LAST-ITERATE CONVERGENCE OF PARAMETER-FREE AND SCALE-INVARIANT REGRET MATCHING$^+$

Anonymous authors
Paper under double-blind review

ABSTRACT

Regret Matching$^+$ (RM$^+$) is widely used as a local regret minimizer in counterfactual regret minimization (CFR) for two-player zero-sum extensive-form games. RM$^+$ and its predictive variants are parameter-free and invariant to positive rescaling of the payoffs, and existing guarantees control their average or best iterates. Their last-iterate behavior, however, has remained theoretically unresolved even in two-player zero-sum matrix games, where their self-play dynamics can be studied in isolation. We study two predictive RM$^+$ dynamics in this setting. The first is the Increasing Regret Extra-Gradient Predictive Regret Matching$^+$ (IREG-PRM$^+$) dynamics of Zhang et al. (2025), and the second is the Increasing-Regret Optimistic PRM$^+$ (IRO-PRM$^+$) dynamics that we introduce. We prove that the played strategy profiles of both dynamics converge to the Nash equilibrium set for every two-player zero-sum matrix game. For IREG-PRM$^+$, we further bound the cumulative duality gap by $\mathcal{O}(\sqrt{T})$. To the best of our knowledge, these are the first last-iterate convergence guarantees for a parameter-free and payoff-scale-invariant RM$^+$-type algorithm. Experiments on structured and random matrix games support the theoretical results.

1 INTRODUCTION

Counterfactual regret minimization (CFR) has reached major milestones in large two-player zero-sum imperfect-information games, including superhuman no-limit poker, by running a local regret minimizer at each information set (Zinkevich et al., 2007; Bowling et al., 2015; Moravčík et al., 2017; Brown & Sandholm, 2018). Regret Matching (RM) and Regret Matching$^+$ (RM$^+$) are the regret minimizers most widely used in this role (Hart & Mas-Colell, 2000; Tammelin, 2014). Predictive Regret Matching$^+$ (PRM$^+$) extends RM$^+$ by incorporating a prediction of the next utility vector (Farina et al., 2021). These algorithms are attractive because they require no step-size parameter and are invariant to positive rescaling of the payoffs. We study the self-play dynamics of these learners in two-player zero-sum matrix games, where each player's strategy set is a single simplex. When two such no-regret learners play such a game against each other, their average strategy profile converges to a Nash equilibrium, a guarantee known as *average-iterate convergence* (Freund & Schapire, 1999). However, this guarantee controls the average strategy profile rather than the strategy actually played at each iteration.

The played strategy profiles of standard no-regret dynamics can cycle around the equilibrium instead of converging (Mertikopoulos et al., 2018; Bailey & Piliouras, 2018). Such cycling is problematic when maintaining an average of the entire trajectory is undesirable or when the final strategy is deployed directly, as in large-scale game solving (Bakhtin et al., 2023). A stronger guarantee that rules out such cycling is *last-iterate convergence*, in which the played strategy profile itself approaches the equilibrium set. For RM$^+$, however, last-iterate convergence is difficult to attain, since vanilla RM$^+$ and predictive RM$^+$ can fail to converge in the last iterate (Farina et al., 2023). Moreover, the existing guarantees for RM$^+$ recover it only by adding explicit regularization and a step size, unlike the parameter-free RM$^+$ updates used in practice (Cai et al., 2025a).

Seeking stronger theoretical guarantees without losing the parameter-free and payoff-scale-invariant properties, Zhang et al. (2025) introduced Increasing Regret Extra-Gradient Predictive Regret Matching$^+$ (IREG-PRM$^+$). Their design addresses an instability of RM$^+$ near the zero regret vector. In RM$^+$, the played strategy is the regret vector normalized to the simplex, so it depends only

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.

on the direction of that vector and is ill-defined near the zero vector. By keeping the Euclidean norm of the regret vector from decreasing, the IREG-PRM$^+$ update prevents this vector from shrinking toward zero and thereby stabilizes the played strategy. In two-player zero-sum matrix games, this update achieves an $\mathcal{O}(T^{-1})$ average-iterate guarantee and an $\mathcal{O}(T^{-1/2})$ best-iterate guarantee. Their experiments further show rapid convergence of the last iterates. These guarantees, however, do not explain that observation. A best-iterate guarantee controls only the best strategy profile among the first T iterations, which does not imply that the played strategy profiles (x^t, y^t) converge as t grows.

Building on the framework of Zhang et al. (2025), we establish last-iterate convergence for two parameter-free and payoff-scale-invariant predictive RM$^+$ dynamics. We prove that for every two-player zero-sum matrix game, the played strategy profiles generated by either dynamics from zero initial regret vectors converge to the set of Nash equilibria. To the best of our knowledge, this is the first last-iterate convergence guarantee for a parameter-free and payoff-scale-invariant RM$^+$-type algorithm.

In proving last-iterate convergence, we must control how far each played strategy moves from its prediction. The RM$^+$ regret bound weights this displacement by the regret-vector norm, rather than by a fixed step-size constant. Both players start with zero regret, so this weight is initially zero, and the bound gives no control. This vanishing weight is why the earlier analysis reaches only a best-iterate guarantee. The increasing-regret construction keeps this norm from decreasing, so the weight stays positive once a player's regret becomes nonzero. A player whose regret stays zero instead keeps playing a best response. In both cases, each player's gain from switching to a best response vanishes, so the duality gap vanishes without a step size tied to the payoff matrix.

We summarize our contributions as follows. First, we prove that the extragradient IREG-PRM$^+$, which uses two gradient evaluations per iteration, converges in the last iterate and that its cumulative duality gap is $\mathcal{O}(\sqrt{T})$. Second, we introduce IRO-PRM$^+$, a single-call variant that predicts with the previous payoff vector and uses only one new gradient evaluation per iteration, and establish the same last-iterate convergence guarantee. Finally, our experiments corroborate both guarantees, with the two dynamics driving the last-iterate duality gap toward zero where vanilla and predictive RM$^+$ do not.

2 Related Literature

Last-iterate convergence in games. In two-player zero-sum games, no-regret learning guarantees convergence of the average strategy profile (Zinkevich et al., 2007), whereas the strategy produced at the current iteration may cycle around equilibrium (Mertikopoulos et al., 2018; Bailey & Piliouras, 2018). Last-iterate convergence has been established for several methods in zero-sum and monotone games, including optimistic and extragradient methods. Representative examples include optimistic multiplicative weights update (Daskalakis & Panageas, 2019; Wei et al., 2021), optimistic mirror descent (Rakhlin & Sridharan, 2013; Mertikopoulos et al., 2019), and extragradient methods (Korpelevich, 1976; Golowich et al., 2020b). For these methods, explicit last-iterate rates are known in both two-player and multi-player settings (Golowich et al., 2020a; Cai et al., 2022; Gorbunov et al., 2022). Related guarantees have also been established for extensive-form games in the sequence-form representation (von Stengel, 1996; Lee et al., 2021). These methods, however, require a step-size or regularization parameter that is generally tied to the payoff operator, and hence do not inherit the payoff-scale invariance of RM$^+$.

Regret matching and CFR for extensive-form games. Regret matching and RM$^+$ are local regret minimizers used by counterfactual regret minimization (CFR) to solve large two-player zero-sum extensive-form games (Hart & Mas-Colell, 2000; Zinkevich et al., 2007; Tammelin, 2014). Both updates derive from Blackwell approachability, which is equivalent to no-regret learning (Blackwell, 1956; Abernethy et al., 2011; Farina et al., 2021). Practical CFR variants include Monte Carlo sampling, discounting, and predictive regret minimization, while related deep-learning methods address large-scale game solving (Lanctot et al., 2009; Brown & Sandholm, 2019; Farina et al., 2021; 2019; 2022; Brown et al., 2019; 2020). The local RM$^+$ update acts on a simplex and is not itself a solver for the composed extensive-form game. For this reason, the last-iterate behavior of RM$^+$ has been studied first in two-player zero-sum matrix games.

Table 1: Convergence guarantees for RM$^+$-type self-play dynamics in two-player zero-sum matrix games. The IREG-PRM$^+$ and IRO-PRM$^+$ dynamics analyzed here are the first that are simultaneously parameter-free, payoff-scale-invariant, and last-iterate convergent.

Dynamics	Parameter-free	Scale-invariant	Convergence guarantee
RM$^+$ (Tammelin, 2014)	✓	✓	average-iterate
PRM$^+$ (Farina et al., 2021)	✓	✓	average-iterate
Smoothed RM$^+$ (Cai et al., 2025a)	×	×	last-iterate
IREG-PRM$^+$ (Zhang et al., 2025)	✓	✓	best-iterate; **last-iterate (this paper)**
IRO-PRM$^+$ (this paper)	✓	✓	**last-iterate**

Last-iterate convergence of RM$^+$ dynamics. Numerical examples show nonconvergent last-iterate behavior for vanilla RM$^+$ and predictive RM$^+$ (Farina et al., 2023; Cai et al., 2025a). Cai et al. (2025a) establish asymptotic last-iterate convergence for extragradient RM$^+$ and smooth predictive RM$^+$, and obtain a linear rate after adding restarts. Their constructions use a smoothed domain and a step size depending on the regret operator. Thus, they do not provide a parameter-free, payoff-scale-invariant counterpart of the RM$^+$ updates used in practice. An alternative black-box reduction converts the average iterates generated by an uncoupled learning algorithm into the last iterates of a new dynamics (Cai et al., 2025b). The reduced dynamics is different from the original RM$^+$ recurrence because its current output encodes the base learner's average. Closest to our work, Zhang et al. (2025) introduced parameter-free and payoff-scale-invariant predictive RM$^+$ dynamics based on an equal-norm prediction shift. Their analysis of IREG-PRM$^+$ establishes average-iterate and best-iterate convergence through a regret bounded by variation in utilities (RVU) inequality (Syrgkanis et al., 2015). The negative squared-displacement term in this inequality is weighted by the regret-vector norm, rather than by a fixed positive constant, and the weight can vanish at the zero-regret state. Consequently, their result does not control every played strategy profile. We analyze the same equal-norm-shift construction and establish asymptotic last-iterate convergence of the played strategy profiles. We further prove the analogous guarantee for a single-call optimistic variant that replaces the extragradient query with the previous payoff vector. Table 1 places these guarantees side by side.

Parameter-free and scale-invariant online learning. A separate line of work in online learning removes the dependence on a step-size parameter. Parameter-free algorithms attain near-optimal regret without tuning a step size to the horizon or to the comparator norm (Orabona & Pál, 2016; Cutkosky & Orabona, 2018). Scale-free algorithms further make the regret invariant to the scale of the loss vectors (Orabona & Pál, 2018). RM$^+$ shares both properties, since its update uses no step size and remains invariant to positive rescaling of the payoffs. The dynamics we study retain these two properties and additionally attain last-iterate convergence. These general parameter-free analyses control regret rather than the last iterate. Hence, they do not yield such a guarantee.

3 Preliminaries

Two-player zero-sum matrix game. In this study, we focus on the two-player zero-sum matrix game

$$\max_{x\in\mathcal{X}} \min_{y\in\mathcal{Y}} x^\mathsf{T} A y, \tag{1}$$

where $A \in \mathbb{R}^{d_x \times d_y}$ is the payoff matrix, and the row player x and the column player y choose their strategies from the probability simplices $\mathcal{X} = \Delta_{d_x}$ and $\mathcal{Y} = \Delta_{d_y}$ over their d_x and d_y actions. Here, $\Delta_d = \{v \in \mathbb{R}^d_+ : \mathbf{1}^\mathsf{T} v = 1\}$, $\mathbb{R}^d_+$ is the nonnegative orthant, and $\mathbf{1}$ is the all-ones vector. We consider an online learning setting, where the following process repeats from iteration $t = 1$ to T: (i) each player chooses a strategy, $x^t \in \mathcal{X}$ and $y^t \in \mathcal{Y}$, based on previously observed feedback; (ii) each player receives its utility vector as feedback, given by

$$u_x^t = Ay^t, \qquad u_y^t = -A^\mathsf{T} x^t. \tag{2}$$

Nash equilibrium and duality gap. We use the Nash equilibrium (Nash, 1951) as the solution concept for the matrix game (1). A strategy profile $(x^\star, y^\star) \in \mathcal{X} \times \mathcal{Y}$ is a Nash equilibrium when

neither player can profit from a unilateral deviation, that is,

$$x^\mathsf{T} A y^\star \le (x^\star)^\mathsf{T} A y^\star \le (x^\star)^\mathsf{T} A y \quad \text{for every } (x, y) \in \mathcal{X} \times \mathcal{Y}.$$

Such a profile always exists by the minimax theorem (von Neumann, 1928). Let $Z^\star$ denote the set of Nash equilibria. The duality gap at (x, y) is

$$\mathrm{Gap}(x, y) = \max_{x' \in \mathcal{X}} (x')^\mathsf{T} A y - \min_{y' \in \mathcal{Y}} x^\mathsf{T} A y'. \tag{3}$$

It is a standard measure of proximity to a Nash equilibrium (Farina et al., 2021). Moreover, it is continuous and nonnegative on $\mathcal{X} \times \mathcal{Y}$, and it vanishes exactly on $Z^\star$.

Other notation. For a vector v, $\|v\|_1$ and $\|v\|_2$ denote its ℓ_1 and Euclidean norms, and $[v]_+$ its coordinatewise positive part. For vectors v and w, $\langle v, w \rangle$ denotes their Euclidean inner product. For a point z and a nonempty closed set Z in a Euclidean space, $\mathrm{dist}(z, Z) = \min_{z' \in Z} \|z - z'\|_2$. We equip $\mathcal{X} \times \mathcal{Y}$ with the Euclidean norm.

4 Predictive RM$^+$ Dynamics

We present the two predictive RM$^+$ dynamics whose last-iterate convergence we establish in Section 5. Both run the same learner on each player's simplex. They differ only in the prediction supplied to that learner and in the number of gradient evaluations per iteration. Section 4.1 describes this common learner and the self-play loop that drives it. Section 4.2 then gives the two predictions that instantiate the loop.

4.1 Common learner update

This common learner is the parameter-free predictive RM$^+$ construction of Zhang et al. (2025). Since it extends RM$^+$, we first recall the RM$^+$ update (Tammelin, 2014). For each player $i \in \{x, y\}$, RM$^+$ maintains a regret vector $q_i^t \in \mathbb{R}^{d_i}_+$, plays its normalization, and updates it from the observed utility u_i^t:

$$z_i^t = N(q_i^t), \qquad q_i^{t+1} = \left[q_i^t + u_i^t - \left\langle u_i^t, z_i^t \right\rangle \mathbf{1} \right]_+, \tag{4}$$

where the normalization

$$N(q) = \frac{q}{\|q\|_1} \tag{5}$$

maps a nonzero $q \in \mathbb{R}^d_+$ onto the simplex and is undefined at $q = 0$. The played strategy z_i^t depends only on the direction of q_i^t, not on its scale. The Euclidean norm $\|q_i^t\|_2$ therefore acts as an inverse step size, since a larger norm moves the played strategy less from one iteration to the next. The induced step size diverges as the regret vector shrinks to zero, where the played strategy is not determined.

This learner keeps the RM$^+$ recurrence and modifies only the vector inside $[\cdot]_+$ in the update (4) in two ways, one to make it predictive and one to keep the regret vector away from zero. First, following predictive RM$^+$, it adds a prediction $\widehat{m}_i^t$ of the next utility before the positive part (Farina et al., 2021). Second, following Zhang et al. (2025), it subtracts a scalar multiple of $\mathbf{1}$, chosen so that taking the positive part preserves the Euclidean norm and the regret vector cannot shrink to zero. For $q \in \mathbb{R}^d_+ \setminus \{0\}$ and $m \in \mathbb{R}^d$, the shift $\gamma(q, m)$ is the scalar for which

$$\left\| [q + m - \gamma(q, m)\mathbf{1}]_+ \right\|_2 = \|q\|_2 . \tag{6}$$

The existence, uniqueness, and sorting-based computation of this shift are due to Zhang et al. (2025, Appendix B). Proposition 4.1 restates them in our notation, together with the positive homogeneity of γ. This homogeneity underlies the payoff-scale invariance of both dynamics, established in Proposition 4.2.

Proposition 4.1 (Zhang et al. (2025))**.** *For every $q \in \mathbb{R}^d_+ \setminus \{0\}$ and $m \in \mathbb{R}^d$, Equation (6) has a unique solution. For every $\lambda > 0$, this solution satisfies*

$$\gamma(\lambda q, \lambda m) = \lambda \gamma(q, m).$$

Moreover, $\gamma(q, 0) = 0$. The shift $\gamma(q, m)$ can also be computed in $O(d \log d)$ time.

Algorithm 1 Predictive RM$^+$ self-play

Require: payoff matrix A
initialize $q_x^1 = 0$, $q_y^1 = 0$, arbitrary $\widetilde{z}_x^1 \in \mathcal{X}$, $\widetilde{z}_y^1 \in \mathcal{Y}$, and $u_x^0 = 0$, $u_y^0 = 0$
for $t = 1, 2, \ldots$ **do**
  form the predictions $(\widehat{m}_x^t, \widehat{m}_y^t)$ ▷ Eq. (9) or (10), Section 4.2
  for $i \in \{x, y\}$ **do**
    $(z_i^t, r_i^t, m_i^t) \leftarrow \textsc{NextStrategy}(q_i^t, \widetilde{z}_i^t, \widehat{m}_i^t)$
  end for
  $(x^t, y^t) \leftarrow (z_x^t, z_y^t)$
  $u_x^t \leftarrow Ay^t$, $u_y^t \leftarrow -A^\top x^t$ ▷ actual utilities, Eq. (2)
  for $i \in \{x, y\}$ **do**
    $(q_i^{t+1}, \widetilde{z}_i^{t+1}) \leftarrow \textsc{UpdateRegret}(u_i^t, r_i^t, m_i^t, z_i^t)$
  end for
end for

procedure $\textsc{NextStrategy}(q_i^t, \widetilde{z}_i^t, \widehat{m}_i^t)$ ▷ Eq. (7)
  if $q_i^t = 0$ **then**
    $m_i^t \leftarrow 0$, $r_i^t \leftarrow 0$, $p_i^t \leftarrow 0$, $z_i^t \leftarrow \widetilde{z}_i^t$ ▷ discard the prediction
  else
    $m_i^t \leftarrow \widehat{m}_i^t$
    $r_i^t \leftarrow q_i^t + m_i^t - \gamma(q_i^t, m_i^t)\mathbf{1}$ ▷ equal-norm shift, Eq. (6)
    $p_i^t \leftarrow [r_i^t]_+$, $z_i^t \leftarrow N(p_i^t)$
  end if
  return (z_i^t, r_i^t, m_i^t)
end procedure
procedure $\textsc{UpdateRegret}(u_i^t, r_i^t, m_i^t, z_i^t)$ ▷ Eq. (8)
  $q_i^{t+1} \leftarrow [r_i^t + (u_i^t - m_i^t) - \langle u_i^t - m_i^t, z_i^t \rangle \mathbf{1}]_+$
  if $q_i^{t+1} \neq 0$ **then**
    $\widetilde{z}_i^{t+1} \leftarrow N(q_i^{t+1})$
  else
    $\widetilde{z}_i^{t+1} \leftarrow z_i^t$ ▷ carry the played strategy
  end if
  return $(q_i^{t+1}, \widetilde{z}_i^{t+1})$
end procedure

We prove Proposition 4.1 in Appendix C.

The learner carries a state across iterations, consisting of a regret vector $q_i^t \in \mathbb{R}_+^{d_i}$ and a strategy $\widetilde{z}_i^t \in \Delta_{d_i}$. The strategy $\widetilde{z}_i^t$ equals $N(q_i^t)$ whenever $q_i^t \neq 0$. Each iteration runs two subprocedures. NEXTSTRATEGY forms the played strategy z_i^t from the state and a prediction $\widehat{m}_i^t$. UPDATEREGRET then produces the next state $(q_i^{t+1}, \widetilde{z}_i^{t+1})$ from the observed utility u_i^t. Algorithm 1 runs NEXTSTRATEGY and UPDATEREGRET for both players in self-play.

NEXTSTRATEGY has two cases. For $q_i^t \neq 0$, it forms the played strategy by

$$\begin{aligned} m_i^t &= \widehat{m}_i^t, \\ r_i^t &= q_i^t + m_i^t - \gamma(q_i^t, m_i^t)\mathbf{1}, \\ p_i^t &= \left[r_i^t\right]_+, \\ z_i^t &= N(p_i^t). \end{aligned} \tag{7}$$

By Equation (6), the positive part p_i^t keeps the Euclidean norm of q_i^t. Hence, $p_i^t \neq 0$, and z_i^t is well-defined. For $q_i^t = 0$, the normalization N is undefined. The learner then discards the prediction and plays the carried strategy, with $m_i^t = r_i^t = p_i^t = 0$ and $z_i^t = \widetilde{z}_i^t$.[1]

[1]The published IREG-PRM$^+$ pseudocode returns from the zero-regret branch before assigning the pre-projection vector used by the subsequent update (Zhang et al., 2025). Our assignments for the case $q_i^t = 0$

The two cases then call UPDATEREGRET. It forms the next state $(q_i^{t+1}, \widetilde{z}_i^{t+1})$ from the observed utility u_i^t,

$$q_i^{t+1} = \left[r_i^t + (u_i^t - m_i^t) - \left\langle u_i^t - m_i^t, z_i^t \right\rangle \mathbf{1}\right]_+,$$
$$\widetilde{z}_i^{t+1} = \begin{cases} N(q_i^{t+1}), & q_i^{t+1} \neq 0, \\ z_i^t, & q_i^{t+1} = 0. \end{cases} \tag{8}$$

The update is the RM$^+$ recurrence (4) on the shifted regret r_i^t and the prediction error $u_i^t - m_i^t$, in place of q_i^t and u_i^t.

We write $(x^t, y^t) = (z_x^t, z_y^t)$ and $(\widetilde{x}^t, \widetilde{y}^t) = (\widetilde{z}_x^t, \widetilde{z}_y^t)$ for the played and carried strategy profiles. The two dynamics differ only in how they form the prediction $(\widehat{m}_x^t, \widehat{m}_y^t)$, specified in Section 4.2.

4.2 THE TWO DYNAMICS WITH EXTRAGRADIENT AND OPTIMISTIC PREDICTIONS

IREG-PRM$^+$ (extragradient prediction). IREG-PRM$^+$ is the extragradient dynamics of Zhang et al. (2025). It evaluates the payoff vectors at the carried profile $(\widetilde{x}^t, \widetilde{y}^t)$,

$$\widehat{m}_x^t = A\widetilde{y}^t, \qquad \widehat{m}_y^t = -A^\mathsf{T}\widetilde{x}^t. \tag{9}$$

This evaluation and the one at the played profile (x^t, y^t) make two gradient evaluations per iteration.

IRO-PRM$^+$ (optimistic prediction). IRO-PRM$^+$ is our variant that instead predicts with the previous observed utility, the standard optimistic prediction in predictive online learning (Rakhlin & Sridharan, 2013; Farina et al., 2021). With $u_x^0 = u_y^0 = 0$, its predictions are

$$\widehat{m}_x^t = u_x^{t-1}, \qquad \widehat{m}_y^t = u_y^{t-1}. \tag{10}$$

This prediction reuses the utilities already computed in Algorithm 1. Hence, this variant makes one new gradient evaluation per iteration after initialization.

Payoff-scale invariance. IREG-PRM$^+$ and IRO-PRM$^+$ are invariant to a positive rescaling of the payoff matrix, following the scale-invariant construction of Zhang et al. (2025). The only nontrivial ingredient is the positive homogeneity of γ from Proposition 4.1. Every other operation in NEXTSTRATEGY and UPDATEREGRET is positively homogeneous of degree one, and the normalization N is invariant to a positive rescaling of its argument.

Proposition 4.2 (Payoff-scale invariance)**.** *Fix one of the two predictions, the extragradient prediction (9) or the optimistic prediction (10). Fix also initial strategies $\widetilde{z}_x^1$ and $\widetilde{z}_y^1$, initial regret vectors $q_x^1 = q_y^1 = 0$, and, for the optimistic prediction, initial utility vectors $u_x^0 = u_y^0 = 0$. For a payoff matrix A, let (x_A^t, y_A^t) be the played profile that Algorithm 1 generates from this initialization. Then, for every $\lambda > 0$ and every $t \geq 1$,*

$$(x_{\lambda A}^t, y_{\lambda A}^t) = (x_A^t, y_A^t).$$

We prove Proposition 4.2 in Appendix D. Note that the extragradient case restates Zhang et al. (2025).

5 LAST-ITERATE CONVERGENCE GUARANTEES

In this section, we establish the last-iterate convergence guarantees for the two dynamics introduced in Section 4. Specifically, we prove that the played strategy profiles of both dynamics converge to the Nash equilibrium set, and we bound the cumulative duality gap of IREG-PRM$^+$ by $\mathcal{O}(\sqrt{T})$. Both guarantees hold for every two-player zero-sum matrix game, without any bound on the payoff matrix norm or any parameter tuning.

complete this branch. We do not claim that they reproduce the iterates of a literal execution of the unmodified branch.

5.1 IREG-PRM$^+$

We first analyze IREG-PRM$^+$, defined by the prediction (9). This is the equal-norm, parameter-free extragradient construction introduced by Zhang et al. (2025), with the query pattern named after the classical extragradient method of Korpelevich (1976). The closest prior last-iterate guarantees for RM$^+$ use smoothing and a tuned update, whereas the present theorem applies to the unsmoothed equal-norm-shift recurrence (Cai et al., 2025a). Theorem 5.1 shows that the played strategy profiles of IREG-PRM$^+$ converge to the Nash equilibrium set, and that the cumulative duality gap grows as $\mathcal{O}(\sqrt{T})$.

Theorem 5.1 (Last-iterate convergence of IREG-PRM$^+$)**.** *Let A define any two-player zero-sum matrix game. Suppose that both players use the update in Section 4.1, the predictions in Equation (9), and initial regret vectors $q_x^1 = q_y^1 = 0$. Then,*

$$\lim_{t\to\infty} \operatorname{dist}\big((x^t, y^t), Z^\star\big) = 0.$$

Moreover,

$$\sum_{t=1}^{T} \operatorname{Gap}(x^t, y^t) = \mathcal{O}(\sqrt{T}).$$

The theorem does not, however, identify a single equilibrium as the limit when this set is not a singleton, nor does it establish a linear rate for the last iterate. The second bound is a dynamic-regret guarantee in the sense of Zinkevich (2003). Here, the sum $\sum_{t=1}^{T} \operatorname{Gap}(x^t, y^t)$ is the total regret of the two players against the best responses to their opponents' current strategies. This regret is controlled in self-play, where both players follow the dynamics, rather than against an adversary.

5.2 IRO-PRM$^+$

We now turn to IRO-PRM$^+$, defined by the prediction (10). It uses the previous observed utility as its prediction, following the predictable-sequence viewpoint of Rakhlin & Sridharan (2013) and the predictive RM$^+$ construction of Farina et al. (2021). Theorem 5.2 shows that the played strategy profiles of IRO-PRM$^+$ also converge to the Nash equilibrium set.

Theorem 5.2 (Last-iterate convergence of IRO-PRM$^+$)**.** *Let A define any two-player zero-sum matrix game. Suppose that both players use the update in Section 4.1, the predictions in Equation (10), and initial regret vectors $q_x^1 = q_y^1 = 0$. Then,*

$$\lim_{t\to\infty} \operatorname{Gap}(x^t, y^t) = 0, \qquad \lim_{t\to\infty} \operatorname{dist}\big((x^t, y^t), Z^\star\big) = 0.$$

In particular, Theorem 5.2 gives a last-iterate guarantee for the single-call counterpart of IREG-PRM$^+$ without a step-size parameter. Unlike Theorem 5.1, however, it does not bound the sum $\sum_{t=1}^{T} \operatorname{Gap}(x^t, y^t)$. Such a bound would require square summability of the duality gap, which our analysis of IRO-PRM$^+$ does not establish.

5.3 Proof sketch of Theorems 5.1 and 5.2

This section sketches the proofs of Theorems 5.1 and 5.2, with the complete proofs deferred to Appendices A and B. Both proofs run the same common update and reduce the last-iterate guarantee to controlling the prediction displacement $\|z^t - \widetilde{z}^t\|_2$. They diverge in how they establish this control.

(1) Common update and the three cases. For a single player, the common update (8) collapses to

$$q^{t+1} = \big[q^t + \big(u^t - \langle u^t, z^t\rangle \mathbf{1}\big) + \big(\langle m^t, z^t\rangle - \gamma(q^t, m^t)\big)\mathbf{1}\big]_+,$$

with the centered prediction error $g^t = (u^t - m^t) - \langle u^t - m^t, z^t\rangle \mathbf{1}$. The equal-norm shift makes the shift correction $\langle m^t, z^t\rangle - \gamma(q^t, m^t)$ nonnegative and keeps the regret-vector norm $\|q^t\|_2$ from decreasing. Hence, this norm plays the role of a step size, staying positive once a player's regret is nonzero. This is the mechanism anticipated in Section 1. The analysis splits into the three cases in which both, one, or neither player attains a nonzero regret vector. An inactive player keeps a constant strategy with zero best-response residual, so only an active player needs further work.

(2) Convergence of IREG-PRM$^+$. The last iterate converges once the prediction displacements are square-summable, $\sum_t \|z^t - \widetilde{z}^t\|_2^2 < \infty$. Indeed, a Lipschitz estimate yields $\mathrm{Gap}(x^t, y^t) \le \mathcal{O}\big(\|x^t - \widetilde{x}^t\|_2 + \|y^t - \widetilde{y}^t\|_2\big)$. Square-summable displacements then make the gap square-summable, so it vanishes. A Hoffman-type error bound $\mathrm{dist}\big((x^t, y^t), Z^\star\big) \le \mathcal{O}\big(\mathrm{Gap}(x^t, y^t)\big)$ (Hoffman, 1952) carries this to the equilibrium set. It remains to prove that the displacements are square-summable. The crux, shown in Lemma A.9, is that the shift correction lower-bounds the displacement with the regret-vector-norm weight,

$$\langle m^t, z^t \rangle - \gamma(q^t, m^t) \ge \frac{\|q^t\|_2}{2d} \|z^t - \widetilde{z}^t\|_2^2 . \tag{11}$$

An RVU inequality (Syrgkanis et al., 2015) bounds each player's regret by the prediction error minus the summed shift correction,

$$\max_{z \in \Delta_d} \sum_{t=1}^{T} \langle u^t, z - z^t \rangle \le \sqrt{\sum_{t=1}^{T} \|g^t\|_2^2} - \sum_{t=1}^{T} \big(\langle m^t, z^t \rangle - \gamma(q^t, m^t)\big).$$

In self-play, the two players' regrets sum to a nonnegative quantity. Combining this inequality with that nonnegativity and the shift bound (11), we control the displacements by the prediction error,

$$\sum_{t=1}^{T} \frac{\|q^t\|_2}{2d} \|z^t - \widetilde{z}^t\|_2^2 \le \sqrt{\sum_{t=1}^{T} \|g^t\|_2^2}.$$

Bilinearity makes the prediction error an action of the payoff matrix, $u_x^t - m_x^t = A(y^t - \widetilde{y}^t)$ and $u_y^t - m_y^t = -A^\mathsf{T}(x^t - \widetilde{x}^t)$. Hence, the right-hand side is at most $\mathcal{O}(\sigma_{\max}(A))\sqrt{\sum_t \|z^t - \widetilde{z}^t\|_2^2}$, where $\sigma_{\max}(A)$ is the spectral norm of the payoff matrix A. An active player keeps $\|q^t\|_2$ bounded below, so the left-hand side is at least a positive multiple of $\sum_t \|z^t - \widetilde{z}^t\|_2^2$. Combining these bounds yields a self-bounding inequality,

$$\sum_t \|z^t - \widetilde{z}^t\|_2^2 \le \mathcal{O}(\sigma_{\max}(A))\sqrt{\sum_t \|z^t - \widetilde{z}^t\|_2^2}.$$

The update has no step size to cancel $\sigma_{\max}(A)$, yet this inequality bounds $\sum_t \|z^t - \widetilde{z}^t\|_2^2$ by $\mathcal{O}(\sigma_{\max}(A)^2)$. This bound is finite for every fixed $\sigma_{\max}(A)$, so no bound on the payoff matrix norm is needed.

(3) Convergence of IRO-PRM$^+$. IRO-PRM$^+$ predicts with the previous payoff vector. Its analysis reuses the shift bound (11) and again closes by a self-bounding inequality. It diverges from the IREG-PRM$^+$ sketch in the reduction and the prediction error. A residual-transfer step reduces convergence to the equilibrium set to the vanishing of both the prediction displacement and the prediction error, which follows once $\sum_{t=1}^{T} \|g^t\|_2^2$ is bounded uniformly in T. The prediction error is now the successive increment,

$$u_x^t - m_x^t = A(y^t - y^{t-1}), \qquad u_y^t - m_y^t = -A^\mathsf{T}(x^t - x^{t-1}),$$

so $\|g^t\|_2$ is at most $\mathcal{O}(\sigma_{\max}(A))$ times the strategy increment $z^t - z^{t-1}$, which splits into the prediction displacement $z^t - \widetilde{z}^t$ and the carried-to-previous displacement $\widetilde{z}^t - z^{t-1}$. The energy bound $\|q^{t+1}\|_2^2 \le \|q^t\|_2^2 + \|g^t\|_2^2$ and the shift bound (11) yield $\sum_t \|z^t - \widetilde{z}^t\|_2^2 \le \mathcal{O}\big(\sqrt{\sum_{t=1}^{T} \|g^t\|_2^2}\big)$ for the displacement part. The main difficulty is the carried-to-previous part, which the same energy bound and a Lipschitz estimate for the normalization map bound by only a logarithm,

$$\sum_t \|\widetilde{z}^t - z^{t-1}\|_2^2 \le \mathcal{O}\Big(\log\big(1 + \textstyle\sum_{t=1}^{T} \|g^t\|_2^2\big)\Big).$$

The two parts yield the self-bounding inequality

$$\sum_{t=1}^{T} \|g^t\|_2^2 \le \mathcal{O}\Big(\sqrt{\textstyle\sum_{t=1}^{T} \|g^t\|_2^2}\Big) + \mathcal{O}\Big(\log\big(1 + \textstyle\sum_{t=1}^{T} \|g^t\|_2^2\big)\Big).$$

Since $\log(1 + x) \le \sqrt{x}$, this sum is bounded uniformly in T. Unlike the IREG-PRM$^+$ route, this controls only the vanishing of the gap, not its square summability, so it yields no dynamic-regret bound.

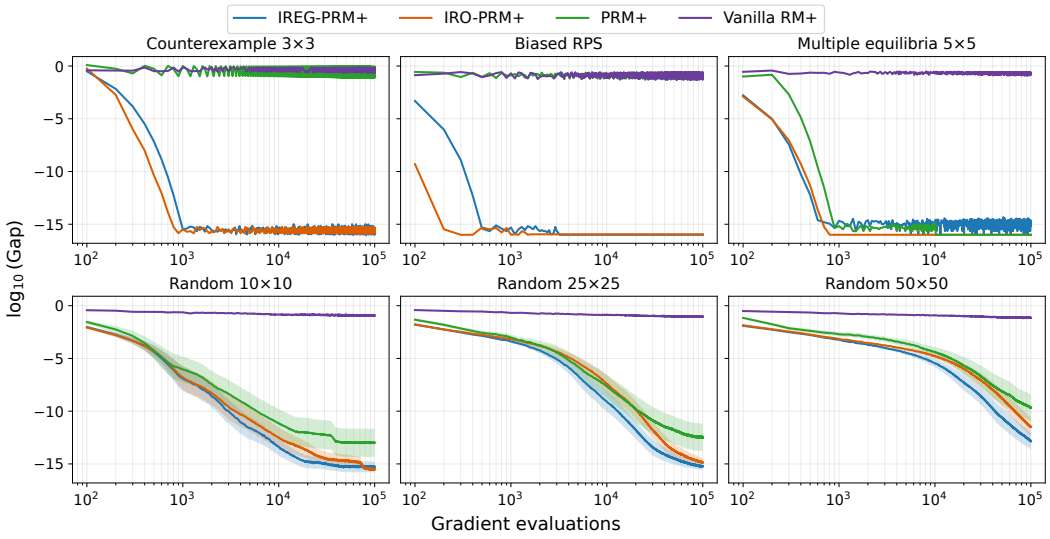

Figure 1: Duality gap of IREG-PRM$^+$, IRO-PRM$^+$, PRM$^+$, and vanilla RM$^+$ on the six benchmarks. The shaded regions show the standard errors across the random-game trials.

6 Experiments

This section empirically compares the last-iterate behavior of our two dynamics, IREG-PRM$^+$ and IRO-PRM$^+$, with that of PRM$^+$ (Farina et al., 2021) and vanilla RM$^+$ (Tammelin, 2014).

We use two classes of two-player zero-sum matrix games. The first class consists of three structured games, namely the counterexample of Farina et al. (2023) on which predictive RM$^+$ fails to converge in the last iterate, the biased Rock–Paper–Scissors (RPS) game, and the 5×5 game with multiple Nash equilibria of Wei et al. (2021), with payoff matrices listed in Appendix E. The second class consists of random payoff games with $d \in \{10, 25, 50\}$ actions per player. Each payoff entry is drawn independently and uniformly from $[-1, 1]$. All runs start from uniform initial strategies.

Each run records the duality gap of (3) at the played strategy profile. We use gradient evaluations, rather than iterations, as the comparison axis, since the four dynamics differ in evaluations per iteration. For each random-game dimension, we average over 20 payoff matrices with distinct random seeds.

Figure 1 shows the logarithm of the duality gap. On every benchmark, the gap of IREG-PRM$^+$ and IRO-PRM$^+$ decreases toward zero, in agreement with Theorems 5.1 and 5.2. On the counterexample and the biased RPS game, PRM$^+$ and vanilla RM$^+$ instead retain a nonzero gap. On the random games, PRM$^+$ converges more slowly, and vanilla RM$^+$ retains a substantially larger gap. These finite-horizon experiments do not, however, establish asymptotic convergence or a linear rate. Appendix F reports the analogous comparison under alternating updates.

7 Conclusion

Motivated by the role of RM$^+$ within CFR, we asked whether its parameter-free, payoff-scale-invariant self-play dynamics converge in the last iterate, a question left open by existing average- and best-iterate analyses. We prove last-iterate convergence to the equilibrium set for two predictive RM$^+$ dynamics, IREG-PRM$^+$ and IRO-PRM$^+$. To our knowledge, these are the first such guarantees for a parameter-free, payoff-scale-invariant RM$^+$-type algorithm. For IREG-PRM$^+$, the analysis also yields an $\mathcal{O}(\sqrt{T})$ self-play dynamic-regret bound on the cumulative duality gap.

The present analysis leaves two distinct questions open. First, it does not establish last-iterate convergence of the composed CFR dynamics in an extensive-form game. Second, neither analysis yields a uniform contraction and therefore neither establishes a linear rate.

AI USE STATEMENT

We used generative AI tools for two purposes. First, we used them to aid and polish the writing, improving the prose and the clarity of exposition in the statements and proofs. Second, we used them to assist in developing and checking the proofs of our theoretical results. We did not use generative AI tools for research ideation, for writing the experiment code, or to draft sections of the paper. The authors verified all definitions, theorems, and proofs, and take full responsibility for their correctness. The experiment code was written and checked by the authors. We have reviewed all AI-assisted text and proofs. We take responsibility for the final content of this work, including all text, claims, and artifacts.

REPRODUCIBILITY STATEMENT

The two dynamics are defined in full in Section 4, and the game setting and assumptions are stated in Section 3. Complete proofs of the last-iterate convergence guarantees are given in Appendices A and B. The experimental setup, including the benchmark games and the initialization, is described in Section 6, and the payoff matrices of the structured games are listed in Appendix E. Source code to reproduce all experiments is included in the supplementary material.

REFERENCES

Jacob Abernethy, Peter L. Bartlett, and Elad Hazan. Blackwell approachability and no-regret learning are equivalent. In *COLT*, pp. 27–46, 2011.

James P. Bailey and Georgios Piliouras. Multiplicative weights update in zero-sum games. In *EC*, pp. 321–338, 2018.

Anton Bakhtin, David J. Wu, Adam Lerer, Jonathan Gray, Athul Paul Jacob, Gabriele Farina, Alexander H. Miller, and Noam Brown. Mastering the game of No-Press Diplomacy via human-regularized reinforcement learning and planning. In *ICLR*, 2023.

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

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. Superhuman AI for heads-up no-limit poker: Libratus beats top professionals. *Science*, 359(6374):418–424, 2018.

Noam Brown and Tuomas Sandholm. Solving imperfect-information games via discounted regret minimization. In *AAAI*, pp. 1829–1836, 2019.

Noam Brown, Adam Lerer, Sam Gross, and Tuomas Sandholm. Deep counterfactual regret minimization. In *ICML*, pp. 793–802, 2019.

Noam Brown, Anton Bakhtin, Adam Lerer, and Qucheng Gong. Combining deep reinforcement learning and search for imperfect-information games. In *NeurIPS*, pp. 17057–17069, 2020.

Yang Cai, Argyris Oikonomou, and Weiqiang Zheng. Finite-time last-iterate convergence for learning in multi-player games. In *NeurIPS*, pp. 33904–33919, 2022.

Yang Cai, Gabriele Farina, Julien Grand-Clément, Christian Kroer, Chung-Wei Lee, Haipeng Luo, and Weiqiang Zheng. Last-iterate convergence properties of regret-matching algorithms in games. In *ICLR*, 2025a.

Yang Cai, Haipeng Luo, Chen-Yu Wei, and Weiqiang Zheng. From average-iterate to last-iterate convergence in games: A reduction and its applications. In *NeurIPS*, 2025b.

Ashok Cutkosky and Francesco Orabona. Black-box reductions for parameter-free online learning in Banach spaces. In *COLT*, pp. 1493–1529, 2018.

Constantinos Daskalakis and Ioannis Panageas. Last-iterate convergence: Zero-sum games and constrained min-max optimization. In *ITCS*, 2019.

Gabriele Farina, Christian Kroer, and Tuomas Sandholm. Optimistic regret minimization for extensive-form games via dilated distance-generating functions. In *NeurIPS*, 2019.

Gabriele Farina, Christian Kroer, and Tuomas Sandholm. Faster game solving via predictive Blackwell approachability: Connecting regret matching and mirror descent. In *AAAI*, pp. 5363–5371, 2021.

Gabriele Farina, Christian Kroer, Chung-Wei Lee, and Haipeng Luo. Clairvoyant regret minimization: Equivalence with Nemirovski's conceptual prox method and extension to general convex games. *arXiv preprint arXiv:2208.14891*, 2022.

Gabriele Farina, Julien Grand-Clément, Christian Kroer, Chung-Wei Lee, and Haipeng Luo. Regret matching$^+$: (in)stability and fast convergence in games. In *NeurIPS*, pp. 61546–61572, 2023.

Yoav Freund and Robert E. Schapire. Adaptive game playing using multiplicative weights. *Games and Economic Behavior*, 29(1-2):79–103, 1999.

Noah Golowich, Sarath Pattathil, and Constantinos Daskalakis. Tight last-iterate convergence rates for no-regret learning in multi-player games. In *NeurIPS*, pp. 20766–20778, 2020a.

Noah Golowich, Sarath Pattathil, Constantinos Daskalakis, and Asuman Ozdaglar. Last iterate is slower than averaged iterate in smooth convex-concave saddle point problems. In *COLT*, pp. 1758–1784, 2020b.

Eduard Gorbunov, Adrien Taylor, and Gauthier Gidel. Last-iterate convergence of optimistic gradient method for monotone variational inequalities. In *NeurIPS*, pp. 21858–21870, 2022.

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

Alan J. Hoffman. On approximate solutions of systems of linear inequalities. *Journal of Research of the National Bureau of Standards*, 49:263–265, 1952.

G. M. Korpelevich. The extragradient method for finding saddle points and other problems. *Ekonomika i Matematicheskie Metody*, 12(4):747–756, 1976.

Marc Lanctot, Kevin Waugh, Martin Zinkevich, and Michael Bowling. Monte Carlo sampling for regret minimization in extensive games. In *NeurIPS*, pp. 1078–1086, 2009.

Chung-Wei Lee, Christian Kroer, and Haipeng Luo. Last-iterate convergence in extensive-form games. In *NeurIPS*, pp. 14293–14305, 2021.

Panayotis Mertikopoulos, Christos Papadimitriou, and Georgios Piliouras. Cycles in adversarial regularized learning. In *SODA*, pp. 2703–2717, 2018.

Panayotis Mertikopoulos, Bruno Lecouat, Houssam Zenati, Chuan-Sheng Foo, Vijay Chandrasekhar, and Georgios Piliouras. Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile. In *ICLR*, 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.

John Nash. Non-cooperative games. *Annals of Mathematics*, 54(2):286–295, 1951.

Francesco Orabona and Dávid Pál. Coin betting and parameter-free online learning. In *NeurIPS*, pp. 577–585, 2016.

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

Alexander Rakhlin and Karthik Sridharan. Online learning with predictable sequences. In *COLT*, pp. 993–1019, 2013.

Vasilis Syrgkanis, Alekh Agarwal, Haipeng Luo, and Robert E. Schapire. Fast convergence of regularized learning in games. In *NeurIPS*, pp. 2989–2997, 2015.

Oskari Tammelin. Solving large imperfect information games using CFR+. *arXiv preprint arXiv:1407.5042*, 2014.

John von Neumann. Zur Theorie der Gesellschaftsspiele. *Mathematische Annalen*, 100(1):295–320, 1928.

Bernhard von Stengel. Efficient computation of behavior strategies. *Games and Economic Behavior*, 14(2):220–246, 1996.

Chen-Yu Wei, Chung-Wei Lee, Mengxiao Zhang, and Haipeng Luo. Linear last-iterate convergence in constrained saddle-point optimization. In *ICLR*, 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.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *ICML*, 2003.

Martin Zinkevich, Michael Johanson, Michael Bowling, and Carmelo Piccione. Regret minimization in games with incomplete information. In *NeurIPS*, 2007.

A PROOFS FOR THEOREM 5.1

A.1 ADDITIONAL NOTATION

For a strategy $z \in \Delta_d$ and a utility vector $u \in \mathbb{R}^d$, define the best-response residual

$$R(z; u) := \max_{1 \le j \le d} u_j - \langle u, z \rangle .$$

For a sequence of utilities and strategies, external regret is

$$\mathrm{Reg}(T) := \max_{z \in \Delta_d} \sum_{t=1}^{T} \langle u^t, z - z^t \rangle = \max_{j} \sum_{t=1}^{T} \big(u_j^t - \langle u^t, z^t \rangle \big).$$

Write $\mathrm{Reg}_x(T)$ and $\mathrm{Reg}_y(T)$ for the two players' regrets. Summing them cancels the realized payoffs. Hence, $\mathrm{Reg}_x(T) + \mathrm{Reg}_y(T)$ equals T times the duality gap of the average strategies. This gap is nonnegative, so

$$\mathrm{Reg}_x(T) + \mathrm{Reg}_y(T) \ge 0. \tag{12}$$

For $m \in \mathbb{R}^d$, write $\mathrm{osc}(m) := \max_j m_j - \min_j m_j$. Recall the normalization N of Equation (5), undefined at $q = 0$. All distances and unqualified norms below are Euclidean.

A.2 DIRECT-FORM IREG-PRM$^+$ SELF-PLAY

For each player $i \in \{x, y\}$, round t begins with a regret vector q_i^t and a strategy $\widetilde{z}_i^t$. When $q_i^t \neq 0$, this strategy satisfies

$$\widetilde{z}_i^t = N(q_i^t). \tag{13}$$

Algorithm 2 writes an entire round in chronological order and contains no query subroutine that advances the regret vector. It is algebraically identical to the clarified two-query implementation. With zero prediction, a learner with $q_i^t \neq 0$ has unique shift $\gamma = 0$ and returns $N(q_i^t) = \widetilde{z}_i^t$, while a learner with $q_i^t = 0$ returns the strategy $\widetilde{z}_i^t$ from the previous round. The subsequent prediction block is therefore exactly the second query, and the final block performs the unique update of q_i^{t+1} and $\widetilde{z}_i^{t+1}$. In particular, at $q_i^t = 0$, the update yields $q_i^{t+1} = [g_i^t]_+$. A zero update reuses the played strategy, and a nonzero update restores (13). No one-step-lag interpretation is used.

This definition is a completion and clarification of the literal pseudocode of Zhang et al. (2025). It is not claimed to be the result of executing the unmodified published branch for $q_i^t = 0$, which returns before assigning r^t.

In the single-player lemmas below, we consider one learner of dimension d governed by Algorithm 2 and suppress the player subscript.

A.3 PROOF OF THEOREM 5.1

Proof. We first record the following dichotomy.

Lemma A.1 (Nondecreasing regret-vector norm and the first round with $q^t \neq 0$)**.** *For every round t,*

$$\|q^{t+1}\|_2 \ge \|q^t\|_2 .$$

Consequently, exactly one of the following occurs:

(i) $q^t = 0$ for every t;

(ii) the first round with nonzero regret

$$\tau := \min\{t : q^t \neq 0\}$$

is finite, and every $t \ge \tau$ satisfies $\|q^t\|_2 \ge \|q^\tau\|_2 > 0$.

Algorithm 2 Clarified IREG-PRM$^+$ self-play in direct form

initialize $q_x^1 = 0$, $q_y^1 = 0$ and arbitrary initial strategies $\widetilde{z}_x^1 = \widetilde{x}^1 \in \mathcal{X}$, $\widetilde{z}_y^1 = \widetilde{y}^1 \in \mathcal{Y}$
for $t = 1, 2, \ldots$ **do**
  read the extrapolated profile $(\widetilde{x}^t, \widetilde{y}^t) \leftarrow (\widetilde{z}_x^t, \widetilde{z}_y^t)$ ▷ zero-prediction query
  $\widehat{m}_x^t \leftarrow A\widetilde{y}^t$, $\widehat{m}_y^t \leftarrow -A^\mathsf{T}\widetilde{x}^t$ ▷ extra-gradient predictions
  for $i \in \{x, y\}$ **do**
    if $q_i^t = 0$ **then**
      $m_i^t \leftarrow 0$, $r_i^t \leftarrow 0$, $p_i^t \leftarrow 0$, $z_i^t \leftarrow \widetilde{z}_i^t$ ▷ reject the supplied prediction
    else
      $m_i^t \leftarrow \widehat{m}_i^t$
      choose the unique $\gamma_i^t \in \mathbb{R}$ satisfying $\left\|[q_i^t + m_i^t - \gamma_i^t \mathbf{1}]_+\right\|_2 = \|q_i^t\|_2$
      $r_i^t \leftarrow q_i^t + m_i^t - \gamma_i^t \mathbf{1}$
      $p_i^t \leftarrow [r_i^t]_+$, $z_i^t \leftarrow N(p_i^t)$
    end if
  end for
  $x^t \leftarrow z_x^t$, $y^t \leftarrow z_y^t$ ▷ actual played strategy profile
  $u_x^t \leftarrow Ay^t$, $u_y^t \leftarrow -A^\mathsf{T}x^t$ ▷ actual utilities
  for $i \in \{x, y\}$ **do**
    $e_i^t \leftarrow u_i^t - m_i^t$
    $g_i^t \leftarrow e_i^t - \langle e_i^t, z_i^t \rangle \mathbf{1}$
    $q_i^{t+1} \leftarrow [r_i^t + g_i^t]_+$
    if $q_i^{t+1} \neq 0$ **then**
      $\widetilde{z}_i^{t+1} \leftarrow N(q_i^{t+1})$
    else
      $\widetilde{z}_i^{t+1} \leftarrow z_i^t$ ▷ reuse the played strategy
    end if
  end for
end for

If $q^t = 0$ for every t, then at every round its played strategy is a best response to its actual utility vector,

$$R(z^t; u^t) = 0,$$

and its external regret satisfies only the inequality needed below,

$$\mathrm{Reg}(T) \leq 0.$$

By Lemma A.1, each player either keeps $q^t = 0$ at every round or attains a nonzero regret vector at a finite round. We treat the three cases in which both, exactly one, or neither player attains a nonzero regret vector.

Suppose first that both players attain a nonzero regret vector.

This case uses the following lemmas.

Lemma A.2 (RVU absorption after both regret vectors are nonzero)**.** *Suppose both players attain a nonzero regret vector. Let T_0 be no smaller than both players' first rounds with nonzero regret. Then, $\sum_{t=T_0}^T \|x^t - \widetilde{x}^t\|_2^2$ and $\sum_{t=T_0}^T \|y^t - \widetilde{y}^t\|_2^2$ are bounded uniformly in T. Consequently,*

$$\sum_{t=1}^{\infty} \left\|x^t - \widetilde{x}^t\right\|_2^2 + \sum_{t=1}^{\infty} \left\|y^t - \widetilde{y}^t\right\|_2^2 < \infty, \qquad \sum_{t=1}^{\infty} \left\|g_x^t\right\|_2^2 + \sum_{t=1}^{\infty} \left\|g_y^t\right\|_2^2 < \infty,$$

and both regret-vector norms are bounded above. No smallness condition on A is required.

Lemma A.3 (Lipschitz transfer to the actual duality gap)**.** *For all $x, x' \in \mathcal{X}$ and $y, y' \in \mathcal{Y}$,*

$$|\operatorname{Gap}(x, y) - \operatorname{Gap}(x', y')| \leq \sigma_{\max}(A) \|x - x'\|_2 + \sigma_{\max}(A) \|y - y'\|_2 .$$

For a fixed utility vector u and strategies $z, s \in \Delta_d$,

$$|R(z; u) - R(s; u)| \leq \|u\|_2 \|z - s\|_2 .$$

Suppose $q_x^t, q_y^t \neq 0$ from some round onward and their regret-vector norms are bounded above. Then, there are finite constants G_x, G_y such that, for all sufficiently large t,

$$\mathrm{Gap}(x^t, y^t) \leq G_x \left\|x^t - \widetilde{x}^t\right\|_2 + G_y \left\|y^t - \widetilde{y}^t\right\|_2 .$$

Consequently, if the two squared displacement sums are finite, then

$$\sum_{t=1}^{\infty} \mathrm{Gap}(x^t, y^t)^2 < \infty.$$

Lemma A.2 proves the two squared displacement sums, the two squared prediction-error sums, and boundedness of both regret-vector norms. Combining this with Lemma A.3, we obtain

$$\sum_{t=1}^{\infty} \mathrm{Gap}(x^t, y^t)^2 < \infty. \tag{14}$$

Suppose next that exactly one player attains a nonzero regret vector. Say Player x attains it at τ and Player y has $q_y^t = 0$ for every t, the other orientation being symmetric.

This case uses the following lemmas.

Lemma A.4 (Corrected RVU inequality with initialization boundary)**.** *Let $q^1 \in \mathbb{R}_+^d$ be arbitrary, and let the learner follow its component of Algorithm 2. Then, for every $T \geq 1$,*

$$\mathrm{Reg}(T) \leq \sqrt{\|q^1\|_2^2 + \sum_{t=1}^{T} \|g^t\|_2^2} - \min_j q_j^1 - \sum_{t=1}^{T} \frac{\|q^t\|_2}{2d} \left\|z^t - \widetilde{z}^t\right\|_2^2 .$$

In particular, when $q^1 = 0$,

$$\mathrm{Reg}(T) \leq \sqrt{\sum_{t=1}^{T} \|g^t\|_2^2} - \sum_{t=1}^{T} \frac{\|q^t\|_2}{2d} \left\|z^t - \widetilde{z}^t\right\|_2^2 .$$

Dropping the nonpositive boundary term $-\min_j q_j^1$, we obtain the arbitrary-initialization form printed in Theorem 3.2 of Zhang et al. (2025).

Lemma A.5 (RM$^+$ energy inequality)**.** *For every round of Algorithm 2,*

$$\left\|q^t\right\|_2^2 \leq \left\|q^{t+1}\right\|_2^2 \leq \left\|q^t\right\|_2^2 + \left\|g^t\right\|_2^2 .$$

Consequently,

$$\left\|q^{T+1}\right\|_2^2 \leq \left\|q^1\right\|_2^2 + \sum_{t=1}^{T} \left\|g^t\right\|_2^2 .$$

In particular, if $\sum_t \|g^t\|_2^2 < \infty$, then $\sup_t \|q^t\|_2 < \infty$.

Lemma A.6 (Prediction-to-best-response geometry)**.** *Let $q \in \mathbb{R}_+^d \setminus \{0\}$, $s = N(q)$, and let*

$$p = [q + m - \gamma \mathbf{1}]_+ , \qquad \|p\|_2 = \|q\|_2 =: \rho > 0, \qquad x = N(p).$$

Then,

$$R(s; m) \leq \left(4\sqrt{d}\,\rho + 2d\ \mathrm{osc}(m)\right) \|x - s\|_2 .$$

The estimate remains valid when a maximizing coordinate of m has $q_j = 0$, when positive weight moves to coordinates outside $\mathrm{supp}(q)$, under arbitrary clipping of positive coordinates of q, and when $d = 1$.

By Lemma A.1, $\mathrm{Reg}_y(T) \leq 0$, which with (12) yields

$$\mathrm{Reg}_x(T) \geq -\mathrm{Reg}_y(T) \geq 0. \tag{15}$$

Because $q_y^t = 0$ for every t, both queries return $\widetilde{y}^t$, so

$$y^t = \widetilde{y}^t \qquad \text{for every } t. \tag{16}$$

For $t \geq \tau$, Player x accepts the supplied prediction, and by (16), we have

$$m_x^t = A\widetilde{y}^t = Ay^t = u_x^t, \qquad g_x^t = 0 \qquad (t \geq \tau).$$

The zero-initialization RVU inequality of Lemma A.4, monotonicity of $\|q_x^t\|_2$, and (15) imply, for $T \geq \tau$,

$$0 \leq \mathrm{Reg}_x(T) \leq \sqrt{\sum_{t=1}^{\tau-1} \|g_x^t\|_2^2} - \frac{\|q_x^\tau\|_2}{2d_x} \sum_{t=\tau}^{T} \left\|x^t - \widetilde{x}^t\right\|_2^2.$$

Therefore,

$$\sum_{t=\tau}^{\infty} \left\|x^t - \widetilde{x}^t\right\|_2^2 \leq \frac{2d_x \sqrt{\sum_{t=1}^{\tau-1} \|g_x^t\|_2^2}}{\|q_x^\tau\|_2} < \infty. \tag{17}$$

Player x's prediction errors are square summable because only a finite prefix is possibly nonzero, and by Lemma A.5, $\sup_t \|q_x^t\|_2 < \infty$. Player y has $\|y^t - \widetilde{y}^t\|_2 = \left\|q_y^t\right\|_2 = 0$ for every t.

It remains to transfer (17) to the gap. By Lemma A.1, we have

$$R(y^t; -A^\mathsf{T} x^t) = 0.$$

For $t \geq \tau$, Lemma A.6, boundedness of $\|q_x^t\|_2$, and

$$\mathrm{osc}(Ay^t) \leq 2\sigma_{\max}(A)$$

yield a finite constant C_x^{br} such that

$$R(\widetilde{x}^t; Ay^t) \leq C_x^{\mathrm{br}} \left\|x^t - \widetilde{x}^t\right\|_2.$$

By the residual Lipschitz bound of Lemma A.3 and

$$\left\|Ay^t\right\|_2 \leq \sigma_{\max}(A),$$

we have

$$\begin{aligned} \mathrm{Gap}(x^t, y^t) &= R(x^t; Ay^t) + R(y^t; -A^\mathsf{T} x^t) \\ &\leq R(\widetilde{x}^t; Ay^t) + \sigma_{\max}(A) \left\|x^t - \widetilde{x}^t\right\|_2 \\ &\leq (C_x^{\mathrm{br}} + \sigma_{\max}(A)) \left\|x^t - \widetilde{x}^t\right\|_2. \end{aligned} \tag{18}$$

Putting (18) into (17), and adding the finite prefix, we obtain (14).

Finally, suppose that neither player attains a nonzero regret vector. By Lemma A.1, $R(x^t; Ay^t) = R(y^t; -A^\mathsf{T} x^t) = 0$ at every round, so $\mathrm{Gap}(x^t, y^t) = R(x^t; Ay^t) + R(y^t; -A^\mathsf{T} x^t) = 0$ and the profile is a Nash equilibrium. Both carried strategies stay constant, so the profile is constant, and

$$\left\|x^t - \widetilde{x}^t\right\|_2 = \left\|y^t - \widetilde{y}^t\right\|_2 = 0, \qquad \left\|q_x^t\right\|_2 = \left\|q_y^t\right\|_2 = 0.$$

Every successive increment is zero, and the gap sequence is identically zero, so (14) holds trivially.

The finite minimax theorem yields the common value

$$v := \max_{x \in \mathcal{X}} \min_{y \in \mathcal{Y}} x^\mathsf{T} A y = \min_{y \in \mathcal{Y}} \max_{x \in \mathcal{X}} x^\mathsf{T} A y.$$

The optimal-strategy sets are

$$\mathcal{X}^\star := \{x \in \mathcal{X} : A^\mathsf{T} x \geq v\mathbf{1}\}, \tag{19}$$

$$\mathcal{Y}^\star := \{y \in \mathcal{Y} : Ay \leq v\mathbf{1}\}, \tag{20}$$

and the Nash set $Z^\star$ of Section 3 decomposes as

$$Z^\star = \mathcal{X}^\star \times \mathcal{Y}^\star. \tag{21}$$

Indeed, if $x \in \mathcal{X}^\star$ and $y \in \mathcal{Y}^\star$, then $x^\mathsf{T} A y \geq v$ by (19) and $x^\mathsf{T} A y \leq v$ by (20). Hence, the payoff equals v, and the two strategies are mutual minimax responses. Conversely, every Nash strategy is minimax optimal, so every Nash profile lies in (21).

The conclusion uses the following error bound.

Lemma A.7 (A Hoffman-type polyhedral error bound)**.** *Let $K \subset \mathbb{R}^n$ be a nonempty compact polytope and let*

$$f(z) := \max_{1 \le k \le M} \ell_k(z),$$

where each ℓ_k is affine. Assume $f \ge 0$ on K and

$$Z := \{z \in K : f(z) = 0\} \neq \varnothing.$$

Then, there is a finite constant H such that

$$\operatorname{dist}(z, Z) \le H f(z) \qquad \text{for all } z \in K.$$

Consequently, there are finite constants H_x, H_y such that

$$\operatorname{dist}(x, \mathcal{X}^\star) \le H_x\big(v - \min_j (A^\mathsf{T} x)_j\big),$$
$$\operatorname{dist}(y, \mathcal{Y}^\star) \le H_y\big(\max_j (Ay)_j - v\big).$$

In particular,

$$\operatorname{dist}\big((x, y), Z^\star\big) \le \sqrt{H_x^2 + H_y^2}\, \operatorname{Gap}(x, y).$$

In every case, (14) holds. A square-summable nonnegative sequence tends to zero, so $\lim_{t\to\infty} \operatorname{Gap}(x^t, y^t) = 0$, and by the error bound of Lemma A.7, $\lim_{t\to\infty} \operatorname{dist}\big((x^t, y^t), Z^\star\big) = 0$. Moreover, by the Cauchy–Schwarz inequality, we have

$$\sum_{t=1}^{T} \operatorname{Gap}(x^t, y^t) \le \sqrt{T \sum_{t=1}^{T} \operatorname{Gap}(x^t, y^t)^2} \le \sqrt{T \sum_{t=1}^{\infty} \operatorname{Gap}(x^t, y^t)^2} = \sqrt{\sum_{t=1}^{\infty} \operatorname{Gap}(x^t, y^t)^2}\, \sqrt{T}.$$

□

A.4 Proof of Lemma A.1

Proof. Let $p^t = [r^t]_+$. When $q^t \neq 0$, NextStrategy imposes

$$\|p^t\|_2 = \|q^t\|_2.$$

When $q^t = 0$, Algorithm 2 yields $p^t = 0$. Hence, the same equality remains true.

By construction,

$$\langle p^t, g^t \rangle = \|p^t\|_1 \langle z^t, g^t \rangle = 0,$$

where the equality is also valid when $p^t = 0$. For real numbers a, b,

$$[a + b]_+^2 \ge [a]_+^2 + 2[a]_+ b. \tag{22}$$

To verify (22), first suppose $a \ge 0$. If $a + b \ge 0$, the difference between the two sides is b^2. If $a + b < 0$, then $b < -a$ and hence $a + 2b < -a \le 0$. The right-hand side $a(a + 2b)$ is therefore nonpositive, whereas the left-hand side is zero. If $a < 0$, then $[a]_+ = 0$, so the right-hand side is zero and the nonnegative left-hand side dominates it.

Applying (22) coordinatewise with $a = r_j^t$ and $b = g_j^t$, we obtain

$$\|q^{t+1}\|_2^2 = \left\| [r^t + g^t]_+ \right\|_2^2 \ge \|p^t\|_2^2 + 2\langle p^t, g^t \rangle = \|q^t\|_2^2.$$

This proves $\|q^{t+1}\|_2 \ge \|q^t\|_2$ and the dichotomy.

Now, suppose $q^t = 0$ for all t. The completion for $q^t = 0$ yields

$$m^t = r^t = 0, \qquad z^t = \widetilde{z}^t,$$

and

$$q^{t+1} = \big[u^t - \langle u^t, z^t \rangle \mathbf{1}\big]_+ = 0.$$

Hence,

$$u_j^t \le \langle u^t, z^t \rangle \qquad \text{for every } j.$$

Because $z^t \in \Delta_d$, its expected utility cannot exceed the largest coordinate:

$$\langle u^t, z^t \rangle \le \max_j u_j^t.$$

The two inequalities imply

$$\max_j u_j^t = \langle u^t, z^t \rangle,$$

which shows $R(z^t; u^t) = 0$. For every fixed action j and every horizon T,

$$\sum_{t=1}^{T} \big(u_j^t - \langle u^t, z^t \rangle\big) \le 0.$$

Taking the maximum over j proves $\mathrm{Reg}(T) \le 0$. □

A.5 Proof of Lemma A.2

Proof. We use the following prediction-error coupling.

Lemma A.8 (Prediction-error coupling when both regret vectors are nonzero)**.** *At a round with $q_x^t, q_y^t \neq 0$, the effective predictions equal the supplied extra-gradient predictions and*

$$\|g_x^t\|_2 \le (1 + \sqrt{d_x})\sigma_{\max}(A)\,\|y^t - \widetilde{y}^t\|_2,$$
$$\|g_y^t\|_2 \le (1 + \sqrt{d_y})\sigma_{\max}(A)\,\|x^t - \widetilde{x}^t\|_2.$$

These bounds are not asserted for a player with $q^t = 0$, because that player replaces the supplied prediction by zero.

Applying the zero-initialization RVU inequality of Lemma A.4, dropping the nonpositive dissipation terms before T_0, using

$$\|q_i^t\|_2 \ge \left\|q_i^{T_0}\right\|_2 \qquad (t \ge T_0),$$

and then applying Lemma A.8, we obtain

$$\mathrm{Reg}_x(T) \le \sqrt{\sum_{t=1}^{T_0-1} \|g_x^t\|_2^2 + (1+\sqrt{d_x})^2\sigma_{\max}(A)^2 \sum_{t=T_0}^{T} \|y^t - \widetilde{y}^t\|_2^2} - \frac{\|q_x^{T_0}\|_2}{2d_x} \sum_{t=T_0}^{T} \|x^t - \widetilde{x}^t\|_2^2, \tag{23}$$

$$\mathrm{Reg}_y(T) \le \sqrt{\sum_{t=1}^{T_0-1} \|g_y^t\|_2^2 + (1+\sqrt{d_y})^2\sigma_{\max}(A)^2 \sum_{t=T_0}^{T} \|x^t - \widetilde{x}^t\|_2^2} - \frac{\|q_y^{T_0}\|_2}{2d_y} \sum_{t=T_0}^{T} \|y^t - \widetilde{y}^t\|_2^2. \tag{24}$$

Combining (23), (24), and (12), we obtain

$$\frac{\|q_x^{T_0}\|_2}{2d_x} \sum_{t=T_0}^{T} \|x^t - \widetilde{x}^t\|_2^2 + \frac{\|q_y^{T_0}\|_2}{2d_y} \sum_{t=T_0}^{T} \|y^t - \widetilde{y}^t\|_2^2 \le \sqrt{\sum_{t=1}^{T_0-1} \|g_x^t\|_2^2 + (1+\sqrt{d_x})^2\sigma_{\max}(A)^2 \sum_{t=T_0}^{T} \|y^t - \widetilde{y}^t\|_2^2} + \sqrt{\sum_{t=1}^{T_0-1} \|g_y^t\|_2^2 + (1+\sqrt{d_y})^2\sigma_{\max}(A)^2 \sum_{t=T_0}^{T} \|x^t - \widetilde{x}^t\|_2^2}. \tag{25}$$

For nonnegative a, b,

$$(\sqrt{a}+\sqrt{b})^2 = a+b+2\sqrt{ab} \geq a+b,$$
$$\sqrt{a+b} \leq \sqrt{a}+\sqrt{b}.$$

Therefore, the right-hand side of (25) is at most

$$\sqrt{\sum_{t=1}^{T_0-1}\|g_x^t\|_2^2}+\sqrt{\sum_{t=1}^{T_0-1}\|g_y^t\|_2^2}$$
$$+(1+\sqrt{d_x})\sigma_{\max}(A)\sqrt{\sum_{t=T_0}^{T}\|y^t-\widetilde{y}^t\|_2^2}+(1+\sqrt{d_y})\sigma_{\max}(A)\sqrt{\sum_{t=T_0}^{T}\|x^t-\widetilde{x}^t\|_2^2}. \tag{26}$$

Young's inequality yields

$$ab \leq \frac{\varepsilon}{2}a^2+\frac{1}{2\varepsilon}b^2 \qquad (a,b\geq 0,\ \varepsilon>0). \tag{27}$$

Applying (27) with $\varepsilon = \left\|q_x^{T_0}\right\|_2/(2d_x)$ to the last term of (26) and with $\varepsilon = \left\|q_y^{T_0}\right\|_2/(2d_y)$ to the third term, we obtain

$$(1+\sqrt{d_y})\sigma_{\max}(A)\sqrt{\sum_{t=T_0}^{T}\|x^t-\widetilde{x}^t\|_2^2} \leq \frac{\left\|q_x^{T_0}\right\|_2}{4d_x}\sum_{t=T_0}^{T}\|x^t-\widetilde{x}^t\|_2^2+\frac{d_x(1+\sqrt{d_y})^2\sigma_{\max}(A)^2}{\left\|q_x^{T_0}\right\|_2}, \tag{28}$$

$$(1+\sqrt{d_x})\sigma_{\max}(A)\sqrt{\sum_{t=T_0}^{T}\|y^t-\widetilde{y}^t\|_2^2} \leq \frac{\left\|q_y^{T_0}\right\|_2}{4d_y}\sum_{t=T_0}^{T}\|y^t-\widetilde{y}^t\|_2^2+\frac{d_y(1+\sqrt{d_x})^2\sigma_{\max}(A)^2}{\left\|q_y^{T_0}\right\|_2}. \tag{29}$$

Substituting (26)–(29) into (25) and moving the two absorbed terms to the left, we obtain

$$\frac{\left\|q_x^{T_0}\right\|_2}{4d_x}\sum_{t=T_0}^{T}\|x^t-\widetilde{x}^t\|_2^2+\frac{\left\|q_y^{T_0}\right\|_2}{4d_y}\sum_{t=T_0}^{T}\|y^t-\widetilde{y}^t\|_2^2 \leq M,$$

where

$$M := \sqrt{\sum_{t=1}^{T_0-1}\|g_x^t\|_2^2}+\sqrt{\sum_{t=1}^{T_0-1}\|g_y^t\|_2^2}+\frac{d_x(1+\sqrt{d_y})^2\sigma_{\max}(A)^2}{\left\|q_x^{T_0}\right\|_2}+\frac{d_y(1+\sqrt{d_x})^2\sigma_{\max}(A)^2}{\left\|q_y^{T_0}\right\|_2}<\infty.$$

Thus,

$$\sum_{t=T_0}^{T}\|x^t-\widetilde{x}^t\|_2^2 \leq \frac{4d_xM}{\left\|q_x^{T_0}\right\|_2}, \qquad \sum_{t=T_0}^{T}\|y^t-\widetilde{y}^t\|_2^2 \leq \frac{4d_yM}{\left\|q_y^{T_0}\right\|_2}$$

uniformly in T, proving that the two squared displacement sums are finite. Lemma A.8 then yields finiteness of the two squared prediction-error sums after adding the finite prefix. Finally, by Lemma A.5, we have

$$\sup_t \left\|q_x^t\right\|_2<\infty, \qquad \sup_t \left\|q_y^t\right\|_2<\infty.$$

No condition such as

$$(1+\sqrt{d_x})(1+\sqrt{d_y})\sigma_{\max}(A)^2 < \frac{\left\|q_x^{T_0}\right\|_2\left\|q_y^{T_0}\right\|_2}{4d_xd_y}$$

appears. The coupling terms in (25) grow as square roots of the path lengths, whereas the dissipation terms grow linearly in those path lengths. Young's inequality absorbs the square-root terms for every finite value of $\sigma_{\max}(A)$. □

A.6 Proof of Lemma A.3

Proof. For real families (a_j) and (b_j),

$$|\max_j a_j - \max_j b_j| \le \max_j |a_j - b_j|, \qquad |\min_j a_j - \min_j b_j| \le \max_j |a_j - b_j|.$$

For example, if $j_\star$ maximizes a_j, then

$$\max_j a_j - \max_j b_j \le a_{j_\star} - b_{j_\star} \le \max_j |a_j - b_j|.$$

Reversing the roles of (a_j) and (b_j), we obtain the corresponding absolute-value inequality. Therefore,

$$|\max_j (Ay)_j - \max_j (Ay')_j| \le \|A(y-y')\|_\infty \le \sigma_{\max}(A)\,\|y-y'\|_2\,,$$
$$|\min_j (A^\mathsf{T}x)_j - \min_j (A^\mathsf{T}x')_j| \le \left\|A^\mathsf{T}(x-x')\right\|_\infty \le \sigma_{\max}(A)\,\|x-x'\|_2\,.$$

This proves the Lipschitz bound for the duality gap. The maximum term in $R(z;u)$ is independent of z, so

$$|R(z;u) - R(s;u)| = |\,\langle u, s-z\rangle\,| \le \|u\|_2\,\|z-s\|_2\,,$$

which is the residual Lipschitz bound.

At a round with $q_x^t, q_y^t \neq 0$, their effective predictions are

$$m_x^t = A\widetilde{y}^t, \qquad m_y^t = -A^\mathsf{T}\widetilde{x}^t.$$

Applying Lemma A.6 with these predictions, we obtain

$$R(\widetilde{x}^t; A\widetilde{y}^t) \le \left(4\sqrt{d_x}\,\left\|q_x^t\right\|_2 + 2d_x\,\mathrm{osc}(A\widetilde{y}^t)\right)\left\|x^t - \widetilde{x}^t\right\|_2\,, \tag{30}$$
$$R(\widetilde{y}^t; -A^\mathsf{T}\widetilde{x}^t) \le \left(4\sqrt{d_y}\,\left\|q_y^t\right\|_2 + 2d_y\,\mathrm{osc}(-A^\mathsf{T}\widetilde{x}^t)\right)\left\|y^t - \widetilde{y}^t\right\|_2\,. \tag{31}$$

For every v,

$$\mathrm{osc}(v) \le 2\,\|v\|_\infty \le 2\,\|v\|_2\,.$$

Since simplex vectors have Euclidean norm at most one,

$$\mathrm{osc}(A\widetilde{y}^t) \le 2\sigma_{\max}(A), \qquad \mathrm{osc}(-A^\mathsf{T}\widetilde{x}^t) \le 2\sigma_{\max}(A). \tag{32}$$

Since the regret-vector norms are bounded above, fix $\overline{\rho}_x, \overline{\rho}_y$ with $\|q_x^t\|_2 \le \overline{\rho}_x$ and $\left\|q_y^t\right\|_2 \le \overline{\rho}_y$. The sum of the left sides of (30) and (31) is exactly

$$\mathrm{Gap}(\widetilde{x}^t, \widetilde{y}^t).$$

Combining this with the oscillation bounds (32) and $\overline{\rho}_x, \overline{\rho}_y$, we have

$$\mathrm{Gap}(\widetilde{x}^t, \widetilde{y}^t) \le (4\sqrt{d_x}\overline{\rho}_x + 4d_x\sigma_{\max}(A))\left\|x^t - \widetilde{x}^t\right\|_2 + (4\sqrt{d_y}\overline{\rho}_y + 4d_y\sigma_{\max}(A))\left\|y^t - \widetilde{y}^t\right\|_2\,. \tag{33}$$

Applying the duality-gap Lipschitz bound between the extrapolated and actual profiles, and combining with (33), we obtain $\mathrm{Gap}(x^t, y^t) \le G_x\,\|x^t - \widetilde{x}^t\|_2 + G_y\,\|y^t - \widetilde{y}^t\|_2$ with

$$G_x := 4\sqrt{d_x}\overline{\rho}_x + 4d_x\sigma_{\max}(A) + \sigma_{\max}(A), \qquad G_y := 4\sqrt{d_y}\overline{\rho}_y + 4d_y\sigma_{\max}(A) + \sigma_{\max}(A).$$

Finally,

$$\mathrm{Gap}(x^t, y^t)^2 \le 2G_x^2\left\|x^t - \widetilde{x}^t\right\|_2^2 + 2G_y^2\left\|y^t - \widetilde{y}^t\right\|_2^2\,.$$

Summation, together with the finite prefix, proves $\sum_{t=1}^\infty \mathrm{Gap}(x^t, y^t)^2 < \infty$. □

A.7 Proof of Lemma A.4

Proof. We use the following clipping identity.

Lemma A.9 (Clipping identity and the $1/(2d)$ coefficient)**.** *Let $q \in \mathbb{R}^d_+ \setminus \{0\}$, let $m \in \mathbb{R}^d$, and choose $\gamma \in \mathbb{R}$ so that*

$$p := [q + m - \gamma \mathbf{1}]_+ , \qquad \|p\|_2 = \|q\|_2 =: \rho > 0.$$

Set

$$s := N(q), \qquad x := N(p), \qquad P := \|p\|_1 , \qquad \delta := \langle m, x \rangle - \gamma.$$

Then,

$$\delta P = \langle p - q, p \rangle = \frac{1}{2} \|p - q\|_2^2 ,$$

and

$$\delta \geq \frac{\rho}{2d} \|x - s\|_2^2 .$$

The conclusion includes $d = 1$.

Define the true cumulative regret vector before round t by

$$\ell^1 := 0, \qquad \ell^{t+1} := \ell^t + u^t - \langle u^t, z^t \rangle \mathbf{1}.$$

Thus,

$$\mathrm{Reg}(T) = \max_j \ell_j^{T+1}.$$

Define the auxiliary predicted true-regret vector

$$\widehat{r}^t := \ell^t + m^t - \langle m^t, z^t \rangle \mathbf{1}.$$

Since

$$g^t = (u^t - m^t) - \langle u^t - m^t, z^t \rangle \mathbf{1},$$

we have

$$\ell^{t+1} - \widehat{r}^t = g^t. \tag{34}$$

The update and (34) yield the coordinatewise inequality

$$\begin{aligned}(q^{t+1} - r^t) - (\ell^{t+1} - \widehat{r}^t) &= \left[r^t + g^t\right]_+ - (r^t + g^t) \\ &\geq 0. \end{aligned} \tag{35}$$

When $q^t \neq 0$,

$$\begin{aligned}(r^t - q^t) - (\widehat{r}^t - \ell^t) &= \left(m^t - \gamma(q^t, m^t)\mathbf{1}\right) - \left(m^t - \langle m^t, z^t \rangle \mathbf{1}\right) \\ &= \delta_t \mathbf{1}, \qquad \delta_t := \langle m^t, z^t \rangle - \gamma(q^t, m^t).\end{aligned}$$

When $q^t = 0$, Algorithm 2 yields

$$q^t = m^t = r^t = 0, \qquad z^t = \widetilde{z}^t,$$

so the same identity holds with $\delta_t := 0$. Lemma A.9 therefore yields, at every round,

$$(r^t - q^t) - (\widehat{r}^t - \ell^t) \geq \frac{\|q^t\|_2}{2d} \left\|z^t - \widetilde{z}^t\right\|_2^2 \mathbf{1}. \tag{36}$$

The right-hand side is zero when $q^t = 0$ because $\|q^t\|_2 = \|z^t - \widetilde{z}^t\|_2 = 0$.

The exact telescoping identity is

$$\begin{aligned}q^{T+1} - q^1 - \ell^{T+1} = \sum_{t=1}^{T} \Big[&(q^{t+1} - r^t) - (\ell^{t+1} - \widehat{r}^t) \\ &+ (r^t - q^t) - (\widehat{r}^t - \ell^t)\Big].\end{aligned} \tag{37}$$

The term q^1 is the arbitrary-initialization boundary vector. Combining (35), (36), and (37), we obtain the coordinatewise inequality

$$\ell^{T+1} + q^1 + \left(\sum_{t=1}^{T} \frac{\|q^t\|_2}{2d} \left\|z^t - \widetilde{z}^t\right\|_2^2\right) \mathbf{1} \leq q^{T+1}.$$

Taking the largest coordinate and using $q^{T+1} \geq 0$ yields

$$\begin{aligned}\mathrm{Reg}(T) &= \max_j \ell_j^{T+1} \\ &\leq \max_j (q_j^{T+1} - q_j^1) - \sum_{t=1}^{T} \frac{\|q^t\|_2}{2d} \left\|z^t - \widetilde{z}^t\right\|_2^2 \\ &\leq \left\|q^{T+1}\right\|_2 - \min_j q_j^1 - \sum_{t=1}^{T} \frac{\|q^t\|_2}{2d} \left\|z^t - \widetilde{z}^t\right\|_2^2 . \end{aligned} \tag{38}$$

Lemma A.5 proves

$$\left\|q^{T+1}\right\|_2^2 \leq \left\|q^1\right\|_2^2 + \sum_{t=1}^{T} \left\|g^t\right\|_2^2 .$$

Substitution in (38) proves the claimed RVU inequality. When $q^1 = 0$, both $\left\|q^1\right\|_2^2$ and $\min_j q_j^1$ vanish, which yields its zero-initialization form. □

A.8 Proof of Lemma A.5

Proof. The lower bound was proved in Lemma A.1. For the upper bound, the following scalar inequality holds for all $a, b \in \mathbb{R}$:

$$[a+b]_+^2 \leq [a]_+^2 + 2[a]_+ b + b^2. \tag{39}$$

If $a \geq 0$ and $a + b \geq 0$, equality holds. If $a \geq 0$ and $a + b < 0$, the difference between the right and left sides is $(a+b)^2 \geq 0$. If $a < 0$ and $a + b < 0$, the left side is zero and the right side is b^2. Finally, if $a < 0$ and $a + b \geq 0$, then $b > 0$ and

$$0 \leq a + b \leq b,$$

so $(a+b)^2 \leq b^2$.

Applying (39) coordinatewise, we obtain

$$\left\|q^{t+1}\right\|_2^2 = \left\|\left[r^t + g^t\right]_+\right\|_2^2 \leq \left\|p^t\right\|_2^2 + 2\left\langle p^t, g^t\right\rangle + \left\|g^t\right\|_2^2 .$$

As in Lemma A.1,

$$\left\|p^t\right\|_2 = \left\|q^t\right\|_2 , \qquad \left\langle p^t, g^t\right\rangle = 0.$$

This proves the upper bound $\left\|q^{t+1}\right\|_2^2 \leq \|q^t\|_2^2 + \|g^t\|_2^2$. Summing it from $t = 1$ to T proves $\left\|q^{T+1}\right\|_2^2 \leq \left\|q^1\right\|_2^2 + \sum_{t=1}^{T} \|g^t\|_2^2$. □

A.9 Proof of Lemma A.6

Proof. Let

$$M := \max_j m_j, \qquad Q := \|q\|_1 .$$

Only coordinates with $q_j > 0$ contribute to $R(s; m)$. Partition them into

$$U := \{j : q_j > 0,\ p_j > 0\}, \qquad C := \{j : q_j > 0,\ p_j = 0\}.$$

Then,

$$R(s; m) = \sum_{j \in U} s_j (M - m_j) + \sum_{j \in C} s_j (M - m_j). \tag{40}$$

For $j \in U$, clipping is inactive, so

$$m_j - \gamma = p_j - q_j.$$

Hence,

$$M - m_j = (M - \gamma) - (p_j - q_j) \leq (M - \gamma)_+ + |p_j - q_j|. \tag{41}$$

Choose an index $j_\star$ with $m_{j_\star} = M$. If $M \le \gamma$, then $(M-\gamma)_+ = 0$. If $M > \gamma$, then

$$q_{j_\star} + M - \gamma > 0$$

even when $q_{j_\star} = 0$, so clipping is inactive at $j_\star$ and

$$p_{j_\star} - q_{j_\star} = M - \gamma.$$

Thus, in all cases,

$$(M-\gamma)_+ \le \|p-q\|_2 . \tag{42}$$

This explicitly covers a maximizing coordinate outside $\mathrm{supp}(q)$ and the creation of new positive weight there.

Using (41), (42), Cauchy–Schwarz, and $Q \ge \rho$, the contribution of surviving coordinates satisfies

$$\begin{aligned}
\sum_{j\in U} s_j(M-m_j) &\le (M-\gamma)_+ \sum_{j\in U} s_j + \frac{1}{Q}\sum_{j\in U} q_j|p_j-q_j| \\
&\le \|p-q\|_2 + \frac{\|q_U\|_2 \, \|p_U - q_U\|_2}{Q} \\
&\le \|p-q\|_2 + \frac{\rho}{Q}\|p-q\|_2 \\
&\le 2\|p-q\|_2 .
\end{aligned} \tag{43}$$

For $j \in C$, one has

$$0 \le M - m_j \le \mathrm{osc}(m), \qquad p_j - q_j = -q_j.$$

Therefore,

$$\begin{aligned}
\sum_{j\in C} s_j(M-m_j) &\le \frac{\mathrm{osc}(m)}{Q}\sum_{j\in C} q_j \\
&\le \frac{\sqrt{|C|}\,\mathrm{osc}(m)}{Q}\|q_C\|_2 \\
&= \frac{\sqrt{|C|}\,\mathrm{osc}(m)}{Q}\|p_C - q_C\|_2 \\
&\le \frac{\sqrt{d}\,\mathrm{osc}(m)}{\rho}\|p-q\|_2 .
\end{aligned} \tag{44}$$

This estimate allows C to contain all positive coordinates of q except those needed to preserve the nonzero norm of p. Combining (40), (43), and (44), we obtain

$$R(s;m) \le \left(2 + \frac{\sqrt{d}\,\mathrm{osc}(m)}{\rho}\right)\|p-q\|_2 . \tag{45}$$

It remains to compare $p-q$ with $x-s$. Since

$$x = \frac{p}{\|p\|_1}, \qquad \|p\|_2 = \rho,$$

we have

$$\frac{x}{\|x\|_2} = \frac{p}{\rho}.$$

Similarly,

$$\frac{s}{\|s\|_2} = \frac{q}{\rho}.$$

For nonzero vectors a, b, one has

$$\left\| \frac{a}{\|a\|_2} - \frac{b}{\|b\|_2} \right\|_2 \le \frac{2\|a-b\|_2}{\min\{\|a\|_2, \|b\|_2\}}. \tag{46}$$

Indeed,

$$\begin{aligned}\left\| \frac{a}{\|a\|_2} - \frac{b}{\|b\|_2} \right\|_2 &\le \frac{\|a-b\|_2}{\|a\|_2} + \|b\|_2 \left| \frac{1}{\|a\|_2} - \frac{1}{\|b\|_2} \right| \\ &= \frac{\|a-b\|_2 + |\,\|b\|_2 - \|a\|_2\,|}{\|a\|_2} \\ &\le \frac{2\,\|a-b\|_2}{\|a\|_2},\end{aligned}$$

and interchanging a and b, we obtain (46).

Every point of Δ_d has Euclidean norm at least $1/\sqrt{d}$, because

$$1 = \|z\|_1 \le \sqrt{d}\,\|z\|_2 .$$

Applying (46) to $x, s \in \Delta_d$, we therefore obtain

$$\|p-q\|_2 = \rho \left\| \frac{x}{\|x\|_2} - \frac{s}{\|s\|_2} \right\|_2 \le 2\rho\sqrt{d}\,\|x-s\|_2 . \tag{47}$$

Substitution of (47) into (45) proves the claimed bound.

If $d = 1$, then

$$p = q = \rho, \qquad x = s = 1, \qquad \operatorname{osc}(m) = R(s;m) = 0,$$

so the estimate is exact. □

A.10 Proof of Lemma A.7

Proof. For each k, define the closed polytope

$$C_k := \{z \in K : \ell_k(z) \ge \ell_j(z) \text{ for every } j\}.$$

The nonempty sets among the C_k cover K, and

$$f = \ell_k \qquad \text{on } C_k.$$

The diameter $\operatorname{diam}(K)$ is finite because K is compact. Fix any $z^\star \in Z$.

Consider one nonempty C_k. If $C_k \cap Z = \varnothing$, compactness yields

$$\mu_k := \min_{z \in C_k} f(z) > 0.$$

For $z \in C_k$,

$$\operatorname{dist}(z, Z) \le \|z - z^\star\|_2 \le \operatorname{diam}(K) \le \frac{\operatorname{diam}(K)}{\mu_k} f(z).$$

Now, suppose

$$F_k := C_k \cap Z \ne \varnothing.$$

Since $f = \ell_k \ge 0$ on C_k,

$$F_k = \{z \in C_k : \ell_k(z) = 0\}$$

is a face of C_k. Let V_k be the finite vertex set of C_k, and partition it as

$$V_k^0 := \{w \in V_k : \ell_k(w) = 0\}, \qquad V_k^+ := \{w \in V_k : \ell_k(w) > 0\}.$$

Every point of C_k is a convex combination of vertices. If a convex combination lies in F_k, then

$$0 = \ell_k(z)$$

is a convex combination of nonnegative numbers $\ell_k(w)$, so every vertex with positive coefficient must lie in V_k^0. Hence,

$$F_k = \operatorname{conv}(V_k^0).$$

If $V_k^+ = \varnothing$, then $C_k \subset Z$ and

$$\operatorname{dist}(z, Z) = 0 \qquad (z \in C_k).$$

Otherwise, set

$$\delta_k := \min_{w \in V_k^+} \ell_k(w) > 0.$$

Take $z \in C_k$ and write

$$z = \sum_{w \in V_k} \lambda_w w, \qquad \lambda_w \geq 0, \qquad \sum_w \lambda_w = 1.$$

Let

$$\alpha := \sum_{w \in V_k^+} \lambda_w.$$

Then,

$$f(z) = \ell_k(z) = \sum_{w \in V_k^+} \lambda_w \ell_k(w) \geq \delta_k \alpha. \tag{48}$$

If $\alpha = 0$, then $z \in F_k$. If $0 < \alpha < 1$, define

$$z_0 := \frac{1}{1-\alpha} \sum_{w \in V_k^0} \lambda_w w \in F_k, \qquad z_+ := \frac{1}{\alpha} \sum_{w \in V_k^+} \lambda_w w \in C_k.$$

Then,

$$z = (1-\alpha) z_0 + \alpha z_+,$$

and therefore

$$\operatorname{dist}(z, Z) \leq \|z - z_0\|_2 = \alpha \|z_+ - z_0\|_2 \leq \alpha \operatorname{diam}(K) \leq \frac{\operatorname{diam}(K)}{\delta_k} f(z),$$

where the last inequality uses (48).

If $\alpha = 1$, choose any $z_0 \in F_k$. Then,

$$f(z) \geq \delta_k$$

and

$$\operatorname{dist}(z, Z) \leq \|z - z_0\|_2 \leq \operatorname{diam}(K) \leq \frac{\operatorname{diam}(K)}{\delta_k} f(z).$$

Thus, each nonempty C_k has a finite local constant. Taking the maximum over the finitely many nonempty cells proves the abstract error bound $\operatorname{dist}(z, Z) \leq H f(z)$ for all $z \in K$. This is the finite polyhedral error-bound mechanism underlying Hoffman's theorem Hoffman (1952).

Apply the abstract error bound on $K = \mathcal{X}$ to

$$f_x(x) := v - \min_j (A^\mathsf{T} x)_j = \max_j \big(v - (A^\mathsf{T} x)_j\big).$$

For every $x \in \mathcal{X}$,

$$\min_j (A^\mathsf{T} x)_j = \min_{y \in \mathcal{Y}} x^\mathsf{T} A y \leq \max_{x' \in \mathcal{X}} \min_{y \in \mathcal{Y}} (x')^\mathsf{T} A y = v,$$

so $f_x \geq 0$. Moreover, $f_x(x) = 0$ if and only if

$$\min_j (A^\mathsf{T} x)_j = v.$$

Because the displayed minimum is always at most v, this is equivalent to

$$A^\mathsf{T} x \geq v \mathbf{1},$$

and hence the zero set is exactly $\mathcal{X}^\star$. This proves $\operatorname{dist}(x, \mathcal{X}^\star) \leq H_x\big(v - \min_j (A^\mathsf{T} x)_j\big)$.

Apply the lemma on $K = \mathcal{Y}$ to

$$f_y(y) := \max_j (Ay)_j - v.$$

For every $y \in \mathcal{Y}$,

$$\max_j (Ay)_j = \max_{x \in \mathcal{X}} x^\mathsf{T} A y \geq \min_{y' \in \mathcal{Y}} \max_{x \in \mathcal{X}} x^\mathsf{T} A y' = v,$$

so $f_y \geq 0$. Its zero set consists exactly of the y satisfying

$$Ay \leq v\mathbf{1},$$

namely $\mathcal{Y}^\star$. This proves $\operatorname{dist}(y, \mathcal{Y}^\star) \leq H_y\big(\max_j (Ay)_j - v\big)$.

Finally,

$$\operatorname{Gap}(x, y) = f_x(x) + f_y(y),$$

and both terms are nonnegative. Euclidean projection onto a Cartesian product separates, so

$$\begin{aligned}\operatorname{dist}\big((x,y), Z^\star\big)^2 &= \operatorname{dist}(x, \mathcal{X}^\star)^2 + \operatorname{dist}(y, \mathcal{Y}^\star)^2 \\ &\leq H_x^2 f_x(x)^2 + H_y^2 f_y(y)^2 \\ &\leq \big(H_x^2 + H_y^2\big) \operatorname{Gap}(x,y)^2.\end{aligned}$$

Taking square roots proves the error bound $\operatorname{dist}\big((x,y), Z^\star\big) \leq \sqrt{H_x^2 + H_y^2}\, \operatorname{Gap}(x,y)$. □

A.11 Proof of Lemma A.8

Proof. We use the following bound on a centered vector.

Lemma A.10 (Euclidean norm of a centered vector)**.** *For every $e \in \mathbb{R}^d$ and $z \in \Delta_d$,*

$$\|e - \langle e, z\rangle\, \mathbf{1}\|_2 \leq (1 + \sqrt{d})\, \|e\|_2 .$$

When $q_x^t, q_y^t \neq 0$,

$$m_x^t = A\widetilde{y}^t, \qquad u_x^t = Ay^t,$$

so

$$e_x^t = A(y^t - \widetilde{y}^t).$$

Since $g_x^t = e_x^t - \langle e_x^t, z_x^t\rangle\, \mathbf{1}$, Lemma A.10 with $d = d_x$, together with

$$\big\|A(y^t - \widetilde{y}^t)\big\|_2 \leq \sigma_{\max}(A)\, \big\|y^t - \widetilde{y}^t\big\|_2 ,$$

proves the bound $\|g_x^t\|_2 \leq (1 + \sqrt{d_x})\sigma_{\max}(A)\, \|y^t - \widetilde{y}^t\|_2$. The proof of the bound $\big\|g_y^t\big\|_2 \leq (1 + \sqrt{d_y})\sigma_{\max}(A)\, \|x^t - \widetilde{x}^t\|_2$ is identical because

$$e_y^t = -A^\mathsf{T}(x^t - \widetilde{x}^t)$$

and

$$\sigma_{\max}(A^\mathsf{T}) = \sigma_{\max}(A).$$

□

A.12 Proof of Lemma A.9

Proof. For each coordinate j with $p_j > 0$, clipping is inactive and

$$p_j = q_j + m_j - \gamma, \qquad m_j - \gamma = p_j - q_j.$$

For each coordinate with $p_j = 0$, including a coordinate for which $q_j > 0$, both

$$(m_j - \gamma)p_j \quad \text{and} \quad (p_j - q_j)p_j$$

are zero. Therefore, no clipping correction is omitted when the coordinate identities are multiplied by p_j and summed:

$$\delta P = \sum_j (m_j - \gamma)p_j = \sum_j (p_j - q_j)p_j = \langle p - q, p\rangle .$$

By the equal-norm condition, we have

$$\langle p - q, p\rangle = \|p\|_2^2 - \langle p, q\rangle = \frac{1}{2}\big(\|p\|_2^2 + \|q\|_2^2 - 2\,\langle p, q\rangle\big) = \frac{1}{2}\,\|p - q\|_2^2 .$$

This proves the clipping identity and, in particular, $\delta \geq 0$.

Normalize

$$\widehat{p} := \frac{p}{\rho}, \qquad \widehat{q} := \frac{q}{\rho}, \qquad a := \|\widehat{p}\|_1 = \frac{P}{\rho}, \qquad b := \|\widehat{q}\|_1, \qquad \theta := \langle \widehat{p}, \widehat{q} \rangle .$$

The vectors $\widehat{p}, \widehat{q}$ are nonnegative and have Euclidean norm one. Hence,

$$a, b \geq 1, \qquad 0 \leq \theta \leq 1.$$

From the clipping identity,

$$\delta = \rho \frac{1-\theta}{a}. \tag{49}$$

Since $x = \widehat{p}/a$ and $s = \widehat{q}/b$,

$$\begin{aligned} \|x - s\|_2^2 &= \frac{1}{a^2} + \frac{1}{b^2} - \frac{2\theta}{ab} \\ &= \frac{(a-b)^2}{a^2 b^2} + \frac{2(1-\theta)}{ab}. \end{aligned} \tag{50}$$

Set

$$\xi := \widehat{p} - \widehat{q}, \qquad \eta := \widehat{p} + \widehat{q}.$$

Equal Euclidean norms imply

$$\langle \xi, \eta \rangle = \|\widehat{p}\|_2^2 - \|\widehat{q}\|_2^2 = 0.$$

Because $\eta \neq 0$, the vector

$$\mathbf{1} - \frac{\langle \mathbf{1}, \eta \rangle}{\|\eta\|_2^2} \eta$$

is the orthogonal projection of $\mathbf{1}$ onto the orthogonal complement of η. Since $\xi \perp \eta$, by the Cauchy–Schwarz inequality, we have

$$\begin{aligned} (a-b)^2 &= \langle \mathbf{1}, \xi \rangle^2 \\ &\leq \|\xi\|_2^2 \left(\|\mathbf{1}\|_2^2 - \frac{\langle \mathbf{1}, \eta \rangle^2}{\|\eta\|_2^2} \right) \\ &= 2(1-\theta) \left(d - \frac{(a+b)^2}{2(1+\theta)} \right). \end{aligned}$$

Here we used

$$\|\xi\|_2^2 = 2(1-\theta), \qquad \|\eta\|_2^2 = 2(1+\theta), \qquad \langle \mathbf{1}, \eta \rangle = a + b, \qquad \|\mathbf{1}\|_2^2 = d.$$

Moreover,

$$\frac{(a+b)^2}{2(1+\theta)} \geq 1,$$

because $(a+b)^2 \geq 4$ while $2(1+\theta) \leq 4$. Thus,

$$(a-b)^2 \leq 2(d-1)(1-\theta). \tag{51}$$

Substituting (51) into (50) yields

$$\begin{aligned} \|x - s\|_2^2 &\leq 2(1-\theta) \left(\frac{d-1}{a^2 b^2} + \frac{1}{ab} \right) \\ &\leq 2(1-\theta) \frac{d}{a}. \end{aligned} \tag{52}$$

For the second inequality, multiplication by $a > 0$ reduces the claim to

$$\frac{d-1}{ab^2} + \frac{1}{b} \leq d,$$

which follows from $a, b \geq 1$. Combining (52) with (49) proves the coefficient bound $\delta \geq \frac{\rho}{2d} \|x - s\|_2^2$.

If $d = 1$, then the nonnegative scalars p and q both equal ρ, so

$$x = s = 1, \qquad \delta = 0.$$

Both sides of the coefficient bound are therefore zero. This also follows from the preceding calculation because $a = b = \theta = 1$. □

A.13 Proof of Lemma A.10

Proof. By the triangle inequality and Cauchy–Schwarz,

$$\begin{aligned}\|e - \langle e, z\rangle \mathbf{1}\|_2 &\le \|e\|_2 + |\langle e, z\rangle| \, \|\mathbf{1}\|_2 \\ &\le \|e\|_2 + \|e\|_2 \|z\|_2 \sqrt{d} \\ &\le (1 + \sqrt{d}) \|e\|_2 ,\end{aligned}$$

where the last step uses $\|z\|_2 \le 1$ for $z \in \Delta_d$. □

B Proofs for Theorem 5.2

B.1 Additional notation

For a player of dimension d, set

$$\delta_t := \langle m^t, z^t \rangle - \gamma(q^t, m^t)$$

when $q^t \neq 0$, and $\delta_t := 0$ when $q^t = 0$. The common learner update (8) is $q^{t+1} = [r^t + g^t]_+$, where $g^t := (u^t - m^t) - \langle u^t - m^t, z^t \rangle \mathbf{1}$.

B.2 Proof of Theorem 5.2

Proof. Suppose first that both players attain a nonzero regret vector.

This case uses the following lemmas.

Lemma B.1 (Optimistic bootstrap when both players attain a nonzero regret vector)**.** *Suppose that both players attain a nonzero regret vector. Then,*

$$\sum_{t=1}^{\infty} \sum_{i \in \{x,y\}} \left(\|g_i^t\|_2^2 + \|z_i^t - \widetilde{z}_i^t\|_2^2 \right) < \infty,$$

both regret-vector norms are bounded, and

$$\sum_{t=2}^{\infty} \sum_{i \in \{x,y\}} \|z_i^t - z_i^{t-1}\|_2^2 < \infty.$$

Lemma B.2 (Residual transfer)**.** *Suppose a player has $q^t \neq 0$ from some round onward, its regret-vector norm and predictions are bounded, and $\lim_{t\to\infty} \|z^t - \widetilde{z}^t\|_2 = 0$. Then,*

$$\lim_{t\to\infty} R(\widetilde{z}^t; m^t) = 0.$$

If also $\lim_{t\to\infty} \|u^t - m^t\|_\infty = 0$, then

$$\lim_{t\to\infty} R(z^t; u^t) = 0.$$

Lemma B.1 implies that the prediction displacements and successive strategy increments converge to zero and that both regret-vector norms are bounded. Since $m_x^t = Ay^{t-1}$ and $m_y^t = -A^\mathsf{T} x^{t-1}$,

$$\|u_x^t - m_x^t\|_\infty \le \sigma_{\max}(A) \|y^t - y^{t-1}\|_2 ,$$

and the symmetric statement holds for Player y. Since the successive strategy increments converge to zero, $\lim_{t\to\infty} \|u_i^t - m_i^t\|_\infty = 0$ for both players. Lemma B.2 then makes both actual residuals converge to zero. Hence, $\lim_{t\to\infty} \mathrm{Gap}(x^t, y^t) = 0$.

Suppose next that Player x has $q_x^t = 0$ for every t and Player y attains a nonzero regret vector at τ.

This case uses the following lemmas.

Lemma B.3 (Nondecreasing regret-vector norm and the first round with $q^t \neq 0$)**.** *The sequence $(\|q^t\|_2)$ is nondecreasing. Hence, a learner either has $q^t = 0$ for every t or has a finite first round with nonzero regret after which its regret-vector norm is bounded below by a positive constant. If $q^t = 0$ for every t, its strategy $\widetilde{z}^t$ is constant and satisfies*

$$R(z^t; u^t) = 0$$

at every iteration.

Lemma B.4 (Collapsed update and regret-vector-norm growth)**.** *At every iteration,*

$$q^{t+1} = \left[q^t + u^t - \left\langle u^t, z^t \right\rangle \mathbf{1} + \delta_t \mathbf{1}\right]_+ .$$

When $q^t \neq 0$, if $p^t = [q^t + m^t - \gamma(q^t, m^t)\mathbf{1}]_+$, then

$$\delta_t \left\|p^t\right\|_1 = \frac{1}{2}\left\|p^t - q^t\right\|_2^2, \qquad \delta_t \geq \frac{\|q^t\|_2}{2d}\left\|z^t - \widetilde{z}^t\right\|_2^2 .$$

Moreover,

$$0 \leq \left\|q^{t+1}\right\|_2^2 - \left\|q^t\right\|_2^2 \leq \left\|g^t\right\|_2^2, \qquad \left\|q^{t+1} - p^t\right\|_2^2 \leq \left\|q^{t+1}\right\|_2^2 - \left\|q^t\right\|_2^2 .$$

Lemma B.5 (Joint bound on the shifts δ_t)**.** *Suppose that $q_x^1 = q_y^1 = 0$. Then, for every T,*

$$\sum_{t=1}^{T}(\delta_x^t + \delta_y^t) \leq \left\|q_x^{T+1}\right\|_2 + \left\|q_y^{T+1}\right\|_2 .$$

Lemma B.3 yields a constant $x^t = \bar{x}$ and $R(x^t; Ay^t) = 0$ at every iteration. The column utility $-A^\mathsf{T}\bar{x}$ is constant. For $t \geq \tau$, the column learner accepts $m_y^t = -A^\mathsf{T}\bar{x} = u_y^t$. Hence, $g_y^t = 0$. The regret-vector-norm growth bound of Lemma B.4 then yields $\left\|q_y^t\right\|_2 = \left\|q_y^\tau\right\|_2$ for every $t \geq \tau$. For every $T \geq \tau$, the shift lower bound of Lemma B.4 and Lemma B.5 yield

$$\frac{\left\|q_y^\tau\right\|_2}{2d_y}\sum_{t=\tau}^{T}\left\|y^t - \widetilde{y}^t\right\|_2^2 \leq \left\|q_y^{T+1}\right\|_2 = \left\|q_y^\tau\right\|_2 .$$

Letting $T \to \infty$ shows that $\sum_{t=\tau}^{\infty}\|y^t - \widetilde{y}^t\|_2^2 < \infty$. Thus, $\lim_{t\to\infty}\|y^t - \widetilde{y}^t\|_2 = 0$, and by Lemma B.2, $\lim_{t\to\infty} R(y^t; -A^\mathsf{T}x^t) = 0$. The case in which only Player x attains a nonzero regret vector is symmetric.

If neither player attains a nonzero regret vector, both strategies are constant and both residuals vanish by Lemma B.3. Thus, the gap is identically zero. All these cases therefore imply $\lim_{t\to\infty}\mathrm{Gap}(x^t, y^t) = 0$.

If convergence to the equilibrium set failed, compactness would yield a convergent subsequence staying a positive distance from $Z^\star$. Continuity of the gap would force the limit to have zero gap. Hence, that limit would lie in $Z^\star$, contradicting that the subsequence stays a positive distance from $Z^\star$. Therefore, $\lim_{t\to\infty}\mathrm{dist}\big((x^t, y^t), Z^\star\big) = 0$. □

B.3 Proof of Lemma B.1

Proof. Choose $T_0 \geq 2$ such that $q_x^t, q_y^t \neq 0$ for every $t \geq T_0$. Since each δ_i^t is nonnegative, the shift lower bound and the regret-vector-norm growth bound of Lemma B.4, together with Lemma B.5,

yield

$$\begin{aligned}
\sum_{t=T_0}^{T}\sum_i \left\|z_i^t-\widetilde{z}_i^t\right\|_2^2 &\le \max\{2d_x/\left\|q_x^{T_0}\right\|_2, 2d_y/\left\|q_y^{T_0}\right\|_2\}\sum_{t=T_0}^{T}\sum_i \delta_i^t\\
&\le \max\{2d_x/\left\|q_x^{T_0}\right\|_2, 2d_y/\left\|q_y^{T_0}\right\|_2\}\sum_{t=1}^{T}\sum_i \delta_i^t\\
&\le \max\{2d_x/\left\|q_x^{T_0}\right\|_2, 2d_y/\left\|q_y^{T_0}\right\|_2\}\sum_i \left\|q_i^{T+1}\right\|_2\\
&\le \max\{2d_x/\left\|q_x^{T_0}\right\|_2, 2d_y/\left\|q_y^{T_0}\right\|_2\}\sum_i \sqrt{\left\|q_i^{T_0}\right\|_2^2+\sum_{t=T_0}^{T}\|g_i^t\|_2^2}\\
&\le \max\{2d_x/\left\|q_x^{T_0}\right\|_2, 2d_y/\left\|q_y^{T_0}\right\|_2\}\sum_i \left(\left\|q_i^{T_0}\right\|_2+\sqrt{\sum_{t=T_0}^{T}\|g_i^t\|_2^2}\right)\\
&\le \max\{2d_x/\left\|q_x^{T_0}\right\|_2, 2d_y/\left\|q_y^{T_0}\right\|_2\}(\left\|q_x^{T_0}\right\|_2+\left\|q_y^{T_0}\right\|_2)\\
&\quad+\sqrt{2}\max\{2d_x/\left\|q_x^{T_0}\right\|_2, 2d_y/\left\|q_y^{T_0}\right\|_2\}\sqrt{\sum_{t=T_0}^{T}\sum_i\|g_i^t\|_2^2},
\end{aligned}\tag{53}$$

The last two bounds use $\sqrt{a+b}\le\sqrt{a}+\sqrt{b}$ and $\sqrt{r}+\sqrt{s}\le\sqrt{2(r+s)}$.

We use the following Lipschitz bound for the normalization map.

Lemma B.6 (Lipschitz bound for the normalization map)**.** *For nonzero nonnegative vectors $a,b\in\mathbb{R}^d$,*

$$\|N(a)-N(b)\|_2\le\frac{1+\sqrt{d}}{\|a\|_1}\|a-b\|_2.$$

For $t\ge T_0+1$, apply Lemma B.6 to q_i^t and p_i^{t-1} with $d=d_i$. Monotonicity and the regret-vector-norm growth bound of Lemma B.4 yield

$$\begin{aligned}
\left\|\widetilde{z}_i^t-z_i^{t-1}\right\|_2^2 &\le \frac{(1+\sqrt{d_i})^2}{\|q_i^t\|_1^2}\left\|q_i^t-p_i^{t-1}\right\|_2^2\\
&\le (1+\sqrt{d_i})^2\frac{\|q_i^t\|_2^2-\left\|q_i^{t-1}\right\|_2^2}{\|q_i^t\|_1^2}\\
&\le (1+\sqrt{d_i})^2\frac{\|q_i^t\|_2^2-\left\|q_i^{t-1}\right\|_2^2}{\|q_i^t\|_2^2}\\
&\le (1+\sqrt{d_i})^2\frac{\|q_i^t\|_2^2-\left\|q_i^{t-1}\right\|_2^2}{\left\|q_i^{t-1}\right\|_2^2}.
\end{aligned}\tag{54}$$

Also, $\left\|\widetilde{z}_i^{T_0}-z_i^{T_0-1}\right\|_2^2\le 2$ because the squared Euclidean diameter of a simplex is 2.

When $q_i^t\ne 0$, $\|u_i^t\|_2,\|m_i^t\|_2\le\sigma_{\max}(A)$. By Lemma A.10, we have

$$\left\|g_i^t\right\|_2\le 2(1+\sqrt{d_i})\sigma_{\max}(A).$$

Thus, for $t\ge T_0+1$,

$$0\le\frac{\|q_i^t\|_2^2-\left\|q_i^{t-1}\right\|_2^2}{\left\|q_i^{t-1}\right\|_2^2}\le\frac{4(1+\sqrt{d_i})^2\sigma_{\max}(A)^2}{\left\|q_i^{T_0}\right\|_2^2}.$$

For any $R > 0$ and $0 \le r \le R$, the derivative of $(1+R)\log(1+r) - r$ is $(R-r)/(1+r) \ge 0$, so $r \le (1+R)\log(1+r)$. Summing Equation (54) and telescoping logarithms, we obtain

$$\sum_{t=T_0}^{T}\sum_{i}\left\|\widetilde{z}_i^t - z_i^{t-1}\right\|_2^2 \le 4 + \sum_i (1+\sqrt{d_i})^2\left(1 + \frac{4(1+\sqrt{d_i})^2\sigma_{\max}(A)^2}{\left\|q_i^{T_0}\right\|_2^2}\right) \times \log\left(1 + \frac{\sum_{t=T_0}^{T}\sum_{i'}\|g_{i'}^t\|_2^2}{\left\|q_i^{T_0}\right\|_2^2}\right). \tag{55}$$

Optimistic prediction and Lemma A.10 yield

$$\left\|g_x^t\right\|_2 \le \sigma_{\max}(A)\max\{1+\sqrt{d_x}, 1+\sqrt{d_y}\}\left\|y^t - y^{t-1}\right\|_2,$$
$$\left\|g_y^t\right\|_2 \le \sigma_{\max}(A)\max\{1+\sqrt{d_x}, 1+\sqrt{d_y}\}\left\|x^t - x^{t-1}\right\|_2.$$

Since

$$z_i^t - z_i^{t-1} = (z_i^t - \widetilde{z}_i^t) + (\widetilde{z}_i^t - z_i^{t-1}),$$

we have

$$\sum_{t=T_0}^{T}\sum_{i}\left\|g_i^t\right\|_2^2 \le 2\sigma_{\max}(A)^2\left(\max\{1+\sqrt{d_x}, 1+\sqrt{d_y}\}\right)^2 \times \left(\sum_{t=T_0}^{T}\sum_{i}\left\|z_i^t - \widetilde{z}_i^t\right\|_2^2 + \sum_{t=T_0}^{T}\sum_{i}\left\|\widetilde{z}_i^t - z_i^{t-1}\right\|_2^2\right). \tag{56}$$

For every $x \ge 0$, $\log(1+x) \le \sqrt{x}$. To verify this, note that the derivative of $\sqrt{x} - \log(1+x)$ has the sign of $(\sqrt{x}-1)^2$ and that the function vanishes at zero. Substituting Equations (53) and (55) into Equation (56), we obtain finite constants C_0, C_1, independent of T, such that

$$\sum_{t=T_0}^{T}\sum_{i}\left\|g_i^t\right\|_2^2 \le C_0 + C_1\sqrt{\sum_{t=T_0}^{T}\sum_{i}\|g_i^t\|_2^2}.$$

Thus,

$$\sqrt{\sum_{t=T_0}^{T}\sum_{i}\|g_i^t\|_2^2} \le \frac{C_1 + \sqrt{C_1^2 + 4C_0}}{2}$$

uniformly in T. Hence, the prediction errors are square summable. Equations (53) and (55) yield square summability of $\|z_i^t - \widetilde{z}_i^t\|_2$ and $\left\|\widetilde{z}_i^t - z_i^{t-1}\right\|_2$. The regret-vector-norm growth bound of Lemma B.4 bounds both regret-vector norms, and the displacement decomposition proves $\sum_{t=2}^{\infty}\sum_{i\in\{x,y\}}\left\|z_i^t - z_i^{t-1}\right\|_2^2 < \infty$. □

B.4 Proof of Lemma B.6

Proof. By the triangle inequality,

$$\begin{aligned}\|N(a) - N(b)\|_2 &\le \frac{\|a-b\|_2}{\|a\|_1} + \|b\|_2\left|\frac{1}{\|a\|_1} - \frac{1}{\|b\|_1}\right| \\ &\le \frac{\|a-b\|_2}{\|a\|_1} + \frac{|\,\|a\|_1 - \|b\|_1\,|}{\|a\|_1} \\ &\le \frac{1+\sqrt{d}}{\|a\|_1}\|a-b\|_2.\end{aligned}$$

Here

$$\|b\|_2 \le \|b\|_1$$

and

$$|\,\|a\|_1 - \|b\|_1\,| \le \|a-b\|_1 \le \sqrt{d}\,\|a-b\|_2.$$

□

B.5 Proof of Lemma B.2

Proof. By Lemma A.6, we have

$$R(\widetilde{z}^t; m^t) \le \left(4\sqrt{d}\,\left\|q^t\right\|_2 + 2d\,\operatorname{osc}(m^t)\right)\left\|z^t - \widetilde{z}^t\right\|_2.$$

The coefficient is bounded, which proves $\lim_{t\to\infty} R(\widetilde{z}^t; m^t) = 0$. For arbitrary $u, m \in \mathbb{R}^d$ and $z, w \in \Delta_d$,

$$\begin{aligned}|R(z;u) - R(w;m)| &\le |\max_j u_j - \max_j m_j| + |\langle u, z\rangle - \langle m, w\rangle| \\ &\le 2\left\|u - m\right\|_\infty + \left\|m\right\|_2 \left\|z - w\right\|_2.\end{aligned}$$

Apply this with $w = \widetilde{z}^t$ to prove $\lim_{t\to\infty} R(z^t; u^t) = 0$. □

B.6 Proof of Lemma B.3

Proof. Monotonicity follows from the regret-vector-norm growth bound of Lemma B.4. If $q^t = 0$, then the collapsed update of Lemma B.4 becomes

$$q^{t+1} = \left[u^t - \left\langle u^t, z^t\right\rangle \mathbf{1}\right]_+.$$

Thus, $q^{t+1} = 0$ exactly when every coordinate of u^t is at most $\langle u^t, z^t\rangle$. Since z^t is a simplex vector, this is equivalent to $\max_j u_j^t = \langle u^t, z^t\rangle$, which proves $R(z^t; u^t) = 0$. When $q^t = 0$, the learner reuses z^t whenever the update remains zero. □

B.7 Proof of Lemma B.4

Proof. When $q^t \neq 0$, substitution yields

$$\begin{aligned}r^t + g^t &= q^t + m^t - \gamma(q^t, m^t)\mathbf{1} + u^t - m^t - \left\langle u^t - m^t, z^t\right\rangle \mathbf{1} \\ &= q^t + u^t - \left\langle u^t, z^t\right\rangle \mathbf{1} + \left(\left\langle m^t, z^t\right\rangle - \gamma(q^t, m^t)\right)\mathbf{1}.\end{aligned}$$

When $q^t = 0$, the algorithm sets $m^t = r^t = p^t = 0$ and $\delta_t = 0$. Hence, the same identity holds. This proves the collapsed update $q^{t+1} = [q^t + u^t - \langle u^t, z^t\rangle \mathbf{1} + \delta_t \mathbf{1}]_+$. The shift bounds $\delta_t \left\|p^t\right\|_1 = \frac{1}{2}\left\|p^t - q^t\right\|_2^2$ and $\delta_t \ge \frac{\left\|q^t\right\|_2}{2d}\left\|z^t - \widetilde{z}^t\right\|_2^2$ follow from the clipping identity and geometric bound of Lemma A.9.

The construction yields $\left\|p^t\right\|_2 = \left\|q^t\right\|_2$ and

$$\left\langle p^t, g^t\right\rangle = \left\|p^t\right\|_1 \left\langle z^t, g^t\right\rangle = 0.$$

The scalar inequalities

$$[a+b]_+^2 \ge [a]_+^2 + 2[a]_+ b, \qquad [a+b]_+^2 \le [a]_+^2 + 2[a]_+ b + b^2$$

are proved coordinatewise in Lemmas A.1 and A.5. Applying them with $a = r_j^t$ and $b = g_j^t$ proves $0 \le \left\|q^{t+1}\right\|_2^2 - \left\|q^t\right\|_2^2 \le \left\|g^t\right\|_2^2$. If $p_j^t > 0$, then $r_j^t = p_j^t$ and

$$q_j^{t+1} - p_j^t = [p_j^t + g_j^t]_+ - p_j^t \ge g_j^t.$$

Multiplying by p_j^t and summing, including the zero coordinates, we obtain

$$\left\langle p^t, q^{t+1} - p^t\right\rangle \ge \left\langle p^t, g^t\right\rangle = 0.$$

Therefore,

$$\begin{aligned}\left\|q^{t+1}\right\|_2^2 - \left\|q^t\right\|_2^2 &= \left\|q^{t+1} - p^t\right\|_2^2 + 2\left\langle p^t, q^{t+1} - p^t\right\rangle \\ &\ge \left\|q^{t+1} - p^t\right\|_2^2.\end{aligned}$$

□

B.8 Proof of Lemma B.5

Proof. The collapsed update of Lemma B.4 and $[a]_+ \geq a$ yield

$$q^{T+1} \geq \sum_{t=1}^{T} \left(u^t - \langle u^t, z^t \rangle \mathbf{1}\right) + \left(\sum_{t=1}^{T} \delta_t\right) \mathbf{1},$$

where $q^1 = 0$ was used. Hence,

$$\mathrm{Reg}(T) := \max_j \sum_{t=1}^{T} \left(u_j^t - \langle u^t, z^t \rangle\right) \leq \left\|q^{T+1}\right\|_2 - \sum_{t=1}^{T} \delta_t. \tag{57}$$

Write $\overline{x}^T = \frac{1}{T}\sum_{t=1}^{T} x^t$ and $\overline{y}^T = \frac{1}{T}\sum_{t=1}^{T} y^t$ for the average strategies. For the zero-sum game, cancellation of the realized payoffs yields

$$\begin{aligned} \mathrm{Reg}_x(T) + \mathrm{Reg}_y(T) &= \max_x \sum_{t=1}^{T} x^\mathsf{T} A y^t - \min_y \sum_{t=1}^{T} (x^t)^\mathsf{T} A y \\ &= T\left(\max_x x^\mathsf{T} A \overline{y}^T - \min_y (\overline{x}^T)^\mathsf{T} A y\right) \geq 0. \end{aligned}$$

Adding Equation (57) for the two players proves $\sum_{t=1}^{T}(\delta_x^t + \delta_y^t) \leq \left\|q_x^{T+1}\right\|_2 + \left\|q_y^{T+1}\right\|_2$. □

C Proof of Proposition 4.1

Proof. Set $w = q + m$ and

$$F(\gamma) := \left\|[w - \gamma \mathbf{1}]_+\right\|_2.$$

The function F is continuous because every coordinate map $\gamma \mapsto [w_j - \gamma]_+$ is continuous. It is nonincreasing. More precisely, if $\gamma_1 < \gamma_2$ and $F(\gamma_2) > 0$, then some coordinate satisfies $w_j > \gamma_2$. For that coordinate,

$$[w_j - \gamma_1]_+ > [w_j - \gamma_2]_+.$$

Every other positive part at γ_1 is at least its value at γ_2. Hence, $F(\gamma_1) > F(\gamma_2)$ whenever $F(\gamma_2) > 0$. Moreover, $F(\gamma) = 0$ for every $\gamma \geq \max_j w_j$. As $\gamma \to -\infty$, we have $F(\gamma) \to \infty$. Since $\|q\|_2 > 0$, continuity yields a solution. Strict decrease on the positive range yields uniqueness.

The following algorithm is the deterministic sorting-based routine of Zhang et al. (2025), written in our notation.

Algorithm 3 Computing the equal-norm shift by sorting

Require: $q \in \mathbb{R}_+^d \setminus \{0\}$ and $m \in \mathbb{R}^d$
1: $(w_1, \ldots, w_d) \leftarrow$ coordinates of $q + m$ sorted in nonincreasing order
2: $\rho \leftarrow \|q\|_2$, $s \leftarrow 0$, $s_2 \leftarrow 0$
3: **for** $k = 1, \ldots, d$ **do**
4: $\quad s \leftarrow s + w_k$, $s_2 \leftarrow s_2 + w_k^2$
5: $\quad \gamma_k \leftarrow \left(s - \sqrt{s^2 - k(s_2 - \rho^2)}\right)/k$
6: $\quad$ **if** $k = d$ or $\gamma_k \geq w_{k+1}$ **then**
7: $\quad\quad$ **return** γ_k
8: $\quad$ **end if**
9: **end for**

At iteration k, the updated sums satisfy

$$s = \sum_{j=1}^{k} w_j, \qquad s_2 = \sum_{j=1}^{k} w_j^2, \qquad \rho = \|q\|_2 > 0.$$

We prove by induction over the iterations reached by the algorithm that

$$s^2 - k(s_2 - \rho^2) > 0, \qquad \gamma_k < w_k, \qquad \sum_{j=1}^{k}(w_j - \gamma_k)^2 = \rho^2.$$

For $k = 1$, these assertions follow from

$$s^2 - (s_2 - \rho^2) = \rho^2 > 0, \qquad \gamma_1 = w_1 - \rho < w_1.$$

Suppose that $k > 1$ is reached. The failed stopping test at $k - 1$ yields $\gamma_{k-1} < w_k$. Since $w_j \geq w_k > \gamma_{k-1}$ for $j < k$, the induction hypothesis implies

$$\begin{aligned}\sum_{j=1}^{k}(w_j - w_k)^2 &= \sum_{j=1}^{k-1}(w_j - w_k)^2 \\ &< \sum_{j=1}^{k-1}(w_j - \gamma_{k-1})^2 = \rho^2.\end{aligned}$$

Also, $s - kw_k \geq 0$ because the coordinates are sorted. Consequently,

$$\begin{aligned}s^2 - k(s_2 - \rho^2) &= (s - kw_k)^2 + k\left(\rho^2 - \sum_{j=1}^{k}(w_j - w_k)^2\right) \\ &> (s - kw_k)^2 \geq 0, \\ \gamma_k &= \frac{s - \sqrt{s^2 - k(s_2 - \rho^2)}}{k} < w_k.\end{aligned}$$

Thus, the square root is real at every reached iteration. The candidate is the smaller root of the quadratic, so

$$\sum_{j=1}^{k}(w_j - \gamma_k)^2 = s_2 - 2s\gamma_k + k\gamma_k^2 = \rho^2.$$

The algorithm stops after at most d iterations because the test $k = d$ always returns. At $k = d$, no value of w_{d+1} is needed. At a return with $k < d$, the stopping condition $\gamma_k \geq w_{k+1}$ holds. Together with the induction bound and the sorted order, this yields

$$w_j > \gamma_k \quad (1 \leq j \leq k), \qquad w_j \leq \gamma_k \quad (k < j \leq d).$$

The second range is empty when $k = d$. Since sorting preserves the Euclidean norm, the returned value satisfies

$$\begin{aligned}\left\|[q + m - \gamma_k \mathbf{1}]_+\right\|_2^2 &= \sum_{j=1}^{d}[w_j - \gamma_k]_+^2 \\ &= \sum_{j=1}^{k}(w_j - \gamma_k)^2 = \rho^2.\end{aligned}$$

Hence, uniqueness shows that the returned value is $\gamma(q, m)$. The non-strict stopping condition also covers $\gamma_k = w_{k+1}$ because the corresponding positive parts are zero.

Computing $q + m$ and ρ takes $O(d)$ arithmetic operations. Sorting takes $O(d \log d)$ comparisons. Each iteration updates s, s_2, evaluates one square root, and checks the stopping condition in constant time under the usual unit-cost arithmetic convention. Thus, the scan takes $O(d)$ time. The total running time is $O(d \log d)$ for $d \geq 2$. For $d = 1$, the algorithm returns $w_1 - \rho$ in constant time.

Fix $\lambda > 0$ and write $c = \gamma(q, m)$. By the positive homogeneity of the coordinatewise positive part, we have

$$\begin{aligned}\left\|[\lambda q + \lambda m - \lambda c \mathbf{1}]_+\right\|_2 &= \lambda \left\|[q + m - c\mathbf{1}]_+\right\|_2 \\ &= \lambda \|q\|_2 = \|\lambda q\|_2 .\end{aligned}$$

Thus, λc solves Equation (6) for $(\lambda q, \lambda m)$. Uniqueness proves $\gamma(\lambda q, \lambda m) = \lambda\gamma(q, m)$. If $m = 0$, then $F(0) = \|q\|_2$. Hence, uniqueness forces $\gamma(q, 0) = 0$. □

D Proof of Proposition 4.2

Proof. We prove the claim by induction on t, writing the two executions with subscripts A and λA. At the beginning of iteration t, consider the induction hypotheses

$$q^t_{i,\lambda A} = \lambda q^t_{i,A}, \qquad \widetilde{z}^t_{i,\lambda A} = \widetilde{z}^t_{i,A} \qquad \text{for each } i \in \{x, y\}. \tag{58}$$

Under the prediction (10), consider the additional hypothesis

$$u^{t-1}_{i,\lambda A} = \lambda u^{t-1}_{i,A} \qquad \text{for each } i \in \{x, y\}. \tag{59}$$

At $t = 1$, the regret vectors are zero and the strategies $\widetilde{z}^1_i$ are identical by assumption. Under the prediction (10), the initial utilities are also zero by assumption. Hence, the hypotheses hold at $t = 1$.

The predictions scale by λ,

$$\widehat{m}^t_{i,\lambda A} = \lambda \widehat{m}^t_{i,A} \qquad \text{for each } i \in \{x, y\}. \tag{60}$$

Under the prediction (9), the carried strategies are identical by (58). Hence, scaling A by λ scales $A\widetilde{y}^t$ and $-A^\mathsf{T}\widetilde{x}^t$ by λ. Under the prediction (10), Equation (59) implies (60) directly.

Fix a player i. If $q^t_{i,A} = 0$, then Equation (58) shows that both executions take the branch for $q^t_i = 0$. Hence, both executions set $m^t_i = r^t_i = p^t_i = 0$ and play the carried strategy $\widetilde{z}^t_i$.

Suppose instead that $q^t_{i,A} \neq 0$. Then, Equation (58) shows that both executions take the branch for $q^t_i \neq 0$. The positive homogeneity of γ from Proposition 4.1 and Equation (60) yield

$$\gamma\big(q^t_{i,\lambda A}, \widehat{m}^t_{i,\lambda A}\big) = \lambda\gamma\big(q^t_{i,A}, \widehat{m}^t_{i,A}\big).$$

By Equation (7), we then have

$$m^t_{i,\lambda A} = \lambda m^t_{i,A}, \qquad r^t_{i,\lambda A} = \lambda r^t_{i,A}, \qquad p^t_{i,\lambda A} = \lambda p^t_{i,A}. \tag{61}$$

By Equation (6), $\big\|p^t_{i,A}\big\|_2 = \big\|q^t_{i,A}\big\|_2 > 0$. Hence, both normalizations are well defined. Equations (5) and (61) then yield $z^t_{i,\lambda A} = z^t_{i,A}$.

In both cases, player i plays the same strategy in the two executions. Applying this to both players, the played profiles are identical. By Equation (2), we therefore have

$$u^t_{i,\lambda A} = \lambda u^t_{i,A} \qquad \text{for each } i \in \{x, y\}. \tag{62}$$

By Equation (8), $q^{t+1}_{i,\lambda A} = \lambda q^{t+1}_{i,A}$. If the next regret vector is nonzero, then Equation (5) yields identical next carried strategies $\widetilde{z}^{t+1}_i$. If the next regret vector is zero, then both executions carry the identical played strategy. Together with Equation (62), this establishes the induction hypotheses at iteration $t + 1$. □

E Benchmark Games

In this section, we list the payoff matrix A of each structured game used in Section 6. In every game, the row player maximizes $x^\mathsf{T}Ay$ and the column player minimizes it, as in Section 3.

The counterexample of Farina et al. (2023) has the payoff matrix

$$A = \begin{bmatrix} 3 & 0 & -3 \\ 0 & 3 & -4 \\ 0 & 0 & 1 \end{bmatrix}.$$

The biased Rock–Paper–Scissors game has the payoff matrix

$$A = \begin{bmatrix} 0 & 1 & -3 \\ -1 & 0 & 1 \\ 3 & -1 & 0 \end{bmatrix}.$$

The 5×5 game with multiple Nash equilibria of Wei et al. (2021) has the payoff matrix

$$A = \begin{bmatrix} 0 & -1 & 1 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 2 & -1 \\ -1 & 1 & 0 & -1 & 2 \end{bmatrix}.$$

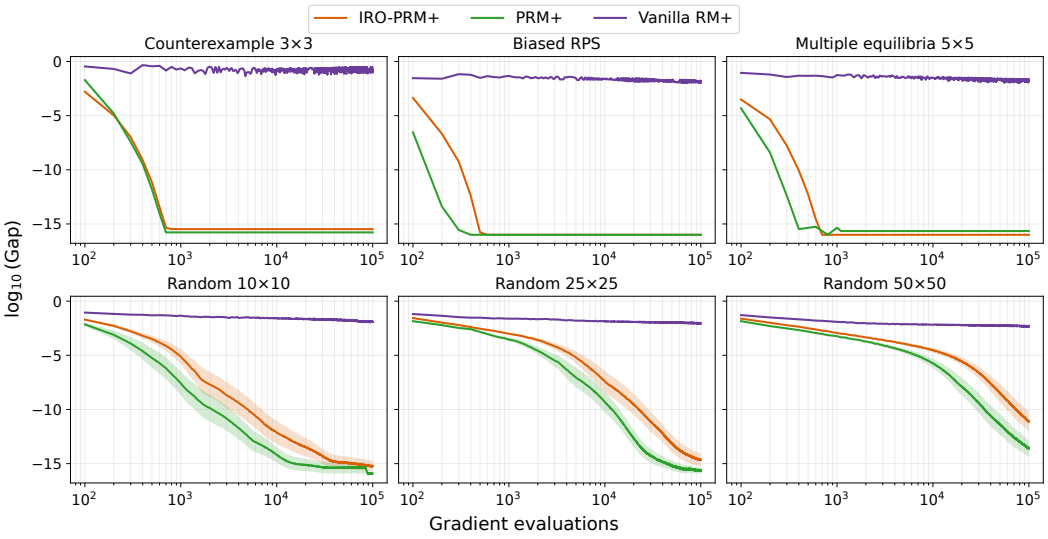

Figure 2: Duality gap of IRO-PRM$^+$, PRM$^+$, and vanilla RM$^+$ under alternating updates. IREG-PRM$^+$ is omitted because it is an extragradient method in our comparison. The curves and shaded regions have the same meaning as in Figure 1.

F ALTERNATING-UPDATE EXPERIMENTS

The experiments in Section 6 use simultaneous updates. Under alternating updates, we instead compare IRO-PRM$^+$, PRM$^+$, and vanilla RM$^+$. IREG-PRM$^+$ is omitted because it uses an extragradient query rather than an alternating update. This comparison also tests whether the empirical benefit of alternation reported by Zhang et al. (2025) extends to the PRM$^+$ baselines. The benchmark suite, the number of gradient evaluations, and the reporting protocol are those of the experiments in Section 6.

Figure 2 displays the results. IRO-PRM$^+$ and PRM$^+$ reach the numerical precision floor on the structured games. On the random games, both methods decrease the gap, and PRM$^+$ is below IRO-PRM$^+$ at 100,000 gradient evaluations for all three dimensions in these runs. Vanilla RM$^+$ remains bounded away from zero throughout the displayed horizon. These observations are consistent with the empirical advantage of alternation reported by Zhang et al. (2025). They do not establish asymptotic convergence or a convergence rate for any alternating variant.