Last-Iterate Convergence with Full and Noisy Feedback in Two-Player Zero-Sum Games

Kenshi Abe
CyberAgent, Inc.

Kaito Ariu
CyberAgent, Inc., KTH Royal Institute of Technology

Mitsuki Sakamoto   **Kentaro Toyoshima**   **Atsushi Iwasaki**
University of Electro-Communications

Abstract

This paper proposes Mutation-Driven Multiplicative Weights Update (M2WU) for learning an equilibrium in two-player zero-sum normal-form games and proves that it exhibits the last-iterate convergence property in both full and noisy feedback settings. In the former, players observe their exact gradient vectors of the utility functions. In the latter, they only observe the noisy gradient vectors. Even the celebrated Multiplicative Weights Update (MWU) and Optimistic MWU (OMWU) algorithms may not converge to a Nash equilibrium with noisy feedback. On the contrary, M2WU exhibits the last-iterate convergence to a stationary point near a Nash equilibrium in both feedback settings. We then prove that it converges to an exact Nash equilibrium by iteratively adapting the mutation term. We empirically confirm that M2WU outperforms MWU and OMWU in exploitability and convergence rates.

1 INTRODUCTION

This paper considers learning algorithms for finding an (approximate) equilibrium in two-player zero-sum games. Motivated by the training for Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) and multi-agent reinforcement learning (Busoniu et al., 2008), many algorithms have been developed to find a near-optimal solution to minimax problems (Blum and Monsour, 2007; Daskalakis et al., 2018) in the form of $\min_x \max_y f(x, y)$.

In this context, no-regret learning, which minimizes regret in repeated decisions, has been extensively studied (Banerjee and Peng, 2005; Zinkevich et al., 2007; Daskalakis et al., 2011). These algorithms, including the well-known Multiplicative Weights Update (MWU), exhibit the *average-iterate convergence* by minimizing the regret of each player; that is, the averaged strategies over iterations converge to the minimax solution (the Nash equilibrium). Still, it has been shown that the actual trajectory of updated strategies diverges or cycles (Mertikopoulos et al., 2018; Bailey and Piliouras, 2018). This feature is unsatisfactory because averaging may require nonnegligible amounts of memory and computation for large games, or averaging introduces additional error when the nonlinear function approximation is used, as in the case of training GANs.

This paper focuses on whether the actual sequence of updated strategies converges to an equilibrium, i.e., the *last-iterate convergence*, which is inevitably a stronger notion than the average-iterate convergence. A series of *optimistic* no-regret learning algorithms is proven to exhibit the last-iterate convergence (Daskalakis et al., 2018; Mertikopoulos et al., 2019). In particular, the Optimistic MWU (OMWU) algorithm is guaranteed to converge to a Nash equilibrium at an exponential rate (Daskalakis and Panageas, 2019; Wei et al., 2021b). However, existing guarantees require that players observe the exact gradient vectors of their utility functions at each iteration, which we call *full feedback*.

We generalize the full feedback setting to the *noisy feedback* setting, where players can only observe the gradient vectors with additive noise at each iteration; this setting is also called *semi-bandit feedback*. For this setting, the celebrated OMWU is not guaranteed to have the last-iterate convergence and may diverge or enter a limit cycle, as shown in Figure 5d. It has been so far guaranteed only in some restricted games, such as those with a strict Nash equilibrium, in the noisy feedback setting (Cohen et al., 2017; Giannou et al., 2021a).
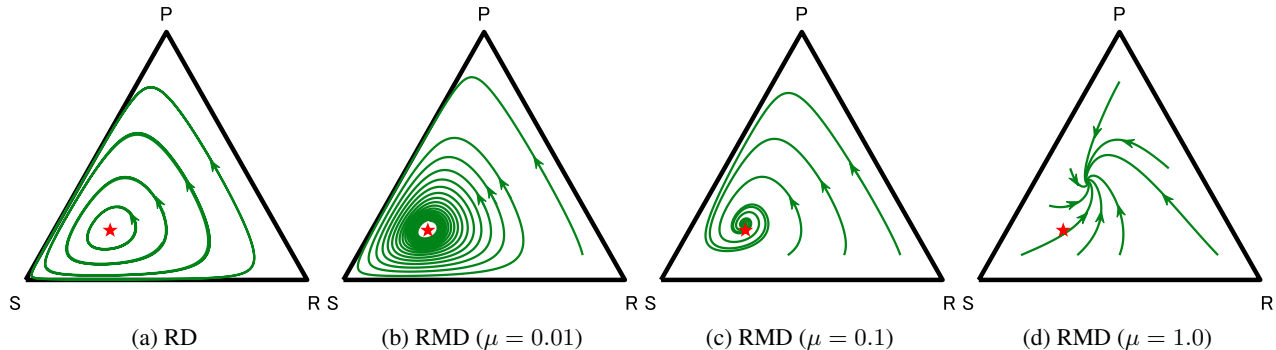
Figure 1: Learning dynamics of RD and RMD in biased Rock-Paper-Scissors (the game matrix is given by $[[0, -3, 1], [3, 0, -1], [-1, 1, 0]]$). The red star represents the Nash equilibrium point of the game.

To this end, we propose *Mutant* MWU[1] (M2WU) as the first learning algorithm that enjoys the last-iterate convergence with noisy feedback. M2WU is inspired by the fact that MWU is tantamount to replicator dynamics (RD), which is widely used in evolutionary game theory (Börgers and Sarin, 1997; Bloembergen et al., 2015). Our M2WU is designed so that it corresponds to replicator-mutator dynamics (RMD) (Hofbauer and Sigmund, 1998; Hofbauer et al., 2009; Zagorsky et al., 2013; Bauer et al., 2019), where each player may mutate his/her action. RMD has a unique stationary point that is asymptotically stable. Then, introducing mutation stabilizes the dynamics and empirically makes numerical errors in computation small (Zagorsky et al., 2013). Figure 1 demonstrates that RMD clearly converges to a near-equilibrium in a biased Rock-Paper-Scissors game, while RD oscillates around an equilibrium. Our M2WU inherits these advantages via an additional mutation term.

Starting with the full feedback case, we show that M2WU with a constant learning rate converges to a stationary point of RMD, which is known to be an approximate Nash equilibrium. The amount of approximation is specified by the mutation rates. Convergence occurs at an exponentially fast rate. Although OMWU achieves a similar convergence rate, it requires that the equilibrium in underlying games must be unique to establish convergence at that rate (Daskalakis and Panageas, 2019; Wei et al., 2021b). The convergence guarantee of M2WU holds with noisy feedback under mild conditions for noise influencing the player's observations (zero-mean martingale noise with tame second-moment tails). Specifically, M2WU converges to the stationary point almost surely. We utilize the fact that M2WU forms a continuous-time dynamics (RMD) and the existence of its Lyapunov function. In contrast, the existing convergence proof of OMWU depends on the path length of the observed gradient vectors (Mertikopoulos et al., 2019; Wei et al., 2021b), making such a guarantee

with noise difficult.

Surprisingly, in both feedback settings, we successfully establish convergence to an exact Nash equilibrium via iteratively adapting the mutation term according to the recently maintained strategy by M2WU. To the best of our knowledge, the proposed M2WU with an appropriate choice of the update interval is the first to exhibit the last-iterate convergence to an exact Nash equilibrium with noisy feedback. We further empirically demonstrate that M2WU outperforms MWU and OMWU in several games in exploitability and convergence rate, regardless of which feedback is applied.

2 RELATED LITERATURE

Last-Iterate Convergence with Full Feedback. Recently, various optimistic learning algorithms (Rakhlin and Sridharan, 2013a,b) such as optimistic Follow the Regularized Leader (FTRL) and optimistic Mirror Descent have been proposed, and their last-iterate convergence guarantees are proven with full feedback. In particular, last-iterate convergence for OMWU (Daskalakis and Panageas, 2019; Wei et al., 2021b; Lei et al., 2021; Farina et al., 2022), Optimistic Gradient Descent Ascent (OGDA) (Daskalakis et al., 2018; Mertikopoulos et al., 2019; Daskalakis and Panageas, 2018; Liang and Stokes, 2019; Golowich et al., 2020a; Wei et al., 2021b; de Montbrun and Renault, 2022), and extra-gradient algorithms (Golowich et al., 2020b; Mokhtari et al., 2020; Cai et al., 2022) have been proven in various settings such as minimax optimization and monotone games. Some studies have proposed alternative approaches that exhibit last-iterate convergence by perturbing each player's utility function via strongly convex functions (Cen et al., 2021; Perolat et al., 2021; Liu et al., 2022; Bernasconi et al., 2022) or by utilizing the asymmetric information assumption (Nguyen et al., 2021). Notably, Abe et al. (2022) analyze a continuous-time version of M2WU. However, the last-iterate convergence properties are guaranteed only with full feedback, not with noisy feedback.

---

[1]An implementation of our method is available at `https://github.com/CyberAgentAILab/m2wu`.

Last-Iterate Convergence with Noisy Feedback. A few studies have been done to prove last-iterate convergence with noisy feedback. Most existing studies discuss last-iterate convergence under the assumption that the game's equilibrium is a pure (or strict). Cohen et al. (2017) prove the convergence of an MWU-based algorithm with noise for the potential game, in which the game always has a pure Nash equilibrium, with the help of the stochastic approximation technique. There are also analyses with FTRL-based algorithms with noise (Giannou et al., 2021a,b). Such results have been obtained under other strong assumptions, such as strict (or strong) monotonicity (Bravo et al., 2018; Hsieh et al., 2019; Kannan and Shanbhag, 2019; Azizian et al., 2021), strict variational stability (Mertikopoulos et al., 2019; Mertikopoulos and Zhou, 2019), and unconstrained action set (Hsieh et al., 2022). Another approach is to use a two-time scaling, i.e., fixing the strategies of the two players to obtain sufficient samples for accurate estimates of the expected value of the utility, e.g., Wei et al. (2021a).

3 PRELIMINARIES

3.1 Two-Player Zero-Sum Normal-Form Game

A two-player normal-form game is defined as $\langle A_1, A_2, u_1, u_2 \rangle$, where A_i and $u_i : A_1 \times A_2 \to [-u_{\max}, u_{\max}]$ denote the finite action set for each player $i \in \{1, 2\}$ and a utility function for player i, respectively. In a two-player zero-sum normal-form game, there are conditions on the utility function: $u_1(a_1, a_2) = -u_2(a_1, a_2)$ for all $a_1 \in A_1$ and $a_2 \in A_2$. We denote $\Delta(A_i) = \{p \in \Delta(A_i) \mid \forall a \in A_i, \ p(a) \geq 0\}$ as a probability simplex on A_i and $\pi_i \in \Delta(A_i)$ as a *mixed strategy* for player i. Further, we denote by $\pi = (\pi_1, \pi_2)$ the *strategy profile*. For a given strategy profile π, the expected value of the utility for player i is written as follows $v_i^\pi = \mathbb{E}_{a \sim \pi}[u_i(a_1, a_2)]$. We also define the conditional expected utility with action $a_i \in A_i$ as $q_i^\pi(a_i) = \mathbb{E}_{a_{-i} \sim \pi_{-i}}[u_i(a_i, a_{-i})|a_i]$, where $-i$ represents the opponent to the player i. We denote $q_i^\pi = (q_i^\pi(a))_{a \in A_i}$ as the conditional expected utility vector.

3.2 Nash Equilibrium and Exploitability

A *Nash equilibrium* (Nash, 1951) is a widely used solution concept for games. In a Nash equilibrium, no player can improve his/her expected utility by deviating from his/her specified strategy. In two-player zero-sum normal-form games, a strategy profile $\pi^* = (\pi_1^*, \pi_2^*)$ is called a Nash equilibrium if for any $\pi_1 \in \Delta(A_1)$ and $\pi_2 \in \Delta(A_2)$,

$$v_1^{\pi_1^*, \pi_2} \geq v_1^{\pi_1^*, \pi_2^*} \geq v_1^{\pi_1, \pi_2^*}.$$

We denote the set of Nash equilibria by Π^*. An ϵ-*Nash equilibrium* (π_1, π_2) is an approximation of a Nash equi-

librium, which satisfies the following inequality:

$$\max_{\tilde{\pi}_1 \in \Delta(A_1)} v_1^{\tilde{\pi}_1, \pi_2} + \max_{\tilde{\pi}_2 \in \Delta(A_2)} v_2^{\pi_1, \tilde{\pi}_2} \leq \epsilon.$$

Furthermore, we define $\mathrm{explt}(\pi) = \max_{\tilde{\pi}_1 \in \Delta(A_1)} v_1^{\tilde{\pi}_1, \pi_2} + \max_{\tilde{\pi}_2 \in \Delta(A_2)} v_2^{\pi_1, \tilde{\pi}_2}$ as *exploitability* of the strategy profile π. Exploitability is widely used to assess how close π is to Nash equilibrium in two-player zero-sum games and always takes a non-negative value (Johanson et al., 2011, 2012; Lockhart et al., 2019; Timbers et al., 2022; Abe and Kaneko, 2021). A strategy profile π has exploitability of 0 if and only if π is a Nash equilibrium.

3.3 Problem Setting

In this study, we consider a setting where the following process is repeated: 1) At each iteration $t \in \mathbb{N}$, each player $i \in \{1, 2\}$ determines the (mixed) strategy $\pi_i^t \in \Delta(A_i)$ based on the previously observed feedback; 2) Each player i observes the new feedback $\hat{q}_i^{\pi^t}$ with respect to the gradient vector of the expected utility function $\nabla_{\pi_i^t} v_i^{\pi^t} = q_i^{\pi^t}$. This study considers two feedback settings: *full feedback* and *noisy feedback*. In the full feedback setting, each player i observes the conditional expected utility vector $q_i^{\pi^t}$ as feedback, i.e., $\hat{q}_i^{\pi^t} = q_i^{\pi^t}$. In the noisy feedback setting, at each iteration t, each player observes the noisy conditional expected utility vector

$$\hat{q}_i^{\pi^t}(a) = q_i^{\pi^t}(a) + \xi^t(a) \quad \text{for } a \in A_i,$$

where the sequence of the noise vectors $(\xi^t(a))_{a \in A_i}$ is independent over a and t. This type of noise-additive setting is standard in recent research (Cohen et al., 2017; Bravo et al., 2018; Giannou et al., 2021a,b).

Multiplicative Weights Update (MWU) is a widely used algorithm for learning a Nash equilibrium. In MWU, each player i updates her strategy π_i^t at iteration t as follows:

$$\pi_i^{t+1}(a) = \frac{\pi_i^t(a) \exp\left(\eta_t \hat{q}_i^{\pi^t}(a)\right)}{\sum_{a' \in A_i} \pi_i^t(a') \exp\left(\eta_t \hat{q}_i^{\pi^t}(a')\right)},$$

where $\eta_t > 0$ is a learning rate.

3.4 Other Notations

We denote the interior of $\Delta(A_i)$ by $\Delta^\circ(A_i) = \{p \in \Delta(A_i) \mid \forall a \in A_i, \ p(a) > 0\}$. The Kullback-Leibler divergence is defined by $\mathrm{KL}(x, y) = \sum_i x_i \ln \frac{x_i}{y_i}$. Besides, with a slight abuse of notation, we denote the sum of Kullback-Leibler divergences as $\mathrm{KL}(\pi, \pi') = \sum_{i=1}^2 \mathrm{KL}(\pi_i, \pi_i')$.

4 MUTANT MWU

This section proposes a mutant Multiplicative Weights Update (M2WU) algorithm. M2WU is a variant of MWU,

which adds a mutation (perturbation) term to the gradient vector. Specifically, M2WU updates each player's strategy by the following update rule:

$$\pi_i^{t+1}(a) = \frac{\pi_i^t(a) \exp\left(\eta_t q_i^{\mu,t}(a)\right)}{\sum_{a' \in A_i} \pi_i^t(a') \exp\left(\eta_t q_i^{\mu,t}(a')\right)}, \quad (1)$$

$$q_i^{\mu,t}(a) = \hat{q}_i^{\pi^t}(a) + \frac{\mu}{\pi_i^t(a)}\left(r_i(a) - \pi_i^t(a)\right),$$

where $\mu \in (0,1]$ is the *mutation rate*, and $r_i \in \Delta^\circ(A_i)$ is the *reference strategy*. We call $q_i^{\mu,t} = (q_i^{\mu,t}(a))_{a \in A_i}$ the *mutation gradient*.

The pseudo-code of M2WU is Algorithm 1 with $N = \infty$.

The mutation gradient $q_i^{\mu,t}$ is inspired by RMD, which is governed by the following ordinary differential equation:

$$\frac{d}{dt}\pi_i^t(a) = \pi_i^t(a)\left(q_i^{\pi^t}(a) - v_i^{\pi^t}\right) + \mu\left(r_i(a) - \pi_i^t(a)\right). \quad \text{(RMD)}$$

RMD is the continuous-time version of M2WU and has been reported to stabilize the learning dynamics (Bomze and Burger, 1995; Bauer et al., 2019). Intuitively, the mutation term $\mu\left(r_i(a) - \pi_i^t(a)\right)$ of RMD has the role of slightly shifting the trajectory of strategies from one of RD. This allows the trajectory to escape from the cyclic orbits and allows it to converge to an approximate Nash equilibrium as a stationary point of RMD. Figure 1 illustrates the trajectories of RD and RMD with $\mu \in \{0.01, 0.1, 1.0\}$ in a biased version of the Rock-Paper-Scissors game. From Figure 1a, the trajectory of RD cycles and fails to converge to a Nash equilibrium since the equilibrium is a mixed strategy with full support. On the other hand, as depicted in Figures 1b-1d, RMD's trajectory converges to a unique stationary point in this game. In fact, Abe et al. (2022) proved this convergence property of RMD for any two-player zero-sum normal-form games.

We note that M2WU with a constant learning rate can be viewed as an instantiation of the discrete-time Mutant FTRL algorithm with entropy regularization (Abe et al., 2022). The squared ℓ^2-norm regularization, as used in OGDA (Daskalakis and Panageas, 2018; Wei et al., 2021b), cannot be used with the mutation term since it allows strategies to run into the boundary of $\Delta(A_i)$ where the mutation term is undefined. It is interesting future work to find appropriate mutation terms for other regularizers as well.

5 CONVERGENCE TO AN APPROXIMATE NASH EQUILIBRIUM

This section mainly shows that the updated strategy profile π^t converges to a stationary point of (RMD). We denote the stationary point of (RMD) with fixed μ and $r = (r_i)_{i=1}^2$ by $\pi^{\mu,r}$.

Algorithm 1 M2WU for player i. The algorithm with $N = \infty$ corresponds to M2WU with a fixed reference strategy.

Require: Learning rate sequence $\{\eta_t\}_{t \geq 0}$, mutation rate μ, update frequency N, initial strategy π_i^0, initial reference strategy r_i^0

1: $k \leftarrow 0$, $\tau \leftarrow 0$
2: **for** $t = 0, 1, 2, \cdots$ **do**
3:     Observe the (noisy) gradient vector $\hat{q}_i^{\pi^t}$.
4:     **for** $a \in A_i$ **do**
5:         Compute next action probability $\pi_i^{t+1}(a)$ by
$$\pi_i^{t+1}(a) = \frac{\pi_i^t(a)\exp\left(\eta_t q_i^{\mu,t}(a)\right)}{\sum_{a' \in A_i}\pi_i^t(a')\exp\left(\eta_t q_i^{\mu,t}(a')\right)},$$
6:         where $q_i^{\mu,t}(a) = \hat{q}_i^{\pi^t}(a) + \frac{\mu}{\pi_i^t(a)}\left(r_i^k(a) - \pi_i^t(a)\right)$
7:     **end for**
8:     $\tau \leftarrow \tau + 1$
9:     **if** $\tau = N$ **then**
10:         $k \leftarrow k + 1$, $\tau \leftarrow 0$
11:         $r_i^k \leftarrow \pi_i^t$
12:     **end if**
13: **end for**

5.1 Full Feedback Setting

First, we establish the last-iterate convergence rate of M2WU with *full feedback*. Recall that in the full feedback setting, each player i observes the conditional expected utility vector $\hat{q}_i^{\pi^t} = q_i^{\pi^t}$ as feedback. The following convergence result for M2WU with a constant learning rate $\eta_t = \eta$ is obtained in the full feedback setting:

Theorem 5.1. *Let $\pi^{\mu,r} \in \prod_{i=1}^2 \Delta(A_i)$ be a stationary point of (RMD). If we use the constant learning rate sequence in M2WU, $\forall t \geq 0 : \eta_t = \eta \in (0, \min(\frac{\mu\alpha}{\mu^2\beta+\gamma}, \zeta))$, the strategy π^t updated by M2WU satisfies that for any initial strategy profile $\pi^0 \in \prod_{i=1}^2 \Delta^\circ(A_i)$ and $t \geq 0$:*

$$\mathrm{KL}(\pi^{\mu,r}, \pi^t) \leq \mathrm{KL}(\pi^{\mu,r}, \pi^0)(1 - \eta(\mu\alpha - \eta(\mu^2\beta + \gamma)))^t,$$

where α, β, γ, and ζ are constants that depend only on π_0, $\pi^{\mu,r}$, and r.

This result means that for a fixed μ and r, π^t converges to $\pi^{\mu,r}$ exponentially fast. From this theorem, π^t converges to a 2μ-Nash equilibrium because $\pi^{\mu,r}$ is a 2μ-Nash equilibrium (Bauer et al., 2019):

Corollary 5.2. *For any constant learning rate $\eta_t = \eta \in (0, \min(\frac{\mu\alpha}{\mu^2\beta+\gamma}, \zeta))$, the exploitability for M2WU is bounded as:*

$$\mathrm{explt}(\pi^t) \leq \mathrm{explt}(\pi^{\mu,r}) + 2u_{\max}\sqrt{\mathrm{KL}(\pi^{\mu,r}, \pi^0)}(1 - C)^{\frac{t}{2}}$$

$$\leq 2\mu + 2u_{\max}\sqrt{\mathrm{KL}(\pi^{\mu,r}, \pi^0)}(1 - C)^{\frac{t}{2}},$$

where $C = \eta(\mu\alpha - \eta(\mu^2\beta + \gamma))$, and α, β, γ and ζ are the same constants used in Theorem 5.1.

The proof of this corollary is shown in Appendix C. We note that from the upper bound on the learning rate $\frac{\mu\alpha}{\mu^2\beta+\gamma}$ in Theorem 5.1 and Corollary 5.2, η should decrease in proportion to the decrease of μ. We will demonstrate this fact empirically in Figure 3 in Section 7.

5.1.1 Proof Sketch of Theorem 5.1

We sketch below the proof of Theorem 5.1. Complete proofs for the theorem and associated lemmas are presented in Appendix B.

(1) Decomposing Single-Step Variation of $\mathrm{KL}(\pi^{\mu,r}, \cdot)$**.** First, we derive the following difference equation for the Kullback-Leibler divergence between $\pi^{\mu,r}$ and π^t:

$$\mathrm{KL}(\pi^{\mu,r}, \pi^{t+1}) - \mathrm{KL}(\pi^{\mu,r}, \pi^t) = \qquad (2)$$

$$\eta \underbrace{\sum_{i=1}^{2}\left(v_i^{\pi_i^t, \pi_{-i}^{\mu,r}} + \mu - \mu \sum_{a \in A_i} r_i(a)\frac{\pi_i^{\mu,r}(a)}{\pi_i^t(a)}\right)}_{(A)} + \underbrace{\mathrm{KL}(\pi^t, \pi^{t+1})}_{(B)}.$$

Equation (2) stems from the fact that for any $\pi \in \prod_{i=1}^{2}\Delta(A_i)$, $\mathrm{KL}(\pi, \pi^t) = \sum_{i=1}^{2}(\eta\langle\sum_{s=1}^{t-1} q_i^{\mu,s}, \pi_i^t - \pi_i\rangle - \psi_i(\pi_i^t) + \psi_i(\pi_i))$, where $\psi_i(p) = \sum_{a \in A_i} p(a)\ln p(a)$. For the details of the proof, see Appendix B. Hereafter, we quantify the terms (A) and (B), respectively.

(2) Equivalence Notation of (A) in Quasi-Metric Form. First, we prove that the term (A) can be rewritten by the (pseudo) metric between $\pi^{\mu,r}$ and π^t.

Lemma 5.3. *Let* $\pi^{\mu,r} \in \prod_{i=1}^{2}\Delta(A_i)$ *be a stationary point of (RMD). Then,* π^t *updated by M2WU satisfies that:*

$$\sum_{i=1}^{2}\left(v_i^{\pi_i^t, \pi_{-i}^{\mu,r}} + \mu - \mu \sum_{a \in A_i} r_i(a)\frac{\pi_i^{\mu,r}(a)}{\pi_i^t(a)}\right)$$

$$= -\mu \sum_{i=1}^{2}\sum_{a \in A_i} r_i(a)\left(\sqrt{\frac{\pi_i^t(a)}{\pi_i^{\mu,r}(a)}} - \sqrt{\frac{\pi_i^{\mu,r}(a)}{\pi_i^t(a)}}\right)^2.$$

This result can be shown by using Lemma 5.6 in Abe et al. (2022).

(3) Quasi-Metric Upper Bound on the Term (B). Next, we upper bound the Kullback-Leibler divergence between π^t and π^{t+1} by the (pseudo) metric between $\pi^{\mu,r}$ and π^t:

Lemma 5.4. *For any fixed learning rate* $\eta_t = \eta \in (0, \zeta)$*, M2WU ensures for any* $t \geq 0$*:*

$$\mathrm{KL}(\pi^t, \pi^{t+1}) \leq 8\eta^2 u_{\max}^2 \sum_{i=1}^{2}\|\pi_i^t - \pi_i^{\mu,r}\|_1^2$$

$$+ 8\eta^2\mu^2 \sum_{i=1}^{2}\sum_{a \in A_i} r_i(a)^2 \left(\frac{1}{\pi_i^{\mu,r}(a)} - \frac{1}{\pi_i^t(a)}\right)^2,$$

where ζ *is the same constant in Theorem 5.1.*

(4) Putting It All Together. By combining (2), Lemma 5.3, and Lemma 5.4, we get:

$$\mathrm{KL}(\pi^{\mu,r}, \pi^{t+1}) - \mathrm{KL}(\pi^{\mu,r}, \pi^t)$$

$$\leq -\eta\mu \sum_{i=1}^{2}\sum_{a \in A_i} r_i(a)\left(\sqrt{\frac{\pi_i^t(a)}{\pi_i^{\mu,r}(a)}} - \sqrt{\frac{\pi_i^{\mu,r}(a)}{\pi_i^t(a)}}\right)^2$$

$$+ 8\eta^2\mu^2 \sum_{i=1}^{2}\sum_{a \in A_i} r_i(a)^2 \left(\frac{1}{\pi_i^{\mu,r}(a)} - \frac{1}{\pi_i^t(a)}\right)^2$$

$$+ 8\eta^2 u_{\max}^2 \sum_{i=1}^{2}\|\pi_i^t - \pi_i^{\mu,r}\|_1^2.$$

From Pinsker's inequality (Tsybakov, 2009), we can upper bound $\sum_{i=1}^{2}\sum_{a \in A_i} r_i(a)^2 \left(\frac{1}{\pi_i^{\mu,r}(a)} - \frac{1}{\pi_i^t(a)}\right)^2$ and $\sum_{i=1}^{2}\|\pi_i^t - \pi_i^{\mu,r}\|_1^2$ by $\mathrm{KL}(\pi^{\mu,r}, \pi^t)$, respectively. Furthermore, from Jensen's inequality, we can lower bound $\sum_{i=1}^{2}\sum_{a \in A_i} r_i(a)\left(\sqrt{\frac{\pi_i^t(a)}{\pi_i^{\mu,r}(a)}} - \sqrt{\frac{\pi_i^{\mu,r}(a)}{\pi_i^t(a)}}\right)^2$ by $\mathrm{KL}(\pi^{\mu,r}, \pi^t)$. Therefore, for $\eta \in (0, \frac{\mu\alpha}{\mu^2\beta+\gamma})$, we have:

$$\mathrm{KL}(\pi^{\mu,r}, \pi^{t+1}) \leq \left(1 - \eta(\mu\alpha - \eta(\mu^2\beta + \gamma))\right)\mathrm{KL}(\pi^{\mu,r}, \pi^t).$$

Thus, by mathematical induction, the statement of the theorem is concluded. $\square$

5.2 Noisy Feedback Setting

Next, we consider a *noisy feedback* setting, where each player's observation is affected by noise. We assume the following mild condition on the noise distribution. Let $\mathcal{F}_t$ be the σ-algebra generated by the random sequence $(\pi_i^s, (\hat{q}_i^{\pi^s}(a))_{a \in A_i})_{s=1,\ldots,t}$.

Assumption 5.5. *For all player* $i \in \{1, 2\}$*, the noise process* $(\xi^t(a))_{a \in A_i}$ *satisfies the following two conditions:*

(i) *Zero-mean:* $\mathbb{E}[\xi^t(a) \mid \mathcal{F}_{t-1}] = 0, \forall a \in A_i, \forall t \geq 1$*, almost surely.*

(ii) *Moderate tails: For any* $x > 0$*,* $\mathbb{P}[|\xi^t(a)|^2 \geq x \mid \mathcal{F}_{t-1}] \leq C/x^\kappa, \forall a \in A_i, \forall t \geq 1$*, almost surely, with some constants* $C > 0$ *and* $\kappa > 2$*.*

Assumption 5.5 (i) means that the observation is unbiased: $\mathbb{E}[\hat{q}_i^{\pi^t}(a) \mid \mathcal{F}_{t-1}] = q_i^{\pi^t}(a)$, almost surely. Assumption 5.5 (ii) is a relatively weak assumption on the noise and is satisfied by a wide range of distributions, including bounded, sub-Gaussian, and sub-exponential distributions (Cohen et al., 2017). Assumption 5.5 (ii) also implies that the variance of the noise is upper bounded by some constant. The following convergence results are obtained in the noisy feedback setting.

Theorem 5.6. *Suppose that there exists a constant* $D > 0$ *such that* $\pi_i^t(a_i) > D$ *for all* $i \in \{1, 2\}$*,* $a_i \in A_i$*, and* $t \geq 0$*. Under Assumption 5.5, the strategy* π^t *updated by M2WU with the step size* $\eta_t \propto t^{-\lambda}$ *for some constant*

$\lambda \in (1/\kappa, 1]$ *converges to the stationary point $\pi^{\mu,r}$ almost surely.*

For the noise process that can take an arbitrarily large value of κ, the value of λ can be arbitrarily close to 0. This suggests that learning is possible with a nearly-constant learning rate sequence for sub-Gaussian, sub-exponential, and bounded distributions. The proof of Theorem 5.6 is based on the method of stochastic approximation (Benaïm, 1999; Borkar, 2009). Here, we only present a sketch of the proof of Theorem 5.6. A complete proof is presented in Appendix D.

5.2.1 Proof Sketch of Theorem 5.6

First, we show that the update rule of the strategy π_i^t by M2WU is an approximate Robbins-Monro algorithm (Robbins and Monro, 1951; Benaïm, 1999). For this purpose, we use Taylor's theorem to rewrite the strategy update formula as the following equation.

$$\pi_i^{t+1}(a_i) = \pi_i^t(a_i) + \eta_t(F(\pi_i^t) + U_t + \hat{\phi}_t).$$

It is relatively easy to check that $F(\pi_i^t)$ is a continuous function and U_t is a martingale difference sequence. The fact that the Hessian of the logit function is bounded by a constant indicates that $\hat{\phi}_t$ is on the order of $\eta_t \|\hat{q}_i^{\mu,t}\|_2^2$. Using Assumption 5.5, we can show that $\phi_t = \mathcal{O}(\eta_t t^p) \to 0$, almost surely, where $p \in (0, 1/\kappa)$. Thus, we can conclude that $\{\pi_i^t\}_{t \geq 1}$ is an approximate Robbins-Monro algorithm and an asymptotic pseudo-trajectory of the replicator mutator dynamics (RMD).

From Theorem 5.2 in Abe et al. (2022), there exists a strict Lyapunov function of (RMD), and the stationary point of (RMD) is unique. These conditions allow us to apply the results of Benaïm (1999) and conclude that π_i^t converges to the stationary point almost surely. □

Remark 5.7 (Comparison to optimistic algorithms). In the previous work on optimistic algorithms such as OMWU and OGDA (Mertikopoulos et al., 2019; Wei et al., 2021b), the proofs for the last-iterate convergence depend heavily on the path length of the gradient vectors $\sum_{t=1}^{T} \sum_{i=1}^{2} \|q_i^{\pi^t} - q_i^{\pi^{t-1}}\|^2$. In the full feedback setting, this term can be canceled out by the path length of strategy profiles $\sum_{t=1}^{T} \sum_{i=1}^{2} (-1/\eta_t^2) \|\pi_i^t - \pi_i^{t-1}\|^2$ with the universal constant learning rate. However, in the noisy feedback setting, the term $\sum_{t=1}^{T} \sum_{i=1}^{2} \|\hat{q}_i^{\pi^t} - \hat{q}_i^{\pi^{t-1}}\|^2$ appears instead of $\sum_{t=1}^{T} \sum_{i=1}^{2} \|q_i^{\pi^t} - q_i^{\pi^{t-1}}\|^2$, and it would grow linearly in T even if π^t is fixed. Therefore, providing the last-iterate convergence results for optimistic algorithms with noisy feedback is challenging. In contrast, the proof of last-iterate convergence with M2WU does not rely on the path length of the gradient vectors. Specifically, it exploits the existence of continuous-time dynamics (RMD) for M2WU and its Lyapunov function.

6 CONVERGENCE TO AN EXACT NASH EQUILIBRIUM

Sections 4 and 5 presented the M2WU with a fixed reference strategy profile r and its convergence results. As shown in Corollary 5.2, π^t updated by M2WU converges to the stationary point $\pi^{\mu,r}$. Therefore, if the exploitability of the stationary point $\pi^{\mu,r}$ goes to zero, the exact Nash equilibrium of the original game can be obtained. To this end, we control the exploitability of $\pi^{\mu,r}$ by adapting the reference strategy r. That is, we copy the updated strategy profile π^t to the reference profile r every N iterations. This technique is similar to the direct convergence method by Perolat et al. (2021). The pseudo-code of M2WU with this technique corresponds to Algorithm 1 with finite N.

Let us define r^k as the k-th reference strategy profile. From Theorem 5.1, π^t converges to $\pi^{\mu,r}$ when N is set to a sufficiently large value. In this case, the following reference strategy r^{k+1} is set to the stationary point π^{μ,r^k} of the (RMD) dynamics with reference strategy r^k. In the remaining part of this section, we show that the sequence of stationary points $\{\pi^{\mu,r^k}\}_{k \geq 0} = \{r^k\}_{k \geq 1}$ converges to Π^* of the original game.

Theorem 6.1. *For any start point $r^0 \in \prod_{i=1}^{2} \Delta^\circ(A_i)$, the sequence of stationary points $\{\pi^{\mu,r^k}\}_{k \geq 0} = \{r^k\}_{k \geq 1}$ converges to the set of equilibria Π^* of the original game.*

This result means that the exploitability of π^{μ,r^k} converges to 0 because $\mathrm{explt}(\pi^{\mu,r^k}) \leq \mathcal{O}\left(\sqrt{\mathrm{KL}(\pi^\star, \pi^{\mu,r^k})}\right)$, where $\pi^\star = \arg\min_{\pi \in \Pi^*} \mathrm{KL}(\pi, \pi^{\mu,r^k})$. We explain the proof sketch here; the complete proof of Theorem 6.1 is in Appendix E.

Proof Sketch of Theorem 6.1 Let $F : \prod_{i=1}^{2} \Delta^\circ(A_i) \to \prod_{i=1}^{2} \Delta^\circ(A_i)$ be a function which maps the reference strategies r to the associated stationary point $\pi^{\mu,r}$. We also define $r^{k+1} = F(r^k)$, and $r^0 \in \prod_{i=1}^{2} \Delta^\circ(A_i)$ as the starting reference strategy profile. Note that the function F is well-defined because the stationary point $\pi^{\mu,r}$ of (RMD) is unique for each $r \in \prod_{i=1}^{2} \Delta^\circ(A_i)$ (Abe et al., 2022). First, we prove that the distance between Π^* and r^k decreases monotonically as k increases:
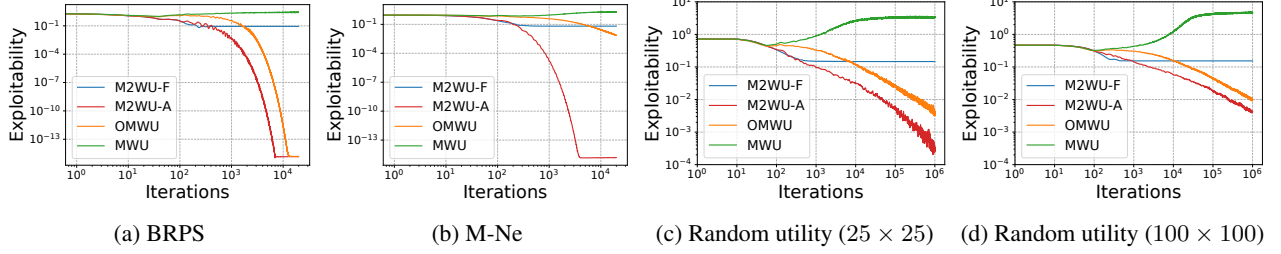
Lemma 6.2. *For any $k \geq 0$, if $r^k \in \prod_{i=1}^{2} \Delta^\circ(A_i) \setminus \Pi^*$, then:*

$$\min_{\pi^* \in \Pi^*} \mathrm{KL}(\pi^*, r^{k+1}) < \min_{\pi^* \in \Pi^*} \mathrm{KL}(\pi^*, r^k).$$

Otherwise, if $r^k \in \Pi^*$, then $r^{k+1} = r^k \in \Pi^*$.

We also show that $F(\cdot)$ is a continuous function:

Lemma 6.3. *Let $F(r) : \prod_{i=1}^{2} \Delta^\circ(A_i) \to \prod_{i=1}^{2} \Delta^\circ(A_i)$ be a function which maps the reference strategies r to the associated stationary point $\pi^{\mu,r}$ of (RMD). Then, $F(\cdot)$ is a continuous function on $\prod_{i=1}^{2} \Delta^\circ(A_i)$.*

(a) BRPS (b) M-Ne (c) Random utility (25×25) (d) Random utility (100×100)

Figure 2: Exploitability of π^t for M2WU, MWU, and OMWU with full feedback.

For these lemmas, we can use Lyapunov arguments to obtain a convergence result for $\{\pi^{\mu,r^k}\}_{k\geq 0}$. □

Remark 6.4. One might think that a simple annealing approach that gradually decrease mutation parameter μ, leads dynamics to reach an exact Nash equilibrium. While this is indeed the case, Theorem 5.1 indicates that the learning rate η must also be significantly reduced. The annealing approach would make the convergence speed very slow.

7 EXPERIMENTS

We here abbreviate M2WU with a fixed reference strategy profile (Algorithm 1 with $N = \infty$) as M2WU-F, while M2WU with adaptive reference strategy profiles as M2WU-A. This section conducts a series of experiments to demonstrate how the four algorithms, i.e., MWU, OMWU, M2WU-F, and M2WU-A, including ours, perform.

We focus on four games: Biased Rock-Paper-Scissors (BRPS), Multiple Nash Equilibria (M-Ne), and two random utility games with 25 and 100 actions. Note that we borrow the M-Ne game from Wei et al. (2021b). Tables 1 and 2 provide the payoff matrices for BRPS and M-Ne, respectively.
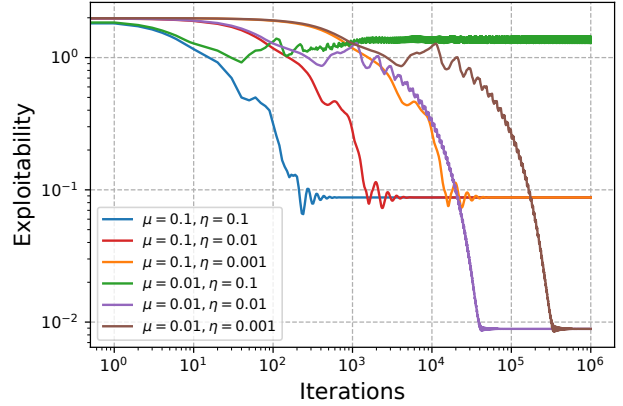
Table 1: Biased RPS game matrix

	R	P	S
R	0	−1	3
P	1	0	−1
S	−3	1	0

Table 2: M-Ne game matrix

	y_1	y_2	y_3	y_4	y_5
x_1	0	1	−1	0	0
x_2	−1	0	1	0	0
x_3	1	−1	0	0	0
x_4	1	−1	0	−2	1
x_5	1	−1	0	1	−2

BRPS has the unique Nash equilibrium $\Pi_i^* = \{(1/5, 3/5, 1/5)\}$. M-Ne has the following set of

Figure 3: Exploitability of π^t for M2WU-F with varying $\mu \in \{0.1, 0.01\}$ and $\eta \in \{0.1, 0.01, 0.001\}$ in BRPS with full feedback.

Nash equilibria:

$$\Pi_1^* = \{(1/3, 1/3, 1/3, 0, 0)\},$$
$$\Pi_2^* = \left\{y \in \Delta^5 \mid y_1 = y_2 = y_3; \; y_5/2 \leq y_4 \leq 2y_5\right\}.$$

Let us proceed to random utility games to consider how our algorithms perform in relatively large games whose numbers of actions are 25 or 100. We draw each utility (or payoff) component from the standard Gaussian distribution $\mathcal{N}(0,1)$ in an i.i.d. manner.

For each game, we average exploitability over 100 instances with different random seeds. We also set the initial strategy profile π^0 uniformly at random in $\prod_{i=1}^2 \Delta^\circ(A_i)$ in each instance for BRPS and M-Ne with full feedback. In other instances, the initial strategy is set to $(1/|A_i|)_{a\in A_i}$ for $i \in \{1, 2\}$.

7.1 Full Feedback

This section examines the algorithms with full feedback. Unless noticed, we use the learning rate η of 0.1 and the mutation rate μ of 0.1. For M2WU-F, we also assume the reference strategy profile r_i^0 is fixed at $(1/|A_i|)_{a\in A_i}$. For M2WU-A, we update it every $N = 100$ iterations.
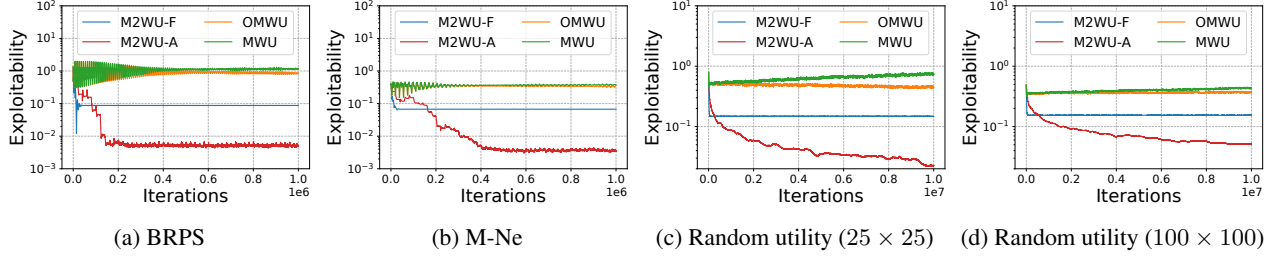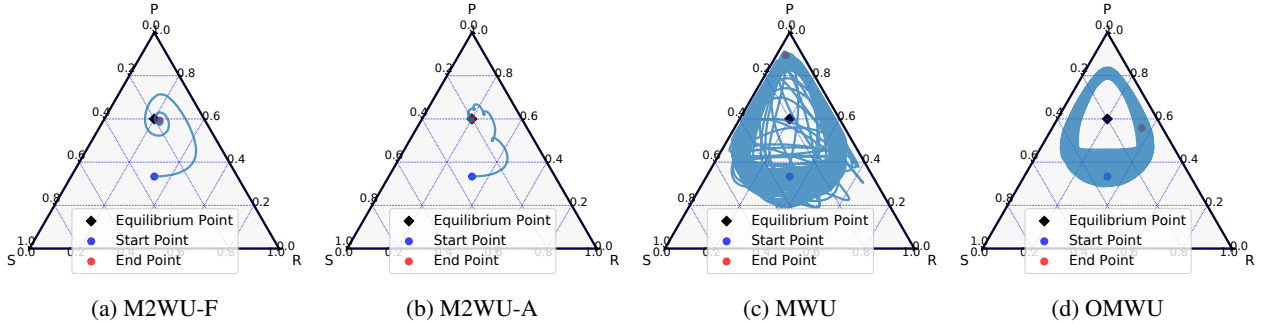
(a) BRPS   (b) M-Ne   (c) Random utility (25×25)   (d) Random utility (100×100)

Figure 4: Exploitability of π^t for M2WU, MWU, and OMWU with noisy feedback.

(a) M2WU-F   (b) M2WU-A   (c) MWU   (d) OMWU

Figure 5: Trajectories of π^t for M2WU, MWU and OMWU in BRPS with noisy feedback. We set the initial strategy to $\pi_i^0 = (1/|A_i|)_{a \in A_i}$ for $i \in \{1, 2\}$. The black point represents the equilibrium strategy. The blue/red points represent the initial/final points, respectively.

Figure 2 averages the exploitability of the last-iterate strategy π^t for the four algorithms in the four games. We observe that, in any game, MWU and M2WU-F do not converge to an equilibrium and the exploitability reaches around 0.1 at best. Both OMWU and M2WU-A exhibit clear convergence. M2WU-A converges faster than OMWU, which may converge to the same exploitability in the long run.

Note that the quick convergence of M2WU-F to a constant value lower than $2\mu = 0.2$ supports Corollary 5.2 in the upper bound on the exploitability of π^t. The best performance of M2WU-A supports Theorem 6.1 and Lemma 6.2, which imply that the sequence of stationary points of (RMD) converges to the Nash equilibrium.

We have so far fixed the learning rate at 0.1 and the mutation rate at 0.1. Figure 3 exhibits the exploitability of M2WU-F in BRPS with varying $\eta \in \{0.1, 0.01, 0.001\}$ and $\mu \in \{0.1, 0.01\}$. The tendencies in the other games are qualitatively similar. We observe that the lower mutation rate makes exploitability low, while the larger mutation rate makes convergence fast. When the mutation rate is low (e.g., 0.01), the learning rate must be lower to enjoy the last-iterate convergence. Otherwise, agents do not properly learn equilibrium (See the green line with $\eta = 0.1$ and $\mu = 0.01$ in Figure 3). This is why we chose $\eta = 0.1$ and $\mu = 0.1$ as a baseline. With them, M2WU-F quickly converges, though the exploitability can be tuned. This result highlights the relationship between μ and the upper bound

on η in Theorem 5.1.

7.2 Noisy Feedback

This section turns to the noisy feedback case, where payoffs observed in each period are perturbed. Players observe noisy estimates $\hat{q}_i^{\pi^t}(a_i) = q_i^{\pi^t}(a_i) + \xi^t(a_i)$ of the gradient vectors at each period, where the noise $\xi^t(a_i)$ is drawn from the Gaussian distribution $\mathcal{N}(0, 0.1^2)$ for $i \in \{1, 2\}$, $a_i \in A_i$ and $t \geq 0$ in an i.i.d. manner. We here use the learning rate η of 0.001 and the mutation rate μ of 0.1 for M2WU-F and 0.5 for M2WU-A. We also update it every $N = 20,000$ period in Algorithm 1, which is much longer than the full feedback case to handle noisy estimates. Note that, for OMWU, we use the noisy gradient vector $\hat{q}_i^{\pi^{t-1}}$ at the previous step $t - 1$ as the prediction vector.

Figure 4 averages the exploitability for the four algorithms in the four games. In all of the games, M2WU-F and -A always outperform MWU and OMWU. The two existing algorithms first oscillate, do not improve much, and remain far from equilibrium. As in the full feedback case, although M2WU-A converges less rapidly than M2WU-F, it admits significantly low exploitability. This tendency does not change even when the action space is larger and even when the learning rate is decayed. We provide additional experimental results when using different learning rates $\eta \in \{0.1, 0.05, 0.01, 0.005\}$ in Appendix F.1. Also,

the case with the decayed learning rate, formally, $\eta_t = t^{-\frac{3}{4}}$, is placed in Appendix F.2.

Figure 5 demonstrates the trajectories of π^t of each algorithm in BRPS. We fix the initial strategy to be uniform. Both M2WU-F and M2WU-A exhibit a clear convergence near the equilibrium (red) point, while MWU and OMWU do not at all. This result strongly supports Figure 4(a). Note that M2WU-A obtains the strategy with lower exploitability than M2WU-F as Theorem 6.1 suggests.

8 CONCLUSION

In this paper, we proposed M2WU, an algorithm that utilizes a simple idea of stabilizing learning dynamics through mutation with a reference strategy. We proved that in both full and noisy feedback settings, the last-iterate strategy converges to the stationary point of RMD. In particular, we showed that such convergence occurs exponentially fast with a constant learning rate in the full feedback setting. Furthermore, last-iterate convergence to an exact Nash equilibrium was also proven by iteratively reusing the converged stationary point as a subsequent reference strategy. The numerical experiments showed that, even with the presence of noise, the strategy updated by M2WU exhibits a lower exploitability than MWU and OMWU. Future research could examine the convergence rate with noisy feedback and extend M2WU and its analyses to extensive-form games.

Acknowledgements

References

Abe, K. and Kaneko, Y. (2021). Off-policy exploitability-evaluation in two-player zero-sum markov games. In *AAMAS*, pages 78–87.

Abe, K., Sakamoto, M., and Iwasaki, A. (2022). Mutation-driven follow the regularized leader for last-iterate convergence in zero-sum games. In *UAI*, pages 1–10.

Azizian, W., Iutzeler, F., Malick, J., and Mertikopoulos, P. (2021). The last-iterate convergence rate of optimistic mirror descent in stochastic variational inequalities. In *COLT*, pages 326–358.

Bailey, J. P. and Piliouras, G. (2018). Multiplicative weights update in zero-sum games. In *Economics and Computation*, pages 321–338.

Banerjee, B. and Peng, J. (2005). Efficient no-regret multi-agent learning. In *AAAI*, pages 41–46.

Bauer, J., Broom, M., and Alonso, E. (2019). The stabilization of equilibria in evolutionary game dynamics through mutation: mutation limits in evolutionary games. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 475(2231):20190355.

Benaïm, M. (1999). Dynamics of stochastic approximation algorithms. In *Séminaire de Probabilités XXXIII*, pages 1–68.

Benaïm, M. and Hirsch, M. W. (1996). Asymptotic pseudotrajectories and chain recurrent flows, with applications. *Journal of Dynamics and Differential Equations*, 8(1):141–176.

Bernasconi, M., Marchesi, A., and Trovò, F. (2022). Last-iterate convergence to trembling-hand perfect equilibria. *arXiv preprint arXiv:2208.08238*.

Bloembergen, D., Tuyls, K., Hennes, D., and Kaisers, M. (2015). Evolutionary dynamics of multi-agent learning: A survey. *Journal of Artificial Intelligence Research*, 53:659–697.

Blum, A. and Monsour, Y. (2007). Learning, regret minimization, and equilibria. In *Algorithmic game theory*, pages 79–102. Cambridge University Press.

Bomze, I. M. and Burger, R. (1995). Stability by mutation in evolutionary games. *Games and Economic Behavior*, 11(2):146–172.

Borkar, V. S. (2009). *Stochastic approximation: a dynamical systems viewpoint*, volume 48. Springer.

Bravo, M., Leslie, D., and Mertikopoulos, P. (2018). Bandit learning in concave n-person games. In *NeurIPS*, pages 5666–5676.

Busoniu, L., Babuska, R., and De Schutter, B. (2008). A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(2):156–172.

Börgers, T. and Sarin, R. (1997). Learning through reinforcement and replicator dynamics. *Journal of Economic Theory*, 77(1):1–14.

Cai, Y., Oikonomou, A., and Zheng, W. (2022). Tight last-iterate convergence of the extragradient method for constrained monotone variational inequalities. *arXiv preprint arXiv:2204.09228*.

Cen, S., Wei, Y., and Chi, Y. (2021). Fast policy extragradient methods for competitive games with entropy regularization. In *NeurIPS*, pages 27952–27964.

Cohen, J., Héliou, A., and Mertikopoulos, P. (2017). Learning with bandit feedback in potential games. In *NeurIPS*, pages 6372–6381.

Daskalakis, C., Deckelbaum, A., and Kim, A. (2011). Near-optimal no-regret algorithms for zero-sum games. In *SODA*, pages 235–254.

Daskalakis, C., Ilyas, A., Syrgkanis, V., and Zeng, H. (2018). Training gans with optimism. In *ICLR*.

Daskalakis, C. and Panageas, I. (2018). The limit points of (optimistic) gradient descent in min-max optimization. In *NeurIPS*, pages 9256–9266.

Daskalakis, C. and Panageas, I. (2019). Last-iterate convergence: Zero-sum games and constrained min-max optimization. In *ITCS*, pages 27:1–27:18.

de Montbrun, É. and Renault, J. (2022). Convergence of optimistic gradient descent ascent in bilinear games. *arXiv preprint arXiv:2208.03085*.

Farina, G., Lee, C.-W., Luo, H., and Kroer, C. (2022). Kernelized multiplicative weights for 0/1-polyhedral games: Bridging the gap between learning in extensive-form and normal-form games. In *ICML*, pages 6337–6357.

Giannou, A., Vlatakis-Gkaragkounis, E.-V., and Mertikopoulos, P. (2021a). On the rate of convergence of regularized learning in games: From bandits and uncertainty to optimism and beyond. In *NeurIPS*, pages 22655–22666.

Giannou, A., Vlatakis-Gkaragkounis, E. V., and Mertikopoulos, P. (2021b). Survival of the strictest: Stable and unstable equilibria under regularized learning with partial information. In *COLT*, pages 2147–2148.

Golowich, N., Pattathil, S., and Daskalakis, C. (2020a). Tight last-iterate convergence rates for no-regret learning in multi-player games. In *NeurIPS*, pages 20766–20778.

Golowich, N., Pattathil, S., Daskalakis, C., and Ozdaglar, A. (2020b). Last iterate is slower than averaged iterate in smooth convex-concave saddle point problems. In *COLT*, pages 1758–1784.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *NeurIPS*, pages 2672–2680.

Hofbauer, J. and Sigmund, K. (1998). *Evolutionary Games and Population Dynamics*. Cambridge University Press, Cambridge.

Hofbauer, J., Sorin, S., and Viossat, Y. (2009). Time average replicator and best-reply dynamics. *Mathematics of Operations Research*, 34(2):263–269.

Hsieh, Y.-G., Antonakopoulos, K., Cevher, V., and Mertikopoulos, P. (2022). No-regret learning in games with noisy feedback: Faster rates and adaptivity via learning rate separation. *arXiv preprint arXiv:2206.06015*.

Hsieh, Y.-G., Iutzeler, F., Malick, J., and Mertikopoulos, P. (2019). On the convergence of single-call stochastic extra-gradient methods. In *NeurIPS*, pages 6938–6948.

Johanson, M., Bard, N., Burch, N., and Bowling, M. (2012). Finding optimal abstract strategies in extensive-form games. In *AAAI*, pages 1371–1379.

Johanson, M., Waugh, K., Bowling, M., and Zinkevich, M. (2011). Accelerating best response calculation in large extensive games. In *IJCAI*, pages 258–265.

Kannan, A. and Shanbhag, U. V. (2019). Optimal stochastic extragradient schemes for pseudomonotone stochastic variational inequality problems and their variants. *Computational Optimization and Applications*, 74(3):779–820.

Lei, Q., Nagarajan, S. G., Panageas, I., et al. (2021). Last iterate convergence in no-regret learning: constrained min-max optimization for convex-concave landscapes. In *AISTATS*, pages 1441–1449.

Liang, T. and Stokes, J. (2019). Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks. In *AISTATS*, pages 907–915.

Liu, M., Ozdaglar, A., Yu, T., and Zhang, K. (2022). The power of regularization in solving extensive-form games. *arXiv preprint arXiv:2206.09495*.

Lockhart, E., Lanctot, M., Pérolat, J., Lespiau, J.-B., Morrill, D., TImbers, F., and Tuyls, K. (2019). Computing approximate equilibria in sequential adversarial games by exploitability descent. In *IJCAI*, pages 464–470.

Mertikopoulos, P., Lecouat, B., Zenati, H., Foo, C.-S., Chandrasekhar, V., and Piliouras, G. (2019). Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile. In *ICLR*.

Mertikopoulos, P., Papadimitriou, C., and Piliouras, G. (2018). Cycles in adversarial regularized learning. In *SODA*, pages 2703–2717.

Mertikopoulos, P. and Zhou, Z. (2019). Learning in games with continuous action sets and unknown payoff functions. *Mathematical Programming*, 173(1):465–507.

Mokhtari, A., Ozdaglar, A., and Pattathil, S. (2020). A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. In *AISTATS*, pages 1497–1507.

Nash, J. (1951). Non-cooperative games. *Annals of mathematics*, pages 286–295.

Nguyen, T.-D., Zemhoho, A. B., Tran-Thanh, L., et al. (2021). Last round convergence and no-dynamic regret in asymmetric repeated games. In *ALT*, pages 553–577.

Perolat, J., Munos, R., Lespiau, J.-B., Omidshafiei, S., Rowland, M., Ortega, P., Burch, N., Anthony, T., Balduzzi, D., De Vylder, B., et al. (2021). From poincaré recurrence to convergence in imperfect information games: Finding equilibrium via regularization. In *ICML*, pages 8525–8535.

Rakhlin, A. and Sridharan, K. (2013a). Online learning with predictable sequences. In *COLT*, pages 993–1019.

Rakhlin, S. and Sridharan, K. (2013b). Optimization, learning, and games with predictable sequences. In *NeurIPS*, pages 3066–3074.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407.

Timbers, F., Bard, N., Lockhart, E., Lanctot, M., Schmid, M., Burch, N., Schrittwieser, J., Hubert, T., and Bowling, M. (2022). Approximate exploitability: Learning a best response. In *IJCAI*, pages 3487–3493.

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer.

Wei, C.-Y., Lee, C.-W., Zhang, M., and Luo, H. (2021a). Last-iterate convergence of decentralized optimistic gradient descent/ascent in infinite-horizon competitive markov games. In *COLT*, pages 4259–4299.

Wei, C.-Y., Lee, C.-W., Zhang, M., and Luo, H. (2021b). Linear last-iterate convergence in constrained saddle-point optimization. In *ICLR*.

Zagorsky, B. M., Reiter, J. G., Chatterjee, K., and Nowak, M. A. (2013). Forgiver triumphs in alternating prisoner's dilemma. *PLOS ONE*, 8(12):1–8.

Zinkevich, M., Johanson, M., Bowling, M., and Piccione, C. (2007). Regret minimization in games with incomplete information. In *NeurIPS*, pages 1729–1736.

A   NOTATIONS

In this section, we summarize the notations we use in Table 3.

Table 3: Notations

Symbol	Description
A_i	Action set for player i
u_i	Utility function for player i
π_i	Strategy for player i
π	Strategy profile
v_i^π	Player i's expected utility for a given strategy profile π
q_i^π	Player i's conditional expected utility vector for a given strategy profile π
$\hat{q}_i^{\pi^t}$	Player i's noisy conditional expected utility vector at iteration t
ξ^t	Noise vector at iteration t
π^*	Nash equilibrium
Π_i^*	Set of Nash equilibria for player i
$\Delta(A_i)$	Probability simplex on A_i
$\Delta^\circ(A_i)$	Interior of $\Delta(A_i)$
$\mathrm{KL}(\cdot, \cdot)$	Kullback-Leibler divergence
η_t	Learning rate at iteration t
μ	Mutation rate
r_i	Reference strategy
$q_i^{\mu,t}$	$\left(\hat{q}_i^{\pi^t}(a) + \frac{\mu}{\pi_i^t(a)}(r_i(a) - \pi_i^t(a)) \right)_{a \in A_i}$
$\pi^{\mu,r}$	Stationary point of (RMD) for given μ and r

B PROOFS FOR THEOREM 5.1

B.1 Proof of Theorem 5.1

Proof of Theorem 5.1. Let us define the following notation:

$$\Omega^{\mu,r} = \left\{ \pi \in \prod_{i=1}^{2} \Delta(A_i) \mid \mathrm{KL}(\pi^{\mu,r}, \pi) \leq \mathrm{KL}(\pi^{\mu,r}, \pi^0) \right\},$$

$$\rho = \min_{\pi \in \Omega^{\mu,r}} \min_{i \in \{1,2\}, a \in A_i} \pi_i(a_i) > 0,$$

$$\zeta = \frac{1}{2u_{\max} + \frac{1}{\rho} \max_{i \in \{1,2\}, a \in A_i} r_i(a)} > 0,$$

$$\alpha = \min_{i \in \{1,2\}, a \in A_i} \frac{r_i(a)}{\pi_i^{\mu,r}(a)} > 0,$$

$$\beta = \frac{16}{\rho^2} \left(\max_{i \in \{1,2\}, a \in A_i} \frac{r_i(a)}{\pi_i^{\mu,r}(a)} \right)^2 > 0,$$

$$\gamma = 16u_{\max}^2 > 0.$$

We prove the statement by mathematical induction. Clearly, for $t = 0$, we have $\mathrm{KL}(\pi^{\mu,r}, \pi^t) \leq \mathrm{KL}(\pi^{\mu,r}, \pi^0)$ and $\pi^0 \in \Omega^{\mu,r}$. Let us assume that $\pi^t \in \Omega^{\mu,r}$, i.e., $\mathrm{KL}(\pi^{\mu,r}, \pi^t) \leq \mathrm{KL}(\pi^{\mu,r}, \pi^0)$. Under the assumption that $\mathrm{KL}(\pi^{\mu,r}, \pi^t) \leq \mathrm{KL}(\pi^{\mu,r}, \pi^0)$, we have $\pi_i^t(a) \geq \rho$ for all $i \in \{1, 2\}$ and $a \in A_i$.

We first derive the difference equation for $\mathrm{KL}(\pi^{\mu,r}, \pi^t)$:

Lemma B.1. *Let $\pi^{\mu,r} \in \prod_{i=1}^{2} \Delta(A_i)$ be a stationary point of (RMD). Then, π^t updated by M2WU satisfies that:*

$$\mathrm{KL}(\pi^{\mu,r}, \pi^{t+1}) - \mathrm{KL}(\pi^{\mu,r}, \pi^t) = \eta_t \sum_{i=1}^{2} \left(v_i^{\pi_i^t, \pi_{-i}^{\mu,r}} + \mu - \mu \sum_{a_i \in A_i} r_i(a_i) \frac{\pi_i^{\mu,r}(a_i)}{\pi_i^t(a_i)} \right) + \mathrm{KL}(\pi^t, \pi^{t+1}).$$

Moreover, under the assumption that $\eta_t = \eta \in (0, \min(\frac{\mu\alpha}{\mu^2\beta+\gamma}, \zeta))$, the statement of Lemma 5.4 holds. By combining Lemmas 5.3, 5.4, and B.1, we get:

$$\mathrm{KL}(\pi^{\mu,r}, \pi^{t+1}) - \mathrm{KL}(\pi^{\mu,r}, \pi^t) \leq - \eta\mu \sum_{i=1}^{2} \sum_{a \in A_i} r_i(a) \left(\sqrt{\frac{\pi_i^t(a)}{\pi_i^{\mu,r}(a)}} - \sqrt{\frac{\pi_i^{\mu,r}(a)}{\pi_i^t(a)}} \right)^2$$

$$+ 8\eta^2\mu^2 \sum_{i=1}^{2} \sum_{a \in A_i} r_i(a)^2 \left(\frac{1}{\pi_i^{\mu,r}(a)} - \frac{1}{\pi_i^t(a)} \right)^2$$

$$+ 8\eta^2 u_{\max}^2 \sum_{i=1}^{2} \|\pi_i^t - \pi_i^{\mu,r}\|_1^2. \tag{3}$$

We prove the lower bound on $\sum_{i=1}^{2} \sum_{a \in A_i} r_i(a) \left(\sqrt{\frac{\pi_i^t(a)}{\pi_i^{\mu,r}(a)}} - \sqrt{\frac{\pi_i^{\mu,r}(a)}{\pi_i^t(a)}} \right)^2$ as follows:

$$\sum_{i=1}^{2} \sum_{a \in A_i} r_i(a) \left(\sqrt{\frac{\pi_i^t(a)}{\pi_i^{\mu,r}(a)}} - \sqrt{\frac{\pi_i^{\mu,r}(a)}{\pi_i^t(a)}} \right)^2 = \sum_{i=1}^{2} \sum_{a \in A_i} \frac{r_i(a)}{\pi_i^{\mu,r}(a)} \frac{(\pi_i^t(a) - \pi_i^{\mu,r}(a))^2}{\pi_i^t(a)}$$

$$\geq \left(\min_{i \in \{1,2\}, a \in A_i} \frac{r_i(a)}{\pi_i^{\mu,r}(a)} \right) \sum_{i=1}^{2} \sum_{a \in A_i} \frac{(\pi_i^t(a) - \pi_i^{\mu,r}(a))^2}{\pi_i^t(a)}$$

$$\geq \left(\min_{i \in \{1,2\}, a \in A_i} \frac{r_i(a)}{\pi_i^{\mu,r}(a)} \right) \sum_{i=1}^{2} \ln \left(1 + \sum_{a \in A_i} \frac{(\pi_i^t(a) - \pi_i^{\mu,r}(a))^2}{\pi_i^t(a)} \right)$$

$$= \left(\min_{i \in \{1,2\}, a \in A_i} \frac{r_i(a)}{\pi_i^{\mu,r}(a)} \right) \sum_{i=1}^{2} \ln \left(\sum_{a \in A_i} \pi_i^{\mu,r}(a) \frac{\pi_i^{\mu,r}(a)}{\pi_i^t(a)} \right)$$

$$\geq \left(\min_{i \in \{1,2\}, a \in A_i} \frac{r_i(a)}{\pi_i^{\mu,r}(a)} \right) \sum_{i=1}^{2} \sum_{a \in A_i} \pi_i^{\mu,r}(a) \ln \left(\frac{\pi_i^{\mu,r}(a)}{\pi_i^t(a)} \right)$$

$$= \left(\min_{i \in \{1,2\}, a \in A_i} \frac{r_i(a)}{\pi_i^{\mu,r}(a)} \right) \mathrm{KL}(\pi^{\mu,r}, \pi^t), \tag{4}$$

where the second inequality follows from $x \geq \ln(1+x)$ for all $x > 0$, and the third inequality follows from the concavity of the $\ln(\cdot)$ function and Jensen's inequality for concave functions. Next, $\sum_{i=1}^{2} \sum_{a \in A_i} r_i(a)^2 \left(\frac{1}{\pi_i^{\mu,r}(a)} - \frac{1}{\pi_i^t(a)} \right)^2$ is upper bounded as follows:

$$\sum_{i=1}^{2} \sum_{a \in A_i} r_i(a)^2 \left(\frac{1}{\pi_i^{\mu,r}(a)} - \frac{1}{\pi_i^t(a)} \right)^2 = \sum_{i=1}^{2} \sum_{a \in A_i} \left(\frac{r_i(a)}{\pi_i^{\mu,r}(a)\pi_i^t(a)} \right)^2 (\pi_i^{\mu,r}(a) - \pi_i^t(a))^2$$

$$\leq \frac{1}{\rho^2} \left(\max_{i \in \{1,2\}, a \in A_i} \frac{r_i(a)}{\pi_i^{\mu,r}(a)} \right)^2 \sum_{i=1}^{2} \|\pi_i^{\mu,r} - \pi_i^t\|_2^2$$

$$\leq \frac{1}{\rho^2} \left(\max_{i \in \{1,2\}, a \in A_i} \frac{r_i(a)}{\pi_i^{\mu,r}(a)} \right)^2 \sum_{i=1}^{2} \|\pi_i^{\mu,r} - \pi_i^t\|_1^2$$

$$\leq \frac{2}{\rho^2} \left(\max_{i \in \{1,2\}, a \in A_i} \frac{r_i(a)}{\pi_i^{\mu,r}(a)} \right)^2 \mathrm{KL}(\pi^{\mu,r}, \pi^t), \tag{5}$$

where the last inequality follows from Pinsker's inequality (Tsybakov, 2009). Similarly, $\sum_{i=1}^{2} \|\pi_i^t - \pi_i^{\mu,r}\|_1^2$ is upper bounded as

$$\sum_{i=1}^{2} \|\pi_i^t - \pi_i^{\mu,r}\|_1^2 \leq 2\mathrm{KL}(\pi^{\mu,r}, \pi^t). \tag{6}$$

By combining (3), (4), (5), and (6), we have:

$$\mathrm{KL}(\pi^{\mu,r}, \pi^{t+1}) - \mathrm{KL}(\pi^{\mu,r}, \pi^t)$$

$$\leq -\eta\mu \left(\min_{i \in \{1,2\}, a \in A_i} \frac{r_i(a)}{\pi_i^{\mu,r}(a)} \right) \mathrm{KL}(\pi^{\mu,r}, \pi^t) + 16\eta^2 \left(\frac{\mu^2}{\rho^2} \left(\max_{i \in \{1,2\}, a \in A_i} \frac{r_i(a)}{\pi_i^{\mu,r}(a)} \right)^2 + u_{\max}^2 \right) \mathrm{KL}(\pi^{\mu,r}, \pi^t)$$

$$= \left(-\eta\mu\alpha + \eta^2(\mu^2\beta + \gamma) \right) \mathrm{KL}(\pi^{\mu,r}, \pi^t).$$

Thus, we get:

$$\mathrm{KL}(\pi^{\mu,r}, \pi^{t+1}) \leq (1 - (\eta\mu\alpha - \eta^2(\mu^2\beta + \gamma)))\mathrm{KL}(\pi^{\mu,r}, \pi^t),$$

and then, for $\eta \in (0, \min(\frac{\mu\alpha}{\mu^2\beta+\gamma}, \zeta))$:

$$\mathrm{KL}(\pi^{\mu,r}, \pi^{t+1}) - \mathrm{KL}(\pi^{\mu,r}, \pi^t) \leq 0.$$

Thus, if $\eta \in (0, \min(\frac{\mu\alpha}{\mu^2\beta+\gamma}, \zeta))$, then $\mathrm{KL}(\pi^{\mu,r}, \pi^{t+1}) \leq \mathrm{KL}(\pi^{\mu,r}, \pi^t) \leq \mathrm{KL}(\pi^{\mu,r}, \pi^0)$ and $\pi^{t+1} \in \Omega^{\mu,r}$ also hold. By mathematical induction, if $\eta \in (0, \min(\frac{\mu\alpha}{\mu^2\beta+\gamma}, \zeta))$, for all $t \geq 0$:

$$\mathrm{KL}(\pi^{\mu,r}, \pi^{t+1}) \leq (1 - (\eta\mu\alpha - \eta^2(\mu^2\beta + \gamma)))\mathrm{KL}(\pi^{\mu,r}, \pi^t) \leq \cdots \leq (1 - (\eta\mu\alpha - \eta^2(\mu^2\beta + \gamma)))^{t+1}\mathrm{KL}(\pi^{\mu,r}, \pi^0).$$

$\square$

B.2 Proof of Lemma 5.3

Proof of Lemma 5.3. First, we introduce the following lemma from Abe et al. (2022):

Lemma B.2 (Lemma 5.6 of Abe et al. (2022)). *Let $\pi_i^{\mu,r} \in \Delta(A_i)$ be a stationary point of (RMD) for $i \in \{1,2\}$. Then, for any $\pi_i' \in \Delta(A_i)$:*

$$v_i^{\pi_i',\pi_{-i}^{\mu,r}} = v_i^{\pi^{\mu,r}} + \mu - \mu \sum_{a_i \in A_i} r_i(a_i) \frac{\pi_i'(a_i)}{\pi_i^{\mu,r}(a_i)}.$$

From this lemma, we have:

$$\sum_{i=1}^{2} v_i^{\pi_i^t,\pi_{-i}^{\mu,r}} + 2\mu - \mu \sum_{i=1}^{2} \sum_{a_i \in A_i} r_i(a_i) \frac{\pi_i^{\mu,r}(a_i)}{\pi_i^t(a_i)} = \sum_{i=1}^{2} v_i^{\pi^{\mu,r}} + 4\mu - \mu \sum_{i=1}^{2} \sum_{a_i \in A_i} r_i(a_i) \left(\frac{\pi_i^t(a_i)}{\pi_i^{\mu,r}(a_i)} + \frac{\pi_i^{\mu,r}(a_i)}{\pi_i^t(a_i)} \right)$$

$$= 4\mu - \mu \sum_{i=1}^{2} \sum_{a_i \in A_i} r_i(a_i) \left(\frac{\pi_i^t(a_i)}{\pi_i^{\mu,r}(a_i)} + \frac{\pi_i^{\mu,r}(a_i)}{\pi_i^t(a_i)} \right)$$

$$= -\mu \sum_{i=1}^{2} \sum_{a_i \in A_i} r_i(a_i) \left(\sqrt{\frac{\pi_i^t(a_i)}{\pi_i^{\mu,r}(a_i)}} - \sqrt{\frac{\pi_i^{\mu,r}(a_i)}{\pi_i^t(a_i)}} \right)^2,$$

where the second equality follows from $\sum_{i=1}^{2} v_i^{\pi^{\mu,r}} = 0$ by the definition of zero-sum games. $\qquad\square$

B.3 Proof of Lemma 5.4

Proof of Lemma 5.4. Let us assume that $\eta \in (0, \zeta)$, where α, β, γ, and ζ are defined in Appendix B.1.

First, we have:

$$\mathrm{KL}(\pi^t, \pi^{t+1}) = \sum_{i=1}^{2} \sum_{a \in A_i} \pi_i^t(a) \ln \frac{\pi_i^t(a)}{\pi_i^{t+1}(a)}$$

$$= 2 \sum_{i=1}^{2} \sum_{a \in A_i} \frac{1}{2} \pi_i^t(a) \ln \frac{\pi_i^t(a)}{\pi_i^{t+1}(a)} \le 2 \ln \left(\frac{1}{2} \sum_{i=1}^{2} \sum_{a \in A_i} \pi_i^t(a) \frac{\pi_i^t(a)}{\pi_i^{t+1}(a)} \right),$$

where the inequality follows from the concavity of the $\ln(\cdot)$ function and Jensen's inequality for concave functions. Here, from the update rule (1):

$$\frac{\pi_i^t(a)}{\pi_i^{t+1}(a)} = \frac{\sum_{a' \in A_i} \pi_i^t(a') \exp \left(\eta \left(q_i^{\pi^t}(a') + \mu \frac{r_i(a')}{\pi_i^t(a')} \right) \right)}{\exp \left(\eta \left(q_i^{\pi^t}(a) + \mu \frac{r_i(a)}{\pi_i^t(a)} \right) \right)},$$

and then we get:

$$\mathrm{KL}(\pi^t, \pi^{t+1}) \le 2 \ln \left(\frac{1}{2} \sum_{i=1}^{2} \sum_{a \in A_i} \pi_i^t(a) \frac{\sum_{a' \in A_i} \pi_i^t(a') \exp \left(\eta \left(q_i^{\pi^t}(a') + \mu \frac{r_i(a')}{\pi_i^t(a')} \right) \right)}{\exp \left(\eta \left(q_i^{\pi^t}(a) + \mu \frac{r_i(a)}{\pi_i^t(a)} \right) \right)} \right)$$

$$= 2 \ln \left(\frac{1}{2} \sum_{i=1}^{2} \sum_{a \in A_i} \sum_{a' \in A_i} \pi_i^t(a) \pi_i^t(a') \exp \left(\eta \left(q_i^{\pi^t}(a') + \mu \frac{r_i(a')}{\pi_i^t(a')} - q_i^{\pi^t}(a) - \mu \frac{r_i(a)}{\pi_i^t(a)} \right) \right) \right). \quad (7)$$

Furthermore, from the assumption for the learning rate $\eta < \frac{1}{2u_{\max} + \frac{\mu}{\rho} \max_{i \in \{1,2\}, a \in A_i} r_i(a)} \le \zeta$, we have $\eta \left(q_i^{\pi^t}(a') + \mu \frac{r_i(a')}{\pi_i^t(a')} - q_i^{\pi^t}(a) - \mu \frac{r_i(a)}{\pi_i^t(a)} \right) \le 1$. Thus, we can use the fact that $\exp(x) \le 1 + x + x^2$ for $x \le 1$, and

then:

$$\frac{1}{2} \sum_{i=1}^{2} \sum_{a \in A_i} \sum_{a' \in A_i} \pi_i^t(a) \pi_i^t(a') \exp\left(\eta\left(q_i^{\pi^t}(a') + \mu\frac{r_i(a')}{\pi_i^t(a')} - q_i^{\pi^t}(a) - \mu\frac{r_i(a)}{\pi_i^t(a)}\right)\right)$$

$$\leq \frac{1}{2} \sum_{i=1}^{2} \sum_{a \in A_i} \sum_{a' \in A_i} \pi_i^t(a) \pi_i^t(a')\left(1 + \eta\left(q_i^{\pi^t}(a') + \mu\frac{r_i(a')}{\pi_i^t(a')} - q_i^{\pi^t}(a) - \mu\frac{r_i(a)}{\pi_i^t(a)}\right)\right)$$

$$+ \frac{1}{2} \sum_{i=1}^{2} \sum_{a \in A_i} \sum_{a' \in A_i} \pi_i^t(a) \pi_i^t(a')\left(\eta^2\left(q_i^{\pi^t}(a') + \mu\frac{r_i(a')}{\pi_i^t(a')} - q_i^{\pi^t}(a) - \mu\frac{r_i(a)}{\pi_i^t(a)}\right)^2\right)$$

$$= 1 + \frac{\eta^2}{2} \sum_{i=1}^{2} \sum_{a \in A_i} \sum_{a' \in A_i} \pi_i^t(a) \pi_i^t(a')\left(q_i^{\pi^t}(a') + \mu\frac{r_i(a')}{\pi_i^t(a')} - q_i^{\pi^t}(a) - \mu\frac{r_i(a)}{\pi_i^t(a)}\right)^2$$

$$\leq \exp\left(\frac{\eta^2}{2} \sum_{i=1}^{2} \sum_{a \in A_i} \sum_{a' \in A_i} \pi_i^t(a) \pi_i^t(a')\left(q_i^{\pi^t}(a') + \mu\frac{r_i(a')}{\pi_i^t(a')} - q_i^{\pi^t}(a) - \mu\frac{r_i(a)}{\pi_i^t(a)}\right)^2\right), \tag{8}$$

where the first equality follows from $\sum_{a \in A_i} \sum_{a' \in A_i} \pi_i^t(a) \pi_i^t(a')\left(\eta\left(q_i^{\pi^t}(a') + \mu\frac{r_i(a')}{\pi_i^t(a')} - q_i^{\pi^t}(a) - \mu\frac{r_i(a)}{\pi_i^t(a)}\right)\right) = 0$, and the last inequality follows from $1 + x \leq \exp(x)$ for $x \in \mathbb{R}$. By combining (7) and (8), we get:

$$\text{KL}(\pi^t, \pi^{t+1}) \leq 2\ln\left(\exp\left(\frac{\eta^2}{2} \sum_{i=1}^{2} \sum_{a \in A_i} \sum_{a' \in A_i} \pi_i^t(a) \pi_i^t(a')\left(q_i^{\pi^t}(a') + \mu\frac{r_i(a')}{\pi_i^t(a')} - q_i^{\pi^t}(a) - \mu\frac{r_i(a)}{\pi_i^t(a)}\right)^2\right)\right)$$

$$= \eta^2 \sum_{i=1}^{2} \sum_{a \in A_i} \sum_{a' \in A_i} \pi_i^t(a) \pi_i^t(a')\left(q_i^{\pi^t}(a') + \mu\frac{r_i(a')}{\pi_i^t(a')} - q_i^{\pi^t}(a) - \mu\frac{r_i(a)}{\pi_i^t(a)}\right)^2. \tag{9}$$

Here, by using the ordinary differential equation (RMD), we have for all $i \in \{1, 2\}$ and $a \in A$:

$$q_i^{\pi^{\mu,r}}(a) = v_i^{\pi^{\mu,r}} - \frac{\mu}{\pi_i^{\mu,r}(a)}\left(r_i(a) - \pi_i^{\mu,r}(a)\right).$$

Thus,

$$q_i^{\pi^t}(a') + \mu\frac{r_i(a')}{\pi_i^t(a')} - q_i^{\pi^t}(a) - \mu\frac{r_i(a)}{\pi_i^t(a)}$$

$$= q_i^{\pi^{\mu,r}}(a') + \mu\frac{r_i(a')}{\pi_i^t(a')} - q_i^{\pi^{\mu,r}}(a) - \mu\frac{r_i(a)}{\pi_i^t(a)} + q_i^{\pi^t}(a') - q_i^{\pi^{\mu,r}}(a') - q_i^{\pi^t}(a) + q_i^{\pi^{\mu,r}}(a)$$

$$= \mu\left(\frac{r_i(a)}{\pi_i^{\mu,r}(a)} - \frac{r_i(a)}{\pi_i^t(a)} - \frac{r_i(a')}{\pi_i^{\mu,r}(a')} + \frac{r_i(a')}{\pi_i^t(a')}\right) + q_i^{\pi^t}(a') - q_i^{\pi^{\mu,r}}(a') - q_i^{\pi^t}(a) + q_i^{\pi^{\mu,r}}(a).$$

Then,

$$\sum_{i=1}^{2} \sum_{a \in A_i} \sum_{a' \in A_i} \pi_i^t(a) \pi_i^t(a')\left(q_i^{\pi^t}(a') + \mu\frac{r_i(a')}{\pi_i^t(a')} - q_i^{\pi^t}(a) - \mu\frac{r_i(a)}{\pi_i^t(a)}\right)^2$$

$$= \sum_{i=1}^{2} \sum_{a \in A_i} \sum_{a' \in A_i} \pi_i^t(a) \pi_i^t(a')\left(\mu\left(\frac{r_i(a)}{\pi_i^{\mu,r}(a)} - \frac{r_i(a)}{\pi_i^t(a)} - \frac{r_i(a')}{\pi_i^{\mu,r}(a')} + \frac{r_i(a')}{\pi_i^t(a')}\right) + q_i^{\pi^t}(a') - q_i^{\pi^{\mu,r}}(a') - q_i^{\pi^t}(a) + q_i^{\pi^{\mu,r}}(a)\right)^2$$

$$\leq 2\mu^2 \sum_{i=1}^{2} \sum_{a \in A_i} \sum_{a' \in A_i} \pi_i^t(a) \pi_i^t(a')\left(\frac{r_i(a)}{\pi_i^{\mu,r}(a)} - \frac{r_i(a)}{\pi_i^t(a)} - \frac{r_i(a')}{\pi_i^{\mu,r}(a')} + \frac{r_i(a')}{\pi_i^t(a')}\right)^2$$

$$+ 2\sum_{i=1}^{2} \sum_{a \in A_i} \sum_{a' \in A_i} \pi_i^t(a) \pi_i^t(a')\left(q_i^{\pi^t}(a') - q_i^{\pi^{\mu,r}}(a') - q_i^{\pi^t}(a) + q_i^{\pi^{\mu,r}}(a)\right)^2$$

$$\leq 4\mu^2 \sum_{i=1}^{2} \sum_{a \in A_i} \sum_{a' \in A_i} \pi_i^t(a) \pi_i^t(a')\left(\left(\frac{r_i(a)}{\pi_i^{\mu,r}(a)} - \frac{r_i(a)}{\pi_i^t(a)}\right)^2 + \left(\frac{r_i(a')}{\pi_i^{\mu,r}(a')} - \frac{r_i(a')}{\pi_i^t(a')}\right)^2\right)$$

$$+ 4 \sum_{i=1}^{2} \sum_{a \in A_i} \sum_{a' \in A_i} \pi_i^t(a) \pi_i^t(a') \left(\left(q_i^{\pi^t}(a') - q_i^{\pi^{\mu,r}}(a') \right)^2 + \left(q_i^{\pi^t}(a) - q_i^{\pi^{\mu,r}}(a) \right)^2 \right)$$

$$= 8\mu^2 \sum_{i=1}^{2} \sum_{a \in A_i} \pi_i^t(a) r_i(a)^2 \left(\frac{1}{\pi_i^{\mu,r}(a)} - \frac{1}{\pi_i^t(a)} \right)^2 + 8 \sum_{i=1}^{2} \sum_{a \in A_i} \pi_i^t(a) \left(\sum_{b \in A_{-i}} \left(\pi_{-i}^t(b) - \pi_{-i}^{\mu,r}(b) \right) u_i(a', b) \right)^2$$

$$\leq 8\mu^2 \sum_{i=1}^{2} \sum_{a \in A_i} \pi_i^t(a) r_i(a)^2 \left(\frac{1}{\pi_i^{\mu,r}(a)} - \frac{1}{\pi_i^t(a)} \right)^2 + 8 \sum_{i=1}^{2} \sum_{a \in A_i} \pi_i^t(a) u_{\max}^2 \| \pi_{-i}^t - \pi_{-i}^{\mu,r} \|_1^2$$

$$\leq 8\mu^2 \sum_{i=1}^{2} \sum_{a \in A_i} r_i(a)^2 \left(\frac{1}{\pi_i^{\mu,r}(a)} - \frac{1}{\pi_i^t(a)} \right)^2 + 8 \sum_{i=1}^{2} u_{\max}^2 \| \pi_{-i}^t - \pi_{-i}^{\mu,r} \|_1^2$$

$$= 8\mu^2 \sum_{i=1}^{2} \sum_{a \in A_i} r_i(a)^2 \left(\frac{1}{\pi_i^{\mu,r}(a)} - \frac{1}{\pi_i^t(a)} \right)^2 + 8 \sum_{i=1}^{2} u_{\max}^2 \| \pi_i^t - \pi_i^{\mu,r} \|_1^2, \tag{10}$$

where the first and second inequalities follow from $(a + b)^2 \leq 2(a^2 + b^2)$ for $a, b \in \mathbb{R}$, and the third inequality follows from Hölder's inequality.

By combining (9) and (10), if $\eta \in (0, \zeta)$, for all $t \geq 0$:

$$\mathrm{KL}(\pi^t, \pi^{t+1}) \leq 8\eta^2 \left(\mu^2 \sum_{i=1}^{2} \sum_{a \in A_i} r_i(a)^2 \left(\frac{1}{\pi_i^{\mu,r}(a)} - \frac{1}{\pi_i^t(a)} \right)^2 + u_{\max}^2 \sum_{i=1}^{2} \| \pi_i^t - \pi_i^{\mu,r} \|_1^2 \right).$$

$\square$

B.4 Proof of Lemma B.1

Proof of Lemma B.1. We introduce the following lemma:

Lemma B.3. *For any $\pi \in \prod_{i=1}^{2} \Delta(A_i)$, π^t updated by M2WU satisfies that:*

$$\mathrm{KL}(\pi, \pi^t) = \sum_{i=1}^{2} \left(\left\langle \sum_{s=1}^{t-1} \eta_s q_i^{\mu,s}, \pi_i^t \right\rangle - \psi_i(\pi_i^t) - \left\langle \sum_{s=1}^{t-1} \eta_s q_i^{\mu,s}, \pi_i \right\rangle + \psi_i(\pi_i) \right),$$

where $\psi_i(p) = \sum_{a \in A_i} p(a) \ln p(a)$.

From Lemma B.3, we have:

$$\mathrm{KL}(\pi^{\mu,r}, \pi^{t+1}) - \mathrm{KL}(\pi^{\mu,r}, \pi^t)$$

$$= \sum_{i=1}^{2} \left(\left\langle \sum_{s=1}^{t} \eta_s q_i^{\mu,s}, \pi_i^{t+1} \right\rangle - \psi_i(\pi_i^{t+1}) - \left\langle \sum_{s=1}^{t} \eta_s q_i^{\mu,s}, \pi_i^{\mu,r} \right\rangle + \psi_i(\pi_i^{\mu,r}) \right)$$

$$- \sum_{i=1}^{2} \left(\left\langle \sum_{s=1}^{t-1} \eta_s q_i^{\mu,s}, \pi_i^t \right\rangle - \psi_i(\pi_i^t) - \left\langle \sum_{s=1}^{t-1} \eta_s q_i^{\mu,s}, \pi_i^{\mu,r} \right\rangle + \psi_i(\pi_i^{\mu,r}) \right)$$

$$= \sum_{i=1}^{2} \left(\left\langle \sum_{s=1}^{t} \eta_s q_i^{\mu,s}, \pi_i^{t+1} \right\rangle - \psi_i(\pi_i^{t+1}) - \left\langle \sum_{s=1}^{t} \eta_s q_i^{\mu,s}, \pi_i^t \right\rangle + \psi_i(\pi_i^t) \right) - \eta_t \sum_{i=1}^{2} \langle q_i^{\mu,t}, \pi_i^{\mu,r} - \pi_i^t \rangle$$

$$= \mathrm{KL}(\pi^t, \pi^{t+1}) - \eta_t \sum_{i=1}^{2} \langle q_i^{\mu,t}, \pi_i^{\mu,r} - \pi_i^t \rangle$$

$$= \mathrm{KL}(\pi^t, \pi^{t+1}) + \eta_t \sum_{i=1}^{2} \sum_{a \in A_i} \left(q_i^{\pi^t}(a) + \frac{\mu}{\pi_i^t(a)} \left(r_i(a) - \pi_i^t(a) \right) \right) \left(\pi_i^t(a) - \pi_i^{\mu,r}(a) \right)$$

$$= \mathrm{KL}(\pi^t, \pi^{t+1}) + \eta_t \sum_{i=1}^{2} \sum_{a \in A_i} \left(\pi_i^t(a) - \pi_i^{\mu,r}(a) \right) \left(q_i^{\pi^t}(a) + \mu \frac{r_i(a)}{\pi_i^t(a)} \right)$$

$$=\mathrm{KL}(\pi^t, \pi^{t+1}) + \eta_t \sum_{i=1}^{2} \left(v_i^{\pi^t} - v_i^{\pi_i^{\mu,r}, \pi_{-i}^t} + \mu - \mu \sum_{a \in A_i} r_i(a) \frac{\pi_i^{\mu,r}(a)}{\pi_i^t(a)} \right)$$

$$=\mathrm{KL}(\pi^t, \pi^{t+1}) + \eta_t \sum_{i=1}^{2} \left(-v_i^{\pi_i^{\mu,r}, \pi_{-i}^t} + \mu - \mu \sum_{a \in A_i} r_i(a) \frac{\pi_i^{\mu,r}(a)}{\pi_i^t(a)} \right)$$

$$=\mathrm{KL}(\pi^t, \pi^{t+1}) + \eta_t \sum_{i=1}^{2} \left(v_i^{\pi_i^t, \pi_{-i}^{\mu,r}} + \mu - \mu \sum_{a \in A_i} r_i(a) \frac{\pi_i^{\mu,r}(a)}{\pi_i^t(a)} \right),$$

where the seventh equality follows from $\sum_{i=1}^{2} v_i^{\pi^t} = 0$, and the last equality follows from $-v_1^{\pi_1^{\mu,r}, \pi_2^t} = v_2^{\pi_1^{\mu,r}, \pi_2^t}$ and $-v_2^{\pi_1^t, \pi_2^{\mu,r}} = v_1^{\pi_1^t, \pi_2^{\mu,r}}$ by the definition of two-player zero-sum games. $\qquad\square$

B.5 Proof of Lemma B.3

Proof of Lemma B.3. From the definition of the Kullback-Leibler divergence, we have:

$$\mathrm{KL}(\pi, \pi^t) = \sum_{i=1}^{2} \mathrm{KL}(\pi_i, \pi_i^t) = \sum_{i=1}^{2} \sum_{a \in A_i} \pi_i(a) \ln \frac{\pi_i(a)}{\pi_i^t(a)}$$

$$= \sum_{i=1}^{2} \left(\sum_{a \in A_i} \left(\pi_i^t(a) - \pi_i(a) \right) \ln \pi_i^t(a) - \sum_{a \in A_i} \pi_i^t(a) \ln \pi_i^t(a) + \sum_{a \in A_i} \pi_i(a) \ln \pi_i(a) \right)$$

$$= \sum_{i=1}^{2} \left(\sum_{a \in A_i} \left(\pi_i^t(a) - \pi_i(a) \right) \ln \pi_i^t(a) - \psi_i(\pi_i^t) + \psi_i(\pi_i) \right). \tag{11}$$

Here, the update rule (1) is equivalent to:

$$\pi_i^t(a) = \frac{\exp\left(\sum_{s=1}^{t-1} \eta_s q_i^{\mu,s}(a) \right)}{\sum_{a' \in A_i} \exp\left(\sum_{s=1}^{t-1} \eta_s q_i^{\mu,s}(a') \right)},$$

and then we have:

$$\sum_{a \in A_i} \left(\pi_i^t(a) - \pi_i(a) \right) \ln \pi_i^t(a) = \sum_{a \in A_i} \left(\pi_i^t(a) - \pi_i(a) \right) \left(\sum_{s=1}^{t-1} \eta_s q_i^{\mu,s}(a) - \ln \left(\sum_{a' \in A_i} \exp \left(\sum_{s=1}^{t-1} \eta_s q_i^{\mu,s}(a') \right) \right) \right)$$

$$= \sum_{a \in A_i} \left(\pi_i^t(a) - \pi_i(a) \right) \sum_{s=1}^{t-1} \eta_s q_i^{\mu,s}(a)$$

$$= \left\langle \sum_{s=1}^{t-1} \eta_s q_i^{\mu,s}, \pi_i^t \right\rangle - \left\langle \sum_{s=1}^{t-1} \eta_s q_i^{\mu,s}, \pi_i \right\rangle. \tag{12}$$

By combining (11) and (12), we get:

$$\mathrm{KL}(\pi, \pi^t) = \sum_{i=1}^{2} \left(\left\langle \sum_{s=1}^{t-1} \eta_s q_i^{\mu,s}, \pi_i^t \right\rangle - \psi_i(\pi_i^t) - \left\langle \sum_{s=1}^{t-1} \eta_s q_i^{\mu,s}, \pi_i \right\rangle + \psi_i(\pi_i) \right).$$

$\qquad\square$

C PROOF OF COROLLARY 5.2

Proof of Corollary 5.2. From the definition of exploitability, we have:

$$
\begin{aligned}
\mathrm{explt}(\pi^t) &= \sum_{i=1}^{2} \max_{\tilde{\pi}_i \in \Delta(A_i)} v_i^{\tilde{\pi}_i, \pi_{-i}^t} \\
&= \sum_{i=1}^{2} \left(\max_{\tilde{\pi}_i \in \Delta(A_i)} v_i^{\tilde{\pi}_i, \pi_{-i}^{\mu,r}} + \max_{\tilde{\pi}_i \in \Delta(A_i)} v_i^{\tilde{\pi}_i, \pi_{-i}^t} - \max_{\tilde{\pi}_i \in \Delta(A_i)} v_i^{\tilde{\pi}_i, \pi_{-i}^{\mu,r}} \right) \\
&= \mathrm{explt}(\pi^{\mu,r}) + \sum_{i=1}^{2} \left(\max_{\tilde{\pi}_i \in \Delta(A_i)} v_i^{\tilde{\pi}_i, \pi_{-i}^t} - \max_{\tilde{\pi}_i \in \Delta(A_i)} v_i^{\tilde{\pi}_i, \pi_{-i}^{\mu,r}} \right) \\
&\leq \mathrm{explt}(\pi^{\mu,r}) + \sum_{i=1}^{2} \left(\max_{\tilde{\pi}_i \in \Delta(A_i)} \left(v_i^{\tilde{\pi}_i, \pi_{-i}^t} - v_i^{\tilde{\pi}_i, \pi_{-i}^{\mu,r}} \right) \right) \\
&\leq \mathrm{explt}(\pi^{\mu,r}) + \sum_{i=1}^{2} \left(\| \pi_i^{\mu,r} - \pi_i^t \|_1 \max_{\tilde{\pi}_{-i} \in \Delta(A_{-i})} \| q_i^{\pi_i^t, \tilde{\pi}_{-i}} \|_\infty \right) \\
&\leq \mathrm{explt}(\pi^{\mu,r}) + \sum_{i=1}^{2} \left(u_{\max} \sqrt{2\mathrm{KL}(\pi_i^{\mu,r}, \pi_i^t)} \right) \\
&\leq \mathrm{explt}(\pi^{\mu,r}) + u_{\max} \sqrt{2} \sqrt{2 \sum_{i=1}^{2} \mathrm{KL}(\pi_i^{\mu,r}, \pi_i^t)} \\
&= \mathrm{explt}(\pi^{\mu,r}) + 2 u_{\max} \sqrt{\mathrm{KL}(\pi^{\mu,r}, \pi^t)}, \tag{13}
\end{aligned}
$$

where the second inequality follows from Hölder's inequality, the third inequality follows from Pinsker's inequality (Tsybakov, 2009), and the fourth inequality follows from $\sqrt{a} + \sqrt{b} \leq \sqrt{2(a+b)}$ for $a, b > 0$. By combining (13) and Theorem 5.1, we have:

$$
\mathrm{explt}(\pi^t) \leq \mathrm{explt}(\pi^{\mu,r}) + 2 u_{\max} \sqrt{\mathrm{KL}(\pi^{\mu,r}, \pi^0)} (1 - C_2)^{\frac{t}{2}}. \tag{14}
$$

Moreover, from Lemma 3.5 of Bauer et al. (2019), a stationary point $\pi^{\mu,r}$ of (RMD) satisfies that for all $i \in \{1, 2\}$ and $a_i \in A_i$, $q_i^{\pi^{\mu,r}}(a_i) - v_i^{\pi^{\mu,r}} \leq \mu$. Therefore, the term of $\mathrm{exploit}(\pi^{\mu,r})$ can be bounded as:

$$
\begin{aligned}
\mathrm{explt}(\pi^{\mu,r}) &= \sum_{i=1}^{2} \max_{\tilde{\pi}_i \in \Delta(A_i)} v_i^{\tilde{\pi}_i, \pi_{-i}^{\mu,r}} \\
&= \sum_{i=1}^{2} \left(\max_{\tilde{\pi}_i \in \Delta(A_i)} v_i^{\tilde{\pi}_i, \pi_{-i}^{\mu,r}} - v_i^{\pi^{\mu,r}} \right) \\
&= \sum_{i=1}^{2} \left(\max_{a_i \in A_i} q_i^{\pi^{\mu,r}}(a_i) - v_i^{\pi^{\mu,r}} \right) \leq 2\mu, \tag{15}
\end{aligned}
$$

where the first equality follows from $\sum_{i=1}^{2} v_i^{\pi^{\mu,r}} = 0$ by the definition of zero-sum games. By combining (14) and (15), we have:

$$
\mathrm{explt}(\pi^t) \leq 2\mu + 2 u_{\max} \sqrt{\mathrm{KL}(\pi^{\mu,r}, \pi^t)},
$$

This concludes the statement. $\qquad\square$

D PROOFS FOR THEOREM 5.6

D.1 Proof of Theorem 5.6

In preparation for the proof, we first define the notion of *approximate Robbins-Monro algorithms* (Robbins and Monro, 1951; Benaïm, 1999).

Definition D.1. *The stochastic approximation algorithm*

$$z(t+1) = z(t) + \eta_t(F(z(t)) + U_t + \beta_t)$$

is refer to as an approximate Robbins-Monro algorithm if the following conditions are satisfied.

- $F : \mathbb{R}^m \to \mathbb{R}^m$ *is a continuous function*
- $(U_t)_{t \geq 1}$ *s.t.* $U_n \in \mathbb{R}^m, \forall n \in \mathbb{N}$ *is a martingale difference noise*
- $(\eta_t)_{t \geq 1}$ *is a given sequence of numbers such that* $\sum_{t=1}^{\infty} \eta_t = \infty$ *and* $\lim_{t \to \infty} \eta_t = 0$
- $\lim_{t \to \infty} \beta_t = 0$ *almost surely*

We provide the definition of the asymptotic pseudo-trajectory.

Definition D.2 ((Benaïm and Hirsch, 1996)). *A flow ϕ on a metric space (M, d) is a continuous mapping*

$$\phi : \mathbb{R} \times M \mapsto M, \qquad (t, x) \mapsto \phi_t(x) \tag{16}$$

such that $\phi_0(x) = x$ and $\phi_{t+\alpha} = \phi_t \circ \phi_\alpha$ for all $t, \alpha \in \mathbb{R}$. For a metric space (M, d), a continuous function $X : \mathbb{R} \mapsto M$ is an asymptotic pseudo trajectory for ϕ if

$$\lim_{t \to \infty} \sup_{s \in [0,T]} d(X(t+s), \phi_s(X(t))) = 0, \tag{17}$$

for every $T > 0$.

For each $i \in \{1, 2\}$, let we define the logit function $g_i : \mathbb{R}^{|A_i|} \mapsto \Delta(A_i)$ as

$$g_i(z_i) = \left(\frac{\exp(z_i(a_i))}{\sum_{a_i' \in A_i} \exp(z_i(a_i'))} \right)_{a_i \in A_i}. \tag{18}$$

We write $g_i(z_i)(a_i) \in \mathbb{R}$ be the a_i-th element of $g_i(z_i)$ and $\nabla g_i(z_i)(a_i) \in \mathbb{R}^{|A_i|}$ be the gradient vector of $g_i(z_i)(a_i)$, respectively. As a first result, we prove that the dynamics of the strategy $\{\pi_i^t\}$ updated by M2WU is an asymptotic pseudo trajectory of a continuous dynamics.

Lemma D.3. *Suppose that the sequence $\{\eta_t\}_{t \geq 1}$ satisfy $\eta_t \propto t^{-\lambda}$ for some $\lambda \in (1/\kappa, 1]$, where κ is a constant defined in Assumption 5.5 (ii). Then, for each $i \in \{1, 2\}$, the sequence of strategies $\{\pi_i^t\}_{t \geq 1}$ updated by M2WU is an asymptotic pseudo trajectory for the replicator mutator dynamics:*

$$\frac{d}{dt} \pi_i^t(a_i) = \pi_i^t(a_i) \left(q_i^{\pi^t}(a_i) - v_i^{\pi^t} \right) + \mu \left(r_i(a_i) - \pi_i^t(a_i) \right). \tag{RMD}$$

We present the proof of Lemma D.3 in Appendix D.2. Furthermore, we have the following exponential convergence result to the stationary point $\pi_i^{\mu,r}$ in the noiseless continuous time setting in Abe et al. (2022).

Theorem D.4 (Theorem 5.2. of Abe et al. (2022)). *Let $\pi_i^{\mu,r} \in \Delta(A_i)$ be a stationary point of (RMD) for all $i \in \{1, 2\}$. Then, for all $\mu > 0$, the continuous-time dynamics π^t updated by (RMD) satisfies the following:*

$$\frac{d}{dt} \text{KL}(\pi^{\mu,r}, \pi^t) = -\mu \sum_{i=1}^{2} \sum_{a_i \in A_i} r_i(a_i) \left(\sqrt{\frac{\pi_i^t(a_i)}{\pi_i^{\mu,r}(a_i)}} - \sqrt{\frac{\pi_i^{\mu,r}(a_i)}{\pi_i^t(a_i)}} \right)^2.$$

Furthermore, π^t satisfies that:

$$\frac{d}{dt} \text{KL}(\pi^{\mu,r}, \pi^t) \leq -\mu \xi \text{KL}(\pi^{\mu,r}, \pi^t),$$

where $\xi = \min_{i \in \{1,2\}, a_i \in A_i} \frac{c_i(a_i)}{\pi_i^{\mu,r}(a_i)}$.

It can be observed that the function $\text{KL}(\pi^{\mu,r}, \cdot)$ is a strict Lyapunov function. Furthermore, the stationary point of (RMD) is unique (Abe et al., 2022). Therefore, from Corollary 6.6 of Benaïm (1999), π_i^t updated by M2WU converges to $\pi_i^{\mu,r}$, almost surely. $\qquad \square$

D.2 Proof of Lemma D.3

For each $i \in \{1, 2\}$, for any $a_i, a_i', \tilde{a}_i \in A_i$,

$$\frac{\partial}{\partial z_i(a_i')}(g_i(z_i)(a_i)) = g_i(z_i)(a_i)(\mathbb{1}_{\{a_i = a_i'\}} - g_i(z_i)(a_i'))$$

and

$$\frac{\partial^2}{\partial z_i(\tilde{a}_i)\partial z_i(a_i')}(g_i(z_i)(a_i))$$
$$= g_i(z_i)(a_i)\left(\mathbb{1}_{\{a_i = a_i' = \tilde{a}_i\}} - g_i(z_i)(\tilde{a}_i)\mathbb{1}_{\{a_i = a_i'\}} - g_i(z_i)(a_i')\left(\mathbb{1}_{\{a_i = \tilde{a}_i\}} + \mathbb{1}_{\{a_i' = \tilde{a}_i\}} - 2g_i(z_i)(\tilde{a}_i)\right)\right). \qquad (19)$$

We write

$$\hat{q}_i^{\mu,t} = \left(\hat{q}_i^{\pi^t}(a_i) + \frac{\mu}{\pi_i^t(a_i)}(r_i(a_i) - \pi_i^t(a_i))\right)_{a_i \in A_i}$$

$$q_i^{\mu,t} = \left(q_i^{\pi^t}(a_i) + \frac{\mu}{\pi_i^t(a_i)}(r_i(a_i) - \pi_i^t(a_i))\right)_{a_i \in A_i}$$

$$\hat{\phi}_t = \frac{\eta_t}{2}(\hat{q}_i^{\mu,t})^{\top}\text{Hess}(g_i(\zeta)(a_i))\hat{q}_i^{\mu,t}.$$

By Taylor's theorem, we get the following computations.

$$\pi_i^{t+1}(a_i) = g_i(z_i^{t+1})(a_i)$$
$$= g_i(z_i^t + \eta_t \hat{q}_i^{\mu,t})(a_i)$$
$$= g_i(z_i^t)(a_i) + \eta_t\left((\nabla g_i(z_i^t)(a_i))^{\top}\hat{q}_i^{\mu,t} + \frac{\eta_t}{2}(\hat{q}_i^{\mu,t})^{\top}\text{Hess}(g_i(\zeta)(a_i))\hat{q}_i^{\mu,t}\right)$$
$$= g_i(z_i^t)(a_i) + \eta_t\left((\nabla g_i(z_i^t)(a_i))^{\top}(q_i^{\mu,t} + \hat{q}_i^{\mu,t} - q_i^{\mu,t}) + \hat{\phi}_t\right),$$

where ζ is a point between z_i^t and z_i^{t+1}.

$$(\nabla g_i(z_i^t)(a_i))^{\top}q_i^{\mu,t} = \sum_{a_i' \in A_i} g_i(z_i^t)(a_i)(\mathbb{1}_{\{a_i = a_i'\}} - g_i(z_i^t)(a_i'))\left(q_i^{\pi^t}(a_i') + \frac{\mu}{\pi_i^t(a_i')}(r_i(a_i') - \pi_i^t(a_i'))\right)$$

$$= g_i(z_i^t)(a_i)\left(q_i^{\mu,t}(a_i) - \sum_{a_i' \in A_i} g_i(z_i^t)(a_i')q_i^{\mu,t}(a_i')\right)$$

$$= \pi_i^t(a_i)\left(q_i^{\mu,t}(a_i) - \sum_{a_i' \in A_i} \pi_i^t(a_i')q_i^{\mu,t}(a_i')\right).$$

We can write the dynamics of $\pi_i^t(a_i)$ as follows

$$\pi_i^{t+1}(a_i) = \pi_i^t(a_i) + \eta_t(F(\pi_i^t) + U_t + \hat{\phi}_t),$$

where $F(\pi_i^t) = \pi_i^t(a_i)\left(q_i^{\mu,t}(a_i) - \sum_{a_i' \in A_i} \pi_i^t(a_i')q_i^{\mu,t}(a_i')\right)$ is a continuous function and $U_t = (\nabla g_i(z_i^t)(a_i))^{\top}(\hat{q}_i^{\mu,t} - q_i^{\mu,t})$ is a martingale difference sequence from Assumption 5.5 and the bound on the utility function. Note that from the assumption on $(\eta_t)_{t \geq 1}$, $\lim_{t \to \infty} \eta_t = 0$ and $\sum_{t=1}^{\infty} \eta_t = \infty$. From the form of (19), the elements of $\text{Hess}(g_i(\zeta)(a_i))$ is bounded by some constant. Thus, the limit of $\hat{\phi}_t$ is determined by the term $\|\hat{q}_i^{\mu,t}\|_2^2$. Let $\mathcal{E}_t$ be an event such that $\|\hat{q}_i^{\mu,t}\|_2^2 \geq t^p$ with $p < 1/\kappa$, where κ is in Assumption 5.5 (ii) and p is a value satisfies $\eta_t = o(t^{-p})$ (note that $\eta_t \propto t^{-\kappa}$ and $1/\kappa < \lambda \leq 1$). Using Assumption 5.5, the assumption that $\pi_i^t(a) > D > 0$, and the boundedness of the utility function,

$$\sum_{t=1}^{\infty} \mathbb{P}(\mathcal{E}_t) = \sum_{t=1}^{\infty} \mathbb{P}(\|\hat{q}_i^{\mu,t}\|_2^2 \geq t^p \mid \mathcal{F}_{t-1}) = \sum_{t=1}^{\infty} \mathcal{O}\left(t^{-\kappa p}\right) < \infty.$$

From the Borel–Cantelli lemma, $\mathbb{P}(\cap_{t=1}^\infty \cup_{s\geq t}^\infty \mathcal{E}_s) = 0$. Therefore, the event $\mathcal{E}_t$ occurs only for a finite number of t, almost surely. Thus, as $\eta_t t^p \propto t^{-\lambda+p} = o(1)$,

$$\hat{\phi}_t = \mathcal{O}(\eta_t \|\hat{q}_i^{\mu,t}\|_2^2) = \mathcal{O}(\eta_t t^p) = o(1),$$

almost surely. Therefore, from Definition D.1, the update of $\{\pi_i^t\}_{t\geq 1}$ is an approximate Robbins-Monro algorithm. From Proposition 4.2 of Benaïm (1999), $\{\pi_i^t\}_{t\geq 1}$ is an asymptotic pseudo-trajectory of the replicator mutator dynamics. $\qquad\square$

E PROOFS FOR THEOREM 6.1

E.1 Proof of Theorem 6.1

Proof of Theorem 6.1. From Lemma 6.2, sequence $\{\min_{\pi^*\in\Pi^*} \mathrm{KL}(\pi^*, r^k)\}_{k\geq 0}$ is a monotonically decreasing sequence and is bounded from below by zero. Hence, $\{\min_{\pi^*\in\Pi^*} \mathrm{KL}(\pi^*, r^k)\}_{k\geq 0}$ converges to some $b \geq 0$. We show that $b = 0$ by a contradiction argument.

Suppose $b > 0$ and let us define $B = \min_{\pi^*\in\Pi^*} \mathrm{KL}(\pi^*, r^0)$. Since $\min_{\pi^*\in\Pi^*} \mathrm{KL}(\pi^*, r^k)$ monotonically decreases, r^k is in the set $\Omega_{b,B} = \{r \in \prod_{i=1}^2 \Delta^\circ(A_i) \mid b \leq \min_{\pi^*\in\Pi^*} \mathrm{KL}(\pi^*, r) \leq B\}$ for all $k \geq 0$. Since $\min_{\pi^*\in\Pi^*} \mathrm{KL}(\pi^*, r)$ is a continuous function on $\prod_{i=1}^2 \Delta^\circ(A_i)$, the preimage $\Omega_{b,B}$ of the closed set $[b, B]$ is also closed. Furthermore, since $\prod_{i=1}^2 \Delta^\circ(A_i)$ is a bounded set, $\Omega_{b,B}$ is a bounded set. Thus, $\Omega_{b,B}$ is a compact set.

From Lemma 6.3, $\min_{\pi^*\in\Pi^*} \mathrm{KL}(\pi^*, F(r)) - \min_{\pi^*\in\Pi^*} \mathrm{KL}(\pi^*, r)$ is also a continuous function. Since a continuous function has a maximum over a compact set, the maximum $\delta = \max_{r\in\Omega_{b,B}} \{\min_{\pi^*\in\Pi^*} \mathrm{KL}(\pi^*, F(r)) - \min_{\pi^*\in\Pi^*} \mathrm{KL}(\pi^*, r)\}$ exists. From Lemma 6.2 and the assumption that $b > 0$, we have $\delta < 0$. It follows that:

$$\min_{\pi^*\in\Pi^*} \mathrm{KL}(\pi^*, r^k) = \min_{\pi^*\in\Pi^*} \mathrm{KL}(\pi^*, r^0) + \sum_{l=0}^{k-1} \left(\min_{\pi^*\in\Pi^*} \mathrm{KL}(\pi^*, r^{l+1}) - \min_{\pi^*\in\Pi^*} \mathrm{KL}(\pi^*, r^l) \right)$$

$$\leq B + \sum_{l=0}^{k-1} \delta = B + k\delta.$$

This implies that $\min_{\pi^*\in\Pi^*} \mathrm{KL}(\pi^*, r^k) < 0$ for $k > \frac{B}{-\delta}$, which is a contradiction because $\min_{\pi^*\in\Pi^*} \mathrm{KL}(\pi^*, r) \geq 0$. Therefore, the sequence of $\min_{\pi^*\in\Pi^*} \mathrm{KL}(\pi^*, r^k)$ converges to 0, and r^k converges to some strategy profile in Π^*. $\qquad\square$

E.2 Proof of Lemma 6.2

Proof of Lemma 6.2. First, we prove the first statement of the lemma using following two lemmas:

Lemma E.1. *Let $\pi^{\mu,r}$ be a stationary point of (RMD) with the reference strategy profile r. Assuming that $r \neq \pi^{\mu,r}$, for any Nash equilibrium π^* of the original game, we have:*

$$\mathrm{KL}(\pi^*, \pi^{\mu,r}) - \mathrm{KL}(\pi^*, r) < 0.$$

Lemma E.2. *Let $\pi^{\mu,r}$ be a stationary point of (RMD) with the reference strategy profile r. If $r = \pi^{\mu,r}$, then r is a Nash equilibrium of the original game.*

From Lemma E.2, when $r \in \prod_{i=1}^2 \Delta^\circ(A_i) \setminus \Pi^*$, $r \neq \pi^{\mu,r}$ always holds. Let us define $\pi^\star = \arg\min_{\pi^*\in\Pi^*} \mathrm{KL}(\pi^*, r)$. From Lemma E.1, if $r \neq \pi^{\mu,r}$ we have:

$$\min_{\pi^*\in\Pi^*} \mathrm{KL}(\pi^*, r) = \mathrm{KL}(\pi^\star, r) > \mathrm{KL}(\pi^\star, \pi^{\mu,r}) \geq \min_{\pi^*\in\Pi^*} \mathrm{KL}(\pi^*, \pi^{\mu,r}).$$

Therefore, if $r \in \prod_{i=1}^2 \Delta^\circ(A_i) \setminus \Pi^*$ then $\min_{\pi^*\in\Pi^*} \mathrm{KL}(\pi^*, \pi^{\mu,r}) < \min_{\pi^*\in\Pi^*} \mathrm{KL}(\pi^*, r)$.

Next, we prove the second statement of the lemma. Assume that $r \in \Pi^*$ implies that $\pi^{\mu,r} \neq r$. In this case, we can apply Lemma E.1, so we have for all $\pi^* \in \Pi^*$, $\mathrm{KL}(\pi^*, \pi^{\mu,r}) < \mathrm{KL}(\pi^*, r)$. On the other hand, since $r \in \Pi^*$, there exists a Nash equilibrium $\pi^\star$ such that $\mathrm{KL}(\pi^\star, r) = 0$. Thus, we have $\mathrm{KL}(\pi^\star, \pi^{\mu,r}) < \mathrm{KL}(\pi^\star, r) = 0$, which contradicts that $\mathrm{KL}(\pi^\star, \pi^{\mu,r}) \geq 0$. Therefore, if $r \in \Pi^*$ then $\pi^{\mu,r} = r$. $\qquad\square$

E.3 Proof of Lemma 6.3

Proof of Lemma 6.3. For a given $r \in \prod_{i=1}^{2} \Delta^{\circ}(A_i)$, let us consider that π^t follows the following (RMD) dynamics with a reference strategy profile $r \in \prod_{i=1}^{2} \Delta^{\circ}(A_i)$:

$$\frac{d}{dt}\pi_i^t(a_i) = \pi_i^t(a_i)\left(q_i^{\pi^t}(a_i) - v_i^{\pi^t}\right) + \mu\left(r_i(a_i) - \pi_i^t(a_i)\right).$$

Therefore, we have for a given $r' \in \prod_{i=1}^{2} \Delta^{\circ}(A_i)$ and the associated stationary point $\pi^{\mu,r'}$:

$$\frac{d}{dt}\mathrm{KL}(\pi^{\mu,r'}, \pi^t)$$

$$= \sum_{i=1}^{2}\sum_{a\in A_i}\pi_i^{\mu,r'}(a)\frac{d}{dt}\ln\left(\frac{\pi_i^{\mu,r'}(a)}{\pi_i^t(a)}\right)$$

$$= -\sum_{i=1}^{2}\sum_{a\in A_i}\pi_i^{\mu,r'}(a)\frac{d}{dt}\ln\pi_i^t(a)$$

$$= -\sum_{i=1}^{2}\sum_{a\in A_i}\frac{\pi_i^{\mu,r'}(a)}{\pi_i^t(a)}\frac{d}{dt}\pi_i^t(a)$$

$$= -\sum_{i=1}^{2}\sum_{a\in A_i}\frac{\pi_i^{\mu,r'}(a)}{\pi_i^t(a)}\left(\pi_i^t(a_i)\left(q_i^{\pi^t}(a_i) - v_i^{\pi^t}\right) + \mu\left(r_i(a_i) - \pi_i^t(a_i)\right)\right)$$

$$= \sum_{i=1}^{2}\sum_{a\in A_i}\left(\pi_i^{\mu,r'}(a)\left(v_i^{\pi^t} - q_i^{\pi^t}(a_i)\right) - \frac{\mu}{\pi_i^t(a)}\left(r_i(a_i) - \pi_i^t(a_i)\right)\pi_i^{\mu,r'}(a)\right)$$

$$= \sum_{i=1}^{2}\sum_{a\in A_i}\left(\left(\pi_i^t(a) - \pi_i^{\mu,r'}(a)\right)q_i^{\pi^t}(a) - \frac{\mu}{\pi_i^t(a)}\left(r_i(a_i) - \pi_i^t(a_i)\right)\pi_i^{\mu,r'}(a) + \frac{\mu}{\pi_i^t(a)}\left(r_i(a_i) - \pi_i^t(a_i)\right)\pi_i^t(a)\right)$$

$$= \sum_{i=1}^{2}\sum_{a\in A_i}\left(q_i^{\pi^t}(a) + \frac{\mu}{\pi_i^t(a)}\left(r_i(a) - \pi_i^t(a)\right)\right)\left(\pi_i^t(a) - \pi_i^{\mu,r'}(a)\right)$$

$$= \sum_{i=1}^{2}\sum_{a\in A_i}\left(q_i^{\pi^t}(a) + \frac{\mu}{\pi_i^t(a)}\left(r_i'(a) - \pi_i^t(a)\right)\right)\left(\pi_i^t(a) - \pi_i^{\mu,r'}(a)\right)$$

$$+ \sum_{i=1}^{2}\sum_{a\in A_i}\left(\frac{\mu}{\pi_i^t(a)}\left(r_i(a) - \pi_i^t(a)\right) - \frac{\mu}{\pi_i^t(a)}\left(r_i'(a) - \pi_i^t(a)\right)\right)\left(\pi_i^t(a) - \pi_i^{\mu,r'}(a)\right), \tag{20}$$

where the sixth equality follows from $v_i^{\pi^t} = \sum_{a\in A_i}\pi_i^t(a)q_i^{\pi^t}(a)$ and $\sum_{a\in A_i}\frac{\mu}{\pi_i^t(a)}\left(r_i(a_i) - \pi_i^t(a_i)\right)\pi_i^t(a) = \mu\sum_{a\in A_i}\left(r_i(a_i) - \pi_i^t(a_i)\right) = 0$.

The first term of (20) can be written as:

$$\sum_{i=1}^{2}\sum_{a\in A_i}\left(q_i^{\pi^t}(a) + \frac{\mu}{\pi_i^t(a)}\left(r_i'(a) - \pi_i^t(a)\right)\right)\left(\pi_i^t(a) - \pi_i^{\mu,r'}(a)\right)$$

$$= \sum_{i=1}^{2}\sum_{a\in A_i}\left(\pi_i^t(a) - \pi_i^{\mu,r'}(a)\right)\left(q_i^{\pi^t}(a) + \mu\left(\frac{r_i'(a)}{\pi_i^t(a)} - 1\right)\right)$$

$$= \sum_{i=1}^{2}\sum_{a\in A_i}\left(\pi_i^t(a) - \pi_i^{\mu,r'}(a)\right)\left(q_i^{\pi^t}(a) + \mu\frac{r_i'(a)}{\pi_i^t(a)}\right)$$

$$= \sum_{i=1}^{2}\left(v_i^{\pi^t} - v_i^{\pi_i^{\mu,r'},\pi_{-i}^t} + \mu - \mu\sum_{a\in A_i}r_i'(a)\frac{\pi_i^{\mu,r'}(a)}{\pi_i^t(a)}\right)$$

$$= \sum_{i=1}^{2} \left(-v_i^{\pi_i^{\mu,r'}, \pi_{-i}^t} + \mu - \mu \sum_{a \in A_i} r'_i(a) \frac{\pi_i^{\mu,r'}(a)}{\pi_i^t(a)} \right)$$

$$= \sum_{i=1}^{2} \left(v_i^{\pi_i^t, \pi_{-i}^{\mu,r'}} + \mu - \mu \sum_{a \in A_i} r'_i(a) \frac{\pi_i^{\mu,r'}(a)}{\pi_i^t(a)} \right),$$

where the fourth equality follows from $\sum_{i=1}^{2} v_i^{\pi^t} = 0$, and the last equality follows from $-v_1^{\pi_1^{\mu,r'}, \pi_2^t} = v_2^{\pi_1^{\mu,r'}, \pi_2^t}$ and $-v_2^{\pi_1^t, \pi_2^{\mu,r'}} = v_1^{\pi_1^t, \pi_2^{\mu,r'}}$ by the definition of two-player zero-sum games. Here, from Lemma B.2, for all $i \in \{1, 2\}$:

$$v_i^{\pi_i^t, \pi_{-i}^{\mu,r'}} = v_i^{\pi^{\mu,r'}} + \mu - \mu \sum_{a \in A_i} r'_i(a) \frac{\pi_i^t(a)}{\pi_i^{\mu,r'}(a)},$$

and then we have:

$$\sum_{i=1}^{2} \sum_{a \in A_i} \left(q_i^{\pi^t}(a) + \frac{\mu}{\pi_i^t(a)} \left(r'_i(a) - \pi_i^t(a) \right) \right) \left(\pi_i^t(a) - \pi_i^{\mu,r'}(a) \right)$$

$$= \sum_{i=1}^{2} v_i^{\pi^{\mu,r'}} + 4\mu - \mu \sum_{i=1}^{2} \sum_{a \in A_i} r'_i(a) \left(\frac{\pi_i^t(a)}{\pi_i^{\mu,r'}(a)} + \frac{\pi_i^{\mu,r'}(a)}{\pi_i^t(a)} \right)$$

$$= 4\mu - \mu \sum_{i=1}^{2} \sum_{a \in A_i} r'_i(a) \left(\frac{\pi_i^t(a)}{\pi_i^{\mu,r'}(a)} + \frac{\pi_i^{\mu,r'}(a)}{\pi_i^t(a)} \right)$$

$$= -\mu \sum_{i=1}^{2} \sum_{a \in A_i} r'_i(a) \left(\sqrt{\frac{\pi_i^t(a)}{\pi_i^{\mu,r'}(a)}} - \sqrt{\frac{\pi_i^{\mu,r'}(a)}{\pi_i^t(a)}} \right)^2, \tag{21}$$

where the second equality follows from $\sum_{i=1}^{2} v_i^{\pi^{\mu,r'}} = 0$ by the definition of zero-sum games.

On the other hand, the second term of (20) is written as:

$$\sum_{i=1}^{2} \sum_{a \in A_i} \left(\frac{\mu}{\pi_i^t(a)} \left(r_i(a) - \pi_i^t(a) \right) - \frac{\mu}{\pi_i^t(a)} \left(r'_i(a) - \pi_i^t(a) \right) \right) \left(\pi_i^t(a) - \pi_i^{\mu,r'}(a) \right)$$

$$= \mu \sum_{i=1}^{2} \sum_{a \in A_i} \frac{1}{\pi_i^t(a)} \left(r_i(a) - r'_i(a) \right) \left(\pi_i^t(a) - \pi_i^{\mu,r'}(a) \right) \leq \mu \sum_{i=1}^{2} \sum_{a \in A_i} \frac{1}{\pi_i^t(a)} \left| r_i(a) - r'_i(a) \right|. \tag{22}$$

Combining (20), (21), and (22), we can obtain:

$$\frac{d}{dt} \text{KL}(\pi^{\mu,r'}, \pi^t) \leq -\mu \sum_{i=1}^{2} \sum_{a \in A_i} r'_i(a) \left(\sqrt{\frac{\pi_i^t(a)}{\pi_i^{\mu,r'}(a)}} - \sqrt{\frac{\pi_i^{\mu,r'}(a)}{\pi_i^t(a)}} \right)^2 + \mu \sum_{i=1}^{2} \sum_{a \in A_i} \frac{1}{\pi_i^t(a)} \left| r_i(a) - r'_i(a) \right|.$$

Setting the start point as $\pi^0 = \pi^{\mu,r}$, we have for all $t \geq 0$, $\pi^t = \pi^{\mu,r}$. In this case, for all $t \geq 0$, we have $\frac{d}{dt} \text{KL}(\pi^{\mu,r'}, \pi^t) = 0$. Thus,

$$\sum_{i=1}^{2} \sum_{a \in A_i} r'_i(a) \left(\sqrt{\frac{\pi_i^{\mu,r}(a)}{\pi_i^{\mu,r'}(a)}} - \sqrt{\frac{\pi_i^{\mu,r'}(a)}{\pi_i^{\mu,r}(a)}} \right)^2 \leq \sum_{i=1}^{2} \sum_{a \in A_i} \frac{1}{\pi_i^{\mu,r}(a)} \left| r_i(a) - r'_i(a) \right|.$$

Here, since r_i is in interior of $\Delta(A_i)$, there exists $\nu_1 > 0$ such that $\forall i, \forall a \in A_i, r_i(a) > \nu_1$. Furthermore, from Lemma C.2 in Abe et al. (2022), $\pi_i^{\mu,r}$ is also in interior of $\Delta(A_i)$. Thus, there exists $\nu_2 > 0$ such that $\forall i, \forall a \in A_i, \pi_i^{\mu,r}(a) > \nu_2$. For a given $\varepsilon > 0$, let us define $\delta = \frac{\varepsilon^2 \nu_1 \nu_2}{4 + \varepsilon^2 \nu_2} \sqrt{\frac{1}{\sum_{i=1}^{2} |A_i|}}$. If $\|r' - r\|_2 < \delta$, then $\|r' - r\|_1 \leq \|r' - r\|_2 \sqrt{\sum_{i=1}^{2} |A_i|} < \frac{\varepsilon^2 \nu_1 \nu_2}{4 + \varepsilon^2 \nu_2}$.

Thus, $\forall i, \forall a \in A_i, r_i'(a) > \left(1 - \frac{\varepsilon^2 \nu_2}{4 + \varepsilon^2 \nu_2}\right) \nu_1 > 0$. So, if $\|r' - r\|_2 < \delta$, we have then:

$$
\begin{aligned}
\sum_{i=1}^{2} \sum_{a \in A_i} r_i'(a) \left(\sqrt{\frac{\pi_i^{\mu, r'}(a)}{\pi_i^{\mu, r}(a)}} - \sqrt{\frac{\pi_i^{\mu, r}(a)}{\pi_i^{\mu, r'}(a)}}\right)^2 &= \sum_{i=1}^{2} \sum_{a \in A_i} r_i'(a) \left(\frac{\pi_i^{\mu, r'}(a)}{\pi_i^{\mu, r}(a)} + \frac{\pi_i^{\mu, r}(a_i)}{\pi_i^{\mu, r'}(a)} - 2\right) \\
&= \sum_{i=1}^{2} \sum_{a \in A_i} \frac{r_i'(a)}{\pi_i^{\mu, r'}(a)} \frac{(\pi_i^{\mu, r}(a) - \pi_i^{\mu, r'}(a))^2}{\pi_i^{\mu, r}(a)} \\
&\geq \left(1 - \frac{\varepsilon^2 \nu_2}{4 + \varepsilon^2 \nu_2}\right) \nu_1 \sum_{i=1}^{2} \sum_{a \in A_i} \frac{(\pi_i^{\mu, r}(a) - \pi_i^{\mu, r'}(a))^2}{\pi_i^{\mu, r}(a)} \\
&\geq \left(1 - \frac{\varepsilon^2 \nu_2}{4 + \varepsilon^2 \nu_2}\right) \nu_1 \sum_{i=1}^{2} \ln\left(1 + \sum_{a \in A_i} \frac{(\pi_i^{\mu, r}(a) - \pi_i^{\mu, r'}(a))^2}{\pi_i^{\mu, r}(a)}\right) \\
&= \left(1 - \frac{\varepsilon^2 \nu_2}{4 + \varepsilon^2 \nu_2}\right) \nu_1 \sum_{i=1}^{2} \ln\left(\sum_{a \in A_i} \pi_i^{\mu, r'}(a) \frac{\pi_i^{\mu, r'}(a)}{\pi_i^{\mu, r}(a)}\right) \\
&\geq \left(1 - \frac{\varepsilon^2 \nu_2}{4 + \varepsilon^2 \nu_2}\right) \nu_1 \sum_{i=1}^{2} \sum_{a \in A_i} \pi_i^{\mu, r'}(a) \ln\left(\frac{\pi_i^{\mu, r'}(a)}{\pi_i^{\mu, r}(a)}\right) \\
&= \left(1 - \frac{\varepsilon^2 \nu_2}{4 + \varepsilon^2 \nu_2}\right) \nu_1 \sum_{i=1}^{2} \mathrm{KL}(\pi_i^{\mu, r'}, \pi_i^{\mu, r}),
\end{aligned}
$$

where the second inequality follow from $x \geq \ln(1 + x)$ for all $x > 0$, and the third inequality follows from the concavity of the $\ln(\cdot)$ function and Jensen's inequality for concave functions. Moreover,

$$
\sum_{i=1}^{2} \mathrm{KL}(\pi_i^{\mu, r'}, \pi_i^{\mu, r}) \geq \frac{1}{2} \sum_{i=1}^{2} \|\pi_i^{\mu, r'} - \pi_i^{\mu, r}\|_1^2 \geq \frac{1}{4} \|\pi^{\mu, r'} - \pi^{\mu, r}\|_1^2,
$$

where we use Pinsker's inequality (Tsybakov, 2009), and the fact that $\sum_{i=1}^{2} x_i^2 \geq \frac{1}{2} \left(\sum_{i=1}^{2} x_i\right)^2$ for $x_i \in \mathbb{R}$. Thus, we get:

$$
\left(1 - \frac{\varepsilon^2 \nu_2}{4 + \varepsilon^2 \nu_2}\right) \frac{\nu_1}{4} \|\pi^{\mu, r'} - \pi^{\mu, r}\|_1^2 \leq \sum_{i=1}^{2} \sum_{a \in A_i} \frac{1}{\pi_i^{\mu, r}(a)} |r_i(a) - r_i'(a)| < \frac{1}{\nu_2} \|r' - r\|_1.
$$

Therefore, if $\|r' - r\| < \delta$, we have then:

$$
\begin{aligned}
\|\pi^{\mu, r'} - \pi^{\mu, r}\|_2 \leq \|\pi^{\mu, r'} - \pi^{\mu, r}\|_1 &< \sqrt{\frac{4}{\left(1 - \frac{\varepsilon^2 \nu_2}{4 + \varepsilon^2 \nu_2}\right) \nu_1 \nu_2} \|r' - r\|_1} \\
&\leq \sqrt{\frac{4}{\left(1 - \frac{\varepsilon^2 \nu_2}{4 + \varepsilon^2 \nu_2}\right) \nu_1 \nu_2} \|r' - r\|_2 \sqrt{\sum_{i=1}^{2} |A_i|}} \\
&< \sqrt{\frac{4}{\left(1 - \frac{\varepsilon^2 \nu_2}{4 + \varepsilon^2 \nu_2}\right) \nu_1 \nu_2} \delta \sqrt{\sum_{i=1}^{2} |A_i|}} \\
&= \sqrt{\frac{4}{\left(1 - \frac{\varepsilon^2 \nu_2}{4 + \varepsilon^2 \nu_2}\right) \nu_1 \nu_2} \frac{\varepsilon^2 \nu_1 \nu_2}{4 + \varepsilon^2 \nu_2}} \\
&= \sqrt{\frac{4}{1 - \frac{\varepsilon^2 \nu_2}{4 + \varepsilon^2 \nu_2}} \frac{\varepsilon^2}{4 + \varepsilon^2 \nu_2}} = \varepsilon.
\end{aligned}
$$

Thus, for every $\varepsilon > 0$, there exists $\delta > 0$ such that for all $r' \in \prod_{i=1}^{2} \Delta^{\circ}(A_i)$, if $\|r' - r\|_2 < \delta$ then $\|\pi^{\mu,r'} - \pi^{\mu,r}\|_2 < \varepsilon$. Therefore, $F(\cdot)$ is a continuous function on $\prod_{i=1}^{2} \Delta^{\circ}(A_i)$. $\qquad\square$

E.4 Proof of Lemma E.1

Proof of Lemma E.1. First, we have:

$$
\begin{aligned}
\mathrm{KL}(\pi^*, \pi^{\mu,r}) - \mathrm{KL}(\pi^*, r) &= \sum_{i=1}^{2} \left(\sum_{a_i \in A_i} \pi_i^*(a_i) \ln \frac{\pi_i^*(a_i)}{\pi_i^{\mu,r}(a_i)} - \sum_{a_i \in A_i} \pi_i^*(a_i) \ln \frac{\pi_i^*(a_i)}{r_i(a_i)} \right) \\
&= \sum_{i=1}^{2} \sum_{a_i \in A_i} \pi_i^*(a_i) \ln \frac{r_i(a_i)}{\pi_i^{\mu,r}(a_i)} \\
&\leq 2 \ln \left(\frac{1}{2} \sum_{i=1}^{2} \sum_{a_i \in A_i} \pi_i^*(a_i) \frac{r_i(a_i)}{\pi_i^{\mu,r}(a_i)} \right),
\end{aligned}
$$

where the inequality follows from the concavity of the $\ln(\cdot)$ function and Jensen's inequality for concave functions. Since $\ln(\cdot)$ is strictly concave, the equality holds if and only if $r = \pi^{\mu,r}$. Therefore, from the assumption that $r \neq \pi^{\mu,r}$, we have:

$$
\mathrm{KL}(\pi^*, \pi^{\mu,r}) - \mathrm{KL}(\pi^*, r) < 2 \ln \left(\frac{1}{2} \sum_{i=1}^{2} \sum_{a_i \in A_i} \pi_i^*(a_i) \frac{r_i(a_i)}{\pi_i^{\mu,r}(a_i)} \right). \tag{23}
$$

Here, by using the ordinary differential equation (RMD), we have for all $i \in \{1, 2\}$ and $a_i \in A_i$:

$$
\pi_i^{\mu,r}(a_i) \left(q_i^{\pi^{\mu,r}}(a_i) - v_i^{\pi^{\mu,r}} \right) + \mu \left(r_i(a_i) - \pi_i^{\mu,r}(a_i) \right) = 0,
$$

and then:

$$
\frac{r_i(a_i)}{\pi_i^{\mu,r}(a_i)} = 1 - \frac{1}{\mu} \left(q_i^{\pi^{\mu,r}}(a_i) - v_i^{\pi^{\mu,r}} \right). \tag{24}
$$

Combining (23) and (24), we have:

$$
\begin{aligned}
\mathrm{KL}(\pi^*, \pi^{\mu,r}) - \mathrm{KL}(\pi^*, r) &< 2 \ln \left(\frac{1}{2} \sum_{i=1}^{2} \sum_{a_i \in A_i} \pi_i^*(a_i) \left(1 - \frac{1}{\mu} \left(q_i^{\pi^{\mu,r}}(a_i) - v_i^{\pi^{\mu,r}} \right) \right) \right) \\
&= 2 \ln \left(\frac{1}{2} \sum_{i=1}^{2} \left(1 - \frac{1}{\mu} \left(v_i^{\pi_i^*,\pi_{-i}^r} - v_i^{\pi^{\mu,r}} \right) \right) \right).
\end{aligned}
$$

Since $\sum_{a_i \in A_i} \pi_i^*(a_i) \frac{r_i(a_i)}{\pi_i^{\mu,r}(a_i)} > 0$, we have $1 - \frac{1}{\mu} \left(v_i^{\pi_i^*,\pi_{-i}^r} - v_i^{\pi^{\mu,r}} \right) > 0$. Also, since π^* is the Nash equilibrium, we get:

$$
\begin{aligned}
\sum_{i=1}^{2} \left(1 - \frac{1}{\mu} \left(v_i^{\pi_i^*,\pi_{-i}^r} - v_i^{\pi^{\mu,r}} \right) \right) &= 2 + \frac{1}{\mu} \sum_{i=1}^{2} \left(v_i^{\pi^{\mu,r}} - v_i^{\pi_i^*,\pi_{-i}^r} \right) \\
&= 2 + \frac{1}{\mu} \sum_{i=1}^{2} \left(-v_i^{\pi_i^*,\pi_{-i}^r} - v_i^{\pi^*} \right) \\
&= 2 + \frac{1}{\mu} \sum_{i=1}^{2} \left(v_i^{\pi_i^{\mu,r},\pi_{-i}^*} - v_i^{\pi^*} \right) \leq 2,
\end{aligned}
$$

where the second equality follows from $\sum_{i=1}^{2} v_i^{\pi^{\mu,r}} = 0$ and $\sum_{i=1}^{2} v_i^{\pi^*} = 0$, and the last equality follows from $-v_1^{\pi_1^*,\pi_2^r} = v_2^{\pi_1^*,\pi_2^r}$ and $-v_2^{\pi_1^r,\pi_2^*} = v_1^{\pi_1^r,\pi_2^*}$ by the definition of two-player zero-sum games. Thus, we get

$$
\mathrm{KL}(\pi^*, \pi^{\mu,r}) - \mathrm{KL}(\pi^*, r) < 2(\ln 1) \leq 0.
$$

$\qquad\square$

E.5 Proof of Lemma E.2

Proof of Lemma E.2. By using the ordinary differential equation (RMD), we have for all $i \in \{1, 2\}$ and $a_i \in A_i$:

$$\pi_i^{\mu,r}(a_i) \left(q_i^{\pi^{\mu,r}}(a_i) - v_i^{\pi^{\mu,r}} \right) + \mu \left(r_i(a_i) - \pi_i^{\mu,r}(a_i) \right) = 0.$$

Since $r = \pi^{\mu,r}$, we have for all $i \in \{1, 2\}$ and $a_i \in A_i$:

$$r_i(a_i) \left(q_i^r(a_i) - v_i^r \right) = 0.$$

From the definition of the reference strategy, we have $r_i(a_i) > 0$, and then $v_i^r = \max_{a_i \in A_i} q_i^r(a_i)$ for all $i \in \{1, 2\}$. Therefore, each player i's strategy r_i is a best response to the other player $-i$'s strategy r_{-i}. Thus, r is a Nash equilibrium of the original game. $\square$

F ADDITIONAL EXPERIMENTAL RESULTS WITH NOISE

F.1 Various Learning Rates

Figures 6 and 7 show the numerical results with varying learning rates $\eta \in \{0.1, 0.05, 0.01, 0.005, 0.001\}$ in BRPS and M-Ne with noisy-information feedback. We observe that M2WU exhibits lower exploitability than MWU and OMWU for all η. Also, we can see that OMWU does not converge for all learning rates in both games (See also Remark 5.7).

(a) $\eta = 0.1$

(b) $\eta = 0.05$

(c) $\eta = 0.01$

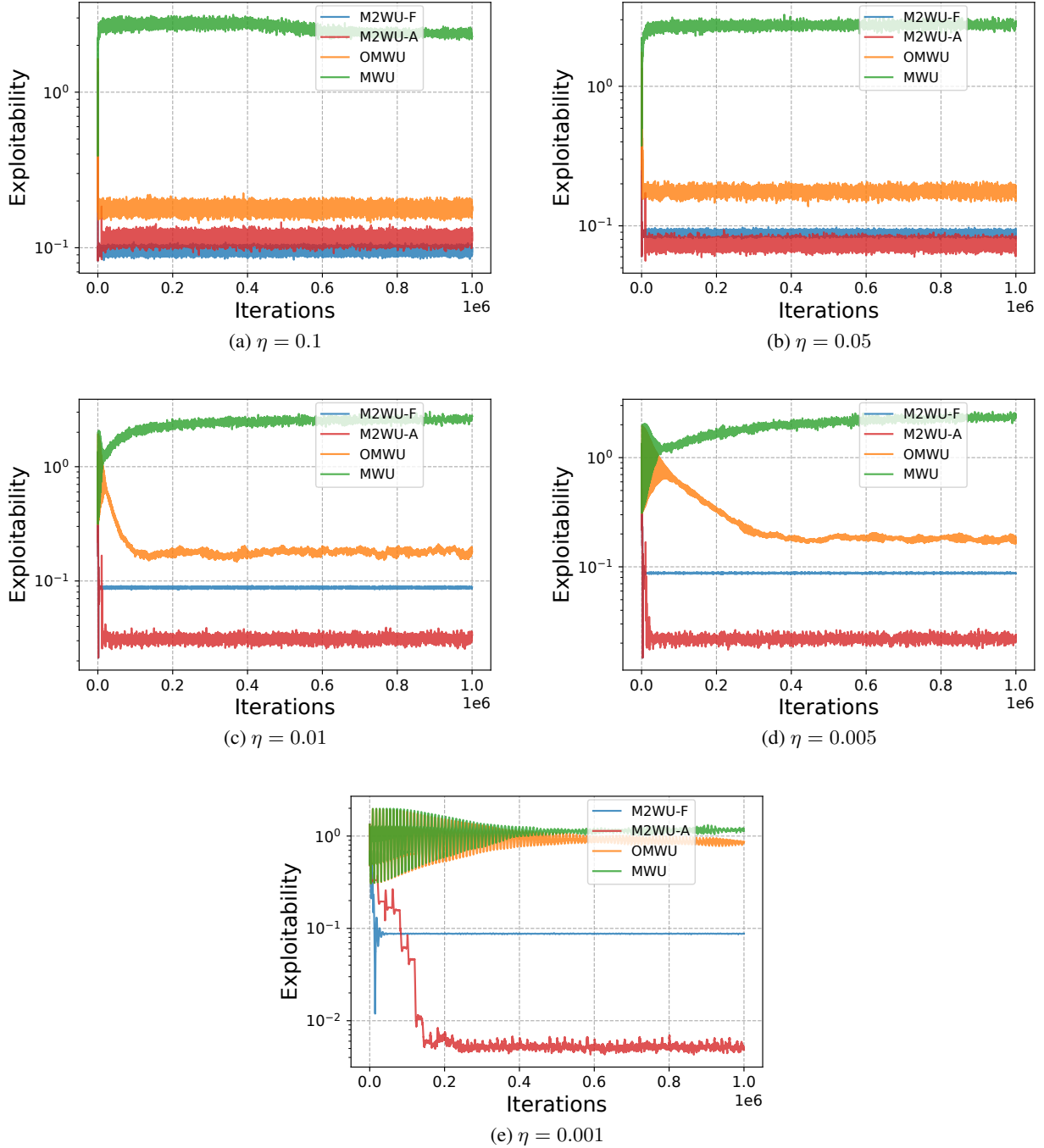(d) $\eta = 0.005$

(e) $\eta = 0.001$

Figure 6: Exploitability of π^t for M2WU, MWU, and OMWU with varying $\eta \in \{0.1, 0.05, 0.01, 0.005, 0.001\}$ in BRPS with noisy-information feedback.
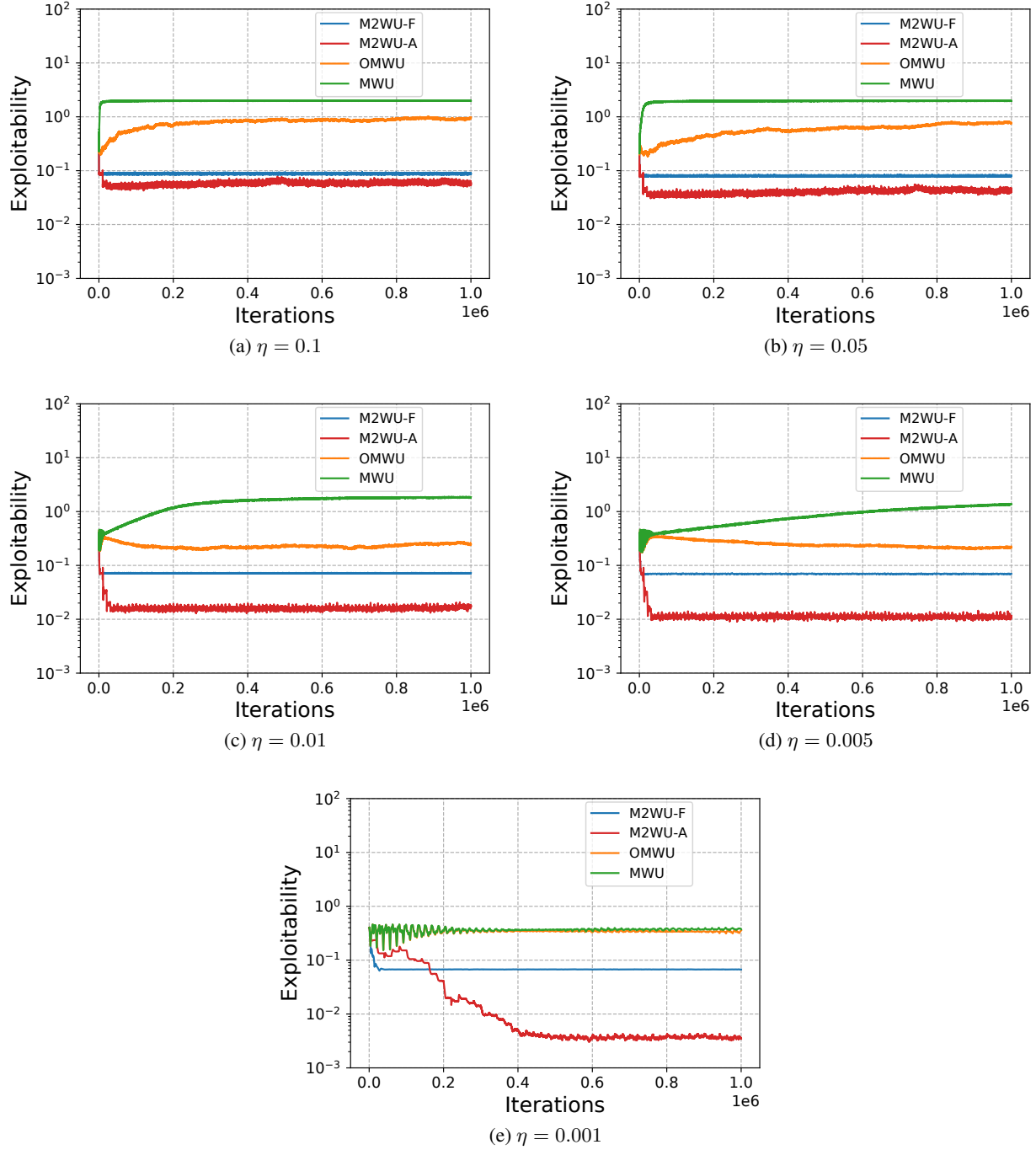
(a) $\eta = 0.1$

(b) $\eta = 0.05$

(c) $\eta = 0.01$

(d) $\eta = 0.005$

(e) $\eta = 0.001$

Figure 7: Exploitability of π^t for M2WU, MWU, and OMWU with varying $\eta \in \{0.1, 0.05, 0.01, 0.005, 0.001\}$ in M-Ne with noisy-information feedback.

F.2 Decreasing Learning Rates

In this section, we investigate the performance of M2WU with decreasing learning rates under the noisy-information feedback setting. We set the learning rates to $\eta_t = t^{-\frac{3}{4}}$ for all algorithms. Other settings are equivalent to the noisy-information feedback experiments in Section 7.2. Figure 8 shows the average exploitability of π^t on 100 instances. Even with the decreasing learning rates, M2WU demonstrates better performance than MWU and OMWU.
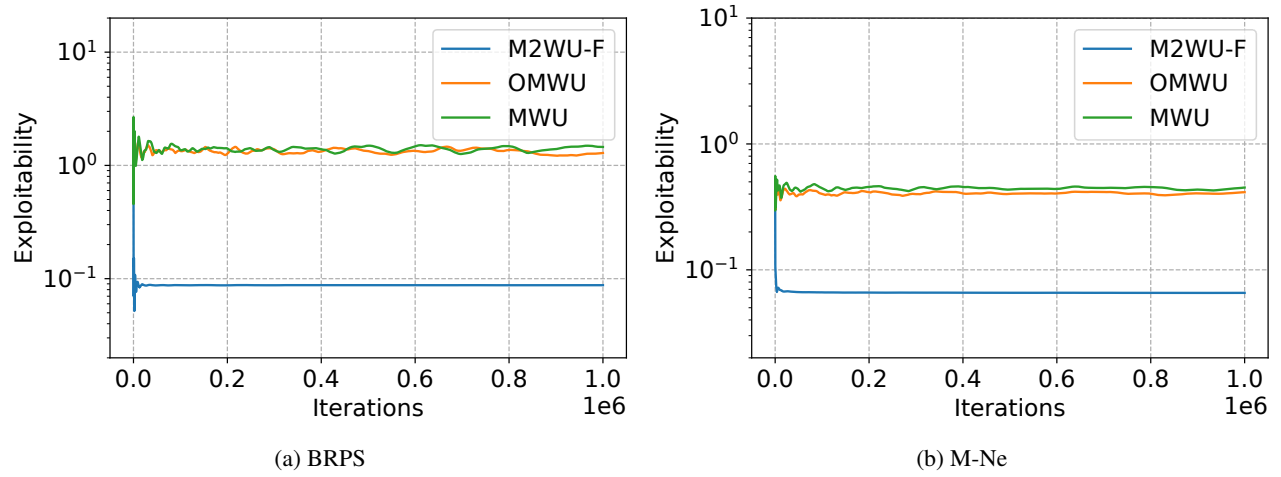
(a) BRPS

(b) M-Ne

Figure 8: Exploitability of π^t for M2WU, MWU, and OMWU with decreasing learning rates under the noisy-information feedback setting.