^{3/2}}{\sqrt{P_\infty^{\tau_x - 1}(\widetilde{\ell}_\mu)}}\right) + 2^{7/2}\left(\sqrt{P_\infty^{\tau_y}(\widetilde{g}_\mu)M} + \frac{2^{9/2}A_{\text{diff}}^2 M^{3/2}}{\sqrt{P_\infty^{\tau_y - 1}(\widetilde{g}_\mu)}}\right) \\
&\qquad - \left(\frac{1}{8\eta_x^{\tau_x - 1}} Q_1^{\tau_x}(\widehat{x}) + \frac{1}{8\eta_y^{\tau_y - 1}} Q_1^{\tau_y}(\widehat{y})\right) + \kappa\,, \qquad (26)
\end{aligned}
$$

where the second inequality follows by considering the worst cases with respect to $Q_1^{\tau_y}(\widehat{y})$ and $Q_1^{\tau_x}(\widehat{x})$ using the inequality $b\sqrt{z} - az \le b^2/(4a)$ that holds for $a > 0$, $b \ge 0$, and $z \ge 0$, and the last inequality follows from the subadditivity of $z \mapsto \sqrt{z}$ for $z \ge 0$ and $\eta_x^{\tau_x - 1} \le \sqrt{M_x / P_\infty^{\tau_x - 1}(\widetilde{\ell}_\mu)}$ and $\eta_y^{\tau_y - 1} \le \sqrt{M_y / P_\infty^{\tau_y - 1}(\widetilde{g}_\mu)}$, which follow from the definitions of the learning rates in (18).

To evaluate the first two terms in the last inequality, define

$$
\begin{aligned}
\widetilde{D}_g &\coloneqq A_{\text{diff}} + 2\max_{t \in [T]} \|\widetilde{c}_x^t\|_\infty \le A_{\text{diff}} + 2\min\{\rho A_{\max}, \widetilde{C}_x\}\,, \\
\widetilde{D}_\ell &\coloneqq A_{\text{diff}} + 2\max_{t \in [T]} \|\widetilde{c}_y^t\|_\infty \le A_{\text{diff}} + 2\min\{\rho A_{\max}, \widetilde{C}_y\}\,, \qquad (27)
\end{aligned}
$$

and choose $\tau_x, \tau_y \in \{2, \dots, T+1\}$ by

$$
\begin{aligned}
\tau_x &= \min\{\widetilde{\tau}^\circ(\widetilde{\ell}_\mu), T+1\}\,, \quad \widetilde{\tau}^\circ(\widetilde{\ell}_\mu) \coloneqq \widetilde{\tau}^\circ(\widetilde{\ell}_\mu; \sqrt{M}, 2^{9/2}A_{\text{diff}}^2 M^{3/2})\,, \\
\tau_y &= \min\{\widetilde{\tau}^\circ(\widetilde{g}_\mu), T+1\}\,, \quad \widetilde{\tau}^\circ(\widetilde{g}_\mu) \coloneqq \widetilde{\tau}^\circ(\widetilde{g}_\mu; \sqrt{M}, 2^{9/2}A_{\text{diff}}^2 M^{3/2})\,, \qquad (28)
\end{aligned}
$$

where $\widetilde{\tau}^\circ$ is the stopping time of Lemma 8, applied to $\widetilde{\ell}^t = A^\top x^t + \widetilde{c}_y^t$ with $\widetilde{D} = \widetilde{D}_\ell$ and to $\widetilde{g}^t = Ay^t + \widetilde{c}_x^t$ with $\widetilde{D} = \widetilde{D}_g$, respectively.

We next derive the inequalities that hold when $\widetilde{\tau}^\circ(\widetilde{\ell}_\mu) < \infty$. From Lemma 8 and $\widetilde{D}_\ell \le A_{\text{diff}} + 2\min\{\rho A_{\max}, \widetilde{C}_y\}$ in (27), the two terms involving $\widetilde{\ell}_\mu$ in the last inequality of (26) are upper bounded as

$$
\sqrt{P_\infty^{\tau_x}(\widetilde{\ell}_\mu)M} + \frac{2^{9/2}A_{\text{diff}}^2 M^{3/2}}{\sqrt{P_\infty^{\tau_x - 1}(\widetilde{\ell}_\mu)}} \le \sqrt{2M}\,\widetilde{D}_\ell + 2^{13/4}A_{\text{diff}}M \le 11A_{\text{diff}}M + 3\sqrt{\rho A_{\max}\widetilde{C}_y M}\,, \quad (29)
$$

where we used $M \ge 1$ and $2\sqrt{2}\min\{\rho A_{\max}, \widetilde{C}_y\} \le 3\sqrt{\rho A_{\max}\widetilde{C}_y}$, which follows from $\min\{a, b\} \le \sqrt{ab}$ for $a, b \ge 0$. From Lemma 8 we also have $P_\infty^{\tau_x - 1}(\widetilde{\ell}_\mu) \ge 2^{9/2}A_{\text{diff}}^2 M$ and thus

$$
\eta_x^{\tau_x - 1} \le \sqrt{\frac{M_x}{P_\infty^{\tau_x - 1}(\widetilde{\ell}_\mu)}} \le \sqrt{\frac{M_x}{2^{9/2}A_{\text{diff}}^2 M}} \le \frac{1}{4.75A_{\text{diff}}}\,. \qquad (30)
$$

Similarly, if $\widetilde{\tau}^\circ(\widetilde{g}_\mu) < \infty$, the two terms involving $\widetilde{g}_\mu$ in the last inequality of (26) are upper bounded as

$$
\sqrt{P_\infty^{\tau_y}(\widetilde{g}_\mu)M} + \frac{2^{9/2}A_{\text{diff}}^2 M^{3/2}}{\sqrt{P_\infty^{\tau_y - 1}(\widetilde{g}_\mu)}} \le 11A_{\text{diff}}M + 3\sqrt{\rho A_{\max}\widetilde{C}_x M}\,, \qquad (31)
$$

and

$$\eta_y^{\tau_y - 1} \le \frac{1}{4.75 A_{\mathrm{diff}}} \,. \tag{32}$$

Case 1: when $\widetilde{\tau}^\circ(\widetilde{\ell}_\mu) < \infty$ and $\widetilde{\tau}^\circ(\widetilde{g}_\mu) < \infty$. Combining (26) with (29) and (31), we have

$$\begin{aligned}\mathsf{Reg}_{x,g}^T + \mathsf{Reg}_{y,\ell}^T \le 2^{7/2}\big(22 A_{\mathrm{diff}} M + 6\sqrt{\rho A_{\max}(\widetilde{C}_x + \widetilde{C}_y) M}\big) + \kappa \\ - \left(\frac{1}{8\eta_x^{\tau_x - 1}} Q_1^{\tau_x}(\widehat{x}) + \frac{1}{8\eta_y^{\tau_y - 1}} Q_1^{\tau_y}(\widehat{y})\right).\end{aligned} \tag{33}$$

Therefore, combining (33) with the fact that $\mathsf{Reg}_{x,g}^T + \mathsf{Reg}_{y,\ell}^T \ge 0$, we have

$$\begin{aligned}&Q_1^{\tau_x}(\widehat{x}) \lesssim \eta_x^{\tau_x - 1}\big(A_{\mathrm{diff}} M + \sqrt{A_{\max}(\widetilde{C}_x + \widetilde{C}_y) M} + \kappa\big) \\ &\lesssim M + \frac{1}{A_{\mathrm{diff}}}\left(\sqrt{A_{\max}(\widetilde{C}_x + \widetilde{C}_y) M} + \sqrt{\big[A_{\mathrm{diff}}^2(\widehat{C}_x + \widehat{C}_y) + A_{\max}(\widetilde{C}_x + \widetilde{C}_y)\big] M} + C_x + C_y\right),\end{aligned} \tag{34}$$

where the second inequality follows from (30) and the definition of κ in (24). Similarly, using (32), we obtain the same upper bound for $Q_1^{\tau_y}(\widehat{y})$. Multiplying by A_{diff}^2 and applying the AM–GM inequalities $A_{\mathrm{diff}}\sqrt{A_{\max}(\widetilde{C}_x + \widetilde{C}_y) M} \le \frac{1}{2} A_{\mathrm{diff}}^2 M + \frac{1}{2} A_{\max}(\widetilde{C}_x + \widetilde{C}_y)$ and $A_{\mathrm{diff}}\sqrt{[A_{\mathrm{diff}}^2(\widehat{C}_x + \widehat{C}_y) + A_{\max}(\widetilde{C}_x + \widetilde{C}_y)] M} \le \frac{1}{2} A_{\mathrm{diff}}^2 M + \frac{1}{2}[A_{\mathrm{diff}}^2(\widehat{C}_x + \widehat{C}_y) + A_{\max}(\widetilde{C}_x + \widetilde{C}_y)]$, we obtain

$$A_{\mathrm{diff}}^2\big(Q_1^{\tau_x}(\widehat{x}) + Q_1^{\tau_y}(\widehat{y})\big) \lesssim A_{\mathrm{diff}}^2 M + A_{\mathrm{diff}}^2(\widehat{C}_x + \widehat{C}_y) + A_{\max}(\widetilde{C}_x + \widetilde{C}_y) + A_{\mathrm{diff}}(C_x + C_y) \,. \tag{35}$$

From the definitions of τ_x, τ_y and Lemma 8 with (27), we also have

$$\begin{aligned}P_\infty^{\tau_x}(\widetilde{\ell}_\mu) &\le 2^{9/2} A_{\mathrm{diff}}^2 M + 2\widetilde{D}_\ell^2 \lesssim A_{\mathrm{diff}}^2 M + A_{\max}\widetilde{C}_y \,, \\ P_\infty^{\tau_y}(\widetilde{g}_\mu) &\le 2^{9/2} A_{\mathrm{diff}}^2 M + 2\widetilde{D}_g^2 \lesssim A_{\mathrm{diff}}^2 M + A_{\max}\widetilde{C}_x \,,\end{aligned} \tag{36}$$

where we used $\widetilde{D}_\ell^2 \le 2A_{\mathrm{diff}}^2 + 8\min\{\rho A_{\max}, \widetilde{C}_y\}^2 \le 2A_{\mathrm{diff}}^2 + 8\rho A_{\max}\widetilde{C}_y$ and the analogous bound for $\widetilde{D}_g^2$.

Finally, by substituting (35) and (36) into (21), we obtain

$$\begin{aligned}\mathsf{Reg}_{\widehat{x},\widetilde{g}}^T &\lesssim \sqrt{\big[A_{\mathrm{diff}}^2 M + A_{\mathrm{diff}}^2(\widehat{C}_x + \widehat{C}_y) + A_{\max}(\widetilde{C}_x + \widetilde{C}_y) + A_{\mathrm{diff}}(C_x + C_y)\big] M_x} \\ &\lesssim A_{\mathrm{diff}}\sqrt{M M_x} + \sqrt{\Big[A_{\max}(\widetilde{C}_x + \widetilde{C}_y) + A_{\mathrm{diff}}^2(\widehat{C}_x + \widehat{C}_y) + A_{\mathrm{diff}}(C_x + C_y)\Big] M_x} \,.\end{aligned}$$

Similarly, substituting (35) and (36) into (23), we obtain

$$\mathsf{Reg}_{\widehat{y},\widetilde{\ell}}^T \lesssim A_{\mathrm{diff}}\sqrt{M M_y} + \sqrt{\Big[A_{\max}(\widetilde{C}_x + \widetilde{C}_y) + A_{\mathrm{diff}}^2(\widehat{C}_x + \widehat{C}_y) + A_{\mathrm{diff}}(C_x + C_y)\Big] M_y} \,.$$

Combining these upper bounds on $\mathsf{Reg}_{\widehat{x},\widetilde{g}}^T$ and $\mathsf{Reg}_{\widehat{y},\widetilde{\ell}}^T$ with the inequalities $\mathsf{Reg}_{x,\widetilde{g}}^T \le \mathsf{Reg}_{\widehat{x},\widetilde{g}}^T + (1+\rho) A_{\max}\widehat{C}_x = \mathsf{Reg}_{\widehat{x},\widetilde{g}}^T + O(A_{\max}\widehat{C}_x)$ and $\mathsf{Reg}_{x,g}^T \le \mathsf{Reg}_{\widehat{x},\widetilde{g}}^T + C_x$ from Proposition 1 and the analogous inequalities for the y-player yields the desired upper bounds in Theorem 3.

Case 2: when $\widetilde{\tau}^\circ(\widetilde{\ell}_\mu) = \infty$ and $\widetilde{\tau}^\circ(\widetilde{g}_\mu) < \infty$. From the choice of $\widetilde{\tau}^\circ(\widetilde{\ell}_\mu)$ in (28), we have $\tau_x = T + 1$, so that $Q_1^{\tau_x}(\widehat{x}) = 0$ and

$$P_\infty^{\tau_x}(\widetilde{\ell}_\mu) = P_\infty^{T+1}(\widetilde{\ell}_\mu) \le 2^{9/2} A_{\mathrm{diff}}^2 M + \widetilde{D}_\ell^2 \lesssim A_{\mathrm{diff}}^2 M + A_{\max}\widetilde{C}_y \,. \tag{37}$$

Thus, setting $Q_1^{\tau_x}(\widehat{x}) = 0$ in (25), we have

$$\begin{aligned}
&\mathsf{Reg}_{x,g}^T + \mathsf{Reg}_{y,\ell}^T \\
&\le \sqrt{128(P_\infty^{\tau_y}(\widetilde{g}_\mu) + P_\infty^{T+1}(\widetilde{\ell}_\mu))M} + \left(\sqrt{128A_{\mathrm{diff}}^2 Q_1^{\tau_y}(\widehat{y})M} - \frac{1}{8\eta_y^{\tau_y-1}} Q_1^{\tau_y}(\widehat{y})\right) \\
&\qquad - \frac{1}{8\eta_y^{\tau_y-1}} Q_1^{\tau_y}(\widehat{y}) + \kappa \\
&\le \sqrt{128P_\infty^{T+1}(\widetilde{\ell}_\mu)M} + 2^{7/2}\left(\sqrt{P_\infty^{\tau_y}(\widetilde{g}_\mu)M} + \frac{2^{9/2}A_{\mathrm{diff}}^2 M^{3/2}}{\sqrt{P_\infty^{\tau_y-1}(\widetilde{g}_\mu)}}\right) - \frac{1}{8\eta_y^{\tau_y-1}} Q_1^{\tau_y}(\widehat{y}) + \kappa \\
&\le O\big(A_{\mathrm{diff}}M + \sqrt{A_{\max}(\widetilde{C}_x + \widetilde{C}_y)M}\big) - \frac{1}{8\eta_y^{\tau_y-1}} Q_1^{\tau_y}(\widehat{y}) + \kappa\,,
\end{aligned}$$

where the second inequality follows from the same argument as in (26) and the last inequality follows from (37), which gives $\sqrt{128P_\infty^{T+1}(\widetilde{\ell}_\mu)M} \lesssim A_{\mathrm{diff}}M + \sqrt{A_{\max}\widetilde{C}_y M}$, together with (31). From this inequality, repeating the argument of (34) and (35) for $Q_1^{\tau_y}(\widehat{y})$ (with $Q_1^{\tau_x}(\widehat{x}) = 0$) yields the bound in (35). Finally, by substituting (35), the bound on $P_\infty^{T+1}(\widetilde{\ell}_\mu)$ in (37), and $P_\infty^{\tau_y}(\widetilde{g}_\mu) \lesssim A_{\mathrm{diff}}^2 M + A_{\max}\widetilde{C}_x$ from (36) into (21) and (23) and using Proposition 1, we obtain the desired bounds as in Case 1.

Case 3: when $\widetilde{\tau}^\circ(\widetilde{\ell}_\mu) < \infty$ and $\widetilde{\tau}^\circ(\widetilde{g}_\mu) = \infty$. Repeating the same analysis as in Case 2 with the roles of the two players swapped gives the desired bounds.

Case 4: when $\widetilde{\tau}^\circ(\widetilde{\ell}_\mu) = \infty$ and $\widetilde{\tau}^\circ(\widetilde{g}_\mu) = \infty$. From the choices of $\widetilde{\tau}^\circ(\widetilde{\ell}_\mu), \widetilde{\tau}^\circ(\widetilde{g}_\mu)$ in (28), we have $P_\infty^{T+1}(\widetilde{\ell}_\mu) \le 2^{9/2}A_{\mathrm{diff}}^2 M + \widetilde{D}_\ell^2$ and $P_\infty^{T+1}(\widetilde{g}_\mu) \le 2^{9/2}A_{\mathrm{diff}}^2 M + \widetilde{D}_g^2$. Combining these bounds with (20) and (22) and using Proposition 1, we obtain the desired bounds, which completes the proof. ■

B.5 Scale-agnostic learning dynamic when no communication is possible

All of the learning dynamics presented in the main body achieve scale-agnostic and scale-free fast convergence by tuning their learning rates using the gradients observed by the opponents. Such communication is not a limitation if one views learning in games as an equilibrium computation problem. However, if one views the players as independent decision makers, it is natural to ask whether one can design learning dynamics that guarantee scale-agnostic or scale-free fast convergence without requiring such gradient communication. Here we show how to construct a scale-agnostic learning dynamic with fast convergence of $O(1/T)$ without such communication, at the cost of a worse dependence on the payoff range A_{diff}. For simplicity, we focus on the honest regime here.

We consider the optimistic Hedge algorithm in (1) with the learning rates η_x^t and η_y^t replaced by

$$\eta_x^t = \sqrt{\frac{M_x}{P_\infty^t(g_\mu) + P_1^t(x)}}\,, \quad \eta_y^t = \sqrt{\frac{M_y}{P_\infty^t(\ell_\mu) + P_1^t(y)}}\,, \tag{38}$$

where we recall the conventions $g_\mu^0 = \ell_\mu^0 = \mathbf{0}$ and $z^0 = \mathbf{0}$, which give $\|x^1 - x^0\|_1 = 1$, and we adopt the same convention as in (1) when the denominators are zero. Each player can compute its denominator from its own observations and strategies, so no communication is needed. The learning dynamic defined by (38) is scale-agnostic but not scale-free in the sense of Definition 2. This is because rescaling the utilities by $c > 0$ scales $P_\infty^t(g_\mu)$ by c^2 while leaving $P_1^t(x)$ unchanged, so the learning rates do not scale as $1/c$.

Theorem 4. *In the honest regime, the sum of the regrets under the learning dynamic defined by (38) is upper bounded as*

$$\mathsf{Reg}_x^T + \mathsf{Reg}_y^T \lesssim \big(1 + A_{\mathrm{diff}}^{3/2}\big)\log m \lesssim \begin{cases} A_{\mathrm{diff}}^{3/2}\log m & \text{if } A_{\mathrm{diff}} \ge 1\,, \\ \log m & \text{if } A_{\mathrm{diff}} < 1\,. \end{cases}$$

Consequently, the average play $\big(\frac{1}{T}\sum_{t=1}^T x^t, \frac{1}{T}\sum_{t=1}^T y^t\big)$ is an $O((1 + A_{\mathrm{diff}}^{3/2})\log m/T)$-approximate Nash equilibrium.

This regret bound is worse than that of Theorem 1 in both cases: for $A_{\mathrm{diff}} \geq 1$ it is larger by a factor of $\sqrt{A_{\mathrm{diff}}}$, and for $A_{\mathrm{diff}} < 1$ it does not vanish as $A_{\mathrm{diff}} \to 0$, in contrast to the $O(A_{\mathrm{diff}} \log m)$ bound of Theorem 1. This dependence arises from the strategy path lengths $P_1^t(x)$ and $P_1^t(y)$ in (38), which do not scale with the utilities. Improving the dependence on A_{diff} without communication, or establishing a matching lower bound for this class of dynamics, is an interesting direction for future work.

Proof. If $A_{\mathrm{diff}} = 0$, then $\mathsf{Reg}_x^T = \mathsf{Reg}_y^T = 0$ as in the proof of Theorem 1. Hence we assume $A_{\mathrm{diff}} > 0$ in what follows. Let $B = 1 + \bar{A}_{\mathrm{diff}}$ for simplicity. Applying Lemma 5 to $h^t = -g_\mu^t$ with $\nu^t = \|x^t - x^{t-1}\|_1^2$ gives

$$\mathsf{Reg}_x^T \leq \sqrt{32\big(P_\infty^{T+1}(g_\mu) + P_1^{T+1}(x)\big)M_x} - \sum_{t=1}^{T} \frac{1}{4\eta_x^t}\|x^{t+1} - x^t\|_1^2 ,$$

and similarly for the y-player. From Lemma 6 and the initial centering, we also have

$$P_\infty^{T+1}(g_\mu) \leq \bar{A}_{\mathrm{diff}}^2 P_1^{T+1}(y) , \qquad P_\infty^{T+1}(\ell_\mu) \leq \bar{A}_{\mathrm{diff}}^2 P_1^{T+1}(x) , \tag{39}$$

where we used $\|g_\mu^1 - g_\mu^0\|_\infty \leq \bar{A}_{\mathrm{diff}} = \bar{A}_{\mathrm{diff}}\|y^1 - y^0\|_1$ and $\|g_\mu^s - g_\mu^{s-1}\|_\infty \leq \bar{A}_{\mathrm{diff}}\|y^s - y^{s-1}\|_1$ for $s \geq 2$ in (11). The former follows from the initial centering as in the proof of Lemma 1, together with $\|y^1\|_1 = 1$ and $y^0 = \mathbf{0}$. The bound for ℓ_μ follows in the same manner. Then, by combining the upper bounds on Reg_x^T and Reg_y^T with (39), the subadditivity of $\sqrt{\cdot}$, and $M_x, M_y \leq M$, we have

$$\begin{aligned}\mathsf{Reg}_x^T + \mathsf{Reg}_y^T \leq{}& B\sqrt{32P_1^{T+1}(x)M} + B\sqrt{32P_1^{T+1}(y)M} \\ &- \sum_{t=1}^{T} \frac{1}{4\eta_x^t}\|x^{t+1} - x^t\|_1^2 - \sum_{t=1}^{T} \frac{1}{4\eta_y^t}\|y^{t+1} - y^t\|_1^2 .\end{aligned} \tag{40}$$

We next lower bound the negative terms in (40). We only discuss the x-player since the same argument applies to the y-player. Since $\eta_x^t \leq \sqrt{M_x/P_1^t(x)} \leq \sqrt{M/P_1^t(x)}$ by (38), and since every increment of $P_1^t(x)$ is at most 4, which gives $\sqrt{P_1^{s-1}(x)} \geq \sqrt{P_1^s(x)} - 2$, we have

$$\begin{aligned}\sum_{t=1}^{T} \frac{1}{4\eta_x^t}\|x^{t+1} - x^t\|_1^2 &\geq \frac{1}{4\sqrt{M}} \sum_{s=2}^{T} \sqrt{P_1^{s-1}(x)}\,\|x^s - x^{s-1}\|_1^2 \\ &\geq \frac{1}{4\sqrt{M}} \sum_{s=2}^{T} \Big(\sqrt{P_1^s(x)} - 2\Big)\|x^s - x^{s-1}\|_1^2 \\ &\geq \frac{1}{4\sqrt{M}} \left(\sum_{s=1}^{T} \sqrt{P_1^s(x)}\,\|x^s - x^{s-1}\|_1^2 - 2P_1^{T+1}(x)\right) \\ &\geq \frac{1}{4\sqrt{M}} \left(\frac{2}{3}\big(P_1^{T+1}(x)\big)^{3/2} - 4P_1^{T+1}(x)\right) ,\end{aligned} \tag{41}$$

where the first inequality follows by keeping only the terms with $t \leq T-1$ and reindexing with $s = t+1$, the second inequality follows from $\sqrt{P_1^{s-1}(x)} \geq \sqrt{P_1^s(x)} - 2$ shown above, the third inequality follows from $P_1^1(x) = 0$ and $\sum_{s=2}^{T}\|x^s - x^{s-1}\|_1^2 \leq P_1^{T+1}(x)$, and the last inequality follows from Lemma 17 in Appendix C applied to the sequence $(P_1^s(x))_{s=1}^{T+1}$, whose increments are at most 4.

Plugging (41) and its y-player analogue into (40) yields

$$\begin{aligned}\mathsf{Reg}_x^T + \mathsf{Reg}_y^T \leq{}& B\sqrt{32P_1^{T+1}(x)M} + \frac{P_1^{T+1}(x)}{\sqrt{M}} - \frac{\big(P_1^{T+1}(x)\big)^{3/2}}{6\sqrt{M}} \\ &+ B\sqrt{32P_1^{T+1}(y)M} + \frac{P_1^{T+1}(y)}{\sqrt{M}} - \frac{\big(P_1^{T+1}(y)\big)^{3/2}}{6\sqrt{M}} .\end{aligned} \tag{42}$$

We now bound the terms involving $P_1^{T+1}(x)$ in (42) by considering two cases. The terms involving $P_1^{T+1}(y)$ are bounded in the same manner. If $P_1^{T+1}(x) \geq 2^7 BM$, then, dividing the sum of the first

two terms in (42) by $\left(P_1^{T+1}(x)\right)^{3/2}/(6\sqrt{M}) > 0$, we obtain

$$\frac{6\sqrt{32}\,BM}{P_1^{T+1}(x)} + \frac{6}{\sqrt{P_1^{T+1}(x)}} \le \frac{6\sqrt{32}}{2^7} + \frac{6}{\sqrt{2^9}} = \frac{3\sqrt{2}}{8} < 1\,,$$

where we used $B \ge 1$ and $M \ge 4$, which give $P_1^{T+1}(x) \ge 2^9$. Hence, the sum of the three terms is nonpositive in this case. Otherwise, we have $P_1^{T+1}(x) < 2^7 BM$, and thus

$$\begin{aligned} B\sqrt{32P_1^{T+1}(x)M} + \frac{P_1^{T+1}(x)}{\sqrt{M}} &\le B\sqrt{32\cdot 2^7 BM^2} + \frac{2^7 BM}{\sqrt{M}} \\ &= 2^6 B^{3/2}M + 2^7 B\sqrt{M} \le 2^7 B^{3/2}M\,, \end{aligned}$$

where we used $2^7\sqrt{M} \le 2^6 M$ for $M \ge 4$. Therefore, we obtain

$$\mathsf{Reg}_x^T + \mathsf{Reg}_y^T \le 2^8 B^{3/2}M \le 2^8(1+A_{\mathrm{diff}})^{3/2}M \le 2^{19/2}\big(1+A_{\mathrm{diff}}^{3/2}\big)M\,,$$

where the last inequality follows from $(1+A_{\mathrm{diff}})^{3/2} \le \big(2\max\{1, A_{\mathrm{diff}}\}\big)^{3/2} \le 2^{3/2}\big(1+A_{\mathrm{diff}}^{3/2}\big)$. This completes the proof of Theorem 4. ∎

C Deferred Proofs for Multiplayer General-Sum Games from Section 4

This section provides the proofs deferred from Section 4. Throughout this section, we assume $T \ge 3$. If $U^1 = 0$, then all players keep playing uniform distributions and observing zero utility vectors since the game is fixed, and thus their swap regrets are zero. Hence, in proving Theorem 2, we assume $U^1 > 0$. We first introduce some notation. We define the maximum of the squared path lengths of the clipped utility vectors and its increments by

$$\mathbb{S}^t \coloneqq \max_{j\in[n]} P_\infty^t(\bar{u}_j)\,, \qquad \Delta^t \coloneqq \mathbb{S}^{t+1} - \mathbb{S}^t\,, \qquad \text{so that} \quad \mathbb{S}^{t+1} = \mathbb{S}^t + \Delta^t \quad \text{and} \quad \mathbb{S}^1 = 0\,,$$

where we recall $\bar{u}_j^0 = \mathbf{0}$ and $U^0 = 0$. Since $P_\infty^{t+1}(\bar{u}_j) = P_\infty^t(\bar{u}_j) + \|\bar{u}_j^t - \bar{u}_j^{t-1}\|_\infty^2$, we have

$$0 \le \Delta^t \le \max_{j\in[n]} \|\bar{u}_j^t - \bar{u}_j^{t-1}\|_\infty^2 \quad \text{for every } t \in [T]\,. \tag{43}$$

We also recall that H is the Shannon entropy and ϕ is the log-barrier regularizer defined in Section 4. For positive probability vectors $p, q \in \Delta_d$, the Bregman divergence induced by ϕ is

$$D_\phi(q,p) = \sum_{b=1}^d \left(\frac{q(b)}{p(b)} - 1 - \log\frac{q(b)}{p(b)}\right). \tag{44}$$

C.1 Stability analysis of stationary distributions

Here we provide a stability analysis of stationary distributions. Recall that the rows of $\mathbf{Q}_i^t$ are the expert distributions $y_{i,a}^t$, while player i uses the stationary distribution x_i^t. To relate changes in these distributions, we begin with the following result.

Lemma 9 (Tsuchiya 2026b, Theorem 3)**.** *Let $d \ge 2$ and let $\mathbf{Q}, \mathbf{Q}' \in \mathcal{M}_d$ have positive entries. Denote their unique stationary distributions by $\pi(\mathbf{Q})$ and $\pi(\mathbf{Q}')$, respectively. Then, it holds that*

$$\|\pi(\mathbf{Q}') - \pi(\mathbf{Q})\|_1^2 \le 16(d-1)\sum_{a=1}^d\sum_{b=1}^d \min\left\{1, \left(\log\frac{\mathbf{Q}'(a,b)}{\mathbf{Q}(a,b)}\right)^2\right\}. \tag{45}$$

The RHS of (45) is expressed in terms of the ratios of transition probabilities. To bound it by the Bregman divergence in (44), we use the following inequality.

Lemma 10 (Tsuchiya 2026b, Lemma 3)**.** *For every $r > 0$, it holds that*

$$\min\{1, (\log r)^2\} \le 6(r - 1 - \log r)\,.$$

Combining Lemmas 9 and 10 gives the following bound, which will be used to evaluate the negative term in the regret bound.

Lemma 11. *Let $\mathbf{Q}, \mathbf{Q}' \in \mathcal{M}_d$ have positive entries, with rows $p_a^\top, q_a^\top$ for $a \in [d]$, respectively. Let $x, x' \in \Delta_d$ be their stationary distributions, that is, $\mathbf{Q}^\top x = x$ and $(\mathbf{Q}')^\top x' = x'$. Then, it holds that*

$$\|x' - x\|_1^2 \le 96d\sum_{a=1}^d D_\phi(q_a, p_a)\,. \tag{46}$$

Proof. If $d = 1$, then $x = x' = (1)$, and the claim is immediate. Suppose that $d \geq 2$. Since $\mathbf{Q}(a, b) = p_a(b)$ and $\mathbf{Q}'(a, b) = q_a(b)$, we have

$$\begin{aligned}\|x' - x\|_1^2 &\leq 16(d-1) \sum_{a=1}^{d} \sum_{b=1}^{d} \min\left\{1, \left(\log \frac{q_a(b)}{p_a(b)}\right)^2\right\} \\ &\leq 96(d-1) \sum_{a=1}^{d} D_\phi(q_a, p_a) \leq 96d \sum_{a=1}^{d} D_\phi(q_a, p_a)\,,\end{aligned}$$

where the first inequality follows from Lemma 9, and the second inequality from Lemma 10 and the definition of D_ϕ in (44). This completes the proof. ■

To use Lemma 11, we need a negative multiple of $D_\phi(y_{i,a}^{t+1}, y_{i,a}^t)$ to remain after bounding the prediction error in Lemma 2. The negative Shannon entropy $-H$ in the regularizer allows us to obtain the following bound for any positive learning rate.

Lemma 12. *Let $p, q \in \Delta_d$ have positive entries, and let $g \in \mathbb{R}^d$, $\eta > 0$, and $\theta > 0$. Let $H(y) = -\sum_{b=1}^{d} y(b) \log y(b)$ and $\phi(y) = -\sum_{b=1}^{d} \log y(b)$. Then, it holds that*

$$\langle p - q, g\rangle - \frac{\theta D_\phi(q, p) + D_{-H}(q, p)}{\eta} \leq \frac{\eta}{2}\|g\|_\infty^2 - \frac{\theta}{\eta} D_\phi(q, p)\,.$$

This follows by taking the worst case with respect to q.

C.2 Upper bounding the swap regret using lemmas from Appendix C.1

Here we upper bound the swap regret using the preliminary lemmas from Appendix C.1 and the clipping in Algorithm 1. We first establish the following two lemmas on the clipping rule.

Lemma 13. *Suppose that $U^1 > 0$. For all $t \in \{2, \ldots, T\}$ and $i \in [n]$, the clipped utility vectors in (5) satisfy $\|\bar{u}_i^t\|_\infty \leq U^{t-1}$. Moreover, $\|\bar{u}_i^1\|_\infty = \|u_i^1\|_\infty \leq U^1$, and $\|\bar{u}_i^t\|_\infty \leq \|u_i^t\|_\infty \leq U_{\max}$ for all $t \in [T]$.*

Proof. For $t \geq 2$, the definition of U^t gives $\|u_i^t\|_\infty \leq U^t$. Hence,

$$\|\bar{u}_i^t\|_\infty = \frac{U^{t-1}}{U^t}\|u_i^t\|_\infty \leq U^{t-1}\,.$$

The remaining claims follow from the initialization $\bar{u}_i^1 = u_i^1$ and from $0 < U^{t-1} \leq U^t$ for $t \geq 2$, which completes the proof. ■

The following lemma shows that the total clipping error is bounded independently of the initial utility norm U^1.

Lemma 14. *Suppose that $U^1 > 0$, and let $E^t := \max_{i\in[n]} \|u_i^t - \bar{u}_i^t\|_\infty$ for $t \in [T]$ and $E^0 = 0$. Then, $E^1 = 0$, $E^t = U^t - U^{t-1}$ for $t \geq 2$, and*

$$\sum_{t=1}^{T} E^t = U^T - U^1 \leq U_{\max}\,, \qquad \sum_{t=1}^{T} (E^t)^2 \leq U_{\max}^2\,.$$

Proof. The initialization gives $E^1 = 0$. For $t \geq 2$, let $v^t = \max_{i\in[n]} \|u_i^t\|_\infty$, so that $U^t = \max\{U^{t-1}, v^t\}$. The definition of the clipped utility vectors in (5) gives

$$E^t = \left(1 - \frac{U^{t-1}}{U^t}\right) v^t = U^t - U^{t-1}\,,$$

where the second equality holds since both sides are zero if $v^t \leq U^{t-1}$, and $U^t = v^t$ otherwise. Summing over t yields

$$\sum_{t=1}^{T} E^t = \sum_{t=2}^{T} (U^t - U^{t-1}) = U^T - U^1 \leq U_{\max}\,.$$

Since the errors are nonnegative, we also have

$$\sum_{t=1}^{T} (E^t)^2 \leq \left(\sum_{t=1}^{T} E^t\right)^2 \leq U_{\max}^2\,,$$

which completes the proof. ■

We also use the following consequence of Lemma 13. For every $t \geq 2$, both $\|\bar{u}_i^t\|_\infty$ and $\|\bar{u}_i^{t-1}\|_\infty$ are at most U^{t-1}, where the case $t = 2$ follows from the initialization. Consequently,

$$\max_{i\in[n]} \|\bar{u}_i^t - \bar{u}_i^{t-1}\|_\infty^2 \leq 4(U^{t-1})^2 \leq \gamma(U^{t-1})^2 \quad \text{for } t \geq 2\,. \tag{47}$$

Combining (43) and (47) with $\gamma = 2^{10}m^3 \log T \geq 4$ gives

$$\gamma(U^{t-1})^2 + \mathbb{S}^t \geq \mathbb{S}^{t+1} \geq P_\infty^{t+1}(\bar{u}_i) \quad \text{for every } i \in [n] \text{ and } t \geq 2\,. \tag{48}$$

We then upper bound the swap regret $\mathsf{SwapReg}_{x_i}^T$ in the following lemma.

Lemma 15. *Suppose that $T \geq 3$, $U^1 > 0$, and every player in the game of Section 2.3 uses Algorithm 1. With $m = \max_{i\in[n]} m_i$ and the parameters α, γ, θ specified in Section 4, every $i \in [n]$ satisfies*

$$\begin{aligned}\mathsf{SwapReg}_{x_i}^T \leq\ & 2\alpha\sqrt{\gamma(U^T)^2 + \mathbb{S}^{T+1}} + 2\alpha\sqrt{P_\infty^{T+1}(\bar{u}_i)} \\ & - \frac{\theta}{192m\alpha}\sum_{t=2}^{T-1}\sqrt{\gamma(U^{t-1})^2 + \mathbb{S}^t}\,\|x_i^{t+1} - x_i^t\|_1^2 + 11U_{\max}\,.\end{aligned} \tag{49}$$

Proof. We first separate the effect of clipping and then apply the general bound for OFTRL to each expert. Fix a player $i \in [n]$ and comparators $y_a \in \Delta_{m_i}$ for $a \in \mathcal{A}_i$. By the definition of the external regrets,

$$\sum_{a\in\mathcal{A}_i} \mathsf{Reg}_{i,a}^T(y_a) = \sum_{a\in\mathcal{A}_i}\sum_{t=1}^T \langle y_a - y_{i,a}^t, \bar{u}_{i,a}^t\rangle + \sum_{a\in\mathcal{A}_i}\sum_{t=1}^T \langle y_a - y_{i,a}^t, u_{i,a}^t - \bar{u}_{i,a}^t\rangle\,. \tag{50}$$

The second term on the RHS of (50) satisfies

$$\begin{aligned}\sum_{a\in\mathcal{A}_i}\sum_{t=1}^T \langle y_a - y_{i,a}^t, u_{i,a}^t - \bar{u}_{i,a}^t\rangle &\leq 2\sum_{t=1}^T\sum_{a\in\mathcal{A}_i} x_i^t(a)\|u_i^t - \bar{u}_i^t\|_\infty \\ &\leq 2\sum_{t=1}^T E^t \leq 4U_{\max}\,,\end{aligned} \tag{51}$$

where we used Hölder's inequality, $\sum_{a\in\mathcal{A}_i} x_i^t(a) = 1$, and Lemma 14. Thus, it remains to bound the regrets on the clipped utility vectors.

Write $\lambda^t = 1/\eta^t$ for $t \geq 2$ and adopt the convention $\lambda^1 = 0$ for the first round, in which $y_{i,a}^1 = \mathbf{1}/m_i$, so that $(\lambda^t)_t$ is nondecreasing since U^{t-1} and $\mathbb{S}^t$ are nondecreasing. For the analysis, define $y_{i,a}^{T+1}$ by the update in (6) using the observations through round T and the same fixed horizon T in the parameters. Let x_i^{T+1} be the stationary distribution of the resulting matrix. For each expert, set $\bar{u}_{i,a}^0 = \mathbf{0}$ and apply Lemma 2 with $h^t = -\bar{u}_{i,a}^t$, $m^t = -\bar{u}_{i,a}^{t-1}$, and $\psi^t = \lambda^t(\theta\phi - H)$ for $t \geq 2$, while $\psi^1 \equiv 0$.

For each $a \in \mathcal{A}_i$, let $\widehat{y}_a = (1 - 1/T)y_a + \mathbf{1}/(m_i T)$. The bound in (7) of Lemma 2 gives

$$\begin{aligned}\sum_{t=1}^T \langle \widehat{y}_a - y_{i,a}^t, \bar{u}_{i,a}^t\rangle \leq\ & \psi^{T+1}(\widehat{y}_a) - \psi^1(y_{i,a}^1) + \sum_{t=1}^T\big(\psi^t(y_{i,a}^{t+1}) - \psi^{t+1}(y_{i,a}^{t+1})\big) \\ & + \sum_{t=1}^T\big(\langle y_{i,a}^t - y_{i,a}^{t+1}, -(\bar{u}_{i,a}^t - \bar{u}_{i,a}^{t-1})\rangle - D_{\psi^t}(y_{i,a}^{t+1}, y_{i,a}^t)\big) \\ & + \langle \widehat{y}_a - y_{i,a}^{T+1}, -\bar{u}_{i,a}^T\rangle\,.\end{aligned} \tag{52}$$

We first bound the three terms in the first line on the RHS of (52). Let $z_i = \mathbf{1}/m_i$, which minimizes both $-H$ and ϕ on Δ_{m_i}, and hence $\theta\phi - H$. Since $\psi^t = \lambda^t(\theta\phi - H)$ and $\lambda^1 = 0$, we have

$$\begin{aligned}
&\psi^{T+1}(\widehat{y}_a) - \psi^1(y^1_{i,a}) + \sum_{t=1}^{T}\big(\psi^t(y^{t+1}_{i,a}) - \psi^{t+1}(y^{t+1}_{i,a})\big) \\
&= \psi^{T+1}(\widehat{y}_a) - \psi^{T+1}(z_i) + \sum_{t=1}^{T}(\lambda^t - \lambda^{t+1})\big(\theta\phi(y^{t+1}_{i,a}) - H(y^{t+1}_{i,a}) - \theta\phi(z_i) + H(z_i)\big) \\
&\le \psi^{T+1}(\widehat{y}_a) - \psi^{T+1}(z_i) \le \frac{\log m_i + \theta m_i \log T}{\eta^{T+1}}\,,
\end{aligned}$$

where the first inequality follows from $\lambda^t \le \lambda^{t+1}$ and the minimality of z_i, and the second inequality from $0 \le H(\widehat{y}_a)$, $H(z_i) = \log m_i$, $\phi(\widehat{y}_a) \le m_i \log(m_i T)$, and $\phi(z_i) = m_i \log m_i$. Summing this bound over the m_i experts gives

$$\frac{m_i \log m_i + \theta m_i^2 \log T}{\eta^{T+1}} \le \frac{2m \log m}{\eta^{T+1}}\,,$$

where we used $m_i \le m$ and $\theta = \log m/(m \log T)$. Replacing y_a by $\widehat{y}_a$ in the first term on the RHS of (50) introduces the difference

$$\sum_{a\in\mathcal{A}_i}\sum_{t=1}^{T}\langle y_a - \widehat{y}_a, \bar{u}^t_{i,a}\rangle \le \frac{2}{T}\sum_{t=1}^{T}\sum_{a\in\mathcal{A}_i}\|\bar{u}^t_{i,a}\|_\infty \le 2U_{\max}\,. \tag{53}$$

The last term on the RHS of (52) and the $t = 1$ term in its second line satisfy

$$\begin{aligned}
\sum_{a\in\mathcal{A}_i}\langle \widehat{y}_a - y^{T+1}_{i,a}, -\bar{u}^T_{i,a}\rangle &\le 2\sum_{a\in\mathcal{A}_i}\|\bar{u}^T_{i,a}\|_\infty \le 2U_{\max}\,, \\
\sum_{a\in\mathcal{A}_i}\langle y^1_{i,a} - y^2_{i,a}, -\bar{u}^1_{i,a}\rangle &\le 2\sum_{a\in\mathcal{A}_i}\|\bar{u}^1_{i,a}\|_\infty \le 2U_{\max}\,,
\end{aligned} \tag{54}$$

where we used $\lambda^1 = 0$, and the last inequalities in these bounds and in (53) follow from $\sum_{a\in\mathcal{A}_i}\|\bar{u}^t_{i,a}\|_\infty = \|\bar{u}^t_i\|_\infty \le U_{\max}$.

For the remaining terms in the second line of (52), since $D_{\psi^t} = \lambda^t(\theta D_\phi + D_{-H})$, applying Lemma 12 with $p = y^t_{i,a}$, $q = y^{t+1}_{i,a}$, $g = -(\bar{u}^t_{i,a} - \bar{u}^{t-1}_{i,a})$, and $\eta = \eta^t$ gives

$$\langle y^t_{i,a} - y^{t+1}_{i,a}, -(\bar{u}^t_{i,a} - \bar{u}^{t-1}_{i,a})\rangle - D_{\psi^t}(y^{t+1}_{i,a}, y^t_{i,a}) \le \frac{\eta^t}{2}\|\bar{u}^t_{i,a} - \bar{u}^{t-1}_{i,a}\|^2_\infty - \frac{\theta}{\eta^t}D_\phi(y^{t+1}_{i,a}, y^t_{i,a})\,.$$

Combining (50) and (52) with the preceding estimates and the identity in Section 4.1, and maximizing over the comparators, we obtain

$$\begin{aligned}
\mathsf{SwapReg}^T_{x_i} \le \frac{2m\log m}{\eta^{T+1}} &+ \frac{1}{2}\sum_{t=2}^{T}\eta^t\sum_{a\in\mathcal{A}_i}\|\bar{u}^t_{i,a} - \bar{u}^{t-1}_{i,a}\|^2_\infty \\
&- \frac{\theta}{96m}\sum_{t=2}^{T}\frac{\|x^{t+1}_i - x^t_i\|^2_1}{\eta^t} + 10U_{\max}\,,
\end{aligned} \tag{55}$$

where the third term on the RHS follows from (46) in Lemma 11 with $d = m_i \le m$, and the last term is $4U_{\max} + 2U_{\max} + 2U_{\max} + 2U_{\max}$ by (51), (53) and (54).

We next bound the second term on the RHS of (55). For every $t \ge 2$, we have

$$\begin{aligned}
&\sum_{a\in\mathcal{A}_i}\|\bar{u}^t_{i,a} - \bar{u}^{t-1}_{i,a}\|^2_\infty \\
&= \sum_{a\in\mathcal{A}_i}\|x^t_i(a)\bar{u}^t_i - x^{t-1}_i(a)\bar{u}^{t-1}_i\|^2_\infty \\
&\le 2\sum_{a\in\mathcal{A}_i}(x^t_i(a))^2\|\bar{u}^t_i - \bar{u}^{t-1}_i\|^2_\infty + 2\sum_{a\in\mathcal{A}_i}(x^t_i(a) - x^{t-1}_i(a))^2\|\bar{u}^{t-1}_i\|^2_\infty \\
&\le 2\|\bar{u}^t_i - \bar{u}^{t-1}_i\|^2_\infty + 2\|\bar{u}^{t-1}_i\|^2_\infty\|x^t_i - x^{t-1}_i\|^2_1\,,
\end{aligned} \tag{56}$$

where we used $\sum_a (x^t_i(a))^2 \le 1$ and $\|\cdot\|_2 \le \|\cdot\|_1$. The second term in the last line of (56) contributes $\eta^t\|\bar{u}^{t-1}_i\|^2_\infty\|x^t_i - x^{t-1}_i\|^2_1$ to the RHS of (55). For $t \ge 3$, half of the $t-1$ term in the negative sum

in (55) absorbs this contribution, since

$$\eta^t \eta^{t-1} \|\bar{u}_i^{t-1}\|_\infty^2 \le \frac{\alpha^2}{\gamma} = \frac{\theta}{1024m} \le \frac{\theta}{192m}\,, \tag{57}$$

where the first inequality follows from $\eta^t \le \eta^{t-1} \le \alpha/(\sqrt{\gamma}\,U^{t-2})$ and $\|\bar{u}_i^{t-1}\|_\infty \le U^{t-2}$ from Lemma 13, and the equality from the choices of α, γ, and θ. At $t = 2$, the corresponding term satisfies

$$\eta^2 \|\bar{u}_i^1\|_\infty^2 \|x_i^2 - x_i^1\|_1^2 \le 4\eta^2 (U^1)^2 \le \frac{\sqrt{\log m}}{8m\sqrt{\log T}} U^1 \le U_{\max}\,,$$

where the first inequality follows from $\|\bar{u}_i^1\|_\infty \le U^1$ and $\|x_i^2 - x_i^1\|_1 \le 2$, the second inequality from $\eta^2 \le \alpha/(\sqrt{\gamma}\,U^1)$ and the choices of α and γ, and the last inequality from $m \ge 2$, $T \ge 3$, $\log m \le m$, and $U^1 \le U_{\max}$. Substituting (56) and (57) and the bound at $t = 2$ into (55), pairing each $t \in \{3, \dots, T\}$ with $t - 1$, and discarding the $t = T$ term of the negative sum on its RHS, we obtain

$$\mathsf{SwapReg}_{x_i}^T \le \frac{2m \log m}{\eta^{T+1}} + \sum_{t=2}^{T} \eta^t \|\bar{u}_i^t - \bar{u}_i^{t-1}\|_\infty^2 - \frac{\theta}{192m} \sum_{t=2}^{T-1} \frac{\|x_i^{t+1} - x_i^t\|_1^2}{\eta^t} + 11 U_{\max}\,. \tag{58}$$

It remains to substitute the learning rate in (6). For the second term on the RHS of (58), (48) gives

$$\sum_{t=2}^{T} \eta^t \|\bar{u}_i^t - \bar{u}_i^{t-1}\|_\infty^2 \le \alpha \sum_{t=2}^{T} \frac{\|\bar{u}_i^t - \bar{u}_i^{t-1}\|_\infty^2}{\sqrt{P_\infty^{t+1}(\bar{u}_i)}} \le 2\alpha \sqrt{P_\infty^{T+1}(\bar{u}_i)}\,, \tag{59}$$

where a term with zero numerator and denominator is interpreted as zero, and the last inequality follows by telescoping $a/\sqrt{b+a} \le 2(\sqrt{b+a} - \sqrt{b})$ for $a, b \ge 0$. For the first and third terms on the RHS of (58), the definition of η^t and $\alpha^2 = m \log m$ give

$$\frac{2m \log m}{\eta^{T+1}} = 2\alpha \sqrt{\gamma (U^T)^2 + \mathbb{S}^{T+1}}\,, \qquad \frac{1}{\eta^t} = \frac{\sqrt{\gamma (U^{t-1})^2 + \mathbb{S}^t}}{\alpha}\,.$$

Substituting these identities and (59) into (58) gives (49), which completes the proof. ■

C.3 Proof of Theorem 2

Here we provide the proof of Theorem 2. We bound $\mathbb{S}^{T+1}$ using the following relation between the squared differences of the utility vectors and those of the strategies.

Lemma 16. *In the multiplayer general-sum game of Section 2.3, for every $t \in \{2, \dots, T\}$, it holds that*

$$\max_{i \in [n]} \|u_i^t - u_i^{t-1}\|_\infty^2 \le U_{\max}^2 (n-1) \sum_{i \in [n]} \|x_i^t - x_i^{t-1}\|_1^2\,.$$

Proof. Let $\mathcal{A}_{-i} = \times_{j \ne i} \mathcal{A}_j$. Then,

$$\begin{aligned}
\|u_i^t - u_i^{t-1}\|_\infty &= \max_{a_i \in \mathcal{A}_i} \left| \sum_{a_{-i} \in \mathcal{A}_{-i}} u_i(a_i, a_{-i}) \left(\prod_{j \ne i} x_j^t(a_j) - \prod_{j \ne i} x_j^{t-1}(a_j) \right) \right| \\
&\le U_{\max} \sum_{a_{-i} \in \mathcal{A}_{-i}} \left| \prod_{j \ne i} x_j^t(a_j) - \prod_{j \ne i} x_j^{t-1}(a_j) \right| \le U_{\max} \sum_{j \ne i} \|x_j^t - x_j^{t-1}\|_1\,,
\end{aligned}$$

where the equality follows from the definition of the expected utility vector u_i^t, the first inequality from $u_i(a_i, a_{-i}) \in [-U_{\max}, U_{\max}]$, and the last inequality from the fact that the total variation distance between two product distributions is at most the sum of the total variation distances between their marginals. Using this inequality and the Cauchy–Schwarz inequality, we have

$$\begin{aligned}
\|u_i^t - u_i^{t-1}\|_\infty^2 &\le U_{\max}^2 \left(\sum_{j \ne i} \|x_j^t - x_j^{t-1}\|_1 \right)^2 \\
&\le U_{\max}^2 \left(\sqrt{(n-1) \sum_{j \ne i} \|x_j^t - x_j^{t-1}\|_1^2} \right)^2 = U_{\max}^2 (n-1) \sum_{j \ne i} \|x_j^t - x_j^{t-1}\|_1^2\,.
\end{aligned}$$

Taking the maximum over $i \in [n]$ in the last display and extending the sum over $j \neq i$ to all $j \in [n]$ gives

$$\max_{i\in[n]} \|u_i^t - u_i^{t-1}\|_\infty^2 \le U_{\max}^2 (n-1) \sum_{j\in[n]} \|x_j^t - x_j^{t-1}\|_1^2 ,$$

which completes the proof. ■

The following lemma gives a lower bound on a weighted sum of the increments of a nondecreasing sequence.

Lemma 17. *Let $(\mathbb{S}^t)_{t=1}^{T+1}$ and $(\Delta^t)_{t=1}^{T}$ be nonnegative reals such that $\mathbb{S}^1 = 0$, $\mathbb{S}^{t+1} = \mathbb{S}^t + \Delta^t$, and $\Delta^t \le D$ for all $t \in [T]$. Then,*

$$\sum_{t=1}^{T} \sqrt{\mathbb{S}^t}\, \Delta^t \ge \frac{2}{3} (\mathbb{S}^{T+1})^{3/2} - \sqrt{D}\, \mathbb{S}^{T+1} .$$

Proof. For each $t \in [T]$ and $x \in [\mathbb{S}^t, \mathbb{S}^{t+1}]$, we have $\sqrt{x} \le \sqrt{\mathbb{S}^{t+1}} \le \sqrt{\mathbb{S}^t + D} \le \sqrt{\mathbb{S}^t} + \sqrt{D}$. Hence,

$$\begin{aligned} \frac{2}{3} (\mathbb{S}^{T+1})^{3/2} &= \int_0^{\mathbb{S}^{T+1}} \sqrt{x}\, \mathrm{d}x = \sum_{t=1}^{T} \int_{\mathbb{S}^t}^{\mathbb{S}^{t+1}} \sqrt{x}\, \mathrm{d}x \\ &\le \sum_{t=1}^{T} \big(\sqrt{\mathbb{S}^t} + \sqrt{D}\big) \Delta^t = \sum_{t=1}^{T} \sqrt{\mathbb{S}^t} \Delta^t + \sqrt{D}\, \mathbb{S}^{T+1} , \end{aligned}$$

which completes the proof. ■

We are now ready to prove Theorem 2.

Proof. We first use the nonnegativity of swap regret to bound the sum in the third term on the RHS of (49) in Lemma 15. To simplify the coefficients, define

$$\Xi := m^2 \sqrt{\log m \log T} , \qquad Z^t := \frac{\mathbb{S}^t}{m^3 \log T} \quad \text{for } t \in [T+1] ,$$

and write

$$W_i := \sum_{t=2}^{T-1} \sqrt{2^{10}(U^{t-1})^2 + Z^t}\, \|x_i^{t+1} - x_i^t\|_1^2 , \qquad W := \sum_{i\in[n]} W_i .$$

The choices of α, γ, and θ give

$$\begin{aligned} \alpha \sqrt{\gamma (U^T)^2 + \mathbb{S}^{T+1}} &= \Xi \sqrt{2^{10}(U^T)^2 + Z^{T+1}} , \\ \frac{\theta}{m\alpha} \sqrt{\gamma (U^{t-1})^2 + \mathbb{S}^t} &= \frac{\Xi}{m^3 \log T} \sqrt{2^{10}(U^{t-1})^2 + Z^t} . \end{aligned}$$

Using these identities in (49), we obtain

$$\begin{aligned} 0 \le \mathsf{SwapReg}_{x_i}^T &\le 2\Xi \sqrt{2^{10}(U^T)^2 + Z^{T+1}} + 2\Xi \sqrt{Z^{T+1}} - \frac{\Xi}{192 m^3 \log T} W_i + 11 U_{\max} \\ &\le \Xi \left(75 U_{\max} + 4\sqrt{Z^{T+1}} - \frac{W_i}{192 m^3 \log T} \right) , \end{aligned} \tag{60}$$

where the first inequality follows from $\mathsf{SwapReg}_{x_i}^T(\mathbf{M}) = 0$ for the identity matrix $\mathbf{M}$, the second inequality from $P_\infty^{T+1}(\bar{u}_i) \le \mathbb{S}^{T+1}$, and the last inequality from $U^T \le U_{\max}$, $\Xi \ge 1$, and $\sqrt{2^{10}(U^T)^2 + Z^{T+1}} \le 32 U_{\max} + \sqrt{Z^{T+1}}$. Rearranging (60) and summing over $i \in [n]$ gives

$$W \le 192 m^3 \log T\, n (75 U_{\max} + 4\sqrt{Z^{T+1}}) . \tag{61}$$

We next relate W to the increments of $\mathbb{S}^t$. Let $e_i^t = u_i^t - \bar{u}_i^t$, and recall $E^t = \max_i \|e_i^t\|_\infty$ from Lemma 14. For $t \geq 3$, (43) and Lemma 16 give

$$\begin{aligned}\Delta^t &\leq \max_{i\in[n]} \|\bar{u}_i^t - \bar{u}_i^{t-1}\|_\infty^2 \\ &\leq 2\max_{i\in[n]} \|u_i^t - u_i^{t-1}\|_\infty^2 + 2\max_{i\in[n]} \|e_i^t - e_i^{t-1}\|_\infty^2 \\ &\leq 2U_{\max}^2(n-1)\sum_{i\in[n]} \|x_i^t - x_i^{t-1}\|_1^2 + 4(E^t)^2 + 4(E^{t-1})^2 \,. \end{aligned} \tag{62}$$

Moreover, (48) applied at round $t-1$ gives

$$\sqrt{Z^t} \leq \sqrt{2^{10}(U^{t-2})^2 + Z^{t-1}} \quad \text{for } t \geq 3 \,. \tag{63}$$

Multiplying (62) by $\sqrt{Z^t}/(m^3 \log T)$, summing over $t = 3, \dots, T$, and using (63), we obtain

$$\begin{aligned}\sum_{t=3}^T \sqrt{Z^t}\,\frac{\Delta^t}{m^3 \log T} &\leq \frac{2U_{\max}^2(n-1)}{m^3 \log T} \sum_{t=3}^T \sqrt{2^{10}(U^{t-2})^2 + Z^{t-1}} \sum_{i\in[n]} \|x_i^t - x_i^{t-1}\|_1^2 \\ &\quad + \frac{4\sqrt{Z^{T+1}}}{m^3 \log T} \sum_{t=3}^T ((E^t)^2 + (E^{t-1})^2) \\ &\leq \frac{2U_{\max}^2(n-1)}{m^3 \log T} W + \frac{16U_{\max}^2}{m^3 \log T}\sqrt{Z^{T+1}} \,, \end{aligned}$$

where the last inequality follows by reindexing the first sum with $s = t-1$ and using $\sum_{t=3}^T ((E^t)^2 + (E^{t-1})^2) \leq 2\sum_{t=1}^T (E^t)^2 \leq 4U_{\max}^2$ from Lemma 14. Since $\|\bar{u}_i^t\|_\infty \leq U_{\max}$, (43) also gives $\Delta^t \leq 4U_{\max}^2$ for every $t \in [T]$. The term at $t = 1$ vanishes since $Z^1 = 0$, and the term at $t = 2$ is at most $4U_{\max}^2\sqrt{Z^{T+1}}/(m^3 \log T)$. Thus,

$$\sum_{t=1}^T \sqrt{Z^t}\,\frac{\Delta^t}{m^3 \log T} \leq \frac{2U_{\max}^2(n-1)}{m^3 \log T} W + \frac{20U_{\max}^2}{m^3 \log T}\sqrt{Z^{T+1}} \,. \tag{64}$$

We next lower bound the LHS of (64). The increments of Z^t equal $\Delta^t/(m^3 \log T)$ and are at most $4U_{\max}^2/(m^3 \log T)$. Hence, Lemma 17 applied to $(Z^t)_t$ gives

$$\sum_{t=1}^T \sqrt{Z^t}\,\frac{\Delta^t}{m^3 \log T} \geq \frac{2}{3}(Z^{T+1})^{3/2} - \frac{2U_{\max}}{\sqrt{m^3 \log T}} Z^{T+1} \,. \tag{65}$$

Combining (61), (64) and (65) yields

$$\begin{aligned}\frac{2}{3}(Z^{T+1})^{3/2} &\leq 384U_{\max}^2 n(n-1)(75U_{\max} + 4\sqrt{Z^{T+1}}) \\ &\quad + \frac{20U_{\max}^2}{m^3 \log T}\sqrt{Z^{T+1}} + \frac{2U_{\max}}{\sqrt{m^3 \log T}} Z^{T+1} \,. \end{aligned} \tag{66}$$

Since $U^1 > 0$ implies $U_{\max} > 0$, we can set $X = \sqrt{Z^{T+1}}/U_{\max}$ and divide (66) by $U_{\max}^3$. Using $n - 1 \leq n$ and $m^3 \log T \geq 1$, we obtain

$$\frac{2}{3}X^3 \leq 28800n^2 + (1536n^2 + 20)X + 2X^2 \,.$$

Since at least one of the three terms on the RHS is at least $2X^3/9$, we obtain

$$X \leq \max\left\{(129600n^2)^{1/3}, \sqrt{\tfrac{9}{2}(1536n^2 + 20)}, 9\right\} \leq 96n \,,$$

where the last inequality follows from $n \geq 2$. Consequently,

$$\mathbb{S}^{T+1} \leq 96^2 U_{\max}^2 n^2 m^3 \log T \,.$$

Finally, substituting $\sqrt{Z^{T+1}} \leq 96nU_{\max}$ into (60) and discarding the third term on its RHS, which is nonpositive, we obtain

$$\mathsf{SwapReg}_{x_i}^T \leq \Xi U_{\max}(75 + 384n) \leq 512U_{\max} n m^2 \sqrt{\log m \log T} \,.$$

This proves the regret bound. The correlated-equilibrium guarantee follows from the result of Foster & Vohra (1997) in Section 2.3, which completes the proof. ■

C.4 Scale-agnostic and scale-free swap regret minimization without fast convergence

Here we show that the regret bound labeled "Scale-free OLO" in Table 1 can be achieved by scale-free online learning. Recall that the swap regret can be expressed as the sum of the external regrets of m_i external regret minimizers, as discussed in Section 4.1. Hence, it suffices to construct an external regret minimization algorithm for each expert a of each player i.

Let $M_i = \max\{4, \log m_i/2^{3/2}\}$ for each $i \in [n]$. Then, for each expert $a \in \mathcal{A}_i$ of player $i \in [n]$, we use the Hedge algorithm to compute $y_{i,a}^t \in \Delta_{m_i}$ by

$$y_{i,a}^t(b) \propto \exp\left(\eta_{i,a}^t \sum_{s=1}^{t-1} u_{i,a}^s(b)\right), \quad \eta_{i,a}^t = \sqrt{\frac{M_i}{\sum_{s=1}^{t-1}\|u_{i,a}^s\|_\infty^2}} \tag{67}$$

for $b \in [m_i]$, where we note that $y_{i,a}^1 = \frac{1}{m_i}\mathbf{1}$. If the denominator of the learning rate $\eta_{i,a}^t$ is zero, we set $\eta_{i,a}^t = \infty$ and $y_{i,a}^t = \mathbf{1}/m_i$.

Proposition 2. *The learning dynamic based on (67) guarantees*

$$\mathsf{SwapReg}_{x_i}^T = O(U_{\max}\sqrt{Tm_i \log m_i})$$

for every player $i \in [n]$.

Proof. If $m_i = 1$, then $\mathsf{SwapReg}_{x_i}^T = 0$, so the claim is immediate. Hence, we assume $m_i \geq 2$ in what follows. We apply Lemma 4 in Appendix A to each expert with $h^t = -u_{i,a}^t$, $m^t = \mathbf{0}$, and $\nu^t = 0$. With $c = \max\{2, \sqrt{\log m_i}/2^{3/4}\}$, the learning rate is $\eta_{i,a}^t$ in (67). Since $\log_+(m_i) \leq 32M_i$, for every $y \in \Delta_{m_i}$ we obtain

$$\mathsf{Reg}_{i,a}^T(y) = \sum_{t=1}^T \langle y - y_{i,a}^t, u_{i,a}^t\rangle \leq \sqrt{32\sum_{t=1}^T \|u_{i,a}^t\|_\infty^2 M_i} \leq \sqrt{32\sum_{t=1}^T x_i^t(a)\|u_i^t\|_\infty^2 M_i}\,, \tag{68}$$

where we used $u_{i,a}^t = x_i^t(a)u_i^t$. Hence, using the reduction from swap regret minimization to the instances of external regret minimization discussed in Section 4.1 and (68), we have

$$\begin{aligned}
\mathsf{SwapReg}_{x_i}^T = \sum_{a\in\mathcal{A}_i} \mathsf{Reg}_{i,a}^T &\leq \sum_{a\in\mathcal{A}_i} \sqrt{32\sum_{t=1}^T x_i^t(a)\|u_i^t\|_\infty^2 M_i} \\
&\leq \sqrt{32m_i \sum_{t=1}^T \sum_{a\in\mathcal{A}_i} x_i^t(a)\|u_i^t\|_\infty^2 M_i} \\
&\leq U_{\max}\sqrt{32m_i M_i T} = O(U_{\max}\sqrt{Tm_i \log m_i})\,,
\end{aligned}$$

where the second inequality follows from the Cauchy–Schwarz inequality and the last inequality from $\sum_{a\in\mathcal{A}_i} x_i^t(a) = 1$ and $\|u_i^t\|_\infty \leq U_{\max}$. This completes the proof. ■

D Additional Related Work

Optimistic learning has been a central tool for achieving fast convergence to equilibrium. Starting from Rakhlin & Sridharan (2013); Syrgkanis et al. (2015), there have been many subsequent developments, *e.g.,* Foster et al. (2016); Chen & Peng (2020); Anagnostides et al. (2022a;b).

Scale-free learning was first studied for prediction with expert advice (Cesa-Bianchi et al., 2007), and has since seen many refinements such as de Rooij et al. (2014); Mhammedi et al. (2019), to name a few. In online convex optimization, which includes online linear optimization as a special case, scale-free and parameter-free methods have been developed in a series of works such as Orabona & Pál (2018); Cutkosky (2019); Chen et al. (2021). See also Orabona (2019) for a comprehensive introduction.

D.1 Comparison with Zhang et al. (2025)

To obtain a fast convergence rate of $O(1/T)$, the regret analysis of Zhang et al. (2025) assumes $\delta = \min\{\|g^1\|, \|\ell^1\|\} > 0$. However, this δ cannot be known to each player in advance and

can be arbitrarily small. Theorem 1 affirmatively answers the question raised in the conclusion of Zhang et al. (2025) of whether one can retain the scale-free fast convergence while removing the dependence on the $1/\delta$ factor in their regret bound. As shown in the proof of Theorem 1 in Appendix B.2, we address this issue by introducing an appropriate stopping time and developing a new analysis that effectively leverages the negative term in the RVU bound.

Note that, if one can assume $\delta > 0$, then simply replacing the constants $\log_+(m_x)$ and $\log_+(m_y)$ in the denominators of the learning rates in Tsuchiya et al. (2025) by $4\delta^2$ yields a scale-free and scale-agnostic learning dynamic with a regret bound of $O(\frac{A_{\mathrm{diff}} \log m}{\delta})$ without prior knowledge of δ, whose rate is better than that of Zhang et al. (2025). As can also be seen in the swap-regret analysis in Appendix C and in Appendix B.5, the main difficulty of scale-agnostic learning arises when the scale parameters are small, and thus assuming $\delta > 0$ allows us to achieve scale-free and scale-agnostic fast convergence more easily. Nevertheless, Zhang et al. (2025) primarily aim to explain the empirical success of regret matching in zero-sum games from a scale-freeness perspective and to propose a new regret-matching variant, making their contribution complementary to ours.

Moreover, their learning dynamic uses optimistic online gradient descent, rather than optimistic Hedge (*i.e.,* optimistic FTRL with the negative Shannon entropy regularizer). As a result, in two-player zero-sum games their dynamic yields a polynomial dependence on the number of actions m. One can replace it with optimistic online mirror descent with the negative Shannon entropy regularizer to resolve this, but an additional $\log T$ factor seems inevitable even with a refined analysis. Thus, it is preferable to use optimistic FTRL.

E Numerical Experiments

This section examines how the payoff scale and payoff range affect the convergence of the learning dynamics in two-player zero-sum games through numerical experiments. Since our dynamics in (1) are scale-free and use gradients centered at the first round, their strategies do not change when all entries of the payoff matrix are multiplied by a positive constant or when a constant is added to all entries. We compare our dynamics with two variants of optimistic Hedge whose learning rates are set for $A_{\max} = 1$ as in most existing work, and refer to these variants as the existing dynamics. We observe that the convergence of our dynamics, measured relative to A_{diff}, does not depend on $A_{\max}$ or A_{diff}, whereas the existing dynamics converge much more slowly when $A_{\max}$ or A_{diff} is small.

E.1 Setup

Games. We generate 30 matrices of size 32×32, one for each random seed, whose entries are drawn independently from the uniform distribution on $[-1, 1]$ (as in Wei et al., 2021, Section 6), and divide each matrix by its maximum absolute entry to obtain A_0. From each A_0, we construct two sets of games.

- Games with varying $A_{\max}$. We multiply A_0 by $A_{\max} \in \{10^{-4}, 10^{-3}, \dots, 10^{4}\}$, which yields games whose maximum absolute entry is $A_{\max}$. The ratio $A_{\mathrm{diff}}/A_{\max}$ does not depend on $A_{\max}$.
- Games with varying A_{diff}. We apply to the entries of A_0 the increasing affine map that sends the minimum entry to $1 - A_{\mathrm{diff}}$ and the maximum entry to 1 for $A_{\mathrm{diff}} \in \{10^{-3}, 10^{-2}, 10^{-1}, 1\}$, which yields games with $A_{\max} = 1$ and payoff range A_{diff}.

Both constructions multiply all entries of A_0 by a positive constant and add a constant, and thus preserve the Nash equilibria.

Compared methods. We compare the following three learning dynamics, all of which are instances of optimistic Hedge.

- Our dynamics in (1), whose learning rates use the squared path lengths of the gradients of both players.
- Optimistic Hedge with the fixed learning rate $\eta = 1/2$ (Syrgkanis et al., 2015, Corollary 8). We also use $\eta \in \{1/8, 1/4, 1, 2\}$ to examine the dependence on the learning rate.

- Optimistic Hedge with the adaptive learning rate (Tsuchiya et al., 2025, Section 4.1). For $m_x = 32$, the learning rate of the x-player at round t is $\sqrt{2/(4 + \sum_{s=1}^{t-1} \|g^s - g^{s-1}\|_\infty^2)} \leq 1/\sqrt{2}$ with $g^0 = \mathbf{0}$, and that of the y-player is defined in the same way with ℓ in place of g.

The latter two use parameters set for $A_{\max} = 1$ and do not adjust them to the actual $A_{\max}$ or A_{diff}. All three dynamics start from the uniform strategies, and the two players update their strategies simultaneously using the exact gradients $g^t = Ay^t$ and $\ell^t = A^\top x^t$ for $T = 10^4$ rounds.

Evaluation. We measure the convergence of the average strategies $\bar{x}^T = \frac{1}{T}\sum_{t=1}^T x^t$ and $\bar{y}^T = \frac{1}{T}\sum_{t=1}^T y^t$ by the Nash gap

$$\mathrm{Gap}(\bar{x}^T, \bar{y}^T) \coloneqq \max_{a \in [m_x]} (A\bar{y}^T)(a) - \min_{b \in [m_y]} (A^\top \bar{x}^T)(b) = \frac{\mathsf{Reg}_x^T + \mathsf{Reg}_y^T}{T}.$$

Since Theorem 1 bounds Reg_x^T and Reg_y^T of our dynamics by $O(A_{\mathrm{diff}} \log m)$, we report the Nash gap and the regrets divided by A_{diff}, which makes the values comparable across different $A_{\max}$ and A_{diff}. The learning dynamics do not use A_{diff}, which is used only for this evaluation. The line plots show the medians and the 25th and 75th percentiles over the random seeds.

E.2 Results

Our dynamics. The Nash gap of our dynamics divided by A_{diff} is the same for all $A_{\max}$ and A_{diff}. Figure 1 shows this for the Nash gap after $T = 10^4$ rounds, and Figures 2 and 3 show that it holds at every round and that the Nash gap decreases in proportion to $1/T$, as predicted by Theorem 1. This is because both constructions of the games leave the strategies of our dynamics unchanged. Indeed, for each random seed, the Nash gaps and regrets of our dynamics divided by A_{diff} agree across all 13 games within a relative difference of 10^{-6} at every round.

Existing dynamics. The existing dynamics converge more slowly than our dynamics when $A_{\max}$ or A_{diff} is small. For $A_{\max} \leq 10^{-3}$, their Nash gaps divided by A_{diff} stay close to the initial values for 10^4 rounds (Figure 2), and their final values are more than 10^3 times larger than ours (Figure 1). This is because their learning rates are bounded by constants set for $A_{\max} = 1$, and thus their updates are too small when the payoffs are small. The same holds as A_{diff} decreases, even though their parameters are set for the correct value $A_{\max} = 1$ (Figures 1 and 3). Since optimistic Hedge depends on the gradients only through the differences between their entries, which are at most A_{diff} in absolute value, the same bounds on the learning rates make the updates too small when A_{diff} is small. When $A_{\max} \geq 10$, the adaptive learning rate becomes small as the squared path length grows, and the Nash gap of optimistic Hedge with this learning rate is comparable to ours. In contrast, the fixed learning rate $\eta = 1/2$ increases the Nash gap in the first few rounds for $A_{\max} \geq 10^2$ (Figure 2) and yields final Nash gaps six to eleven times larger than ours for $A_{\max} \geq 10$ (Figure 1).

Fixed learning rates. None of the tested fixed learning rates performs comparably to our dynamics for all $A_{\max}$. Figure 4 shows the medians of the Nash gaps divided by A_{diff} after $T = 10^4$ rounds for all tested fixed learning rates, together with those of our dynamics and the adaptive learning rate. A fixed learning rate achieves a Nash gap within a factor of two of ours only for $A_{\max} \in \{1, 10\}$ and for $A_{\mathrm{diff}} = 1$, and different values of η are needed for $A_{\max} = 1$ and $A_{\max} = 10$. Thus, choosing a fixed learning rate that performs well requires prior knowledge of the payoff scale.

Longer runs. To examine whether the existing dynamics catch up with ours after more rounds, we also run the three dynamics for $T = 10^5$ rounds at $A_{\max} = 10^{-3}$ and at $A_{\mathrm{diff}} = 10^{-3}$ with $A_{\max} = 1$, using the first 10 of the 30 matrices. Figure 5 shows that the Nash gaps of the existing dynamics start to decrease only after about 10^4 rounds and are still more than 10^3 times larger than ours at $T = 10^5$. Figure 6 shows that the individual regrets of our dynamics stay below $0.5A_{\mathrm{diff}}$ at all rounds, in line with Theorem 1, whereas those of the existing dynamics grow almost linearly in T up to more than $10^3 A_{\mathrm{diff}}$.

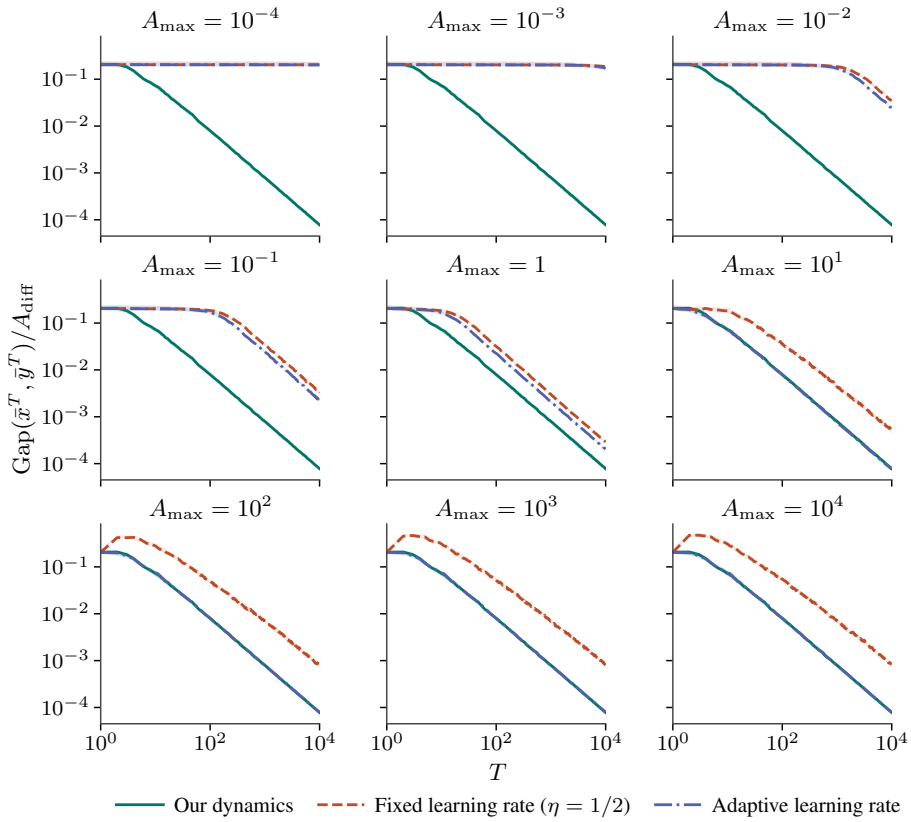

Figure 2: Nash gap divided by A_{diff} for the games with varying $A_{\max}$. Lines and shaded regions show the medians and the 25th and 75th percentiles over 30 random seeds. All panels use the same axis limits.

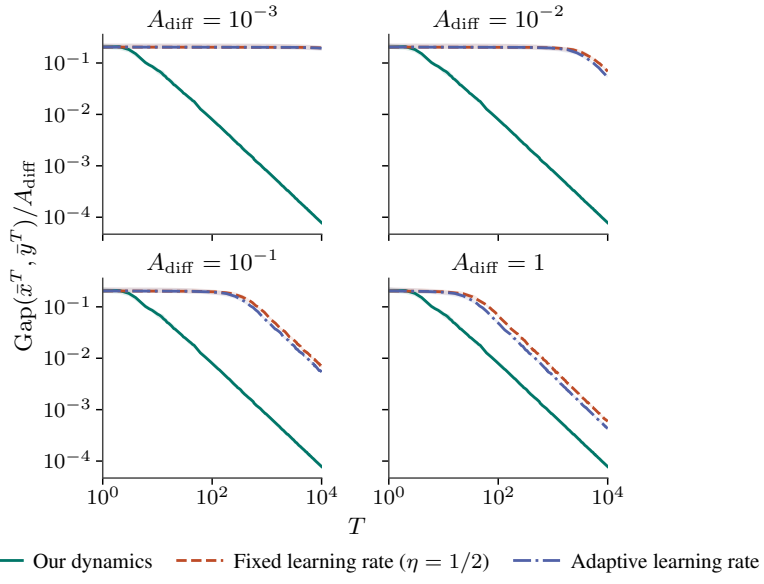

Figure 3: Nash gap divided by A_{diff} for the games with varying A_{diff} and $A_{\max} = 1$. Lines and shaded regions show the medians and the 25th and 75th percentiles over 30 random seeds. All panels use the same axis limits.

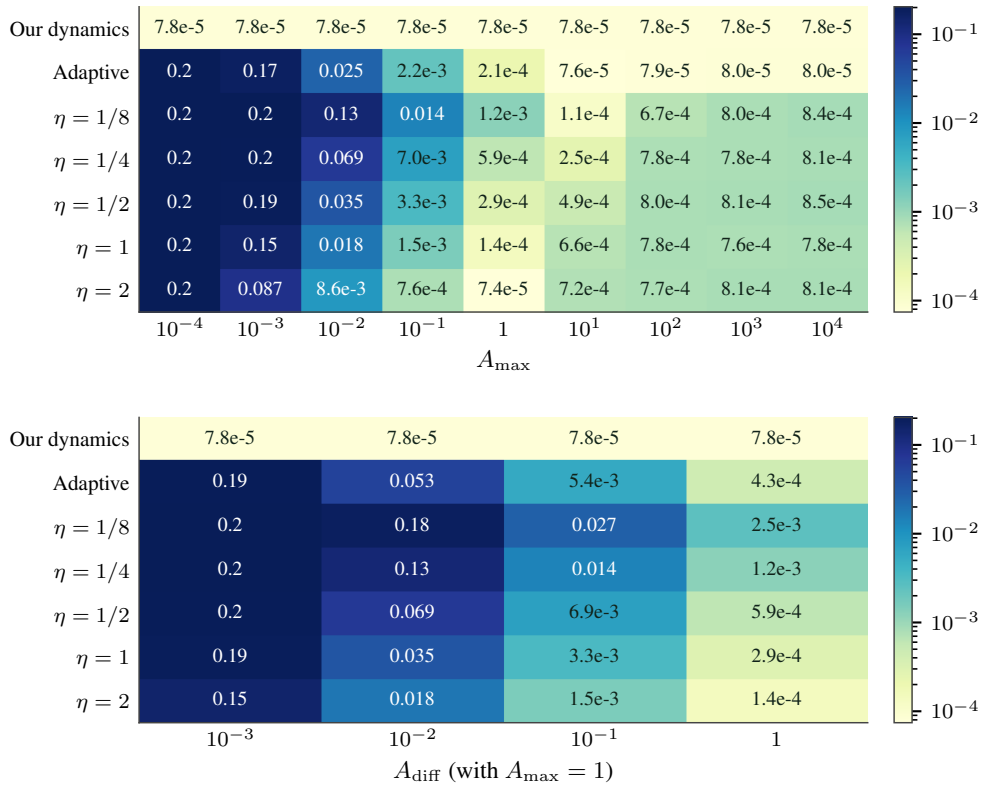

Figure 4: Medians of the Nash gaps divided by A_{diff} after $T = 10^4$ rounds over 30 random seeds, for the games with varying $A_{\max}$ (top) and those with varying A_{diff} and $A_{\max} = 1$ (bottom). The first two rows show our dynamics and optimistic Hedge with the adaptive learning rate, and the other rows show optimistic Hedge with the indicated fixed learning rate. Both panels use the same logarithmic color scale.

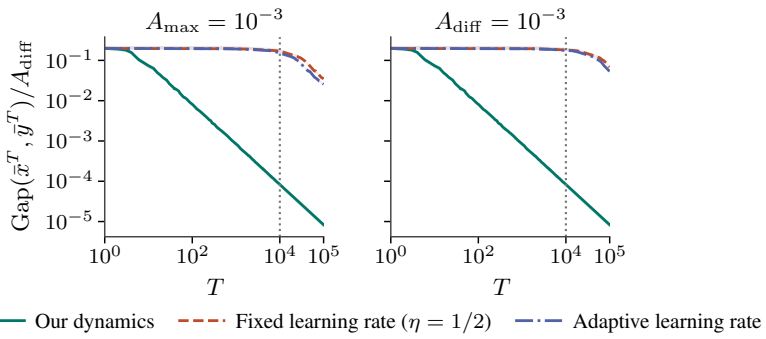

Figure 5: Nash gap divided by A_{diff} for $T = 10^5$ rounds at $A_{\max} = 10^{-3}$ (left) and at $A_{\text{diff}} = 10^{-3}$ with $A_{\max} = 1$ (right). Lines and shaded regions show the medians and the 25th and 75th percentiles over the first 10 random seeds. The dotted vertical lines mark $T = 10^4$.

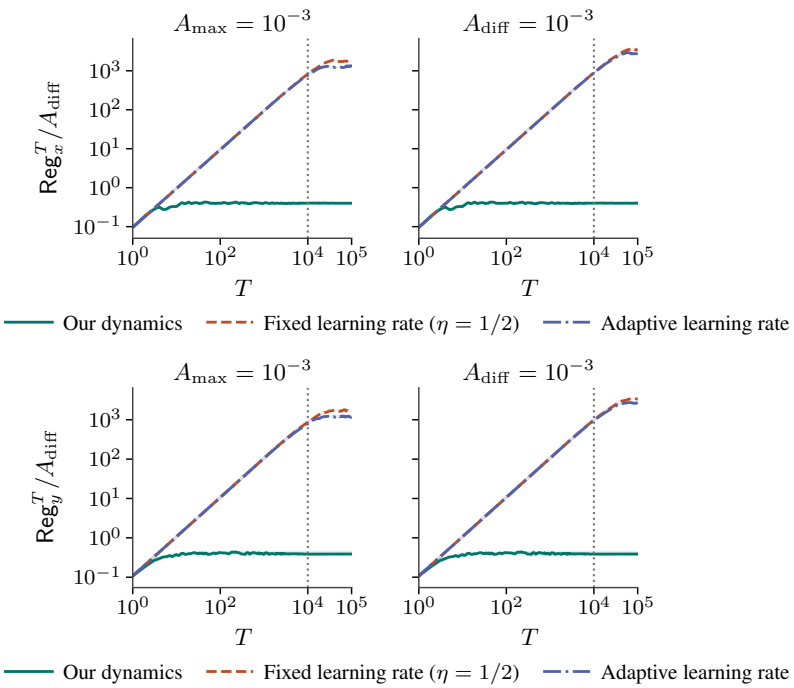

Figure 6: $\text{Reg}_x^T / A_{\text{diff}}$ (top) and $\text{Reg}_y^T / A_{\text{diff}}$ (bottom) in the runs of Figure 5. Lines and shaded regions show the medians and the 25th and 75th percentiles over the first 10 random seeds.