No-Regret Learning in Time-Varying Zero-Sum Games

Mengxiao Zhang [*1] **Peng Zhao** [*2] **Haipeng Luo** [1] **Zhi-Hua Zhou** [2]

Abstract

Learning from repeated play in a fixed two-player zero-sum game is a classic problem in game theory and online learning. We consider a variant of this problem where the game payoff matrix changes over time, possibly in an adversarial manner. We first present three performance measures to guide the algorithmic design for this problem: 1) the well-studied *individual regret*, 2) an extension of *duality gap*, and 3) a new measure called *dynamic Nash Equilibrium regret*, which quantifies the cumulative difference between the player's payoff and the minimax game value. Next, we develop a single parameter-free algorithm that *simultaneously* enjoys favorable guarantees under all these three performance measures. These guarantees are adaptive to different non-stationarity measures of the payoff matrices and, importantly, recover the best known results when the payoff matrix is fixed. Our algorithm is based on a two-layer structure with a meta-algorithm learning over a group of black-box base-learners satisfying a certain property, along with several novel ingredients specifically designed for the time-varying game setting. Empirical results further validate the effectiveness of our algorithm.

1. Introduction

Repeated play in a fixed two-player zero-sum game, a fundamental problem in the interaction between game theory and online learning, has been extensively studied in recent decades. In particular, many efforts have been devoted to designing online algorithms such that both players achieve small individual regret (that is, difference between one's cumulative payoff and that of the best fixed action) while at the same time the dynamics of the players' strategy leads to a Nash equilibrium, a pair of strategies that neither player has incentive to deviate from; see for example (Freund & Schapire, 1999; Rakhlin & Sridharan, 2013; Daskalakis et al., 2015; Syrgkanis et al., 2015; Chen & Peng, 2020; Wei et al., 2021; Hsieh et al., 2021; Daskalakis et al., 2021).

In contrast to this large body of studies for learning over a *fixed* zero-sum game, repeated play over a sequence of *time-varying* games, the focus of this paper and a ubiquitous scenario in practice, is much less explored. While minimizing individual regret still makes perfect sense in this case, it is not immediately clear what other desirable game-theoretic guarantees are that generalize the concept of approaching a Nash equilibrium when the game is fixed. As far as we know, Cardoso et al. (2019) are the first to explicitly consider this problem. They proposed the notion of Nash-Equilibrium regret (NE-regret) as the performance measure, which quantifies the difference between the learners' cumulative payoff and the minimax value of the cumulative payoff matrix. The authors proposed an algorithm with $\widetilde{\mathcal{O}}(\sqrt{T})$ NE-regret after T rounds of play and, importantly, proved that no algorithm can simultaneously achieve sublinear NE-regret and sublinear individual regret for both players.

Our work starts by questioning whether the NE-regret of Cardoso et al. (2019) is indeed a good performance measure for the problem of learning in time-varying games, especially given its incompatibility with the arguably most standard goal of having small individual regret. We then discover that measuring performance with NE-regret can in fact be highly unreasonable: we show an example (in Section 3) where even the two players perform perfectly (in the sense that they play the corresponding Nash equilibrium in every round), the resulting NE-regret is still *linear* in T!

Motivated by this observation, we revisit the basic problem of how to measure the algorithm's performance in such a time-varying game setting. Concretely, we consider three performance measures that we believe are appropriate and natural: 1) the standard individual regret; 2) the direct generalization of cumulative duality gap from a fixed game to a varying game; and 3) a new measure called *dynamic NE-regret*, which quantifies the difference between the learner's cumulative payoff and the cumulative minimax game value (instead of the minimax value of the cumulative payoff ma-

[*]Equal contribution, in alphabetical order. [1]University of Southern California [2]National Key Laboratory for Novel Software Technology, Nanjing University. Correspondence to: Mengxiao Zhang <mengxiao.zhang@usc.edu>, Peng Zhao <zhaop@lamda.nju.edu.cn>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

Table 1. Summary of our results. The first column indicates the three performance measures considered in this work. The second column presents our main results for a sequence of time-varying payoff matrices $\{A_t\}_{t=1}^T$, and notably all results are simultaneously achieved by one single parameter-free algorithm. These guarantees are expressed in terms of three different (unknown) non-stationarity measures of the payoff matrices: P_T, V_T, and W_T, all of which are $\Theta(T)$ in the worst-case and zero in the stationary case (when $A_t = A$ for all $t \in [T]$); see Section 3 for definitions. Additionally, for duality gap we use notation Q_T as a shorthand for $V_T + \min\{P_T, W_T\}$. Substituting all non-stationarity measures with zero leads to our corollaries for the stationary case shown in the third column, which match the state-of-the-art for all three performance measures (up to logarithmic factors) as shown in the last column.

Measure	**Time-Varying Game** ($\{A_t\}_{t=1}^T$, general)	**Stationary Game** ($A_t = A$, fixed)	
Individual Regret	$\widetilde{\mathcal{O}}\big(\sqrt{1 + V_T + \min\{P_T, W_T\}}\big)$ [Theorem 4]	$\widetilde{\mathcal{O}}(1)$ [Corollary 5]	$\mathcal{O}(1)$ (Hsieh et al., 2021)
Dynamic NE-Regret	$\widetilde{\mathcal{O}}\big(\min\{\sqrt{(1+V_T)(1+P_T)} + P_T, 1 + W_T\}\big)$ [Theorem 6]	$\widetilde{\mathcal{O}}(1)$ [Corollary 7]	$\mathcal{O}(1)$ (Hsieh et al., 2021)[1]
Duality Gap	$\widetilde{\mathcal{O}}\big(\min\{T^{\frac{3}{4}}(1+Q_T)^{\frac{1}{4}}, T^{\frac{1}{2}}(1+Q_T^{\frac{3}{2}} + P_T Q_T)^{\frac{1}{2}}\}\big)$ [Theorem 8]	$\widetilde{\mathcal{O}}(\sqrt{T})$ [Corollary 9]	$\mathcal{O}(\sqrt{T})$ (Wei et al., 2021)

trix, as in NE-regret). We argue that dynamic NE-regret is a better measure compared to NE-regret: first, in the earlier example where both players play perfectly in each round using the corresponding Nash equilibrium, their dynamic NE-regret is exactly zero (while their NE-regret can be linear in T); second, having small dynamic NE-regret does not prevent one from enjoying small individual regret or duality gap (as will become clear soon).

With these performance measures in mind, our main contribution is to develop *one single parameter-free algorithm* that simultaneously enjoys favorable guarantees under all measures. These guarantees are adaptive to some unknown non-stationarity measures of the payoff matrices — naturally, the bounds worsen as the non-stationarity becomes larger. More specifically, the individual regret is always at most $\widetilde{\mathcal{O}}(\sqrt{T})$, the well-known worst-case bound, but could be much smaller if the non-stationarity measures are sublinear; on the other hand, the duality gap and dynamic NE-regret are sublinear as long as the non-stationarity measures are sublinear. In the special case of a fixed payoff matrix, all non-stationarity measures become zero and our results immediately recover the state-of-the-art results (up to logarithmic factors); see Table 1 for details. Notice that the best known results for a fixed game are not necessarily achieved by the same algorithm, while again, our results are all achieved by one adaptive algorithm. We also conduct empirical studies to further support our theoretical findings.

Techniques. For a fixed game, Syrgkanis et al. (2015) proposed the "Regret bounded by Variation in Utilities" (RVU) property as the key condition for an algorithm to achieve good performance. On the other hand, one of the key tools for achieving our results is to ensure a small gap between each player's cumulative payoff and that of a sequence of changing comparators, known as *dynamic regret* in the literature (Zinkevich, 2003). Therefore, our first step is to generalize the RVU property to "Dynamic Regret bounded by Variation in Utilities" (DRVU) property, and to show that many existing algorithms indeed satisfy DRVU.

Furthermore, to achieve strong guarantees for all performance measures without any prior knowledge, we also need to deploy a two-layer structure, with a meta-algorithm learning over and combining decisions of a group of base-learners, each of which satisfies the DRVU property but uses a different step size. Although such a framework has been used in many prior works in online learning (see for example the latest advances (Chen et al., 2021; Zhao et al., 2021) and references therein), several new ingredients are required to achieve our results. First, when updating the meta-algorithm, a correction term related to the stability of each base-algorithm is injected into the loss for the corresponding base-algorithm, which plays a key role in the analysis. More specifically, we show (in Lemma 10) an explicit bound for the stability of the meta-algorithm's decisions, whose proof requires a careful analysis using the correction terms above and the unique game structure. Second, we also introduce a set of additional "dummy" base-algorithms that always play some fixed action. This plays a key role in controlling the dynamics of the base-learners' outputs and turns out to be critical when bounding the duality gap.

Related Work. Two-player zero-sum game is one of the most fundamental problems in game theory, whose studies date back to the seminal work of von Neumann (1928). Freund & Schapire (1999) discovered the profound connections between zero-sum games and no-regret online learning, and since then there have been extensive studies on designing no-regret algorithms to solve games in the stationary set-

[1] This is implicitly implied by the results of Hsieh et al. (2021), as our Lemma 17 shows that in the stationary case dynamic NE-regret is bounded by the individual regret.

ting (Rakhlin & Sridharan, 2013; Daskalakis et al., 2015; Syrgkanis et al., 2015; Chen & Peng, 2020; Wei et al., 2021; Daskalakis et al., 2021). We refer the reader to (Daskalakis et al., 2021) for a more thorough discussion on the literature. Several recent works start considering the problem of learning over a sequence of non-stationary payoffs under different structures, including zero-sum matrix games (Mai et al., 2018; Cardoso et al., 2019; Fiez et al., 2021), convex-concave games (Roy et al., 2019) and strongly monotone games (Duvocelle et al., 2021). For zero-sum games, Fiez et al. (2021); Mai et al. (2018) focus on the periodic case and proves divergence results for a class of learning algorithms; (Cardoso et al., 2019) is the closest to our work, but as mentioned, we argue that their proposed measure (NE-regret) is not always appropriate (see Section 3.1). Learning in time-varying games is also related to bandits with knapsack (Badanidiyuru et al., 2018; Immorlica et al., 2019).

Organization. We formulate the problem set in Section 2, then present the performance measures in Section 3 and our algorithm in Section 4. Next, we provide theoretical guarantees in Section 5. We finally report the empirical results in Section 6 and conclude the paper in Section 7.

2. Problem Setup and Notations

We consider the following problem of two players (called x-player and y-player) repeatedly playing a zero-sum game for T rounds, with m fixed actions for x-player and n fixed actions for y-player. At each round $t \in [T] \triangleq \{1, \dots, T\}$, the environment first chooses a payoff matrix $A_t \in [-1, 1]^{m \times n}$, whose (i, j) entry denotes the loss/reward for x-player/y-player when they play action i and action j respectively. Without knowing A_t, x-player (y-player) decides her own mixed strategy (that is, a distribution over actions) $x_t \in \Delta_m$ ($y_t \in \Delta_n$), where Δ_k denotes the probability simplex $\Delta_k = \{u \in \mathbb{R}^k_{\geq 0} \mid \sum_{i=1}^k u_i = 1\}$. At the end of this round, x-player suffers expected loss $x_t^\top A_t y_t$ and observes the loss vector $A_t y_t$, while y-player receives the expected reward $x_t^\top A_t y_t$ and observes the reward vector $x_t^\top A_t$. Note that neither player observes the matrix A_t itself.

When A_t is fixed for all t, this exactly recovers the standard stationary setting considered in for example (Syrgkanis et al., 2015). Having a time-varying A_t allows us to capture various possible sources of non-stationarity. In fact, A_t can even be decided by an adaptive adversary who makes the decision knowing the players' algorithm and their decisions in earlier rounds. Our setting is almost the same as (Cardoso et al., 2019), except that the feedback they consider is either the entire matrix A_t (stronger than ours) or just one entry of A_t sampled according to (x_t, y_t) (weaker than ours).

For each game matrix A_t, define the set of minimax strategies for x-player as $\mathcal{X}_t^* = \operatorname{argmin}_{x \in \Delta_m} \max_{y \in \Delta_n} x^\top A_t y$ and similarly the set of maximin strategies for y-player as $\mathcal{Y}_t^* = \operatorname{argmax}_{y \in \Delta_n} \min_{x \in \Delta_m} x^\top A_t y$. It is well-known that any pair $(x_t^*, y_t^*) \in \mathcal{X}_t^* \times \mathcal{Y}_t^*$ forms a Nash equilibrium of A_t with the following saddle-point property: $x_t^{*\top} A_t y \leq x_t^{*\top} A_t y_t^* \leq x^\top A_t y_t^*$ holds for any $x \in \Delta_m$ and $y \in \Delta_n$. Throughout the paper, (x_t^*, y_t^*) denotes an arbitrary Nash equilibrium of A_t.

Notations. For a real-valued matrix $A \in \mathbb{R}^{m \times n}$, its infinity norm is defined as $\|A\|_\infty \triangleq \max_{i,j} |A_{ij}|$. We use $\mathbf{1}_N$ and $\mathbf{0}_N$ to denote the all-one and all-zero vectors of length N. For conciseness, we often hide polynomial dependence on the size of the game (that is, m and n) in the $\mathcal{O}(\cdot)$-notation. The $\widetilde{\mathcal{O}}(\cdot)$-notation further omits logarithmic dependence on T. We sometimes write $\min_{x \in \Delta_m}$ ($\min_{y \in \Delta_n}$) simply as $\min_x$ ($\min_y$) when there is no confusion.

3. How to Measure the Performance?

With the learning protocol specified, the next pressing question is to determine what the goal is when designing algorithms for the two players. When A_t is fixed, most studies consider minimizing individual regret for each player and some form of convergence to a Nash equilibrium of the fixed game as the two primary goals. While minimizing individual regret is still naturally defined when A_t is changing over time, it is less clear what other desirable game-theoretic guarantees are in this case. In Section 3.1, we formally discuss three performance measures that we think are reasonable for this problem. Then in Section 3.2, we further discuss how to measure the non-stationarity of the sequence $\{A_t\}_{1:T}$ that will play a role in how well the players can do under some of the performance measures.

3.1. Performance Measures

① Individual Regret. The first measure we consider is the standard individual regret. For x-player, this is defined as

$$\mathrm{Reg}_T^x \triangleq \sum_{t=1}^T x_t^\top A_t y_t - \min_{x \in \Delta_m} \sum_{t=1}^T x^\top A_t y_t, \qquad (1)$$

that is, the difference between her total loss and that of the best fixed strategy (assuming the same behavior from the opponent). Similarly, the regret for y-player is defined as $\mathrm{Reg}_T^y \triangleq \max_{y \in \Delta_n} \sum_{t=1}^T x_t^\top A_t y - \sum_{t=1}^T x_t^\top A_t y_t$. Achieving sublinear (in T) individual regret implies that on average each player performs almost as well as their best fixed strategy, and this is arguably the most standard and basic goal for online learning problems.

② Duality Gap. For a game matrix A_t, the duality gap of a pair of strategy (x_t, y_t) is defined as $\max_{y \in \Delta_n} x_t^\top A_t y - \min_{x \in \Delta_m} x^\top A_t y_t$. It is always nonnegative since $\max_{y \in \Delta_n} x_t^\top A_t y \geq x_t^\top A_t y_t^* \geq x_t^{*\top} A_t y_t^* \geq x_t^{*\top} A_t y_t \geq \min_{x \in \Delta_m} x^\top A_t y_t$, and it is zero if and only

if (x_t, y_t) is a Nash equilibrium of A_t. Thus, the duality gap measures how close (x_t, y_t) is to the equilibria in some sense. We thus naturally use the cumulative duality gap:

$$\text{Dual-Gap}_T \triangleq \sum_{t=1}^{T} \left(\max_{y \in \Delta_n} x_t^\top A_t y - \min_{x \in \Delta_m} x^\top A_t y_t \right), \quad (2)$$

as another performance measure. When A_t is fixed, this measure is considered in (Wei et al., 2021) for example.

③ Dynamic Nash Equilibrium (NE)-Regret. Before introducing this last measure, we first review what Cardoso et al. (2019) proposed as the goal for this problem, that is, ensuring small Nash Equilibrium (NE)-regret, defined as

$$\text{NE-Reg}_T \triangleq \left| \sum_{t=1}^{T} x_t^\top A_t y_t - \min_{x \in \Delta_m} \max_{y \in \Delta_n} \sum_{t=1}^{T} x^\top A_t y \right|. \quad (3)$$

In words, this is the difference between the cumulative loss of x-player (or equivalently the cumulative reward of y-player) and the minimax value of the cumulative payoff matrix ($\sum_{t=1}^{T} A_t$). While this might appear to be a reasonable generalization of individual regret for a central controller who decides x_t and y_t jointly, we argue below that this measure is in fact often inappropriate for two reasons.

The first reason is in fact already hinted in (Cardoso et al., 2019): they proved that no algorithm can always ensure sublinear NE-regret and simultaneously sublinear individual regret for both players. Given that minimizing individual regret selfishly is a natural impulse and the standard goal for each player, NE-regret can only make sense when both players are controlled by a centralized algorithm.

The second reason is perhaps more profound. Consider the following two-phase example: when $t \le T/2$, $A_t = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$; when $t > T/2$, $A_t = \begin{pmatrix} 1 & -1 \\ 1 & -1 \end{pmatrix}$.[2] It is straightforward to verify that: when $t \le T/2$, the equilibrium for A_t is the uniform distribution for both players, leading to game value 0; when $t > T/2$, the equilibrium is such that y-player always picks the first column, leading to game value 1; and the equilibrium for the cumulative game matrix $\sum_{t=1}^{T} A_t = \begin{pmatrix} T & -T \\ 0 & 0 \end{pmatrix}$ is x-player picking the second row while y-player picking the first column, leading to game value 0. To sum up, even if both players play perfectly in each round using the equilibrium, their NE-regret is still $|T/2 - 0| = T/2$, which is a vacuous bound linear in T!

Motivated by the observations above, we propose a variant of NE-regret as the third performance measure, called *dynamic NE-regret*:[3]

$$\text{DynNE-Reg}_T \triangleq \left| \sum_{t=1}^{T} x_t^\top A_t y_t - \sum_{t=1}^{T} \min_{x \in \Delta_m} \max_{y \in \Delta_n} x^\top A_t y \right|.$$

Compared to NE-regret, here we move the minimax operation inside the summation, making it the cumulative difference between x-player's loss and the minimax game value in each round. In other words, similarly to duality gap, dynamic NE-regret provides yet another way to measure in each round, how close (x_t, y_t) is to the equilibria of A_t from the game value perspective.

The connection between NE-regret and Dynamic NE-regret is on the surface analogous to that between standard regret and dynamic regret (Zinkevich, 2003) (see Appendix B.1 for definitions and more related discussions). However, while dynamic regret is always no less than standard regret, *Dynamic NE-regret could be smaller than NE-regret* — simply consider our earlier two-phase example: the perfect players (who always play an equilibrium) clearly have 0 dynamic NE-regret, but their NE-regret is $T/2$ as discussed. This example also shows that dynamic NE-regret is more reasonable compared to NE-regret. Moreover, as will become clear soon, dynamic NE-regret is compatible with individual regret (and also duality gap), in the sense there are algorithms that provably perform well under all these measures.

We conclude this section with the following two remarks.

Remark 1 (Comparisons of the three measures). Both individual regret and dynamic NE-regret are bounded by duality gap (see proofs in Appendix B.2), but the latter could be much larger. On the other hand, individual regret and dynamic NE-regret are generally incomparable.

Remark 2 (Other possibilities). The three measures we consider are by no mean the only possibilities. Another reasonable one is the tracking error $\sum_{t=1}^{T} (\|x_t - x_t^*\|_1 + \|y_t - y_t^*\|_1)$ that directly measures the distance between (x_t, y_t) and the equilibrium (x_t^*, y_t^*) (assuming unique equilibrium for simplicity). This is considered in (Roy et al., 2019; Balasubramanian & Ghadimi, 2021) (for different problems). However, tracking error bounds are in fact not well studied even when A_t is fixed — the best known results still depend on some problem-dependent constant that can be arbitrarily large (Daskalakis & Panageas, 2019; Wei et al., 2021). Deriving tracking error bounds in our setting is thus beyond the scope of this paper. Note that in many optimization studies, one often only cares about finding a point that is close to the optimal solution in terms of their function value instead of their absolute distance. Our dynamic NE-regret and duality gap are both in this same sprite by looking at the game value instead of the actual distance as in tracking error.

[2] The same example is in fact also used by Cardoso et al. (2019) to prove the incompatibility of individual regret and NE-regret.

[3] In fact, a preprint by Roy et al. (2019) also considers a similar measure for general convex-concave problem, but we believe that their results are incorrect. Specifically, they claim (in their Theorem 4.3) that an $\widetilde{\mathcal{O}}(\sqrt{T})$ bound is always achievable for dynamic NE-regret, but this is clearly impossible because when A_t always has identical columns (so y-player does not play any role), dynamic NE-regret becomes the dynamic regret (Zinkevich, 2003) for x-player, which is well-known to be $\Omega(T)$ in the worst case.

3.2. Non-stationarity Measures

For duality gap and dynamic NE-regret, it is not difficult to see that if A_t changes drastically over time, then no meaningful guarantees are possible. This is similar to dynamic regret in standard online learning problems, where guarantees are always expressed in terms of some non-stationarity measure of the environment and are meaningful only when the non-stationarity is reasonably small. In our setting, we consider the following three different ways to measure non-stationarity of the sequence $\{A_t\}_{t=1}^T$.

Variation of Nash Equilibria. Recall the notation $\mathcal{X}_t^* \times \mathcal{Y}_t^*$, the set of Nash equilibria for matrix A_t. Define the variation of Nash equilibria as:

$$P_T \triangleq \min_{\forall t,(x_t^*,y_t^*)\in\mathcal{X}_t^*\times\mathcal{Y}_t^*} \sum_{t=2}^{T} \left(\|x_t^* - x_{t-1}^*\|_1 + \|y_t^* - y_{t-1}^*\|_1\right),$$

which quantifies the drift of the Nash equilibria of the game matrices in ℓ_1-norm.

Variation/Variance of Game Matrices. The path-length variation and the variance of $\{A_t\}_{t=1}^T$ are respectively defined as

$$V_T \triangleq \sum_{t=2}^{T} \|A_t - A_{t-1}\|_\infty^2, \quad W_T \triangleq \sum_{t=1}^{T} \|A_t - \bar{A}\|_\infty,$$

where $\bar{A} = \frac{1}{T}\sum_{t=1}^T A_t$ is the averaged game matrix.

Clearly, P_T, V_T, and W_T are all $\Theta(T)$ in the worst case, and 0 when A_t is fixed over time. For dynamic regret and duality gap, the natural goal is to enjoy sublinear bounds whenever (some of) these non-stationarity measures are sublinear (which we indeed achieve).

We conclude by pointing out some connections between these non-stationarity measures. First, $V_T \le 8W_T$ holds but the former could be much smaller. Second, P_T is generally not comparable with V_T and W_T, and there are examples where $P_T = 0$ and $V_T = W_T = \Theta(T)$, or $P_T = \Theta(T)$ and $V_T = W_T = \mathcal{O}(1)$. We defer all details to Appendix B.

4. Proposed Algorithm

In this section, we present our proposed algorithm for time-varying games, which provably achieves favorable guarantees under all three performance measures. To illustrate the ideas behind our algorithm design, we first review how (Syrgkanis et al., 2015) achieves fast convergence results for a fixed game, followed by a detailed discussion on how to generalize their idea and overcome the difficulties brought by time-varying games. For conciseness, throughout the section we focus on the x-player; how the y-player should behave is completely symmetric.

For a fixed game $A_t = A$, Syrgkanis et al. (2015) proposed that each player should deploy an online learning algorithm that satisfies a specific property called "Regret bounded by Variation in Utilities" (RVU). More specifically, an online learning algorithm proposes $x_t \in \Delta_m$ at the beginning of round t, and then receives a loss vector $g_t \in \mathbb{R}^m$ and suffers loss $\langle x_t, g_t\rangle$. Its regret against a comparator $u \in \Delta_m$ after T rounds is naturally $\sum_{t=1}^T \langle x_t - u, g_t\rangle$, and the RVU property states that this should be bounded by $\alpha + \beta\sum_{t=2}^T \|g_t - g_{t-1}\|_\infty^2 - \gamma\sum_{t=2}^T \|x_t - x_{t-1}\|_1^2$ for some parameters $\alpha, \beta, \gamma > 0$.[4] To see why RVU property is useful, consider x-player deploying such an algorithm with g_t set to $A_t y_t = A y_t$. Then her regret is further bounded as $\alpha + \beta\sum_{t=2}^T \|Ay_t - Ay_{t-1}\|_\infty^2 - \gamma\sum_{t=2}^T \|x_t - x_{t-1}\|_1^2 \le \alpha + \beta\sum_{t=2}^T \|y_t - y_{t-1}\|_1^2 - \gamma\sum_{t=2}^T \|x_t - x_{t-1}\|_1^2$. Therefore, as long as y-player also deploys the same algorithm, by symmetry, the sum of their regret is at most $\alpha + (\beta - \gamma)(\sum_{t=2}^T \|x_t - x_{t-1}\|_1^2 + \sum_{t=2}^T \|y_t - y_{t-1}\|_1^2)$, which can be simply bounded by (a constant) α as long as $\beta \le \gamma$. Many useful guarantees can then be obtained as a corollary of the fact that the sum of regret is small.

In our setting where A_t is changing over time, our first observation is that instead of the sum of the two players' regret, what we need to control is the sum of their dynamic regret (Zinkevich, 2003), which plays an important role when deriving guarantees for all the three measures (*including individual regret*). Specifically, for an online learning algorithm producing x_t and receiving g_t, its dynamic regret against a sequence of comparators $u_1, \ldots, u_T \in \Delta_m$ is defined as $\sum_{t=1}^T \langle x_t - u_t, g_t\rangle$. Generalizing RVU, we naturally introduce the following "Dynamic Regret bounded by Variation in Utilities" (DRVU) property.

Definition 3 (DRVU Property)**.** Denote by $\mathcal{A}(\eta)$ an online learning algorithm with a parameter $\eta > 0$. We say that it satisfies the *Dynamic Regret bounded by Variation in Utilities* property (abbreviated as DRVU(η)) with parameters $\alpha, \beta, \gamma > 0$, if its dynamic regret $\sum_{t=1}^T \langle x_t - u_t, g_t\rangle$ on any loss sequence $g_1, \ldots, g_T$ with respect to any comparator sequence $u_1, \ldots, u_T$ is bounded by

$$\frac{\alpha}{\eta}(1 + P_T^u) + \eta\beta\sum_{t=1}^{T} \|g_t - g_{t-1}\|_\infty^2 - \frac{\gamma}{\eta}\sum_{t=2}^{T} \|x_t - x_{t-1}\|_1^2,$$

where $P_T^u \triangleq \sum_{t=2}^T \|u_t - u_{t-1}\|_1$ is the path-length of the comparator sequence.

Compared to RVU, DRVU naturally replaces the first constant term in the regret bound with a term depending on the path-length of the comparator sequence. We also add another step size parameter η (whose role will become clear

[4] Without loss of generality, we here focus on $(\|\cdot\|_1, \|\cdot\|_\infty)$ norm, and it is straightforward to generalize the argument to general primal-dual norm as in (Syrgkanis et al., 2015).

soon). Recent studies in dynamic regret (Zhao et al., 2020; 2021) show that variants of optimistic Online Mirror Descent (such as Optimistic Gradient Descent and Optimistic Hedge) indeed satisfy DRVU with $\alpha, \beta, \gamma = \widetilde{\Theta}(1)$; see Appendix D for formal statements and proofs.

Now, if x-player deploys an algorithm satisfying DRVU and feeds it with loss vector $g_t = A_t y_t$ (and similarly y-player does the same), we can indeed prove a desired guarantee for each of the three performance measures. However, the tuning of η will require the knowledge of the unknown parameters P_T, V_T, W_T and, perhaps more importantly, be different for each different measures. To obtain an adaptive algorithm that performs well under all three measures without any prior knowledge, we further propose a two-layer structure with a meta-algorithm learning over and combining decisions of a set of base-learners, each of which satisfies DRVU(η) but with a different step size η. While this idea of "learning over learning algorithms" is not new in online learning, we will discuss below what extra difficulties show up in our case and how we address them.

4.1. Base-learners

Define $N = \lfloor \frac{1}{2} \log_2 T \rfloor + 1$. Our algorithm maintains $N + m$ base-learners: $\forall i \in [N]$, the i-th base-learner is any algorithm that satisfies DRVU(η_i^x), where $\eta_i^x = \frac{2^{i-1}}{L\sqrt{T}}$ and

$$L = \max\left\{4, \sqrt{16c\beta}, \sqrt{8c\beta/\gamma}\right\} \tag{4}$$

(β and γ are the parameters from DRVU and $c = \widetilde{\mathcal{O}}(1)$ is a constant whose exact value can be found in the proof); the last m base-learners are dummy learners, with the $(j+N)$-th one always outputting the basis vector $e_j \in \Delta_m$ (that is, always choosing the j-th action). We note that the dummy base-learners are important in controlling the duality gap (but not the other two measures). We let $\mathcal{S}_x \triangleq \mathcal{S}_{1,x} \cup \mathcal{S}_{2,x}$ with $\mathcal{S}_{1,x} = [N]$ and $\mathcal{S}_{2,x} = \{N+1, \ldots, N+m\}$ denote the set of indices of base-learners.

At round t, each base-learner i submits her decision $x_{t,i} \in \Delta_m$ to the meta-algorithm, who decides the final decision x_t. Upon receiving the feedback $A_t y_t$, the meta-algorithm sends the same (as the loss vector g_t) to each base-learner $i \in \mathcal{S}_{1,x}$ (no updates needed for the dummy base-learners).

4.2. Meta-algorithm

With all the decisions $\{x_{t,i}\}_{i \in \mathcal{S}_x}$ collected from the base-learners, the meta-algorithm outputs the final decision $x_t = \sum_{i \in \mathcal{S}_x} p_{t,i} x_{t,i}$,[5] where $p_t \in \Delta_{|\mathcal{S}_x|}$ is a distribution over the base-learners updated according to a version of Optimistic Online Gradient Descent (OOGD) (Rakhlin & Sridharan, 2013; Syrgkanis et al., 2015):

$$\begin{aligned} p_t &= \operatorname*{argmin}_{p \in \Delta_{|\mathcal{S}_x|}} \left\{ \varepsilon_t^x \langle p, m_t^x \rangle + \|p - \widehat{p}_t\|_2^2 \right\}, \\ \widehat{p}_{t+1} &= \operatorname*{argmin}_{p \in \Delta_{|\mathcal{S}_x|}} \left\{ \varepsilon_t^x \langle p, \ell_t^x \rangle + \|p - \widehat{p}_t\|_2^2 \right\}. \end{aligned} \tag{5}$$

Here, $\varepsilon_t^x > 0$ is a time-varying learning rate, $\{\widehat{p}_t\}_{t=1,2,\ldots}$ is an auxiliary sequence (starting with $\widehat{p}_1$ as the uniform distribution) updated via projected gradient descent using some loss vector sequence $\ell_1^x, \ell_2^x, \ldots \in \mathbb{R}^{|\mathcal{S}_x|}$, and p_t is updated via projected gradient descent from the distribution $\widehat{p}_t$ and using a loss predictor $m_t^x \in \mathbb{R}^{|\mathcal{S}_x|}$. It remains to specify what ℓ_t^x and m_t^x are (the tuning of the learning rate will be specified in the final algorithm).

Since base-learner i predicts $x_{t,i}$ and receives loss vector $A_t y_t$, it is natural to set its loss $\ell_{t,i}^x$ as $x_{t,i}^\top A_t y_t$ from the meta-algorithm's perspective. In light of standard OOGD, m_t^x should then be set to $x_{t,i}^\top A_{t-1} y_{t-1}$, meaning that the last loss vector $A_{t-1} y_{t-1}$ is used to predict the current one (that is unknown yet when computing p_t). However, this setup leads to the following issue. When applying DRVU(η_i^x) to this base-learner, we see that a negative term related to $\|x_{t,i} - x_{t-1,i}\|_1^2$ and a positive term related to $\|y_t - y_{t-1}\|_1^2$ arise (the latter is from $\|A_t y_t - A_{t-1} y_{t-1}\|_\infty^2 \le 2\|A_t - A_{t-1}\|_\infty^2 + 2\|y_t - y_{t-1}\|_1^2$, with the first term only related to the non-stationarity of game matrices). By symmetry, y-player contributes a positive term $\|x_t - x_{t-1}\|_1^2$, which now cannot be canceled by $\|x_{t,i} - x_{t-1,i}\|_1^2$, unlike the case with only one learner for each player discussed earlier.

To address this issue, we propose to add a stability correction term to both ℓ_t^x and m_t^x. Concretely, they are defined as $\ell_{1,i}^x = x_{1,i}^\top A_1 y_1$ and $m_{1,i}^x = 0, \forall i$, and for all $t \ge 2$:

$$\ell_{t,i}^x = x_{t,i}^\top A_t y_t + \lambda \|x_{t,i} - x_{t-1,i}\|_1^2, \tag{6}$$

$$m_{t,i}^x = x_{t,i}^\top A_{t-1} y_{t-1} + \lambda \|x_{t,i} - x_{t-1,i}\|_1^2, \tag{7}$$

where $\lambda = \frac{\gamma L}{2}$ (γ is the parameter from DRVU). From a technical perspective, this introduces to the regret a nega-

Algorithm 1 Algorithm for the x-player

Input: a base-algorithm $\mathcal{A}(\eta)$ satisfying DRVU(η).
Initialize: a set of base-learners $\mathcal{S}_x$ as described in Section 4.1, $\widehat{p}_1 = \frac{1}{|\mathcal{S}_x|}\mathbf{1}_{|\mathcal{S}_x|}$, learning rate $\varepsilon_1^x = \frac{1}{L}$ (*c.f.* Eq. (4)).
for $t = 1, \ldots, T$ **do**
  Receive $x_{t,i} \in \Delta_m$ from each base-learner $i \in \mathcal{S}_x$.
  Compute m_t^x based on Eq. (7) and p_t based on Eq. (5).
  Play the final decision $x_t = \sum_{i \in \mathcal{S}_x} p_{t,i} x_{t,i}$.
  Suffer loss $x_t^\top A_t y_t$ and observe the loss vector $A_t y_t$.
  Compute ℓ_t^x based on Eq. (6) and $\widehat{p}_{t+1}$ based on Eq. (5).
  Update $\varepsilon_{t+1}^x = 1/\sqrt{L^2 + \sum_{s=2}^t \|A_t y_t - A_{t-1} y_{t-1}\|_\infty^2}$.
  Send $A_t y_t$ as the feedback to each base-learner.
end

[5] Note the slight abuse of notations here: while $p_{t,i}$ represents the i-th entry of vector p_t, $x_{t,i}$ is *not* the i-th entry of x_t.

tive term $\sum_{i\in\mathcal{S}_x} p_{t,i}\|x_{t,i}-x_{t-1,i}\|_1^2$, and a positive term $\|x_{t,i}-x_{t-1,i}\|_1^2$ which can be canceled by the aforementioned negative term from DRVU(η_i^x). To see why the extra negative term is useful, notice that the troublesome term $\|x_t-x_{t-1}\|_1^2$ from DRVU(η_i^x) can be bounded as

$$\|x_t-x_{t-1}\|_1^2=\left\|\sum_{i\in\mathcal{S}_x}p_{t,i}x_{t,i}-\sum_{i\in\mathcal{S}_x}p_{t-1,i}x_{t-1,i}\right\|_1^2$$
$$\le 2\sum_{i\in\mathcal{S}_x}p_{t,i}\|x_{t,i}-x_{t-1,i}\|_1^2+2\|p_t-p_{t-1}\|_1^2,$$

where the first term can exactly be canceled by the extra negative term introduced by the correction term, and the second term can in fact also be canceled in a standard way since the meta-algorithm itself can be shown to satisfy RVU. This explains the design of our correction terms from a technical level. Intuitively, injecting this correction term guides the meta-algorithm to bias toward the more stable base-learners, hence also stabilizing the final decision x_t.

We note that a similar technique was used in analyzing gradient-variation dynamic regret for online convex optimization (Zhao et al., 2021). Our approach is different from theirs in the sense that there is only one player in their setting and the correction term is used to cancel the additional gradient variation introduced by the variation of her own decision. In contrast, in our setting the correction term is used to cancel the opponent's gradient variation.

To summarize, our final algorithm (for the x-player) is presented in Algorithm 1. We also include the symmetric version for the y-player in Algorithm 2 (Appendix A) for completeness. We emphasize again that this is a parameter-free algorithm that does not require any prior knowledge of the environment.

5. Theoretical Guarantees and Analysis

In this section, we first provide the guarantees of our algorithm under each of the three performance measures, and then highlight several key ideas in the analysis, with the full proofs deferred to Appendix E. Recall that our guarantees are all expressed in terms of the non-stationarity measures P_T, V_T, and W_T, defined in Section 3.2. Also, to avoid showing the cumbersome dependence on the DRVU parameters (α, β, γ) in all our bounds, we will simply assume that they are all $\Theta(1)$, which, as mentioned earlier and shown in Appendix D, is indeed the case for standard algorithms.

5.1. Performance Guarantees

We state our results for each performance measure separately below, but emphasize again that they hold simultaneously. First, we show the individual regret bound.

Theorem 4 (Individual Regret)**.** *When the x-player uses Algorithm 1, irrespective of y-player's strategies, we have*

$$\text{Reg}_T^x=\sum_{t=1}^T x_t^\top A_t y_t-\min_x\sum_{t=1}^T x^\top A_t y_t=\widetilde{\mathcal{O}}(\sqrt{T}).$$

Furthermore, if x-player follows Algorithm 1 and y-player follows Algorithm 2, then individual regret satisfies:

$$\max\{\text{Reg}_T^x,\text{Reg}_T^y\}=\widetilde{\mathcal{O}}\Big(\sqrt{1+V_T+\min\{P_T,W_T\}}\Big).$$

The first statement of Theorem 4 provides a robustness guarantee for our algorithm — no matter how non-stationary the game matrices are and no matter how the opponent behaves, following our algorithm always ensures $\widetilde{\mathcal{O}}(\sqrt{T})$ individual regret, the standard worst-case regret bound. On the other hand, when both players follow our algorithm, their individual regret could be even smaller depending on the non-stationarity. In particular, as long as $V_T+\min\{P_T,W_T\}=o(T)$ (that is, not the worst case scenario), our bound becomes $o(\sqrt{T})$. Also note that P_T and W_T are generally incomparable (see Appendix B), but our bound achieves the minimum of them, thus achieving the best of both worlds.

When the game matrix is fixed, we have $P_T=V_T=W_T=0$, immediately leading to the following corollary.

Corollary 5. *When x-player follows Algorithm 1 and y-player follows Algorithm 2, if $A_t=A$ for all $t\in[T]$, then* $\max\{\text{Reg}_T^x,\text{Reg}_T^y\}=\widetilde{\mathcal{O}}(1)$.

The best known individual regret bound for learning in a fixed two-player zero-sum game is $\mathcal{O}(1)$ (Hsieh et al., 2021). Our result matches theirs up to logarithmic factors.

The next theorem presents the dynamic NE-regret bound.

Theorem 6 (Dynamic NE-Regret)**.** *When x-player follows Algorithm 1 and y-player follows Algorithm 2, we have the following dynamic NE-regret bound:*

$$\text{DynNE-Reg}_T=\left|\sum_{t=1}^T x_t^\top A_t y_t-\sum_{t=1}^T\min_{x\in\Delta_m}\max_{y\in\Delta_n}x^\top A_t y\right|$$
$$=\widetilde{\mathcal{O}}\big(\min\{\sqrt{(1+V_T)(1+P_T)}+P_T,1+W_T\}\big).$$

Similarly, our dynamic NE-regret bound is $o(T)$ as long as P_T *or* W_T is $o(T)$. When the game matrix is fixed, we again obtain the following direct corollary by noticing $P_T=V_T=W_T=0$ in this case.

Corollary 7. *When x-player follows Algorithm 1 and y-player follows Algorithm 2, if $A_t=A$ for all $t\in[T]$, then* $\text{DynNE-Reg}_T=\widetilde{\mathcal{O}}(1)$.

In fact, when the game is fixed, dynamic NE-regret degenerates to NE-regret of Cardoso et al. (2019)

as $\sum_{t=1}^T \min_x \max_y x^\top A y = \min_x \max_y \sum_{t=1}^T x^\top A y$. Their algorithm would achieve $\widetilde{\mathcal{O}}(\sqrt{T})$ (dynamic) NE-regret in this case. A better $\mathcal{O}(1)$ result is implicitly implied by the aforementioned work of Hsieh et al. (2021), as we show (in Lemma 17) that (dynamic) NE-regret is bounded by the individual regret in this stationary case. Our result again matches theirs up to logarithmic factors.

The last theorem provides an upper bound for duality gap.

Theorem 8 (Duality Gap). *When x-player follows Algorithm 1 and y-player follows Algorithm 2, we have*

$$\text{Dual-Gap}_T = \sum_{t=1}^T \max_{y\in\Delta_n} x_t^\top A_t y - \sum_{t=1}^T \min_{x\in\Delta_m} x^\top A_t y_t$$
$$= \widetilde{\mathcal{O}}\Big(\min\{T^{\frac{3}{4}}(1+Q_T)^{\frac{1}{4}}, T^{\frac{1}{2}}(1+Q_T^{\frac{3}{2}} + P_T Q_T)^{\frac{1}{2}}\}\Big),$$

where $Q_T \triangleq V_T + \min\{P_T, W_T\}$.

Once again the bound is $o(T)$ whenever $Q_T = o(T)$, and it implies the following corollary.

Corollary 9. *When x-player follows Algorithm 1 and y-player follows Algorithm 2, if $A_t = A$ for all $t \in [T]$, then* $\text{Dual-Gap}_T = \widetilde{\mathcal{O}}(\sqrt{T})$.

Notably, the best known result of duality gap for a fixed game is $\mathcal{O}(\sqrt{T})$ (Wei et al., 2021), and our result again matches theirs up to logarithmic factors.

5.2. Key Ideas for Analysis

We now highlight some key components and novelty of our analysis. As mentioned in Section 4, to bound all the three metrics, the key is to bound the sum of the two players' dynamic regret, which further requires controlling the stability of the strategies between consecutive rounds. The following key lemma shows how such stability is controlled by the non-stationarity measures of $\{A_t\}_{t=1}^T$.

Lemma 10. *When x-player follows Algorithm 1 and y-player follows Algorithm 2, we have both $\sum_{t=2}^T \|x_t - x_{t-1}\|_1^2$ and $\sum_{t=2}^T \|y_t - y_{t-1}\|_1^2$ bounded by*

$$\widetilde{\mathcal{O}}\Big(\min\big\{\sqrt{(1+V_T)(1+P_T)} + P_T, 1+W_T\big\}\Big).$$

This lemma implies an $\widetilde{\mathcal{O}}(1)$ stability bound when the game is fixed, which is first proven in (Hsieh et al., 2021) where both players run the Optimistic Hedge algorithm with an adaptive learning rate. Our result generalizes theirs but requires a novel analysis due to both the time-varying matrices and the two-layer structure of our algorithm. As another note, this lemma also highlights another difference of our method compared to (Zhao et al., 2021) — as mentioned in Section 4 our algorithm shares some similarity with theirs, but no explicit stability bound is proven or required in their problem, while stability is crucial for our whole analysis. We next present the proof sketch for Lemma 10. More details can be found in Appendix F.

Proof Sketch. We show in Lemma 15 that the sum of the two players' dynamic regret (against a sequence $u_1, \ldots, u_T \in \Delta_m$ for x-player and a sequence $v_1, \ldots, v_T \in \Delta_n$ for y-player) can be bounded by

$$\sum_{t=1}^T \left(x_t^\top A_t v_t - u_t^\top A_t y_t\right) = \widetilde{\mathcal{O}}\left(\frac{1+P_T}{\eta} + \eta(1+V_T)\right)$$
$$- \Omega\left(\sum_{t=1}^T (\|x_t - x_{t-1}\|_1^2 + \|y_t - y_{t-1}\|_1^2)\right),$$

for *any* step size $0 < \eta \le \widetilde{\mathcal{O}}(1)$. Here, $P_T \triangleq P_T^u + P_T^v$, $P_T^u \triangleq \sum_{t=2}^T \|u_t - u_{t-1}\|_1$ and $P_T^v \triangleq \sum_{t=2}^T \|v_t - v_{t-1}\|_1$ are the path-length of comparators. Then, Lemma 10 can be proven by taking different choices of η and the comparator sequence. For example, consider picking $(u_t, v_t) = (x_t^*, y_t^*)$. Since the saddle point property ensures $x_t^\top A_t y_t^* - x_t^{*\top} A_t y_t \ge 0$, rearranging and picking the optimal η thus gives the first bound $\widetilde{\mathcal{O}}(\sqrt{(1+V_T)(1+P_T)} + P_T)$ on the stability. To prove the second bound, pick $(u_t, v_t) = (\bar{u}^*, \bar{v}^*)$ where $(\bar{u}^*, \bar{v}^*)$ is a Nash equilibrium of the averaged game matrix. Then, we have $P_T^u = P_T^v = 0$ and $\sum_{t=1}^T x_t^\top A_t v_t - \sum_{t=1}^T u_t^\top A_t y_t \ge -\mathcal{O}(W_T)$. Rearranging, picking the optimal η, and using $V_T \le \mathcal{O}(W_T)$ then proves the $\widetilde{\mathcal{O}}(1+W_T)$ bound. □

We finally briefly mention two more new ideas when bounding the duality gap. First, we apply a reduction from general dynamic regret that competes with any comparator sequence to its worst-case variant, which in some place helps bound the duality gap by the aforementioned stability. Second, we show how the extra set of "dummy" base-learners enables the meta-algorithm to have a direct control on the duality gap. We refer the reader to Appendix E.3 for more details.

6. Experiment

In this section, we provide empirical studies on the performance of our proposed algorithm in time-varying games.

We construct an environment such that $P_T = \Theta(\sqrt{T})$, $W_T = \Theta(T^{\frac{3}{4}})$, and $V_T = \Theta(\sqrt{T})$. Under this environment, our theoretical results indicate that $\max\{\text{Reg}_T^x, \text{Reg}_T^y\} \le \widetilde{\mathcal{O}}(T^{\frac{1}{4}})$, $\text{NE-Reg}_T \le \widetilde{\mathcal{O}}(\sqrt{T})$ and $\text{Dual-Gap}_T \le \widetilde{\mathcal{O}}(T^{\frac{7}{8}})$. Our empirical results validate the effectiveness of our algorithm in this environment, and in fact its performance is even better than the theoretical upper bounds, which also encourage us to investigate better guarantees in the future.

The environment setup is as follows. We set the size of game matrix to be $m \times n$ with $m = 2$ and $n = 2$. The total

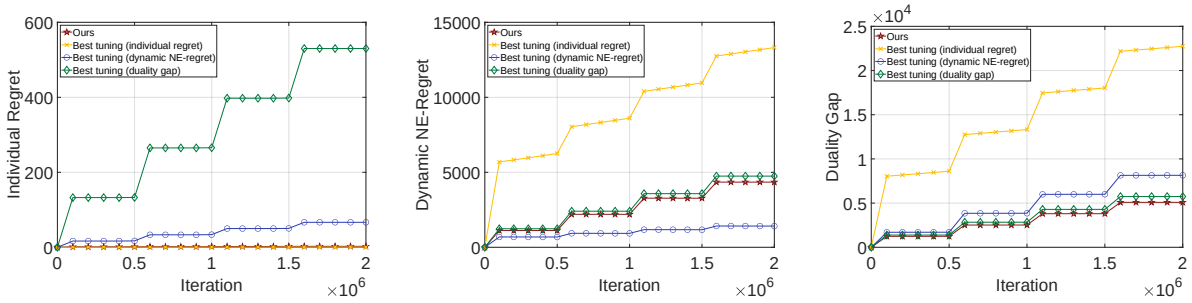

Figure 1. Empirical results of our algorithm (red line) compared with the base-learners with different step size choices. "Best tuning (measure)" denotes the curve of the base-learner with the step size choice that performs the best with respect to this "measure". The three figures show that our algorithm's performance on all three measures is comparable to (or even better than) the base-learner with the best step size tuning, while the base-learners specifically tuned for a single measure cannot perform well on all other measures simultaneously.

time horizon is set as $T = 2 \times 10^6$. Define

$$A_0 = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & -\frac{1}{2} \end{pmatrix}, A_1 = \begin{pmatrix} -1 & -1 \\ 1 & 1 \end{pmatrix}, E = \begin{pmatrix} \frac{1}{3} & -\frac{1}{2} \\ \frac{1}{3} & -\frac{1}{2} \end{pmatrix}.$$

Set $T_0 = 2\lfloor T^{1/2} \rfloor$. The scheduling of the game matrices is separated into $K = 4$ epochs and during each epoch k,

$$A_t = \begin{cases} A_0 + (-1)^t \cdot E, & t \in \left[\frac{(k-1)T}{K} + 1, \frac{(k-1)T}{K} + T_0\right], \\ \left(\frac{1}{2} - (-1)^t \cdot T^{-\frac{1}{4}}\right) A_1, & t \in \left[\frac{(k-1)T}{K} + T_0 + 1, \frac{kT}{K}\right]. \end{cases}$$

Specifically, during the first phase of each epoch, when t is even, $A_t = A_0 + E$, in which x-player's Nash equilibrium is $x_t^* = (0, 1)$ and $y_t^* = (1, 0)$; when t is odd, $A_t = A_0 - E$, where $x_t^* = (0, 1)$ but $y_t^* = (0, 1)$. Hence, the variation of Nash equilibrium of the first phase is $\Theta(1)$ per consecutive rounds. Also, the variation of the game matrix is $\Theta(1)$ per consecutive rounds. During the second phase of each epoch, the Nash equilibrium of A_t keeps the same but the variation of the game matrix is $\Theta(T^{-\frac{1}{2}})$ per consecutive rounds. Thus, over the T rounds, $P_T = \Theta(T_0) = \Theta(\sqrt{T})$, $V_T = \Theta(\sqrt{T})$. Direct calculation shows $W_T = \Theta(T^{\frac{3}{4}})$.

To show the necessity of the two-layer structure, we compare the performance of our two-layer algorithm with one single base-learner with a fixed step size chosen specifically to minimize each measure. Concretely, we choose the base-learner as optimistic Hedge with a fixed-share update, which satisfies DRVU(η) property as we prove in Appendix D.1. As mentioned in Section 4, this base-learner with a specific choice of the step size can indeed achieve a favorable bound for a specific measure. In our environment setup, to achieve the best individual regret bound, the step size needs to be chosen as $\Theta(1/\sqrt{1 + P_T + V_T}) = \Theta(T^{-\frac{1}{4}})$, while to achieve the best dynamic NE-regret bound, the step size should be chosen as $\Theta(\sqrt{P_T/(1 + P_T + V_T)}) = \Theta(1)$, which means that the base-learner cannot guarantee the desired bounds for all the three measures simultaneously.

We implement Algorithm 1 for x-player and Algorithm 2 for y-player with $L = 4$ and step size pool $\eta_i = \frac{2^{i-1}}{4\sqrt{T}}$ for both players. The number of base-learners (i.e., the size of step size pool) is $N = \lfloor \frac{1}{2} \log_2 T \rfloor + 1 = 11$. We also run our base-learner with each single η_i separately and pick the base-learner with best step size for each measure respectively to see their performance in all the three measures.

Figure 1 plots the results with respect to all the three measures (individual regret, dynamic NE-regret, and duality gap). We can observe that our algorithm's performance on all three measures is comparable to (or even better than) the base-learner with the best step size tuning, while the base-learners specifically tuned for a single measure cannot perform well on all other measures simultaneously, which supports our theoretical results and also validate the necessity of a two-layer structure of our proposed algorithm.

7. Discussions and Future Directions

Our work is among the first few to study learning in time-varying games, and we believe that our proposed performance measures and algorithm are important first steps in this direction. Our results can also be directly extended to general convex-concave games over a bounded convex domain (details omitted). We also conduct experiments with synthetic data to show the effectiveness of our algorithm.

One missing part in our work is the tightness of each bound — even though they match the best known results for a fixed game, it is unclear whether they can be further improved in the general case. We leave this as a future direction. Another interesting direction would be to consider extending the results to time-varying multi-player general-sum games.

Acknowledgements

Peng Zhao and Zhi-Hua Zhou are supported by National Science Foundation of China (61921006). Haipeng Luo and Mengxiao Zhang are supported by NSF Award IIS-1943607.

References

Badanidiyuru, A., Kleinberg, R., and Slivkins, A. Bandits with knapsacks. *Journal of ACM*, 65(3):13:1–13:55, 2018.

Balasubramanian, K. and Ghadimi, S. Zeroth-order nonconvex stochastic optimization: Handling constraints, high dimensionality, and saddle points. *Foundations of Computational Mathematics*, pp. 1–42, 2021.

Besbes, O., Gur, Y., and Zeevi, A. J. Non-stationary stochastic optimization. *Operations Research*, 63(5):1227–1244, 2015.

Cardoso, A. R., Abernethy, J. D., Wang, H., and Xu, H. Competing against Nash equilibria in adversarially changing zero-sum games. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pp. 921–930, 2019.

Cesa-bianchi, N., Gaillard, P., Lugosi, G., and Stoltz, G. Mirror descent meets fixed share (and feels no regret). *Advances in Neural Information Processing Systems*, 25: 980–988, 2012.

Cesa-Bianchi, N., Gaillard, P., Lugosi, G., and Stoltz, G. Mirror descent meets fixed share (and feels no regret). In *Advances in Neural Information Processing Systems 25 (NIPS)*, pp. 989–997, 2012.

Chen, G. and Teboulle, M. Convergence analysis of a proximal-like minimization algorithm using bregman functions. *SIAM Journal on Optimization*, 3(3):538–543, 1993.

Chen, L., Luo, H., and Wei, C. Impossible tuning made possible: A new expert algorithm and its applications. In *Proceedings of the 34th Conference on Learning Theory (COLT)*, pp. 1216–1259, 2021.

Chen, X. and Peng, B. Hedging in games: Faster convergence of external and swap regrets. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, pp. 18990–18999, 2020.

Chiang, C.-K., Yang, T., Lee, C.-J., Mahdavi, M., Lu, C.-J., Jin, R., and Zhu, S. Online optimization with gradual variations. In *Proceedings of the 25th Conference On Learning Theory (COLT)*, pp. 6.1–6.20, 2012.

Daskalakis, C. and Panageas, I. Last-iterate convergence: Zero-sum games and constrained min-max optimization. In *Proceedings of the 10th Innovations in Theoretical Computer Science (ITCS) Conference*, 2019.

Daskalakis, C., Deckelbaum, A., and Kim, A. Near-optimal no-regret algorithms for zero-sum games. *Games and Economic Behavior*, 92:327–348, 2015.

Daskalakis, C., Fishelson, M., and Golowich, N. Near-optimal no-regret learning in general games. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, pp. to appear, 2021.

Duvocelle, B., Mertikopoulos, P., Staudigl, M., and Vermeulen, D. Multi-agent online learning in time-varying games. *Mathematics of Operations Research*, to appear, 2021.

Fiez, T., Sim, R., Skoulakis, S., Piliouras, G., and Ratliff, L. J. Online learning in periodic zero-sum games. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, pp. to appear, 2021.

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

Herbster, M. and Warmuth, M. K. Tracking the best expert. *Machine learning*, 32(2):151–178, 1998.

Hsieh, Y.-G., Antonakopoulos, K., and Mertikopoulos, P. Adaptive learning in continuous games: Optimal regret bounds and convergence to Nash equilibrium. In *Proceedings of the 34th Conference on Learning Theory (COLT)*, pp. 2388–2422, 2021.

Immorlica, N., Sankararaman, K. A., Schapire, R. E., and Slivkins, A. Adversarial bandits with knapsacks. In *Proceedings of the 60th IEEE Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 202–219, 2019.

Luo, H. and Schapire, R. E. Achieving all with no parameters: AdaNormalHedge. In *Proceedings of the 28th Annual Conference Computational Learning Theory (COLT)*, pp. 1286–1304, 2015.

Mai, T., Mihail, M., Panageas, I., Ratcliff, W., Vazirani, V. V., and Yunker, P. Cycles in zero-sum differential games and biological diversity. In *Proceedings of the 2018 ACM Conference on Economics and Computation (EC)*, pp. 339–350, 2018.

Pogodin, R. and Lattimore, T. On first-order bounds, variance and gap-dependent bounds for adversarial bandits. In *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 894–904, 2019.

Rakhlin, A. and Sridharan, K. Optimization, learning, and games with predictable sequences. In *Advances in Neural Information Processing Systems 26 (NIPS)*, pp. 3066–3074, 2013.

Roy, A., Chen, Y., Balasubramanian, K., and Mohapatra, P. Online and bandit algorithms for nonstationary stochastic saddle-point optimization. *arXiv preprint arXiv:1912.01698*, 2019.

Syrgkanis, V., Agarwal, A., Luo, H., and Schapire, R. E. Fast convergence of regularized learning in games. In *Advances in Neural Information Processing Systems 28 (NIPS)*, pp. 2989–2997, 2015.

von Neumann, J. Zur theorie der gesellschaftsspiele. *Mathematische Annalen*, 100:295–320, 1928.

Wei, C.-Y., Hong, Y.-T., and Lu, C.-J. Tracking the best expert in non-stationary stochastic environments. In *Advances in Neural Information Processing Systems 29 (NIPS)*, pp. 3972–3980, 2016.

Wei, C.-Y., Lee, C.-W., Zhang, M., and Luo, H. Linear last-iterate convergence in constrained saddle-point optimization. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, 2021.

Yang, T., Zhang, L., Jin, R., and Yi, J. Tracking slowly moving clairvoyant: Optimal dynamic regret of online learning with true and noisy gradient. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pp. 449–457, 2016.

Zhang, L., Lu, S., and Zhou, Z.-H. Adaptive online learning in dynamic environments. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, pp. 1330–1340, 2018.

Zhang, Y.-J., Zhao, P., and Zhou, Z.-H. A simple online algorithm for competing with dynamic comparators. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 390–399, 2020.

Zhao, P. and Zhang, L. Improved analysis for dynamic regret of strongly convex and smooth functions. In *Proceedings of the 3rd Conference on Learning for Dynamics and Control (L4DC)*, pp. 48–59, 2021.

Zhao, P., Zhang, Y.-J., Zhang, L., and Zhou, Z.-H. Dynamic regret of convex and smooth functions. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, pp. 12510–12520, 2020.

Zhao, P., Zhang, Y.-J., Zhang, L., and Zhou, Z.-H. Adaptivity and non-stationarity: Problem-dependent dynamic regret for online convex optimization. *ArXiv preprint*, arXiv:2112.14368, 2021.

Zinkevich, M. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning (ICML)*, pp. 928–936, 2003.

A. Algorithm for y-player

For completeness, in this section, we show the algorithm run by y-player as follows. Our algorithm for y-player maintains $N+n$ base-learners: for $i \in [N]$, the i-th base-learner is any algorithm that satisfies DRVU(η_i^y) where $\eta_i^y = \frac{2^{i-1}}{L\sqrt{T}}$ and L is defined in Eq. (4); the last n base-learners are dummy learners, in which the $(j+N)$-th one always outputting the basis vector $e_j \in \Delta_n$. Let $\mathcal{S}_y \triangleq \mathcal{S}_{1,y} \cup \mathcal{S}_{2,y}$ with $\mathcal{S}_{1,y} = [N]$ and $\mathcal{S}_{2,y} = \{N+1, \ldots, N+n\}$ denote the set of indices of base-learners.

At round t, each base-learner j submits her decision $y_{t,j} \in \Delta_n$ to the meta-algorithm, who decides the final decision y_t. After receiving the feedback $A_t^\top x_t$, the meta-algorithm sends this feedback to each base-learner $j \in \mathcal{S}_{1,y}$.

The meta-algorithm of y-player performs the following update:

$$\begin{aligned} q_t &= \operatorname*{argmin}_{q \in \Delta_{|\mathcal{S}_y|}} \left\{ \varepsilon_t^y \langle q, m_t^y \rangle + \|q - \widehat{q}_t\|_2^2 \right\}, \\ \widehat{q}_{t+1} &= \operatorname*{argmin}_{q \in \Delta_{|\mathcal{S}_y|}} \left\{ \varepsilon_t^y \langle q, \ell_t^y \rangle + \|q - \widehat{q}_t\|_2^2 \right\}, \end{aligned} \tag{8}$$

where ε_t^y is the dynamic learning rate for the y-player. The loss vector $\ell_t^y \in \Delta_{|\mathcal{S}_y|}$ and loss predictor vector $m_t^y \in \Delta_{|\mathcal{S}_y|}$ is defined as follows: for any $j \in \mathcal{S}_y$,

$$\ell_{t,j}^y = -y_{t,j}^\top A_t^\top x_t + \lambda \|y_{t,j} - y_{t-1,j}\|_1^2, \tag{9}$$

$$m_{t,j}^y = -y_{t,j}^\top A_{t-1}^\top x_{t-1} + \lambda \|y_{t,j} - y_{t-1,j}\|_1^2. \tag{10}$$

The full pseudo code of the algorithm run by y-player is shown in Algorithm 2.

Algorithm 2 Algorithm for the y-player

Input: a base-algorithm $\mathcal{A}(\eta)$ satisfying DRVU(η).
Initialize: a set of base-learners $\mathcal{S}_y$ as described in Appendix A, $\widehat{p}_1 = \frac{1}{|\mathcal{S}_y|}\mathbf{1}_{|\mathcal{S}_y|}$, learning rate $\varepsilon_1^y = \frac{1}{L}$ (*c.f.* Eq. (4)).

for $t = 1, \ldots, T$ **do**
- Receive $y_{t,j} \in \Delta_n$ from each base-learner $j \in \mathcal{S}_y$.
- Compute m_t^y based on Eq. (10) and q_t based on Eq. (8).
- Play the final decision $y_t = \sum_{j \in \mathcal{S}_y} q_{t,j} y_{t,j}$.
- Suffer loss $-x_t^\top A_t y_t$ and observe the loss vector $-A_t^\top x_t$.
- Compute ℓ_t^y based on Eq. (9) and $\widehat{q}_{t+1}$ based on Eq. (8).
- Update $\varepsilon_{t+1}^y = 1/\sqrt{L^2 + \sum_{s=2}^t \|A_t^\top x_t - A_{t-1}^\top x_{t-1}\|_\infty^2}$.
- Send $-A_t^\top x_t$ as the feedback to each base-learner.

end

B. Discussions on Performance Measure

In this section, we include more discussions on the performance measures presented in Section 3.1.

B.1. Relationship between Dynamic NE-Regret and NE-Regret

Before discussing the relationship between dynamic NE-regret and NE-regret for the game setting, we first review the notion of dynamic regret and static regret for the online convex optimization (OCO) setting. Then we show that in contrast to the case in OCO that the worst-case dynamic regret is always larger than static regret, in the online game setting, dynamic NE-regret is not necessarily larger than the standard NE-regret due to the different structure of the minimax operation.

Dynamic Regret for OCO. OCO can be regarded as an iterative game between the player and the environment. At each round $t \in [T]$, the player makes the decision x_t from a convex feasible domain $\mathcal{X} \subseteq \mathbb{R}^d$ and simultaneously the environment chooses the loss function $f_t : \mathcal{X} \mapsto \mathbb{R}$, then the player suffers an instantaneous loss $f_t(x_t)$ and observe the full information about the loss function. The standard regret measure is defined as the difference between the cumulative loss of the player

and that of the best action in hindsight:

$$\text{Reg}_T = \sum_{t=1}^{T} f_t(x_t) - \min_{x\in\mathcal{X}} \sum_{t=1}^{T} f_t(x). \tag{11}$$

Note that the measure only competes with a single fixed decision over the time. A stronger measure proposed for OCO problems is called *general dynamic regret* (Zinkevich, 2003; Zhang et al., 2018; Zhao et al., 2020; 2021), defined as

$$\text{D-Reg}_T(u_1, \ldots, u_T) = \sum_{t=1}^{T} f_t(x_t) - \sum_{t=1}^{T} f_t(u_t), \tag{12}$$

which benchmarks the player's performance against an arbitrary sequence of comparators $u_1, \ldots, u_T \in \mathcal{X}$. The measure is also studied in the prediction with expert advice setting (Cesa-Bianchi et al., 2012; Luo & Schapire, 2015; Wei et al., 2016). We emphasize that one of the key tools to achieve our results for time-varying games is to derive a favorable bound for the above general dynamic regret for each player. See Lemma 15 for the details of our derived bound.

In addition, there is a variant of the above general dynamic regret called the *worst-case dynamic regret*, defined as

$$\text{D-Reg}_T^* = \text{D-Reg}_T(x_1^*, \ldots, x_T^*) = \sum_{t=1}^{T} f_t(x_t) - \sum_{t=1}^{T} f_t(x_t^*), \tag{13}$$

where $x_t^* \in \operatorname{argmin}_{x\in\mathcal{X}} f_t(x)$ is the minimizer of the online loss function f_t. The worst-case dynamic regret is extensively studied in the literature (Besbes et al., 2015; Yang et al., 2016; Zhang et al., 2020; Zhao & Zhang, 2021). It is worth noting that both standard regret in Eq. (11) and the worst-case dynamic regret in Eq. (13) are special cases of the general dynamic regret in Eq. (12). In fact, by choosing the comparators as $u_1 = \ldots = u_T \in \operatorname{argmin}_{x\in\mathcal{X}} \sum_{t=1}^{T} f_t(x)$, the general dynamic regret recovers the standard static regret; and by choosing the comparators as $u_t = x_t^* \in \operatorname{argmin}_{x\in\mathcal{X}} f_t(x)$ for $t \in [T]$, the general dynamic regret recovers the worst-case dynamic regret.

Notice that the worst-case dynamic regret in Eq. (13) is strictly larger than the static regret in Eq. (11), whereas the general dynamic regret in Eq. (12) is not necessarily larger than the static regret due to the flexibility of the comparator sequence.

Dynamic NE-Regret of Online Two-Player Zero-Sum Game. In this part, we aim to show that, different from the relationships between the (worst-case) dynamic regret and static regret in OCO setting, dynamic NE-regret is not necessarily larger than the NE-regret in the game setting. For a better readability, we here restate the definitions of NE-regret and dynamic NE-regret. Specifically, NE-regret is defined as the absolute value of the difference between the learners' cumulative payoff and the minimax value of the time-averaged payoff matrix, namely,

$$\text{NE-Reg}_T \triangleq \left| \sum_{t=1}^{T} x_t^\top A_t y_t - \min_{x\in\Delta_m} \max_{y\in\Delta_n} \sum_{t=1}^{T} x^\top A_t y \right|. \tag{14}$$

The dynamic NE-regret proposed by this paper is defined as absolute value of the difference between the cumulative payoff of the two players against the sum of the minimax game value at each round, namely,

$$\text{DynNE-Reg}_T \triangleq \left| \sum_{t=1}^{T} x_t^\top A_t y_t - \sum_{t=1}^{T} \min_{x\in\Delta_m} \max_{y\in\Delta_n} x^\top A_t y \right|. \tag{15}$$

Comparing to the original NE-regret in Eq. (14), we here move the minimax operation inside the summation of the benchmark. The operation is similar to that of the worst-case dynamic regret in Eq. (13), which moves the minimization operation inside the summation of the benchmark compared to the standard static regret in Eq. (11). However, the important point to note here is: worst-case dynamic regret is always no smaller than the static regret in online convex optimization setting, whereas the dynamic NE-regret is not necessarily larger than the NE-regret. Recall the example of two-phase online games in Section 3.1: the online matrix is set as $A_t = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$ when $t \le T/2$, and set as $A_t = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}$ when $t > T/2$. In this case, when both players are indeed using the Nash equilibrium strategy at each round, they will suffer 0 dynamic NE-regret, while still incur a linear NE-regret as $\text{NE-Reg}_T = |T/2 - 0| = T/2$.

B.2. Relationships among Individual Regret, Duality Gap, and Dynamic NE-Regret

In this subsection, we discuss the relationship among the three performance measures considered in this work: individual regret, duality gap, and dynamic NE-regret. As mentioned in Section 3.1, both the individual regret and the dynamic NE-regret are bounded by duality gap. In the following, we present a formal statement and provide the proof.

Proposition 11. *Consider any strategy sequence* $\{x_t\}_{t=1}^T$ *and* $\{y_t\}_{t=1}^T$*, where* $x_t \in \Delta_m$*,* $y_t \in \Delta_n$*,* $t \in [T]$*. We have*

$$\text{Reg}_T^x \le \text{Dual-Gap}_T,\ \text{Reg}_T^y \le \text{Dual-Gap}_T,\ \textit{and}\ \ \text{DynNE-Reg}_T \le \text{Dual-Gap}_T, \tag{16}$$

where all these measures are defined in Section 3.1.

Proof. First, we show that $\text{Reg}_T^x \le \text{Dual-Gap}_T$ as follows.

$$\begin{aligned}
\text{Reg}_T^x &= \sum_{t=1}^T x_t^\top A_t y_t - \min_{x\in\Delta_m} \sum_{t=1}^T x^\top A_t y_t \\
&\le \sum_{t=1}^T \max_{y\in\Delta_n} x_t^\top A_t y - \min_{x\in\Delta_m} \sum_{t=1}^T x^\top A_t y_t \\
&\le \sum_{t=1}^T \max_{y\in\Delta_n} x_t^\top A_t y - \sum_{t=1}^T \min_{x\in\Delta_m} x^\top A_t y_t = \text{Dual-Gap}_T.
\end{aligned}$$

The inequality of $\text{Reg}_T^y \le \text{Dual-Gap}_T$ can be obtained in the same way as shown above.

For the relationship between DynNE-Reg_T and Dual-Gap_T, actually we have

$$\begin{aligned}
\sum_{t=1}^T x_t^\top A_t y_t - \sum_{t=1}^T \min_{x\in\Delta_m}\max_{y\in\Delta_n} x^\top A_t y &\le \sum_{t=1}^T \max_{y\in\Delta_n} x_t^\top A_t y - \sum_{t=1}^T \min_{x\in\Delta_m}\max_{y\in\Delta_n} x^\top A_t y \\
&= \sum_{t=1}^T \max_{y\in\Delta_n} x_t^\top A_t y - \sum_{t=1}^T x_t^{*\top} A_t y_t^* \qquad ((x_t^*, y_t^*) \in \mathcal{X}_t^* \times \mathcal{Y}_t^*) \\
&\le \sum_{t=1}^T \max_{y\in\Delta_n} x_t^\top A_t y - \sum_{t=1}^T x_t^{*\top} A_t y_t \\
&\le \sum_{t=1}^T \max_{y\in\Delta_n} x_t^\top A_t y - \sum_{t=1}^T \min_{x\in\Delta_m} x^\top A_t y_t.
\end{aligned} \tag{17}$$

In addition,

$$\begin{aligned}
\sum_{t=1}^T \min_{x\in\Delta_m}\max_{y\in\Delta_n} x^\top A_t y - \sum_{t=1}^T x_t^\top A_t y_t &\le \sum_{t=1}^T x_t^{*\top} A_t y_t^* - \sum_{t=1}^T \min_{x\in\Delta_m} x^\top A_t y_t \\
&\le \sum_{t=1}^T x_t^\top A_t y_t^* - \sum_{t=1}^T \min_{x\in\Delta_m} x^\top A_t y_t \\
&\le \sum_{t=1}^T \max_{y\in\Delta_n} x_t^\top A_t y - \sum_{t=1}^T \min_{x\in\Delta_m} x^\top A_t y_t.
\end{aligned} \tag{18}$$

Combining Eq. (17) and Eq. (18) shows that $\text{DynNE-Reg}_T \le \text{Dual-Gap}_T$. □

C. Discussions on Non-Stationarity Measure

In the following, we present more discussions on the relationships among all three non-stationarity measures (P_T, V_T, and W_T) proposed in Section 3.2.

Comparison between Nash non-stationarity P_T and game matrix non-stationarity W_T, V_T. Here we present two specific cases to show that the non-stationarity on Nash equilibrium is not comparable to the one on game matrix.

- Case 1. Let $A_t = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$. Consider the time-varying games with $A_t = \frac{1}{T}A$ when t is odd and $A_t = \frac{T-1}{T}A$ when t is even. Notice that all A_t's have the same (unique) Nash equilibrium $x_t^* = y_t^* = (\frac{1}{2}, \frac{1}{2})$ in this case, which implies that the path-length of Nash equilibria is $P_T = 0$. By contrast, the other two non-stationarity measures related to game payoff matrices are large, concretely, $W_T = T(\frac{1}{2} - \frac{1}{T}) = \Theta(T)$ and $V_T = T \cdot \frac{(T-2)^2}{T^2} = \Theta(T)$.
- Case 2. Let $A' = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ and $E = \begin{pmatrix} \varepsilon & \varepsilon \\ -\varepsilon & -\varepsilon \end{pmatrix}$ for some $\varepsilon > 0$. Consider $A_t = A' + (-1)^t E$. In this case, we have $x_t^* = (1, 0)$ when t is odd and $x_t^* = (0, 1)$ when t is even; $y_t^* = (\frac{1}{2}, \frac{1}{2})$ for all rounds. Then the path-length of Nash equilibria is large, $P_T = \Theta(T)$. By contrast, the other two measures can be small, specifically, $W_T = \Theta(T\varepsilon) = \mathcal{O}(1)$ and $V_T = \Theta(T\varepsilon^2) = \mathcal{O}(1/T)$ when choosing $\varepsilon = \mathcal{O}(1/T)$.

Comparison between two game matrix non-stationarity measures V_T and W_T. Here we show the relationship between two non-stationarity measures regarding game matrix. First, we have $V_T \le \mathcal{O}(W_T)$ as $\sum_{t=2}^T \|A_t - A_{t-1}\|_\infty^2 \le 2\sum_{t=2}^T (\|A_t - \bar{A}\|_\infty^2 + \|A_{t-1} - \bar{A}\|_\infty^2) \le \mathcal{O}(W_T)$. Indeed, V_T can be much smaller than W_T in some cases, for instance when $A_t = \frac{t}{T}A$ with $A' = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$, $V_T = T \cdot \frac{1}{T^2} = \Theta(\frac{1}{T})$ whereas $W_T = \sum_{t=1}^T |\frac{t}{T} - \frac{T+1}{2T}| = \Theta(T)$.

D. Verifying DRVU Property

In this section, we present two instantiations of Optimistic Online Mirror Descent (Optimistic OMD) (Rakhlin & Sridharan, 2013) and prove that both of them satisfy the DRVU property in Definition 3 with $\alpha, \beta, \gamma = \widetilde{\Theta}(1)$.

Consider the general protocol of online linear optimization over the linear function sequence $\{f_1, \ldots, f_T\}$ with $f_t(x) = \langle x, g_t \rangle$ over a convex feasible set $\mathcal{X} \subseteq \mathbb{R}^d$. Optimistic OMD is a generic algorithmic framework parametrized by a sequence of optimistic vectors $M_1, \ldots, M_T \in \mathbb{R}^d$ and a regularizer ψ that is 1-strongly convex with respect to a certain norm $\|\cdot\|$. Optimistic OMD starts from an initial point $x_1 \in \mathcal{X}$ and then makes the following two-step update at each round:

$$\begin{aligned} x_t &= \operatorname*{argmin}_{x \in \mathcal{X}} \eta_t \langle M_t, x \rangle + D_\psi(x, \widehat{x}_t), \\ \widehat{x}_{t+1} &= \operatorname*{argmin}_{x \in \mathcal{X}} \eta_t \langle g_t, x \rangle + D_\psi(x, \widehat{x}_t). \end{aligned} \tag{19}$$

In above, $\eta_t > 0$ is the step size at round t, and $D_\psi(\cdot, \cdot)$ is the Bregman divergence induced by the regularizer ψ. Zhao et al. (2021) prove the following general result for the dynamic regret of Optimistic OMD, and we present the proof in Appendix D.3 for completeness.

Theorem 12 (Theorem 1 of Zhao et al. (2021))**.** *The dynamic regret of Optimistic OMD whose update rule is specified in Eq. (19) is bounded by*

$$\begin{aligned} \sum_{t=1}^T \langle x_t, g_t \rangle - \sum_{t=1}^T \langle u_t, g_t \rangle \le \sum_{t=1}^T \eta_t \|g_t - M_t\|_*^2 + \sum_{t=1}^T \frac{1}{\eta_t} \Big(D_\psi(u_t, \widehat{x}_t) - D_\psi(u_t, \widehat{x}_{t+1}) \Big) \\ - \sum_{t=1}^T \frac{1}{\eta_t} \Big(D_\psi(\widehat{x}_{t+1}, x_t) + D_\psi(x_t, \widehat{x}_t) \Big), \end{aligned}$$

which holds for any comparator sequence $u_1, \ldots, u_T \in \mathcal{X}$.

Note that the theorem is very general due to the flexibility in choosing the comparator sequence $u_1, \ldots, u_T$ and the regularizer ψ. In the following, we present two instantiations of Optimistic OMD: Optimistic Hedge with a fixed-share update (Herbster & Warmuth, 1998; Cesa-bianchi et al., 2012) and Optimistic Online Gradient Descent (Chiang et al., 2012), and then we use the above general theorem to prove that the two algorithms indeed satisfy the DRVU property defined in Definition 3.

D.1. Optimistic Hedge with a Fixed-share Update

In this subsection, we show that Optimistic Hedge with a fixed-shared update indeed satisfies the DRVU property.

Consider the following online convex optimization with linear loss functions: for $t = 1, \ldots, T$, an online learning algorithm proposes $x_t \in \Delta_m$ at the beginning of round t, and then receives a loss vector $g_t \in \mathbb{R}^m$ and suffer loss $\langle g_t, x_t \rangle$.

We first present the algorithmic procedure. Optimistic Hedge with a fixed-share update starts from an initial distribution $\widetilde{x}_1 \in \Delta_m$ and updates according to

$$\begin{aligned} x_t &= \operatorname*{argmin}_{x \in \Delta_m} \eta \langle x, h_t \rangle + D_\psi(x, \widetilde{x}_t), \\ \widehat{x}_{t+1} &= \operatorname*{argmin}_{x \in \Delta_m} \eta \langle x, g_t \rangle + D_\psi(x, \widetilde{x}_t), \\ \widetilde{x}_{t+1} &= (1-\xi)\widehat{x}_{t+1} + \frac{\xi}{m}\mathbf{1}_m, \end{aligned} \tag{20}$$

where $\eta > 0$ is a fixed step size, $\psi(x) = \sum_{i=1}^m x_i \log x_i$ is the negative-entropy regularizer, $D_\psi(\cdot, \cdot)$ is the induced Bregman divergence, and $0 \le \xi \le 1$ is the fixed-share coefficient. The first step updates by the optimistic vector $h_t \in \mathbb{R}^m$ that serving as a guess of the next-round loss, the second step updates by the received loss $g_t \in \mathbb{R}^m$, and the final step admits a fixed-share update. We have the following result on the dynamic regret of Optimistic Hedge with a fixed-share update.

Lemma 13. *Set the fixed-share coefficient as $\xi = 1/T$. The dynamic regret of Optimistic Hedge with a fixed-share update is at most*

$$\sum_{t=1}^T \langle g_t, x_t \rangle - \sum_{t=1}^T \langle g_t, u_t \rangle \le \frac{(3 + \log(mT))(1 + P_T^u)}{\eta} + \eta \sum_{t=1}^T \|g_t - h_t\|_\infty^2 - \frac{1}{4\eta}\|x_t - x_{t-1}\|_1^2, , \tag{21}$$

where $u_1, \ldots, u_T \in \Delta_m$ is any comparator sequence and $P_T = \sum_{t=2}^T \|u_t - u_{t-1}\|_1$ denotes the path-length of comparators. Therefore, when choosing the optimism as $h_t = g_{t-1}$, the algorithm satisfies the DRVU(η) property with parameters $\alpha = 3 + \log(mT)$, $\beta = 1$, and $\gamma = \frac{1}{4}$.

Proof. First we note that the chosen regularizer, $\psi(x) = \sum_{i=1}^m x_i \log x_i$, is 1-strongly convex in $\|\cdot\|_1$, because for any $x, x' \in \Delta_m$ it holds that

$$\psi(x) - \psi(x') - \langle \psi(x'), x - x' \rangle = \sum_{i=1}^m x_i \log \frac{x_i}{x_i'} \ge \frac{1}{2}\|x - x'\|_1^2,$$

where the last inequality is by Pinsker's inequality. Therefore, we can apply the general result of Theorem 12 with $f_t(x) = \langle g_t, x \rangle$ and $M_t = h_t$ and achieve the following result,

$$\begin{aligned} \sum_{t=1}^T \langle g_t, x_t \rangle - \sum_{t=1}^T \langle g_t, u_t \rangle \le \eta \sum_{t=1}^T \|g_t - h_t\|_\infty^2 + \frac{1}{\eta}\sum_{t=1}^T \left(D_\psi(u_t, \widetilde{x}_t) - D_\psi(u_t, \widehat{x}_{t+1})\right) \\ - \frac{1}{\eta}\sum_{t=1}^T \left(D_\psi(\widehat{x}_{t+1}, x_t) + D_\psi(x_t, \widetilde{x}_t)\right). \end{aligned} \tag{22}$$

We now evaluate the right-hand side. For the second term $\sum_{t=1}^T (D_\psi(u_t, \widetilde{x}_t) - D_\psi(u_t, \widehat{x}_{t+1}))$, we have

$$\begin{aligned} &D_\psi(u_t, \widetilde{x}_t) - D_\psi(u_t, \widehat{x}_{t+1}) \\ &= \sum_{i=1}^m u_{t,i} \log \frac{u_{t,i}}{\widetilde{x}_{t,i}} - \sum_{i=1}^m u_{t,i} \log \frac{u_{t,i}}{\widehat{x}_{t+1,i}} = \sum_{i=1}^m u_{t,i} \log \frac{\widehat{x}_{t+1,i}}{\widetilde{x}_{t,i}} \\ &= \left(\sum_{i=1}^m u_{t,i} \log \frac{1}{\widetilde{x}_{t,i}} - \sum_{i=1}^m u_{t-1,i} \log \frac{1}{\widehat{x}_{t,i}}\right) + \left(\sum_{i=1}^m u_{t-1,i} \log \frac{1}{\widehat{x}_{t,i}} - \sum_{i=1}^m u_{t,i} \log \frac{1}{\widehat{x}_{t+1,i}}\right). \end{aligned} \tag{23}$$

Notice that the first term above can be further upper bounded by

$$\sum_{i=1}^m u_{t,i} \log \frac{1}{\widetilde{x}_{t,i}} - \sum_{i=1}^m u_{t-1,i} \log \frac{1}{\widehat{x}_{t,i}}$$

$$= \sum_{i=1}^{m}(u_{t,i} - u_{t-1,i}) \log \frac{1}{\widetilde{x}_{t,i}} + \sum_{i=1}^{m} u_{t-1,i} \log \frac{\widehat{x}_{t,i}}{\widetilde{x}_{t,i}}$$
$$\leq \log \frac{m}{\xi} \cdot \|u_t - u_{t-1}\|_1 + \log \frac{1}{1-\xi},$$

where the inequality comes from the fixed-share update procedure, where we have $\log \frac{1}{\widetilde{x}_{t,i}} \leq \log \frac{m}{\xi}$ and $\log \frac{\widehat{x}_{t,i}}{\widetilde{x}_{t,i}} \leq \log \frac{1}{1-\xi}$ for any $i \in [m]$ and $t \in [T]$. Then, taking summation of Eq. (23) over $t = 2$ to T and combining the fact that $(D_\psi(u_1, \widetilde{x}_1) - D_\psi(u_1, \widehat{x}_2)) = \sum_{i=1}^{m} u_{1,i} \log \frac{\widehat{x}_{2,i}}{\widetilde{x}_{1,i}}$ and $\widetilde{x}_{1,i} \geq \frac{\xi}{m}$ for any $i \in [m]$, we get

$$\begin{aligned}
&\frac{1}{\eta}\sum_{t=1}^{T}(D_\psi(u_t, \widetilde{x}_t) - D_\psi(u_t, \widehat{x}_{t+1})) \\
&\leq \frac{1}{\eta}\left(\log \frac{m}{\xi}\sum_{t=2}^{T}\|u_t - u_{t-1}\|_1 + (T-1)\log\frac{1}{1-\xi} + \sum_{i=1}^{m} u_{1,i}\log\frac{1}{\widehat{x}_{2,i}} + \sum_{i=1}^{m} u_{1,i}\log\frac{\widehat{x}_{2,i}}{\widetilde{x}_{1,i}}\right) \\
&\leq \frac{1}{\eta}\left(\log \frac{m}{\xi}\left(1 + \sum_{t=2}^{T}\|u_t - u_{t-1}\|_1\right) + (T-1)\log\frac{1}{1-\xi}\right). \qquad (24)
\end{aligned}$$

Next, we proceed to analyze the negative term , i.e., the third term of the right-hand side in Eq. (22). Indeed,

$$\begin{aligned}
&\sum_{t=2}^{T}(D_\psi(\widehat{x}_t, x_{t-1}) + D_\psi(x_t, \widetilde{x}_t)) \\
&\geq \frac{1}{2}\sum_{t=2}^{T}\left(\|\widehat{x}_t - x_{t-1}\|_1^2 + \|x_t - \widetilde{x}_t\|_1^2\right) && \text{(Pinsker's inequality)} \\
&\geq \frac{1}{4}\sum_{t=2}^{T}\left(\|x_t - x_{t-1} + \widehat{x}_t - \widetilde{x}_t\|_1^2\right) && (\|x\|_1^2 + \|y\|_1^2 \geq \tfrac{1}{2}\|x+y\|_1^2) \\
&= \frac{1}{4}\sum_{t=2}^{T}\|x_t - x_{t-1} + \xi\big(\widehat{x}_t - \frac{1}{m}\mathbf{1}_m\big)\|_1^2 && \text{(due to the fixed-share update)} \\
&\geq \frac{1}{4}\sum_{t=2}^{T}\|x_t - x_{t-1}\|_1^2 - \sum_{t=2}^{T}\frac{\xi}{2}\|x_t - x_{t-1}\|_1\left\|\widehat{x}_t - \frac{1}{m}\mathbf{1}_m\right\|_1 && (\|a-b\|_1^2 \geq \|a\|_1^2 - 2\|a\|_1 \cdot \|b\|_1) \\
&\geq \frac{1}{4}\sum_{t=2}^{T}\|x_t - x_{t-1}\|_1^2 - 2\xi(T-1). && (25)
\end{aligned}$$

Substituting Eq. (24) and Eq. (25) into the general dynamic regret upper bound in Eq. (22), we achieve

$$\begin{aligned}
\sum_{t=1}^{T}\langle g_t, x_t - u_t\rangle &\leq \eta\sum_{t=1}^{T}\|g_t - h_t\|_\infty^2 + \frac{1}{\eta}\log\frac{m}{\xi}\cdot(1+P_T^u) + \frac{1}{\eta}(T-1)\log\frac{1}{1-\xi} + \frac{2\xi}{\eta}(T-1) - \frac{1}{4\eta}\sum_{t=2}^{T}\|x_t - x_{t-1}\|_1^2 \\
&\leq \eta\sum_{t=1}^{T}\|g_t - h_t\|_\infty^2 + \frac{1}{\eta}\left(3 + \log(mT)(1+P_T^u)\right) - \frac{1}{4\eta}\sum_{t=2}^{T}\|x_t - x_{t-1}\|_1^2,
\end{aligned}$$

where the last step holds because we set $\xi = \frac{1}{T}$ and

$$\frac{1}{\eta}(T-1)\log\frac{1}{1-\xi} + \frac{2\xi}{\eta}(T-1) = \frac{1}{\eta}(T-1)\log\left(1 + \frac{\xi}{1-\xi}\right) + \frac{2\xi}{\eta}(T-1) \leq \frac{1}{\eta}(T-1)\frac{\xi}{1-\xi} + \frac{2\xi}{\eta}(T-1) \leq \frac{3}{\eta}.$$

When choosing the optimism as $h_t = g_{t-1}$, we then have

$$\sum_{t=1}^{T}\langle g_t, x_t - u_t\rangle \leq \eta\sum_{t=1}^{T}\|g_t - g_{t-1}\|_\infty^2 + \frac{3+\log(mT)}{\eta}(1+P_T^u) - \frac{1}{4\eta}\sum_{t=2}^{T}\|x_t - x_{t-1}\|_1^2,$$

which verifies the DRVU property of Definition 3, with $\alpha = 3 + \log(mT)$, $\beta = 1$, and $\gamma = \frac{1}{4}$. This ends the proof. □

D.2. Optimistic Online Gradient Descent

In this subsection, we show that Optimistic Online Gradient Descent (Optimistic OGD) with a fixed-shared update indeed satisfies the DRVU property.

Consider the following online convex optimization with linear loss functions: for $t = 1, \dots, T$, an online learning algorithm proposes $x_t \in \Delta_m$ at the beginning of round t, and then receives a loss vector $g_t \in \mathbb{R}^m$ and suffer loss $\langle g_t, x_t \rangle$.

We first present the algorithmic procedure. Optimistic OGD starts from an initial distribution $\widehat{x}_1 \in \Delta_m$ and updates according to

$$\begin{aligned} x_t &= \operatorname*{argmin}_{x \in \Delta_m} \eta \langle x, h_t \rangle + D_\psi(x, \widehat{x}_t), \\ \widehat{x}_{t+1} &= \operatorname*{argmin}_{x \in \Delta_m} \eta \langle x, g_t \rangle + D_\psi(x, \widehat{x}_t), \end{aligned} \tag{26}$$

where $\eta > 0$ is a fixed step size and $\psi(x) = \frac{1}{2}\|x\|_2^2$ is the Euclidean regularizer and $D_\psi(\cdot, \cdot)$ is the induced Bregman divergence. Compared to Eq. (20). The first step updates by the optimistic vector $h_t \in \mathbb{R}^m$ that serving as a guess of the next-round loss, the second step updates by the received loss $g_t \in \mathbb{R}^m$. We note that Optimistic OGD does not require a fixed-share mixing operation to achieve dynamic regret.

Then, we have the following result on the dynamic regret of Optimistic OGD.

Lemma 14. *The dynamic regret of Optimistic OGD is at most*

$$\sum_{t=1}^{T} \langle g_t, x_t \rangle - \sum_{t=1}^{T} \langle g_t, u_t \rangle \leq \frac{(m+2)P_T^u}{\eta} + \frac{\eta m}{2} \sum_{t=2}^{T} \|g_t - h_t\|_\infty^2 - \frac{1}{4\eta m} \sum_{t=2}^{T} \|x_t - x_{t-1}\|_1^2 + \mathcal{O}(1), \tag{27}$$

where $u_1, \dots, u_T \in \Delta_m$ is any comparator sequence and $P_T = \sum_{t=2}^{T} \|u_t - u_{t-1}\|_1$ denotes the path-length of comparators. Therefore, when choosing the optimism as $h_t = g_{t-1}$, the algorithm satisfies the DRVU(η) property with parameters $\alpha = m + 2$, $\beta = \frac{m}{2}$, and $\gamma = \frac{1}{4m}$.

Proof. From the general result of Theorem 12, we have the following dynamic regret bound for Optimistic OGD:

$$\sum_{t=1}^{T} \langle g_t, x_t - u_t \rangle \leq \frac{\eta}{2} \sum_{t=1}^{T} \|g_t - h_t\|_2^2 + \frac{1}{2\eta} \sum_{t=1}^{T} \left(\|\widehat{x}_t - u_t\|_2^2 - \|\widehat{x}_{t+1} - u_t\|_2^2 \right) - \frac{1}{2\eta} \sum_{t=1}^{T} \left(\|\widehat{x}_{t+1} - x_t\|_2^2 + \|x_t - \widehat{x}_t\|_2^2 \right).$$

Besides, we have

$$\begin{aligned} &\sum_{t=1}^{T} \left(\|\widehat{x}_t - u_t\|_2^2 - \|\widehat{x}_{t+1} - u_t\|_2^2 \right) \\ &\leq \sum_{t=2}^{T} \|\widehat{x}_t - u_t\|_2^2 - \sum_{t=2}^{T} \|\widehat{x}_t - u_{t-1}\|_2^2 + D^2 \\ &= \sum_{t=2}^{T} \left(\|u_t - u_{t-1}\|_2 \cdot \|\widehat{x}_t + \widehat{x}_t - u_t - u_{t-1}\|_2 \right) + D^2 \\ &\leq \sum_{t=2}^{T} \left(\|u_t - u_{t-1}\|_2 \cdot \|\widehat{x}_t + \widehat{x}_t - u_t - u_{t-1}\|_1 \right) + D^2 \\ &\leq D^2 + 4 \sum_{t=2}^{T} \|u_t - u_{t-1}\|_2 \\ &\leq D^2 + 4 \sum_{t=2}^{T} \|u_t - u_{t-1}\|_1, \end{aligned}$$

where we introduce the notation $D = \sup_{x,y\in\Delta_m}\|x-y\|_2$, and it can be verified that $D \le \sqrt{2m}$. Further we have

$$\sum_{t=1}^{T}\left(\|\widehat{x}_{t+1}-x_t\|_2^2+\|x_t-\widehat{x}_t\|_2^2\right) \ge \sum_{t=2}^{T}\left(\|\widehat{x}_t-x_{t-1}\|_2^2+\|x_t-\widehat{x}_t\|_2^2\right) \ge \frac{1}{2}\sum_{t=2}^{T}\|x_t-x_{t-1}\|_2^2 \ge \frac{1}{2m}\sum_{t=2}^{T}\|x_t-x_{t-1}\|_1^2.$$

Combining the above three inequalities, we achieve that

$$\begin{aligned}\sum_{t=1}^{T}\langle g_t, x_t-u_t\rangle &\le \frac{\eta}{2}\sum_{t=1}^{T}\|g_t-h_t\|_2^2+\frac{1}{\eta}\left(m+2P_T^u\right)-\frac{1}{4\eta m}\sum_{t=2}^{T}\|x_t-x_{t-1}\|_1^2\\ &\le \frac{\eta m}{2}\sum_{t=1}^{T}\|g_t-h_t\|_\infty^2+\frac{m+2}{\eta}\left(1+P_T^u\right)-\frac{1}{4\eta m}\sum_{t=2}^{T}\|x_t-x_{t-1}\|_1^2.\end{aligned}$$

Therefore, choosing the optimism as $h_t = g_{t-1}$, we then verify the DRVU property of Definition 3 for Optimistic OGD, with $\alpha = m+2$, $\beta = \frac{m}{2}$, and $\gamma = \frac{1}{4m}$. This ends the proof. □

D.3. Proof of Theorem 12

Proof. We decompose the instantaneous dynamic regret into three terms and bound each one respectively. Specifically,

$$f_t(x_t)-f_t(u_t) \le \langle\nabla f_t(x_t), x_t-u_t\rangle = \langle\nabla f_t(x_t)-M_t, x_t-\widehat{x}_{t+1}\rangle+\langle M_t, x_t-\widehat{x}_{t+1}\rangle+\langle\nabla f_t(x_t),\widehat{x}_{t+1}-u_t\rangle.$$

The first term can be controlled by Lemma 22, which guarantees that the OMD update satisfies $\|x_t-\widehat{x}_{t+1}\| \le \eta_t\|\nabla f_t(x_t)-M_t\|_*$ and thus,

$$\langle\nabla f_t(x_t)-M_t, x_t-\widehat{x}_{t+1}\rangle \le \|\nabla f_t(x_t)-M_t\|_*\cdot\|x_t-\widehat{x}_{t+1}\| \le \eta_t\|\nabla f_t(x_t)-M_t\|_*^2.$$

We now analyze the remaining two terms on the right-hand side. By the Bregman proximal inequality in Lemma 21 and the OMD update step $x_t = \operatorname{argmin}_{x\in\mathcal{X}}\eta_t\langle M_t, x\rangle + D_\psi(x,\widehat{x}_t)$, we have

$$\langle M_t, x_t-\widehat{x}_{t+1}\rangle \le \frac{1}{\eta_t}\Big(D_\psi(\widehat{x}_{t+1},\widehat{x}_t)-D_\psi(\widehat{x}_{t+1},x_t)-D_\psi(x_t,\widehat{x}_t)\Big).$$

Similarly, the OMD update step $\widehat{x}_{t+1} = \operatorname{argmin}_{x\in\mathcal{X}}\eta_t\langle\nabla f_t(x_t), x\rangle + D_\psi(x,\widehat{x}_t)$ implies

$$\langle\nabla f_t(x_t),\widehat{x}_{t+1}-u_t\rangle \le \frac{1}{\eta_t}\Big(D_\psi(u_t,\widehat{x}_t)-D_\psi(u_t,\widehat{x}_{t+1})-D_\psi(\widehat{x}_{t+1},\widehat{x}_t)\Big).$$

Combining the above three inequalities yields an upper bound for the instantaneous dynamic regret:

$$f_t(x_t)-f_t(u_t) \le \eta_t\|\nabla f_t(x_t)-M_t\|_*^2+\frac{1}{\eta_t}\Big(D_\psi(u_t,\widehat{x}_t)-D_\psi(u_t,\widehat{x}_{t+1})-D_\psi(\widehat{x}_{t+1},x_t)-D_\psi(x_t,\widehat{x}_t)\Big). \tag{28}$$

Taking the summation over all iterations completes the proof. □

E. Proofs for Section 5

In this section, we provide the proofs for the main results presented in Section 5, including individual regret of Theorem 4, dynamic NE-regret of Theorem 6, and duality gap of Theorem 8.

E.1. Proof of Theorem 4 (Individual Regret)

Proof. In the following, we focus on the individual regret of x-player, and the result for y-player can be proven in a similar way. The proof is split into three parts.

(1) First, we prove the $\widetilde{\mathcal{O}}(\sqrt{T})$ individual regret bound for x-player no matter whether the y-player follows the strategy suggested by Algorithm 2.

(2) Second, we prove the $\widetilde{\mathcal{O}}(\sqrt{1+V_T+P_T})$ bound, which depends on V_T (the variation of the payoff matrices) and P_T (the path-length of the Nash equilibrium sequence).
(3) Finally, we prove the $\widetilde{\mathcal{O}}(\sqrt{1+V_T+W_T})$ bound, which depends on V_T and W_T, the variation and variance of the payoff matrices.

Our analysis is mainly based on the general dynamic regret bound proven in Lemma 15. Specifically, using Eq. (35), setting a fixed comparator, i.e., $u_1 = \ldots = u_T \in \operatorname{argmin}_{x\in\Delta_m} \sum_{t=1}^T x^\top A_t y_t$ (then the path-length $P_T^u = 0$), and also dropping the last three negative terms in the regret upper bound, for any $i \in \mathcal{S}_{1,x}$ we have

$$\sum_{t=1}^T x_t^\top A_t y_t - \min_x \sum_{t=1}^T x^\top A_t y_t \le \widetilde{\mathcal{O}}\left(\frac{1}{\eta_i^x} + \eta_i^x \sum_{t=2}^T \|A_t - A_{t-1}\|_\infty^2 + \eta_i^x \sum_{t=2}^T \|y_t - y_{t-1}\|_1^2\right).$$

In the following, we further bound the right-hand side in three different ways to achieve different individual regret bounds.

The $\widetilde{\mathcal{O}}(\sqrt{T})$ robustness bound. First of all, we prove that for x-player, her individual regret against y-player's actions is at most $\widetilde{\mathcal{O}}(\sqrt{T})$, which holds even when y-player does not follow strategies suggested by Algorithm 2. Note that $\sum_{t=2}^T \|A_t - A_{t-1}\|_\infty^2 + \sum_{t=2}^T \|y_t - y_{t-1}\|_1^2 \le \mathcal{O}(T)$. Therefore, we have

$$\sum_{t=1}^T x_t^\top A_t y_t - \min_x \sum_{t=1}^T x^\top A_t y_t \le \widetilde{\mathcal{O}}\left(\frac{1}{\eta_i^x} + \eta_i^x T\right) \le \widetilde{\mathcal{O}}(\sqrt{T}),$$

where the last inequality is achieved by taking $i = i^\dagger$ such that $\eta_{i^\dagger} = \Theta(1/\sqrt{T})$. Note that the choice is viable due to the configuration of the step size pool. Similarly, we can also attain an $\widetilde{\mathcal{O}}(\sqrt{T})$ robustness bound for y-player.

Next, we demonstrate two adaptive bounds of the individual regret when both players follow our prescribed strategy (namely, x-player is using Algorithm 1 and y-player is using Algorithm 2). They are both in the worst case $\widetilde{\mathcal{O}}(\sqrt{T})$ but can be much smaller if the sequence of online payoff matrices is less non-stationary.

The $\widetilde{\mathcal{O}}(\sqrt{1+V_T+P_T})$ bound. We first consider the individual regret bound that scales with the variation of Nash equilibria denoted by $P_T \triangleq \min_{\forall t,(x_t^*,y_t^*)\in\mathcal{X}_t^*\times\mathcal{Y}_t^*} \sum_{t=2}^T \left(\|x_t^* - x_{t-1}^*\|_1 + \|y_t^* - y_{t-1}^*\|_1\right)$ and $V_T = \sum_{t=2}^T \|A_t - A_{t-1}\|_\infty^2$. According to Lemma 16, which proves the stability of the dynamics with respect to P_T and V_T when both players are following the suggested strategy, we have $\sum_{t=2}^T \|y_t - y_{t-1}\|_1^2 \le \widetilde{\mathcal{O}}(\sqrt{(1+V_T)(1+P_T)} + P_T)$. Therefore, we achieve

$$\begin{aligned}
\sum_{t=1}^T x_t^\top A_t y_t - \min_{x\in\Delta_m} x^\top A_t y_t &\le \widetilde{\mathcal{O}}\left(\frac{1}{\eta_i^x} + \eta_i^x V_T + \eta_i^x\left(\sqrt{(1+V_T)(1+P_T)} + P_T\right)\right) \\
&\le \widetilde{\mathcal{O}}\left(\frac{1}{\eta_i^x} + \eta_i^x(1+V_T+P_T)\right) && \text{(by AM-GM inequality)} \\
&\le \widetilde{\mathcal{O}}\left(\sqrt{1+V_T+P_T}\right),
\end{aligned}$$

where in the last inequality, we choose $i = i^\ddagger$ such that $\eta_{i^\ddagger}^x \in [\frac{1}{2}\eta_*^x, 2\eta_*^x]$ where $\eta_*^x = \min\{\frac{1}{\sqrt{1+V_T+P_T}}, \frac{1}{L}\}$. The choice of $\eta_{i^\ddagger}^x$ is also viable due to the configuration of the step size pool.

The $\widetilde{\mathcal{O}}(\sqrt{1+V_T+W_T})$ bound. We then consider the individual regret bound that scales with V_T and the variance of the game matrices denoted by $W_T = \sum_{t=1}^T \|A_t - \bar{A}\|_\infty$ with $\bar{A} = \frac{1}{T}\sum_{t=1}^T A_t$ being the averaged game matrix. Then according to Lemma 18, which proves the stability of the dynamics with respect to W_T when both players are following the suggested strategy, we have $\sum_{t=2}^T \|y_t - y_{t-1}\|_1^2 \le \widetilde{\mathcal{O}}(1+W_T)$. Therefore, we achieve

$$\sum_{t=1}^T x_t^\top A_t y_t - \min_{x\in\Delta_m} x^\top A_t y_t \le \widetilde{\mathcal{O}}\left(\frac{1}{\eta_i^x} + \eta_i^x(V_T+1+W_T)\right) \le \widetilde{\mathcal{O}}\left(\sqrt{1+V_T+W_T}\right),$$

where in the last inequality, we choose $i = i^*$ such that $\eta_{i^*}^x \in [\frac{1}{2}\eta_*^x, 2\eta_*^x]$ where $\eta_*^x = \min\{\frac{1}{\sqrt{1+V_T+W_T}}, \frac{1}{L}\}$. The choice of $\eta_{i^*}^x$ is also viable due to the configuration of the step size pool.

Combining above three upper bounds finishes the proof of Theorem 4. □

E.2. Proof of Theorem 6 (Dynamic NE-Regret)

Proof. The proof for the dynamic NE-regret measure consists of two parts. We first prove the $\widetilde{\mathcal{O}}(\sqrt{(1+V_T)(1+P_T)}+P_T)$ bound and then prove the $\widetilde{\mathcal{O}}(1+W_T)$ bound.

The $\widetilde{\mathcal{O}}(\sqrt{(1+V_T)(1+P_T)}+P_T)$ bound. Let (x_t^*, y_t^*) be the Nash equilibrium of the online game matrix A_t. We consider the upper bound in terms of the non-stationarity measure $P_T \triangleq \sum_{t=2}^T \left(\|x_t^* - x_{t-1}^*\|_1 + \|y_t^* - y_{t-1}^*\|_1\right)$.[6] As (x_t^*, y_t^*) is the Nash equilibrium of A_t, the inequality $x_t^{*\top} A_t y \le x_t^{*\top} A_t y_t^* \le x^\top A_t y_t^*$ holds for any $x \in \Delta_m$ and $y \in \Delta_n$. We notice that

$$\min_x \max_y x^\top A_t y = x_t^{*\top} A_t y_t^* \ge x_t^{*\top} A_t y_t \ge \min_x x^\top A_t y_t,$$
$$\min_x \max_y x^\top A_t y = x_t^{*\top} A_t y_t^* \le x_t^\top A_t y_t^* \le \max_y x_t^\top A_t y.$$

Therefore, we have

$$\sum_{t=1}^T x_t^\top A_t y_t - \sum_{t=1}^T \min_x \max_y x^\top A_t y \le \sum_{t=1}^T x_t^\top A_t y_t - \sum_{t=1}^T x_t^{*\top} A_t y_t,$$
$$-\sum_{t=1}^T x_t^\top A_t y_t + \sum_{t=1}^T \min_x \max_y x^\top A_t y \le -\sum_{t=1}^T x_t^\top A_t y_t + \sum_{t=1}^T x_t^\top A_t y_t^*,$$

which means that the dynamic NE-regret is upper bounded by the maximum of the following two dynamic regret bounds:

$$\left|\sum_{t=1}^T x_t^\top A_t y_t - \sum_{t=1}^T \min_x \max_y x^\top A_t y\right| \le \max\left\{\sum_{t=1}^T x_t^\top A_t y_t - \sum_{t=1}^T x_t^{*\top} A_t y_t, -\sum_{t=1}^T x_t^\top A_t y_t + \sum_{t=1}^T x_t^\top A_t y_t^*\right\}.$$

Moreover, according to the general dynamic regret analysis in Lemma 15 with the choice of $\{(u_t, v_t)\}_{t=1}^T = \{(x_t^*, y_t^*)\}_{t=1}^T$ and dropping the three negative terms, we have

$$\sum_{t=1}^T x_t^\top A_t y_t - \sum_{t=1}^T x_t^{*\top} A_t y_t \le \widetilde{\mathcal{O}}\left(\frac{1+P_T^x}{\eta_i^x} + \eta_i^x V_T + \eta_i^x \sum_{t=2}^T \|y_t - y_{t-1}\|_1^2\right).$$

where $P_T^x = \sum_{t=2}^T \|x_t^* - x_{t-1}^*\|_1$ denotes the path-length of the Nash equilibria of the x-player. According to Lemma 16, we have $\sum_{t=2}^T \|y_t - y_{t-1}\|_1^2 \le \widetilde{\mathcal{O}}(\sqrt{(1+V_T)(1+P_T)} + P_T)$, so using AM-GM inequality achieves

$$\begin{aligned}
\sum_{t=1}^T x_t^\top A_t y_t - \sum_{t=1}^T x_t^{*\top} A_t y_t &\le \widetilde{\mathcal{O}}\left(\frac{1+P_T^x}{\eta_i^x} + \eta_i^x V_T + \eta_i^x(\sqrt{(1+V_T)(1+P_T)} + P_T)\right) \\
&\le \widetilde{\mathcal{O}}\left(\frac{1+P_T}{\eta_i^x} + \eta_i^x(1+V_T+P_T)\right) \\
&\le \widetilde{\mathcal{O}}\left(\sqrt{(1+P_T)(1+V_T+P_T)} + P_T\right) \\
&\le \widetilde{\mathcal{O}}\left(\sqrt{(1+P_T)(1+V_T)} + P_T\right).
\end{aligned}$$

where the last inequality is by choosing $i = i^*$ such that $\eta_{i^*}^x \in [\frac{1}{2}\eta_*^x, 2\eta_*^x]$ where $\eta_*^x = \min\left\{\sqrt{\frac{1+P_T}{1+V_T+P_T}}, \frac{1}{L}\right\}$.

[6] Strictly speaking, the path-length non-stationarity measure is $P_T \triangleq \min_{\forall t, (x_t^*, y_t^*) \in \mathcal{X}_t^* \times \mathcal{Y}_t^*} \sum_{t=2}^T \left(\|x_t^* - x_{t-1}^*\|_1 + \|y_t^* - y_{t-1}^*\|_1\right)$ as the Nash equilibrium of each round may not unique. Fortunately, our analysis holds for any Nash equilibrium, so we can in particular take the sequence of Nash equilibria making $\sum_{t=2}^T \left(\|x_t^* - x_{t-1}^*\|_1 + \|y_t^* - y_{t-1}^*\|_1\right)$ smallest possible. The quantity P_T is only used in the analysis, and our algorithm does not require any prior knowledge about it.

The $\widetilde{\mathcal{O}}(1+W_T)$ bound. According to Lemma 17, we have the NE-regret is bounded by the maximum of two individual regret upper bounds plus the variance of the game matrices.

$$\left|\sum_{t=1}^{T} x_t^\top A_t y_t - \sum_{t=1}^{T} \min_x \max_y x^\top A_t y\right| \le \max\left\{\sum_{t=1}^{T} x_t^\top A_t y_t - \sum_{t=1}^{T} x^{*\top} A_t y_t, \sum_{t=1}^{T} x_t^\top A_t y^* - \sum_{t=1}^{T} x_t^\top A_t y_t\right\} + 2W_T.$$

Using the $\widetilde{\mathcal{O}}(\sqrt{1+V_T+W_T})$ individual regret bound proven in Theorem 4 and the fact that $V_T \le \mathcal{O}(W_T)$, we have

$$\left|\sum_{t=1}^{T} x_t^\top A_t y_t - \sum_{t=1}^{T} \min_x \max_y x^\top A_t y\right| \le \widetilde{\mathcal{O}}\left(\sqrt{1+V_T+W_T} + W_T\right) \le \widetilde{\mathcal{O}}\left(1+W_T\right).$$

To summarize, combining the above two upper bounds for dynamic NE-regret, we finally achieve the following guarantee:

$$\left|\sum_{t=1}^{T} x_t^\top A_t y_t - \sum_{t=1}^{T} \min_x \max_y x^\top A_t y\right| \le \widetilde{\mathcal{O}}\left(\min\left\{\sqrt{(1+V_T)(1+P_T)} + P_T, 1+W_T\right\}\right),$$

which completes the proof of Theorem 6. □

E.3. Proof of Theorem 8 (Duality Gap)

Proof. In Theorem 8, there are indeed two different upper bounds for duality gap of our approach, as restated below.

$$\sum_{t=1}^{T} x_t^\top A_t \bar{y}_t^* - \sum_{t=1}^{T} \bar{x}_t^{*\top} A_t y_t \le \widetilde{\mathcal{O}}\Big(\min\{T^{\frac{3}{4}}\left(1+Q_T\right)^{\frac{1}{4}}, T^{\frac{1}{2}}(1+Q_T^{\frac{3}{2}} + P_T Q_T)^{\frac{1}{2}}\}\Big),$$

where $Q_T \triangleq V_T + \min\{P_T, W_T\}$ is introduced to simplify the notation. Now we will prove the two bounds respectively.

The $\widetilde{\mathcal{O}}(T^{\frac{3}{4}}\left(1+V_T+\min\{P_T,W_T\}\right)^{\frac{1}{4}})$ bound. For convenience of the following proof, we introduce the notation of $f_t(x) \triangleq x^\top A_t y_t$, and then the best response is essentially the minimizer of the function, namely, $\bar{x}_t^* = \operatorname{argmin}_{x\in\Delta_m} f_t(x)$. We now investigate the following worst-case dynamic regret of the x-player,

$$\sum_{t=1}^{T} x_t^\top A_t y_t - \sum_{t=1}^{T} \bar{x}_t^{*\top} A_t y_t = \sum_{t=1}^{T} f_t(x_t) - \sum_{t=1}^{T} f_t(\bar{x}_t^*),$$

which benchmarks the cumulative loss of x-player's actions with the best response at each round. We decompose the quantity into the following two terms:

$$\sum_{t=1}^{T} f_t(x_t) - \sum_{t=1}^{T} f_t(\bar{x}_t^*) = \underbrace{\sum_{t=1}^{T} f_t(x_t) - \sum_{t=1}^{T} f_t(u_t)}_{\texttt{term (i)}} + \underbrace{\sum_{t=1}^{T} f_t(u_t) - \sum_{t=1}^{T} f_t(\bar{x}_t^*)}_{\texttt{term (ii)}},$$

where we insert a term $\sum_{t=1}^{T} f_t(u_t)$ as an anchor quantity. Notably, this comparator sequence $\{u_t\}_{t=1}^T$ can be arbitrarily set without affecting the above equation. In particular, we choose it as a piecewise-stationary comparator sequence such that $\mathcal{I}_1, \mathcal{I}_2, \ldots, \mathcal{I}_K$ is an even partition of the total horizon $[T]$ with $|\mathcal{I}_k| = \Delta$ for $k = 1, \ldots, K$ (for simplicity, suppose the time horizon T is divisible by epoch length Δ), and for any $t \in \mathcal{I}_k$, $u_t \triangleq \operatorname{argmin}_{x\in\Delta_m} \sum_{t\in\mathcal{I}_k} f_t(x)$. Then, following the general dynamic regret bound proven in Lemma 15, for this particular comparator sequence and for any $i \in \mathcal{S}_{1,x}$, we have the following upper bound for term (i):

$$\begin{aligned}
&\sum_{t=1}^{T} f_t(x_t) - \sum_{t=1}^{T} f_t(u_t) \\
&\le \mathcal{O}\left(\frac{\alpha(1+P_T^u)}{\eta_i^x}\right) + \eta_i^x c\beta \sum_{t=2}^{T} \|A_t - A_{t-1}\|_\infty^2 + \eta_i^x c\beta \sum_{t=2}^{T} \|y_t - y_{t-1}\|_1^2 + \left(\lambda - \frac{\gamma}{\eta_i^x}\right) \sum_{t=2}^{T} \|x_{t,i} - x_{t-1,i}\|_1^2
\end{aligned}$$

$$
\begin{aligned}
&- L\sum_{t=1}^{T}(\|p_t-\widehat{p}_{t+1}\|_2^2+\|p_t-\widehat{p}_t\|_2^2)-\lambda\sum_{t=1}^{T}\sum_{i\in\mathcal{S}_x}p_{t,i}\|x_{t,i}-x_{t-1,i}\|_1^2+\widetilde{\mathcal{O}}(1)\\
&\le\widetilde{\mathcal{O}}\left(\frac{1+P_T^u}{\eta_i^x}+\eta_i^x\sum_{t=2}^{T}\|A_t-A_{t-1}\|_\infty^2\right)+\eta_i^x c\beta\sum_{t=2}^{T}\|y_t-y_{t-1}\|_1^2\\
&\quad-\frac{L}{2}\sum_{t=2}^{T}\|p_t-p_{t-1}\|_1^2-\lambda\sum_{t=2}^{T}\sum_{i=1}^{N}p_{t,i}\|x_{t,i}-x_{t-1,i}\|_1^2 \qquad (\lambda-\tfrac{\gamma}{\eta_i^x}=\tfrac{\gamma L}{2}-\tfrac{\gamma}{\eta_i^x}\le 0)\\
&\le\widetilde{\mathcal{O}}\left(\frac{1+P_T^u}{\eta_i^x}+\eta_i^x\sum_{t=2}^{T}\|A_t-A_{t-1}\|_\infty^2\right)+2\eta_i^x c\beta\sum_{t=2}^{T}\|q_t-q_{t-1}\|_1^2+2\eta_i^x c\beta\sum_{t=2}^{T}\sum_{j\in\mathcal{S}_y}q_{t,j}\|y_{t,j}-y_{t-1,j}\|_1^2\\
&\quad-\frac{L}{2}\sum_{t=2}^{T}\|p_t-p_{t-1}\|_1^2-\lambda\sum_{t=2}^{T}\sum_{i=1}^{N}p_{t,i}\|x_{t,i}-x_{t-1,i}\|_1^2 \qquad \text{(by Eq. (43))}\\
&\le\widetilde{\mathcal{O}}\left(\frac{T/\Delta}{\eta_i^x}+\eta_i^x\sum_{t=2}^{T}\|A_t-A_{t-1}\|_\infty^2\right)+2\eta_i^x c\beta\sum_{t=2}^{T}\|q_t-q_{t-1}\|_1^2+2\eta_i^x c\beta\sum_{t=2}^{T}\sum_{j\in\mathcal{S}_y}q_{t,j}\|y_{t,j}-y_{t-1,j}\|_1^2\\
&\quad-\frac{L}{2}\sum_{t=2}^{T}\|p_t-p_{t-1}\|_1^2-\lambda\sum_{t=2}^{T}\sum_{i=1}^{N}p_{t,i}\|x_{t,i}-x_{t-1,i}\|_1^2,
\end{aligned}
$$

where the last step holds because $P_T^u=\sum_{t=2}^{T}\|u_t-u_{t-1}\|_1=\mathcal{O}(K)=\mathcal{O}(T/\Delta)$ by the specific construction of the comparator sequence.

Moreover, in Lemma 20 we present a general result to relate the function-value difference between the sequence of piecewise minimizers and the sequence of each-round minimizers, so term (ii) can be well upper bounded as follows.

$$
\sum_{t=1}^{T}f_t(u_t)-\sum_{t=1}^{T}f_t(\bar{x}_t^*)\le 2\Delta\sum_{t=2}^{T}\|A_ty_t-A_{t-1}y_{t-1}\|_\infty.
$$

Combining the above two inequalities, we get the worst-case dynamic regret for the x-player: for any $i\in\mathcal{S}_{1,x}$, we have

$$
\begin{aligned}
&\sum_{t=1}^{T}x_t^\top A_ty_t-\sum_{t=1}^{T}\bar{x}_t^{*\top}A_ty_t\\
&\le\widetilde{\mathcal{O}}\left(\frac{T/\Delta}{\eta_i^x}+\eta_i^xV_T\right)+2\eta_i^x c\beta\sum_{t=2}^{T}\|q_t-q_{t-1}\|_1^2+2\eta_i^x c\beta\sum_{t=2}^{T}\sum_{j\in\mathcal{S}_y}q_{t,j}\|y_{t,j}-y_{t-1,j}\|_1^2\\
&\quad+2\Delta\sum_{t=2}^{T}\|A_ty_t-A_{t-1}y_{t-1}\|_\infty-\frac{L}{2}\sum_{t=2}^{T}\|p_t-p_{t-1}\|_1^2-\lambda\sum_{t=2}^{T}\sum_{i\in\mathcal{S}_x}p_{t,i}\|x_{t,i}-x_{t-1,i}\|_1^2.
\end{aligned}
$$

Similarly, we can also obtain the worst-case dynamic regret for the y-player: for any $j\in\mathcal{S}_{1,y}$, we have

$$
\begin{aligned}
&\sum_{t=1}^{T}x_t^\top A_t\bar{y}_t^*-\sum_{t=1}^{T}x_t^\top A_ty_t\\
&\le\widetilde{\mathcal{O}}\left(\frac{T/\Delta}{\eta_j^y}+\eta_j^yV_T\right)+2\eta_j^y c\beta\sum_{t=2}^{T}\|p_t-p_{t-1}\|_1^2+2\eta_j^y c\beta\sum_{t=2}^{T}\sum_{i\in\mathcal{S}_x}p_{t,i}\|x_{t,i}-x_{t-1,i}\|_1^2\\
&\quad+2\Delta\sum_{t=2}^{T}\|x_t^\top A_t-x_{t-1}^\top A_{t-1}\|_\infty-\frac{L}{2}\sum_{t=2}^{T}\|q_t-q_{t-1}\|_1^2-\lambda\sum_{t=2}^{T}\sum_{j\in\mathcal{S}_y}q_{t,j}\|y_{t,j}-y_{t-1,j}\|_1^2.
\end{aligned}
$$

Combining the two dynamic regret bounds yields: for any $i\in\mathcal{S}_{1,x}$ and any $j\in\mathcal{S}_{1,y}$,

$$
\sum_{t=1}^{T}x_t^\top A_t\bar{y}_t^*-\sum_{t=1}^{T}\bar{x}_t^{*\top}A_ty_t
$$

$$
\begin{aligned}
&\leq \widetilde{\mathcal{O}}\left(\frac{T/\Delta}{\eta_i^x}+\frac{T/\Delta}{\eta_j^y}+\eta_i^x V_T+\eta_j^y V_T\right)\\
&\quad+2\Delta\sum_{t=2}^{T}\|A_t y_t-A_{t-1}y_{t-1}\|_\infty+2\Delta\sum_{t=2}^{T}\|x_t^\top A_t-x_{t-1}^\top A_{t-1}\|_\infty\\
&\quad+\left(2\eta_j^y c\beta-\frac{L}{2}\right)\sum_{t=2}^{T}\|q_t-q_{t-1}\|_1^2+\left(2\eta_i^x c\beta-\frac{L}{2}\right)\sum_{t=2}^{T}\|p_t-p_{t-1}\|_1^2\\
&\quad+(2\eta_j^y c\beta-\lambda)\sum_{t=2}^{T}\sum_{j\in\mathcal{S}_y}q_{t,j}\|y_{t,j}-y_{t-1,j}\|_1^2+(2\eta_i^x c\beta-\lambda)\sum_{t=2}^{T}\sum_{i\in\mathcal{S}_x}p_{t,i}\|x_{t,i}-x_{t-1,i}\|_1^2\\
&\leq \widetilde{\mathcal{O}}\left(\frac{T/\Delta}{\eta_i^x}+\frac{T/\Delta}{\eta_j^y}+\eta_i^x V_T+\eta_j^y V_T\right)+2\Delta\sum_{t=2}^{T}\|A_t y_t-A_{t-1}y_{t-1}\|_\infty+2\Delta\sum_{t=2}^{T}\|x_t^\top A_t-x_{t-1}^\top A_{t-1}\|_\infty, \qquad (29)
\end{aligned}
$$

where the last inequality is because $2\eta_j^y c\beta-\frac{L}{2}\leq 0$, $2\eta_i^x c\beta-\frac{L}{2}\leq 0$, $2\eta_i^x c\beta-\gamma\leq 0$, $2\eta_i^x c\beta-\gamma\leq 0$ based on the fact that $\eta_j^y\leq\frac{1}{L}$, $\eta_i^x\leq\frac{1}{L}$ and $L=\max\{4,\sqrt{16c\beta},\sqrt{\frac{8c\beta}{\gamma}}\}$ and $\lambda=\frac{\gamma L}{2}$.

Next, we bound the last two terms in the right-hand side of Eq. (29). Indeed,

$$
\begin{aligned}
&\sum_{t=2}^{T}\|A_t y_t-A_{t-1}y_{t-1}\|_\infty+\sum_{t=2}^{T}\|x_t^\top A_t-x_{t-1}^\top A_{t-1}\|_\infty\\
&\leq\sqrt{T}\left(\sum_{t=2}^{T}\|A_t y_t-A_{t-1}y_{t-1}\|_\infty^2+\sum_{t=2}^{T}\|x_t^\top A_t-x_{t-1}^\top A_{t-1}\|_\infty^2\right)^{\frac{1}{2}}\\
&\leq\widetilde{\mathcal{O}}\left(\sqrt{T(1+V_T+\min\{P_T,W_T\})}\right),
\end{aligned}
$$

where the first inequality is by Cauchy-Schwarz inequality and the second inequality uses the gradient-variation bound in Lemma 19. Plugging this into Eq. (29) and choosing $\eta_i^x,\eta_j^y\in[\frac{1}{2}\eta_*,2\eta_*]$ with $\eta_*=\min\left\{\frac{1}{L},\sqrt{\frac{T}{\Delta(1+V_T)}}\right\}$, we have

$$
\begin{aligned}
\sum_{t=1}^{T}x_t^\top A_t\bar{y}_t^*-\sum_{t=1}^{T}\bar{x}_t^{*\top}A_t y_t&\leq\widetilde{\mathcal{O}}\left(\sqrt{\frac{T(1+V_T)}{\Delta}}+\frac{T}{\Delta}+\Delta\sqrt{T(1+V_T+\min\{P_T,W_T\})}\right)\\
&\leq\widetilde{\mathcal{O}}\left(\frac{T}{\Delta}+\Delta\sqrt{T(1+V_T+\min\{P_T,W_T\})}\right)\\
&=\widetilde{\mathcal{O}}\left(T^{\frac{3}{4}}\left(1+V_T+\min\{P_T,W_T\}\right)^{\frac{1}{4}}\right).
\end{aligned}
$$

where the last inequality is by setting the epoch length Δ optimally. We remark that the above choice of η_i^x,η_j^y is viable due to the construction of step size pool and the fact $\sqrt{T/(\Delta(1+V_T))}\geq\Theta(1/T)$; besides, the setting of epoch length Δ is also feasible, and notably the epoch length is only used in the analysis and our algorithm does not require its information.

The $\widetilde{\mathcal{O}}(T^{\frac{1}{2}}(1+Q_T^{\frac{3}{2}}+P_TQ_T)^{\frac{1}{2}})$ bound. From the update rule of the meta-algorithm and Eq. (28) proven in Theorem 12 with $\psi(x)=\frac{1}{2}\|x\|_2^2$, we have the following instantaneous regret bound for any $p\in\Delta_{|\mathcal{S}_x|}$ and $q\in\Delta_{|\mathcal{S}_y|}$.

$$
\begin{aligned}
\langle p_t-p,\ell_t^x\rangle&\leq\varepsilon_t^x\|\ell_t^x-m_t^x\|_2^2+\frac{1}{\varepsilon_t^x}\Big(\|\widehat{p}_t-p\|_2^2-\|\widehat{p}_{t+1}-p\|_2^2-\|p_t-\widehat{p}_{t+1}\|_2^2-\|p_t-\widehat{p}_t\|_2^2\Big),\\
\langle q_t-q,\ell_t^y\rangle&\leq\varepsilon_t^y\|\ell_t^y-m_t^y\|_2^2+\frac{1}{\varepsilon_t^y}\Big(\|\widehat{q}_t-q\|_2^2-\|\widehat{q}_{t+1}-q\|_2^2-\|q_t-\widehat{q}_{t+1}\|_2^2-\|q_t-\widehat{q}_t\|_2^2\Big).
\end{aligned}
\qquad (30)
$$

Recall that the feedback loss and optimism are set as follows. For x-player, we have $\ell_t^x=X_t^\top A_t y_t+\lambda X_t^\delta$ and $m_t^x=X_t^\top A_{t-1}y_{t-1}+\lambda X_t^\delta$, where $X_t=[x_{t,1};x_{t,2};\ldots,x_{t,|\mathcal{S}_x|}]$ and $X_t^\delta=[\|x_{t,1}-x_{t-1,1}\|_1^2;\|x_{t,2}-x_{t-1,2}\|_1^2;\ldots;\|x_{t,|\mathcal{S}_x|}-x_{t-1,|\mathcal{S}_x|}\|_1^2]$. For y-player, we have $\ell_t^y=-Y_t^\top A_t^\top x_t+\lambda Y_t^\delta$, $m_t^y=-Y_t^\top A_{t-1}^\top x_{t-1}+\lambda Y_t^\delta$, $Y_t=[y_{t,1};y_{t,2};\ldots;y_{t,|\mathcal{S}_y|}]$

and $Y_t^\delta = [\|y_{t,1} - y_{t-1,1}\|_1^2; \|y_{t,2} - y_{t-1,2}\|_1^2; \ldots; \|y_{t,|\mathcal{S}_y|} - y_{t-1,|\mathcal{S}_y|}\|_1^2]$. Note that $\mathcal{S}_x \triangleq \mathcal{S}_{1,x} \cup \mathcal{S}_{2,x}$ with $\mathcal{S}_{1,x} = [N]$ and $\mathcal{S}_{2,x} = \{N+1, \ldots, N+m\}$; besides, $\mathcal{S}_y \triangleq \mathcal{S}_{1,y} \cup \mathcal{S}_{2,y}$ with $\mathcal{S}_{1,y} = [N]$ and $\mathcal{S}_{2,y} = \{N+1, \ldots, N+n\}$. $N = \lfloor \frac{1}{2}\log_2 T \rfloor + 1$.

Then using Cauchy-Schwarz inequality and noticing that the dimensions of X_t, Y_t are $\widetilde{\Theta}(1)$, we have

$$\begin{aligned}
\|\ell_t^x - m_t^x\|_2^2 &= \|X_t^\top A_t y_t - X_t^\top A_{t-1} y_{t-1}\|_2^2 \le c' \left(\|A_t - A_{t-1}\|_\infty^2 + \|y_t - y_{t-1}\|_2^2\right), \\
\|\ell_t^y - m_t^y\|_2^2 &= \| - Y_t^\top A_t^\top x_t + -Y_t^\top A_{t-1}^\top x_{t-1}\|_2^2 \le c' \left(\|A_t - A_{t-1}\|_\infty^2 + \|x_t - x_{t-1}\|_2^2\right),
\end{aligned} \tag{31}$$

where $c' > 0$ is a universal constant independent with the time horizon and the non-stationarity measures (ignoring the dependence on poly-logarithmic factors in T).

In the following, we will specify the choice of the compared weight vectors. Concretely, let (x_t^*, y_t^*) be any Nash equilibrium of the payoff matrix A_t. We pick the compared weight distribution $p = p_t^* \in \Delta_{|\mathcal{S}_x|}$ and $q = q_t^* \in \Delta_{|\mathcal{S}_y|}$ such that both p_t^* and q_t^* have supports only on the additional dummy base-learners and the supports finally result in a Nash equilibrium of A_t, namely, $p_{t,i}^* = q_{t,j}^* = 0$ for $i = 1, \ldots, |\mathcal{S}_{1,x}|$ and $j = 1, \ldots, |\mathcal{S}_{1,y}|$, $p_{t,i+|\mathcal{S}_{1,x}|}^* = x_{t,i}^*$ for $i = 1, \ldots, m$ and $q_{t,j+|\mathcal{S}_{1,y}|}^* = y_{t,j}^*$ for $j = 1, \ldots, n$. By definition, we have

$$\left\langle p_t - p_t^*, X_t^\top A_t y_t \right\rangle + \left\langle q_t - q_t^*, -Y_t^\top A_t x_t \right\rangle = -x_t^{*\top} A_t y_t + x_t^\top A_t y_t^* \ge 0. \tag{32}$$

Moreover, combining Eq. (30) and Eq. (31) yields the following results:

$$\begin{aligned}
\left\langle p_t - p_t^*, X_t^\top A_t y_t \right\rangle \le \frac{1}{\varepsilon_t^x} &\left(\|\widehat{p}_t - p_t^*\|_2^2 - \|\widehat{p}_{t+1} - p_t^*\|_2^2 - \|p_t - \widehat{p}_{t+1}\|_2^2 - \|p_t - \widehat{p}_t\|_2^2\right) \\
&+ c' \cdot \varepsilon_t^x \left(\|A_t - A_{t-1}\|_\infty^2 + \|y_t - y_{t-1}\|_2^2\right) + \lambda \left\langle p_t^* - p_t, X_t^\delta \right\rangle.
\end{aligned}$$

Notice that in fact we have $\left\langle p_t^*, X_t^\delta \right\rangle = 0$ due to the choice of p_t^*. More specifically, p_t^* has support only on the additional dummy base-learners and for those dummy learners their stability quantity is zero (namely, $X_{t,i}^\delta = 0$ for $i = |\mathcal{S}_{1,x}| + 1, \ldots, |\mathcal{S}_{1,x}| + m$). In addition, it is clear that $\left\langle p_t, X_t^\delta \right\rangle \ge 0$, so we have

$$\begin{aligned}
&\left\langle p_t - p_t^*, X_t^\top A_t y_t \right\rangle \\
&\le \frac{1}{\varepsilon_t^x} \left(\|\widehat{p}_t - p_t^*\|_2^2 - \|\widehat{p}_{t+1} - p_t^*\|_2^2 - \|p_t - \widehat{p}_{t+1}\|_2^2 - \|p_t - \widehat{p}_t\|_2^2\right) + c' \cdot \varepsilon_t^x \left(\|A_t - A_{t-1}\|_\infty^2 + \|y_t - y_{t-1}\|_2^2\right).
\end{aligned}$$

Similarly, we get

$$\begin{aligned}
&\left\langle q_t - q_t^*, -Y_t^\top A_t x_t \right\rangle \\
&\le \frac{1}{\varepsilon_t^y} \left(\|\widehat{q}_t - q_t^*\|_2^2 - \|\widehat{q}_{t+1} - q_t^*\|_2^2 - \|q_t - \widehat{q}_{t+1}\|_2^2 - \|q_t - \widehat{q}_t\|_2^2\right) + c' \cdot \varepsilon_t^y \left(\|A_t - A_{t-1}\|_\infty^2 + \|x_t - x_{t-1}\|_2^2\right).
\end{aligned}$$

Adding the above two inequalities and rearranging the terms, based on Eq. (32), we have

$$\begin{aligned}
&\frac{1}{\varepsilon_t^x} \left(\|p_t - \widehat{p}_{t+1}\|_2^2 + \|p_t - \widehat{p}_t\|_2^2\right) + \frac{1}{\varepsilon_t^y} \left(\|q_t - \widehat{q}_{t+1}\|_2^2 + \|q_t - \widehat{q}_t\|_2^2\right) \\
&\le \frac{1}{\varepsilon_t^x} \left(\|\widehat{p}_t - p_t^*\|_2^2 - \|\widehat{p}_{t+1} - p_t^*\|_2^2\right) + c' \cdot \varepsilon_t^x (\|A_t - A_{t-1}\|_\infty^2 + \|y_t - y_{t-1}\|_2^2) \\
&\quad + \frac{1}{\varepsilon_t^y} \left(\|\widehat{q}_t - q_t^*\|_2^2 - \|\widehat{q}_{t+1} - q_t^*\|_2^2\right) + c' \cdot \varepsilon_t^y (\|A_t - A_{t-1}\|_\infty^2 + \|x_t - x_{t-1}\|_2^2) + \lambda D(\|X_t^\delta\|_2 + \|Y_t^\delta\|_2) \\
&\le \frac{1}{\varepsilon_t^x} \left(\|\widehat{p}_t - p_t^*\|_2^2 - \|\widehat{p}_{t+1} - p_{t+1}^*\|_2^2\right) + c' \cdot \varepsilon_t^x (\|A_t - A_{t-1}\|_\infty^2 + \|y_t - y_{t-1}\|_2^2) \\
&\quad + \frac{1}{\varepsilon_t^y} \left(\|\widehat{q}_t - q_t^*\|_2^2 - \|\widehat{q}_{t+1} - q_{t+1}^*\|_2^2\right) + c' \cdot \varepsilon_t^y (\|A_t - A_{t-1}\|_\infty^2 + \|x_t - x_{t-1}\|_2^2) \\
&\quad + 2D \left(\frac{1}{\varepsilon_t^x} + \frac{1}{\varepsilon_t^y}\right) \cdot (\|p_t^* - p_{t+1}^*\|_1 + \|q_t^* - q_{t+1}^*\|_1).
\end{aligned} \tag{33}$$

In above, $D = \sqrt{2(N + \max\{m, n\})} = \widetilde{\Theta}(1)$ is another universal constant (ignoring the dependence on logarithmic factors in T) serving as the upper bound of $\|p - p'\|_2$ and $\|q - q'\|_2$ for any $p, p' \in \Delta_{|\mathcal{S}_x|}$ and for any $q, q' \in \Delta_{|\mathcal{S}_y|}$.

Next, we show that the desired duality gap upper bound can be related to the terms on the left-hand side of above inequality, namely, $\frac{1}{\varepsilon_t^x}\left(\|p_t - \widehat{p}_{t+1}\|_2^2 + \|p_t - \widehat{p}_t\|_2^2\right) + \frac{1}{\varepsilon_t^y}\left(\|q_t - \widehat{q}_{t+1}\|_2^2 + \|q_t - \widehat{q}_t\|_2^2\right)$. To see this, we first have the following inequalities from the update rule of the meta-algorithm as well as the first-order optimality condition: for any $p' \in \Delta_{|\mathcal{S}_x|}$ and $q' \in \Delta_{|\mathcal{S}_y|}$,

$$
\begin{aligned}
&(\widehat{p}_{t+1} - \widehat{p}_t + \varepsilon_t^x X_t^\top A_t y_t + \varepsilon_t^x \lambda X_t^\delta)^\top (p' - \widehat{p}_{t+1}) \geq 0, \\
&(\widehat{q}_{t+1} - \widehat{q}_t - \varepsilon_t^y Y_t^\top A_t^\top x_t + \varepsilon_t^y \lambda Y_t^\delta)^\top (q' - \widehat{q}_{t+1}) \geq 0.
\end{aligned}
$$

Rearranging the terms and introducing the notations $\widetilde{x}_t \triangleq \sum_{i \in \mathcal{S}_x} \widehat{p}_{t+1,i} x_{t,i}$ and $\widetilde{y}_t \triangleq \sum_{j \in \mathcal{S}_y} \widehat{q}_{t+1,j} y_{t,j}$, we then have for any $p' \in \Delta_{|\mathcal{S}_x|}$ and $q' \in \Delta_{|\mathcal{S}_y|}$,

$$
\begin{aligned}
&(\widehat{p}_{t+1} - \widehat{p}_t)^\top (p' - \widehat{p}_{t+1}) \\
&\geq \varepsilon_t^x (\widehat{p}_{t+1} - p')^\top (X_t^\top A_t y_t + \lambda X_t^\delta) \\
&= \varepsilon_t^x (\widehat{p}_{t+1} - p')^\top (X_t^\top A_t \widetilde{y}_t + X_t^\top A_t (y_t - \widetilde{y}_t) + \lambda X_t^\delta) \\
&\geq \varepsilon_t^x (\widehat{p}_{t+1} - p')^\top X_t^\top A_t \widetilde{y}_t - c'' \cdot \varepsilon_t^x \|\widehat{p}_{t+1} - p'\|_2 \cdot \|q_t - \widehat{q}_{t+1}\|_2 + \lambda \varepsilon_t^x \left\langle \widehat{p}_{t+1} - p', X_t^\delta \right\rangle,
\end{aligned}
$$

and also

$$
\begin{aligned}
&(\widehat{q}_{t+1} - \widehat{q}_t)^\top (q' - \widehat{q}_{t+1}) \\
&\geq \varepsilon_{t,y} (\widehat{q}_{t+1} - q')^\top (-Y_t^\top A_t x_t + \lambda Y_t^\delta) \\
&= \varepsilon_t^y (\widehat{q}_{t+1} - q')^\top (-Y_t^\top A_t \widetilde{x}_t - Y_t^\top A_t (x_t - \widetilde{x}_t) + \lambda Y_t^\delta) \\
&\geq \varepsilon_t^y (\widehat{q}_{t+1} - q')^\top (-Y_t^\top A_t \widetilde{x}_t) - c'' \cdot \varepsilon_t^y \|\widehat{q}_{t+1} - q'\|_2 \cdot \|p_t - \widehat{p}_{t+1}\|_2 + \lambda \varepsilon_t^y \left\langle \widehat{q}_{t+1} - q', Y_t^\delta \right\rangle,
\end{aligned}
$$

where $c'' > 0$ is also a universal constant independent with the time horizon and the non-stationarity measures (ignoring the dependence on poly-logarithmic factors in T). Rearranging the terms arrives that

$$
\begin{aligned}
(\widehat{p}_{t+1} - p')^\top X_t^\top A_t \widetilde{y}_t &\leq \|p' - \widehat{p}_{t+1}\|_2 \cdot \widetilde{\mathcal{O}}\left(\frac{1}{\varepsilon_t^x}\|\widehat{p}_{t+1} - \widehat{p}_t\|_2 + \|q_t - \widehat{q}_{t+1}\|_2\right) + \lambda \left\langle p' - \widehat{p}_{t+1}, X_t^\delta \right\rangle, \\
(\widehat{q}_{t+1} - q')^\top (-Y_t^\top A_t \widetilde{x}_t) &\leq \|q' - \widehat{q}_{t+1}\|_2 \cdot \widetilde{\mathcal{O}}\left(\frac{1}{\varepsilon_t^y}\|\widehat{q}_{t+1} - \widehat{q}_t\|_2 + \|p_t - \widehat{p}_{t+1}\|_2\right) + \lambda \left\langle q' - \widehat{q}_{t+1}, Y_t^\delta \right\rangle.
\end{aligned}
$$

Let $(\bar{x}_t^*, \bar{y}_t^*)$ be the corresponding best response for the strategy $(\widetilde{x}_t, \widetilde{y}_t)$ with respect to the payoff A_t, i.e., $\bar{x}_t^* = \operatorname{argmin}_{x \in \Delta_m} x^\top A_t \widetilde{y}_t$ and $\bar{y}_t^* = \operatorname{argmax}_{y \in \Delta_n} \widetilde{x}_t^\top A_t y$. Denote the duality gap bound of $(\widetilde{x}_t, \widetilde{y}_t)$ as $\alpha_t(\widetilde{x}_t, \widetilde{y}_t) \triangleq \max_y \widetilde{x}_t^\top A_t y - \min_x x^\top A_t \widetilde{y}_t$. Now we pick the comparator vectors $p' \in \Delta_{|\mathcal{S}_x|}$ and $q' \in \Delta_{|\mathcal{S}_y|}$ such that both p' and q' have supports only on the additional dummy base-learners and the supports finally form the best response of $\widetilde{x}_t, \widetilde{y}_t$, namely, $p'_i = q'_j = 0$ for $i = 1, \ldots, |\mathcal{S}_{1,x}|$, $j = 1, \ldots, |\mathcal{S}_{1,y}|$ and $p'_{i+|\mathcal{S}_{1,x}|} = \bar{x}_{t,i}^*$ for $i = 1, \ldots, m$ and $q'_{j+|\mathcal{S}_{1,y}|} = \bar{y}_{t,j}^*$ for $j = 1, \ldots, n$. Due to the construction, we confirm that $\left\langle p', X_t^\delta \right\rangle = 0$ and $\left\langle q', Y_t^\delta \right\rangle = 0$.

As a result, combining the two inequalities on the above gives the following upper bound for duality gap:

$$
\begin{aligned}
\alpha_t(\widetilde{x}_t, \widetilde{y}_t) &= (\widehat{p}_{t+1} - p')^\top X_t^\top A_t \widetilde{y}_t + (\widehat{q}_{t+1} - q')^\top (-Y_t^\top A_t \widetilde{x}_t) \\
&\leq \widetilde{\mathcal{O}}\left(\frac{1}{\varepsilon_t^x}\|\widehat{p}_{t+1} - \widehat{p}_t\|_2 + \|q_t - \widehat{q}_{t+1}\|_2 + \frac{1}{\varepsilon_t^y}\|\widehat{q}_{t+1} - \widehat{q}_t\|_2 + \|p_t - \widehat{p}_{t+1}\|_2\right) \\
&\leq \widetilde{\mathcal{O}}\left(\frac{1}{\varepsilon_t^x}(\|\widehat{p}_{t+1} - \widehat{p}_t\|_2 + \|p_{t+1} - \widehat{p}_{t+1}\|_2) + \frac{1}{\varepsilon_t^y}(\|\widehat{q}_{t+1} - \widehat{q}_t\|_2 + \|q_t - \widehat{q}_{t+1}\|_2)\right).
\end{aligned}
$$

The first inequality holds because $\|p' - \widehat{p}_{t+1}\|_2 \leq D = \widetilde{\Theta}(1)$ and $\|q' - \widehat{q}_{t+1}\|_2 \leq D = \widetilde{\Theta}(1)$ hold for any $t \in [T]$. The second inequality is obtained by scaling two terms with a factor of $\frac{1}{\varepsilon_t^x}$ and $\frac{1}{\varepsilon_t^y}$ respectively, where we notice that $1 \leq \frac{1}{L\varepsilon_t^x}$ and $1 \leq \frac{1}{L\varepsilon_t^y}$ are true as $\varepsilon_t^x \leq \frac{1}{L}$ and $\varepsilon_t^y \leq \frac{1}{L}$ holds for all $t \in [T]$.

Then, by Cauchy-Schwarz inequality, we obtain the following upper bound for the square of duality gap bound of $(\widetilde{x}_t, \widetilde{y}_t)$:

$$
\begin{aligned}
&\alpha_t^2(\widetilde{x}_t, \widetilde{y}_t) \\
&\leq \widetilde{\mathcal{O}}\left(\left(\frac{1}{\varepsilon_t^x}(\|\widehat{p}_{t+1} - \widehat{p}_t\|_2 + \|p_{t+1} - \widehat{p}_{t+1}\|_2) + \frac{1}{\varepsilon_t^y}(\|\widehat{q}_{t+1} - \widehat{q}_t\|_2 + \|q_t - \widehat{q}_{t+1}\|_2)\right)^2\right) \\
&\leq \widetilde{\mathcal{O}}\left(\left(\frac{1}{\varepsilon_t^x} + \frac{1}{\varepsilon_t^y}\right)\left(\frac{1}{\varepsilon_t^x}(\|\widehat{p}_{t+1} - \widehat{p}_t\|_2^2 + \|p_{t+1} - \widehat{p}_{t+1}\|_2^2) + \frac{1}{\varepsilon_t^y}(\|\widehat{q}_{t+1} - \widehat{q}_t\|_2^2 + \|q_t - \widehat{q}_{t+1}\|_2^2)\right)\right) \\
&\leq \widetilde{\mathcal{O}}\left(\left(\frac{1}{\varepsilon_t^x} + \frac{1}{\varepsilon_t^y}\right)\left(\frac{1}{\varepsilon_t^x}(\|\widehat{p}_t - p_t^*\|_2^2 - \|\widehat{p}_{t+1} - p_{t+1}^*\|_2^2) + \varepsilon_t^x(\|A_t - A_{t-1}\|_\infty^2 + \|y_t - y_{t-1}\|_2^2)\right)\right) \\
&\quad + \widetilde{\mathcal{O}}\left(\left(\frac{1}{\varepsilon_t^x} + \frac{1}{\varepsilon_t^y}\right)\left(\frac{1}{\varepsilon_t^y}(\|\widehat{q}_t - q_t^*\|_2^2 - \|\widehat{q}_{t+1} - q_{t+1}^*\|_2^2) + \varepsilon_t^y(\|A_t - A_{t-1}\|_\infty^2 + \|x_t - x_{t-1}\|_2^2)\right)\right) \\
&\quad + \widetilde{\mathcal{O}}\left(\left(\frac{1}{\varepsilon_t^x} + \frac{1}{\varepsilon_t^y}\right)^2 \cdot (\|p_t^* - p_{t+1}^*\|_1 + \|q_t^* - q_{t+1}^*\|_1)\right).
\end{aligned}
$$

Notably, the last step makes use of the inequality in Eq. (33) and $\lambda = \frac{\gamma L}{2} = \widetilde{\Theta}(1)$. For simplicity, we introduce the notation $\frac{1}{\varepsilon_t} \triangleq \frac{1}{\varepsilon_t^x} + \frac{1}{\varepsilon_t^y}$. Taking a summation on the squared duality gap over all rounds and using the fact that $\varepsilon_t^x, \varepsilon_t^y \leq \widetilde{\mathcal{O}}(1)$ and $\varepsilon_t^x, \varepsilon_t^y$ are non-increasing in t, we have (omitting all dimension and $\mathrm{poly}(\log T)$ factors)

$$
\begin{aligned}
&\sum_{t=1}^T \alpha_t^2(\widetilde{x}_t, \widetilde{y}_t) \\
&\leq \sum_{t=1}^T \widetilde{\mathcal{O}}\left(\left(\frac{1}{\varepsilon_{t+1}^x \cdot \varepsilon_{t+1}} - \frac{1}{\varepsilon_t^x \cdot \varepsilon_t}\right)\|\widehat{p}_{t+1} - p_{t+1}^*\|_2^2\right) + \sum_{t=1}^T \widetilde{\mathcal{O}}\left(\left(\frac{1}{\varepsilon_{t+1}^y \cdot \varepsilon_{t+1}} - \frac{1}{\varepsilon_t^y \cdot \varepsilon_t}\right)\|\widehat{q}_{t+1} - q_{t+1}^*\|_2^2\right) \\
&\quad + \frac{1}{\varepsilon_T^2}\widetilde{\mathcal{O}}(P_T) + \frac{1}{\varepsilon_T}\sum_{t=2}^T \widetilde{\mathcal{O}}\Big(\|A_t - A_{t-1}\|_\infty^2 + \|y_t - y_{t-1}\|_2^2 + \|x_t - x_{t-1}\|_2^2\Big) \\
&\leq \widetilde{\mathcal{O}}\left(\frac{1 + P_T}{\varepsilon_{T+1}^2}\right) + \frac{1}{\varepsilon_{T+1}}\sum_{t=2}^T \widetilde{\mathcal{O}}\Big(\|A_t - A_{t-1}\|_\infty^2 + \|x_t - x_{t-1}\|_2^2 + \|y_t - y_{t-1}\|_2^2\Big),
\end{aligned}
$$

where the last inequality uses $\|\widehat{p}_{t+1} - p_{t+1}^*\|_2 \leq \widetilde{\mathcal{O}}(1)$, $\|\widehat{q}_{t+1} - q_{t+1}^*\|_2 \leq \widetilde{\mathcal{O}}(1)$. According to Lemma 16 and Lemma 18,

$$
\sum_{t=2}^T \|x_t - x_{t-1}\|_2^2 + \sum_{t=2}^T \|y_t - y_{t-1}\|_2^2 \leq \widetilde{\mathcal{O}}\left(\min\left\{\sqrt{(1 + V_T)(1 + P_T)} + P_T, 1 + W_T\right\}\right).
$$

In addition, according to the definition of ε_t^x and ε_t^y, we have

$$
\begin{aligned}
\frac{1}{\varepsilon_{T+1}^x} &= \sqrt{L^2 + \sum_{t=2}^T \|A_t y_t - A_{t-1} y_{t-1}\|_\infty^2} \leq \widetilde{\mathcal{O}}\left(\sqrt{1 + V_T + \min\{P_T, W_T\}}\right), \\
\frac{1}{\varepsilon_{T+1}^y} &= \sqrt{L^2 + \sum_{t=2}^T \|x_t^\top A_t - x_{t-1}^\top A_{t-1}\|_\infty^2} \leq \widetilde{\mathcal{O}}\left(\sqrt{1 + V_T + \min\{P_T, W_T\}}\right),
\end{aligned}
$$

where the last inequality is due to the gradient-variation bound in Lemma 19. Therefore, combining all above inequalities can achieve the following result on the squared duality gap:

$$
\begin{aligned}
\sum_{t=1}^T \alpha_t^2(\widetilde{x}_t, \widetilde{y}_t) &\leq \widetilde{\mathcal{O}}\Big((1 + P_T)(1 + V_T + \min\{P_T, W_T\})\Big) + \widetilde{\mathcal{O}}\Big((1 + V_T + \min\{W_T, P_T\})^{\frac{3}{2}}\Big) \\
&= \widetilde{\mathcal{O}}\left((1 + V_T + \min\{P_T, W_T\})\left(\sqrt{1 + V_T + \min\{P_T, W_T\}} + P_T\right)\right).
\end{aligned}
$$

We further introduce the notation $Q_T \triangleq V_t + \min\{P_T, W_T\}$ to simplify the presentation. Then, by Cauchy-Schwarz inequality we have

$$\sum_{t=1}^{T} \alpha_t(\widetilde{x}_t, \widetilde{y}_t) \leq \widetilde{\mathcal{O}}\left(\sqrt{T(1+Q_T)(\sqrt{1+Q_T}+P_T)}\right) = \widetilde{\mathcal{O}}\left(T^{\frac{1}{2}}(1+Q_T^{\frac{3}{2}}+P_TQ_T)^{\frac{1}{2}}\right). \tag{34}$$

We finally transform the above bound back to $\alpha_t(x_t, y_t)$ by noticing that

$$\begin{aligned}
&\sum_{t=1}^{T} \alpha_t(x_t, y_t) \\
&= \sum_{t=1}^{T} \left(\max_{y\in\Delta_n} x_t^\top A_t y - \min_{x\in\Delta_m} x^\top A_t y_t\right) \\
&= \sum_{t=1}^{T} \left(\max_{y\in\Delta_n} \widetilde{x}_t^\top A_t y - \min_{x\in\Delta_m} x^\top A_t \widetilde{y}_t\right) + \sum_{t=1}^{T} \left(\max_{y\in\Delta_n} x_t^\top A_t y - \max_{y\in\Delta_n} \widetilde{x}_t^\top A_t y\right) + \sum_{t=1}^{T} \left(\min_{x\in\Delta_m} x^\top A_t \widetilde{y}_t - \min_{x\in\Delta_m} x^\top A_t y_t\right) \\
&\leq \sum_{t=1}^{T} \alpha_t(\widetilde{x}_t, \widetilde{y}_t) + \sum_{t=1}^{T} \widetilde{\mathcal{O}}\left(\|p_t - \widehat{p}_{t+1}\|_2 + \|q_t - \widehat{q}_{t+1}\|_2\right) \\
&\leq \sum_{t=1}^{T} \alpha_t(\widetilde{x}_t, \widetilde{y}_t) + \widetilde{\mathcal{O}}\left(\sqrt{T\sum_{t=1}^{T}(\|p_t - \widehat{p}_{t+1}\|_2^2 + \|q_t - \widehat{q}_{t+1}\|_2^2)}\right) && \text{(by Cauchy-Schwarz inequality)} \\
&\leq \sum_{t=1}^{T} \alpha_t(\widetilde{x}_t, \widetilde{y}_t) + \widetilde{\mathcal{O}}\left(\sqrt{T\min\{\sqrt{(1+V_T)(1+P_T)}+P_T, 1+W_T\}}\right). && \text{(by Lemma 16 and Lemma 18)} \\
&\leq \widetilde{\mathcal{O}}\left(T^{\frac{1}{2}}\left(1+Q_T^{\frac{3}{2}}+P_TQ_T\right)^{\frac{1}{2}}\right) + \widetilde{\mathcal{O}}\left(\sqrt{T(1+Q_T)}\right) && \text{(by Eq. (34) and Cauchy-Schwarz inequality)} \\
&\leq \widetilde{\mathcal{O}}\left(T^{\frac{1}{2}}\left(1+Q_T^{\frac{3}{2}}+P_TQ_T\right)^{\frac{1}{2}}\right).
\end{aligned}$$

To summarize, combining the both types of upper bounds for duality gap, we finally achieve the following guarantee:

$$\sum_{t=1}^{T} \max_{y\in\Delta_n} x_t^\top A_t y - \sum_{t=1}^{T} \min_{x\in\Delta_m} x^\top A_t y_t \leq \widetilde{\mathcal{O}}\Big(\min\{T^{\frac{3}{4}}\left(1+Q_T\right)^{\frac{1}{4}}, T^{\frac{1}{2}}(1+Q_T^{\frac{3}{2}}+P_TQ_T)^{\frac{1}{2}}\}\Big),$$

which completes the proof of Theorem 8. □

F. Key Lemmas

This section presents several key lemmas used in proving our theoretical results.

We first provide an analysis for the general dynamic regret of the meta-base two-layer approach, which serves as one of the key technical tools for proving upper bounds for the three performance measures. The result is shown in Lemma 15, and we emphasize that the regret bounds hold for *any* comparator sequence, which is crucial and useful in the subsequent analysis.

Lemma 15 (General dynamic regret)**.** *Algorithm 1 guarantees that x-player's dynamic regret with respect to any comparator sequence $u_1, \ldots, u_T \in \Delta_m$ is bounded by*

$$\begin{aligned}
&\sum_{t=1}^{T} x_t^\top A_t y_t - \sum_{t=1}^{T} u_t^\top A_t y_t \\
&\leq \mathcal{O}\left(\frac{\alpha(1+P_T^u)}{\eta_i^x}\right) + \eta_i^x c\beta \sum_{t=2}^{T} \|A_t - A_{t-1}\|_\infty^2 + \eta_i^x c\beta \sum_{t=2}^{T} \|y_t - y_{t-1}\|_1^2 + \left(\lambda - \frac{\gamma}{\eta_i^x}\right) \sum_{t=2}^{T} \|x_{t,i} - x_{t-1,i}\|_1^2 \\
&\quad - L\sum_{t=1}^{T}(\|p_t - \widehat{p}_{t+1}\|_2^2 + \|p_t - \widehat{p}_t\|_2^2) - \lambda \sum_{t=2}^{T} \sum_{i\in\mathcal{S}_x} p_{t,i}\|x_{t,i} - x_{t-1,i}\|_1^2 + \widetilde{\mathcal{O}}(1),
\end{aligned} \tag{35}$$

for a specific $c = \widetilde{\Theta}(1)$ and any compared base-learner's index $i \in \mathcal{S}_{1,x}$.

Similarly, Algorithm 2 guarantees that y-player's dynamic regret with respect to any comparator sequence $v_1, \ldots, v_T \in \Delta_n$ is at most

$$\begin{aligned}
&-\sum_{t=1}^{T} x_t^\top A_t y_t + \sum_{t=1}^{T} x_t^\top A_t v_t \\
&\le \mathcal{O}\left(\frac{\alpha(1+P_T^v)}{\eta_j^y}\right) + \eta_j^y c\beta \sum_{t=2}^{T} \|A_t - A_{t-1}\|_\infty^2 + \eta_j^y c\beta \sum_{t=2}^{T} \|x_t - x_{t-1}\|_1^2 + \left(\lambda - \frac{\gamma}{\eta_j^y}\right) \sum_{t=2}^{T} \|y_{t,i} - y_{t-1,i}\|_1^2 \qquad (36)\\
&\quad - L\sum_{t=1}^{T}(\|q_t - \widehat{q}_{t+1}\|_2^2 + \|q_t - \widehat{q}_t\|_2^2) - \lambda \sum_{t=2}^{T}\sum_{i\in\mathcal{S}_x} q_{t,i}\|y_{t,i} - y_{t-1,i}\|_1^2 + \widetilde{\mathcal{O}}(1),
\end{aligned}$$

which also holds for any compared base-learner's index $j \in \mathcal{S}_{1,y}$.

Proof. We consider the dynamic regret for x-player and similar results hold for y-player. First, we decompose the dynamic regret for x-player into the sum of the meta-regret and base-regret. Specifically, for any $i \in \mathcal{S}_{1,x}$, we have

$$\sum_{t=1}^{T} x_t^\top A_t y_t - \sum_{t=1}^{T} u_t A_t y_t = \underbrace{\sum_{t=1}^{T} x_t^\top A_t y_t - \sum_{t=1}^{T} x_{t,i}^\top A_t y_t}_{\texttt{meta-regret}} + \underbrace{\sum_{t=1}^{T} x_{t,i}^\top A_t y_t - \sum_{t=1}^{T} u_t^\top A_t y_t}_{\texttt{base-regret}}.$$

We now give upper bounds for the meta-regret and base-regret respectively.

First, we consider the meta-regret, which is essentially the static regret with respect to any base-learner with an index $i \in \mathcal{S}_{1,x}$. Recall several notations introduced in the algorithm. For x-player, $\ell_t^x = X_t^\top A_t y_t + \lambda X_t^\delta$ and $m_t^x = X_t^\top A_{t-1} y_{t-1} + \lambda X_t^\delta$, where $X_t = [x_{t,1}; x_{t,2}; \ldots, x_{t,|\mathcal{S}_x|}]$ and $X_t^\delta = [\|x_{t,1} - x_{t-1,1}\|_1^2; \|x_{t,2} - x_{t-1,2}\|_1^2; \ldots; \|x_{t,|\mathcal{S}_x|} - x_{t-1,|\mathcal{S}_x|}\|_1^2]$. Note that $\mathcal{S}_x \triangleq \mathcal{S}_{1,x} \cup \mathcal{S}_{2,x}$ with $\mathcal{S}_{1,x} = [N]$ and $\mathcal{S}_{2,x} = \{N+1, \ldots, N+m\}$, where $N = \lfloor \frac{1}{2}\log_2 T \rfloor + 1$. The notations for y-player are similarly defined and we do not restate here for conciseness. According to the general result of Theorem 12 with $f_t(p_t) = \langle p_t, X_t^\top A_t y_t + \lambda X_t^\delta \rangle$, $M_t = X_t^\top A_{t-1} y_{t-1} + \lambda X_t^\delta$, and $u_t = e_i \in \Delta_{|\mathcal{S}_x|}$ for all $t \in [T]$, we have the following regret bound for the meta-algorithm,

$$\begin{aligned}
&\sum_{t=1}^{T} \langle p_t - e_i, X_t^\top A_t y_t + \lambda X_t^\delta \rangle \\
&\le \sum_{t=2}^{T} \varepsilon_t^x \|X_t^\top A_t y_t - X_t^\top A_{t-1} y_{t-1}\|_2^2 \\
&\qquad + \sum_{t=1}^{T} \frac{1}{\varepsilon_t^x}\left(\|\widehat{p}_t - e_i\|_2^2 - \|\widehat{p}_{t+1} - e_i\|_2^2\right) - \sum_{t=1}^{T} \frac{1}{\varepsilon_t^x}\left(\|p_t - \widehat{p}_{t+1}\|_2^2 + \|p_t - \widehat{p}_t\|_2^2\right) + \widetilde{\mathcal{O}}(1) \\
&\le c_1 \sum_{t=2}^{T} \varepsilon_t^x \|A_t y_t - A_{t-1} y_{t-1}\|_\infty^2 \\
&\qquad + \frac{1}{\varepsilon_T^x} \sum_{t=1}^{T} \left(\|\widehat{p}_t - e_i\|_2^2 - \|\widehat{p}_{t+1} - e_i\|_2^2\right) - \sum_{t=1}^{T} \frac{1}{\varepsilon_t^x}\left(\|p_t - \widehat{p}_{t+1}\|_2^2 + \|p_t - \widehat{p}_t\|_2^2\right) + \widetilde{\mathcal{O}}(1) \\
&\le c_1 \sum_{t=2}^{T} \frac{\|A_t y_t - A_{t-1} y_{t-1}\|_\infty^2}{\sqrt{L^2 + \sum_{s=2}^{t-1} \|A_s y_s - A_{s-1} y_{s-1}\|_\infty^2}} + \frac{\widetilde{\mathcal{O}}(1)}{\varepsilon_T^x} - L\sum_{t=1}^{T}\left(\|p_t - \widehat{p}_{t+1}\|_2^2 + \|p_t - \widehat{p}_t\|_2^2\right) \\
&\qquad\qquad \text{(by definition of } \varepsilon_t^x \text{ and } \max_{p\in\Delta_{|\mathcal{S}_x|}} \|p - e_i\|_2^2 \le \widetilde{\mathcal{O}}(1)) \\
&\le c_2 \sqrt{L^2 + \sum_{t=2}^{T} \|A_t y_t - A_{t-1} y_{t-1}\|_\infty^2} + \widetilde{\mathcal{O}}(1) - L\sum_{t=1}^{T}(\|p_t - \widehat{p}_{t+1}\|_2^2 + \|p_t - \widehat{p}_t\|_2^2),
\end{aligned}$$

where $c_1, c_2 = \widetilde{\Theta}(1)$ and the last step holds by Lemma 23.

Next, we consider the base-regret. Since the base-algorithm $\mathcal{B}_i$ satisfies the DRVU property, the base-regret is upper bounded as follows:

$$\sum_{t=1}^{T} x_{t,i}^\top A_t y_t - \sum_{t=1}^{T} u_t^\top A_t y_t \le \frac{\alpha(1+P_T^u)}{\eta_i^x} + \eta_i^x \beta \sum_{t=2}^{T} \|A_t y_t - A_{t-1} y_{t-1}\|_\infty^2 - \frac{\gamma}{\eta_i^x} \sum_{t=2}^{T} \|x_{t,i} - x_{t-1,i}\|_1^2.$$

Summing up the above two inequalities, we achieve the following dynamic regret guarantee for the x-player:

$$\begin{aligned}
&\sum_{t=1}^{T} x_t^\top A_t y_t - \sum_{t=1}^{T} u_t^\top A_t y_t \\
&\le c_2 \sqrt{L^2 + \sum_{t=2}^{T} \|A_t y_t - A_{t-1} y_{t-1}\|_\infty^2} + \widetilde{\mathcal{O}}(1) - L \sum_{t=1}^{T} (\|p_t - \widehat{p}_{t+1}\|_2^2 + \|p_t - \widehat{p}_t\|_2^2) \\
&\quad + \frac{\alpha(1+P_T^u)}{\eta_i^x} + \eta_i^x \beta \sum_{t=2}^{T} \|A_t y_t - A_{t-1} y_{t-1}\|_\infty^2 - \frac{\gamma}{\eta_i^x} \sum_{t=2}^{T} \|x_{t,i} - x_{t-1,i}\|_1^2 \\
&\quad + \lambda \sum_{t=2}^{T} \|x_{t,i} - x_{t-1,i}\|_1^2 - \lambda \sum_{t=2}^{T} \sum_{i \in \mathcal{S}_x} p_{t,i} \|x_{t,i} - x_{t-1,i}\|_1^2 + \widetilde{\mathcal{O}}(1) \\
&\le \mathcal{O}\left(\frac{\alpha(1+P_T^u)}{\eta_i^x}\right) + \eta_i^x (c_2^2 + \beta) \sum_{t=2}^{T} \|A_t y_t - A_{t-1} y_{t-1}\|_\infty^2 + \left(\lambda - \frac{\gamma}{\eta_i^x}\right) \sum_{t=2}^{T} \|x_{t,i} - x_{t-1,i}\|_1^2 \\
&\quad - L \sum_{t=1}^{T} (\|p_t - \widehat{p}_{t+1}\|_2^2 + \|p_t - \widehat{p}_t\|_2^2) - \lambda \sum_{t=2}^{T} \sum_{i \in \mathcal{S}_x} p_{t,i} \|x_{t,i} - x_{t-1,i}\|_1^2 + \widetilde{\mathcal{O}}(1) \\
&\le \mathcal{O}\left(\frac{\alpha(1+P_T^u)}{\eta_i^x}\right) + \eta_i^x c \beta \sum_{t=2}^{T} \|A_t - A_{t-1}\|_\infty^2 + \eta_i^x c \beta \sum_{t=2}^{T} \|y_t - y_{t-1}\|_1^2 + \left(\lambda - \frac{\gamma}{\eta_i^x}\right) \sum_{t=2}^{T} \|x_{t,i} - x_{t-1,i}\|_1^2 \\
&\quad - L \sum_{t=1}^{T} (\|p_t - \widehat{p}_{t+1}\|_2^2 + \|p_t - \widehat{p}_t\|_2^2) - \lambda \sum_{t=2}^{T} \sum_{i \in \mathcal{S}_x} p_{t,i} \|x_{t,i} - x_{t-1,i}\|_1^2 + \widetilde{\mathcal{O}}(1),
\end{aligned}$$

where $c = \widetilde{\Theta}(1)$. This proves Eq. (35). Repeating the above analysis for y-player proves Eq. (36). □

The following lemma presents the stability lemma in terms of the non-stationarity measure P_T (that is, the NE variation). We give the stability upper bounds from the aspects of meta-algorithm and final decisions. For simplicity, we assume that the DRVU parameters (α, β, γ) are all $\widetilde{\Theta}(1)$, which is indeed the case for standard algorithms as proven in Appendix D.

Lemma 16 (NE-variation stability)**.** *Suppose that x-player follows Algorithm 1 and y-player follows Algorithm 2. Then, the following inequalities hold simultaneously. In the meta-algorithm aspect, we have*

$$\sum_{t=1}^{T} \|p_t - \widehat{p}_{t+1}\|_2^2 \le \widetilde{\mathcal{O}}\left(\sqrt{(1+V_T)(1+P_T)} + P_T\right), \tag{37}$$

$$\sum_{t=1}^{T} \|q_t - \widehat{q}_{t+1}\|_2^2 \le \widetilde{\mathcal{O}}\left(\sqrt{(1+V_T)(1+P_T)} + P_T\right); \tag{38}$$

in the final decision aspect, we have

$$\sum_{t=2}^{T} \|x_t - x_{t-1}\|_1^2 \le \widetilde{\mathcal{O}}\left(\sqrt{(1+V_T)(1+P_T)} + P_T\right), \tag{39}$$

$$\sum_{t=2}^{T} \|y_t - y_{t-1}\|_1^2 \le \widetilde{\mathcal{O}}\left(\sqrt{(1+V_T)(1+P_T)} + P_T\right). \tag{40}$$

Proof. Let (x_t^*, y_t^*) denote the Nash equilibrium of the game matrix A_t at round t. Based on Lemma 15 with a choice of $\{u_t\}_{t=1}^T = \{x_t^*\}_{t=1}^T$ and $\{v_t\}_{t=1}^T = \{y_t^*\}_{t=1}^T$, we have

$$
\begin{aligned}
&\sum_{t=1}^T x_t^\top A_t y_t^* - \sum_{t=1}^T x_t^{*\top} A_t y_t \\
&= \sum_{t=1}^T x_t^\top A_t y_t^* - \sum_{t=1}^T x_t^\top A_t y_t + \sum_{t=1}^T x_t^\top A_t y_t - \sum_{t=1}^T x_t^{*\top} A_t y_t \\
&\le \widetilde{\mathcal{O}}\left(\frac{\alpha(1+P_T^x)}{\eta_i^x} + \frac{\alpha(1+P_T^y)}{\eta_j^y} + \beta(\eta_i^x+\eta_j^y)V_T\right) + \eta_i^x c\beta \sum_{t=2}^T \|y_t - y_{t-1}\|_1^2 + \eta_j^y c\beta \sum_{t=2}^T \|x_t - x_{t-1}\|_1^2 \\
&\quad + \left(\lambda - \frac{\gamma}{\eta_i^x}\right)\sum_{t=2}^T \|x_{t,i} - x_{t-1,i}\|_1^2 + \left(\lambda - \frac{\gamma}{\eta_j^y}\right)\sum_{t=2}^T \|y_{t,j} - y_{t-1,j}\|_1^2 \\
&\quad - L\sum_{t=1}^T (\|p_t - \widehat{p}_{t+1}\|_2^2 + \|p_t - \widehat{p}_t\|_2^2 + \|q_t - \widehat{q}_{t+1}\|_2^2 + \|q_t - \widehat{q}_t\|_2^2) \\
&\quad - \lambda \sum_{t=2}^T \sum_{i\in S_x} p_{t,i}\|x_{t,i} - x_{t-1,i}\|_1^2 - \lambda \sum_{t=2}^T \sum_{j\in S_y} q_{t,j}\|y_{t,j} - y_{t-1,j}\|_1^2 \\
&\le \widetilde{\mathcal{O}}\left(\frac{\alpha(1+P_T^x)}{\eta_i^x} + \frac{\alpha(1+P_T^y)}{\eta_j^y} + \beta(\eta_i^x+\eta_j^y)V_T\right) + 2\eta_i^x c\beta \sum_{t=2}^T \|q_t - q_{t-1}\|_1^2 + 2\eta_i^x c\beta \sum_{t=2}^T \sum_{j\in\mathcal{S}_y} \|y_{t,j} - y_{t-1,j}\|_1^2 \\
&\quad + 2\eta_j^y c\beta \sum_{t=2}^T \|p_t - p_{t-1}\|_1^2 + 2\eta_j^y c\beta \sum_{t=2}^T \sum_{i\in\mathcal{S}_x} \|x_{t,i} - x_{t-1,i}\|_1^2 \\
&\quad + \left(\lambda - \frac{\gamma}{\eta_i^x}\right)\sum_{t=2}^T \|x_{t,i} - x_{t-1,i}\|_1^2 + \left(\lambda - \frac{\gamma}{\eta_j^y}\right)\sum_{t=2}^T \|y_{t,j} - y_{t-1,j}\|_1^2 \\
&\quad - \frac{L}{2}\sum_{t=1}^T (\|p_t - \widehat{p}_{t+1}\|_2^2 + \|p_t - \widehat{p}_t\|_2^2 + \|q_t - \widehat{q}_{t+1}\|_2^2 + \|q_t - \widehat{q}_t\|_2^2) - \frac{L}{4}\sum_{t=2}^T (\|p_t - p_{t-1}\|_2^2 + \|q_t - q_{t-1}\|_2^2) \\
&\quad - \lambda \sum_{t=2}^T \sum_{i\in S_x} p_{t,i}\|x_{t,i} - x_{t-1,i}\|_1^2 - \lambda \sum_{t=2}^T \sum_{j\in S_y} q_{t,j}\|y_{t,j} - y_{t-1,j}\|_1^2 \\
&\le \widetilde{\mathcal{O}}\left(\frac{\alpha(1+P_T^x)}{\eta_i^x} + \frac{\alpha(1+P_T^y)}{\eta_j^y} + \beta(\eta_i^x+\eta_j^y)V_T\right) \\
&\quad + \left(2\eta_i^x c\beta - \frac{L}{4}\right)\sum_{t=2}^T \|q_t - q_{t-1}\|_1^2 + \left(2\eta_j^y c\beta - \frac{L}{4}\right)\sum_{t=2}^T \|p_t - p_{t-1}\|_1^2 \\
&\quad + (2\eta_i^x c\beta - \lambda)\sum_{t=2}^T \sum_{j\in\mathcal{S}_y} q_{t,j}\|y_{t,j} - y_{t-1,j}\|_1^2 + \big(2\eta_j^y c\beta - \lambda\big)\sum_{t=2}^T \sum_{i\in\mathcal{S}_x} p_{t,i}\|x_{t,i} - x_{t-1,i}\|_1^2 \\
&\quad + \left(\lambda - \frac{\gamma}{\eta_i^x}\right)\sum_{t=2}^T \|x_{t,i} - x_{t-1,i}\|_1^2 + \left(\lambda - \frac{\gamma}{\eta_j^y}\right)\sum_{t=2}^T \|y_{t,j} - y_{t-1,j}\|_1^2 \\
&\quad - \frac{L}{2}\sum_{t=1}^T \Big(\|p_t - \widehat{p}_{t+1}\|_2^2 + \|p_t - \widehat{p}_t\|_2^2 + \|q_t - \widehat{q}_{t+1}\|_2^2 + \|q_t - \widehat{q}_t\|_2^2\Big).
\end{aligned}
$$

According to the choice of $L = \max\left\{4, \sqrt{16c\beta}, \sqrt{\frac{8c\beta}{\gamma}}\right\} = \widetilde{\Theta}(1)$, $\lambda = \frac{\gamma L}{2}$, and $\eta_i^x, \eta_j^y \le \frac{1}{L}$, it can be verified that

$$
2\eta_i^x c\beta - \frac{L}{4} \le \frac{2c\beta}{L} - \frac{L}{4} \le -\frac{L}{8}; \qquad 2\eta_j^y c\beta - \frac{L}{4} \le \frac{2c\beta}{L} - \frac{L}{4} \le -\frac{L}{8};
$$

$$2\eta_i^x c\beta - \lambda = \frac{2c\beta}{L} - \frac{\gamma L}{2} \le -\frac{\gamma L}{4}; \qquad 2\eta_j^y c\beta - \lambda = \frac{2c\beta}{L} - \frac{\gamma L}{2} \le -\frac{\gamma L}{4};$$
$$\lambda - \frac{\gamma}{\eta_i^x} = \frac{\gamma L}{2} - \frac{\gamma}{\eta_i^x} \le \frac{\gamma}{2}\left(L - \frac{2}{\eta_i^x}\right) \le -\frac{\gamma}{2\eta_i^x}; \qquad \lambda - \frac{\gamma}{\eta_j^y} = \frac{\gamma L}{2} - \frac{\gamma}{\eta_j^y} \le \frac{\gamma}{2}\left(L - \frac{2}{\eta_j^y}\right) \le -\frac{\gamma}{2\eta_j^y}. \tag{41}$$

In addition, since (x_t^*, y_t^*) is the Nash equilibrium of A_t, it follows that

$$x_t^\top A_t y_t^* - x_t^{*\top} A_t y_t \ge x_t^{*\top} A_t y_t^* - x_t^{*\top} A_t y_t^* = 0.$$

Therefore, as α, β, $\gamma = \widetilde{\Theta}(1)$, we have the following inequalities simultaneously.

$$\sum_{t=2}^T \|p_t - p_{t-1}\|_1^2 \le \frac{8}{L} \cdot \widetilde{\mathcal{O}}\left(\frac{1+P_T^x}{\eta_i^x} + \frac{1+P_T^y}{\eta_j^y} + (\eta_i^x + \eta_j^y)V_T\right) \le \widetilde{\mathcal{O}}\left(\sqrt{(1+V_T)(1+P_T)} + P_T\right),$$

where the last inequality is because we pick η_i^x and η_j^y to be the one such that $\eta_i^x \in [\frac{1}{2}\eta_*^x, 2\eta_*^x]$, $\eta_j^y \in [\frac{1}{2}\eta_*^y, 2\eta_*^y]$ where $\eta_*^x = \min\left\{\sqrt{\frac{1+P_T^x}{1+V_T}}, \frac{1}{L}\right\}$ and $\eta_*^y = \min\left\{\sqrt{\frac{1+P_T^y}{1+V_T}}, \frac{1}{L}\right\}$. This is achievable based on the choice of our step size pool. Similarly, we have

$$\sum_{t=2}^T \|q_t - q_{t-1}\|_1^2 \le \widetilde{\mathcal{O}}\left(\sqrt{(1+V_T)(1+P_T)} + P_T\right),$$
$$\sum_{t=1}^T \|p_t - \widehat{p}_{t+1}\|_1^2 \le \widetilde{\mathcal{O}}\left(\sqrt{(1+V_T)(1+P_T)} + P_T\right),$$
$$\sum_{t=1}^T \|q_t - \widehat{q}_{t+1}\|_1^2 \le \widetilde{\mathcal{O}}\left(\sqrt{(1+V_T)(1+P_T)} + P_T\right),$$
$$\sum_{t=2}^T \sum_{i\in\mathcal{S}_x} p_{t,i}\|x_{t,i} - x_{t-1,i}\|_1^2 \le \widetilde{\mathcal{O}}\left(\sqrt{(1+V_T)(1+P_T)} + P_T\right),$$
$$\sum_{t=2}^T \sum_{j\in\mathcal{S}_y} q_{t,j}\|y_{t,j} - y_{t-1,j}\|_1^2 \le \widetilde{\mathcal{O}}\left(\sqrt{(1+V_T)(1+P_T)} + P_T\right).$$

In addition, note that

$$\begin{aligned}
&\sum_{t=2}^T \|x_t - x_{t-1}\|_1^2 \\
&= \sum_{t=2}^T \left\| \sum_{i\in\mathcal{S}_x} p_{t,i}x_{t,i} - \sum_{i\in\mathcal{S}_x} p_{t-1,i}x_{t-1,i} \right\|_1^2 \\
&= \sum_{t=2}^T \left\| \sum_{i\in\mathcal{S}_x} p_{t,i}(x_{t,i} - x_{t-1,i}) - \sum_{i\in\mathcal{S}_x} (p_{t-1,i} - p_{t,i})x_{t-1,i} \right\|_1^2 \\
&\le 2\sum_{t=2}^T \left\| \sum_{i\in\mathcal{S}_x} p_{t,i}(x_{t,i} - x_{t-1,i}) \right\|_1^2 + 2\sum_{t=2}^T \left\| \sum_{i\in\mathcal{S}_x} (p_{t,i} - p_{t-1,i})x_{t-1,i} \right\|_1^2 \\
&\le 2\sum_{t=2}^T \left(\sum_{i\in\mathcal{S}_x} p_{t,i} \|x_{t,i} - x_{t-1,i}\|_1 \right)^2 + 2\sum_{t=2}^T \left(\sum_{i\in\mathcal{S}_x} |p_{t,i} - p_{t-1,i}| \cdot \|x_{t-1,i}\|_1 \right)^2 \\
&\le 2\sum_{t=2}^T \left(\sum_{i\in\mathcal{S}_x} p_{t,i} \|x_{t,i} - x_{t-1,i}\|_1 \right)^2 + 2\sum_{t=2}^T \left(\sum_{i\in\mathcal{S}_x} |p_{t,i} - p_{t-1,i}| \cdot \|x_{t-1,i}\|_1 \right)^2 \\
&\le 2\sum_{t=2}^T \sum_{i\in\mathcal{S}_x} p_{t,i}\|x_{t,i} - x_{t-1,i}\|_1^2 + 2\sum_{t=2}^T \|p_t - p_{t-1}\|_1^2,
\end{aligned} \tag{42}$$

where the last inequality is by Cauchy-Schwarz inequality and $\|x_{t,i}\|_1 = 1$ for all $i \in \mathcal{S}_x$. Similarly, we have for y-player,

$$\sum_{t=2}^{T}\|y_t - y_{t-1}\|_1^2 \leq 2\sum_{t=2}^{T}\sum_{j\in\mathcal{S}_y} q_{t,j}\|x_{t,j} - x_{t-1,j}\|_1^2 + 2\sum_{t=2}^{T}\|q_t - q_{t-1}\|_1^2. \tag{43}$$

Based on the above results, we further have

$$\begin{aligned}
\sum_{t=2}^{T}\|x_t - x_{t-1}\|_1^2 \leq& 2\sum_{t=2}^{T}\|p_t - p_{t-1}\|_1^2 + 2\sum_{t=2}^{T}\sum_{i=1}^{N} p_{t,i}\|x_{t,i} - x_{t-1,i}\|_1^2 \leq \widetilde{\mathcal{O}}\left(\sqrt{(1+V_T)(1+P_T)} + P_T\right),\\
\sum_{t=2}^{T}\|y_t - y_{t-1}\|_1^2 \leq& 2\sum_{t=2}^{T}\|q_t - q_{t-1}\|_1^2 + 2\sum_{t=2}^{T}\sum_{i=1}^{N} q_{t,i}\|y_{t,i} - y_{t-1,i}\|_1^2 \leq \widetilde{\mathcal{O}}\left(\sqrt{(1+V_T)(1+P_T)} + P_T\right),
\end{aligned}$$

which completes the proof. □

The following lemma shows the relationship between dynamic NE-regret and the individual regret.

Lemma 17 (Dynamic-NE-regret-to-individual-regret conversation)**.** *For arbitrary sequences of $\{x_t\}_{t=1}^T$, $\{y_t\}_{t=1}^T$ and $\{A_t\}_{t=1}^T$, where $x_t \in \Delta_m$, $y_t \in \Delta_n$, and $A_t \in \mathbb{R}^{m\times n}$, $\forall t \in [T]$, we have*

$$\left|\sum_{t=1}^{T} x_t^\top A_t y_t - \sum_{t=1}^{T}\min_x\max_y x^\top A_t y\right| \leq \max\left\{\sum_{t=1}^{T} x_t^\top A_t y_t - \sum_{t=1}^{T} x^{*\top} A_t y_t, \sum_{t=1}^{T} x_t^\top A_t y^* - \sum_{t=1}^{T} x_t^\top A_t y_t\right\} + 2W_T, \tag{44}$$

where $W_T = \sum_{t=1}^{T}\|A_t - \bar{A}\|_\infty$ is the variance of the game matrices with $\bar{A} = \frac{1}{T}\sum_{t=1}^{T} A_t$ being the average game matrix and (x^*, y^*) is a pair of Nash equilibrium of $\bar{A}$. In the special case where $A_t = A$ for all $t \in [T]$, we have dynamic NE-regret bounded by the maximum of the two individual regrets as $W_T = 0$.

Proof. Suppose that $\sum_{t=1}^{T} x_t^\top A_t y_t - \sum_{t=1}^{T}\min_x\max_y x^\top A_t y \geq 0$, then the dynamic NE-regret can be upper bounded as

$$\begin{aligned}
&\left|\sum_{t=1}^{T} x_t^\top A_t y_t - \sum_{t=1}^{T}\min_x\max_y x^\top A_t y\right| && \text{(let } (x_t^*, y_t^*) \text{ be the Nash equilibrium of } A_t)\\
=&\sum_{t=1}^{T} x_t^\top A_t y_t - \sum_{t=1}^{T} x_t^{*\top} A_t y_t^* && \text{(let } (x^*, y^*) \text{ be the Nash equilibrium of } \bar{A})\\
\leq&\sum_{t=1}^{T} x_t^\top A_t y_t - \sum_{t=1}^{T} x_t^{*\top} A_t y^* && \text{(changing } y_t^* \text{ to } y^* \text{ decreases the game value w.r.t. } A_t)\\
\leq&\sum_{t=1}^{T} x_t^\top A_t y_t - \sum_{t=1}^{T} x_t^{*\top} \bar{A} y^* + W_T && \text{(shifting the payoff matrix from } A_t \text{ to } \bar{A})\\
\leq&\sum_{t=1}^{T} x_t^\top A_t y_t - \sum_{t=1}^{T} x^{*\top} \bar{A} y^* + W_T && \text{(changing } x_t^* \text{ to } x^* \text{ decreases the game value w.r.t. } \bar{A})\\
\leq&\sum_{t=1}^{T} x_t^\top A_t y_t - \sum_{t=1}^{T} x^{*\top} \bar{A} y_t + W_T && \text{(changing } y^* \text{ to } y_t \text{ decreases the game value w.r.t. } \bar{A})\\
\leq&\sum_{t=1}^{T} x_t^\top A_t y_t - \sum_{t=1}^{T} x^{*\top} A_t y_t + 2W_T. && \text{(shifting the payoff matrix from } \bar{A} \text{ to } A_t)
\end{aligned}$$

Similarly, when $\sum_{t=1}^{T} x_t^\top A_t y_t - \sum_{t=1}^{T}\min_x\max_y x^\top A_t y \leq 0$, we can verify that

$$\left|\sum_{t=1}^{T} x_t^\top A_t y_t - \sum_{t=1}^{T}\min_x\max_y x^\top A_t y\right| \qquad \text{(let } (x_t^*, y_t^*) \text{ be the Nash equilibrium of } A_t)$$

$$
\begin{aligned}
&= \sum_{t=1}^{T} x_t^{*\top} A_t y_t^* - \sum_{t=1}^{T} x_t^\top A_t y_t && \text{(let } (x^*, y^*) \text{ be the Nash equilibrium of } \bar{A}) \\
&\le \sum_{t=1}^{T} x^{*\top} A_t y_t^* - \sum_{t=1}^{T} x_t^\top A_t y_t && \text{(changing } x_t^* \text{ to } x^* \text{ increases the game value w.r.t. } A_t) \\
&\le \sum_{t=1}^{T} x^{*\top} \bar{A} y_t^* - \sum_{t=1}^{T} x_t^\top A_t y_t + W_T && \text{(shifting the payoff matrix from } A_t \text{ to } \bar{A}) \\
&\le \sum_{t=1}^{T} x^{*\top} \bar{A} y^* - \sum_{t=1}^{T} x_t^\top A_t y_t + W_T && \text{(changing } y_t^* \text{ to } y^* \text{ increases the game value w.r.t. } \bar{A}) \\
&\le \sum_{t=1}^{T} x_t^\top \bar{A} y^* - \sum_{t=1}^{T} x_t^\top A_t y_t + W_T && \text{(changing } x^* \text{ to } x_t \text{ increases the game value w.r.t. } \bar{A}) \\
&\le \sum_{t=1}^{T} x_t^\top A_t y^* - \sum_{t=1}^{T} x_t^\top A_t y_t + 2W_T. && \text{(shifting the payoff matrix from } \bar{A} \text{ to } A_t)
\end{aligned}
$$

Combining the two cases yields the desired result. □

Next, we present the following stability lemma in terms of the non-stationarity measure W_T (that is, the payoff variance). We give the stability upper bounds from the aspects of meta-algorithm and final decisions. Again, we assume that the DRVU parameters (α, β, γ) are all $\widetilde{\Theta}(1)$.

Lemma 18 (Payoff-variance stability)**.** *Suppose that x-player follows Algorithm 1 and y-player follows Algorithm 2. Then, the following inequalities hold simultaneously. In the meta-algorithm aspect, we have*

$$
\sum_{t=1}^{T} \|p_t - \widehat{p}_{t+1}\|_2^2 \le \widetilde{\mathcal{O}}\left(1 + W_T\right), \quad \sum_{t=1}^{T} \|q_t - \widehat{q}_{t+1}\|_2^2 \le \widetilde{\mathcal{O}}\left(1 + W_T\right); \tag{45}
$$

in the final decision aspect, we have

$$
\sum_{t=2}^{T} \|x_t - x_{t-1}\|_1^2 \le \widetilde{\mathcal{O}}\left(1 + W_T\right), \quad \sum_{t=2}^{T} \|y_t - y_{t-1}\|_1^2 \le \widetilde{\mathcal{O}}\left(1 + W_T\right). \tag{46}
$$

Proof. Let $\bar{A} = \frac{1}{T}\sum_{t=1}^{T} A_t$ denote the average game matrix, and let (x^*, y^*) be the Nash equilibrium of $\bar{A}$. Then according to the saddle point property of (x^*, y^*), we have for any $x_t \in \Delta_m$ and $y_t \in \Delta_n$, $x_t^\top \bar{A} y^* - x^{*\top} \bar{A} y_t \ge 0$. Therefore, based on Lemma 15 with $u_t = y^*$ and $v_t = x^*$ for all $t \in [T]$, we have

$$
\begin{aligned}
0 &\le \sum_{t=1}^{T} x_t^\top \bar{A} y^* - \sum_{t=1}^{T} x^{*\top} \bar{A} y_t \\
&\le \sum_{t=1}^{T} x_t^\top A_t y^* - \sum_{t=1}^{T} x_t^\top A_t y_t + \sum_{t=1}^{T} x_t^\top A_t y_t - \sum_{t=1}^{T} x^{*\top} A_t y_t + 2W_T \\
&\le \widetilde{\mathcal{O}}\left(\frac{\alpha}{\eta_i^x} + \frac{\alpha}{\eta_j^y} + \beta(\eta_i^x + \eta_j^y) V_T\right) + \eta_i^x c\beta \sum_{t=2}^{T} \|y_t - y_{t-1}\|_1^2 + \eta_j^y c\beta \sum_{t=2}^{T} \|x_t - x_{t-1}\|_1^2 \\
&\quad + \left(\lambda - \frac{\gamma}{\eta_i^x}\right) \sum_{t=2}^{T} \|x_{t,i} - x_{t-1,i}\|_1^2 + \left(\lambda - \frac{\gamma}{\eta_j^y}\right) \sum_{t=2}^{T} \|y_{t,j} - y_{t-1,j}\|_1^2 \\
&\quad - L \sum_{t=1}^{T} (\|p_t - \widehat{p}_{t+1}\|_2^2 + \|p_t - \widehat{p}_t\|_2^2 + \|q_t - \widehat{q}_{t+1}\|_2^2 + \|q_t - \widehat{q}_t\|_2^2)
\end{aligned}
$$

$$
\begin{aligned}
&- \lambda \sum_{t=2}^{T} \sum_{i \in S_x} p_{t,i} \|x_{t,i} - x_{t-1,i}\|_1^2 - \lambda \sum_{t=2}^{T} \sum_{j \in S_y} q_{t,j} \|y_{t,j} - y_{t-1,j}\|_1^2 + 2W_T \\
\leq\ & \widetilde{\mathcal{O}}\left(\frac{\alpha}{\eta_i^x} + \frac{\alpha}{\eta_j^y} + \beta(\eta_i^x + \eta_j^y) V_T \right) + 2\eta_i^x c\beta \sum_{t=2}^{T} \|q_t - q_{t-1}\|_1^2 + 2\eta_i^x c\beta \sum_{t=2}^{T} \sum_{j \in \mathcal{S}_y} \|y_{t,j} - y_{t-1,j}\|_1^2 \\
&+ 2\eta_j^y c\beta \sum_{t=2}^{T} \|p_t - p_{t-1}\|_1^2 + 2\eta_j^y c\beta \sum_{t=2}^{T} \sum_{i \in \mathcal{S}_x} \|x_{t,i} - x_{t-1,i}\|_1^2 \\
&+ \left(\lambda - \frac{\gamma}{\eta_i^x} \right) \sum_{t=2}^{T} \|x_{t,i} - x_{t-1,i}\|_1^2 + \left(\lambda - \frac{\gamma}{\eta_j^y} \right) \sum_{t=2}^{T} \|y_{t,j} - y_{t-1,j}\|_1^2 \\
&- \frac{L}{2} \sum_{t=1}^{T} (\|p_t - \widehat{p}_{t+1}\|_2^2 + \|p_t - \widehat{p}_t\|_2^2 + \|q_t - \widehat{q}_{t+1}\|_2^2 + \|q_t - \widehat{q}_t\|_2^2) - \frac{L}{4} \sum_{t=2}^{T} (\|p_t - p_{t-1}\|_2^2 + \|q_t - q_{t-1}\|_2^2) \\
&- \lambda \sum_{t=2}^{T} \sum_{i \in S_x} p_{t,i} \|x_{t,i} - x_{t-1,i}\|_1^2 - \lambda \sum_{t=2}^{T} \sum_{j \in S_y} q_{t,j} \|y_{t,j} - y_{t-1,j}\|_1^2 + 2W_T \\
\leq\ & \widetilde{\mathcal{O}}\left(\frac{\alpha}{\eta_i^x} + \frac{\alpha}{\eta_j^y} + \beta(\eta_i^x + \eta_j^y) V_T \right) + \left(2\eta_i^x c\beta - \frac{L}{4} \right) \sum_{t=2}^{T} \|q_t - q_{t-1}\|_1^2 + \left(2\eta_j^y c\beta - \frac{L}{4} \right) \sum_{t=2}^{T} \|p_t - p_{t-1}\|_1^2 \\
&+ (2\eta_i^x c\beta - \lambda) \sum_{t=2}^{T} \sum_{j \in \mathcal{S}_y} q_{t,j} \|y_{t,j} - y_{t-1,j}\|_1^2 + \left(2\eta_j^y c\beta - \lambda \right) \sum_{t=2}^{T} \sum_{i \in \mathcal{S}_x} p_{t,i} \|x_{t,i} - x_{t-1,i}\|_1^2 \\
&+ \left(\lambda - \frac{\gamma}{\eta_i^x} \right) \sum_{t=2}^{T} \|x_{t,i} - x_{t-1,i}\|_1^2 + \left(\lambda - \frac{\gamma}{\eta_j^y} \right) \sum_{t=2}^{T} \|y_{t,j} - y_{t-1,j}\|_1^2 \\
&- \frac{L}{2} \sum_{t=1}^{T} (\|p_t - \widehat{p}_{t+1}\|_2^2 + \|p_t - \widehat{p}_t\|_2^2 + \|q_t - \widehat{q}_{t+1}\|_2^2 + \|q_t - \widehat{q}_t\|_2^2) + 2W_T.
\end{aligned}
$$

Based on the choice of L and λ, we can again verify the condition of Eq. (41), which leads to the following inequalities:

$$
\sum_{t=2}^{T} \|p_t - p_{t-1}\|_1^2 \leq \frac{8}{L} \cdot \widetilde{\mathcal{O}}\left(\frac{1}{\eta_i^x} + \frac{1}{\eta_j^y} + (\eta_i^x + \eta_j^y) V_T + W_T \right) \leq \widetilde{\mathcal{O}}\left(\sqrt{1 + V_T} + W_T \right) \leq \widetilde{\mathcal{O}}\left(1 + W_T \right),
$$

where the second last inequality is because we pick η_i^x and η_j^y to be the one such that $\eta_i^x \in [\frac{1}{2}\eta_*^x, 2\eta_*^x]$, $\eta_j^y \in [\frac{1}{2}\eta_*^y, 2\eta_*^y]$ where $\eta_*^x = \min\left\{ \sqrt{\frac{1}{1+V_T}}, \frac{1}{L} \right\}$ and $\eta_*^y = \min\left\{ \sqrt{\frac{1}{1+V_T}}, \frac{1}{L} \right\}$. This is achievable based on the choice of our step size pool. The last inequality holds because of AM-GM inequality and $V_T \leq \mathcal{O}(W_T)$. Similarly, we have

$$
\begin{aligned}
&\sum_{t=2}^{T} \|q_t - q_{t-1}\|_1^2 \leq \widetilde{\mathcal{O}}\left(1 + W_T \right), \\
&\sum_{t=1}^{T} \|p_t - \widehat{p}_{t+1}\|_1^2 \leq \widetilde{\mathcal{O}}\left(1 + W_T \right), \\
&\sum_{t=1}^{T} \|q_t - \widehat{q}_{t+1}\|_1^2 \leq \widetilde{\mathcal{O}}\left(1 + W_T \right), \\
&\sum_{t=2}^{T} \sum_{i \in \mathcal{S}_x} p_{t,i} \|x_{t,i} - x_{t-1,i}\|_1^2 \leq \widetilde{\mathcal{O}}\left(1 + W_T \right), \\
&\sum_{t=2}^{T} \sum_{j \in \mathcal{S}_y} q_{t,j} \|y_{t,j} - y_{t-1,j}\|_1^2 \leq \widetilde{\mathcal{O}}\left(1 + W_T \right).
\end{aligned}
$$

Based on the above results, we further have

$$\sum_{t=2}^{T}\|x_t - x_{t-1}\|_1^2 \leq 2\sum_{t=2}^{T}\|p_t - p_{t-1}\|_1^2 + 2\sum_{t=2}^{T}\sum_{i=1}^{N} p_{t,i}\|x_{t,i} - x_{t-1,i}\|_1^2 \leq \widetilde{\mathcal{O}}\left(1 + W_T\right),$$
$$\sum_{t=2}^{T}\|y_t - y_{t-1}\|_1^2 \leq 2\sum_{t=2}^{T}\|q_t - q_{t-1}\|_1^2 + 2\sum_{t=2}^{T}\sum_{i=1}^{N} q_{t,i}\|y_{t,i} - y_{t-1,i}\|_1^2 \leq \widetilde{\mathcal{O}}\left(1 + W_T\right),$$

which completes the proof. □

Building upon the stability of the decisions of both x-player and y-player proven in Lemma 16 and Lemma 18, we further show the variation of the (gradient) feedback received by both x-player and y-player.

Lemma 19. *Suppose x-player follows Algorithm 1 and y-player follows Algorithm 2. Then, the gradient variation can be bounded as follows:*

$$\begin{aligned}\sum_{t=2}^{T}\|A_t y_t - A_{t-1}y_{t-1}\|_\infty^2 &\leq \widetilde{\mathcal{O}}\Big(1 + V_T + \min\{P_T, W_T\}\Big),\\ \sum_{t=2}^{T}\|x_t^\top A_t - x_{t-1}^\top A_{t-1}\|_\infty^2 &\leq \widetilde{\mathcal{O}}\Big(1 + V_T + \min\{P_T, W_T\}\Big).\end{aligned} \tag{47}$$

Proof. The gradient variation of the x-player can be upper bounded as follows:

$$\begin{aligned}\sum_{t=2}^{T}\|A_t y_t - A_{t-1}y_{t-1}\|_\infty^2 &\leq 2\sum_{t=2}^{T}\|A_t y_t - A_{t-1}y_t\|_\infty^2 + 2\sum_{t=2}^{T}\|A_{t-1}y_t - A_{t-1}y_{t-1}\|_\infty^2\\ &\leq 2\sum_{t=2}^{T}\|A_t - A_{t-1}\|_\infty^2 + \mathcal{O}\left(\sum_{t=2}^{T}\|y_t - y_{t-1}\|_1^2\right)\\ &\leq \mathcal{O}(V_T) + \widetilde{\mathcal{O}}\left(\min\{\sqrt{1 + V_T + P_T} + P_T, 1 + W_T\}\right)\\ &\leq \widetilde{\mathcal{O}}\left(\min\{1 + V_T + P_T, 1 + V_T + W_T\}\right)\end{aligned}$$

where the second last step holds by Lemma 16 and Lemma 18, and the last step makes use of AM-GM inequality. A similar argument can be applied to upper bound $\sum_{t=2}^{T}\|x_t^\top A_t - x_{t-1}^\top A_{t-1}\|_\infty^2$. This ends the proof. □

The following lemma establishes a general result to relate the function-value difference between the sequence of piecewise minimizers and the sequence of each-round minimizers.

Lemma 20. *Let $A_t \in \mathbb{R}^{m\times n}$ and $y_t \in \Delta_n$ for all $t \in [T]$ with $\max_{t\in[T]}\|A_t\|_\infty \leq 1$. Let $\mathcal{I}_1, \mathcal{I}_2, \ldots, \mathcal{I}_K$ be an even partition of the total horizon $[T]$ with $|\mathcal{I}_k| = \Delta$ for $k = 1, \ldots, K$ (for simplicity, suppose the time horizon T is divisible by epoch length Δ). Denote $\bar{x}_t^* = \operatorname{argmin}_{x\in\Delta_m} x^\top A_t y_t$ for any $t \in [T]$ and denote $u_t \triangleq \operatorname{argmin}_{x\in\Delta_m}\sum_{\tau\in\mathcal{I}_k} x^\top A_\tau y_\tau$ for any $t \in \mathcal{I}_k$. Then, we have*

$$\sum_{t=1}^{T} u_t^\top A_t y_t - \sum_{t=1}^{T}\bar{x}_t^{*\top} A_t y_t \leq 2\Delta\sum_{t=2}^{T}\|A_t y_t - A_{t-1}y_{t-1}\|_\infty. \tag{48}$$

Proof. The proof follows the analysis of function-variation type worst-case dynamic regret (Zhang et al., 2020). For convenience, we introduce the notation $f_t(x) \triangleq x^\top A_t y_t$ to denote the each-round online function of x-player. Then,

$$\begin{aligned}&\sum_{i=1}^{T} f_t(u_t) - \sum_{i=1}^{T} f_t(\bar{x}_t^*) = \sum_{k=1}^{K}\sum_{t\in\mathcal{I}_k}\left(f_t(u_t) - f_t(\bar{x}_t^*)\right)\\ &\leq 2\Delta\cdot\sum_{k=1}^{K}\sum_{t\in\mathcal{I}_k}\|A_t y_t - A_{t-1}y_{t-1}\|_\infty = 2\Delta\sum_{t=2}^{T}\|A_t y_t - A_{t-1}y_{t-1}\|_\infty,\end{aligned}$$

where the inequality holds because of the setting of $|\mathcal{I}_k| = \Delta$ and the following fact about the instantaneous quantity:

$$\begin{aligned}
f_t(u_t) - f_t(\bar{x}_t^*) &\le f_t(\bar{x}_{t_1}^*) - f_t(\bar{x}_t^*) && \text{(denote by } t_1 \text{ the starting time stamp of } \mathcal{I}_k\text{)} \\
&= f_t(\bar{x}_{t_1}^*) - f_{t_1}(\bar{x}_{t_1}^*) + f_{t_1}(\bar{x}_{t_1}^*) - f_t(\bar{x}_t^*) \\
&\le f_t(\bar{x}_{t_1}^*) - f_{t_1}(\bar{x}_{t_1}^*) + f_{t_1}(\bar{x}_t^*) - f_t(\bar{x}_t^*) && \text{(by optimality of } \bar{x}_{t_1}^*\text{)} \\
&\le 2 \sum_{t \in \mathcal{I}_k} \sup_x |f_t(x) - f_{t-1}(x)| \\
&= 2 \sum_{t \in \mathcal{I}_k} \sup_x \left| x^\top (A_t y_t - A_{t-1} y_{t-1}) \right| \\
&\le 2 \sum_{t \in \mathcal{I}_k} \|A_t y_t - A_{t-1} y_{t-1}\|_\infty.
\end{aligned}$$

Hence we finish the proof. □

G. Technical Lemmas

Lemma 21 (Bregman proximal inequality (Chen & Teboulle, 1993, Lemma 3.2)). *Let $\mathcal{X}$ be a convex set in a Banach space. Let $f : \mathcal{X} \mapsto \mathbb{R}$ be a closed proper convex function on $\mathcal{X}$. Given a convex regularizer $\psi : \mathcal{X} \mapsto \mathbb{R}$, we denote its induced Bregman divergence by $D_\psi(\cdot, \cdot)$. Then, any update of the form $x_k = \operatorname{argmin}_{x \in \mathcal{X}} \{f(x) + D_\psi(x, x_{k-1})\}$ satisfies the following inequality for any $u \in \mathcal{X}$,*

$$f(x_k) - f(u) \le D_\psi(u, x_{k-1}) - D_\psi(u, x_k) - D_\psi(x_k, x_{k-1}). \tag{49}$$

Lemma 22 (stability lemma (Chiang et al., 2012, Proposition 7)). *Let $x_* = \operatorname{argmin}_{x \in \mathcal{X}} \langle a, x \rangle + D_\psi(x, c)$ and $x'_* = \operatorname{argmin}_{x \in \mathcal{X}} \langle a', x \rangle + D_\psi(x, c)$. When the regularizer $\psi : \mathcal{X} \mapsto \mathbb{R}$ is a 1-strongly convex function with respect to the norm $\|\cdot\|$, we have $\|x_* - x'_*\| \le \|(\nabla\psi(c) - a) - (\nabla\psi(c) - a')\|_* = \|a - a'\|_*$.*

Lemma 23 (variant of self-confident tuning (Pogodin & Lattimore, 2019, Lemma 4.8)). *Let $a_1, a_2, \ldots, a_T$ be non-negative real numbers. Then*

$$\sum_{t=1}^{T} \frac{a_t}{\sqrt{1 + \sum_{s=1}^{t-1} a_s}} \le 4 \sqrt{1 + \sum_{t=1}^{T} a_t} + \max_{t \in [T]} a_t.$$