Optimistic Gradient Descent Ascent (OGDA) algorithm for saddle-point optimization has received growing attention due to its favorable last-iterate convergence. However, its behavior for simple two-player matrix games is still not fully understood — previous analysis lacks explicit convergence rates, only applies to exponentially small learning rate, or requires additional conditions such as uniqueness of the optimal solution. In this work, we significantly expand the understanding of OGDA, introducing a set of sufficient conditions under which OGDA exhibits concrete last-iterate convergence rates with a constant learning rate. Specifically, we show that matrix games satisfy these conditions and OGDA converges exponentially fast without any additional assumptions. More generally, our conditions hold for smooth bilinear functions and strongly-convex-strongly-concave functions over a constrained set. We provide experimental results to further support our theory.

To further demonstrate the significance of our results for matrix games, we greatly generalize the ideas to finite-horizon stochastic/Markov games and provide the first algorithm that simultaneously ensures 1) linear last-iterate convergence when playing against itself and 2) low regret when playing against an arbitrary slowly-changing opponent.

1 Introduction

Saddle-point optimization in the form of $\min_x \max_y f(x, y)$ dates back to [32], where the celebrated minimax theorem was discovered. Due to advances of Generative Adversarial Networks (GANs) [19] (which itself is a saddle-point problem), the question of how to find a good approximation of the saddle point, especially via an efficient iterative algorithm, has recently gained significant research interest. Simple algorithms such as Gradient Descent Ascent (GDA) and Multiplicative Weight Updates (MWU) is known to cycle and fails to converge even in simple bilinear cases (see e.g., [5]).

Many recent works consider resolving this issue via simple modifications of standard algorithms, usually in the form of some extra gradient descent/ascent steps. This includes Extra-Gradient methods (EG) [27, 31], Optimistic Gradient Descent Ascent (OGDA) [12, 17, 30], Optimistic Multiplicative Weight Updates (OMWU) [10, 26], and others. In particular, OGDA and OMWU are especially suitable for a repeated game setting where two players repeatedly propose $x_t$ and $y_t$ and receive only $\nabla_x f(x_t, y_t)$ and $\nabla_y f(x_t, y_t)$ respectively as feedback, with the goal of converging to a saddle point or equivalently a Nash equilibrium using game theory terminology.

Despite considerable progress, especially those for the unconstrained setting, the behavior of these algorithms for the constrained setting, where $x$ and $y$ are restricted to compact convex sets $\mathcal{X}$ and $\mathcal{Y}$
respectively, is still not fully understood. This is even true when \( f \) is a bilinear function and \( \mathcal{X} \) and \( \mathcal{Y} \) are simplex, known as the classic two-player zero-sum games in normal form, or simply matrix games. Indeed, existing convergence results on the last iterate of OGDA or OMWU for matrix games are unsatisfactory — they lack explicit convergence rates [33], only apply to exponentially small learning rate thus not reflecting the behavior of the algorithms in practice [10], or require additional conditions such as uniqueness of the equilibrium or a good initialization of the algorithms [10].

Motivated by this fact, in this work, we significantly expand the understanding of OGDA for constrained and smooth convex-concave saddle-point problems. Specifically, we start with proving an average duality gap convergence of OGDA at the rate of \( O(1/\sqrt{T}) \) after \( T \) iterations. Then, to obtain a more favorable last-iterate convergence in terms of distance to the set of equilibria, we propose two general sufficient conditions called Generalized Saddle-Point Restricted Secant Inequality (SP-RSI), under which we prove concrete last-iterate convergence rates for OGDA, all with a constant learning rate and without further assumptions.

Our last-iterate convergence results greatly generalize that of [21, Theorem 2], which itself is a consolidated version of results from several earlier works. The key implication of our new results is that, by showing that matrix games satisfy our SP-RSI condition, we provide by far the most general last-iterate convergence guarantee with a linear rate for this problem with OGDA. Our result does not suffer from the drawbacks of existing works mentioned earlier — it holds even when there are multiple equilibria and it is achieved with a constant learning rate and any initialization.

More generally, the same linear last-iterate convergence holds for any bilinear games over polytopes since they also satisfy the SP-RSI condition as we show. To complement this result, we construct an example of a bilinear game with a non-polytope feasible set where OGDA provably does not ensure linear convergence, indicating that the shape of the feasible set matters. We also provide experimental results to support our theory, which in particular show that OGDA indeed converges faster than OMWU.

Finally, to further showcase the significance of our results for matrix games, we greatly generalize the ideas to episodic stochastic games (a.k.a. Markov games). We show that using OGDA at each state with appropriate feedback ensures both linear convergence to an equilibrium when playing against itself and at the same time a worst-case regret guarantee when playing against an arbitrary slowly-changing opponent. As far as we know, this is the first algorithm for stochastic games with these two properties simultaneously and both with concrete finite-time bounds. Our regret bound is also better than the recent work of Radanovic et al. [34], which considers a variant of OMWU.

2 Related Work

Average-iterate convergence. While showing last-iterate convergence has been a challenging task, it is well-known that the average-iterate of many standard algorithms such as GDA and MWU enjoys a converging duality gap at the rate of \( O(1/\sqrt{T}) \) [15]. A line of works show that the rate can be improved to \( O(1/T) \) using the “optimistic” version of these algorithms such as OGDA and OMWU [35, 11, 36]. For tasks such as training GANs, however, average-iterate convergence is unsatisfactory since averaging large neural networks is usually prohibited.

Extra-Gradient (EG) algorithms. The saddle-point problem fits into the more general variational inequality framework [20]. A classic algorithm for variational inequalities is EG, first introduced in [25]. Tseng [37] is the first to prove last-iterate convergence for EG in various settings such as bilinear or smooth strongly-convex-strongly-concave problems. Recent works significantly expand the understanding of EG and its variants for unconstrained bilinear problems [27], unconstrained strongly-convex-strongly-concave problems [31], and more [42, 28, 18].

The original EG is not applicable to a repeated game setting where only one gradient evaluation is possible in each iteration. However, there are variants of EG that allow so. In fact, some of these versions coincide with the OGDA algorithm under different names such as modified Arrow–Hurwicz method [33] and “extrapolation from the past” [17]. Nevertheless, to best of our knowledge, none of the existing results covers the constrained bilinear case (which is one of our key contributions).

OGDA and OMWU. Recently, last-iterate convergence for OGDA has been proven in various settings such as convex-concave problems [12], unconstrained bilinear problems [8, 27], strongly-
We consider the following constrained saddle-point problem:
\[
\min_{x \in X} \max_{y \in Y} f(x, y),
\]
where \( X \) and \( Y \) are compact convex sets, and \( f \) is a continuous differentiable function that is convex in \( x \) for any fixed \( y \) and concave in \( y \) for any fixed \( x \). By the celebrated minimax theorem [32], we have
\[
\min_{x \in X} \max_{y \in Y} f(x, y) = \max_{y \in Y} \min_{x \in X} f(x, y).
\]
The set of minimax optimal strategy is denoted by \( X^* = \arg\min_{x \in X} \max_{y \in Y} f(x, y) \), and the set of maximin optimal strategy is denoted by \( Y^* = \arg\max_{y \in Y} \min_{x \in X} f(x, y) \). It is well-known that \( X^* \) and \( Y^* \) are compact and convex, and any pair \((x^*, y^*) \in X^* \times Y^*\) is a Nash equilibrium satisfying \( f(x^*, y^*) \leq f(x^*, y) \leq f(x^*, y^*) \leq f(x, y^*) \) for any \((x, y) \in X \times Y\).

For notational convenience, we define \( Z = X \times Y \) and similarly \( Z^* = X^* \times Y^* \). For a point \( z = (x, y) \in Z \), we further define \( f(z) = f(x, y) \) and \( F(z) = (\nabla_x f(x, y), -\nabla_y f(x, y)) \).

Our goal is to find a point \( z \in Z \) that is close to the set of Nash equilibria \( Z^* \), and we consider two ways of measuring the closeness. The first one is the \textit{duality gap}, defined as
\[
\alpha_f(z) = \max_{y' \in Y} f(x, y') - \min_{x' \in X} f(x', y),
\]
which is always non-negative since \( \max_{y' \in Y} f(x, y') \geq f(x, y) \geq \min_{x' \in X} f(x', y) \). The second one is the distance between \( z \) and \( Z^* \). Specifically, for any compact set \( A \), we define the projection operator \( \Pi_A(a) = \arg\min_{a' \in A} \|a - a'\| \) (ties broken arbitrarily; all norms in this work are 2-norm unless specified otherwise). The squared distance between \( z \) and \( Z^* \) is then defined as
\[
\text{dist}(z, Z^*) = \|z - \Pi_{Z^*}(z)\|^2.
\]
Throughout the paper, we make the following two regularity assumptions. The first one is without loss of generality and assumes that the diameter of the feasible set \( Z \) is normalized to 1.

**Assumption 1.** For any \( z, z' \in Z \), \( \|z - z'\| \leq 1 \) holds.

The second one assumes that \( f \) is \( L \)-smooth, which can be written as the following.

**Assumption 2.** For any \( z, z' \in Z \), \( \|F(z) - F(z')\| \leq L\|z - z'\| \) holds.
Finally, we state a useful lemma related to the duality gap, used in several places of our proofs.

**Lemma 1.** For any \( z \in \mathcal{Z} \), we have \( \alpha_f(z) \leq \max_{z' \in \mathcal{Z}} F(z)^\top (z - z') \).

This is a direct consequence of the convexity of \( f(\cdot, y) \) and the concavity of \( f(x, \cdot) \):

\[
\alpha_f(z) = \max_{(x', y') \in \mathcal{X} \times \mathcal{Y}} (f(x, y') - f(x, y) + f(x, y) - f(x', y)) \\
\leq \max_{(x', y') \in \mathcal{X} \times \mathcal{Y}} \left( \nabla_y f(x, y)^\top (y' - y) + \nabla_x f(x, y)^\top (x - x') \right) = \max_{z' \in \mathcal{Z}} F(z)^\top (z - z').
\]

Note that the lemma also indicates that \( \max_{z' \in \mathcal{Z}} F(z)^\top (z - z') \) is always non-negative due to the non-negativity of the duality gap.

### 3.1 Optimistic Gradient Descent Ascent

We consider solving the problem via an iterative algorithm known as Optimistic (projected) Gradient Descent Ascent (OGDA). Starting from two arbitrary points \( z_0 = (\hat{x}_0, \hat{y}_0) \) and \( z_0 = (x_0, y_0) \) from \( \mathcal{Z} \), OGDA with step size \( \eta > 0 \) iteratively computes the following for \( t = 1, 2, \ldots \),

\[
\begin{align*}
\hat{x}_{t+1} & = \Pi_X \left( \hat{x}_t - \eta \nabla_x f(x_t, y_t) \right), \\
x_{t+1} & = \Pi_X \left( \hat{x}_{t+1} - \eta \nabla_x f(x_t, y_t) \right), \\
\hat{y}_{t+1} & = \Pi_Y \left( \hat{y}_t + \eta \nabla_y f(x_t, y_t) \right), \\
y_{t+1} & = \Pi_Y \left( \hat{y}_{t+1} + \eta \nabla_y f(x_t, y_t) \right),
\end{align*}
\]

which can be compactly written as

\[
\begin{align*}
\hat{z}_{t+1} & = \Pi_{\mathcal{Z}} \left( \hat{z}_t - \eta F(z_t) \right), \\
z_{t+1} & = \Pi_{\mathcal{Z}} \left( \hat{z}_{t+1} - \eta F(z_t) \right).
\end{align*}
\]

Note that there are several slightly different versions of the algorithm in the literature, which differ in the timing of performing the projection. Our version is same as those in [6, 35]. It is also referred to as “single-call extra-gradient” in [21], but it does not belong to the class of “extra-gradient” methods discussed in [37, 27, 18] for example.

Also note that OGDA only requires accessing \( f \) via its gradient, and in fact, only one gradient at the point \( z_t \) for each iteration. This aspect makes it especially suitable for a repeated game setting, where in each round, one player proposes \( x_t \) while another player proposes \( y_t \). With only the information of the gradient from the environment \( \nabla_x f(x_t, y_t) \) for the first player and \( \nabla_y f(x_t, y_t) \) for the other), both players can execute the algorithm.

### 4 Convergence Results

In this section, we provide our main convergence results for the iterate \( \hat{z}_t \) of OGDA. Specifically, in Section 4.1 we first provide a convergence guarantee in terms of the duality gap. Then in Section 4.2, we propose a general condition subsuming many well-studied cases, under which OGDA is shown to ensure a last-iterate convergence guarantee in terms of the distance between \( \hat{z}_t \) and \( \mathcal{Z}^* \).

Before presenting the main theorems, we first show two important lemmas. The first one shows the connection between \( \|\hat{z}_{t+1} - z\|^2 \) and \( \|\hat{z}_t - z\|^2 \) for any \( z \in \mathcal{Z} \), and the proof follows standard analysis of OGDA (see e.g., [35, Lemma 1]). For completeness, we provide a proof in Appendix B.

**Lemma 2.** For any \( t \geq 1 \) and \( z \in \mathcal{Z} \), OGDA with \( \eta \leq \frac{1}{8\|\nabla f\|} \) ensures

\[
2\eta F(z_t)^\top (z_t - z) \leq \|\hat{z}_t - z\|^2 - \|\hat{z}_{t+1} - z\|^2 - \|\hat{z}_{t+1} - z_t\|^2 - \frac{15}{16} \|z_t - z\|^2 + \frac{1}{16} \|\hat{z}_t - z_{t-1}\|^2.
\]

Choosing any \( z \in \mathcal{Z}^* \) in this lemma, we have \( F(z_t)^\top (z_t - z) \geq f(x_t, y_t) - f(x, y_t) + f(x_t, y) - f(x_t, y_t) = f(x_t, y) - f(x, y_t) \geq 0 \) by convexity/concavity and the optimality of \( z \). Further rearranging then shows that \( \hat{z}_{t+1} \) is closer to \( z \) than \( \hat{z}_t \) as long as \( \|\hat{z}_{t+1} - z_t\|^2 + \frac{15}{16} \|z_t - \hat{z}_t\|^2 - \frac{1}{16} \|\hat{z}_t - z_{t-1}\|^2 \) is positive, which drives the convergence of \( \hat{z}_t \) towards \( \mathcal{Z}^* \). To this end, in the following lemma, we provide a lower bound for the main term \( \|\hat{z}_{t+1} - z_t\|^2 + \|z_t - \hat{z}_t\|^2 \).
Lemma 3. For any $t \geq 0$ and $z' \in \mathcal{Z}$ with $z' \neq \hat{z}_{t+1}$, OGDA with $\eta \leq \frac{1}{81\alpha}$ ensures
\[
\|\hat{z}_{t+1} - z_t\|^2 + \|z_t - \hat{z}_t\|^2 \geq \frac{32}{81\eta^2} \left( \frac{F(\hat{z}_{t+1})^\top(\hat{z}_{t+1} - z')}{\|\hat{z}_{t+1} - z\|^2} \right)_+^2,
\]
where $[a]_+ \triangleq \max\{a, 0\}$.

See Appendix B for the proof. Note that the right-hand side of Eq. (7) can be further related to the duality gap of $\hat{z}_{t+1}$ through Lemma 1. Thus, the high-level idea of proving last-iterate convergence emerges: when $\hat{z}_{t+1}$ has a large duality gap, the term $\|\hat{z}_{t+1} - z_t\|^2 + \|z_t - \hat{z}_t\|^2$ has to be large also, which in turn leads to a large decrease in $\|\hat{z}_{t+1} - z\|^2$ compared to $\|\hat{z}_t - z\|^2$ by Lemma 2.

### 4.1 Duality Gap Convergence

We are now ready to present our first convergence result on the duality gap.

Theorem 4. OGDA with $\eta \leq \frac{1}{81\alpha}$ ensures $\frac{1}{T} \sum_{t=1}^{T} \alpha_f(\hat{z}_t) = O \left( \frac{1}{\sqrt{T}} \right)$ for any $T$.

See Appendix C for the proof. This theorem indicates that $\alpha_f(\hat{z}_t)$ is converging to zero. A convergence rate of $\alpha_f(\hat{z}_t) = O \left( \frac{1}{\sqrt{T}} \right)$ would be compatible with the theorem, but is not directly implied by it. We are only able to prove a concrete convergence rate on $\alpha_f(\hat{z}_t)$ under further conditions; see Section 4.2. In a recent work, Golowich et al. [18] consider the unconstrained setting and show that the extra-gradient algorithm obtains the rate $O \left( \frac{1}{\sqrt{T}} \right)$, under an extra assumption that the Hessian of $f$ is also Lipschitz. We also note that the extra-gradient algorithm requires more cooperation between the two players compared to OGDA and is less suitable for a repeated game setting.

### 4.2 Pointwise Convergence

Recall the discussion after Lemma 3 that a large duality gap, or rather a large right-hand side of Eq. (7), ensures a sufficient decrease in $\|\hat{z}_{t+1} - z_t\|^2$. To obtain a last-iterate convergence result, however, we still need to connect $\text{dist}(\hat{z}_t, \mathcal{Z}^*)$ back to the duality gap. Motivated by this fact, we propose the following two general conditions on $f$ and $\mathcal{Z}$ to achieve so.

**Definition 1** (Generalized Saddle-Point Restricted Secant Inequality (SP-RSI)). The SP-RSI-1 and SP-RSI-2 conditions are defined as: for any $z \in \mathcal{Z}$ with $z^* = \Pi_{\mathcal{Z}^*}(z)$,

\[
\text{(SP-RSI-1)} \quad \max_{z' \in \mathcal{Z}} F(z')^\top(z - z') \geq C\|z - z^*\|^\beta + 1,
\]

\[
\text{(SP-RSI-2)} \quad F(z)^\top(z - z^*) \geq C\|z - z^*\|^\beta + 2,
\]

holds for some parameter $\beta \geq 0$ and $C > 0$.

We call these conditions Generalized Saddle-Point Restricted Secant Inequality (SP-RSI) for their resemblance to the Restricted Secant Inequality [41, 24]. The latter leads to linear convergence in convex optimization. To the best of our knowledge, our proposed conditions are new for saddle-point problems. The closest condition in the literature is from a recent work by Hsieh et al. [21] for variational inequalities (their Assumption 3(s)). In our context, their condition is $(F(z) - F(z'))^\top(z - z') \geq C\|z - z\|^2$ for all $z, z' \in \mathcal{Z}$. This in fact implies SP-RSI-2 with $\beta = 0$. To see this, simply set $z' = z^*$ and notice that $F(z^*)^\top(z - z^*) \geq 0$ holds by the first-order optimality condition of $z^*$. Therefore, our condition SP-RSI-2 is only less stringent and covers more problems. In fact, under their stronger condition the equilibrium is necessarily unique, but we generally allow multiple equilibria.

These two conditions might be hard to interpret at first glance, but they subsume many standard settings studied in the literature. The first and perhaps the most important example is bilinear games with a polytope feasible set, which in particular includes the classic two-player matrix games as a special case (since they are bilinear games over the simplex).

**Theorem 5.** A bilinear game $f(x, y) = x^\top Gy$ with $\mathcal{X} \subseteq \mathbb{R}^M$ and $\mathcal{Y} \subseteq \mathbb{R}^N$ being polytopes and $G \in \mathbb{R}^{M \times N}$ satisfies SP-RSI-1 with $\beta = 0$.  

5
Note that we have not provided the concrete form of the parameter $C$ in the theorem (which depends on $\mathcal{X}$, $\mathcal{Y}$, and $G$), but it can be found in the proof (see Appendix D). The next example shows that strongly-convex-strongly-concave problems are also special cases of our condition.

**Theorem 6.** If $f$ is strongly convex in $x$ and strongly concave in $y$, then SP-RSI-2 holds with $\beta = 0$.

Next, we provide a toy example where SP-RSI-2 only holds with $\beta > 0$.

**Theorem 7.** Let $\mathcal{X} \equiv \mathcal{Y} \triangleq \{(a, b) : 0 \leq a, b \leq 1, a + b = 1\}$, $n > 2$ be an integer, and $f(x, y) = x(1)^{2n} - x(1)y(1) - y(1)^{2n}$. Then SP-RSI-2 holds with $\beta = 2n - 2$ but not with $\beta = 0$.

Under either condition with any value of $\beta$, we show the following last-iterate convergence guarantee of OGDA, which is one of the main contributions of this work.

**Theorem 8.** For any $\eta \leq \frac{1}{8L}$, if either SP-RSI-1 or SP-RSI-2 holds with $\beta = 0$, then OGDA guarantees linear last-iterate convergence rate:

$$\text{dist}(\tilde{z}_t, Z^*) \leq 3(1 + C)^{-t};$$

(8)

on the other hand, if either condition holds with $\beta > 0$, then we have

$$\text{dist}(\tilde{z}_t, Z^*) \leq \left(2 + \left(\frac{2}{C'\beta}\right)^\frac{1}{2}\right)t^{-\frac{1}{4}},$$

(9)

where $C' \triangleq \frac{16n^2 \min \{C^2, L^2\}}{81(1 + \beta)^2}$.

We defer the proof to Appendix F and make several remarks. First, note that based on a convergence result on $\text{dist}(\tilde{z}_t, Z^*)$, one can immediately obtain a convergence guarantee for the duality gap $\alpha_f(\tilde{z}_t)$ as long as $f$ is also Lipschitz. This is because $\alpha_f(\tilde{z}_t) \leq \max_{x', y'} f(\tilde{x}_t, y') - f(x^*, y') + f(x', y^*) - f(x^*, y^*) \leq O(\|\tilde{x}_t - x^*\| + \|y_t - y^*\|) = O(\sqrt{\text{dist}(\tilde{z}_t, Z^*)})$ where $(x^*, y^*) = \Pi_{\mathcal{Z}}(\tilde{z}_t)$. While this leads to stronger guarantees compared to Theorem 4, we emphasize that the latter holds even without the SP-RSI conditions.

Second, our results significantly generalize [21, Theorem 2] which itself is a consolidated version of several earlier works and also shows a linear convergence rate of OGDA under a condition stronger than our SP-RSI-2 with $\beta = 0$ as discussed earlier. More specifically, our results show that linear convergence rate holds for a much broader set of problems, especially under SP-RSI-1 which was not discovered before. Furthermore, we also show slower sublinear convergence rates for any value of $\beta > 0$ under either condition, which is also new as far as we know. In particular, we empirically verify that OGDA indeed does not converge exponentially fast for the toy example defined in Theorem 7 (see Appendix A).

Last but not least, the most significant implication of Theorem 8 is that it provides the by far most general linear convergence rate for the classic two-player matrix games, or more generally bilinear games with polytope constraints, according to Theorem 5 and Eq. (8). Compared to recent works of [8, 10] for matrix games (on OGDA or OMWU), our result is considerably stronger: 1) we do not require a unique equilibrium while they do; 2) our linear convergence rate holds for any initial points $z_0$ and $\tilde{z}_0$, while their only holds if the initial points are in a small neighborhood of the unique equilibrium (otherwise the rate is sublinear); 3) our only requirement on the step size is $\eta \leq \frac{1}{8L}$, while they require an exponentially small $\eta$, which does not reflect the behavior of the algorithms in practice. In Appendix A, we empirically show that OGDA indeed outperforms OMWU when both tuned with a constant learning rate.

One may wonder what happens if a bilinear game has a non-polytope constraint. It turns out that in this case, SP-RSI-1 may only hold with $\beta > 0$, due to the following example showing that linear convergence provably does not hold for OGDA when the feasible set has a curved boundary.

**Theorem 9.** There exist bilinear games with a non-polytope feasible set such that $\text{dist}(\tilde{z}_t, Z) = \Omega(1/t^2)$ holds for OGDA.

This example indicates that the shape of the feasible set plays an important role in last-iterate convergence, which may be an interesting future direction to investigate. This is also verified empirically in our experiments (see Appendix A).
5 Episodic Stochastic Games

In this section, we switch the focus to two-player episodic stochastic games, which can be viewed as multi-stage matrix games and are often used for modeling multi-agent reinforcement learning problems. We propose a variant of OGDA for this problem and extend our last-iterate convergence analysis to this much more challenging setting.

Specifically, a stochastic game is defined by a tuple \((S, A, B, \ell, p_0, p, H)\), where: 1) \(S\) is a finite state space, which is a union of \(H\) disjoint layers \(S_1, \ldots, S_H\); 2) \(A\) and \(B\) are finite action space for Player 1 and Player 2 respectively; 3) \(\ell : S \times A \times B \to [-1, 1]\) is the loss (payoff) function for Player 1 (Player 2), with \(\ell(s, a, b)\) specifying how much Player 1 pays to Player 2 if they are at state \(s\) and select actions \(a\) and \(b\) respectively; 4) \(p_0 \in \Delta_S\) is the initial state distribution; 5) \(p : S \times A \times B \to \Delta_S\) is the transition function, with \(p(s' | s, a, b)\) specifying the probability of transferring to state \(s'\) after actions \(a\) and \(b\) are taken by the two players respectively at state \(s\). One restriction on \(p\) is that for any \(i < H, a, b, s \in S_i, p(\cdot | s, a, b)\) is supported on \(S_{i+1}\). In other words, transition only happens from one layer to the next layer.

We consider a full-information setting (studied in e.g., [34]) where both players know the game parameters \((S, A, B, \ell, p_0, p, H)\) and interact through \(T\) episodes with the goal of either converging to an equilibrium when cooperating with each other or ensuring low regret otherwise. Specifically, in each episode \(t = 1, \ldots, T\), Player 1 (Player 2) decides her policy \(x_t (y_t)\), which is a collection of distributions \(\{x^t_s\}_{s \in S}\) (\(\{y^t_s\}_{s \in S}\)) where \(x^t_s \in \Delta_A (y^t_s \in \Delta_B)\) is her action distribution at state \(s\). Then starting from an initial state \(s \in S_1\) drawn from \(p_0\), the two players repeat the following: sample actions \(a\) from \(x^t_s\) and \(b\) from \(y^t_s\); Player 1 pays \(\ell(s, a, b)\) to Player 2; if \(s \notin S_H\), then the state \(s\) is updated as a new sample drawn from \(p(\cdot | s, a, b)\), otherwise this episode ends. At the end of the episode, the two players inform each other of their polices \(x_t\) and \(y_t\) used in this episode.

According to this protocol, for a fixed policy pair \((x, y)\), it is clear that the expected loss (payoff) of Player 1 (Player 2) is \(\rho_{x,y} = \mathbb{E}_{s \sim p_0} V^s_{x,y}\) where \(V^s_{x,y}\) is defined as (with the help of another auxiliary function \(Q^s_{x,y}\)):

\[
V^s_{x,y} = \sum_{a \in A} \sum_{b \in B} x^s(a) y^s(b) Q^s_{x,y}(a, b) = x^s V^s_{x,y},
\]

\[
Q^s_{x,y}(a, b) = \begin{cases} 
\ell(s, a, b) & \text{if } s \in S_H, \\
\ell(s, a, b) + \sum_{s'} p(s' | s, a, b) V^{s'}_{x,y} & \text{else}
\end{cases}
\]

In words, \(Q^s_{x,y}(a, b)\) is the expected loss of Player 1 if the two players start from state \(s\), take actions \(a\) and \(b\) respectively, and then follow polices \(x\) and \(y\) for the rest of the episode. Similarly, \(V_t^s\) is the expected loss of Player 1 if the two players start from state \(s\) and follow polices \(x\) and \(y\). Clearly, both of them can be computed efficiently via dynamic programming.

Note that \(\rho_{x,y}\) is non-convex in \(x\) and non-concave in \(y\) if \(H > 1\). However, due to the layered structure of the problem, it is clear that an equilibrium \((x_*, y_*)\) for this game is such that for each \(s \in S_1, (x^*_s, y^*_s)\) is an equilibrium for the matrix game with matrix \(Q^s_{x, y_*}\). For conciseness, we write \(Q^s_{x, y_*} \triangleq Q^s_{x, y^*_s}\). Further define \(X^*_x = \arg\min_{x^s} \max_{y^s} x^s V^s_{x, y^s} \) and \(Y^*_y = \arg\max_{y^s} \min_{x^s} x^s V^s_{x, y}\), which are the sets minimax and maximin strategies for the respective player at state \(s\).

5.1 Last-iterate Convergence of Policies

We now present an algorithm so that if both players deploy it, their policies converge to \(X^*_x\) and \(Y^*_y\) respectively for all state \(s\). Based on previous discussions, a naive approach would be to first find an approximate equilibrium for all states in the last layer \(s \in S_H\), using for example OGDA based on our results in Section 4 since this is a standard matrix game, then similarly find an approximate equilibrium for all states in the previous layer \(s \in S_{H-1}\), since we now have a good approximation for \(Q_*^s\). Repeating this process one can ensure that the entire policy converges to the optimal set. This is, however, clearly a wasteful and impractical approach.

To improve the efficiency, our idea is basically to learn all the layers simultaneously by running (a variant of) OGDA at each state with appropriate feedback. Specifically, with notation shorthands
Theorem 10. For any opponent generating policy \( y_t \), our algorithm update \( x_t \) and \( y_t \) in the following way:

\[
\begin{align*}
\tilde{x}^s_{t+1} &= \Pi_{\Delta_A} \left( \tilde{x}^s_t - \eta Q^s_{t,t} y^s_t \right), \\
x^s_{t+1} &= \Pi_{\Delta_A} \left( \tilde{x}^s_{t+1} - \eta Q^s_{t+1,t} y^s_t \right), \\
\tilde{y}^s_{t+1} &= \Pi_{\Delta_B} \left( \tilde{y}^s_t + \eta Q^r_{t,t} x^s_t \right), \\
y^s_{t+1} &= \Pi_{\Delta_B} \left( \tilde{y}^s_{t+1} + \eta Q^r_{t+1,t} x^s_t \right),
\end{align*}
\]

for each \( s \) starting from the last layer back to the first layer. Several remarks on the algorithm are in order. First, note that in contrast to the version of OGDA described in Section 3.1, here we use different feedback to update \( \tilde{x}^s_{t+1} \) and \( x^s_{t+1} \) (similarly \( \tilde{y}^s_{t+1} \) and \( y^s_{t+1} \)). This is only for a better regret guarantee described in Section 5.2. For the last-iterate convergence result, changing \( Q^s_{t+1,t} \) and \( Q^s_{t,t+1} \) to \( Q^s_{t,t} \) does not matter.

Second, in update rule Eq. (11) (similarly for Eq. (13)), although \( Q^s_{t+1,t} \) depends on \( x^s_{t+1} \) and thus the update appears to be recursive, by definition the dependence is only through \( x^s_{t+1} \) for \( s' \) from layers after \( s \) where \( x^s_{t+1} \) has already been computed. Therefore, the algorithm is well-defined and can be implemented efficiently. Also note that to compute \( x^s_{t+1} \) and \( y^s_{t+1} \), each player indeed only needs to know the opponent’s policy \( (y^s_t \text{ and } x^s_t) \) respectively from the last episode.

Finally, while the idea of applying OGDA at each state to deal with a matrix game is natural, the fact that this matrix \( Q^s_{t,t} \) (or \( Q^s_{t+1,t} \) and \( Q^s_{t,t+1} \)) is actually changing over time introduces extra complication in the analysis. Fortunately, since the time-varying component is only due to the changing policy for states after the current layer, if they can be shown to converge, then the change of the matrix is getting smaller and smaller, making the convergence of the current layer possible as well. Based on this idea, we prove the following theorem showing the linear last-iterate convergence of our algorithm. Since at each state the learner faces a matrix game, we define \( C > 0 \) to be a constant that ensures the following for all states and all \((x^s, y^s)\) and its projection \((x^s_*, y^s_*)\) to the set \( X^s_* \times Y^s_*\):

\[
\max_{\tilde{x}^s, \tilde{y}^s} \left( x^s - Q^s_{t} \tilde{y}^s - \tilde{x}^s Q^r_{t} y^s \right) \geq C \left( \|x^s - x^s_*\| + \|y^s - y^s_*\| \right).
\]

This corresponds to SP-RSI-1, and by the same argument as in Theorem 5, such \( C \) always exists.

Theorem 10. With any \( \eta \leq \frac{1}{16H\sqrt{2|A||B|}} \), update rules (10)-(13) ensure for all \( h, t, \) and \( s \in S_h, \)

\[
\text{dist} \left( \tilde{z}^s_t, Z^s_* \right) \leq (1 + C')^{-t+1} + O \left( \frac{H^2 \log \frac{H|A||B|}{C'} c}{c'} \right),
\]

where \( \tilde{z}^s_t = (\tilde{x}^s_t, \tilde{y}^s_t) \), \( Z^s_* = X^s_* \times Y^s_* \), and \( C' = \frac{\eta^2 H^2}{1000} \).

The linear convergence guarantee in Theorem 10 is similar to Eq. (8), except that for layer \( h, t \) is offset by \( \tilde{O} \left( \frac{H-h}{c} \right) \), reflecting the fact that the earlier layers converge slightly slower than the later ones. As far as we know, this is the first last-iterate convergence result for stochastic games.

5.2 Low Regret against Slowly-changing Opponents

Next, we consider a setup where we control the behavior of Player 1, while Player 2 is an arbitrary opponent. In this case, a standard measure of the performance of Player 1 is her regret, defined as

\[
\text{Reg}_T = \sum_{t=1}^T \rho_{x,y} - \min_{x} \sum_{t=1}^T \rho_{x,y},
\]

which is the difference between the expected total loss of Player 1 and that of the best fixed policy in hindsight. Achieving sublinear regret in general is known to be at least as hard as agnostic learning of parity functions [40, 34], for which no polynomial-time algorithm is known. Instead, following [34], we aim to achieve a regret bound in terms of the variation of the opponent’s policy. Indeed, it is a common situation that the opponent is slowly changing over time, for example, when the opponent is also applying some relatively stable learning algorithm.

It turns out that the very same algorithm specified in Eq. (10) and Eq. (11), achieves such a desirable regret bound already, as we show in the following theorem.

Theorem 11. For any opponent generating policy \( y_1, \ldots, y_T \), the update rules (10) and (11) with any \( \eta > 0 \) ensure

\[
\text{Reg}_T = O \left( \frac{H}{\eta} \nu_1 + \eta H^2 |A| \nu_2 \right),
\]

\[
\nu_1 = \nu_2 = 0
\]
where $V_1 = 1 + \sum_{t=2}^{T} \sum_{h=1}^{H} \max_{s \in S_h} \|y^s_t - y^s_{t-1}\|_1$ and $V_2 = \frac{1}{T} + \sum_{t=2}^{T} \sum_{h=1}^{H} \max_{s \in S_h} \|y^s_t - y^s_{t-1}\|_2^2$. In particular, if there exists $\alpha > 0$ such that $\|y^s_t - y^s_{t-1}\|_1 = O(T^{-\alpha})$ for all $t > 1$ and $s$, then tuning the best step size gives $\text{Reg} = O(\sqrt{HT\|A\| \cdot T_{\text{max}} \{1 - 3\frac{\alpha}{2}, 1\}})$.

Under the same setting, we are only aware of a recent related work by Radanovic et al. [34] analyzing an algorithm based on OMWU with some sophisticated tricks, and providing a worse bound of $O(T_{\text{max}}^{\max(1 - \frac{\alpha}{2}, \frac{1}{4})})$ (ignoring dependence on other parameters) when $\|y^s_t - y^s_{t-1}\|_1 = O(T^{-\alpha})$). The work by Gajane et al. [16] studies a more challenging measure called dynamic regret in a setting with bandit-feedback, which is not comparable to our result. We emphasize that only our algorithm enjoys a regret bound and simultaneously a last-iterate convergence guarantee when playing against itself.

Acknowledgments and Disclosure of Funding

We are grateful for the support of NSF Awards IIS-1755781 and IIS-1943607.

References

[1] Yu Bai and Chi Jin. Provable self-play algorithms for competitive reinforcement learning. arXiv preprint arXiv:2002.04017, 2020.

[2] Michael Bowling. Convergence and no-regret in multiagent learning. In Advances in neural information processing systems, pages 209–216, 2005.

[3] Michael Bowling and Manuela Veloso. Rational and convergent learning in stochastic games. In Proceedings of the 17th international joint conference on Artificial intelligence-Volume 2, pages 1021–1026, 2001.

[4] Ronen I Brafman and Moshe Tennenholtz. R-max-a general polynomial time algorithm for near-optimal reinforcement learning. Journal of Machine Learning Research, 3(Oct):213–231, 2002.

[5] Yun Kuen Cheung and Georgios Piliouras. Vortices instead of equilibria in minmax optimization: Chaos and butterfly effects of online learning in zero-sum games. In Conference on Learning Theory, pages 807–834, 2019.

[6] Chao-Kai Chiang, Tianbao Yang, Chia-Jung Lee, Mehrdad Mahdavi, Chi-Jen Lu, Rong Jin, and Shenghuo Zhu. Online optimization with gradual variations. In Conference on Learning Theory, pages 6–1, 2012.

[7] Vincent Conitzer and Tuomas Sandholm. Awesome: A general multiagent learning algorithm that converges in self-play and learns a best response against stationary opponents. Machine Learning, 67(1-2):23–43, 2007.

[8] Constantinos Daskalakis and Ioannis Panageas. The limit points of (optimistic) gradient descent in min-max optimization. In Advances in Neural Information Processing Systems, pages 9236–9246, 2018.

[9] Constantinos Daskalakis and Ioannis Panageas. Last-iterate convergence: Zero-sum games and constrained min-max optimization. Smooth Games Optimization and Machine Learning Workshop (NeurIPS 2019), 2019.

[10] Constantinos Daskalakis and Ioannis Panageas. Last-iterate convergence: Zero-sum games and constrained min-max optimization. Innovations in Theoretical Computer Science, 2019.

[11] Constantinos Daskalakis, Alan Deckelbaum, Anthony Kim, et al. Near-optimal no-regret algorithms for zero-sum games. Games and Economic Behavior, 2015.

[12] Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training gans with optimism. In International Conference on Learning Representations, 2018.
[13] Damek Davis. Lecture 5, mathematical programming I, 2016. Available at people.orie.cornell.edu/dsd95/teaching/orie6300/lec05.pdf.

[14] Damek Davis. Lecture 6, mathematical programming I, 2016. Available at people.orie.cornell.edu/dsd95/teaching/orie6300/lec06.pdf.

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

[16] Pratik Gajane, Ronald Ortner, and Peter Auer. Variational regret bounds for reinforcement learning. arXiv preprint arXiv:1905.05857, 2019.

[17] Gauthier Gidel, Hugo Berard, Gaëtan Vignoud, Pascal Vincent, and Simon Lacoste-Julien. A variational inequality perspective on generative adversarial networks. International Conference on Learning Representations, 2019.

[18] Noah Golowich, Sarath Pattathil, Constantinos Daskalakis, and Asuman Ozdaglar. Last iterate is slower than averaged iterate in smooth convex-concave saddle point problems. arXiv preprint arXiv:2002.00057, 2020.

[19] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Advances in neural information processing systems, 2014.

[20] Patrick T Harker and Jong-Shi Pang. Finite-dimensional variational inequality and nonlinear complementarity problems: a survey of theory, algorithms and applications. Mathematical programming, 1990.

[21] Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. On the convergence of single-call stochastic extra-gradient methods. In Advances in Neural Information Processing Systems, 2019.

[22] Junling Hu and Michael P Wellman. Nash q-learning for general-sum stochastic games. Journal of machine learning research, 4(Nov):1039–1069, 2003.

[23] Zeyu Jia, Lin F Yang, and Mengdi Wang. Feature-based q-learning for two-player stochastic games. arXiv preprint arXiv:1906.00423, 2019.

[24] Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-Łojasiewicz condition. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pages 795–811. Springer, 2016.

[25] G. M. Korpelevich. The extragradient method for finding saddle points and other problems. 1976.

[26] Qi Lei, Sai Ganesh Nagarajan, Ioannis Panageas, and Xiao Wang. Last iterate convergence in no-regret learning: constrained min-max optimization for convex-concave landscapes. arXiv preprint arXiv:2002.06768, 2020.

[27] Tengyuan Liang and James Stokes. Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks. In The 22nd International Conference on Artificial Intelligence and Statistics, 2019.

[28] Tianyi Lin, Chi Jin, Michael Jordan, et al. Near-optimal algorithms for minimax optimization. arXiv preprint arXiv:2002.02417, 2020.

[29] Michael L Littman. Markov games as a framework for multi-agent reinforcement learning. In Machine learning proceedings 1994, pages 157–163. Elsevier, 1994.

[30] Panayotis Mertikopoulos, Houssam Zenati, Bruno Lecouat, Chuan-Sheng Foo, Vijay Chandrasekhar, and Georgios Piliouras. Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile. In International Conference on Learning Representations, 2019.
[31] Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. arXiv preprint arXiv:1901.08511, 2019.

[32] John von Neumann. Zur theorie der gesellschaftsspiele. Mathematische annalen, 1928.

[33] Leonid Denisovich Popov. A modification of the arrow-hurwicz method for search of saddle points. Mathematical notes of the Academy of Sciences of the USSR, 1980.

[34] Goran Radanovic, Rati Devidze, David Parkes, and Adish Singla. Learning to collaborate in markov decision processes. In International Conference on Machine Learning, pages 5261–5270, 2019.

[35] Sasha Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. In Advances in Neural Information Processing Systems, pages 3066–3074, 2013.

[36] Vasilis Syrgkanis, Alekh Agarwal, Haipeng Luo, and Robert E Schapire. Fast convergence of regularized learning in games. In Advances in Neural Information Processing Systems, 2015.

[37] Paul Tseng. On linear convergence of iterative methods for the variational inequality problem. 1995.

[38] Chen-Yu Wei, Yi-Te Hong, and Chi-Jen Lu. Online reinforcement learning in stochastic games. In Advances in Neural Information Processing Systems, pages 4987–4997, 2017.

[39] Qiaomin Xie, Yudong Chen, Zhaoran Wang, and Zhuoran Yang. Learning zero-sum simultaneous-move markov games using function approximation and correlated equilibrium. arXiv preprint arXiv:2002.07066, 2020.

[40] Yasin Abbasi Yadkori, Peter L Bartlett, Varun Kanade, Yevgeny Seldin, and Csaba Szepesvári. Online learning in markov decision processes with adversarially chosen transition probability distributions. In Advances in neural information processing systems, pages 2508–2516, 2013.

[41] Hui Zhang and Wotao Yin. Gradient methods for convex minimization: better rates under weaker conditions. arXiv preprint arXiv:1303.4645, 2013.

[42] Junyu Zhang, Mingyi Hong, and Shuzhong Zhang. On lower iteration complexity bounds for the saddle point problems. arXiv preprint arXiv:1912.07481, 2019.
A Experiments

A.1 Matrix Games on Simplex

Here we provide empirical results on the performance of OGDA for matrix games on simplex. We set the size of the game matrix to be $32 \times 32$, then generate a random matrix with each entry $G_{ij}$ drawn uniformly at random from $[-1, 1]$, and finally rescale its operator norm to 1.

We compare the performances of OGDA with three other algorithms: Gradient Descent Ascent (GDA), Optimistic Multiplicative Weight Update (OMWU), and Multiplicative Weight Update (MWU). For all algorithms, we use a constant learning rate $\eta = \frac{1}{T}$. We set $T = 10^6$ and plot curves for $t = 1, 2, \ldots, 10^6$. Figure 1 shows the results. The $x$-axis of both plots represents time step $t$. The $y$-axis of the left plot represents $\ln(\|\hat{x}_t - z^*\|)$, the $\ell_2$-norm between $\hat{x}_t$ and $z^*$ after taking the natural logarithm. On the other hand, the $y$-axis of the right plot represents the logarithmic duality gap at each round $t$, which is

$$
\ln(\alpha_f(\hat{z}_t)) = \ln \left( \max_j (G^T \hat{x}_t)_j - \min_i (G^T \hat{y}_t)_i \right).
$$

Note that here we use the limit point of $\hat{z}_t$ of OGDA to be $z^*$. This is approximately calculated by running OGDA much more than $T$ iterations (e.g. $10T$ iterations). In fact, with probability 1, the game has a unique Nash Equilibrium [10]. In other words, $\text{dist}(z, \mathcal{Z}^*) = \|z - z^*\|^2$ with probability 1. We also empirically verify that the iterates of OMWU converge to the same point as OGDA. So the distance measure we use is a fair comparison.

From Figure 1, we have the following observations. First, considering the convergence to Nash Equilibrium, the curve of OGDA is a straight line when $t > 2 \times 10^5$, supporting the linear convergence of the distance shown in Theorem 8. Although the curve of OMWU is also decreasing, the rate of its convergence is slower than the one of OGDA. On the other hand, the curves of the other two non-optimistic algorithms (MWU and GDA) are almost parallel to the $x$-axis, which means these two algorithms do not enjoy last-iterate convergence. Second, for the duality gap, the graph is roughly the same as the one for point-wise convergence. One minor difference for the plot of duality gap is that the curve for OGDA is slightly oscillating, although it still converges linearly overall.

A.2 Matrix Game on Curved Regions

Next, we do experiments on a bilinear game similar to the one constructed in the proof of Theorem 9. Specifically, the bilinear game is defined by

$$
f(x, y) = x(2)y(1) - x(1)y(2), \quad \mathcal{X} = \mathcal{Y} \triangleq \{(a, b), 0 \leq a, b \leq 1, \ a^n \leq b\}.
$$

For any positive integer $n$, the equilibrium point of this game is $(0, 0)$ for both $x$ and $y$. Note that in Theorem 9, we prove that OGDA only converges at a rate no better than $\Omega(1/t^2)$ in this game with $\mathcal{X} = \mathcal{Y} \triangleq \{(a, b), 0 \leq a, b \leq \frac{1}{2}, \ a^2 \leq b\}$.

Figure 2 shows the empirical results for various values of $n$. In this figure, we plot $\|\hat{z}_t - z^*\|$ versus the time step $t$ in the log-log scale. Note that in a log-log plot, a straight line with slope $s$ implies a convergence rate of order $O(t^s)$, that is, a sublinear convergence rate. It is clear from Figure 2 that OGDA indeed converges sublinearly for all $n$, supporting our Theorem 9.

A.3 Strongly-convex-strongly-concave Games

In this section, we redo the experiments for strongly-convex-strongly-concave games in [26], where they consider the following case:

$$
f(x, y) = x(1)^2 - y(1)^2 + 2x(1)y(1), \quad \mathcal{X} = \mathcal{Y} \triangleq \{(a, b), 0 \leq a, b \leq 1, \ a + b = 1\}.
$$

The equilibrium point is $(0, 1)$ for both $x$ and $y$. In Figure 3, we present the log-log plot of $\|\hat{z}_t - z^*\|$ versus time step $t$ and compare OGDA to the other three algorithms as in Appendix A.1. The strictly concave curve of OGDA implies that the algorithm converges linearly, supporting Theorem 6 and Theorem 8. Also note that here, OGDA and GDA outperform MWU and OMWU, which is different from the empirical results shown in [26]. We hypothesize that this is because they use a different version of OGDA and GDA.

---

1 All source codes can be found at https://github.com/bahh723/OGDA-last-iterate.
Figure 1: Experiments of different algorithms on a matrix game $f(x, y) = x^\top G y$, where we generate $G \in \mathbb{R}^{32 \times 32}$ with each entry $G_{ij}$ drawn uniformly at random from $[-1, 1]$ and then rescale $G$'s operator norm to 1. The left graph shows the point-wise convergence results of MWU, OMWU, GDA and OGDA. The right graph shows the duality-gap convergence results of MWU, OMWU, GDA and OGDA. Both graphs show the linear convergence of OGDA.

Figure 2: Experiments of OGDA on matrix games with curved regions where $f(x, y) = x(2)y(1) - x(1)y(2)$, $\mathcal{X} = \{ (a, b), 0 \leq a, b \leq 1, a^n \leq b \}$, and $n = 2, 4, 6, 8$. This figure is a log-log plot of $\|z_t - z^*\|$ versus $t$, and it indicates sublinear convergence rates of OGDA in all these games.

Figure 3: Experiments on a strongly-convex-strongly-concave game where $f(x, y) = x(1)^2 - y(1)^2 + 2x(1)y(1)$, $\mathcal{X} = \{ (a, b), 0 \leq a, b \leq 1, a + b = 1 \}$. The figure is a log-log graph of $\|\tilde{z}_t - z^*\|$ versus the time step $t$. The result shows that GDA and OGDA outperform OMWU and MWU in this case.
Figure 4: Experiments of OGDA on a set of games satisfying SP-RSI-2 with \( \beta > 0 \), where \( f(x, y) = x(1)^2n - x(1)y(1) - y(1)^2n \) for some integer \( n \geq 2 \) and \( \mathcal{X} = \mathcal{Y} = \{(a, b), 0 \leq a, b \leq 1, a + b = 1\} \). The result shows that OGDA converges to the Nash equilibrium with sublinear rates in these instances.

### A.4 An Example with \( \beta > 0 \) for SP-RSI-2

We also consider the toy example in Theorem 7, where \( f(x, y) = x(1)^2n - x(1)y(1) - y(1)^2n \) for some integer \( n \geq 2 \) and \( \mathcal{X} = \mathcal{Y} = \{(a, b), 0 \leq a, b \leq 1, a + b = 1\} \). The equilibrium point is \((0, 1)\) for both \( x \) and \( y \). We prove in Theorem 7 that SP-RSI-2 does not hold for \( \beta = 0 \) but does hold for \( \beta = 2n - 2 \).

The point-wise convergence result is shown in Figure 4, which is again a log-log plot of \( \|\hat{z}_t - z^*\| \) versus time step \( t \). One can observe that the convergence rate of OGDA is sublinear, supporting our theory again.

### A.5 Stochastic Games

Finally, we show results for stochastic games. The algorithm we use is the original version of OGDA, that is, updating Eq. (11) and Eq. (13) with \( Q_{s,t} \) instead of \( Q_{s,t+1} \) and \( Q_{s,t+1} \). As discussed in Section 5, for the last-iterate convergence result, this does not matter.

In the following we use \([K]\) to denote \(\{1, 2, \ldots, K\}\) for some positive integer \(K\). In this experiment, we set \(H = 4\), \(A = B = [16]\), and \(S_i = [8]\) for each \(i \in [H]\). To expedite the convergence, we use \(\eta = \frac{1}{8}\) instead of the learning rate suggested in Theorem 10. Both \(\ell\) and \(p\) are randomly generated. Specifically, each \(\ell(s, a, b)\) is uniformly at random drawn from \([0, 1]\). To generate \(p(s'|s, a, b)\), we first choose \(p'(s'|s, a, b)\) from \([0, 1]\) uniformly at random, and then normalize it so that \(p' (\cdot | s, a, b)\) is a distribution.

The results are shown in Figure 5. Similar to Appendix A.1, we compare OGDA with MWU, OMWU, and GDA. Also, we show the convergence of the duality gap and the distance. The duality gap is defined as

\[
\alpha(\hat{z}_t) = \alpha(\hat{x}_t, \hat{y}_t) = \max_y \rho_{\hat{x}_t,y} - \min_x \rho_{x,\hat{y}_t},
\]

where \(\rho_{x,y} = \mathbb{E}_{s \sim p_0} V_{x,y}^s\) and \(p_0\) is set to be the uniform distribution supported on \(S_1\). On the other hand, the distance is defined as the sum of the distance between \(\hat{z}_t^*\) and \(z^*_t\) over all states, where \(z^*_t\) is calculated in a similar way as in Appendix A.1. The conclusion of the results is also similar to that in Appendix A.1, with OGDA outperforming others and enjoying a linear rate.
Figure 5: Experiments of OGDA and other algorithms on stochastic games. We set $H = 4$, $A = B = [16]$, and $S_i = [8]$ for each $i$. Each $\ell(s, a, b)$ is uniformly drawn from $[0, 1]$ at random. As for $p(s'|s, a, b)$, we first choose $p(s'|s, a, b)$ from $[0, 1]$ uniformly at random, and normalize them to form probability distributions. The result of OGDA is compared with MWU, OMWU, and GDA. Both of the point-wise convergence (left) and the duality gap convergence (right) are presented.

B Proofs of Lemma 2 and Lemma 3

We first provide two auxiliary lemmas.

Lemma 12. Let $A$ be a convex set, and let $u' = \Pi_A(u - g)$. Then for any $u^* \in A$,  
\[ 2(u' - u^*, g) \leq \|u^* - u\|^2 - \|u^* - u'\|^2 - \|u' - u\|^2. \]  
(15)

Proof. Since $u' = \text{argmin}_{u' \in A} \|u' - u + g\|^2$, by the first-order optimality condition for $u'$, we have  
\[ (u' - u + g)^\top (u^* - u') \geq 0. \]

Note that the right-hand side minus the left-hand side of Eq. (15) is exactly equal to $2(u' - u + g)^\top (u^* - u')$. Thus Eq. (15) holds. \hfill \square

Lemma 13. Let $u, u_1, u_2 \in A$ (a convex set) be related by the following:  
\[ u_1 = \Pi_A(u - g_1), \]
\[ u_2 = \Pi_A(u - g_2). \]

Then we have  
\[ \|u_1 - u_2\| \leq \|g_1 - g_2\|. \]

Proof. By the first-order optimality conditions of $u_1$ and $u_2$, we have  
\[ (u_1 - u + g_1) \cdot (u_2 - u_1) \geq 0, \]
\[ (u_2 - u + g_2) \cdot (u_1 - u_2) \geq 0. \]

Summing them up and rearranging, we get  
\[ -\|u_1 - u_2\|^2 + \langle u_2 - u_1, g_1 - g_2 \rangle \geq 0. \]

Rearranging, we get  
\[ \|u_1 - u_2\|^2 \leq \langle u_2 - u_1, g_1 - g_2 \rangle \leq \|u_1 - u_2\| \|g_1 - g_2\|. \]

Therefore,  
\[ \|u_1 - u_2\| \leq \|g_1 - g_2\|. \]  
\hfill \square

Proof of Lemma 2. Considering Eq. (5), and using Lemma 12 with $u = \tilde{z}_t$, $u' = \tilde{z}_{t+1}$, $u^* = z$, and $g = \eta F(z_t)$, we get  
\[ 2\eta F(z_t)^\top (\tilde{z}_{t+1} - z) \leq \|z - \tilde{z}_t\|^2 - \|z - \tilde{z}_{t+1}\|^2 - \|\tilde{z}_{t+1} - \tilde{z}_t\|^2. \]
Considering Eq. (6) (with \( t \) replaced with \( t - 1 \)), and using Lemma 12 with \( u = \tilde{z}_t, u' = z_t, u'' = \tilde{z}_{t+1}, \) and \( g = \eta F(z_{t-1}) \), we get
\[
2\eta F(z_{t-1})^\top (z_t - \tilde{z}_{t+1}) \leq \|z_{t-1} - \tilde{z}_t\|^2 - \|z_{t+1} - z_t\|^2 - \|z_t - \tilde{z}_t\|^2.
\]
Summing up the above two inequalities, and adding \( 2\eta (F(z_t) - F(z_{t-1}))^\top (z_t - \tilde{z}_{t+1}) \) to both sides, we get
\[
2\eta F(z_t)^\top (z_t - z) \leq \|z_t - \tilde{z}_t\|^2 - \|z - \tilde{z}_{t+1}\|^2 - \|z_{t+1} - z_t\|^2 - \|z_t - \tilde{z}_t\|^2 + 2\eta (F(z_t) - F(z_{t-1}))^\top (z_t - \tilde{z}_{t+1}).
\]
Using Lemma 13 with \( u = \tilde{z}_t, u_1 = z_t, u_2 = \tilde{z}_{t+1}, g_1 = \eta F(z_{t-1}) \) and \( g_2 = \eta F(z_t) \), we get
\[
\|z_t - \tilde{z}_{t+1}\| \leq \eta \|F(z_{t-1}) - F(z_t)\|.
\]
Therefore,
\[
2\eta (F(z_t) - F(z_{t-1}))^\top (z_t - \tilde{z}_{t+1}) \leq 2\eta^2 \|F(z_t) - F(z_{t-1})\|^2 \leq 2\eta^2 L^2 \|z_t - z_{t-1}\|^2 \quad \text{(by the smoothness assumption)}
\]
\[
\leq \frac{1}{32} \|z_t - z_{t-1}\|^2. \quad \text{(by our choice of \( \eta \))}
\]
Continuing from Eq. (16),
\[
2\eta F(z_t)^\top (z_t - z) \leq \|z_t - \tilde{z}_t\|^2 - \|z - \tilde{z}_{t+1}\|^2 - \|z_{t+1} - z_t\|^2 - \|z_t - \tilde{z}_t\|^2 + \frac{1}{32} \|z_t - z_{t-1}\|^2
\]
\[
\leq \|z_t - \tilde{z}_t\|^2 - \|z - \tilde{z}_{t+1}\|^2 - \|z_{t+1} - z_t\|^2 - \|z_t - \tilde{z}_t\|^2 + \frac{1}{16} \|z_t - z_{t-1}\|^2 + \frac{1}{16} \|z_t - z_{t-1}\|^2
\]
(\text{using } (a + b)^2 \leq 2(a^2 + b^2))
\[
\leq \|z_t - \tilde{z}_t\|^2 - \|z - \tilde{z}_{t+1}\|^2 - \|z_{t+1} - z_t\|^2 - \frac{15}{16} \|z_t - z_{t-1}\|^2 + \frac{1}{16} \|z_t - z_{t-1}\|^2
\]
This concludes the proof. \( \Box \)

Proof of Lemma 3. Below we consider any \( z' \neq \tilde{z}_{t+1} \in Z \). Considering Eq. (5), and using the first-order optimality condition of \( \tilde{z}_{t+1} \), we have
\[
(\tilde{z}_{t+1} - \tilde{z}_t + \eta F(z_t))^\top (z' - \tilde{z}_{t+1}) \geq 0.
\]
Rearranging it we get
\[
(\tilde{z}_{t+1} - \tilde{z}_t)^\top (z' - \tilde{z}_{t+1}) \geq \eta F(z_t)^\top (\tilde{z}_{t+1} - z')
\]
\[
= \eta F(\tilde{z}_{t+1})^\top (\tilde{z}_{t+1} - z') + \eta (F(z_t) - F(\tilde{z}_{t+1}))^\top (\tilde{z}_{t+1} - z')
\]
\[
\geq \eta F(\tilde{z}_{t+1})^\top (\tilde{z}_{t+1} - z') - \eta L \|z_t - \tilde{z}_{t+1}\| \|z_{t+1} - z'\|
\]
\[
\geq \eta F(\tilde{z}_{t+1})^\top (\tilde{z}_{t+1} - z') - \frac{1}{8} \|z_t - \tilde{z}_{t+1}\| \|z_{t+1} - z'\|
\]
where the third step uses Hölder’s inequality and the smoothness condition, and the last steps uses the condition \( \eta \leq 1/(8L) \). Upper bounding the left-hand side by \( \|\tilde{z}_{t+1} - \tilde{z}_t\| \|\tilde{z}_{t+1} - z'\| \) and rearranging, we get
\[
\|\tilde{z}_{t+1} - z'\| \left( \|\tilde{z}_{t+1} - \tilde{z}_t\| + \frac{1}{8} \|z_t - \tilde{z}_{t+1}\| \right) \geq \eta F(\tilde{z}_{t+1})^\top (\tilde{z}_{t+1} - z'),
\]
and thus
\[
\left( \|\tilde{z}_{t+1} - \tilde{z}_t\| + \frac{1}{8} \|z_t - \tilde{z}_{t+1}\| \right)^2 \geq \frac{\eta^2 \|F(\tilde{z}_{t+1})^\top (\tilde{z}_{t+1} - z')\|^2}{\|\tilde{z}_{t+1} - z'\|^2}
\]
Finally, noticing that
\[
\left( \|\tilde{z}_{t+1} - \tilde{z}_t\| + \frac{1}{8} \|z_t - \tilde{z}_{t+1}\| \right)^2 \leq \left( \|z_t - \tilde{z}_t\| + \frac{9}{8} \|z_t - \tilde{z}_{t+1}\| \right)^2
\]
\[
\leq \left( \frac{9}{8} \|z_t - \tilde{z}_t\| + \frac{9}{8} \|z_t - \tilde{z}_{t+1}\| \right)^2 \leq \frac{81}{32} \left( \|z_t - \tilde{z}_t\|^2 + \|z_t - \tilde{z}_{t+1}\|^2 \right)
\]
finishes the proof. \( \Box \)
C Proof of Theorem 4

Proof of Theorem 4. By Lemma 2, summing over \( t \) (from 1 to \( T - 1 \)), telescoping, and realizing that \( \mathcal{Z} \) has diameter at most 1 according to Assumption 2, we get

\[
\sum_{t=1}^{T-1} 2\eta F(z_t)^\top (z_t - z) \leq \frac{17}{16} - \frac{15}{16} \sum_{t=1}^{T-1} (\|\hat{z}_{t+1} - z_t\|^2 + \|z_t - \hat{z}_t\|^2),
\]

for any \( z \in \mathcal{Z} \). In particular, let \( z \) be any element from \( \mathcal{Z}^* \). Then the left-hand side is lower bounded by zero since,

\[
F(z_t)^\top (z_t - z) \geq f(x_t, y_t) - f(x, y_t) + f(x_t, y) - f(x_t, y_t) = f(x_t, y) - f(x, y_t) \geq 0 \tag{17}
\]

by convexity/concavity and the optimality of \( z \). Rearranging then leads to

\[
\sum_{t=1}^{T-1} (\|\hat{z}_{t+1} - z_t\|^2 + \|z_t - \hat{z}_t\|^2) \leq \frac{17}{15}.
\]

On the other hand, by Lemma 3, Lemma 1, and again the bounded diameter of \( \mathcal{Z} \), we have

\[
\|\hat{z}_{t+1} - z_t\|^2 + \|z_t - \hat{z}_t\|^2 \geq \frac{32}{81}\eta^2 \max_{z \in \mathcal{Z}} \left( \frac{[F(\hat{z}_{t+1})]^\top (\hat{z}_{t+1} - z')]^2}{\|\hat{z}_{t+1} - z\|^2} \right) \geq \frac{32\eta^2}{81} \alpha_f(\hat{z}_{t+1})^2.
\]

Summing over \( t \) from 0 to \( T - 1 \) and combining with the earlier upper bound we arrive at

\[
\sqrt{T} \sum_{t=1}^{T} \alpha_f(\hat{z}_t)^2 = O\left(1/\eta^2\right).
\]

Finally using Cauchy-Schwarz’s inequality \( \sum_{t=1}^{T} \alpha_f(\hat{z}_t)^2 \leq \sqrt{T} \sqrt{\sum_{t=1}^{T} \alpha_f(\hat{z}_t)^2} \) finishes the proof. \( \square \)

D Proof of Theorem 5

Proof of Theorem 5. Let \( \rho = \min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} x^\top G y = \max_{y \in \mathcal{Y}} \min_{x \in \mathcal{X}} x^\top G y \) be the game value. In this proof, we will prove that there exists some \( C > 0 \) such that

\[
\max_{y \in \mathcal{Y}} x^\top G y' - \rho \geq C\|x - \Pi_{\mathcal{X}^*}(x)\| \tag{18}
\]

for all \( x \in \mathcal{X} \). Similarly we can prove

\[
\max_{x \in \mathcal{X}} \rho - x^\top G y \geq C\|y - \Pi_{\mathcal{Y}^*}(y)\|, \tag{19}
\]

for all \( y \in \mathcal{Y} \). Then combining the two proves

\[
\max_{z'} F(z)^\top (z - z') = \max_{y'} x^\top G y' - \min_{x'} x'^\top G y \geq C (\|y - \Pi_{\mathcal{Y}^*}(y)\| + \|x - \Pi_{\mathcal{X}^*}(x)\|)
\]

\[
\geq C \|z - \Pi_{\mathcal{Z}^*}(z)\|,
\]

meaning that SP-RSI-1 holds with \( \beta = 0 \).

We break the proof into following several claims.

Claim 1. If \( \mathcal{X}, \mathcal{Y} \) are polytopes, then \( \mathcal{X}^* \) and \( \mathcal{Y}^* \) are also polytopes.

Proof of Claim 1. Note that \( \mathcal{X}^* = \{x \in \mathcal{X} : \max_{y \in \mathcal{Y}} x^\top G y \leq \rho\} \). Since \( \mathcal{Y} \) is a polytope, the maximum is attained at vertices of \( \mathcal{Y} \). Therefore, \( \mathcal{X}^* \) can be equivalently written as \( \{x \in \mathcal{X} : \max_{y \in \mathcal{V}(\mathcal{Y})} x^\top G y \leq \rho\} \) where \( \mathcal{V}(\mathcal{Y}) \) is the set of vertices of \( \mathcal{Y} \). Since the constraints of \( \mathcal{X}^* \) are all linear constraints, \( \mathcal{X}^* \) is a polytope. \( \square \)

With Claim 1, we can without loss of generality write \( \mathcal{X}^* \) as

\[
\mathcal{X}^* = \{x \in \mathbb{R}^M : a_i^\top x \leq b_i, \text{ for } i = 1, \ldots, L, c_i^\top x \leq d_i, \text{ for } i = 1, \ldots, K\},
\]

where the \( a_i^\top x \leq b_i \) constraints come from \( x \in \mathcal{X} \) and the \( c_i^\top x \leq d_i \) constraints come from \( \max_{y \in \mathcal{V}(\mathcal{Y})} x^\top G y \leq \rho \). Below, we refer to \( a_i^\top x \leq b_i \) as the feasibility constraints, and \( c_i^\top x \leq d_i \) as
as the optimality constraints. In fact, one can identify the $i$-th optimality constraint as $c_i = G y^{(i)}$ and $d_i = \rho$ where $y^{(i)}$ is the $i$-th vertex of $\mathcal{Y}$. This is based on our construction of $X^*$ in the proof of Claim 1. Therefore, $K = |\mathcal{Y}|$.

Since Eq. (18) clearly holds for $x \in X^*$, below, we focus on an $x \in X \setminus X^*$, and let $x^* \triangleq \Pi_{X^*}(x)$.

We say a constraint is tight we also have the additional constraints that $c_i x^* = d_i$Below we assume that there are $\ell$ tight feasibility constraints at and $k$ tight optimality constraints at $x^*$. Without loss of generality, we assume these tight constraints correspond to $i = 1, \ldots, \ell$ and $i = 1, \ldots, k$ respectively. That is,

$$a_i^T x^* = b_i, \quad \text{for } i = 1, \ldots, \ell,$$

$$c_i^T x^* = d_i, \quad \text{for } i = 1, \ldots, k.$$

**Claim 2.** $x$ violates at least one of the tight optimality constraint at $x^*$.

**Proof of Claim 2.** We prove this by contradiction. Suppose that $x$ satisfies all $k$ tight optimality constraints at $x^*$. Then $x$ must violates some of the remaining $K - k$ optimality constraints (otherwise $x \in X^*$). Assume that it violates constraints $K - n + 1, \ldots, K$ for some $1 \leq n \leq K - k$. Thus we have the following:

$$c_i^T x \leq d_i \quad \text{for } i = 1, \ldots, K - n;$$

$$c_i^T x > d_i \quad \text{for } i = K - n + 1, \ldots, K.$$

Recall that $c_i^T x^* \leq d_i$ for $i = 1, \ldots, K - n$ and $c_i^T x^* < d_i$ for all $i = K - n + 1, \ldots, K$. Thus there exists some $x'$ that lies strictly between $x$ and $x^*$ that makes all constraints hold (notice that $x$ and $x^*$ both satisfy all feasibility constraints), which contradicts with $\Pi_{X^*}(x) = x^*$. \hfill \Box

**Claim 3.** $\max_{y' \in \mathcal{Y}} (x^T G y' - \rho) \geq \max_{i \in \{1, \ldots, k\}} c_i^T (x - x^*)$.

**Proof of Claim 3.** Recall that we identify $c_i$ with $G y^{(i)}$ and $d_i = \rho$. Therefore,

$$\max_{y' \in \mathcal{Y}} (x^T G y' - \rho) = \max_{i \in \{1, \ldots, |\mathcal{Y}|\}} (c_i^T x - d_i) \geq \max_{i \in \{1, \ldots, k\}} (c_i^T x - d_i) = \max_{i \in \{1, \ldots, k\}} c_i^T (x - x^*),$$

where the last equality is because $c_i^T x^* = d_i$ for $i = 1, \ldots, k$. \hfill \Box

Recall from linear programming literature [13, 14] that the normal cone of $X^*$ at $x^*$ can be expressed as follows:

$$\mathcal{N}_{x^*} = \left\{ x' - x^* : x' \in \mathbb{R}^M, \quad \Pi_{X^*}(x') = x^* \right\} = \left\{ \sum_{i=1}^\ell p_i a_i + \sum_{i=1}^k q_i c_i : p_i \geq 0, \quad q_i \geq 0 \right\}.$$

The normal cone of $X^*$ at $x^*$ consists of all outgoing normal vectors of $X^*$ originated from $x^*$. Clearly, $x - x^*$ belongs to $\mathcal{N}_{x^*}$. However, besides the fact that $x - x^*$ is a normal vector of $X^*$, we also have the additional constraints that $x \in \mathcal{X}$. We claim that in our case, $x - x^*$ lies in the following smaller cone (which is a subset of $\mathcal{N}_{x^*}$):

**Claim 4.** $x - x^*$ belongs to

$$\mathcal{M}_{x^*} = \left\{ \sum_{i=1}^\ell p_i a_i + \sum_{i=1}^k q_i c_i : p_i \geq 0, \quad q_i \geq 0, \quad a_j^T \left( \sum_{i=1}^\ell p_i a_i + \sum_{i=1}^k q_i c_i \right) \leq 0, \quad \forall j = 1, \ldots, \ell \right\}.$$

**Proof of Claim 4.**

As argued above, $x - x^* \in \mathcal{N}_{x^*}$, and thus $x - x^*$ can be expressed as $\sum_{i=1}^\ell p_i a_i + \sum_{i=1}^k q_i c_i$ with $p_i \geq 0, q_i \geq 0$. To prove that $x - x^* \in \mathcal{M}_{x^*}$, we only need to prove that it satisfies the additional constraints, that is,

$$a_i^T (x - x^*) \leq 0, \quad \forall i = 1, \ldots, \ell.$$
This can be shown by noticing that for all $i = 1, \ldots, \ell$,
\[
\begin{align*}
  a_i^\top (x - x^*) &= (a_i^\top x^* - b_i) + a_i^\top (x - x^*) \\
  &= a_i^\top (x^* + x - x^*) - b_i \\
  &= a_i^\top x - b_i \leq 0. 
\end{align*}
\]  
(the $i$-th constraint is tight at $x^*$)
\[x \in \mathcal{X}\]  
\[\square\]

**Claim 5.** $x - x^*$ can be written as $\sum_{i=1}^\ell p_i a_i + \sum_{i=1}^k q_i c_i$ with $0 \leq p_i, q_i \leq C'\|x - x^*\|$ for all $i$ and some problem-dependent constant $C' < \infty$.

**Proof of Claim 5.** Notice that for any $x \notin \mathcal{X}^*$, $\frac{x - x^*}{\|x - x^*\|} \in \mathcal{M}_{x^*}$ (because $x - x^* \in \mathcal{M}_{x^*}$ and $\mathcal{M}_{x^*}$ is a cone). Furthermore, $\frac{x - x^*}{\|x - x^*\|} \in \{ v \in \mathbb{R}^M : \|v\|_\infty \leq 1 \}$. Therefore $\frac{x - x^*}{\|x - x^*\|} \in \mathcal{M}_{x^*} \cap \{ v \in \mathbb{R}^M : \|v\|_\infty \leq 1 \}$, which is a bounded subset of the cone $\mathcal{M}_{x^*}$.

Below we argue that there exists a large enough $C' > 0$ such that
\[
\left\{ \sum_{i=1}^\ell p_i a_i + \sum_{i=1}^k q_i c_i : 0 \leq p_i, q_i \leq C', \forall i \right\} \supseteq \mathcal{M}_{x^*} \cap \{ v \in \mathbb{R}^M : \|v\|_\infty \leq 1 \} \triangleq \mathcal{P}. 
\]

To see this, first note that $\mathcal{P}$ is a polytope. For every vertex $\tilde{v}$ of $\mathcal{P}$, the smallest $C'$ such that $\tilde{v}$ belongs to the left-hand side is the solution of the following linear programming:
\[
\min_{p_i, q_i, C'_0} C'_0 \quad \text{s.t.} \quad \tilde{v} = \sum_{i=1}^\ell p_i a_i + \sum_{i=1}^k q_i c_i, \quad 0 \leq p_i, q_i \leq C'_0.
\]

Since $\tilde{v} \in \mathcal{M}_{x^*}$, this linear programming is always feasible and admits a finite solution $C'_0 < \infty$. Now let $C' = \max_{v \in \mathcal{V}(\mathcal{P})} C'_v$ where $\mathcal{V}(\mathcal{P})$ is the set of all vertices of $\mathcal{P}$. Then since any $v \in \mathcal{P}$ can be expressed as a convex combination of points in $\mathcal{V}(\mathcal{P})$, $v$ can be also be expressed as $\sum_{i=1}^\ell p_i a_i + \sum_{i=1}^k q_i c_i$ with $0 \leq p_i, q_i \leq C'$.

To sum up, $\frac{x - x^*}{\|x - x^*\|}$ can be represented as $\sum_{i=1}^\ell p_i a_i + \sum_{i=1}^k q_i c_i$ with $0 \leq p_i, q_i \leq C'$. This further implies that $x - x^*$ can be represented as $\sum_{i=1}^\ell p_i a_i + \sum_{i=1}^k q_i c_i$ with $0 \leq p_i, q_i \leq C'\|x - x^*\|$. Notice that $C'$ only depends on the set of tight constraints at $x^*$.

Finally, we are ready to combine all previous claims and prove the desired inequality.

Define $A_i \triangleq a_i^\top (x - x^*)$ and $C_i \triangleq c_i^\top (x - x^*)$. By Claim 5, we can write $x - x^*$ as $\sum_{i=1}^\ell p_i a_i + \sum_{i=1}^k q_i c_i$ with $0 \leq p_i, q_i \leq C'\|x - x^*\|$, and thus,
\[
\sum_{i=1}^\ell p_i A_i + \sum_{i=1}^k q_i C_i = \left( \sum_{i=1}^\ell p_i a_i + \sum_{i=1}^k q_i c_i \right)^\top (x - x^*) = \|x - x^*\|^2.
\]

On the other hand, since $x - x^* \in \mathcal{M}_{x^*}$, by Claim 4, we have
\[
\sum_{i=1}^\ell p_i A_i = \sum_{i=1}^\ell p_i a_i^\top (x - x^*) \leq 0
\]
and
\[
\sum_{i=1}^k q_i C_i \leq \left( \max_{i \in \{1, \ldots, k\}} C_i \right) \sum_{i=1}^k q_i \leq \left( \max_{i \in \{1, \ldots, k\}} C_i \right) k C'\|x - x^*\|,
\]
where in the first inequality we use the fact $p_i \geq 0$, and in the second inequality we use the fact $\max_{i \in \{1, \ldots, k\}} C_i > 0$ (by Claim 2) and $0 \leq q_i \leq k C'\|x - x^*\|$.
Combining the above three inequalities, we get
\[
\max_{i \in \{1, \ldots, k\}} C_i \geq \frac{1}{kC'}\|x - x^*\|.
\]

Then by Claim 3,
\[
\max_{y' \in Y} (x^T Gy' - \rho) \geq \max_{i \in \{1, \ldots, k\}} C_i \geq \frac{1}{kC'}\|x - x^*\|.
\]

Note that \(k\) and \(C'\) only depend on the set of tight constraints at the projection point \(x^*\), and there are only finitely many different sets of tight constraints. Therefore, we conclude that there exists a constant \(C > 0\) such that \(\max_{y' \in Y} (x^T Gy' - \rho) \geq C\|x - x^*\|\) holds for all \(x\) and \(x^*\), which completes the proof. \(\square\)

\[\text{E Proof of Theorem 6 and Theorem 7}\]

\[\text{Proof of Theorem 6.}\] Suppose that \(f\) is \(\gamma\)-strongly-convex in \(x\), and \(\gamma\)-strongly-concave in \(y\), and let \((x^*, y^*) \in Z^*\). Then for any \((x, y)\) we have
\[
f(x, y) - f(x^*, y) \leq \nabla_x f(x, y)^	op (x - x^*) - \frac{\gamma}{2}\|x - x^*\|^2;
\]
\[
f(x, y^*) - f(x, y) \leq \nabla_y f(x, y)^	op (y^* - y) - \frac{\gamma}{2}\|y - y^*\|^2.
\]

Summing up the two inequalities, and noticing that \(f(x, y^*) - f(x^*, y) \geq 0\) for any \((x^*, y^*) \in Z^*\), we get
\[
F(z) (z - z^*) \geq \frac{\gamma}{2}\|z - z^*\|^2,
\]
which implies \(\text{SP-RSI-2 with } \beta = 0\) and \(C = \gamma/2\). \(\square\)

\[\text{Proof of Theorem 7.}\] First, we show that \(f\) has a unique Nash Equilibrium \(z^* = (x^*, y^*) = ((0, 1); (0, 1))\). First, as \(f\) is a strictly monotone decreasing function with respect to \(g(1)\), we must have \(y^*(1) = 0 \text{ and } y^*(2) = 1\). In addition, if \(x = (0, 1), \max_{y \in Y} f(x, y) = -\min_{y \in Y} y(1)^{2n} = 0\). If \(x \neq (0, 1), \) then by choosing \(y^* = (0, 1), f(x, y^*) = x(1)^{2n} > 0\). Therefore, we have \(x^* = (0, 1), \) which proves that the unique Nash Equilibrium is \(x^* = (0, 1), y^* = (0, 1)\).

Second, we show that \(f\) satisfies \(\text{SP-RSI-2 with } \beta = 2n - 2\). In fact, for any \(z = (x, y)\), we have
\[
F(z) (z - z^*) = \begin{bmatrix} 2nx(1)^{2n-1} - y(1) \\ 2ny(1)^{2n-1} + x(1) \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} x(1) \\ x(2) - 1 \\ y(1) \\ y(2) - 1 \end{bmatrix} = 2n\left(x(1)^{2n} + y(1)^{2n}\right) \geq 4n \cdot \frac{(x(1)^2 + y(1)^2)^n}{2^n} \quad \text{(Jensen's inequality)}
\]

Note that \(\|z - z^*\| = \sqrt{x(1)^2 + (1 - x(2))^2 + y(1)^2 + (1 - y(2))^2} = \sqrt{2x(1)^2 + 2y(1)^2}.\) Therefore, we have \(F(z) (z - z^*) \geq \frac{n}{2^{n-2}}\|z - z^*\|^2\). This shows that \(f\) satisfies \(\text{SP-RSI-2 with } \beta = 2n - 2\) and \(C = \frac{n}{2^{n-2}}\). Clearly, the same cannot always hold true for \(\beta = 0\). \(\square\)

\[\text{F Proof of Theorem 8}\]

\[\text{Proof of Theorem 8.}\] We use Lemma 2, and simplify the notations with \(\Omega_t \triangleq \text{dist}(\tilde{z}_t, Z^*) = \|\Pi_{Z^*}(\tilde{z}_t) - \tilde{z}_t\|^2, \Omega_{t+1} \triangleq \|\Pi_{Z^*}(\tilde{z}_t) - \tilde{z}_{t+1}\|^2\), and \(\theta_t = \frac{1}{16}\|\tilde{z}_t - z_{t-1}\|^2\). By the definition of projection, we have \(\Omega_{t+1} \leq \Omega_t\).
Notice that for any $z \in \mathbb{Z}^*$, $F(z_t) \top (z_t - z) \geq 0$ (see Eq. (17)), so by Lemma 2 we have

$$\Omega_{t+1} \leq \Omega_{t+1}^{'}$$

$$= \Omega_t - \frac{1}{2} \left( \| \tilde{z}_{t+1} - z_t \|^2 + \| z_t - \hat{z}_t \|^2 \right) - \frac{7}{16} \| \tilde{z}_t - z_{t-1} \|^2 + \frac{1}{16} \| \tilde{z}_t - z_{t-1} \|^2$$

$$\leq \Omega_t - \frac{1}{2} \left( \| \tilde{z}_{t+1} - z_t \|^2 + \| z_t - \hat{z}_t \|^2 \right) - \frac{7}{16} \| \tilde{z}_t - z_{t-1} \|^2 + \frac{1}{16} \| \tilde{z}_t - z_{t-1} \|^2$$

(20)

**Case 1. SP-RSI-1.** In this case, we apply Lemma 3 with the fact $\| \tilde{z}_{t+1} - z \| \leq 1$ for any $z' \in \mathbb{Z}$ to arrive at

$$\| \tilde{z}_{t+1} - z_t \|^2 + \| z_t - \hat{z}_t \|^2 \geq \frac{32\eta^2}{81} \max_{z' \in \mathbb{Z}} \left[ F(\tilde{z}_{t+1}) \top (\tilde{z}_{t+1} - z') \right]^2$$

$$\geq \frac{32\eta^2 C^2}{81} \| \tilde{z}_{t+1} - \Pi \mathbb{Z}^* (\tilde{z}_{t+1}) \|^2 (\beta+1) = \frac{32\eta^2 C^2}{81} \Omega_{t+1}^{\beta+1}$$

where in the second line we use the definition of SP-RSI-1. Combining this with Eq. (20), we get

$$\Omega_{t+1} + 8\theta_{t+1} \leq \Omega_t - \frac{16\eta^2 C^2}{81} \Omega_{t+1}^{\beta+1} + \theta_t \leq \Omega_t - C' \Omega_{t+1}^{\beta+1} + \theta_t,$$

(21)

where $C' = \frac{16\eta^2 \min \{ C^2, \beta \} \mathbb{L}^2}{81 (\beta+1) 2^\beta} \leq 1$ is defined in the statement of the theorem. Eq. (21) is of the form specified in Lemma 14 and Lemma 15 for the case of $\beta = 0$ and $\beta > 0$ respectively. Therefore, when $\beta = 0$, we apply Lemma 14 with $q = C'$, proving Eq. (8); and when $\beta > 0$, we apply Lemma 15 with $p = \beta, q = C'$ and note that the required condition $C' (\beta+1) 2^\beta \leq 1$ is satisfied, proving Eq. (9).

**Case 2. SP-RSI-2.** In this case, if $\tilde{z}_{t+1} \notin \mathbb{Z}^*$, we apply Lemma 3 with $z' = \Pi \mathbb{Z}^* (\tilde{z}_{t+1}) \neq \tilde{z}_{t+1}$ and get

$$\| \tilde{z}_{t+1} - z_t \|^2 + \| z_t - \hat{z}_t \|^2 \geq \frac{32\eta^2}{81} \left[ F(\tilde{z}_{t+1}) \top (\tilde{z}_{t+1} - \Pi \mathbb{Z}^* (\tilde{z}_{t+1})) \right]^2$$

$$\geq \frac{32\eta^2 C^2}{81} \Omega_{t+1}^{\beta+1} \leq \frac{32\eta^2 C^2}{81} \Omega_{t+1}^{\beta+1},$$

where the second step uses the definition of SP-RSI-2. Note that this also trivially holds when $\tilde{z}_{t+1} \in \mathbb{Z}^*$ since in this case $\Omega_{t+1} = 0$. We have thus arrived at the exact same bound as in Case 1, and the rest of the proof is identical.

**Lemma 14.** Consider non-negative sequences $\{ A_t \}$ and $\{ \theta_t \}$ that satisfy $A_{t+1} + 8\theta_{t+1} \leq A_t - qA_{t+1} + \theta_t, A_t \leq 1, \theta_t \leq \frac{1}{16}$ for all $t \geq 1$ and some $0 < q \leq 1$. Then we have

$$A_t \leq 3 \left( \frac{1}{1+q} \right)^t.$$

**Proof.** Rearranging the recursion formula we get

$$A_{t+1} + \frac{8}{1+q} \theta_{t+1} \leq \frac{1}{1+q} A_t + \frac{1}{1+q} \theta_t \leq \frac{1}{1+q} \left( A_t + \frac{8}{1+q} \theta_t \right),$$

where we use $\frac{1}{1+q} \leq \frac{8}{(1+q)^2}$ because $0 < q \leq 1$. By induction we can prove that $A_t \leq A_t + \frac{8}{1+q} \theta_t \leq \left( \frac{1}{1+q} \right)^{t-1} \left( A_1 + \frac{8}{1+q} \theta_1 \right) \leq 3 \left( \frac{1}{1+q} \right)^t.$

**Lemma 15.** Consider non-negative sequences $\{ A_t \}$ and $\{ \theta_t \}$ that satisfy $A_{t+1} + 8\theta_{t+1} \leq A_t - qA_{t+1} + \theta_t, A_t \leq 1, \theta_t \leq \frac{1}{16}$ for all $t \geq 1$ and some $p > 0, 0 < q \leq 1$ such that $q(1+p)2^p \leq 1$. Then we have

$$A_t \leq \left( 2 + \left( \frac{2}{qp} \right)^{\frac{1}{p}} \right)^{t-\frac{1}{p}}.$$
Proof. We will invoke the following formula that is given by the fundamental theorem of calculus: for \(a \geq b \geq 0\),
\[
d^p - b^p + 1 = \int_b^a \left( \frac{d}{dx}x^p \right) dx = (p + 1) \int_b^a x^p dx \leq (p + 1)(a - b)a^p. \quad (22)
\]
Using it, we have
\[
(A_t + 4\theta_t)^p + 1 \leq 4(1 + p)\theta_t(A_t + 4\theta_t)^p \leq (p + 1)\theta_t2^{p+2}.
\]
Therefore, using the recursion condition and \(q(1 + p)2^p \leq 1\) implies
\[
A_{t+1} + 8\theta_{t+1} \leq A_t - qA_{t+1}^p + \theta_t
\leq A_t - q(A_{t+1} + 4\theta_{t+1})^p + q(1 + p)2^{p+2}\theta_{t+1} + \theta_t
\leq A_t - q(A_{t+1} + 4\theta_{t+1})^p + 4\theta_{t+1} + \theta_t,
\]
which further gives
\[
A_{t+1} + 4\theta_{t+1} \leq A_t - 4\theta_{t+1} - q(A_{t+1} + 4\theta_{t+1})^p + 1.
\]
Below we define \(B_t = A_t + 4\theta_{t+1}\) such that \(B_{t+1} \leq B_t - qB_{t+1}^p\). Next we will prove \(B_{t+1} \leq B_t - \frac{q}{2}B_t^p\) by considering the following two cases.

Case 1. \(B_{t+1} \geq B_t\). In this case, clearly we have
\[
B_{t+1} \leq B_t - qB_{t+1}^p \leq B_t - qB_t^p \leq B_t - \frac{q}{2}B_t^p.
\]

Case 2. \(B_{t+1} \leq B_t\). In this case, we use Eq. (22) and get
\[
B_{t+1} \leq B_t - qB_{t+1}^p \leq B_t - qB_t^p + q(p + 1)(B_t - B_{t+1})B_t^p.
\]
By rearranging, we get
\[
B_{t+1} \leq \left(1 - \frac{qB_t^p}{1 + q(p + 1)B_t^p}\right) B_t.
\]
Using the fact \(q(p + 1)B_t^p \leq q(p + 1)2^p\), which is at most 1 by the condition of the lemma, we further get
\[
B_{t+1} \leq \left(1 - \frac{qB_t^p}{2}\right) B_t.
\]
Combining the two cases we have shown \(B_{t+1} \leq B_t - \frac{q}{2}B_t^p\). Below we use induction to prove \(B_t \leq ct^{-\frac{1}{p}}\) where \(c = \max \left\{2, \left(\frac{2}{q(p + 1)}\right)^{\frac{1}{p}}\right\}\). When \(t = 1\), we have \(B_1 \leq 2 \leq c\). Suppose that it holds for \(1, \ldots, t\). Note that the function \(f(B_t) = (1 - \frac{q}{2}B_t^p)B_t\) is increasing in \(B_t\) as \(f'(B_t) = 1 - \frac{q(p + 1)}{2}B_t^p \geq 0\) by the condition \(q(p + 1)2^p \leq 1\) and the fact \(B_t \leq 2\). Therefore, we apply the induction hypothesis and get
\[
B_{t+1} \leq \left(1 - \frac{q}{2}B_t^p\right) B_t
\leq \left(1 - \frac{q}{2}e^{pt-1}\right)ct^{-\frac{1}{p}}
= ct^{-\frac{1}{p}} - \frac{q}{2}ct^{-1-\frac{1}{p}}
\leq ct^{-\frac{1}{p}} - \frac{c}{p}t^{-1-\frac{1}{p}}
\leq c(t - 1)^{-\frac{1}{p}},
\]
where the last inequality is by the fundamental theorem of calculus:
\[
t^{-\frac{1}{p}} - (1 + t)^{-\frac{1}{p}} = \int_{1+t}^{t} \left( \frac{d}{dx}x^{-\frac{1}{p}} \right) dx = \int_{1+t}^{t} \left( -\frac{1}{p} \right) x^{-1-\frac{1}{p}} dx
= \int_{t}^{t+1} \frac{1}{p}x^{-1-\frac{1}{p}} dx \leq \frac{1}{p}t^{-1-\frac{1}{p}}.
\]
Since \(B_t\) is an upper bound of \(A_t\), we have also obtained the same upper bound for \(A_t\), completing the proof. \(\square\)
G Proof of Theorem 9

Proof of Theorem 9. Consider the following $2 \times 2$ bilinear game with curved feasible sets (we use $x(i)$ to denote the i-th coordinate of $x$):

$$f(x, y) = x^T G y = [x(1), x(2)] \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} y(1) \\ y(2) \end{bmatrix}$$

$$\mathcal{X} = \{ x : 0 \leq x(1) \leq \frac{1}{2}, \quad 0 \leq x(2) \leq \frac{1}{4}, \quad x(2) \geq x(1)^2 \}.$$

$$\mathcal{Y} = \{ y : 0 \leq y(1) \leq \frac{1}{2}, \quad 0 \leq y(2) \leq \frac{1}{4}, \quad y(2) \geq y(1)^2 \}.$$

Below, we use Claim 1 - Claim 4 to argue that if the two players start from $x_0 = y_0 = \hat{x}_0 = \hat{y}_0 = \left(\frac{1}{2}, \frac{1}{4}\right)$, and use any constant learning rate $\eta$, then the convergence is sublinear in the sense that $\|z_t - z^*\| \geq \Omega(1/t)$.

Claim 1. The unique equilibrium is $x^* = 0$, $y^* = 0$.

When $x = 0$, clearly $\max_{y' \in \mathcal{Y}} f(x, y') = 0$. Below we prove that if $x \neq 0$, then $\max_{y' \in \mathcal{Y}} f(x, y') > 0$. If $x(1) \neq 0$, we let $y'(1) = \frac{1}{2} x(1)$ and $y'(2) = \frac{1}{4} x(1)^2$ (which satisfies $y' \in \mathcal{Y}$), and thus

$$f(x, y') = x(2) y'(1) - x(1) y'(2) = x(1)^2 \cdot \frac{1}{2} x(1) - x(1) \cdot \frac{1}{4} x(1)^2 = \frac{1}{4} x(1)^3 > 0.$$ 

If $x(1) = 0$ but $x(2) \neq 0$, we let $y'(1) = \frac{1}{2}$, $y'(2) = \frac{1}{4}$, and thus

$$f(x, y') = x(2) y'(1) - x(1) y'(2) = \frac{1}{2} x(2) > 0.$$ 

Thus $\max_{y' \in \mathcal{Y}} f(x, y') > 0$ if $x \neq 0$ and thus $x^* = 0$ is the unique optimal solution for $x$. By the symmetry between $x$ and $y$ (because $G = -G^T$), we can also prove that the unique optimal solution for $y$ is $y^* = 0$.

Claim 2. Suppose that $x_0 = y_0 = \hat{x}_0 = \hat{y}_0 = \left(\frac{1}{2}, \frac{1}{4}\right)$. Then, at any step $t \in [T]$, we have $x_t = y_t$ and $\hat{x}_t \equiv \hat{y}_t$ and all $x_t, y_t, \hat{x}_t, \hat{y}_t$ belong to $\{ u \in \mathbb{R}^2 : u(2) = u(1)^2 \}$.

We prove this by induction. The base case trivially holds. Suppose that for step $t$, we have $x_t = y_t$ and $\hat{x}_t = \hat{y}_t$ and $x_t, y_t, \hat{x}_t, \hat{y}_t \in \{ u \in \mathbb{R}^2 : u(2) = u(1)^2 \}$. Then consider step $t + 1$. According to the dynamic of OGDA, we have

$$\hat{x}_{t+1} = \Pi_{\mathcal{X}} \left\{ \hat{x}_t - \eta \begin{bmatrix} -y_t(2) \\ y_t(1) \end{bmatrix} \right\} = \Pi_{\mathcal{X}} \left\{ \begin{bmatrix} \hat{x}_t(1) + \eta y_t(2) \\ \hat{x}_t(2) - \eta y_t(1) \end{bmatrix} \right\},$$

$$x_{t+1} = \Pi_{\mathcal{X}} \left\{ \hat{x}_{t+1} - \eta \begin{bmatrix} -y_t(2) \\ y_t(1) \end{bmatrix} \right\} = \Pi_{\mathcal{X}} \left\{ \begin{bmatrix} \hat{x}_{t+1}(1) + \eta y_t(2) \\ \hat{x}_{t+1}(2) - \eta y_t(1) \end{bmatrix} \right\},$$

$$\hat{y}_{t+1} = \Pi_{\mathcal{Y}} \left\{ \hat{y}_t + \eta \begin{bmatrix} -x_t(2) \\ x_t(1) \end{bmatrix} \right\} = \Pi_{\mathcal{Y}} \left\{ \begin{bmatrix} \hat{y}_t(1) + \eta x_t(2) \\ \hat{y}_t(2) - \eta x_t(1) \end{bmatrix} \right\},$$

$$y_{t+1} = \Pi_{\mathcal{Y}} \left\{ \hat{y}_{t+1} + \eta \begin{bmatrix} -x_t(2) \\ x_t(1) \end{bmatrix} \right\} = \Pi_{\mathcal{Y}} \left\{ \begin{bmatrix} \hat{y}_{t+1}(1) + \eta x_t(2) \\ \hat{y}_{t+1}(2) - \eta x_t(1) \end{bmatrix} \right\}.$$ 

According to induction hypothesis, we have $\hat{x}_{t+1} = \hat{y}_{t+1}$, which further leads to $x_{t+1} = y_{t+1}$.

Now we prove that for any $\begin{bmatrix} x(1) \\ x(2) \end{bmatrix}$ such that $x(1) \geq 0$, $x(2) \leq \frac{1}{4}$ and $x(2) < x(1)^2$, we have $\Pi_{\mathcal{X}} \left\{ \begin{bmatrix} x(1) \\ x(2) \end{bmatrix} \right\}$ satisfies that $\pi(1)^2 = \pi(2)$. Otherwise, suppose that $\pi(1)^2 < \pi(2)$. Then according to the intermediate value theorem, there exists $\begin{bmatrix} \pi(1) \\ \pi(2) \end{bmatrix}$ that lies in the line segment of $\begin{bmatrix} x(1) \\ x(2) \end{bmatrix}$ and $\begin{bmatrix} \pi(1) \\ \pi(2) \end{bmatrix}$.
\[
\begin{bmatrix}
\pi(1) \\
\pi(2)
\end{bmatrix}
\] such that \( \bar{x}(1)^2 = \bar{x}(2) \). Moreover, as \( x(1) \geq 0, \bar{x}(1) \geq 0, x(2) \leq \frac{1}{4}, \bar{x}(2) \leq \frac{1}{4} \), we know that \( \begin{bmatrix}
\bar{x}(1) \\
\bar{x}(2)
\end{bmatrix} \in \mathcal{X} \). Therefore, we have \( \| \bar{x} - x \| < \| x - x \| \), which leads to contradiction.

Now consider \( \hat{x}_{t+1} \). According to induction hypothesis, we have \( (\hat{x}_t(1) + \eta y_t(2))^2 \geq \hat{x}_t(1)^2 = \hat{x}_t(2) \geq \hat{x}_t(2) - \eta y_t(1) \). If equalities hold, trivially we have \( \hat{x}_{t+1}(1)^2 = \hat{x}_t(1)^2 = \hat{x}_t(2) = \hat{x}_{t+1}(2) \) according to Eq. (23). Otherwise, as \( \hat{x}_t(1) + \eta y_t(2) \geq 0, \hat{x}_t(2) - \eta y_t(1) \leq \frac{1}{4} \), according to above analysis, we also have \( \hat{x}_{t+1}(1)^2 = \hat{x}_{t+1}(2) \). Applying similar analysis to \( \hat{y}_{t+1}, x_{t+1} \) and \( y_{t+1} \) finishes the induction proof.

**Claim 3.** With \( \eta \leq \frac{1}{2} \), the following holds for all \( t \geq 1 \),
\[
x_t(1) \in \left[ \frac{1}{2} \hat{x}_t(1), 2\hat{x}_t(1) \right],
\]
\[
\hat{x}_t(1) \in \left[ \hat{x}_{t-1}(1) - 4\eta \hat{x}_{t-1}(1)^2, \hat{x}_{t-1}(1) + 4\eta \hat{x}_{t-1}(1)^2 \right].
\]

We prove the claim by induction on \( t \). The case \( t = 1 \) trivially holds. Suppose that Eq. (27) and Eq. (28) hold at step \( t \). Now consider step \( t + 1 \).

**Induction to get Eq. (28).** According to Claim 2, we have
\[
\hat{x}_{t+1} = \Pi_X \left\{ \hat{x}_t - \eta \begin{bmatrix}
y_t(2) \\
y_t(1)
\end{bmatrix} \right\} = \Pi_X \left\{ \begin{bmatrix}
\hat{x}_t(1) + \eta x_t(1)^2 \\
\hat{x}_t(1)^2 - \eta x_t(1)
\end{bmatrix} \right\},
\]
and \( \hat{x}_{t+1} = (u, u^2) \) for some \( u \in [0, 1/2] \). Using the definition of the projection function, we have
\[
\hat{x}_{t+1}(1) = \underset{u \in [0, \frac{1}{2}]}{\arg \min} \left\{ (\hat{x}_t(1) + \eta x_t(1)^2 - u)^2 + (\hat{x}_t(1)^2 - \eta x_t(1) - u^2)^2 \right\} \triangleq \arg \min \ g(u).
\]

Now we show that \( \arg \min_{u \in [0, \frac{1}{2}]} g(u) = \arg \min_{u \in \mathbb{R}} g(u) \). Note that
\[
\nabla g(u) = 2(u - \hat{x}_t(1) - \eta x_t(1)^2) + 4u (u^2 + \eta x_t(1) - \hat{x}_t(1)^2),
\]
which means \( g(u) > g(\frac{1}{2}) \). On the other hand, when \( u < 0 \), using \( \hat{x}_t(1) \leq \frac{1}{2} \), we have
\[
\nabla g(u) < 2u - 4u \hat{x}_t(1)^2 \leq u < 0,
\]
which means \( g(u) > g(0) \). Combining Eq. (31) and Eq. (32), we know that \( \arg \min_{u \in [0, \frac{1}{2}]} g(u) = \arg \min_{u \in \mathbb{R}} g(u) \). Therefore, \( \hat{x}_{t+1}(1) \) is the unconstrained minimizer of convex function \( g(u) \), which means \( \nabla g(\hat{x}_{t+1}(1)) = 0 \). Below we use contradiction to prove that \( \hat{x}_{t+1}(1) < \hat{x}_t(1) - 4n\hat{x}_t(1)^2 \). If \( \hat{x}_{t+1}(1) < \hat{x}_t(1) - 4n\hat{x}_t(1)^2 \), we use Eq. (30) and get
\[
\nabla g(\hat{x}_{t+1}(1)) = 2(\hat{x}_{t+1}(1) - \hat{x}_t(1) - \eta x_t(1)^2) + 4\hat{x}_{t+1}(1) (\hat{x}_{t+1}(1)^2 + \eta x_t(1) - \hat{x}_t(1)^2)
\]
\[
< 2(-4n\hat{x}_t(1)^2 - \eta x_t(1)^2) + 4\hat{x}_{t+1}(1) (\eta x_t(1) - 8n\hat{x}_t(1)^3 + 16n^2 \hat{x}_t(1)^4)
\]
\[
\leq -\frac{17}{2} \eta \hat{x}_t(1)^2 + 4\hat{x}_{t+1}(1) (2\eta \hat{x}_t(1) - 8n\hat{x}_t(1)^3 + 16n^2 \hat{x}_t(1)^4) \quad (\text{Eq. (27)})
\]
\[
\leq -\frac{17}{2} \eta \hat{x}_t(1)^2 + 4\hat{x}_{t+1}(1) (2\eta \hat{x}_t(1) + 16n^2 \hat{x}_t(1)^4)
\]
\[
\leq -\frac{17}{2} \eta \hat{x}_t(1)^2 + 4(\hat{x}_t(1) - 4n\hat{x}_t(1)^2) (2\eta \hat{x}_t(1) + 16n^2 \hat{x}_t(1)^4)
\]
\[
(\hat{x}_{t+1}(1) < \hat{x}_t(1) - 4n\hat{x}_t(1)^2)
\]
\[
= -\frac{1}{2} \eta \hat{x}_t(1)^2 + 64n^2 \hat{x}_t(1)^5 - 32n^2 \hat{x}_t(1)^3 - 256n^3 \hat{x}_t(1)^6
\]
\[
\leq -\frac{1}{2} \eta \hat{x}_t(1)^2 - 16n^2 \hat{x}_t(1)^3 - 256n^3 \hat{x}_t(1)^6
\]
\[
(\hat{x}_t(1) \leq \frac{1}{2})
\]
\[
\leq 0,
\]
which leads to contradiction. Similarly, if \( \hat{x}_{t+1}(1) > \hat{x}_t(1) + 4\eta \hat{x}_t(1)^2 \), we have

\[
\nabla g(\hat{x}_{t+1}(1)) = 2(\hat{x}_{t+1}(1) - \hat{x}_t(1) - \eta x_t(1)^2) + 4\hat{x}_{t+1}(1) (\hat{x}_{t+1}(1)^2 + \eta x_t(1) - \hat{x}_t(1)^2) \\
> 2(4\eta \hat{x}_t(1)^2 - \eta x_t(1)^2) + 4\hat{x}_{t+1}(1) (\eta x_t(1) + 8\eta \hat{x}_t(1)^3 + 16\eta^2 \hat{x}_t(1)^4) \\
\geq 0. \quad \text{(Eq. (27))}
\]

The above calculations conclude that

\[
\hat{x}_{t+1}(1) \in [\hat{x}_t(1) - 4\eta \hat{x}_t(1)^2, \hat{x}_t(1) + 4\eta \hat{x}_t(1)^2] .
\] (33)

**Induction to get Eq. (27).** Similarly, we have

\[
x_{t+1}(1) = \arg \min_{u \in [0, \frac{1}{4}]} \left\{ (\hat{x}_{t+1}(1) + \eta x_t(1)^2 - u)^2 + (\hat{x}_{t+1}(1)^2 - \eta x_t(1) - u^2)^2 \right\} \triangleq \arg \min h(u),
\]

and

\[
\nabla h(u) = 2(u - \hat{x}_{t+1}(1) - \eta x_t(1)^2) + 4u(u^2 + \eta x_t(1) - \hat{x}_{t+1}(1)^2),
\]

and \( \nabla h(x_{t+1}(1)) = 0 \). If \( x_{t+1}(1) < \frac{1}{2} \hat{x}_{t+1}(1) \), we have

\[
\nabla h(x_{t+1}(1)) = 2(x_{t+1}(1) - \hat{x}_{t+1}(1) - \eta x_t(1)^2) + 4x_{t+1}(1) (x_{t+1}(1)^2 + \eta x_t(1) - \hat{x}_{t+1}(1)^2) \\
< -\hat{x}_{t+1}(1) - 2\eta x_t(1)^2 - 3x_{t+1}(1)\hat{x}_{t+1}(1)^2 + 2\eta \hat{x}_{t+1}(1)x_t(1) \quad (x_{t+1}(1) < \frac{1}{2} \hat{x}_{t+1}(1)) \\
\leq 0. \quad (\eta \leq \frac{1}{64t} , x_t(1) \leq \frac{1}{2})
\]

If \( x_{t+1}(1) > 2\hat{x}_{t+1}(1) \), we also have

\[
\nabla h(x_{t+1}(1)) = 2(x_{t+1}(1) - \hat{x}_{t+1}(1) - \eta x_t(1)^2) + 4x_{t+1}(1) (x_{t+1}(1)^2 + \eta x_t(1) - \hat{x}_{t+1}(1)^2) \\
> 2\hat{x}_{t+1}(1) - 2\eta x_t(1)^2 + 24\hat{x}_{t+1}(1)^3 + 8\eta \hat{x}_{t+1}(1)x_t(1) \quad (x_{t+1}(1) > \frac{1}{2} \hat{x}_{t+1}(1)) \\
\geq 2\hat{x}_{t+1}(1) - 2\eta x_t(1)^2 + 24\hat{x}_{t+1}(1)^3 + 8\eta (\frac{1}{2} x_t(1) - 4\eta \hat{x}_t(1)^2)x_t(1) \quad \text{(Eq. (33))} \\
\geq 2\hat{x}_{t+1}(1) + 2\eta x_t(1)^2 + 24\hat{x}_{t+1}(1)^3 + 8\eta (\frac{1}{2} x_t(1) - 4\eta \hat{x}_t(1)^2)x_t(1) \quad \text{(Eq. (27))} \\
= 2\hat{x}_{t+1}(1) + 2\eta x_t(1)^2 + 24\hat{x}_{t+1}(1)^3 + 32\eta^2 \hat{x}_t(1)^3x_t(1) \\
\geq 0. \quad (\eta \leq \frac{1}{64t} , x_t(1) \leq \frac{1}{2})
\]

Both lead to contradiction. Therefore, we conclude that \( x_{t+1}(1) \in [\frac{1}{2} \hat{x}_{t+1}(1), 2\hat{x}_{t+1}(1)] \), which finishes the induction proof.

**Claim 4.** If \( \eta \leq \frac{1}{64t} \), we have \( \| \tilde{z}_t - z^* \| \geq \Omega(1/t) \).

Now we are ready to prove \( \| \tilde{z}_t - z^* \| \geq \Omega(1/t) \). First we show \( \tilde{z}_t(1) \geq \frac{1}{2t} \) for all \( t \geq 1 \) by induction. The case \( t = 1 \) trivially holds. Suppose that it holds at step \( t \). Considering step \( t + 1 \), we have

\[
\tilde{z}_{t+1}(1) \geq \tilde{x}_t(1) - 4\eta \tilde{x}_t(1)^2 \quad \text{(Claim 3)} \\
\geq \tilde{x}_t(1) - \frac{1}{16} \tilde{x}_t(1)^2 \quad (\eta \leq \frac{1}{64t}) \\
\geq \frac{1}{2t} - \frac{1}{64t^2} \quad (\frac{1}{2t} \leq \tilde{x}_t(1) \leq \frac{1}{2} \text{ and function } x - \frac{1}{16} x^2 \text{ is increasing when } x \leq 8) \\
\geq \frac{1}{2(t+1)}. \quad (t \geq 1)
\]

Note that according to Claim 1, \( z^* = 0 \). Therefore, we have \( \| \tilde{z}_t - z^* \| \geq \tilde{x}_{t+1}(1) \geq \frac{1}{2t} \), which finishes the proof. \( \square \)
H Auxiliary Lemmas for Stochastic Games

Definition 2. Define $\mu_{s, y}^s$ as the probability of visiting state $s$ given that the players use policies $(x, y)$. By the game structure, it satisfies the following:

$$
\mu_{s, y}^x = \begin{cases} 
    p_0(s') & \text{if } s' \in S_1, \\
    \sum_{a, b \in \Delta} \sigma^x(a) \gamma^y(b) p(s' | s, a, b) \mu_{s, y}^s & \text{if } s' \notin S_1.
\end{cases}
$$

Also, $\sum_{s \in S_h} \mu_{s, y}^x = 1$ for any $h, x, y$.

Lemma 16 (Value difference lemma). For any policies $x, y, \tilde{x}, \tilde{y}$, we have

$$
\rho_{x, y} - \rho_{\tilde{x}, \tilde{y}} = \sum_{s \in S} \sum_{a, b} \mu_{x, y}^s \left( x^s(a) y^s(b) - \tilde{x}^s(a) \tilde{y}^s(b) \right) Q_{x, y}^s(a, b).
$$

Proof. By definitions, we have the following equalities:

$$
\rho_{x, y} = \sum_{s \in S} \sum_{a, b} \mu_{x, y}^s \tilde{x}^s(a) \tilde{y}^s(b) \ell(s, a, b)
$$

$$
= \sum_{s \in S} \sum_{a, b} \mu_{x, y}^s \tilde{x}^s(a) \tilde{y}^s(b) \left( Q_{x, y}^s(a, b) - \sum_{s' \in S} p(s' | s, a, b) V_{x, y}^{s'} \right) 
$$

(by the definition of $Q_{x, y}^s$)

$$
= \sum_{s \in S} \sum_{a, b} \mu_{x, y}^s \tilde{x}^s(a) \tilde{y}^s(b) Q_{x, y}^s(a, b) - \sum_{s' \in S \setminus S_1} \mu_{x, y}^{s'} V_{x, y}^{s'} 
$$

(by Definition 2)

$$
= \sum_{s \in S} \sum_{a, b} \mu_{x, y}^s \tilde{x}^s(a) \tilde{y}^s(b) Q_{x, y}^s(a, b) - \sum_{s \in S} \sum_{s' \in S_1} \mu_{x, y}^s V_{x, y}^{s'} + \sum_{s \in S_1} \mu_{x, y}^s V_{x, y}^s 
$$

(by the definition of $V_{x, y}^s$)

$$
= \sum_{s \in S} \sum_{a, b} \mu_{x, y}^s \left( \tilde{x}^s(a) \tilde{y}^s(b) - x^s(a) y^s(b) \right) Q_{x, y}^s(a, b) + \sum_{s \in S} p(s) V_{x, y}^s 
$$

Rearranging the above finishes the proof. $\square$

Lemma 17. Let $s \in S_h$ for some $h$. Then for any actions $a, b$ and any policies $x, y, \tilde{x}, \tilde{y}$, we have

$$
|Q_{x, y}^s(a, b) - Q_{x, y}^s(a, b)| \leq H \left( \sum_{i=h+1}^H \max_{s' \in S_i} \|x^{s'} - x^s\|_1 + \|\tilde{y}^{s'} - y^s\|_1 \right).
$$

Proof. Notice that for $s \in S_1$, by definition we have

$$
|Q_{x, y}^s(a, b) - Q_{x, y}^s(a, b)| \leq \sum_{s'} p(s' | s, a, b) \left| V_{x, y}^{s'} - V_{x, y}^s \right|
$$

\[
\leq \max_{s' \in S_{h+1}} \left| V_{x, y}^{s'} - V_{x, y}^s \right|
\]
where in the second to last step we use the fact that \( Q \) is at most \( H \) clearly. Using the above two relations repeatedly for \( i = h, h + 1, \ldots, H - 1 \) proves the result.

\[ \]
Proof. By Definition 2, we proceed with
\[
\sum_{s' \in S_h} \left| \mu^s_{x, y} - \mu^{s'}_{x, y} \right|
\]
\[
= \sum_{s' \in S_h} \left| \sum_{s \in S_{h-1}} \sum_{a, b} (\tilde{x}^s(a)\tilde{y}^s(b)p(s' \mid s, a, b)\mu^s_{x, y} - x^s(a)y^s(b)p(s' \mid s, a, b)\mu^{s'}_{x, y}) \right|
\]
\[
\leq \sum_{s \in S_{h-1}} \sum_{a, b} \left( \sum_{s' \in S_h} p(s') \left| (\tilde{x}^s(a)\tilde{y}^s(b)\mu^s_{x, y} - x^s(a)y^s(b)\mu^{s'}_{x, y}) \right| \right)
\]
\[
= \sum_{s \in S_{h-1}} \sum_{a, b} \left| \tilde{x}^s(a)\tilde{y}^s(b)\mu^s_{x, y} - x^s(a)y^s(b)\mu^{s'}_{x, y} \right|
\]
\[
\leq \sum_{s \in S_{h-1}} \sum_{a, b} \left| \tilde{x}^s(a) - x^s(a) \right| \left| \tilde{y}^s(b) - y^s(b) \right| \left| \mu^s_{x, y} \right| + \sum_{s \in S_{h-1}} \sum_{a, b} x^s(a) \left| \tilde{y}^s(b) - y^s(b) \right| \left| \mu^s_{x, y} \right|
\]
\[
\leq \max_{s \in S_{h-1}} \| \tilde{x}^s - x^s \|_1 + \max_{s \in S_{h-1}} \| \tilde{y}^s - y^s \|_1 + \sum_{s \in S_{h-1}} \left| \mu^s_{x, y} - \mu^{s'}_{x, y} \right|.
\]
Applying the above inequality repeatedly finishes the proof. \qed

I Proofs of Theorem 10

We follow our convention that $z^s \triangleq (x^s, y^s)$. Also, we define the following notation:

Definition 3. Define $\Phi_{t, h} = \max_{s \geq h} x^s \| z^s - z^s \|^2$.

Lemma 20. For any $s \in S_h$ and any $\tilde{z}^s = (\tilde{x}^s, \tilde{y}^s)$ such that $\left( \tilde{x}^s_{t+1}Q^s_{t+1}\tilde{y}^s - \tilde{x}^s_{t+1} \right) > 0$, we have
\[
\| z^s_t - \tilde{z}^s_t \|^2 + \| \tilde{z}^s_{t+1} - z^s_t \|^2 > \frac{64}{1875} \eta^2 \left( \tilde{x}^s_{t+1}Q^s_{t+1}\tilde{y}^s - \tilde{x}^s_{t+1} \right)^2 - H^2 \Phi_{t, h+1}.
\]

Proof. The proof follows similar arguments of the proof of Lemma 3. By the optimality of $\tilde{x}^s_{t+1}$ we have for any $\tilde{x}^s$,
\[
(\tilde{x}^s_{t+1} - \tilde{x}^s + \eta Q^s_{t, t}y^s_t)^\top (\tilde{x}^s - \tilde{x}^s_{t+1}) \geq 0.
\]
Rearranging, we get
\[
(\tilde{x}^s_{t+1} - \tilde{x}^s)^\top (\tilde{x}^s - \tilde{x}^s_{t+1}) \geq \eta (\tilde{x}^s_{t+1} - \tilde{x}^s)^\top Q^s_{t, t}y^s_t. \tag{34}
\]
Because $\| \tilde{x}^s - \tilde{x}^s_{t+1} \| \leq 2$, for the left-hand side we have $(\tilde{x}^s_{t+1} - \tilde{x}^s)^\top (\tilde{x}^s - \tilde{x}^s_{t+1}) \leq 2 \| \tilde{x}^s_{t+1} - \tilde{x}^s \|$. For the right-hand side, we have
\[
\eta (\tilde{x}^s_{t+1} - \tilde{x}^s)^\top Q^s_{t, t}y^s_t
\]
\[
= \eta (\tilde{x}^s_{t+1} - \tilde{x}^s)^\top Q^s_{t, t}\tilde{y}^s + \eta (\tilde{x}^s_{t+1} - \tilde{x}^s)^\top (Q^s_{t, t} - Q^s_{t, t})\tilde{y}^s + \eta (\tilde{x}^s_{t+1} - \tilde{x}^s)^\top Q^s_{t, t}(y^s_t - \tilde{y}^s_t)
\]
\[
\geq \eta (\tilde{x}^s_{t+1} - \tilde{x}^s)^\top Q^s_{t, t}\tilde{y}^s + \eta (\tilde{x}^s_{t+1} - \tilde{x}^s)^\top (y^s_t - \tilde{y}^s_t) \| Q^s_{t, t} \|_1 \| y^s_t - \tilde{y}^s_t \| \| Q^s_{t, t} \|_1 \| y^s_t - \tilde{y}^s_t \| \| Q^s_{t, t} \|_1 \| y^s_t - \tilde{y}^s_t \|
\]
\[
\geq \eta (\tilde{x}^s_{t+1} - \tilde{x}^s)^\top Q^s_{t, t}\tilde{y}^s + 2\eta H^2 \sqrt{A + B \sqrt{\Phi_{t, h+1}}} - 2\eta H \| y^s_t - \tilde{y}^s_t \| \| y^s_t - \tilde{y}^s_t \|
\]
\[
\geq \eta (\tilde{x}^s_{t+1} - \tilde{x}^s)^\top Q^s_{t, t}\tilde{y}^s + 2\eta H^2 \sqrt{A + B \sqrt{\Phi_{t, h+1}}} - 2\eta H \sqrt{B \| y^s_t - \tilde{y}^s_t \|}.
\]
Therefore, with Eq. (34), we have
\[
2 \| \tilde{x}^s_{t+1} - \tilde{x}^s \| + 2\eta H^2 \sqrt{A + B \sqrt{\Phi_{t, h+1}}} + 2\eta H \sqrt{B \| y^s_t - \tilde{y}^s_t \|}
\]
\[
\geq \eta (\tilde{x}^s_{t+1} - \tilde{x}^s)^\top Q^s_{t, t}\tilde{y}^s_t.
\]
Similarly for η we can obtain
\[
2\|\hat{y}_{t+1} - \tilde{y}_t\|^2 + 2\eta H^2\sqrt{A + B} \sqrt{\Phi_{t,h+1}} + 2\eta H \sqrt{A}\|x_t^* - \hat{x}_{t+1}\|
\geq \eta \tilde{x}_{t+1}^T Q_s^* (\tilde{y}^* - \tilde{y}_{t+1}).
\]

Summing up these two inequalities we get
\[
\eta \left( \tilde{x}_{t+1}^T Q_s^* \tilde{y}^* - \tilde{x}^T \tilde{Q}_s^* \tilde{y}_{t+1} \right)
\leq 2\sqrt{2}\|\hat{z}_t^* - \hat{z}_{t+1}^*\| + 2\eta H \sqrt{2AB}\|z_t^* - \hat{z}_{t+1}^*\| + 4\eta H^2 \sqrt{A + B} \sqrt{\Phi_{t,h+1}}
\leq 3\|\hat{z}_t^* - \hat{z}_{t+1}^*\| + 2\eta H \sqrt{2AB}\|z_t^* - \hat{z}_{t+1}^*\| + 4\eta H^2 \sqrt{2AB} \sqrt{\Phi_{t,h+1}}
\leq 3\|z_t^* - \hat{z}_{t+1}^*\| + \frac{1}{8}\|z_t^* - \hat{z}_{t+1}^*\| + \frac{H}{4} \sqrt{\Phi_{t,h+1}}
\leq 3\|z_t^* - \hat{z}_{t+1}^*\| + \frac{25}{8}\|\hat{z}_t^* - \hat{z}_{t+1}^*\| + \frac{H}{4} \sqrt{\Phi_{t,h+1}}.
\]

Squaring both sides and using \((a + b + c)^2 \leq 3(a^2 + b^2 + c^2)\), we get
\[
\eta^2 \left( \tilde{x}_{t+1}^T Q_s^* \tilde{y}^* - \tilde{x}^T \tilde{Q}_s^* \tilde{y}_{t+1} \right)^2
\leq \frac{1875}{64} \left( \|z_t^* - \hat{z}_{t+1}^*\|^2 + \|z_t^* - \hat{z}_{t+1}^*\|^2 \right) + \frac{3H^2}{16} \Phi_{t,h+1}.
\]

Rearranging finishes the proof. □

**Proof of Theorem 10.** Consider a state \(s \in S_t\). By the same analysis as in Lemma 2 (specifically Eq. (16) and Lemma 13), we have
\[
2\eta(x_t^* - \hat{x}_t^*)^T Q_{t, s} y_t^*
\leq \|x_t^* - \hat{x}_t^*\|^2 - \|x_t^* - \hat{x}_{t+1}^*\|^2 + 2\eta^2\|Q_{t, s} y_t^* - Q_{t, t-1} y_{t-1}^*\|^2 - \|\hat{x}_{t+1}^* - x_t^*\|^2 - \|x_t^* - \hat{x}_t^*\|^2.
\]

We proceed with
\[
2\eta^2\|Q_{t, s} y_t^* - Q_{t, t-1} y_{t-1}^*\|^2
\leq 4\eta^2\|Q_{t, s} (y_t^* - y_{t-1}^*)\|^2 + 4\eta^2\| (Q_{t, s} - Q_{t, t-1}) y_{t-1}^*\|^2
\leq 4\eta^2 H^2 A\|y_t^* - y_{t-1}^*\|^2 + 4\eta^2 H^4 B \max_{i > h, s' \in S_i} \|y_{t-1}^* - y_{t-1}^*\|^2
\leq 4\eta^2 H^2 A B\|y_t^* - y_{t-1}^*\|^2 + 8\eta^2 H^4 A B (\Phi_{t-1, h+1} + \Phi_{t, h+1})
\leq \frac{1}{128}\|y_t^* - y_{t-1}^*\|^2 + \frac{H^2}{64} (\Phi_{t-1, h+1} + \Phi_{t, h+1})
\leq \frac{1}{64}\|y_t^* - y_{t-1}^*\|^2 + \frac{1}{64}\|y_t^* - y_{t-1}^*\|^2 + \frac{H^2}{64} (\Phi_{t-1, h+1} + \Phi_{t, h+1}).
\]
We can get the $y$ counterpart of Eq. (35) and Eq. (36) similarly. Combining all these inequalities leads to
\[
2\eta(x_t^* - x_{t-1}^*)^\top Q_t^* y_t^* + 2\eta x_t^* x_t^\top Q_{c,t}^* (y_t^* - y_t^*) \\
\leq \|z_t^* - z_{t-1}^*\|^2 + \frac{1}{64} \|z_t^* - z_{t-1}^*\|^2 - \frac{63}{64} \|z_t^* - z_{t-1}^*\|^2 - \|z_t^* - z_{t-1}^*\|^2 \\
+ \frac{H^2}{32} (\Phi_{t-1,h+1} + 2\eta\|t^*\|Q_{c,t}^* y_t^* - x_t^* y_t^*) \\
\leq \|z_t^* - z_{t-1}^*\|^2 + \frac{1}{64} \|z_t^* - z_{t-1}^*\|^2 - \frac{31}{64} \|z_{t+1}^* - z_t^*\|^2 \\
- \frac{1}{2} \cdot \frac{64\eta^2}{1875} \max_{z,y} \left(\hat{x}_t^\top Q_{c,t}^* y - \hat{x}_t^\top Q_{c,t}^* y_{t+1}^*\right)^2 + \frac{1}{2} \frac{H^2}{16} \Phi_{t-1,h+1} + \frac{H^2}{16} (\Phi_{t-1,h+1} + \Phi_{t,h+1}) \\
\leq \|z_t^* - z_{t-1}^*\|^2 + \frac{1}{64} \|z_t^* - z_{t-1}^*\|^2 - \frac{31}{64} \|z_{t+1}^* - z_t^*\|^2 \\
- \frac{32\eta^2 C^2}{1875} \|z_{t+1}^* - z_t^*\|^2 + H^2 (\Phi_{t-1,h+1} + \Phi_{t,h+1}). \\
\text{(by Theorem 5 for some constant $C$)} \\
(37)
\]

Note that the left-hand side of Eq. (37) can be lower bounded as below:
\[
2\eta(x_t^* - x_{t-1}^*)^\top Q_t^* y_t^* + 2\eta x_t^* x_t^\top Q_{c,t}^* (y_t^* - y_t^*) \\
= 2\eta(x_t^* - x_{t-1}^*)^\top Q_t^* y_t^* + 2\eta x_t^* x_t^\top Q_{c,t}^* (y_t^* - y_t^*) \\
+ 2\eta(x_t^* - x_{t-1}^*)^\top (Q_t^* - Q_{c,t}^*) y_t^* + 2\eta x_t^* (Q_t^* - Q_{c,t}^*) (y_t^* - y_t^*) \\
\geq 2\eta(x_t^* - x_{t-1}^*)^\top (Q_t^* - Q_{c,t}^*) y_t^* + 2\eta x_t^* (Q_t^* - Q_{c,t}^*) (y_t^* - y_t^*) \\
\geq -2\eta\|x_t^* - x_{t-1}^*\|\|(Q_t^* - Q_{c,t}^*) y_t^*\| - 2\eta\|x_t^* (Q_t^* - Q_{c,t}^*)\|\|y_t^* - y_t^*\| \\
\text{(Cauchy-Schwarz inequality)} \\
\geq -\frac{\eta^2 C^2}{1000} \|x_t^* - x_{t-1}^*\|^2 - \frac{1000}{C^2} \|(Q_t^* - Q_{c,t}^*) y_t^*\|^2 - \frac{\eta^2 C^2}{1000} \|y_t^* - y_{t-1}^*\|^2 - \frac{1000}{C^2} \|(Q_t^* - Q_{c,t}^*) x_t^*\|^2 \\
\text{(AM-GM inequality)} \\
\geq -\frac{\eta^2 C^2}{1000} \|z_t^* - z_{t-1}^*\|^2 - \frac{1000}{C^2} \times H^4 (A + B)^2 \Phi_{t,h+1} \\
\geq -\frac{\eta^2 C^2}{500} \|z_t^* - z_{t-1}^*\|^2 - \frac{1000}{C^2} \times H^4 (A + B)^2 \Phi_{t,h+1}. \\
\text{Combining with Eq. (37), we get}
\]
\[
\|z_{t+1}^* - z_t^*\|^2 \leq \|z_{t+1}^* - z_t^*\|^2 + \frac{14}{32} \|z_{t+1}^* - z_t^*\|^2 + \frac{15}{32} \|z_{t+1}^* - z_t^*\|^2 - \frac{28\eta^2 C^2}{1875} \|z_{t+1}^* - z_t^*\|^2 \\
\leq \frac{1}{1 + \frac{28\eta^2 C^2}{1875}} \left(\|z_{t+1}^* - z_t^*\|^2 + \frac{1}{64} \|z_{t+1}^* - z_{t-1}^*\|^2\right) + \frac{1001 H^4 (A + B)^2}{C^2} (\Phi_{t-1,h+1} + \Phi_{t,h+1}). \\
\text{Without loss of generality, we assume $C \leq 1$ (otherwise we can choose $C = 1$). By rearranging terms, we have}
\]
\[
\|z_{t+1}^* - z_t^*\|^2 + \frac{14}{32} \|z_{t+1}^* - z_t^*\|^2 \leq \|z_{t+1}^* - z_t^*\|^2 + \frac{15}{32} \|z_{t+1}^* - z_{t-1}^*\|^2 - \frac{28\eta^2 C^2}{1875} \|z_{t+1}^* - z_t^*\|^2 \\
\leq \frac{1}{1 + \frac{28\eta^2 C^2}{1875}} \left(\|z_{t+1}^* - z_t^*\|^2 + \frac{1}{64} \|z_{t+1}^* - z_{t-1}^*\|^2\right) + \frac{1001 H^4 (A + B)^2}{C^2} (\Phi_{t-1,h+1} + \Phi_{t,h+1}). \\
\text{Note that the above holds for any $z_t^* \in Z_t^*$. We choose $z_t^* = \Pi z_t^* (z_t^*)$. In the following, we use the notation $\Omega_t^* = \|z_t^* - \Pi z_t^* (z_t^*)\|^2$ and $\theta_t^* = \|z_t^* - z_{t-1}^*\|^2$. Using the fact $\|z_{t+1}^* - \Pi z_t^* (z_{t+1}^*)\| \leq$
\]

30
We use induction to prove that
\[ \|z_{t+1} - \Pi_{z_t}(z_t')\|, \] the above implies
\[ \Omega_{t+1} + \frac{7}{16} \theta_{t+1} \leq \frac{1}{1 + \frac{28}{1875} \eta^2 C^2} \left( \Omega_{t} + \frac{7}{16} \theta_{t} \right) + \frac{1001 \times 32 H^4 (A + B)^2}{7 C^2} \left[ \max_{i > h, s' \in S_t} (\Omega_{s}' + \frac{7}{16} \theta_{s}' + 1) + \max_{i > h, s' \in S_t} (\Omega_{s}' + \frac{7}{16} \theta_{s}' + 1) \right] \]
because \( \Phi_{t,h} = \max_{i > h, s' \in S_t} \|z_{t}' - z_{s}'\|^2 \leq \max_{i \geq h, s' \in S_t} \left( 2 \|z_{t}' - z_{t+1}'\|^2 + 2 \|z_{t+1}' - z_{s}'\|^2 \right) \)
\[ = 2 \max_{i \geq h, s' \in S_t} \left( \Omega_{t+1} + \theta_{t+1}' \right) \leq \frac{32}{\eta^2} \max_{i \geq h, s' \in S_t} \left( \Omega_{t+1} + \frac{7}{16} \theta_{t+1}' \right), \] where we choose \( z_{t}' = \Pi_{z_t}(z_t') \).

Below we further define \( \zeta_{t,h} \triangleq \max_{s \in S_{h}} (\Omega_{s} + \frac{7}{16} \theta_{s}'). \) Then we can further write
\[ \zeta_{t+1,h} \leq \frac{\zeta_{t,h}}{1 + \frac{28}{1875} \eta^2 C^2} + \frac{5000 H^4 (A + B)^2}{C^2} \left( \max_{i > h} \zeta_{t+1,i} + \max_{i > h} \zeta_{t,i} \right). \] (38)

We use induction to prove that
\[ \zeta_{t,h} \leq 2 \left( \frac{1}{1 + \frac{1}{1000} \eta^2 C^2} \right)^{t-1} \left( \frac{2 \times 10^6 H^4 (A + B)^2}{\eta^2 C^4} \right)^{H-h}. \] (39)

It is clear that \( t = 1 \) holds for every layer \( h \).

The last layer. If \( h = H \), then the recursion Eq. (38) implies, for all \( t \in [T] \),
\[ \zeta_{t,h} \leq 2 \left( \frac{1}{1 + \frac{28}{1875} \eta^2 C^2} \right)^{t-1} \leq 2 \left( \frac{1}{1 + \frac{1}{1000} \eta^2 C^2} \right)^{t-1}. \]

Previous layers. Suppose that the Eq. (39) holds for layers \( h+1, \ldots, H \) and for all \( t \), and suppose that it holds for time \( 1, \ldots, t \) for layer \( h \). Then by Eq. (38),
\[ \zeta_{t+1,h} \leq 2 \left( \frac{1}{1 + \frac{28}{1875} \eta^2 C^2} \right)^{t-1} \left( \frac{2 \times 10^6 H^4 (A + B)^2}{\eta^2 C^4} \right)^{H-h} \times \left( \frac{1}{1 + \frac{1}{1000} \eta^2 C^2} \right)^{t-1} \left( \frac{2 \times 10^6 H^4 (A + B)^2}{\eta^2 C^4} \right)^{H-h} \times \left( \frac{1}{1 + \frac{1}{1000} \eta^2 C^2} \right)^{t-1} \left( \frac{2 \times 10^6 H^4 (A + B)^2}{\eta^2 C^4} \right)^{H-h} \leq 2 \left( \frac{1}{1 + \frac{28}{1875} \eta^2 C^2} \right)^{t-1} \left( \frac{2 \times 10^6 H^4 (A + B)^2}{\eta^2 C^4} \right)^{H-h} , \] (\( C < 1 \))
which finishes the induction.

Note that Eq. (39) implies that
\[ \Omega_{t}' \leq \zeta_{t,h} \leq 2 \left( \frac{1}{1 + \frac{1}{1000} \eta^2 C^2} \right)^{t-1} \left( \log \left( \frac{2 \times 10^6 H^4 (A + B)^2}{\eta^2 C^4} \right) \right) \left( \log \left( 1 + \frac{1}{1000} \eta^2 C^2 \right) \right) \leq \left( \frac{1}{1 + \frac{1}{1000} \eta^2 C^2} \right)^{t-O(H-h) \log \frac{H+B}{\eta^2 C^4} } , \]
which leads to the desired inequality. \( \square \)
J  Proofs of Theorem 11

Proof of Theorem 11. By Lemma 16, for any $x$, 
\[ \rho_{x_t,y_t} - \rho_{x,y} = \sum_{s \in \mathcal{S}} \mu_{x_t,y_t}^s (x^s_t - x^s) \mathcal{T} Q^s_{t,i} y^i_t. \]  
(40)

Note that $(x^s_t - x^s) \mathcal{T} Q^s_{t,i} y^i_t$ is essentially the instantaneous regret of OGDA on the state $s$ at episode $t$. By Eq. (35), we have 
\[ (x^s_t - x^s) \mathcal{T} Q^s_{t,i} y^i_t \leq \frac{\| x^s_t - \tilde{x}^s_t \|^2 - \| x^s_t - \tilde{x}^s_{t+1} \|^2}{2\eta} + \eta \| Q^s_{t,i} y^i_t - Q^s_{t+1} y^i_{t+1} \|^2. \]  
(41)

By Lemma 17, for $t \geq 2$, 
\[ \| Q^s_{t,i} y^i_t - Q^s_{t-1} y^i_{t-1} \|^2 \leq 2\| Q^s_{t,i} (y^i_t - y^i_{t-1}) \|^2 + 2\| (Q^s_{t,i} - Q^s_{t-1}) y^i_{t-1} \|^2 \]
\[ \leq 2H^2 A\| y^i_t - y^i_{t-1} \|^2 + 2H^2 A \left( \sum_{i=h+1}^H \max_{s \in \mathcal{S}_i} \| y^i_t - y^i_{t-1} \| \right)^2 \]  
(By Lemma 17; $A \triangleq |\mathcal{A}|$)
\[ \leq 2H^2 A \Theta^2_{t,h} + 2H^3 A \sum_{i=h+1}^H \Theta^2_{i,i} \]  
where in the last step we use Cauchy-Schwarz inequality. Also, for $t = 1$, $\| Q^s_{t,i} y^i_t - Q^s_{t-1} y^i_{t-1} \|^2 = O(H^2 A)$. Plugging these bounds into Eq. (41), we have for any $h$, 
\[ \sum_{t=1}^T \sum_{s \in \mathcal{S}_h} \mu_{x_t,y_t}^s (x^s - x^s_t) \mathcal{T} Q^s_{t,i} y^i_t \]
\[ \leq \sum_{t=1}^T \sum_{s \in \mathcal{S}_h} \mu_{x_t,y_t}^s \| x^s_t - \tilde{x}^s_t \|^2 - \mu_{x^s_t,y^s_t} \| x^s_t - \tilde{x}^s_{t+1} \|^2 \]  
\[ \leq 2\| Q^s_{t,i} (y^i_t - y^i_{t-1}) \|^2 + 2\| (Q^s_{t,i} - Q^s_{t-1}) y^i_{t-1} \|^2 \]
\[ \leq 2H^2 A\| y^i_t - y^i_{t-1} \|^2 + 2H^2 A \left( \sum_{i=h+1}^H \max_{s \in \mathcal{S}_i} \| y^i_t - y^i_{t-1} \| \right)^2 \]  
(41)

\[ \leq 2H^2 A \Theta^2_{t,h} + 2H^3 A \sum_{i=h+1}^H \Theta^2_{i,i} \]  
(by Lemma 19)

Finally summing over $h$ and combining with Eq. (40) we have 
\[ \sum_{t=1}^T (\rho_{x_t,y_t} - \rho_{x,y}) \]
\[ \leq 2\sum_{t=2}^T \sum_{h=1}^{H-1} \beta_{t,h} + \sum_{t=2}^T \left( 2H^2 A \sum_{h=1}^H \Theta^2_{t,h} + 2H^3 A \sum_{h=1}^H \Theta^2_{t,h} \right) + O \left( \frac{H}{\eta} + \eta H^3 A \right) \]
\[ = O \left( \frac{H}{\eta} \sum_{t=2}^{H} \sum_{h=1}^{H} \Theta_{t,h} + \eta H^4 A \sum_{t=2}^{H} \sum_{h=1}^{H} \Theta^2_{t,h} \right) + O \left( \frac{H}{\eta} + \eta H^3 A \right), \]
completing the proof. \(\square\)