Boosting Perturbed Gradient Ascent for Last-Iterate Convergence in Games

Anonymous Author(s)
Affiliation
Address
email

Abstract

This paper introduces a payoff perturbation technique, introducing a strong convexity to players' payoff functions in games. This technique is specifically designed for first-order methods to achieve last-iterate convergence in games where the gradient of the payoff functions is monotone in the strategy profile space, potentially containing additive noise. Although perturbation is known to facilitate the convergence of learning algorithms, the magnitude of perturbation requires careful adjustment to ensure last-iterate convergence. Previous studies have proposed a scheme in which the magnitude is determined by the distance from an anchoring or reference strategy, which is periodically re-initialized. In response, this paper proposes Gradient Ascent with Boosting Payoff Perturbation, which incorporates a novel perturbation into the underlying payoff function, maintaining the periodically re-initializing anchoring strategy scheme. This innovation empowers us to provide faster last-iterate convergence rates against the existing payoff perturbed algorithms, even in the presence of additive noise.

1 Introduction

This study considers online learning in monotone games, where the gradient of the payoff function is monotone in the strategy profile space. Monotone games encompassed diverse well-studied games as special instances, such as concave-convex games, zero-sum polymatrix games [Cai and Daskalakis, 2011, Cai et al., 2016], λ-cocoercive games [Lin et al., 2020], and Cournot competition [Bravo et al., 2018]. Due to their wide-ranging applications, there has been growing interest in developing learning algorithms to compute Nash equilibria in monotone games.

Typical learning algorithms such as Gradient Ascent [Zinkevich, 2003] and Multiplicative Weights Update [Bailey and Piliouras, 2018] have been extensively studied and shown to converge to equilibria in an average-iterate sense, which is termed *average-iterate convergence*. However, averaging the strategies can be undesirable because it can lead to additional memory or computational costs in the context of training Generative Adversarial Networks [Goodfellow et al., 2014] and preference-based fine-tuning of large language models [Munos et al., 2023, Swamy et al., 2024]. In contrast, *last-iterate convergence*, in which the updated strategy profile itself converges to a Nash equilibrium, has emerged as a stronger notion than average-iterate convergence.

Payoff-perturbed algorithms have recently been regaining attention in this context [Sokota et al., 2023, Liu et al., 2023]. Payoff perturbation is a classical technique, e.g., [Facchinei and Pang, 2003] and introduces a strongly convex penalty to the players' payoff functions to stabilize learning, which leads to convergence to approximate equilibria, not only in the *full feedback* setting where the perfect gradient vector of the payoff function can be used to update strategies, but also in the *noisy feedback* setting where the gradient vector is contaminated by noise.

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

However, to ensure convergence toward a Nash equilibrium of the underlying game, the magnitude of perturbation requires careful adjustment. As a remedy, it is adjusted by the distance from an anchoring or reference strategy. Koshal et al. [2010] and Tatarenko and Kamgarpour [2019] simply decay the magnitude in each iteration, and their methods asymptotically converge, since the perturbed function gradually loses strong convexity. In response to this, recent studies [Perolat et al., 2021, Abe et al., 2023, 2024] re-initialize the anchoring strategies periodically, or in a predefined interval, so that they keep the perturbed function strongly convex and achieve non-asymptotic convergence.

We should also mention the *optimistic* family of learning algorithms, which incorporates recency bias and exhibits last-iterate convergence [Daskalakis et al., 2018, Daskalakis and Panageas, 2019, Mertikopoulos et al., 2019, Wei et al., 2021]. Unfortunately, the property has mainly been proven in the full feedback setting. Although it might empirically work with noisy feedback, the convergence is slower, as demonstrated in Section 6. The fast convergence in the noisy feedback setting is another reason why payoff-perturbed algorithms have been gaining renewed interest.

This paper, in particular, focuses on *Adaptively Perturbed Mirror Descent* (APMD) [Abe et al., 2024], which achieves $\tilde{\mathcal{O}}(1/\sqrt{T})$[1] and $\tilde{\mathcal{O}}(1/T^{\frac{1}{10}})$ last-iterate convergence rates in the full/noisy feedback setting, respectively. The motivation of this study lies in improving the convergence rates of APMD. We propose an elegant one-line modification of APMD, which effectively accelerates convergence. In fact, we just add the difference between the current anchoring strategy and the initial anchoring strategy to the payoff perturbation function in APMD.

Our contributions are manifold. Firstly, we propose a novel payoff-perturbed learning algorithm named *Gradient Ascent with Boosting Payoff Perturbation* (GABP). This method incorporates a unique perturbation payoff function, enabling it to achieve faster convergence rates than APMD. Subsequently, we prove that GABP exhibits accelerated $\tilde{\mathcal{O}}(1/T)$ and $\tilde{\mathcal{O}}(1/T^{\frac{1}{7}})$ last-iterate convergence rates to a Nash equilibrium with full and noisy feedback, respectively. We further show that each player's individual regret is at most $\mathcal{O}\left((\ln T)^2\right)$ in the full feedback setting, provided all players play according to GABP. Finally, through our experiments, we demonstrate the competitive or superior performance of GABP over Optimistic Gradient Ascent [Daskalakis et al., 2018, Wei et al., 2021] and APMD in concave-convex games, irrespective of the presence of noise.

2 Preliminaries

Monotone games. In this study, we focus on a continuous multi-player game, which is denoted as $\left([N], (\mathcal{X}_i)_{i\in[N]}, (v_i)_{i\in[N]}\right)$. $[N] = \{1, 2, \cdots, N\}$ denotes the set of N players. Each player $i \in [N]$ chooses a *strategy* π_i from a d_i-dimensional compact convex strategy space $\mathcal{X}_i$, and we write $\mathcal{X} = \prod_{i\in[N]} \mathcal{X}_i$. Each player i aims to maximize her payoff function $v_i : \mathcal{X} \to \mathbb{R}$, which is differentiable on $\mathcal{X}$. We denote $\pi_{-i} \in \prod_{j\neq i} \mathcal{X}_j$ as the strategies of all players except player i, and $\pi = (\pi_i)_{i\in[N]} \in \mathcal{X}$ as the *strategy profile*. This paper particularly studies learning in *smooth monotone games*, where the gradient operator $V(\cdot) = (\nabla_{\pi_i} v_i(\cdot))_{i\in[N]}$ of the payoff functions is monotone: $\forall \pi, \pi' \in \mathcal{X}$,

$$\langle V(\pi) - V(\pi'), \pi - \pi' \rangle \leq 0, \tag{1}$$

and L-Lipschitz for $L > 0$

$$\|V(\pi) - V(\pi')\| \leq L \|\pi - \pi'\|, \tag{2}$$

where $\|\cdot\|$ denotes the ℓ_2-norm.

Many common and well-studied games, such as concave-convex games, zero-sum polymatrix games [Cai et al., 2016], λ-cocoercive games [Lin et al., 2020], and Cournot competition [Bravo et al., 2018], are included in the class of monotone games.

Example 2.1 (Concave-Convex Games)**.** Consider a game defined by $(\{1, 2\}, (\mathcal{X}_1, \mathcal{X}_2), (v, -v))$, where $v : \mathcal{X}_1 \times \mathcal{X}_2 \to \mathbb{R}$. In this game, player 1 wishes to maximize v, while player 2 aims to minimize v. If v is concave in $x_1 \in \mathcal{X}_1$ and convex in $x_2 \in \mathcal{X}_2$, the game is called a concave-convex game or minimax optimization problem, and it is not hard to see that this game is a special case of monotone games.

[1] We use $\tilde{\mathcal{O}}$ to denote a Landau notation that disregards a polylogarithmic factor.

Nash equilibrium and gap function. A *Nash equilibrium* [Nash, 1951] is a widely used solution concept for a game, which is a strategy profile where no player can gain by changing her own strategy. Formally, a strategy profile $\pi^* \in \mathcal{X}$ is called a Nash equilibrium, if and only if π^* satisfies the following condition:

$$\forall i \in [N], \forall \pi_i \in \mathcal{X}_i,\ v_i(\pi_i^*, \pi_{-i}^*) \geq v_i(\pi_i, \pi_{-i}^*).$$

We define the set of all Nash equilibria to be Π^*. It has been shown that there exists at least one Nash equilibrium [Debreu, 1952] for any smooth monotone games.

To quantify the proximity to Nash equilibrium for a given strategy profile $\pi \in \mathcal{X}$, we use the *gap function*, which is defined as:

$$\mathrm{GAP}(\pi) := \max_{\tilde{\pi} \in \mathcal{X}} \langle V(\pi), \tilde{\pi} - \pi \rangle .$$

Additionally, we use another measure of proximity to Nash equilibrium, referred to as the *tangent residual*. This measure is defined as:

$$r^{\tan}(\pi) := \min_{a \in N_{\mathcal{X}}(\pi)} \|-V(\pi) + a\| ,$$

where $N_{\mathcal{X}}(\pi) = \{(a_i)_{i\in[N]} \in \prod_{i=1}^N \mathbb{R}^{d_i} \mid \sum_{i=1}^N \langle a_i, \pi_i' - \pi_i \rangle \leq 0,\ \forall \pi' \in \mathcal{X}\}$ is the normal cone of $\pi \in \mathcal{X}$. It is easy to see that $\mathrm{GAP}(\pi) \geq 0$ (resp. $r^{\tan}(\pi) \geq 0$) for any $\pi \in \mathcal{X}$, and the equality holds if and only if π is a Nash equilibrium. Defining $D := \sup_{\pi, \pi' \in \mathcal{X}} \|\pi - \pi'\|$ as the diameter of $\mathcal{X}$, the gap function for any given strategy profile $\pi \in \mathcal{X}$ is upper bounded by its tangent residual.

Lemma 2.2 (Lemma 2 of Cai et al. [2022a])**.** *For any $\pi \in \mathcal{X}$, we have:*

$$\mathrm{GAP}(\pi) \leq D \cdot r^{\tan}(\pi).$$

The gap function and the tangent residual are standard measures of proximity to Nash equilibrium; e.g., it has been used in Cai and Zheng [2023], Abe et al. [2024].

Problem setting. This study focuses on the online learning setting in which the following process repeats from iterations $t = 1$ to T: (i) Each player $i \in [N]$ chooses her strategy $\pi_i^t \in \mathcal{X}_i$, based on previously observed feedback; (ii) Each player i receives the (noisy) gradient vector $\widehat{\nabla}_{\pi_i} v_i(\pi^t)$ as feedback. This study examines two feedback models: *full feedback* and *noisy feedback*. In the full feedback setting, each player observes the perfect gradient vector $\widehat{\nabla}_{\pi_i} v_i(\pi^t) = \nabla_{\pi_i} v_i(\pi^t)$. In the noisy feedback setting, each player's gradient feedback $\nabla_{\pi_i} v_i(\pi^t)$ is contaminated by an additive noise vector ξ_i^t, i.e., $\widehat{\nabla}_{\pi_i} v_i(\pi^t) = \nabla_{\pi_i} v_i(\pi^t) + \xi_i^t$, where $\xi_i^t \in \mathbb{R}^{d_i}$. Throughout the paper, we assume that ξ_i^t is the zero-mean and bounded-variance noise vector at each iteration t.

Adaptively perturbed Mirror Descent. To facilitate the convergence in the online learning setting, recent studies have utilized a *payoff perturbation* technique, where payoff functions are perturbed by strongly convex functions [Sokota et al., 2023, Liu et al., 2023, Abe et al., 2022]. However, while the addition of these strongly convex functions leads learning algorithms to converge to a stationary point, this stationary point may be significantly distant from a Nash equilibrium. Therefore, the magnitude of perturbation requires careful adjustment. Perolat et al. [2021], Abe et al. [2023, 2024] have introduced a scheme in which the magnitude is determined by the distance (or divergence function) from an anchoring strategy σ_i, which is periodically re-initialized. Specifically, Adaptively Perturbed Mirror Descent (APMD) [Abe et al., 2024] perturbs each player's payoff function by a strongly convex divergence function $G(\pi_i, \sigma_i) : \mathcal{X}_i \times \mathcal{X}_i \to [0, \infty)$, where the anchoring strategy σ_i is periodically replaced by the current strategy π_i^t every predefined iterations T_σ.

Let us define $\sigma_i^{k(t)}$ as the anchoring strategy after $k(t)$ updates. Since σ_i is overwritten every T_σ iterations, we can write $k(t) = \lfloor (t-1)/T_\sigma \rfloor + 1$ and $\sigma_i^{k(t)} = \pi_i^{T_\sigma(k(t)-1)+1}$. Except for the payoff perturbation and the update of the anchor strategy, APMD updates each player i's strategy in the same way as standard Mirror Descent algorithms:

$$\pi_i^{t+1} = \arg\max_{x \in \mathcal{X}_i} \left\{ \eta_t \left\langle \widehat{\nabla}_{\pi_i} v_i(\pi^t) - \mu \nabla_{\pi_i} G(\pi_i^t, \sigma_i^{k(t)}), x \right\rangle - D_\psi(x, \pi_i^t) \right\} ,$$

Algorithm 1 GABP for player i.

Require: Learning rates $\{\eta_t\}_{t\geq 0}$, perturbation strength μ, update interval T_σ, initial strategy π_i^1

$k \leftarrow 1,\ \tau \leftarrow 0$

$\sigma_i^1 \leftarrow \pi_i^1$

for $t = 1, 2, \cdots, T$ **do**

  Receive the gradient feedback $\widehat{\nabla}_{\pi_i} v_i(\pi^t)$

  Update the strategy by

$$\pi_i^{t+1} = \arg\max_{x \in \mathcal{X}_i} \left\{ \eta_t \left\langle \widehat{\nabla}_{\pi_i} v_i(\pi^t) - \mu \frac{\sigma_i^k - \sigma_i^1}{k+1} - \mu \left(\pi_i^t - \sigma_i^k \right), x \right\rangle - \frac{1}{2} \left\| x - \pi_i^t \right\|^2 \right\}$$

  $\tau \leftarrow \tau + 1$

  if $\tau = T_\sigma$ **then**

    $k \leftarrow k + 1,\ \tau \leftarrow 0$

    $\sigma_i^k \leftarrow \pi_i^{t+1}$

  end if

end for

where η_t is the learning rate at iteration t, $\mu \in (0, \infty)$ is the *perturbation strength*, and $D_\psi(\pi_i, \pi_i') = \psi(\pi_i) - \psi(\pi_i') - \langle \nabla \psi(\pi_i'), \pi_i - \pi_i' \rangle$ as the Bregman divergence associated with a strictly convex function $\psi : \mathcal{X}_i \to \mathbb{R}$. When both G and D_ψ is set to the squared ℓ^2-distance, this algorithm can be equivalently written as:

$$\pi_i^{t+1} = \arg\max_{x \in \mathcal{X}_i} \left\{ \eta_t \left\langle \widehat{\nabla}_{\pi_i} v_i(\pi^t) - \mu \left(\pi_i^t - \sigma_i^{k(t)} \right), x \right\rangle - \frac{1}{2} \left\| x - \pi_i^t \right\|^2 \right\}. \tag{3}$$

We refer to this version of APMD as Adaptively Perturbed Gradient Ascent (APGA). Abe et al. [2024] have shown that APGA exhibits the convergence rates of $\tilde{\mathcal{O}}(1/\sqrt{T})$ and $\tilde{\mathcal{O}}(1/T^{\frac{1}{10}})$ with full and noisy feedback, respectively.

3 Gradient ascent with boosting payoff perturbation

This section proposes an accelerated version of APGA, Gradient Ascent with Boosting Payoff Perturbation (GABP). The pseudo-code of GABP is outlined in Algorithm 1. In order to obtain faster last-iterate convergence rates compared to APGA, GABP introduces a novel payoff perturbation term in addition to APGA's original payoff perturbation term, $\mu \left(\pi_i^t - \sigma_i^{k(t)} \right)$. Formally, GABP updates each player's strategy as follows:

$$\pi_i^{t+1} = \arg\max_{x \in \mathcal{X}_i} \left\{ \eta_t \left\langle \widehat{\nabla}_{\pi_i} v_i(\pi^t) - \underbrace{\mu \frac{\sigma_i^{k(t)} - \sigma_i^1}{k(t)+1}}_{(*)} - \mu \left(\pi_i^t - \sigma_i^{k(t)} \right), x \right\rangle - \frac{1}{2} \left\| x - \pi_i^t \right\|^2 \right\}. \tag{4}$$

The term $(*)$ is our proposed additional perturbation term. It shrinks as $k(t)$, the number of updates of $\sigma_i^{k(t)}$, increases.

For a more intuitive explanation of the proposed perturbation term, we present the following update rule, which is equivalent to (4):

$$\pi_i^{t+1} = \arg\max_{x \in \mathcal{X}_i} \left\{ \eta_t \left\langle \widehat{\nabla}_{\pi} v_i(\pi^t) - \mu \left(\pi_i^t - \frac{k(t)\sigma_i^{k(t)} + \sigma_i^1}{k(t)+1} \right), x \right\rangle - \frac{1}{2} \left\| x - \pi_i^t \right\|^2 \right\}.$$

From this formula, it appears that GABP replaces the reference strategy $\sigma_i^{k(t)}$ for the perturbation term in (3) of APGA with $\frac{k(t)\sigma_i^{k(t)} + \sigma_i^1}{k(t)+1}$. As a result, the anchoring strategy in GABP evolves more gradually than in APGA, leading to further stabilization of the learning dynamics. There is a tradeoff between the shrinking speed of the term (*) and the stabilizing impact on the last-iterate convergence rate of GABP. The shrinking speed of $1/(k(t)+1)$ achieves a faster convergence rate, and we believe that this represents the optimal balance for this trade-off. Although one might think that the term $(*)$ is closely related to Accelerated Optimistic Gradient (AOG) [Cai and Zheng, 2023], we discuss the detail in Appendix F to be concise and avoid a complicated explanation.

4 Last-iterate convergence rates

This section provides the last-iterate convergence rates of GABP in the full/noisy feedback setting, respectively.

4.1 Full feedback setting

First, we demonstrate the last-iterate convergence rate of GABP with *full feedback* where each player receives the perfect gradient vector as feedback at each iteration t, i.e., $\widehat{\nabla}_{\pi_i} v_i(\pi^t) = \nabla_{\pi_i} v_i(\pi^t)$. Theorem 4.1 shows that the last-iterate strategy profile π^T updated by GABP converges to a Nash equilibrium with an $\tilde{\mathcal{O}}(1/T)$ rate in the full feedback setting.

Theorem 4.1. *If we use the constant learning rate $\eta_t = \eta \in (0, \frac{\mu}{(L+\mu)^2})$ and the constant perturbation strength $\mu > 0$, and set $T_\sigma = c \cdot \max(1, \frac{6 \ln 3(T+1)}{\ln(1+\eta\mu)})$ for some constant $c \geq 1$, then the strategy π^t updated by GABP satisfies for any $t \in \{2, 3, \cdots, T+1\}$:*

$$\mathrm{GAP}(\pi^t) \leq \frac{17cD^2\left(\frac{6\ln 3(T+1)}{\ln(1+\eta\mu)}+1\right)}{t-1}\left(\mu + \frac{1+\eta L}{\eta}\right), \textit{and}$$

$$r^{\tan}(\pi^t) \leq \frac{17cD\left(\frac{6\ln 3(T+1)}{\ln(1+\eta\mu)}+1\right)}{t-1}\left(\mu + \frac{1+\eta L}{\eta}\right).$$

This rate is significantly faster than APGA's rate of $\tilde{\mathcal{O}}(1/\sqrt{T})$. Moreover, it is a competitive rate compared to the previous state-of-the-art rate of $\mathcal{O}(1/T)$ [Yoon and Ryu, 2021, Cai and Zheng, 2023]. Note that the rate in Theorem 4.1 holds for any constant perturbation strength $\mu > 0$.

4.1.1 Proof sketch of Theorem 4.1

To derive the bound of the gap function $\mathrm{GAP}(\pi^t)$, it is sufficient to derive that of $r^{\tan}(\pi^t)$ due to Lemma 2.2. This section provides the proof sketch of Theorem 4.1. The complete proof is placed in Appendix B.

(1) Decomposition of the tangent residual of the last-iterate strategy profile. From the first-order optimality condition for π^t, we can see that $V(\pi^{t-1}) - \mu\left(\pi^{t-1} - \frac{k(t-1)\sigma^{k(t-1)}+\sigma^1}{k(t-1)+1}\right) - \frac{1}{\eta}\left(\pi^t - \pi^{t-1}\right) \in N_{\mathcal{X}}(\pi^t)$. Therefore, from the triangle inequality and L-smoothness (2) of the gradient operator, the tangent residual $r^{\tan}(\pi^t)$ can be bounded as:

$$\begin{aligned}
r^{\tan}(\pi^t) &= \min_{a \in N_{\mathcal{X}}(\pi^t)} \left\|-V(\pi^t) + a\right\| \\
&\leq \mathcal{O}\left(\left\|\pi^t - \pi^{t-1}\right\|\right) + \mathcal{O}\left(\left\|\pi^{t-1} - \sigma^{k(t-1)}\right\|\right) + \mathcal{O}\left(\frac{1}{k(t-1)+1}\right).
\end{aligned}$$

Let us define the stationary point $\pi^{\mu,\sigma^{k(t)}}$, which satisfies the following condition: $\forall i \in [N]$,

$$\pi_i^{\mu,\sigma^{k(t)}} = \underset{x \in \mathcal{X}_i}{\arg\max}\left\{v_i(x, \pi_{-i}^{\mu,\sigma^{k(t)}}) - \frac{\mu}{2}\left\|x - \hat{\sigma}^{k(t)}\right\|^2\right\},$$

where $\hat{\sigma}_i^{k(t)} = \frac{k(t)\sigma_i^{k(t)}+\sigma_i^1}{k(t)+1}$. We will show that π^t converges to the stationary point $\pi^{\mu,\sigma^{k(t)}}$ at an exponential rate later. By using $\pi^{\mu,\sigma^{k(t)}}$ and applying the triangle inequality to $\|\pi^t - \pi^{t-1}\|$, we decompose the term of $\mathcal{O}(\|\pi^t - \pi^{t-1}\|)$ into $\mathcal{O}(\|\pi^t - \pi^{\mu,\sigma^{k(t-1)}}\|)$ and $\mathcal{O}(\|\pi^{\mu,\sigma^{k(t-1)}} - \pi^{t-1}\|)$.

Similarly, the term of $\mathcal{O}(\|\pi^{t-1} - \sigma^{k(t-1)}\|)$ is decomposed into $\mathcal{O}(\|\pi^{t-1} - \pi^{\mu,\sigma^{k(t)-1}}\|)$ and $\mathcal{O}(\|\pi^{\mu,\sigma^{k(t)-1}} - \sigma^{k(t-1)}\|)$. Then, the tangent residual is bounded as follows:

$$r^{\tan}(\pi^t) \leq \mathcal{O}\left(\left\|\pi^{\mu,\sigma^{k(t-1)}} - \pi^t\right\|\right) + \mathcal{O}\left(\left\|\pi^{\mu,\sigma^{k(t-1)}} - \pi^{t-1}\right\|\right) + \mathcal{O}\left(\left\|\pi^{\mu,\sigma^{k(t-1)}} - \sigma^{k(t-1)}\right\|\right) + \mathcal{O}\left(\frac{1}{k(t-1)+1}\right). \tag{5}$$

Therefore, it is enough to derive the convergence rate on $\|\pi^{\mu,\sigma^{k(t)-1}} - \pi^t\|$ and $\|\pi^{\mu,\sigma^{k(t-1)}} - \sigma^{k(t-1)}\|$.

(2) Convergence rate of π^t to the stationary point $\pi^{\mu,\sigma^{k(t)}}$. Using the strong convexity of the perturbation payoff function, $\frac{\mu}{2}\|x - \hat{\sigma}_i^{k(t)}\|^2$, we show that π^t converges to $\pi^{\mu,\sigma^{k(t)}}$ exponentially fast (in Lemma B.1). That is, we have for any $t \geq 1$:

$$\left\|\pi^{\mu,\sigma^{k(t)}} - \pi^t\right\|^2 \leq \left(\frac{1}{1+\eta\mu}\right)^{t-(k(t)-1)T_\sigma - 1} \left\|\pi^{\mu,\sigma^{k(t)}} - \sigma^{k(t)}\right\|^2. \tag{6}$$

Since the first and second terms of the right-hand side of (5) are bounded by the distance between the stationary point and the anchoring strategy by using (6), we have:

$$r^{\tan}(\pi^t) \leq \mathcal{O}\left(\left\|\pi^{\mu,\sigma^{k(t-1)}} - \sigma^{k(t-1)}\right\|\right) + \mathcal{O}\left(\frac{1}{k(t-1)+1}\right). \tag{7}$$

(3) Potential function for bounding the distance between $\pi^{\mu,\sigma^{k(t)-1}}$ and $\sigma^{k(t)-1}$. To derive the upper bound on $\left\|\pi^{\mu,\sigma^{k(t-1)}} - \sigma^{k(t-1)}\right\|$, we define the following potential function $P^{k(t)}$:

$$P^{k(t)} := \frac{k(t)(k(t)+1)}{2}\left\|\pi^{\mu,\sigma^{k(t)-1}} - \hat{\sigma}^{k(t)-1}\right\|^2 + k(t)(k(t)+1)\left\langle \hat{\sigma}^{k(t)} - \pi^{\mu,\sigma^{k(t)-1}}, \pi^{\mu,\sigma^{k(t)-1}} - \hat{\sigma}^{k(t)-1}\right\rangle.$$

By some algebra, we can see that $P^{k(t)}$ is approximately non-increasing (in Lemma B.3). That is, we have for any $t \geq 1$ such that $k(t) \geq 2$:

$$P^{k(t)+1} \leq P^{k(t)} + (k(t)+1)^2 \cdot \mathcal{O}\left(\left\|\pi^{\mu,\sigma^{k(t)}} - \sigma^{k(t)+1}\right\| + \left\|\pi^{\mu,\sigma^{k(t)-1}} - \sigma^{k(t)}\right\|\right). \tag{8}$$

Using (6) again, it is easy to show that $\left\|\pi^{\mu,\sigma^{k(t)}} - \sigma^{k(t)+1}\right\| + \left\|\pi^{\mu,\sigma^{k(t)-1}} - \sigma^{k(t)}\right\| \leq \mathcal{O}\left(\frac{1}{(k(t)+1)^3}\right)$ for a sufficiently large T_σ. Therefore, under the assumption that $T_\sigma \geq \Omega(\ln T)$, by telescoping of (8) and some algebra, we can derive the following upper bound on $\left\|\pi^{\mu,\sigma^{k(t)}} - \sigma^{k(t)}\right\|$ (in Lemma B.2):

$$\left\|\pi^{\mu,\sigma^{k(t)}} - \sigma^{k(t)}\right\| \leq \mathcal{O}\left(\frac{1}{k(t)+1}\right). \tag{9}$$

(4) Putting it all together: last-iterate convergence rate of π^t. By combining (7) and (9), we get $r^{\tan}(\pi^t) \leq \mathcal{O}\left(\frac{1}{k(t-1)+1}\right)$. Therefore, since $k(t) = \lfloor \frac{t-1}{T_\sigma} \rfloor + 1$, it holds that $r^{\tan}(\pi^t) \leq \mathcal{O}\left(\frac{T_\sigma}{t+T_\sigma-2}\right)$. Finally, taking $T_\sigma = \Theta(\ln T)$, we have:

$$r^{\tan}(\pi^t) \leq \mathcal{O}\left(\frac{\ln T}{t-1}\right).$$

The upper bound on the gap function is immediately obtained since we have Lemma 2.2. □

4.2 Noisy feedback setting

Next, we establish the last-iterate convergence rate in the *noisy feedback* setting, where each player i observes a noisy gradient vector contaminated by an additive noise vector $\xi_i^t \in \mathbb{R}^{d_i}$: $\widehat{\nabla}_{\pi_i} v_i(\pi^t) + \xi_i^t$. We assume that the noisy vector ξ_i^t is zero-mean and its variance is bounded. Formally, defining the sigma-algebra generated by the history of the observations as $\mathcal{F}_t := \sigma\left((\widehat{\nabla}_{\pi_i} v_i(\pi^1))_{i\in[N]}, \ldots, (\widehat{\nabla}_{\pi_i} v_i(\pi^{t-1}))_{i\in[N]}\right)$, $\forall t \geq 1$, the noisy vector ξ_i^t is assumed to satisfy the following conditions:

Assumption 4.2. *For all $t \geq 1$ and $i \in [N]$, the noise vector ξ_i^t satisfies the following properties: (a) Zero-mean: $\mathbb{E}[\xi_i^t | \mathcal{F}_t] = (0, \cdots, 0)^\top$; (b) Bounded variance: $\mathbb{E}[\|\xi_i^t\|^2 | \mathcal{F}_t] \leq C^2$ with some constant $C > 0$.*

Assumption 4.2 is standard in online learning in games with noisy feedback [Mertikopoulos and Zhou, 2019, Hsieh et al., 2019, Abe et al., 2024] and stochastic optimization [Nemirovski et al., 2009, Nedić and Lee, 2014]. Under Assumption 4.2 and a decreasing learning rate sequence η_t, we can obtain a faster last convergence rate $\tilde{\mathcal{O}}(1/T^{\frac{1}{7}})$ than the convergence rate $\tilde{\mathcal{O}}(1/T^{\frac{1}{10}})$ of APGA.

Theorem 4.3. *Let $\kappa = \frac{\mu}{2}, \theta = \frac{3\mu^2 + 8L^2}{2\mu}$. Suppose that Assumption 4.2 holds and $V(\pi) \leq \zeta$ for any $\pi \in \mathcal{X}$. We also assume that T_σ is set to satisfy $T_\sigma = c \cdot \max(T^{\frac{6}{7}}, 1)$ for some constant $c \geq 1$. If we use the constant perturbation strength $\mu > 0$ and the decreasing learning rate sequence $\eta_t = \frac{1}{\kappa(t - T_\sigma(k(t)-1)) + 2\theta}$, then the strategy π^{T+1} satisfies:*

$$\begin{aligned}&\mathbb{E}\left[\mathrm{GAP}(\pi^{T+1})\right] \\ &\leq \frac{26c\left(D(\mu+L)+\zeta\right)\sqrt{(D+1)(D+\theta)+\kappa}}{T^{\frac{1}{7}}}\left(\sqrt{\frac{1}{\kappa}\left(D^2 + \frac{C^2}{\kappa\theta}\ln\left(\frac{\kappa T}{2\theta}+1\right)\right)}+1\right).\end{aligned}$$

Note that the non-increasing property, as described in (8), of the potential function holds regardless of the presence of noise. This implies that a proof technique similar to the one used with the potential function in the full feedback setting can also be applied in the noisy feedback setting. The detailed proof can be found in Appendix C.

5 Individual regret bound

In this section, we present an upper bound on an individual regret for each player. Specifically, we examine two performance measures in our study: the *external regret* and the *dynamic regret* [Zinkevich, 2003]. The external regret is a conventional measure in online learning. In online learning in games, the external regret for player i is defined as the gap between the player's realized cumulative payoff and the cumulative payoff of the best fixed strategy in hindsight:

$$\mathrm{Reg}_i(T) := \max_{x \in \mathcal{X}_i} \sum_{t=1}^{T} \left(v_i(x, \pi_{-i}^t) - v_i(\pi^t)\right).$$

The dynamics regret is a much stronger performance metric, which is given by:

$$\mathrm{DynamicReg}_i(T) := \sum_{t=1}^{T} \left(\max_{x \in \mathcal{X}_i} v_i(x, \pi_{-i}^t) - v_i(\pi^t)\right).$$

We show in Theorem 5.1 that the individual regret is at most $\mathcal{O}\left((\ln T)^2\right)$ if each player $i \in [N]$ plays according to GABP in the full feedback setting:

Theorem 5.1. *In the same setup of Theorem 4.1, we have for any player $i \in [N]$ and $T \geq 3$:*

$$\mathrm{Reg}_i(T) \leq \mathrm{DynamicReg}_i(T) \leq \mathcal{O}\left((\ln T)^2\right).$$

This regret bound is significantly superior to the $\mathcal{O}(\sqrt{T})$ regret bound of Optimistic Gradient Ascent, and it is slightly inferior to the $\mathcal{O}(\ln T)$ regret bound of AOG [Cai and Zheng, 2023]. The proof is given in Appendix D.

6 Experiments

In this section, we present the empirical results of our GABP, comparing its performance with Adaptively Perturbed Gradient Ascent (APGA) [Abe et al., 2024] and Optimistic Gradient Ascent (OGA) [Daskalakis et al., 2018, Wei et al., 2021]. We conduct experiments on two classes of concave-convex games. The first experiment is carried out on random payoff games, which are two-player zero-sum normal-form games with payoff matrices of size d. In this game, each player's strategy

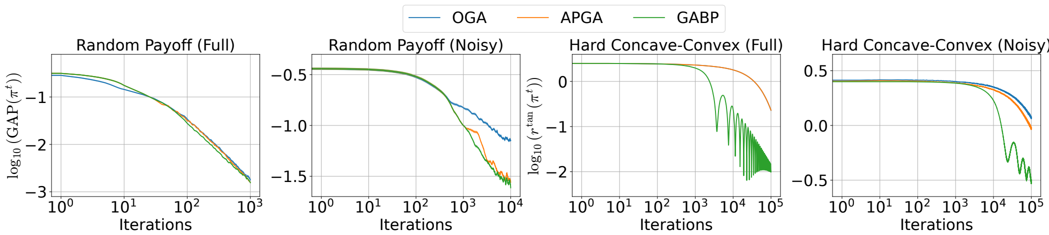

Figure 1: Performance of π^t for GABP, APGA, and OGA with full and noisy feedback in the random payoff and hard concave-convex games, respectively. The shaded area represents the standard errors. Note that we report the gap function for the random payoff game, while the tangent residual is reported for the hard concave-convex game.

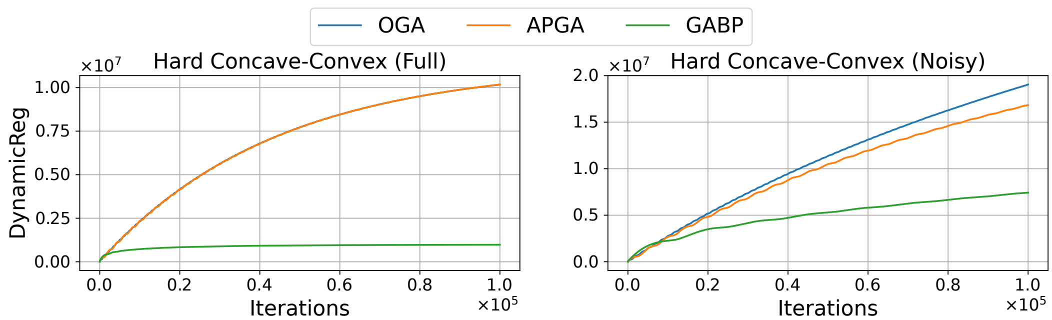

Figure 2: Dynamic regret for GABP, APGA, and OGA with full and noisy feedback.

space is represented by the d-dimensional probability simplex, i.e., $\mathcal{X}_1 = \mathcal{X}_2 = \Delta^d$. All entries of the payoff matrix are drawn independently from a uniform distribution over the interval $[-1, 1]$. We set $d = 50$ and the initial strategies are set to $\pi_1^1 = \pi_2^1 = \frac{1}{d}\mathbf{1}$. The second instance is a *hard concave-convex game* [Ouyang and Xu, 2021], formulated as the following max-min optimization problem: $\max_{x \in \mathcal{X}_1} \min_{y \in \mathcal{X}_2} f(x, y)$, where $f(x, y) = -\frac{1}{2}x^\top Hx + h^\top x + \langle Ax - b, y\rangle$. Following the setup in Cai and Zheng [2023], we choose $\mathcal{X}_1 = \mathcal{X}_2 = [-200, 200]^d$ with $d = 100$. The precise terms of $H \in \mathbb{R}^{d\times d}$, $A \in \mathbb{R}^{d\times d}$, $b \in \mathbb{R}^d$, and $h \in \mathbb{R}^d$ are provided in Appendix E.2. All algorithms are executed with initial strategies $\pi_1^1 = \pi_2^1 = \frac{1}{n}\mathbf{1}$. The detailed hyperparameters of the algorithms, tuned for best performance, are shown in Table 1 in Appendix E.3.

The numerical results of the random payoff game and the hard concave-convex game are shown in Figure 1. Both the full feedback and noisy feedback experiments in the random payoff game were conducted with 50 different random seeds, which corresponds to using 50 different payoff matrices. For experiments on the hard concave-convex game with noisy feedback, we use 10 different random seeds. We assume that the noise vector ξ_i^t is generated from the multivariate Gaussian distribution $\mathcal{N}(0,\ 0.1^2\mathbf{I})$ in an i.i.d. manner for both games. We observe that GABP exhibits competitive or faster performance over APGA and OGA in all experiments.

Figure 2 illustrates the dynamic regret in the hard concave-convex game. GABP exhibits lower regret than APGA and OGA in both settings, demonstrating its efficiency and robustness. Note that APGA and OGA exhibit almost identical trajectories with full feedback, with their plots overlapping completely.

7 Related literature

No-regret learning algorithms have been extensively studied with the intent of achieving key objectives such as average-iterate convergence or last-iterate convergence. Recently, learning algorithms introducing optimism [Rakhlin and Sridharan, 2013a,b], such as optimistic Follow the Regularized Leader [Shalev-Shwartz and Singer, 2006] and optimistic Mirror Descent [Zhou et al., 2017, Hsieh et al., 2021], have been introduced to admit last-iterate convergence in a broad spectrum of game settings. These optimistic algorithms with full feedback have been shown to achieve last-iterate convergence in various classes of games, including bilinear games [Daskalakis et al., 2018, Daskalakis and Panageas, 2019, Liang and Stokes, 2019, de Montbrun and Renault, 2022], cocoercive games [Lin et al., 2020], and saddle point problems [Daskalakis and Panageas, 2018, Mertikopoulos et al., 2019, Golowich et al., 2020b, Wei et al., 2021, Lei et al., 2021, Yoon and Ryu, 2021, Lee and Kim, 2021, Cevher et al., 2023]. Recent studies have provided finite convergence rates for monotone games [Golowich et al., 2020a, Cai et al., 2022a,b, Gorbunov et al., 2022, Cai and Zheng, 2023].

Compared to the full feedback setting, there are significant challenges in learning with noisy feedback. For example, a learning algorithm must estimate the gradient from feedback that is contaminated by noise. Despite the challenge, a vast literature has successfully achieved last-iterate convergence with noisy feedback in specific classes of games, including potential games [Cohen et al., 2017], strongly monotone games [Giannou et al., 2021b,a], and two-player zero-sum games [Abe et al., 2023]. These results have often leveraged unique structures of their payoff functions, such as strict (or strong) monotonicity [Bravo et al., 2018, Kannan and Shanbhag, 2019, Hsieh et al., 2019, Anagnostides and Panageas, 2022] and strict variational stability [Mertikopoulos et al., 2019, Azizian et al., 2021, Mertikopoulos and Zhou, 2019, Mertikopoulos et al., 2022]. Without these restrictions, convergence is mainly demonstrated in an asymptotic manner, with no quantification of the rate [Koshal et al., 2010, 2013, Yousefian et al., 2017, Tatarenko and Kamgarpour, 2019, Hsieh et al., 2020, 2022, Abe et al., 2023]. Consequently, an exceedingly large number of iterations might be necessary to reach an equilibrium.

There have been several studies focusing on payoff-regularized learning, where each player's payoff or utility function is perturbed or regularized via strongly convex functions [Cen et al., 2021, 2023, Pattathil et al., 2023]. Previous studies have successfully achieved convergence to stationary points, which are approximate equilibria. For instance, Sokota et al. [2023] have demonstrated that their perturbed mirror descent algorithm converges to a quantal response equilibrium [McKelvey and Palfrey, 1995, 1998]. Similar results have been obtained with the Boltzmann Q-learning dynamics [Tuyls et al., 2006] and penalty-regularized dynamics [Coucheney et al., 2015] in continuous-time settings [Leslie and Collins, 2005, Abe et al., 2022, Hussain et al., 2023]. To ensure convergence toward a Nash equilibrium of the underlying game, the magnitude of perturbation requires careful adjustment. Several learning algorithms have been proposed to gradually reduce the perturbation strength μ in response to this [Bernasconi et al., 2022, Liu et al., 2023, Cai et al., 2023]. These include well-studied methods such as iterative Tikhonov regularization [Facchinei and Pang, 2003, Koshal et al., 2010, Tatarenko and Kamgarpour, 2019]. Alternatively, Perolat et al. [2021] and Abe et al. [2023] have employed a payoff perturbation scheme, where the magnitude of perturbation is determined by the distance from an anchoring strategy, which is periodically re-initialized by the current strategy. Recently, Abe et al. [2024] have established $\tilde{\mathcal{O}}(1/\sqrt{T})$ and $\tilde{\mathcal{O}}(1/T^{\frac{1}{10}})$ last-iterate convergence rates for the payoff perturbation scheme in the full/noisy feedback setting, respectively. Our algorithm achieves faster $\tilde{\mathcal{O}}(1/T)$ and $\tilde{\mathcal{O}}(1/T^{\frac{1}{7}})$ last-iterate convergence rates by modifying the periodically re-initializing anchoring strategy scheme so that the anchoring strategy evolves more gradually.

8 Conclusion

This study proposes a novel payoff-perturbed algorithm, Gradient Ascent with Boosting Payoff Perturbation, which achieves $\tilde{\mathcal{O}}(1/T)$ and $\tilde{\mathcal{O}}(1/T^{\frac{1}{7}})$ last-iterate convergence rates in monotone games with full/noisy feedback, respectively. Extending our results in settings where each player only observes bandit feedback is an intriguing and challenging future direction.

References

Kenshi Abe, Mitsuki Sakamoto, and Atsushi Iwasaki. Mutation-driven follow the regularized leader for last-iterate convergence in zero-sum games. In *UAI*, pages 1–10, 2022.

Kenshi Abe, Kaito Ariu, Mitsuki Sakamoto, Kentaro Toyoshima, and Atsushi Iwasaki. Last-iterate convergence with full and noisy feedback in two-player zero-sum games. In *AISTATS*, pages 7999–8028, 2023.

Kenshi Abe, Kaito Ariu, Mitsuki Sakamoto, and Atsushi Iwasaki. Adaptively perturbed mirror descent for learning in games. In *ICML*, 2024.

Ioannis Anagnostides and Ioannis Panageas. Frequency-domain representation of first-order methods: A simple and robust framework of analysis. In *SOSA*, pages 131–160, 2022.

Waïss Azizian, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. The last-iterate convergence rate of optimistic mirror descent in stochastic variational inequalities. In *COLT*, pages 326–358, 2021.

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

Martino Bernasconi, Alberto Marchesi, and Francesco Trovò. Last-iterate convergence to trembling-hand perfect equilibria. *arXiv preprint arXiv:2208.08238*, 2022.

Mario Bravo, David Leslie, and Panayotis Mertikopoulos. Bandit learning in concave N-person games. In *NeurIPS*, pages 5666–5676, 2018.

Yang Cai and Constantinos Daskalakis. On minmax theorems for multiplayer games. In *SODA*, pages 217–234, 2011.

Yang Cai and Weiqiang Zheng. Doubly optimal no-regret learning in monotone games. In *ICML*, pages 3507–3524, 2023.

Yang Cai, Ozan Candogan, Constantinos Daskalakis, and Christos Papadimitriou. Zero-sum polymatrix games: A generalization of minmax. *Mathematics of Operations Research*, 41(2):648–655, 2016.

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

Yang Cai, Argyris Oikonomou, and Weiqiang Zheng. Tight last-iterate convergence of the extragradient method for constrained monotone variational inequalities. *arXiv preprint arXiv:2204.09228*, 2022b.

Yang Cai, Haipeng Luo, Chen-Yu Wei, and Weiqiang Zheng. Uncoupled and convergent learning in two-player zero-sum markov games with bandit feedback. In *NeurIPS*, pages 36364–36406, 2023.

Shicong Cen, Yuting Wei, and Yuejie Chi. Fast policy extragradient methods for competitive games with entropy regularization. In *NeurIPS*, pages 27952–27964, 2021.

Shicong Cen, Yuejie Chi, Simon S Du, and Lin Xiao. Faster last-iterate convergence of policy optimization in zero-sum Markov games. In *ICLR*, 2023.

Volkan Cevher, Georgios Piliouras, Ryann Sim, and Stratis Skoulakis. Min-max optimization made simple: Approximating the proximal point method via contraction maps. In *Symposium on Simplicity in Algorithms (SOSA)*, pages 192–206, 2023.

Johanne Cohen, Amélie Héliou, and Panayotis Mertikopoulos. Learning with bandit feedback in potential games. In *NeurIPS*, pages 6372–6381, 2017.

Pierre Coucheney, Bruno Gaujal, and Panayotis Mertikopoulos. Penalty-regulated dynamics and robust learning procedures in games. *Mathematics of Operations Research*, 40(3):611–633, 2015.

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

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

Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training GANs with optimism. In *ICLR*, 2018.

Étienne de Montbrun and Jérôme Renault. Convergence of optimistic gradient descent ascent in bilinear games. *arXiv preprint arXiv:2208.03085*, 2022.

Gerard Debreu. A social equilibrium existence theorem. *Proceedings of the National Academy of Sciences*, 38(10):886–893, 1952.

Francisco Facchinei and Jong-Shi Pang. *Finite-dimensional variational inequalities and complementarity problems*. Springer, 2003.

Angeliki Giannou, Emmanouil Vasileios Vlatakis-Gkaragkounis, and Panayotis Mertikopoulos. Survival of the strictest: Stable and unstable equilibria under regularized learning with partial information. In *COLT*, pages 2147–2148, 2021a.

Angeliki Giannou, Emmanouil-Vasileios Vlatakis-Gkaragkounis, and Panayotis Mertikopoulos. On the rate of convergence of regularized learning in games: From bandits and uncertainty to optimism and beyond. In *NeurIPS*, pages 22655–22666, 2021b.

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

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

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, pages 2672–2680, 2014.

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

Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. On the convergence of single-call stochastic extra-gradient methods. In *NeurIPS*, pages 6938–6948, 2019.

Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. Explore aggressively, update conservatively: Stochastic extragradient methods with variable stepsize scaling. In *NeurIPS*, pages 16223–16234, 2020.

Yu-Guan Hsieh, Kimon Antonakopoulos, and Panayotis Mertikopoulos. Adaptive learning in continuous games: Optimal regret bounds and convergence to Nash equilibrium. In *COLT*, pages 2388–2422, 2021.

Yu-Guan Hsieh, Kimon Antonakopoulos, Volkan Cevher, and Panayotis Mertikopoulos. No-regret learning in games with noisy feedback: Faster rates and adaptivity via learning rate separation. In *NeurIPS*, pages 6544–6556, 2022.

Aamal Abbas Hussain, Francesco Belardinelli, and Georgios Piliouras. Asymptotic convergence and performance of multi-agent Q-learning dynamics. *arXiv preprint arXiv:2301.09619*, 2023.

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

Jayash Koshal, Angelia Nedić, and Uday V Shanbhag. Single timescale regularized stochastic approximation schemes for monotone nash games under uncertainty. In *CDC*, pages 231–236. IEEE, 2010.

Jayash Koshal, Angelia Nedic, and Uday V. Shanbhag. Regularized iterative stochastic approximation methods for stochastic variational inequality problems. *IEEE Transactions on Automatic Control*, 58(3):594–609, 2013.

Sucheol Lee and Donghwan Kim. Fast extra gradient methods for smooth structured nonconvex-nonconcave minimax problems. In *NeurIPS*, pages 22588–22600, 2021.

Qi Lei, Sai Ganesh Nagarajan, Ioannis Panageas, et al. Last iterate convergence in no-regret learning: constrained min-max optimization for convex-concave landscapes. In *AISTATS*, pages 1441–1449, 2021.

David S Leslie and Edmund J Collins. Individual q-learning in normal form games. *SIAM Journal on Control and Optimization*, 44(2):495–514, 2005.

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

Tianyi Lin, Zhengyuan Zhou, Panayotis Mertikopoulos, and Michael Jordan. Finite-time last-iterate convergence for multi-agent learning in games. In *ICML*, pages 6161–6171, 2020.

Mingyang Liu, Asuman Ozdaglar, Tiancheng Yu, and Kaiqing Zhang. The power of regularization in solving extensive-form games. In *ICLR*, 2023.

Richard D McKelvey and Thomas R Palfrey. Quantal response equilibria for normal form games. *Games and economic behavior*, 10(1):6–38, 1995.

Richard D McKelvey and Thomas R Palfrey. Quantal response equilibria for extensive form games. *Experimental economics*, 1:9–41, 1998.

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

Panayotis Mertikopoulos, Bruno Lecouat, Houssam Zenati, Chuan-Sheng Foo, Vijay Chandrasekhar, and Georgios Piliouras. Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile. In *ICLR*, 2019.

Panayotis Mertikopoulos, Ya-Ping Hsieh, and Volkan Cevher. Learning in games from a stochastic approximation viewpoint. *arXiv preprint arXiv:2206.03922*, 2022.

Rémi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Andrea Michi, et al. Nash learning from human feedback. *arXiv preprint arXiv:2312.00886*, 2023.

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

Angelia Nedić and Soomin Lee. On stochastic subgradient mirror-descent algorithm with weighted averaging. *SIAM Journal on Optimization*, 24(1):84–107, 2014.

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

Yuyuan Ouyang and Yangyang Xu. Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems. *Mathematical Programming*, 185(1):1–35, 2021.

Sarath Pattathil, Kaiqing Zhang, and Asuman Ozdaglar. Symmetric (optimistic) natural policy gradient for multi-agent learning with parameter convergence. In *AISTATS*, pages 5641–5685, 2023.

Julien Perolat, Remi Munos, Jean-Baptiste Lespiau, Shayegan Omidshafiei, Mark Rowland, Pedro Ortega, Neil Burch, Thomas Anthony, David Balduzzi, Bart De Vylder, et al. From Poincaré recurrence to convergence in imperfect information games: Finding equilibrium via regularization. In *ICML*, pages 8525–8535, 2021.

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

Sasha Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. In *NeurIPS*, pages 3066–3074, 2013b.

Shai Shalev-Shwartz and Yoram Singer. Convex repeated games and fenchel duality. *Advances in neural information processing systems*, 19, 2006.

Samuel Sokota, Ryan D'Orazio, J Zico Kolter, Nicolas Loizou, Marc Lanctot, Ioannis Mitliagkas, Noam Brown, and Christian Kroer. A unified approach to reinforcement learning, quantal response equilibria, and two-player zero-sum games. In *ICLR*, 2023.

Gokul Swamy, Christoph Dann, Rahul Kidambi, Zhiwei Steven Wu, and Alekh Agarwal. A minimaximalist approach to reinforcement learning from human feedback. *arXiv preprint arXiv:2401.04056*, 2024.

Tatiana Tatarenko and Maryam Kamgarpour. Learning Nash equilibria in monotone games. In *CDC*, pages 3104–3109. IEEE, 2019.

Karl Tuyls, Pieter Jan Hoen, and Bram Vanschoenwinkel. An evolutionary dynamical analysis of multi-agent learning in iterated games. *Autonomous Agents and Multi-Agent Systems*, 12(1): 115–153, 2006.

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

TaeHo Yoon and Ernest K Ryu. Accelerated algorithms for smooth convex-concave minimax problems with $\mathcal{O}(1/k^2)$ rate on squared gradient norm. In *ICML*, pages 12098–12109, 2021.

Farzad Yousefian, Angelia Nedić, and Uday V Shanbhag. On smoothing, regularization, and averaging in stochastic approximation methods for stochastic variational inequality problems. *Mathematical Programming*, 165:391–431, 2017.

Zhengyuan Zhou, Panayotis Mertikopoulos, Aris L Moustakas, Nicholas Bambos, and Peter Glynn. Mirror descent learning in continuous games. In *CDC*, pages 5776–5783. IEEE, 2017.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *ICML*, pages 928–936, 2003.

A Broader impact

Our study can bring about a positive impact on society by contributing to the advancement of the Game AI industry. However, as far as we can envision, there are no conceivable negative social impacts.

B Proofs for Theorem 4.1

B.1 Proof of Theorem 4.1

Proof of Theorem 4.1. From the first-order optimality condition for π^t, we have for any $x \in \mathcal{X}$:

$$\left\langle V(\pi^{t-1}) - \mu\left(\pi^{t-1} - \frac{k(t-1)\sigma^{k(t-1)} + \sigma^1}{k(t-1)+1}\right) - \frac{1}{\eta}\left(\pi^t - \pi^{t-1}\right), \pi^t - x\right\rangle \geq 0,$$

and then $V(\pi^{t-1}) - \mu\left(\pi^{t-1} - \frac{k(t-1)\sigma^{k(t-1)}+\sigma^1}{k(t-1)+1}\right) - \frac{1}{\eta}\left(\pi^t - \pi^{t-1}\right) \in N_{\mathcal{X}}(\pi^t)$. Thus, the tangent residual for π^t can be bounded as:

$$\begin{aligned} r^{\tan}(\pi^t) &= \min_{a \in N_{\mathcal{X}}(\pi^t)} \left\| -V(\pi^t) + a \right\| \\ &\leq \left\| -V(\pi^t) + V(\pi^{t-1}) - \mu\left(\pi^{t-1} - \frac{k(t-1)\sigma^{k(t-1)} + \sigma^1}{k(t-1)+1}\right) - \frac{1}{\eta}\left(\pi^t - \pi^{t-1}\right)\right\|. \end{aligned}$$

Letting us define

$$\pi_i^{\mu,\sigma^k} = \underset{\pi_i \in \mathcal{X}_i}{\arg\max} \left\{ v_i(\pi_i, \pi_{-i}^{\mu,\sigma^k}) - \frac{\mu}{2}\left\| \pi_i - \frac{k\sigma_i^k + \sigma_i^1}{k+1}\right\|^2 \right\},$$

then we get by triangle inequality:

$$\begin{aligned}
r^{\tan}(\pi^t) &\leq \left\| -V(\pi^t) + V(\pi^{t-1}) - \frac{\mu}{k(t-1)+1}(\sigma^{k(t-1)} - \sigma^1) \right. \\
&\qquad \left. - \mu(\pi^{\mu,\sigma^{k(t-1)}} - \pi^{\mu,\sigma^{k(t-1)}} + \pi^{t-1} - \sigma^{k(t-1)}) - \frac{1}{\eta}(\pi^t - \pi^{t-1}) \right\| \\
&\leq \left\| -V(\pi^t) + V(\pi^{t-1}) \right\| + \frac{\mu}{k(t-1)+1} \left\| \sigma^{k(t-1)} - \sigma^1 \right\| \\
&\qquad + \mu \left\| \pi^{\mu,\sigma^{k(t-1)}} - \sigma^{k(t-1)} \right\| + \mu \left\| \pi^{\mu,\sigma^{k(t-1)}} - \pi^{t-1} \right\| + \frac{1}{\eta} \left\| \pi^t - \pi^{t-1} \right\| \\
&\leq \frac{1+\eta L}{\eta} \left\| \pi^t - \pi^{t-1} \right\| + \frac{\mu D}{k(t-1)+1} \\
&\qquad + \mu \left\| \pi^{\mu,\sigma^{k(t-1)}} - \sigma^{k(t-1)} \right\| + \mu \left\| \pi^{\mu,\sigma^{k(t-1)}} - \pi^{t-1} \right\| \\
&\leq \frac{1+\eta L}{\eta} \left\| \pi^{\mu,\sigma^{k(t-1)}} - \pi^t \right\| + \frac{\mu D}{k(t-1)+1} + \mu \left\| \pi^{\mu,\sigma^{k(t-1)}} - \sigma^{k(t-1)} \right\| \\
&\qquad + \left(\mu + \frac{1+\eta L}{\eta} \right) \left\| \pi^{\mu,\sigma^{k(t-1)}} - \pi^{t-1} \right\|. \qquad (10)
\end{aligned}$$

In terms of upper bound on $\left\| \pi^{\mu,\sigma^{k(t-1)}} - \pi^t \right\|$ and $\left\| \pi^{\mu,\sigma^{k(t-1)}} - \pi^{t-1} \right\|$, we introduce the following lemma:

Lemma B.1. *If we use the constant learning rate $\eta_t = \eta \in (0, \frac{\mu}{(L+\mu)^2})$, we have for any $t \geq 1$:*

$$\begin{aligned}
\left\| \pi^{\mu,\sigma^{k(t)}} - \pi^t \right\|^2 &\leq \left(\frac{1}{1+\eta\mu} \right)^{t-(k(t)-1)T_\sigma - 1} \left\| \pi^{\mu,\sigma^{k(t)}} - \sigma^{k(t)} \right\|^2, \\
\left\| \pi^{\mu,\sigma^{k(t)}} - \pi^{t+1} \right\|^2 &\leq \left(\frac{1}{1+\eta\mu} \right)^{t-(k(t)-1)T_\sigma} \left\| \pi^{\mu,\sigma^{k(t)}} - \sigma^{k(t)} \right\|^2.
\end{aligned}$$

Combining (10) and Lemma B.1, we have:

$$r^{\tan}(\pi^t) \leq 2\left(\mu + \frac{1+\eta L}{\eta} \right) \left\| \pi^{\mu,\sigma^{k(t-1)}} - \sigma^{k(t-1)} \right\| + \frac{\mu D}{k(t-1)+1}. \qquad (11)$$

Next, we derive the following upper bound on $\left\| \pi^{\mu,\sigma^{k(t-1)}} - \sigma^{k(t-1)} \right\|$:

Lemma B.2. *If we set $\eta_t = \eta \in (0, \frac{\mu}{(L+\mu)^2})$ and $T_\sigma \geq \max(1, \frac{6 \ln 3(T+1)}{\ln(1+\eta\mu)})$, we have for any $t \geq 1$:*

$$\left\| \pi^{\mu,\sigma^{k(t)}} - \sigma^{k(t)} \right\| \leq \frac{8D}{k(t)+1}.$$

By combining (11) and Lemma B.2, we get:

$$\begin{aligned}
r^{\tan}(\pi^t) &\leq \frac{16D}{k(t-1)+1} \left(\mu + \frac{1+\eta L}{\eta} \right) + \frac{\mu D}{k(t-1)+1} \\
&\leq \frac{17D}{k(t-1)+1} \left(\mu + \frac{1+\eta L}{\eta} \right).
\end{aligned}$$

Therefore, since $k(t) = \lfloor \frac{t-1}{T_\sigma} \rfloor + 1$, it holds that:

$$r^{\tan}(\pi^t) \leq \frac{17DT_\sigma}{t+T_\sigma-2} \left(\mu + \frac{1+\eta L}{\eta} \right).$$

Finally, taking $T_\sigma = c \cdot \max(1, \frac{6 \ln 3(T+1)}{\ln(1+\eta\mu)})$, we have:

$$r^{\tan}(\pi^t) \leq \frac{17cD\left(\frac{6 \ln 3(T+1)}{\ln(1+\eta\mu)} + 1 \right)}{t-1} \left(\mu + \frac{1+\eta L}{\eta} \right).$$

$\square$

B.2 Proof of Lemma B.1

Proof of Lemma B.1. First, we have for any three vectors a, b, c:

$$\frac{1}{2}\|a-b\|^2 - \frac{1}{2}\|a-c\|^2 + \frac{1}{2}\|b-c\|^2 = \langle c-b, a-b\rangle.$$

Thus, we have for any $t \geq 1$:

$$\frac{1}{2}\left\|\pi^{\mu,\sigma^{k(t)}} - \pi^{t+1}\right\|^2 - \frac{1}{2}\left\|\pi^{\mu,\sigma^{k(t)}} - \pi^{t}\right\|^2 + \frac{1}{2}\left\|\pi^{t+1} - \pi^{t}\right\|^2 = \left\langle \pi^t - \pi^{t+1}, \pi^{\mu,\sigma^{k(t)}} - \pi^{t+1}\right\rangle. \tag{12}$$

Here, let us define $\hat{\sigma}^{k(t)} = \frac{k(t)\sigma^{k(t)}+\sigma^1}{k(t)+1}$. Then, from the first-order optimality condition for π^{t+1}, we have for any $t \geq 1$:

$$\left\langle \eta\left(V(\pi^t) - \mu\left(\pi^t - \hat{\sigma}^{k(t)}\right)\right) - \pi^{t+1} + \pi^t, \pi^{t+1} - \pi^{\mu,\sigma^{k(t)}}\right\rangle \geq 0. \tag{13}$$

Similarly, from the first-order optimality condition for $\pi^{\mu,\sigma^{k(t)}}$, we get:

$$\left\langle V(\pi^{\mu,\sigma^{k(t)}}) - \mu\left(\pi^{\mu,\sigma^{k(t)}} - \hat{\sigma}^{k(t)}\right), \pi^{\mu,\sigma^{k(t)}} - \pi^{t+1}\right\rangle \geq 0. \tag{14}$$

Combining (12), (13), and (14) yields:

$$\begin{aligned}
&\frac{1}{2}\left\|\pi^{\mu,\sigma^{k(t)}} - \pi^{t+1}\right\|^2 - \frac{1}{2}\left\|\pi^{\mu,\sigma^{k(t)}} - \pi^{t}\right\|^2 + \frac{1}{2}\left\|\pi^{t+1} - \pi^{t}\right\|^2 \\
&\leq \eta\left\langle V(\pi^t) - \mu\left(\pi^t - \hat{\sigma}^{k(t)}\right), \pi^{t+1} - \pi^{\mu,\sigma^{k(t)}}\right\rangle \\
&= \eta\left\langle V(\pi^{t+1}) - \mu\left(\pi^{t+1} - \hat{\sigma}^{k(t)}\right), \pi^{t+1} - \pi^{\mu,\sigma^{k(t)}}\right\rangle \\
&\quad + \eta\left\langle V(\pi^t) - V(\pi^{t+1}) - \mu\left(\pi^t - \pi^{t+1}\right), \pi^{t+1} - \pi^{\mu,\sigma^{k(t)}}\right\rangle \\
&\leq \eta\left\langle V(\pi^{\mu,\sigma^{k(t)}}) - \mu\left(\pi^{t+1} - \hat{\sigma}^{k(t)}\right), \pi^{t+1} - \pi^{\mu,\sigma^{k(t)}}\right\rangle \\
&\quad + \eta\left\langle V(\pi^t) - V(\pi^{t+1}) - \mu\left(\pi^t - \pi^{t+1}\right), \pi^{t+1} - \pi^{\mu,\sigma^{k(t)}}\right\rangle \\
&= \eta\left\langle V(\pi^{\mu,\sigma^{k(t)}}) - \mu\left(\pi^{\mu,\sigma^{k(t)}} - \hat{\sigma}^{k(t)}\right), \pi^{t+1} - \pi^{\mu,\sigma^{k(t)}}\right\rangle - \eta\mu\left\|\pi^{t+1} - \pi^{\mu,\sigma^{k(t)}}\right\|^2 \\
&\quad + \eta\left\langle V(\pi^t) - V(\pi^{t+1}) - \mu\left(\pi^t - \pi^{t+1}\right), \pi^{t+1} - \pi^{\mu,\sigma^{k(t)}}\right\rangle \\
&\leq -\eta\mu\left\|\pi^{t+1} - \pi^{\mu,\sigma^{k(t)}}\right\|^2 + \eta\left\langle V(\pi^t) - V(\pi^{t+1}) - \mu\left(\pi^t - \pi^{t+1}\right), \pi^{t+1} - \pi^{\mu,\sigma^{k(t)}}\right\rangle,
\end{aligned} \tag{15}$$

where the second inequality follows from (1). From Cauchy-Schwarz inequality and Young's inequality, the second term in the right-hand side of this inequality can be bounded by:

$$\begin{aligned}
&\eta\left\langle V(\pi^t) - V(\pi^{t+1}) - \mu\left(\pi^t - \pi^{t+1}\right), \pi^{t+1} - \pi^{\mu,\sigma^{k(t)}}\right\rangle \\
&= \eta\left\langle V(\pi^t) - V(\pi^{t+1}), \pi^{t+1} - \pi^{\mu,\sigma^{k(t)}}\right\rangle - \eta\mu\left\langle \pi^t - \pi^{t+1}, \pi^{t+1} - \pi^{\mu,\sigma^{k(t)}}\right\rangle \\
&\leq \eta\left(\left\|V(\pi^t) - V(\pi^{t+1})\right\| + \mu\left\|\pi^t - \pi^{t+1}\right\|\right)\cdot\left\|\pi^{t+1} - \pi^{\mu,\sigma^{k(t)}}\right\| \\
&\leq \eta(L+\mu)\left\|\pi^t - \pi^{t+1}\right\|\cdot\left\|\pi^{t+1} - \pi^{\mu,\sigma^{k(t)}}\right\| \\
&\leq \frac{1}{2}\left\|\pi^t - \pi^{t+1}\right\|^2 + \frac{\eta^2(L+\mu)^2}{2}\left\|\pi^{t+1} - \pi^{\mu,\sigma^{k(t)}}\right\|^2 \\
&\leq \frac{1}{2}\left\|\pi^t - \pi^{t+1}\right\|^2 + \frac{\eta\mu}{2}\left\|\pi^{t+1} - \pi^{\mu,\sigma^{k(t)}}\right\|^2,
\end{aligned} \tag{16}$$

where the second inequality follow from (2), and the last inequality follows from the assumption that $\eta \le \frac{\mu}{(L+\mu)^2}$. By combining (15) and (16), we get:

$$\frac{1}{2}\left\|\pi^{\mu,\sigma^{k(t)}} - \pi^{t+1}\right\|^2 - \frac{1}{2}\left\|\pi^{\mu,\sigma^{k(t)}} - \pi^{t}\right\|^2 + \frac{1}{2}\left\|\pi^{t+1} - \pi^{t}\right\|^2$$
$$\le -\frac{\eta\mu}{2}\left\|\pi^{t+1} - \pi^{\mu,\sigma^{k(t)}}\right\|^2 + \frac{1}{2}\left\|\pi^{t} - \pi^{t+1}\right\|^2.$$

Thus,

$$\frac{1+\eta\mu}{2}\left\|\pi^{\mu,\sigma^{k(t)}} - \pi^{t+1}\right\|^2 \le \frac{1}{2}\left\|\pi^{\mu,\sigma^{k(t)}} - \pi^{t}\right\|^2.$$

Therefore, we have for any $t \ge 1$:

$$\left\|\pi^{\mu,\sigma^{k(t)}} - \pi^{t+1}\right\|^2 \le \frac{1}{1+\eta\mu}\left\|\pi^{\mu,\sigma^{k(t)}} - \pi^{t}\right\|^2.$$

Furthermore, since $k(s) = k(t)$ for $s \in [(k(t)-1)T_\sigma + 1, t]$, we have for such s that:

$$\left\|\pi^{\mu,\sigma^{k(t)}} - \pi^{s+1}\right\|^2 \le \frac{1}{1+\eta\mu}\left\|\pi^{\mu,\sigma^{k(t)}} - \pi^{s}\right\|^2.$$

Therefore, by applying this inequality from $t, t-1, \cdots, (k(t)-1)T_\sigma + 1$, we get for any $t \ge 1$:

$$\begin{aligned}\left\|\pi^{\mu,\sigma^{k(t)}} - \pi^{t+1}\right\|^2 &\le \left(\frac{1}{1+\eta\mu}\right)^{t-(k(t)-1)T_\sigma}\left\|\pi^{\mu,\sigma^{k(t)}} - \pi^{(k(t)-1)T_\sigma+1}\right\|^2 \\ &= \left(\frac{1}{1+\eta\mu}\right)^{t-(k(t)-1)T_\sigma}\left\|\pi^{\mu,\sigma^{k(t)}} - \sigma^{k(t)}\right\|^2. \end{aligned} \tag{17}$$

Here, since $k(t) = k(t+1)$ when t satisfies that $t \ne T_\sigma \left\lfloor \frac{t}{T_\sigma}\right\rfloor$, we have for such t that:

$$\left\|\pi^{\mu,\sigma^{k(t+1)}} - \pi^{t+1}\right\|^2 \le \left(\frac{1}{1+\eta\mu}\right)^{t-(k(t+1)-1)T_\sigma}\left\|\pi^{\mu,\sigma^{k(t+1)}} - \sigma^{k(t+1)}\right\|^2. \tag{18}$$

On the other hand, when t satisfies that $t = T_\sigma \left\lfloor \frac{t}{T_\sigma}\right\rfloor$:

$$\begin{aligned} k(t+1) &= \left\lfloor \frac{T_\sigma \left\lfloor \frac{t}{T_\sigma}\right\rfloor + 1 - 1}{T_\sigma}\right\rfloor + 1 = \left\lfloor \frac{t}{T_\sigma}\right\rfloor + 1 \\ &\Rightarrow (k(t+1)-1)T_\sigma = T_\sigma\left\lfloor \frac{t}{T_\sigma}\right\rfloor = t \\ &\Rightarrow \pi^{t+1} = \pi^{(k(t+1)-1)T_\sigma+1} = \sigma^{k(t+1)}. \end{aligned}$$

Therefore, we have for any $t \ge 1$ such that $t = T_\sigma \left\lfloor \frac{t}{T_\sigma}\right\rfloor$:

$$\begin{aligned}\left\|\pi^{\mu,\sigma^{k(t+1)}} - \pi^{t+1}\right\|^2 &= \left\|\pi^{\mu,\sigma^{k(t+1)}} - \sigma^{k(t+1)}\right\|^2 \\ &= \left(\frac{1}{1+\eta\mu}\right)^{t-(k(t+1)-1)T_\sigma}\left\|\pi^{\mu,\sigma^{k(t+1)}} - \sigma^{k(t+1)}\right\|^2. \end{aligned} \tag{19}$$

By combining (17), (18), and (19), we have for any $t \ge 1$:

$$\left\|\pi^{\mu,\sigma^{k(t)}} - \pi^{t+1}\right\|^2 \le \left(\frac{1}{1+\eta\mu}\right)^{t-(k(t)-1)T_\sigma}\left\|\pi^{\mu,\sigma^{k(t)}} - \sigma^{k(t)}\right\|^2,$$
$$\left\|\pi^{\mu,\sigma^{k(t+1)}} - \pi^{t+1}\right\|^2 \le \left(\frac{1}{1+\eta\mu}\right)^{t-(k(t+1)-1)T_\sigma}\left\|\pi^{\mu,\sigma^{k(t+1)}} - \sigma^{k(t+1)}\right\|^2.$$

□

B.3 Proof of Lemma B.2

Proof of Lemma B.2. First, we have for any Nash equilibrium $\pi^* \in \Pi^*$ and $t \geq 1$ such that $k(t) \geq 1$:

$$\begin{aligned}
&\frac{(k(t)+1)(k(t)+2)}{2}\left\|\pi^{\mu,\sigma^{k(t)}} - \hat{\sigma}^{k(t)}\right\|^2 + (k(t)+1)(k(t)+2)\left\langle \hat{\sigma}^{k(t)+1} - \pi^{\mu,\sigma^{k(t)}}, \pi^{\mu,\sigma^{k(t)}} - \hat{\sigma}^{k(t)}\right\rangle \\
&= \frac{(k(t)+1)(k(t)+2)}{2}\left\|\pi^{\mu,\sigma^{k(t)}} - \hat{\sigma}^{k(t)}\right\|^2 \\
&\quad + (k(t)+1)\left\langle (k(t)+1)\sigma^{k(t)+1} + \sigma^1 - (k(t)+2)\pi^{\mu,\sigma^{k(t)}}, \pi^{\mu,\sigma^{k(t)}} - \hat{\sigma}^{k(t)}\right\rangle \\
&= \frac{(k(t)+1)(k(t)+2)}{2}\left\|\pi^{\mu,\sigma^{k(t)}} - \hat{\sigma}^{k(t)}\right\|^2 + (k(t)+1)\left\langle \sigma^1 - \sigma^{k(t)+1}, \pi^{\mu,\sigma^{k(t)}} - \hat{\sigma}^{k(t)}\right\rangle \\
&\quad + (k(t)+1)(k(t)+2)\left\langle \sigma^{k(t)+1} - \pi^{\mu,\sigma^{k(t)}}, \pi^{\mu,\sigma^{k(t)}} - \hat{\sigma}^{k(t)}\right\rangle \\
&= \frac{(k(t)+1)(k(t)+2)}{2}\left\|\pi^{\mu,\sigma^{k(t)}} - \hat{\sigma}^{k(t)}\right\|^2 + (k(t)+1)\left\langle \sigma^1 - \pi^{\mu,\sigma^{k(t)}}, \pi^{\mu,\sigma^{k(t)}} - \hat{\sigma}^{k(t)}\right\rangle \\
&\quad + (k(t)+1)^2\left\langle \sigma^{k(t)+1} - \pi^{\mu,\sigma^{k(t)}}, \pi^{\mu,\sigma^{k(t)}} - \hat{\sigma}^{k(t)}\right\rangle \\
&= \frac{(k(t)+1)(k(t)+2)}{2}\left\|\pi^{\mu,\sigma^{k(t)}} - \hat{\sigma}^{k(t)}\right\|^2 + (k(t)+1)\left\langle \sigma^1 - \pi^*, \pi^{\mu,\sigma^{k(t)}} - \hat{\sigma}^{k(t)}\right\rangle \\
&\quad + (k(t)+1)\left\langle \pi^* - \pi^{\mu,\sigma^{k(t)}}, \pi^{\mu,\sigma^{k(t)}} - \hat{\sigma}^{k(t)}\right\rangle + (k(t)+1)^2\left\langle \sigma^{k(t)+1} - \pi^{\mu,\sigma^{k(t)}}, \pi^{\mu,\sigma^{k(t)}} - \hat{\sigma}^{k(t)}\right\rangle.
\end{aligned}$$

Here, the first-order optimality condition for $\pi^{\mu,\sigma^{k(t)}}$:

$$\begin{aligned}
&\left\langle V(\pi^{\mu,\sigma^{k(t)}}) - \mu\left(\pi^{\mu,\sigma^{k(t)}} - \hat{\sigma}^{k(t)}\right), \pi^{\mu,\sigma^{k(t)}} - \pi^*\right\rangle \geq 0 \\
&\Rightarrow \left\langle \pi^{\mu,\sigma^{k(t)}} - \hat{\sigma}^{k(t)}, \pi^* - \pi^{\mu,\sigma^{k(t)}}\right\rangle \geq \frac{1}{\mu}\left\langle V(\pi^{\mu,\sigma^{k(t)}}), \pi^* - \pi^{\mu,\sigma^{k(t)}}\right\rangle \geq \frac{1}{\mu}\left\langle V(\pi^*), \pi^* - \pi^{\mu,\sigma^{k(t)}}\right\rangle \geq 0,
\end{aligned}$$

where we use (1) and the fact that π^* is a Nash equilibrium. Combining these inequalities yields:

$$\begin{aligned}
&\frac{(k(t)+1)(k(t)+2)}{2}\left\|\pi^{\mu,\sigma^{k(t)}} - \hat{\sigma}^{k(t)}\right\|^2 + (k(t)+1)(k(t)+2)\left\langle \hat{\sigma}^{k(t)+1} - \pi^{\mu,\sigma^{k(t)}}, \pi^{\mu,\sigma^{k(t)}} - \hat{\sigma}^{k(t)}\right\rangle \\
&\geq \frac{(k(t)+1)(k(t)+2)}{2}\left\|\pi^{\mu,\sigma^{k(t)}} - \hat{\sigma}^{k(t)}\right\|^2 + (k(t)+1)\left\langle \sigma^1 - \pi^*, \pi^{\mu,\sigma^{k(t)}} - \hat{\sigma}^{k(t)}\right\rangle \\
&\quad + (k(t)+1)^2\left\langle \sigma^{k(t)+1} - \pi^{\mu,\sigma^{k(t)}}, \pi^{\mu,\sigma^{k(t)}} - \hat{\sigma}^{k(t)}\right\rangle.
\end{aligned}$$

From Young's inequality, we have for any $\rho_1, \rho_2 > 0$:

$$\begin{aligned}
&\frac{(k(t)+1)(k(t)+2)}{2}\left\|\pi^{\mu,\sigma^{k(t)}} - \hat{\sigma}^{k(t)}\right\|^2 + (k(t)+1)(k(t)+2)\left\langle \hat{\sigma}^{k(t)+1} - \pi^{\mu,\sigma^{k(t)}}, \pi^{\mu,\sigma^{k(t)}} - \hat{\sigma}^{k(t)}\right\rangle \\
&\geq \frac{(k(t)+1)(k(t)+2)}{2}\left\|\pi^{\mu,\sigma^{k(t)}} - \hat{\sigma}^{k(t)}\right\|^2 - \frac{\rho_1(k(t)+1)}{2}\left\|\sigma^1 - \pi^*\right\|^2 - \frac{(k(t)+1)}{2\rho_1}\left\|\pi^{\mu,\sigma^{k(t)}} - \hat{\sigma}^{k(t)}\right\|^2 \\
&\quad - \frac{\rho_2(k(t)+1)^2}{2}\left\|\sigma^{k(t)+1} - \pi^{\mu,\sigma^{k(t)}}\right\|^2 - \frac{(k(t)+1)^2}{2\rho_2}\left\|\pi^{\mu,\sigma^{k(t)}} - \hat{\sigma}^{k(t)}\right\|^2 \\
&= \left(\frac{(k(t)+1)(k(t)+2)}{2} - \frac{k(t)+1}{2\rho_1} - \frac{(k(t)+1)^2}{2\rho_2}\right)\left\|\pi^{\mu,\sigma^{k(t)}} - \hat{\sigma}^{k(t)}\right\|^2 \\
&\quad - \frac{\rho_1(k(t)+1)}{2}\left\|\sigma^1 - \pi^*\right\|^2 - \frac{\rho_2(k(t)+1)^2}{2}\left\|\sigma^{k(t)+1} - \pi^{\mu,\sigma^{k(t)}}\right\|^2.
\end{aligned}$$

By setting $\rho_1 = \frac{4}{k(t)+2}, \rho_2 = \frac{4(k(t)+1)}{k(t)+2}$, we obtain:

$$\begin{aligned}
&\frac{(k(t)+1)(k(t)+2)}{2}\left\|\pi^{\mu,\sigma^{k(t)}} - \hat{\sigma}^{k(t)}\right\|^2 + (k(t)+1)(k(t)+2)\left\langle \hat{\sigma}^{k(t)+1} - \pi^{\mu,\sigma^{k(t)}}, \pi^{\mu,\sigma^{k(t)}} - \hat{\sigma}^{k(t)}\right\rangle \\
&\geq \frac{(k(t)+1)(k(t)+2)}{4}\left\|\pi^{\mu,\sigma^{k(t)}} - \hat{\sigma}^{k(t)}\right\|^2 - \frac{2(k(t)+1)}{k(t)+2}\left\|\sigma^1 - \pi^*\right\|^2
\end{aligned}$$

$$
- \frac{2(k(t)+1)^3}{k(t)+2} \left\|\sigma^{k(t)+1} - \pi^{\mu,\sigma^{k(t)}}\right\|^2
$$

$$
\geq \frac{(k(t)+1)(k(t)+2)}{4} \left\|\pi^{\mu,\sigma^{k(t)}} - \hat{\sigma}^{k(t)}\right\|^2 - 2\left\|\sigma^1 - \pi^*\right\|^2 - 2(k(t)+1)^2 \left\|\sigma^{k(t)+1} - \pi^{\mu,\sigma^{k(t)}}\right\|^2. \tag{20}
$$

Here, we introduce the following lemma:

Lemma B.3. *For any $t \geq 1$ such that $k(t) \geq 2$, we have:*

$$
\begin{aligned}
&\frac{(k(t)+1)(k(t)+2)}{2} \left\|\pi^{\mu,\sigma^{k(t)}} - \hat{\sigma}^{k(t)}\right\|^2 + (k(t)+1)(k(t)+2)\left\langle \hat{\sigma}^{k(t)+1} - \pi^{\mu,\sigma^{k(t)}}, \pi^{\mu,\sigma^{k(t)}} - \hat{\sigma}^{k(t)}\right\rangle \\
&\leq \frac{k(t)(k(t)+1)}{2} \left\|\pi^{\mu,\sigma^{k(t)-1}} - \hat{\sigma}^{k(t)-1}\right\|^2 + k(t)(k(t)+1)\left\langle \hat{\sigma}^{k(t)} - \pi^{\mu,\sigma^{k(t)-1}}, \pi^{\mu,\sigma^{k(t)-1}} - \hat{\sigma}^{k(t)-1}\right\rangle \\
&\quad + (k(t)+1)\left\langle (k(t)+1)(\pi^{\mu,\sigma^{k(t)}} - \sigma^{k(t)+1}) + k(t)(\sigma^{k(t)} - \pi^{\mu,\sigma^{k(t)-1}}), \hat{\sigma}^{k(t)} - \pi^{\mu,\sigma^{k(t)}}\right\rangle.
\end{aligned}
$$

By combining ([20]) and Lemma B.3, we get:

$$
\begin{aligned}
&\frac{(k(t)+1)(k(t)+2)}{4} \left\|\pi^{\mu,\sigma^{k(t)}} - \hat{\sigma}^{k(t)}\right\|^2 \\
&\leq \frac{(k(t)+1)(k(t)+2)}{2} \left\|\pi^{\mu,\sigma^{k(t)}} - \hat{\sigma}^{k(t)}\right\|^2 + (k(t)+1)(k(t)+2)\left\langle \hat{\sigma}^{k(t)+1} - \pi^{\mu,\sigma^{k(t)}}, \pi^{\mu,\sigma^{k(t)}} - \hat{\sigma}^{k(t)}\right\rangle \\
&\quad + 2\left\|\sigma^1 - \pi^*\right\|^2 + 2(k(t)+1)^2 \left\|\sigma^{k(t)+1} - \pi^{\mu,\sigma^{k(t)}}\right\|^2 \\
&\leq 3\left\|\pi^{\mu,\sigma^1} - \hat{\sigma}^1\right\|^2 + 6\left\langle \hat{\sigma}^2 - \pi^{\mu,\sigma^1}, \pi^{\mu,\sigma^1} - \hat{\sigma}^1\right\rangle + 2\left\|\sigma^1 - \pi^*\right\|^2 + 2(k(t)+1)^2 \left\|\sigma^{k(t)+1} - \pi^{\mu,\sigma^{k(t)}}\right\|^2 \\
&\quad + \sum_{l=2}^{k(t)} (l+1)\left\langle (l+1)(\pi^{\mu,\sigma^l} - \sigma^{l+1}) + l(\sigma^l - \pi^{\mu,\sigma^{l-1}}), \hat{\sigma}^l - \pi^{\mu,\sigma^l}\right\rangle \\
&= 3\left\|\pi^{\mu,\sigma^1} - \sigma^1\right\|^2 + 2\left\langle 2\sigma^2 + \sigma^1 - 3\pi^{\mu,\sigma^1}, \pi^{\mu,\sigma^1} - \sigma^1\right\rangle + 2\left\|\sigma^1 - \pi^*\right\|^2 \\
&\quad + 2(k(t)+1)^2 \left\|\sigma^{k(t)+1} - \pi^{\mu,\sigma^{k(t)}}\right\|^2 + \sum_{l=2}^{k(t)} (l+1)\left\langle (l+1)(\pi^{\mu,\sigma^l} - \sigma^{l+1}) + l(\sigma^l - \pi^{\mu,\sigma^{l-1}}), \hat{\sigma}^l - \pi^{\mu,\sigma^l}\right\rangle \\
&= 3\left\|\pi^{\mu,\sigma^1} - \sigma^1\right\|^2 + 2\left\langle \sigma^1 - \pi^{\mu,\sigma^1}, \pi^{\mu,\sigma^1} - \sigma^1\right\rangle + 4\left\langle \sigma^2 - \pi^{\mu,\sigma^1}, \pi^{\mu,\sigma^1} - \sigma^1\right\rangle \\
&\quad + 2\left\|\sigma^1 - \pi^*\right\|^2 + 2(k(t)+1)^2 \left\|\sigma^{k(t)+1} - \pi^{\mu,\sigma^{k(t)}}\right\|^2 \\
&\quad + \sum_{l=2}^{k(t)} (l+1)\left\langle (l+1)(\pi^{\mu,\sigma^l} - \sigma^{l+1}) + l(\sigma^l - \pi^{\mu,\sigma^{l-1}}), \hat{\sigma}^l - \pi^{\mu,\sigma^l}\right\rangle \\
&= \left\|\pi^{\mu,\sigma^1} - \sigma^1\right\|^2 + 4\left\langle \sigma^2 - \pi^{\mu,\sigma^1}, \pi^{\mu,\sigma^1} - \sigma^1\right\rangle + 2\left\|\sigma^1 - \pi^*\right\|^2 + 2(k(t)+1)^2 \left\|\sigma^{k(t)+1} - \pi^{\mu,\sigma^{k(t)}}\right\|^2 \\
&\quad + \sum_{l=2}^{k(t)} (l+1)\left\langle (l+1)(\pi^{\mu,\sigma^l} - \sigma^{l+1}) + l(\sigma^l - \pi^{\mu,\sigma^{l-1}}), \hat{\sigma}^l - \pi^{\mu,\sigma^l}\right\rangle \\
&= \left\|\pi^{\mu,\sigma^1} - \sigma^1\right\|^2 + 2\left\|\sigma^1 - \pi^*\right\|^2 + 2(k(t)+1)^2 \left\|\sigma^{k(t)+1} - \pi^{\mu,\sigma^{k(t)}}\right\|^2 \\
&\quad + \sum_{l=1}^{k(t)} (l+1)^2\left\langle \pi^{\mu,\sigma^l} - \sigma^{l+1}, \hat{\sigma}^l - \pi^{\mu,\sigma^l}\right\rangle + \sum_{l=2}^{k(t)} l(l+1)\left\langle \sigma^l - \pi^{\mu,\sigma^{l-1}}, \hat{\sigma}^l - \pi^{\mu,\sigma^l}\right\rangle \\
&\leq 3D^2 + 2(k(t)+1)^2 \left\|\sigma^{k(t)+1} - \pi^{\mu,\sigma^{k(t)}}\right\|^2 + 2D(k(t)+1)^2 \sum_{l=1}^{k(t)} \left\|\pi^{\mu,\sigma^l} - \sigma^{l+1}\right\|.
\end{aligned}
$$

Therefore, we have for any $t \geq 1$ such that $k(t) \geq 2$:

$$\left\|\pi^{\mu,\sigma^{k(t)}} - \hat{\sigma}^{k(t)}\right\|^2 \leq \frac{12D^2}{(k(t)+1)^2} + 8\left\|\sigma^{k(t)+1} - \pi^{\mu,\sigma^{k(t)}}\right\|^2 + 8D\sum_{l=1}^{k(t)}\left\|\pi^{\mu,\sigma^l} - \sigma^{l+1}\right\|.$$

By the definition of $\hat{\sigma}^{k(t)}$,

$$\begin{aligned}&\left\|\pi^{\mu,\sigma^{k(t)}} - \sigma^{k(t)}\right\|^2 + \frac{\left\|\sigma^{k(t)} - \sigma^1\right\|^2}{(k(t)+1)^2} + \frac{2}{k(t)+1}\left\langle \pi^{\mu,\sigma^{k(t)}} - \sigma^{k(t)}, \sigma^{k(t)} - \sigma^1\right\rangle \\ &\leq \frac{12D^2}{(k(t)+1)^2} + 8\left\|\sigma^{k(t)+1} - \pi^{\mu,\sigma^{k(t)}}\right\|^2 + 8D\sum_{l=1}^{k(t)}\left\|\pi^{\mu,\sigma^l} - \sigma^{l+1}\right\|.\end{aligned}$$

Therefore, from Cauchy-Schwarz inequality, we have:

$$\begin{aligned}&\left\|\pi^{\mu,\sigma^{k(t)}} - \sigma^{k(t)}\right\|^2 \\ &\leq \frac{2}{k(t)+1}\left\langle \pi^{\mu,\sigma^{k(t)}} - \sigma^{k(t)}, \sigma^1 - \sigma^{k(t)}\right\rangle + \frac{12D^2}{(k(t)+1)^2} \\ &\quad + 8\left\|\sigma^{k(t)+1} - \pi^{\mu,\sigma^{k(t)}}\right\|^2 + 8D\sum_{l=1}^{k(t)}\left\|\pi^{\mu,\sigma^l} - \sigma^{l+1}\right\| \\ &\leq \frac{2D}{k(t)+1}\left\|\pi^{\mu,\sigma^{k(t)}} - \sigma^{k(t)}\right\| + \frac{12D^2}{(k(t)+1)^2} + 8\left\|\sigma^{k(t)+1} - \pi^{\mu,\sigma^{k(t)}}\right\|^2 + 8D\sum_{l=1}^{k(t)}\left\|\pi^{\mu,\sigma^l} - \sigma^{l+1}\right\|.\end{aligned} \tag{21}$$

Furthermore, from Lemma B.1, we have for any $k \geq 1$:

$$\left\|\pi^{\mu,\sigma^k} - \sigma^{k+1}\right\|^2 \leq \left(\frac{1}{1+\eta\mu}\right)^{T_\sigma}\left\|\pi^{\mu,\sigma^k} - \sigma^k\right\|^2. \tag{22}$$

Combining (21) nad (22), we have for any $t \geq 1$ such that $k(t) \geq 2$:

$$\begin{aligned}\left\|\pi^{\mu,\sigma^{k(t)}} - \sigma^{k(t)}\right\|^2 &\leq \frac{2D}{k(t)+1}\left\|\pi^{\mu,\sigma^{k(t)}} - \sigma^{k(t)}\right\| + \frac{12D^2}{(k(t)+1)^2} \\ &\quad + 8\left(\frac{1}{1+\eta\mu}\right)^{T_\sigma}\left\|\pi^{\mu,\sigma^{k(t)}} - \sigma^{k(t)}\right\|^2 + 8D^2k(t)\left(\frac{1}{1+\eta\mu}\right)^{\frac{T_\sigma}{2}}.\end{aligned}$$

Therefore, since $T_\sigma \geq \max(1, \frac{6\ln 3(T+1)}{\ln(1+\eta\mu)}) \Rightarrow \left(\frac{1}{1+\eta\mu}\right)^{T_\sigma} \leq \frac{(k(t)+1)^3}{(1+\eta\mu)^{T_\sigma}} \leq \frac{1}{16}$, we have for $k(t) \geq 2$:

$$\frac{1}{2}\left(\left\|\pi^{\mu,\sigma^{k(t)}} - \sigma^{k(t)}\right\| - \frac{2D}{k(t)+1}\right)^2 \leq \frac{2D^2}{(k(t)+1)^2} + \frac{12D^2}{(k(t)+1)^2} + \frac{D^2}{2(k(t)+1)^2} \leq \frac{16D^2}{(k(t)+1)^2},$$

and then:

$$\left\|\pi^{\mu,\sigma^{k(t)}} - \sigma^{k(t)}\right\| \leq \frac{2D}{k(t)+1} + \frac{4\sqrt{2}D}{k(t)+1} \leq \frac{8D}{k(t)+1}.$$

On the other hand, for $k(t) = 1$, we have:

$$\left\|\pi^{\mu,\sigma^1} - \sigma^1\right\| \leq D \leq \frac{8D}{1+1}.$$

In summary, for any $t \geq 1$, we have:

$$\left\|\pi^{\mu,\sigma^{k(t)}} - \sigma^{k(t)}\right\| \leq \frac{8D}{k(t)+1}.$$

□

B.4 Proof of Lemma B.3

Proof of Lemma B.3. From the first-order optimality condition for $\pi^{\mu,\sigma^{k(t)}}$, we have:

$$\left\langle V(\pi^{\mu,\sigma^{k(t)}}) - \mu(\pi^{\mu,\sigma^{k(t)}} - \hat{\sigma}^{k(t)}), \pi^{\mu,\sigma^{k(t)}} - \pi^{\mu,\sigma^{k(t)-1}} \right\rangle \geq 0.$$

Similarly, from the first-order optimality condition for $\pi^{\mu,\sigma^{k(t)-1}}$, we have:

$$\left\langle V(\pi^{\mu,\sigma^{k(t)-1}}) - \mu(\pi^{\mu,\sigma^{k(t)-1}} - \hat{\sigma}^{k(t)-1}), \pi^{\mu,\sigma^{k(t)-1}} - \pi^{\mu,\sigma^{k(t)}} \right\rangle \geq 0.$$

Summing up these inequalities, we get for any $t \geq 1$ such that $k(t) \geq 2$:

$$\begin{aligned}
0 &\leq \left\langle V(\pi^{\mu,\sigma^{k(t)}}) - V(\pi^{\mu,\sigma^{k(t)-1}}), \pi^{\mu,\sigma^{k(t)}} - \pi^{\mu,\sigma^{k(t)-1}} \right\rangle - \mu \left\langle \pi^{\mu,\sigma^{k(t)}} - \hat{\sigma}^{k(t)}, \pi^{\mu,\sigma^{k(t)}} - \pi^{\mu,\sigma^{k(t)-1}} \right\rangle \\
&\quad + \mu \left\langle \hat{\sigma}^{k(t)-1} - \pi^{\mu,\sigma^{k(t)-1}}, \pi^{\mu,\sigma^{k(t)-1}} - \pi^{\mu,\sigma^{k(t)}} \right\rangle \\
&\leq -\mu \left\langle \pi^{\mu,\sigma^{k(t)}} - \hat{\sigma}^{k(t)}, \pi^{\mu,\sigma^{k(t)}} - \pi^{\mu,\sigma^{k(t)-1}} \right\rangle + \mu \left\langle \hat{\sigma}^{k(t)-1} - \pi^{\mu,\sigma^{k(t)-1}}, \pi^{\mu,\sigma^{k(t)-1}} - \pi^{\mu,\sigma^{k(t)}} \right\rangle \\
&= -\mu \left\langle \pi^{\mu,\sigma^{k(t)}} - \sigma^{k(t)} + \sigma^{k(t)} - \hat{\sigma}^{k(t)}, \pi^{\mu,\sigma^{k(t)}} - \sigma^{k(t)} + \sigma^{k(t)} - \pi^{\mu,\sigma^{k(t)-1}} \right\rangle \\
&\quad + \mu \left\langle \hat{\sigma}^{k(t)-1} - \pi^{\mu,\sigma^{k(t)-1}}, \pi^{\mu,\sigma^{k(t)-1}} - \pi^{\mu,\sigma^{k(t)}} \right\rangle \\
&= -\mu \left\| \pi^{\mu,\sigma^{k(t)}} - \sigma^{k(t)} \right\|^2 - \mu \left\langle \pi^{\mu,\sigma^{k(t)}} - \sigma^{k(t)}, \sigma^{k(t)} - \pi^{\mu,\sigma^{k(t)-1}} \right\rangle \\
&\quad - \mu \left\langle \sigma^{k(t)} - \hat{\sigma}^{k(t)}, \pi^{\mu,\sigma^{k(t)}} - \pi^{\mu,\sigma^{k(t)-1}} \right\rangle + \mu \left\langle \hat{\sigma}^{k(t)-1} - \pi^{\mu,\sigma^{k(t)-1}}, \pi^{\mu,\sigma^{k(t)-1}} - \pi^{\mu,\sigma^{k(t)}} \right\rangle \\
&= -\mu \left\| \pi^{\mu,\sigma^{k(t)}} - \sigma^{k(t)} \right\|^2 - \mu \left\langle \pi^{\mu,\sigma^{k(t)}} - \sigma^{k(t)}, \sigma^{k(t)} - \pi^{\mu,\sigma^{k(t)-1}} \right\rangle \\
&\quad + \mu \left\langle \pi^{\mu,\sigma^{k(t)}} - \pi^{\mu,\sigma^{k(t)-1}}, \pi^{\mu,\sigma^{k(t)-1}} - \sigma^{k(t)} \right\rangle + \mu \left\langle \hat{\sigma}^{k(t)-1} - \hat{\sigma}^{k(t)}, \pi^{\mu,\sigma^{k(t)-1}} - \pi^{\mu,\sigma^{k(t)}} \right\rangle.
\end{aligned}$$

Here, for any vectors a, b, c, it holds that:

$$\begin{aligned}
\langle a - b, b - c \rangle &= \frac{1}{2}\|a - c\|^2 - \frac{1}{2}\|b - c\|^2 - \frac{1}{2}\|a - b\|^2, \\
\langle a - b, c - d \rangle &= \frac{1}{2}\|a - b\|^2 + \frac{1}{2}\|c - d\|^2 - \frac{1}{2}\|d - c + a - b\|^2.
\end{aligned}$$

Thus, we have:

$$\begin{aligned}
0 &\leq -\mu \left\| \pi^{\mu,\sigma^{k(t)}} - \sigma^{k(t)} \right\|^2 - \frac{\mu}{2} \left\| \pi^{\mu,\sigma^{k(t)}} - \pi^{\mu,\sigma^{k(t)-1}} \right\|^2 \\
&\quad + \frac{\mu}{2} \left\| \pi^{\mu,\sigma^{k(t)-1}} - \sigma^{k(t)} \right\|^2 + \frac{\mu}{2} \left\| \pi^{\mu,\sigma^{k(t)}} - \sigma^{k(t)} \right\|^2 \\
&\quad + \frac{\mu}{2} \left\| \pi^{\mu,\sigma^{k(t)}} - \sigma^{k(t)} \right\|^2 - \frac{\mu}{2} \left\| \pi^{\mu,\sigma^{k(t)-1}} - \sigma^{k(t)} \right\|^2 - \frac{\mu}{2} \left\| \pi^{\mu,\sigma^{k(t)}} - \pi^{\mu,\sigma^{k(t)-1}} \right\|^2 \\
&\quad + \frac{\mu}{2} \left\| \hat{\sigma}^{k(t)-1} - \hat{\sigma}^{k(t)} \right\|^2 + \frac{\mu}{2} \left\| \pi^{\mu,\sigma^{k(t)}} - \pi^{\mu,\sigma^{k(t)-1}} \right\|^2 \\
&\quad - \frac{\mu}{2} \left\| \pi^{\mu,\sigma^{k(t)}} - \pi^{\mu,\sigma^{k(t)-1}} + \hat{\sigma}^{k(t)-1} + \hat{\sigma}^{k(t)} \right\|^2 \\
&= -\frac{\mu}{2} \left\| \pi^{\mu,\sigma^{k(t)}} - \pi^{\mu,\sigma^{k(t)-1}} \right\|^2 + \frac{\mu}{2} \left\| \hat{\sigma}^{k(t)} - \hat{\sigma}^{k(t)-1} \right\|^2 \\
&\quad - \frac{\mu}{2} \left\| \pi^{\mu,\sigma^{k(t)}} - \pi^{\mu,\sigma^{k(t)-1}} + \hat{\sigma}^{k(t)-1} + \hat{\sigma}^{k(t)} \right\|^2 \\
&\leq -\frac{\mu}{2} \left\| \pi^{\mu,\sigma^{k(t)}} - \pi^{\mu,\sigma^{k(t)-1}} \right\|^2 + \frac{\mu}{2} \left\| \hat{\sigma}^{k(t)} - \hat{\sigma}^{k(t)-1} \right\|^2 \\
&= -\frac{\mu}{2} \left\| \pi^{\mu,\sigma^{k(t)}} - \pi^{\mu,\sigma^{k(t)-1}} \right\|^2 + \frac{\mu}{2} \left\| \hat{\sigma}^{k(t)} - \pi^{\mu,\sigma^{k(t)-1}} + \pi^{\mu,\sigma^{k(t)-1}} - \hat{\sigma}^{k(t)-1} \right\|^2
\end{aligned}$$

$$
\begin{aligned}
&= \frac{\mu}{2}\left\|\pi^{\mu,\sigma^{k(t)-1}} - \hat{\sigma}^{k(t)-1}\right\|^2 + \frac{\mu}{2}\left\|\hat{\sigma}^{k(t)} - \pi^{\mu,\sigma^{k(t)-1}}\right\|^2 - \frac{\mu}{2}\left\|\pi^{\mu,\sigma^{k(t)}} - \pi^{\mu,\sigma^{k(t)-1}}\right\|^2 \\
&\quad + \mu\left\langle \hat{\sigma}^{k(t)} - \pi^{\mu,\sigma^{k(t)-1}}, \pi^{\mu,\sigma^{k(t)-1}} - \hat{\sigma}^{k(t)-1}\right\rangle. \qquad (23)
\end{aligned}
$$

Here, from the definition of $\hat{\sigma}^{k(t)}$, we have:

$$
\begin{aligned}
&\frac{1}{2}\left\|\hat{\sigma}^{k(t)} - \pi^{\mu,\sigma^{k(t)-1}}\right\|^2 - \frac{1}{2}\left\|\pi^{\mu,\sigma^{k(t)}} - \pi^{\mu,\sigma^{k(t)-1}}\right\|^2 \\
&= \frac{1}{2}\left\|\frac{k(t)\sigma^{k(t)} + \sigma^1}{k(t)+1} - \pi^{\mu,\sigma^{k(t)-1}}\right\|^2 - \frac{1}{2}\left\|\pi^{\mu,\sigma^{k(t)}} - \pi^{\mu,\sigma^{k(t)-1}}\right\|^2 \\
&= \frac{1}{2}\left\langle \frac{k(t)\sigma^{k(t)} + \sigma^1}{k(t)+1} - \pi^{\mu,\sigma^{k(t)-1}} + \pi^{\mu,\sigma^{k(t)}} - \pi^{\mu,\sigma^{k(t)-1}}, \frac{k(t)\sigma^{k(t)} + \sigma^1}{k(t)+1} - \pi^{\mu,\sigma^{k(t)-1}} - \pi^{\mu,\sigma^{k(t)}} + \pi^{\mu,\sigma^{k(t)-1}}\right\rangle \\
&= \frac{1}{2}\left\langle \frac{\sigma^1 + (k(t)+1)\pi^{\mu,\sigma^{k(t)}} - 2(k(t)+1)\pi^{\mu,\sigma^{k(t)-1}} + k(t)\sigma^{k(t)}}{k(t)+1}, \hat{\sigma}^{k(t)} - \pi^{\mu,\sigma^{k(t)}}\right\rangle \\
&= \frac{1}{2k(t)}\left\langle 2(k(t)+1)\sigma^{k(t)+1} + 2\sigma^1 - 2(k(t)+2)\pi^{\mu,\sigma^{k(t)}}, \hat{\sigma}^{k(t)} - \pi^{\mu,\sigma^{k(t)}}\right\rangle \\
&\quad + \frac{1}{2k(t)}\left\langle -\frac{k(t)+2}{k(t)+1}\sigma^1 + (3k(t)+4)\pi^{\mu,\sigma^{k(t)}} - 2(k(t)+1)\sigma^{k(t)+1}\hat{\sigma}^{k(t)} - \pi^{\mu,\sigma^{k(t)}}\right\rangle \\
&\quad + \frac{1}{2k(t)}\left\langle -2k(t)\pi^{\mu,\sigma^{k(t)-1}} + \frac{k(t)^2}{k(t)+1}\sigma^{k(t)}, \hat{\sigma}^{k(t)} - \pi^{\mu,\sigma^{k(t)}}\right\rangle \\
&= \frac{k(t)+2}{k(t)}\left\langle \hat{\sigma}^{k(t)+1} - \pi^{\mu,\sigma^{k(t)}}, \hat{\sigma}^{k(t)} - \pi^{\mu,\sigma^{k(t)}}\right\rangle \\
&\quad + \frac{1}{2k(t)}\left\langle -\frac{k(t)+2}{k(t)+1}\sigma^1 - \frac{k(t)(k(t)+2)}{k(t)+1}\sigma^{k(t)} + (k(t)+2)\pi^{\mu,\sigma^{k(t)}}, \hat{\sigma}^{k(t)} - \pi^{\mu,\sigma^{k(t)}}\right\rangle \\
&\quad + \frac{1}{2k(t)}\left\langle 2(k(t)+1)(\pi^{\mu,\sigma^{k(t)}} - \sigma^{k(t)+1}) + 2k(t)(\sigma^{k(t)} - \pi^{\mu,\sigma^{k(t)-1}}), \hat{\sigma}^{k(t)} - \pi^{\mu,\sigma^{k(t)}}\right\rangle \\
&= -\frac{k(t)+2}{k(t)}\left\langle \hat{\sigma}^{k(t)+1} - \pi^{\mu,\sigma^{k(t)}}, \pi^{\mu,\sigma^{k(t)}} - \hat{\sigma}^{k(t)}\right\rangle \\
&\quad - \frac{k(t)+2}{2k(t)}\left\langle \frac{k(t)\sigma^{k(t)} + \sigma^1}{k(t)+1} - \pi^{\mu,\sigma^{k(t)}}, \hat{\sigma}^{k(t)} - \pi^{\mu,\sigma^{k(t)}}\right\rangle \\
&\quad + \frac{1}{k(t)}\left\langle (k(t)+1)(\pi^{\mu,\sigma^{k(t)}} - \sigma^{k(t)+1}) + k(t)(\sigma^{k(t)} - \pi^{\mu,\sigma^{k(t)-1}}), \hat{\sigma}^{k(t)} - \pi^{\mu,\sigma^{k(t)}}\right\rangle \\
&= -\frac{k(t)+2}{k(t)}\left\langle \hat{\sigma}^{k(t)+1} - \pi^{\mu,\sigma^{k(t)}}, \pi^{\mu,\sigma^{k(t)}} - \hat{\sigma}^{k(t)}\right\rangle - \frac{k(t)+2}{2k(t)}\left\|\hat{\sigma}^{k(t)} - \pi^{\mu,\sigma^{k(t)}}\right\|^2 \\
&\quad + \frac{1}{k(t)}\left\langle (k(t)+1)(\pi^{\mu,\sigma^{k(t)}} - \sigma^{k(t)+1}) + k(t)(\sigma^{k(t)} - \pi^{\mu,\sigma^{k(t)-1}}), \hat{\sigma}^{k(t)} - \pi^{\mu,\sigma^{k(t)}}\right\rangle. \qquad (24)
\end{aligned}
$$

Combining (23) and (24) yields for any $t \geq 1$ such that $k(t) \geq 2$:

$$
\begin{aligned}
&\frac{k(t)+2}{2k(t)}\left\|\pi^{\mu,\sigma^{k(t)}} - \hat{\sigma}^{k(t)}\right\|^2 + \frac{k(t)+2}{k(t)}\left\langle \hat{\sigma}^{k(t)+1} - \pi^{\mu,\sigma^{k(t)}}, \pi^{\mu,\sigma^{k(t)}} - \hat{\sigma}^{k(t)}\right\rangle \\
&\leq \frac{1}{2}\left\|\pi^{\mu,\sigma^{k(t)-1}} - \hat{\sigma}^{k(t)-1}\right\|^2 + \left\langle \hat{\sigma}^{k(t)} - \pi^{\mu,\sigma^{k(t)-1}}, \pi^{\mu,\sigma^{k(t)-1}} - \hat{\sigma}^{k(t)-1}\right\rangle \\
&\quad + \frac{1}{k(t)}\left\langle (k(t)+1)(\pi^{\mu,\sigma^{k(t)}} - \sigma^{k(t)+1}) + k(t)(\sigma^{k(t)} - \pi^{\mu,\sigma^{k(t)-1}}), \hat{\sigma}^{k(t)} - \pi^{\mu,\sigma^{k(t)}}\right\rangle.
\end{aligned}
$$

Multiplying both sides by $k(t)(k(t)+1)$, we have:

$$
\frac{(k(t)+1)(k(t)+2)}{2}\left\|\pi^{\mu,\sigma^{k(t)}} - \hat{\sigma}^{k(t)}\right\|^2 + (k(t)+1)(k(t)+2)\left\langle \hat{\sigma}^{k(t)+1} - \pi^{\mu,\sigma^{k(t)}}, \pi^{\mu,\sigma^{k(t)}} - \hat{\sigma}^{k(t)}\right\rangle
$$

$$\leq \frac{k(t)(k(t)+1)}{2}\left\|\pi^{\mu,\sigma^{k(t)-1}}-\hat{\sigma}^{k(t)-1}\right\|^2+k(t)(k(t)+1)\left\langle\hat{\sigma}^{k(t)}-\pi^{\mu,\sigma^{k(t)-1}},\pi^{\mu,\sigma^{k(t)-1}}-\hat{\sigma}^{k(t)-1}\right\rangle$$
$$+(k(t)+1)\left\langle(k(t)+1)(\pi^{\mu,\sigma^{k(t)}}-\sigma^{k(t)+1})+k(t)(\sigma^{k(t)}-\pi^{\mu,\sigma^{k(t)-1}}),\hat{\sigma}^{k(t)}-\pi^{\mu,\sigma^{k(t)}}\right\rangle.$$

□

C Proofs for Theorem 4.3

C.1 Proof of Theorem 4.3

Proof of Theorem 4.3. Let us define $K := \frac{T}{T_\sigma}$. We can decompose the gap function for π^{T+1} as follows:

$$\begin{aligned}
&\mathrm{GAP}(\pi^{T+1})\\
&=\max_{x\in\mathcal{X}}\langle V(\pi^{T+1}),x-\pi^{T+1}\rangle\\
&=\max_{x\in\mathcal{X}}\left(\left\langle V(\pi^{\mu,\sigma^K}),x-\pi^{\mu,\sigma^K}\right\rangle-\left\langle V(\pi^{\mu,\sigma^K}),x-\pi^{\mu,\sigma^K}\right\rangle+\langle V(\pi^{T+1}),x-\pi^{T+1}\rangle\right)\\
&=\max_{x\in\mathcal{X}}\left(\left\langle V(\pi^{\mu,\sigma^K}),x-\pi^{\mu,\sigma^K}\right\rangle-\left\langle V(\pi^{\mu,\sigma^K})-V(\pi^{T+1}),x-\pi^{T+1}\right\rangle+\left\langle V(\pi^{\mu,\sigma^K}),\pi^{\mu,\sigma^K}-\pi^{T+1}\right\rangle\right)\\
&\leq\max_{x\in\mathcal{X}}\left(\left\langle V(\pi^{\mu,\sigma^K}),x-\pi^{\mu,\sigma^K}\right\rangle+D\left\|V(\pi^{\mu,\sigma^K})-V(\pi^{T+1})\right\|+\zeta\left\|\pi^{\mu,\sigma^K}-\pi^{T+1}\right\|\right)\\
&\leq\mathrm{GAP}(\pi^{\mu,\sigma^K})+(LD+\zeta)\left\|\pi^{\mu,\sigma^K}-\pi^{T+1}\right\|\\
&\leq D\cdot\min_{c\in N_{\mathcal{X}}(\pi^{\mu,\sigma^K})}\left\|-V(\pi^{\mu,\sigma^K})+c\right\|+(LD+\zeta)\left\|\pi^{\mu,\sigma^K}-\pi^{T+1}\right\|,
\end{aligned}$$

where the last inequality follows from Lemma 2.2. From the first-order optimality condition for π^{μ,σ^K}, we have for any $x\in\mathcal{X}$:

$$\left\langle V(\pi^{\mu,\sigma^K})-\mu\left(\pi^{\mu,\sigma^K}-\frac{K\sigma^K+\sigma^1}{K+1}\right),\pi^{\mu,\sigma^K}-x\right\rangle\geq 0,$$

and then $V(\pi^{\mu,\sigma^K})-\mu\left(\pi^{\mu,\sigma^K}-\frac{K\sigma^K+\sigma^1}{K+1}\right)\in N_{\mathcal{X}}(\pi^{\mu,\sigma^K})$. Thus, the gap function for π^{T+1} can be bounded by:

$$\begin{aligned}
\mathrm{GAP}(\pi^{T+1})&\leq\mu D\cdot\left\|\pi^{\mu,\sigma^K}-\frac{K\sigma^K+\sigma^1}{K+1}\right\|+(LD+\zeta)\left\|\pi^{\mu,\sigma^K}-\pi^{T+1}\right\|\\
&=\mu D\cdot\left\|\frac{\sigma^K-\sigma^1}{K+1}+\pi^{\mu,\sigma^K}-\sigma^K\right\|+(LD+\zeta)\left\|\pi^{\mu,\sigma^K}-\pi^{T+1}\right\|\\
&\leq\mu D\cdot\left(\frac{D}{K+1}+\left\|\pi^{\mu,\sigma^K}-\sigma^K\right\|\right)+(LD+\zeta)\left\|\pi^{\mu,\sigma^K}-\pi^{T+1}\right\|.
\end{aligned}$$

Taking its expectation yields:

$$\begin{aligned}
\mathbb{E}\left[\mathrm{GAP}(\pi^{T+1})\right]&\leq\frac{\mu D^2}{K+1}+\mu D\cdot\mathbb{E}\left[\left\|\pi^{\mu,\sigma^K}-\sigma^K\right\|\right]+(LD+\zeta)\cdot\mathbb{E}\left[\left\|\pi^{\mu,\sigma^K}-\pi^{T+1}\right\|\right]\\
&\leq\frac{\mu D^2}{K+1}+\mu D\cdot\mathbb{E}\left[\left\|\pi^{\mu,\sigma^K}-\sigma^K\right\|\right]+(LD+\zeta)\cdot\sqrt{\mathbb{E}\left[\left\|\pi^{\mu,\sigma^K}-\pi^{T+1}\right\|^2\right]}.
\end{aligned}\tag{25}$$

Here, we derive the following upper bound on $\mathbb{E}\left[\left\|\pi^{\mu,\sigma^{k(t)}}-\pi^{t+1}\right\|^2\right]$:

Lemma C.1. *Let $\kappa=\frac{\mu}{2},\theta=\frac{3\mu^2+8L^2}{2\mu}$. Suppose that Assumption 4.2 holds. If we set $\eta_t=\frac{1}{\kappa(t-T_\sigma(k(t)-1))+2\theta}$, we have for any $t\geq 1$:*

$$\mathbb{E}\left[\left\|\pi^{\mu,\sigma^{k(t)}}-\pi^{t+1}\right\|^2\right]\leq\frac{2\theta}{\kappa\left(t-(k(t)-1)T_\sigma\right)+2\theta}\left(D^2+\frac{C^2}{\kappa\theta}\ln\left(\frac{\kappa\left(t-(k(t)-1)T_\sigma\right)}{2\theta}+1\right)\right).$$

Setting $t = T = KT_\sigma$, we can write $k(t) = \lfloor \frac{KT_\sigma - 1}{T_\sigma} \rfloor + 1 = K$. Therefore, from Lemma C.1, we have:

$$\mathbb{E}\left[\left\|\pi^{\mu,\sigma^K} - \pi^{T+1}\right\|^2\right] \leq \frac{2\theta}{\kappa T_\sigma + 2\theta}\left(D^2 + \frac{C^2}{\kappa\theta}\ln\left(\frac{\kappa T_\sigma}{2\theta} + 1\right)\right). \tag{26}$$

On the other hand, in terms of $\mathbb{E}\left[\left\|\pi^{\mu,\sigma^{k(t)}} - \sigma^{k(t)}\right\|\right]$, we introduce the following lemma:

Lemma C.2. *If we set $\eta_t = \frac{1}{\kappa(t - T_\sigma(k(t)-1)) + 2\theta}$ and $T_\sigma \geq \max(1, T^{\frac{6}{7}})$, we have for any $t \geq 1$:*

$$\mathbb{E}\left[\left\|\pi^{\mu,\sigma^{k(t)}} - \sigma^{k(t)}\right\|\right] \leq \frac{6\left(\sqrt{\kappa} + \sqrt{\theta} + \sqrt{D\theta} + \sqrt{D}\right)}{k(t)}\left(\sqrt{\frac{1}{\kappa}\left(D^2 + \frac{C^2}{\kappa\theta}\ln\left(\frac{\kappa T}{2\theta} + 1\right)\right)} + 1\right).$$

By setting $t = KT_\sigma$ in this lemma, we get:

$$\mathbb{E}\left[\left\|\pi^{\mu,\sigma^K} - \sigma^K\right\|\right] \leq \frac{6\left(\sqrt{\kappa} + \sqrt{\theta} + \sqrt{D\theta} + \sqrt{D}\right)}{K}\left(\sqrt{\frac{1}{\kappa}\left(D^2 + \frac{C^2}{\kappa\theta}\ln\left(\frac{\kappa T}{2\theta} + 1\right)\right)} + 1\right). \tag{27}$$

Combining (25), (26), and (27), we have:

$$\begin{aligned}
&\mathbb{E}\left[\mathrm{GAP}(\sigma^{K+1})\right] \\
&\leq \frac{\mu D^2}{K+1} + \mu D \cdot \frac{6\left(\sqrt{\kappa} + \sqrt{\theta} + \sqrt{D\theta} + \sqrt{D}\right)}{K}\left(\sqrt{\frac{1}{\kappa}\left(D^2 + \frac{C^2}{\kappa\theta}\ln\left(\frac{\kappa T}{2\theta} + 1\right)\right)} + 1\right) \\
&\quad + (LD + \zeta) \cdot \sqrt{\frac{2\theta}{\kappa T_\sigma + 2\theta}\left(D^2 + \frac{C^2}{\kappa\theta}\ln\left(\frac{\kappa T_\sigma}{2\theta} + 1\right)\right)} \\
&\leq \mu D^2 \frac{T_\sigma}{T} + \mu D \cdot \frac{6T_\sigma\left(\sqrt{\kappa} + \sqrt{\theta} + \sqrt{D\theta} + \sqrt{D}\right)}{T}\left(\sqrt{\frac{1}{\kappa}\left(D^2 + \frac{C^2}{\kappa\theta}\ln\left(\frac{\kappa T}{2\theta} + 1\right)\right)} + 1\right) \\
&\quad + (LD + \zeta) \cdot \sqrt{\frac{2\theta}{\kappa T_\sigma}\left(D^2 + \frac{C^2}{\kappa\theta}\ln\left(\frac{\kappa T}{2\theta} + 1\right)\right)},
\end{aligned}$$

where the second inequality follows from $K = \frac{T}{T_\sigma}$. Finally, since $T_\sigma = c \cdot \max(1, T^{\frac{6}{7}})$, we have for any $T \geq T_\sigma$:

$$\begin{aligned}
&\mathbb{E}\left[\mathrm{GAP}(\sigma^{K+1})\right] \\
&\leq \frac{c\mu D^2}{T^{\frac{1}{7}}} + \frac{6c\mu D\left(\sqrt{\kappa} + \sqrt{\theta} + \sqrt{D\theta} + \sqrt{D}\right)}{T^{\frac{1}{7}}}\left(\sqrt{\frac{1}{\kappa}\left(D^2 + \frac{C^2}{\kappa\theta}\ln\left(\frac{\kappa T}{2\theta} + 1\right)\right)} + 1\right) \\
&\quad + \frac{(LD + \zeta)}{T^{\frac{3}{7}}}\sqrt{\frac{2\theta}{\kappa}\left(D^2 + \frac{C^2}{\kappa\theta}\ln\left(\frac{\kappa T}{2\theta} + 1\right)\right)} \\
&\leq \frac{6c\mu D\left(\sqrt{\kappa} + \sqrt{\theta} + \sqrt{D\theta} + \sqrt{D} + D\right)}{T^{\frac{1}{7}}}\left(\sqrt{\frac{1}{\kappa}\left(D^2 + \frac{C^2}{\kappa\theta}\ln\left(\frac{\kappa T}{2\theta} + 1\right)\right)} + 1\right) \\
&\quad + \frac{(LD + \zeta)\sqrt{2\theta}}{T^{\frac{1}{7}}}\left(\sqrt{\frac{1}{\kappa}\left(D^2 + \frac{C^2}{\kappa\theta}\ln\left(\frac{\kappa T}{2\theta} + 1\right)\right)} + 1\right) \\
&\leq \frac{9c\left(\mu D + LD + \zeta\right)\left(\sqrt{\kappa} + \sqrt{\theta} + \sqrt{D\theta} + \sqrt{D} + D\right)}{T^{\frac{1}{7}}}\left(\sqrt{\frac{1}{\kappa}\left(D^2 + \frac{C^2}{\kappa\theta}\ln\left(\frac{\kappa T}{2\theta} + 1\right)\right)} + 1\right).
\end{aligned}$$

Since $T = T_\sigma K$, we have finally:

$$
\begin{aligned}
&\mathbb{E}\left[\mathrm{GAP}(\pi^{T+1})\right] \\
&\leq \frac{9c\left(\mu D + LD + \zeta\right)\left(\sqrt{\kappa} + \sqrt{\theta} + \sqrt{D\theta} + \sqrt{D} + D\right)}{T^{\frac{1}{7}}}\left(\sqrt{\frac{1}{\kappa}\left(D^2 + \frac{C^2}{\kappa\theta}\ln\left(\frac{\kappa T}{2\theta} + 1\right)\right)} + 1\right) \\
&= \frac{9c\left(D(\mu + L) + \zeta\right)\left(\sqrt{\kappa} + (\sqrt{D} + 1)(\sqrt{D} + \sqrt{\theta})\right)}{T^{\frac{1}{7}}}\left(\sqrt{\frac{1}{\kappa}\left(D^2 + \frac{C^2}{\kappa\theta}\ln\left(\frac{\kappa T}{2\theta} + 1\right)\right)} + 1\right) \\
&\leq \frac{18c\left(D(\mu + L) + \zeta\right)\left(\sqrt{\kappa} + \sqrt{(D+1)(D+\theta)}\right)}{T^{\frac{1}{7}}}\left(\sqrt{\frac{1}{\kappa}\left(D^2 + \frac{C^2}{\kappa\theta}\ln\left(\frac{\kappa T}{2\theta} + 1\right)\right)} + 1\right) \\
&\leq \frac{26c\left(D(\mu + L) + \zeta\right)\sqrt{(D+1)(D+\theta) + \kappa}}{T^{\frac{1}{7}}}\left(\sqrt{\frac{1}{\kappa}\left(D^2 + \frac{C^2}{\kappa\theta}\ln\left(\frac{\kappa T}{2\theta} + 1\right)\right)} + 1\right).
\end{aligned}
$$

□

C.2 Proof of Lemma C.1

Proof of Lemma C.1. From the first-order optimality condition for π^{t+1}, we have for $t \geq 1$:

$$
\left\langle \eta_t\left(\hat{V}(\pi^t) - \mu(\pi^t - \hat{\sigma}^{k(t)})\right) - \pi^{t+1} + \pi^t, \pi^{t+1} - \pi^{\mu,\sigma^{k(t)}}\right\rangle \geq 0. \tag{28}
$$

Combining (28), (12), and (14), we have:

$$
\begin{aligned}
&\frac{1}{2}\left\|\pi^{\mu,\sigma^{k(t)}} - \pi^{t+1}\right\|^2 - \frac{1}{2}\left\|\pi^{\mu,\sigma^{k(t)}} - \pi^t\right\|^2 + \frac{1}{2}\left\|\pi^{t+1} - \pi^t\right\|^2 \\
&\leq \eta_t\left\langle \hat{V}(\pi^t) - \mu(\pi^t - \hat{\sigma}^{k(t)}), \pi^{t+1} - \pi^{\mu,\sigma^{k(t)}}\right\rangle \\
&= \eta_t\left\langle V(\pi^{t+1}) - \mu(\pi^{t+1} - \hat{\sigma}^{k(t)}), \pi^{t+1} - \pi^{\mu,\sigma^{k(t)}}\right\rangle + \eta_t\left\langle \hat{V}(\pi^t) - V(\pi^{t+1}) - \mu(\pi^t - \pi^{t+1}), \pi^{t+1} - \pi^{\mu,\sigma^{k(t)}}\right\rangle \\
&\leq \eta_t\left\langle V(\pi^{\mu,\sigma^{k(t)}}) - \mu(\pi^{t+1} - \hat{\sigma}^{k(t)}), \pi^{t+1} - \pi^{\mu,\sigma^{k(t)}}\right\rangle + \eta_t\left\langle \hat{V}(\pi^t) - V(\pi^{t+1}) - \mu(\pi^t - \pi^{t+1}), \pi^{t+1} - \pi^{\mu,\sigma^{k(t)}}\right\rangle \\
&= \eta_t\left\langle V(\pi^{\mu,\sigma^{k(t)}}) - \mu(\pi^{\mu,\sigma^{k(t)}} - \hat{\sigma}^{k(t)}), \pi^{t+1} - \pi^{\mu,\sigma^{k(t)}}\right\rangle - \eta_t\mu\left\|\pi^{\mu,\sigma^{k(t)}} - \pi^{t+1}\right\|^2 \\
&\quad + \eta_t\left\langle V(\pi^t) - V(\pi^{t+1}), \pi^{t+1} - \pi^{\mu,\sigma^{k(t)}}\right\rangle - \eta_t\mu\left\langle \pi^t - \pi^{t+1}, \pi^{t+1} - \pi^{\mu,\sigma^{k(t)}}\right\rangle + \eta_t\left\langle \xi^t, \pi^{t+1} - \pi^{\mu,\sigma^{k(t)}}\right\rangle \\
&\leq -\eta_t\mu\left\|\pi^{\mu,\sigma^{k(t)}} - \pi^{t+1}\right\|^2 + \eta_t\mu\left\langle \pi^{t+1} - \pi^t, \pi^{t+1} - \pi^{\mu,\sigma^{k(t)}}\right\rangle \\
&\quad + \eta_t\left\langle V(\pi^t) - V(\pi^{t+1}), \pi^{t+1} - \pi^{\mu,\sigma^{k(t)}}\right\rangle + \eta_t\left\langle \xi^t, \pi^{t+1} - \pi^{\mu,\sigma^{k(t)}}\right\rangle \\
&= -\eta_t\mu\left\|\pi^{\mu,\sigma^{k(t)}} - \pi^{t+1}\right\|^2 + \frac{\eta_t\mu}{2}\left\|\pi^{t+1} - \pi^t\right\|^2 + \frac{\eta_t\mu}{2}\left\|\pi^{t+1} - \pi^{\mu,\sigma^{k(t)}}\right\|^2 - \frac{\eta_t\mu}{2}\left\|\pi^t - \pi^{\mu,\sigma^{k(t)}}\right\|^2 \\
&\quad + \eta_t\left\langle V(\pi^t) - V(\pi^{t+1}), \pi^{t+1} - \pi^{\mu,\sigma^{k(t)}}\right\rangle + \eta_t\left\langle \xi^t, \pi^{t+1} - \pi^{\mu,\sigma^{k(t)}}\right\rangle \\
&= -\frac{\eta_t\mu}{2}\left\|\pi^{t+1} - \pi^{\mu,\sigma^{k(t)}}\right\|^2 - \frac{\eta_t\mu}{2}\left\|\pi^t - \pi^{\mu,\sigma^{k(t)}}\right\|^2 + \frac{\eta_t\mu}{2}\left\|\pi^{t+1} - \pi^t\right\|^2 \\
&\quad + \eta_t\left\langle V(\pi^t) - V(\pi^{t+1}), \pi^{t+1} - \pi^{\mu,\sigma^{k(t)}}\right\rangle + \eta_t\left\langle \xi^t, \pi^{t+1} - \pi^{\mu,\sigma^{k(t)}}\right\rangle,
\end{aligned} \tag{29}
$$

where the third inequality follows from (1). From Cauchy-Schwarz inequality and Young's inequality, the fourth term on the right-hand side of this inequality can be bounded by:

$$
\begin{aligned}
&\left\langle V(\pi^t) - V(\pi^{t+1}), \pi^{t+1} - \pi^{\mu,\sigma^{k(t)}}\right\rangle \\
&\leq \left\|V(\pi^t) - V(\pi^{t+1})\right\| \cdot \left\|\pi^{t+1} - \pi^{\mu,\sigma^{k(t)}}\right\| \\
&\leq L\left\|\pi^t - \pi^{t+1}\right\| \cdot \left\|\pi^{t+1} - \pi^{\mu,\sigma^{k(t)}}\right\|
\end{aligned}
$$

$$
\begin{aligned}
&\leq \frac{2L^2}{\mu}\left\|\pi^t - \pi^{t+1}\right\|^2 + \frac{\mu}{8}\left\|\pi^{t+1} - \pi^{\mu,\sigma^{k(t)}}\right\|^2 \\
&\leq \frac{2L^2}{\mu}\left\|\pi^t - \pi^{t+1}\right\|^2 + \frac{\mu}{4}\left\|\pi^t - \pi^{\mu,\sigma^{k(t)}}\right\|^2 + \frac{\mu}{4}\left\|\pi^{t+1} - \pi^t\right\|^2 \\
&= \left(\frac{4L^2}{\mu} + \frac{\mu}{2}\right)\frac{\left\|\pi^t - \pi^{t+1}\right\|^2}{2} + \frac{\mu}{2}\frac{\left\|\pi^t - \pi^{\mu,\sigma^{k(t)}}\right\|^2}{2}. \qquad (30)
\end{aligned}
$$

By combining (29) and (30), we have:

$$
\begin{aligned}
\left\|\pi^{\mu,\sigma^{k(t)}} - \pi^{t+1}\right\|^2 &\leq -\eta_t\mu\left\|\pi^{t+1} - \pi^{\mu,\sigma^{k(t)}}\right\|^2 + \left(1 - \frac{\eta_t\mu}{2}\right)\left\|\pi^t - \pi^{\mu,\sigma^{k(t)}}\right\|^2 \\
&\quad - \left(1 - \eta_t\left(\frac{3\mu}{2} + \frac{4L^2}{\mu}\right)\right)\left\|\pi^{t+1} - \pi^t\right\|^2 + 2\eta_t\left\langle \xi^t, \pi^{t+1} - \pi^{\mu,\sigma^{k(t)}}\right\rangle \\
&\leq \left(1 - \frac{\eta_t\mu}{2}\right)\left\|\pi^t - \pi^{\mu,\sigma^{k(t)}}\right\|^2 - \left(1 - \eta_t\left(\frac{3\mu}{2} + \frac{4L^2}{\mu}\right)\right)\left\|\pi^{t+1} - \pi^t\right\|^2 \\
&\quad + 2\eta_t\left\langle \xi^t, \pi^t - \pi^{\mu,\sigma^{k(t)}}\right\rangle + 2\eta_t\left\langle \xi^t, \pi^{t+1} - \pi^t\right\rangle \\
&= (1 - \eta_t\kappa)\left\|\pi^t - \pi^{\mu,\sigma^{k(t)}}\right\|^2 - (1 - \eta_t\theta)\left\|\pi^{t+1} - \pi^t\right\|^2 \\
&\quad + 2\eta_t\left\langle \xi^t, \pi^t - \pi^{\mu,\sigma^{k(t)}}\right\rangle + 2\eta_t\left\langle \xi^t, \pi^{t+1} - \pi^t\right\rangle.
\end{aligned}
$$

By taking the expectation conditioned on $\mathcal{F}_t$ for both sides and using Assumption 4.2 (a) and (b),

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\pi^{\mu,\sigma^{k(t)}} - \pi^{t+1}\right\|^2 \mid \mathcal{F}_t\right] \\
&\leq (1 - \eta_t\kappa)\,\mathbb{E}\left[\left\|\pi^t - \pi^{\mu,\sigma^{k(t)}}\right\|^2 \mid \mathcal{F}_t\right] - (1 - \eta_t\theta)\,\mathbb{E}\left[\left\|\pi^{t+1} - \pi^t\right\|^2 \mid \mathcal{F}_t\right] \\
&\quad + 2\eta_t\left\langle \mathbb{E}\left[\xi^t \mid \mathcal{F}_t\right], \pi^t - \pi^{\mu,\sigma^{k(t)}}\right\rangle + 2\eta_t\mathbb{E}\left[\left\langle \xi^t, \pi^{t+1} - \pi^t\right\rangle \mid \mathcal{F}_t\right] \\
&= (1 - \eta_t\kappa)\left\|\pi^t - \pi^{\mu,\sigma^{k(t)}}\right\|^2 - (1 - \eta_t\theta)\,\mathbb{E}\left[\left\|\pi^{t+1} - \pi^t\right\|^2 \mid \mathcal{F}_t\right] + 2\eta_t\mathbb{E}\left[\left\langle \xi^t, \pi^{t+1} - \pi^t\right\rangle \mid \mathcal{F}_t\right] \\
&\leq (1 - \eta_t\kappa)\left\|\pi^t - \pi^{\mu,\sigma^{k(t)}}\right\|^2 - (1 - \eta_t\theta)\,\mathbb{E}\left[\left\|\pi^{t+1} - \pi^t\right\|^2 \mid \mathcal{F}_t\right] \\
&\quad + \frac{\eta_t^2}{1 - \eta_t\theta}\mathbb{E}\left[\left\|\xi^t\right\|^2 \mid \mathcal{F}_t\right] + (1 - \eta_t\theta)\,\mathbb{E}\left[\left\|\pi^{t+1} - \pi^t\right\|^2 \mid \mathcal{F}_t\right] \\
&\leq (1 - \eta_t\kappa)\left\|\pi^t - \pi^{\mu,\sigma^{k(t)}}\right\|^2 + \frac{\eta_t^2}{1 - \eta_t\theta}\mathbb{E}\left[\left\|\xi^t\right\|^2 \mid \mathcal{F}_t\right] \\
&\leq (1 - \eta_t\kappa)\left\|\pi^t - \pi^{\mu,\sigma^{k(t)}}\right\|^2 + 2\eta_t^2\mathbb{E}\left[\left\|\xi^t\right\|^2 \mid \mathcal{F}_t\right] \\
&\leq (1 - \eta_t\kappa)\left\|\pi^t - \pi^{\mu,\sigma^{k(t)}}\right\|^2 + 2\eta_t^2 C^2.
\end{aligned}
$$

Therefore, under the setting where $\eta_t = \frac{1}{\kappa(t - T_\sigma(k(t)-1)) + 2\theta}$, we have for any $t \geq 1$:

$$
\mathbb{E}\left[\left\|\pi^{\mu,\sigma^{k(t)}} - \pi^{t+1}\right\|^2 \mid \mathcal{F}_t\right] \leq \left(1 - \frac{1}{t - T_\sigma(k(t)-1) + 2\theta/\kappa}\right)\left\|\pi^t - \pi^{\mu,\sigma^{k(t)}}\right\|^2 + 2\eta_t^2 C^2.
$$

Rearranging and taking the expectations, we get:

$$
\begin{aligned}
&(t - T_\sigma(k(t)-1) + 2\theta/\kappa)\,\mathbb{E}\left[\left\|\pi^{\mu,\sigma^{k(t)}} - \pi^{t+1}\right\|^2\right] \\
&\leq (t - 1 - T_\sigma(k(t)-1) + 2\theta/\kappa)\,\mathbb{E}\left[\left\|\pi^{\mu,\sigma^{k(t)}} - \pi^t\right\|^2\right] + \frac{2C^2}{\kappa\left(\kappa(t - T_\sigma(k(t)-1)) + 2\theta\right)}.
\end{aligned}
$$

Since $k(s) = k(t)$ for any $s \in [(k(t)-1)T_\sigma + 1, T]$, telescoping the sum yields:

$$(t - T_\sigma(k(t)-1) + 2\theta/\kappa)\, \mathbb{E}\left[\left\|\pi^{\mu,\sigma^{k(t)}} - \pi^{t+1}\right\|^2\right]$$
$$\leq (s - 1 - T_\sigma(k(t)-1) + 2\theta/\kappa)\, \mathbb{E}\left[\left\|\pi^{\mu,\sigma^{k(t)}} - \pi^{s}\right\|^2\right] + \sum_{m=s}^{t} \frac{2C^2}{\kappa\left(\kappa(m - T_\sigma(k(t)-1)) + 2\theta\right)}.$$

Defining $s = (k(t)-1)T_\sigma + 1$,

$$(t - T_\sigma(k(t)-1) + 2\theta/\kappa)\, \mathbb{E}\left[\left\|\pi^{\mu,\sigma^{k(t)}} - \pi^{t+1}\right\|^2\right]$$
$$\leq \frac{2\theta}{\kappa}\mathbb{E}\left[\left\|\pi^{\mu,\sigma^{k(t)}} - \pi^{(k(t)-1)T_\sigma+1}\right\|^2\right] + \frac{2C^2}{\kappa} \sum_{m=(k(t)-1)T_\sigma+1}^{t} \frac{1}{\kappa(m - T_\sigma(k(t)-1)) + 2\theta}.$$

Therefore,

$$\mathbb{E}\left[\left\|\pi^{\mu,\sigma^{k(t)}} - \pi^{t+1}\right\|^2\right] \leq \frac{2\theta}{\kappa\left(t - T_\sigma(k(t)-1)\right) + 2\theta}\mathbb{E}\left[\left\|\pi^{\mu,\sigma^{k(t)}} - \pi^{(k(t)-1)T_\sigma+1}\right\|^2\right]$$
$$+ \frac{2C^2}{\kappa\left(t - T_\sigma(k(t)-1)\right) + 2\theta} \sum_{m=1}^{t-(k(t)-1)T_\sigma} \frac{1}{\kappa m + 2\theta}. \tag{31}$$

Here, we have:

$$\sum_{m=1}^{t-(k(t)-1)T_\sigma} \frac{1}{\kappa m + 2\theta} \leq \int_0^{t-(k(t)-1)T_\sigma} \frac{1}{\kappa x + 2\theta} dx = \frac{1}{\kappa}\ln\left(\frac{\kappa\left(t - (k(t)-1)T_\sigma\right)}{2\theta} + 1\right). \tag{32}$$

Combining (31), (32), and the fact that $\pi^{(k(t)-1)T_\sigma+1} = \sigma^{k(t)}$, we have:

$$\mathbb{E}\left[\left\|\pi^{\mu,\sigma^{k(t)}} - \pi^{t+1}\right\|^2\right]$$
$$\leq \frac{2\theta}{\kappa\left(t - (k(t)-1)T_\sigma\right) + 2\theta}\left(\mathbb{E}\left[\left\|\pi^{\mu,\sigma^{k(t)}} - \sigma^{k(t)}\right\|^2\right] + \frac{C^2}{\kappa\theta}\ln\left(\frac{\kappa\left(t - (k(t)-1)T_\sigma\right)}{2\theta} + 1\right)\right)$$
$$\leq \frac{2\theta}{\kappa\left(t - (k(t)-1)T_\sigma\right) + 2\theta}\left(D^2 + \frac{C^2}{\kappa\theta}\ln\left(\frac{\kappa\left(t - (k(t)-1)T_\sigma\right)}{2\theta} + 1\right)\right).$$

□

C.3 Proof of Lemma C.2

Proof of Lemma C.2. First, from Lemma C.1, we have for any $k \geq 1$:

$$\mathbb{E}\left[\left\|\pi^{\mu,\sigma^k} - \sigma^{k+1}\right\|^2\right] \leq \frac{2\theta}{\kappa T_\sigma + 2\theta}\left(D^2 + \frac{C^2}{\kappa\theta}\ln\left(\frac{\kappa T_\sigma}{2\theta} + 1\right)\right).$$

Moreover, by taking the expectation of (21), we have for any $t \geq 1$ such that $k(t) \geq 2$:

$$\mathbb{E}\left[\left\|\pi^{\mu,\sigma^{k(t)}} - \sigma^{k(t)}\right\|^2\right] \leq \frac{2D}{k(t)+1}\mathbb{E}\left[\left\|\pi^{\mu,\sigma^{k(t)}} - \sigma^{k(t)}\right\|\right] + \frac{12D^2}{(k(t)+1)^2}$$
$$+ 8\mathbb{E}\left[\left\|\sigma^{k(t)+1} - \pi^{\mu,\sigma^{k(t)}}\right\|^2\right] + 8D\sum_{l=1}^{k(t)}\mathbb{E}\left[\left\|\pi^{\mu,\sigma^l} - \sigma^{l+1}\right\|\right].$$

Combining these inequalities, we get for any $t \geq 1$ such that $k(t) \geq 2$:

$$\mathbb{E}\left[\left\|\pi^{\mu,\sigma^{k(t)}} - \sigma^{k(t)}\right\|^2\right] \leq \frac{2D}{k(t)+1}\mathbb{E}\left[\left\|\pi^{\mu,\sigma^{k(t)}} - \sigma^{k(t)}\right\|\right] + \frac{12D^2}{(k(t)+1)^2}$$
$$+ \frac{16\theta}{\kappa T_\sigma}\left(D^2 + \frac{C^2}{\kappa\theta}\ln\left(\frac{\kappa T_\sigma}{2\theta} + 1\right)\right) + 8Dk(t)\sqrt{\frac{2\theta}{\kappa T_\sigma}\left(D^2 + \frac{C^2}{\kappa\theta}\ln\left(\frac{\kappa T_\sigma}{2\theta} + 1\right)\right)}.$$

Since $T_\sigma \geq \max(1, T^{\frac{6}{7}}) \Rightarrow \frac{k(t)^3}{\sqrt{T_\sigma}} \leq 1$, we have:

$$\mathbb{E}\left[\left(\left\|\pi^{\mu,\sigma^{k(t)}} - \sigma^{k(t)}\right\| - \frac{D}{k(t)+1}\right)^2\right] \leq \frac{13D^2}{k(t)^2} + \frac{16\theta}{\kappa k(t)^2}\left(D^2 + \frac{C^2}{\kappa\theta}\ln\left(\frac{\kappa T}{2\theta}+1\right)\right) + \frac{8D}{k(t)^2}\sqrt{\frac{2\theta}{\kappa}\left(D^2 + \frac{C^2}{\kappa\theta}\ln\left(\frac{\kappa T}{2\theta}+1\right)\right)}.$$

Since $\mathbb{E}[X]^2 \leq \mathbb{E}[X^2]$ for any random variable X, we get:

$$\begin{aligned}
&\frac{13D^2}{k(t)^2} + \frac{16\theta}{\kappa k(t)^2}\left(D^2 + \frac{C^2}{\kappa\theta}\ln\left(\frac{\kappa T}{2\theta}+1\right)\right) + \frac{8D}{k(t)^2}\sqrt{\frac{2\theta}{\kappa}\left(D^2 + \frac{C^2}{\kappa\theta}\ln\left(\frac{\kappa T}{2\theta}+1\right)\right)} \\
&\geq \mathbb{E}\left[\left(\left\|\pi^{\mu,\sigma^{k(t)}} - \sigma^{k(t)}\right\| - \frac{D}{k(t)+1}\right)^2\right] \\
&\geq \mathbb{E}\left[\left\|\pi^{\mu,\sigma^{k(t)}} - \sigma^{k(t)}\right\| - \frac{D}{k(t)+1}\right]^2 \\
&= \left(\mathbb{E}\left[\left\|\pi^{\mu,\sigma^{k(t)}} - \sigma^{k(t)}\right\|\right] - \frac{D}{k(t)+1}\right)^2.
\end{aligned}$$

Then, we have:

$$\begin{aligned}
&\mathbb{E}\left[\left\|\pi^{\mu,\sigma^{k(t)}} - \sigma^{k(t)}\right\|\right] \\
&\leq \frac{D}{k(t)} + \frac{4D}{k(t)} + \frac{4\sqrt{\theta}}{\sqrt{\kappa}k(t)}\sqrt{D^2 + \frac{C^2}{\kappa\theta}\ln\left(\frac{\kappa T}{2\theta}+1\right)} + \frac{3\sqrt{D}}{k(t)}\left(\frac{2\theta}{\kappa}\left(D^2 + \frac{C^2}{\kappa\theta}\ln\left(\frac{\kappa T}{2\theta}+1\right)\right)\right)^{\frac{1}{4}} \\
&\leq \frac{5(\sqrt{\kappa}+\sqrt{\theta})}{k(t)\sqrt{\kappa}}\sqrt{D^2 + \frac{C^2}{\kappa\theta}\ln\left(\frac{\kappa T}{2\theta}+1\right)} + \frac{6\sqrt{D}(\sqrt{\theta}+1)}{k(t)}\left(\sqrt{\frac{1}{\kappa}\left(D^2 + \frac{C^2}{\kappa\theta}\ln\left(\frac{\kappa T}{2\theta}+1\right)\right)}+1\right) \\
&\leq \frac{6\left(\sqrt{\kappa}+\sqrt{\theta}+\sqrt{D\theta}+\sqrt{D})\right)}{k(t)}\left(\sqrt{\frac{1}{\kappa}\left(D^2 + \frac{C^2}{\kappa\theta}\ln\left(\frac{\kappa T}{2\theta}+1\right)\right)}+1\right).
\end{aligned}$$

Furthermore, for $k(t) = 1$, we have:

$$\mathbb{E}\left[\left\|\pi^{\mu,\sigma^1} - \sigma^1\right\|\right] \leq D \leq \frac{6\left(\sqrt{\kappa}+\sqrt{\theta}+\sqrt{D\theta}+\sqrt{D})\right)}{1}\left(\sqrt{\frac{1}{\kappa}\left(D^2 + \frac{C^2}{\kappa\theta}\ln\left(\frac{\kappa T}{2\theta}+1\right)\right)}+1\right).$$

Therefore, we have for any $t \geq 1$:

$$\mathbb{E}\left[\left\|\pi^{\mu,\sigma^{k(t)}} - \sigma^{k(t)}\right\|\right] \leq \frac{6\left(\sqrt{\kappa}+\sqrt{\theta}+\sqrt{D\theta}+\sqrt{D})\right)}{k(t)}\left(\sqrt{\frac{1}{\kappa}\left(D^2 + \frac{C^2}{\kappa\theta}\ln\left(\frac{\kappa T}{2\theta}+1\right)\right)}+1\right).$$

□

D Proof of Theorem 5.1

Proof of Theorem 5.1. By the definition of dynamic regret, we have:

$$\begin{aligned}
\mathrm{DynamicReg}_i(T) &= \sum_{t=1}^{T}\left(\max_{x\in\mathcal{X}_i} v_i(x, \pi_{-i}^t) - v_i(\pi^t)\right) \\
&\leq \mathcal{O}(1) + \sum_{t=3}^{T}\sum_{i=1}^{N}\left(\max_{x\in\mathcal{X}_i} v_i(x, \pi_{-i}^t) - v_i(\pi^t)\right).
\end{aligned}$$

Here, we introduce the following lemma:

Lemma D.1 (Lemma 2 of Cai et al. [2022a])**.** *For any $\pi \in \mathcal{X}$, we have:*

$$\sum_{i=1}^{N} \left(\max_{\tilde{\pi}_i \in \mathcal{X}_i} v_i(\tilde{\pi}_i, \pi_{-i}) - v_i(\pi) \right) \leq \mathrm{GAP}(\pi) \leq D \cdot \max_{\tilde{\pi} \in \mathcal{X}} \langle V(\pi), \tilde{\pi} - \pi \rangle.$$

Therefore, we have:

$$\mathrm{DynamicReg}_i(T) \leq \mathcal{O}(1) + \sum_{t=3}^{T} \mathrm{GAP}(\pi^t).$$

Thus, from Theorem 4.1:

$$\begin{aligned} \mathrm{DynamicReg}_i(T) &\leq \mathcal{O}(1) + \sum_{t=3}^{T} \mathcal{O}\left(\frac{\ln T}{t}\right) \\ &\leq \mathcal{O}\left((\ln T)^2\right). \end{aligned}$$

□

E Experimental details

E.1 Information on the computer resources

The experiments were conducted on macOS Sonoma 14.4.1 with Apple M2 Max and 32GB RAM.

E.2 Hard concave-convex game

Following the setup in Ouyang and Xu [2021], Cai and Zheng [2023], we choose

$$A = \frac{1}{4} \begin{bmatrix} & & & -1 & 1 \\ & & \cdots & \cdots & \\ & -1 & 1 & & \\ -1 & 1 & & & \\ 1 & & & & \end{bmatrix} \in \mathbb{R}^{n \times n}, \quad b = \frac{1}{4} \begin{bmatrix} 1 \\ 1 \\ \cdots \\ 1 \\ 1 \end{bmatrix} \in \mathbb{R}^n, \quad h = \frac{1}{4} \begin{bmatrix} 0 \\ 0 \\ \cdots \\ 0 \\ 1 \end{bmatrix} \in \mathbb{R}^n,$$

and $H = 2A^\top A$.

E.3 Hyperparameters

For each game, we carefully tuned the hyperparameters for each algorithm to ensure optimal performance. The specific parameters for each game and setting are summarized in Table 1.

Game	Algorithm	η	T_σ	μ
Random Payoff (Full Feedback)	OGA	0.05	-	-
	APGA	0.05	20	1.0
	GABP	0.05	10	1.0
Random Payoff (Noisy Feedback)	OGA	0.001	-	-
	APGA	0.001	2000	1.0
	GABP	0.001	1000	1.0
Hard Concave-Convex (Full Feedback)	OGA	1.0	-	-
	APGA	1.0	20	0.1
	GABP	1.0	20	0.1
Hard Concave-Convex (Noisy Feedback)	OGA	0.5	-	-
	APGA	0.5	50	0.1
	GABP	0.1	100	0.1

Table 1: Hyperparameters

F Relationship with accelerated optimistic gradient algorithm

Our GABP bears some relation to Accelerated Optimistic Gradient (AOG) [Cai and Zheng, 2023], which updates the strategy by:

$$\pi_i^{t+\frac{1}{2}} = \arg\max_{x \in \mathcal{X}_i} \left\{ \left\langle \eta \widehat{\nabla}_{\pi_i} v_i(\pi^{t-\frac{1}{2}}) + \frac{\pi_i^1 - \pi_i^t}{t+1}, x \right\rangle - \frac{1}{2} \left\| x - \pi_i^t \right\|^2 \right\},$$
$$\pi_i^{t+1} = \arg\max_{x \in \mathcal{X}_i} \left\{ \left\langle \eta \widehat{\nabla}_{\pi_i} v_i(\pi^{t+\frac{1}{2}}) + \frac{\pi_i^1 - \pi_i^t}{t+1}, x \right\rangle - \frac{1}{2} \left\| x - \pi_i^t \right\|^2 \right\}.$$

This can be equivalently written as:

$$\pi_i^{t+\frac{1}{2}} = \arg\max_{x \in \mathcal{X}_i} \left\{ \eta \left\langle \widehat{\nabla}_{\pi_i} v_i(\pi^{t-\frac{1}{2}}), x \right\rangle - \frac{1}{2} \left\| x - \frac{t\pi_i^t + \pi_i^1}{t+1} \right\|^2 \right\},$$
$$\pi_i^{t+1} = \arg\max_{x \in \mathcal{X}_i} \left\{ \eta \left\langle \widehat{\nabla}_{\pi_i} v_i(\pi^{t+\frac{1}{2}}), x \right\rangle - \frac{1}{2} \left\| x - \frac{t\pi_i^t + \pi_i^1}{t+1} \right\|^2 \right\}.$$

This means that AOG employs a convex combination $\frac{t\pi_i^t + \pi_i^1}{t+1}$ of the current strategy π_i^t and initial strategy π_i^1 as the proximal point in gradient ascent. However, our GABP diverges from AOG in that it uses a convex combination $\frac{k(t)\sigma_i^{k(t)} + \sigma_i^1}{k(t)+1}$ of $\sigma_i^{k(t)}$ and σ_i^1 as the reference strategy for the perturbation term.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We have clearly stated the contributions and scope of this study.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In "Introduction" and "Conclusion", we have reiterated the limitation of this study.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: Please see the theoretical results and their proofs in the Appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We have provided descriptions of experimental setups in the experiments section.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We have included the experimental code in the supplementary material.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We have provided descriptions of experimental setups in the experiments section.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Please see Figures.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We have shown the computer resources for this study in Appendix E.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have carefully reviewed and followed the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We have described the potential societal impacts of our study in Appendix A.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: There are no such risks associated with the paper.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: This study does not use existing assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: This paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.