Journal of Machine Learning Research 27 (2026) 1-69 Submitted 6/25; Revised 4/26; Published 4/26

Finite-Time Decoupled Convergence in Nonlinear Two-Time-Scale Stochastic Approximation

Yuze Han HANYUZE97@RUC.EDU.CN
Center for Applied Statistics and School of Statistics
Renmin University of China
Beijing, China

Xiang Li LX10077@PKU.EDU.CN
School of Mathematical Sciences
Peking University
Beijing, China

Zhihua Zhang* ZHZHANG@MATH.PKU.EDU.CN
School of Mathematical Sciences
Peking University
Beijing, China

Editor: Peter Richtarik

Abstract

In two-time-scale stochastic approximation (SA), two iterates are updated at varying speeds using different step sizes, with each update influencing the other. Previous studies on linear two-time-scale SA have shown that the convergence rates of the mean-square errors for these updates depend solely on their respective step sizes, a phenomenon termed decoupled convergence. However, achieving decoupled convergence in nonlinear SA remains less understood. Our research investigates the potential for finite-time decoupled convergence in nonlinear two-time-scale SA. We demonstrate that, under a nested local linearity assumption, finite-time decoupled convergence rates can be achieved with suitable step size selection. To derive this result, we conduct a convergence analysis of the matrix cross term between the iterates and leverage fourth-order moment convergence rates to control the higher-order error terms induced by local linearity. To further investigate the necessity of local linearity for decoupled convergence, we also construct an example showing that, even when the fast-time-scale update is linear, the nonlinearity of the slow-time-scale update alone can destroy decoupled convergence.

Keywords: two-time-scale stochastic approximation, finite-time convergence, decoupled convergence

1. Introduction

Stochastic approximation (SA), initially introduced by Robbins and Monro (1951), is an iterative method for finding the root of an unknown operator based on noisy observations. This method has gained substantial attention over the past few decades, finding applications in stochastic optimization and reinforcement learning (Kushner and Yin, 2003; Borkar, 2009; Mou et al., 2020, 2022, 2023; Li et al., 2023b,a). However, certain scenarios require

*. Corresponding Author

managing two iterates updated at different time scales using varying step sizes. Examples include stochastic bilevel optimization (Ghadimi and Wang, 2018; Chen et al., 2021; Hong et al., 2023), temporal difference learning (Sutton et al., 2009; Xu et al., 2019; Xu and Liang, 2021; Wang et al., 2021) and actor-critic methods (Borkar and Konda, 1997; Konda and Tsitsiklis, 2003; Wu et al., 2020; Xu et al., 2020). These cases highlight the limitations of traditional SA and underscore the need for a two-time-scale SA approach to better capture the complexities involved.

In this paper, we study the two-time-scale SA (Borkar, 1997), a variant of the classical SA algorithm, designed to identify the roots of systems comprising two coupled, potentially nonlinear equations. We focus on two unknown Lipschitz operators, $F : \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \to \mathbb{R}^{d_x}$ and $G : \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \to \mathbb{R}^{d_y}$, with the aim of finding the root pair $(x^\star, y^\star)$ satisfying

$$\begin{cases} F(x^\star, y^\star) = 0, \\ G(x^\star, y^\star) = 0. \end{cases} \tag{1}$$

Given that F and G are unknown, we assume access to a stochastic oracle that provides noisy evaluations of $F(x, y)$ and $G(x, y)$ at any input pair (x, y). Specifically, for any x and y, the oracle returns $F(x, y) + \xi$ and $G(x, y) + \psi$ where ξ and ψ represent noise components. Utilizing this stochastic oracle, we apply the nonlinear two-time-scale SA to solve the problem defined in (1) by iteratively updating estimates x_t and y_t for $x^\star$ and $y^\star$, respectively, as follows

$$x_{t+1} = x_t - \alpha_t \left(F(x_t, y_t) + \xi_t\right), \tag{2a}$$
$$y_{t+1} = y_t - \beta_t \left(G(x_t, y_t) + \psi_t\right), \tag{2b}$$

where x_0 and y_0 are initialized arbitrarily in $\mathbb{R}^{d_x}$ and $\mathbb{R}^{d_y}$. The noise components ξ_t and ψ_t are modeled as martingale difference sequences. The step sizes α_t and β_t are chosen such that $\beta_t \ll \alpha_t$, making y_t "quasi-static" relative to x_t. This difference in step sizes simulates a setting where y_t remains nearly fixed while x_t undergoes rapid updates, as discussed in Konda and Tsitsiklis (2004).

Assuming that for a given y, the equation $F(x, y) = 0$ has a unique solution $H(y)$, where H is a Lipschitz operator. Under this assumption, we can regard two-time-scale SA as a single-loop approximation of the following double-loop process

$$\begin{aligned} &\text{Inner loop: compute } x = H(y), \text{ or equivalently, solve } F(x, y) = 0 \text{ for a fixed } y, \\ &\text{Outer loop: iteratively find the root of } G(H(y), y) = 0. \end{aligned} \tag{3}$$

This formulation ensures that the solution to (1) satisfies $x^\star = H(y^\star)$; in other words, for $y = y^\star$, $x^\star$ is precisely the root of the inner loop. We designate the update of x_t as the "fast-time-scale" update and that of y_t as the "slow-time-scale" update, referring x_t and y_t as the *fast iterate* and *slow iterate*, respectively.

Our primary interest lies in the behavior of the *slow iterate* y_t, a focus we illustrate through a classic example. Consider the case of stochastic gradient descent (SGD) with Polyak-Ruppert averaging, which is an instance of two-time-scale SA. To minimize a strongly convex objective f, the SGD with Polyak-Ruppert averaging updates are defined as follows

$$x_{t+1} = x_t - \alpha_t(\nabla f(x_t) + \xi_t) \text{ with } \alpha_t = \frac{\alpha_0}{(t+1)^a} \ (0 < a \leq 1), \tag{4a}$$

$$y_{t+1} = \frac{1}{t+1}\sum_{\tau=0}^{t} x_\tau = y_t - \beta_t(y_t - x_t) \text{ with } \beta_t = \frac{1}{t+1}, \tag{4b}$$

where (4a) is the classic SGD update rule (Robbins and Monro, 1951) and (4b) is the Polyak-Ruppert averaging step (Polyak and Juditsky, 1992; Ruppert, 1988). This setup aligns with the formulation (3), where we have $F(x, y) = \nabla f(x)$ and $G(x, y) = y - x$, resulting in $x^\star = y^\star = \arg\min_x f(x)$ and $H(y) \equiv x^\star$. As shown in Moulines and Bach (2011), the convergence rate for $\mathbb{E}\|x_t - x^\star\|^2$ is $\mathcal{O}(\alpha_t)$, achieving the optimal rate $\mathcal{O}(1/t)$ only when the step size parameter $a = 1$ and the initial step size α_0 is sufficiently large. In contrast, with two-time-scale SA, the slow-time-scale update satisfies $\mathbb{E}\|y_t - y^\star\|^2 = \mathcal{O}(1/t)$ as long as $a \geq 0.5$. This result implies that using the slow-time-scale update (4b) enables optimal convergence rates with greater flexibility in choosing step sizes for the fast-time-scale update in (4a). Since the convergence rate of y_t is the desired result, the behavior of x_t, which serves as an auxiliary iterate, becomes secondary.

In the above example, each iterate's convergence rate depends only on its own step size. Similar results were achieved by Kaledin et al. (2020) for linear operators F and G, yielding

$$\mathbb{E}\|y_t - y^\star\|^2 = \mathcal{O}(\beta_t) \qquad \text{and} \qquad \mathbb{E}\|x_t - H(y_t)\|^2 = \mathcal{O}(\alpha_t). \tag{5}$$

Here $y_t - y^\star$ and $x_t - H(y_t)$ represent the errors of the outer and inner loops in (3), respectively. We refer to the phenomenon where each iterate's convergence rate depends solely on its own step size as *decoupled convergence*.

However, when F and G are nonlinear operators, the interactions between the two iterates become more complex, making the path to (5) less straightforward. The asymptotic analysis in Mokkadem and Pelletier (2006), under an additional local linearity assumption, provides evidence supporting decoupled convergence for sufficiently large t. Nevertheless, there is no corresponding non-asymptotic guarantee for any finite t. Consequently, our research goal is to

Establish the *finite-time decoupled convergence* in (5) for nonlinear two-time-scale SA.

The significance of decoupled convergence rates is twofold. First, as analyzed for the SGD example in (4), decoupled convergence allows greater flexibility in selecting step sizes for the fast iterate without affecting the convergence behavior of the main focus, the slow iterate. Additionally, decoupled convergence rates offer a more refined analysis than previous work (Doan, 2022; Shen and Chen, 2022), especially in strict two-time-scale scenarios. Such refined rates are valuable for further asymptotic trajectory analysis (Liang et al., 2023) and online statistical inference (Li et al., 2022, 2023a). Moreover, we emphasize that our main objective is not merely to show that the rate $\mathbb{E}\|y_t - y^\star\|^2 = \Theta(t^{-1})$ can be attained under some particular choice of step sizes. Rather, our goal is to show that, over a broad regime of step-size choices, the convergence rate of the slow iterate can be made essentially *independent* of the step size on the fast time scale.

Contributions. In this paper, we focus on nonlinear two-time-scale SA under the assumptions of strong monotonicity and martingale difference noise, as specified in Assumptions 3 and 6. Our contributions are summarized as follows.

- **Theoretical contribution**: Under the nested local linearity assumption (Assumption 5), we derive detailed convergence rates for $\mathbb{E}\|x_t - H(y_t)\|^2$, $\mathbb{E}\|y_t - y^\star\|^2$ and

$\|\mathbb{E}(x_t - H(y_t))(y_t - y^\star)^\top\|$ in Theorem 3, establishing finite-time decoupled convergence in Corollary 4 with appropriate step size selection. We further investigate the necessity of local linearity for decoupled convergence. In particular, we construct an example and prove in Proposition 5 that, even when F and H are linear, the nonlinearity of G alone can already slow down the convergence rate of the slow iterate. Taken together, these upper and lower bounds provide a relatively complete characterization of when decoupled convergence can be achieved in the nonlinear setting. Moreover, this lower bound also complements the approximation perspective in (3): although two-time-scale SA can be viewed as solving $F(x, y) = 0$ (for fixed y) and $G(H(y), y) = 0$, the original form of $G(x, y)$ may still influence the convergence rates.

- **Technical contribution**: We develop a systematic proof framework for establishing decoupled convergence in the nonlinear setting. A key ingredient is the treatment of the matrix cross term $\|\mathbb{E}(x_t - H(y_t))(y_t - y^*)^\top\|$, which is crucial for obtaining a sharp convergence characterization of the interacting sequences $\{x_t\}_{t=0}^\infty$ and $\{y_t\}_{t=0}^\infty$. While the use of such a cross term is related to the linear-case analysis in Kaledin et al. (2020), the nonlinear setting requires several additional ingredients: an initial coarse convergence-rate analysis, in the spirit of Doan (2022); local linear approximations of F, G, and H; control of the resulting higher-order error terms; a convergence analysis of fourth-order moments; and a final integration of these ingredients to derive the decoupled rates. This proof framework could provide a useful foundation for future finite-time analyses of nonlinear interacting stochastic approximation schemes with multiple time scales and variable step sizes.

1.1 Related Work

Our research investigates the finite-time convergence rate of nonlinear two-time-scale SA and endeavors to establish the decoupled convergence. To contextualize our results, we provide more background of two-time-scale SA.

Decoupled convergence for the linear case. When both F and G are linear, H is also linear. Leveraging this linear structure, Konda and Tsitsiklis (2004) focused on a linearly transformed error $z_t = x_t - H(y_t) + L_t(y_t - y^\star)$, demonstrating that its update does not depend on y_t. Here $\{L_t\}$ is a matrix sequence converging to zero. Based on this insight, they prove that $\beta_t^{-1/2}(y_t - y^\star)$ converges in distribution to a normal distribution under martingale difference noise. It can be extrapolated from their analysis that the $\alpha_t^{-1/2}(x_t - x^\star)$ and $\alpha_t^{-1/2}(x_t - H(y_t))$ also converge in distribution to a normal distribution. Drawing inspiration from this technique, Kaledin et al. (2020) derived finite-time convergence rates, establishing that $\mathbb{E}\|y_t - y^\star\|^2 = \mathcal{O}(\beta_t)$ and $\mathbb{E}\|x_t - H(y_t)\|^2 = \mathcal{O}(\alpha_t)$ hold for both martingale difference and Markovian noise. Concurrent to our work, Haque et al. (2023) achieved the same rates with asymptotically optimal leading terms; Kwon et al. (2025) examined constant stepsize schemes and provided a refined characterization of the bias and variance terms. For a variant of two-time-scale SA with sparse projection, Dalal et al. (2020) derived that $\|y_t - y^\star\| = \widetilde{\mathcal{O}}(\beta_t)$ and $\|x_t - x^\star\| = \widetilde{\mathcal{O}}(\alpha_t)$ hold with high probability.

Decoupled convergence for the nonlinear case. When either F or G is nonlinear, the coupling between x_t and y_t is in a more complex nonlinear form, complicating the

analysis of nonlinear two-time-scale SA. Decoupled convergence analysis remains largely in the domain of asymptotic results. Under local linearity condition of F and G around $(x^\star, y^\star)$ and assuming stability, Mokkadem and Pelletier (2006) demonstrated that $\begin{pmatrix} \alpha_t^{-1/2}(x_t - x^\star) \\ \beta_t^{-1/2}(y_t - y^\star) \end{pmatrix}$ converges weakly to a normal distribution. Recently, this central limit theorem has been extended to settings with Markovian noise by Hu et al. (2024) and to continuous-time dynamics by Sharrock (2022). To the best of our knowledge, our work provides the first finite-time (non-asymptotic) decoupled convergence rate for the nonlinear case.

Other convergence rates for the nonlinear case. Under the strongly monotone and Lipschitz conditions, Doan (2022) and Doan (2021) achieved the $\mathcal{O}(t^{-2/3})$ and $\widetilde{\mathcal{O}}(t^{-2/3})$ convergence rates for the slow iterate with martingale difference noise and Markovian noise, respectively. Further assuming the Lipschitz continuity of ∇H, Shen and Chen (2022) improved this rate to $\mathcal{O}(1/t)$. Huang et al. (2025) and Chandak (2025a) obtained the same rate without this additional condition by introducing auxiliary sequences. An alternative approach to achieve the $\mathcal{O}(1/t)$ rate without relying on the Lipschitz condition of ∇H is to leverage an averaging step to improve the estimates of the operators (Zeng and Doan, 2024; Doan, 2025). Related developments include convergence rates for state- and time-dependent noises (Chen et al., 2025), concentration bounds (Borkar and Pattathil, 2018), functional central limit theorems (Faizal and Borkar, 2023; Han et al., 2024), and finite-time analysis under arbitrary norms with Markovian noise (Chandak et al., 2025). In the absence of the strong monotonicity in the outer loop of (3), two-time-scale SA has been studied within the framework of bilevel optimization (Hong et al., 2023; Zeng et al., 2024) or non-expansive mappings (Chandak, 2025b).

Notation. For a vector x, $\|x\|$ denotes the Euclidean norm; for a matrix A, $\|A\|$ to denote the spectral norm. We use $o(\cdot)$, $\mathcal{O}(\cdot)$, $\Omega(\cdot)$, and $\Theta(\cdot)$ to hide universal constants and $\widetilde{\mathcal{O}}(\cdot)$ to hide both universal constants and log factors. We denote $\max\{a, b\}$ as $a \vee b$ and $\min\{a, b\}$ as $a \wedge b$. The ceiling function $\lceil \cdot \rceil$ denotes the smallest integer greater than or equal to the input number. For two non-negative numbers a and b, $a \lesssim b$ ($a \propto b$) indicates the existence of a positive number C such that $a \leq Cb$ ($a = Cb$) with C depending on parameters of no interest. For two positive sequence $\{a_n\}$ and $\{b_n\}$, $a_n \sim b_n$ signifies $\lim_{n\to\infty} a_n/b_n = 1$. By $\overset{a.s.}{\to}$ we denote almost sure convergence; by $\overset{d}{\to}$ we denote the convergence in distribution. We abbreviate $\{1, 2, \ldots, n\}$ as $[n]$.

Organization. The remainder of this paper is organized as follows. Section 2 introduces several motivating examples, and Section 3 states the basic assumptions. The main theoretical results are presented in Section 4, with the proof framework outlined in Section 5. Section 6 illustrates some numerical results, and Section 7 concludes the paper. Detailed proofs are provided in the appendices.

2. Motivating Examples

In this section, we present several examples of two-time-scale SA. In most of these examples, our primary focus is on evaluating the performance of the slow iterate y_t, with the fast iterate x_t playing a secondary role as an auxiliary sequence.

In the next three examples, we assume $f(\cdot)\colon \mathbb{R}^{d_x} \to \mathbb{R}$ is a strongly convex function and the unique minimizer is $x_o^\star = \arg\min_{x\in\mathbb{R}^{d_x}} f(x)$.

Example 1 (SGD with Polyak-Ruppert averaging) *SGD with Polyak-Ruppert averaging has been introduced in Section 1. Suppose that we want to minimize a function f with access only to the noisy observations of the true gradients. To find the true minimizer, the stochastic gradient method (SGD) (Robbins and Monro, 1951) iteratively updates the iterate x_t by* (4a)*. In order to improve the convergence of SGD, an additional averaging step* (4b) *is often used (Polyak and Juditsky, 1992; Ruppert, 1988). Obviously, these two updates are a special case of the nonlinear two-time-scale SA in* (2) *with $F(x,y) = \nabla f(x), G(x,y) = y - x$ and $H(y) \equiv x^\star$. It follows that $G(H(y), y) = y - x^\star$ and $y^\star = x^\star = x_o^\star$.*

Now we contextualize this example within the framework of (3)*. In the inner loop, our goal is to reduce the norm of $\nabla f(x)$, while in the outer loop, we strive to approach $x_o^\star$. For a strongly convex objective, the two objectives coincide. Moulines and Bach (2011) have demonstrated that $\mathbb{E}\|x_t - x^\star\|^2 = \mathcal{O}(\alpha_t)$ and $\mathbb{E}\|y_t - y^\star\|^2 = \mathcal{O}(1/t)$ as long as $a \geq 1/2$. Given the improved convergence rate of the averaged sequence, our primary interest lies in the slow iterate y_t.*

Example 2 (SGD with momentum) *Stochastic heavy ball (SHB) is a variant of SGD based on momentum and adaptive step sizes and has been shown to be effective (Gadat et al., 2018). A "normalized" version of SHB (Gupal and Bazhenov, 1972; Gitman et al., 2019) iteratively runs*

$$\begin{aligned} x_{t+1} &= x_t - \alpha_t(x_t - \nabla f(y_t) + \xi_t), \\ y_{t+1} &= y_t - \beta_t x_t. \end{aligned} \tag{6}$$

Here one should interpret x_t as a (stochastic) search direction that is defined to be a combination of the current stochastic gradient $\nabla f(y_t) + \xi_t$ and past search direction x_t.[1] *These two updates are a special case of the nonlinear two-time-scale SA in* (2) *with $F(x,y) = x - \nabla f(y), G(x,y) = x$ and $H(y) = \nabla f(y)$. It follows that $G(H(y), y) = \nabla f(y)$, $y^\star = x_o^\star$ and $x^\star = 0$.*

Now, we integrate this example into the framework of (3)*. In the inner loop, our objective is to approximate the gradient of a fixed y, while in the outer loop, we aim to locate the stationary point (also the minimizer) $x_o^\star$, Thus, our primary focus lies in the slow iterate y_t.*

Example 3 (Constrained optimization with Lagrangian multipliers) *Consider the linearly constrained optimization $\min_{x \in \mathbb{R}^{d_x}} f(x)$, s.t. $Ax = b$. This problem can be solved by introducing Lagrange function $L(x,y) = f(x) + y^\top(Ax - b)$ and applying the following primal-dual method (Platt and Barr, 1987)*

$$\begin{aligned} x_{t+1} &= x_t - \alpha_t(\nabla f(x_t) + A^\top y_t + \xi_t), \\ y_{t+1} &= y_t + \beta_t(Ax_t - b), \end{aligned} \tag{7}$$

where y_t denotes the Lagrange multiplier. This update scheme can be viewed as an approximation to dual ascent (Boyd et al., 2011) and is a special case of the nonlinear two-time-scale SA in (2) *with $F(x,y) = \nabla f(x) + A^\top y, G(x,y) = -Ax + b$ and $H(y) = [\nabla f]^{-1}(-A^\top y)$. The equations $F(x,y) = 0$ and $G(x,y) = 0$ correspond precisely to the KKT conditions. Therefore, $x^\star$ and $y^\star$, satisfying $x^\star = H(y^\star)$, are the solutions to the primal and dual problems, respectively.*

1. Here we employ a slightly different update rule for y_t to be consistent with Doan (2022).

We now integrate this example into the framework of (3)*. In the inner loop, for a fixed* y*, the goal is to solve* $\min_{x\in\mathbb{R}^{d_y}} L(x,y)$*, where* $H(y)$ *specifies the corresponding solution. In the outer loop, the aim is to solve the dual problem* $\max_{y\in\mathbb{R}^{d_y}} \min_{x\in\mathbb{R}^{d_x}} L(x,y)$*.*

The final example is bilevel optimization, a topic with a long history in the optimization literature (Bracken and McGill, 1973; Colson et al., 2007). With some abuse of notation, we examine the following (unconstrained) bilevel optimization problem

$$\min_{y\in\mathbb{R}^{d_y}} \ell(y) := g(\tilde{x}^\star(y), y) \quad \text{s.t.} \quad \tilde{x}^\star(y) := \arg\min_{x\in\mathbb{R}^{d_x}} f(x,y), \tag{8}$$

where $f(x,y)$ is the *inner objective function* and $\ell(y)$ is the *outer objective function*. We refer to $\min_{x\in\mathbb{R}^{d_x}} f(x,y)$ as the *inner problem* and $\min_{y\in\mathbb{R}^{d_y}} \ell(y)$ as the *outer problem*. Under some regularization conditions (Ghadimi and Wang, 2018; Chen et al., 2021; Shen and Chen, 2022), we have

$$\nabla\ell(y) = \nabla_y g(\tilde{x}^\star(y), y) - \nabla^2_{yx} f(\tilde{x}^\star(y), y)[\nabla^2_{xx} f(\tilde{x}^\star(y), y)]^{-1}\nabla_x g(\tilde{x}^\star(y), y)$$

and the solution of the inner problem is unique for any y. Thus $\tilde{x}^\star(y)$ is well-defined.

Example 4 (Stochastic bilevel optimization) *Suppose that we want to solve the problem* (8) *with access only to the noisy observations of the true gradients and Hessians. To employ two-time-scale SA* (2)*, we first give the definitions of* $F(x,y)$ *and* $G(x,y)$

$$F(x,y) = \nabla_x f(x,y),\ G(x,y) = \nabla_y g(x,y) - \nabla^2_{yx} f(x,y)[\nabla^2_{xx} f(x,y)]^{-1}\nabla_x g(x,y), \tag{9}$$

where $G(x,y)$ *is a surrogate of* $\nabla\ell(y)$ *by replacing* $\tilde{x}^\star(y)$ *with* x*. It follows that* $H(y) = \tilde{x}^\star(y)$ *and* $G(H(y),y) = \nabla\ell(y)$*. One can apply the following approximation method (Shen and Chen, 2022)*

$$\begin{aligned} x_{t+1} &= x_t - \alpha_t h^t_F = x_t - \alpha_t \left(F(x_t,y_t) + \xi_t\right), \\ y_{t+1} &= y_t - \beta_t h^t_G = y_t - \beta_t \left(G(x_t,y_t) + \psi_t\right), \end{aligned} \tag{10}$$

where h^t_F *and* h^t_G *are estimations of* $F(x_t,y_t)$ *and* $G(x_t,y_t)$*, respectively. For their explicit forms, please refer to Hong et al. (2023). Given the two-level structure inherent in bilevel optimization, it is convenient to contextualize this example within the framework of* (3)*, with the outer problem being the ultimate goal.*

3. Assumptions

In this section, we present the main assumptions required for our analysis.

Assumption 1 (Lipschitz conditions of F and H) *For any fixed* $y\in\mathbb{R}^{d_y}$*, there exists an operator* $H:\mathbb{R}^{d_y}\to\mathbb{R}^{d_x}$ *such that* $x = H(y)$ *is the unique solution of* $F(x,y) = 0$*.* H *and* F *satisfy that for* $\forall\, x\in\mathbb{R}^{d_x}, y_1, y_2\in\mathbb{R}^{d_y}$*,*

$$\|H(y_1) - H(y_2)\| \le L_H\|y_1 - y_2\|, \tag{11}$$

$$\|F(x,y_1) - F(H(y_1),y_1)\| \le L_F\|x - H(y_1)\|. \tag{12}$$

Condition (12) introduces a star-type Lipschitz condition, which is less restrictive and allows for greater flexibility in modeling and analysis.

Assumption 2 (Nested Lipschitz condition of G) *The operator G satisfies that $\forall\, x \in \mathbb{R}^{d_x}, y \in \mathbb{R}^{d_y}$,*

$$\|G(x,y) - G(H(y),y)\| \le L_{G,x}\|x - H(y)\|, \tag{13}$$
$$\|G(H(y),y) - G(H(y^\star),y^\star)\| \le L_{G,y}\|y - y^\star\|. \tag{14}$$

Here $y^\star$ is the unique solution to $G(H(y),y) = 0$.

The nested structure within Assumption 2 aligns with the conceptual framework that views a two-time-scale SA as an approximation of the formulation presented in (3). This type of nested Lipschitz condition is also adopted in Shen and Chen (2022).

Assumption 3 (Star-type strong monotonicity) *For any fixed $y \in \mathbb{R}^{d_y}$, $F(\cdot, y)$ is strongly monotone at $H(y)$, i.e., there exists a constant $\mu_F > 0$ such that*

$$\langle x - H(y), F(x,y) - F(H(y),y)\rangle \ge \mu_F\|x - H(y)\|^2, \quad \forall\, x \in \mathbb{R}^{d_x}. \tag{15}$$

$G(H(\cdot),\cdot)$ is strongly monotone at $y^\star$, i.e., there exists a constant $\mu_G > 0$ such that

$$\langle y - y^\star, G(H(y),y) - G(H(y^\star),y^\star)\rangle \ge \mu_G\|y - y^\star\|^2, \quad \forall\, y \in \mathbb{R}^{d_y}. \tag{16}$$

The star-type strong monotonicity assumption of F and G also appears in previous work (Doan, 2022; Shen and Chen, 2022). A direct consequence of the above assumptions is $L_F \ge \mu_F$ and $L_{G,y} \ge \mu_G$.

Assumption 4 (Uniform local linearity of H) *Assume that H is differentiable and there exist constants $S_H \ge 0$ and $\delta_H \in [0.5, 1]$ such that $\forall\, y_1, y_2 \in \mathbb{R}^{d_y}$,*

$$\|H(y_1) - H(y_2) - \nabla H(y_2)(y_1 - y_2)\| \le S_H\|y_1 - y_2\|^{1+\delta_H}.$$

An assumption closely related to Assumption 4 is the Hölder continuity of ∇H.

Assumption 4† (δ_H-Hölder continuity of ∇H) *We assume that H is differentiable and there exists constants $\widetilde{S}_H \ge 0$ and $\delta_H \in [0.5, 1]$ such that $\forall\, y_1, y_2 \in \mathbb{R}^{d_y}$,*

$$\|\nabla H(y_1) - \nabla H(y_2)\| \le \widetilde{S}_H\|y_1 - y_2\|^{\delta_H}. \tag{17}$$

Assumption 4† relaxes the requirement of Lipschitz continuity for ∇H used in Shen and Chen (2022). Assumptions 4 and 4† are equivalent, as shown in Proposition 1. Therefore, in this paper, we do not distinguish between these two assumptions. Proposition 1 extends Berger et al. (2020, Theorem 4.1, Euclidean case) from scalar-valued functions to vector-valued functions. The proof is provided in Appendix A.1.

Proposition 1 *Under Assumption 4†, Assumption 4 holds with* $S_H = \frac{\widetilde{S}_H}{1+\delta_H}$*; under Assumption 4, Assumption 4† holds with* $\widetilde{S}_H = 2^{1-\delta_H}\sqrt{1+\delta_H}\left(\frac{1+\delta_H}{\delta_H}\right)^{\frac{\delta_H}{2}} S_H$.[2] *For this equivalence, we do not require* $\delta_H \geq 0.5$.

A direct conclusion of Assumptions 4 and 1 is that $\forall\, y_1, y_2 \in \mathbb{R}^{d_y}$, with $R_H = \frac{2L_H}{S_H}$, $\|\nabla H(y_1)\| \leq L_H$ and

$$\|H(y_1) - H(y_2) - \nabla H(y_2)(y_1 - y_2)\| \leq S_H \|y_1 - y_2\| \cdot \min\left\{\|y_1 - y_2\|^{\delta_H}, R_H\right\}. \tag{18}$$

Although a finite-time decoupled convergence rate for nonlinear two-time-scale SA is not established in the literature, asymptotic decoupled convergence has been explored under a local linearity assumption (Mokkadem and Pelletier, 2006). Inspired by this work, we consider the following nested local linearity assumption.

Assumption 5 (Nested local linearity up to order $(1+\delta_F, 1+\delta_G)$**)** *There exist matrices* B_1, B_2, B_3 *with compatible dimensions, constants* $S_{B,F}, S_{B,G} \geq 0$ *and* $\delta_F, \delta_G \in (0, 1]$ *such that*

$$\|F(x,y) - B_1(x - H(y))\| \leq S_{B,F}\left(\|x - H(y)\|^{1+\delta_F} + \|y - y^\star\|^{1+\delta_F}\right), \tag{19}$$

$$\|G(x,y) - B_2(x - H(y)) - B_3(y - y^\star)\| \leq S_{B,G}\left(\|x - H(y)\|^{1+\delta_G} + \|y - y^\star\|^{1+\delta_G}\right). \tag{20}$$

This assumption follows the spirit that two-time-scale SA can be viewed as an approximation of the two-loop procedure (3). The parameters δ_F and δ_G quantify the order of errors in this linear approximation condition. To demonstrate how Assumption 5 can be guaranteed, we introduce the following standard local linearity assumption.

Assumption 5† (Standard local linearity up to order $(1+\delta_F, 1+\delta_G)$**)** *There exists matrices* $A_{11}, A_{12}, A_{21}, A_{22}$ *with compatible dimensions, constants* $S_{A,F}, S_{A,G} \geq 0$ *and* $\delta_F, \delta_G \in (0, 1]$ *such that*

$$\|F(x,y) - A_{11}(x - x^\star) - A_{12}(y - y^\star)\| \leq S_{A,F}\left(\|x - x^\star\|^{1+\delta_F} + \|y - y^\star\|^{1+\delta_F}\right), \tag{21}$$

$$\|G(x,y) - A_{21}(x - x^\star) - A_{22}(y - y^\star)\| \leq S_{A,G}\left(\|x - x^\star\|^{1+\delta_G} + \|y - y^\star\|^{1+\delta_G}\right). \tag{22}$$

Assumption 5† implies both F and G are differentiable at $(x^\star, y^\star)$, leading to $A_{11} = \nabla_x F(x^\star, y^\star)$, $A_{12} = \nabla_y F(x^\star, y^\star)$, $A_{21} = \nabla_x G(x^\star, y^\star)$, $A_{22} = \nabla_y G(x^\star, y^\star)$. Notably, Assumption 5† with $\delta_F = \delta_G = 1$ is the local linearity assumption in Mokkadem and Pelletier (2006). The following proposition provides some properties of the existing assumptions. In particular, it shows that if the operators F and G can be locally approximated around the root $(x^\star, y^\star)$ by linear functions of $x - x^\star$ and $y - y^\star$, up to a higher-order error, then the nested local linearity condition follows naturally—assuming certain prior assumptions are satisfied. The proof is provided in Appendix A.2.

Proposition 2 *Suppose that Assumptions 1–4 hold.*

2. We make the contention that $\left(\frac{1+\delta_H}{\delta_H}\right)^{\frac{\delta_H}{2}} = 1$ if $\delta_H = 0$.

(i) If Assumption 5† holds, then $\|A_{11}\| \le L_F$, $\|A_{21}\| \le L_{G,x}$, $\frac{A_{11}+A_{11}^\top}{2} \succeq \mu_F I$, *and* $\nabla H(y^\star) = -A_{11}^{-1}A_{12}$. *If we further assume* $\delta_H \ge \delta_F \vee \delta_G$, *then Assumption 5 holds with parameters*

$$B_1 = A_{11}, \quad B_2 = A_{21}, \quad B_3 = A_{22} - A_{21}A_{11}^{-1}A_{12},$$
$$S_{B,F} = S_{A,F} + L_F(S_H \vee 2L_H), \ S_{B,G} = S_{A,G} + L_{G,x}(S_H \vee 2L_H).$$

(ii) If Assumption 5 holds, then $\|B_1\| \le L_F, \|B_2\| \le L_{G,x}, \|B_3\| \le L_{G,y}$, $\frac{B_1+B_1^\top}{2} \succeq \mu_F I$, *and* $\frac{B_3+B_3^\top}{2} \succeq \mu_G I$.

The inclusion of the condition $\delta_H \ge \delta_F \vee \delta_G$ in Proposition 2 (i) stems from the nested structure in Assumption 5. This structure is introduced to capture the nested nature of the two-loop procedure in (3), where H is a crucial link between the inner and outer loops. Transforming Assumption 5† into a nested form necessitates a more refined smoothness condition on H to preserve local linearity. Additionally, B_3 equals the Schur complement of A_{11} in the matrix $(A_{11}, A_{12}; A_{21}, A_{22})$.

Moreover, if we allow H to be non-smooth or non-differentiable, Assumption 5 can cover a broader class of problems in which the nonlinearity is "absorbed" into the solution map H, whereas Assumption 5† cannot. Typical examples include cases where H is a projection-type map (e.g., under simplex or box constraints) or a proximal map (e.g., soft-thresholding for ℓ_1 regularization). Although these scenarios are beyond the scope of this paper, we believe that our nested local linearity condition is more consistent with the perspective in (3) and may inspire future research.

For the examples in Section 2, Assumptions 1–5 can be verified under standard regularity conditions, such as the strong convexity of the objective function and the Lipschitz continuity of the Hessian. Details on the verification of the local linearity conditions are deferred to Appendix A.3.

Next, we focus on noise models. We denote by $\mathcal{F}_t$ the filtration containing all the history generated by (2) before time t, i.e., $\mathcal{F}_t = \sigma\{x_0, y_0, \xi_0, \psi_0, \xi_1, \psi_1, \ldots, \xi_{t-1}, \psi_{t-1}\}$.

Assumption 6 (Conditions on the noise) *The sequences of random variables* $\{\xi_t\}_{t=0}^\infty$ *and* $\{\psi_t\}_{t=0}^\infty$ *are martingale difference sequences satisfying*

$$\begin{aligned} &\mathbb{E}[\xi_t \,|\, \mathcal{F}_t] = 0, \quad \mathbb{E}[\psi_t \,|\, \mathcal{F}_t] = 0, \quad \mathbb{E}[\|\xi_t\|^4 | \mathcal{F}_t] \le \Gamma_{11}^2, \quad \mathbb{E}[\|\psi_t\|^4 | \mathcal{F}_t] \le \Gamma_{22}^2, \\ &\|\mathbb{E}[\xi_t \xi_t^\top \,|\, \mathcal{F}_t]\| \le \Sigma_{11}, \quad \|\mathbb{E}[\psi_t \psi_t^\top \,|\, \mathcal{F}_t]\| \le \Sigma_{22}, \quad \|\mathbb{E}[\xi_t \psi_t^\top \,|\, \mathcal{F}_t]\| \le \Sigma_{12}. \end{aligned} \tag{23}$$

In Assumption 6, the boundedness of the fourth-order moments is imposed to control the higher-order error terms introduced by Assumption 5.

The final assumption is the requirements for the step sizes.

Assumption 7 (Conditions on step sizes) *With the constants* μ_F, μ_G, δ_F *and* δ_G *defined in Assumptions 3 and 5, the step sizes* $\{\alpha_t\}_{t=0}^\infty$ *and* $\{\beta_t\}_{t=0}^\infty$ *satisfy the following conditions:*

(i) Constant bounds: $\alpha_t \le \iota_1, \beta_t \le \iota_2, \frac{\beta_t}{\alpha_t} \le \kappa, \frac{\beta_t^2}{\alpha_t} \le \rho$, *where* ι_1, ι_2, κ *and* ρ *are problem-dependent constants with specific forms defined in* (49) *in Appendix A.*

(ii) Growth conditions: $1 \le \frac{\alpha_{t-1}}{\alpha_t} \le 1 + (\frac{\delta_F \mu_F}{16}\alpha_t) \wedge (\frac{\delta_F \mu_G}{16}\beta_t) \wedge (\frac{\delta_G \mu_G}{8}\beta_t)$ *and* $1 \le \frac{\beta_{t-1}}{\beta_t} \le 1 + \frac{\mu_G}{64}\beta_t$ *for any* $t \ge 1$.

(iii) $\frac{\beta_t}{\alpha_t}$ *is non-increasing in* t, *and* $\prod_{\tau=0}^{t}\left(1 - \frac{\mu_G \beta_\tau}{4}\right) = \mathcal{O}(\alpha_t^2)$.

Conditions (i) and (ii) are similar to Kaledin et al. (2020, Assumption A2), though our conditions are more intricate to handle non-linearity. Condition (iii) is a technical requirement introduced to simplify the proof. Notably, these conditions are naturally satisfied as long as both α_t and β_t decrease with t, with β_t decreasing faster than α_t. Additionally, our setup includes single-time-scale SA as a special case if a constant ratio $\beta_t/\alpha_t \equiv \kappa' \le \kappa$ that satisfies Assumption 7 is adopted.

Remark 1 (Discussion on Assumption 7) *We make the following remarks on Assumption 7.*

- *The growth conditions imply that* α_t *and* β_t *are non-increasing. Moreover,* $\alpha_t^{-1} \le \alpha_{t-1}^{-1} + \frac{\delta_F \mu_F}{16}\alpha_t/\alpha_{t-1} \le \alpha_{t-1}^{-1} + \frac{\delta_F \mu_F}{16} = \mathcal{O}(t)$ *and hence* $\alpha_t = \Omega(t^{-1})$. *Similarly,* $\beta_t = \mathcal{O}(t^{-1})$. *Consequently,* $\sum_{t=0}^{\infty}\alpha_t = \infty$ *and* $\sum_{t=0}^{\infty}\beta_t = \infty$, *which is a standard condition in the study of SA to ensure convergence (Borkar, 2009).*

- *This assumption is formulated for a broad class of operators so as to cover all examples satisfying our operator assumptions. As a consequence, the associated step-size conditions are intentionally conservative, and the constants appearing therein (e.g.,* $1/64$*) are chosen mainly for technical convenience in the proofs. If one restricts attention to a narrower problem class or to a specific example, these constants can often be significantly improved.*

- *As indicated by Kaledin et al. (2020), this assumption encompasses diminishing, piecewise constant, and constant step size schedules. Here we focus on diminishing step sizes, which are a standard and widely used choice. A typical example is* $\alpha_t = \Theta(t^{-a})$ *and* $\beta_t = \Theta(t^{-b})$ *with* $0 < a \le b < 1$. *In practice, it is sufficient to impose Assumption 7 only for* $t \ge t_0$, *where* t_0 *is a prescribed integer, since the early stage of optimization is usually dominated by transient behavior. For the purpose of establishing finite-time convergence rates, it is therefore not necessary to require the assumption to hold from the start. In particular, when* $a < b$, *corresponding to the strict two-time-scale regime, all the constant bounds in (i) are of order* $o(1)$ *as* $t_0 \to \infty$. *Treating these terms as* $o(1)$ *can greatly simplify the analysis of the constants; see Remark 4.*

4. Theoretical Analysis

In Section 4.1, we establish upper bounds for $\mathbb{E}\|y_t - y^\star\|^2$ and $\mathbb{E}\|x_t - H(y_t)\|^2$, and prove finite-time decoupled convergence under the local linearity condition. To demonstrate the necessity of local linearity for decoupled convergence, we also construct an example in which local linearity fails and derive a corresponding lower bound in Section 4.2.

4.1 Upper bounds and Decoupled Convergence

To analyze the convergence rate of (x_t, y_t) to $(x^\star, y^\star)$, it is common to examine the mean square errors $\mathbb{E}[\|x_t - x^\star\|^2]$ and $\mathbb{E}[\|y_t - y^\star\|^2]$. Recall that two-time-scale SA can be viewed as an approximation of the two-loop procedure in (3). It is thus more fundamental to consider the following residual variables (Doan, 2022):

$$\hat{x}_t = x_t - H(y_t), \quad \hat{y}_t = y_t - y^\star. \tag{24}$$

Here, $\hat{x}_t$ and $\hat{y}_t$ represent the errors of the inner and outer loops in (3), respectively. Furthermore, given the Lipschitz continuity of H in Assumption 1, we can bound $\|x_t - x^\star\|$ as $\|x_t - x^\star\| \le \|\hat{x}_t\| + L_H\|\hat{y}_t\|$. Thus, it suffices to focus on $(\hat{x}_t, \hat{y}_t)$.

Our main results, based on the assumptions in Section 3, are stated as follows.

Theorem 3 *Suppose that Assumptions 1–7 hold. Then we have for all $t \ge 0$,*

$$\mathbb{E}\|\hat{x}_{t+1}\|^2 \le \mathcal{C}_x\, \alpha_t, \tag{25}$$

$$\|\mathbb{E}\hat{x}_{t+1}\hat{y}_{t+1}^\top\| \le \mathcal{C}_{xy,1}\, \beta_t + \mathcal{C}_{xy,2}\, \alpha_t\beta_t \left(\frac{\alpha_t}{\beta_t}\right)^{\frac{2}{\delta_F}}, \tag{26}$$

$$\mathbb{E}\|\hat{y}_{t+1}\|^2 \le \mathcal{C}_{y,1}\, \beta_t + \mathcal{C}_{y,2}\, \alpha_t\beta_t \left(\frac{\alpha_t}{\beta_t}\right)^{\frac{2}{\delta_F}} + \mathcal{C}_{y,3}\, \alpha_t\beta_t \left(\frac{\alpha_t}{\beta_t}\right)^{\frac{1}{\delta_G}}. \tag{27}$$

The exact constants are given in (86) *in Appendix C.*

In Theorem 3, observe that the step size α_t of the fast iterate also influences the convergence rate of the slow iterate through terms involving the parameters δ_F and δ_G. In the following discussion, we focus on polynomially diminishing step sizes and examine the conditions required for achieving decoupled convergence.

Corollary 4 (Decoupled convergence rates) *Under the same setting of Theorem 3, if we use polynomially diminishing step sizes $\alpha_t = \frac{\alpha_0}{(t+T_0)^a}$ and $\beta_t = \frac{\beta_0}{(t+T_0)^b}$ with $a, b \in (0, 1]$, $1 \le \frac{b}{a} \le 1 + \frac{\delta_F}{2} \wedge \delta_G$ and properly chosen α_0, β_0 and T_0, then we have*

$$\mathbb{E}\|\hat{x}_t\|^2 = \mathcal{O}(\alpha_t), \quad \|\mathbb{E}\hat{x}_t\hat{y}_t^\top\| = \mathcal{O}(\beta_t), \quad \textit{and} \quad \mathbb{E}\|\hat{y}_t\|^2 = \mathcal{O}(\beta_t).$$

For an example choice of the constants, please refer to (87) *in Appendix C.*

Remark 2 (Comparison with previous work) *Note that $x_t - x^\star = \hat{x}_t + H(y_t) - H(y^\star)$. Theorem 4 and the Lipschitz continuity of H, we have $\mathbb{E}\|x_t - x^\star\|^2 = \mathcal{O}(\alpha_t)$ and $\|\mathbb{E}(x_t - x^\star)(y_t - y^\star)^\top\| = \mathcal{O}(\beta_t)$. This result is consistent with the central limit theorem established in Mokkadem and Pelletier (2006):*

$$\begin{pmatrix} \alpha_t^{-1/2}(x_t - x^\star) \\ \beta_t^{-1/2}(y_t - y^\star) \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(0, \begin{pmatrix} \Sigma_x & 0 \\ 0 & \Sigma_y \end{pmatrix}\right). \tag{28}$$

Moreover, our analysis provides a more refined characterization of the matrix cross term, as (28) *only implies $(x_t - x^\star)(y_t - y^\star)^\top = o(\sqrt{\alpha_t\beta_t})$ in probability.*

To establish the convergence rate in (27)*, the convergence rate of the matrix cross term in* (26) *is an essential intermediate result. Neither decoupled convergence for the strict two-time-scale case nor an analysis of the matrix cross term is present in prior work on general nonlinear cases (Shen and Chen, 2022; Doan, 2022).*

Remark 3 (Step size selection for the optimal convergence rate) *To achieve the optimal convergence rate of the slow iterate* $\mathbb{E}\|\hat{y}_t\|^2 = \mathcal{O}(1/t)$*, we could choose* $\beta_t \sim \beta_0 t^{-1}$ *and* $\alpha_t \sim \alpha_0 t^{-a}$ *with* $(1+\frac{\delta_F}{2} \wedge \delta_G)^{-1} \leq a \leq 1$*. In particular, when* $\delta_F = \delta_G = 1$*, the feasible range for* a *in* $\alpha_t \sim \alpha_0 t^{-a}$ *is* $2/3 \leq a \leq 1$*. Our results ensure that achieving* $\mathcal{O}(1/t)$ *convergence for the slow iterate allows greater flexibility in the step size selection for the fast iterate, extending beyond the single-time-scale case considered in Shen and Chen (2022).*

Remark 4 (Leading terms in the constants) *The complete expressions of the constants in* (86) *are fairly complicated and difficult to analyze directly. To simplify the analysis and isolate the most essential parameter dependence, we focus on diminishing step sizes of the form* $\alpha_t = \Theta(t^{-a})$ *and* $\beta_t = \Theta(t^{-b})$*, where* $a, b \in (0, 1]$ *and* $1 < \frac{b}{a} < 1 + \frac{\delta_F}{2} \wedge \delta_G$*. Compared with the requirement in Corollary 4, we additionally require the inequality to be strict. The condition* $\frac{b}{a} < 1 + \frac{\delta_F}{2} \wedge \delta_G$ *ensures that the leading terms in* (25)–(27) *are all given by the first terms*

$$\mathbb{E}\|\hat{x}_{t+1}\|^2 \leq \mathcal{C}_x\, \alpha_t, \quad \|\mathbb{E}\hat{x}_{t+1}\hat{y}_{t+1}^\top\| \leq \mathcal{C}_{xy,1}\, \beta_t + o(\beta_t), \quad \mathbb{E}\|\hat{y}_{t+1}\|^2 \leq \mathcal{C}_{y,1}\, \beta_t + o(\beta_t).$$

Consequently we only need to focus on the constants $\mathcal{C}_x$*,* $\mathcal{C}_{xy,1}$*, and* $\mathcal{C}_{y,1}$*. We also require* $b > a$*, corresponding to the "strict" two-time-scale case, in order to highlight the role of decoupled convergence, since decoupled convergence is trivial when* $a = b$*.*

As discussed in Remark 1, it suffices to focus on $t \geq t_0$ *for a prescribed* t_0*. Under strict two-time-scale diminishing step sizes, the constants* $\iota_1, \iota_2, \kappa, \rho$ *in Assumption 7 are all of order* $o(1)$ *as* $t_0 \to \infty$*. Therefore, in the expressions of* $\mathcal{C}_x$*,* $\mathcal{C}_{xy,1}$*, and* $\mathcal{C}_{y,1}$ *in* (86)*, all terms involving these constants can be treated as higher-order infinitesimals. In particular, the leading terms in* $\mathcal{C}_x$*,* $\mathcal{C}_{xy,1}$*, and* $\mathcal{C}_{y,1}$ *can be summarized as*

$$\begin{aligned} \mathcal{C}_x &\lesssim \frac{\Gamma_{11}}{\mu_F} + o(1), \quad \mathcal{C}_{xy,1} \lesssim \frac{\Sigma_{12}}{\mu_F} + \frac{L_{G,x}\Gamma_{11}}{\mu_F^2} + o(1), \\ \mathcal{C}_{y,1} &\lesssim \frac{\Gamma_{22}}{\mu_G} + \frac{d_y L_{G,x}\Sigma_{12}}{\mu_F \mu_G} + \frac{d_y L_{G,x}^2 \Gamma_{11}}{\mu_F^2 \mu_G} + o(1). \end{aligned} \tag{29}$$

For the detailed derivation, see Appendix C.10.

Meanwhile, from the CLT in (28)*, we can obtain* $\lim_{t\to\infty} \alpha_t^{-1}\mathbb{E}\|x_t\|^2 = \mathrm{tr}(\Sigma_x)$ *and* $\lim_{t\to\infty} \beta_t^{-1}\mathbb{E}\|y_t\|^2 = \mathrm{tr}(\Sigma_y)$ *under additional regular conditions. By analyzing the detailed expression of* Σ_x *and* Σ_y*, we obtain*

$$\mathrm{tr}(\Sigma_x) \leq \frac{\Gamma_{11}}{2\mu_F}, \quad \mathrm{tr}(\Sigma_y) \leq \frac{1}{\mu_G}\left[\Gamma_{22} + 2d_y \frac{L_{G,x}}{\mu_F}\Sigma_{12} + \Big(\frac{L_{G,x}}{\mu_F}\Big)^2 \Gamma_{11}\right]. \tag{30}$$

Moreover, let $\Sigma_{x,y} = \lim_{t\to\infty} \beta_t^{-1}\mathbb{E}[x_t y_t^\top]$*. Then* $\lim_{t\to\infty} \beta_t^{-1}\|\mathbb{E}x_t y_t^\top\| = \|\Sigma_{x,y}\|$*. Konda and Tsitsiklis (2004, Theorem 2.6) provide a characterization of* $\Sigma_{x,y}$ *in the linear case, from which we obtain*

$$\|\Sigma_{x,y}\| \leq \frac{1}{\mu_F}\left(\Sigma_{12} + \frac{L_{G,x}\Gamma_{11}}{2\mu_F}\right). \tag{31}$$

The detailed derivations of (30) *and* (31) *are also deferred to Appendix C.10.*

We now compare the above results. The upper bounds in (29) *come from the non-asymptotic analysis. Accordingly, we omit parameter-independent constants, since the primary goal of the non-asymptotic analysis is to establish the convergence rates rather than to optimize the constants. By contrast, the upper bounds in* (30) *and* (31) *are derived from the asymptotic analysis and may be viewed as nearly tight upper bounds. We find that all parameter dependencies match except for the third term in the expression of* $\mathcal{C}_{y,1}$.

Dimension dependence*. In* (29)*, the third term involving* Γ_{11} *in the upper bound of* $\mathcal{C}_{y,1}$ *depends on the dimension, whereas in* (30)*, the third term involving* Γ_{11} *in the upper bound of* $\mathrm{tr}(\Sigma_y)$ *does not exhibit such dimensional dependence. By contrast, the dimensional dependence in the term involving* Σ_{12} *is unavoidable. Recall that Assumption 6 implies* $\mathbb{E}[\|\xi_t\|^2|\mathcal{F}_t] \le \Gamma_{11}$, $\mathbb{E}[\|\psi_t\|^2|\mathcal{F}_t] \le \Gamma_{22}$, *and* $\|\mathbb{E}[\xi_t\psi_t^\top \,|\, \mathcal{F}_t]\| \le \Sigma_{12}$. *This discrepancy comes from the fact that* Σ_{12} *is an upper bound on the operator norm of the cross-covariance matrix, whereas* $\mathbb{E}[\|\xi_t\|^4|\mathcal{F}_t] \le \Gamma_{11}^2$ *implies* $\mathrm{tr}(\mathbb{E}[\xi_t\xi_t^\top|\mathcal{F}_t]) = \mathbb{E}[\|\xi_t\|^2|\mathcal{F}_t] \le \Gamma_{11}$, *so that* Γ_{11} *serves as an upper bound on the trace of the covariance matrix. Clearly, an upper bound on the trace is also an upper bound on the operator norm; however, there may be a dimensional gap between the trace and the operator norm.*

We believe that the third term of $\mathcal{C}_{y,1}$ *in* (29) *can be improved to* $\frac{d_y L_{G,x}^2 \Sigma_{11}}{\mu_F^2 \mu_G}$ *by carrying out a more refined analysis of* $\mathbb{E}\hat{x}_t\hat{x}_t^\top$, *rather than* $\mathbb{E}\|\hat{x}_t\|^2$, *under a tighter upper bound* $\|\mathbb{E}[\xi_t\xi_t^\top|\mathcal{F}_t]\| \le \Sigma_{11}$. *However, the dimensional dependence may still be unavoidable under the current proof framework. To illustrate this point, consider the special case in which all eigenvalues of* $\mathbb{E}[\xi_t\xi_t^\top|\mathcal{F}_t]$ *are equal and* $d_x = d_y$. *In this case, we have* $\Gamma_{11} = d_x\Sigma_{11} = d_y\Sigma_{11}$. *Such dimensional dependence also appears in the finite-time analysis of the linear case (Kaledin et al., 2020).*

Inverse dependence on the strong monotonicity parameters*. Recall that* μ_F *and* μ_G *are the strong monotonicity parameters of the inner operator* $F(\cdot, y)$ *and the outer operator* $G(H(\cdot),\cdot)$, *respectively. The leading term in the upper bound for* $\mathbb{E}\|\hat{x}_t\|^2$ *scales as* μ_F^{-1}, *and the leading term in the upper bound for* $\mathbb{E}\|\hat{y}_t\|^2$ *scales as* μ_G^{-1}. *This is consistent with the classical behavior of SGD for strongly convex optimization, where the leading constant is proportional to the inverse of the strong convexity parameter; see, for example, Moulines and Bach (2011, Theorem 1).*

Amplification factor $L_{G,x}/\mu_F$. *In addition to the fact that the convergence rates for* $\mathbb{E}\|\hat{x}_t\|^2$ *and* $\mathbb{E}\|\hat{y}_t\|^2$ *depend only on their respective step sizes, the leading constant term for* $\mathbb{E}\|\hat{x}_t\|^2$ *is also almost unaffected by the noise in the slow-time-scale update. By contrast, the leading constant term for* $\mathbb{E}\|\hat{y}_t\|^2$ *is affected by both the fast- and slow-time-scale updates. The effect of the cross-covariance between the fast- and slow-time-scale noises, denoted by* Σ_{12}, *is amplified by a factor of* $L_{G,x}/\mu_F$, *and the effect of the covariance of the fast-time-scale noise, denoted by* Γ_{11}, *is amplified by a factor of* $(L_{G,x}/\mu_F)^2$.

This amplification can be understood from the asymptotic viewpoint in Han et al. (2024, Remark 4.1): the slow iterate behaves asymptotically like a standard SA iterate for the operator $G(H(\cdot),\cdot)$ *with a modified noise term* $\breve{\psi}_t = \psi_t - B_2B_1^{-1}\xi_t$. *By Proposition 2, the ratio* $L_{G,x}/\mu_F$ *upper bounds* $\|B_2B_1^{-1}\|$. *The amplification effect therefore comes from controlling the covariance of* $\breve{\psi}_t$.

Behavior of the matrix cross term. *The leading constant term in the bound for* $\|\mathbb{E}\hat{x}_t\hat{y}_t^\top\|$ *is proportional to* μ_F^{-1}*. This suggests that the behavior of* $\hat{x}_t\hat{y}_t^\top$ *is governed primarily by the strong monotonicity of the inner operator* $F(\cdot, y)$*, rather than that of the outer operator* $G(H(\cdot),\cdot)$*. The key reason is the separation of step sizes, namely* $\alpha_t \gg \beta_t$*. Moreover, the leading term depends on both* Σ_{12} *and* Γ_{11}*, with the effect of* Γ_{11} *amplified by a factor of* $L_{G,x}/\mu_F$*. Under the definition of the modified noise term* $\breve{\psi}_t = \psi_t - B_2B_1^{-1}\xi_t$*, this amplification arises from controlling the cross-covariance between* ξ_t *and* $\breve{\psi}_t$*.*

4.2 A Lower Bound without Local Linearity

As shown in the previous subsection, decoupled convergence can be achieved under the local linearity condition with appropriately chosen step sizes. This naturally raises the following question: *Is local linearity essential for decoupled convergence?* In this subsection, we answer this question in the affirmative and show that the local linearity condition in Assumption 5 is necessary for decoupled convergence.

To this end, we construct the following example in which the local linearity condition fails.

Example 5 *Consider the following nonlinear SA problem with the operators* $F, G\colon \mathbb{R}\times\mathbb{R}\to\mathbb{R}$ *given by*

$$F(x,y) = x - y,\ G(x,y) = -|x-y|\cdot\operatorname{sign}(y) + y, \tag{32}$$

where $\operatorname{sign}(x) = 1_{x>0} - 1_{x<0}$ *is the sign function.*

In this example, $G(x,y)$ involves both the sign and absolute value functions, and therefore does not satisfy the local linearity condition in Assumption 5. However, since $F(x,y) = x-y$ is linear, the induced solution map $H(y) = y$ is also linear. Consequently, the reduced operator $G(H(y),y) = y$ is linear as well. That is, the nonlinearity appears only in $G(x,y)$ before substituting $x = H(y)$. Meanwhile, this example satisfies Assumptions 1–3 with all corresponding parameters equal to 1, as well as Assumption 4 with $S_H = 0$. The unique root is $(x^\star, y^\star) = (0,0)$. Applying the two-time-scale SA algorithm (2) to this example, we obtain the following lower bound, whose proof is deferred to Appendix D.

Proposition 5 (Lower bound for Example 5) *Suppose that: (a) the noise terms satisfy* $\psi_t = 0$*, and the* ξ_t *are i.i.d. with* $\mathbb{E}\xi_t = 0$ *and* $\mathbb{E}\xi_t^2 = \Sigma_1 > 0$*; (b) the step sizes satisfy* $\beta_t/\alpha_t \to 0$ *and Assumption 7 with the parameters* δ_F *and* δ_G *in (ii) replaced by* 1*;*[3] *(c) the initialization satisfies* $y_0 \neq y^\star$*. Then we have* $\mathbb{E}|\hat{x}_t|^2 = \Omega(\alpha_t)$ *and* $\mathbb{E}|\hat{y}_t|^2 = \Omega(\alpha_t)$*.*

The condition on the noise is imposed to simplify the analysis. The requirement $\beta_t/\alpha_t \to 0$ on the step sizes restricts attention to the strict two-time-scale regime, since in the single-time-scale regime, where $\beta_t = \Theta(\alpha_t)$, decoupled convergence is trivial.

Proposition 5 shows that, without local linearity of $G(x,y)$, the convergence rate on the slow time scale is indeed degraded by the larger step size associated with the fast-time-scale update, even though $F(x,y)$, $H(y)$, and $G(H(y),y)$ are all linear. This complements the approximation perspective in (3):

3. For $d_x = d_y = 1$, after this replacement, Assumption 7 reduces to the weaker version, Assumption 7†, introduced in Section 5.3.

Although the two iterates in two-time-scale SA can be interpreted as solving $F(x, y) = 0$ (for fixed y) and $G(H(y), y) = 0$, the detailed form of $G(x, y)$ before substituting $x = H(y)$ still affects the convergence rates.

Moreover, in Proposition 5, the slow-time-scale update is deterministic. This indicates that the main obstacle to decoupled convergence on the slow time scale is the interdependence between the two time scales, rather than the noise in the slow-time-scale update.

This observation also has a possible algorithmic implication. Consider $\widetilde{G}_\alpha(x, y) = \alpha(x - y) + y$ for $\alpha \in \mathbb{R}$. Then $\widetilde{G}_\alpha(H(y), y) = G(H(y), y)$. In other words, the linear operator $\widetilde{G}_\alpha(x, y)$ yields the same outer operator $\widetilde{G}_\alpha(H(y), y)$ as in Example 5. This suggests that, when there are multiple possible choices of the operators F and G leading to the same reduced operator $G(H(\cdot), \cdot)$, it is preferable to choose linear or nearly linear operators in order to ensure a faster convergence rate.

5. Proof Framework for Our Main Theorem

In this section, we outline the framework of our proof for Theorem 3. We first present the high-level idea and comparison with prior works, with the detailed procedure developed in Sections 5.1–5.4.

High-level idea. Our main technical contribution is a systematic framework for handling the cross term $\|\mathbb{E}\hat{x}_t\hat{y}_t^\top\|$ in the nonlinear setting. It's crucial for the sharp convergence characterization of the interacting sequences $\{x_t\}_{t=0}^\infty$ and $\{y_t\}_{t=0}^\infty$. While similar analysis exists for the linear case (e.g., Kaledin et al. 2020), the nonlinear case is harder due to the nonlinearity of F, G, and H. We approximate these mappings by their linear parts, which introduces residual errors. The key challenge is that the cross term's dynamics intertwine with these (higher-order) residuals. Our framework systematically tracks this interaction and controls the error accumulation via a fourth-moment analysis, showing the residuals are indeed higher-order.

The whole proof is organized into the following four steps:

Step 1: Derive Convergence Rates without Local Linearity of F and G

Step 2: Introduce the Matrix Cross Term and Derive Refined One-Step Descent Lemmas

Step 3: Analyze the Convergence Rates of Fourth-Order Moments

Step 4: Integrate the Above Ingredients and Derive Decoupled Convergence Rates

Comparison with prior works. Overall, the nonlinear setting requires several additional components beyond those in Doan (2022) and Kaledin et al. (2020), even though a few individual steps are similar in spirit to these earlier analyses. The first step is close in spirit to Doan (2022), but it only provides a preliminary convergence rate and serves as a starting point for our later analysis. The second step is partly inspired by Kaledin et al. (2020), where we introduce the matrix cross term $\|\mathbb{E}\hat{x}_t\hat{y}_t^\top\|$ to refine the one-step descent analysis for the slow iterate; however, in the nonlinear setting, the local linear approximations of F, G, and H also produce higher-order error terms, which require additional control arguments. The third step, namely the convergence analysis of fourth-order moments, does not appear in either Doan (2022) or Kaledin et al. (2020), and is needed to handle the higher-order terms

caused by nonlinearity. Finally, the last step integrates all these ingredients to establish decoupled convergence, while simultaneously handling the nonlinear remainder terms and the matrix cross term.

5.1 Step 1: Derive Convergence Rates without Local Linearity of F and G.

We first establish a coarse convergence rate without local linearity as a starting point. To this end, we first present the one-step descent lemmas for the squared errors The analysis in this step does not require the local linearity of F and G in Assumption 5, nor does it require the bounded fourth-order moments in Assumption 6. The requirement on the step sizes in Assumption 7 can also be weakened. Thus, the results in this step is based on the weaker version of Assumptions 6 and 7, denoted as Assumptions 8 and 9, with the details deferred to Appendix B.

Lemma 6 (One-step descent of the fast iterate $\hat{x}_t$) *Suppose that Assumptions 1–4 and 8–9 hold. For any $t \geq 0$, we have*

$$\begin{aligned}\mathbb{E}\left[\|\hat{x}_{t+1}\|^2 \,|\, \mathcal{F}_t\right] \leq (1-\mu_F\alpha_t)\,\|\hat{x}_t\|^2 + c_{x,1}\beta_t^2\|\hat{y}_t\|^2 + 2\Gamma_{11}\alpha_t^2 \\ + c_{x,2}\beta_t\sqrt{1-\alpha_t\mu_F}\|\hat{x}_t\|\|\hat{y}_t\| + c_{x,3}\beta_t^2 + c_{x,4}\frac{\beta_t^{2+2\delta_H}}{\alpha_t},\end{aligned} \tag{33}$$

where $\{c_{x,i}\}_{i\in[4]}$ are problem-dependent constants defined in (59).

Lemma 7 (One-step descent of the slow iterate $\hat{y}_t$) *Suppose that Assumptions 1–4 and 8–9 hold. For any $t \geq 0$, we have*

$$\begin{aligned}\mathbb{E}\left[\|\hat{y}_{t+1}\|^2 \,|\, \mathcal{F}_t\right] \leq (1-\mu_G\beta_t)\,\|\hat{y}_t\|^2 + L_{G,x}^2\beta_t^2\|\hat{x}_t\|^2 \\ + 2L_{G,x}\beta_t\sqrt{1-\mu_G\beta_t}\|\hat{x}_t\|\|\hat{y}_t\| + \Gamma_{22}\beta_t^2.\end{aligned} \tag{34}$$

By combining Lemmas 6 and 7 and using carefully designed Lyapunov functions, we can achieve the following convergence results.

Theorem 8 (Convergence rates without local linearity of F and G) *Suppose that Assumptions 1–4 and 8–9 hold. Then we have*

$$\begin{aligned}\mathbb{E}\|\hat{x}_{t+1}\|^2 \leq \prod_{\tau=0}^{t}\left(1-\frac{\mu_G\beta_\tau}{4}\right)\left(3\mathbb{E}\|\hat{x}_0\|^2 + \frac{7L_HL_{G,y}\mathbb{E}\|\hat{y}_0\|^2}{L_{G,x}}\right) + \frac{8\Gamma_{11}}{\mu_F}\alpha_t \\ + c_{x,5}\beta_t + c_{x,6}\frac{\beta_t^2}{\alpha_t} + c_{x,7}\frac{\beta_t^{2+2\delta_H}}{\alpha_t^2}, \\ \mathbb{E}\|\hat{y}_{t+1}\|^2 \leq \prod_{\tau=0}^{t}\left(1-\frac{\mu_G\beta_\tau}{4}\right)\left(\mathbb{E}\|\hat{y}_0\|^2 + \frac{2L_{G,x}\mathbb{E}\|\hat{x}_0\|^2}{7L_HL_{G,y}}\right) + \frac{128L_{G,x}^2\Gamma_{11}}{\mu_F\mu_G^2}\alpha_t \\ + c_{y,1}\beta_t + c_{y,2}\frac{\beta_t^{2+2\delta_H}}{\alpha_t^2},\end{aligned}$$

where $\{c_{x,i}\}_{i\in[7]\setminus[4]}$ and $\{c_{y,i}\}_{i\in[2]}$ are problem-dependent constants defined in (75) *and* (76). *In particular, when $\delta_H \geq 0.5$,*

$$\mathbb{E}\|\hat{x}_{t+1}\|^2 + \mathbb{E}\|\hat{y}_{t+1}\|^2 = \mathcal{O}(\alpha_t). \tag{35}$$

The details proofs of Lemmas 6, 7 and Theorem 8 are given in Appendix B. Although the results in this subsection follow Doan (2022) at a high level, our analysis imposes refined conditions (e.g., Assumption 2.4), leading to explicit dependence on δ_H, whereas Doan's result corresponds to the special case $\delta_H = 0$. As a result, the proof must be reworked despite the similar overall idea.

5.2 Step 2: Introduce the Matrix Cross Term and Derive Refined One-Step Descent Lemmas

With Assumption 5, we could replace $\mathbb{E}\|\hat{x}_t\|\|\hat{y}_t\|$ in Lemma 6 and 7 with the matrix cross term $\|\mathbb{E}\hat{x}_t\hat{y}_t^\top\|$ at the cost of introducing the higher-order residual terms. We also need to derive the one-step descent lemma of the matrix cross term $\|\mathbb{E}\hat{x}_t\hat{y}_t^\top\|$. To simplify the notation, we define

$$\mathcal{Z}_{t,\delta} := \mathbb{E}\|\hat{x}_t\|^\delta + \mathbb{E}\|\hat{y}_t\|^\delta.$$

Lemma 9 (Refined one-step descent of the fast iterate $\hat{x}_t$) *Suppose Assumptions 1–7 hold. We have for any $t \geq 0$,*

$$\begin{aligned}\mathbb{E}\|\hat{x}_{t+1}\|^2 \leq (1-\mu_F\alpha_t)\,\mathbb{E}\|\hat{x}_t\|^2 + c_{x,1}^{de}\beta_t^2\mathbb{E}\|\hat{y}_t\|^2 + c_{x,2}^{de}\beta_t\|\mathbb{E}\hat{x}_t\hat{y}_t^\top\| + 2\Gamma_{11}\alpha_t^2 \\ + c_{x,3}^{de}\beta_t^2 + c_{x,4}^{de}\frac{\beta_t^{2+2\delta_H}}{\alpha_t} + \Delta_{x,t}.\end{aligned} \tag{36}$$

where $\Delta_{x,t}$ is a higher-order residual given in the following

$$\Delta_{x,t} = c_{x,5}^{de}\beta_t\mathcal{Z}_{t,2+\delta_H} + c_{x,6}^{de}\alpha_t\beta_t\mathcal{Z}_{t,2+\delta_F} + c_{x,7}^{de}\beta_t\mathcal{Z}_{t,2+2\delta_G} + c_{x,8}^{de}\beta_t^{1+\delta_H}\mathcal{Z}_{t,2+2\delta_H}, \tag{37}$$

and $\{c_{x,i}^{de}\}_{i\in[8]}$ are problem-dependent constants defined in (88).

Lemma 10 (Refined one-step descent of the slow iterate $\hat{y}_t$) *Suppose that Assumptions 2–3 and 5–7 hold. We have for any $t \geq 0$,*

$$\begin{aligned}\mathbb{E}\|\hat{y}_{t+1}\|^2 \leq \left(1 - \frac{2\mu_G\beta_t}{3}\right)\mathbb{E}\|\hat{y}_t\|^2 + 2L_{G,x}^2\beta_t^2\mathbb{E}\|\hat{x}_t\|^2 + 2d_yL_{G,x}\beta_t\|\mathbb{E}\hat{x}_t\hat{y}_t^\top\| \\ + \Gamma_{22}\beta_t^2 + \Delta_{y,t},\end{aligned} \tag{38}$$

where $\Delta_{y,t}$ is a higher-order residual given in the following

$$\Delta_{y,t} = S_{B,G}^2\beta_t\left(15d_y^2/\mu_G + d_y^2\beta_t + 8d_y\beta_t\right)\mathcal{Z}_{t,2+2\delta_G}. \tag{39}$$

Lemma 11 (One-step descent of the matrix cross term $\|\mathbb{E}\hat{x}_t\hat{y}_t^\top\|$) *Suppose that Assumptions 1–7 hold. We have that for any $t \geq 0$,*

$$\begin{aligned}\|\mathbb{E}\hat{x}_{t+1}\hat{y}_{t+1}^\top\| \leq \left(1 - \frac{\mu_F\alpha_t}{2}\right)\|\mathbb{E}\hat{x}_t\hat{y}_t^\top\| + \beta_t\left(L_{G,x}\mathbb{E}\|\hat{x}_t\|^2 + c_{xy,1}^{de}\mathbb{E}\|\hat{y}_t\|^2\right) \\ + \Sigma_{12}\alpha_t\beta_t + c_{xy,2}^{de}\beta_t^2 + c_{xy,3}^{de}\beta_t^{1+2\delta_H} + \Delta_{xy,t},\end{aligned} \tag{40}$$

where $\Delta_{xy,t}$ is a higher-order residual given in the following

$$\begin{aligned}\Delta_{xy,t} = 2\alpha_tS_{B,F}\mathcal{Z}_{t,2+\delta_F} + \beta_tS_{B,G}(1+2L_H)\mathcal{Z}_{t,2+\delta_G} \\ + \beta_tc_{xy,4}^{de}\mathcal{Z}_{t,2+\delta_H} + 2\alpha_t\beta_tS_{B,F}S_{B,G}\mathcal{Z}_{t,2+\delta_F+\delta_G},\end{aligned} \tag{41}$$

and $\{c_{xy,i}^{de}\}_{i\in[4]}$ are problem-dependent constants defined in (99).

The proof of these lemmas can be found in Appendices C.1, C.2 and C.3.

We conclude this step by briefly outlining the proof idea. Upon incorporating the update rules of $\hat{x}_{t+1}$ and $\hat{y}_{t+1}$ into our desired error metrics (e.g., $\mathbb{E}\|\hat{x}_{t+1}\|^2$), we decompose the errors into primary components and higher-order terms. For each component, we determine individual upper bounds and then aggregate them accordingly. Specifically, we summarize the higher-order terms into single quantities, $\Delta_{x,t}$, $\Delta_{y,t}$, and $\Delta_{xy,t}$.

5.3 Step 3: Analyze the Convergence Rates of Fourth-Order Moments

To analyze the residual terms $\Delta_{x,t}$, $\Delta_{y,t}$ and $\Delta_{xy,t}$, we focus on a single quantity $\mathcal{Z}_{t,4}$ due to the observation from Jensen's inequality: $\mathcal{Z}_{t,\delta} \le (\mathcal{Z}_{t,4})^{\delta/4}$ if $\delta \le 4$.[4] It motivates us to analyze fourth-order moments of errors, i.e., $\mathbb{E}\|\hat{x}_{t+1}\|^4$ and $\mathbb{E}\|\hat{y}_{t+1}\|^4$. To that end, we derive one-step recursions for the conditional fourth-order moments. The derivation process closely parallels that of Lemmas 9 and 10. Moreover, we emphasize that the analysis of fourth-order moments does not require the local linearity of F and G in Assumption 5. Because Assumption 7 involves the parameter δ_F and δ_G in Assumption 5. The analysis relies on the following weaker version of Assumption 7 instead.

Assumption 7† (Conditions on step sizes) *The conditions in Assumption 7 holds with $\mu_F = \mu_G = d_x = d_y = 1$, where d_x and d_y appear in in* (49).

Lemma 12 (One-step descent of the fourth-order moment of the fast iterate $\hat{x}_t$) *Suppose that Assumptions 1–4, 6, and 7† hold. We have for any $t \ge 0$,*

$$\begin{aligned}\mathbb{E}\left[\|\hat{x}_{t+1}\|^4 \,|\, \mathcal{F}_t\right] \le{}& (1-\mu_F\alpha_t)\,\|\hat{x}_t\|^4 + c^{de}_{xx,1}\beta_t\|\hat{x}_t\|^3\|\hat{y}_t\| + c^{de}_{xx,2}\beta_t^2\|\hat{x}_t\|^2\|\hat{y}_t\|^2 \\ &+ c^{de}_{xx,3}\beta_t^4\|\hat{y}_t\|^4 + 32\alpha_t^4\Gamma_{11}^2 + 224L_H^4\beta_t^4\Gamma_{22}^2 \\ &+ \left(20\Gamma_{11}\alpha_t^2 + 20L_H^2\Gamma_{22}\beta_t^2 + c^{de}_{xx,4}\frac{\beta_t^{2+2\delta_H}}{\alpha_t}\right)\|\hat{x}_t\|^2,\end{aligned} \tag{42}$$

where $\{c_{xx,i}\}_{i\in[5]}$ are problem-dependent constants defined in (107).

Lemma 13 (One-step descent of the fourth-order moment of the slow iterate $\hat{y}_t$) *Suppose that Assumptions 2, 3, 6, and 7† hold. We have for any $t \ge 0$,*

$$\begin{aligned}\mathbb{E}\left[\|\hat{y}_{t+1}\|^4 \,|\, \mathcal{F}_t\right] \le{}& \left(1 - \frac{3\mu_G\beta_t}{2}\right)\|\hat{y}_t\|^4 + 4L_{G,x}\beta_t\|\hat{x}_t\|\|\hat{y}_t\|^3 + 18L_{G,x}^2\beta_t^2\|\hat{x}_t\|^2\|\hat{y}_t\|^2 \\ &+ 20L_{G,x}^4\beta_t^4\|\hat{x}_t\|^4 + 18\Gamma_{22}\beta_t^2\|\hat{y}_t\|^2 + 28\beta_t^4\Gamma_{22}^2.\end{aligned} \tag{43}$$

The proof of these lemmas can be found in Appendices C.4 and C.5.

By integrating Lemmas 12 and 13 and using a carefully designed Lyapunov function $V_t = \varrho_3\frac{\beta_t}{\alpha_t}\|\hat{x}_t\|^4 + \|\hat{y}_t\|^4$ for a properly specified ϱ_3, we can apply the results of Theorem 8 to determine the convergence rates for the fourth-order moments.

4. Here we choose the fourth-order for simplicity. It is feasible and natural to use an order smaller than 4, which, however, would increase the complexity of the proof.

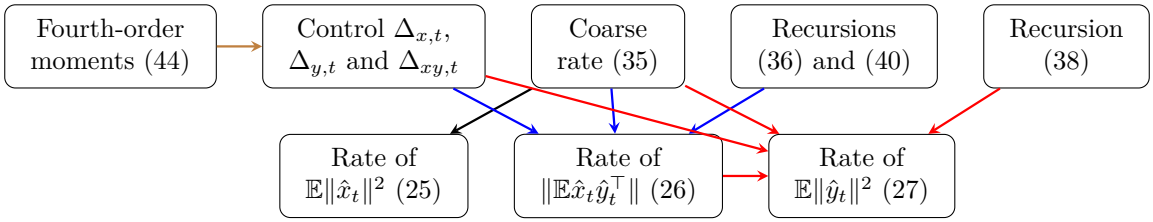

Figure 1: Illustration for Step 4 of the proof sketch.

Lemma 14 (Convergence rates of the fourth-order moments) *Suppose that Assumptions 1–4, 6, and 7† hold. Then we have for all $t \geq 0$,*

$$\mathbb{E}\|\hat{x}_{t+1}\|^4 \leq \prod_{\tau=0}^{t}\left(1-\frac{\mu_G\beta_\tau}{4}\right)\left(2\mathbb{E}\|\hat{x}_0\|^4+\frac{\mu_G^4\,\mathbb{E}\|\hat{y}_0\|^4}{27L_{G,x}^4}\right)+\frac{4c_{xx,7}^{de}}{\mu_F}\alpha_t^2+\frac{3c_{xx,8}^{de}}{\mu_F}\frac{\beta_t^{4+4\delta_H}}{\alpha_t^4},$$
$$\mathbb{E}\|\hat{y}_{t+1}\|^4 \leq \prod_{\tau=0}^{t}\left(1-\frac{\mu_G\beta_\tau}{4}\right)\left(\frac{L_{G,x}^3\mathbb{E}\|\hat{x}_0\|^4}{3\mu_G^2L_HL_{G,y}}+\mathbb{E}\|\hat{y}_0\|^4\right)+\frac{8c_{yy,1}^{de}}{\mu_G}\alpha_t^2+\frac{10c_{yy,2}^{de}}{\mu_G}\frac{\beta_t^{4+4\delta_H}}{\alpha_t^4},$$

where $\{c_{xx,i}^{de}\}_{i\in[7,8]}$ and $\{c_{yy,i}^{de}\}_{i\in[2]}$ are problem-dependent constants defined in (139) *and* (142)*. Moreover, when $\delta_H \geq 0.5$, then*

$$\mathbb{E}\|\hat{x}_{t+1}\|^4+\mathbb{E}\|\hat{y}_{t+1}\|^4=\mathcal{O}(\alpha_t^2). \tag{44}$$

The proof can be found in Appendix C.6.

5.4 Step 4: Integrate the Above Ingredients and Derive Decoupled Convergence Rates

With the aforementioned lemmas, we could integrate them to derive the convergence rates in Theorem 3. Figure 1 provides a visual representation of the process. The integration follows the following procedure.

- Black arrow: The convergence rate of $\mathbb{E}\|\hat{x}_t\|^2$ in (82) directly follows from the coarse rate (35).
- Brown arrow: Using the fourth-order convergence rates in (44) we could manage the higher-order residual terms $\Delta_{x,t}$, $\Delta_{y,t}$, and $\Delta_{xy,t}$. For the detailed upper bounds, refer to (146), (150) and (152).
- Blue arrows: Combining the recursions (36) and (40) with a properly chosen Lyapunov function and applying the coarse rate in (35), we derive the convergence rate of $\|\mathbb{E}\hat{x}_t\hat{y}_t^\top\|$ in (26).
- Red arrows: Substituting the convergence rates of $\mathbb{E}\|\hat{x}_t\|^2$ and $\|\mathbb{E}\hat{x}_t\hat{y}_t^\top\|$ into the recursion (38) yields the convergence rate of $\mathbb{E}\|\hat{y}_t\|^2$ in (27).

The details of proof can be found in Appendix C.7.

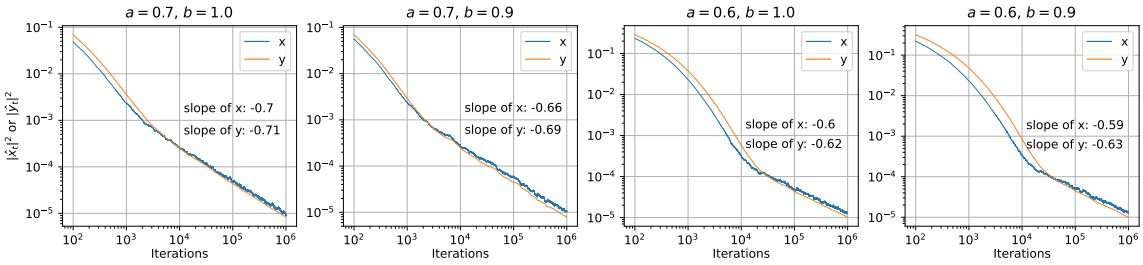

Figure 2: The convergence results within Example 5. We calculate the line slopes using data from the 2×10^5 to 10^6 iteration range.

6. Numerical Experiments

This section presents the numerical experiments. In Section 6.1, we report the results for Example 5 and its locally linear variant, illustrating the necessity of local linearity for decoupled convergence. In Sections 6.2 and 6.3, we consider one-dimensional toy examples and logistic regression, respectively, to illustrate the decoupled convergence rates established in Section 4.1.

6.1 Example 5 and Its Locally Linear Variant

In Section 4.2, we have devised Example 5 to show that local linearity is necessary for decoupled convergence. To illustrate this necessity, we start from $(x_0, y_0) = (2, 1)$, consider noise terms $\xi_t \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$ and $\psi_t = 0$, and step sizes $\alpha_t = \alpha_0(t+1)^{-a}$ and $\beta_t = \beta_0(t+1)^{-b}$. To find the optimal values of (α_0, β_0), we perform a grid search on $\{10, 3, 1, 0.3, 0.1\}^2$ for each pair, running 10^5 iterations. For each $(a, b) \in \{0.7, 0.6\} \times \{1.0, 0.9\}$, we run 10^6 steps of (2) across 10^3 repetitions.

Figure 2 presents results on a log-log scale, plotting the averaged values of $|\hat{x}_t|^2$ and $|\hat{y}_t|^2$ against the number of iterations for different (a, b) pairs. The slope of each line in the log-log plot reflects the convergence rate, as a relationship of the form $y = r\,x^{-s}$ corresponds to $\log y = -s \log x + \log r$. Despite using distinctly different time scales, the slope of the orange line (representing the slow iterate y_t) nearly matches that of the blue line (representing the fast iterate x_t). This indicates that the nonlinear interaction in Example 5 hinders the convergence of the slow iterate y_t, preventing decoupled convergence, consistent with the theoretical results in Proposition 5.

Next, we consider a local linear variant of Example 5. Define the auxiliary function

$$\tilde{h}_\delta(x) = \begin{cases} \operatorname{sign}(x)|x|^\delta/\delta, & |x| \le 1, \\ \operatorname{sign}(x)(|x| - 1 + 1/\delta), & |x| > 1. \end{cases} \tag{45}$$

One can check that when $\delta \ge 1$, $\tilde{h}_\delta$ is 1-Lipschitz continuous.

A local linear variant of Example 5. Consider the following variant of Example 5

$$F(x, y) = x - y, \;\; G(x, y) = -\tilde{h}_{1.5}(|x - y|) \cdot \operatorname{sign}(y) + y. \tag{46}$$

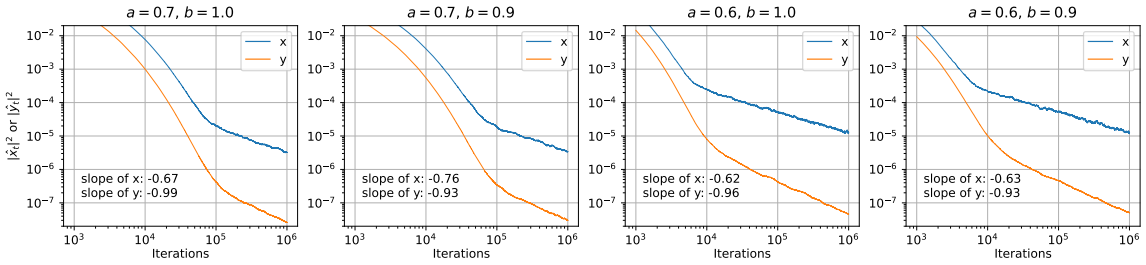

Figure 3: The convergence results for the example in (46). We calculate the line slopes using data from the 3×10^5 to 10^6 iteration range.

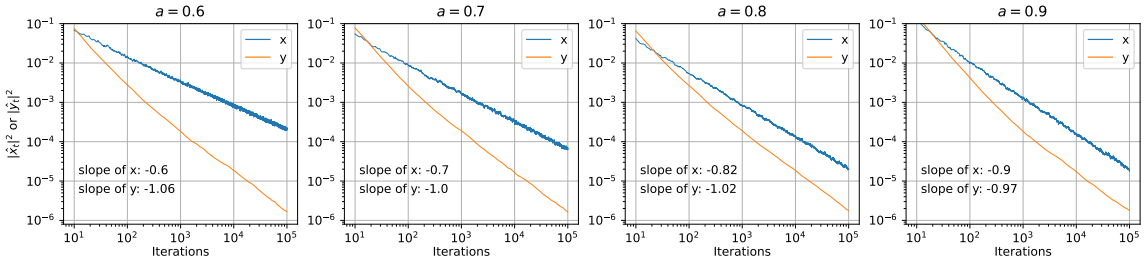

Figure 4: The convergence results for SGD with Polyak-Ruppert averaging (4). We calculate the line slopes using data from the 10^4 to 10^5 iteration range.

Assumption 5 holds with $S_{B,F} = 0$ (allowing δ_F to be set to 1) and $\delta_G = 0.5$. We start from $(x_0, y_0) = (2, 2)$, consider noise terms $\xi_t \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$ and $\psi_t \overset{i.i.d.}{\sim} \mathcal{N}(0, 0.01)$. Other settings are the same as Figure 2. The results are shown in Figure 3. Based on Corollary 4, decoupled convergence is expected within the range $1 \le b/a \le 1.5$. Interestingly, as Figure 3 demonstrates, decoupled convergence is observed even when $b/a = 1/0.6 > 1.5$.

6.2 Toy Examples

In this subsection, we illustrate the decoupled convergence in Section 4 through numerical results on one-dimensional toy examples. To reduce fluctuation, all experiments are repeated 1000 times. The errors $|\hat{x}_t|^2$ and $|\hat{y}_t|^2$ are averaged over these 1000 repetitions.

SGD with Polyak-Ruppert averaging. We employ SGD with Polyak-Ruppert averaging (4) to minimize $f(x) = x^2 + \sin x$ with $(x_0, y_0) = (2, 2)$, $\alpha_t = \alpha_0(t+1)^{-a}$, $\beta_t = (t+1)^{-1}$, $\xi_t \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$ and $a \in \{0.6, 0.7, 0.8, 0.9\}$. For each a, a grid search is performed on $\{10, 3, 1, 0.3, 0.1\}$ to find the optimal choice for α_0 and each grid search is conducted with 10^4 iterations. The results are depicted in Figure 4. The value of a does not affect the convergence rate of $|\hat{y}_t|^2$, which is roughly $\Theta(1/t)$.

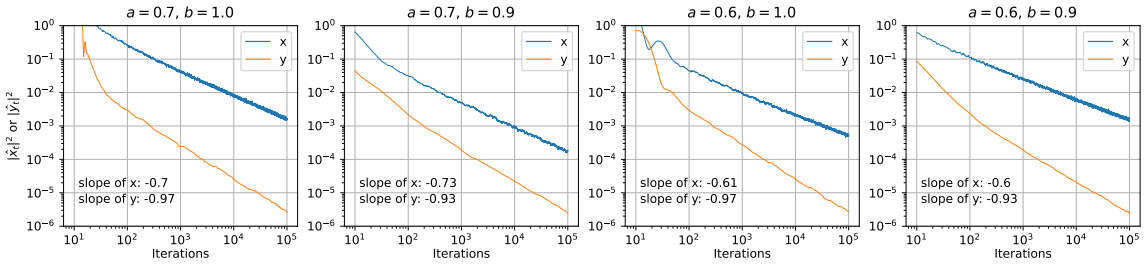

Figure 5: The convergence results for SHB (6). We calculate the line slopes using data from the 10^4 to 10^5 iteration range.

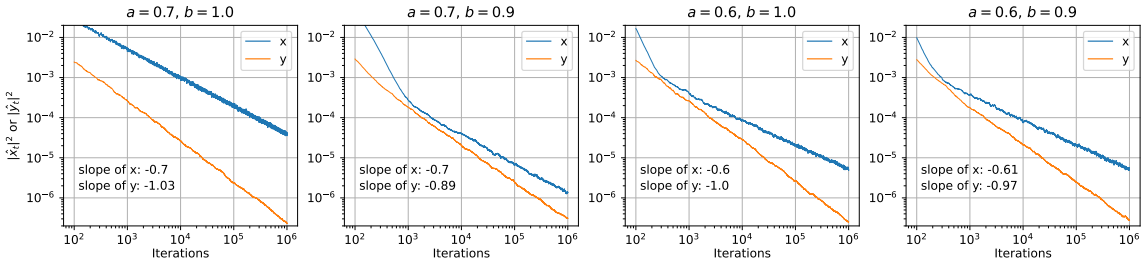

Figure 6: The convergence results for two-time-scale SA to solve (47). We calculate the line slopes using data from the 3×10^5 to 10^6 iteration range.

SGD with momentum. We employ SHB (6) to minimize $f(x) = x^2 + \sin x$ with $(x_0, y_0) = (2, 2)$, $\alpha_t = \alpha_0(t+1)^{-a}$, $\beta_t = \beta_0(t+1)^{-b}$, $\xi_t \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ and $(a, b) \in \{0.7, 0.6\} \times \{1.0, 0.9\}$. For each (a, b), a grid search is performed on $\{10, 3, 1, 0.3, 0.1\}^2$ to find the optimal choices for (α_0, β_0) and each grid search is conducted with 10^4 iterations. The results are depicted in Figure 5. Decoupled convergence is achieved.

Stochastic bilevel optimization. Consider the following problem with $f(x, y) = (x + \tilde{h}_2(y))^2 + \sin(x + \tilde{h}_2(y))$ and $g(x, y) = (x + \tilde{h}_2(y))^2 + y^2 + \sin(y)$, with $\tilde{h}_2$ defined in (45):

$$\begin{aligned} &\min_{y \in \mathbb{R}} (\tilde{x}^\star(y) + \tilde{h}_2(y))^2 + y^2 + \sin(y), \\ &\text{s.t. } \tilde{x}^\star(y) := \arg\min_{x \in \mathbb{R}} (x + \tilde{h}_2(y))^2 + \sin(x + \tilde{h}_2(y)). \end{aligned} \tag{47}$$

We apply two-time-scale SA, with F and G defined in (9) to solve this problem, with $(x_0, y_0) = (2, 2)$, $\alpha_t = \alpha_0(t+1)^{-a}$, $\beta_t = \beta_0(t+1)^{-b}$, $\xi_t, \psi_t \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ and $(a, b) \in \{0.7, 0.6\} \times \{1.0, 0.9\}$. For each (a, b), a grid search is performed on $\{10, 3, 1, 0.3, 0.1\}^2$ to find the optimal choices for (α_0, β_0) and each grid search is conducted with 10^5 iterations. The results, depicted in Figure 6, illustrate decoupled convergence for different (a, b) pairs.

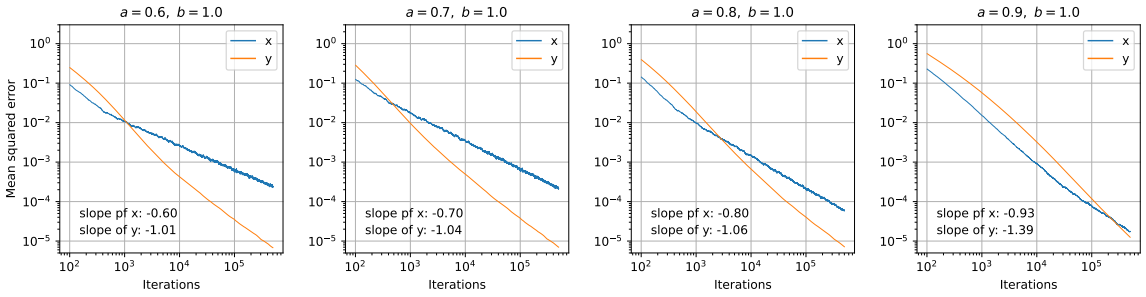

Figure 7: The convergence results for SGD with Polyak-Ruppert averaging (4) to solve (48). We calculate the line slopes using data from the 10^5 to 5×10^5 iteration range.

6.3 Logistic Regression

In this subsection, we consider the following ℓ_2-regularized logistic regression problem:

$$\min_{x \in \mathbb{R}^d} f(x) := \frac{1}{n} \sum_{i=1}^{n} \log\bigl(1 + \exp(-b_i a_i^\top x)\bigr) + \frac{\lambda}{2} \|x\|^2, \tag{48}$$

where $a_i \in \mathbb{R}^d$ is the covariate and $b_i \in \{-1, 1\}$ is the label. We set the regularization parameter to $\lambda = 0.01$, so that the objective is strongly convex and therefore admits a unique minimizer.

Data generation. We use a synthetic logistic regression model with dimension $d = 20$ and sample size $n = 1000$. To generate the dataset, we first sample a ground-truth parameter $w_{\text{true}} \in \mathbb{R}^{20}$ from a standard Gaussian distribution. Then we generate covariates $a_i \sim \mathcal{N}(0, I_{20})$ independently, and produce binary labels according to the logistic model $\mathbb{P}(b_i = 1 \mid a_i) = \sigma(a_i^\top w_{\text{true}})$, $\sigma(z) = \frac{1}{1+e^{-z}}$. Equivalently, $b_i \in \{-1, 1\}$ is sampled with $\mathbb{P}(b_i = 1 \mid a_i) = \sigma(a_i^\top w_{\text{true}})$, $\mathbb{P}(b_i = -1 \mid a_i) = 1 - \sigma(a_i^\top w_{\text{true}})$. After generation, the dataset is fixed throughout the whole experiment.

SGD with Polyak-Ruppert averaging. We employ SGD with Polyak-Ruppert averaging (4) to minimize (48), with $(x_0, y_0) = (0, 0)$, $\alpha_t = \alpha_0(t+1)^{-a}$, $\beta_t = (t+1)^{-1}$, and $a \in \{0.6, 0.7, 0.8, 0.9\}$. For each a, we perform a grid search over $\{10, 3, 1, 0.3, 0.1\}$ to determine the optimal choice of α_0, and each grid search is conducted for 5×10^4 iterations. The noise for the fast iterate comes from minibatch sampling with batch size 32. The results are shown in Figure 7. The y-axis represents the average of $\|\hat{x}_t\|^2$ or $\|\hat{y}_t\|^2$ over 100 repetitions. The figure shows that decoupled convergence can be achieved.

SGD with momentum. We employ SHB (6) to minimize (48), with $(x_0, y_0) = (0, 0)$, $\alpha_t = \alpha_0(t+1)^{-a}$, $\beta_t = \beta_0(t+100)^{-b}$, and $(a, b) \in \{0.7, 0.6\} \times \{1.0, 0.9\}$. For each (a, b), we perform a grid search over $\{10, 3, 1, 0.3, 0.1\} \times \{1000, 300, 100, 30, 10\}$ to determine the optimal choices of (α_0, β_0), and each grid search is conducted for 5×10^4 iterations. The noise for the fast iterate again comes from minibatch sampling with batch size 32. The results are shown in Figure 8. Unlike Figure 7, the y-axis here represents the average of $\|x_t - x^\star\|^2$ or $\|\hat{y}_t\|^2$ over 100 repetitions. We plot $\|x_t - x^\star\|^2$ instead of $\|\hat{x}_t\|^2$ to reduce the

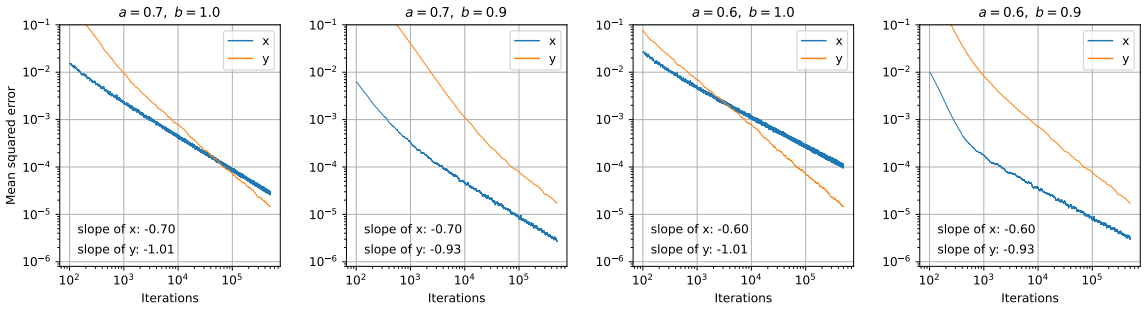

Figure 8: The convergence results for SHB (6) to solve (48). We calculate the line slopes using data from the 10^5 to 5×10^5 iteration range.

computational cost, because computing $H(y) = \nabla f(y)$ requires passing through the entire dataset, whereas $x^\star = H(y^\star) = 0$. As discussed in Remark 2, $\mathbb{E}\|x_t - x^\star\|^2$ is also of order $\mathcal{O}(\alpha_t)$. Figure 8 again shows that decoupled convergence can be achieved.

7. Concluding Remarks

In this paper, we have investigated the potential for finite-time decoupled convergence in nonlinear two-time-scale SA under the strongly monotone condition, wherein the mean-square errors of different iterates depend solely on their respective step sizes. Viewing the two-time-scale SA as an approximation of a two-loop procedure, our primary focus is on the outer-loop iterate, i.e., the slow iterate y_t.

Under a nested local linearity assumption, we have established the first finite-time decoupled convergence for nonlinear two-time-scale SA with appropriate step size selection. This decoupled convergence offers greater flexibility in choosing the step size for the fast iterate x_t, without impacting the convergence rate of the main focus, the slow iterate y_t. Our analytical framework advances the approach for the linear operators in Kaledin et al. (2020) to adapt complexities introduced by non-linearity. In particular, we have derived a refined characterization of the matrix cross term, surpassing previous asymptotic results (Mokkadem and Pelletier, 2006), and applied fourth-order moment convergence rates to manage higher-order error terms induced by local linearity. In addition, we provide an example showing that decoupled convergence may fail even when the fast-time-scale update is linear, as long as the slow-time-scale update remains nonlinear. Together with our upper bound, this lower-bound result helps clarify when decoupled convergence should be expected in the nonlinear setting. It also sheds further light on the approximation perspective in (3): even if two-time-scale SA can be viewed as solving $F(x, y) = 0$ (for fixed y) and $G(H(y), y) = 0$, the original form of $G(x, y)$ may still affect the convergence behavior. We hope that this observation may also be useful in inspiring future algorithm design.

Despite the progress made in our paper, several avenues for future research remain. First, our results could be extended to include scenarios with Markovian noise or non-strongly monotone operators, broadening the applicability of our approach. Second, investigating

the asymptotic trajectory behavior and developing online statistical inference methods for two-time-scale SA based on our non-asymptotic convergence results would be interesting future directions. Finally, another natural direction is to generalize our framework to more complex algorithms, such as those involving multiple iterates or multiple time scales.

Acknowledgments

We are grateful to the Action Editor and the anonymous reviewers for their careful reading and constructive comments, which helped improve the presentation of this paper. This research was supported by the National Natural Science Foundation of China (NSFC) [Grants 12501417, 12350001, 12271011].

Appendix A. Omitted Details in Section 3

In this section, we present the omitted details in Section 3.

First, we give the detailed definition of the constants in Assumption 7.

$$\begin{aligned}\iota_1 &= \frac{\mu_F}{4L_F^2} \wedge \frac{1}{12\mu_F},\ \iota_2 = \frac{\mu_G}{L_{G,x}^2} \wedge \frac{1}{14\mu_G},\ \kappa = \frac{\mu_F\mu_G}{(28d_x \vee 200L_{G,y})L_H L_{G,x}} \wedge \frac{\mu_F}{5\mu_G},\\ \rho &= \frac{\mu_F}{(16d_x \vee 200)L_H^2 L_{G,x}^2}.\end{aligned} \tag{49}$$

As long as $\beta_t/\alpha_t = o(1)$ and $\beta_t = o(1)$, e.g., $\alpha_t \sim \alpha_0 t^{-a}$ and $\beta_t \sim \beta_0 \sim t^{-b}$, Assumption 7 will hold for sufficiently large t, regardless of the initial values α_0 and β_0. However, if β_t/α_t remains constant, α_0 and β_0 must be appropriately chosen to satisfy Assumption 7. This comparison highlights the advantage of using different time scales over the single-time-scale case in terms of flexible parameter selection.

In the remaining part, the proofs of Propositions 1 and 2 are given in Appendices A.1 and A.2, respectively. The verification of Assumptions 4 and 5 is provided in Appendix A.3.

A.1 Proof of Proposition 1

Proof [Proof of Proposition 1] The proof is divided into two parts. The first part is straightforward while the second part follows a similar procedure as the proof of Berger et al. (2020, Theorem 4.1)

Assumption 4† $\implies$ **Assumption 4** . For any unit vector $e \in \mathbb{R}^{d_x}$, we define $f(y) = \langle e, H(y)\rangle$. Then we have $f'(y) = \nabla H(y)^\top e$ and

$$\begin{aligned}&f(y_1) - f(y_2) - \langle f'(y_2), y_1 - y_2\rangle = \int_0^1 \langle f'(y_2 + t(y_1 - y_2)) - f'(y_2), y_1 - y_2\rangle \mathrm{d}t\\ &\le \|y_1 - y_2\| \int_0^1 \|\nabla H(y_2 + t(y_1 - y_2)) - \nabla H(y_2)\| \mathrm{d}t\\ &\overset{(17)}{\le} \widetilde{S}_H \|y_1 - y_2\|^{1+\delta_H} \int_0^1 t^{\delta_H} \mathrm{d}t \le \frac{\widetilde{S}_H}{1+\delta_H} \|y_1 - y_2\|^{1+\delta_H}.\end{aligned}$$

From the definition of f, we have $\langle e, H(y_1)-H(y_2)-\nabla H(y_2)(y_1-y_2)\rangle \le \frac{\widetilde{S}_H}{1+\delta_H}\|y_1-y_2\|^{1+\delta_H}$. Since e is an arbitrary unit vector, Assumption 4 holds with $S_H = \frac{\widetilde{S}_H}{1+\delta_H}$.

Assumption 4 $\implies$ Assumption 4† . Under Assumption 4, we have for any $y_1, y_2 \in \mathbb{R}^{d_y}$ and any unit vector $e \in \mathbb{R}^{d_x}$, it holds that

$$-S_H\|y_1-y_2\|^{1+\delta_H} \le \langle e, H(y_1)-H(y_2)-\nabla H(y_2)(y_1-y_2)\rangle \le S_H\|y_1-y_2\|^{1+\delta_H}. \quad (50)$$

Now we fix two arbitrary points $\bar{y}_1$ and $\bar{y}_2 \in \mathbb{R}^{d_y}$ with $\bar{y}_1 \neq \bar{y}_2$. Without loss of generality, we assume $H(\bar{y}_1) = 0$ and $\nabla H(\bar{y}_1) = 0$. Otherwise, we could replace $H(y)$ by $\widetilde{H}(y) =: H(y) - H(\bar{y}_1) - \nabla H(\bar{y}_1)(y-\bar{y}_1)$ in (50). For any $z_1, z_2 \in \mathbb{R}^{d_y}$, setting $(y_1, y_2) = (z_1, \bar{y}_1)$ and $(z_1, \bar{y}_2)$ in (50) yields

$$-S_H\|z_1-\bar{y}_1\|^{1+\delta_H} \le \langle e, H(z_1)\rangle \le \langle e, H(\bar{y}_2) + \nabla H(\bar{y}_2)(z_1-\bar{y}_2)\rangle + S_H\|z_1-\bar{y}_2\|^{1+\delta_H}. \quad (51)$$

Similarly, setting $(y_1, y_2) = (z_2, \bar{y}_1)$ and $(z_2, \bar{y}_2)$ in (50) yields

$$S_H\|z_2-\bar{y}_1\|^{1+\delta_H} \ge \langle e, H(z_2)\rangle \ge \langle e, H(\bar{y}_2) + \nabla H(\bar{y}_2)(z_2-\bar{y}_2)\rangle - S_H\|z_2-\bar{y}_2\|^{1+\delta_H}. \quad (52)$$

Subtracting the rightmost and leftmost sides of (51) from (52), we obtain

$$\langle e, \nabla H(\bar{y}_2)(z_2-z_1)\rangle \le S_H\left(\|z_1-\bar{y}_1\|^{1+\delta_H} + \|z_2-\bar{y}_1\|^{1+\delta_H} + \|z_1-\bar{y}_2\|^{1+\delta_H} + \|z_2-\bar{y}_2\|^{1+\delta_H}\right).$$

Let $\tilde{e} \in \mathbb{R}^{d_y}$ be an arbitrary unit vector, $z_1 = \frac{\bar{y}_1+\bar{y}_2-\alpha\tilde{e}}{2}$ and $z_2 = \frac{\bar{y}_1+\bar{y}_2+\alpha\tilde{e}}{2}$ with α a positive constant determined later. Then we have

$$\alpha\langle e, \nabla H(\bar{y}_2)\,\tilde{e}\rangle \le 2S_H\left(\left\|\frac{\bar{y}_2-\bar{y}_1+\alpha\tilde{e}}{2}\right\|^{1+\delta_H} + \left\|\frac{\bar{y}_2-\bar{y}_1-\alpha\tilde{e}}{2}\right\|^{1+\delta_H}\right). \quad (53)$$

Since the function $h(x) = x^{\frac{1+\delta_H}{2}}$ is concave, Jensen's inequality together with the parallelogram identity implies

$$\left\|\frac{\bar{y}_2-\bar{y}_1+\alpha\tilde{e}}{2}\right\|^{1+\delta_H} + \left\|\frac{\bar{y}_2-\bar{y}_1-\alpha\tilde{e}}{2}\right\|^{1+\delta_H} \le 2^{\frac{1-\delta_H}{2}}\left(\frac{\|\bar{y}_2-\bar{y}_1\|^2+\alpha^2}{2}\right)^{\frac{1+\delta_H}{2}}, \quad (54)$$

Plugging (54) into (53), dividing both sides by α and setting $\alpha = k\,\|\bar{y}_2-\bar{y}_1\|$, we obtain $\langle e, \nabla H(\bar{y}_2)\,\tilde{e}\rangle \le 2^{1-\delta_H}S_H\frac{(1+k^2)^{\frac{1+\delta_H}{2}}}{k}\|\bar{y}_2-\bar{y}_1\|^{\delta_H}$.

If $\delta_H = 0$, letting $k \to \infty$ yields $\langle e, \nabla H(\bar{y}_2)\,\tilde{e}\rangle \le 2^{1-\delta_H}S_H\|\bar{y}_2-\bar{y}_1\|^{\delta_H}$. If $\delta_H > 0$, setting $k = 1/\sqrt{\delta_H}$ yields $\langle e, \nabla H(\bar{y}_2)\,\tilde{e}\rangle \le 2^{1-\delta_H}S_H\sqrt{1+\delta_H}\left(\frac{1+\delta_H}{\delta_H}\right)^{\frac{\delta_H}{2}}\|\bar{y}_2-\bar{y}_1\|^{\delta_H}$. Since e and $\tilde{e}$ are two arbitrary unit vectors and $\nabla H(\bar{y}_1) = 0$, summarizing the two cases shows that Assumption 4† holds with $\widetilde{S}_H = 2^{1-\delta_H}\sqrt{1+\delta_H}\left(\frac{1+\delta_H}{\delta_H}\right)^{\frac{\delta_H}{2}}S_H$. ■

A.2 Proof of Proposition 2

Proof [Proof of Proposition 2] We prove the results step by step.

Proof of Part (i). We first prove $\|A_{11}\| \le L_F$ and $\|A_{21}\| \le L_{G,x}$. Setting $y = y^\star$ in (21) yields $\|F(x, y^\star) - A_{11}(x - x^\star)\| \le S_{A,F}\|x - x^\star\|^{1+\delta_F}$. Then by Condition (12), we have $L_F\|x - H(y^\star)\| \ge \|F(x, y^\star)\| \ge \|A_{11}(x - x^\star)\| - \|F(x, y^\star) - A_{11}(x - x^\star)\|$. Recall that $x^\star = H(y^\star)$. It follows that

$$\|A_{11}(x - x^\star)\| \le L_F\|x - x^\star\| + S_{A,F}\|x - x^\star\|^{1+\delta_F}.$$

Let $x - x^\star = te_1$ where $e_1 \in \mathbb{R}^{d_x}$ is an arbitrary unit vector and $t > 0$. Dividing both sides by t and letting $t \to 0$ yields $\|A_{11}e_1\| \le L_F$. Since e_1 is arbitrary, we obtain $\|A_{11}\| \le L_F$. Setting $y = y^\star$ in (22) and applying Condition (13), we can obtain $\|A_{21}(x - x^\star)\| \le L_{G,x}\|x - x^\star\| + S_{A,G}\|x - x^\star\|^{1+\delta_G}$. Similarly, we can obtain $\|A_{21}\| \le L_{G,x}$.

Next, we prove $A_{11}\nabla H(y^\star) + A_{12} = 0$. Recall that the first inequality of Assumption 5† is $\|F(x, y) - A_{11}(x - x^\star) - A_{12}(y - y^\star)\| \le S_{A,F}\left(\|x - x^\star\|^{1+\delta_F} + \|y - y^\star\|^{1+\delta_F}\right)$. Since $F(H(y), y) = 0$, setting $x = H(y)$ yields

$$\|A_{11}(H(y) - H(y^\star)) + A_{12}(y - y^\star)\| \le S_{A,F}(1 + L_H^{1+\delta_F})\|y - y^\star\|^{1+\delta_F}.$$

By Assumption 4 and the triangle inequality, we have

$$\|(A_{11}H(y^\star) + A_{12})(y - y^\star)\| \le S_{A,F}(1 + L_H^{1+\delta_F})\|y - y^\star\|^{1+\delta_F} + L_F S_H\|y - y^\star\|^{1+\delta_H}.$$

Setting $y - y^* = te_2$ for an arbitrary unit vector $e_2 \in \mathbb{R}^{d_y}$ and letting $t \to 0$ yields $\|(A_{11}H(y^\star) + A_{12})e_2\| = 0$. Since e_2 is arbitrary, we have $A_{11}\nabla H(y^\star) + A_{12} = 0$.

In the following, we prove that the first inequality of Assumption 5 holds. Since $\|A_{11}\| \le L_F$ and $\|A_{21}\| \le L_{G,x}$. we have

$$\begin{aligned}
&\|F(x, y) - A_{11}(x - H(y))\| \\
&\le \|F(x, y) - A_{11}(x - x^\star) - A_{12}(y - y^\star)\| + \|A_{11}\left(H(y) - H(y^\star) - \nabla H(y^\star)(y - y^\star)\right)\| \\
&\overset{(a)}{\le} S_{A,F}\left(\|x - x^\star\|^{1+\delta_F} + \|y - y^\star\|^{1+\delta_F}\right) + L_F S_H\|y - y^\star\| \cdot \min\left\{\|y - y^\star\|^{\delta_H}, R_H\right\},
\end{aligned}$$

where (a) also uses (18) and $R_H = \frac{2L_H}{S_H}$. If $\|y - y^\star\| \le 1$, then we have

$$\begin{aligned}
\|F(x, y) - A_{11}(x - H(y))\| &\le S_{A,F}\left(\|x - x^\star\|^{1+\delta_F} + \|y - y^\star\|^{1+\delta_F}\right) + L_F S_H\|y - y^\star\|^{1+\delta_H} \\
&\le S_{A,F}\|x - x^\star\|^{1+\delta_F} + (S_{A,F} + L_F S_H)\|y - y^\star\|^{1+\delta_F}.
\end{aligned}$$

Otherwise, we have

$$\begin{aligned}
\|F(x, y) - A_{11}(x - H(y))\| &\le S_{A,F}\left(\|x - x^\star\|^{1+\delta_F} + \|y - y^\star\|^{1+\delta_F}\right) + 2L_F L_H\|y - y^\star\| \\
&\le S_{A,F}\|x - x^\star\|^{1+\delta_F} + (S_{A,F} + 2L_F L_H)\|y - y^\star\|^{1+\delta_F}.
\end{aligned}$$

Combining the two cases yields

$$\|F(x, y) - A_{11}(x - H(y))\| \le S_{A,F}\|x - x^\star\|^{1+\delta_F} + (S_{A,F} + L_F \max\{S_H, 2L_H\})\|y - y^\star\|^{1+\delta_F}. \tag{55}$$

Before proving the second inequality in Assumption 5†, we first check the lower bound for $\frac{A_{11}+A_{11}^\top}{2}$. By Assumption 3 and (55), it follows that

$$\begin{aligned}\mu_F\|x-x^\star\|^2 \le \langle x-x^\star, F(x,y^\star)\rangle &\le \langle x-x^\star, A_{11}(x-x^\star)\rangle + S_{B,F}\|x-x^\star\|^{2+\delta_F}\\ &= \left\langle x-x^\star, \frac{A_{11}+A_{11}^\top}{2}(x-x^\star)\right\rangle + S_{B,F}\|x-x^\star\|^{2+\delta_F}.\end{aligned}$$

Dividing $\|x-x^\star\|^2$ on the both sides of the last inequality and letting x converge to $x^\star$ along the direction v, we then have that $\mu_F \le v^\top \frac{A_{11}+A_{11}^\top}{2} v$ for any unit one vector v, which implies that $\frac{A_{11}+A_{11}^\top}{2} \succeq \mu_F I$. This condition also implies A_{11} is non-singular and consequently $H(y^\star) = -A_{11}^{-1}A_{12}$.

Finally, we prove the second inequality in Assumption 5†. We have

$$\begin{aligned}&\|G(x,y) - A_{21}(x-H(y)) - (A_{22}-A_{21}A_{11}^{-1}A_{12})(y-y^\star)\|\\ &\le \|G(x,y) - A_{21}(x-x^\star) - A_{22}(y-y^\star))\| + \|A_{21}\left(H(y)-x^\star + A_{11}^{-1}A_{12}(y-y^\star)\right)\|\\ &\le S_{A,G}\left(\|x-x^\star\|^{1+\delta_G} + \|y-y^\star\|^{1+\delta_G}\right) + L_{G,x}S_H\|y-y^\star\|\cdot\min\{\|y-y^\star\|^{\delta_H}, R_H\}.\end{aligned}$$

Similar to the proof of (55), we can obtain

$$\begin{aligned}&\|G(x,y) - A_{21}(x-H(y)) - (A_{22}-A_{21}A_{11}^{-1}A_{12})(y-y^\star)\|\\ &\quad\le S_{A,G}\|x-x^\star\|^{1+\delta_G} + (S_{A,G} + L_{G,x}\max\{S_H, 2L_H\})\|y-y^\star\|^{1+\delta_G}.\end{aligned}$$

Hence, the near linearity conditions in Assumption 5 follow with $B_1 = A_{11}$, $B_2 = A_{21}$, $B_3 = A_{22} - A_{21}A_{11}^{-1}A_{12}$, $S_{B,F} = S_{A,F} + L_F(S_H \vee 2L_H)$ and $S_{B,G} = S_{A,G} + L_{G,x}(S_H \vee 2L_H)$.

Proof of Part (ii). Setting $y = y^\star$ in (19) and applying Condition (12), we obtain

$$\|B_1(x-H(y^\star))\| \le L_F\|x-H(y^\star)\| + S_{B,F}\|x-H(y^\star)\|^{1+\delta_F}.$$

Similar to the proof in Part (i), we can obtain $\|B_1\| \le L_F$. Setting $y = y^\star$ in (20) and applying Condition (13), we can obtain $\|B_2\| \le L_{G,x}$. Setting $x = H(y)$ in (20) and applying Condition (14), we can obtain $\|B_3\| \le L_{G,y}$. The proof of the lower bounds for $\frac{B_1+B_1^\top}{2}$ and $\frac{B_3+B_3^\top}{2}$ is also similar ot that of the lower bound for $\frac{A_{11}+A_{11}^\top}{2}$ in Part (i). ■

A.3 Verification of Local Linearity Conditions on Examples

In this subsection, we present the verification of local linearity conditions in Assumptions 4 and 5 on the examples in Section 2.

First, we would like to emphasize an important simplification in verifying Assumption 5. Under the Lipschitz continuity of F, G, and H in Assumptions 1 and 2, together with the uniform local linearity of H in Assumption 4, if either (19)–(20) or (21)–(22) holds in a neighborhood of the solution $(x^\star, y^\star)$, then these inequalities automatically extends to

all (x, y), thereby yielding Assumption 5; see Han et al. (2024, Propositions A.2 and A.3). Therefore, for the examples, it is sufficient to verify (19)–(20) or (21)–(22) only locally around $(x^\star, y^\star)$.

For Examples 1–3, we assume that $f\colon \mathbb{R}^{d_x} \to \mathbb{R}$ is strongly convex, with unique minimizer $x_o^\star = \arg\min_{x\in\mathbb{R}^{d_x}} f(x)$.

Example 1: SGD with Polyak-Ruppert averaging. SGD with Polyak-Ruppert averaging in (4) an example of two-time-scale SA (2) with $F(x, y) = \nabla f(x)$, $G(x, y) = y - x$, and $H(y) \equiv x^\star$. It follows that $G(H(y), y) = y - x^\star$, and hence $y^\star = x^\star = x_o^\star$. If $\nabla f(x)$ is Lipschitz continuous, then F, G, and H are all Lipschitz continuous. It therefore remains only to verify (19) and (20) locally. Since H is constant, Assumption 4 holds trivially with $S_H = 0$ and $\delta_H = 1$. If $\nabla f(x)$ is differentiable in a neighborhood of $x_o^\star$ and $\nabla^2 f(x)$ is δ-Hölder continuous there, then $\|F(x, y) - \nabla^2 f(x^\star)(x - H(y))\| = \|\nabla f(x) - \nabla^2 f(x^\star)(x - x_o^\star)\| = \mathcal{O}(\|x - x^\star\|^{1+\delta})$ in that neighborhood, while $\|G(x, y) + (x - H(y)) - (y - y^\star)\| = 0$. Thus, Assumption 5 holds with $B_1 = \nabla^2 f(x_o^\star)$, $B_2 = -I$, $B_3 = I$, $\delta_F = \delta$, $S_{B,G} = 0$, and $\delta_G = 1$.

Example 2: SGD with momentum. SGD with momentum in (6) is an example of two-time-scale SA (2) with $F(x, y) = x - \nabla f(y)$, $G(x, y) = x$, and $H(y) = \nabla f(y)$. It follows that $G(H(y), y) = \nabla f(y)$, $y^\star = x_o^\star$, and $x^\star = 0$. If $\nabla f(x)$ is Lipschitz continuous, then F, G, and H are all Lipschitz continuous. It therefore remains only to verify (19) and (20) locally. If $\nabla^2 f(x)$ is δ-Hölder continuous, then Assumption 4 holds with $\delta_H = \delta$. Moreover, $\|F(x, y) - (x - H(y))\| = 0$, and $\|G(x, y) - (x - H(y)) - \nabla^2 f(x_o^\star)(y - y^\star)\| = \mathcal{O}(\|y - y^\star\|^{1+\delta})$. Therefore, Assumption 5 holds with $B_1 = I$, $B_2 = I$, $B_3 = \nabla^2 f(x_o^\star)$, $S_{B,F} = 0$, $\delta_F = 1$, and $\delta_G = \delta$.

Example 3: Constrained optimization with Lagrange multipliers. The algorithm in (7) is an example of two-time-scale SA (2) with $F(x, y) = \nabla f(x) + A^\top y$, $G(x, y) = -Ax + b$, and $H(y) = [\nabla f]^{-1}(-A^\top y)$. If $\nabla f(x)$ and $[\nabla f]^{-1}(x)$ are Lipschitz continuous, then F, G, and H are all Lipschitz continuous. For this example, it is more convenient to first verify (21) and (22) locally. If $\nabla^2 f(x)$ is δ-Hölder continuous, then (21) and (22) hold locally with $A_{11} = \nabla^2 f(x^\star)$, $A_{12} = A^\top$, $A_{21} = -A$, $A_{22} = 0$, $S_{A,G} = 0$, and $\delta_F = \delta_G = \delta$. Since $H(y) = [\nabla f]^{-1}(-A^\top y)$, we have $\nabla H(y) = -[\nabla^2 f(H(y))]^{-1} A^\top$. We now show that $\nabla H(y)$ is also δ-Hölder continuous.

Because f is strongly convex, there exists $m > 0$ such that $\nabla^2 f(x) \succeq mI$, and hence $\|[\nabla^2 f(x)]^{-1}\| \le 1/m$. Therefore, $\|\nabla H(y_1) - \nabla H(y_2)\| \le \|A\| \, \|[\nabla^2 f(H(y_1))]^{-1} - [\nabla^2 f(H(y_2))]^{-1}\|$. Using the matrix inverse identity $M^{-1} - N^{-1} = M^{-1}(N - M)N^{-1}$, we obtain $\|M^{-1} - N^{-1}\| \le \|M^{-1}\| \, \|N^{-1}\| \, \|M - N\|$. Taking $M = \nabla^2 f(H(y_1))$ and $N = \nabla^2 f(H(y_2))$ yields

$$\|[\nabla^2 f(H(y_1))]^{-1} - [\nabla^2 f(H(y_2))]^{-1}\| \le m^{-2} \|\nabla^2 f(H(y_1)) - \nabla^2 f(H(y_2))\|. \tag{56}$$

Combining the δ-Hölder continuity of $\nabla^2 f$ with the Lipschitz continuity of H, we obtain $\|\nabla H(y_1) - \nabla H(y_2)\| \lesssim \|y_1 - y_2\|^\delta$. By Proposition 1, Assumption 4 holds with $\delta_H = \delta$. Therefore, by Proposition 2 and Han et al. (2024, Proposition A.3), Assumption 5 holds with $B_1 = A_{11} = \nabla^2 f(x^\star)$, $B_2 = A_{21} = -A$, $B_3 = A_{22} - A_{21} A_{11}^{-1} A_{12} = A[\nabla^2 f(x^\star)]^{-1} A^\top$, and $\delta_F = \delta_G = \delta$. For the verification of the strong monotonicity condition, please refer to Chandak (2025b).

Example 4: Stochastic bilevel optimization. With a slight abuse of notation, consider the unconstrained bilevel optimization problem in (8). Suppose that $\ell(y)$ is strongly

convex and that $f(x,y)$ is strongly convex in x for each fixed y. To apply two-time-scale SA, $F(x,y)$ and $G(x,y)$ are of the following form

$$F(x,y) = \nabla_x f(x,y),$$
$$G(x,y) = \nabla_y g(x,y) - \nabla^2_{yx} f(x,y)[\nabla^2_{xx} f(x,y)]^{-1} \nabla_x g(x,y).$$

Then $H(y) = \tilde{x}^\star(y)$ and the solution $(x^\star, y^\star)$ satisfies $y^\star = \arg\min_{y\in\mathbb{R}^{d_y}} \ell(y)$ and $x^\star = \tilde{x}^\star(y^\star)$. Shen and Chen (2022, Lemma 1) provide conditions under which F, G, H, and ∇H are Lipschitz continuous. For simplicity, we impose the slightly stronger assumption that $\nabla^2 f(x,y)$, $\nabla f(x,y)$, $\nabla g(x,y)$, and $g(x,y)$ are all Lipschitz continuous. Then, by Proposition 1, Assumption 4 holds with $\delta_H = 1$. Under these conditions, $F(x,y)$ satisfies (21) with $A_{11} = \nabla_{xx} f(x^\star, y^\star)$, $A_{12} = \nabla_{xy} f(x^\star, y^\star)$, and $\delta_F = 1$.

For G, we further assume that $\nabla^2 g(x,y)$ is Lipschitz continuous in a neighborhood of $(x^\star, y^\star)$. For convenience, define

$$A(x,y) := \nabla_y g(x,y),\ B(x,y) := \nabla^2_{yx} f(x,y),\ C(x,y) := \nabla^2_{xx} f(x,y),\ D(x,y) := \nabla_x g(x,y),$$

so that $G(x,y) = A(x,y) - B(x,y)C(x,y)^{-1}D(x,y)$. Since $f(x,y)$ is strongly convex in x, there exists $\mu > 0$ such that $C(x,y) \succeq \mu I$, and hence $\|C(x,y)^{-1}\| \le \mu^{-1}$. Let $z := (x,y)$. Arguing as in the derivation of (56), we obtain $\|C(z_1)^{-1} - C(z_2)^{-1}\| \le \mu^{-2}\|C(z_2) - C(z_1)\| \lesssim \|z_1 - z_2\|$, which shows that C^{-1} is Lipschitz continuous. For the gradient of G, the product rule together with the derivative of the inverse gives

$$\nabla G = \nabla A - (\nabla B)C^{-1}D + BC^{-1}(\nabla C)C^{-1}D - BC^{-1}\nabla D. \tag{57}$$

Here, ∇A, ∇B, ∇C, and ∇D denote derivatives with respect to the full variable $z = (x,y)$. Under our regularity conditions, ∇A, ∇B, ∇C, and ∇D are all Lipschitz continuous in a neighborhood of $(x^\star, y^\star)$, while B, C^{-1}, and D are bounded there. Hence ∇G is Lipschitz continuous in a neighborhood of $(x^\star, y^\star)$. This implies that (22) holds with $A_{21} = \nabla_x G(x^\star, y^\star)$, $A_{22} = \nabla_y G(x^\star, y^\star)$, and $\delta_G = 1$. Therefore, by Proposition 2, Assumption 5 holds with $B_1 = A_{11} = \nabla_{xx} f(x^\star, y^\star)$, $B_2 = A_{21} = \nabla_x G(x^\star, y^\star)$, $B_3 = A_{22} - A_{21}A_{11}^{-1}A_{12} = \nabla_y G(x^\star, y^\star) - \nabla_x G(x^\star, y^\star)[\nabla_{xx} f(x^\star, y^\star)]^{-1}\nabla_{xy} f(x^\star, y^\star)$, and $\delta_F = \delta_G = 1$, where ∇G is given by (57).

Appendix B. Convergence Rates without Local Linearity

In this section, we focus on the convergence rates of nonlinear two-time-scale SA without imposing local linearity on F and G, i.e., without Assumption 5. For Assumption 4, we consider the more general case $\delta_H \in [0,1]$. Moreover, we only need the following weaker version of Assumptions 6 and 7 throughout this section.

Assumption 8 (Martingale difference noise with bounded variance) *The sequences of random variables $\{\xi_t\}_{t=0}^\infty$ and $\{\psi_t\}_{t=0}^\infty$ are martingale difference sequences satisfying*

$$\mathbb{E}[\xi_t \,|\, \mathcal{F}_t] = 0, \quad \mathbb{E}[\psi_t \,|\, \mathcal{F}_t] = 0, \quad \mathbb{E}[\|\xi_t\|^2 \,|\, \mathcal{F}_t] \le \Gamma_{11}, \quad \mathbb{E}[\|\psi_t\|^2 \,|\, \mathcal{F}_t] \le \Gamma_{22}. \tag{58}$$

Assumption 9 (Conditions on step sizes) *The step sizes $\{\alpha_t\}_{t=0}^\infty$ and $\{\beta_t\}_{t=0}^\infty$ satisfy the following conditions that for $\forall t \ge 1$:*

- *Constant bounds:* $\alpha_t \le \iota_1, \beta_t \le \iota_2, \frac{\beta_t}{\alpha_t} \le \kappa, \frac{\beta_t^2}{\alpha_t} \le \rho$ *with*

$$\iota_1 = \frac{\mu_F}{4L_F^2},\ \iota_2 = \frac{\mu_G}{L_{G,y}^2},\ \kappa = \frac{\mu_F\mu_G}{28L_HL_{G,x}(1 \vee L_{G,y})} \wedge \frac{\mu_F}{\mu_G},\ \rho = \frac{\mu_F}{16L_H^2L_{G,x}^2}.$$

- *Growth conditions:* $1 \le \frac{\alpha_{t-1}}{\alpha_t} \le 1 + (\frac{\mu_F}{4}\alpha_t) \wedge (\frac{\mu_G}{8}\beta_t)$ *and* $1 \le \frac{\beta_{t-1}}{\beta_t} \le 1 + \frac{\mu_G}{32}\beta_t$.
- $\frac{\beta_t}{\alpha_t}$ *is non-increasing in* t, *and* $\prod_{\tau=0}^{t}\left(1 - \frac{\mu_G\beta_\tau}{4}\right) = \mathcal{O}(\alpha_t)$

B.1 Proof of Lemma 6

Proof [Proof of Lemma 6] We first present the specific forms of the constants.

$$c_{x,1} = 4L_H^2L_{G,y}^2,\ c_{x,2} = 7L_HL_{G,y},\ c_{x,3} = 2L_H^2\Gamma_{22},\ c_{x,4} = \frac{16S_H^2\Gamma_{22}^{1+\delta_H}}{\mu_F}. \tag{59}$$

Recall that $\hat{x}_t = x_t - H(y_t)$. We then consider

$$\hat{x}_{t+1} = x_{t+1} - H(y_{t+1}) = \hat{x}_t - \alpha_t F(x_t, y_t) - \alpha_t\xi_t + H(y_t) - H(y_{t+1}), \tag{60}$$

which implies

$$\begin{aligned}\|\hat{x}_{t+1}\|^2 = \|\hat{x}_t - \alpha_t F(x_t, y_t)\|^2 + \|H(y_t) - H(y_{t+1}) - \alpha_t\xi_t\|^2 \\ + 2\langle \hat{x}_t - \alpha_t F(x_t, y_t), H(y_t) - H(y_{t+1})\rangle - 2\alpha_t\langle \hat{x}_t - \alpha_t F(x_t, y_t), \xi_t\rangle.\end{aligned} \tag{61}$$

We then analyze each term on the right-hand side of (61).

For the first term, noting that $F(H(y_t), y_t) = 0$, we have

$$\begin{aligned}&\|\hat{x}_t - \alpha_t F(x_t, y_t)\|^2 \\ &= \|\hat{x}_t\|^2 - 2\alpha_t\langle \hat{x}_t, F(x_t, y_t) - F(H(y_t), y_t)\rangle + \alpha_t^2\|F(x_t, y_t) - F(H(y_t), y_t)\|^2 \\ &\le \|\hat{x}_t\|^2 - 2\mu_F\alpha_t\|\hat{x}_t\|^2 + L_F^2\alpha_t^2\|\hat{x}_t\|^2 \le \left(1 - \frac{7\mu_F\alpha_t}{4}\right)\|\hat{x}_t\|^2,\end{aligned} \tag{62}$$

where the inequality uses strong monotone and Lispchitz continuity of F in (15) and (12) respectively.

For the second term, recall that $\mathcal{F}_t = \sigma\{x_0, y_0, \xi_0, \psi_0, \xi_1, \psi_1, \ldots, \xi_{t-1}, \psi_{t-1}\}$. We then take the conditional expectation of the second term on the right-hand side of (61) w.r.t. $\mathcal{F}_t$ and using Assumption 8 to have

$$\begin{aligned}&\mathbb{E}\left[\|H(y_t) - H(y_{t+1}) - \alpha_t\xi_t\|^2 \mid \mathcal{F}_t\right] \\ &\le 2L_H^2\beta_t^2\|G(x_t, y_t)\|^2 + 2L_H^2\beta_t^2\mathbb{E}\left[\|\psi_t\|^2 \mid \mathcal{F}_t\right] + 2\alpha_t^2\mathbb{E}\left[\|\xi_t\|^2 \mid \mathcal{F}_t\right] \\ &\le 4L_H^2L_{G,x}^2\beta_t^2\|\hat{x}_t\|^2 + 4L_H^2L_{G,y}^2\beta_t^2\|\hat{y}_t\|^2 + 2\beta_t^2L_H^2\Gamma_{22} + 2\alpha_t^2\Gamma_{11},\end{aligned} \tag{63}$$

where the last inequality uses the following inequality that depends on the fact $G(H(y^\star), y^\star) = 0$ and Assumption 2

$$\|G(x_t, y_t)\| \le \|G(x_t, y_t) - G(H(y_t), y_t)\| + \|G(H(y_t), y_t) - G(H(y^\star), y^\star)\|$$

$$\leq L_{G,x}\|\hat{x}_t\| + L_{G,y}\|\hat{y}_t\|, \tag{64}$$

and thus $\|G(x_t, y_t)\|^2 \leq 2L_{G,x}^2\|\hat{x}_t\|^2 + 2L_{G,y}^2\|\hat{y}_t\|^2$.

For the last term, we have that

$$\begin{aligned}\langle \hat{x}_t - \alpha_t F(x_t, y_t), H(y_t) - H(y_{t+1})\rangle = &\underbrace{\langle \hat{x}_t - \alpha_t F(x_t, y_t), \nabla H(y_t)(y_{t+1} - y_t)\rangle}_{\spadesuit_1} \\ &+ \underbrace{\langle \hat{x}_t - \alpha_t F(x_t, y_t), \nabla H(y_t)(y_t - y_{t+1}) + H(y_t) - H(y_{t+1})\rangle}_{\spadesuit_2}.\end{aligned} \tag{65}$$

For one thing, Assumptions 1 and 4 imply that $\|\nabla H(y_t)\| \leq L_H$. Then we have

$$\begin{aligned}\mathbb{E}\left[\spadesuit_1|\mathcal{F}_t\right] &= \beta_t\langle \hat{x}_t - \alpha_t F(x_t, y_t), \nabla H(y_t)G(x_t, y_t)\rangle \\ &\leq \beta_t L_H\|\hat{x}_t - \alpha_t F(x_t, y_t)\|\|G(x_t, y_t)\| \overset{(62)}{\leq} \beta_t\sqrt{1-\alpha_t\mu_F}L_H\|\hat{x}_t\|\|G(x_t, y_t)\|.\end{aligned}$$

For another thing, by Assumption 4, it follows that

$$\begin{aligned}&\|\nabla H(y_t)(y_t - y_{t+1}) + H(y_t) - H(y_{t+1})\| \leq S_H\|y_t - y_{t+1}\| \cdot \min\left\{\|y_t - y_{t+1}\|^{\delta_H}, R_H\right\} \\ &\overset{(a)}{\leq} S_H\beta_t\left(\|G(x_t, y_t)\| + \|\psi_t\|\right) \cdot \min\left\{\beta_t^{\delta_H}\left(\|G(x_t, y_t)\|^{\delta_H} + \|\psi_t\|^{\delta_H}\right), R_H\right\} \\ &\overset{(b)}{\leq} S_H\beta_t\left(\|G(x_t, y_t)\| + \|\psi_t\|\right) \cdot \left(\min\left\{\beta_t^{\delta_H}\|G(x_t, y_t)\|^{\delta_H}, R_H\right\} + \min\left\{\beta_t^{\delta_H}\|\psi_t\|^{\delta_H}, R_H\right\}\right) \\ &\overset{(c)}{\leq} S_H\Bigg(2\beta_t\|G(x_t, y_t)\|R_H + \frac{1}{1+\delta_H}\beta_t^{1+\delta_H}\|\psi_t\|^{1+\delta_H} + \frac{\delta_H}{1+\delta_H}\min\left\{\beta_t^{\delta_H}\|G(x_t, y_t)\|^{\delta_H}, R_H\right\}^{\frac{1+\delta_H}{\delta_H}} \\ &\qquad + \beta_t^{1+\delta_H}\|\psi_t\|^{1+\delta_H}\Bigg) \\ &\leq S_H\left(\frac{2+3\delta_H}{1+\delta_H}\beta_t\|G(x_t, y_t)\|R_H + \frac{2+\delta_H}{1+\delta_H}\beta_t^{1+\delta_H}\|\psi_t\|^{1+\delta_H}\right),\end{aligned} \tag{66}$$

where (a) uses $(a+b)^{\delta_H} \leq a^{\delta_H} + b^{\delta_H}$ for any $a, b > 0$ and $\delta_H \in [0, 1]$, (b) uses the inequality that $\min\{a+b, c\} \leq \min\{a, c\} + \min\{b, c\}$ for any non-negative a, b, c, and (c) follows from Young's inequality $ab \leq \frac{a^p}{p} + \frac{b^q}{q}$ for any $a, b, p, q \geq 0$ and $\frac{1}{p} + \frac{1}{q} = 1$. Note that Jensen's inequality implies $\mathbb{E}[\|\phi_t\|^{1+\delta_H}|\mathcal{F}_t] \leq \left(\mathbb{E}[\|\phi_t\|^2|\mathcal{F}_t]\right)^{\frac{1+\delta_H}{2}} \leq \Gamma_{22}^{\frac{1+\delta_H}{2}}$. Hence,

$$\mathbb{E}\left[\|\nabla H(y_t)(y_t - y_{t+1}) + H(y_t) - H(y_{t+1})\||\mathcal{F}_t\right] \leq 6L_H\beta_t\|G(x_t, y_t)\| + 4S_H\beta_t^{1+\delta_H}\Gamma_{22}^{\frac{1+\delta_H}{2}},$$

$$\begin{aligned}\text{and } \mathbb{E}\left[\spadesuit_2|\mathcal{F}_t\right] &\leq \|\hat{x}_t - \alpha_t F(x_t, y_t)\| \cdot \mathbb{E}\left[\|\nabla H(y_t)(y_t - y_{t+1}) + H(y_t) - H(y_{t+1})\||\mathcal{F}_t\right] \\ &\overset{(62)}{\leq} \sqrt{1-\alpha_t\mu_F}\|\hat{x}_t\|\left(6L_H\beta_t\|G(x_t, y_t)\| + 4S_H\beta_t^{1+\delta_H}\Gamma_{22}^{\frac{1+\delta_H}{2}}\right).\end{aligned}$$

Combing these two inequalities, we then obtain the preceding relation

$$\mathbb{E}\left[\langle \hat{x}_t - \alpha_t F(x_t, y_t), H(y_t) - H(y_{t+1})\rangle|\mathcal{F}_t\right]$$

$$\leq 7L_H\beta_t\sqrt{1-\alpha_t\mu_F}\|\hat{x}_t\|\|G(x_t,y_t)\| + 4S_H\sqrt{1-\alpha_t\mu_F}\beta_t^{1+\delta_H}\|\hat{x}_t\|\Gamma_{22}^{\frac{1+\delta_H}{2}}$$
$$\overset{(64)}{\leq} 7L_H\beta_t\left(L_{G,x}\sqrt{1-\alpha_t\mu_F}\|\hat{x}_t\|^2 + L_{G,y}\sqrt{1-\alpha_t\mu_F}\|\hat{x}_t\|\|\hat{y}_t\|\right) + 4S_H\sqrt{1-\alpha_t\mu_F}\beta_t^{1+\delta_H}\|\hat{x}_t\|\Gamma_{22}^{\frac{1+\delta_H}{2}}$$

$$\leq \frac{\mu_F\alpha_t}{2}\|\hat{x}_t\|^2 + 7L_HL_{G,y}\beta_t\sqrt{1-\alpha_t\mu_F}\|\hat{x}_t\|\|\hat{y}_t\| + \frac{16S_H^2\Gamma_{22}^{1+\delta_H}}{\mu_F}\frac{\beta_t^{2+2\delta_H}}{\alpha_t} \tag{67}$$

where the last inequality uses the relation obtained from the choices of step sizes $7L_HL_{G,x}\beta_t \leq \frac{\mu_F\alpha_t}{4}$ and Cauchy–Schwarz inequality.

Combing the three bounds in (62), (63), and (67), we obtain that

$$\begin{aligned}\mathbb{E}\left[\|\hat{x}_{t+1}\|^2 \,|\, \mathcal{F}_t\right] &\leq \left(1 - \frac{7\mu_F\alpha_t}{4}\right)\|\hat{x}_t\|^2 + 4L_H^2L_{G,x}^2\beta_t^2\|\hat{x}_t\|^2 + 4L_H^2L_{G,y}^2\beta_t^2\|\hat{y}_t\|^2 + 2L_H^2\Gamma_{22}\beta_t^2 \\ &\quad + 2\Gamma_{11}\alpha_t^2 + \frac{\mu_F\alpha_t}{2}\|\hat{x}_t\|^2 + 7L_HL_{G,y}\beta_t\sqrt{1-\alpha_t\mu_F}\|\hat{x}_t\|\|\hat{y}_t\| + \frac{16S_H^2\Gamma_{22}^{1+\delta_H}}{\mu_F}\frac{\beta_t^{2+2\delta_H}}{\alpha_t} \\ &\leq (1-\mu_F\alpha_t)\|\hat{x}_t\|^2 + c_{x,1}\beta_t^2\|\hat{y}_t\|^2 + c_{x,2}\beta_t\sqrt{1-\alpha_t\mu_F}\|\hat{x}_t\|\|\hat{y}_t\| + 2\alpha_t^2\Gamma_{11} \\ &\quad + c_{x,3}\beta_t^2 + c_{x,4}\frac{\beta_t^{2+2\delta_H}}{\alpha_t},\end{aligned}$$

where the second inequality uses the fact that $4L_H^2L_{G,x}^2\beta_t^2 \leq \frac{\mu_F\alpha_t}{4}$ and the constants are defined as the last inequality uses the notations in (59). ■

B.2 Proof of Lemma 7

Proof [Proof of Lemma 7] Recall that $\hat{y} = y - y^\star$. Using (2) we consider

$$\hat{y}_{t+1} = \hat{y}_t - \beta_t G(H(y_t), y_t) + \beta_t\left(G(H(y_t), y_t) - G(x_t, y_t)\right) - \beta_t\psi_t, \tag{68}$$

which implies that

$$\begin{aligned}\|\hat{y}_{t+1}\|^2 &= \|\hat{y}_t - \beta_t G(H(y_t), y_t)\|^2 + \|\beta_t\left(G(H(y_t), y_t) - G(x_t, y_t)\right) - \beta_t\psi_t\|^2 \\ &\quad + 2\beta_t\langle\hat{y}_t - \beta_t G(H(y_t), y_t), G(H(y_t), y_t) - G(x_t, y_t)\rangle - 2\beta_t\langle\hat{y}_t - \beta_t G(H(y_t), y_t), \psi_t\rangle.\end{aligned} \tag{69}$$

We next analyze each term on the right-hand side of (69).

For the first term, we have that

$$\begin{aligned}\|\hat{y}_t - \beta_t G(H(y_t), y_t)\|^2 &\overset{(16)}{\leq} \|\hat{y}_t\|^2 - 2\mu_G\beta_t\|\hat{y}_t\|^2 + \beta_t^2\|G(H(y_t), y_t) - G(H(y^\star), y^\star)\|^2 \\ &\leq (1 - 2\mu_G\beta_t)\|\hat{y}_t\|^2 + L_{G,y}^2\beta_t^2\|\hat{y}_t\|^2 \leq (1-\mu_G\beta_t)\|\hat{y}_t\|^2,\end{aligned} \tag{70}$$

where the first inequality also uses $G(H(y^\star), y^\star) = 0$ and the last inequality uses the relation $L_{G,y}^2\beta_t^2 \leq \mu_G\beta_t$.

For the second term, taking the conditional expectation on its both sides w.r.t $\mathcal{F}_t$ and using Assumptions 8 and 2, we have

$$\mathbb{E}\left[\|\beta_t\left(G(H(y_t), y_t) - G(x_t, y_t)\right) - \beta_t\psi_t\|^2 \mid \mathcal{F}_t\right] \leq L_{G,x}^2\beta_t^2\|\hat{x}_t\|^2 + \beta_t^2\Gamma_{22}. \tag{71}$$

For the third term, it follows that

$$\begin{aligned}&2\beta_t\langle \hat{y}_t - \beta_t G(H(y_t), y_t), G(H(y_t), y_t) - G(x_t, y_t)\rangle \\ &\overset{(13)}{\leq} 2\beta_t\|\hat{y}_t - \beta_t G(H(y_t), y_t)\| \cdot L_{G,x}\|\hat{x}_t\| \overset{(70)}{\leq} 2L_{G,x}\beta_t\sqrt{1-\mu_G\beta_t}\|\hat{x}_t\|\|\hat{y}_t\|. \end{aligned} \tag{72}$$

Finally, taking the conditional expectation of (69) w.r.t $\mathcal{F}_t$ and using (70)–(72) yields

$$\mathbb{E}\left[\|\hat{y}_{t+1}\|^2 \,|\, \mathcal{F}_t\right] \leq (1-\mu_G\beta_t)\,\|\hat{y}_t\|^2 + L_{G,x}^2\beta_t^2\|\hat{x}_t\|^2 + \beta_t^2\Gamma_{22} + 2L_{G,x}\beta_t\sqrt{1-\mu_G\beta_t}\|\hat{x}_t\|\|\hat{y}_t\|,$$

which concludes our proof. ■

B.3 Proof of Theorem 8

Before proving Theorem 8, we first present a refined one-step descent lemma by applying Cauchy-Schwarz inequality to the cross term $\mathcal{O}(\beta_t\|\hat{x}_t\|\|\hat{y}_t\|)$.

Lemma 15 (One-step descent lemma) *Under Assumptions 1 – 4, 8 and 9, it follows that*

$$\mathbb{E}[|\hat{x}_{t+1}\|^2|\mathcal{F}_t] \leq \left(1-\frac{\mu_F\alpha_t}{2}\right)\|\hat{x}_t\|^2 + \frac{c_{x,2}^2}{2\mu_F}\frac{\beta_t^2}{\alpha_t}\|\hat{y}_t\|^2 + 2\Gamma_{11}\alpha_t^2 + c_{x,3}\beta_t^2 + c_{x,4}\frac{\beta_t^{2+2\delta_H}}{\alpha_t}, \tag{73}$$

$$\mathbb{E}[\|\hat{y}_{t+1}\|^2|\mathcal{F}_t] \leq \left(1-\frac{\mu_G\beta_t}{2}\right)\|\hat{y}_t\|^2 + \frac{2L_{G,x}^2}{\mu_G}\beta_t\|\hat{x}_t\|^2 + \Gamma_{22}\beta_t^2, \tag{74}$$

where $c_{x,2}$ and $c_{x,3}$ are defined in Lemma 6.

Proof [Proof of Lemma 15] By Lemma 6, using the Cauchy-Schwarz inequality, we have

$$\begin{aligned}\mathbb{E}\left[\|\hat{x}_{t+1}\|^2 \,|\, \mathcal{F}_t\right] &\leq \left(1-\frac{\mu_F\alpha_t}{2}\right)\|\hat{x}_t\|^2 + \left(c_{x,1}\beta_t^2 + \frac{c_{x,2}^2\beta_t^2(1-\mu_F\alpha_t)}{2\mu_F\alpha_t}\right)\|\hat{y}_t\|^2 + 2\Gamma_{11}\alpha_t^2 \\ &\quad + c_{x,3}\beta_t^2 + c_{x,4}\frac{\beta_t^{2+2\delta_H}}{\alpha_t} \\ &\leq \left(1-\frac{\mu_F\alpha_t}{2}\right)\|\hat{x}_t\|^2 + \frac{c_{x,2}^2}{2\mu_F}\frac{\beta_t^2}{\alpha_t}\|\hat{y}_t\|^2 + 2\Gamma_{11}\alpha_t^2 + c_{x,3}\beta_t^2 + c_{x,4}\frac{\beta_t^{2+2\delta_H}}{\alpha_t},\end{aligned}$$

where the last inequality uses the fact that $2c_{x,1} - c_{x,2}^2 \leq 0$. Similarly, by Lemma 7, using the Cauchy-Schwarz inequality, we have

$$\begin{aligned}\mathbb{E}\left[\|\hat{y}_{t+1}\|^2 \,|\, \mathcal{F}_t\right] &\leq \left(1-\frac{\mu_G\beta_t}{2}\right)\|\hat{y}_t\|^2 + \left(L_{G,x}^2\beta_t^2 + \frac{2L_{G,x}^2(1-\mu_G\beta_t)\beta_t}{\mu_G}\right)\|\hat{x}_t\|^2 + \Gamma_{22}\beta_t^2 \\ &\leq \left(1-\frac{\mu_G\beta_t}{2}\right)\|\hat{y}_t\|^2 + \frac{2L_{G,x}^2}{\mu_G}\beta_t\|\hat{x}_t\|^2 + \Gamma_{22}\beta_t^2.\end{aligned}$$

We then complete the proof. ■

Lemma 15 is akin to Doan (2022, Lemmas 1 and 2). While there are minor differences, the key distinction lies in the incorporation of δ_H in the term $\mathcal{O}(\beta_t^{2+2\delta_H}/\alpha_t)$, which is $\mathcal{O}(\beta_t^2/\alpha_t)$ in Doan (2022). Combining (73) and (74) through the careful construction of Lyapunov functions, we can establish both almost sure convergence and L_2-convergence.

Proof [Proof of Theorem 8] We first present the specific forms of the constants:

$$c_{y,1} = \frac{8\Gamma_{22}(2L_H L_{G,x} + 3L_{G,y})}{3\mu_G L_{G,y}}, \quad c_{y,2} = \frac{1280 L_{G,x}^2 S_H^2 \Gamma_{22}^{1+\delta_H}}{\mu_F^2 \mu_G^2}, \tag{75}$$

$$\begin{aligned} c_{x,5} &= \frac{1120 L_H L_{G,x} L_{G,y} \Gamma_{11}}{\mu_F^2 \mu_G}, \qquad c_{x,6} = \frac{21\mu_G L_H L_{G,y} c_{y,1}}{4\mu_F L_{G,x}} + \frac{8L_H^2 \Gamma_{22}}{\mu_F}, \\ c_{x,7} &= \frac{7\mu_G^2 c_{y,2}}{2L_{G,x}^2} + \frac{48 S_H^2 \Gamma_{22}^{1+\delta_H}}{\mu_F^2}. \end{aligned} \tag{76}$$

We characterize the L_2-convergence rates by introducing the Lyapunov function $\widetilde{U}_t = \varrho_1 \frac{\beta_t}{\alpha_t}\|\hat{x}_t\|^2 + \|\hat{y}_t\|^2$ with $\varrho_1 = \frac{8L_{G,x}^2}{\mu_F \mu_G}$. Note that $\frac{\beta_t}{\alpha_t} \le \frac{\mu_G \mu_F}{4c_{x,2} L_{G,x}} \implies \frac{\mu_G}{4}\beta_t \ge \varrho_1 \frac{c_{x,2}^2}{2\mu_F}\frac{\beta_t^3}{\alpha_t^2}$. Since $\frac{\beta_{t+1}}{\alpha_{t+1}} \le \frac{\beta_t}{\alpha_t}$, it follows that

$$\begin{aligned} \mathbb{E}\widetilde{U}_{t+1} &\le \varrho_1 \frac{\beta_t}{\alpha_t} \cdot \left[\left(1 - \frac{\mu_F \alpha_t}{2}\right) \mathbb{E}\|\hat{x}_t\|^2 + \frac{c_{x,2}^2}{2\mu_F}\frac{\beta_t^2}{\alpha_t}\mathbb{E}\|\hat{y}_t\|^2 + \left(2\Gamma_{11}\alpha_t^2 + c_{x,3}\beta_t^2 + c_{x,4}\frac{\beta_t^{2+2\delta_H}}{\alpha_t}\right)\right] \\ &\quad + \left(1 - \frac{\mu_G \beta_t}{2}\right)\mathbb{E}\|\hat{y}_t\|^2 + \frac{2L_{G,x}^2}{\mu_G}\beta_t \mathbb{E}\|\hat{x}_t\|^2 + \Gamma_{22}\beta_t^2 \\ &\le \left(1 - \frac{\mu_G \beta_t}{4}\right)\mathbb{E}\widetilde{U}_t + 2\varrho_1 \Gamma_{11}\alpha_t \beta_t + \left(\Gamma_{22} + \frac{c_{x,3} L_{G,x}}{3L_H L_{G,y}}\right)\beta_t^2 + \varrho_1 c_{x,4}\frac{\beta_t^{3+2\delta_H}}{\alpha_t^2}, \end{aligned} \tag{77}$$

where the last inequality uses $\frac{\beta_t}{\alpha_t} \le \kappa \le \frac{\mu_F}{\mu_G}$. For ease of notation, we set $\alpha_{j,t} = \prod_{\tau=j}^t \left(1 - \frac{\mu_F \alpha_\tau}{2}\right)$ and $\beta_{j,t} = \prod_{\tau=j}^t \left(1 - \frac{\mu_G \beta_\tau}{4}\right)$. Then iterating (77) yields

$$\begin{aligned} \mathbb{E}\widetilde{U}_{t+1} &\le \beta_{0,t}\mathbb{E}\widetilde{U}_0 + 2\varrho_1\Gamma_{11} \cdot \sum_{\tau=0}^t \beta_{\tau+1,t}\alpha_\tau\beta_\tau + \left(\Gamma_{22} + \frac{c_{x,3}L_{G,x}}{3L_H L_{G,y}}\right) \cdot \sum_{\tau=0}^t \beta_{\tau+1,t}\beta_\tau^2 \\ &\quad + \varrho_1 c_{x,4} \cdot \sum_{\tau=0}^t \beta_{\tau+1,t}\frac{\beta_\tau^{3+2\delta_H}}{\alpha_\tau^2} \\ &\le \beta_{0,t}\mathbb{E}\widetilde{U}_0 + \frac{16\varrho_1\Gamma_{11}}{\mu_G}\alpha_t + \frac{8\Gamma_{22}(2L_H L_{G,x} + 3L_{G,y})}{3\mu_G L_{G,y}}\beta_t + \frac{10\varrho_1 c_{x,4}}{\mu_G}\frac{\beta_t^{2+2\delta_H}}{\alpha_t^2}. \end{aligned} \tag{78}$$

where the last inequality follows from Lemma 16, whose proof is deferred to Appendix B.4.

Lemma 16 (Step sizes inequalities) *Under Assumption 9, it follows that*

(i) $\sum_{\tau=0}^t \beta_{\tau+1,t}\beta_\tau^2 \le \frac{8\beta_t}{\mu_G}$ *and* $\sum_{\tau=0}^t \alpha_{\tau+1,t}\alpha_\tau^2 \le \frac{4\alpha_t}{\mu_F}$.

(ii) $\sum_{\tau=0}^t \alpha_{\tau+1,t}\beta_\tau^2 \le \frac{4}{\mu_F}\frac{\beta_t^2}{\alpha_t}$ *and* $\sum_{\tau=0}^t \alpha_{\tau+1,t}\beta_\tau\beta_{\tau-1} \le \frac{6}{\mu_F}\frac{\beta_t^2}{\alpha_t}$.

(iii) $\sum_{\tau=0}^{t}\beta_{\tau+1,t}\alpha_\tau\beta_\tau \le \frac{8\alpha_t}{\mu_G}$ *and* $\sum_{\tau=0}^{t}\alpha_{\tau+1,t}\beta_\tau\alpha_{\tau-1} \le \frac{10}{\mu_F}\beta_t$.

(iv) $\sum_{\tau=0}^{t}\alpha_{\tau+1,t}\beta_\tau\beta_{0,\tau-1} \le \frac{8}{\mu_G}\beta_{0,t}$.

(v) $\sum_{\tau=0}^{t}\beta_{\tau+1,t}\frac{\beta_\tau^{3+2\delta_H}}{\alpha_\tau^2} \le \frac{10}{\mu_G}\frac{\beta_t^{2+2\delta_H}}{\alpha_t^2}$.

(vi) $\sum_{\tau=0}^{t}\alpha_{\tau+1,t}\frac{\beta_\tau^{2+2\delta_H}}{\alpha_\tau} \le \frac{3}{\mu_F}\frac{\beta_t^{2+2\delta_H}}{\alpha_t^2}$ *and* $\sum_{\tau=0}^{t}\alpha_{\tau+1,t}\beta_\tau\frac{\beta_{\tau-1}^{1+2\delta_H}}{\alpha_{\tau-1}} \le \frac{4}{\mu_F}\frac{\beta_t^{2+2\delta_H}}{\alpha_t^2}$.

As a result of (78), it follows that

$$\mathbb{E}\|\hat{y}_{t+1}\|^2 \le \mathbb{E}\widetilde{U}_{t+1} \le \beta_{0,t}\mathbb{E}\widetilde{U}_0 + \frac{16\varrho_1\Gamma_{11}}{\mu_G}\alpha_t + c_{y,1}\beta_t + c_{y,2}\frac{\beta_t^{2+2\delta_H}}{\alpha_t^2}, \tag{79}$$

where $c_{y,1}$ and $c_{y,2}$ are defined in (75). Recall that

$$\mathbb{E}\|\hat{x}_{t+1}\|^2 \le \left(1-\frac{\mu_F\alpha_t}{2}\right)\|\hat{x}_t\|^2 + \frac{c_{x,2}^2}{2\mu_F}\frac{\beta_t^2}{\alpha_t}\|\hat{y}_t\|^2 + 2\Gamma_{11}\alpha_t^2 + c_{x,3}\beta_t^2 + c_{x,4}\frac{\beta_t^{2+2\delta_H}}{\alpha_t}.$$

Iterating this inequality yields that

$$\begin{aligned}
\mathbb{E}\|\hat{x}_{t+1}\|^2 &\le \alpha_{0,t}\mathbb{E}\|\hat{x}_0\|^2 + \frac{c_{x,2}^2\kappa}{2\mu_F}\sum_{\tau=0}^{t}\alpha_{\tau+1,t}\beta_\tau\mathbb{E}\|\hat{y}_\tau\|^2 + \frac{8\Gamma_{11}}{\mu_F}\alpha_t + \frac{4c_{x,3}}{\mu_F}\frac{\beta_t^2}{\alpha_t} + \frac{3c_{x,4}}{\mu_F}\frac{\beta_t^{2+2\delta_H}}{\alpha_t^2} \\
&\overset{(79)}{\le} \frac{c_{x,2}^2\kappa}{2\mu_F}\sum_{\tau=0}^{t}\alpha_{\tau+1,t}\beta_\tau\left(\beta_{0,\tau-1}\mathbb{E}\widetilde{U}_0 + \frac{16\varrho_1\Gamma_{11}}{\mu_G}\alpha_{\tau-1} + c_{y,1}\beta_{\tau-1} + c_{y,2}\frac{\beta_{\tau-1}^{2+2\delta_H}}{\alpha_{\tau-1}^2}\right) \\
&\quad + \alpha_{0,t}\mathbb{E}\|\hat{x}_0\|^2 + \frac{8\Gamma_{11}}{\mu_F}\alpha_t + \frac{4c_{x,3}}{\mu_F}\frac{\beta_t^2}{\alpha_t} + \frac{3c_{x,4}}{\mu_F}\frac{\beta_t^{2+2\delta_H}}{\alpha_t^2} \\
&\le \frac{7L_HL_{G,y}}{L_{G,x}}\beta_{0,t}\mathbb{E}\widetilde{U}_0 + \alpha_{0,t}\mathbb{E}\|\hat{x}_0\|^2 + \frac{8\Gamma_{11}}{\mu_F}\alpha_t + c_{x,5}\beta_t + c_{x,6}\frac{\beta_t^2}{\alpha_t} + c_{x,7}\frac{\beta_t^{2+2\delta_H}}{\alpha_t^2},
\end{aligned} \tag{80}$$

where the first and last inequality use Lemma 16, $\frac{\beta_\tau}{\alpha_\tau} \le \kappa$, $\kappa \le \frac{\mu_F\mu_G}{28L_HL_{G,x}L_{G,y}}$ and the constants are defined in (76).

Note that

$$\mathbb{E}\widetilde{U}_0 = \frac{8L_{G,x}^2\beta_0}{\mu_F\mu_G\alpha_0}\mathbb{E}\|\hat{x}_0\|^2 + \mathbb{E}\|\hat{y}_0\|^2 \le \frac{8L_{G,x}^2\kappa}{\mu_F\mu_G}\mathbb{E}\|\hat{x}_0\|^2 + \mathbb{E}\|\hat{y}_0\|^2 = \frac{2L_{G,x}\mathbb{E}\|\hat{x}_0\|^2}{7L_HL_{G,y}} + \mathbb{E}\|\hat{y}_0\|^2. \tag{81}$$

Using (81), we then simplify (80) into

$$\mathbb{E}\|\hat{x}_{t+1}\|^2 \le (\alpha_{0,t}+2\beta_{0,t})\,\mathbb{E}\|\hat{x}_0\|^2 + \frac{7L_HL_{G,y}}{L_{G,x}}\beta_{0,t}\mathbb{E}\|\hat{y}_0\|^2 + \frac{8\Gamma_{11}}{\mu_F}\alpha_t + c_{x,5}\beta_t + c_{x,6}\frac{\beta_t^2}{\alpha_t} + c_{x,7}\frac{\beta_t^{2+2\delta_H}}{\alpha_t^2},$$

which together with $\alpha_{0,t} \le \beta_{0,t}$ completes the proof. ■

B.4 Proof of Lemma 16

Proof [Proof of Lemma 16]

(i) This mainly follows from (i) in Lemma 17.

(ii) We first show that $\left(\frac{\beta_{t-1}}{\beta_t}\right)^2 \le 1+\frac{\mu_F}{8\kappa}\beta_t$. This is easy to prove by using $\left(\frac{\beta_{t-1}}{\beta_t}\right)^2 \le \left(1+\frac{\mu_G}{32}\beta_t\right)^2 \le \left(1+\frac{\mu_F}{32\kappa}\beta_t\right)^2 \le 1+\frac{\mu_F}{8\kappa}\beta_t$. where the second inequality uses $\kappa \le \frac{\mu_F}{\mu_G}$ and the last inequality uses $\beta_t \le \iota_2 < \frac{64\kappa}{\mu_F}$ and $\frac{4\kappa}{\mu_F} = \frac{\iota_2}{3}$. Then the first inequality follows from (iii) in Lemma 17. For the second inequality, note that $\sum_{\tau=0}^t \alpha_{\tau+1,t}\beta_\tau\beta_{\tau-1} \le \left(1+\frac{\mu_G\iota_2}{32}\right)\sum_{\tau=0}^t \alpha_{\tau+1,t}\beta_\tau^2 \le \left(1+\frac{\mu_G\iota_2}{32}\right)\frac{4}{\mu_F}\frac{\beta_t^2}{\alpha_t} \le \frac{6}{\mu_F}\frac{\beta_t^2}{\alpha_t}$.

(iii) Note that $\frac{\beta_t}{\alpha_t} \le \frac{\mu_F}{\mu_G}$, $1-\frac{\mu_F\alpha_t}{2} \le 1-\frac{\mu_G\beta_t}{2} \le \left(1-\frac{\mu_G}{4}\beta_t\right)^2$ and $1-\frac{\mu_G}{4}\beta_t \ge 1-\frac{\mu_G}{2}\iota_2 \ge 1-\frac{1}{2}=\frac{1}{2}$. It follows that

$$\sum_{\tau=0}^t \alpha_{\tau+1,t}\beta_\tau\beta_{0,\tau-1} \le \sum_{\tau=0}^t \beta_{\tau+1,t}^2\beta_\tau\beta_{0,\tau-1} \le 2\beta_{0,t}\sum_{\tau=0}^t \beta_{\tau+1,t}\beta_\tau \le \frac{8\beta_{0,t}}{\mu_G}.$$

The proof of remaining parts follows a similar pattern. ■

Lemma 17 *Let $\{\alpha_t,\beta_t\}$ be nonincreasing positive numbers.*

(i) If $\beta_t \le \frac{1}{a}$ and $\frac{\beta_{t-1}}{\beta_t} \le 1+\frac{a}{2}\beta_t$ for all $t \ge 1$ and some $a>0$, $\sum_{j=0}^t \beta_j^2\prod_{\tau=j+1}^t(1-a\beta_\tau) \le \frac{2}{a}\beta_t$.

(ii) If $\beta_t \le \kappa\alpha_t, \alpha_t \le \frac{1}{a}$, and $\left(\frac{\beta_{t-1}}{\beta_t}\right)^2 \le 1+\frac{a}{2\kappa}\beta_t$ for all $t \ge 1$ and some $a>0$, $\sum_{j=0}^t \beta_j^2\prod_{\tau=j+1}^t(1-a\alpha_\tau) \le \frac{2}{a}\frac{\beta_t^2}{\alpha_t}$.

(iii) If $\beta_t \le \frac{1}{a}$ and $\frac{\alpha_{t-1}}{\alpha_t} \le 1+\frac{a}{2}\beta_t$ for all $t \ge 1$ and some $a>0$, $\sum_{j=0}^t \beta_j\alpha_j\prod_{\tau=j+1}^t(1-a\beta_\tau) \le \frac{2}{a}\alpha_t$.

Proof [Proof of Lemma 17] See the proof of Kaledin et al. (2020, Lemma 14). ■

Appendix C. Omitted Proofs in Section 4

In this section, we give the detailed proof of Theorem 3. For generality, with δ_H define in Assumption 4, we assume $\delta_H \in (0,1]$ instead of $\delta_H \in [0.5,1]$ throughout this section. We first present the formal version of Theorem 3.

Theorem 18 (Formal version of Theorem 3) *Suppose that Assumptions 1–7 hold. In particular, we assume $\delta_H \in (0,1]$ instead of $\delta_H \in [0.5,1]$ in Assumption 4. Then we have for all $t \ge 0$,*

$$\mathbb{E}\|\hat{x}_{t+1}\|^2 \le C_{x,0}^{de}\prod_{\tau=0}^t\left(1-\frac{\mu_G\beta_\tau}{4}\right)+C_{x,1}^{de}\,\alpha_t+C_{x,2}^{de}\frac{\beta_t^{2+2\delta_H}}{\alpha_t^2}, \tag{82}$$

$$\|\mathbb{E}\hat{x}_{t+1}\hat{y}_{t+1}^\top\| \le C_{xy,0}^{de}\prod_{\tau=0}^{t}\left(1-\frac{\mu_G\beta_\tau}{4}\right)+C_{xy,1}^{de}\,\beta_t+C_{xy,2}^{de}\frac{\beta_t^{1+2\delta_H}}{\alpha_t} + \left(C_{xy,3}^{de}\,\alpha_t\beta_t+C_{xy,4}^{de}\frac{\beta_t^{5+4\delta_H}}{\alpha_t^5}\right)\left(\frac{\alpha_t}{\beta_t}\right)^{\frac{2}{\delta_F}}, \tag{83}$$

$$\begin{aligned}\mathbb{E}\|\hat{y}_{t+1}\|^2 \le{}& C_{y,0}^{de}\prod_{\tau=0}^{t}\left(1-\frac{\mu_G\beta_\tau}{4}\right)+C_{y,1}^{de}\,\beta_t+C_{y,2}^{de}\frac{\beta_t^{1+2\delta_H}}{\alpha_t}\\ &+\left(C_{y,3}^{de}\,\alpha_t\beta_t+C_{y,4}^{de}\frac{\beta_t^{5+4\delta_H}}{\alpha_t^5}\right)\left(\frac{\alpha_t}{\beta_t}\right)^{\frac{2}{\delta_F}}\\ &+\left(C_{y,5}^{de}\,\alpha_t\beta_t+C_{y,6}^{de}\frac{\beta_t^{5+4\delta_H}}{\alpha_t^5}\right)\left(\frac{\alpha_t}{\beta_t}\right)^{\frac{1}{\delta_G}},\end{aligned} \tag{84}$$

where $\{C_{x,i}^{de}\}_{i\in[2]\cup\{0\}}$, $\{C_{xy,i}^{de}\}_{i\in[4]\cup\{0\}}$ *and* $\{C_{y,i}^{de}\}_{i\in[6]\cup\{0\}}$ *are problem-dependent constants defined in* (143), (161) *and* (166).

Under Assumption 7 (iii), we have

$$C_{prod} := \sup_{t\ge 0}\frac{\prod_{\tau=0}^{t}\left(1-\frac{\mu_G\beta_\tau}{4}\right)}{\alpha_t^2} < \infty. \tag{85}$$

With $\delta_H \ge 0.5$ and the constants bounds in Assumption 7 (i), the constants in (25), (26) and (27) can be defined as

$$\begin{aligned}\mathcal{C}_x &= C_{x,0}^{de}\,C_{prod}\,\iota_1 + C_{x,1}^{de} + C_{x,2}^{de}\,\iota_2^{2\delta_H-1}\kappa^3,\\ \mathcal{C}_{xy,1} &= C_{xy,1}^{de} + C_{xy,2}^{de}\,\iota_2^{2\delta_H-1}\kappa, \quad \mathcal{C}_{xy,2} = C_{xy,0}^{de}\,C_{prod}\,\kappa^{\frac{2}{\delta_F}-1} + C_{xy,3}^{de} + C_{xy,4}^{de}\,\iota_2^{4\delta_H-2}\,\kappa^6,\\ \mathcal{C}_{y,1} &= C_{y,1}^{de} + C_{y,2}^{de}\,\iota_2^{2\delta_H-1}\kappa, \quad \mathcal{C}_{y,2} = C_{y,0}^{de}\,C_{prod}\,\kappa^{\frac{2}{\delta_F}-1} + C_{y,3}^{de} + C_{y,4}^{de}\,\iota_2^{4\delta_H-2}\,\kappa^6,\\ \mathcal{C}_{y,3} &= C_{y,5}^{de} + C_{y,6}^{de}\,\iota_2^{4\delta_H-2}\,\kappa^6.\end{aligned} \tag{86}$$

Next, we give an example choice of the step sizes in Corollary 4.

$$\begin{aligned}&\alpha_t = \frac{128}{(\delta_F\wedge\delta_G)\mu_G\kappa(t+T_0)^a} \text{ and } \beta_t = \frac{128}{(\delta_F\wedge\delta_G)\mu_G(t+T_0)^b}\\ &\text{with } a,b\in(0,1],\ 1\le\frac{b}{a}\le 1+\frac{\delta_F}{2}\wedge\delta_G, \mathcal{T}_1 = \frac{128}{\mu_G(\kappa\iota_1\wedge\iota_2\wedge\frac{\rho}{\kappa})} \text{ and } T_0 \ge \left\lceil \mathcal{T}_1^{1/a}\right\rceil.\end{aligned} \tag{87}$$

We emphasize that in the proof, our primary focus is on the order of the mean squared error, rather than optimizing the constants. The constants in (86) serve only to provide an upper bound that ensures Theorem 18 holds, and may be quite loose in certain cases, such as SGD with averaging in (4). The step size choice in (87) is thus presented as a feasible example for general cases. For specific examples, however, the optimal values of α_0, β_0 and T_0, can vary significantly. For example, in (4), β_0 and T_0 can both be set as 1.

Before present the detailed proof, we first give an operator decomposition lemma, which states that with Assumption 5, we could decompose the nonlinear operators F and G as the linear parts plus higher-order error terms.

Lemma 19 (Operator decomposition) *Suppose that Assumption 5 holds. With $\hat{x}_t$ and $\hat{y}_t$ defined in (24), we have the following results.*

(i) $F(x_t, y_t) = B_1\hat{x}_t + R_t^F$ *with* $\|R_t^F\| \le S_{B,F}(\|\hat{x}_t\|^{1+\delta_F} + \|\hat{y}_t\|^{1+\delta_F})$. *It follows that* $\hat{x}_t - \alpha_t F(x_t, y_t) = (I - \alpha_t B_1)\hat{x}_t - \alpha_t R_t^F$. *We further have* $\|I - \alpha_t B_1\| \le 1 - \mu_F\alpha_t$ *if* $0 \le \alpha_t \le \frac{\mu_F}{L_F^2}$.

(ii) $G(H(y_t), y_t) = B_3\hat{y}_t + R_t^{GH}$ *with* $\|R_t^{GH}\| \le S_{B,G}\|\hat{y}_t\|^{1+\delta_G}$. *It follows that* $\hat{y}_t - \beta_t G(H(y_t), y_t) = (I - \beta_t B_3)\hat{y}_t - \beta_t R_t^{GH}$. *We further have* $\|I - \beta_t B_3\| \le 1 - \mu_G\beta_t$ *if* $0 \le \beta_t \le \frac{\mu_G}{L_{G,y}^2}$.

(iii) $G(x_t, y_t) = B_2\hat{x}_t + B_3\hat{y}_t + R_t^G$ *with* $\|R_t^G\| \le S_{B,G}(\|\hat{x}_t\|^{1+\delta_G} + \|\hat{y}_t\|^{1+\delta_G})$.

C.1 Proof of Lemma 9

Proof [Proof of Lemma 9] We first present the specific forms of constants:

$$\begin{aligned}
&c_{x,1}^{de} = 4L_H^2L_{G,y}^2,\ c_{x,2}^{de} = 2d_xL_HL_{G,y},\ c_{x,3}^{de} = 2L_H^2\Gamma_{22},\ c_{x,4}^{de} = \frac{96d_xS_H^2\Gamma_{22}^{1+\delta_H}}{\mu_F},\\
&c_{x,5}^{de} = 4d_x\widetilde{S}_H(L_{G,x}\wedge L_{G,y}) \propto S_H,\ c_{x.6}^{de} = 4d_xS_{B,F}L_H(L_{G,x}\wedge L_{G,y}) \propto S_{B,F},\\
&c_{x,7}^{de} = \frac{d_xL_HS_{B,G}^2}{L_{G,x}} \propto S_{B,G}^2,\ c_{x,8}^{de} = \frac{8d_xS_H^2(L_{G,x}^{2+2\delta_H}\wedge L_{G,y}^{2+2\delta_H})}{L_H^{1+\delta_H}L_{G,x}^{1+\delta_H}} \propto S_H^2.
\end{aligned} \tag{88}$$

The decomposition in (61) implies that

$$\begin{aligned}
\mathbb{E}\|\hat{x}_{t+1}\|^2 \overset{(62)+(63)}{\le}& \left(1 - \frac{3\mu_F\alpha_t}{2}\right)\mathbb{E}\|\hat{x}_t\|^2 + c_{x,1}\beta_t^2\mathbb{E}\|\hat{y}_t\|^2 + 2\beta_t^2L_H^2\Gamma_{22} + 2\alpha_t^2\Gamma_{11}\\
&+ 2\mathbb{E}\underbrace{\langle\hat{x}_t - \alpha_tF(x_t, y_t), H(y_t) - H(y_{t+1})\rangle}_{\spadesuit},
\end{aligned} \tag{89}$$

where we have also used the fact that $4L_H^2L_{G,x}^2\beta_t^2 \le \frac{\mu_F\alpha_t}{4}$. The proof is almost identical to that of Lemma 6 except that we take additional care on the cross term $\mathbb{E}\spadesuit$.

By Lemma 19, we have $\hat{x}_t - \alpha_tF(x_t, y_t) = (1-\alpha_tB_1)\hat{x}_t - \alpha_tR_t^F$ with $\|R_t^F\| \le S_{B,F}(\|\hat{x}_t\|^{1+\delta_F} + \|\hat{y}_t\|^{1+\delta_F})$ and $y_t - y_{t+1} = \hat{y}_t - \hat{y}_{t+1} = \beta_t(B_2\hat{x}_t + B_3\hat{y}_t + R_t^G + \psi_t)$ with $\|R_t^G\| \le S_{B,G}(\|\hat{x}_t\|^{1+\delta_G} + \|\hat{y}_t\|^{1+\delta_G})$. By Assumption 4, it follows that

$$H(y_{t+1}) - H(y_t) = \nabla H(y_t)(y_{t+1} - y_t) + R_t^H = \nabla H(y^\star)(y_{t+1} - y_t) + R_t^{\nabla H} + R_t^H$$

where $\|R_t^H\| \le S_H\|y_{t+1} - y_t\|^{1+\delta_H}$ and $R_t^{\nabla H} = (\nabla H(y_t) - \nabla H(y^\star))(y_{t+1} - y_t)$. Therefore, we have

$$\begin{aligned}
\mathbb{E}\spadesuit &= \mathbb{E}\mathrm{tr}(\spadesuit) = -\mathrm{tr}\left(\mathbb{E}(H(y_{t+1}) - H(y_t))(\hat{x}_t - \alpha_tF(x_t, y_t))^\top\right)\\
&\le d_x\left\|\mathbb{E}(\nabla H(y^\star)(y_{t+1} - y_t) + R_t^{\nabla H} + R_t^H)((I - \alpha_tB_1)\hat{x}_t - \alpha_tR_t^F)^\top\right\|
\end{aligned}$$

$$\begin{aligned}\leq d_x \cdot \Bigg[&\underbrace{\left\|\nabla H(y^\star)\mathbb{E}(y_{t+1}-y_t)\hat{x}_t^\top(1-\alpha_t B_1)^\top\right\|}_{\spadesuit_1} + \underbrace{\mathbb{E}\|R_t^H\|\|\hat{x}_t-\alpha_t F(x_t,y_t)\|}_{\spadesuit_2} \\ &+ \alpha_t \underbrace{\left\|\nabla H(y^\star)\mathbb{E}(y_{t+1}-y_t)(R_t^F)^\top\right\|}_{\spadesuit_3} + \underbrace{\|\mathbb{E}R_t^{\nabla H}\left(\hat{x}_t-\alpha_t F(x_t,y_t)\right)^\top\|}_{\spadesuit_4}\Bigg].\end{aligned} \tag{90}$$

We then analyze the four terms $\{\spadesuit_i\}_{i\in[4]}$ on the right-hand side of the last inequality.

Proposition 20 *For any random variable $X \geq 0$ and real number $a > 0$, it follows that*

$$2\mathbb{E}X^3 \leq a\mathbb{E}X^2 + \frac{1}{a}\mathbb{E}X^4.$$

- For the term $\spadesuit_1$, using $\|I-\alpha_t B_1\| \leq 1-\mu_F\alpha_t \leq 1$ and $\|\nabla H(y^\star)\| \leq L_H$, we have

$$\begin{aligned}\spadesuit_1 &= \beta_t\left\|\mathbb{E}\nabla H(y^\star)(B_2\hat{x}_t+B_3\hat{y}_t+R_t^G)\hat{x}_t^\top(I-\alpha_t B_1)^\top\right\| \\ &\leq \beta_t L_H\left(L_{G,x}\mathbb{E}\|\hat{x}_t\|^2 + L_{G,y}\|\mathbb{E}\hat{y}_t\hat{x}_t^\top\| + \mathbb{E}\|R_t^G\|\|\hat{x}_t\|\right) \\ &\overset{(92)}{\leq} \beta_t L_H\left(2L_{G,x}\mathbb{E}\|\hat{x}_t\|^2 + L_{G,y}\|\mathbb{E}\hat{y}_t\hat{x}_t^\top\| + \frac{S_{B,G}^2}{2L_{G,x}}\left(\mathbb{E}\|\hat{x}_t\|^{2+2\delta_G}+\mathbb{E}\|\hat{y}_t\|^{2+2\delta_G}\right)\right).\end{aligned} \tag{91}$$

  Here the last inequality uses the following result which could be obtained by Proposition 20,

$$\mathbb{E}\|R_t^G\|\|\hat{x}_t\| \leq L_{G,x}\mathbb{E}\|\hat{x}_t\|^2 + \frac{S_{B,G}^2}{2L_{G,x}}\left(\mathbb{E}\|\hat{x}_t\|^{2+2\delta_G}+\mathbb{E}\|\hat{y}_t\|^{2+2\delta_G}\right). \tag{92}$$

- For the term $\spadesuit_2$, it follows that

$$\begin{aligned}\spadesuit_2 &\leq S_H\cdot\mathbb{E}\|y_{t+1}-y_t\|^{1+\delta_H}\|\hat{x}_t-\alpha_t F(x_t,y_t)\| \\ &\overset{(62)+(64)}{\leq} 2S_H\beta_t^{1+\delta_H}\left(2L_{G,x}^{1+\delta_H}\mathbb{E}\|\hat{x}_t\|^{2+\delta_H} + 2L_{G,y}^{1+\delta_H}\mathbb{E}\|\hat{y}_t\|^{1+\delta_H}\|\hat{x}_t\| + \Gamma_{22}^{\frac{1+\delta_H}{2}}\mathbb{E}\|\hat{x}_t\|\right) \\ &\leq 2L_H^{1+\delta_H}L_{G,x}^{1+\delta_H}\beta_t^{1+\delta_H}\mathbb{E}\|\hat{x}_t\|^2 + \frac{4S_H^2\beta_t^{1+\delta_H}}{L_H^{1+\delta_H}L_{G,x}^{1+\delta_H}}(L_{G,x}^{2+2\delta_H}\mathbb{E}\|\hat{x}_t\|^{2+2\delta_H} + L_{G,y}^{2+2\delta_H}\mathbb{E}\|\hat{y}_t\|^{2+2\delta_H}) \\ &\quad + \frac{\mu_F\alpha_t}{12d_x}\mathbb{E}\|\hat{x}_t\|^2 + \frac{48d_xS_H^2\Gamma_{22}^{1+\delta_H}}{\mu_F}\frac{\beta_t^{2+2\delta_H}}{\alpha_t} \\ &\leq \frac{\mu_F\alpha_t}{6d_x}\mathbb{E}\|\hat{x}_t\|^2 + \frac{4S_H^2\beta_t^{1+\delta_H}}{L_H^{1+\delta_H}L_{G,x}^{1+\delta_H}}\left(L_{G,x}^{2+2\delta_H}\mathbb{E}\|\hat{x}_t\|^{2+2\delta_H} + L_{G,y}^{2+2\delta_H}\mathbb{E}\|\hat{y}_t\|^{2+2\delta_H}\right) \\ &\quad + \frac{48d_xS_H^2\Gamma_{22}^{1+\delta_H}}{\mu_F}\frac{\beta_t^{2+2\delta_H}}{\alpha_t},\end{aligned} \tag{93}$$

  where the second inequality also use $(a+b)^{1+\delta} \leq 2(a^{1+\delta}+b^{1+\delta})$ for any $a, b \geq 0$ and $\delta \in [0,1]$, the third inequality uses Cauchy–Schwarz inequality, and the last inequality uses $(L_H L_{G,x}\beta_t)^{1+\delta_H} \leq L_H L_{G,x}\beta_t \leq \frac{\mu_F\alpha_t}{24d_x} < 1$.

- For the term $\spadesuit_3$, it follows that

$$\begin{aligned}\spadesuit_3 &\le \beta_t L_H S_{B,F} \mathbb{E}\|G(x_t,y_t)\| \cdot \left(\|\hat{x}_t\|^{1+\delta_F} + \|\hat{y}_t\|^{1+\delta_F}\right) \\ &\overset{(64)}{\le} \beta_t L_H S_{B,F} \mathbb{E}\left(L_{G,x}\|\hat{x}_t\| + L_{G,y}\|\hat{y}_t\|\right) \cdot \left(\|\hat{x}_t\|^{1+\delta_F} + \|\hat{y}_t\|^{1+\delta_F}\right) \\ &\le 2\beta_t S_{B,F} L_H (L_{G,x} \wedge L_{G,y}) \left(\mathbb{E}\|\hat{x}_t\|^{2+\delta_F} + \mathbb{E}\|\hat{y}_t\|^{2+\delta_F}\right). \end{aligned} \tag{94}$$

- For the term $\spadesuit_4$, by the definition of $R_t^{\nabla H}$, it follows that

$$\begin{aligned}\spadesuit_4 &\overset{(17)+(62)}{\le} \beta_t \widetilde{S}_H \mathbb{E}\|\hat{y}_t\|^{\delta_H} \|G(x_t,y_t)\| \|\hat{x}_t\| \\ &\overset{(64)}{\le} \beta_t \widetilde{S}_H \left(L_{G,x}\mathbb{E}\|\hat{x}_t\|^2 \|\hat{y}_t\|^{\delta_H} + L_{G,y}\mathbb{E}\|\hat{y}_t\|^{1+\delta_H}\|\hat{x}_t\|\right) \\ &\le 2\beta_t \widetilde{S}_H (L_{G,x} \wedge L_{G,y}) \left(\mathbb{E}\|\hat{x}_t\|^{2+\delta_H} + \mathbb{E}\|\hat{y}_t\|^{2+\delta_H}\right), \end{aligned} \tag{95}$$

  where the last step uses Young's inequality.

Plugging (91), (93), (94), and (95) into (90), we have that

$$\mathbb{E}\spadesuit \overset{(88)+(37)}{\le} \frac{\mu_F \alpha_t}{4}\mathbb{E}\|\hat{x}_t\|^2 + \frac{c_{x,2}^{de}\beta_t}{2}\|\mathbb{E}\hat{x}_t\hat{y}_t^\top\| + \frac{c_{x,4}^{de}}{2}\frac{\beta_t^{2+2\delta_H}}{\alpha_t} + \frac{\Delta_{x,t}}{2}, \tag{96}$$

where the inequality also uses $L_H L_{G,x}\beta_t \le \frac{\mu_F\alpha_t}{24 d_x}$ and we use the constants $c_{x,5}^{de}$ to $c_{x,8}^{de}$ defined in (88) to hide the problem-dependent coefficients to yield the expression of $\Delta_{x,t}$ in (37). Moreover, by Proposition 2, we have $c_{x,5}^{de} \propto S_H$. Then plugging (96) into (89) gives

$$\begin{aligned}\mathbb{E}\|\hat{x}_{t+1}\|^2 &\overset{(89)}{\le} \left(1 - \frac{3\mu_F\alpha_t}{2}\right)\mathbb{E}\|\hat{x}_t\|^2 + c_{x,1}^{de}\beta_t^2\mathbb{E}\|\hat{y}_t\|^2 + 2\beta_t^2 L_H^2 \Gamma_{22} + 2\alpha_t^2\Gamma_{11} + 2\mathbb{E}\spadesuit \\ &\overset{(88)+(96)}{\le} (1-\mu_F\alpha_t)\,\mathbb{E}\|\hat{x}_t\|^2 + c_{x,1}^{de}\beta_t^2\mathbb{E}\|\hat{y}_t\|^2 + c_{x,2}^{de}\beta_t\|\mathbb{E}\hat{x}_t\hat{y}_t^\top\| + 2\Gamma_{11}\alpha_t^2 \\ &\quad + c_{x,3}^{de}\beta_t^2 + c_{x,4}^{de}\frac{\beta_t^{2+2\delta_H}}{\alpha_t} + \Delta_{x,t}, \end{aligned}$$

which is the desired result. ■

C.2 Proof of Lemma 10

Proof [Proof of Lemma 10] The decomposition in (69) implies that

$$\begin{aligned}\mathbb{E}\|\hat{y}_{t+1}\|^2 &\overset{(70)+(71)}{\le} (1-\mu_G\beta_t)\mathbb{E}\|\hat{y}_t\|^2 + L_{G,x}^2\beta_t^2\mathbb{E}\|\hat{x}_t\|^2 + \beta_t^2\Gamma_{22} \\ &\quad + 2\beta_t\mathbb{E}\langle \hat{y}_t - \beta_t G(H(y_t),y_t), G(H(y_t),y_t) - G(x_t,y_t)\rangle \end{aligned} \tag{97}$$

The proof is almost identical to that of Lemma 7 except that we take additional care on the last term.

By Lemma 19, we have

$$\begin{aligned}\hat{y}_t - \beta_t G(H(y_t), y_t) &= (I - \beta_t B_3)\hat{y}_t - \beta_t R_t^{GH} \\ G(H(y_t), y_t) &= B_3\hat{y}_t + R_t^{GH} \\ G(x_t, y_t) &= B_2\hat{x}_t + B_3\hat{y}_t + R_t^{G}\end{aligned}$$

with $\|R_t^{GH}\| \le S_{B,G}\|\hat{y}_t\|^{1+\delta_G}$ and $\|R_t^{G}\| \le S_{B,G}(\|\hat{x}_t\|^{1+\delta_G} + \|\hat{y}_t\|^{1+\delta_G})$. Therefore, we have

$$\begin{aligned}&\mathbb{E}\langle \hat{y}_t - \beta_t G(H(y_t), y_t), G(H(y_t), y_t) - G(x_t, y_t)\rangle \\ &\le d_y \left\| \mathbb{E}((I - \beta_t B_3)\hat{y}_t - \beta_t R_t^{GH})(-B_2\hat{x}_t + R_t^{GH} - R_t^{G})^\top \right\| \\ &\le d_y \Big[(1 - \mu_G\beta_t)\|B_2\| \|\mathbb{E}\hat{y}_t\hat{x}_t^\top\| + (1 - \mu_G\beta_t)\mathbb{E}\|\hat{y}_t\|(\|R_t^{GH}\| + \|R_t^{G}\|) \\ &\quad + \beta_t\|B_2\|\mathbb{E}\|R_t^{GH}\|\|\hat{x}_t\| + \beta_t\mathbb{E}\|R_t^{GH}\|(\|R_t^{GH}\| + \|R_t^{G}\|)\Big] \\ &\le d_y L_{G,x}\|\mathbb{E}\hat{x}_t\hat{y}_t^\top\| + \frac{\mu_G}{6}\mathbb{E}\|\hat{y}_t\|^2 + \frac{L_{G,x}^2}{2}\beta_t\mathbb{E}\|\hat{x}_t\|^2 \\ &\quad + S_{B,G}^2\left(\frac{15d_y^2}{2\mu_G} + \frac{d_y^2}{2}\beta_t + 4d_y\beta_t\right)\left(\mathbb{E}\|\hat{x}_t\|^{2+2\delta_G} + \mathbb{E}\|\hat{y}_t\|^{2+2\delta_G}\right),\end{aligned} \tag{98}$$

where the second inequality uses $\|I - \beta_t B_3\| \le 1 - \mu_G\beta_t$ and the last inequality uses Cauchy–Schwarz inequality. Plugging (98) into (97) yields

$$\begin{aligned}\mathbb{E}\|\hat{y}_{t+1}\|^2 &\overset{(97)}{\le} (1 - \mu_G\beta_t)\mathbb{E}\|\hat{y}_t\|^2 + L_{G,x}^2\beta_t^2\mathbb{E}\|\hat{x}_t\|^2 + \beta_t^2\Gamma_{22} \\ &\quad + 2\beta_t\mathbb{E}\langle \hat{y}_t - \beta_t G(H(y_t), y_t), G(H(y_t), y_t) - G(x_t, y_t)\rangle \\ &\overset{(98)}{\le} \left(1 - \frac{2\mu_G\beta_t}{3}\right)\mathbb{E}\|\hat{y}_t\|^2 + 2L_{G,x}^2\beta_t^2\mathbb{E}\|\hat{x}_t\|^2 + 2d_y L_{G,x}\beta_t\|\mathbb{E}\hat{x}_t\hat{y}_t^\top\| + \Gamma_{22}\beta_t^2 + \Delta_{y,t},\end{aligned}$$

where we use $\Delta_{y,t}$ defined in (39) to collect the remaining terms. ■

C.3 Proof of Lemma 11

Proof [Proof of Lemma 11] We first present the specific forms of the constants:

$$\begin{aligned}&c_{xy,1}^{de} = L_H L_{G,x} + \Gamma_{22}^{\frac{1+\delta_H}{2}} S_H + 6\iota_2 L_H L_{G,y}^2,\ c_{xy,2}^{de} = (2 + L_H)\Sigma_{22} + 4\iota_2^{\delta_H}\Gamma_{22}^{\frac{2+\delta_H}{2}},\ c_{xy,3}^{de} = S_H\Gamma_{22}^{\frac{1+\delta_H}{2}}, \\ &c_{xy,4}^{de} = 2\widetilde{S}_H(L_{G,x} \wedge L_{G,y}) + 8\iota_2^{\delta_H} S_H(L_{G,x}^{1+\delta_H} \wedge L_{G,y}^{1+\delta_H}) + 16\iota_2^{1+\delta_H} S_H(L_{G,x}^{2+\delta_H} + L_{G,y}^{2+\delta_H}) \propto S_H.\end{aligned} \tag{99}$$

Using the update rules (60) and (68), we have

$$\mathbb{E}\left[\hat{x}_{t+1}\hat{y}_{t+1}^\top\right] = \underbrace{\mathbb{E}\left(\hat{x}_t - \alpha_t F(x_t, y_t)\right)\left(\hat{y}_t - \beta_t G(H(y_t), y_t)\right)^\top}_{\blacklozenge_1} + \alpha_t\beta_t\mathbb{E}\left[\xi_t\psi_t^\top\right]$$

$$+\mathbb{E}\underbrace{\beta_t\left(\hat{x}_t-\alpha_t F(x_t,y_t)\right)\left(G(H(y_t),y_t)-G(x_t,y_t)\right)^\top}_{\blacklozenge_2}+\mathbb{E}\underbrace{\left(H(y_t)-H(y_{t+1})\right)^\top\hat{y}_{t+1}}_{\blacklozenge_3}.$$

It then follows that

$$\|\mathbb{E}\hat{x}_{t+1}\hat{y}_{t+1}^\top\|\le\|\mathbb{E}\blacklozenge_1\|+\|\mathbb{E}\blacklozenge_2\|+\|\mathbb{E}\blacklozenge_3\|+\alpha_t\beta_t\Sigma_{12}. \tag{100}$$

We then analyze $\{\|\mathbb{E}\blacklozenge_i\|\}_{i\in[3]}$ in the following respectively.

To analyze $\blacklozenge_1$, we make use of Lemma 19 and obtain that

$$\begin{aligned}\blacklozenge_1=&(1-\alpha_tB_1)\hat{x}_t\hat{y}_t^\top(1-\beta_tB_3)^\top-\beta_t(1-\alpha_tB_1)\hat{x}_t\left(R_t^{GH}\right)^\top\\&-\alpha_tR_t^F\hat{y}_t^\top(1-\beta_tB_3)^\top+\alpha_t\beta_tR_t^F\left(R_t^{GH}\right)^\top.\end{aligned}$$

Taking expectation and then the spectrum norm on both sides, we have

$$\begin{aligned}\|\mathbb{E}\blacklozenge_1\|\le&(1-\mu_F\alpha_t)(1-\mu_G\beta_t)\|\mathbb{E}\hat{x}_t\hat{y}_t^\top\|+\beta_t(1-\mu_F\alpha_t)S_{B,G}\mathbb{E}\|\hat{x}_t\|\|\hat{y}_t\|^{1+\delta_G}\\&+\alpha_t(1-\mu_G\beta_t)S_{B,F}\mathbb{E}\|\hat{y}_t\|\left(\|\hat{x}_t\|^{1+\delta_F}+\|\hat{y}_t\|^{1+\delta_F}\right)\\&+\alpha_t\beta_tS_{B,F}S_{B,G}\mathbb{E}\left(\|\hat{x}_t\|^{1+\delta_F}+\|\hat{y}_t\|^{1+\delta_F}\right)\|\hat{y}_t\|^{1+\delta_G}\\\le&(1-\mu_F\alpha_t)(1-\mu_G\beta_t)\|\mathbb{E}\hat{x}_t\hat{y}_t^\top\|+\Delta_{xy,t}^{(1)},\end{aligned}\tag{101}$$

Here we use $\Delta_{xy,t}^{(1)}$ to denote the higher-order residual collecting all the remaining terms for simplicity. More specifically, it follows that

$$\begin{aligned}\Delta_{xy,t}^{(1)}=&\beta_t(1-\mu_F\alpha_t)S_{B,G}\mathbb{E}\|\hat{x}_t\|\|\hat{y}_t\|^{1+\delta_G}+\alpha_t(1-\mu_G\beta_t)S_{B,F}\mathbb{E}\|\hat{y}_t\|\left(\|\hat{x}_t\|^{1+\delta_F}+\|\hat{y}_t\|^{1+\delta_F}\right)\\&+\alpha_t\beta_tS_{B,F}S_{B,G}\mathbb{E}\|\hat{y}_t\|^{1+\delta_G}\left(\|\hat{x}_t\|^{1+\delta_F}+\|\hat{y}_t\|^{1+\delta_F}\right).\end{aligned}$$

Now we can apply Young's inequality to decouple the cross terms and obtain

$$\begin{aligned}\Delta_{xy,t}^{(1)}\le&\beta_tS_{B,G}(\mathbb{E}\|\hat{x}_t\|^{2+\delta_G}+\mathbb{E}\|\hat{y}_t\|^{2+\delta_G})+2\alpha_tS_{B,F}(\mathbb{E}\|\hat{x}_t\|^{2+\delta_F}+\mathbb{E}\|\hat{y}_t\|^{2+\delta_F})\\&+2\alpha_t\beta_tS_{B,F}S_{B,G}(\mathbb{E}\|\hat{x}_t\|^{2+\delta_F+\delta_G}+\mathbb{E}\|\hat{y}_t\|^{2+\delta_F+\delta_G}).\end{aligned}\tag{102}$$

To analyze $\blacklozenge_2$, we have

$$\|\blacklozenge_2\|\overset{(62)}{\le}\beta_t\sqrt{1-\mu_F\alpha_t}\|\hat{x}_t\|\cdot\|G(H(y_t),y_t)-G(x_t,y_t)\|\overset{(13)}{\le}\beta_t\sqrt{1-\mu_F\alpha_t}\cdot L_{G,x}\|\hat{x}_t\|^2. \tag{103}$$

To analyze $\blacklozenge_3$, we will use the near linearity in Assumption 5 again. By Assumption 4, we have that

$$H(y_{t+1})-H(y_t)=\nabla H(y_t)(y_{t+1}-y_t)+R_t^H=\nabla H(y^\star)(y_{t+1}-y_t)+R_t^{\nabla H}+R_t^H,$$

where R_t^H and $R_t^{\nabla H}$ are defined by

$$R_t^H:=H(y_{t+1})-H(y_t)-\nabla H(y_t)(y_{t+1}-y_t),$$

$$R_t^{\nabla H} := (\nabla H(y_t) - \nabla H(y^\star))\,(y_{t+1} - y_t).$$

with R_t^H satisfies $\|R_t^H\| \le S_H \|y_{t+1} - y_t\|^{1+\delta_H}$. Note that $y_{t+1} - y_t = \beta_t(G(x_t, y_t) + \psi_t)$. Then

$$\begin{aligned}
\|\mathbb{E}\blacklozenge_3\| &\le \beta_t \|\nabla H(y^\star)\mathbb{E}G(x_t,y_t)\hat{y}_t^\top\| + \beta_t^2\|\nabla H(y^\star)\mathbb{E}G(x_t,y_t)G(x_t,y_t)^\top\| \\
&\quad + \beta_t^2\|\nabla H(y^\star)\mathbb{E}\psi_t\psi_t^\top\| + \|\mathbb{E}R_t^{\nabla H}\hat{y}_{t+1}^\top\| + S_H\mathbb{E}\left[\|y_{t+1}-y_t\|^{1+\delta_H}\|\hat{y}_{t+1}\|\right] \\
&\le \beta_t L_H \|\mathbb{E}\underbrace{G(x_t,y_t)\hat{y}_t^\top}_{\diamondsuit_1}\| + \beta_t^2 L_H(\mathbb{E}\underbrace{\|G(x_t,y_t)\|^2}_{\diamondsuit_2} + \Sigma_{22}) \\
&\quad + \|\mathbb{E}\underbrace{R_t^{\nabla H}\hat{y}_{t+1}^\top}_{\diamondsuit_3}\| + S_H\mathbb{E}\underbrace{\|y_{t+1}-y_t\|^{1+\delta_H}\|\hat{y}_{t+1}\|}_{\diamondsuit_4}.
\end{aligned}$$

To proceed with the proof, we then analyze $\{\diamondsuit_i\}_{i\in[4]}$ respectively in the following.

- For the term $\diamondsuit_1$, we use Lemma 19 and obtain

$$\|\mathbb{E}\diamondsuit_1\| \le L_{G,x}\|\mathbb{E}\hat{x}_t\hat{y}_t^\top\| + L_{G,y}\mathbb{E}\|\hat{y}_t\|^2 + S_{B,G}(\mathbb{E}\|\hat{x}_t\|^{1+\delta_G}\|\hat{y}_t\| + \mathbb{E}\|\hat{y}_t\|^{2+\delta_G}).$$

- For the term $\diamondsuit_2$, the inequality (64) implies that

$$\mathbb{E}\diamondsuit_2 = \mathbb{E}\|G(x_t,y_t)\|^2 \le 2L_{G,x}^2\mathbb{E}\|\hat{x}_t\|^2 + 2L_{G,y}^2\mathbb{E}\|\hat{y}_t\|^2. \tag{104}$$

- For the term $\diamondsuit_3$, (11) and (17) imply $\|\nabla H(y_t) - \nabla H(y^\star)\| \le \min\{\widetilde{S}_H\|\hat{y}_t\|^{\delta_H}, 2L_H\}$, it follows

$$\begin{aligned}
\|\mathbb{E}\diamondsuit_3\| &\le \beta_t\|\mathbb{E}\left(\nabla H(y_t) - \nabla H(y^\star)\right)G(x_t,y_t)\hat{y}_t^\top\| + \beta_t^2\|\mathbb{E}\left(\nabla H(y_t) - \nabla H(y^\star)\right)\psi_t\psi_t^\top\| \\
&\quad + \beta_t^2\|\mathbb{E}\left(\nabla H(y_t) - \nabla H(y^\star)\right)G(x_t,y_t)G(x_t,y_t)^\top\| \\
&\le \beta_t\widetilde{S}_H\mathbb{E}\|G(x_t,y_t)\|\|\hat{y}_t\|^{1+\delta_H} + 2L_H\beta_t^2(\mathbb{E}\|G(x_t,y_t)\|^2 + \Sigma_{22}) \\
&\overset{(64)}{\le} \beta_t\widetilde{S}_H(L_{G,x}\mathbb{E}\|\hat{x}_t\|\|\hat{y}_t\|^{1+\delta_H} + L_{G,y}\mathbb{E}\|\hat{y}_t\|^{2+\delta_H}) + 2L_H\beta_t^2\Sigma_{22} \\
&\quad + 4\beta_t^2 L_H\left(L_{G,x}^2\mathbb{E}\|\hat{x}_t\|^2 + L_{G,y}^2\mathbb{E}\|\hat{y}_t\|^2\right).
\end{aligned}$$

- For the term $\diamondsuit_4$, it follows

$$\begin{aligned}
\mathbb{E}\diamondsuit_4 &\le \beta_t^{1+\delta_H}\mathbb{E}\left(\|G(x_t,y_t)\| + \|\psi_t\|\right)^{1+\delta_H}\|\hat{y}_t\| + \beta_t^{2+\delta_H}\mathbb{E}\left(\|G(x_t,y_t)\| + \|\psi_t\|\right)^{2+\delta_H} \\
&\le 2\beta_t^{1+\delta_H}\mathbb{E}\|G(x_t,y_t)\|^{1+\delta_H}\|\hat{y}_t\| + 2\Gamma_{22}^{\frac{1+\delta_H}{2}}\beta_t^{1+\delta_H}\mathbb{E}\|\hat{y}_t\| \\
&\quad + 4\beta_t^{2+\delta_H}\mathbb{E}\left(\|G(x_t,y_t)\|^{2+\delta_H} + \|\psi_t\|^{2+\delta_H}\right) \\
&\overset{(64)}{\le} 4\beta_t^{1+\delta_H}(L_{G,x}^{1+\delta_H}\mathbb{E}\|\hat{x}_t\|^{1+\delta_H}\|\hat{y}_t\| + L_{G,y}^{1+\delta_H}\mathbb{E}\|\hat{y}_t\|^{2+\delta_H}) + \Gamma_{22}^{\frac{1+\delta_H}{2}}(\beta_t\mathbb{E}\|\hat{y}_t\|^2 + \beta_t^{1+2\delta_H}) \\
&\quad + 4\beta_t^{2+\delta_H}\left(4L_{G,x}^{2+\delta_H}\mathbb{E}\|\hat{x}_t\|^{2+\delta_H} + 4L_{G,y}^{2+\delta_H}\mathbb{E}\|\hat{y}_t\|^{2+\delta_H} + \Gamma_{22}^{\frac{2+\delta_H}{2}}\right),
\end{aligned}$$

where we have used $(a+b)^\gamma \le 2^{\gamma-1}(a^\gamma + b^\gamma)$ for any non-negative a, b and $\gamma \ge 1$.

Putting these pieces together and noting $\beta_t \leq \iota_2$, we have that

$$\begin{aligned}\|\mathbb{E}\blacklozenge_3\| \leq \beta_t L_H L_{G,x}\|\mathbb{E}\hat{x}_t\hat{y}_t^\top\| + 6\beta_t^2 L_H L_{G,x}^2\mathbb{E}\|\hat{x}_t\|^2 + \beta_t c_{xy,1}^{de}\mathbb{E}\|\hat{y}_t\|^2 + c_{xy,2}^{de}\beta_t^2 \\ + c_{xy,3}^{de}\beta_t^{1+2\delta_H} + \Delta_{xy,t}^{(2)},\end{aligned} \tag{105}$$

where $c_{xy,1}^{de}, c_{xy,2}^{de}, c_{xy,3}^{de}$ are constants defined in (99) and $\Delta_{xy,t}^{(2)}$ is a higher-order residual covering all the remaining terms in (105). By Young's inequality, we can derive the following upper bound for $\Delta_{xy,t}^{(2)}$

$$\begin{aligned}\Delta_{xy,t}^{(2)} &\leq 2\beta_t L_H S_{B,G}(\mathbb{E}\|\hat{x}_t\|^{2+\delta_G} + \mathbb{E}\|\hat{y}_t\|^{2+\delta_G}) + 16\beta_t^{2+\delta_H} S_H L_{G,x}^{2+\delta_H}\mathbb{E}\|\hat{x}_t\|^{2+\delta_H} \\ &\quad + \beta_t(\widetilde{S}_H L_{G,x} + 4\beta_t^{\delta_H} S_H L_{G,x}^{1+\delta_H})(\mathbb{E}\|\hat{x}_t\|^{2+\delta_H} + \mathbb{E}\|\hat{y}_t\|^{2+\delta_H}) \\ &\quad + \beta_t\left(\widetilde{S}_H L_{G,y} + 4\beta_t^{\delta_H} S_H L_{G,y}^{1+\delta_H} + 16\beta_t^{1+\delta_H} S_H L_{G,y}^{2+\delta_H}\right)\mathbb{E}\|\hat{y}_t\|^{2+\delta_H} \\ &\leq 2\beta_t L_H S_{B,G}(\mathbb{E}\|\hat{x}_t\|^{2+\delta_G} + \mathbb{E}\|\hat{y}_t\|^{2+\delta_G}) + \beta_t c_{xy,4}^{de}(\mathbb{E}\|\hat{x}_t\|^{2+\delta_H} + \mathbb{E}\|\hat{y}_t\|^{2+\delta_H}),\end{aligned} \tag{106}$$

where $c_{xy,4}^{de}$ is defined in (99).

Now, we are ready to establish this lemma. Plugging (101), (103) and (105) into (100), we have that

$$\begin{aligned}\|\mathbb{E}\hat{x}_{t+1}\hat{y}_{t+1}^\top\| \leq \left(1 - \frac{\mu_F\alpha_t}{2}\right)\|\mathbb{E}\hat{x}_t\hat{y}_t^\top\| + \beta_t\left(L_{G,x}\mathbb{E}\|\hat{x}_t\|^2 + c_{xy,1}^{de}\mathbb{E}\|\hat{y}_t\|^2\right) + \Sigma_{12}\alpha_t\beta_t \\ + c_{xy,2}^{de}\beta_t^2 + c_{xy,3}^{de}\beta_t^{1+2\delta_H} + \Delta_{xy,t},\end{aligned}$$

where the inequality uses the following facts

- Since $\frac{\beta_t}{\alpha_t} \leq \kappa \leq \frac{\mu_F}{2L_H L_{G,x}}$, $(1-\mu_F\alpha_t)(1-\mu_G\beta_t) + \beta_t L_H L_{G,x} \leq 1 - \mu_F\alpha_t + \beta_t L_H L_{G,x} \leq 1 - \frac{\mu_F\alpha_t}{2}$.
- Since $\frac{\beta_t}{\alpha_t} \leq \kappa \leq \frac{\mu_F}{12L_H L_{G,x}}$, $\sqrt{1-\mu_F\alpha_t} + 6\beta_t L_H L_{G,x} \leq 1 - \frac{\mu_F\alpha_t}{2} + 6\beta_t L_H L_{G,x} \leq 1$.
- Combining (102) and (106), we have

$$\begin{aligned}&\Delta_{xy,t}^{(1)} + \Delta_{xy,t}^{(2)} \\ &\leq 2\alpha_t S_{B,F}(\mathbb{E}\|\hat{x}_t\|^{2+\delta_F} + \mathbb{E}\|\hat{y}_t\|^{2+\delta_F}) + \beta_t S_{B,G}(1+2L_H)(\mathbb{E}\|\hat{x}_t\|^{2+\delta_G} + \mathbb{E}\|\hat{y}_t\|^{2+\delta_G}) \\ &\quad + \beta_t c_{xy,4}^{de}(\mathbb{E}\|\hat{x}_t\|^{2+\delta_H} + \mathbb{E}\|\hat{y}_t\|^{2+\delta_H}) + 2\alpha_t\beta_t S_{B,F}S_{B,G}(\mathbb{E}\|\hat{x}_t\|^{2+\delta_F+\delta_G} + \mathbb{E}\|\hat{y}_t\|^{2+\delta_F+\delta_G}) \\ &=: \Delta_{xy,t}.\end{aligned}$$

We complete the proof. ■

C.4 Proof of Lemma 12

Proof [Proof of Lemma 12] We first present the specific forms of the constants:

$$c_{xx,1}^{de} = 24L_H L_{G,y},\ c_{xx,2}^{de} = 40L_H^2 L_{G,y}^2,\ c_{xx,3}^{de} = 1280L_H^4 L_{G,y}^4,\ c_{xx,4}^{de} = \frac{64S_H^2\Gamma_{22}^{1+\delta_H}}{\mu_F}. \tag{107}$$

From the proof of Lemma 6, we have

$$\|\hat{x}_{t+1}\|^2 \overset{(61)}{=} \underbrace{\|\hat{x}_t - \alpha_t F(x_t, y_t)\|^2}_{\mathcal{C}_1} + \underbrace{\|H(y_t) - H(y_{t+1}) - \alpha_t \xi_t\|^2}_{\mathcal{C}_2} + \underbrace{2\langle \hat{x}_t - \alpha_t F(x_t, y_t), H(y_t) - H(y_{t+1})\rangle}_{\mathcal{C}_3} + \underbrace{2\alpha_t \langle \hat{x}_t - \alpha_t F(x_t, y_t), -\xi_t\rangle}_{\mathcal{C}_4}.$$

Taking the square of both sides yields

$$\|\hat{x}_{t+1}\|^4 \leq \mathcal{C}_1^2 + 4\mathcal{C}_2^2 + 4\mathcal{C}_3^2 + 4\mathcal{C}_4^2 + 2\mathcal{C}_1(\mathcal{C}_2 + \mathcal{C}_3 + \mathcal{C}_4), \tag{108}$$

where the last inequality is due to $(a+b+c)^2 \leq 3(a^2+b^2+c^2)$ for any $a, b \in \mathbb{R}$. We then analyze these terms respectively in the following.

- For $\mathcal{C}_1^2$, we have

$$\mathcal{C}_1^2 \overset{(62)}{\leq} \left(1 - \frac{7\mu_F \alpha_t}{2} + \frac{49\mu_F^2 \alpha_t^2}{16}\right) \|\hat{x}_t\|^4. \tag{109}$$

- For $\mathcal{C}_2^2$, taking the expectation w.r.t. $\mathcal{F}_t$, we have

$$\begin{aligned}\mathbb{E}\left[\mathcal{C}_2^2 \,|\, \mathcal{F}_t\right] &\overset{(11)}{\leq} 8\left(L_H^4 \beta_t^4 \mathbb{E}\left[\|G(x_t, y_t) - \psi_t\|^4 \,|\, \mathcal{F}_t\right] + \alpha_t^4 \Gamma_{11}^2\right) \\ &\overset{(111)}{\leq} 8\left(5L_H^4 \beta_t^4 \|G(x_t, y_t)\|^4 + 7L_H^4 \beta_t^4 \Gamma_{22}^2 + \alpha_t^4 \Gamma_{11}^2\right),\end{aligned} \tag{110}$$

  where the first inequality also uses $(a+b)^4 \leq 8(a^4+b^4)$ for any $a, b \in \mathbb{R}$ and the last inequality uses (111) below

$$\mathbb{E}\left[\|G(x_t, y_t) - \psi_t\|^4\right] \leq 5\|G(x_t, y_t)\|^4 + 7\Gamma_{22}^2. \tag{111}$$

  To derive (111), we first notice that

$$\begin{aligned}\|G(x_t, y_t) - \psi_t\|^4 &\leq \|G(x_t, y_t)\|^4 + 6\|G(x_t, y_t)\|^2 \|\psi_t\|^2 + 4\|G(x_t, y_t)\| \|\psi_t\|^3 + \|\psi_t\|^4 \\ &\quad - 4\|G(x_t, y_t)\|^2 \langle G(x_t, y_t), \psi_t\rangle \\ &\leq 5\|G(x_t, y_t)\|^4 + 7\|\psi_t\|^4 - 4\|G(x_t, y_t)\|^2 \langle G(x_t, y_t), \psi_t\rangle,\end{aligned}$$

  where the last inequality follows from Young's inequality. Taking the conditional expectation gives (111).

  Then we plug (64) into (110) and obtain

$$\mathbb{E}\left[\mathcal{C}_2^2 \,|\, \mathcal{F}_t\right] \leq 320 L_H^4 L_{G,x}^4 \beta_t^4 \|\hat{x}_t\|^4 + 320 L_H^4 L_{G,y}^4 \beta_t^4 \|\hat{y}_t\|^4 + 56 L_H^4 \beta_t^4 \Gamma_{22}^2 + 8\alpha_t^4 \Gamma_{11}^2, \tag{112}$$

  where the inequality also uses $(a+b)^4 \leq 8(a^4+b^4)$ for any $a, b \in \mathbb{R}$.

- For $\mathcal{C}_3^2$, taking the expectation w.r.t. $\mathcal{F}_t$, we have

$$\begin{aligned}\mathbb{E}\left[\mathcal{C}_3^2 \,|\, \mathcal{F}_t\right] &\overset{(11)+(62)}{\leq} 4L_H^2\|\hat{x}_t\|^2\mathbb{E}\left[\|y_t - y_{t+1}\|^2 \,|\, \mathcal{F}_t\right] \\ &\overset{(64)}{\leq} 4L_H^2\beta_t^2\|\hat{x}_t\|^2\left(2L_{G,x}^2\|\hat{x}_t\|^2 + 2L_{G,y}^2\|\hat{y}_t\|^2 + \Gamma_{22}\right) \\ &= 8L_H^2L_{G,x}^2\beta_t^2\|\hat{x}_t\|^4 + 8L_H^2L_{G,y}^2\beta_t^2\|\hat{x}_t\|^2\|\hat{y}_t\|^2 + 4L_H^2\Gamma_{22}\beta_t^2\|\hat{x}_t\|^2. \end{aligned} \tag{113}$$

- For $\mathcal{C}_4^2$, taking the expectation w.r.t. $\mathcal{F}_t$, we have

$$\mathbb{E}\left[\mathcal{C}_4^2 \,|\, \mathcal{F}_t\right] \leq 4\alpha_t^2\|\hat{x}_t - \alpha_t F(x_t, y_t)\|^2\mathbb{E}\left[\|\xi_t\|^2 \,|\, \mathcal{F}_t\right] \overset{(62)}{\leq} 4\alpha_t^2\Gamma_{11}\|\hat{x}_t\|^2. \tag{114}$$

- For $\mathcal{C}_1\mathcal{C}_2$, taking the expectation w.r.t. $\mathcal{F}_t$, we have

$$\begin{aligned}\mathbb{E}\left[\mathcal{C}_1\mathcal{C}_2 \,|\, \mathcal{F}_t\right] &= \mathcal{C}_1\mathbb{E}\left[\mathcal{C}_2 \,|\, \mathcal{F}_t\right] \\ &\overset{(62)+(63)}{\leq} \|\hat{x}_t\|^2\left(4L_H^2L_{G,x}^2\beta_t^2\|\hat{x}_t\|^2 + 4L_H^2L_{G,y}^2\beta_t^2\|\hat{y}_t\|^2 + 2\beta_t^2L_H^2\Gamma_{22} + 2\alpha_t^2\Gamma_{11}\right) \\ &\leq 4L_H^2L_{G,x}^2\beta_t^2\|\hat{x}_t\|^4 + 4L_H^2L_{G,y}^2\beta_t^2\|\hat{x}_t\|^2\|\hat{y}_t\|^2 + 2L_H^2\Gamma_{22}\beta_t^2\|\hat{x}_t\|^2 + 2\Gamma_{11}\alpha_t^2\|\hat{x}_t\|^2. \end{aligned} \tag{115}$$

- For $\mathcal{C}_1\mathcal{C}_3$, taking the expectation w.r.t. $\mathcal{F}_t$, we have

$$\begin{aligned}\mathbb{E}\left[\mathcal{C}_1\mathcal{C}_3 \,|\, \mathcal{F}_t\right] &= \mathcal{C}_1\mathbb{E}\left[\mathcal{C}_3 \,|\, \mathcal{F}_t\right] \\ &\overset{(62)+(67)}{\leq} \|\hat{x}_t\|^2\left(\mu_F\alpha_t\|\hat{x}_t\|^2 + 14L_HL_{G,y}\beta_t\|\hat{x}_t\|\|\hat{y}_t\| + \frac{32S_H^2\Gamma_{22}^{1+\delta_H}}{\mu_F}\frac{\beta_t^{2+\delta_H}}{\alpha_t}\right) \\ &\leq \mu_F\alpha_t\|\hat{x}_t\|^4 + 14L_HL_{G,y}\beta_t\|\hat{x}_t\|^3\|\hat{y}_t\| + \frac{32S_H^2\Gamma_{22}^{1+\delta_H}}{\mu_F}\frac{\beta_t^{2+2\delta_H}}{\alpha_t}\|\hat{x}_t\|^2. \end{aligned} \tag{116}$$

- For $\mathcal{C}_1\mathcal{C}_4$, taking the expectation w.r.t. $\mathcal{F}_t$, we have

$$\mathbb{E}\left[\mathcal{C}_1\mathcal{C}_4 \,|\, \mathcal{F}_t\right] = \mathcal{C}_1\mathbb{E}\left[\mathcal{C}_4 \,|\, \mathcal{F}_t\right] = 0. \tag{117}$$

Moreover, since $\alpha_t \leq \frac{1}{12\mu_F}$ and $\frac{\beta_t^2}{\alpha_t} \leq \frac{\mu_F}{200L_H^2L_{G,x}^2}$, we have

$$1 - \frac{3\mu_F\alpha_t}{2} + \frac{49\mu_F^2\alpha_t^2}{16} + 40L_H^2L_{G,x}^2\beta_t^2 + 1280L_H^4L_{G,x}^4\beta_t^4 \leq 1 - \mu_F\alpha_t.$$

Taking the expectation on both sides of (108) w.r.t. $\mathcal{F}_t$ and plugging (109), (112) to (117) into it, together with the above inequality and the definition of $\{c_{xx,i}\}_{i\in[4]}$ in (107), we obtain (42).

■

C.5 Proof of Lemma 13

Proof [Proof of Lemma 13] From the proof of Lemma 7, we have

$$\|\hat{y}_{t+1}\|^2 \overset{(69)}{=} \underbrace{\|\hat{y}_t - \beta_t G(H(y_t), y_t)\|^2}_{\mathcal{D}_1} + \underbrace{\|\beta_t \left(G(H(y_t), y_t) - G(x_t, y_t)\right) - \beta_t \psi_t\|^2}_{\mathcal{D}_2} + \underbrace{2\beta_t \langle \hat{y}_t - \beta_t G(H(y_t), y_t), G(H(y_t), y_t) - G(x_t, y_t)\rangle}_{\mathcal{D}_3} + \underbrace{2\beta_t \langle \hat{y}_t - \beta_t G(H(y_t), y_t), -\psi_t\rangle}_{\mathcal{D}_4}.$$

Taking the square of both sides yields

$$\|\hat{y}_{t+1}\|^4 \leq \mathcal{D}_1^2 + 4\mathcal{D}_2^2 + 4\mathcal{D}_3^2 + 4\mathcal{D}_4^2 + 2\mathcal{D}_1(\mathcal{D}_2 + \mathcal{D}_3 + \mathcal{D}_4), \tag{118}$$

where the last inequality is due to $(a+b+c)^2 \leq 3(a^2+b^2+c^2)$ for any $a, b \in \mathbb{R}$. We then analyze these terms respectively in the following.

- For $\mathcal{D}_1^2$, we have

$$\mathcal{D}_1^2 \overset{(70)}{\leq} (1 - 2\mu_G\beta_t + \mu_G^2\beta_t^2)\|\hat{y}_t\|^4 \leq \left(1 - \frac{3\mu_G\beta_t}{2}\right)\|\hat{y}_t\|^4, \tag{119}$$

  where the last inequality is due to $\beta_t \leq \frac{1}{2\mu_G}$.

- For $\mathcal{D}_2^2$, taking the expectation w.r.t. $\mathcal{F}_t$, we have

$$\mathbb{E}\left[\mathcal{D}_2^2 \,|\, \mathcal{F}_t\right] = \beta_t^4 \mathbb{E}\left[\|G(H(y_t), y_t) - G(x_t, y_t) - \psi_t\|^4 \,|\, \mathcal{F}_t\right].$$

  Similar to the derivation of (111), we can obtain

$$\mathbb{E}\left[\|G(H(y_t), y_t) - G(x_t, y_t) - \psi_t\|^4\right] \leq 5\|G(x_t, y_t) - G(H(y_t), y_t)\|^4 + 7\Gamma_{22}^2.$$

  It follows that

$$\mathbb{E}\left[\mathcal{D}_2^2 \,|\, \mathcal{F}_t\right] \leq 5\beta_t^4\|G(x_t, y_t) - G(H(y_t), y_t)\|^4 + 7\beta_t^4\Gamma_{22}^2 \overset{(13)}{\leq} 5L_{G,x}^4\beta_t^4\|\hat{x}_t\|^4 + 7\beta_t^4\Gamma_{22}^2. \tag{120}$$

- For $\mathcal{D}_3^2$, we have

$$\mathcal{D}_3^2 \leq 4\beta_t^2\|\hat{y}_t - \beta_t G(H(y_t), y_t)\|^2\|G(H(y_t), y_t) - G(x_t, y_t)\|^2 \overset{(13)+(70)}{\leq} 4L_{G,x}^2\beta_t^2\|\hat{y}_t\|^2\|\hat{x}_t\|^2. \tag{121}$$

- For $\mathcal{D}_4^2$, taking the expectation w.r.t. $\mathcal{F}_t$, we have

$$\mathbb{E}\left[\mathcal{D}_4^2 \,|\, \mathcal{F}_t\right] \leq 4\beta_t^2\|\hat{y}_t - \beta_t G(H(y_t), y_t)\|^2 \mathbb{E}\left[\|\psi_t\|^2 \,|\, \mathcal{F}_t\right] \overset{(70)}{\leq} 4\Gamma_{22}\beta_t^2\|\hat{y}_t\|^2. \tag{122}$$

- For $\mathcal{D}_1\mathcal{D}_2$, taking the expectation w.r.t. $\mathcal{F}_t$, we have

$$\mathbb{E}\left[\mathcal{D}_1\mathcal{D}_2 \,|\, \mathcal{F}_t\right] = \mathcal{D}_1\mathbb{E}\left[\mathcal{D}_2 \,|\, \mathcal{F}_t\right] \overset{(70)+(71)}{\leq} L_{G,x}^2\beta_t^2\|\hat{x}_t\|^2\|\hat{y}_t\|^2 + \Gamma_{22}\beta_t^2\|\hat{y}_t\|^2. \tag{123}$$

- For $\mathcal{D}_1\mathcal{D}_3$, taking the expectation w.r.t. $\mathcal{F}_t$, we have

$$\mathbb{E}\left[\mathcal{D}_1\mathcal{D}_3 \,|\, \mathcal{F}_t\right] = \mathcal{D}_1\mathbb{E}\left[\mathcal{D}_3 \,|\, \mathcal{F}_t\right] \overset{(70)+(72)}{\leq} 2L_{G,x}\beta_t\|\hat{x}_t\|\|\hat{y}_t\|^3. \tag{124}$$

- For $\mathcal{D}_1\mathcal{D}_4$, taking the expectation w.r.t. $\mathcal{F}_t$, we have

$$\mathbb{E}\left[\mathcal{D}_1\mathcal{D}_4 \,|\, \mathcal{F}_t\right] = \mathcal{D}_1\mathbb{E}\left[\mathcal{D}_4 \,|\, \mathcal{F}_t\right] = 0. \tag{125}$$

Taking the expectation on both sides of (118) w.r.t. $\mathcal{F}_t$ and plugging (119) to (125) into it yields (43). ■

C.6 Proof of Lemma 14

Proof [Proof of Lemma 14] To characterize the L_4-convergence rate, we define the Lyapunov function $V_t = \varrho_3\frac{\beta_t}{\alpha_t}\|\hat{x}_t\|^4 + \|\hat{y}_t\|^4$ with $\varrho_3 = \frac{54L_{G,x}^4}{\mu_F\mu_G^3}$.

Derive the one-step descent. We first employ Lemmas 12 and 13 to derive the one-step descent of V_t. Since $\frac{\beta_t}{\alpha_t} \leq \kappa$, we have

$$20\Gamma_{11}\alpha_t^2 + 20L_H^2\Gamma_{22}\beta_t^2 \leq c_{xx,5}^{de}\alpha_t^2, \quad 32\alpha_t^4\Gamma_{11}^2 + 224L_H^4\beta_t^4\Gamma_{22}^2 \leq c_{xx,6}^{de}\alpha_t^4. \tag{126}$$

with

$$c_{xx,5}^{de} = 20\Gamma_{11} + 20L_H\Gamma_{22}\kappa^2,\ c_{xx,6}^{de} = 32\Gamma_{11}^2 + 224L_H^4\Gamma_{22}^2\kappa^4.$$

As a result of $\frac{\beta_{t+1}}{\alpha_{t+1}} \leq \frac{\beta_t}{\alpha_t}$, we have

$$\begin{aligned}
&\mathbb{E}\left[V_{t+1} \,|\, \mathcal{F}_t\right] \\
&\overset{(42)+(43)}{\leq} V_t - \underbrace{\varrho_3\mu_F\beta_t\|\hat{x}_t\|^4}_{\mathcal{E}_0} - \frac{3\mu_G\beta_t}{2}\|\hat{y}_t\|^4 + \underbrace{4L_{G,x}\beta_t\|\hat{x}_t\|\|\hat{y}_t\|^3}_{\mathcal{E}_1} + \underbrace{\varrho_3 c_{xx,1}^{de}\frac{\beta_t^2}{\alpha_t}\|\hat{x}_t\|^3\|\hat{y}_t\|}_{\mathcal{E}_2} \\
&\quad + \underbrace{18L_{G,x}^2\beta_t^2\|\hat{x}_t\|^2\|\hat{y}_t\|^2}_{\mathcal{E}_3} + \underbrace{\varrho_3 c_{xx,2}^{de}\frac{\beta_t^3}{\alpha_t}\|\hat{x}_t\|^2\|\hat{y}_t\|^2}_{\mathcal{E}_4} + \underbrace{20L_{G,x}^4\beta_t^4\|\hat{x}_t\|^4}_{\mathcal{E}_5} + \underbrace{\varrho_3 c_{xx,3}^{de}\frac{\beta_t^5}{\alpha_t}\|\hat{y}_t\|^4}_{\mathcal{E}_6} \\
&\quad + \varrho_3 c_{xx,4}^{de}\frac{\beta_t^{3+2\delta_H}}{\alpha_t^2}\|\hat{x}_t\|^2 + \varrho_3 c_{xx,5}^{de}\alpha_t\beta_t\|\hat{x}_t\|^2 + 18\Gamma_{22}\beta_t^2\|\hat{y}_t\|^2 + \varrho_3 c_{xx,6}^{de}\alpha_t^3\beta_t + 28\Gamma_{22}^2\beta_t^4.
\end{aligned} \tag{127}$$

We then analyze $\mathcal{E}_0$ to $\mathcal{E}_6$ respectively in the following.

- For $\mathcal{E}_0$, we have $\frac{\beta_t}{\alpha_t} \leq \frac{\mu_F}{5\mu_G}$. It follows that

$$\mathcal{E}_0 \geq \frac{19}{20}\varrho_3\mu_F\beta_t\|\hat{x}_t\|^4 + \frac{\varrho_3\mu_G\beta_t^2}{4\alpha_t}\|\hat{x}_t\|^4. \tag{128}$$

- For $\mathcal{E}_1$, by Young's inequality, we have

$$\mathcal{E}_1 \leq \frac{L_{G,x}^4 \beta_t}{\lambda_1^4 \mu_G^3} \|\hat{x}_t\|^4 + 3\lambda_1^{4/3} \mu_G \beta_t \|\hat{y}_t\|^4 \leq \frac{\varrho_3 \mu_F \beta_t}{2} \|\hat{x}_t\|^4 + \mu_G \beta_t \|\hat{y}_t\|^4, \tag{129}$$

  where the last inequality is by setting $\lambda_1^{4/3} = 1/3$.

- For $\mathcal{E}_2$, by Young's inequality, we have

$$\mathcal{E}_2 = \varrho_3 \left(\beta_t^{3/4} \lambda_2 \|\hat{x}_t\|^3\right) \left(c_{xx,1} \lambda_2^{-1} \beta_t^{5/4} \alpha_t^{-1} \|\hat{y}_t\|\right) \leq \frac{3\lambda_2^{4/3} \varrho_3 \beta_t}{4} \|\hat{x}_t\|^4 + \frac{\varrho_3 (c_{xx,1}^{de})^4 \beta_t^5}{4\lambda_2^4 \alpha_t^4} \|\hat{y}_t\|^4.$$

  Since $\frac{\beta_t}{\alpha_t} \leq \frac{\mu_F \mu_G}{200 L_H L_{G,x} L_{G,y}}$, we have $\frac{\beta_t^4}{\alpha_t^4} = \frac{\mu_F^4 \mu_G^4}{200^4 L_H^4 L_{G,x}^4 L_{G,y}^4} \leq \frac{54 \cdot 24^4 \mu_F^3 \mu_G}{200^4 \varrho_3 (c_{xx,1}^{de})^4}$. Then setting $\lambda_2^{4/3} = \mu_F/4$ yields

$$\mathcal{E}_2 \leq \frac{3\varrho_3 \mu_F \beta_t}{16} \|\hat{x}_t\|^4 + \frac{4^2 \cdot 54 \cdot 24^4 \mu_G \beta_t}{200^4} \|\hat{y}_t\|^4 \leq \frac{3\varrho_3 \mu_F \beta_t}{16} \|\hat{x}_t\|^4 + 0.18 \mu_G \beta_t \|\hat{y}_t\|^4. \tag{130}$$

- For $\mathcal{E}_3$, by Cauchy-Schwarz inequality, we have $\mathcal{E}_3 \leq \frac{\varrho_3 \mu_F \beta_t}{8} \|\hat{x}_t\|^4 + \frac{8 \cdot 81 L_{G,x}^4 \beta_t^3}{\mu_F \varrho_3} \|\hat{y}_t\|^4$. Since $\beta_t \leq \frac{1}{14\mu_G}$, we have $\beta_t^2 \leq \frac{1}{14^2 \mu_G^2} = \frac{\varrho_3 \mu_F \mu_G}{14^2 \cdot 54 L_{G,x}^4}$. It follows that

$$\mathcal{E}_3 \leq \frac{\varrho_3 \mu_F \beta_t}{8} \|\hat{x}_t\|^4 + \frac{8 \cdot 81 \mu_G \beta_t}{14^2 \cdot 54} \|\hat{y}_t\|^4 \leq \frac{\varrho_3 \mu_F \beta_t}{8} \|\hat{x}_t\|^4 + 0.062 \mu_G \beta_t \|\hat{y}_t\|^4. \tag{131}$$

- For $\mathcal{E}_4$, by AM-GM inequality, we have $\mathcal{E}_4 \leq \frac{\varrho_3 \mu_F \beta_t}{8} \|\hat{x}_t\|^4 + \frac{2\varrho_3 (c_{xx,2}^{de})^2 \beta_t^5}{\mu_F \alpha_t^2} \|\hat{y}_t\|^4$. Since $\alpha_t \leq \frac{1}{12\mu_F}$ and $\frac{\beta_t}{\alpha_t} \leq \frac{\mu_F \mu_G}{200 L_H L_{G,x} L_{G,y}}$, we have $\frac{\beta_t^4}{\alpha_t^2} \leq \frac{\mu_F^2 \mu_G^4}{12^2 \cdot 200^4 L_H^4 L_{G,x}^4 L_{G,y}^4} \leq \frac{54 \cdot 40^2 \mu_F \mu_G}{12^2 \cdot 200^4 \varrho_3 (c_{xx,2}^{de})^2}$. It follows that

$$\mathcal{E}_4 \leq \frac{\varrho_3 \mu_F \beta_t}{8} \|\hat{x}_t\|^4 + \frac{108 \cdot 40^2 \mu_G \beta_t}{12^2 \cdot 200^4} \|\hat{y}_t\|^4 \leq \frac{\varrho_3 \mu_F \beta_t}{8} \|\hat{x}_t\|^4 + 0.001 \mu_G \beta_t \|\hat{y}_t\|^4. \tag{132}$$

- For $\mathcal{E}_5$, since $\beta_t \leq \frac{1}{14\mu_G}$, we have $\beta_t^3 \leq \frac{1}{14^3 \mu_G^3} \leq \frac{\varrho_3 \mu_F}{14^3 \cdot 54 L_{G,x}^4}$. It follows that

$$\mathcal{E}_5 \leq 0.001 \varrho_3 \mu_F \beta_t \|\hat{x}_t\|^4. \tag{133}$$

- For $\mathcal{E}_6$, since $\alpha_t \leq \frac{1}{12\mu_F}$ and $\frac{\beta_t}{\alpha_t} \leq \frac{\mu_F \mu_G}{200 L_H L_{G,x} L_{G,y}}$, we have $\frac{\beta_t^4}{\alpha_t} \leq \frac{\mu_F \mu_G^4}{12^3 \cdot 200^4 L_H^4 L_{G,x}^4 L_{G,y}^4} \leq \frac{54 \cdot 1280 \mu_G}{12^3 \cdot 200^4 \varrho_3 c_{xx,3}^{de}} \leq 0.001 \frac{\mu_G}{\varrho_3 c_{xx,3}^{de}}$. It follows that

$$\mathcal{E}_6 \leq 0.001 \mu_G \beta_t \|\hat{y}_t\|^4. \tag{134}$$

Plugging (128) to (134) into (127) and taking the expectation yields

$$\mathbb{E}V_{t+1} \leq \left(1 - \frac{\mu_G \beta_t}{4}\right) \mathbb{E}V_t + \left(\varrho_3 c_{xx,4}^{de} \frac{\beta_t^{3+2\delta_H}}{\alpha_t^2} + \varrho_3 c_{xx,5}^{de} \alpha_t \beta_t + 18\Gamma_{22}\beta_t^2\right) \left(\mathbb{E}\|\hat{x}_t\|^2 + \mathbb{E}\|\hat{y}_t\|^2\right) + \varrho_3 c_{xx,6}^{de} \alpha_t^3 \beta_t + 28\Gamma_{22}^2 \beta_t^4. \tag{135}$$

Since we assume $\prod_{\tau=0}^{t} \left(1 - \frac{\mu_G \beta_\tau}{4}\right) = \mathcal{O}(\alpha_t)$, then Theorem 8 with the definitions of $c_{x,7}$ in (76) and $c_{y,2}$ in (75) implies

$$\mathbb{E}\|\hat{x}_t\|^2 + \mathbb{E}\|\hat{y}_t\|^2 \leq c_+ \alpha_t + c_- \frac{\beta_t^{2+2\delta_H}}{\alpha_t^2} \text{ with } c_- \propto S_H^2 \Gamma_{22}^{1+\delta_H}. \tag{136}$$

Also note that $\frac{\beta_t}{\alpha_t} \leq \kappa$. Then we have

$$\begin{aligned} &\left(\varrho_3 c_{xx,4}^{de} \frac{\beta_t^{3+2\delta_H}}{\alpha_t^2} + \varrho_3 c_{xx,5}^{de} \alpha_t \beta_t + 18\Gamma_{22}\beta_t^2\right) \left(\mathbb{E}\|\hat{x}_t\|^2 + \mathbb{E}\|\hat{y}_t\|^2\right) \\ &\leq \beta_t \left[\varrho_3 c_{xx,4}^{de} \frac{\beta_t^{2+2\delta_H}}{\alpha_t^2} + (\varrho_3 c_{xx,5}^{de} + 18\Gamma_{22}\kappa)\alpha_t\right] \left(c_+ \alpha_t + c_- \frac{\beta_t^{2+2\delta_H}}{\alpha_t^2}\right). \end{aligned} \tag{137}$$

Plugging (137) into (135) and using $\alpha_t \leq \iota_1$, $\beta_t \leq \iota_2$, $\frac{\beta_t}{\alpha_t} \leq \kappa$ and $(c_1 a + c_2 b)(c_3 a + c_4 b) \leq c_1 c_3 a^2 + c_2 c_4 b^2 + (c_2 c_3 + c_1 c_4)(a^2 + b^2)/2$ for $a, b, c_1, c_2, c_3, c_4 \geq 0$, we obtain

$$\mathbb{E}V_{t+1} \leq \left(1 - \frac{\mu_G \beta_t}{4}\right) \mathbb{E}V_t + c_{yy,1}^{de} \alpha_t^2 \beta_t + c_{yy,2}^{de} \frac{\beta_t^{5+4\delta_H}}{\alpha_t^4}, \tag{138}$$

where the constants $c_{yy,1}^{de}$ and $c_{yy,2}^{de}$ are defined as

$$\begin{aligned} c_{yy,1}^{de} &= (\varrho_3 c_{xx,5}^{de} + 18\Gamma_{22}\kappa)c_+ + \frac{\varrho_3 c_{xx,4}^{de} c_+ + (\varrho_3 c_{xx,5}^{de} + 18\Gamma_{22}\kappa)c_-}{2} + \varrho_3 c_{xx,6}^{de} \iota_1 + 28\Gamma_{22}^2 \kappa^2 \iota_2, \\ c_{yy,2}^{de} &= \varrho_3 c_{xx,4}^{de} c_- + \frac{\varrho_3 c_{xx,4}^{de} c_+ + (\varrho_3 c_{xx,5}^{de} + 18\Gamma_{22}\kappa)c_-}{2} = S_H^2 \Gamma_{22}^{1+\delta_H} \cdot h_1(S_H, \Gamma_{22}) \end{aligned} \tag{139}$$

for some function $h_1(\cdot, \cdot)$. Here the last inequality follows from the definitions of $c_{xx,4}^{de}$ in (107) and c_- in (136).

Establish the convergence rates. With the one-step descent inequality (138), we could establish the convergence rates of $\mathbb{E}\|\hat{x}_t\|^4$ and $\mathbb{E}\|\hat{y}_t\|^4$. For simplicity, we introduce (as we did in the proof of Theorem 8) $\alpha_{j,t} = \prod_{\tau=j}^{t} \left(1 - \frac{\mu_F \alpha_\tau}{2}\right)$ and $\beta_{j,t} = \prod_{\tau=j}^{t} \left(1 - \frac{\mu_G \beta_\tau}{4}\right)$. Iterating (138) and applying Lemma 21 (i) yields

$$\mathbb{E}V_{t+1} \leq \beta_{0,t}\, \mathbb{E}V_0 + \frac{8c_{yy,1}^{de}}{\mu_G} \alpha_t^2 + \frac{10 c_{yy,2}^{de}}{\mu_G} \frac{\beta_t^{4+4\delta_H}}{\alpha_t^4}.$$

The proof of Lemma 21 is similar to that of Lemma 16 and is omitted.

Lemma 21 (Step sizes inequalities) *We define the product as* $\alpha_{j,t} = \prod_{\tau=j}^{t} \left(1 - \frac{\mu_F \alpha_\tau}{2}\right)$ *and* $\beta_{j,t} = \prod_{\tau=j}^{t} \left(1 - \frac{\mu_G \beta_\tau}{4}\right)$. *Under Assumption 7, it holds that*

(i) $\sum_{j=0}^{t}\beta_{j+1,t}\,\alpha_j^2\beta_j \le \frac{8\alpha_t^2}{\mu_G}$ *and* $\sum_{j=0}^{t}\beta_{j+1,t}\frac{\beta_j^{5+4\delta_H}}{\alpha_j^4} \le \frac{10\beta_t^{4+4\delta_H}}{\mu_G\alpha_t^4}$.

(ii) $\sum_{j=0}^{t}\alpha_{j+1,t}\,\alpha_j^3 \le \frac{4\alpha_t^2}{\mu_F}$, $\sum_{j=0}^{t}\alpha_{j+1,t}\,\alpha_j\beta_{0,j-1} \le \frac{8\beta_{0,t}}{\mu_F}$ *and* $\sum_{j=0}^{t}\alpha_{j+1,t}\frac{\beta_j^{4+4\delta_H}}{\alpha_j^3} \le \frac{3\beta_t^{4+4\delta_H}}{\mu_F\alpha_t^4}$

Then we obtain the convergence rate of $\mathbb{E}\|\hat{y}_t\|^4$ as shown in the following

$$\mathbb{E}\|\hat{y}_t\|^4 \le \mathbb{E}V_t \le \beta_{0,t-1}\,\mathbb{E}V_0 + \frac{8c_{yy,1}^{de}}{\mu_G}\alpha_{t-1}^2 + \frac{10c_{yy,2}^{de}}{\mu_G}\frac{\beta_t^{4+4\delta_H}}{\alpha_t^4}. \tag{140}$$

Since $\kappa \le \frac{\mu_F\mu_G}{200L_HL_{G,x}L_{G,y}}$, we have

$$\mathbb{E}V_0 = \varrho_3\frac{\beta_0}{\alpha_0}\mathbb{E}\|\hat{x}_0\|^4 + \mathbb{E}\|\hat{y}_0\|^4 \le \frac{54L_{G,x}^4\kappa}{\mu_F\mu_G^3}\mathbb{E}\|\hat{x}_0\|^4 + \mathbb{E}\|\hat{y}_0\|^4 \le c_{yy,2}^{de}\mathbb{E}\|\hat{x}_0\|^4 + \mathbb{E}\|\hat{y}_0\|^4,$$

where $c_{yy,2}^{de} = \frac{L_{G,x}^3}{3\mu_G^2L_HL_{G,y}}$. As a result,

$$\mathbb{E}\|\hat{y}_t\|^4 \le \beta_{0,t-1}\left(c_{yy,2}^{de}\mathbb{E}\|\hat{x}_0\|^4 + \mathbb{E}\|\hat{y}_0\|^4\right) + \frac{8c_{yy,1}^{de}}{\mu_G}\alpha_{t-1}^2 + \frac{10c_{yy,2}^{de}}{\mu_G}\frac{\beta_t^{4+4\delta_H}}{\alpha_t^4}.$$

Plugging (130), (132), (134) and (126) into (42) and taking the expectation, we obtain

$$\begin{aligned}\mathbb{E}\|\hat{x}_{t+1}\|^4 &\le \left(1-\frac{\mu_F\alpha_t}{2}\right)\mathbb{E}\|\hat{x}_t\|^4 + \frac{\mu_G\alpha_t}{5\varrho_3}\mathbb{E}\|\hat{y}_t\|^4 + \left(c_{xx,4}^{de}\frac{\beta_t^{2+2\delta_H}}{\alpha_t} + c_{xx,5}^{de}\alpha_t^2\right)\mathbb{E}\|\hat{x}_t\|^2 + c_{xx,6}^{de}\alpha_t^4\\ &\overset{(136)+(140)}{\le} \left(1-\frac{\mu_F\alpha_t}{2}\right)\mathbb{E}\|\hat{x}_t\|^4 + \frac{\mu_G\mathbb{E}V_0}{5\varrho_3}\beta_{0,t-1}\alpha_t + c_{xx,7}^{de}\,\alpha_t^3 + c_{xx,8}^{de}\frac{\beta_t^{4+4\delta_H}}{\alpha_t^3},\end{aligned} \tag{141}$$

where the last inequality also uses $\alpha_t \le \iota_1$, $\beta_t \le \iota_2$, $\frac{\alpha_{t-1}}{\alpha_t} \le 1+\frac{\mu_F\alpha_t}{8} \le 1+\frac{\mu_F\iota_1}{8}$, $\frac{\beta_{t-1}}{\beta_t} \le 1+\frac{\mu_G\beta_t}{64} \le 1+\frac{\mu_G\iota_2}{64}$, $\frac{\beta_{t-1}^a\alpha_{t-1}^{-b}}{\beta_t^a\alpha_t^{-b}} \le \left(\frac{\beta_{t-1}}{\beta_t}\right)^a$ for any $a,b>0$, and the constants are defined as

$$\begin{aligned}c_{xx,7}^{de} &= \frac{8c_{yy,1}^{de}}{5\varrho_3}\left(1+\frac{\mu_F\iota_1}{8}\right)^2 + \left(c_{xx,5}^{de}c_+ + \frac{c_{xx,5}^{de}c_- + c_{xx,4}^{de}c_+}{2}\right)\left[1+\frac{\mu_F\iota_1}{8}+\left(1+\frac{\mu_G\iota_2}{64}\right)^4\right] + c_{xx,6}^{de}\iota_1,\\ c_{xx,8}^{de} &= \frac{10c_{yy,2}^{de}}{\mu_G}\left(1+\frac{\mu_G\iota_2}{64}\right)^8 + \left(c_{xx,4}^{de}c_- + \frac{c_{xx,5}^{de}c_- + c_{xx,4}^{de}c_+}{2}\right)\left[1+\frac{\mu_F\iota_1}{8}+\left(1+\frac{\mu_G\iota_2}{64}\right)^4\right]\\ &\lesssim S_H^2\Gamma_{22}^{1+\delta_H}\cdot h_2(S_H,\Gamma_{22})\end{aligned} \tag{142}$$

for some function $h_2(\cdot,\cdot)$. Here the last inequality follows from the definitions of $c_{xx,4}^{de}$ in (107), c_- in (136) and $c_{yy,2}^{de}$ in (139). Iterating (141) yields

$$\mathbb{E}\|\hat{x}_{t+1}\|^4 \le \alpha_{0,t}\mathbb{E}\|\hat{x}_0\|^4 + \frac{8\mu_G}{5\varrho_3\mu_F}\beta_{0,t}\mathbb{E}V_0 + \frac{4c_{xx,7}^{de}}{\mu_F}\alpha_t^2 + \frac{3c_{xx,8}^{de}}{\mu_F}\frac{\beta_t^{4+4\delta_H}}{\alpha_t^4},$$

where the inequality also applies Lemma 21 (ii). Since $\frac{\mu_G \beta_j}{\mu_F \alpha_j} \leq \frac{\mu_G}{\mu_F}\kappa \leq \frac{1}{2}$ for any j, it holds that $\alpha_{0,t} \leq \beta_{0,t}$. Recall the definition of V_0 and $\varrho_3 = \frac{54 L_{G,x}^4}{\mu_F \mu_G^3}$. Then we can obtain

$$\mathbb{E}\|\hat{x}_{t+1}\|^4 \leq 2\beta_{0,t}\mathbb{E}\|\hat{x}_0\|^4 + \frac{\mu_G^4}{27 L_{G,x}^4}\beta_{0,t}\mathbb{E}\|\hat{y}_0\|^4 + \frac{4c_{xx,7}^{de}}{\mu_F}\alpha_t^2 + \frac{3c_{xx,8}^{de}}{\mu_F}\frac{\beta_t^{4+4\delta_H}}{\alpha_t^4}.$$

This completes the proof. ■

C.7 Proof of Theorem 18

Proof [Proof of Theorem 18] Since the conditions of Theorem 18 are stronger than those of Theorem 8. The first inequality follows from Theorem 8 with

$$C_{x,0}^{de} = 3\mathbb{E}\|\hat{x}_0\|^2 + \frac{7L_H L_{G,y}\mathbb{E}\|\hat{y}_0\|^2}{L_{G,x}},\ C_{x,1}^{de} = \frac{8\Gamma_{11}}{\mu_F} + c_{x,5}\kappa + c_{x,6}\kappa^2,\ C_{x,2}^{de} = c_{x,7} \propto S_H^2 \Gamma_{22}^{1+\delta_H}, \tag{143}$$

where the constants $\{c_{x,i}\}_{i\in[7]\setminus[4]}$ are defined in (76).

The proof of the other two inequalities is divided into three parts. We first use Lemma 14 to analyze the high-order terms defined in Lemmas 9 to 11, then derive the convergence rate of $\|\mathbb{E}\hat{x}_t\hat{y}_t^\top\|$ and finally use this to establish the rate of $\mathbb{E}\|\hat{y}_t\|^2$.

Analyze the high-order terms. Since we assume $\prod_{\tau=0}^t \left(1 - \frac{\mu_G\beta_\tau}{4}\right) = \mathcal{O}(\alpha_t^2)$, then Lemma 14 with the definitions of $C_{xx,8}^{de}$ in (142) and $C_{yy,2}^{de}$ in (139) implies

$$\mathbb{E}\|\hat{x}_t\|^4 + \mathbb{E}\|\hat{y}_t\|^4 \leq c_{+,2}\,\alpha_t^2 + c_{-,2}\frac{\beta_t^{4+4\delta_H}}{\alpha_t^4} \text{ with } c_{-,2} = S_H^2\Gamma_{22}^{1+\delta_H}\cdot h_3(S_H,\Gamma_{22}) \tag{144}$$

for some function $h_3(\cdot,\cdot)$. Note that the terms in $\Delta_{x,t}, \Delta_{y,t}$ and $\Delta_{xy,t}$ all have their exponents lying between 2 and 4. Then we could apply Jensen's inequality to control them by the fourth-order terms or apply Young's inequality to bound them by second-order and fourth-order terms.

We first focus on $\Delta_{x,t}$, which is defined in (37). By Jensen's inequality and Assumption 7, for $z = \|\hat{x}_t\|$ or $\|\hat{y}_t\|$ and $\gamma \in [0,2]$, we have

$$\mathbb{E}z^{2+\gamma} \leq (\mathbb{E}z^4)^{\frac{2+\gamma}{4}} \leq c_{+,2}^{\frac{2+\gamma}{4}}\,\alpha_t^{1+\frac{\gamma}{2}} + c_{-,2}^{\frac{2+\gamma}{4}}\left(\frac{\beta_t^{2+2\delta_H}}{\alpha_t^2}\right)^{1+\frac{\gamma}{2}} \leq c_{+,2}^{\frac{2+\gamma}{4}}\,\iota_1^{\frac{\gamma}{2}}\alpha_t + c_{-,2}^{\frac{2+\gamma}{4}}\,\iota_2^{\delta_H\gamma}\kappa^\gamma\frac{\beta_t^{2+2\delta_H}}{\alpha_t^2}. \tag{145}$$

Substituting (145) into (37), we obtain

$$\Delta_{x,t} \leq c_{x,9}^{de}\alpha_t\beta_t + c_{x,10}^{de}\frac{\beta_t^{3+2\delta_H}}{\alpha_t^2}, \tag{146}$$

where the constants are defined as

$$\begin{aligned} c_{x,9}^{de} &= 2c_{x,5}^{de} c_{+,2}^{\frac{2+\delta_H}{4}} \iota_1^{\frac{\delta_H}{2}} + 2c_{x,6}^{de} c_{+,2}^{\frac{2+\delta_F}{4}} \iota_1^{1+\frac{\delta_F}{2}} + 2c_{x,7}^{de} c_{+,2}^{\frac{1+\delta_G}{2}} \iota_1^{\delta_G} + 2c_{x,8}^{de} c_{+,2}^{\frac{1+\delta_H}{2}} \iota_1^{\delta_H} \iota_2^{\delta_H}, \\ c_{x,10}^{de} &= 2c_{x,5}^{de} c_{-,2}^{\frac{2+\delta_H}{4}} \iota_2^{\delta_H^2} \kappa^{\delta_H} + 2c_{x,6}^{de} c_{-,2}^{\frac{2+\delta_F}{4}} \iota_1 \iota_2^{\delta_H \delta_F} \kappa^{\delta_F} + 2c_{x,7}^{de} c_{-,2}^{\frac{1+\delta_G}{2}} \iota_2^{2\delta_H \delta_G} \kappa^{2\delta_G} \\ &\quad + 2c_{x,8}^{de} c_{-,2}^{\frac{1+\delta_H}{2}} \iota_2^{2\delta_H^2+\delta_H} \kappa^{2\delta_H} \\ &= S_H \Gamma_{22}^{\frac{1+\delta_H}{2}} \cdot h_4(S_H, \Gamma_{22}) \end{aligned} \tag{147}$$

for some function $h_4(\cdot,\cdot)$. Here the last inequality follows from the definitions of $c_{x,5}^{de}$ to $c_{x,8}^{de}$ in (88) and $c_{-,2}$ in (144).

For $\Delta_{y,t}$ defined in (39), we have

$$\Delta_{y,t} \le c_{y,1}^{de} \beta_t (\mathbb{E}\|\hat{x}_t\|^{2+2\delta_G} + \mathbb{E}\|\hat{y}_t\|^{2+2\delta_G}) \text{ with } c_{y,1}^{de} = S_{B,G}^2 \left(\frac{15 d_y^2}{\mu_G} + d_y^2 \iota_2 + 8 d_y \iota_2 \right). \tag{148}$$

By Young's inequality with $p = \frac{1}{1-\delta_G}$ and $q = \frac{1}{\delta_G}$, for $z = \|\hat{x}_t\|$ or $\|\hat{y}_t\|$, we have

$$c_{y,1}^{de} \beta_t z^{2+2\delta_G} \le \frac{4 L_H L_{G,x} \beta_t^2}{\mu_F \alpha_t} z^2 + c_{y,2}^{de} \frac{\beta_t^2}{\alpha_t} \left(\frac{\alpha_t}{\beta_t} \right)^{\frac{1}{\delta_G}} z^4 \text{ with } c_{y,2}^{de} \propto S_{B,G}^{\frac{2}{\delta_G}}, \tag{149}$$

where we use $c_{y,2}^{de}$ to hide the problem-dependent coefficients. Plugging this inequality and (144) into (148) and noting that $\frac{\beta_t}{\alpha_t} \le \frac{\mu_F \mu_G}{24 L_H L_{G,x}}$, we obtain

$$\Delta_{y,t} \le \frac{\mu_G \beta_t}{6} \mathbb{E}\|\hat{y}_t\|^2 + \frac{4 L_H L_{G,x} \beta_t^2}{\mu_F \alpha_t} \mathbb{E}\|\hat{x}_t\|^2 + c_{y,2}^{de} \left(c_{+,2} \alpha_t \beta_t^2 + c_{-,2} \frac{\beta_t^{6+4\delta_H}}{\alpha_t^5} \right) \left(\frac{\alpha_t}{\beta_t} \right)^{\frac{1}{\delta_G}}. \tag{150}$$

Finally, we concentrate on $\Delta_{xy,t}$, which is defined in (41). We first tackle the first term. By Young's inequality with $p = \frac{2}{2-\delta_F}$ and $q = \frac{2}{\delta_F}$, for $z = \|\hat{x}_t\|$ or $\|\hat{y}_t\|$, we have

$$2\alpha_t S_{B,F} z^{2+\delta_F} \le \frac{L_{G,x} \beta_t}{4} z^2 + c_{xy,5}^{de} \beta_t \left(\frac{\alpha_t}{\beta_t} \right)^{\frac{2}{\delta_F}} z^4 \text{ and } c_{xy,5}^{de} \propto S_{B,F}^{\frac{2}{\delta_F}}, \tag{151}$$

where we use $c_{xy,5}^{de}$ to hide the problem-dependent coefficients. Other terms can be controlled by (145). Recall that $\alpha_t \le \iota_1$ and $\frac{\beta_t}{\alpha_t} \le \kappa$. Plugging the above inequality and (145) into (41) yields

$$\begin{aligned} \Delta_{xy,t} &\le L_{G,x} \beta_t (\mathbb{E}\|\hat{x}_t\|^2 + \mathbb{E}\|\hat{y}_t\|^2) + c_{xy,6}^{de} \alpha_t \beta_t + c_{xy,7}^{de} \frac{\beta_t^{3+2\delta_H}}{\alpha_t^2} \\ &\quad + \left(c_{xy,8}^{de} \alpha_t^2 \beta_t + c_{xy,9}^{de} \frac{\beta_t^{5+4\delta_H}}{\alpha_t^4} \right) \left(\frac{\alpha_t}{\beta_t} \right)^{\frac{2}{\delta_F}}, \end{aligned} \tag{152}$$

where the constants are defined as

$$
\begin{aligned}
c_{xy,6}^{de} &= 2S_{B,G}(1+2L_H)c_{+,2}^{\frac{2+\delta_G}{4}}\iota_1^{\frac{\delta_G}{2}} + 2c_{xy,4}^{de}c_{+,2}^{\frac{2+\delta_H}{4}}\iota_1^{\frac{\delta_H}{2}} + 4S_{B,F}S_{B,G}\,c_{+,2}^{\frac{2+\delta_F+\delta_G}{4}}\iota_1^{1+\frac{\delta_F+\delta_G}{2}},\\
c_{xy,7}^{de} &= 2S_{B,G}(1+2L_H)c_{-,2}^{\frac{2+\delta_G}{4}}\iota_2^{\delta_H\delta_G}\kappa^{\delta_G} + 2c_{xy,4}^{de}c_{-,2}^{\frac{2+\delta_H}{4}}\iota_2^{\delta_H^2}\kappa^{\delta_H}\\
&\quad + 4S_{B,F}S_{B,G}\,c_{-,2}^{\frac{2+\delta_F+\delta_G}{4}}\iota_1\iota_2^{\delta_H(\delta_F+\delta_G)}\kappa^{\delta_F+\delta_G} \qquad (153)\\
&= (S_{B,G}+S_H+S_{B,F}S_{B,G})S_H\Gamma_{22}^{\frac{1+\delta_H}{2}}\cdot h_5(S_H,\Gamma_{22}),\\
c_{xy,8}^{de} &= c_{xy,5}^{de}c_{+,2} \propto S_{B,F}^{\frac{2}{\delta_F}},\quad c_{xy,9}^{de} = c_{xy,5}^{de}c_{-,2} = S_{B,F}^{\frac{2}{\delta_F}}S_H\Gamma_{22}^{\frac{1+\delta_H}{2}}\cdot h_6(S_H,\Gamma_{22})
\end{aligned}
$$

for some functions $h_5(\cdot,\cdot)$ and $h_6(\cdot,\cdot)$. Here the inequalities follows from the definitions of $c_{xy,4}^{de}$ in (88), $c_{xy,5}^{de}$ in (151) and $c_{-,2}$ in (144).

Derive the upper bound of $\|\mathbb{E}\hat{x}_t\hat{y}_t^\top\|$. With the upper bounds of the higher-order terms, we could derive the upper bound of $\|\mathbb{E}\hat{x}_t\hat{y}_t^\top\|$. Substituting (146), (150) and (152) into (36), (38) and (40) respectively yields

$$
\begin{aligned}
\mathbb{E}\|\hat{x}_{t+1}\|^2 &\le (1-\mu_F\alpha_t)\,\mathbb{E}\|\hat{x}_t\|^2 + c_{x,1}^{de}\beta_t^2\mathbb{E}\|\hat{y}_t\|^2 + c_{x,2}^{de}\beta_t\|\mathbb{E}\hat{x}_t\hat{y}_t^\top\| + 2\Gamma_{11}\alpha_t^2\\
&\quad + c_{x,3}^{de}\,\beta_t^2 + c_{x,4}^{de}\frac{\beta_t^{2+2\delta_H}}{\alpha_t} + c_{x,9}^{de}\,\alpha_t\beta_t + c_{x,10}^{de}\frac{\beta_t^{3+2\delta_H}}{\alpha_t^2}, \qquad (154)\\
\mathbb{E}\|\hat{y}_{t+1}\|^2 &\le \left(1-\frac{\mu_G\beta_t}{2}\right)\mathbb{E}\|\hat{y}_t\|^2 + \left(\frac{4L_HL_{G,x}\beta_t^2}{\mu_F\alpha_t} + 2L_{G,x}^2\beta_t^2\right)\mathbb{E}\|\hat{x}_t\|^2 + \Gamma_{22}\beta_t^2\\
&\quad + 2d_yL_{G,x}\beta_t\|\mathbb{E}\hat{x}_t\hat{y}_t^\top\| + c_{y,2}^{de}\left(c_{+,2}\,\alpha_t\beta_t^2 + c_{-,2}\frac{\beta_t^{6+4\delta_H}}{\alpha_t^5}\right)\left(\frac{\alpha_t}{\beta_t}\right)^{\frac{1}{\delta_G}}, \qquad (155)\\
\|\mathbb{E}\hat{x}_{t+1}\hat{y}_{t+1}^\top\| &\le \left(1-\frac{\mu_F\alpha_t}{2}\right)\|\mathbb{E}\hat{x}_t\hat{y}_t^\top\| + 2L_{G,x}\beta_t\mathbb{E}\|\hat{x}_t\|^2 + (c_{xy,1}^{de}+L_{G,x})\beta_t\mathbb{E}\|\hat{y}_t\|^2 + \Sigma_{12}\alpha_t\beta_t\\
&\quad + c_{xy,2}^{de}\beta_t^2 + c_{xy,3}^{de}\beta_t^{1+2\delta_H} + c_{xy,6}^{de}\,\alpha_t\beta_t + c_{xy,7}^{de}\frac{\beta_t^{3+2\delta_H}}{\alpha_t^2}\\
&\quad + \left(c_{xy,8}^{de}\,\alpha_t^2\beta_t + c_{xy,9}^{de}\frac{\beta_t^{5+4\delta_H}}{\alpha_t^4}\right)\left(\frac{\alpha_t}{\beta_t}\right)^{\frac{2}{\delta_F}}. \qquad (156)
\end{aligned}
$$

Define the Lyapunov function $W_t = \varrho_4\frac{\beta_t}{\alpha_t}\mathbb{E}\|\hat{x}_t\|^2 + \|\mathbb{E}\hat{x}_t\hat{y}_t^\top\|$ with $\varrho_4 = \frac{6L_{G,x}}{\mu_F}$. Since $\frac{\beta_t}{\alpha_t} \le \kappa \le \frac{\mu_F\mu_G}{24d_xL_HL_{G,x}L_{G,y}} \wedge \frac{\mu_F}{5\mu_G}$ and $c_{x,4} = 2d_xL_HL_{G,y}$, one can check

$$
\varrho_4c_{x,4}\frac{\beta_t^2}{\alpha_t} \le \frac{72d_xL_HL_{G,x}L_{G,y}}{\mu_F^2}\cdot\kappa^2\cdot\frac{\mu_F\alpha_t}{6} \le \frac{\mu_F\alpha_t}{6} \text{ and } 2L_{G,x} = \frac{\mu_F\varrho_4}{3}. \qquad (157)
$$

Combining (154), (156), (157) and $\frac{\beta_{t+1}}{\alpha_{t+1}} \le \frac{\beta_t}{\alpha_t}$, we can obtain

$$
\begin{aligned}
W_{t+1} &\le \left(1-\frac{\mu_F\alpha_t}{3}\right)W_t + c_{xy,10}^{de}\beta_t\mathbb{E}\|\hat{y}_t\|^2 + c_{xy,11}^{de}\alpha_t\beta_t + c_{xy,12}^{de}\beta_t^{1+2\delta_H}\\
&\quad + \left(c_{xy,8}^{de}\alpha_t^2\beta_t + c_{xy,9}^{de}\frac{\beta_t^{5+4\delta_H}}{\alpha_t^4}\right)\left(\frac{\alpha_t}{\beta_t}\right)^{\frac{2}{\delta_F}}, \qquad (158)
\end{aligned}
$$

where we also use Assumption 7 and $\varrho_4 = \frac{6L_{G,x}}{\mu_F}$, and the constants are defined as

$$\begin{aligned}
c^{de}_{xy,10} &= c^{de}_{xy,1} + L_{G,x} + \varrho_4 c^{de}_{x,1}\rho, \\
c^{de}_{xy,11} &= 2\varrho_4\Gamma_{11} + \Sigma_{12} + \varrho_4 c^{de}_{x,3}\kappa^2 + \varrho_4 c^{de}_{x,9}\kappa + c^{de}_{xy,2}\kappa + c^{de}_{xy,6}, \\
c^{de}_{xy,12} &= c^{de}_{xy,3} + \varrho_4 c^{de}_{x,4}\kappa^2 + \varrho_4 c^{de}_{x,10}\kappa^3 + c^{de}_{xy,7}\kappa^2 = S_H\Gamma_{22}^{\frac{1+\delta_H}{2}} \cdot h_7(S_H, \Gamma_{22})
\end{aligned} \tag{159}$$

for some function $h_7(\cdot,\cdot)$, and the last step follows from the definitions of $c^{de}_{xy,3}$ in (99), $c^{de}_{x,4}$ in (88), $c^{de}_{x,10}$ in (147) and $c^{de}_{xy,7}$ in (153).

For simplicity, we let $\tilde{\alpha}_{j,t} = \prod_{\tau=j}^{t}\left(1 - \frac{\mu_F\alpha_\tau}{3}\right)$, $\beta_{j,t} = \prod_{\tau=j}^{t}\left(1 - \frac{\mu_G\beta_\tau}{4}\right)$ and $\tilde{\beta}_{j,t} = \prod_{\tau=j}^{t}\left(1 - \frac{\mu_G\beta_\tau}{2}\right)$. Note that under our assumptions, (136) holds, i.e.,

$$\mathbb{E}\|\hat{x}_t\|^2 + \mathbb{E}\|\hat{y}_t\|^2 \le c_+\alpha_t + c_-\frac{\beta_t^{2+2\delta_H}}{\alpha_t^2} \le c_+\alpha_t + c_-\kappa^2\beta_t^{2\delta_H} \text{ with } c_- \propto S_H^2\Gamma_{22}^{1+\delta_H}. \tag{160}$$

Plugging (160) into (158) and iterating, we obtain

$$W_{t+1} \le C^{de}_{xy,0}\,\beta_{0,t} + C^{de}_{xy,1}\,\beta_t + C^{de}_{xy,2}\frac{\beta_t^{1+2\delta_H}}{\alpha_t} + \left(C^{de}_{xy,3}\,\alpha_t\beta_t + C^{de}_{xy,4}\frac{\beta_t^{5+4\delta_H}}{\alpha_t^5}\right)\left(\frac{\alpha_t}{\beta_t}\right)^{\frac{2}{\delta_F}},$$

where the inequality applies (i) and (ii) in Lemma 22, whose proof is deferred to Appendix C.9, and also uses $\tilde{\alpha}_{0,t} \le \beta_{0,t}$, and the constants are defined as

$$\begin{aligned}
C^{de}_{xy,0} &= \varrho_4\kappa\mathbb{E}\|\hat{x}_0\|^2 + \|\mathbb{E}\hat{x}_0\hat{y}_0^\top\|,\ C^{de}_{xy,1} = \frac{6(c^{de}_{xy,10}c_+ + c^{de}_{xy,11})}{\mu_F}, \\
C^{de}_{xy,2} &= \frac{4(c^{de}_{xy,10}c_-\kappa^2 + c^{de}_{xy,12})}{\mu_F} = S_H\Gamma_{22}^{\frac{1+\delta_H}{2}} \cdot h_{xy,2}(S_H, \Gamma_{22}), \\
C^{de}_{xy,3} &= \frac{12c^{de}_{xy,8}}{\mu_F} \propto S_{B,F}^{\frac{2}{\delta_F}},\ C^{de}_{xy,4} = \frac{9c^{de}_{xy,9}}{\mu_F} = S_{B,F}^{\frac{2}{\delta_F}}S_H\Gamma_{22}^{\frac{1+\delta_H}{2}} \cdot h_{xy,4}(S_H, \Gamma_{22})
\end{aligned} \tag{161}$$

for some functions $\{h_{xy,i}(\cdot,\cdot)\}_{i=2,4}$. Here we have used the definitions of c_- in (136), $c^{de}_{xy,8}$, $c^{de}_{xy,9}$ in (153) and $c^{de}_{xy,12}$ in (159).

Lemma 22 (Step sizes inequalities) *We define the product as $\tilde{\alpha}_{j,t} = \prod_{\tau=j}^{t}(1 - \frac{\mu_F\alpha_\tau}{3})$, $\tilde{\beta}_{j,t} = \prod_{\tau=j}^{t}\left(1 - \frac{\mu_G\beta_\tau}{2}\right)$ and $\beta_{j,t} = \prod_{\tau=j}^{t}\left(1 - \frac{\mu_G\beta_\tau}{4}\right)$. Under Assumption 7, it holds that*

(i) $\sum_{j=0}^{t}\tilde{\alpha}_{j+1,t}\,\alpha_j\beta_j \le \frac{6\beta_t}{\mu_F}$, $\sum_{j=0}^{t}\tilde{\alpha}_{j+1,t}\,\beta_j^{1+2\delta_H} \le \frac{4}{\mu_F}\frac{\beta_t^{1+2\delta_H}}{\alpha_t}$.

(ii) $\sum_{j=0}^{t}\tilde{\alpha}_{j+1,t}\,\alpha_j^2\beta_j\left(\frac{\alpha_j}{\beta_j}\right)^{\frac{2}{\delta_F}} \le \frac{12\alpha_t\beta_t}{\mu_F}\left(\frac{\alpha_t}{\beta_t}\right)^{\frac{2}{\delta_F}}$, $\sum_{j=0}^{t}\tilde{\alpha}_{j+1,t}\,\frac{\beta_j^{5+4\delta_H}}{\alpha_j^4}\left(\frac{\alpha_j}{\beta_j}\right)^{\frac{2}{\delta_F}} \le \frac{9\beta_t^{5+4\delta_H}}{\mu_F\alpha_t^5}\left(\frac{\alpha_t}{\beta_t}\right)^{\frac{2}{\delta_F}}$.

(iii) $\sum_{j=0}^{t}\tilde{\beta}_{j+1,t}\beta_j^2 \le \frac{4\beta_t}{\mu_G}$, $\sum_{j=0}^{t}\tilde{\beta}_{j+1,t}\beta_j\beta_{0,j-1} \le \frac{8\beta_{0,t}}{\mu_G}$, $\sum_{j=0}^{t}\tilde{\beta}_{j+1,t}\frac{\beta_t^{2+2\delta_H}}{\alpha_t} = \frac{3\beta_t^{1+2\delta_H}}{\mu_G\alpha_t}$.

(iv) $\sum_{j=0}^{t}\tilde{\beta}_{j+1,t}\,\alpha_j\beta_j^2\left(\frac{\alpha_j}{\beta_j}\right)^{\frac{2}{\delta_F}} \le \frac{6\alpha_t\beta_t}{\mu_G}\left(\frac{\alpha_t}{\beta_t}\right)^{\frac{2}{\delta_F}}$, $\sum_{j=0}^{t}\tilde{\beta}_{j+1,t}\,\frac{\beta_j^{6+4\delta_H}}{\alpha_j^5}\left(\frac{\alpha_j}{\beta_j}\right)^{\frac{2}{\delta_F}} \le \frac{6\beta_t^{5+4\delta_H}}{\mu_G\alpha_t^5}\left(\frac{\alpha_t}{\beta_t}\right)^{\frac{2}{\delta_F}}$.

(v) $\sum_{j=0}^{t}\tilde{\beta}_{j+1,t}\,\alpha_j\beta_j^2\left(\frac{\alpha_j}{\beta_j}\right)^{\frac{1}{\delta_G}}\leq\frac{6\alpha_t\beta_t}{\mu_G}\left(\frac{\alpha_t}{\beta_t}\right)^{\frac{1}{\delta_G}}$, $\sum_{j=0}^{t}\tilde{\beta}_{j+1,t}\,\frac{\beta_j^{6+4\delta_H}}{\alpha_j^5}\left(\frac{\alpha_j}{\beta_j}\right)^{\frac{2}{\delta_F}}\leq\frac{6\beta_t^{5+4\delta_H}}{\mu_G\alpha_t^5}\left(\frac{\alpha_t}{\beta_t}\right)^{\frac{1}{\delta_G}}$.

As a result, $\|\mathbb{E}\hat{x}_{t+1}\hat{y}_{t+1}^\top\|\leq W_{t+1}$ implies

$$\|\mathbb{E}\hat{x}_{t+1}\hat{y}_{t+1}^\top\|\leq C_{xy,0}^{de}\,\beta_{0,t}+C_{xy,1}^{de}\,\beta_t+C_{xy,2}^{de}\frac{\beta_t^{1+2\delta_H}}{\alpha_t}+\left(C_{xy,3}^{de}\,\alpha_t\beta_t+C_{xy,4}^{de}\frac{\beta_t^{5+4\delta_H}}{\alpha_t^5}\right)\left(\frac{\alpha_t}{\beta_t}\right)^{\frac{2}{\delta_F}}. \tag{162}$$

Since $\kappa\leq\frac{\mu_F\mu_G}{200L_HL_{G,x}L_{G,y}}$, we have $C_{xy,0}^{de}\leq\frac{\mu_G}{33L_HL_{G,y}}\mathbb{E}\|\hat{x}_0\|^2+\|\mathbb{E}\hat{x}_0\hat{y}_0^\top\|$. Thus we obtain the second inequality.

Derive the upper bound of $\mathbb{E}\|\hat{y}_t\|^2$. Now we are prepared to establish the convergence rate of $\mathbb{E}\|\hat{y}_t\|^2$. From (162) we can obtain

$$\begin{aligned}\|\mathbb{E}\hat{x}_t\hat{y}_t^\top\|\leq{}&C_{xy,0}^{de}\,\beta_{0,t-1}+C_{xy,1}^{de}\,\zeta\beta_t+C_{xy,2}^{de}\,\zeta^3\frac{\beta_t^{1+2\delta_H}}{\alpha_t}\\&+\left(C_{xy,3}^{de}\,\zeta^{\frac{2}{\delta_F}+2}\alpha_t\beta_t+C_{xy,4}^{de}\,\zeta^{\frac{2}{\delta_F}+9}\frac{\beta_t^{5+4\delta_H}}{\alpha_t^5}\right)\left(\frac{\alpha_t}{\beta_t}\right)^{\frac{2}{\delta_F}},\end{aligned} \tag{163}$$

where $\zeta=1+\frac{\mu_F\iota_1}{16}+\frac{\mu_G\iota_2}{64}$. Plugging (163) and (160) into (155) and applying the upper bounds in Assumption 7, with $c_{y,3}^{de}:=\frac{4L_HL_{G,x}}{\mu_F}+2L_{G,x}^2\iota_1$, we have

$$\begin{aligned}\mathbb{E}\|\hat{y}_{t+1}\|^2\leq{}&\left(1-\frac{\mu_G\beta_t}{2}\right)\mathbb{E}\|\hat{y}_t\|^2+2d_yL_{G,x}\beta_t\beta_{0,t-1}+\left(c_{y,3}^{de}c_++2d_yL_{G,x}C_{xy,1}^{de}\zeta+\Gamma_{22}\right)\beta_t^2\\&+\left(c_{y,3}^{de}c_-\kappa^2+2d_yL_{G,x}C_{xy,2}^{de}\zeta^3\right)\frac{\beta_t^{2+2\delta_H}}{\alpha_t}\\&+2d_yL_{G,x}\left(C_{xy,3}^{de}\zeta^{\frac{2}{\delta_F}+2}\alpha_t\beta_t^2+C_{xy,4}^{de}\zeta^{\frac{2}{\delta_F}+9}\frac{\beta_t^{6+4\delta_H}}{\alpha_t^5}\right)\left(\frac{\alpha_t}{\beta_t}\right)^{\frac{2}{\delta_F}}\\&+c_{y,2}^{de}\left(c_{+,2}\alpha_t\beta_t^2+c_{-,2}\frac{\beta_t^{6+4\delta_H}}{\alpha_t^5}\right)\left(\frac{\alpha_t}{\beta_t}\right)^{\frac{1}{\delta_G}}.\end{aligned} \tag{164}$$

Iterating (164) and applying (iii), (iv) and (v) in Lemma 22 yields

$$\begin{aligned}\mathbb{E}\|\hat{y}_{t+1}\|^2\leq{}&C_{y,0}^{de}\,\beta_{0,t}+C_{y,1}^{de}\,\beta_t+C_{y,2}^{de}\frac{\beta_t^{1+2\delta_H}}{\alpha_t}+\left(C_{y,3}^{de}\,\alpha_t\beta_t+C_{y,4}^{de}\frac{\beta_t^{5+4\delta_H}}{\alpha_t^5}\right)\left(\frac{\alpha_t}{\beta_t}\right)^{\frac{2}{\delta_F}}\\&+\left(C_{y,5}^{de}\,\alpha_t\beta_t+C_{y,6}^{de}\frac{\beta_t^{5+4\delta_H}}{\alpha_t^5}\right)\left(\frac{\alpha_t}{\beta_t}\right)^{\frac{1}{\delta_G}},\end{aligned} \tag{165}$$

where the constants are defined as

$$
\begin{aligned}
C_{y,0}^{de} &= \mathbb{E}\|\hat{y}_0\|^2 + \frac{16 d_y L_{G,x} C_{xy,0}^{de}}{\mu_G},\ C_{y,1}^{de} = \frac{4\left(c_{y,3}^{de} c_+ + 2 d_y L_{G,x} C_{xy,1}^{de}\zeta + \Gamma_{22}\right)}{\mu_G},\\
C_{y,2}^{de} &= \frac{3\left(c_{y,3}^{de} c_- \kappa^2 + 2 d_y L_{G,x} C_{xy,2}^{de}\zeta^3\right)}{\mu_G} = S_H \Gamma_{22}^{\frac{1+\delta_H}{2}} \cdot h_{y,2}(S_H, \Gamma_{22}),\\
C_{y,3}^{de} &= \frac{6(2 d_y L_{G,x} C_{xy,3}^{de}\zeta^{\frac{2}{\delta_F}+2})}{\mu_G} \propto S_{B,F}^{\frac{2}{\delta_F}},\\
C_{y,4}^{de} &= \frac{6(2 d_y L_{G,x} C_{xy,4}^{de}\zeta^{\frac{2}{\delta_F}+9})}{\mu_G} = S_{B,F}^{\frac{2}{\delta_F}} S_H \Gamma_{22}^{\frac{1+\delta_H}{2}} \cdot h_{y,4}(S_H, \Gamma_{22}),\\
C_{y,5}^{de} &= \frac{6 c_{y,2}^{de} c_{+,2}}{\mu_G} \propto S_{B,G}^{\frac{2}{\delta_G}},\ C_{y,6}^{de} = \frac{6 c_{y,2}^{de} c_{-,2}}{\mu_G} = S_{B,G}^{\frac{2}{\delta_G}} S_H \Gamma_{22}^{\frac{1+\delta_H}{2}} \cdot h_{y,6}(S_H, \Gamma_{22})
\end{aligned}
\tag{166}
$$

for some functions $\{h_{y,i}(\cdot,\cdot)\}_{i=2,4,6}$. Here we have used the definitions of c_- in (136), $C_{xy,2}^{de}$ to $C_{xy,4}^{de}$ in (161), $c_{y,2}^{de}$ in (149) and $c_{-,2}$ in (144). ■

C.8 Proof of Corollary 4

Proof [Proof of Corollay 4] Define $\beta_{0,T} = \prod_{t=0}^{T}\left(1 - \frac{\mu_G \beta_t}{4}\right)$. Clearly, $\frac{\beta_t}{\alpha_t}$ is non-increasing and upper bounded by κ. By setting α_t, β_t, T_0 as in (87), we have $\alpha_t \le \iota_1 \wedge \frac{\iota_2}{\kappa} \wedge \frac{\rho}{\kappa}^2$ for all $t \ge 0$, so the constant bounds in Assumption 7 hold. Since $(1+x)^\gamma \le 1 + \gamma x$ for $x \ge 0$ and $\gamma \in (0,1]$, then for any $t \ge 1$, we have

$$
\frac{1}{\alpha_t}\left(\frac{\alpha_{t-1}}{\alpha_t} - 1\right) \le \frac{1}{\alpha_t} \cdot \frac{a}{t - 1 + T_0} \le \frac{\delta_F \mu_G \kappa a}{128} \frac{(t+T_0)^a}{t - 1 + T_0} \le \frac{\delta_F \mu_G \kappa}{64} \frac{1}{(t+T_0)^{1-a}} \le \frac{\delta_F \mu_F}{16}.
$$

Similarly, we can obtain $\frac{1}{\beta_t}\left(\frac{\alpha_{t-1}}{\alpha_t} - 1\right) \le \frac{(\delta_F \wedge \delta_G)\mu_G}{16}$ and $\frac{1}{\beta_t}\left(\frac{\beta_{t-1}}{\beta_t} - 1\right) \le \frac{\mu_G}{64}$. Thus the growth condition in Assumption 7 holds.

For $\beta_{0,T}$, if $b = 1$, we have $\beta_{0,T} \le \prod_{t=0}^{T}\left(1 - \frac{\mu_G \beta_t}{64}\right) \le \prod_{t=0}^{T}\left(1 - \frac{2}{t+T_0}\right) \le \frac{T_0^2}{(T+T_0)(T+T_0-1)} = \mathcal{O}(\alpha_T^2)$. If $b < 1$, we have $\beta_{0,T} \le \exp\left(-32\sum_{t=0}^{T}(t+T_0)^{-b}\right) \le \exp\left(-\frac{(T+T_0+1)^{1-b} - T_0^{1-b}}{(1-b)/32}\right) = o(\alpha_T^2)$. Then the conditions in Theorem 18 hold. Note that $\delta_H \ge 0.5$ implies $\frac{\beta_t^{1+2\delta_H}}{\alpha_t} = \mathcal{O}(\beta_t)$ and $\frac{\beta_t^{5+4\delta_H}}{\alpha_t^5} = \mathcal{O}(\alpha_t \beta_t)$; $\frac{b}{a} \le 1 + \frac{\delta_F}{2} \wedge \delta_G \le 2$ implies $\left(\frac{\alpha_t}{\beta_t}\right)^{\frac{2}{\delta_F} \vee \frac{1}{\delta_G}} = \mathcal{O}(\frac{1}{\alpha_t})$ and $\beta_{0,T} = \mathcal{O}(\alpha_T^2) = \mathcal{O}(\beta_T)$. It follows that $\|\mathbb{E}\hat{x}_T \hat{y}_T^\top\| = \mathcal{O}(\beta_T)$ and $\mathbb{E}\|\hat{y}_T\|^2 = \mathcal{O}(\beta_T)$ The upper bound of $\mathbb{E}\|\hat{x}_T\|^2$ is from Theorem 3. ■

C.9 Proof of Lemma 22

Proof [Proof of Lemma 22] The overall proof is similar to that of Lemma 16 and Kaledin et al. (2020, Lemma 14). We only provide a detailed proof for one slightly different case, the first inequality in (ii).

Define $\zeta_\tau = \frac{\alpha_\tau}{\beta_\tau}$. Then we have $\frac{\zeta_{\tau-1}}{\zeta_\tau} \le \frac{\alpha_{\tau-1}}{\alpha_\tau}$. For $x \in (0.0.2)$, one can check $\exp(x) \le 1 + 1.2x$. Since $\delta_F \le 1$, $\frac{\beta_\tau}{\alpha_\tau} \le \frac{\mu_F}{5\mu_G}$ and $\mu_F\alpha_\tau \le \mu_F\iota_1 \le 1$, the growth condition implies

$$\begin{aligned}
&\frac{\alpha_{\tau-1}}{\alpha_\tau}\frac{\beta_{\tau-1}}{\beta_\tau}\left(\frac{\zeta_{\tau-1}}{\zeta_\tau}\right)^{\frac{2}{\delta_F}}\left(1-\frac{\mu_F\alpha_\tau}{3}\right) \le \left(1+\frac{\mu_G\beta_\tau}{64}\right)\left(1+\frac{\delta_F\mu_F\alpha_\tau}{16}\right)^{\frac{3}{\delta_F}}\left(1-\frac{\mu_F\alpha_\tau}{3}\right) \\
&\le \left(1+\frac{\mu_F\alpha_\tau}{80}\right)\exp\left(\frac{3\mu_F\alpha_\tau}{16}\right)\left(1-\frac{\mu_F\alpha_\tau}{3}\right) \le \left(1+\frac{\mu_F\alpha_\tau}{80}\right)\left(1+\frac{3.6\mu_F\alpha_\tau}{16}\right)\left(1-\frac{\mu_F\alpha_\tau}{3}\right) \\
&= \left(1+\frac{\mu_F\alpha_\tau}{4}\right)\left(1-\frac{\mu_F\alpha_\tau}{3}\right) \le 1-\frac{\mu_F\alpha_\tau}{12}. \qquad (167)
\end{aligned}$$

Following the derivation of Kaledin et al. (2020, Lemma 14), we can obtain the desired result. ■

C.10 Analysis of Constants in Leading Terms

In this subsection, we provide the details for Remark 4.

C.10.1 Derivation for (29)

we first analyze the constants $\mathcal{C}_x$, $\mathcal{C}_{xy,1}$ and $\mathcal{C}_{y,1}$ appearing in Theorem 3, with their detailed expression in (86). As mentioned in Remark 4, we focus on the diminishing step sizes $\alpha_t = \Theta(t^{-a})$ and $\beta_t = \Theta(t^{-b})$ with $0 < a < b \le 1$ and $\frac{b}{a} < 1 + \frac{\delta_F}{2} \wedge \delta_F$ to capture the most essential dependence on the parameters. Moreover, as analyzed in Remark 1, we can focus on $t \ge t_0$ with a prescribed t_0, then the constants $\iota_1, \iota_2, \kappa, \rho$ in Assumption 7 are of the order $o(1)$ as $t_0 \to \infty$. Then, in the expression of the $\mathcal{C}_x$, $\mathcal{C}_{xy,1}$ and $\mathcal{C}_{y,1}$, all terms involving these constants can be viewed as higher-order infinitesimal (as $t_0 \to \infty$), as shown below

$$\mathcal{C}_x = C^{de}_{x,1} + o(1), \quad \mathcal{C}_{xy,1} = C^{de}_{xy,1} + o(1), \quad \mathcal{C}_{y,1} = C^{de}_{y,1} + o(1).$$

With $C^{de}_{x,1}$ defined in (143), we have

$$\mathcal{C}_x = \frac{8\Gamma_{11}}{\mu_F} + o(1). \qquad (168)$$

Next, we analyze $\mathcal{C}_{xy,1} = C^{de}_{xy,1} + o(1)$. Although $C^{de}_{xy,1} = \frac{6}{\mu_F}(c^{de}_{xy,10}c_+ + c^{de}_{xy,11})$ as defined in (161), we can improve the constant c_+ to $o(1)$ by applying the results in Theorem 3. Note that with $1 < \frac{b}{a} < 1 + \frac{\delta_F}{2} \wedge \delta_F$, we have $\mathbb{E}\|\hat{y}_t\|^2 = o(\alpha_t)$ from the proof of Corollary 4. If we plug this upper bound into (158) instead of plugging (160) into (158) as we did in the proof of Theorem 18, we can obtain

$$C^{de}_{xy,1} = \frac{6}{\mu_F}\left(c^{de}_{xy,10}\cdot o(1) + c^{de}_{xy,11}\right) = \frac{6c^{de}_{xy,11}}{\mu_F} + o(1).$$

Then with $c^{de}_{xy,11}$ defined in (159) and $c^{de}_{xy,6}$ defined in (153), we have

$$\mathcal{C}_{xy,1} = C^{de}_{xy,1} + o(1) = \frac{6c^{de}_{xy,11}}{\mu_F} + o(1) = \frac{6}{\mu_F}\left(\frac{12L_{G,x}\Gamma_{11}}{\mu_F} + \Sigma_{12}\right) + o(1). \qquad (169)$$

Finally, we analyze $\mathcal{C}_{y,1} = C_{y,1}^{de} + o(1)$. Although $C_{y,1}^{de} = \frac{4}{\mu_G}\left(c_{y,3}^{de} c_+ + 2d_y L_{G,x} C_{xy,1}^{de}\zeta + \Gamma_{22}\right)$ as defined in (166), we can also improve the constant c_+. To that end, the expression of $\mathcal{C}_x$ in (168) implies $\mathbb{E}\|\hat{x}_t\|^2 = \frac{8\Gamma_{11}}{\mu_F}\alpha_{t-1} + o(\alpha_{t-1}) = \frac{8\Gamma_{11}}{\mu_F}\alpha_t + o(\alpha_t)$, where we have used $1 \le \frac{\alpha_{t-1}}{\alpha_t} \le 1 + \frac{\mu_F}{16}\alpha_t = 1 + o(1)$. If we plug (163) and this upper bound into (155) when we deriving (164) in the proof of Theorem 18, we can obtain

$$C_{y,1}^{de} = \frac{4}{\mu_G}\left[c_{y,3}^{de}\Big(\frac{8\Gamma_{11}}{\mu_F} + o(1)\Big) + 2d_y L_{G,x} C_{xy,1}^{de}\zeta + \Gamma_{22}\right]$$

Moreover, with $\zeta = 1 + \frac{\mu_F \iota_1}{16} + \frac{\mu_G \iota_2}{64} = 1 + o(1)$ defined below (163), $c_{y,3}^{de} = \frac{4L_H L_{G,x}}{\mu_F} + 2L_{G,x}^2 \iota_1 = \frac{4L_H L_{G,x}}{\mu_F} + o(1)$ defined above (164), and the expression of $C_{xy,1}^{de}$ in (169), we have

$$\mathcal{C}_{y,1} = C_{y,1}^{de} + o(1) = \frac{4}{\mu_G}\left[\frac{32 L_H L_{G,x}\Gamma_{11}}{\mu_F^2} + \frac{12 d_y L_{G,x}}{\mu_F}\left(\frac{12 L_{G,x}\Gamma_{11}}{\mu_F} + \Sigma_{12}\right) + \Gamma_{22}\right] + o(1).$$

C.10.2 Derivation for (30) and (31)

We first give the explicit expressions for Σ_x, Σ_y, and $\Sigma_{x,y}$. Mokkadem and Pelletier (2006) assume that $\alpha_n = \Theta(n^{-a})$, $\beta_n = \Theta(n^{-b})$ with $1/2 < a < b \le 1$ and

$$\mathbb{E}\left[\begin{pmatrix}\xi_t\xi_t^\top & \xi_t\psi_t^\top \\ \psi_t\xi_t^\top & \psi_t\psi_t^\top\end{pmatrix}\middle|\mathcal{F}_t\right] \overset{a.s.}{\to} \begin{pmatrix}\Sigma_\xi & \Sigma_{\xi,\psi} \\ \Sigma_{\xi,\psi}^\top & \Sigma_\psi\end{pmatrix}.$$

For brevity, we omit other regular conditions. Then the asymptotic covariance matrices Σ_x and Σ_y in (28) have the following expressions

$$\Sigma_x = \int_0^\infty \exp(-B_1 s)\Sigma_\xi \exp(-B_1^\top s) ds \tag{170}$$

$$\Sigma_y = \int_0^\infty \exp\left(-\Big(B_3 - \frac{\tilde{\beta} I}{2}\Big)s\right)\widetilde{\Sigma}_\psi \exp\left(-\Big(B_3^\top - \frac{\tilde{\beta} I}{2}\Big)s\right) ds. \tag{171}$$

Here $\tilde{\beta} = \lim_{n\to\infty}\beta_{n+1}^{-1} - \beta_n^{-1}$ and $\tilde{\beta} > 0$ only when $b = 1$. $\widetilde{\Sigma}_\psi$ is the asymptotic covariance of the modified noise $\breve{\psi}_t = \psi_t - B_2 B_1^{-1}\xi_t$ mentioned in Remark 4 and has the following expression

$$\widetilde{\Sigma}_\psi := \Sigma_\psi - B_2 B_1^{-1}\Sigma_{\xi,\psi} - \Sigma_{\xi,\psi}^\top B_1^{-\top} B_2^\top + B_2 B_1^{-1}\Sigma_\xi B_1^{-\top} B_2^\top. \tag{172}$$

Since $x^\star = H(y^\star)$, $\beta_t = o(\alpha_t)$, and H is L_H-Lipschitz continuous, we have $\|H(y_t) - H(y^\star)\| \le L_H\|y_t - y^\star\| = o_p(\alpha_t^{1/2})$ and consequently $\alpha_t^{-1/2}\hat{x}_t = \alpha_t^{-1/2}(x_t - x^\star) - \alpha_t^{-1/2}(H(y_t) - H(y^\star)) = \alpha_t^{-1/2}(x_t - x^\star) = o_p(1)$. Then $\alpha_t^{-1/2}\hat{x}_t$ has the same asymptotic distribution as $\alpha_t^{-1/2}(x_t - x^\star)$ in (28). If we further assume that $\{\alpha_t^{-1}\|\hat{x}_t\|^2\}_{t=1}^\infty$ and $\{\beta_t^{-1}\|\hat{y}_t\|^2\}_{t=1}^\infty$ are asymptotically uniformly integrable, then we can obtain $\lim_{t\to\infty}\alpha_t^{-1}\mathbb{E}\|x_t\|^2 = \mathrm{tr}(\Sigma_x)$ and $\lim_{t\to\infty}\beta_t^{-1}\mathbb{E}\|y_t\|^2 = \mathrm{tr}(\Sigma_y)$ (Van der Vaart, 2000, Theorem 2.20). For $\Sigma_{x,y} = \lim_{t\to\infty}\beta_t^{-1}\mathbb{E}[x_t y_t^\top]$, Konda and Tsitsiklis (2004, Theorem 2.6) show that for the linear case, $\Sigma_{x,y}$ satisfies the following equation

$$B_1\Sigma_{x,y} + \Sigma_x B_2^\top = \Sigma_{\xi,\psi} \tag{173}$$

Under the aforementioned uniform integrability condition, we have $\lim_{t\to\infty} \beta_t^{-1}\|\mathbb{E}x_t y_t^\top\| = \|\Sigma_{x,y}\|$. Next, we derive the upper bounds for $\mathrm{tr}(\Sigma_x)$, $\mathrm{tr}(\Sigma_y)$ and $\|\Sigma_{x,y}\|$.

The upper bound for $\mathrm{tr}(\Sigma_x)$. For Σ_x define in (170), we first derive an exponential upper bound for $\|e^{-B_1 s}\|$ and $\|e^{-B_1^\top s}\|$. By Proposition 2, we have $\frac{B_1+B_1^\top}{2} \succeq \mu_F I$. For any $v \in \mathbb{R}^d$, define $w(s) := e^{-B_1^\top s}v$. Then we have

$$\frac{d}{ds}\|w(s)\|^2 = \frac{d}{ds}\, v^\top e^{-B_1 s}e^{-B_1^\top s}v = -v^\top e^{-B_1 s}(B_1 + B_1^\top)e^{-B_1^\top s}v \le -2\mu_F\|w(s)\|^2.$$

By Grönwall's inequality, we obtain $\|w(s)\|^2 \le \|v\|^2 e^{-2\mu_F s}$. Since v is arbitrary, this implies $\|e^{-B_1^\top s}\| \le e^{-\mu_F s}$, $\forall s \ge 0$. Similarly, we have $\|e^{-B_1 s}\| \le e^{-\mu_F s}$, $\forall s \ge 0$. Next, we bound the trace of the integrand. Using the fact that $\mathrm{tr}(AB) \le \|A\|\,\mathrm{tr}(B)$ when $B \succeq 0$, we obtain

$$\begin{aligned}\mathrm{tr}\big(e^{-B_1 s}\Sigma_\xi e^{-B_1^\top s}\big) &= \mathrm{tr}\big(e^{-B_1^\top s}e^{-B_1 s}\Sigma_\xi\big) \le \big\|e^{-B_1^\top s}e^{-B_1 s}\big\|\,\mathrm{tr}(\Sigma_\xi)\\ &\le \|e^{-B_1^\top s}\|\,\|e^{-B_1 s}\|\,\mathrm{tr}(\Sigma_\xi) \le e^{-2\mu_F s}\,\mathrm{tr}(\Sigma_\xi).\end{aligned}$$

By Assumption 6, we have $\mathrm{tr}(\Sigma_\xi) \le \Gamma_{11}$. Then integrating over $s \in [0,\infty)$ yields

$$\mathrm{tr}(\Sigma_x) = \int_0^\infty \mathrm{tr}\big(e^{-B_1 s}\Sigma_\xi e^{-B_1^\top s}\big)\,ds \le \mathrm{tr}(\Sigma_\xi)\int_0^\infty e^{-2\mu_F s}\,ds = \frac{\mathrm{tr}(\Sigma_\xi)}{2\mu_F} \le \frac{\Gamma_{11}}{2\mu_F}.$$

The upper bound for $\mathrm{tr}(\Sigma_y)$. Recall that Σ_y is defined in (171). Under Assumption 7, we have $\tilde\beta < \mu_G/2$ and thus $\frac{B_3+B_3^\top-\tilde\beta I}{2} \succeq \frac{\mu_G}{2}I$. Similar to the former derivation, we can obtain $\mathrm{tr}(\Sigma_y) \le \frac{\mathrm{tr}(\widetilde\Sigma_\psi)}{\mu_G}$. It remains to bound $\mathrm{tr}(\widetilde\Sigma_\psi)$. Recall that $\widetilde\Sigma_\psi$ is defined in (172). Letting $A := B_2 B_1^{-1}$, then we have $\widetilde\Sigma_\psi = \Sigma_\psi - A\Sigma_{\xi,\psi} - \Sigma_{\xi,\psi}^\top A^\top + A\Sigma_\xi A^\top$. By Assumption 6, we have $\mathrm{tr}(\Sigma_\xi) \le \Gamma_{11}$, $\mathrm{tr}(\Sigma_\psi) \le \Gamma_{22}$ and $\|\Sigma_{\xi,\psi}\| \le \Sigma_{12}$. Then we have

$$\begin{aligned}\mathrm{tr}(\widetilde\Sigma_\psi) &= \mathrm{tr}(\Sigma_\psi) - \mathrm{tr}(A\Sigma_{\xi,\psi}) - \mathrm{tr}(\Sigma_{\xi,\psi}^\top A^\top) + \mathrm{tr}(A\Sigma_\xi A^\top)\\ &= \mathrm{tr}(\Sigma_\psi) - 2\,\mathrm{tr}(A\Sigma_{\xi,\psi}) + \mathrm{tr}(A\Sigma_\xi A^\top) \le \Gamma_{22} + 2\big|\mathrm{tr}(A\Sigma_{\xi,\psi})\big| + \mathrm{tr}(A\Sigma_\xi A^\top).\end{aligned}$$

For the last term, since $\Sigma_\xi \succeq 0$, we have

$$\mathrm{tr}(A\Sigma_\xi A^\top) = \mathrm{tr}(A^\top A\,\Sigma_\xi) \le \|A^\top A\|\,\mathrm{tr}(\Sigma_\xi) = \|A\|^2\,\mathrm{tr}(\Sigma_\xi) \le \|A\|^2\Gamma_{11}.$$

For the second term, using $|\mathrm{tr}(X)| \le d_y\|X\|$ for $X \in \mathbb{R}^{d_y\times d_y}$, we obtain

$$\big|\mathrm{tr}(A\Sigma_{\xi,\psi})\big| \le d_y\,\|A\Sigma_{\xi,\psi}\| \le d_y\,\|A\|\,\|\Sigma_{\xi,\psi}\| \le d_y\,\|A\|\,\Sigma_{12}.$$

Combining the above yields

$$\mathrm{tr}(\widetilde\Sigma_\psi) \le \Gamma_{22} + 2d_y\,\|A\|\,\Sigma_{12} + \|A\|^2\Gamma_{11}.$$

Finally, we bound $\|A\| = \|B_2B_1^{-1}\|$. By Proposition 2, we have $\|B_2\| \le L_{G,x}$, $\frac{B_1+B_1^\top}{2} \succeq \mu_F I$, and $\frac{B_3+B_3^\top}{2} \succeq \mu_G I$. For any $x \ne 0$, $\mu_F\|x\|^2 \le x^\top\frac{B_1+B_1^\top}{2}x = x^\top B_1 x \le \|B_1 x\|\,\|x\|$. It follows that $\|B_1^{-1}B_1x\| = \|x\| \le \frac{1}{\mu_F}\|B_1x\|$. Because B_1 is invertible and x is arbitrary, we have

$\|B_1^{-1}\| \leq \frac{1}{\mu_F}$. Therefore, $\|A\| = \|B_2 B_1^{-1}\| \leq \|B_2\| \, \|B_1^{-1}\| \leq \frac{L_{G,x}}{\mu_F}$. Combining the above analysis yields

$$\operatorname{tr}(\Sigma_y) \leq \frac{1}{\mu_G} \left[\Gamma_{22} + 2d_y \frac{L_{G,x}}{\mu_F} \Sigma_{12} + \Big(\frac{L_{G,x}}{\mu_F} \Big)^2 \Gamma_{11} \right].$$

The upper bound for $\|\Sigma_{x,y}\|$. Because $\Sigma_{x,y}$ satisfies (173), we have $\Sigma_{x,y} = B_1^{-1}(\Sigma_{\xi,\psi} - \Sigma_x B_2^\top)$. We have established $\|B_1^{-1}\| \leq \frac{1}{\mu_F}$ and $\operatorname{tr}(\Sigma_x) \leq \frac{\Gamma_{11}}{2\mu_F}$. By Assumption 6 and Proposition 2, $\|\Sigma_{\xi,\psi}\| \leq \Sigma_{12}$ and $\|B_2\| \leq L_{G,x}$. With $\|\Sigma_x\| \leq \operatorname{tr}(\Sigma_x)$, we have

$$\|\Sigma_{x,y}\| \leq \frac{1}{\mu_F} \Big(\|\Sigma_{\xi,\psi}\| + \|\Sigma_x\| \|B_2^\top\| \Big) \leq \frac{1}{\mu_F} \left(\Sigma_{12} + \frac{L_{G,x}\Gamma_{11}}{2\mu_F} \right).$$

Appendix D. Proof for the Lower Bound

In this section, we present the proof of Proposition 5. This proof also relies on the convergence rates for the MSE and fourth-order moments without local linearity in Theorem 8 and Lemma 14. Under the conditions in Proposition 5, one can check that the conditions of Theorem 8 and Lemma 14, especially Assumptions 7† and 9, are satisfied. Moreover, Proposition 5 together with Theorem 8 implies both $\mathbb{E}|\hat{x}_t|^2$ and $\mathbb{E}|\hat{y}_t|^2$ are of the order $\Theta(\alpha_t)$.

Proof [Proof of Proposition 5] Under the conditions of Proposition 5, the update rule of two-time-scale SA becomes

$$\begin{aligned} x_{t+1} &= x_t - \alpha_t \left(x_t - y_t + \xi_t \right), \\ y_{t+1} &= y_t - \beta_t \left(y_t - |x_t - y_t| \operatorname{sign}(y_t) \right). \end{aligned}$$

Note that for this example, $x^\star = y^\star = 0 \in \mathbb{R}$, $\hat{x}_t = x_t - y_t$ and $\hat{y}_t = y_t$. Correspondingly, the update for the errors term becomes

$$\begin{aligned} \hat{x}_{t+1} &= (1-\alpha_t)\hat{x}_t - \alpha_t \xi_t + y_t - y_{t+1} \\ &= (1-\alpha_t)\hat{x}_t - \alpha_t \xi_t + \beta_t \hat{y}_t - \beta_t |\hat{x}_t| \operatorname{sign}(\hat{y}_t), \qquad (174) \\ \hat{y}_{t+1} &= (1-\beta_t)\hat{y}_t + \beta_t |\hat{x}_t| \operatorname{sign}(\hat{y}_t). \qquad (175) \end{aligned}$$

Because this example satisfies Assumptions 1–4 and 6 with $L_H = L_F = L_{G,x} = L_{G,y} = \mu_F = \mu_G = 1$, $S_H = 0$ and $\delta_H = 1$, then by Theorem 8 and Lemma 14, we have $\mathbb{E}|\hat{x}_t|^2 + \mathbb{E}|\hat{y}_t|^2 = \mathcal{O}(\alpha_t)$ and $\mathbb{E}|\hat{x}_t|^4 = \mathcal{O}(\alpha_t^2)$. By Assumption 7†, we have $\alpha_t \leq 1/12$ and $\beta_t \leq 1/14$. We also have $\beta_t/\alpha_t \leq 1/200$.

The remaining proof proceeds in three steps. First, we show that $\mathbb{E}|\hat{x}_t|^2 = \Omega(\alpha_t)$. Second, we establish that $\mathbb{E}|\hat{x}_t| = \Omega(\sqrt{\alpha_t})$. Finally, we prove that $\mathbb{E}|\hat{y}_t| = \Omega(\sqrt{\beta_t})$, and consequently $\mathbb{E}|\hat{y}_t|^2 \geq (\mathbb{E}|\hat{y}_t|)^2 = \Omega(\beta_t)$.

Step 1: Prove $\mathbb{E}|\hat{x}_t|^2 = \Omega(\alpha_t)$**.** Squaring both sides of (174) and taking the expectation yields

$$\begin{aligned} \mathbb{E}|\hat{x}_{t+1}|^2 &= (1-\alpha_t)^2 \mathbb{E}|\hat{x}_t|^2 + \alpha_t^2 \Sigma_t + \beta_t^2 \mathbb{E}|\hat{y}_t|^2 + \beta_t^2 \mathbb{E}|\hat{x}_t|^2 + 2\beta_t(1-\alpha_t)\mathbb{E}\hat{x}_t\hat{y}_t \\ &\quad - 2\beta_t(1-\alpha_t)\mathbb{E}\hat{x}_t|\hat{x}_t| \operatorname{sign}(\hat{y}_t) - 2\beta_t^2 \mathbb{E}|\hat{x}_t\hat{y}_t| \\ &\geq (1 - 2\alpha_t - 2\beta_t)\mathbb{E}|\hat{x}_t|^2 + \alpha_t^2 \Sigma_1 - 2\beta_t(1+\beta_t)\mathbb{E}|\hat{x}_t\hat{y}_t|. \end{aligned}$$

By the AM-GM inequality, we have $2\beta_t(1+\beta_t)\mathbb{E}|\hat{x}_t\hat{y}_t| \le \alpha_t\mathbb{E}|\hat{x}_t|^2 + \frac{\beta_t^2(1+\beta_t)^2}{\alpha_t}\mathbb{E}|\hat{y}_t|^2$. Recall that $\beta_t \le 1/14$ and $\beta_t/\alpha_t \le 1/200$. It follows that

$$\mathbb{E}|\hat{x}_{t+1}|^2 \ge (1-4\alpha_t)\mathbb{E}|\hat{x}_t|^2 + \alpha_t^2\Sigma_1 - \frac{2\beta_t^2}{\alpha_t}\mathbb{E}|\hat{y}_t|^2.$$

Recall that $\mathbb{E}|\hat{y}_t|^2 = \mathcal{O}(\alpha_t)$ and $\beta_t/\alpha_t \to 0$. Then there exists t_0 such that $\forall t \ge t_0$, we have

$$\begin{aligned}\mathbb{E}|\hat{x}_{t+1}|^2 &\ge (1-4\alpha_t)\mathbb{E}|\hat{x}_t|^2 + \frac{\alpha_t^2\Sigma_1}{2} \\ &\ge \mathbb{E}|\hat{x}_{t_0}|^2\prod_{i=t_0}^{t}(1-4\alpha_i) + \frac{\Sigma_1}{2}\sum_{i=t_0}^{t}\alpha_i^2\prod_{j=i+1}^{t}(1-4\alpha_j).\end{aligned}$$

Assumption 7† implies α_i is non-increasing and $\alpha_t^{-1} \le \alpha_{t-1}^{-1} + \mu_F/16$. Then we can obtain $\alpha_t^{-1} = \mathcal{O}(t)$ and consequently $\alpha_t = \Omega(t^{-1})$. By telescoping, we have

$$\begin{aligned}\sum_{i=t_0}^{t}\alpha_i^2\prod_{j=i+1}^{t}(1-4\alpha_j) \ge \alpha_t\sum_{i=t_0}^{t}\alpha_i\prod_{j=i+1}^{t}(1-4\alpha_j) &= \frac{\alpha_t}{4}\left[1-\prod_{j=t_0}^{t}(1-4\alpha_j)\right] \\ &\ge \frac{\alpha_t}{4}\left[1-\exp\Big(-4\sum_{j=t_0}^{t}\alpha_j\Big)\right].\end{aligned}$$

Since $\alpha_t = \Omega(t^{-1})$, then there exists t_1 such that $\forall t \ge t_1$, $\sum_{i=t_0}^{t}\alpha_i^2\prod_{j=i+1}^{t}(1-4\alpha_j) \ge \alpha_t/8$. Thus, $\mathbb{E}|\hat{x}_t|^2 = \Omega(\alpha_t)$.

Step 2: Prove $\mathbb{E}|\hat{x}_t| = \Omega(\sqrt{\alpha_t})$. Combining the results of Theorem 8 and Step 1 yields $\mathbb{E}|\hat{x}_t|^2 = \Theta(\alpha_t)$. Moreover, combining the result of Lemma 14 and $\mathbb{E}|\hat{x}_t|^4 \ge (\mathbb{E}|\hat{x}_t|^2)^2$ also yields $\mathbb{E}|\hat{x}_t|^4 = \Theta(\alpha_t^2)$. Thus, there exists $\gamma_1 > 0$ such that $\forall t, (\mathbb{E}|\hat{x}_t|^2)^2/\mathbb{E}|\hat{x}_t|^4 \ge \gamma_1$. To apply this result to give a lower bound for $\mathbb{E}|\hat{x}_t|$, we need the following inequality.

Proposition 23 (Paley–Zygmund inequality) *If $Z \ge 0$ is a random variable with finite variance and $\theta \in [0,1]$, then $\mathbb{P}(Z > \theta\,\mathbb{E}Z) \ge (1-\theta)^2(\mathbb{E}Z)^2/\mathbb{E}Z^2$.*

Let $Z = |\hat{x}_t|^2$ and $\theta \in (0,1)$. Then

$$\mathbb{P}\left(|\hat{x}_t| \ge \sqrt{\theta\,\mathbb{E}|\hat{x}_t|^2}\right) = \mathbb{P}\left(|\hat{x}_t|^2 \ge \theta\,\mathbb{E}|\hat{x}_t|^2\right) \ge (1-\theta)^2\gamma_1.$$

Thus,

$$\mathbb{E}|\hat{x}_t| \ge \sqrt{\theta\,\mathbb{E}|\hat{x}_t|^2}\,\mathbb{P}\left(|\hat{x}_t| \ge \sqrt{\theta\,\mathbb{E}|\hat{x}_t|^2}\right) \ge (1-\theta)^2\sqrt{\theta}\,\gamma_1\sqrt{\mathbb{E}|\hat{x}_t|^2} = \Omega(\sqrt{\alpha_t}).$$

Step 3: Prove $\mathbb{E}|\hat{y}_t| = \Omega(\sqrt{\beta_t})$. Since $\mathbb{E}|\hat{x}_t| = \Omega(\sqrt{\alpha_t})$, there exists $\gamma_2 > 0$ such that $\forall t,\ \mathbb{E}|\hat{x}_t| \ge \gamma_2\sqrt{\alpha_t}$. Now we apply this result to derive the lower bound for $\mathbb{E}|\hat{y}_t|$. First, we note that the two terms on the right-hand side of (175) have the same sign, leading

to $\text{sign}(\hat{y}_{t+1}) = \text{sign}(\hat{y}_t)$. Because $y_0 \neq y^\star$, we have $\text{sign}(\hat{y}_0) \neq 0$. Using $\beta_t \leq 1/14$, which follows from Assumption 7†, we obtain that $\text{sign}(\hat{y}_t) \neq 0$ for all $t \geq 0$. Thus, we have

$$\begin{aligned}\mathbb{E}|\hat{y}_{t+1}| &= \mathbb{E}\hat{y}_{t+1}\,\text{sign}(\hat{y}_t) = (1-\beta_t)\mathbb{E}|\hat{y}_t| + \beta_t\mathbb{E}|\hat{x}_t| \geq (1-\beta_t)\mathbb{E}|\hat{y}_t| + \gamma_2\beta_t\sqrt{\alpha_t} \\ &\geq \mathbb{E}|\hat{y}_0|\prod_{i=0}^{t}(1-\beta_i) + \gamma_2\sum_{i=0}^{t}\beta_i\sqrt{\alpha_i}\prod_{j=i+1}^{t}(1-\beta_j).\end{aligned}$$

Similar to the analysis in Step 1, for the second term, we have

$$\sum_{i=0}^{t}\beta_i\sqrt{\alpha_i}\prod_{j=i+1}^{t}(1-\beta_i) \geq \sqrt{\alpha_t}\sum_{i=0}^{t}\beta_i\prod_{j=i+1}^{t}(1-\beta_i) \geq \sqrt{\alpha_t}\left[1-\exp\Big(-\sum_{j=0}^{t}\beta_j\Big)\right]$$

Since $\beta_t = \Omega(t^{-1})$, then there exists t_2 such that $\forall t \geq t_2$, $\sum_{i=0}^{t}\beta_i\sqrt{\alpha_i}\prod_{j=i+1}^{t}(1-\beta_i) \geq \sqrt{\alpha_t}/2$. Thus, $\mathbb{E}|\hat{y}_t| = \Omega(\sqrt{\alpha_t})$ and $\mathbb{E}|\hat{y}_t|^2 \geq \mathbb{E}(|\hat{y}_t|)^2 = \Omega(\alpha_t)$.

■

References

Guillaume O Berger, P-A Absil, Raphaël M Jungers, and Yurii Nesterov. On the quality of first-order approximation of functions with hölder continuous gradient. *Journal of Optimization Theory and Applications*, 185:17–33, 2020.

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

Vivek S Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*, volume 48. Springer, 2009.

Vivek S Borkar and Vijaymohan R Konda. The actor-critic algorithm as multi-time-scale stochastic approximation. *Sadhana*, 22:525–543, 1997.

Vivek S Borkar and Sarath Pattathil. Concentration bounds for two time scale stochastic approximation. In *2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 504–511. IEEE, 2018.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

Jerome Bracken and James T McGill. Mathematical programs with optimization problems in the constraints. *Operations Research*, 21(1):37–44, 1973.

Siddharth Chandak. $O(1/k)$ finite-time bound for non-linear two-time-scale stochastic approximation. *arXiv preprint arXiv:2504.19375*, 2025a.

Siddharth Chandak. Non-expansive mappings in two-time-scale stochastic approximation: Finite-time analysis. *arXiv preprint arXiv:2501.10806*, 2025b.

Siddharth Chandak, Shaan Ul Haque, and Nicholas Bambos. Finite-time bounds for two-time-scale stochastic approximation with arbitrary norm contractions and Markovian noise. In *2025 IEEE 64th Conference on Decision and Control (CDC)*, pages 6095–6101. IEEE, 2025.

Tianyi Chen, Yuejiao Sun, and Wotao Yin. Closing the gap: Tighter analysis of alternating stochastic gradient methods for bilevel problems. In *Advances in Neural Information Processing Systems*, volume 34, pages 25294–25307, 2021.

Zixi Chen, Yumin Xu, and Ruixun Zhang. Convergence rate in a nonlinear two-time-scale stochastic approximation with state (time)-dependence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 15993–16000, 2025.

Benoît Colson, Patrice Marcotte, and Gilles Savard. An overview of bilevel optimization. *Annals of Operations Research*, 153:235–256, 2007.

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

Thinh T Doan. Finite-time convergence rates of nonlinear two-time-scale stochastic approximation under Markovian noise. *arXiv preprint arXiv:2104.01627*, 2021.

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

Thinh T Doan. Fast nonlinear two-time-scale stochastic approximation: Achieving $\mathcal{O}(1/k)$ finite-sample complexity. *IEEE Transactions on Automatic Control*, 2025.

Fathima Zarin Faizal and Vivek Borkar. Functional central limit theorem for two timescale stochastic approximation. *arXiv preprint arXiv:2306.05723*, 2023.

Sébastien Gadat, Fabien Panloup, and Sofiane Saadane. Stochastic heavy ball. *Electronic Journal of Statistics*, 12:461–529, 2018.

Saeed Ghadimi and Mengdi Wang. Approximation methods for bilevel programming. *arXiv preprint arXiv:1802.02246*, 2018.

Igor Gitman, Hunter Lang, Pengchuan Zhang, and Lin Xiao. Understanding the role of momentum in stochastic gradient methods. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

A. M. Gupal and L. T. Bazhenov. A stochastic analog of the conjugate gradient method. *Cybernetics*, 8:138–140, 1972.

Yuze Han, Xiang Li, Jiadong Liang, and Zhihua Zhang. Decoupled functional central limit theorems for two-time-scale stochastic approximation. *arXiv preprint arXiv:2412.17070*, 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.

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.

Jie Hu, Vishwaraj Doshi, et al. Central limit theorem for two-timescale stochastic approximation with Markovian noise: Theory and applications. In *International Conference on Artificial Intelligence and Statistics*, pages 1477–1485. PMLR, 2024.

Yue Huang, Zhaoxian Wu, Shiqian Ma, and Qing Ling. Single-timescale multi-sequence stochastic approximation without fixed point smoothness: Theories and applications. *IEEE Transactions on Signal Processing*, 73:1939–1953, 2025.

Maxim Kaledin, Eric Moulines, Alexey Naumov, Vladislav Tadic, and Hoi-To Wai. Finite time analysis of linear two-timescale stochastic approximation with Markovian noise. In *Conference on Learning Theory*, pages 2144–2203. PMLR, 2020.

Vijay R Konda and John N Tsitsiklis. On actor-critic algorithms. *SIAM journal on Control and Optimization*, 42(4):1143–1166, 2003.

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

Harold Kushner and G George Yin. *Stochastic Approximation and Recursive Algorithms and Applications*, volume 35. Springer Science & Business Media, 2003.

Jeongyeol Kwon, Luke Dotson, Yudong Chen, and Qiaomin Xie. Two-timescale linear stochastic approximation: Constant stepsizes go a long way. In *International Conference on Artificial Intelligence and Statistics*, pages 3781–3789. PMLR, 2025.

Xiang Li, Jiadong Liang, Xiangyu Chang, and Zhihua Zhang. Statistical estimation and online inference via Local SGD. In *Conference on Learning Theory*, pages 1613–1661. PMLR, 2022.

Xiang Li, Jiadong Liang, and Zhihua Zhang. Online statistical inference for nonlinear stochastic approximation with Markovian data. *arXiv preprint arXiv:2302.07690*, 2023a.

Xiang Li, Wenhao Yang, Zhihua Zhang, and Michael I Jordan. A statistical analysis of Polyak-Ruppert averaged Q-learning. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2261. PMLR, 2023b.

Jiadong Liang, Yuze Han, Xiang Li, and Zhihua Zhang. Asymptotic behaviors and phase transitions in projected stochastic approximation: A jump diffusion approach. *arXiv preprint arXiv:2304.12953*, 2023.

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

Wenlong Mou, Chris Junchi Li, Martin J Wainwright, Peter L Bartlett, and Michael I Jordan. On linear stochastic approximation: Fine-grained Polyak-Ruppert and non-asymptotic concentration. In *Conference on Learning Theory*, pages 2947–2997. PMLR, 2020.

Wenlong Mou, Koulik Khamaru, Martin J Wainwright, Peter L Bartlett, and Michael I Jordan. Optimal variance-reduced stochastic approximation in Banach spaces. *arXiv preprint arXiv:2201.08518*, 2022.

Wenlong Mou, Ashwin Pananjady, and Martin J. Wainwright. Optimal oracle inequalities for projected fixed-point equations, with applications to policy evaluation. *Mathematics of Operations Research*, 48(4):2308–2336, 2023.

Eric Moulines and Francis Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, volume 24, 2011.

John Platt and Alan Barr. Constrained differential optimization. In *Neural Information Processing Systems*, 1987.

Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407, 1951.

David Ruppert. Efficient estimations from a slowly convergent robbins-monro process. Technical report, Cornell University Operations Research and Industrial Engineering, 1988.

Louis Sharrock. Two-timescale stochastic approximation for bilevel optimisation problems in continuous-time models. *arXiv preprint arXiv:2206.06995*, 2022.

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

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.

Aad W Van der Vaart. *Asymptotic Statistics*, volume 3. Cambridge university press, 2000.

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

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

Tengyu Xu and Yingbin Liang. Sample complexity bounds for two timescale value-based reinforcement learning algorithms. In *International Conference on Artificial Intelligence and Statistics*, pages 811–819. PMLR, 2021.

Tengyu Xu, Shaofeng Zou, and Yingbin Liang. Two time-scale off-policy TD learning: Non-asymptotic analysis over Markovian samples. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Tengyu Xu, Zhe Wang, and Yingbin Liang. Non-asymptotic convergence analysis of two time-scale (natural) actor-critic algorithms. *arXiv preprint arXiv:2005.03557*, 2020.

Sihan Zeng and Thinh T Doan. Fast two-time-scale stochastic gradient method with applications in reinforcement learning. In *Conference on Learning Theory*, pages 5166–5212. PMLR, 2024.

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.