Fast Two-Time-Scale Stochastic Gradient Method with Applications in Reinforcement Learning

Sihan Zeng SIHAN.ZENG@JPMCHASE.COM
JPMorgan AI Research, United States
Thinh T. Doan THINHDOAN@VT.EDU
Virginia Tech, United States

Editors: Shipra Agrawal and Aaron Roth

Abstract

Two-time-scale optimization is a framework introduced in Zeng et al. (2024) that abstracts a range of policy evaluation and policy optimization problems in reinforcement learning (RL). Akin to bi-level optimization under a particular type of stochastic oracle, the two-time-scale optimization framework has an upper level objective whose gradient evaluation depends on the solution of a lower level problem, which is to find the root of a strongly monotone operator. In this work, we propose a new method for solving two-time-scale optimization that achieves significantly faster convergence than the prior arts. The key idea of our approach is to leverage an averaging step to improve the estimates of the operators in both lower and upper levels before using them to update the decision variables. These additional averaging steps eliminate the direct coupling between the main variables, enabling the accelerated performance of our algorithm. We characterize the finite-time convergence rates of the proposed algorithm under various conditions of the underlying objective function, including strong convexity, convexity, Polyak-Lojasiewicz condition, and general non-convexity. These rates significantly improve over the best-known complexity of the standard two-time-scale stochastic approximation algorithm. When applied to RL, we show how the proposed algorithm specializes to novel online sample-based methods that surpass or match the performance of the existing state of the art. Finally, we support our theoretical results with numerical simulations in RL.

Keywords: Bi-level optimization, accelerated stochastic optimization, reinforcement learning

1. Introduction

Inspired by the two-time-scale optimization framework for reinforcement learning (RL) introduced in Zeng et al. (2024), we study the following optimization problem

$$\theta^\star = \operatorname{argmin}_{\theta\in\mathbb{R}^d} \; h(\theta), \tag{1}$$

where $h : \mathbb{R}^d \to \mathbb{R}$ is a continuously differentiable (possibly nonconvex) function. The gradient of h can only be accessed through a special stochastic oracle $F : \mathbb{R}^d \times \mathbb{R}^r \times \mathcal{X} \to \mathbb{R}$. The three arguments to F are the decision variable $\theta \in \mathbb{R}^d$, an auxiliary variable $\omega \in \mathbb{R}^r$, and a random variable X over the sample space $\mathcal{X}$. Given any θ, $F(\theta, \omega, X)$ returns an unbiased estimate of the gradient $\nabla h(\theta)$ only under a single "correct" setting of ω when the random variable X follows some distribution ξ

$$\nabla_\theta h(\theta) = \mathbb{E}_{X\sim\xi}[F(\theta, \omega^\star(\theta), X)], \tag{2}$$

where $\omega^\star(\theta) \in \mathbb{R}^r$ is defined through another stochastic sampling operator $G : \mathbb{R}^d \times \mathbb{R}^r \times \mathcal{X} \to \mathbb{R}$, as the solution to the equation

$$\mathbb{E}_{X\sim\xi}[G(\theta, \omega^\star(\theta), X] = 0. \tag{3}$$

In this work, the operator G is assumed to be strongly monotone with respect to ω, implying that $\omega^\star(\theta)$ exists and is unique for any θ. Given any $\omega \neq \omega^\star(\theta)$, the stochastic gradient $F(\theta, \omega, X)$ may be arbitrarily biased. By the first-order optimality condition, it is obvious that solving (1) is equivalent to finding a pair of solutions $(\theta^\star, \omega^\star)$ that solve the following two coupled nonlinear equations

$$\begin{cases} \mathbb{E}_{X\sim\xi}[F(\theta^\star, \omega^\star, X)] = 0, \\ \mathbb{E}_{X\sim\xi}[G(\theta^\star, \omega^\star, X)] = 0. \end{cases} \tag{4}$$

As detailed in Section 4, a wide range of RL methods, such as gradient-based temporal difference learning (Sutton et al., 2008, 2009) and actor-critic algorithms for policy optimization (Yang et al., 2019; Wu et al., 2020; Chen and Zhao, 2024), aim to solve problems that are special cases of (4). Thus, by developing and analyzing algorithms for the unified framework (4), we can apply them to various RL problems and immediately deduce performance guarantees, without having to tailor analysis for each specific problem. Hong et al. (2023); Zeng et al. (2024) pioneer the effort to construct such a unified framework and study solving (4) with the classic two-time-scale stochastic approximation (SA) algorithm, in which the iterates θ_k, ω_k are introduced to track $\theta^\star, \omega^\star$ and updated in the following simple manner

$$\theta_{k+1} = \theta_k - \alpha_k F(\theta_k, \omega_k, X_k), \quad \omega_{k+1} = \omega_k - \beta_k G(\theta_k, \omega_k, X_k), \tag{5}$$

where X_k are independently drawn from ξ in every iteration. The step sizes α_k and β_k need to be chosen carefully according to the structure of h, usually one much smaller than the other, hence leading to the name "two-time-scale". Despite the simplicity and popularity of this algorithm, its convergence rate is (at least as known so far) sub-optimal when compared to standard SGD under the most common oracle where we have access to $H(\theta, X) \in \mathbb{R}^d$ such that

$$\mathbb{E}_{X\in\xi}[H(\theta, X)] = \nabla h(\theta). \tag{6}$$

This is due to the complex coupling between θ_k, ω_k, and the noise X_k, Taking strongly convex function h as an example, standard SGD with k queries of the stochastic gradient H is able to reduce the optimality error to $\mathcal{O}(1/k)$, while the two-time-scale SA algorithm can only achieve $\mathcal{O}(1/k^{2/3})$.

Remark 1 *Under a Lipschitz/Holder continuous gradient assumption on $\omega^\star$, Shen and Chen (2022); Han et al. (2024) show that the standard two-time-scale algorithm achieves $\mathcal{O}(1/k)$ for strongly convex functions. We do not make this assumption in our work.*

1.1. Main Contributions

First, our work proposes a new variant of the two-time-scale SA algorithm, presented in Algorithm 1, which for the first time achieves the optimal convergence rates when solving (4) under function h satisfying various common structural properties, namely, strong convexity, convexity, Polyak-Lojasiewicz (PL) condition, and general non-convexity. We summarize the rates as well as the convergence metrics and choices of step sizes in Table 1. Our key idea behind the innovation is to leverage an averaging technique to estimate the unknown operators F, G through their samples before updating the decision variables. In other words, we "denoise" the estimate of the stochastic operators F, G, which leads to more stable updates of θ_k and ω_k. This seemingly simple modification effectively decouples the direct interaction between θ_k and ω_k and allows us to choose more aggressive step sizes, resulting in the accelerated convergence rates as well as simplified and elegant proofs.

Second, we show how the results derived for the general framework apply to policy evaluation and optimization problem in reinforcement learning. We discuss three specific applications, the first of which is temporal difference learning with gradient correction (TDC) under linear and non-linear function approximation. In the linear setting, the objective is strongly convex and Algorithm 1 specializes to a variant of the TDC algorithm guaranteed to converge with rate $\widetilde{\mathcal{O}}(1/k)$. In the non-linear setting, the objective is non-convex and our convergence rate is $\widetilde{\mathcal{O}}(1/k^{1/2})$. These results match the best-known rates of TDC in both cases. The second application is sample-based policy optimization in the linear-quadratic regulator, which has an objective function satisfying the PL condition. The state-of-the-art online actor-critic algorithm for this problem has a complexity of $\widetilde{\mathcal{O}}(1/k^{2/3})$ while our algorithm achieves the much faster rate $\widetilde{\mathcal{O}}(1/k)$. The final application is sample-based policy optimization in an entropy-regularized MDP, the objective of which also observes the PL condition. The results in Zeng et al. (2024) imply that the classic two-time-scale SA algorithm finds the globally optimal policy with rate $\widetilde{\mathcal{O}}(1/k^{2/3})$. Our algorithm in this context enjoys a finite-time rate of $\widetilde{\mathcal{O}}(1/k)$, again significantly elevating the state of the art.

Table 1: *Summary of finite-time complexity under various structural assumptions on h. $\bar{\theta}_k$ is a weighted average of history iterates defined in Theorem 2.*

Structural property	Metric	Complexity in this paper	Order of step sizes	Best known complexity of standard two-time-scale SA
Strong Convexity	$\mathbb{E}[\|\theta_k - \theta^\star\|^2]$	$\mathcal{O}(k^{-1})$	k^{-1}	$\mathcal{O}(k^{-\frac{2}{3}})$
Convexity	$\mathbb{E}[h(\bar{\theta}_k) - \theta^\star]$	$\mathcal{O}(k^{-\frac{1}{2}})$	$k^{-\frac{1}{2}}$	$\mathcal{O}(k^{-\frac{1}{4}})$
PL Condition	$\mathbb{E}[h(\theta_k) - h(\theta^\star)]$	$\mathcal{O}(k^{-1})$	k^{-1}	$\mathcal{O}(k^{-\frac{2}{3}})$
Non-Convexity	$\min_{t\leqslant k} \mathbb{E}[\|\nabla h(\theta_t)\|^2]$	$\mathcal{O}(k^{-\frac{1}{2}})$	$k^{-\frac{1}{2}}$	$\mathcal{O}(k^{-\frac{2}{5}})$

1.2. Related Work

This paper is closely related to the existing works on two-time-scale stochastic approximation, stochastic bi-level and composite optimization, as well as those that analyze abstracted or specific RL algorithms. We discuss the latest advances in these subjects to give context to our contributions.

Two-time-scale stochastic approximation. The aim of two-time-scale SA, pioneered by works including Borkar (1997); Konda and Borkar (1999); Mokkadem and Pelletier (2006b), is to solve a system of equations in the form of (4). Existing works in this domain primarily focus on the settings where F and G are linear operators (Konda and Tsitsiklis, 2004; Dalal et al., 2018; Doan and Romberg, 2019; Dalal et al., 2020; Gupta et al., 2019) or strongly monotone operators (Mokkadem and Pelletier, 2006a; Doan, 2022; Shen and Chen, 2022). Our work can be seen as a generalization of this line of work to the case where F is not a strongly monotone operator but the gradient map of a function h exhibiting various structure.

Stochastic bi-level and composite optimization. Bi-level optimization (Colson et al., 2007; Chen et al., 2022a, 2023) studies solve an optimization program of the form

$$\min_x f_1(x, y^\star(x)) \quad \text{subject to } y^\star(x) \in \operatorname{argmin}_y f_2(x, y), \tag{7}$$

or equivalently (by the first-order condition), finding a pair of stationary points $(x^\star, y^\star)$ that observes

$$\nabla_x f_1(x^\star, y^\star) = 0, \quad \nabla_y f_2(x^\star, y^\star) = 0.$$

This is a special case of our objective (4) with G being a gradient map. In RL problems G usually abstracts the Bellman backup operator associated with policy evaluation, which is well-known to not be the gradient of any function. In this sense, our framework is more general and suitable for modeling applications in RL.

As another related subject, stochastic composite optimization (Ghadimi and Lan, 2012; Wang et al., 2017; Chen et al., 2021) studies problems structured as

$$\min_x g_1(g_2(x)).$$

At a first glance, this optimization objective reduces to (7) by choosing $g_1 = f_1$, $g_2(x) = (x, y^\star(x))$, and $y^\star(x)$ to be the minimizer of $f_2(x, \cdot)$. Nevertheless, there is a key difference in the stochastic oracle. The assumption in stochastic composite optimization is that g_2 is continuous differentiable and that we can access an oracle that returns a direct stochastic estimate of ∇g_2. This assumption is restrictive in most RL applications, in which there is no guarantee on the differentiability of g_2 and only indirect information of g_2 is accessible.

Finite-Time Analysis of RL algorithms. Last but not least, we note the connection of our paper to the vast volume of recent literature that present finite-time and finite-sample analysis for various RL algorithms including temporal difference learning (Wang et al., 2021; Haque et al., 2023), actor-critic algorithms for standard MDP (Wu et al., 2020; Olshevsky and Gharesifard, 2023; Chen and Zhao, 2024), and actor-critic algorithms for the linear-quadratic regulator (Yang et al., 2019; Ju et al., 2022; Zhou and Lu, 2023). As samples in RL are sequentially drawn from a Markov chain, many of these works drop the i.i.d. noise assumption and realistically assume that the samples X_k are Markovian. Markovian samples can either be "time-invariant" in the sense that the stationary distribution of the Markov chain does not change over time (for example in policy evaluation or off-policy learning) or "controlled" if the samples are obtained from an Markov chain whose stationary distribution changes as the policy is updated (for example in actor-critic algorithms). To focus on the main contribution which is the acceleration due to stable estimates of operators F and G, our work assumes i.i.d. noise for simplicity. We note that the techniques to handle "time-invariant" and "controlled" Markovian noise have been well-known since Srikant and Ying (2019) and Zou et al. (2019). Extending the analysis of Algorithm 1 to the Markovian noise setting can be done in a straightforward way and only makes the bounds worse by a $\log(k)$ factor. We discuss this extension in more details in Remark 2.

Outline of the paper. The rest of the paper is organized as follows. In Section 2 we present the proposed fast two-time-scale optimization algorithm. In Section 3 we study the finite-time convergence rates of the algorithm under various functional structure. Section 4 discusses how the optimization framework (1)-(3) applies to RL problems and how the proposed algorithm solves these problems faster than (or as fast as) the prior arts. Finally, numerical simulations verifying the superiority of the proposed algorithm are presented in Section 5.

2. Faster Two-Time-Scale Stochastic Gradient Method

We formally present the fast two-time-scale stochastic gradient method in Algorithm 1, in which we maintain and iteratively update four variables: θ_k and ω_k are used to track $\theta^\star$ and $\omega^\star(\theta^\star)$, while f_k

and g_k estimate the stochastic operators F and G at the current iterates (θ_k, ω_k). An averaging step is carried out on the estimate of the stochastic operators. Note that Algorithm 1 reduces to the standard two-time-scale SA algorithm (5) if we set $\lambda_k = 1$ for all k. With a large λ_k the stochastic operator $F(\theta_k, \omega_k, X_k)$ estimates the true gradient $\nabla h(\theta_k)$ with high variance. By carefully choosing λ_k to decay sufficiently fast, the algorithm updates the main variables with much more stable gradient estimates, which is the key driver behind the accelerated convergence rates.

Algorithm 1 Fast Two-Time-Scale Stochastic Gradient Method

1: **Initialization:** decision variables $\theta_0 \in \mathbb{R}^d$ and ω_0, operator estimation variables $f_0 \in \mathbb{R}^d$ and $g_0 \in \mathbb{R}^r$, step size sequences $\{\alpha_k, \beta_k, \lambda_k\}$
2: **for** $k = 0, 1, \cdots, K-1$ **do**
3: Independent sample draw $X_k \sim \xi$
4: Decision variable update:

$$\theta_{k+1} = \theta_k - \alpha_k f_k, \quad \omega_{k+1} = \omega_k - \beta_k g_k. \tag{8}$$

5: Gradient estimation variable update:

$$f_{k+1} = (1-\lambda_k) f_k + \lambda_k F(\theta_k, \omega_k, X_k), \quad g_{k+1} = (1-\lambda_k)\, g_k + \lambda_k G(\theta_k, \omega_k, X_k). \tag{9}$$

6: **end for**

3. Finite-Time Complexity

In this section we study the finite-time complexity of Algorithm 1 under various structural properties of the objective function, namely, strong convexity, convexity, PL condition, and general non-convexity. Since the algorithm draws exactly one sample in each iteration, the time complexity translates to an equivalent sample complexity. We first state the main technical assumptions which are all fairly standard in the existing literature.

Assumption 1 (Lipschitz Continuity) *There exists a positive constant L such that for any $\theta_1, \theta_2 \in \mathbb{R}^d$, $\omega_1, \omega_2 \in \mathbb{R}^r$, and $X \in \mathcal{X}$*

$$\begin{aligned}
\|F(\theta_1, \omega_1, X) - F(\theta_2, \omega_2, X)\| &\leqslant L\left(\|\theta_1 - \theta_2\| + \|\omega_1 - \omega_2\|\right), \\
\|G(\theta_1, \omega_1, X) - G(\theta_2, \omega_2, X)\| &\leqslant L\left(\|\theta_1 - \theta_2\| + \|\omega_1 - \omega_2\|\right), \\
\|\nabla h(\theta_1) - \nabla h(\theta_2)\| &\leqslant L\|\theta_1 - \theta_2\|, \\
\|\omega^\star(\theta_1) - \omega^\star(\theta_2)\| &\leqslant L\|\theta_1 - \theta_2\|.
\end{aligned}$$

Assumption 1 states that the operators F, G, and $\omega^\star$ are Lipschitz continuous and the function h has Lipschitz gradients. This assumption is very common among works on two-time-scale SA (Zeng et al., 2024; Doan, 2022; Hong et al., 2023), and holds true in the RL applications to be discussed in Section 4. Without any loss of generality, we assume $L \geqslant 1$. We slightly abuse the notation to write

$$F(\theta, \omega) \triangleq \mathbb{E}_{X\sim\xi}[F(\theta, \omega, X)], \quad G(\theta, \omega) \triangleq \mathbb{E}_{X\sim\xi}[G(\theta, \omega, X)].$$

Assumption 1 obviously implies

$$\|F(\theta_1,\omega_1)-F(\theta_2,\omega_2)\| \leqslant L\left(\|\theta_1-\theta_2\|+\|\omega_1-\omega_2\|\right),$$
$$\|G(\theta_1,\omega_1)-G(\theta_2,\omega_2)\| \leqslant L\left(\|\theta_1-\theta_2\|+\|\omega_1-\omega_2\|\right).$$

Assumption 2 (Bounded Variance) *There exists a positive constant B such that for all $\theta \in \mathbb{R}^d$, $\omega \in \mathbb{R}^r$, and $X \in \mathcal{X}$*

$$\mathbb{E}_{X\sim\xi}[\|F(\theta,\omega,X)-F(\theta,\omega)\|^2] \leqslant B, \quad \mathbb{E}_{X\sim\xi}[\|G(\theta,\omega,X)-G(\theta,\omega)\|^2] \leqslant B.$$

The second assumption states that the operators F and G have bounded variance. This assumption is again standard in the literature on stochastic gradient algorithms under i.i.d. noise (Rakhlin et al., 2012; Shamir and Zhang, 2013; Hazan and Kale, 2014). We use $\mathcal{H}_k = \{X_0, X_1, \cdots, X_k\}$ to denote the randomness information observed until iteration k.

Assumption 3 (Strong Monotonicity) *There exists a constant $\mu_G > 0$ such that*

$$\langle \mathbb{E}_{X\sim\xi}[G(\theta,\omega,X)], \omega-\omega^\star(\theta)\rangle \geqslant \mu_G\|\omega-\omega^\star(\theta)\|^2, \quad \forall\theta\in\mathbb{R}^d, \omega\in\mathbb{R}^r. \tag{10}$$

Assumption 3 requires the operator G to be strongly monotone in expectation. If G is the gradient of a function ζ, i.e. there exists a function $\zeta : \mathbb{R}^d \times \mathbb{R}^r \to \mathbb{R}$ such that $\nabla_\omega \zeta(\theta,\omega) = G(\theta,\omega)$ for all θ,ω, then (10) amounts to the strong convexity of ζ. Again, this assumption can be verified to hold in all RL applications discussed in Section 4.

To simplify notations we introduce the residual variables, which will all be driven to 0

$$\Delta f_k = f_k - F\left(\theta_k,\omega_k\right), \ \Delta g_k = g_k - G\left(\theta_k,\omega_k\right), \ y_k = \|\omega_k-\omega^\star(\theta_k)\|^2. \tag{11}$$

3.1. Strong Convexity

We first analyze the convergence of Algorithm 1 when the upper level objective is strongly convex, under the following step sizes

$$\lambda_k = \frac{c_\lambda}{k+\tau+1}, \quad \alpha_k = \frac{c_\alpha}{k+\tau+1}, \quad \beta_k = \frac{c_\beta}{k+\tau+1}. \tag{12}$$

Here c_α, c_β, c_α, and $\tau \geqslant 1$ are properly selected constants depending only on the structure of h and G. Their exact choices are presented in Appendix A.

Assumption 4 (Strong Convexity) *The upper level objective h is strongly convex, i.e. there exists a constant $\mu_h > 0$ such that*

$$\langle \nabla h(\theta_1)-\nabla h(\theta_2), \theta_1-\theta_2\rangle \geqslant \mu_h\|\theta_1-\theta_2\|^2, \quad \forall\theta_1,\theta_2\in\mathbb{R}^d. \tag{13}$$

Theorem 1 *We recall the definition of residual variables in (11) and denote $z_k = \|\theta_k-\theta^\star\|^2$. Under Assumptions 1-4 and the step sizes in (12), the iterates of Algorithm 1 satisfy for all k*

$$\mathbb{E}[z_{k+1}] \leqslant \frac{B\tau^2}{k+\tau+1} + \frac{\tau^2}{(k+\tau+1)^2}\left(\|\Delta f_0\|^2 + \|\Delta g_0\|^2 + z_0 + y_0\right). \tag{14}$$

The theorem states that when h is strongly convex the iterate of Algorithm 1 converge to the globally optimal solution $\theta^\star$ in the sense of $\|\theta_k - \theta^\star\|^2$ with a rate of $\mathcal{O}(1/k)$, which matches the complexity of the standard SGD under the oracle H in (6) (Rakhlin et al., 2012). This rate is much better than the best-known complexity of the standard two-time-scale SA algorithm established in Hong et al. (2023); Zeng et al. (2024), which is $\mathcal{O}(1/k^{2/3})$. (1)-(3) under a strongly convex h abstracts the objective of gradient-based temporal difference learning in RL, which we present in Section 4.1.

Remark 2 *We note that Markovian/i.i.d. sampling results in a difference in the analysis of f_k, g_k in Lemma 2 and 3. In the proof of Lemma 2 in Appendix A.3 for example, it affects how we handle the cross term*

$$b_k = \langle \Delta f_k, F(\theta, \omega_k, X_k) - F(\theta, \omega_k) \rangle \Rightarrow \begin{cases} \mathbb{E}[b_k] = 0 & \textit{under i.i.d samples} \\ \mathbb{E}[b_k] \neq 0 & \textit{under Markovian samples} \end{cases}$$

When X_k is from a Markov chain with ξ as the stationary distribution, $\mathbb{E}[b_k]$ captures the gap between ξ and the distribution of X_k. This gap decays exponentially, under the assumption that the Markov chain is geometrically ergodic, which is a standard assumption in the literature (Xiong et al., 2021; Chen et al., 2022b; Zeng et al., 2022). Using techniques first developed in Srikant and Ying (2019), we can show that $\mathbb{E}[\|b_k\|] \leqslant \mathcal{O}(\alpha_k \log(1/\alpha_k))$. The rest of the proof of Theorem 1 can proceed without modification, and eventually we have $\mathbb{E}[z_{k+1}] \leqslant \mathcal{O}(\log(k+1)/(k+1))$, only differing from (14) by a logarithm factor. The same extension can be made similarly on the results to be presented later in this section for convex, PL, and non-convex functions.

3.2. Convexity

When h is convex, we show the convergence to the globally optimal value function. The step sizes are

$$\lambda_k = \frac{1}{4(K+1)^{1/2}}, \quad \alpha_k = \frac{c_\alpha}{(K+1)^{1/2}}, \quad \beta_k = \frac{c_\beta}{(K+1)^{1/2}}, \tag{15}$$

where the parameters c_α, c_β need to be small enough that $c_\alpha \leqslant c_\beta$, $c_\beta \leqslant \frac{1}{72(L+1)^6}$, $c_\beta \leqslant 1/4$, $c_\beta \leqslant 1/\mu_G$, and $(16 + \frac{23(L+1)^6}{\mu_G})\frac{c_\alpha^2}{c_\beta} + 12(L+1)^6 c_\alpha^2 \leqslant 1$. Note that we use horizon-aware constant step sizes in this case, selected a priori with respect to the total number of iterations K.

Theorem 2 *Suppose the function h is convex in θ. We recall the definition of residual variables in (11) and denote $z_k = \|\theta_k - \theta^\star\|^2$. We define $w_k = \frac{1}{(1+\frac{1}{K+1})^k}$ and $\bar{\theta}_K = (\sum_{k=0}^K w_k \theta_k)/(\sum_{k'=0}^K w_{k'})$. Under Assumptions 1-3 and the step sizes in (15), the iterates of Algorithm 1 satisfy*

$$\mathbb{E}[h(\bar{\theta}_K) - h(\theta^\star)] \leqslant \frac{1}{8c_\alpha^2\sqrt{K+1}}\left(\|\Delta f_0\|^2 + \|\Delta g_0\|^2 + z_0 + y_0\right) + \frac{B\lambda_0^2}{8c_\alpha\sqrt{K+1}}.$$

Under a convex objective, Algorithm 1 is guaranteed to find the globally optimal objective function with rate $\mathcal{O}(1/K^{1/2})$, again matching the complexity of standard SGD under the oracle (6) (Lan, 2020). Compared with the standard two-time-scale SA algorithm which has been shown to converge with rate $\mathcal{O}(1/K^{1/4})$ (Hong et al., 2023; Zeng et al., 2024), our proposed algorithm enjoys significantly improved time and sample complexity.

3.3. Non-Convexity with Polyak-Lojasiewicz Condition

We now shift to the non-convex regime. The optimal solution to (1) may not be unique without the strong convexity assumption. Let $\Theta^\star \subseteq \mathbb{R}^d$ denote the set of optimal solutions to (1) and $h^\star$ denote the optimal function value. We study the complexity of Algorithm 1 when the upper-level objective function of (1) satisfies the PL condition below.

Assumption 5 (Polyak-Lojasiewicz Condition) *There exists a constant $\mu_h > 0$ such that*

$$\frac{1}{2}\|\nabla h(\theta)\|^2 \geqslant \mu_h \left(h(\theta) - h^\star\right), \quad \forall \theta \in \mathbb{R}^d.$$

We study Algorithm 1 under the following choice of step sizes

$$\lambda_k = \frac{c_\lambda}{k+\tau+1}, \quad \alpha_k = \frac{c_\alpha}{k+\tau+1}, \quad \beta_k = \frac{c_\beta}{k+\tau+1},$$

where c_α, c_β, c_λ, and $\tau \geqslant 1$ again need to be properly selected according to the structure of h and G. Detailed choice of the parameters can be found in Appendix C.

Theorem 3 *We recall the residual variables defined in (11) and denote $x_k = h(\theta_k) - h^\star$. Under Assumptions 1-3, Assumption 5, and the step sizes above, the iterates of Algorithm 1 satisfy for all k*

$$\mathbb{E}[x_{k+1}] \leqslant \frac{B\tau^2}{k+\tau+1} + \frac{\tau^2}{(k+\tau+1)^2}\left(\|\Delta f_0\|^2 + \|\Delta g_0\|^2 + x_0 + y_0\right).$$

The theorem states that under the PL condition $h(\theta_k)$ converges with rate $\mathcal{O}(1/k)$ to the globally optimal function value $h^\star$. This again outperforms the standard two-time-scale SA algorithm which has a complexity of $\mathcal{O}(1/k^{2/3})$. Functions observing the PL condition include the policy optimization objectives in the linear-quadratic regulator and entropy-regularized MDP, which are the subjects of discussion in Section 4.2 and 4.3.

3.4. General Nonconvexity

For non-convex functions without additional structure, we consider step sizes that decay as $1/\sqrt{k}$

$$\lambda_k = \frac{1}{4(k+1)^{1/2}}, \quad \alpha_k = \frac{\alpha_0}{(k+1)^{1/2}}, \quad \beta_k = \frac{\beta_0}{(k+1)^{1/2}}, \tag{16}$$

with α_0, β_0 chosen such that $\alpha_0 \leqslant \beta_0$, $\beta_0 \leqslant \frac{1}{72L}$, $\beta_0 \leqslant \lambda_0$, $\alpha_0 \leqslant \frac{1}{72L^2}$.

Theorem 4 *Under Assumptions 1-3 and the step sizes in (16), we have for all k*

$$\min_{t\leqslant k} \mathbb{E}\left[\|\nabla h(\theta_t)\|^2\right] \leqslant \frac{32}{\alpha_0(k+1)^{1/2}}\left(\|\Delta f_0\|^2 + \|\Delta g_0\|^2 + (h(\theta_0) - h^\star) + y_0\right) + \frac{4\log(k+2)}{\alpha_0(k+1)^{1/2}}.$$

The convergence in the non-convex case is not global, but rather to a stationary point of h. The state-of-the-art convergence rate of the standard two-time-scale SA algorithm is $\mathcal{O}(1/k^{2/5})$ (Hong et al., 2023; Zeng et al., 2024), which we improve to $\mathcal{O}(1/k^{1/2})$. Applications of the optimization framework for non-convex functions include gradient-based policy evaluation under non-linear function approximation which we present in Section 4.1 and sample-based policy optimization under the standard MDP (Wu et al., 2020; Olshevsky and Gharesifard, 2023; Chen and Zhao, 2024).

4. Motivating Applications

4.1. Gradient-Based Policy Evaluation in Discounted-Reward MDP

Temporal difference learning with gradient correction (TDC) is a popular algorithm for policy evaluation in the setting where samples are collected by a behavior policy different from the one to be evaluated. Originally designed to work with linear function approximation (Sutton et al., 2008, 2009), this method has later been extended to the non-linear function approximation setting in Maei et al. (2009).

Consider a Markov decision process (MDP) characterized by $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma)$, where $\mathcal{S}$, $\mathcal{A}$ are the state and action space, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \Delta_{\mathcal{S}}$ is the transition kernel, $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the reward function, and $\gamma \in (0, 1)$ is the discount factor. In the linear function approximation case, each state $s \in \mathcal{S}$ is associated with a feature vector $\phi(s) \in \mathbb{R}^d$. We denote by Φ the stacked feature matrix

$$\Phi = \begin{bmatrix} - & \phi(s_1)^T & - \\ \cdots & \cdots & \cdots \\ - & \phi(s_{|\mathcal{S}|})^T & - \end{bmatrix} \in \mathbb{R}^{|\mathcal{S}| \times d}.$$

Policy evaluation studies finding the value function of a policy $\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}$. Given a parameter $\theta \in \mathbb{R}^d$, the approximated value function is $\Phi\theta \in \mathbb{R}^{|\mathcal{S}|}$. Let Π denote the orthogonal projection onto the subspace spanned by Φ. The objective of TDC is to find a parameter θ that minimizes the mean square projected Bellman error

$$\min_{\theta} J(\theta) \triangleq \|\Phi\theta - \Pi T^{\pi} \Phi\theta\|_{\mu_\pi}^2, \tag{17}$$

where μ_π denotes the stationary distribution of states under policy π and $\|v\|_{\mu_\pi}^2 = \sum_{s \in \mathcal{S}} \mu_\pi(s) v(s)^2$ for any $v \in \mathbb{R}^{|\mathcal{S}|}$.

As noted in Sutton et al. (2009), unbiased stochastic gradients of J cannot be directly obtained using samples. Instead, we need to introduce an auxiliary variable $\omega \in \mathbb{R}^d$ and solve

$$\nabla_\theta J(\theta) = -2\mathbb{E}_\pi[r(s,a) + \gamma\phi(s')^\top \theta\phi(s) - \phi(s)^\top \theta\phi(s)] + 2\gamma\mathbb{E}_\pi[\phi(s')\phi(s)^\top]\,\omega = 0, \tag{18}$$

$$\mathbb{E}_\pi[\phi(s)\phi(s)^\top]\,\omega = \mathbb{E}_\pi[r(s,a) + \gamma\phi(s')^\top \theta\phi(s) - \phi(s)^\top \theta\phi(s)]. \tag{19}$$

It is straightforward to see that the objective (18)-(19) and stochastic oracles of the TDC algorithm exactly matches (1)-(3), where the operator in the lower level problem (19) is strongly monotone and the upper level objective is strongly convex. As a result, Algorithm 1 specializes to a variant of the TDC algorithm guaranteed to converge with rate $\widetilde{\mathcal{O}}(1/k)$, which matches the best known convergence rate of TDC under linear function approximation (Haque et al., 2023). When non-linear function approximation is considered, TDC solves a system of equations that resemble (18)-(19), where the lower level operator is still strongly monotone but the upper level objective is non-convex (Maei et al., 2009). The best known convergence rate for TDC under non-linear function approximation is $\widetilde{\mathcal{O}}(1/k^{-1/2})$ to a stationary point of J (Wang et al., 2021), which we again match by our result in Theorem 4.

4.2. Policy Optimization for Linear-Quadratic Regulators

Linear-quadratic regulator (LQR) studies finding an optimal control sequence $\{u_k\}$ that stabilizes a linear system with the minimum cost

$$\begin{aligned} \{u_k^\star\} \;=\; \min_{\{u_k\}} \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\Big[\sum_{k=0}^{T}\left(x_k^\top Q x_k + u_k^\top R u_k\right) \mid x_0\Big] \\ \text{subject to} \quad x_{k+1} = Ax_k + Bu_k + w_k, \end{aligned} \tag{20}$$

where $x_k \in \mathbb{R}^{d_1}$, $u_k \in \mathbb{R}^{d_2}$ are the state and the control variables, $w_k \in \mathbb{R}^{d_1}$ is the system noise, $A \in \mathbb{R}^{d_1\times d_1}$ and $B \in \mathbb{R}^{d_1\times d_2}$ are the transition matrices, and $Q \in \mathbb{R}^{d_1\times d_1}$, $R \in \mathbb{R}^{d_2\times d_2}$ are positive-definite cost matrices. We assume that $\{w_k\}$ is independently and identically distributed according to $N(0,\Psi)$ for some covariance matrix $\Psi \in \mathbb{R}^{d_1\times d_1}$.

It is a classic result in the control literature that the optimal control sequence $\{u_k^\star\}$ is a time-invariant linear function of the state

$$u_k^\star = -K^\star x_k, \tag{21}$$

for a matrix $K^\star \in \mathbb{R}^{d_2\times d_1}$ as a function of A, B, Q, R (Bertsekas, 2012). Exploiting this fact allows us to re-formulate the LQR objective as finding the optimal control matrix K, which results in an optimization problem of the following form

$$\begin{aligned} \min_{K} \quad & \operatorname{trace}(P_K\Psi)) \\ \text{subject to} \quad & P_K = Q + K^\top R K + (A-BK)^\top P_K (A-BK). \end{aligned} \tag{22}$$

To encourage exploration, we consider the regularized objective with $\sigma \geqslant 0$ and $\Psi_\sigma = \Psi + \sigma^2 BB^\top$

$$\min_K J(K) \triangleq \operatorname{trace}(P_K\Psi_\sigma) + \sigma^2 \operatorname{trace}(R) \quad \text{subject to (22)}, \tag{23}$$

which is known to have the same optimal solution as the unregularized problem (Gravell et al., 2020).

We are interested in finding the solution to (23) under unknown transition matrices A and B, by taking a gradient-based approach. The natural gradient of J with respect to K, which we denote by $\widetilde{\nabla}_K J$ and which can be regarded as $\nabla_K J$ on a slightly transformed system of coordinates, has the following closed-form expression

$$\widetilde{\nabla}_K J = 2\left(R + B^\top P_K B\right)K - 2B^\top P_K A.$$

To obtain a reliable stochastic estimate of $\widetilde{\nabla}_K J$ the key is to estimate $R + B^\top P_K B$ and $B^\top P_K A$. For simplicity of notation, we define

$$\Omega_K = \begin{pmatrix} \Omega_K^{11} & \Omega_K^{12} \\ \Omega_K^{21} & \Omega_K^{22} \end{pmatrix} = \begin{pmatrix} Q + A^\top P_K A & A^\top P_K B \\ B^\top P_K A & R + B^\top P_K B \end{pmatrix}, \tag{24}$$

of which $R + B^\top P_K B$ and $B^\top P_K A$ are sub-matrices. We also define $\operatorname{svec}(\cdot)$ as the operation to vectorize the upper triangular sub-matrix of a given symmetric matrix with the off-diagonal entries weighted by $\sqrt{2}$. We further define $\phi(x,u) = \operatorname{svec}(\left[x^\top, u^\top\right]^\top \left[x^\top, u^\top\right])$ for any $x \in \mathbb{R}^{d_1}, u \in \mathbb{R}^{d_2}$,

which can be seen as the feature vector for the system in state (x, u). It is shown in Yang et al. (2019) that Ω_K and $J(K)$ jointly satisfy

$$\mathbb{E}_{x\sim\mu_K, u\sim N(-Kx,\sigma^2 I)}\left[M_{x,u,x',u'}\right]\begin{bmatrix} J(K) \\ \operatorname{svec}(\Omega_K)\end{bmatrix} = \mathbb{E}_{x\sim\mu_K, u\sim N(-Kx,\sigma^2 I)}\left[c_{x,u}\right], \tag{25}$$

where the matrix $M_{x,u,x',u'}$ and vector $c_{x,u}$ are

$$M_{x,u,x',u'} = \begin{bmatrix} 1 & 0 \\ \phi(x,u) & \phi(x,u)\left[\phi(x,u) - \phi\left(x',u'\right)\right]^\top\end{bmatrix},\ c_{x,u} = \begin{bmatrix} x^\top Q x + u^\top R u \\ \left(x^\top Q x + u^\top R u\right)\phi(x,u)\end{bmatrix}.$$

Equation (25) maps to the lower level problem (3). Under the assumption that K is a stable controller with respect to A and B (Yang et al., 2019), the linear operator $\mathbb{E}[M_{x,u,x',u'}]$ is strongly monotone (Yang et al., 2019). The upper level problem that corresponds to (2) is to solve $\widetilde{\nabla}_K J = 0$, which is known to observe the PL condition (Fazel et al., 2018). This means that we can apply Algorithm 1 to this problem, which specializes to an single-looped "actor-critic" algorithm with a finite-time convergence rate of $\widetilde{\mathcal{O}}(1/k)$ to the globally optimal solution measured by $J(K_k) - J(\theta^\star)$. The complexity of the current state-of-the-art single-looped algorithm for LQR is $\widetilde{\mathcal{O}}(1/k^{2/3})$, which we improve over. We also note the recent few works on nested-loop actor-critic algorithms that rely on batched multi-trajectory samples (Ju et al., 2022; Zhou and Lu, 2023). These algorithms achieve the complexity $\widetilde{\mathcal{O}}(1/k)$ as well, but in practice are much less convenient to implement than our algorithm which is single-looped and only requires a single online trajectory of samples.

4.3. Policy Optimization for Entropy-Regularized MDPs

The optimization framework (1)-(3) also applies to the policy optimization problem in the entropy-regularized MDP. Consider the MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma)$ introduced in Section 4.1. The regularized value function of a policy $\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}$ is

$$V_\tau^\pi(s) = \mathbb{E}_{a_k\sim\pi(\cdot|s_k), s_{k+1}\sim\mathcal{P}(\cdot|s_k,a_k)}\Big[\sum_{k=0}^{\infty}\gamma^k\big(r(s_k,a_k) - \tau\log\pi(a_k\mid s_k)\big)\mid s_0 = s\Big],$$

where τ is a positive regularization weight. Under initial state distribution $\rho \in \Delta_{\mathcal{S}}$, the expected cumulative reward under policy π is $J_\tau(\pi) \triangleq \mathbb{E}_{s\sim\rho}[V_\tau^\pi(s)]$. We consider the tabular setting where the policy is represented through the softmax function by a parameter $\theta \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$

$$\pi_\theta(a\mid s) = \frac{\exp\left(\theta_{s,a}\right)}{\sum_{a'\in\mathcal{A}}\exp\left(\theta_{s,a'}\right)}.$$

Policy optimization studies finding the parameter $\theta^\star$ that maximizes the cumulative reward $J_\tau(\pi_\theta)$. By the first-order condition, this is equivalent to solving

$$\nabla_\theta J_\tau(\pi_\theta) = \frac{1}{1-\gamma}\mathbb{E}_{\pi_\theta}\big[\left(r(s,a) - \tau\log\pi_\theta(a\mid s) + \gamma V_\tau^{\pi_\theta}(s') - V_\tau^{\pi_\theta}(s)\right)\nabla_\theta\log\pi_\theta(a\mid s)\big] = 0.$$

Evaluating the gradient requires computing the regularized value function $V_\tau^{\pi_\theta} \in \mathbb{R}^{|\mathcal{S}|}$, which satisfies the regularized Bellman equation

$$\mathbb{E}_{\pi_\theta}\Big[r(s,a) - \tau\log\pi_\theta(a\mid s) + \gamma V_\tau^{\pi_\theta}(s') - V_\tau^{\pi_\theta}(s)\Big] = 0. \tag{26}$$

Algorithm 2 Fast Single-Loop Actor-Critic Algorithm for LQR

1: **Initialization:** Control matrix K_0, auxiliary variable $\hat{\Omega}_0, \hat{J}_0$, gradient estimation variable f_0, g_0^J, g_0^Ω, step size sequences $\{\alpha_k, \beta_k, \lambda_k\}$

2: Sample an initial state $x_0 \sim \mathcal{D}$, takes an initial control $u_0 \sim N(-K_0 x_0, \sigma^2 I)$, and observe cost $c_0 = x_0^\top Q x_0 + u_0^\top R u_0$ and next state $x_1 \sim N(Ax_0 + Bu_0, \Psi)$

3: **for** $k = 0, 1, 2, ...$ **do**

4: Take the control $u_{k+1} \sim N(-K_k x_{k+1}, \sigma^2 I)$. Observe cost $c_{k+1} = x_{k+1}^\top Q x_{k+1} + u_{k+1}^\top R u_{k+1}$ and next state $x_{k+2} \sim N(Ax_{k+1} + Bu_{k+1}, \Psi)$

5: Control matrix update:

$$K_{k+1} = K_k - \alpha_k f_k$$

6: Auxiliary variable update:

$$\hat{J}_{k+1} = \hat{J}_k - \beta_k g_k^J, \quad \hat{\Omega}_{k+1} = \hat{\Omega}_k - \beta_k g_k^\Omega$$

7: Gradient estimation variable update:

$$f_{k+1} = (1-\lambda_k) f_k + \lambda_k (\hat{\Omega}_k^{22} K_k - \hat{\Omega}_k^{21})$$
$$g_{k+1}^J = (1-\lambda_k) g_k^J + \lambda_k (\hat{J}_k - x_k^\top Q x_k - u_k^\top R u_k)$$
$$g_{k+1}^\Omega = (1-\lambda_k) g_k^\Omega - \lambda_k \begin{bmatrix} x_k \\ u_k \end{bmatrix} \begin{bmatrix} x_k \\ u_k \end{bmatrix}^\top \left(\begin{bmatrix} x_k^\top & u_k^\top \end{bmatrix} \hat{\Omega}_k \begin{bmatrix} x_k \\ u_k \end{bmatrix} + \hat{J}_k - x_k^\top Q x_k - u_k^\top R u_k \right)$$

8: **end for**

This objective again falls under the general two-time-scale optimization framework, with the operator in lower level problem (26) being strongly monotone. The upper level objective function J_τ satisfies the PL condition (Mei et al., 2020). Applying Algorithm 1 to this problem results in a fast actor-critic algorithm (Algorithm 2) that is able to reduce the optimality gap $J_\tau(\theta_k) - J_\tau(\theta^\star)$ with rate $\widetilde{\mathcal{O}}(1/k)$. The analysis in Zeng et al. (2024) implies that the standard two-time-scale algorithm (5) solves this problem with a complexity of $\widetilde{\mathcal{O}}(1/k^{2/3})$, which we vastly improve over.

5. Numerical Simulations

We experimentally verify that Algorithm 1 converges faster than (5) in solving 1) a policy evaluation problem under linear function approximation, and 2) policy optimization for the LQR.

Policy evaluation problem under linear function approximation. As we have discussed in Section 4.1, TDC is the state-of-the-art policy evaluation algorithm under off-policy samples which enjoys a convergence rate of $\widetilde{\mathcal{O}}(1/k)$ under linear function approximation. Numerically we evaluate Algorithm 1 against TDC on the task of evaluating the value function of a randomly generated policy in a randomly generated environment with $|\mathcal{S}| = |\mathcal{A}| = 50$ and the feature dimension $d = 10$. Figure 1 shows the empirical performance gap between the standard TDC and Algorithm 1, though

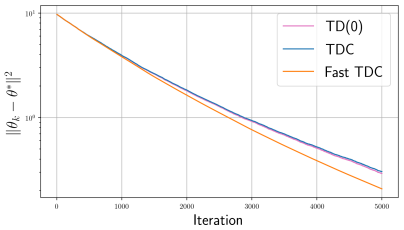

Figure 1: Fast TDC Algorithm for Random Policy Evaluation

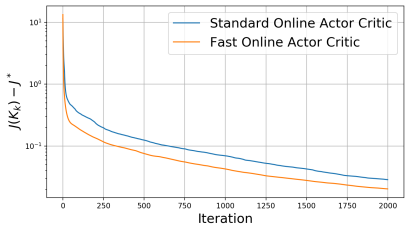

Figure 2: Performance of Fast Actor-Critic Algorithm for LQR

both algorithm have the same complexity in theory. For comparison we also include TD(0) learning under linear function approximation. The simulation results in Figure 1 numerically confirms the superiority of the proposed algorithm. Details of the experimental setup can be found in Appendix E.

Policy optimization for LQR. We experiment with the small-dimensional environment considered in Zeng et al. (2024) where the transition and costs matrices are chosen to be

$$A = \begin{bmatrix} 0.5 & 0.01 & 0 \\ 0.01 & 0.5 & 0.01 \\ 0 & 0.01 & 0.5 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 0.1 \\ 0 & 0.1 \\ 0 & 0.1 \end{bmatrix}, \quad Q = I_3, \quad R = I_2. \tag{27}$$

We apply Algorithm 2 to this problem along with online actor-critic algorithm studied in Zeng et al. (2024), which is equivalent to (5) applied to this context. In Figure 2, we plot the error decay under the two algorithms, which again shows that the proposed algorithm enjoys faster convergence.

Disclaimer

This paper was prepared for informational purposes in part by the Artificial Intelligence Research group of JP Morgan Chase & Co and its affiliates ("JP Morgan"), and is not a product of the Research Department of JP Morgan. JP Morgan makes no representation and warranty whatsoever and disclaims all liability, for the completeness, accuracy or reliability of the information contained herein. This document is not intended as investment research or investment advice, or a recommendation, offer or solicitation for the purchase or sale of any security, financial instrument, financial product or service, or to be used in any way for evaluating the merits of participating in any transaction, and shall not constitute a solicitation under any jurisdiction or to any person, if such solicitation under such jurisdiction or to such person would be unlawful.

Acknowledgement

This work was supported by the National Science Foundation under awards ECCS-CAREER-2339509 and CCF-2343599.

References

Dimitri Bertsekas. *Dynamic programming and optimal control: Volume I*, volume 1. Athena scientific, 2012.

Vivek S Borkar. Stochastic approximation with two time scales. *Systems & Control Letters*, 29(5): 291–294, 1997.

Tianyi Chen, Yuejiao Sun, and Wotao Yin. Solving stochastic compositional optimization is nearly as easy as solving stochastic optimization. *IEEE Transactions on Signal Processing*, 69:4937–4948, 2021.

Tianyi Chen, Yuejiao Sun, Quan Xiao, and Wotao Yin. A single-timescale method for stochastic bilevel optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 2466–2488. PMLR, 2022a.

Xuxing Chen, Tesi Xiao, and Krishnakumar Balasubramanian. Optimal algorithms for stochastic bilevel optimization under relaxed smoothness conditions. *arXiv preprint arXiv:2306.12067*, 2023.

Xuyang Chen and Lin Zhao. Finite-time analysis of single-timescale actor-critic. *Advances in Neural Information Processing Systems*, 36, 2024.

Zaiwei Chen, Sheng Zhang, Thinh T Doan, John-Paul Clarke, and Siva Theja Maguluri. Finite-sample analysis of nonlinear stochastic approximation with applications in reinforcement learning. *Automatica*, 146:110623, 2022b.

Benoît Colson, Patrice Marcotte, and Gilles Savard. An overview of bilevel optimization. *Annals of operations research*, 153(1):235–256, 2007.

G. Dalal, G. Thoppe, B. Szörényi, and S. Mannor. Finite sample analysis of two-timescale stochastic approximation with applications to reinforcement learning. In *COLT*, 2018.

Gal Dalal, Balazs Szorenyi, and Gugan Thoppe. A tale of two-timescale reinforcement learning with the tightest finite-time bound. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34 (04):3701–3708, Apr. 2020.

Thinh T Doan. Nonlinear two-time-scale stochastic approximation convergence and finite-time performance. *IEEE Transactions on Automatic Control*, 2022.

Thinh T Doan and Justin Romberg. Linear two-time-scale stochastic approximation a finite-time analysis. In *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 399–406. IEEE, 2019.

Maryam Fazel, Rong Ge, Sham Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *International conference on machine learning*, pages 1467–1476. PMLR, 2018.

Saeed Ghadimi and Guanghui Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization i: A generic algorithmic framework. *SIAM Journal on Optimization*, 22(4):1469–1492, 2012.

Benjamin Gravell, Peyman Mohajerin Esfahani, and Tyler Summers. Learning optimal controllers for linear systems with multiplicative noise via policy gradient. *IEEE Transactions on Automatic Control*, 66(11):5283–5298, 2020.

Harsh Gupta, Rayadurgam Srikant, and Lei Ying. Finite-time performance bounds and adaptive learning rate selection for two time-scale reinforcement learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Yuze Han, Xiang Li, and Zhihua Zhang. Finite-time decoupled convergence in nonlinear two-time-scale stochastic approximation. *arXiv preprint arXiv:2401.03893*, 2024.

Shaan Ul Haque, Sajad Khodadadian, and Siva Theja Maguluri. Tight finite time bounds of two-time-scale linear stochastic approximation with markovian noise. *arXiv preprint arXiv:2401.00364*, 2023.

Elad Hazan and Satyen Kale. Beyond the regret minimization barrier: optimal algorithms for stochastic strongly-convex optimization. *The Journal of Machine Learning Research*, 15(1): 2489–2512, 2014.

Mingyi Hong, Hoi-To Wai, Zhaoran Wang, and Zhuoran Yang. A two-timescale stochastic algorithm framework for bilevel optimization: Complexity analysis and application to actor-critic. *SIAM Journal on Optimization*, 33(1):147–180, 2023.

Caleb Ju, Georgios Kotsalis, and Guanghui Lan. A model-free first-order method for linear quadratic regulator with o $(1/\varepsilon)$ sampling complexity. *arXiv preprint arXiv:2212.00084*, 2022.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part I 16*, pages 795–811. Springer, 2016.

V. R. Konda and J. N. Tsitsiklis. Convergence rate of linear two-time-scale stochastic approximation. *The Annals of Applied Probability*, 14(2):796–819, 2004.

Vijaymohan R Konda and Vivek S Borkar. Actor-critic–type learning algorithms for markov decision processes. *SIAM Journal on control and Optimization*, 38(1):94–123, 1999.

Guanghui Lan. *First-order and stochastic optimization methods for machine learning*, volume 1. Springer, 2020.

Hamid Maei, Csaba Szepesvari, Shalabh Bhatnagar, Doina Precup, David Silver, and Richard S Sutton. Convergent temporal-difference learning with arbitrary smooth function approximation. *Advances in neural information processing systems*, 22, 2009.

Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods. In *International Conference on Machine Learning*, pages 6820–6829. PMLR, 2020.

A. Mokkadem and M. Pelletier. Convergence rate and averaging of nonlinear two-time-scale stochastic approximation algorithms. *The Annals of Applied Probability*, 16(3):1671–1702, 2006a.

Abdelkader Mokkadem and Mariane Pelletier. Convergence rate and averaging of nonlinear two-time-scale stochastic approximation algorithms. *Ann. Appl. Probab.*, 16(1):1671–1702, 2006b.

Alex Olshevsky and Bahman Gharesifard. A small gain analysis of single timescale actor critic. *SIAM Journal on Control and Optimization*, 61(2):980–1007, 2023.

Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 1571–1578, 2012.

Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *International conference on machine learning*, pages 71–79. PMLR, 2013.

Han Shen and Tianyi Chen. A single-timescale analysis for stochastic approximation with multiple coupled sequences. *Advances in Neural Information Processing Systems*, 35:17415–17429, 2022.

Rayadurgam Srikant and Lei Ying. Finite-time error bounds for linear stochastic approximation and TD learning. In *Conference on Learning Theory*, pages 2803–2830. PMLR, 2019.

Richard S Sutton, Csaba Szepesvári, and Hamid Reza Maei. A convergent o (n) algorithm for off-policy temporal-difference learning with linear function approximation. *Advances in neural information processing systems*, 21(21):1609–1616, 2008.

Richard S Sutton, Hamid Reza Maei, Doina Precup, Shalabh Bhatnagar, David Silver, Csaba Szepesvári, and Eric Wiewiora. Fast gradient-descent methods for temporal-difference learning with linear function approximation. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 993–1000, 2009.

Mengdi Wang, Ethan X Fang, and Han Liu. Stochastic compositional gradient descent: algorithms for minimizing compositions of expected-value functions. *Mathematical Programming*, 161(1): 419–449, 2017.

Yue Wang, Shaofeng Zou, and Yi Zhou. Non-asymptotic analysis for two time-scale tdc with general smooth function approximation. *Advances in Neural Information Processing Systems*, 34: 9747–9758, 2021.

Yue Frank Wu, Weitong Zhang, Pan Xu, and Quanquan Gu. A finite-time analysis of two time-scale actor-critic methods. *Advances in Neural Information Processing Systems*, 33:17617–17628, 2020.

Huaqing Xiong, Tengyu Xu, Yingbin Liang, and Wei Zhang. Non-asymptotic convergence of adam-type reinforcement learning algorithms under markovian sampling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10460–10468, 2021.

Zhuoran Yang, Yongxin Chen, Mingyi Hong, and Zhaoran Wang. Provably global convergence of actor-critic: A case for linear quadratic regulator with ergodic cost. *Advances in neural information processing systems*, 32, 2019.

Sihan Zeng, Thinh T Doan, and Justin Romberg. Finite-time convergence rates of decentralized stochastic approximation with applications in multi-agent and multi-task learning. *IEEE Transactions on Automatic Control*, 2022.

Sihan Zeng, Thinh T Doan, and Justin Romberg. A two-time-scale stochastic optimization framework with applications in control and reinforcement learning. *SIAM Journal on Optimization*, 34(1): 946–976, 2024.

Mo Zhou and Jianfeng Lu. Single timescale actor-critic method to solve the linear quadratic regulator with convergence guarantees. *Journal of Machine Learning Research*, 24(222):1–34, 2023.

Shaofeng Zou, Tengyu Xu, and Yingbin Liang. Finite-sample analysis for sarsa with linear function approximation. *Advances in neural information processing systems*, 32, 2019.

Appendix A. Analysis for Strongly Convex Functions

The exact choice of the step size parameters is made such that $\alpha_k \leqslant \beta_k \leqslant \lambda_k \leqslant 1/4$, $c_\alpha \geqslant \frac{8}{\mu_h}$, and

$$\alpha_k \leqslant \min\{\frac{\lambda_k}{\mu_h}, \frac{\mu_h}{6(L+1)^4}, \frac{8\mu_G}{\mu_h}\beta_k, \frac{\mu_h\mu_G}{152(L+1)^6}\beta_k, \frac{\mu_G\beta_k}{8(8L^4+\frac{9L^2}{\mu_G})}, \frac{\lambda_k}{4}(48L^2+\frac{10L}{\mu_G})^{-1}\},$$
$$\beta_k \leqslant \min\{\frac{1}{240(L+1)^6}, \frac{1}{\mu_G}, \frac{\lambda_k}{4}(56L^2+\frac{9}{2\mu_h})^{-1}, \frac{\lambda_k}{4}(48L^2+\frac{10L}{\mu_G})^{-1}, \frac{\mu_G}{168L^4}\},$$

which requires β_0 to be selected sufficiently small and τ sufficiently large. Note that parameters c_α, β_0, and τ always exist and only depend on the structural constants of h and G.

The proof of Theorem 1 is built on the lemmas below. Their proofs can be found at the end of the section.

Lemma 1 *Under Assumption 1-3, the sequence of variables $\{\theta_k, \omega_k, f_k, g_k\}$ satisfy for all k*

$$\begin{aligned}
\|f_k\| &\leqslant \|\Delta f_k\| + L\sqrt{y_k} + L(L+1)\sqrt{z_k},\\
\|g_k\| &\leqslant \|\Delta g_k\| + L\sqrt{y_k},\\
\|\theta_{k+1}-\theta_k\| &\leqslant \alpha_k\left(\|\Delta f_k\| + L\sqrt{y_k} + L(L+1)\sqrt{z_k}\right),\\
\|\omega_{k+1}-\omega_k\| &\leqslant \beta_k\left(\|\Delta g_k\| + L\sqrt{y_k}\right).
\end{aligned}$$

Lemma 2 *Suppose that the step size λ_k satisfy $\lambda_k \leqslant 1/4$ for all k. Then, under Assumption 1-3, we have the following bound on the iteration-wise convergence of $\mathbb{E}[\|\Delta f_k\|^2]$.*

$$\begin{aligned}
\mathbb{E}[\|\Delta f_{k+1}\|^2] \leqslant{}& (1-\lambda_k)\mathbb{E}[\|\Delta f_k\|^2] - (\frac{\lambda_k}{4} - \frac{24L^2\beta_k^2}{\lambda_k})\mathbb{E}[\|\Delta f_k\|^2]\\
&+ \frac{4}{\lambda_k}\mathbb{E}\left[6L^2\beta_k^2\|\Delta g_k\|^2 + 10L^4\beta_k^2 y_k + 6(L+1)^6\alpha_k^2 z_k\right] + 2B\lambda_k^2.
\end{aligned}$$

Lemma 3 *Suppose that the step size λ_k satisfy $\lambda_k \leqslant 1/4$ for all k. Then, under Assumption 1-3, we have the following bound on the iteration-wise convergence of $\mathbb{E}[\|\Delta g_k\|^2]$*

$$\begin{aligned}
\mathbb{E}[\|\Delta g_{k+1}\|^2] \leqslant{}& (1-\lambda_k)\mathbb{E}[\|\Delta g_k\|^2] - (\frac{\lambda_k}{4} - \frac{24L^2\beta_k^2}{\lambda_k})\mathbb{E}[\|\Delta g_k\|^2]\\
&+ \frac{4}{\lambda_k}\mathbb{E}\left[6L^2\beta_k^2\|\Delta f_k\|^2 + 10L^4\beta_k^2 y_k + 6(L+1)^6\alpha_k^2 z_k\right] + 2B\lambda_k^2.
\end{aligned}$$

Lemma 4 *Suppose the step size α_k is chosen such that $\alpha_k \leqslant \frac{\mu_h}{6(L+1)^4}$ and $\alpha_k \leqslant \frac{1}{2\mu_h}$ for all k. Under Assumption 1-4, we have*

$$z_{k+1} \leqslant (1-\frac{\mu_h\alpha_k}{4})z_k + \frac{9L^2\alpha_k}{\mu_h}y_k + \frac{9\alpha_k}{2\mu_h}\|\Delta f_k\|^2.$$

Lemma 5 *Suppose that the step sizes satisfy $\alpha_k \leqslant \beta_k \leqslant \frac{1}{72(L+1)^6}$ and $\beta_k \leqslant 1/\mu_G$ for all k. Then, under Assumption 1-3, we have*

$$\begin{aligned}
y_{k+1} \leqslant{}& (1-\mu_G\beta_k)y_k - \left(\frac{\mu_G}{4}\beta_k - 8L^4\alpha_k - L^2\beta_k^2\right)y_k\\
&+ \left(\frac{19(L+1)^6\alpha_k^2}{\mu_G\beta_k} + 3(L+1)^6\alpha_k\beta_k\right)z_k + \left(\frac{8\beta_k}{\mu_G} + 2L\beta_k^2\right)\|\Delta g_k\|^2 + 8L\alpha_k\|\Delta f_k\|^2.
\end{aligned}$$

A.1. Proof of Theorem 1

We consider the Lyapunov function

$$V_k \triangleq \mathbb{E}[\|\Delta f_k\|^2 + \|\Delta g_k\|^2 + z_k + y_k].$$

Combining Lemma 2-5, we get

$$\begin{aligned}
V_{k+1} &= \mathbb{E}[\|\Delta f_{k+1}\|^2 + \|\Delta g_{k+1}\|^2 + z_{k+1} + y_{k+1}] \\
&\leqslant (1-\lambda_k)\mathbb{E}[\|\Delta f_k\|^2] - (\frac{\lambda_k}{4} - \frac{24L^2\beta_k^2}{\lambda_k})\mathbb{E}[\|\Delta f_k\|^2] \\
&\quad + \frac{4}{\lambda_k}\mathbb{E}\left[6L^2\beta_k^2\|\Delta g_k\|^2 + 10L^4\beta_k^2 y_k + 6(L+1)^6\alpha_k^2 z_k\right] + 2B\lambda_k^2 \\
&\quad + (1-\lambda_k)\mathbb{E}[\|\Delta g_k\|^2] - (\frac{\lambda_k}{4} - \frac{24L^2\beta_k^2}{\lambda_k})\mathbb{E}[\|\Delta g_k\|^2] \\
&\quad + \frac{4}{\lambda_k}\mathbb{E}\left[6L^2\beta_k^2\|\Delta f_k\|^2 + 10L^4\beta_k^2 y_k + 6(L+1)^6\alpha_k^2 z_k\right] + 2B\lambda_k^2 \\
&\quad + \mathbb{E}[(1-\frac{\mu_h\alpha_k}{4})z_k + \frac{9L^2\alpha_k}{\mu_h}y_k + \frac{9\alpha_k}{2\mu_h}\|\Delta f_k\|^2] \\
&\quad + \mathbb{E}[(1-\mu_G\beta_k)y_k - \left(\frac{\mu_G}{4}\beta_k - 8L^4\alpha_k - L^2\beta_k^2\right)y_k \\
&\quad + \left(\frac{19(L+1)^6\alpha_k^2}{\mu_G\beta_k} + 3(L+1)^6\alpha_k\beta_k\right)z_k + \left(\frac{8\beta_k}{\mu_G} + 2L\beta_k^2\right)\|\Delta g_k\|^2 + 8L\alpha_k\|\Delta f_k\|^2] \\
&\leqslant (1-\lambda_k)\mathbb{E}[\|\Delta f_k\|^2] - (\frac{\lambda_k}{4} - \frac{48L^2\beta_k^2}{\lambda_k} - \frac{9\alpha_k}{2\mu_h} - 8L\alpha_k)\mathbb{E}[\|\Delta f_k\|^2] \\
&\quad + (1-\lambda_k)\mathbb{E}[\|\Delta g_k\|^2] - (\frac{\lambda_k}{4} - \frac{48L^2\beta_k^2}{\lambda_k} - \frac{8\beta_k}{\mu_G} - 2L\beta_k^2)\mathbb{E}[\|\Delta g_k\|^2] \\
&\quad + (1-\frac{\mu_h\alpha_k}{8})\mathbb{E}[z_k] - (\frac{\mu_h\alpha_k}{8} - 15(L+1)^6\alpha_k\beta_k - \frac{19(L+1)^6\alpha_k^2}{\mu_G\beta_k})\mathbb{E}[z_k] \\
&\quad + (1-\mu_G\beta_k)\mathbb{E}[y_k] - \left(\frac{\mu_G}{4}\beta_k - 8L^4\alpha_k - 21L^4\beta_k^2 - \frac{9L^2\alpha_k}{\mu_h}\right)\mathbb{E}[y_k] + 4B\lambda_k^2 \\
&\leqslant (1-\frac{\mu_h\alpha_k}{8})V_k + 4B\lambda_k^2 - (\frac{\lambda_k}{4} - \frac{48L^2\beta_k^2}{\lambda_k} - \frac{9\alpha_k}{2\mu_h} - 8L\alpha_k)\mathbb{E}[\|\Delta f_k\|^2] \\
&\quad - (\frac{\lambda_k}{4} - \frac{48L^2\beta_k^2}{\lambda_k} - \frac{8\beta_k}{\mu_G} - 2L\beta_k^2)\mathbb{E}[\|\Delta g_k\|^2] \\
&\quad - (\frac{\mu_h\alpha_k}{8} - 15(L+1)^6\alpha_k\beta_k - \frac{19(L+1)^6\alpha_k^2}{\mu_G\beta_k})\mathbb{E}[z_k] \\
&\quad - \left(\frac{\mu_G}{4}\beta_k - 8L^4\alpha_k - 21L^4\beta_k^2 - \frac{9L^2\alpha_k}{\mu_h}\right)\mathbb{E}[y_k], \qquad (28)
\end{aligned}$$

where the third inequality simplifies the terms with the conditions $\alpha_k \leqslant \frac{\lambda_k}{\mu_h}$ and $\alpha_k \leqslant \frac{8\mu_G\beta_k}{\mu_h}$.

The terms on the right hand side of (28) except for the first two are non-positive under the step size conditions $\beta_k \leqslant \frac{1}{\mu_G}$, $\beta_k \leqslant \frac{\lambda_k}{4}(56L^2 + \frac{9}{2\mu_h})^{-1}$, $\beta_k \leqslant \frac{\lambda_k}{4}(48L^2 + \frac{10L}{\mu_G})^{-1}$, $\alpha_k \leqslant \frac{\mu_h\mu_G}{152(L+1)^6}\beta_k$,

$\beta_k \leqslant \frac{1}{240(L+1)^6}$, $\beta_k \leqslant \frac{\mu_G}{168L^4}$, and $\alpha_k \leqslant \frac{\mu_G \beta_k}{8(8L^4+\frac{9L^2}{\mu_G})}$. This leads to

$$V_{k+1} \leqslant (1-\frac{\mu_h \alpha_k}{8})V_k + 4B\lambda_k^2 \leqslant (1-\frac{2}{k+\tau+1})V_k + 4B\lambda_k^2,$$

where the second inequality follows from $c_\alpha \geqslant \frac{16}{\mu_h}$.

Multiplying by $(k+\tau+1)^2$ on both sides, we have

$$\begin{aligned}(k+\tau+1)^2 V_{k+1} &\leqslant (k+\tau+1)(k+\tau-1)V_k + 4B(k+\tau+1)^2\lambda_k^2 \\ &\leqslant (k+\tau)^2 V_k + \frac{B(k+\tau+1)^2}{4(k+1)^2} \\ &\leqslant (k+\tau)^2 V_k + (\frac{B}{2} + \frac{B\tau^2}{2}) \\ &\leqslant \tau^2 V_0 + B\tau^2(k+1)\end{aligned}$$

Dividing by $(k+\tau+1)^2$ now, we have

$$V_{k+1} \leqslant \frac{\tau^2 V_0}{(k+\tau+1)^2} + \frac{B\tau^2(k+1)}{(k+\tau+1)^2} \leqslant \frac{\tau^2 V_0}{(k+\tau+1)^2} + \frac{B\tau^2}{(k+\tau+1)},$$

which obviously implies

$$\mathbb{E}[\|\theta_{k+1} - \theta^\star\|^2] \leqslant \frac{\tau^2 V_0}{(k+\tau+1)^2} + \frac{B\tau^2}{(k+\tau+1)}.$$

■

A.2. Proof of Lemma 1

By definition,

$$F(\theta^\star, \omega^\star(\theta)) = \nabla h(\theta^\star) = 0,$$

which implies

$$\begin{aligned}\|f_k\| &= \|\Delta f_k + F(\theta_k, \omega_k) - F(\theta_k, \omega^\star(\theta_k)) + F(\theta_k, \omega^\star(\theta_k)) - F(\theta^\star, \omega^\star(\theta^\star))\| \\ &\leqslant \|\Delta f_k\| + \|F(\theta_k, \omega_k) - F(\theta_k, \omega^\star(\theta_k))\| + \|F(\theta_k, \omega^\star(\theta_k)) - F(\theta^\star, \omega^\star(\theta^\star))\| \\ &\leqslant \|\Delta f_k\| + L\|\omega_k - \omega^\star(\theta_k)\| + L(L+1)\|\theta_k - \theta^\star\| \\ &\leqslant \|\Delta f_k\| + L\sqrt{y_k} + L(L+1)\sqrt{z_k},\end{aligned}$$

where the second inequality follows from the Lipschitz continuity of F.

Similarly, since $G(\theta, \omega^\star(\theta)) = 0$ for any θ, we can derive

$$\begin{aligned}\|g_k\| &= \|\Delta g_k + G(\theta_k, \omega_k) - G(\theta_k, \omega^\star(\theta))\| \\ &\leqslant \|\Delta g_k\| + \|G(\theta_k, \omega_k) - G(\theta_k, \omega^\star(\theta))\| \\ &\leqslant \|\Delta g_k\| + L\sqrt{y_k}.\end{aligned}$$

The bounds on $\|\theta_{k+1} - \theta_k\|$ and $\|\omega_{k+1} - \omega_k\|$ easily follow from the two inequalities above and (8).

■

A.3. Proof of Lemma 2

By the update rule in (9), we have

$$\begin{aligned}\Delta f_{k+1} &= f_{k+1} - F(\theta_{k+1},\omega_{k+1})\\ &= (1-\lambda_k)f_k + \lambda_k F(\theta_k,\omega_k,X_k) - F(\theta_{k+1},\omega_{k+1})\\ &= (1-\lambda_k)f_k + \lambda_k F(\theta_k,\omega_k) - F(\theta_{k+1},\omega_{k+1}) + \lambda_k\left(F(\theta_k,\omega_k) - F(\theta_k,\omega_k,X_k)\right)\\ &= (1-\lambda_k)\Delta f_k + F(\theta_k,\omega_k) - F(\theta_{k+1},\omega_{k+1}) + \lambda_k\left(F(\theta_k,\omega_k) - F(\theta_k,\omega_k,X_k)\right).\end{aligned}$$

This leads to

$$\begin{aligned}&\|\Delta f_{k+1}\|^2\\ &= (1-\lambda_k)^2\|\Delta f_k\|^2 + \left\|\left(F(\theta_k,\omega_k) - F(\theta_{k+1},\omega_{k+1})\right) + \lambda_k\left(F(\theta_k,\omega_k) - F(\theta_k,\omega_k,X_k)\right)\right\|^2\\ &\quad + 2(1-\lambda_k)\Delta f_k^\top\left(F(\theta_k,\omega_k) - F(\theta_{k+1},\omega_{k+1})\right)\\ &\quad + 2(1-\lambda_k)\lambda_k\Delta f_k^\top\left(F(\theta_k,\omega_k) - F(\theta_k,\omega_k,X_k)\right)\\ &\leqslant (1-\lambda_k)^2\|\Delta f_k\|^2 + 2\|F(\theta_k,\omega_k) - F(\theta_{k+1},\omega_{k+1})\|^2 + 2\lambda_k^2\|F(\theta_k,\omega_k) - F(\theta_k,\omega_k,X_k)\|^2\\ &\quad + \frac{\lambda_k}{2}\|\Delta f_k\|^2 + \frac{2}{\lambda_k}\|F(\theta_k,\omega_k) - F(\theta_{k+1},\omega_{k+1})\|^2\\ &\quad + 2(1-\lambda_k)\lambda_k\Delta f_k^\top\left(F(\theta_k,\omega_k) - F(\theta_k,\omega_k,X_k)\right)\\ &\leqslant (1-\lambda_k)\|\Delta f_k\|^2 + 2\lambda_k^2\|F(\theta_k,\omega_k) - F(\theta_k,\omega_k,X_k)\|^2\\ &\quad + (\lambda_k^2 - \frac{\lambda_k}{2})\|\Delta f_k\|^2 + \frac{4}{\lambda_k}\|F(\theta_k,\omega_k) - F(\theta_{k+1},\omega_{k+1})\|^2\\ &\quad + 2(1-\lambda_k)\lambda_k\Delta f_k^\top\left(F(\theta_k,\omega_k) - F(\theta_k,\omega_k,X_k)\right),\end{aligned} \tag{29}$$

where the last inequality follows from $\lambda_k \leqslant 1$.

The second term on the right hand side of (29) is bounded in expectation, i.e.

$$\mathbb{E}_{X_k\sim\xi}[\|F(\theta_k,\omega_k,X_k) - F(\theta_k,\omega_k)\|^2] \leqslant B.$$

In addition, the last term of (29) is zero in expectation since $\mathbb{E}[\Delta f_k^\top(F(\theta_k,\omega_k) - F(\theta_k,\omega_k,X_k)) \mid \mathcal{H}_{k-1}] = \Delta f_k^\top\mathbb{E}[(F(\theta_k,\omega_k) - F(\theta_k,\omega_k,X_k)) \mid \mathcal{H}_{k-1}] = 0$. As a result, we have

$$\begin{aligned}\mathbb{E}[\|\Delta f_{k+1}\|^2] \leqslant (1-\lambda_k)\mathbb{E}[\|\Delta f_k\|^2] - (\frac{\lambda_k}{2} - \lambda_k^2)\mathbb{E}[\|\Delta f_k\|^2]\\ + \frac{4}{\lambda_k}\mathbb{E}[\|F(\theta_k,\omega_k) - F(\theta_{k+1},\omega_{k+1})\|^2] + 2B\lambda_k^2.\end{aligned} \tag{30}$$

By Lemma 1 and the Lipschitz condition of F,

$$\begin{aligned}&\|F(\theta_k,\omega_k) - F(\theta_{k+1},\omega_{k+1})\|^2\\ &\leqslant 2L^2\|\theta_{k+1} - \theta_k\|^2 + 2L^2\|\omega_{k+1} - \omega_k\|^2\\ &\leqslant 6L^2\alpha_k^2(\|\Delta f_k\|^2 + L^2 y_k + (L+1)^4 z_k) + 4L^2\beta_k^2(\|\Delta g_k\|^2 + L^2 y_k)\\ &\leqslant 6L^2\beta_k^2\|\Delta f_k\|^2 + 6L^2\beta_k^2\|\Delta g_k\|^2 + 10L^4\beta_k^2 y_k + 6(L+1)^6\alpha_k^2 z_k.\end{aligned}$$

Plugging this bound into (30), we get

$$\begin{aligned}
&\mathbb{E}[\|\Delta f_{k+1}\|^2] \\
&\leqslant (1-\lambda_k)\mathbb{E}[\|\Delta f_k\|^2] - (\frac{\lambda_k}{2} - \lambda_k^2)\mathbb{E}[\|\Delta f_k\|^2] \\
&\quad + \frac{4}{\lambda_k}\mathbb{E}\left[6L^2\beta_k^2\|\Delta f_k\|^2 + 6L^2\beta_k^2\|\Delta g_k\|^2 + 10L^4\beta_k^2 y_k + 6(L+1)^6\alpha_k^2 z_k\right] + 2B\lambda_k^2 \\
&\leqslant (1-\lambda_k)\mathbb{E}[\|\Delta f_k\|^2] - (\frac{\lambda_k}{4} - \frac{24L^2\beta_k^2}{\lambda_k})\mathbb{E}[\|\Delta f_k\|^2] \\
&\quad + \frac{4}{\lambda_k}\mathbb{E}\left[6L^2\beta_k^2\|\Delta g_k\|^2 + 10L^4\beta_k^2 y_k + 6(L+1)^6\alpha_k^2 z_k\right] + 2B\lambda_k^2,
\end{aligned}$$

where the last inequality follows from $\lambda_k \leqslant 1/4$.

■

A.4. Proof of Lemma 3

By the update rule in (9), we have

$$\begin{aligned}
\Delta g_{k+1} &= g_{k+1} - G(\theta_{k+1}, \omega_{k+1}) \\
&= (1-\lambda_k)g_k + \lambda_k G(\theta_k, \omega_k, X_k) - G(\theta_{k+1}, \omega_{k+1}) \\
&= (1-\lambda_k)g_k + \lambda_k G(\theta_k, \omega_k) - G(\theta_{k+1}, \omega_{k+1}) + \lambda_k\left(G(\theta_k, \omega_k) - G(\theta_k, \omega_k, X_k)\right) \\
&= (1-\lambda_k)\Delta g_k + G(\theta_k, \omega_k) - G(\theta_{k+1}, \omega_{k+1}) + \lambda_k\left(G(\theta_k, \omega_k) - G(\theta_k, \omega_k, X_k)\right).
\end{aligned}$$

This leads to

$$\begin{aligned}
&\|\Delta g_{k+1}\|^2 \\
&= (1-\lambda_k)^2\|\Delta g_k\|^2 + \left\|\left(G(\theta_k, \omega_k) - G(\theta_{k+1}, \omega_{k+1})\right) + \lambda_k\left(G(\theta_k, \omega_k) - G(\theta_k, \omega_k, X_k)\right)\right\|^2 \\
&\quad + 2(1-\lambda_k)\Delta g_k^\top\left(G(\theta_k, \omega_k) - G(\theta_{k+1}, \omega_{k+1})\right) \\
&\quad + 2(1-\lambda_k)\lambda_k\Delta g_k^\top\left(G(\theta_k, \omega_k) - G(\theta_k, \omega_k, X_k)\right) \\
&\leqslant (1-\lambda_k)^2\|\Delta g_k\|^2 + 2\|G(\theta_k, \omega_k) - G(\theta_{k+1}, \omega_{k+1})\|^2 + 2\lambda_k^2\|G(\theta_k, \omega_k) - G(\theta_k, \omega_k, X_k)\|^2 \\
&\quad + \frac{\lambda_k}{2}\|\Delta g_k\|^2 + \frac{2}{\lambda_k}\|G(\theta_k, \omega_k) - G(\theta_{k+1}, \omega_{k+1})\|^2 \\
&\quad + 2(1-\lambda_k)\lambda_k\Delta g_k^\top\left(G(\theta_k, \omega_k) - G(\theta_k, \omega_k, X_k)\right) \\
&\leqslant (1-\lambda_k)\|\Delta g_k\|^2 + 2\lambda_k^2\|G(\theta_k, \omega_k) - G(\theta_k, \omega_k, X_k)\|^2 \\
&\quad + (\lambda_k^2 - \frac{\lambda_k}{2})\|\Delta g_k\|^2 + \frac{4}{\lambda_k}\|G(\theta_k, \omega_k) - G(\theta_{k+1}, \omega_{k+1})\|^2 \\
&\quad + 2(1-\lambda_k)\lambda_k\Delta g_k^\top\left(G(\theta_k, \omega_k) - G(\theta_k, \omega_k, X_k)\right). \qquad (31)
\end{aligned}$$

The second term on the right hand side of (31) is bounded in expectation, i.e.

$$\mathbb{E}_{X_k\sim\xi}[\|G(\theta_k, \omega_k, X_k) - G(\theta_k, \omega_k)\|^2] \leqslant B.$$

In addition, the last term of (29) is zero in expectation since $\mathbb{E}[\Delta g_k^\top (G(\theta_k,\omega_k) - G(\theta_k,\omega_k,X_k)) \mid \mathcal{H}_{k-1}] = \Delta g_k^\top \mathbb{E}[(G(\theta_k,\omega_k) - G(\theta_k,\omega_k,X_k)) \mid \mathcal{H}_{k-1}] = 0$. As a result, we have

$$\mathbb{E}[\|\Delta g_{k+1}\|^2] \leqslant (1-\lambda_k)\mathbb{E}[\|\Delta g_k\|^2] - (\frac{\lambda_k}{2} - \lambda_k^2)\mathbb{E}[\|\Delta g_k\|^2] + \frac{4}{\lambda_k}\mathbb{E}[\|G(\theta_k,\omega_k) - G(\theta_{k+1},\omega_{k+1})\|^2] + 2B\lambda_k^2. \tag{32}$$

By Lemma 1 and the Lipschitz condition of G,

$$\begin{aligned}
&\|G(\theta_k,\omega_k) - G(\theta_{k+1},\omega_{k+1})\|^2 \\
&\leqslant 2L^2\|\theta_{k+1}-\theta_k\|^2 + 2L^2\|\omega_{k+1}-\omega_k\|^2 \\
&\leqslant 6L^2\alpha_k^2(\|\Delta f_k\|^2 + L^2 y_k + (L+1)^4 z_k) + 4L^2\beta_k^2(\|\Delta g_k\|^2 + L^2 y_k) \\
&\leqslant 6L^2\beta_k^2\|\Delta f_k\|^2 + 6L^2\beta_k^2\|\Delta g_k\|^2 + 10L^4\beta_k^2 y_k + 6(L+1)^6\alpha_k^2 z_k.
\end{aligned}$$

Plugging this bound into (32), we get

$$\begin{aligned}
\mathbb{E}[\|\Delta g_{k+1}\|^2] &\leqslant (1-\lambda_k)\mathbb{E}[\|\Delta g_k\|^2] - (\frac{\lambda_k}{2} - \lambda_k^2)\mathbb{E}[\|\Delta g_k\|^2] + 2B\lambda_k^2 \\
&\quad + \frac{4}{\lambda_k}\mathbb{E}\left[6L^2\beta_k^2\|\Delta f_k\|^2 + 6L^2\beta_k^2\|\Delta g_k\|^2 + 10L^4\beta_k^2 y_k + 6(L+1)^6\alpha_k^2 z_k\right] \\
&\leqslant (1-\lambda_k)\mathbb{E}[\|\Delta g_k\|^2] - (\frac{\lambda_k}{4} - \frac{24L^2\beta_k^2}{\lambda_k})\mathbb{E}[\|\Delta g_k\|^2] \\
&\quad + \frac{4}{\lambda_k}\mathbb{E}\left[6L^2\beta_k^2\|\Delta f_k\|^2 + 10L^4\beta_k^2 y_k + 6(L+1)^6\alpha_k^2 z_k\right] + 2B\lambda_k^2,
\end{aligned}$$

where the last inequality follows from $\lambda_k \leqslant 1/4$.

■

A.5. Proof of Lemma 4

We have by the update rule in (8) and the definition of the residual variables in (11)

$$\begin{aligned}
z_{k+1} &= \|\theta_{k+1} - \theta^\star\|^2 \\
&= \|\theta_k - \alpha_k f_k - \theta^\star\|^2 \\
&= \|\theta_k - \alpha_k F(\theta_k,\omega_k) - \alpha_k \Delta f_k - \theta^\star\|^2 \\
&= \|\theta_k - \theta^\star - \alpha_k F(\theta_k,\omega_k)\|^2 - 2\alpha_k\langle \theta_k - \theta^\star - \alpha_k F(\theta_k,\omega_k), \Delta f_k\rangle + \alpha_k^2\|\Delta f_k\|^2.
\end{aligned} \tag{33}$$

To bound the first term on the right hand side of (33),

$$\begin{aligned}
&\|\theta_k - \theta^\star - \alpha_k F(\theta_k,\omega_k)\|^2 \\
&= \|\theta_k - \theta^\star - \alpha_k\left(F(\theta_k,\omega^\star(\theta_k)) - F(\theta^\star,\omega^\star(\theta^\star))\right) + \alpha_k\left(F(\theta_k,\omega^\star(\theta_k)) - F(\theta_k,\omega_k)\right)\|^2 \\
&= z_k + \alpha_k^2\|F(\theta_k,\omega^\star(\theta_k)) - F(\theta^\star,\omega^\star(\theta^\star))\|^2 + \alpha_k^2\|F(\theta_k,\omega^\star(\theta_k)) - F(\theta_k,\omega_k)\|^2 \\
&\quad - 2\alpha_k\langle \theta_k - \theta^\star, F(\theta_k,\omega^\star(\theta_k)) - F(\theta^\star,\omega^\star(\theta^\star))\rangle \\
&\quad + 2\alpha_k\langle \theta_k - \theta^\star, F(\theta_k,\omega^\star(\theta_k)) - F(\theta_k,\omega_k)\rangle
\end{aligned}$$

$$
\begin{aligned}
&- 2\alpha_k^2\langle F(\theta_k,\omega^\star(\theta_k)) - F(\theta^\star,\omega^\star(\theta^\star)), F(\theta_k,\omega^\star(\theta_k)) - F(\theta_k,\omega_k)\rangle \\
&\leqslant z_k + \alpha_k^2\|F(\theta_k,\omega^\star(\theta_k)) - F(\theta^\star,\omega^\star(\theta^\star))\|^2 + \alpha_k^2\|F(\theta_k,\omega^\star(\theta_k)) - F(\theta_k,\omega_k)\|^2 \\
&\quad - 2\mu_h\alpha_k z_k + \mu_h\alpha_k z_k + \frac{\alpha_k}{\mu_h}\|F(\theta_k,\omega^\star(\theta_k)) - F(\theta_k,\omega_k)\|^2 \\
&\quad + \alpha_k^2 z_k + \alpha_k^2\|F(\theta_k,\omega^\star(\theta_k)) - F(\theta_k,\omega_k)\|^2 \\
&\leqslant (1 - \mu_h\alpha_k + \alpha_k^2)z_k + 2L^2(L+1)^2\alpha_k^2 z_k + (2L^2\alpha_k^2 + \frac{L^2\alpha_k}{\mu_h})y_k \\
&\leqslant (1 - \mu_h\alpha_k + 3(L+1)^4\alpha_k^2)z_k + (2L^2\alpha_k^2 + \frac{L^2\alpha_k}{\mu_h})y_k,
\end{aligned}
$$

where the first equality follows from $F(\theta^\star,\omega^\star(\theta^\star)) = 0$, the first inequality plugs in (13) and uses the fact that $2\|\boldsymbol{a}\|\|\boldsymbol{b}\| \leqslant c\|\boldsymbol{a}\|^2 + \frac{1}{c}\|\boldsymbol{b}\|^2$ for any vector $\boldsymbol{a}, \boldsymbol{b}$ and scalar $c > 0$, and the second inequality follows from the Lipschitz continuity of F and $\omega^\star$. Choosing the step size such that $\alpha_k \leqslant \frac{\mu_h}{6(L+1)^4}$ and $\alpha_k \leqslant \frac{1}{2\mu_h}$ results in

$$
\|\theta_k - \theta^\star - \alpha_k F(\theta_k,\omega_k)\|^2 \leqslant (1 - \frac{\mu_h\alpha_k}{2})z_k + \frac{2L^2\alpha_k}{\mu_h}y_k. \tag{34}
$$

The second term on the right hand side of (33) can be treated based on (34)

$$
\begin{aligned}
&- 2\alpha_k\langle\theta_k - \theta^\star - \alpha_k F(\theta_k,\omega_k), \Delta f_k\rangle \\
&\leqslant \frac{\mu_h\alpha_k}{4}\|\theta_k - \theta^\star - \alpha_k F(\theta_k,\omega_k)\|^2 + \frac{4\alpha_k}{\mu_h}\|\Delta f_k\|^2 \\
&\leqslant \frac{\mu_h\alpha_k}{4}(1 - \frac{\mu_h\alpha_k}{2})z_k + \frac{\mu_h\alpha_k}{4}\cdot\frac{2L^2\alpha_k}{\mu_h}y_k + \frac{4\alpha_k}{\mu_h}\|\Delta f_k\|^2 \\
&\leqslant (\frac{\mu_h\alpha_k}{4} - \frac{\mu_h^2\alpha_k^2}{8})z_k + \frac{L^2\alpha_k^2}{2}y_k + \frac{4\alpha_k}{\mu_h}\|\Delta f_k\|^2.
\end{aligned} \tag{35}
$$

Plugging (34) and (35) into (33), we get

$$
\begin{aligned}
z_{k+1} &\leqslant (1 - \frac{\mu_h\alpha_k}{2})z_k + \frac{2L^2\alpha_k}{\mu_h}y_k \\
&\quad + (\frac{\mu_h\alpha_k}{4} - \frac{\mu_h^2\alpha_k^2}{8})z_k + \frac{L^2\alpha_k^2}{2}y_k + \frac{4\alpha_k}{\mu_h}\|\Delta f_k\|^2 + \alpha_k^2\|\Delta f_k\|^2 \\
&\leqslant (1 - \frac{\mu_h\alpha_k}{4})z_k + \frac{9L^2\alpha_k}{\mu_h}y_k + \frac{9\alpha_k}{2\mu_h}\|\Delta f_k\|^2,
\end{aligned}
$$

where again we use $\alpha_k \leqslant \frac{1}{2\mu_h}$ to simplify the terms.

■

A.6. Proof of Lemma 5

By the update rule in (8),

$$
\omega_{k+1} - \omega^\star(\theta_{k+1}) = \omega_{k+1} - \omega^\star(\theta_{k+1})
$$

$$
\begin{aligned}
&= (\omega_k - \omega^\star(\theta_k)) - \beta_k g_k - (\omega^\star(\theta_{k+1}) - \omega^\star(\theta_k)) \\
&= (\omega_k - \omega^\star(\theta_k)) - \beta_k G(\theta_k, \omega_k) - \beta_k \Delta g_k - (\omega^\star(\theta_{k+1}) - \omega^\star(\theta_k)) \,.
\end{aligned}
$$

Taking the norm yields

$$
\begin{aligned}
y_{k+1} &= \|\omega_{k+1} - \omega^\star(\theta_{k+1})\|^2 \\
&= \underbrace{\|\omega_k - \omega^\star(\theta_k) - \beta_k G(\theta_k, \omega_k)\|^2}_{T_1} + \beta_k^2 \|\Delta g_k\|^2 + \underbrace{\|\omega^\star(\theta_{k+1}) - \omega^\star(\theta_k)\|^2}_{T_2} \\
&\quad \underbrace{-2 \left(\omega_k - \omega^\star(\theta_k) - \beta_k G(\theta_k, \omega_k)\right)^\top \left(\omega^\star(\theta_{k+1}) - \omega^\star(\theta_k)\right)}_{T_3} \\
&\quad \underbrace{-2\beta_k \left(\omega_k - \omega^\star(\theta_k) - \beta_k G(\theta_k, \omega_k)\right)^\top \Delta g_k}_{T_4} + \underbrace{2\beta_k \Delta g_k^\top \left(\omega^\star(\theta_{k+1}) - \omega^\star(\theta_k)\right)}_{T_5} \,. \qquad (36)
\end{aligned}
$$

We bound each term of (36) individually. First, since by definition $G(\theta, \omega^\star(\theta)) = 0$ for all θ, we have

$$
\begin{aligned}
T_1 &= \|\omega_k - \omega^\star(\theta_k)\|^2 - 2\beta_k \left(\omega_k - \omega^\star(\theta_k)\right)^\top G(\theta_k, \omega_k) + \beta_k^2 \|G(\theta_k, \omega_k)\|^2 \\
&= y_k - 2\beta_k \left(\omega_k - \omega^\star(\theta_k)\right)^\top G(\theta_k, \omega_k) + \beta_k^2 \|G(\theta_k, \omega_k) - G(\theta_k, \omega^\star(\theta_k))\|^2 \\
&\leqslant y_k - 2\mu_G \beta_k \|\omega_k - \omega^\star(\theta_k)\|^2 + L^2 \beta_k^2 \|\omega_k - \omega^\star(\theta_k)\|^2 \\
&= (1 - 2\mu_G \beta_k + L^2 \beta_k^2) y_k,
\end{aligned}
$$

where the first inequality follows from Assumption 3 and the Lipschitz continuity of G.

The term T_2 can be simply treated using the Lipschitz condition of $\omega^\star$

$$
\begin{aligned}
T_2 &= \|\omega^\star(\theta_{k+1}) - \omega^\star(\theta_k)\|^2 \leqslant L^2 \|\theta_{k+1} - \theta_k\|^2 \leqslant L^2 \alpha_k^2 \left(\|\Delta f_k\| + L\sqrt{y_k} + L(L+1)\sqrt{z_k}\right)^2 \\
&\leqslant 3L^2 \alpha_k^2 \left(\|\Delta f_k\|^2 + L^2 y_k + (L+1)^4 z_k\right) .
\end{aligned}
$$

By the Cauchy-Schwarz inequality,

$$
\begin{aligned}
&T_3 \\
&\leqslant 2\|\omega_k - \omega^\star(\theta_k) - \beta_k G(\theta_k, \omega_k)\| \|\omega^\star(\theta_{k+1}) - \omega^\star(\theta_k)\| \\
&\leqslant 2\|\omega_k - \omega^\star(\theta_k)\| \|\omega^\star(\theta_{k+1}) - \omega^\star(\theta_k)\| + 2\beta_k \|G(\theta_k, \omega_k) - G(\theta_k, \omega^\star(\theta_k))\| \|\omega^\star(\theta_{k+1}) - \omega^\star(\theta_k)\| \\
&\leqslant (2L + 2L^2 \beta_k) \sqrt{y_k} \|\theta_{k+1} - \theta_k\| \\
&\leqslant 4L\alpha_k \sqrt{y_k} \left(\|\Delta f_k\| + L\sqrt{y_k} + L(L+1)\sqrt{z_k}\right) .
\end{aligned}
$$

where the second inequality follows from $G(\theta, \omega^\star(\theta)) = 0$ for all θ, and the fourth inequality applies Lemma 1 and the step size rule $\beta_k \leqslant 1/L$.

We can bound T_4 following a similar line of analysis

$$
T_4 \leqslant 2\beta_k \|\omega_k - \omega^\star(\theta_k) - \beta_k G(\theta_k, \omega_k)\| \|\Delta g_k\| \leqslant 4\beta_k \sqrt{y_k} \|\Delta g_k\| .
$$

Finally, we treat T_5 again by the Cauchy-Schwarz inequality

$$
T_5 \leqslant 2\beta_k \|\Delta g_k\| \|\omega^\star(\theta_{k+1}) - \omega^\star(\theta_k)\|
$$

$$\leqslant 2L\alpha_k\beta_k\|\Delta g_k\|\Big(\|\Delta f_k\| + L\sqrt{y_k} + L(L+1)\sqrt{z_k}\Big),$$

where the second inequality again uses Lemma 1.

Applying the bounds on T_1-T_5 to (36), we get

$$\begin{aligned}
y_{k+1} &\leqslant (1 - 2\mu_G\beta_k + L^2\beta_k^2)y_k + \beta_k^2\|\Delta g_k\|^2 + 3L^2\alpha_k^2\left(\|\Delta f_k\|^2 + L^2 y_k + (L+1)^4 z_k\right) \\
&\quad + 4L\alpha_k\sqrt{y_k}\left(\|\Delta f_k\| + L\sqrt{y_k} + L(L+1)\sqrt{z_k}\right) + 4\beta_k\sqrt{y_k}\|\Delta g_k\| \\
&\quad + 2L\alpha_k\beta_k\|\Delta g_k\|\Big(\|\Delta f_k\| + L\sqrt{y_k} + L(L+1)\sqrt{z_k}\Big) \\
&\leqslant (1 - 2\mu_G\beta_k + L^2\beta_k^2)y_k + \beta_k^2\|\Delta g_k\|^2 + 3L^2\alpha_k^2\left(\|\Delta f_k\|^2 + L^2 y_k + (L+1)^4 z_k\right) \\
&\quad + 2L\alpha_k y_k + 2L\alpha_k\|\Delta f_k\|^2 + 4L^2\alpha_k y_k + \frac{\mu_G}{4}\beta_k y_k + \frac{16(L+1)^6\alpha_k^2}{\mu_G\beta_k} z_k + \frac{\mu_G\beta_k}{2} y_k \\
&\quad + \frac{8\beta_k}{\mu_G}\|\Delta g_k\|^2 + L\alpha_k\beta_k\|\Delta g_k\|^2 + 3L\alpha_k\beta_k\left(\|\Delta f_k\|^2 + L^2 y_k + (L+1)^4 z_k\right) \\
&\leqslant (1 - \mu_G\beta_k)y_k - \left(\frac{\mu_G}{4}\beta_k - 8L^4\alpha_k - L^2\beta_k^2\right) y_k \\
&\quad + \left(\frac{19(L+1)^6\alpha_k^2}{\mu_G\beta_k} + 3(L+1)^6\alpha_k\beta_k\right) z_k + \left(\frac{8\beta_k}{\mu_G} + 2L\beta_k^2\right)\|\Delta g_k\|^2 + 8L\alpha_k\|\Delta f_k\|^2,
\end{aligned}$$

where the second inequality follows from the fact that $2\|\boldsymbol{a}\|\|\boldsymbol{b}\| \leqslant c\|\boldsymbol{a}\|^2 + \frac{1}{c}\|\boldsymbol{b}\|^2$ for any vector $\boldsymbol{a}, \boldsymbol{b}$ and scalar $c > 0$. We have also simplified terms using the conditions $\alpha_k \leqslant \beta_k \leqslant \frac{1}{72(L+1)^6}$ and $\beta_k \leqslant 1/\mu_G$.

■

Appendix B. Analysis for Convex Functions

Our analysis in this section relies on Lemma 2, 3, and 5, which we have established for the proof of Theorem 1. Note that these lemmas do not require the function h to be strongly convex and therefore can be applied here as well. We also introduce the following additional lemma on the iteration-wise drift of z_k. The proof of this lemma can be found at the end of this section.

Lemma 6 *Under Assumptions made in Theorem 2 and the step sizes in (15), we have*

$$z_{k+1} \leqslant z_k + (16 + \frac{4L^2}{\mu_G})\frac{\alpha_k^2 z_k}{\beta_k} - 2\alpha_k\Big(h(\theta_k) - h(\theta^\star)\Big) + 3\alpha_k^2\|\Delta f_k\|^2 + 3L^2\alpha_k^2 y_k.$$

B.1. Proof of Theorem 2

We consider the Lyapunov function

$$V_k \triangleq \mathbb{E}[\|\Delta f_k\|^2 + \|\Delta g_k\|^2 + z_k + y_k].$$

Combining the results from Lemma 2, 3, 5, and 6, and simplifying the terms with the choice of step sizes given in (15),

V_{k+1}

$$
\begin{aligned}
&= \mathbb{E}[\|\Delta f_{k+1}\|^2 + \|\Delta g_{k+1}\|^2 + z_{k+1} + y_{k+1}] \\
&\leqslant (1-\lambda_k)\mathbb{E}[\|\Delta f_k\|^2] - (\frac{\lambda_k}{4} - \frac{24L^2\beta_k^2}{\lambda_k})\mathbb{E}[\|\Delta f_k\|^2] \\
&\quad + \frac{4}{\lambda_k}\mathbb{E}\left[6L^2\beta_k^2\|\Delta g_k\|^2 + 10L^4\beta_k^2 y_k + 6(L+1)^6\alpha_k^2 z_k\right] + 2B\lambda_k^2 \\
&\quad + (1-\lambda_k)\mathbb{E}[\|\Delta g_k\|^2] - (\frac{\lambda_k}{4} - \frac{24L^2\beta_k^2}{\lambda_k})\mathbb{E}[\|\Delta g_k\|^2] \\
&\quad + \frac{4}{\lambda_k}\mathbb{E}\left[6L^2\beta_k^2\|\Delta f_k\|^2 + 10L^4\beta_k^2 y_k + 6(L+1)^6\alpha_k^2 z_k\right] + 2B\lambda_k^2 \\
&\quad + \mathbb{E}\Big[z_k + (16 + \frac{4L^2}{\mu_G})\frac{\alpha_k^2 z_k}{\beta_k} - 2\alpha_k\Big(h(\theta_k) - h(\theta^\star)\Big) + 3\alpha_k^2\|\Delta f_k\|^2 + 3L^2\alpha_k^2 y_k \\
&\quad + (1-\mu_G\beta_k)y_k - \left(\frac{\mu_G}{4}\beta_k - 8L^4\alpha_k - L^2\beta_k^2\right)y_k \\
&\quad + \left(\frac{19(L+1)^6\alpha_k^2}{\mu_G\beta_k} + 3(L+1)^6\alpha_k\beta_k\right)z_k + \left(\frac{8\beta_k}{\mu_G} + 2L\beta_k^2\right)\|\Delta g_k\|^2 + 8L\alpha_k\|\Delta f_k\|^2\Big] \\
&\leqslant (1-\lambda_k)\mathbb{E}[\|\Delta f_k\|^2] - (\frac{\lambda_k}{4} - \frac{48L^2\beta_k^2}{\lambda_k} - 11L\alpha_k)\mathbb{E}[\|\Delta f_k\|^2] \\
&\quad + (1-\lambda_k)\mathbb{E}[\|\Delta g_k\|^2] - (\frac{\lambda_k}{4} - (50L^2 + \frac{8}{\mu_G})\beta_k)\mathbb{E}[\|\Delta g_k\|^2] \\
&\quad + (1 + (16 + \frac{23(L+1)^6}{\mu_G})\frac{\alpha_k^2}{\beta_k} + 12(L+1)^6\alpha_k^2)\mathbb{E}[z_k] - 2\alpha_k\mathbb{E}[h(\theta_k) - h(\theta^\star)] \\
&\quad + (1-\mu_G\beta_k)\mathbb{E}[y_k] - \left(\frac{\mu_G}{2}\beta_k - 11L^4\alpha_k - L^2\beta_k^2 - \frac{80L^4\beta_k^2}{\lambda_k}\right)\mathbb{E}[y_k] + 4B\lambda_k^2 \\
&\leqslant (1-\lambda_k)\mathbb{E}[\|\Delta f_k\|^2] + (1-\lambda_k)\mathbb{E}[\|\Delta g_k\|^2] + (1 + (16 + \frac{23(L+1)^6}{\mu_G})\frac{\alpha_k^2}{\beta_k} + 12(L+1)^6\alpha_k^2)\mathbb{E}[z_k] \\
&\quad + (1-\mu_G\beta_k)\mathbb{E}[y_k] - 2\alpha_k\mathbb{E}[h(\theta_k) - h(\theta^\star)] + 4B\lambda_k^2 \\
&\leqslant (1 + (16 + \frac{23(L+1)^6}{\mu_G})\frac{\alpha_k^2}{\beta_k} + 12(L+1)^6\alpha_k^2)V_k - 2\alpha_k\mathbb{E}[h(\theta_k) - h(\theta^\star)] + 4B\lambda_k^2,
\end{aligned}
$$

where we use the step size conditions $\alpha_k \leqslant \frac{1}{44L^4}\lambda_k$, $\beta_k \leqslant \frac{\lambda_k}{4}\min\{(11L+48L^2)^{-1}, (50L^2+\frac{8}{\mu_G})^{-1}\}$, $\beta_k \leqslant \frac{\mu_G}{640L^4}\lambda_k$, and $\beta_k \leqslant \frac{\mu_G}{16L^2}$.

Re-arranging the inequality, we have for all $k = 0, 1, \cdots, K-1$

$$
\begin{aligned}
2\alpha_k\mathbb{E}[h(\theta_k) - h(\theta^\star)] &\leqslant (1 + (16 + \frac{23(L+1)^6}{\mu_G})\frac{\alpha_k^2}{\beta_k} + 12(L+1)^6\alpha_k^2)V_k - V_{k+1} + 4B\lambda_k^2 \\
&\leqslant (1 + \frac{1}{K+1})V_k - V_{k+1} + \frac{B\lambda_0^2}{4(K+1)},
\end{aligned}
$$

since c_α and c_β are chosen such that $(16 + \frac{23(L+1)^6}{\mu_G})\frac{c_\alpha^2}{c_\beta} + 12(L+1)^6 c_\alpha^2 \leqslant 1$.

Defining $w_k = \frac{1}{(1+\frac{1}{K+1})^k}$, we have

$$
\frac{2c_\alpha}{\sqrt{K+1}}\sum_{k=0}^{K} w_k\mathbb{E}[h(\theta_k) - h(\theta^\star)] \leqslant \sum_{k=0}^{K}\left((1 + \frac{1}{K+1})w_k V_k - w_k V_{k+1} + \frac{B\lambda_0^2}{4(K+1)}w_k\right)
$$

$$\leqslant V_0 + \frac{B\lambda_0^2}{4(K+1)} \sum_{k=0}^{K} w_k. \tag{37}$$

The sum of the normalizing factor can be bounded as

$$\sum_{k=0}^{K} w_k = \sum_{k=0}^{K} \frac{1}{(1+\frac{1}{K+1})^k} \geqslant 2c_\alpha(K+1)(1+\frac{1}{K+1})^K \geqslant 4c_\alpha(K+1), \tag{38}$$

where the last inequality follows from $(1+\frac{1}{K+1})^K \geqslant 2$ for any $K \geqslant 1$.

Combining (37) and (38), we have

$$\begin{aligned}\frac{1}{\sum_{k'=0}^{K} w_{k'}} \sum_{k=0}^{K} w_k \mathbb{E}[h(\theta_k) - h(\theta^\star)] &\leqslant \frac{\sqrt{K+1}}{2c_\alpha \sum_{k'=0}^{K} w_{k'}} \left(V_0 + \frac{B\lambda_0^2}{4(K+1)} \sum_{k=0}^{K} w_k \right) \\ &\leqslant \frac{V_0}{8c_\alpha^2\sqrt{K+1}} + \frac{B\lambda_0^2}{8c_\alpha\sqrt{K+1}}. \end{aligned} \tag{39}$$

Let $\bar{\theta}_K = \frac{\sum_{k=0}^{K} w_k \theta_k}{\sum_{k'=0}^{K} w_{k'}}$. Since the function h is convex, the claimed result easily follows from (39). ■

B.2. Proof of Lemma 6

From the update rule (8), we have

$$\begin{aligned} z_{k+1} &= \|\theta_{k+1} - \theta^\star\|^2 \\ &= \|\theta_k - \alpha_k f_k - \theta^\star\|^2 \\ &= \|\theta_k - \theta^\star\|^2 + \alpha_k^2\|f_k\|^2 - 2\alpha_k\langle\theta_k - \theta^\star, f_k\rangle \\ &= \|\theta_k - \theta^\star\|^2 + \alpha_k^2\|f_k\|^2 - 2\alpha_k\langle\theta_k - \theta^\star, \Delta f_k\rangle - 2\alpha_k\langle\theta_k - \theta^\star, F(\theta_k, \omega_k)\rangle \\ &= z_k + \alpha_k^2\|f_k\|^2 - 2\alpha_k\langle\theta_k - \theta^\star, \Delta f_k\rangle - 2\alpha_k\langle\theta_k - \theta^\star, F(\theta_k, \omega^\star(\theta_k))\rangle \\ &\quad - 2\alpha_k\langle\theta_k - \theta^\star, F(\theta_k, \omega_k) - F(\theta_k, \omega^\star(\theta_k))\rangle \\ &\leqslant z_k - 2\alpha_k\Big(h(\theta_k) - h(\theta^\star)\Big) + \alpha_k^2\|f_k\|^2 - 2\alpha_k\langle\theta_k - \theta^\star, \Delta f_k\rangle \\ &\quad - 2\alpha_k\langle\theta_k - \theta^\star, F(\theta_k, \omega_k) - F(\theta_k, \omega^\star(\theta_k))\rangle, \end{aligned} \tag{40}$$

where the last inequality follows from the convexity of h, i.e. $\langle\theta_k - \theta^\star, \nabla h(\theta_k)\rangle \geqslant h(\theta_k) - h(\theta^\star)$.

To bound the fourth term on the right hand side of (40),

$$-2\alpha_k\langle\theta_k - \theta^\star, \Delta f_k\rangle \leqslant 2\alpha_k\|\theta_k - \theta^\star\|\|\Delta f_k\| \leqslant \frac{16\alpha_k^2}{\lambda_k} z_k + \frac{\lambda_k}{16}\|\Delta f_k\|^2, \tag{41}$$

where the final inequality follows from the fact that $2\|\boldsymbol{a}\|\|\boldsymbol{b}\| \leqslant c\|\boldsymbol{a}\|^2 + \frac{1}{c}\|\boldsymbol{b}\|^2$ for any vector $\boldsymbol{a}, \boldsymbol{b}$ and scalar $c > 0$.

The fifth term on the right hand side of (40) can be treated in a similar manner

$$-2\alpha_k\langle\theta_k - \theta^\star, F(\theta_k, \omega^\star(\theta_k)) - F(\theta_k, \omega_k)\rangle \leqslant 2\alpha_k\|\theta_k - \theta^\star\|\|F(\theta_k, \omega^\star(\theta_k)) - F(\theta_k, \omega_k)\|$$

$$\leqslant 2L\alpha_k\|\theta_k-\theta^\star\|\|\omega^\star(\theta_k)-\omega_k\| \\ \leqslant \frac{4L^2\alpha_k^2}{\mu_G\beta_k}z_k+\frac{\mu_G}{4}\beta_k y_k. \tag{42}$$

Plugging (41) and (42) into (40), we get

$$\begin{aligned} z_{k+1} &\leqslant z_k-2\alpha_k\Big(h(\theta_k)-h(\theta^\star)\Big)+\alpha_k^2\|f_k\|^2-2\alpha_k\langle\theta_k-\theta^\star,\Delta f_k\rangle \\ &\quad -2\alpha_k\langle\theta_k-\theta^\star,F(\theta_k,\omega_k)-F(\theta_k,\omega^\star(\theta_k))\rangle \\ &\leqslant z_k-2\alpha_k\Big(h(\theta_k)-h(\theta^\star)\Big)+3\alpha_k^2\left(\|\Delta f_k\|^2+L^2y_k+(L+1)^4z_k\right) \\ &\quad +\frac{16\alpha_k^2}{\lambda_k}z_k+\frac{\lambda_k}{16}\|\Delta f_k\|^2+\frac{4L^2\alpha_k^2}{\mu_G\beta_k}z_k+\frac{\mu_G}{4}\beta_k y_k \\ &\leqslant z_k+(16+\frac{4L^2}{\mu_G})\frac{\alpha_k^2 z_k}{\beta_k}-2\alpha_k\Big(h(\theta_k)-h(\theta^\star)\Big)+3\alpha_k^2\|\Delta f_k\|^2+3L^2\alpha_k^2y_k. \end{aligned}$$

■

Appendix C. Analysis for Functions Satisfying Polyak-Lojasiewicz Condition

The exact requirement on step sizes is

$$\lambda_k=\frac{c_\lambda}{k+\tau+1},\quad \alpha_k=\frac{c_\alpha}{k+\tau+1},\quad \beta_k=\frac{c_\beta}{k+\tau+1}, \tag{43}$$

where $c_\alpha, c_\beta, c_\lambda$ need to be carefully balanced and τ selected sufficiently large such that $\alpha_k \leqslant \beta_k \leqslant \lambda_k \leqslant 1/4$, $c_\beta \leqslant 1/L$, $c_\beta \leqslant \frac{\mu_G}{320L^4}c_\lambda$, $c_\beta \leqslant \frac{\mu_h^3}{480(L+1)^6}$, $c_\beta \leqslant \frac{\mu_G}{8L^2}$, $\beta_k \leqslant \frac{\lambda_k}{4}\min\{(\frac{2L^2}{\mu_h^3}+58L^2)^{-1},(50L^2+\frac{8}{\mu_G})^{-1}\}$, $\alpha_k \leqslant \frac{\mu_G\beta_k}{4\left(14L^3+\frac{132(L+1)^6}{\mu_h^3}\right)}$, $\alpha_k \leqslant \frac{\mu_h^2}{3072(L+1)^6}\lambda_k$, $\alpha_k \leqslant \beta_k$, and $c_\alpha \geqslant \frac{2}{\min\{\mu_h,\mu_G\}}$.

The PL condition implies the following quadratic growth inequality, where $\text{Proj}_{\Theta^\star}(\theta)$ denotes the projection of θ to $\Theta^\star$ (Karimi et al., 2016).

$$\frac{2}{\mu_h}\left(h(\theta)-h^\star\right)\geqslant\|\text{Proj}_{\Theta^\star}(\theta)-\theta\|^2. \tag{44}$$

Equation (44) plays an important role in the proof of a lemma that we now introduce.

Lemma 7 *Under Assumption 1-3 and Assumption 5, the sequence of variables $\{\theta_k,\omega_k,f_k,g_k\}$ satisfy for all k*

$$\begin{aligned} \|f_k\| &\leqslant \|\Delta f_k\|+L\sqrt{y_k}+\frac{2L(L+1)}{\mu_h}\sqrt{x_k}, \\ \|g_k\| &\leqslant \|\Delta g_k\|+L\sqrt{y_k}, \\ \|\theta_{k+1}-\theta_k\| &\leqslant \alpha_k\left(\|\Delta f_k\|+L\sqrt{y_k}+\frac{2L(L+1)}{\mu_h}\sqrt{x_k}\right), \\ \|\omega_{k+1}-\omega_k\| &\leqslant \beta_k\left(\|\Delta g_k\|+L\sqrt{y_k}\right). \end{aligned}$$

Our analysis in this section also relies on a few more lemmas which we present below. The proofs of all lemmas can be found at the end of the section.

Lemma 8 *Under Assumption 1-3 and Assumption 5 and the step sizes given in (43), we have the following bound on* $\|\Delta f_{k+1}\|^2$

$$\mathbb{E}[\|\Delta f_{k+1}\|^2] \leqslant (1-\lambda_k)\mathbb{E}[\|\Delta f_k\|^2] - (\frac{\lambda_k}{4} - \frac{24L^2\beta_k^2}{\lambda_k})\mathbb{E}[\|\Delta f_k\|^2] + \frac{4}{\lambda_k}\mathbb{E}\left[6L^2\beta_k^2\|\Delta g_k\|^2 + 10L^4\beta_k^2 y_k + \frac{24(L+1)^6}{\mu_h}\alpha_k^2 x_k\right] + 2B\lambda_k^2.$$

Lemma 9 *Under Assumption 1-3 and Assumption 5 and the step sizes given in (43), we have the following bound on* $\|\Delta g_{k+1}\|^2$

$$\mathbb{E}[\|\Delta g_{k+1}\|^2] \leqslant (1-\lambda_k)\mathbb{E}[\|\Delta g_k\|^2] - (\frac{\lambda_k}{4} - \frac{24L^2\beta_k^2}{\lambda_k})\mathbb{E}[\|\Delta g_k\|^2] + \frac{4}{\lambda_k}\mathbb{E}\left[6L^2\beta_k^2\|\Delta f_k\|^2 + 10L^4\beta_k^2 y_k + \frac{24(L+1)^6}{\mu_h}\alpha_k^2 x_k\right] + 2B\lambda_k^2.$$

Lemma 10 *Under Assumption 1-3 and Assumption 5 and the step sizes given in (43), we have*

$$x_{k+1} \leqslant (1-\mu_h\alpha_k)x_k - (\frac{\mu_h}{4}\alpha_k - \frac{6(L+1)^5}{\mu_h^2}\alpha_k^2)x_k + (\frac{4L^4}{\mu_h^3} + 2L)\alpha_k y_k + (\frac{2L^2\alpha_k}{\mu_h^3} + 2L\alpha_k^2)\|\Delta f_k\|^2.$$

Lemma 11 *Under Assumption 1-3 and Assumption 5 and the step sizes given in (43), we have*

$$y_{k+1} \leqslant (1-\mu_G\beta_k)y_k - \left(\frac{\mu_G}{2}\beta_k - \left(12L^3 + \frac{128(L+1)^6}{\mu_h^3}\right)\alpha_k - L^2\beta_k^2\right)y_k + \left(\frac{\mu_h}{8}\alpha_k + \frac{24(L+1)^6}{\mu_h^2}\alpha_k\beta_k\right)x_k + \left(\frac{8\beta_k}{\mu_G} + 2L\beta_k^2\right)\|\Delta g_k\|^2 + 8L\alpha_k\|\Delta f_k\|^2.$$

C.1. Proof of Theorem 3

Combining the results from Lemma 8-11 and simplifying the terms with the choice of step sizes given in (43),

$$\begin{aligned}
&\mathbb{E}[V_{k+1}] \\
&= \mathbb{E}[\|\Delta f_{k+1}\|^2 + \|\Delta g_{k+1}\|^2 + x_{k+1} + y_{k+1}] \\
&\leqslant (1-\lambda_k)\mathbb{E}[\|\Delta f_k\|^2] - (\frac{\lambda_k}{4} - \frac{24L^2\beta_k^2}{\lambda_k})\mathbb{E}[\|\Delta f_k\|^2] \\
&\quad + \frac{4}{\lambda_k}\mathbb{E}\left[6L^2\beta_k^2\|\Delta g_k\|^2 + 10L^4\beta_k^2 y_k + \frac{24(L+1)^6}{\mu_h}\alpha_k^2 x_k\right] + 2B\lambda_k^2 \\
&\quad + (1-\lambda_k)\mathbb{E}[\|\Delta g_k\|^2] - (\frac{\lambda_k}{4} - \frac{24L^2\beta_k^2}{\lambda_k})\mathbb{E}[\|\Delta g_k\|^2]
\end{aligned}$$

$$
\begin{aligned}
&+\frac{4}{\lambda_k}\mathbb{E}\left[6L^2\beta_k^2\|\Delta f_k\|^2+10L^4\beta_k^2 y_k+\frac{24(L+1)^6}{\mu_h}\alpha_k^2 x_k\right]+2B\lambda_k^2\\
&+\mathbb{E}\Big[(1-\mu_h\alpha_k)x_k-(\frac{\mu_h}{4}\alpha_k-\frac{6(L+1)^5}{\mu_h^2}\alpha_k^2)x_k+(\frac{4L^4}{\mu_h^3}+2L)\alpha_k y_k+(\frac{2L^2\alpha_k}{\mu_h^3}+2L\alpha_k^2)\|\Delta f_k\|^2\\
&+(1-\mu_G\beta_k)y_k-\left(\frac{\mu_G}{2}\beta_k-\left(12L^3+\frac{128(L+1)^6}{\mu_h^3}\right)\alpha_k-L^2\beta_k^2\right)y_k\\
&+\left(\frac{\mu_h}{8}\alpha_k+\frac{24(L+1)^6}{\mu_h^2}\alpha_k\beta_k\right)x_k+\left(\frac{8\beta_k}{\mu_G}+2L\beta_k^2\right)\|\Delta g_k\|^2+8L\alpha_k\|\Delta f_k\|^2\Big]\\
\leqslant\;&(1-\lambda_k)\mathbb{E}[\|\Delta f_k\|^2]-(\frac{\lambda_k}{4}-\frac{48L^2\beta_k^2}{\lambda_k}-(\frac{2L^2}{\mu_h^3}+10L)\alpha_k)\mathbb{E}[\|\Delta f_k\|^2]\\
&+(1-\lambda_k)\mathbb{E}[\|\Delta g_k\|^2]-(\frac{\lambda_k}{4}-(50L^2+\frac{8}{\mu_G})\beta_k)\mathbb{E}[\|\Delta g_k\|^2]\\
&+\frac{8}{\lambda_k}\mathbb{E}\left[10L^4\beta_k^2 y_k+\frac{24(L+1)^6}{\mu_h}\alpha_k^2 x_k\right]+4B\lambda_k^2\\
&+(1-\mu_h\alpha_k)\mathbb{E}[x_k]-(\frac{\mu_h}{4}\alpha_k-\frac{6(L+1)^5}{\mu_h^2}\alpha_k^2)\mathbb{E}[x_k]+(\frac{4L^4}{\mu_h^3}+2L)\alpha_k\mathbb{E}[y_k]\\
&+(1-\mu_G\beta_k)\mathbb{E}[y_k]-\left(\frac{\mu_G}{2}\beta_k-\left(12L^3+\frac{128(L+1)^6}{\mu_h^3}\right)\alpha_k-L^2\beta_k^2\right)\mathbb{E}[y_k]\\
&+\left(\frac{\mu_h}{8}\alpha_k+\frac{24(L+1)^6}{\mu_h^2}\alpha_k\beta_k\right)\mathbb{E}[x_k]\\
\leqslant\;&(1-\lambda_k)\mathbb{E}[\|\Delta f_k\|^2]+(1-\lambda_k)\mathbb{E}[\|\Delta g_k\|^2]+(1-\mu_h\alpha_k)\mathbb{E}[x_k]+(1-\mu_G\beta_k)\mathbb{E}[y_k]+4B\lambda_k^2\\
&-(\frac{\lambda_k}{4}-(\frac{2L^2}{\mu_h^3}+58L^2)\beta_k)\mathbb{E}[\|\Delta f_k\|^2]-(\frac{\lambda_k}{4}-(50L^2+\frac{8}{\mu_G})\beta_k)\mathbb{E}[\|\Delta g_k\|^2]\\
&-(\frac{\mu_h}{16}\alpha_k-\frac{30(L+1)^6}{\mu_h^2}\alpha_k\beta_k)\mathbb{E}[x_k]-\left(\frac{\mu_G}{4}\beta_k-\left(14L^3+\frac{132(L+1)^6}{\mu_h^3}\right)\alpha_k-L^2\beta_k^2\right)\mathbb{E}[y_k],
\end{aligned}
$$

where we use the step size conditions $\beta_k \leqslant \frac{\mu_G}{320L^4}\lambda_k$ and $\alpha_k \leqslant \frac{\mu_h^2}{3072(L+1)^6}\lambda_k$.

Again, due to carefully selected step sizes, (specifically, $\beta_k \leqslant \frac{\mu_h^3}{480(L+1)^6}$, $\beta_k \leqslant \frac{\mu_G}{8L^2}$, $\beta_k \leqslant \frac{\lambda_k}{4}\min\{(\frac{2L^2}{\mu_h^3}+58L^2)^{-1},(50L^2+\frac{8}{\mu_G})^{-1}\}$, and $\alpha_k \leqslant \frac{\mu_G\beta_k}{4\left(14L^3+\frac{132(L+1)^6}{\mu_h^3}\right)}$), we can make the last four terms of the inequality above non-positive. As a result, we have

$$
\begin{aligned}
&\mathbb{E}[V_{k+1}]\\
&\leqslant(1-\lambda_k)\mathbb{E}[\|\Delta f_k\|^2]+(1-\lambda_k)\mathbb{E}[\|\Delta g_k\|^2]+(1-\mu_h\alpha_k)\mathbb{E}[x_k]+(1-\mu_G\beta_k)\mathbb{E}[y_k]+4B\lambda_k^2\\
&\leqslant(1-\min\{\mu_h,\mu_G\}\alpha_k)\mathbb{E}[V_k]+4B\lambda_k^2\\
&=(1-\frac{\min\{\mu_h,\mu_G\}c_\alpha}{k+\tau+1})\mathbb{E}[V_k]+\frac{B}{4(k+1)^2}\\
&\leqslant(1-\frac{2}{k+\tau+1})\mathbb{E}[V_k]+\frac{B}{4(k+1)^2}\\
&\leqslant\frac{k+\tau-1}{k+\tau+1}\mathbb{E}[V_k]+\frac{B}{4(k+1)^2}.
\end{aligned}
$$

Multiplying by $(k+\tau+1)^2$ on both sides of the inequality, we have

$$\begin{aligned}(k+\tau+1)^2\mathbb{E}[V_{k+1}] &\leqslant (k+\tau+1)(k+\tau-1)\mathbb{E}[V_k] + \frac{B(k+\tau+1)^2}{4(k+1)^2} \\ &\leqslant (k+\tau)^2\mathbb{E}[V_k] + \left(\frac{B}{2} + \frac{B\tau^2}{2}\right) \\ &\leqslant \tau^2 V_0 + \left(\frac{B}{2} + \frac{B\tau^2}{2}\right)(k+1).\end{aligned}$$

As a result,

$$\begin{aligned}\mathbb{E}[V_{k+1}] &\leqslant \frac{B\tau^2(k+1)}{(k+\tau+1)^2} + \frac{\tau^2 V_0}{(k+\tau+1)^2} \\ &\leqslant \frac{B\tau^2}{k+\tau+1} + \frac{\tau^2 V_0}{(k+\tau+1)^2},\end{aligned}$$

which obviously leads to the claimed result.

■

C.2. Proof of Lemma 7

By definition,

$$F(\theta, \omega^\star(\theta)) = \nabla h(\theta) = 0, \quad \forall \theta \in \Theta^\star$$

which implies

$$\begin{aligned}&\|f_k\| \\ &= \|\Delta f_k + F(\theta_k, \omega_k) - F(\theta_k, \omega^\star(\theta_k)) + F(\theta_k, \omega^\star(\theta_k)) - F(\text{Proj}_{\Theta^\star}(\theta_k), \omega^\star(\text{Proj}_{\Theta^\star}(\theta_k)))\| \\ &\leqslant \|\Delta f_k\| + \|F(\theta_k, \omega_k) - F(\theta_k, \omega^\star(\theta_k))\| + \|F(\theta_k, \omega^\star(\theta_k)) - F(\text{Proj}_{\Theta^\star}(\theta_k), \omega^\star(\text{Proj}_{\Theta^\star}(\theta_k)))\| \\ &\leqslant \|\Delta f_k\| + L\|\omega_k - \omega^\star(\theta_k)\| + L(L+1)\|\theta_k - \text{Proj}_{\Theta^\star}(\theta_k)\| \\ &\leqslant \|\Delta f_k\| + L\sqrt{y_k} + \frac{2L(L+1)}{\mu_h}\sqrt{x_k},\end{aligned}$$

where the second inequality follows from the Lipschitz continuity of F and $\omega^\star$ and the final inequality applies (44).

Similarly, since $G(\theta, \omega^\star(\theta)) = 0$ for any θ, we can derive

$$\begin{aligned}\|g_k\| &= \|\Delta g_k + G(\theta_k, \omega_k) - G(\theta_k, \omega^\star(\theta))\| \\ &\leqslant \|\Delta g_k\| + \|G(\theta_k, \omega_k) - G(\theta_k, \omega^\star(\theta))\| \\ &\leqslant \|\Delta g_k\| + L\sqrt{y_k}.\end{aligned}$$

The bounds on $\|\theta_{k+1} - \theta_k\|$ and $\|\omega_{k+1} - \omega_k\|$ easily follow from the two inequalities above and (9).

■

C.3. Proof of Lemma 8

By the update rule in (9), we have

$$\begin{aligned}\Delta f_{k+1} &= f_{k+1} - F(\theta_{k+1}, \omega_{k+1}) \\ &= (1-\lambda_k) f_k + \lambda_k F(\theta_k, \omega_k, X_k) - F(\theta_{k+1}, \omega_{k+1}) \\ &= (1-\lambda_k) f_k + \lambda_k F(\theta_k, \omega_k) - F(\theta_{k+1}, \omega_{k+1}) + \lambda_k \left(F(\theta_k, \omega_k) - F(\theta_k, \omega_k, X_k)\right) \\ &= (1-\lambda_k)\Delta f_k + F(\theta_k, \omega_k) - F(\theta_{k+1}, \omega_{k+1}) + \lambda_k \left(F(\theta_k, \omega_k) - F(\theta_k, \omega_k, X_k)\right).\end{aligned}$$

This leads to

$$\begin{aligned}&\|\Delta f_{k+1}\|^2 \\ &= (1-\lambda_k)^2 \|\Delta f_k\|^2 + \left\| \left(F(\theta_k, \omega_k) - F(\theta_{k+1}, \omega_{k+1})\right) + \lambda_k \left(F(\theta_k, \omega_k) - F(\theta_k, \omega_k, X_k)\right) \right\|^2 \\ &\quad + 2(1-\lambda_k)\Delta f_k^\top \left(F(\theta_k, \omega_k) - F(\theta_{k+1}, \omega_{k+1})\right) \\ &\quad + 2(1-\lambda_k)\lambda_k \Delta f_k^\top \left(F(\theta_k, \omega_k) - F(\theta_k, \omega_k, X_k)\right) \\ &\leqslant (1-\lambda_k)^2 \|\Delta f_k\|^2 + 2\|F(\theta_k, \omega_k) - F(\theta_{k+1}, \omega_{k+1})\|^2 + 2\lambda_k^2 \|F(\theta_k, \omega_k) - F(\theta_k, \omega_k, X_k)\|^2 \\ &\quad + \frac{\lambda_k}{2}\|\Delta f_k\|^2 + \frac{2}{\lambda_k}\|F(\theta_k, \omega_k) - F(\theta_{k+1}, \omega_{k+1})\|^2 \\ &\quad + 2(1-\lambda_k)\lambda_k \Delta f_k^\top \left(F(\theta_k, \omega_k) - F(\theta_k, \omega_k, X_k)\right) \\ &\leqslant (1-\lambda_k)\|\Delta f_k\|^2 + 2\lambda_k^2 \|F(\theta_k, \omega_k) - F(\theta_k, \omega_k, X_k)\|^2 \\ &\quad + (\lambda_k^2 - \frac{\lambda_k}{2})\|\Delta f_k\|^2 + \frac{4}{\lambda_k}\|F(\theta_k, \omega_k) - F(\theta_{k+1}, \omega_{k+1})\|^2 \\ &\quad + 2(1-\lambda_k)\lambda_k \Delta f_k^\top \left(F(\theta_k, \omega_k) - F(\theta_k, \omega_k, X_k)\right), \end{aligned} \tag{45}$$

where the last inequality follows from $\lambda_k \leqslant 1$.

The second term on the right hand side of (45) is bounded in expectation, i.e.

$$\mathbb{E}_{X_k \sim \xi}[\|F(\theta_k, \omega_k, X_k) - F(\theta_k, \omega_k)\|^2] \leqslant B.$$

In addition, the last term of (45) is zero in expectation since $\mathbb{E}[\Delta f_k^\top (F(\theta_k, \omega_k) - F(\theta_k, \omega_k, X_k)) \mid \mathcal{H}_{k-1}] = \Delta f_k^\top \mathbb{E}[(F(\theta_k, \omega_k) - F(\theta_k, \omega_k, X_k)) \mid \mathcal{H}_{k-1}] = 0$. As a result, we have

$$\begin{aligned}\mathbb{E}[\|\Delta f_{k+1}\|^2] \leqslant (1-\lambda_k)\mathbb{E}[\|\Delta f_k\|^2] - (\frac{\lambda_k}{2} - \lambda_k^2)\mathbb{E}[\|\Delta f_k\|^2] \\ + \frac{4}{\lambda_k}\mathbb{E}[\|F(\theta_k, \omega_k) - F(\theta_{k+1}, \omega_{k+1})\|^2] + 2B\lambda_k^2.\end{aligned} \tag{46}$$

By Lemma 7 and the Lipschitz condition of F,

$$\begin{aligned}&\|F(\theta_k, \omega_k) - F(\theta_{k+1}, \omega_{k+1})\|^2 \\ &\leqslant 2L^2\|\theta_{k+1} - \theta_k\|^2 + 2L^2\|\omega_{k+1} - \omega_k\|^2 \\ &\leqslant 6L^2\alpha_k^2(\|\Delta f_k\|^2 + L^2 y_k + \frac{4(L+1)^4}{\mu_h^2}x_k) + 4L^2\beta_k^2(\|\Delta g_k\|^2 + L^2 y_k) \\ &\leqslant 6L^2\beta_k^2\|\Delta f_k\|^2 + 6L^2\beta_k^2\|\Delta g_k\|^2 + 10L^4\beta_k^2 y_k + \frac{24(L+1)^6}{\mu_h}\alpha_k^2 x_k.\end{aligned}$$

Plugging this bound into (46), we get

$$\begin{aligned}\mathbb{E}[\|\Delta f_{k+1}\|^2] &\leqslant (1-\lambda_k)\mathbb{E}[\|\Delta f_k\|^2] - (\frac{\lambda_k}{2} - \lambda_k^2)\mathbb{E}[\|\Delta f_k\|^2] + 2B\lambda_k^2 \\ &\quad + \frac{4}{\lambda_k}\mathbb{E}\left[6L^2\beta_k^2\|\Delta f_k\|^2 + 6L^2\beta_k^2\|\Delta g_k\|^2 + 10L^4\beta_k^2 y_k + \frac{24(L+1)^6}{\mu_h}\alpha_k^2 x_k\right] \\ &\leqslant (1-\lambda_k)\mathbb{E}[\|\Delta f_k\|^2] - (\frac{\lambda_k}{4} - \frac{24L^2\beta_k^2}{\lambda_k})\mathbb{E}[\|\Delta f_k\|^2] \\ &\quad + \frac{4}{\lambda_k}\mathbb{E}\left[6L^2\beta_k^2\|\Delta g_k\|^2 + 10L^4\beta_k^2 y_k + \frac{24(L+1)^6}{\mu_h}\alpha_k^2 x_k\right] + 2B\lambda_k^2,\end{aligned}$$

where the last inequality follows from $\lambda_k \leqslant 1/4$.

■

C.4. Proof of Lemma 9

By the update rule in (9), we have

$$\begin{aligned}\Delta g_{k+1} &= g_{k+1} - G(\theta_{k+1}, \omega_{k+1}) \\ &= (1-\lambda_k)g_k + \lambda_k G(\theta_k, \omega_k, X_k) - G(\theta_{k+1}, \omega_{k+1}) \\ &= (1-\lambda_k)g_k + \lambda_k G(\theta_k, \omega_k) - G(\theta_{k+1}, \omega_{k+1}) + \lambda_k\left(G(\theta_k, \omega_k) - G(\theta_k, \omega_k, X_k)\right) \\ &= (1-\lambda_k)\Delta g_k + G(\theta_k, \omega_k) - G(\theta_{k+1}, \omega_{k+1}) + \lambda_k\left(G(\theta_k, \omega_k) - G(\theta_k, \omega_k, X_k)\right).\end{aligned}$$

This leads to

$$\begin{aligned}&\|\Delta g_{k+1}\|^2 \\ &= (1-\lambda_k)^2\|\Delta g_k\|^2 + \left\|\left(G(\theta_k, \omega_k) - G(\theta_{k+1}, \omega_{k+1})\right) + \lambda_k\left(G(\theta_k, \omega_k) - G(\theta_k, \omega_k, X_k)\right)\right\|^2 \\ &\quad + 2(1-\lambda_k)\Delta g_k^\top\left(G(\theta_k, \omega_k) - G(\theta_{k+1}, \omega_{k+1})\right) \\ &\quad + 2(1-\lambda_k)\lambda_k\Delta g_k^\top\left(G(\theta_k, \omega_k) - G(\theta_k, \omega_k, X_k)\right) \\ &\leqslant (1-\lambda_k)^2\|\Delta g_k\|^2 + 2\|G(\theta_k, \omega_k) - G(\theta_{k+1}, \omega_{k+1})\|^2 + 2\lambda_k^2\|G(\theta_k, \omega_k) - G(\theta_k, \omega_k, X_k)\|^2 \\ &\quad + \frac{\lambda_k}{2}\|\Delta g_k\|^2 + \frac{2}{\lambda_k}\|G(\theta_k, \omega_k) - G(\theta_{k+1}, \omega_{k+1})\|^2 \\ &\quad + 2(1-\lambda_k)\lambda_k\Delta g_k^\top\left(G(\theta_k, \omega_k) - G(\theta_k, \omega_k, X_k)\right) \\ &\leqslant (1-\lambda_k)\|\Delta g_k\|^2 + 2\lambda_k^2\|G(\theta_k, \omega_k) - G(\theta_k, \omega_k, X_k)\|^2 \\ &\quad + (\lambda_k^2 - \frac{\lambda_k}{2})\|\Delta g_k\|^2 + \frac{4}{\lambda_k}\|G(\theta_k, \omega_k) - G(\theta_{k+1}, \omega_{k+1})\|^2 \\ &\quad + 2(1-\lambda_k)\lambda_k\Delta g_k^\top\left(G(\theta_k, \omega_k) - G(\theta_k, \omega_k, X_k)\right). \qquad (47)\end{aligned}$$

The second term on the right hand side of (47) is bounded in expectation, i.e.

$$\mathbb{E}_{X_k\sim\xi}[\|G(\theta_k, \omega_k, X_k) - G(\theta_k, \omega_k)\|^2] \leqslant B.$$

In addition, the last term of (45) is zero in expectation since $\mathbb{E}[\Delta g_k^\top(G(\theta_k, \omega_k) - G(\theta_k, \omega_k, X_k)) \mid \mathcal{H}_{k-1}] = \Delta g_k^\top\mathbb{E}[(G(\theta_k, \omega_k) - G(\theta_k, \omega_k, X_k)) \mid \mathcal{H}_{k-1}] = 0$. As a result, we have

$$\mathbb{E}[\|\Delta g_{k+1}\|^2] \leqslant (1-\lambda_k)\mathbb{E}[\|\Delta g_k\|^2] - (\frac{\lambda_k}{2} - \lambda_k^2)\mathbb{E}[\|\Delta g_k\|^2]$$

$$+\frac{4}{\lambda_k}\mathbb{E}[\|G(\theta_k,\omega_k)-G(\theta_{k+1},\omega_{k+1})\|^2]+2B\lambda_k^2. \tag{48}$$

By Lemma 7 and the Lipschitz condition of G,

$$\begin{aligned}
&\|G(\theta_k,\omega_k)-G(\theta_{k+1},\omega_{k+1})\|^2\\
&\leqslant 2L^2\|\theta_{k+1}-\theta_k\|^2+2L^2\|\omega_{k+1}-\omega_k\|^2\\
&\leqslant 6L^2\alpha_k^2(\|\Delta f_k\|^2+L^2y_k+\frac{4(L+1)^4}{\mu_h^2}x_k)+4L^2\beta_k^2(\|\Delta g_k\|^2+L^2y_k)\\
&\leqslant 6L^2\beta_k^2\|\Delta f_k\|^2+6L^2\beta_k^2\|\Delta g_k\|^2+10L^4\beta_k^2y_k+\frac{24(L+1)^6}{\mu_h}\alpha_k^2x_k.
\end{aligned}$$

Plugging this bound into (48), we get

$$\begin{aligned}
&\mathbb{E}[\|\Delta g_{k+1}\|^2]\\
&\leqslant(1-\lambda_k)\mathbb{E}[\|\Delta g_k\|^2]-(\frac{\lambda_k}{2}-\lambda_k^2)\mathbb{E}[\|\Delta g_k\|^2]\\
&\quad+\frac{4}{\lambda_k}\mathbb{E}\left[6L^2\beta_k^2\|\Delta f_k\|^2+6L^2\beta_k^2\|\Delta g_k\|^2+10L^4\beta_k^2y_k+\frac{24(L+1)^6}{\mu_h}\alpha_k^2x_k\right]+2B\lambda_k^2\\
&\leqslant(1-\lambda_k)\mathbb{E}[\|\Delta g_k\|^2]-(\frac{\lambda_k}{4}-\frac{24L^2\beta_k^2}{\lambda_k})\mathbb{E}[\|\Delta g_k\|^2]\\
&\quad+\frac{4}{\lambda_k}\mathbb{E}\left[6L^2\beta_k^2\|\Delta f_k\|^2+10L^4\beta_k^2y_k+\frac{24(L+1)^6}{\mu_h}\alpha_k^2x_k\right]+2B\lambda_k^2,
\end{aligned}$$

where the last inequality follows from $\lambda_k\leqslant 1/4$. ■

C.5. Proof of Lemma 10

From the update rule (8) and the Lipschitz gradient condition in Assumption 1, we have

$$\begin{aligned}
h(\theta_{k+1})&\leqslant h(\theta_k)+\langle\nabla h(\theta_k),\theta_{k+1}-\theta_k\rangle+\frac{L}{2}\|\theta_{k+1}-\theta_k\|^2\\
&=h(\theta_k)-\alpha_k\langle\nabla h(\theta_k),f_k\rangle+\frac{L\alpha_k^2}{2}\|f_k\|^2\\
&=h(\theta_k)-\alpha_k\langle\nabla h(\theta_k),\Delta f_k\rangle-\alpha_k\langle\nabla h(\theta_k),F(\theta_k,\omega_k)\rangle+\frac{L\alpha_k^2}{2}\|f_k\|^2\\
&\leqslant h(\theta_k)-\alpha_k\langle\nabla h(\theta_k),\Delta f_k\rangle-\alpha_k\langle\nabla h(\theta_k),F(\theta_k,\omega^\star(\theta_k))\rangle\\
&\quad+\alpha_k\langle\nabla h(\theta_k),F(\theta_k,\omega^\star(\theta_k))-F(\theta_k,\omega_k)\rangle\\
&\quad+\frac{3L\alpha_k^2}{2}\left(\|\Delta f_k\|^2+Ly_k+\frac{4(L+1)^4}{\mu_h^2}x_k\right)\\
&=h(\theta_k)-\alpha_k\|\nabla h(\theta_k)\|^2-\alpha_k\langle\nabla h(\theta_k),\Delta f_k\rangle\\
&\quad+\alpha_k\langle\nabla h(\theta_k),F(\theta_k,\omega^\star(\theta_k))-F(\theta_k,\omega_k)\rangle\\
&\quad+\frac{3L\alpha_k^2}{2}\left(\|\Delta f_k\|^2+Ly_k+\frac{4(L+1)^4}{\mu_h^2}x_k\right),
\end{aligned}$$

where the second inequality applies Lemma 7.

As a result of Assumption 5, the inequality implies

$$\begin{aligned}
x_{x+1} &= h(\theta_{k+1}) - h(\theta^\star) \\
&\leqslant h(\theta_k) - h(\theta^\star) - \alpha_k \|\nabla h(\theta_k)\|^2 - \alpha_k \langle \nabla h(\theta_k), \Delta f_k \rangle \\
&\quad + \alpha_k \langle \nabla h(\theta_k), F(\theta_k, \omega^\star(\theta_k)) - F(\theta_k, \omega_k) \rangle + \frac{3L\alpha_k^2}{2} \left(\|\Delta f_k\|^2 + L y_k + \frac{4(L+1)^4}{\mu_h^2} x_k \right) \\
&\leqslant h(\theta_k) - h(\theta^\star) - 2\mu_h \alpha_k \left(h(\theta_k) - h(\theta^\star) \right) - \alpha_k \langle \nabla h(\theta_k), \Delta f_k \rangle \\
&\quad + \alpha_k \langle \nabla h(\theta_k), F(\theta_k, \omega^\star(\theta_k)) - F(\theta_k, \omega_k) \rangle + \frac{3L\alpha_k^2}{2} \left(\|\Delta f_k\|^2 + L y_k + \frac{4(L+1)^4}{\mu_h^2} x_k \right) \\
&\leqslant (1 - 2\mu_h \alpha_k) x_k - \alpha_k \langle \nabla h(\theta_k), \Delta f_k \rangle \\
&\quad + \alpha_k \langle \nabla h(\theta_k), F(\theta_k, \omega^\star(\theta_k)) - F(\theta_k, \omega_k) \rangle + \frac{3L\alpha_k^2}{2} \left(\|\Delta f_k\|^2 + L y_k + \frac{4(L+1)^4}{\mu_h^2} x_k \right).
\end{aligned} \tag{49}$$

To bound the second term on the right hand side of (49),

$$\begin{aligned}
-\alpha_k \langle \nabla h(\theta_k), \Delta f_k \rangle &\leqslant \alpha_k \|\nabla h(\theta_k)\| \|\Delta f_k\| \\
&= \alpha_k \|\nabla h(\theta_k) - \nabla h(\text{Proj}_{\Theta^\star}(\theta_k))\| \|\Delta f_k\| \\
&\leqslant L\alpha_k \|\theta_k - \text{Proj}_{\Theta^\star}(\theta_k)\| \|\Delta f_k\| \\
&\leqslant \frac{2L\alpha_k}{\mu_h} \sqrt{x_k} \|\Delta f_k\| \\
&\leqslant \frac{\mu_h}{2} \alpha_k x_k + \frac{2L^2 \alpha_k}{\mu_h^3} \|\Delta f_k\|^2,
\end{aligned} \tag{50}$$

where the third inequality plugs in the quadratic growth inequality (44) and the final inequality follows from the fact that $2\|\boldsymbol{a}\|\|\boldsymbol{b}\| \leqslant c\|\boldsymbol{a}\|^2 + \frac{1}{c}\|\boldsymbol{b}\|^2$ for any vector $\boldsymbol{a}, \boldsymbol{b}$ and scalar $c > 0$.

The third term on the right hand side of (49) can be treated in a similar manner

$$\begin{aligned}
\alpha_k \langle \nabla h(\theta_k), F(\theta_k, \omega^\star(\theta_k)) - F(\theta_k, \omega_k) \rangle &\leqslant \alpha_k \|\nabla h(\theta_k)\| \|F(\theta_k, \omega^\star(\theta_k)) - F(\theta_k, \omega_k)\| \\
&\leqslant \frac{2L\alpha_k}{\mu_h} \sqrt{x_k} \cdot (L\|\omega^\star(\theta_k) - \omega_k\|) \\
&= \frac{2L^2 \alpha_k}{\mu_h} \sqrt{x_k} \sqrt{y_k} \\
&\leqslant \frac{\mu_h}{4} \alpha_k x_k + \frac{4L^4 \alpha_k}{\mu_h^3} y_k.
\end{aligned} \tag{51}$$

Plugging (50) and (51) into (49), we get

$$\begin{aligned}
&x_{k+1} \\
&\leqslant (1 - 2\mu_h \alpha_k) x_k - \alpha_k \langle \nabla h(\theta_k), \Delta f_k \rangle \\
&\quad + \alpha_k \langle \nabla h(\theta_k), F(\theta_k, \omega^\star(\theta_k)) - F(\theta_k, \omega_k) \rangle + \frac{3L\alpha_k^2}{2} \left(\|\Delta f_k\|^2 + L y_k + \frac{4(L+1)^4}{\mu_h^2} x_k \right)
\end{aligned}$$

$$\leqslant (1-2\mu_h\alpha_k)x_k + \frac{\mu_h}{2}\alpha_k x_k + \frac{2L^2\alpha_k}{\mu_h^3}\|\Delta f_k\|^2$$
$$+ \frac{\mu_h}{4}\alpha_k x_k + \frac{4L^4\alpha_k}{\mu_h^3}y_k + \frac{3L\alpha_k^2}{2}\left(\|\Delta f_k\|^2 + Ly_k + \frac{4(L+1)^4}{\mu_h^2}x_k\right)$$
$$\leqslant (1-\mu_h\alpha_k)x_k - (\frac{\mu_h}{4}\alpha_k - \frac{6(L+1)^5}{\mu_h^2}\alpha_k^2)x_k + (\frac{4L^4}{\mu_h^3} + 2L)\alpha_k y_k + (\frac{2L^2\alpha_k}{\mu_h^3} + 2L\alpha_k^2)\|\Delta f_k\|^2.$$

■

C.6. Proof of Lemma 11

By the update rule in (8),

$$\begin{aligned}\omega_{k+1} - \omega^\star(\theta_{k+1}) &= \omega_{k+1} - \omega^\star(\theta_{k+1}) \\ &= (\omega_k - \omega^\star(\theta_k)) - \beta_k g_k - (\omega^\star(\theta_{k+1}) - \omega^\star(\theta_k)) \\ &= (\omega_k - \omega^\star(\theta_k)) - \beta_k G(\theta_k,\omega_k) - \beta_k \Delta g_k - (\omega^\star(\theta_{k+1}) - \omega^\star(\theta_k))\,.\end{aligned}$$

Taking the norm yields

$$\begin{aligned}y_{k+1} &= \|\omega_{k+1} - \omega^\star(\theta_{k+1})\|^2 \\ &= \underbrace{\|\omega_k - \omega^\star(\theta_k) - \beta_k G(\theta_k,\omega_k)\|^2}_{T_1} + \beta_k^2\|\Delta g_k\|^2 + \underbrace{\|\omega^\star(\theta_{k+1}) - \omega^\star(\theta_k)\|^2}_{T_2} \\ &\quad \underbrace{-2\left(\omega_k - \omega^\star(\theta_k) - \beta_k G(\theta_k,\omega_k)\right)^\top \left(\omega^\star(\theta_{k+1}) - \omega^\star(\theta_k)\right)}_{T_3} \\ &\quad \underbrace{-2\beta_k\left(\omega_k - \omega^\star(\theta_k) - \beta_k G(\theta_k,\omega_k)\right)^\top \Delta g_k}_{T_4} + \underbrace{2\beta_k \Delta g_k^\top \left(\omega^\star(\theta_{k+1}) - \omega^\star(\theta_k)\right)}_{T_5}. \qquad (52)\end{aligned}$$

We bound each term of (52) individually. First, since by definition $G(\theta,\omega^\star(\theta)) = 0$ for all θ, we have

$$\begin{aligned}T_1 &= \|\omega_k - \omega^\star(\theta_k)\|^2 - 2\beta_k\left(\omega_k - \omega^\star(\theta_k)\right)^\top G(\theta_k,\omega_k) + \beta_k^2\|G(\theta_k,\omega_k)\|^2 \\ &= y_k - 2\beta_k\left(\omega_k - \omega^\star(\theta_k)\right)^\top G(\theta_k,\omega_k) + \beta_k^2\|G(\theta_k,\omega_k) - G(\theta_k,\omega^\star(\theta_k))\|^2 \\ &\leqslant y_k - 2\mu_G\beta_k\|\omega_k - \omega^\star(\theta_k)\|^2 + L^2\beta_k^2\|\omega_k - \omega^\star(\theta_k)\|^2 \\ &= (1 - 2\mu_G\beta_k + L^2\beta_k^2)y_k,\end{aligned}$$

where the first inequality follows from Assumption 3 and the Lipschitz continuity of G.

The term T_2 can be simply treated using the Lipschitz condition of $\omega^\star$

$$T_2 = \|\omega^\star(\theta_{k+1}) - \omega^\star(\theta_k)\|^2 \leqslant L^2\|\theta_{k+1} - \theta_k\|^2 \leqslant L^2\alpha_k^2\left(\|\Delta f_k\| + L\sqrt{y_k} + \frac{2L(L+1)}{\mu_h}\sqrt{x_k}\right)^2$$
$$\leqslant 3L^2\alpha_k^2\left(\|\Delta f_k\|^2 + L^2 y_k + \frac{4(L+1)^4}{\mu_h^2}x_k\right).$$

By the Cauchy-Schwarz inequality,

T_3

$$
\begin{aligned}
&\leqslant 2\|\omega_k - \omega^\star(\theta_k) - \beta_k G(\theta_k, \omega_k)\|\|\omega^\star(\theta_{k+1}) - \omega^\star(\theta_k)\| \\
&\leqslant 2\|\omega_k - \omega^\star(\theta_k)\|\|\omega^\star(\theta_{k+1}) - \omega^\star(\theta_k)\| + 2\beta_k\|G(\theta_k, \omega_k) - G(\theta_k, \omega^\star(\theta_k))\|\|\omega^\star(\theta_{k+1}) - \omega^\star(\theta_k)\| \\
&\leqslant (2L + 2L^2\beta_k)\sqrt{y_k}\|\theta_{k+1} - \theta_k\| \\
&\leqslant 4L\alpha_k\sqrt{y_k}\left(\|\Delta f_k\| + L\sqrt{y_k} + \frac{2L(L+1)}{\mu_h}\sqrt{x_k}\right).
\end{aligned}
$$

where the second inequality follows from $G(\theta, \omega^\star(\theta)) = 0$ for all θ, and the fourth inequality applies Lemma 7 and the step size rule $\beta_k \leqslant 1/L$.

We can bound T_4 following a similar line of analysis

$$
T_4 \leqslant 2\beta_k\|\omega_k - \omega^\star(\theta_k) - \beta_k G(\theta_k, \omega_k)\|\|\Delta g_k\| \leqslant 4\beta_k\sqrt{y_k}\|\Delta g_k\|.
$$

Finally, we treat T_5 again by the Cauchy-Schwarz inequality

$$
\begin{aligned}
T_5 &\leqslant 2\beta_k\|\Delta g_k\|\|\omega^\star(\theta_{k+1}) - \omega^\star(\theta_k)\| \\
&\leqslant 2L\alpha_k\beta_k\|\Delta g_k\|\left(\|\Delta f_k\| + L\sqrt{y_k} + \frac{2L(L+1)}{\mu_h}\sqrt{x_k}\right),
\end{aligned}
$$

where the second inequality again uses Lemma 7.

Applying the bounds on T_1-T_5 to (52), we get

$$
\begin{aligned}
y_{k+1} &\leqslant (1 - 2\mu_G\beta_k + L^2\beta_k^2)y_k + \beta_k^2\|\Delta g_k\|^2 + 3L^2\alpha_k^2\left(\|\Delta f_k\|^2 + L^2 y_k + \frac{4(L+1)^4}{\mu_h^2}x_k\right) \\
&\quad + 4L\alpha_k\sqrt{y_k}\left(\|\Delta f_k\| + L\sqrt{y_k} + \frac{2L(L+1)}{\mu_h}\sqrt{x_k}\right) + 4\beta_k\sqrt{y_k}\|\Delta g_k\| \\
&\quad + 2L\alpha_k\beta_k\|\Delta g_k\|\left(\|\Delta f_k\| + L\sqrt{y_k} + \frac{2L(L+1)}{\mu_h}\sqrt{x_k}\right) \\
&\leqslant (1 - 2\mu_G\beta_k + L^2\beta_k^2)y_k + \beta_k^2\|\Delta g_k\|^2 + 3L^2\alpha_k^2\left(\|\Delta f_k\|^2 + L^2 y_k + \frac{4(L+1)^4}{\mu_h^2}x_k\right) \\
&\quad + 2L\alpha_k y_k + 2L\alpha_k\|\Delta f_k\|^2 + 4L^2\alpha_k y_k + \frac{128(L+1)^6\alpha_k}{\mu_h^3}y_k + \frac{\mu_h\alpha_k}{8}x_k + \frac{\mu_G\beta_k}{2}y_k \\
&\quad + \frac{8\beta_k}{\mu_G}\|\Delta g_k\|^2 + L\alpha_k\beta_k\|\Delta g_k\|^2 + 3L\alpha_k\beta_k\left(\|\Delta f_k\|^2 + L^2 y_k + \frac{4(L+1)^4}{\mu_h^2}x_k\right) \\
&\leqslant (1 - \mu_G\beta_k)y_k - \left(\frac{\mu_G}{2}\beta_k - \left(12L^3 + \frac{128(L+1)^6}{\mu_h^3}\right)\alpha_k - L^2\beta_k^2\right)y_k \\
&\quad + \left(\frac{\mu_h}{8}\alpha_k + \frac{24(L+1)^6}{\mu_h^2}\alpha_k\beta_k\right)x_k + \left(\frac{8\beta_k}{\mu_G} + 2L\beta_k^2\right)\|\Delta g_k\|^2 + 8L\alpha_k\|\Delta f_k\|^2
\end{aligned}
$$

where the second inequality follows from the fact that $2\|\boldsymbol{a}\|\|\boldsymbol{b}\| \leqslant c\|\boldsymbol{a}\|^2 + \frac{1}{c}\|\boldsymbol{b}\|^2$ for any vector $\boldsymbol{a}, \boldsymbol{b}$ and scalar $c > 0$.

■

Appendix D. Analysis for Non-Convex Functions

Our analysis in this section relies on the following lemmas, the proofs of which can be found at the end of the section.

Lemma 12 *Under Assumption 1-3, the sequence of variables $\{\theta_k, \omega_k, f_k, g_k\}$ satisfy for all k*

$$\begin{aligned}\|f_k\| &\leqslant \|\Delta f_k\| + L\sqrt{y_k} + \|\nabla h(\theta_k)\|,\\ \|g_k\| &\leqslant \|\Delta g_k\| + L\sqrt{y_k},\\ \|\theta_{k+1}-\theta_k\| &\leqslant \alpha_k\left(\|\Delta f_k\| + L\sqrt{y_k} + \|\nabla h(\theta_k)\|\right),\\ \|\omega_{k+1}-\omega_k\| &\leqslant \beta_k\left(\|\Delta g_k\| + L\sqrt{y_k}\right).\end{aligned}$$

Lemma 13 *Under Assumption 1-3 and the step sizes given in (16), we have the following bound on $\|\Delta f_{k+1}\|^2$*

$$\begin{aligned}\mathbb{E}[\|\Delta f_{k+1}\|^2] \leqslant (1-\lambda_k)\mathbb{E}[\|\Delta f_k\|^2] - (\frac{\lambda_k}{4} - \frac{24L^2\beta_k^2}{\lambda_k})\mathbb{E}[\|\Delta f_k\|^2]\\ + \frac{4}{\lambda_k}\mathbb{E}\left[6L^2\beta_k^2\|\Delta g_k\|^2 + 10L^4\beta_k^2 y_k + 6L^2\alpha_k^2\|\nabla h(\theta_k)\|^2\right] + 2B\lambda_k^2.\end{aligned}$$

Lemma 14 *Under Assumption 1-3 and the step sizes given in (16), we have the following bound on $\|\Delta g_{k+1}\|^2$*

$$\begin{aligned}\mathbb{E}[\|\Delta g_{k+1}\|^2] \leqslant (1-\lambda_k)\mathbb{E}[\|\Delta g_k\|^2] - (\frac{\lambda_k}{4} - \frac{24L^2\beta_k^2}{\lambda_k})\mathbb{E}[\|\Delta g_k\|^2]\\ + \frac{4}{\lambda_k}\mathbb{E}\left[6L^2\beta_k^2\|\Delta f_k\|^2 + 10L^4\beta_k^2 y_k + 6L^2\alpha_k^2\|\nabla h(\theta_k)\|^2\right] + 2B\lambda_k^2.\end{aligned}$$

Lemma 15 *Under Assumption 1-3 and the step sizes in (16), we have*

$$h(\theta_{k+1}) \leqslant h(\theta_k) - \frac{\alpha_k}{4}\|\nabla h(\theta_k)\|^2 + \frac{5\alpha_k}{4}\|\Delta f_k\|^2 + \frac{5L^2\alpha_k}{4}y_k.$$

Lemma 16 *Under Assumption 1-3 and the step sizes in (16), we have*

$$\begin{aligned}y_{k+1} \leqslant (1-\mu_G\beta_k)y_k - \left(\frac{\mu_G}{2}\beta_k - 108L^4\alpha_k - L^2\beta_k^2\right)y_k\\ + \frac{\alpha_k}{8}\|\nabla h(\theta_k)\|^2 + \left(\frac{8\beta_k}{\mu_G} + 2L\beta_k^2\right)\|\Delta g_k\|^2 + 8L\alpha_k\|\Delta f_k\|^2.\end{aligned}$$

D.1. Proof of Theorem 4

Combining the results from Lemma 13-16 and simplifying the terms with the choice of step sizes given in (16), we have

$$\begin{aligned}&\mathbb{E}[\|\Delta f_{k+1}\|^2 + \|\Delta g_{k+1}\|^2 + h(\theta_{k+1}) + y_{k+1}]\\ &\leqslant (1-\lambda_k)\mathbb{E}[\|\Delta f_k\|^2] - (\frac{\lambda_k}{4} - \frac{24L^2\beta_k^2}{\lambda_k})\mathbb{E}[\|\Delta f_k\|^2]\\ &\quad + \frac{4}{\lambda_k}\mathbb{E}\left[6L^2\beta_k^2\|\Delta g_k\|^2 + 10L^4\beta_k^2 y_k + 6L^2\alpha_k^2\|\nabla h(\theta_k)\|^2\right] + 2B\lambda_k^2\\ &\quad + (1-\lambda_k)\mathbb{E}[\|\Delta g_k\|^2] - (\frac{\lambda_k}{4} - \frac{24L^2\beta_k^2}{\lambda_k})\mathbb{E}[\|\Delta g_k\|^2]\end{aligned}$$

$$
\begin{aligned}
&+ \frac{4}{\lambda_k}\mathbb{E}\left[6L^2\beta_k^2\|\Delta f_k\|^2 + 10L^4\beta_k^2 y_k + 6L^2\alpha_k^2\|\nabla h(\theta_k)\|^2\right] + 2B\lambda_k^2 \\
&+ \mathbb{E}\Big[h(\theta_k) - \frac{\alpha_k}{4}\|\nabla h(\theta_k)\|^2 + \frac{5\alpha_k}{4}\|\Delta f_k\|^2 + \frac{5L^2\alpha_k}{4}y_k \\
&+ (1-\mu_G\beta_k)y_k - \left(\frac{\mu_G}{2}\beta_k - 108L^4\alpha_k - L^2\beta_k^2\right)y_k \\
&+ \frac{\alpha_k}{8}\|\nabla h(\theta_k)\|^2 + \left(\frac{8\beta_k}{\mu_G} + 2L\beta_k^2\right)\|\Delta g_k\|^2 + 8L\alpha_k\|\Delta f_k\|^2\Big] \\
\leqslant\ & (1-\lambda_k)\mathbb{E}[\|\Delta f_k\|^2] - (\frac{\lambda_k}{4} - \frac{48L^2\beta_k^2}{\lambda_k} - \frac{37L}{4}\alpha_k)\mathbb{E}[\|\Delta f_k\|^2] \\
&+ (1-\lambda_k)\mathbb{E}[\|\Delta g_k\|^2] - (\frac{\lambda_k}{4} - (50L^2 + \frac{8}{\mu_G})\beta_k)\mathbb{E}[\|\Delta g_k\|^2] \\
&+ \mathbb{E}[h(\theta_k)] - (\frac{\alpha_k}{8} - \frac{48L^2\alpha_k^2}{\lambda_k})\mathbb{E}[\|\nabla h(\theta_k)\|^2] \\
&+ (1-\mu_G\beta_k)\mathbb{E}[y_k] - \left(\frac{\mu_G}{2}\beta_k - 110L^4\alpha_k - L^2\beta_k^2 - \frac{80L^4\beta_k^2}{\lambda_k}\right)\mathbb{E}[y_k] + 4B\lambda_k^2 \\
\leqslant\ & (1-\lambda_k)\mathbb{E}[\|\Delta f_k\|^2] + (1-\lambda_k)\mathbb{E}[\|\Delta g_k\|^2] + \mathbb{E}[h(\theta_k)] - \frac{\alpha_k}{16}\mathbb{E}[\|\nabla h(\theta_k)\|^2] \\
&+ (1-\mu_G\beta_k)\mathbb{E}[y_k] + 4B\lambda_k^2,
\end{aligned}
$$

where we use the step size conditions $\beta_k \leqslant \frac{\lambda_k}{4}\min\{(\frac{37L}{4} + 48L^2)^{-1}, (50L^2 + \frac{8}{\mu_G})^{-1}\}$, $\alpha_k \leqslant \frac{1}{768L^2}\lambda_k$, $\beta_k \leqslant \frac{\mu_G}{320L^4}\lambda_k$, $\beta_k \leqslant \frac{\mu_G}{8L^2}$, and $\alpha_k \leqslant \frac{\mu_G}{880L^4}\beta_k$.

This inequality obviously implies

$$
\begin{aligned}
\alpha_k\mathbb{E}[\|\nabla h(\theta_k)\|^2] \leqslant\ & 16\mathbb{E}[\|\Delta f_k\|^2 - \|\Delta f_{k+1}\|^2] + 16\mathbb{E}[\|\Delta g_k\|^2 - \|\Delta g_{k+1}\|^2] \\
&+ 16\mathbb{E}[h(\theta_k) - h(\theta_{k+1})] + 16\mathbb{E}[y_k - y_{k+1}] + 64B\lambda_k^2.
\end{aligned}
$$

Since $\|\Delta f_k\|^2$, $\|\Delta g_k\|^2$, and y_k are non-negative and $h(\theta_k) \geqslant h(\theta^\star)$, we have the following inequality by summing up over k

$$
\sum_{t=0}^{k}\alpha_t\mathbb{E}[\|\nabla h(\theta_t)\|^2] \leqslant 16\left(\|\Delta f_0\|^2 + \|\Delta g_0\|^2 + (h(\theta_0) - h(\theta^\star)) + y_0\right) + 64B\sum_{t=0}^{k}\lambda_t^2
$$

It is a standard result that (see, for example, Lemma 3 of Zeng et al. (2024))

$$
\sum_{t=0}^{k}\alpha_t = \alpha_0\sum_{t=0}^{k}\frac{1}{(t+1)^{1/2}} \geqslant \frac{\alpha_0(k+1)^{1/2}}{2},
$$

$$
\sum_{t=0}^{k}\lambda_t^2 = \lambda_0^2\sum_{t=0}^{k}\frac{1}{t+1} \leqslant 2\lambda_0^2\log(k+2)
$$

Finally,

$$
\begin{aligned}
&\min_{t\leqslant k}\mathbb{E}\left[\|\nabla h(\theta_t)\|^2\right] \\
&\leqslant \frac{1}{\sum_{t'=0}^{k}\alpha_{t'}}\sum_{t=0}^{k}\alpha_t\mathbb{E}\left[\|\nabla h(\theta_t)\|^2\right]
\end{aligned}
$$

$$\leqslant \frac{32}{\alpha_0(k+1)^{1/2}}\left(\|\Delta f_0\|^2 + \|\Delta g_0\|^2 + (h(\theta_0) - h(\theta^\star)) + y_0\right) + \frac{4\log(k+2)}{\alpha_0(k+1)^{1/2}}.$$

■

D.2. Proof of Lemma 12

By the update rule,

$$\begin{aligned}
\|f_k\| &= \|\Delta f_k + F(\theta_k, \omega_k) - F(\theta_k, \omega^\star(\theta_k)) + F(\theta_k, \omega^\star(\theta_k))\| \\
&\leqslant \|\Delta f_k\| + \|F(\theta_k, \omega_k) - F(\theta_k, \omega^\star(\theta_k))\| + \|F(\theta_k, \omega^\star(\theta_k))\| \\
&= \|\Delta f_k\| + \|F(\theta_k, \omega_k) - F(\theta_k, \omega^\star(\theta_k))\| + \|\nabla h(\theta_k)\| \\
&\leqslant \|\Delta f_k\| + L\|\omega_k - \omega^\star(\theta_k)\| + \|\nabla h(\theta_k)\| \\
&\leqslant \|\Delta f_k\| + L\sqrt{y_k} + \|\nabla h(\theta_k)\|,
\end{aligned}$$

where the second inequality follows from the Lipschitz continuity of F.

Similarly, since $G(\theta, \omega^\star(\theta)) = 0$ for any θ, we can derive

$$\begin{aligned}
\|g_k\| &= \|\Delta g_k + G(\theta_k, \omega_k) - G(\theta_k, \omega^\star(\theta))\| \\
&\leqslant \|\Delta g_k\| + \|G(\theta_k, \omega_k) - G(\theta_k, \omega^\star(\theta))\| \\
&\leqslant \|\Delta g_k\| + L\sqrt{y_k}.
\end{aligned}$$

The bounds on $\|\theta_{k+1} - \theta_k\|$ and $\|\omega_{k+1} - \omega_k\|$ easily follow from the two inequalities above and (9).

■

D.3. Proof of Lemma 13

By the update rule in (9), we have

$$\begin{aligned}
\Delta f_{k+1} &= f_{k+1} - F(\theta_{k+1}, \omega_{k+1}) \\
&= (1-\lambda_k) f_k + \lambda_k F(\theta_k, \omega_k, X_k) - F(\theta_{k+1}, \omega_{k+1}) \\
&= (1-\lambda_k) f_k + \lambda_k F(\theta_k, \omega_k) - F(\theta_{k+1}, \omega_{k+1}) + \lambda_k \left(F(\theta_k, \omega_k) - F(\theta_k, \omega_k, X_k)\right) \\
&= (1-\lambda_k)\Delta f_k + F(\theta_k, \omega_k) - F(\theta_{k+1}, \omega_{k+1}) + \lambda_k \left(F(\theta_k, \omega_k) - F(\theta_k, \omega_k, X_k)\right).
\end{aligned}$$

This leads to

$$\begin{aligned}
&\|\Delta f_{k+1}\|^2 \\
&= (1-\lambda_k)^2\|\Delta f_k\|^2 + \left\| \left(F(\theta_k, \omega_k) - F(\theta_{k+1}, \omega_{k+1})\right) + \lambda_k \left(F(\theta_k, \omega_k) - F(\theta_k, \omega_k, X_k)\right) \right\|^2 \\
&\quad + 2(1-\lambda_k)\Delta f_k^\top \left(F(\theta_k, \omega_k) - F(\theta_{k+1}, \omega_{k+1})\right) \\
&\quad + 2(1-\lambda_k)\lambda_k \Delta f_k^\top \left(F(\theta_k, \omega_k) - F(\theta_k, \omega_k, X_k)\right) \\
&\leqslant (1-\lambda_k)^2\|\Delta f_k\|^2 + 2\|F(\theta_k, \omega_k) - F(\theta_{k+1}, \omega_{k+1})\|^2 + 2\lambda_k^2\|F(\theta_k, \omega_k) - F(\theta_k, \omega_k, X_k)\|^2 \\
&\quad + \frac{\lambda_k}{2}\|\Delta f_k\|^2 + \frac{2}{\lambda_k}\|F(\theta_k, \omega_k) - F(\theta_{k+1}, \omega_{k+1})\|^2
\end{aligned}$$

$$
\begin{aligned}
&+ 2(1-\lambda_k)\lambda_k \Delta f_k^\top \left(F(\theta_k,\omega_k) - F(\theta_k,\omega_k,X_k)\right) \\
&\leqslant (1-\lambda_k)\|\Delta f_k\|^2 + 2\lambda_k^2 \|F(\theta_k,\omega_k) - F(\theta_k,\omega_k,X_k)\|^2 \\
&\quad + (\lambda_k^2 - \frac{\lambda_k}{2})\|\Delta f_k\|^2 + \frac{4}{\lambda_k}\|F(\theta_k,\omega_k) - F(\theta_{k+1},\omega_{k+1})\|^2 \\
&\quad + 2(1-\lambda_k)\lambda_k \Delta f_k^\top \left(F(\theta_k,\omega_k) - F(\theta_k,\omega_k,X_k)\right), \qquad (53)
\end{aligned}
$$

where the last inequality follows from $\lambda_k \leqslant 1$.

The second term on the right hand side of (53) is bounded in expectation, i.e.

$$\mathbb{E}_{X_k\sim\xi}[\|F(\theta_k,\omega_k,X_k) - F(\theta_k,\omega_k)\|^2] \leqslant B.$$

In addition, the last term of (53) is zero in expectation since $\mathbb{E}[\Delta f_k^\top (F(\theta_k,\omega_k) - F(\theta_k,\omega_k,X_k)) \mid \mathcal{H}_{k-1}] = \Delta f_k^\top \mathbb{E}[(F(\theta_k,\omega_k) - F(\theta_k,\omega_k,X_k)) \mid \mathcal{H}_{k-1}] = 0$. As a result, we have

$$
\begin{aligned}
\mathbb{E}[\|\Delta f_{k+1}\|^2] &\leqslant (1-\lambda_k)\mathbb{E}[\|\Delta f_k\|^2] - (\frac{\lambda_k}{2} - \lambda_k^2)\mathbb{E}[\|\Delta f_k\|^2] \\
&\quad + \frac{4}{\lambda_k}\mathbb{E}[\|F(\theta_k,\omega_k) - F(\theta_{k+1},\omega_{k+1})\|^2] + 2B\lambda_k^2. \qquad (54)
\end{aligned}
$$

By Lemma 12 and the Lipschitz condition of F,

$$
\begin{aligned}
&\|F(\theta_k,\omega_k) - F(\theta_{k+1},\omega_{k+1})\|^2 \\
&\leqslant 2L^2\|\theta_{k+1} - \theta_k\|^2 + 2L^2\|\omega_{k+1} - \omega_k\|^2 \\
&\leqslant 6L^2\alpha_k^2(\|\Delta f_k\|^2 + L^2 y_k + \|\nabla h(\theta_k)\|^2) + 4L^2\beta_k^2(\|\Delta g_k\|^2 + L^2 y_k) \\
&\leqslant 6L^2\beta_k^2\|\Delta f_k\|^2 + 6L^2\beta_k^2\|\Delta g_k\|^2 + 10L^4\beta_k^2 y_k + 6L^2\alpha_k^2\|\nabla h(\theta_k)\|^2.
\end{aligned}
$$

Plugging this bound into (54), we get

$$
\begin{aligned}
&\mathbb{E}[\|\Delta f_{k+1}\|^2] \\
&\leqslant (1-\lambda_k)\mathbb{E}[\|\Delta f_k\|^2] - (\frac{\lambda_k}{2} - \lambda_k^2)\mathbb{E}[\|\Delta f_k\|^2] \\
&\quad + \frac{4}{\lambda_k}\mathbb{E}\left[6L^2\beta_k^2\|\Delta f_k\|^2 + 6L^2\beta_k^2\|\Delta g_k\|^2 + 10L^4\beta_k^2 y_k + 6L^2\alpha_k^2\|\nabla h(\theta_k)\|^2\right] + 2B\lambda_k^2 \\
&\leqslant (1-\lambda_k)\mathbb{E}[\|\Delta f_k\|^2] - (\frac{\lambda_k}{4} - \frac{24L^2\beta_k^2}{\lambda_k})\mathbb{E}[\|\Delta f_k\|^2] \\
&\quad + \frac{4}{\lambda_k}\mathbb{E}\left[6L^2\beta_k^2\|\Delta g_k\|^2 + 10L^4\beta_k^2 y_k + 6L^2\alpha_k^2\|\nabla h(\theta_k)\|^2\right] + 2B\lambda_k^2,
\end{aligned}
$$

where the last inequality follows from $\lambda_k \leqslant 1/4$.

■

D.4. Proof of Lemma 14

By the update rule in (9), we have

$$\Delta g_{k+1} = g_{k+1} - G(\theta_{k+1},\omega_{k+1})$$

$$
\begin{aligned}
&= (1-\lambda_k)g_k + \lambda_k G(\theta_k,\omega_k,X_k) - G(\theta_{k+1},\omega_{k+1}) \\
&= (1-\lambda_k)g_k + \lambda_k G(\theta_k,\omega_k) - G(\theta_{k+1},\omega_{k+1}) + \lambda_k \left(G(\theta_k,\omega_k) - G(\theta_k,\omega_k,X_k)\right) \\
&= (1-\lambda_k)\Delta g_k + G(\theta_k,\omega_k) - G(\theta_{k+1},\omega_{k+1}) + \lambda_k \left(G(\theta_k,\omega_k) - G(\theta_k,\omega_k,X_k)\right).
\end{aligned}
$$

This leads to

$$
\begin{aligned}
&\|\Delta g_{k+1}\|^2 \\
&= (1-\lambda_k)^2\|\Delta g_k\|^2 + \left\| \left(G(\theta_k,\omega_k) - G(\theta_{k+1},\omega_{k+1})\right) + \lambda_k \left(G(\theta_k,\omega_k) - G(\theta_k,\omega_k,X_k)\right) \right\|^2 \\
&\quad + 2(1-\lambda_k)\Delta g_k^\top \left(G(\theta_k,\omega_k) - G(\theta_{k+1},\omega_{k+1})\right) \\
&\quad + 2(1-\lambda_k)\lambda_k \Delta g_k^\top \left(G(\theta_k,\omega_k) - G(\theta_k,\omega_k,X_k)\right) \\
&\leqslant (1-\lambda_k)^2\|\Delta g_k\|^2 + 2\|G(\theta_k,\omega_k) - G(\theta_{k+1},\omega_{k+1})\|^2 + 2\lambda_k^2\|G(\theta_k,\omega_k) - G(\theta_k,\omega_k,X_k)\|^2 \\
&\quad + \frac{\lambda_k}{2}\|\Delta g_k\|^2 + \frac{2}{\lambda_k}\|G(\theta_k,\omega_k) - G(\theta_{k+1},\omega_{k+1})\|^2 \\
&\quad + 2(1-\lambda_k)\lambda_k \Delta g_k^\top \left(G(\theta_k,\omega_k) - G(\theta_k,\omega_k,X_k)\right) \\
&\leqslant (1-\lambda_k)\|\Delta g_k\|^2 + 2\lambda_k^2\|G(\theta_k,\omega_k) - G(\theta_k,\omega_k,X_k)\|^2 \\
&\quad + (\lambda_k^2 - \frac{\lambda_k}{2})\|\Delta g_k\|^2 + \frac{4}{\lambda_k}\|G(\theta_k,\omega_k) - G(\theta_{k+1},\omega_{k+1})\|^2 \\
&\quad + 2(1-\lambda_k)\lambda_k \Delta g_k^\top \left(G(\theta_k,\omega_k) - G(\theta_k,\omega_k,X_k)\right). \qquad (55)
\end{aligned}
$$

The second term on the right hand side of (55) is bounded in expectation, i.e.

$$\mathbb{E}_{X_k\sim\xi}[\|G(\theta_k,\omega_k,X_k) - G(\theta_k,\omega_k)\|^2] \leqslant B.$$

In addition, the last term of (53) is zero in expectation since $\mathbb{E}[\Delta g_k^\top (G(\theta_k,\omega_k) - G(\theta_k,\omega_k,X_k)) \mid \mathcal{H}_{k-1}] = \Delta g_k^\top \mathbb{E}[(G(\theta_k,\omega_k) - G(\theta_k,\omega_k,X_k)) \mid \mathcal{H}_{k-1}] = 0$. As a result, we have

$$
\begin{aligned}
\mathbb{E}[\|\Delta g_{k+1}\|^2] \leqslant (1-\lambda_k)\mathbb{E}[\|\Delta g_k\|^2] - (\frac{\lambda_k}{2} - \lambda_k^2)\mathbb{E}[\|\Delta g_k\|^2] \\
+ \frac{4}{\lambda_k}\mathbb{E}[\|G(\theta_k,\omega_k) - G(\theta_{k+1},\omega_{k+1})\|^2] + 2B\lambda_k^2. \qquad (56)
\end{aligned}
$$

By Lemma 12 and the Lipschitz condition of G,

$$
\begin{aligned}
&\|G(\theta_k,\omega_k) - G(\theta_{k+1},\omega_{k+1})\|^2 \\
&\leqslant 2L^2\|\theta_{k+1} - \theta_k\|^2 + 2L^2\|\omega_{k+1} - \omega_k\|^2 \\
&\leqslant 6L^2\alpha_k^2(\|\Delta f_k\|^2 + L^2 y_k + \|\nabla h(\theta_k)\|^2) + 4L^2\beta_k^2(\|\Delta g_k\|^2 + L^2 y_k) \\
&\leqslant 6L^2\beta_k^2\|\Delta f_k\|^2 + 6L^2\beta_k^2\|\Delta g_k\|^2 + 10L^4\beta_k^2 y_k + 6L^2\alpha_k^2\|\nabla h(\theta_k)\|^2.
\end{aligned}
$$

Plugging this bound into (56), we get

$$
\begin{aligned}
&\mathbb{E}[\|\Delta g_{k+1}\|^2] \\
&\leqslant (1-\lambda_k)\mathbb{E}[\|\Delta g_k\|^2] - (\frac{\lambda_k}{2} - \lambda_k^2)\mathbb{E}[\|\Delta g_k\|^2]
\end{aligned}
$$

$$+\frac{4}{\lambda_k}\mathbb{E}\left[6L^2\beta_k^2\|\Delta f_k\|^2+6L^2\beta_k^2\|\Delta g_k\|^2+10L^4\beta_k^2 y_k+6L^2\alpha_k^2\|\nabla h(\theta_k)\|^2\right]+2B\lambda_k^2$$
$$\leqslant(1-\lambda_k)\mathbb{E}[\|\Delta g_k\|^2]-(\frac{\lambda_k}{4}-\frac{24L^2\beta_k^2}{\lambda_k})\mathbb{E}[\|\Delta g_k\|^2]$$
$$+\frac{4}{\lambda_k}\mathbb{E}\left[6L^2\beta_k^2\|\Delta f_k\|^2+10L^4\beta_k^2 y_k+6L^2\alpha_k^2\|\nabla h(\theta_k)\|^2\right]+2B\lambda_k^2,$$

where the last inequality follows from $\lambda_k \leqslant 1/4$.

■

D.5. Proof of Lemma 15

From the update rule (8) and the Lipschitz gradient condition in Assumption 1, we have

$$\begin{aligned}
h(\theta_{k+1}) &\leqslant h(\theta_k)+\langle\nabla h(\theta_k),\theta_{k+1}-\theta_k\rangle+\frac{L}{2}\|\theta_{k+1}-\theta_k\|^2\\
&= h(\theta_k)-\alpha_k\langle\nabla h(\theta_k),f_k\rangle+\frac{L\alpha_k^2}{2}\|f_k\|^2\\
&= h(\theta_k)-\alpha_k\langle\nabla h(\theta_k),\Delta f_k\rangle-\alpha_k\langle\nabla h(\theta_k),F(\theta_k,\omega^\star(\theta_k))\rangle\\
&\quad+\alpha_k\langle\nabla h(\theta_k),F(\theta_k,\omega^\star(\theta_k))-F(\theta_k,\omega_k)\rangle+\frac{L\alpha_k^2}{2}\|f_k\|^2\\
&= h(\theta_k)-\alpha_k\|\nabla h(\theta_k)\|^2-\alpha_k\langle\nabla h(\theta_k),\Delta f_k\rangle\\
&\quad+\alpha_k\langle\nabla h(\theta_k),F(\theta_k,\omega^\star(\theta_k))-F(\theta_k,\omega_k)\rangle+\frac{L\alpha_k^2}{2}\|f_k\|^2. \qquad (57)
\end{aligned}$$

To bound the second term on the right hand side of (57),

$$-\alpha_k\langle\nabla h(\theta_k),\Delta f_k\rangle\leqslant\alpha_k\|\nabla h(\theta_k)\|\|\Delta f_k\|\leqslant\frac{\alpha_k}{4}\|\nabla h(\theta_k)\|^2+\alpha_k\|\Delta f_k\|^2, \qquad (58)$$

where the final inequality follows from the fact that $2\|\boldsymbol{a}\|\|\boldsymbol{b}\|\leqslant c\|\boldsymbol{a}\|^2+\frac{1}{c}\|\boldsymbol{b}\|^2$ for any vector $\boldsymbol{a},\boldsymbol{b}$ and scalar $c>0$.

The third term on the right hand side of (57) can be treated in a similar manner

$$\begin{aligned}
\alpha_k\langle\nabla h(\theta_k),F(\theta_k,\omega^\star(\theta_k))-F(\theta_k,\omega_k)\rangle &\leqslant\alpha_k\|\nabla h(\theta_k)\|\|F(\theta_k,\omega^\star(\theta_k))-F(\theta_k,\omega_k)\|\\
&\leqslant L\alpha_k\|\nabla h(\theta_k)\|\|\omega^\star(\theta_k)-\omega_k\|\\
&\leqslant\frac{\alpha_k}{4}\|\nabla h(\theta_k)\|^2+L^2\alpha_k y_k. \qquad (59)
\end{aligned}$$

Plugging (58) and (59) into (57), we get

$$\begin{aligned}
&h(\theta_{k+1})\\
&\leqslant h(\theta_k)-\alpha_k\|\nabla h(\theta_k)\|^2-\alpha_k\langle\nabla h(\theta_k),\Delta f_k\rangle\\
&\quad+\alpha_k\langle\nabla h(\theta_k),F(\theta_k,\omega^\star(\theta_k))-F(\theta_k,\omega_k)\rangle+\frac{L\alpha_k^2}{2}\|f_k\|^2\\
&\leqslant h(\theta_k)-\alpha_k\|\nabla h(\theta_k)\|^2+\frac{\alpha_k}{4}\|\nabla h(\theta_k)\|^2+\alpha_k\|\Delta f_k\|^2+\frac{\alpha_k}{4}\|\nabla h(\theta_k)\|^2
\end{aligned}$$

$$
\begin{aligned}
&+ L^2\alpha_k y_k + \frac{L\alpha_k^2}{2}\|f_k\|^2 \\
&\leqslant h(\theta_k) - \frac{\alpha_k}{2}\|\nabla h(\theta_k)\|^2 + \alpha_k\|\Delta f_k\|^2 + L^2\alpha_k y_k + \frac{3L\alpha_k^2}{2}\left(\|\Delta f_k\|^2 + L^2 y_k + \|\nabla h(\theta_k)\|^2\right) \\
&\leqslant h(\theta_k) - \frac{\alpha_k}{4}\|\nabla h(\theta_k)\|^2 + \frac{5\alpha_k}{4}\|\Delta f_k\|^2 + \frac{5L^2\alpha_k}{4} y_k,
\end{aligned}
$$

where the third inequality plugs in the result of Lemma 12, and the final inequality uses $\alpha_k \leqslant \frac{1}{6L}$. ■

D.6. Proof of Lemma 16

By the update rule in (8),

$$
\begin{aligned}
\omega_{k+1} - \omega^\star(\theta_{k+1}) &= \omega_{k+1} - \omega^\star(\theta_{k+1}) \\
&= (\omega_k - \omega^\star(\theta_k)) - \beta_k g_k - (\omega^\star(\theta_{k+1}) - \omega^\star(\theta_k)) \\
&= (\omega_k - \omega^\star(\theta_k)) - \beta_k G(\theta_k, \omega_k) - \beta_k \Delta g_k - (\omega^\star(\theta_{k+1}) - \omega^\star(\theta_k)).
\end{aligned}
$$

Taking the norm yields

$$
\begin{aligned}
y_{k+1} &= \|\omega_{k+1} - \omega^\star(\theta_{k+1})\|^2 \\
&= \underbrace{\|\omega_k - \omega^\star(\theta_k) - \beta_k G(\theta_k, \omega_k)\|^2}_{T_1} + \beta_k^2\|\Delta g_k\|^2 + \underbrace{\|\omega^\star(\theta_{k+1}) - \omega^\star(\theta_k)\|^2}_{T_2} \\
&\quad \underbrace{-2\left(\omega_k - \omega^\star(\theta_k) - \beta_k G(\theta_k, \omega_k)\right)^\top \left(\omega^\star(\theta_{k+1}) - \omega^\star(\theta_k)\right)}_{T_3} \\
&\quad \underbrace{-2\beta_k\left(\omega_k - \omega^\star(\theta_k) - \beta_k G(\theta_k, \omega_k)\right)^\top \Delta g_k}_{T_4} + \underbrace{2\beta_k \Delta g_k^\top \left(\omega^\star(\theta_{k+1}) - \omega^\star(\theta_k)\right)}_{T_5}.
\end{aligned} \tag{60}
$$

We bound each term of (60) individually. First, since by definition $G(\theta, \omega^\star(\theta)) = 0$ for all θ, we have

$$
\begin{aligned}
T_1 &= \|\omega_k - \omega^\star(\theta_k)\|^2 - 2\beta_k\left(\omega_k - \omega^\star(\theta_k)\right)^\top G(\theta_k, \omega_k) + \beta_k^2\|G(\theta_k, \omega_k)\|^2 \\
&= y_k - 2\beta_k\left(\omega_k - \omega^\star(\theta_k)\right)^\top G(\theta_k, \omega_k) + \beta_k^2\|G(\theta_k, \omega_k) - G(\theta_k, \omega^\star(\theta_k))\|^2 \\
&\leqslant y_k - 2\mu_G\beta_k\|\omega_k - \omega^\star(\theta_k)\|^2 + L^2\beta_k^2\|\omega_k - \omega^\star(\theta_k)\|^2 \\
&= (1 - 2\mu_G\beta_k + L^2\beta_k^2) y_k,
\end{aligned}
$$

where the first inequality follows from Assumption 3 and the Lipschitz continuity of G.

The term T_2 can be simply treated using the Lipschitz condition of $\omega^\star$

$$
\begin{aligned}
T_2 &= \|\omega^\star(\theta_{k+1}) - \omega^\star(\theta_k)\|^2 \leqslant L^2\|\theta_{k+1} - \theta_k\|^2 \leqslant L^2\alpha_k^2\left(\|\Delta f_k\| + L\sqrt{y_k} + \|\nabla h(\theta_k)\|\right)^2 \\
&\leqslant 3L^2\alpha_k^2\left(\|\Delta f_k\|^2 + L^2 y_k + \|\nabla h(\theta_k)\|^2\right).
\end{aligned}
$$

By the Cauchy-Schwarz inequality,

T_3

$$
\begin{aligned}
&\leqslant 2\|\omega_k - \omega^\star(\theta_k) - \beta_k G(\theta_k,\omega_k)\|\|\omega^\star(\theta_{k+1}) - \omega^\star(\theta_k)\| \\
&\leqslant 2\|\omega_k - \omega^\star(\theta_k)\|\|\omega^\star(\theta_{k+1}) - \omega^\star(\theta_k)\| + 2\beta_k\|G(\theta_k,\omega_k) - G(\theta_k,\omega^\star(\theta_k))\|\|\omega^\star(\theta_{k+1}) - \omega^\star(\theta_k)\| \\
&\leqslant (2L + 2L^2\beta_k)\sqrt{y_k}\|\theta_{k+1} - \theta_k\| \\
&\leqslant 4L\alpha_k\sqrt{y_k}\left(\|\Delta f_k\| + L\sqrt{y_k} + \|\nabla h(\theta_k)\|\right).
\end{aligned}
$$

where the second inequality follows from $G(\theta, \omega^\star(\theta)) = 0$ for all θ, and the fourth inequality applies Lemma 12 and the step size rule $\beta_k \leqslant 1/L$.

We can bound T_4 following a similar line of analysis

$$
T_4 \leqslant 2\beta_k\|\omega_k - \omega^\star(\theta_k) - \beta_k G(\theta_k,\omega_k)\|\|\Delta g_k\| \leqslant 4\beta_k\sqrt{y_k}\|\Delta g_k\|.
$$

Finally, we treat T_5 again by the Cauchy-Schwarz inequality

$$
\begin{aligned}
T_5 &\leqslant 2\beta_k\|\Delta g_k\|\|\omega^\star(\theta_{k+1}) - \omega^\star(\theta_k)\| \\
&\leqslant 2L\alpha_k\beta_k\|\Delta g_k\|\Big(\|\Delta f_k\| + L\sqrt{y_k} + \|\nabla h(\theta_k)\|\Big),
\end{aligned}
$$

where the second inequality again uses Lemma 12.

Applying the bounds on T_1-T_5 to (60), we get

$$
\begin{aligned}
&y_{k+1} \\
&\leqslant (1 - 2\mu_G\beta_k + L^2\beta_k^2)y_k + \beta_k^2\|\Delta g_k\|^2 + 3L^2\alpha_k^2\left(\|\Delta f_k\|^2 + L^2 y_k + \|\nabla h(\theta_k)\|^2\right) \\
&\quad + 4L\alpha_k\sqrt{y_k}\left(\|\Delta f_k\| + L\sqrt{y_k} + \|\nabla h(\theta_k)\|\right) + 4\beta_k\sqrt{y_k}\|\Delta g_k\| \\
&\quad + 2L\alpha_k\beta_k\|\Delta g_k\|\Big(\|\Delta f_k\| + L\sqrt{y_k} + \|\nabla h(\theta_k)\|\Big) \\
&\leqslant (1 - 2\mu_G\beta_k + L^2\beta_k^2)y_k + \beta_k^2\|\Delta g_k\|^2 + 3L^2\alpha_k^2\left(\|\Delta f_k\|^2 + L^2 y_k + \|\nabla h(\theta_k)\|^2\right) \\
&\quad + 2L\alpha_k y_k + 2L\alpha_k\|\Delta f_k\|^2 + 4L^2\alpha_k y_k + 96L^2\alpha_k y_k + \frac{\alpha_k}{24}\|\nabla h(\theta_k)\|^2 + \frac{\mu_G\beta_k}{2}y_k + \frac{8\beta_k}{\mu_G}\|\Delta g_k\|^2 \\
&\quad + L\alpha_k\beta_k\|\Delta g_k\|^2 + 3L\alpha_k\beta_k\left(\|\Delta f_k\|^2 + L^2 y_k + \|\nabla h(\theta_k)\|^2\right) \\
&\leqslant (1 - \mu_G\beta_k)y_k - \left(\frac{\mu_G}{2}\beta_k - 108L^4\alpha_k - L^2\beta_k^2\right)y_k \\
&\quad + \frac{\alpha_k}{8}\|\nabla h(\theta_k)\|^2 + \left(\frac{8\beta_k}{\mu_G} + 2L\beta_k^2\right)\|\Delta g_k\|^2 + 8L\alpha_k\|\Delta f_k\|^2,
\end{aligned}
$$

where the second inequality follows from the fact that $2\|\boldsymbol{a}\|\|\boldsymbol{b}\| \leqslant c\|\boldsymbol{a}\|^2 + \frac{1}{c}\|\boldsymbol{b}\|^2$ for any vector $\boldsymbol{a}, \boldsymbol{b}$ and scalar $c > 0$. We have also simplified terms using the conditions $\alpha_k \leqslant \frac{1}{72L^2}$ and $\beta_k \leqslant \frac{1}{72L}$. ■

Appendix E. Simulations Details

We generate completely random MDPs with transition probability matrix $\mathcal{P} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|\times|\mathcal{S}|}$ drawn i.i.d. from $\text{Unif}(0,1)$ and then normalized such that

$$
\sum_{s'\in\mathcal{S}} \mathcal{P}(s' \mid s, a) = 1, \quad \forall s, a.
$$

The behavior policy π_b is chosen to be uniform for all states. The policy π to be evaluated is generated randomly by first drawing $\psi \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ such that $\psi_{s,a} \sim N(0,1)$, which produces the policy π through the softmax function

$$\pi(a \mid s) = \frac{\exp(\psi_{s,a})}{\sum_{a'} \exp(\psi_{s,a'})}, \quad \forall s, a.$$

We draw the feature matrix $\Phi \in \mathbb{R}^{|\mathcal{S}|\times dx}$ entry-wise i.i.d. from $N(0,1)$. The value function $V \in \mathbb{R}^{|\mathcal{S}|}$ is the solution to the Bellman equation

$$V = R + \gamma P^{\pi} V,$$

where the matrix $P^{\pi} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ is the state transition matrix under policy π

$$P_{s,s'} = \sum_{a} \mathcal{P}(s' \mid s, a)\pi(a \mid s)$$

We would like to design the experiments such that the value function $V \in \mathbb{R}^{|\mathcal{S}|}$ lies in the span of the feature matrix. To achieve, we determine the optimal parameter $\theta^{\star}$ first, by sampling it entry-wise i.i.d. from $N(0,1)$. This produces the value function $V = \Phi\theta$, from which we reverse engineer the reward function

$$R = (I - \gamma P^{\pi})V.$$

The discount factor γ is 0.5.

All algorithms start with zero initialization, i.e. $\theta_0 = \omega_0 = f_0 = g_0 = 0$. The step sizes α_k and β_k are set to constant values $5e-4$ and $2e-3$ for simplicity, while the step size λ_k is selected as

$$\lambda_k = \frac{4}{5(k+10)}.$$