A Single-Timescale Method for Stochastic Bilevel Optimization

Tianyi Chen[⋆] **Yuejiao Sun**[†] **Quan Xiao**[⋆] **Wotao Yin**[†]

[⋆]Rensselaer Polytechnic Institute [†]University of California, Los Angeles

Abstract

Stochastic bilevel optimization generalizes the classic stochastic optimization from the minimization of a single objective to the minimization of an objective function that depends on the solution of another optimization problem. Recently, bilevel optimization is regaining popularity in emerging machine learning applications such as hyper-parameter optimization and model-agnostic meta learning. To solve this class of optimization problems, existing methods require either double-loop or two-timescale updates, which are sometimes less efficient. This paper develops a new optimization method for a class of stochastic bilevel problems that we term Single-Timescale stochAstic BiLevEl optimization (**STABLE**) method. STABLE runs in a single loop fashion, and uses a single-timescale update with a fixed batch size. To achieve an ϵ-stationary point of the bilevel problem, STABLE requires $\mathcal{O}(\epsilon^{-2})$ samples in total; and to achieve an ϵ-optimal solution in the strongly convex case, STABLE requires $\mathcal{O}(\epsilon^{-1})$ samples. To the best of our knowledge, when STABLE was proposed, it is the *first* bilevel optimization algorithm achieving the same order of sample complexity as SGD for single-level stochastic optimization.

1 Introduction

In this paper, we consider solving the stochastic optimization problems of the following form

$$\min_{x\in\mathcal{X}} \quad F(x) := \mathbb{E}_{\xi}\left[f\left(x, y^*(x); \xi\right)\right] \qquad \text{(upper)} \quad \text{(1a)}$$

$$\text{s.t.} \quad y^*(x) \in \arg\min_{y\in\mathbb{R}^{d_y}} \mathbb{E}_{\phi}[g(x, y; \phi)] \quad \text{(lower)} \quad \text{(1b)}$$

where f and g are differentiable functions; ξ and ϕ are random variables; and $\mathcal{X} \subset \mathbb{R}^d$ is closed and convex set. The problem (1) is often referred to as the stochastic *bilevel* problem, where the upper-level optimization problem depends on the solution of the lower-level optimization over $y \in \mathbb{R}^{d_y}$, denoted as $y^*(x)$, which depends on the value of upper-level variable $x \in \mathcal{X}$.

Bilevel optimization can be viewed as a generalization of the classic two-stage stochastic programming (Shapiro et al., 2009), in which the upper-level objective function depends on the optimal lower-level objective value rather than the lower-level solution. Earlier works have studied applications in portfolio management and game theory (Stackelberg, 1952); see two recent surveys (Dempe and Zemkoho, 2020; Liu et al., 2021). Recently, bilevel optimization has gained growing popularity in a number of machine learning applications such as meta-learning (Rajeswaran et al., 2019), reinforcement learning (Konda and Borkar, 1999; Hong et al., 2020), hyper-parameter optimization (Franceschi et al., 2018), continual learning (Borsos et al., 2020), and image processing (Kunisch and Pock, 2013). In some of these applications, when the lower-level problem admits a closed-form solution, bilevel optimization also reduces to the recently studied stochastic compositional optimization (Wang et al., 2017a; Ghadimi et al., 2020; Chen et al., 2020).

Unlike single-level stochastic problems, algorithms tailored for solving bilevel stochastic problems are much less explored. This is partially because solving this class of problems via traditional optimization techniques faces a number of challenges. A key difficulty due to the nested structure is that (stochastic) gradient, a basic element in continuous optimization machinery, is prohibitively expensive or even impossible to compute. As we will show later, since computing an unbiased stochastic gradient of $F(x)$ requires solving the lower-level problem once, running stochastic gradient descent (SGD) on the upper-level problem essentially results in a double-loop algorithm which uses an iterative algorithm to solve the lower-level problem thousands or even millions of times.

Proceedings of the 25[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

Table 1: Sample complexity of several state-of-the-art algorithms (BSA in (Ghadimi and Wang, 2018), TTSA in (Hong et al., 2020), stocBiO in (Ji et al., 2020)) to achieve an ϵ-stationary point of $F(x)$ in the nonconvex setting and an ϵ-optimal solution of $F(x)$ in the strongly convex setting; the notation $\widetilde{\mathcal{O}}(\cdot)$ hides logarithmic terms of ϵ^{-1}.

	STABLE	**BSA**	**TTSA**	**stocBiO**
batch size	$\mathcal{O}(1)$	$\widetilde{\mathcal{O}}(1)$	$\widetilde{\mathcal{O}}(1)$	$\widetilde{\mathcal{O}}(\epsilon^{-1})$
# of loops	Single	Double	Single	Double
# of samples (nonconvex)	$\mathcal{O}(\epsilon^{-2})$ in ξ $\mathcal{O}(\epsilon^{-2})$ in ϕ	$\mathcal{O}(\epsilon^{-2})$ in ξ $\widetilde{\mathcal{O}}(\epsilon^{-3})$ in ϕ	$\mathcal{O}(\epsilon^{-2.5})$ in ξ $\widetilde{\mathcal{O}}(\epsilon^{-2.5})$ in ϕ	$\mathcal{O}(\epsilon^{-2})$ in ξ $\widetilde{\mathcal{O}}(\epsilon^{-2})$ in ϕ
# of samples (strongly convex)	$\mathcal{O}(\epsilon^{-1})$ in ξ $\mathcal{O}(\epsilon^{-1})$ in ϕ	$\mathcal{O}(\epsilon^{-1})$ in ξ $\widetilde{\mathcal{O}}(\epsilon^{-2})$ in ϕ	$\mathcal{O}(\epsilon^{-1.5})$ in ξ $\widetilde{\mathcal{O}}(\epsilon^{-1.5})$ in ϕ	/
complexity of y update	$\mathcal{O}(d_y^3)$	$\mathcal{O}(d_y^2)$	$\mathcal{O}(d_y^2)$	$\mathcal{O}(d_y^2)$

1.1 Prior art

To put our work in context, we review prior art that we group in the following two categories.

Bilevel optimization. Bilevel optimization has a long history in operations research, where the lower level problem is served as the constraint of the upper level problem (Bracken and McGill, 1973; Ye and Zhu, 1995; Vicente and Calamai, 1994; Colson et al., 2007). Many recent efforts have been made to solve the bilevel optimization problems. One successful approach is to reformulate the bilevel problem as a single-level problem by replacing the lower-level problem by its optimality conditions (Colson et al., 2007; Kunapuli et al., 2008). Recently, gradient-based first-order methods for bilevel optimization have gained popularity, where the idea is to iteratively approximate the (stochastic) gradient of the upper-level problem either in forward or backward manner (Sabach and Shtern, 2017; Franceschi et al., 2018; Shaban et al., 2019; Grazzi et al., 2020). While most of these works assume the unique solution of the lower-level problem, cases where this assumption does not hold have been tackled in the recent work (Liu et al., 2020). All these algorithms have excellent empirical performance, but many of them either provide no theoretical guarantees or only focus on the asymptotic performance analysis.

The non-asymptotic analysis of bilevel optimization algorithms has been recently studied in some *pioneering* works, e.g., (Ghadimi and Wang, 2018; Hong et al., 2020; Ji et al., 2020), just to name a few. In both (Ghadimi and Wang, 2018; Ji et al., 2020), bilevel stochastic optimization algorithms have been developed that run in a double-loop manner. To achieve an ϵ-stationary point, they only need the sample complexity $\mathcal{O}(\epsilon^{-2})$ that is comparable to the complexity of SGD for the single-level case. Recently, a single-loop two-timescale stochastic approximation algorithm has been developed in (Hong et al., 2020) for the bilevel problem (1). Due to the nature of two-timescale update, it incurs the sub-optimal sample complexity $\mathcal{O}(\epsilon^{-2.5})$. Therefore, the existing single-loop solvers for bilevel problems are significantly slower than those for problems without bilevel compositions, but otherwise share many structures and properties.

Concurrent work. After our STABLE was developed and released, its rate of convergence was improved to $\mathcal{O}(\epsilon^{-1.5})$ by momentum accelerations in (Khanduri et al., 2021; Guo and Yang, 2021; Yang et al., 2021). The adaptive gradient variant has been studied in (Huang and Huang, 2021). Besides, a tighter analysis for alternating stochastic gradient descent (ALSET) method was proposed in (Chen et al., 2021). The contributions compared to ALSET are: (a) ALSET uses SGD on the lower level but STABLE has a correction term, so STABLE has a reduced stochastic oracle complexity; (b) STABLE can handle the *constrained* upper-level problem using *Moreau envelop*.

Stochastic compositional optimization. When the lower-level problem in (1b) admits a smooth closed-form solution, the bilevel problem (1) reduces to stochastic compositional optimization. Popular approaches tackling this class of problems use two sequences of variables being updated in two different time scales (Wang et al., 2017a,b). However, the complexity of (Wang et al., 2017a) and (Wang et al., 2017b) is worse than $\mathcal{O}(\epsilon^{-2})$ of SGD for the non-compositional case. Building upon recent variance-reduction techniques, variance-reduced methods have been developed to solve a special class of the stochastic compositional problem with the *finite-sum structure*, e.g., (Lian et al., 2017; Zhang and Xiao, 2019), but they usually operate in a double-loop manner. Other related compositional algorithms also include (Tran-Dinh et al., 2020; Hu et al., 2020).

While most of existing algorithms rely on either two-timescale or double-loop updates, the single-timescale

single-loop approaches have been recently developed in (Ghadimi et al., 2020; Chen et al., 2020), which achieve the sample complexity $\mathcal{O}(\epsilon^{-2})$. These encouraging recent results imply that solving stochastic compositional optimization is nearly as easy as solving stochastic optimization.

Our work is also related to the stochastic min-max optimization; see e.g., (Daskalakis and Panageas, 2018; Luo et al., 2020; Rafique et al., 2021; Mokhtari et al., 2020; Lin et al., 2020; Nouiehed et al., 2019). However, whether the techniques used in compositional and min-max optimization permeate to solving more challenging bilevel problems remains unknown. This paper is devoted to answering this question.

1.2 Our contributions

To this end, this paper aims to develop a *single-loop single-timescale* stochastic algorithm, which, for the class of smooth bilevel problems, can match the sample complexity of SGD for single-level stochastic optimization problems. In the context of existing methods, our contributions can be summarized as follows.

C1) We develop a new stochastic gradient estimator tailored for a certain class of stochastic bilevel problems, which is motivated by an ODE analysis for the corresponding continuous-time deterministic problems. Our new stochastic bilevel gradient estimator is flexible to combine with any existing stochastic optimization algorithms for the single-level problems, and solve this class of stochastic bilevel problems as sample-efficient as single-level problems.

C2) When we combine this stochastic gradient estimator with SGD for the upper-level update, we term it as the Single-Timescale stochAstic BiLevEl optimization (**STABLE**) method. In the nonconvex case, to achieve ϵ-stationary point of (1), STABLE only requires $\mathcal{O}(\epsilon^{-2})$ samples in total. In the strongly convex case, to achieve ϵ-optimal solution of (1), STABLE only requires $\mathcal{O}(\epsilon^{-1})$ samples. This is achieved by designing a new Lyapunov function. To the best of our knowledge, when STABLE was proposed, it is the *first* bilevel algorithm achieving the order of sample complexity as SGD. See the sample complexity of state-of-the-art algorithms in Table 1.

Trade-off and limitations. While our new bilevel algorithm significantly improves the sample complexity of existing algorithms, it pays the price of additional computation per iteration. Specifically, in order to better estimate the stochastic bilevel gradient, a matrix inversion and an eigenvalue truncation are needed per iteration, which cost $\mathcal{O}(d^3)$ computation for a $d \times d$ matrix. In contrast, some of recent works (Ghadimi and Wang, 2018; Hong et al., 2020; Ji et al., 2020) reduce matrix inversion to more efficient computations of matrix-vector products, which cost $\mathcal{O}(d^2)$ computation per iteration. Therefore, our algorithm is preferable in the regime where the sampling is more costly than computation or the dimension d is relatively small.

2 A Single-timescale Optimization Method for Bilevel Problems

In this section, we will first provide background of bilevel problems, and then present our stochastic bilevel gradient method, followed by an ODE analysis to highlight the intuition of our design.

2.1 Preliminaries

We use $\|\cdot\|$ to denote the ℓ_2 norm for vectors and Frobenius norm for matrices. We use $\mathcal{F}^k$ to denote the collection of random variables, i.e., $\mathcal{F}^k := \{\phi^0, \ldots, \phi^{k-1}, \xi^0, \ldots, \xi^{k-1}\}$. We define the deterministic version of (1) without constraint on $\mathcal{X}$ as

$$\begin{aligned} &\min_{x\in\mathbb{R}^d} F(x) := f\left(x, y^*(x)\right) \\ &\text{s.t. } y^*(x) \in \arg\min_{y\in\mathbb{R}^{d_y}} g(x,y) \end{aligned} \tag{2}$$

where the functions are defined as $g(x,y) := \mathbb{E}_\phi[g(x,y;\phi)]$ and $f(x,y) := \mathbb{E}_\xi[f(x,y;\xi)]$.

We also define $\nabla^2_{yy} g\left(x,y\right)$ as the Hessian matrix of g with respect to y and define $\nabla^2_{xy} g\left(x,y\right)$ as

$$\nabla^2_{xy} g\left(x,y\right) := \begin{bmatrix} \frac{\partial^2}{\partial x_1 \partial y_1} g\left(x,y\right) & \cdots & \frac{\partial^2}{\partial x_1 \partial y_{d_y}} g\left(x,y\right) \\ & \cdots & \\ \frac{\partial^2}{\partial x_d \partial y_1} g\left(x,y\right) & \cdots & \frac{\partial^2}{\partial x_d \partial y_{d_y}} g\left(x,y\right) \end{bmatrix}.$$

We make the following standard assumptions that are commonly used in stochastic bilevel optimization literature (Ghadimi and Wang, 2018; Hong et al., 2020; Ji et al., 2020; Khanduri et al., 2021; Guo and Yang, 2021).

Assumption 1 (Lipschitz continuity). *For any x, $\nabla_x f(x,\cdot)$, $\nabla_y f(x,\cdot)$, $\nabla_y g(x,y)$, $\nabla^2_{xy} g(x,\cdot;\phi)$, $\nabla^2_{yy} g(x,\cdot;\phi)$ are $L_{f_x}, L_{f_y}, L_g, L_{g_{xy}}, L_{g_{yy}}$-Lipschitz continuous. For any fixed y, $\nabla_x f(\cdot,y;\xi)$, $\nabla_y f(\cdot,y;\xi)$, $\nabla^2_{xy} g(\cdot,y;\phi)$, $\nabla^2_{yy} g(\cdot,y;\phi)$ are $\bar{L}_{f_x}, \bar{L}_{f_y}, \bar{L}_{g_{xy}}, \bar{L}_{g_{yy}}$-Lipschitz continuous.*

Assumption 2 (strong convexity of lower-level objective). *For any fixed x, $g(x,y)$ is μ_g-strongly convex in y, that is, $\nabla^2_{yy} g(x,y) \succeq \mu_g I$.*

Assumptions 1 and 2 together ensure that the first- and second-order derivations of $f(x,y), g(x,y)$ as well as the solution mapping $y^*(x)$ are well-behaved.

Assumption 3 (stochastic derivatives). *The stochastic derivatives $\nabla_x f(x,y;\xi)$, $\nabla_y f(x,y;\xi)$, $\nabla_y g(x,y;\phi)$, $\nabla^2_{xy} g(x,y,\phi)$, and $\nabla^2_{yy} g(x,y,\phi)$ are unbiased estimators of $\nabla_x f(x,y)$, $\nabla_y f(x,y)$, $\nabla_y g(x,y)$, $\nabla^2_{xy} g(x,y)$, and $\nabla^2_{yy} g(x,y)$, respectively; and their variances are bounded by $\sigma^2_{f_x}, \sigma^2_{f_y}$, $\sigma^2_{g_y}$, $\sigma^2_{g_{xy}}, \sigma^2_{g_{yy}}$, respectively. Moreover, their moments are bounded by*

$$\mathbb{E}_\xi[\|\nabla_x f(x,y;\xi)\|^p] \le C^p_{f_x},\ p=2,4$$
$$\mathbb{E}_\xi[\|\nabla_y f(x,y;\xi)\|^p] \le C^p_{f_y},\ p=2,4 \tag{3a}$$
$$\mathbb{E}_\phi[\|\nabla^2_{xy} g(x,y;\phi)\|^2] \le C^2_{g_{xy}},$$
$$\mathbb{E}_\phi[\|\nabla^2_{yy} g(x,y;\phi)\|^2] \le C^2_{g_{yy}}. \tag{3b}$$

Assumption 3 is the counterpart of the unbiasedness and bounded variance assumption in the single-level stochastic optimization, which are standard also in (Ghadimi and Wang, 2018; Hong et al., 2020). In addition, the bounded moments in Assumption 3 ensure the Lipschitz continuity of the upper-level gradient $\nabla F(x)$.

We first highlight the inherent challenge of directly applying the single-level SGD method (Robbins and Monro, 1951) to the bilevel problem (1). To illustrate this point, we derive the gradient of the upper-level function $F(x)$ in the next proposition by analyzing the lower-level optimality condition; see the proof in Appendix C.

Proposition 1 *Under Assumption 2, we have the gradients*

$$\nabla_x y^*(x)^\top := -\nabla^2_{xy} g(x,y^*(x))\left[\nabla^2_{yy} g(x,y^*(x))\right]^{-1} \tag{4a}$$
$$\nabla F(x) = \nabla_x f(x,y^*(x)) + \nabla_x y^*(x)^\top \nabla_y f(x,y^*(x)). \tag{4b}$$

Note that the gradient $\nabla F(x)$ contains the second-order information of the lower-level problem $g(x,y)$ since it depends on the sensitivity of the lower-level solution $y^*(x)$. The sensitivity of the solution $y^*(x)$ for a strongly-convex program has also been explored in the time-varying convex optimization literature through the lens of perturbation analysis; see e.g., (Simonetto et al., 2016). Therefore, we hope that the sample complexity results of bilevel optimization in this paper will also stimulate future research in time-varying convex optimization.

In addition, notice that obtaining an unbiased stochastic estimate of $\nabla F(x)$ and applying SGD on x face two main difficulties: **(D1)** the gradient $\nabla F(x)$ at x depends on the minimizer of the lower-level problem $y^*(x)$; **(D2)** even if $y^*(x)$ is known, it is hard to apply the stochastic approximation to obtain an unbiased estimate of $\nabla F(x)$ since $\nabla F(x)$ is nonlinear in $\nabla^2_{yy} g(x,y^*(x))$; see the discussion of (D2) in stochastic compositional optimization literature, e.g., (Wang et al., 2017a; Chen et al., 2020).

Similar to some existing algorithms for bilevel problems, our method addresses (D1) by evaluating $\nabla F(x)$ on a certain vector y in place of $y^*(x)$, but it differs in how to recursively update y and how to address (D2). Resembling the definition (4) with $y^*(x)$ replaced by y, we introduce the notation

$$\overline{\nabla}_x f(x,y) := \nabla_x f(x,y) - \nabla^2_{xy} g(x,y) \times \left[\nabla^2_{yy} g(x,y)\right]^{-1} \nabla_y f(x,y). \tag{5}$$

As we will show in Lemma 5 of Appendix, Assumptions 1-3 ensure that $\nabla F(\cdot)$, $\overline{\nabla}_x f(x,\cdot)$, and $y^*(\cdot)$ are all Lipschitz continuous with constants L_F, L_f, L_y, respectively.

2.2 A single-timescale bilevel method

Before we present our method, we first review a successful recent effort. To overcome the difficulty of applying plain-vanilla SGD, a *two-timescale* stochastic approximation (TTSA) algorithm has been recently developed in (Hong et al., 2020). TTSA is a single-loop algorithm and amenable to efficient implementation. It consists of two sequences $\{x^k\}$ and $\{y^k\}$: for a given x^k, y^k estimates the minimizer $y^*(x^k)$; and, x^k estimates the minimizer x^*. For notational brevity, we define

$$h^k_g := \nabla_y g(x^k,y^k;\phi^k), \quad h^k_{yy}(\phi) := \nabla^2_{yy} g(x^k,y^k;\phi),$$
$$h^k_{xy}(\phi) := \nabla^2_{xy} g(x^k,y^k;\phi). \tag{6}$$

With α_k and β_k denoting two sequences of stepsizes, the TTSA recursion is given by

$$y^{k+1} = y^k - \beta_k h^k_g \tag{7a}$$
$$x^{k+1} = \mathcal{P}_{\mathcal{X}}\left(x^k - \alpha_k\left(\nabla_x f(x^k,y^k;\xi^k) - h^k_{xy}(\phi^k)\nabla^{-1}_{yy}\nabla_y f(x^k,y^k;\xi^k)\right)\right) \tag{7b}$$

where ∇^{-1}_{yy} is a mini-batch approximation of $\left[\nabla^2_{yy} g(x^k,y^k)\right]^{-1}$. The timescale separation refers to the different order of stepsizes used in updating multiple variables. To ensure convergence, TTSA requires y^k to be updated in a timescale faster than that of x^k so that x^k is relatively static with respect to y^k; i.e., $\lim_{k\to\infty} \alpha_k/\beta_k = 0$ (Hong et al., 2020). This is termed the two-timescale update. However, this prevents TTSA from choosing the stepsize $\mathcal{O}(1/\sqrt{k})$ as SGD, and also results in its *suboptimal complexity*.

We find that the key reason preventing TTSA from using a single-timescale update is its undesired stochastic upper-level gradient estimator (7b) that uses an inaccurate lower-level variable y^k to approximate $y^*(x^k)$.

Algorithm 1 STABLE for stochastic bilevel problems

initialize: $x^0, y^0, H_{xy}^0, H_{yy}^0$, stepsizes $\{\alpha_k, \beta_k\}$.
for $k = 0, 1, \ldots, K-1$ **do**
  compute $h_{xy}^{k-1}(\phi^k)$ and $h_{xy}^k(\phi^k)$
  ▷ randomly select datum ϕ^k
  update H_{xy}^k via (9a)
  compute $h_{yy}^{k-1}(\phi^k)$ and $h_{yy}^k(\phi^k)$
  update H_{yy}^k via (9b)
  compute $\nabla_x f\left(x^k, y^k; \xi^k\right)$, $\nabla_y f\left(x^k, y^k; \xi^k\right)$
  ▷ randomly select datum ξ^k
  update x^k and y^k via (8)
end for

With more insights given in Section 2.3, we propose a new stochastic bilevel optimization method based on a new stochastic bilevel gradient estimator, which we term Single-Timescale stochAstic BiLevEl optimization (**STABLE**) method. Its recursion is given by

$$x^{k+1} = \mathcal{P}_{\mathcal{X}}\left(x^k - \alpha_k\left(\nabla_x f(x^k, y^k; \xi^k) - H_{xy}^k (H_{yy}^k)^{-1} \nabla_y f(x^k, y^k; \xi^k)\right)\right) \quad (8a)$$

$$y^{k+1} = y^k - \beta_k h_g^k - (H_{yy}^k)^{-1}(H_{xy}^k)^\top (x^{k+1} - x^k). \quad (8b)$$

where $\mathcal{P}_{\mathcal{X}}$ denotes the projection on set $\mathcal{X}$. In (8), the estimates of second-order derivatives are updated as (with stepsize $\tau_k > 0$)

$$H_{xy}^k = \overline{\mathcal{P}}\left((1-\tau_k)\left(H_{xy}^{k-1} - h_{xy}^{k-1}(\phi^k)\right) + h_{xy}^k(\phi^k)\right) \quad (9a)$$

$$H_{yy}^k = \underline{\mathcal{P}}\left((1-\tau_k)\left(H_{yy}^{k-1} - h_{yy}^{k-1}(\phi^k)\right) + h_{yy}^k(\phi^k)\right) \quad (9b)$$

where $\overline{\mathcal{P}}$ is the projection to set $\{X : \|X\| \le C_{g_{xy}}\}$ and $\underline{\mathcal{P}}$ is the projection to set $\{X : X \succeq \mu_g I\}$.

Compared with (7) and other existing algorithms, the unique features of STABLE lie in: **(F1)** its y^k-update that will be shown to better "predict" the next $y^*(x^{k+1})$; and, **(F2)** a recursive update of H_{xy}^k, H_{yy}^k that is motivated by the advanced variance reduction techniques for single-level nonconvex optimization problems such as STORM (Cutkosky and Orabona, 2019), Hybrid SGD(Tran-Dinh et al., 2021) and the recent stochastic compositional optimization method (Chen et al., 2020). The marriage of (F1)-(F2) enables STABLE to have a better estimate of $\nabla F(x^k)$, which is responsible for its improved convergence. Note that we use three stepsizes α_k, β_k and τ_k in (8), we call our method a single-timescale algorithm because the upper- and lower-level variables use the same order of stepsizes that decrease at the same rate as that of SGD. As we will show later, for a class of bilevel problems, the single-timescale recursion (8) achieves the same convergence rate as SGD for single-level problems. See a summary of STABLE in Algorithm 1.

Remark. The projection in (9) is introduced for our current analysis. However, projection in (9a) is not uncommon in stochastic algorithms to ensure stability, and the eigenvalue truncation in (9b) is a usual subroutine in Newton-based methods, which is also referred to the positive definite truncation (Nocedal and Wright, 2006; Paternain et al., 2019). One potential way to avoid it is to replace (9) with a trust region computation.

2.3 Continuous-time ODE analysis

Similar to the stochastic compositional optimization (Chen et al., 2020), we provide some intuition of our algorithm design via an ODE for the deterministic problem (2). To minimize $F(x)$, we use an ODE analysis to design a continuous dynamic

$$\dot{x}(t) = -\alpha \mathcal{T}(x(t), y(x(t))) \quad (10)$$

by choosing an operator $\mathcal{T}$. For single-level minimization of a smooth function $h(x(t))$, one can use the gradient flow $\dot{x}(t) = -\alpha \nabla h(x(t))$. For bilevel minimization (2), however, we shall avoid $\mathcal{T}(x, y) = \nabla_x \left(f(x, y^*(x))\right.$ and instead use y to approximate $y^*(x)$. Here note that we have dropped (t) for conciseness. Hence, define the operator as

$$\mathcal{T}(x,y) := \nabla_x f(x,y) - \nabla_{xy}^2 g(x,y)[\nabla_{yy}^2 g(x,y)]^{-1} \times \nabla_y f(x,y) \overset{(5)}{=} \overline{\nabla}_x f(x,y). \quad (11)$$

Here, the variable y follows another dynamic that we specify below, which accompanies the x-dynamic (10). We will also find a Lyapunov function V such that

(C1) $\dot{V} < 0$;
(C2) $\dot{V} = 0$ if and only if $\nabla F(x) = 0$ and $y = y^*(x)$.

If the $\dot{x}$ and $\dot{y}$ dynamics drive an appropriate Lyapunov function V satisfying (C1) and (C2), then x converges to a stationary point of the upper-level problem and y converges to the solution of the lower-level problem.

We first state the results for the continuous-time dynamics below.

Theorem 1 (Continuous-time dynamics) *If we define the x- and y-dynamics as*

$$\dot{x} = -\alpha \nabla_x f(x,y) - \alpha \nabla_{xy}^2 g(x,y)[\nabla_{yy}^2 g(x,y)]^{-1} \nabla_y f(x,y)$$
$$\dot{y} = -\beta \nabla_y g(x,y) - \left[\nabla_{yy}^2 g\left(x,y\right)\right]^{-1} \nabla_{yx}^2 g\left(x,y\right) \dot{x} \quad (12)$$

and choose the constants α and β appropriately, then there exists a Lyapunov function V of the x- and y-dynamics that satisfies (C1) and (C2).

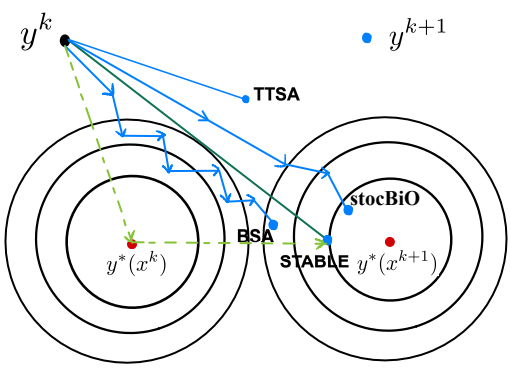

Figure 1: A geometric illustration of the y^k update under the state-of-the-art algorithms; black dot represents y^k, red dots represent the lower-level solution $y^*(x^k)$ and $y^*(x^{k+1})$, blue dots represent y^{k+1} under different algorithms, and blue arrow denotes the inner loop updates. **STABLE** updates y^k by linearly combining the stochastic gradient direction towards $y^*(x^k)$ and the moving direction from $y^*(x^k)$ to $y^*(x^{k+1})$. In contrast, **BSA** (Ghadimi and Wang, 2018) runs multiple stochastic gradient steps; **TTSA** (Hong et al., 2020) runs one stochastic gradient step with a smaller stepsize; **stocBiO** (Ji et al., 2020) runs multiple stochastic gradient steps with an increasing batch size.

Proof: To highlight the intuition, we provide a constructive proof of this theorem. We first try $V_0 := f(x, y^*(x))$. To clarify, we can use $y^*(x)$ in a Lyapunov function but not in a dynamic to evolve a quantity. In this case, we have

$$\begin{aligned}\dot{V}_0 &= \langle \nabla_x f(x, y^*(x)), \dot{x}\rangle + \langle \nabla_y f(x, y^*(x)), \nabla_x y^*(x)\dot{x}\rangle \\ &= \langle \nabla_x f(x, y^*(x)) + \nabla_x y^*(x)^\top \nabla_y f(x, y^*(x)), \dot{x}\rangle.\end{aligned}$$

Recall the definition in (4). Then we have

$$\begin{aligned}\dot{V}_0 &= -\alpha\langle \mathcal{T}(x, y^*(x)), \mathcal{T}(x, y)\rangle \\ &\overset{(a)}{\le} -\alpha\|\mathcal{T}(x, y^*(x))\|^2 \\ &\quad + \alpha\|\overline{\nabla}_x f(x, y) - \overline{\nabla}_x f(x, y^*(x))\|\|\mathcal{T}(x, y^*(x))\| \\ &\overset{(b)}{\le} -\alpha\|\mathcal{T}(x, y^*(x))\|^2 \\ &\quad + \alpha L_f\|y - y^*(x)\|\|\mathcal{T}(x, y^*(x))\| \\ &\overset{(c)}{\le} -\frac{\alpha}{2}\|\mathcal{T}(x, y^*(x))\|^2 + \frac{\alpha L_f^2}{2}\|y - y^*(x)\|^2 \qquad (13)\end{aligned}$$

where (a) uses the Cauchy-Schwarz inequality, (b) follows from the L_f-Lipschitz continuity of $\overline{\nabla}_x f(x, \cdot)$ established in Lemma 5, and (c) is due to the Young's inequality.

To satisfy (C1), we have $\dot{V}_0 \le 0$ only if $L_f\|y - y^*(x)\| \le \|\mathcal{T}(x, y^*(x))\|$, thus, requiring the information of $\|y - y^*(x)\|$ — not doable without knowing $y^*(x)$.

Let us try to mitigate the term $\|y(x) - y^*(x)\|^2$ by defining the following new Lyapunov function:

$$V := f(x, y^*(x)) + \frac{1}{2}\|y - y^*(x)\|^2 \qquad (14)$$

which implies that

$$\begin{aligned}\dot{V} &= -\alpha\langle \mathcal{T}(x, y^*(x)), \mathcal{T}(x, y)\rangle + \langle y - y^*(x), \dot{y} - \nabla_x y^*(x)\dot{x}\rangle \\ &\overset{(13)}{\le} -\frac{\alpha}{2}\|\mathcal{T}(x, y^*(x))\|^2 + \frac{\alpha L_f^2}{2}\|y - y^*(x)\|^2 \\ &\quad + \langle y - y^*(x), \dot{y} - \nabla_x y^*(x)\dot{x}\rangle \qquad (15) \\ &\le -\frac{\alpha}{2}\|\mathcal{T}(x, y^*(x))\|^2 - \left(\beta - \frac{\alpha L_f^2}{2}\right)\|y - y^*(x)\|^2 \\ &\quad + \langle y - y^*(x), \dot{y} + \beta(y - y^*(x)) - \nabla_x y^*(x)\dot{x}\rangle \qquad (16)\end{aligned}$$

where $\beta > 0$ is a fixed constant. The first two terms in the RHS of (16) are non-positive given that $\alpha \ge 0$ and $\beta \ge \alpha L_f^2/2$, but the last term can be either positive or negative. To control the last term and thus ensure the descent of $V(t)$, we are motivated to use a y-dynamic like

$$\dot{y} \approx -\beta(y - y^*(x)) + \nabla_x y^*(x)\dot{x}. \qquad (17)$$

To avoid using y^* in a dynamic, we approximate $y - y^*(x)$ by $\nabla_y g(x, y)$ and $\nabla_x y^*(x)$ by (cf. (4a))

$$\nabla_x y(x) := -\left[\nabla_{yy}^2 g(x, y)\right]^{-1}\nabla_{xy}^2 g(x, y). \qquad (18)$$

These choices lead to the y-dynamics:

$$\dot{y} = -\beta\nabla_y g(x, y) + \nabla_x y(x)\dot{x}. \qquad (19)$$

Although we approximate (17) by (19), we will plug y-dynamics (19) into (16) and show that V satisfies (C1). Specifically, plugging (19) into (15) leads to

$$\begin{aligned}&\langle y - y^*(x), \dot{y} - \nabla_x y^*(x)\dot{x}\rangle \\ =& -\langle y - y^*(x), \beta\nabla_y g(x, y) \\ &- \nabla_x y(x)\dot{x} + \nabla_x y^*(x)\dot{x}\rangle. \qquad (20)\end{aligned}$$

As $g(x, \cdot)$ is μ_g-strongly convex by Assumption 2, we have

$$\langle y - y^*(x), \nabla_y g(x, y) - \nabla_y g(x, y^*(x))\rangle \ge \mu_g\|y - y^*(x)\|^2 \qquad (21)$$

where $\nabla_y g(x, y^*(x)) = 0$ as $y^*(x)$ minimizes $g(x, y)$.

Therefore, plugging (21) into (20), we have

$$\begin{aligned}&\langle y - y^*(x), \dot{y} - \nabla_x y^*(x)\dot{x}\rangle \\ \le& -\langle y - y^*(x), (\nabla_x y^*(x) - \nabla_x y(x))\dot{x}\rangle - \beta\mu_g\|y - y^*(x)\|^2 \\ \le& \|y - y^*(x)\|\|\nabla_x y^*(x) - \nabla_x y(x)\|\|\dot{x}\| - \beta\mu_g\|y - y^*(x)\|^2 \\ \le& \alpha B_x L_y\|y - y^*(x)\|^2 - \beta\mu_g\|y - y^*(x)\|^2 \qquad (22)\end{aligned}$$

where the second inequality uses the Cauchy-Schwarz inequality, and the last inequality follows the bound

B_x of $\|\dot{x}\|$ and the Lipschitz constant L_y of $\nabla_x y(x)$, both of which can be derived from Assumptions 1–3.

Now plugging (22) into (15), we have

$$\dot{V} \leq -\frac{\alpha}{2}\|\mathcal{T}(x, y^*(x))\|^2 - \left(\beta\mu_g - \frac{\alpha L_f^2}{2} - \alpha B_x L_y\right)\|y - y^*(x)\|^2. \quad (23)$$

Now let us check (C1) and (C2). To ensure $\dot{V} \leq 0$ in (C1), we can set $\alpha \leq \frac{2\mu_g\beta}{L_f^2+2B_xL_y}$. For (C2), we have $\dot{V} = 0$ if and only if $y = y^*(x)$ and $\mathcal{T}(x, y^*(x)) = \nabla F(x) = 0$.

With the insights gained from the continuous-time update (12), our stochastic update (8) essentially discretizes time t into iteration k, and replaces the first- and second-order derivatives in $\dot{x}$ and $\dot{y}$ by their recursive (variance-reduced) stochastic values in (9).

Remark. The key ingredient of our STABLE method is the design of the lower-level update on y^k, which leads to a more accurate stochastic estimate of $\nabla F(x^k)$. See a comparison of the y-update with other algorithms in Figure 1. In the update (8), we implement the SGD-like update for the upper-level variable x^k. With the lower-level y^k update unchanged, it is easy to apply SGD-improvement techniques such as momentum and variance reduction, to accelerate the convergence of STABLE. This will help STABLE achieve state-of-the-art performance for stochastic bilevel optimization.

3 Convergence Analysis

In this section, we establish the convergence rate of our single-timescale STABLE algorithm. We will highlight the key steps of the proof and leave the detailed analysis in Appendix.

Moreau Envelop. Different from problem (2), (1) tackles the constraint on a convex and closed set $\mathcal{X}$. To levarage the ODE analysis to the constraint case, for fixed $\rho > 0$, we define the Moreau envelop and proximal map as follows.

$$\Phi_{1/\rho}(z) := \min_{x\in\mathcal{X}} \left\{F(x) + (\rho/2)\|x - z\|^2\right\}$$
$$\widehat{x}(z) := \arg\min_{x\in\mathcal{X}} \left\{F(x) + (\rho/2)\|x - z\|^2\right\} \quad (24)$$

For any $\epsilon > 0$, we use the definition in (Davis and Drusvyatskiy, 2018) that $x^k \in \mathcal{X}$ is an ϵ-nearly stationary solution if x^k satisfies the following condition

$$\mathbb{E}\left[\|\widehat{x}(x^k) - x^k\|^2\right] \leq \rho^2\epsilon. \quad (25)$$

In Section 3.1, we will utilize the near-stationarity condition (25) as a tool to quantify the convergence of STABLE when $F(x)$ is non-convex.

3.1 Main results

We first present the result of our algorithm when the upper-level function $F(x)$ is nonconvex in x. We need the following additional assumption.

Assumption 4 (weak convexity). *Function $F(x)$ is μ_F-weakly convex in x, that is, $\nabla^2_{xx}F(x) \succeq \mu_F I$.* Note that μ_F is not necessarily positive.

For simplicity of the convergence analysis, we define the following Lyapunov function

$$\begin{aligned}\mathbb{V}^k :=& \Phi_{1/\rho}(x^k) + \|y^k - y^*(x^k)\|^2 \\ &+ \|H^k_{yy} - \nabla^2_{yy}g(x^k, y^k)\|^2 \\ &+ \|H^k_{xy} - \nabla^2_{xy}g(x^k, y^k)\|^2 \end{aligned} \quad (26)$$

which mimics the continuous-time Lyapunov function (14) for the deterministic problem. Similar to the ODE analysis, we need to quantify the difference between two Lyapunov functions $\mathbb{V}^{k+1} - \mathbb{V}^k$. We will first analyze the descent of the Moreau Envelop of the upper-level objective in the next lemma.

Lemma 2 (Descent of the upper level) *Under Assumptions 1–4, the sequence of x^k satisfies*

$$\begin{aligned}&\mathbb{E}[\Phi_{1/\rho}(x^{k+1})] - \mathbb{E}[\Phi_{1/\rho}(x^k)] \\ &\leq -\frac{(\mu_F+\rho)\rho\alpha_k}{4}\|\widehat{x}(x^k) - x^k\|^2 + 2\rho\alpha_k^2\left(C_{f_x}^2 + \frac{C_{g_{xy}^2}C_{f_y}^2}{\mu_g^2}\right) \\ &\quad + \frac{2L_f^2\rho\alpha_k}{\mu_F+\rho}\|y^k - y^*(x^k)\|^2 \\ &\quad + \frac{4C_{f_y}^2C_{g_{xy}}^2\rho\alpha_k}{(\mu_F+\rho)\mu_g^4}\mathbb{E}[\|H^k_{yy} - \nabla^2_{yy}(x^k, y^k)\|^2] \\ &\quad + \frac{4C_{f_y}^2\rho\alpha_k}{(\mu_F+\rho)\mu_g^2}\mathbb{E}[\|H^k_{xy} - \nabla^2_{xy}(x^k, y^k)\|^2] \end{aligned} \quad (27)$$

where L_f, L_F are defined in Lemma 5 of Appendix, and $C_{g_{xy}}$ is the projection radius in (9a).

Lemma 2 implies that the descent of the Moreau Envelop of the upper-level objective functions depends on the error of the lower-level variable y^k, and the estimation errors of H^k_{yy} and H^k_{xy}. After bounding all of them in Lemmas 3 and 4 in Appendix, we can get the following convergence result.

Theorem 2 (Nonconvex) *Under Assumptions 1–4*

and setting $\rho > |\mu_F|$, if we choose the stepsizes as

$$\beta_k \le \min\left\{\frac{1}{\sqrt{K}}, \frac{\mu_g/L_g}{32(\mu_g+L_g)c}\right\} \tag{28a}$$

$$\alpha_k \le \min\left\{\beta_k, \frac{(c+4\rho C_{g_{xy}}^2 C_{f_y}^2/(\mu_F+\rho)\mu_g^4)^{-1}}{\sqrt{K}}, \frac{(c+4\rho C_{f_y}^2/(\mu_F+\rho)\mu_g^2)^{-1}}{\sqrt{K}}, \frac{\mu_g L_g \beta_k/(\mu_g+L_g)}{2(c+2\rho L_f^2/(\mu_F+\rho))}\right\} \tag{28b}$$

and $\tau_k = \frac{1}{\sqrt{K}}$, then the iterates $\{x^k\}$ and $\{y^k\}$ satisfy

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\widehat{x}(x^k)-x^k\right\|^2\right] = \mathcal{O}\left(\frac{1}{\sqrt{K}}\right) \quad \text{and}$$
$$\mathbb{E}\left[\left\|y^K - y^*(x^K)\right\|^2\right] = \mathcal{O}\left(\frac{1}{\sqrt{K}}\right) \tag{29}$$

where $y^*(x^K)$ is the minimizer of the problem (1b)*, and $c>0$ is a constant that is independent of the stepsizes $\alpha_k, \beta_k, \tau_k$ and the number of iterations K.*

Theorem 2 implies that the convergence rate of STABLE to the stationary point of (1) is $\mathcal{O}(K^{-\frac{1}{2}})$. Since each iteration of STABLE only uses two samples (see Algorithm 1), the sample complexity to achieve an ϵ-stationary point of (1) is $\mathcal{O}(\epsilon^{-2})$, which is on the same order of SGD's sample complexity for the single-level nonconvex problems (Ghadimi and Lan, 2013), and significantly improves the state-of-the-art single-loop TTSA's convergence rate $\mathcal{O}(\epsilon^{-2.5})$ (Hong et al., 2020). In addition, this convergence rate is not directly comparable to other recently developed bilevel optimization methods, e.g., (Ghadimi and Wang, 2018; Ji et al., 2020) since STABLE does not need the increasing batchsize nor double-loop. Regarding the sample complexity, however, STABLE improves over (Ghadimi and Wang, 2018; Ji et al., 2020) by at least the order of $\mathcal{O}(\log(\epsilon^{-1}))$.

We next present the result in the strongly convex case for completeness, where the following additional assumption is needed.

Assumption 5 (strong convexity). *Function $F(x)$ is μ-strongly convex in x, that is, $\nabla_{xx}^2 F(x) \succeq \mu I$.*

Notice that Assumption 5 does not contradict with Assumption 3 since for the constrained upper-level problem (1), only in the constraint set $\mathcal{X}$ do the gradients need to be bounded. Regarding applications, hyperparameter optimization for linear regression or SVM satisfies this assumption.

Theorem 3 (Strongly convex) *Under Assumptions 1–3, 5, if we choose the stepsizes as*

$$\beta_k = \tau_k \le \min\left\{\frac{\mu_g/L_g}{32(\mu_g+L_g)}, \frac{1}{K_0+k}\right\} \tag{30a}$$

$$\alpha_k \le \min\left\{\sqrt{\frac{\mu_g L_g}{4c(\mu_g+L_g)}}, \frac{\mu\mu_g L_g}{2L_f^2(\mu_g+L_g)}, \frac{1}{\sqrt{4c}}, \frac{\mu\mu_g^4}{8C_{g_{xy}}^2 C_{f_y}^2}, \frac{\mu\mu_g^2}{8C_{f_y}^2}\right\}\beta_k \tag{30b}$$

where $K_0 > 0$ is a sufficiently large constant and $c>0$ is an absolute constant that is independent of $\alpha_k, \beta_k, \tau_k$, then the iterates $\{x^k\}$ and $\{y^k\}$ satisfy

$$\mathbb{E}\left[\left\|x^k - x^*\right\|^2\right] = \mathcal{O}\left(\frac{1}{k}\right) \quad \text{and}$$
$$\mathbb{E}\left[\left\|y^k - y^*(x^k)\right\|^2\right] = \mathcal{O}\left(\frac{1}{k}\right) \tag{31}$$

where the solution x^* is defined as $x^* = \arg\min_{x\in\mathcal{X}} F(x)$ and $y^*(x^k)$ is the minimizer of the lower-level problem in (1b)*.*

Theorem 3 implies that to achieve an ϵ-optimal solution for both the lower-level and upper-level problems, the sample complexity of STABLE is $\mathcal{O}(\epsilon^{-1})$. This complexity is on the same order of SGD's complexity for the single-level strongly convex problems (Ghadimi and Lan, 2013), and improves the state-of-the-art single-loop TTSA's sample complexity $\mathcal{O}(\epsilon^{-2})$ for an ϵ-optimal upper-level solution and $\mathcal{O}(\epsilon^{-1.5})$ for an ϵ-optimal lower-level solution (Hong et al., 2020). Compared with double-loop bilevel algorithms in this strong-convex case, STABLE also improves over the BSA's query complexity $\mathcal{O}(\epsilon^{-1})$ in terms of the stochastic upper-level function and $\mathcal{O}(\epsilon^{-2})$ in terms of the stochastic lower-level function (Ghadimi and Wang, 2018).

4 Numerical Tests

This section evaluates the empirical performance of our STABLE. For all compared algorithms, we follow the order of stepsizes suggested in the original papers, and the stepsizes are chosen from $\{1, 0.5, 0.1, 0.05, 0.01\}$, e.g., the best one for each algorithm. In our numerical experiments, we compared our method with several state-of-the-art algorithms such as BSA in (Ghadimi and Wang, 2018), and TTSA in (Hong et al., 2020). We did not include other recent algorithms such as (Ji et al., 2020; Guo and Yang, 2021; Khanduri et al., 2021) which either require increasing the batch size or adding the acceleration of x-update. All the algorithms are implemented using Python 3.6 and run on the same laptop.

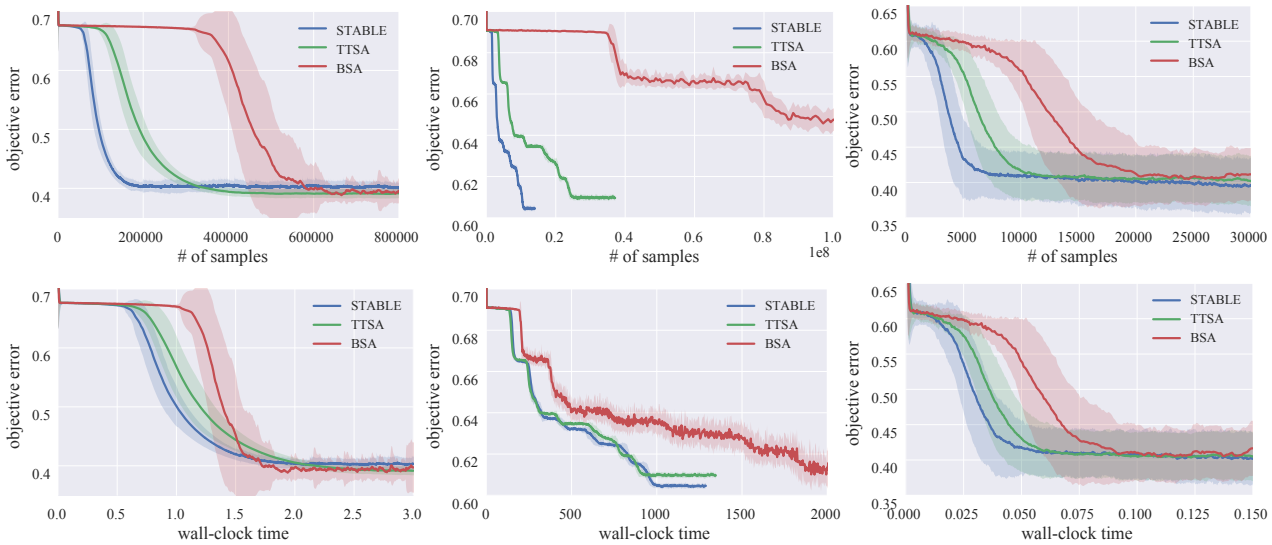

Figure 2: The hyper-parameter optimization task on ijcnn1, covtype and australian datasets. The solid line shows the results averaged over 50 independent trials with random initialization, and the shaded region denotes the standard deviation of results over random trials.

We test all the algorithms in a hyper-parameter optimization task which aims to find the optimal hyper-parameter $x \in \mathbb{R}^d$ (e.g., regularization coefficient), which is used in training a model $y \in \mathbb{R}^d$ on the training set, such that the learned model achieves the low risk on the validation set. Let $\ell(y;\xi)$ denote the logistic loss of the model y on datum ξ, and $\mathcal{D}_{\rm val}$ and $\mathcal{D}_{\rm tra}$ denote, respectively, the training and validation datasets. Specifically, we aim to solve

$$\min_{x\in\mathbb{R}^d} \ \mathbb{E}_{\xi\sim\mathcal{D}_{\rm val}}[\ell(y^*(x);\xi)]$$
$$\text{s.t. } y^*(x) \in \arg\min_{y\in\mathbb{R}^d} \mathbb{E}_{\phi\sim\mathcal{D}_{\rm tra}}[\ell(y;\phi)] + \sum_{i=1}^d x_i y_i^2. \quad (32)$$

In Figure 2, we compare the performance of three algorithms on ijcnn1, covtype and australian datasets (Chang and Lin, 2011) and report their objective errors versus number of samples and the walk-clock time. In all tested datasets, STABLE has sizeable gain in terms of sample complexity compared with the double-loop or two-timescale algorithms since it uses single-loop and single-timescale update. In addition, although TTSA has more efficient y-update, STABLE enjoys the better overall wall-clock time in our simulated setting. This suggests that our STABLE algorithm is preferable in the regime where the sampling is more costly than computation or the dimension d is relatively small, for example in hyperparameter optimization in quantitative trading.

5 Conclusions

This paper develops a new stochastic gradient estimator for bilevel optimization problems. When running SGD on top of this stochastic bilevel gradient, the resultant STABLE algorithm runs in a single loop fashion, and uses a single-timescale update. In both the nonconvex and strongly-convex cases, STABLE matches the sample complexity of SGD for single-level stochastic problems. One possible extension is to apply SGD-improvement techniques to accelerate STABLE, which helps STABLE achieve state-of-the-art performance for bilevel problems. Another natural extension is to apply our bilevel optimization method to the general two-timescale stochastic approximation case, in a similar fashion to (Dalal et al., 2018; Kaledin et al., 2020). Improving the sample complexity of such general case can be of great interest to the reinforcement learning community.

Acknowledgment

The work of T. Chen and Q. Xiao was partially supported by and the Rensselaer-IBM AI Research Collaboration (`http://airc.rpi.edu`), part of the IBM AI Horizons Network (`http://ibm.biz/AIHorizons`) and NSF 2134168.

References

Zalán Borsos, Mojmír Mutnỳ, and Andreas Krause. Coresets via bilevel optimization for continual learning and streaming. In *Proc. Advances in Neural Info. Process. Syst.*, Virtual, December 2020.

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

Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A

library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2: 27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Tianyi Chen, Yuejiao Sun, and Wotao Yin. Solving stochastic compositional optimization is nearly as easy as solving stochastic optimization. *arXiv preprint:2008.10847*, August 2020.

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

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

Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in non-convex sgd. *Proc. Advances in Neural Info. Process. Syst.*, 32, December 2019.

Gal Dalal, Gugan Thoppe, Balázs Szörényi, and Shie Mannor. Finite sample analysis of two-timescale stochastic approximation with applications to reinforcement learning. In *Conference On Learning Theory*, pages 1199–1233, Graz, Austria, July 2018.

Constantinos Daskalakis and Ioannis Panageas. The limit points of (optimistic) gradient descent in min-max optimization. In *Proc. Advances in Neural Info. Process. Syst.*, pages 9256–9266, Montreal, Canada, December 2018.

Damek Davis and Dmitriy Drusvyatskiy. Stochastic subgradient method converges at the rate $\mathcal{O}(k^{-1/4})$ on weakly convex functions. *arXiv preprint arXiv:1802.02988*, 2018.

Stephan Dempe and Alain Zemkoho. *Bilevel Optimization*. Springer, 2020.

Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazzi, and Massimiliano Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *Proc. Intl. Conf. Machine Learn.*, pages 1568–1577, Vienna, Austria, June 2018.

Saeed Ghadimi and Guanghui Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23 (4):2341–2368, 2013.

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

Saeed Ghadimi, Andrzej Ruszczynski, and Mengdi Wang. A single timescale stochastic approximation method for nested stochastic optimization. *SIAM Journal on Optimization*, 30(1):960–979, March 2020.

Riccardo Grazzi, Luca Franceschi, Massimiliano Pontil, and Saverio Salzo. On the iteration complexity of hypergradient computation. In *Proc. Intl. Conf. Machine Learn.*, pages 3748–3758, virtual, July 2020.

Zhishuai Guo and Tianbao Yang. Randomized stochastic variance-reduced methods for stochastic bilevel optimization. *arXiv preprint arXiv:2105.02266*, May 2021.

Mingyi Hong, Hoi-To Wai, Zhaoran Wang, and Zhuoran Yang. A two-timescale framework for bilevel optimization: Complexity analysis and application to actor-critic. *arXiv preprint:2007.05170*, 2020.

Yifan Hu, Siqi Zhang, Xin Chen, and Niao He. Biased stochastic gradient descent for conditional stochastic optimization. *arXiv preprint:2002.10790*, February 2020.

Feihu Huang and Heng Huang. Biadam: Fast adaptive bilevel optimization methods. *arXiv preprint arXiv:2106.11396*, 2021.

Kaiyi Ji, Junjie Yang, and Yingbin Liang. Provably faster algorithms for bilevel optimization and applications to meta-learning. *arXiv preprint:2010.07962*, 2020.

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, Virtual, July 2020.

Prashant Khanduri, Siliang Zeng, Mingyi Hong, Hoi-To Wai, Zhaoran Wang, and Zhuoran Yang. A momentum-assisted single-timescale stochastic approximation algorithm for bilevel optimization. *arXiv preprintarXiv:2102.07367*, February 2021.

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

Gautam Kunapuli, Kristin P Bennett, Jing Hu, and Jong-Shi Pang. Classification model selection via bilevel programming. *Optimization Methods & Software*, 23(4):475–489, 2008.

Karl Kunisch and Thomas Pock. A bilevel optimization approach for parameter learning in variational models. *SIAM Journal on Imaging Sciences*, 6(2): 938–983, 2013.

Xiangru Lian, Mengdi Wang, and Ji Liu. Finite-sum composition optimization via variance reduced gradient descent. In *Proc. Intl. Conf. on Artif. Intell. and Stat.*, Fort Lauderdale, FL, April 2017.

Tianyi Lin, Chi Jin, and Michael Jordan. On gradient descent ascent for nonconvex-concave minimax prob-

lems. In *Proc. Intl. Conf. Machine Learn.*, pages 6083–6093, Virtual, July 2020.

Risheng Liu, Pan Mu, Xiaoming Yuan, Shangzhi Zeng, and Jin Zhang. A generic first-order algorithmic framework for bi-level programming beyond lower-level singleton. In *Proc. Intl. Conf. Machine Learn.*, pages 6305–6315, Virtual, July 2020.

Risheng Liu, Jiaxin Gao, Jin Zhang, Deyu Meng, and Zhouchen Lin. Investigating bi-level optimization for learning and vision from a unified perspective: A survey and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

Luo Luo, Haishan Ye, Zhichao Huang, and Tong Zhang. Stochastic recursive gradient descent ascent for stochastic nonconvex-strongly-concave minimax problems. In *Proc. Advances in Neural Info. Process. Syst.*, Virtual, December 2020.

Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. In *Proc. Intl. Conf. on Artif. Intell. and Stat.*, pages 1497–1507, Palermo, Italy, August 2020.

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A basic course*, volume 87. Springer, Berlin, Germany, 2013.

Jorge Nocedal and Stephen Wright. *Numerical Optimization*. Springer, Berlin, Germany, 2006.

Maher Nouiehed, Maziar Sanjabi, Tianjian Huang, Jason D Lee, and Meisam Razaviyayn. Solving a class of non-convex min-max games using iterative first order methods. In *Proc. Advances in Neural Info. Process. Syst.*, pages 14934–14942, Vancouver, Canada, December 2019.

Santiago Paternain, Aryan Mokhtari, and Alejandro Ribeiro. A Newton-based method for nonconvex optimization with fast evasion of saddle points. *SIAM Journal on Optimization*, 29(1):343–368, January 2019.

Hassan Rafique, Mingrui Liu, Qihang Lin, and Tianbao Yang. Non-convex min-max optimization: Provable algorithms and applications in machine learning. *Optimization Methods and Software*, March 2021.

Aravind Rajeswaran, Chelsea Finn, Sham M Kakade, and Sergey Levine. Meta-learning with implicit gradients. In *Proc. Advances in Neural Info. Process. Syst.*, pages 113–124, Vancouver, Canada, December 2019.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22(3):400–407, September 1951.

Shoham Sabach and Shimrit Shtern. A first order method for solving convex bilevel optimization problems. *SIAM Journal on Optimization*, 27(2):640–660, 2017.

Amirreza Shaban, Ching-An Cheng, Nathan Hatch, and Byron Boots. Truncated back-propagation for bilevel optimization. In *Proc. Intl. Conf. on Artif. Intell. and Stat.*, pages 1723–1732, Naha, Okinawa, Japan, April 2019.

Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński. *Lectures on Stochastic Programming: Modeling and Theory*. SIAM, Philadelphia, PA, 2009.

Andrea Simonetto, Aryan Mokhtari, Alec Koppel, Geert Leus, and Alejandro Ribeiro. A class of prediction-correction methods for time-varying convex optimization. *IEEE Transactions on Signal Processing*, 64(17):4576–4591, May 2016.

Heinrich Von Stackelberg. *The Theory of Market Economy*. Oxford University Press, 1952.

Quoc Tran-Dinh, Nhan Pham, and Lam Nguyen. Stochastic Gauss-Newton algorithms for nonconvex compositional optimization. In *Proc. Intl. Conf. Machine Learn.*, pages 9572–9582, Virtual, July 2020.

Quoc Tran-Dinh, Nhan H Pham, Dzung T Phan, and Lam M Nguyen. A hybrid stochastic optimization framework for composite nonconvex optimization. *Mathematical Programming*, pages 1–67, 2021.

Luis N Vicente and Paul H Calamai. Bilevel and multilevel programming: A bibliography review. *Journal of Global optimization*, 5(3):291–306, 1994.

Mengdi Wang, Ethan X Fang, and Han Liu. Stochastic compositional gradient descent: algorithms for minimizing compositions of expected-value functions. *Mathematical Programming*, 161(1-2):419–449, January 2017a.

Mengdi Wang, Ji Liu, and Ethan Fang. Accelerating stochastic composition optimization. *Journal Machine Learning Research*, 18(1):3721–3743, 2017b.

Junjie Yang, Kaiyi Ji, and Yingbin Liang. Provably faster algorithms for bilevel optimization. *arXiv preprint arXiv:2106.04692*, June 2021.

JJ Ye and DL Zhu. Optimality conditions for bilevel programming problems. *Optimization*, 33(1):9–27, 1995.

Junyu Zhang and Lin Xiao. A stochastic composite gradient method with incremental variance reduction. In *Proc. Advances in Neural Info. Process. Syst.*, pages 9075–9085, Vancouver, Canada, December 2019.

Supplementary Material for "A Single-Timescale Method for Stochastic Bilevel Optimization"

A Proof sketch

In this section, we highlight the key steps of the proof towards Theorem 2. The proof for the strongly convex case in Theorem 3 will follow similar steps.

For simplicity of the convergence analysis, we define the following Lyapunov function

$$\mathbb{V}^k := \Phi_{1/\rho}(x^k) + \|y^k - y^*(x^k)\|^2 + \|H_{yy}^k - \nabla_{yy}^2 g(x^k, y^k)\|^2 + \|H_{xy}^k - \nabla_{xy}^2 g(x^k, y^k)\|^2 \tag{33}$$

which mimics the continuous-time Lyapunov function (14) for the deterministic problem.

Similar to the ODE analysis, we first quantify the difference between two Lyapunov functions as

$$\begin{aligned}\mathbb{V}^{k+1} - \mathbb{V}^k =& \underbrace{\Phi_{1/\rho}(x^{k+1}) - \Phi_{1/\rho}(x^k)}_{\text{Lemma 2}} + \underbrace{\|y^{k+1} - y^*(x^{k+1})\|^2 - \|y^k - y^*(x^k)\|^2}_{\text{Lemma 3}} \\ &+ \underbrace{\|H_{yy}^{k+1} - \nabla_{yy}^2 g(x^{k+1}, y^{k+1})\|^2 - \|H_{yy}^k - \nabla_{yy}^2 g(x^k, y^k)\|^2}_{\text{Lemma 4}} \\ &+ \underbrace{\|H_{xy}^{k+1} - \nabla_{xy}^2 g(x^{k+1}, y^{k+1})\|^2 - \|H_{xy}^k - \nabla_{xy}^2 g(x^k, y^k)\|^2}_{\text{Lemma 4}}.\end{aligned} \tag{34}$$

The difference in (34) consists of four difference terms: the first term quantifies the descent of the Moreau Envelope of the upper-level objective functions; the second term characterizes the descent of the lower-level optimization errors; and, the third and fourth terms measure the estimation error of the second-order quantities. Since Lemma 2 stated in the main body bounded the Moreau Envelop of the upper level, we will bound the rest, respectively, in the ensuing lemmas.

We will analyze the error of the lower-level variable, which is the key step to improving the existing results.

Lemma 3 (Error of lower level) *Suppose that Assumptions 1–3 hold, and y^{k+1} is generated by running iteration (8) given x^k. If we choose $\beta_k \le \frac{2}{\mu_g + L_g}$, then y^{k+1} satisfies*

$$\begin{aligned}\mathbb{E}\left[\|y^*(x^{k+1}) - y^{k+1}\|^2 | \mathcal{F}^k\right] \le& \left(1 - \frac{\mu_g L_g \beta^k}{\mu_g + L_g} + \frac{c\alpha_k^2}{\beta_k}\right)\|y^k - y^*(x^k)\|^2 + \left(1 + \frac{\mu_g L_g \beta^k}{\mu_g + L_g}\right)\beta_k^2 \sigma_{g_y}^2 \\ &+ \frac{c\alpha_k^4}{\beta_k} + \mathbb{E}\left[\|H_{yy}^k - \nabla_{yy}^2 g(x^k, y^k)\|^2 | \mathcal{F}^k\right]\frac{c\alpha_k^2}{\beta_k} + \mathbb{E}\left[\|H_{xy}^k - \nabla_{xy}^2 g(x^k, y^k)\|^2 | \mathcal{F}^k\right]\frac{c\alpha_k^2}{\beta_k}.\end{aligned} \tag{35}$$

Roughly speaking, Lemma 3 implies that if the stepsizes α_k^2 and β_k^2 and the estimation errors of H_{yy}^k and H_{xy}^k are decreasing fast enough, the error of y^{k+1} will also decrease.

Since the RHS of both Lemmas 2 and 3 critically depend on the quality of H_{yy}^k and H_{xy}^k, we will next build upon the results in (Chen et al., 2020, Lemma 2) to analyze the estimation errors.

Lemma 4 (Estimation errors of H_{xy}^k and H_{yy}^k) *Suppose Assumptions 1–3 hold, and H_{xy}^k and H_{yy}^k are generated by running (9). The mean square error of H_{xy}^k satisfies*

$$\begin{aligned}\mathbb{E}\Big[\|H_{xy}^k - \nabla_{xy}^2 g(x^k, y^k)\|^2 \mid \mathcal{F}^k\Big] \le& (1 - \tau_k)^2 \|H_{xy}^{k-1} - \nabla_{xy}^2 g(x^{k-1}, y^{k-1})\|^2 + 2\tau_k^2 \sigma_{g_{xy}}^2 \\ &+ 2(1 - \tau_k)^2 (\bar{L}_{g_{xy}}^2 + L_{g_{xy}}^2)\|x^k - x^{k-1}\|^2 + 2(1 - \tau_k)^2 (\bar{L}_{g_{xy}}^2 + L_{g_{xy}}^2)\|y^k - y^{k-1}\|^2\end{aligned} \tag{36}$$

where the constants $L_{g_{xy}}, L_{g_{yy}}, \bar{L}_{g_{xy}}, \bar{L}_{g_{yy}}, \sigma_{g_{xy}}, \sigma_{g_{yy}}$ are defined in Assumptions 1 and 3. And likewise, the mean square error of H_{yy}^k satisfies

$$\mathbb{E}\Big[\|H_{yy}^k - \nabla_{yy}^2 g(x^k, y^k)\|^2 \mid \mathcal{F}^k\Big] \le (1-\tau_k)^2 \|H_{yy}^{k-1} - \nabla_{yy}^2 g(x^{k-1}, y^{k-1})\|^2 + 2\tau_k^2 \sigma_{g_{yy}}^2 + 2(1-\tau_k)^2(\bar{L}_{g_{yy}}^2 + L_{g_{yy}}^2)\|x^k - x^{k-1}\|^2 + 2(1-\tau_k)^2(\bar{L}_{g_{yy}}^2 + L_{g_{yy}}^2)\|y^k - y^{k-1}\|^2. \tag{37}$$

Intuitively, the update of x^k is bounded and so is the update of y^k, and thus $\|x^k - x^{k-1}\|^2 = \mathcal{O}(\alpha_{k-1}^2)$ and $\|y^k - y^{k-1}\|^2 = \mathcal{O}(\beta_{k-1}^2)$. Plugging them into the RHS of Lemma 4, it suggests that if the stepsizes $\alpha_k^2, \beta_k^2, \tau_k^2$ are decreasing, then the estimation errors of H_{xy}^k and H_{yy}^k also decrease.

Applying Lemmas 2–4 to (34) and rearranging terms, we will be able to get

$$\mathbb{E}[\mathbb{V}^{k+1}] - \mathbb{E}[\mathbb{V}^k] \le -c_1 \mathbb{E}[\|y^k - y^*(x^k)\|^2] - c_2 \mathbb{E}[\|\widehat{x}(x^k) - x^k\|^2] + c_3 \tag{38}$$

where the constants are $c_1 = \mathcal{O}(\beta_k)$, $c_2 = \mathcal{O}(\alpha_k)$ and $c_3 = \mathcal{O}(\alpha_k^2 + \beta_k^2 + \tau_k^2)$. By choosing stepsizes $\alpha_k, \beta_k, \tau_k$ as (28) and telescoping both sides of (38), we obtain the main results in Theorem 2.

B Auxiliary Lemmas

In this section, we present some auxiliary lemmas that will be used frequently in the proof.

Lemma 5 ((Ghadimi and Wang, 2018, Lemma 2.2)) *Under Assumptions 1 and 2, we have*

$$\|\overline{\nabla}_x f(x, y^*(x)) - \overline{\nabla}_x f(x, y)\| \le L_f \|y^*(x) - y\| \tag{39a}$$

$$\|\nabla F(x_1) - \nabla F(x_2)\| \le L_F \|x_1 - x_2\| \tag{39b}$$

$$\|y^*(x_1) - y^*(x_2)\| \le L_y \|x_1 - x_2\| \tag{39c}$$

and the constants L_f, L_y, L_F are defined as

$$L_f := L_{f_x} + \frac{C_{g_{xy}} L_{f_y}}{\mu_g} + \frac{C_{f_y}}{\mu_g}\left(L_{fxy} + \frac{C_{g_{xy}} L_{g_{yy}}}{\mu_g}\right), \quad L_y := \frac{C_{g_{xy}}}{\mu_g}$$

$$L_F := \bar{L}_{f_x} + \frac{C_{g_{xy}}(\bar{L}_{f_y} + L_f)}{\mu_g} + \frac{C_{f_y}}{\mu_g}\left(\bar{L}_{fxy} + \frac{C_{g_{xy}} \bar{L}_{g_{yy}}}{\mu_g}\right)$$

where the constants are defined in Assumptions 1–3.

C Proof of Proposition 1

Proof: Define the Jacobian matrix

$$\nabla_x y(x) = \begin{bmatrix} \frac{\partial}{\partial x_1} y_1(x) & \cdots & \frac{\partial}{\partial x_d} y_1(x) \\ & \cdots & \\ \frac{\partial}{\partial x_1} y_{d_y}(x) & \cdots & \frac{\partial}{\partial x_d} y_{d_y}(x) \end{bmatrix}.$$

By the chain rule, it follows that

$$\nabla F(x) := \nabla_x f\left(x, y^*(x)\right) + \nabla_x y^*(x)^\top \nabla_y f\left(x, y^*(x)\right). \tag{40}$$

The minimizer $y^*(x)$ satisfies

$$\nabla_y g(x, y^*(x)) = 0, \quad \text{thus} \quad \nabla_x \left(\nabla_y g(x, y^*(x))\right) = 0. \tag{41}$$

By the chain rule again, it follows that

$$\nabla_{xy}^2 g\left(x, y^*(x)\right) + \nabla_x y^*(x)^\top \nabla_{yy}^2 g\left(x, y^*(x)\right) = 0.$$

By Assumption 2, $\nabla_{yy}^2 g\left(x, y^*(x)\right)$ is invertible, so

$$\nabla_x y^*(x)^\top := -\nabla_{xy}^2 g\left(x, y^*(x)\right) \left[\nabla_{yy}^2 g\left(x, y^*(x)\right)\right]^{-1}. \tag{42}$$

By substituting (42) into (40), we arrive at (4).

D Proof of Lemma 2

Proof: Now we turn to analyze the update of x. For convenience, we define the update in (8a) as

$$x^{k+1} = \mathcal{P}_{\mathcal{X}}\left(x^k - \alpha_k \bar{h}_f^k\right) \qquad \text{with} \qquad \bar{h}_f^k := \nabla_x f\left(x^k, y^k; \xi^k\right) - H_{xy}^k (H_{yy}^k)^{-1} \nabla_y f\left(x^k, y^k; \xi^k\right) \tag{43}$$

and let $\widehat{x}^k$ and $\widehat{x}$ denote $\widehat{x}(x^k)$ and $\widehat{x}(x)$.

For $\forall x \in \mathcal{X}$, using the weakly convexity of F, we know that

$$F(\hat{x}) \geq F(x) + \langle \nabla F(x), \hat{x} - x\rangle + \frac{\mu_F}{2}\|\hat{x} - x\|^2$$

On the other hand, by the definition of $\hat{x}$, $\forall x \in \mathcal{X}$, it holds that

$$F(x) + \frac{\rho}{2}\|x - x\|^2 - F(\hat{x}) - \frac{\rho}{2}\|\hat{x} - x\|^2 = F(x) - F(\hat{x}) - \frac{\rho}{2}\|\hat{x} - x\|^2 \geq 0.$$

Adding above two inequalities, we get that

$$\langle \nabla F(x), \hat{x} - x\rangle \leq -\frac{\mu_F + \rho}{2}\|\hat{x} - x\|^2.$$

If we choose μ_F such that $\mu_F + \rho > 0$ and using the definition of Moreau Envelop, we have that

$$\begin{aligned}
\Phi_{1/\rho}(x^{k+1}) &= F(\hat{x}^{k+1}) + \frac{\rho}{2}\|x^{k+1} - \hat{x}^{k+1}\|^2 \leq F(\hat{x}^k) + \frac{\rho}{2}\|x^{k+1} - \hat{x}^k\|^2 \\
&= F(\hat{x}^k) + \frac{\rho}{2}\|x^k - \hat{x}^k\|^2 + \frac{\rho}{2}\|x^{k+1} - x^k\|^2 + \rho\langle x^k - \hat{x}^k, x^{k+1} - x^k\rangle \\
&= \Phi_{1/\rho}(x^k) + \frac{\rho}{2}\|x^{k+1} - x^k\|^2 + \rho\langle x^{k+1} - \hat{x}^k, x^{k+1} - x^k\rangle - \rho\|x^k - x^{k+1}\|^2 \\
&\leq \Phi_{1/\rho}(x^k) + \rho\alpha_k\langle \hat{x}^k - x^k, h_f^k\rangle + \rho\alpha_k\langle h_f^k, x^k - x^{k+1}\rangle
\end{aligned} \tag{44}$$

where the fourth inequality is due to $\langle x^k - \alpha_k h_f^k - x^{k+1}, \hat{x}^k - x^{k+1}\rangle \leq 0$ using the definition of $\mathcal{P}_{\mathcal{X}}$. Then taking the conditional expectation of both sides in (44), we have that

$$\begin{aligned}
\mathbb{E}\left[\Phi_{1/\rho}(x^{k+1})|\mathcal{F}^k\right] &\leq \Phi_{1/\rho}(x^k) + \rho\alpha_k\mathbb{E}\left[\left\langle \widehat{x}^k - x^k, \bar{h}_f^k\right\rangle|\mathcal{F}_k\right] + \alpha_k^2\rho\mathbb{E}[\left\|\bar{h}_f^k\right\|^2|\mathcal{F}^k] \\
&\leq \Phi_{1/\rho}(x^k) + 2\rho\alpha_k^2\left(C_{f_x}^2 + \left(\frac{C_{g_{xy}}}{\mu_g}\right)^2 C_{f_y}^2\right) + \rho\alpha_k\mathbb{E}\left[\left\langle \widehat{x}^k - x^k, \bar{h}_f^k\right\rangle|\mathcal{F}_k\right]
\end{aligned} \tag{45}$$

where the second inequality comes from (61). Then we bound the third term in (45) and get that

$$\begin{aligned}
\mathbb{E}\left[\left\langle \widehat{x}^k - x^k, \bar{h}_f^k\right\rangle|\mathcal{F}_k\right] &\leq \mathbb{E}\left[\left\langle \widehat{x}^k - x^k, \bar{h}_f^k - \bar{\nabla}_x f(x^k, y^k) + \bar{\nabla}_x f(x^k, y^k) - \nabla F(x^k) + \nabla F(x^k)\right\rangle|\mathcal{F}_k\right] \\
&\leq \left\langle \widehat{x}^k - x^k, \mathbb{E}\left[\bar{h}_f^k|\mathcal{F}^k\right] - \bar{\nabla}_x f(x^k, y^k)\right\rangle + \mathbb{E}\left[\left\langle \widehat{x}^k - x^k, \bar{\nabla}_x f(x^k, y^k) - \nabla F(x^k)\right\rangle|\mathcal{F}_k\right] \\
&\quad + \mathbb{E}\left[\left\langle \widehat{x}^k - x^k, \nabla F(x^k)\right\rangle|\mathcal{F}_k\right] \\
&\leq \frac{\gamma_k}{4}\|\widehat{x}^k - x^k\|^2 + \frac{\|\nabla_y f(x^k, y^k)\|^2}{\gamma_k}\mathbb{E}\left[\left\|(H_{yy}^k)^{-1}H_{xy}^k - H_{yy}(x^k, y^k)^{-1}H_{xy}(x^k, y^k)\right\|^2|\mathcal{F}^k\right] \\
&\quad + \frac{\gamma_k}{4}\|\widehat{x}^k - x^k\|^2 + \frac{1}{\gamma_k}\|\bar{\nabla}_x f(x^k, y^k) - \nabla F(x^k)\|^2 - \frac{\mu_F + \rho}{2}\|\widehat{x}^k - x^k\|^2 \\
&\leq \frac{\gamma_k}{2}\|\widehat{x}^k - x^k\|^2 + \frac{2C_{f_y}^2}{\gamma_k\mu_g^2}\left[\frac{C_{g_{xy}}^2}{\mu_g^2}\mathbb{E}[\|H_{yy}^k - H_{yy}(x^k, y^k)\|^2|\mathcal{F}^k] + \mathbb{E}[\|H_{xy}^k - H_{xy}(x^k, y^k)\|^2|\mathcal{F}^k]\right] \\
&\quad + \frac{L_f^2}{\gamma_k}\|y^k - y^*(x^k)\|^2 - \frac{\mu_F + \rho}{2}\|\widehat{x}^k - x^k\|^2
\end{aligned}$$

where the third inequality uses Young's inequality with parameter γ_k, (61) in (Hong et al., 2020) and the fact that

$$\mathbb{E}_{\xi^k}[\bar{h}_f^k|\mathcal{F}^k] = \nabla_x f\left(x^k, y^k\right) - (H_{yy}^k)^{-1}H_{xy}^k\nabla_y f\left(x^k, y^k\right); \tag{46}$$

and the last inequality follows the same steps of (57) and Assumption 3. We choose $\gamma_k = \frac{\mu_F+\rho}{2}$, then we get

$$\mathbb{E}\left[\langle \widehat{x}^k - x^k, \bar{h}_f^k\rangle |\mathcal{F}_k\right] \le -\frac{\mu_F+\rho}{4}\|\widehat{x}^k - x^k\|^2 + \frac{4C_{f_y}^2 C_{g_{xy}}^2}{(\mu_F+\rho)\mu_g^4}\mathbb{E}[\|H_{yy}^k - H_{yy}(x^k,y^k)\|^2|\mathcal{F}^k]$$
$$+ \frac{4C_{f_y}^2}{(\mu_F+\rho)\mu_g^2}\mathbb{E}[\|H_{xy}^k - H_{xy}(x^k,y^k)\|^2|\mathcal{F}^k] + \frac{2L_f^2}{\mu_F+\rho}\|y^k - y^*(x^k)\|^2 \tag{47}$$

Plugging (47) into (45) and taking expectation over all the randomness lead to the lemma.

E Proof of Lemma 3

Proof: We start by decomposing the error of the lower level variable as

$$\begin{aligned}
&\mathbb{E}\left[\|y^{k+1} - y^*(x^{k+1})\|^2|\mathcal{F}^k\right]\\
&= \mathbb{E}\left[\|y^k - \beta_k h_g^k - y^*(x^k) + y^*(x^k) - y^*(x^{k+1}) - (H_{yy}^k)^{-1}(H_{xy}^k)^\top(x^{k+1}-x^k)\|^2|\mathcal{F}^k\right]\\
&\le (1+\varepsilon)\underbrace{\mathbb{E}[\|y^k - \beta_k h_g^k - y^*(x^k)\|^2|\mathcal{F}^k]}_{I_1}\\
&\quad + (1+\varepsilon^{-1})\underbrace{\mathbb{E}[\|y^*(x^k) - y^*(x^{k+1}) - (H_{yy}^k)^{-1}(H_{xy}^k)^\top(x^{k+1}-x^k)\|^2|\mathcal{F}^k]}_{I_2}.
\end{aligned} \tag{48}$$

The upper bound of I_1 can be derived as

$$\begin{aligned}
I_1 &= \|y^k - y^*(x^k)\|^2 - 2\beta_k\mathbb{E}[\langle y^k - y^*(x^k), h_g^k\rangle|\mathcal{F}^k] + \beta_k^2\mathbb{E}[\|h_g^k\|^2|\mathcal{F}^k]\\
&\overset{(a)}{\le} \|y^k - y^*(x^k)\|^2 - 2\beta_k\langle y^k - y^*(x^k), \nabla_y g(x^k,y^k)\rangle + \beta_k^2\|\nabla_y g(x^k,y^k)\|^2 + \beta_k^2\sigma_{g_y}^2\\
&\overset{(b)}{\le} \left(1 - \frac{2\mu_g L_g}{\mu_g+L_g}\beta^k\right)\|y^k - y^*(x^k)\|^2 + \beta_k\left(\beta_k - \frac{2}{\mu_g+L_g}\right)\|\nabla_y g(x^k,y^k)\|^2 + \beta_k^2\sigma_{g_y}^2\\
&\overset{(c)}{\le} \left(1 - \frac{2\mu_g L_g}{\mu_g+L_g}\beta^k\right)\|y^k - y^*(x^k)\|^2 + \beta_k^2\sigma_{g_y}^2
\end{aligned} \tag{49}$$

where (a) comes from the fact that $\mathrm{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$, (b) follows from the μ_g-strong convexity and L_g smoothness of $g(x,y)$ (Nesterov, 2013, Theorem 2.1.11), and (c) follows from the choice of stepsize $\beta_k \le \frac{\mu_g/L_g}{32(\mu_g+L_g)} \le \frac{2}{\mu_g+L_g}$ in (28a).

The upper bound of I_2 can be derived as

$$\begin{aligned}
I_2 &= \mathbb{E}\left[\left\|y^*(x^k) - y^*(x^{k+1}) - (H_{yy}^k)^{-1}(H_{xy}^k)^\top(x^{k+1}-x^k)\right\|^2|\mathcal{F}^k\right]\\
&\le 3\mathbb{E}\left[\left\|y^*(x^{k+1}) - y^*(x^k) - \nabla_x y^*(x^k)(x^{k+1}-x^k)\right\|^2|\mathcal{F}^k\right]\\
&\quad + 3\mathbb{E}\left[\left\|\left(\nabla_x y^*(x^k) - H_{yy}(x^k,y^k)^{-1}H_{xy}(x^k,y^k)^\top\right)(x^{k+1}-x^k)\right\|^2|\mathcal{F}^k\right]\\
&\quad + 3\mathbb{E}\left[\left\|\left(H_{yy}(x^k,y^k)^{-1}H_{xy}(x^k,y^k)^\top - (H_{yy}^k)^{-1}(H_{xy}^k)^\top\right)(x^{k+1}-x^k)\right\|^2|\mathcal{F}^k\right].
\end{aligned} \tag{50}$$

We first bound the first approximation error in the RHS of (50) by

$$\begin{aligned}
&\left\|y^*(x^{k+1}) - y^*(x^k) - \nabla_x y^*(x^k)(x^{k+1}-x^k)\right\|^2\\
&= \left\|\int_0^1 \nabla_x y^*(x^k + t(x^{k+1}-x^k))(x^{k+1}-x^k)dt - \nabla_x y^*(x^k)(x^{k+1}-x^k)\right\|^2\\
&\le \int_0^1 \left\|\nabla_x y^*(x^k + t(x^{k+1}-x^k)) - \nabla_x y^*(x^k)\right\|^2\|x^{k+1}-x^k\|^2 dt \le \frac{L_y^2}{2}\|x^{k+1}-x^k\|^4
\end{aligned} \tag{51}$$

where the first inequality follows from the Cauchy-Schwarz inequality, and the second inequality follows from the L_y-Lipschitz continuity of $\nabla_x y^*(x)$ in Lemma 5.

Next we bound the second term in the RHS of (50) as

$$\begin{aligned}&\mathbb{E}\left[\left\|\left(\nabla_x y^*(x^k) - H_{yy}(x^k,y^k)^{-1}H_{xy}(x^k,y^k)^\top\right)(x^{k+1}-x^k)\right\|^2|\mathcal{F}^k\right]\\ \le&\mathbb{E}\left[\left\|\nabla_x y^*(x^k) - H_{yy}(x^k,y^k)^{-1}H_{xy}(x^k,y^k)^\top\right\|^2\left\|x^{k+1}-x^k\right\|^2|\mathcal{F}^k\right]\end{aligned}\tag{52}$$

and likewise, the third term of (50) as

$$\begin{aligned}&\mathbb{E}\left[\left\|\left(H_{yy}(x^k,y^k)^{-1}H_{xy}(x^k,y^k)^\top - (H_{yy}^k)^{-1}(H_{xy}^k)^\top\right)(x^{k+1}-x^k)\right\|^2|\mathcal{F}^k\right]\\ \le&\mathbb{E}\left[\left\|H_{yy}(x^k,y^k)^{-1}H_{xy}(x^k,y^k)^\top - (H_{yy}^k)^{-1}(H_{xy}^k)^\top\right\|^2\left\|x^{k+1}-x^k\right\|^2|\mathcal{F}^k\right].\end{aligned}\tag{53}$$

We then bound the approximation error of $H_{yy}(x^k,y^k)^{-1}H_{xy}(x^k,y^k)^\top$ in (52) by

$$\begin{aligned}&\left\|\nabla_x y^*(x^k) - H_{yy}(x^k,y^k)^{-1}H_{xy}(x^k,y^k)^\top\right\|^2\\ =&\left\|H_{yy}\left(x^k,y^*(x^k)\right)^{-1}H_{xy}\left(x^k,y^*(x^k)\right)^\top - H_{yy}(x^k,y^k)^{-1}H_{xy}(x^k,y^k)^\top\right\|^2\\ =&\Big\|H_{yy}\left(x^k,y^*(x^k)\right)^{-1}H_{xy}\left(x^k,y^*(x^k)\right)^\top - H_{yy}(x^k,y^k)^{-1}H_{xy}\left(x^k,y^*(x^k)\right)^\top\\ &\quad + H_{yy}(x^k,y^k)^{-1}H_{xy}\left(x^k,y^*(x^k)\right)^\top - H_{yy}(x^k,y^k)^{-1}H_{xy}(x^k,y^k)^\top\Big\|^2\\ \le&2C_{g_{xy}}^2\left\|H_{yy}\left(x^k,y^*(x^k)\right)^{-1} - H_{yy}(x^k,y^k)^{-1}\right\|^2 + \frac{2}{\mu_g^2}\left\|H_{xy}\left(x^k,y^*(x^k)\right) - H_{xy}(x^k,y^k)\right\|^2\end{aligned}\tag{54}$$

where the inequality follows from $\|H_{xy}(x,y)\| \le C_{g_{xy}}$ and $H_{yy}(x,y) \succeq \mu_g I$.

Note that

$$\begin{aligned}&\left\|H_{yy}\left(x^k,y^*(x^k)\right)^{-1} - H_{yy}(x^k,y^k)^{-1}\right\|^2\\ =&\left\|H_{yy}\left(x^k,y^*(x^k)\right)^{-1}\left(H_{yy}\left(x^k,y^*(x^k)\right) - H_{yy}(x^k,y^k)\right)H_{yy}(x^k,y^k)^{-1}\right\|^2\\ \le&\left\|H_{yy}\left(x^k,y^*(x^k)\right)^{-1}\right\|^2\left\|H_{yy}\left(x^k,y^*(x^k)\right) - H_{yy}(x^k,y^k)\right\|^2\left\|H_{yy}(x^k,y^k)^{-1}\right\|^2\\ \le&\frac{1}{\mu_g^4}\left\|H_{yy}\left(x^k,y^*(x^k)\right) - H_{yy}(x^k,y^k)\right\|^2\end{aligned}\tag{55}$$

where the last inequality follows from $H_{yy}(x,y) \succeq \mu_g I$.

Therefore, we have

$$\begin{aligned}&\left\|\nabla_x y^*(x^k) - H_{yy}(x^k,y^k)^{-1}H_{xy}(x^k,y^k)^\top\right\|^2\\ \le&\frac{2C_{g_{xy}}^2}{\mu_g^4}\left\|H_{yy}\left(x^k,y^*(x^k)\right) - H_{yy}(x^k,y^k)\right\|^2 + \frac{2}{\mu_g^2}\left\|H_{xy}\left(x^k,y^*(x^k)\right) - H_{xy}(x^k,y^k)\right\|^2.\end{aligned}\tag{56}$$

Following the steps towards (56), we bound the error of $(H_{yy}^k)^{-1}(H_{xy}^k)^\top$ in (53) by

$$\begin{aligned}&\left\|(H_{yy}^k)^{-1}(H_{xy}^k)^\top - H_{yy}(x^k,y^k)^{-1}H_{xy}(x^k,y^k)^\top\right\|^2\\ =&\left\|(H_{yy}^k)^{-1}(H_{xy}^k)^\top - H_{yy}(x^k,y^k)^{-1}(H_{xy}^k)^\top + H_{yy}(x^k,y^k)^{-1}(H_{xy}^k)^\top - H_{yy}(x^k,y^k)^{-1}H_{xy}(x^k,y^k)^\top\right\|^2\\ \le&2\left\|(H_{yy}^k)^{-1}(H_{xy}^k)^\top - H_{yy}(x^k,y^k)^{-1}(H_{xy}^k)^\top\right\|^2 + 2\left\|H_{yy}(x^k,y^k)^{-1}(H_{xy}^k)^\top - H_{yy}(x^k,y^k)^{-1}H_{xy}(x^k,y^k)^\top\right\|^2\\ \le&\frac{2C_{g_{xy}}^2}{\mu_g^4}\left\|H_{yy}^k - H_{yy}(x^k,y^k)\right\|^2 + \frac{2}{\mu_g^2}\left\|H_{xy}^k - H_{xy}(x^k,y^k)\right\|^2\end{aligned}\tag{57}$$

where the second inequality follows from $\|H_{xy}^k\| \le C_{g_{xy}}$ and $H_{yy}^k \succeq \mu_g I$.

Plugging (51)-(57) back to (50), we have

$$\begin{aligned} I_2 \le & \frac{3L_y^2}{2}\mathbb{E}[\|x^{k+1}-x^k\|^4|\mathcal{F}^k] + \frac{6C_{g_{xy}}^2}{\mu_g^4}\|H_{yy}(x^k,y^*(x^k)) - H_{yy}(x^k,y^k)\|^2\mathbb{E}[\|x^{k+1}-x^k\|^2|\mathcal{F}^k] \\ & + \frac{6}{\mu_g^2}\|H_{xy}(x^k,y^*(x^k)) - H_{xy}(x^k,y^k)\|^2\mathbb{E}[\|x^{k+1}-x^k\|^2|\mathcal{F}^k] \\ & + \frac{6C_{g_{xy}}^2}{\mu_g^4}\mathbb{E}[\|H_{yy}^k - H_{yy}(x^k,y^k)\|^2\|x^{k+1}-x^k\|^2|\mathcal{F}^k] \\ & + \frac{6}{\mu_g^2}\mathbb{E}[\|H_{xy}^k - H_{xy}(x^k,y^k)\|^2\|x^{k+1}-x^k\|^2|\mathcal{F}^k]. \end{aligned} \tag{58}$$

Using the Lipschitz continuity of $H_{xy}(x,y)$ and $H_{yy}(x,y)$ in Assumption 1, from (58), we have

$$\begin{aligned} I_2 \le & \frac{3L_y^2}{2}\mathbb{E}[\|x^{k+1}-x^k\|^4|\mathcal{F}^k] + \frac{6}{\mu_g^2}\left(\frac{C_{g_{xy}}^2 L_{g_{yy}}}{\mu_g^2} + L_{g_{xy}}\right)\|y^k - y^*(x^k)\|^2\mathbb{E}[\|x^{k+1}-x^k\|^2|\mathcal{F}^k] \\ & + \frac{6C_{g_{xy}}^2}{\mu_g^4}\mathbb{E}[\|H_{yy}^k - H_{yy}(x^k,y^k)\|^2\|x^{k+1}-x^k\|^2|\mathcal{F}^k] \\ & + \frac{6}{\mu_g^2}\mathbb{E}[\|H_{xy}^k - H_{xy}(x^k,y^k)\|^2\|x^{k+1}-x^k\|^2|\mathcal{F}^k]. \end{aligned} \tag{59}$$

For any $p = 2, 4$, we next analyze quantity $\mathbb{E}[\|x^{k+1}-x^k\|^p|\mathcal{F}^k]$ in (59). Recall the simplified update (43). Therefore, we have $\|x^{k+1}-x^k\| \le \alpha_k\|\bar{h}_f^k\|$ and

$$\begin{aligned} \|\bar{h}_f^k\| &= \left\|\nabla_x f\left(x^k,y^k;\xi^k\right) - (H_{yy}^k)^{-1}H_{xy}^k\nabla_y f\left(x^k,y^k;\xi^k\right)\right\| \\ &\le \left\|\nabla_x f(x^k,y^k;\xi^k)\right\| + \left\|(H_{yy}^k)^{-1}H_{xy}^k\nabla_y f\left(x^k,y^k;\xi^k\right)\right\| \\ &\overset{(a)}{\le} \left\|\nabla_x f(x^k,y^k;\xi^k)\right\| + \frac{C_{g_{xy}}}{\mu_g}\left\|\nabla_y f(x^k,y^k;\xi^k)\right\| \end{aligned} \tag{60}$$

where (a) follows from the upper and lower projections of H_{xy}^k and H_{yy}^k in (9).

Therefore, for $p = 2, 4$, we have

$$\begin{aligned} \mathbb{E}[\|\bar{h}_f^k\|^p|\mathcal{F}^k, H_{x,y}^k, H_{yy}^k] &\le 2^{p-1}\mathbb{E}\Big[\|\nabla_x f(x^k,y^k;\xi^k)\|^p|\mathcal{F}^k, H_{x,y}^k, H_{yy}^k\Big] \\ &\quad + 2^{p-1}\left(\frac{C_{g_{xy}}}{\mu_g}\right)^p\mathbb{E}\Big[\|\nabla_y f(x^k,y^k;\xi^k)\|^p|\mathcal{F}^k, H_{x,y}^k, H_{yy}^k\Big] \\ &\le 2^{p-1}\left(C_{f_x}^p + \left(\frac{C_{g_{xy}}}{\mu_g}\right)^p C_{f_y}^p\right) \end{aligned} \tag{61}$$

where the last inequality from Assumption 3. And thus

$$\mathbb{E}\left[\|x^{k+1}-x^k\|^p|\mathcal{F}^k, H_{xy}^k, H_{yy}^k\right] \le 2^{p-1}\left(C_{f_x}^p + \left(\frac{C_{g_{xy}}}{\mu_g}\right)^p C_{f_y}^p\right)\alpha_k^p. \tag{62}$$

Plugging (62) into (59), we have

$$\begin{aligned} I_2 \leq & 12L_y^2\left(C_{f_x}^4+\left(\frac{C_{g_{xy}}}{\mu_g}\right)^4 C_{f_y}^4\right)\alpha_k^4 \\ & +\frac{12}{\mu_g^2}\left(\frac{C_{g_{xy}}^2 L_{g_{yy}}}{\mu_g^2}+L_{g_{xy}}\right)\left(C_{f_x}^2+\left(\frac{C_{g_{xy}}}{\mu_g}\right)^2 C_{f_y}^2\right)\|y^k-y^*(x^k)\|^2\alpha_k^2 \\ & +\frac{12C_{g_{xy}}^2}{\mu_g^4}\left(C_{f_x}^2+\left(\frac{C_{g_{xy}}}{\mu_g}\right)^2 C_{f_y}^2\right)\mathbb{E}[\|H_{yy}^k-H_{yy}(x^k,y^k)\|^2|\mathcal{F}^k]\alpha_k^2 \\ & +\frac{12}{\mu_g^2}\left(C_{f_x}^2+\left(\frac{C_{g_{xy}}}{\mu_g}\right)^2 C_{f_y}^2\right)\mathbb{E}[\|H_{xy}^k-H_{xy}(x^k,y^k)\|^2|\mathcal{F}^k]\alpha_k^2. \end{aligned} \tag{63}$$

Now let us define the constants as

$$\begin{aligned} \tilde{c}_1 := \max\Bigg\{ & 12L_y^2\left(C_{f_x}^4+\left(\frac{C_{g_{xy}}}{\mu_g}\right)^4 C_{f_y}^4\right), \frac{12}{\mu_g^2}\left(\frac{C_{g_{xy}}^2 L_{g_{yy}}}{\mu_g^2}+L_{g_{xy}}\right)\left(C_{f_x}^2+\left(\frac{C_{g_{xy}}}{\mu_g}\right)^2 C_{f_y}^2\right), \\ & \frac{12C_{g_{xy}}^2}{\mu_g^4}\left(C_{f_x}^2+\left(\frac{C_{g_{xy}}}{\mu_g}\right)^2 C_{f_y}^2\right), \frac{12}{\mu_g^2}\left(C_{f_x}^2+\left(\frac{C_{g_{xy}}}{\mu_g}\right)^2 C_{f_y}^2\right)\Bigg\} \\ \tilde{c}_2 := & \frac{2}{\mu_g+L_g}+\frac{\mu_g+L_g}{\mu_g L_g}, \quad c := \tilde{c}_1\tilde{c}_2. \end{aligned}$$

Plugging the upper bounds of I_1 in (49) and I_2 in (63) into (48) with $\epsilon = \frac{\mu_g L_g}{\mu_g+L_g}\beta^k$, we have

$$\begin{aligned} & \mathbb{E}\left[\|y^{k+1}-y^*(x^{k+1})\|^2|\mathcal{F}^k\right] \\ \leq & \left(1-\frac{\mu_g L_g}{\mu_g+L_g}\beta^k\right)\|y^k-y^*(x^k)\|^2+\left(1+\frac{\mu_g L_g}{\mu_g+L_g}\beta^k\right)\beta_k^2\sigma_{g_y}^2+\tilde{c}_1\tilde{c}_2\frac{\alpha_k^4}{\beta_k} \\ & +\tilde{c}_1\tilde{c}_2\frac{\alpha_k^2}{\beta_k}\|y^k-y^*(x^k)\|^2+\tilde{c}_1\tilde{c}_2\mathbb{E}\left[\|H_{yy}^k-H_{yy}(x^k,y^k)\|^2|\mathcal{F}^k\right]\frac{\alpha_k^2}{\beta_k} \\ & +\tilde{c}_1\tilde{c}_2\mathbb{E}\left[\|H_{xy}^k-H_{xy}(x^k,y^k)\|^2|\mathcal{F}^k\right]\frac{\alpha_k^2}{\beta_k} \end{aligned} \tag{64}$$

where we have used the fact that

$$\begin{aligned} \left(1+\frac{\mu_g L_g}{\mu_g+L_g}\beta^k\right)\left(1-\frac{2\mu_g L_g}{\mu_g+L_g}\beta^k\right) & \leq 1-\frac{\mu_g L_g}{\mu_g+L_g}\beta^k \\ \left(1+\left(\frac{\mu_g L_g}{\mu_g+L_g}\beta^k\right)^{-1}\right) & \leq \frac{1}{\beta_k}\left(\frac{2}{\mu_g+L_g}+\frac{\mu_g+L_g}{\mu_g L_g}\right)=\frac{\tilde{c}_2}{\beta_k} \end{aligned}$$

where the last inequality uses $\beta_k \leq \frac{2}{\mu_g+L_g}$ in (28a). The proof is complete by defining $c := \tilde{c}_1\tilde{c}_2$.

F Proof of Lemma 4

Proof: Recall that $g(x,y) = \mathbb{E}_\phi[g(x,y,\phi)]$. We only have access to the stochastic estimates of $\nabla_{xy}^2 g(x,y), \nabla_{yy}^2 g(x,y)$, that is

$$h_{yy}^k(\phi) := \nabla_{yy}^2 g\left(x^k,y^k;\phi\right), \qquad h_{xy}^k(\phi) := \nabla_{xy}^2 g\left(x^k,y^k;\phi\right). \tag{65}$$

For notational brevity in the analysis, we define

$$H_{xy}(x,y) := \nabla_{xy}^2 g(x,y), \qquad H_{yy}(x,y) := \nabla_{yy}^2 g(x,y). \tag{66}$$

and rewrite the update of (9) as

$$H_{xy}^k := \mathcal{P}_{\{X:\|X\|\le C_{g_{xy}}\}}\left\{\hat{H}_{xy}^k\right\} \quad \text{with} \quad \hat{H}_{xy}^k := (1-\tau_k)(H_{xy}^{k-1} - h_{xy}^{k-1}(\phi^k)) + h_{xy}^k(\phi^k) \tag{67a}$$

$$H_{yy}^k := \mathcal{P}_{\{X:X\succeq \mu_g I\}}\left\{\hat{H}_{yy}^k\right\} \quad \text{with} \quad \hat{H}_{yy}^k := (1-\tau_k)\big(H_{yy}^{k-1} - h_{yy}^{k-1}(\phi^k)\big) + h_{yy}^k(\phi^k). \tag{67b}$$

To analyze the approximation error of H_{xy}^k, we decompose it into

$$\begin{aligned}\mathbb{E}\Big[\|H_{xy}^k - H_{xy}(x^k,y^k)\|^2\big|\mathcal{F}^k\Big] &\le \mathbb{E}\left[\|\hat{H}_{xy}^k - H_{xy}(x^k,y^k)\|^2\big|\mathcal{F}^k\right] \\ &= \left\|\mathbb{E}\left[\hat{H}_{xy}^k - H_{xy}(x^k,y^k)|\mathcal{F}^k\right]\right\|^2 + \sum_{i,j}\mathrm{Var}\left[(\hat{H}_{xy}^k - H_{xy}(x^k,y^k))_{i,j}|\mathcal{F}^k\right]\end{aligned} \tag{68}$$

where the inequality holds since the projection onto the convex set $\{X : X \succeq \mu_g I\}$ is non-expansive, and the equality comes from the bias-variance decomposition that $\mathrm{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$ for any random variable X.

We first analyze the bias term in (68) by

$$\begin{aligned}&\mathbb{E}\left[\hat{H}_{xy}^k - H_{xy}(x^k,y^k)|\mathcal{F}^k\right] \\ \overset{(9)}{=}\ &\mathbb{E}\left[(1-\tau_k)\left(H_{xy}^{k-1} + h_{xy}^k(\phi^k) - h_{xy}^{k-1}(\phi^k)\right) + \tau_k h_{xy}^k(\phi^k) - H_{xy}(x^k,y^k)|\mathcal{F}^k\right] \\ =\ &(1-\tau_k)\left(H_{xy}^{k-1} + H_{xy}(x^k,y^k) - H_{xy}(x^{k-1},y^{k-1})\right) + \tau_k H_{xy}(x^k,y^k) - H_{xy}(x^k,y^k) \\ =\ &(1-\tau_k)\left(H_{xy}^{k-1} - H_{xy}(x^{k-1},y^{k-1})\right).\end{aligned} \tag{69}$$

The variance term in (68) follows

$$\begin{aligned}&\sum_{i,j}\mathrm{Var}\left[(\hat{H}_{xy}^k - H_{xy}(x^k,y^k))_{i,j}|\mathcal{F}^k\right] = \sum_{i,j}\mathrm{Var}\left[(\hat{H}_{xy}^k)_{i,j}|\mathcal{F}^k\right] \\ \overset{(67a)}{=}\ &\sum_{i,j}\mathrm{Var}\left[(1-\tau_k)(h_{xy}^k(\phi^k) - h_{xy}^{k-1}(\phi^k))_{i,j} + \tau_k(h_{xy}^k(\phi^k))_{i,j}|\mathcal{F}^k\right] \\ \le\ &2(1-\tau_k)^2\sum_{i,j}\mathrm{Var}\left[(h_{xy}^k(\phi^k) - h_{xy}^{k-1}(\phi^k))_{i,j}|\mathcal{F}^k\right] + 2\tau_k^2\sum_{i,j}\mathrm{Var}\left[(h_{xy}^k(\phi^k))_{i,j}|\mathcal{F}^k\right] \\ \overset{(a)}{\le}\ &2(1-\tau_k)^2\mathbb{E}\left[\|h_{xy}^k(\phi^k) - h_{xy}^{k-1}(\phi^k)\|^2|\mathcal{F}^k\right] + 2\tau_k^2\sum_{i,j}\mathrm{Var}\left[(h_{xy}^k(\phi^k))_{i,j}|\mathcal{F}^k\right] \\ \overset{(b)}{\le}\ &2(1-\tau_k)^2\left(\bar{L}_{g_{xy}}^2 + L_{g_{xy}}^2\right)\left(\|x^k - x^{k-1}\|^2 + \|y^k - y^{k-1}\|^2\right) + 2\tau_k^2\sigma_{g_{xy}}^2\end{aligned} \tag{70}$$

where (a) uses $\mathrm{Var}[X] \le \mathbb{E}[X]^2$ and (b) follows from Assumptions 1 and 3.

Therefore, plugging (69) and (70) into (68), we have

$$\begin{aligned}\mathbb{E}[\|H_{xy}^k - H_{xy}(x^k,y^k)\|^2|\mathcal{F}^k] \le\ &(1-\tau_k)^2\left\|H_{xy}^{k-1} - H_{xy}(x^{k-1},y^{k-1})\right\|^2 + 2\tau_k^2\sigma_{g_{xy}}^2 \\ &+ 2(1-\tau_k)^2\left(\bar{L}_{g_{xy}}^2 + L_{g_{xy}}^2\right)\left(\|x^k - x^{k-1}\|^2 + \|y^k - y^{k-1}\|^2\right).\end{aligned}$$

Similarly, we can derive the approximation error of H_{yy}^k as

$$\begin{aligned}\mathbb{E}[\|H_{yy}^k - H_{yy}(x^k,y^k)\|^2|\mathcal{F}^k] \le\ &(1-\tau_k)^2\|H_{yy}^{k-1} - H_{yy}(x^{k-1},y^{k-1})\|^2 + 2\tau_k^2\sigma_{g_{yy}}^2 \\ &+ 2(1-\tau_k)^2\left(\bar{L}_{g_{yy}}^2 + L_{g_{yy}}^2\right)\left(\|x^k - x^{k-1}\|^2 + \|y^k - y^{k-1}\|^2\right).\end{aligned}$$

The proof is then complete.

G Proof of Theorem 2

Proof: Using Lemmas 2-4, we, respectively, bound the four difference terms in (34) and obtain

$$\begin{aligned}\mathbb{E}[\mathbb{V}^{k+1}]-\mathbb{E}[\mathbb{V}^k] \le & -\frac{(\mu_F+\rho)\rho\alpha_k}{4}\mathbb{E}[\|\widehat{x}(x^k)-x^k\|^2]-\left(\frac{\mu_g L_g}{\mu_g+L_g}\beta_k-\tilde{c}_1\tilde{c}_2\frac{\alpha_k^2}{\beta_k}-\frac{2L_f^2\rho\alpha_k}{\mu_F+\rho}\right)\mathbb{E}[\|y^k-y^*(x^k)\|^2]\\ & -\left(\tau_{k+1}-\tilde{c}_1\tilde{c}_2\frac{\alpha_k^2}{\beta_k}-\frac{4\rho\alpha_k C_{g_{xy}}^2 C_{f_y}^2}{(\mu_F+\rho)\mu_g^4}\right)\mathbb{E}[\|H_{yy}^k-H_{yy}(x^k,y^k)\|^2]\\ & -\left(\tau_{k+1}-\tilde{c}_1\tilde{c}_2\frac{\alpha_k^2}{\beta_k}-\frac{4\rho\alpha_k C_{f_y}^2}{(\mu_F+\rho)\mu_g^2}\right)\mathbb{E}[\|H_{xy}^k-H_{xy}(x^k,y^k)\|^2]\\ & +2\rho\alpha_k^2\left(C_{f_x}^2+\left(\frac{C_{g_{xy}}}{\mu_g}\right)^2 C_{f_y}^2\right)+\left(1+\frac{\mu_g L_g}{\mu_g+L_g}\beta^k\right)\beta_k^2\sigma_{g_y}^2+\tilde{c}_1\tilde{c}_2\frac{\alpha_k^4}{\beta_k}+4\tau_{k+1}^2\sigma_{g_y}^2\\ & +4(1-\tau_{k+1})^2\tilde{c}_3\alpha_k^2\left(C_{f_x}^2+\left(\frac{C_{g_{xy}}}{\mu_g}\right)^2 C_{f_y}^2\right)+2(1-\tau_{k+1})^2\tilde{c}_3\mathbb{E}[\|y^{k+1}-y^k\|^2]. \end{aligned} \tag{71}$$

where the constant is defined as $\tilde{c}_3 := \bar{L}_{g_{xy}}^2+L_{g_{xy}}^2+\bar{L}_{g_{yy}}^2+L_{g_{yy}}^2$.

Note that using the y-update (8b), we also have

$$\begin{aligned}\mathbb{E}[\|y^{k+1}-y^k\|^2] &= \mathbb{E}\big[\big\|\beta_k h_g^k-(H_{yy}^k)^{-1}H_{xy}^k(x^{k+1}-x^k)\big\|^2\big]\\ &\le 2\beta_k^2\mathbb{E}\big[\|h_g^k\|^2\big]+2\mathbb{E}\big[\|(H_{yy}^k)^{-1}\|^2\|H_{xy}^k\|^2\big\|x^{k+1}-x^k\big\|^2\big]\\ &\overset{(a)}{\le} 2\beta_k^2\mathbb{E}\big[\|\nabla_y g(x^k,y^k)\|^2\big]+2\beta_k^2\sigma_{g_y}^2+2\mathbb{E}\big[\|(H_{yy}^k)^{-1}\|^2\|H_{xy}^k\|^2\big\|x^{k+1}-x^k\big\|^2\big]\\ &\overset{(b)}{\le} 2\beta_k^2\mathbb{E}\big[\|\nabla_y g(x^k,y^k)\|^2\big]+2\beta_k^2\sigma_{g_y}^2+2\left(\frac{C_{g_{xy}}}{\mu_g}\right)^2\mathbb{E}\big[\big\|x^{k+1}-x^k\big\|^2\big]\\ &\overset{(c)}{\le} 4\beta_k^2L_g^2\mathbb{E}[\|y^k-y^*(x^k)\|^2]+2\beta_k^2\sigma_{g_y}^2+2\left(\frac{C_{g_{xy}}}{\mu_g}\right)^2\mathbb{E}[\|x^{k+1}-x^k\|^2] \end{aligned} \tag{72}$$

where (a) follows from $\mathbb{E}[X^2]=\mathrm{Var}[X]+\mathbb{E}[X]^2$ and Assumption 3, (b) uses the upper and lower projections of H_{xy}^k and H_{yy}^k in (9), and (c) is due to $\nabla_y g(x^k,y^*(x^k))=0$ as well as Assumption 1.

Selecting parameter $\tau_k=\frac{1}{\sqrt{K}}$, using (62) to bound $\mathbb{E}[\|x^{k+1}-x^k\|^2]$ and using (71)-(72), we have

$$\begin{aligned}\mathbb{E}[\mathbb{V}^{k+1}]-\mathbb{E}[\mathbb{V}^k] \le & -\frac{(\mu_F+\rho)\rho\alpha_k}{4}\mathbb{E}[\|\widehat{x}(x^k)-x^k\|^2]+\alpha_k^2\left(2\rho+4\tilde{c}_3+8\tilde{c}_3\left(\frac{C_{g_{xy}}}{\mu_g}\right)^2\right)\left(C_{f_x}^2+\left(\frac{C_{g_{xy}}}{\mu_g}\right)^2 C_{f_y}^2\right)\\ & -\left(\frac{\mu_g L_g}{\mu_g+L_g}\beta_k-\tilde{c}_1\tilde{c}_2\frac{\alpha_k^2}{\beta_k}-\frac{2L_f^2\rho\alpha_k}{\mu_F+\rho}-8\tilde{c}_3\beta_k^2L_g^2\right)\mathbb{E}[\|y^k-y^*(x^k)\|^2]\\ & -\left(\frac{1}{\sqrt{K}}-\tilde{c}_1\tilde{c}_2\frac{\alpha_k^2}{\beta_k}-\frac{4\rho C_{g_{xy}}^2C_{f_y}^2\alpha_k}{(\mu_F+\rho)\mu_g^4}\right)\mathbb{E}[\|H_{yy}^k-H_{yy}(x^k,y^k)\|^2]\\ & -\left(\frac{1}{\sqrt{K}}-\tilde{c}_1\tilde{c}_2\frac{\alpha_k^2}{\beta_k}-\frac{4\rho C_{f_y}^2\alpha_k}{(\mu_F+\rho)\mu_g^2}\right)\mathbb{E}[\|H_{xy}^k-H_{xy}(x^k,y^k)\|^2]\\ & +\left(1+\frac{\mu_g L_g}{\mu_g+L_g}\beta^k\right)\beta_k^2\sigma_{g_y}^2+\tilde{c}_1\tilde{c}_2\frac{\alpha_k^4}{\beta_k}+\frac{4\sigma_{g_y}^2}{K}+4\tilde{c}_3\beta_k^2\sigma_{g_y}^2. \end{aligned} \tag{73}$$

Choosing the stepsize α_k as (28), it will lead to (cf. $c := \tilde{c}_1\tilde{c}_2$)

$$\frac{1}{\sqrt{K}} - \tilde{c}_1\tilde{c}_2\frac{\alpha_k^2}{\beta_k} - \frac{4\rho C_{g_{xy}}^2 C_{f_y}^2 \alpha_k}{(\mu_F+\rho)\mu_g^4} \overset{(a)}{\geq} \frac{1}{\sqrt{K}} - \tilde{c}_1\tilde{c}_2\alpha_k - \frac{4\rho C_{g_{xy}}^2 C_{f_y}^2 \alpha_k}{(\mu_F+\rho)\mu_g^4} \overset{(b)}{\geq} 0 \tag{74a}$$

$$\frac{1}{\sqrt{K}} - \tilde{c}_1\tilde{c}_2\frac{\alpha_k^2}{\beta_k} - \frac{4\rho C_{f_y}^2 \alpha_k}{(\mu_F+\rho)\mu_g^2} \overset{(c)}{\geq} \frac{1}{\sqrt{K}} - \tilde{c}_1\tilde{c}_2\alpha_k - \frac{4\rho C_{f_y}^2 \alpha_k}{(\mu_F+\rho)\mu_g^2} \overset{(d)}{\geq} 0 \tag{74b}$$

where both (a) and (c) follow from $\alpha_k \leq \beta_k$ in (28b); and (b) and (d) follow from the second and the third terms in (28b). In addition, choosing the stepsize β_k as (28) will lead to

$$\begin{aligned}\frac{\mu_g L_g}{\mu_g+L_g}\beta_k - \tilde{c}_1\tilde{c}_2\frac{\alpha_k^2}{\beta_k} - \frac{2L_f^2\rho\alpha_k}{\mu_F+\rho} - 8\tilde{c}_3\beta_k^2 L_g^2 &\overset{(e)}{\geq} \frac{\mu_g L_g}{\mu_g+L_g}\beta_k - (\tilde{c}_1\tilde{c}_2 + \frac{2L_f^2\rho}{\mu_F+\rho})\alpha_k - 8\tilde{c}_3\beta_k^2 L_g^2 \\ &\overset{(f)}{\geq} \frac{\mu_g L_g \beta_k}{2(\mu_g+L_g)} - 8\tilde{c}_3\beta_k^2 L_g^2 \overset{(g)}{\geq} \frac{\mu_g L_g \beta_k}{4(\mu_g+L_g)}\end{aligned} \tag{74c}$$

where (e) follows from $\alpha_k \leq \beta_k$ in (28b), (f) is due to the last terms in (28b), and (g) uses (28a).

Using (74) to cancel terms in (73), we are able to get

$$\mathbb{E}[\mathbb{V}^{k+1}] - \mathbb{E}[\mathbb{V}^k] \leq -\frac{\mu_g L_g \beta_k}{4(\mu_g+L_g)}\mathbb{E}[\|y^k - y^*(x^k)\|^2] - \frac{(\mu_F+\rho)\rho\alpha_k}{4}\mathbb{E}[\|\widehat{x}(x^k) - x^k\|^2] + \mathcal{O}\left(\frac{1}{K}\right) \tag{75}$$

from which we can reach Theorem 2 after telescoping the both sides of (75).

H Proof of Theorem 3

Slightly different from the Lyapunov function (26), we define the following Lyapunov function

$$\mathbb{V}^k := \|x^k - x^*\|^2 + \|y^k - y^*(x^k)\|^2 + \|H_{yy}^k - \nabla_{yy}^2 g(x^k, y^k)\|^2 + \|H_{xy}^k - \nabla_{xy}^2 g(x^k, y^k)\|^2.$$

Lemma 6 *Suppose Assumptions 1–3 hold and $F(x)$ is μ-strongly convex. Then x^k satisfies*

$$\begin{aligned}\mathbb{E}[\|x^{k+1} - x^*\|^2] \leq (1-\mu\alpha_k)\mathbb{E}[\|x^k - x^*\|^2] + \frac{2L_f^2}{\mu}\alpha_k\mathbb{E}[\|y^k - y^*(x^k)\|^2] + \alpha_k^2\mathbb{E}[\|\bar{h}_f^k\|^2] \\ + \frac{4C_{g_{xy}}^2 C_{f_y}^2}{\mu_g^4\mu}\alpha_k\mathbb{E}[\|H_{yy}^k - H_{yy}(x^k, y^k)\|^2] + \frac{4C_{f_y}^2}{\mu_g^2\mu}\alpha_k\mathbb{E}[\|H_{xy}^k - H_{xy}(x^k, y^k)\|^2]\end{aligned} \tag{76}$$

where L_f, L_F are defined in Lemma 5, and $C_{g_{xy}}$ is the projection radius of H_{xy}^k in (9a).

Proof: We start with

$$\begin{aligned}\mathbb{E}[\|x^{k+1} - x^*\|^2|\mathcal{F}^k] &\overset{(a)}{\leq} \mathbb{E}[\|x^k - \alpha_k\bar{h}_f^k - x^*\|^2|\mathcal{F}^k] \\ &= \|x^k - x^*\|^2 - 2\alpha_k\langle x^k - x^*, \mathbb{E}[\bar{h}_f^k|\mathcal{F}^k]\rangle + \alpha_k^2\mathbb{E}[\|\bar{h}_f^k\|^2|\mathcal{F}^k] \\ &= \|x^k - x^*\|^2 - 2\alpha_k\langle x^k - x^*, \nabla F(x^k)\rangle \\ &\quad + 2\alpha_k\langle x^k - x^*, \nabla F(x^k) - \mathbb{E}[\bar{h}_f^k|\mathcal{F}^k]\rangle + \alpha_k^2\mathbb{E}[\|\bar{h}_f^k\|^2|\mathcal{F}^k] \\ &\overset{(b)}{\leq} \|x^k - x^*\|^2 - 2\alpha_k\langle x^k - x^*, \nabla F(x^k) - \nabla F(x^*)\rangle \\ &\quad + 2\alpha_k\langle x^k - x^*, \nabla F(x^k) - \mathbb{E}[\bar{h}_f^k|\mathcal{F}^k]\rangle + \alpha_k^2\mathbb{E}[\|\bar{h}_f^k\|^2|\mathcal{F}^k]\end{aligned} \tag{77}$$

where (a) follows the fact that $\mathcal{P}_{\mathcal{X}}$ is non-expansive, and (b) follows the optimality condition that $\langle\nabla F(x^*), x - x^*\rangle \geq 0$ for any $x \in \mathcal{X}$.

Using the μ-strong convexity of $F(x)$, it follows that

$$-\langle x^k - x^*, \nabla F(x^k) - \nabla F(x^*)\rangle \leq -\mu\|x^k - x^*\|^2 \tag{78}$$

plugging which into (77) leads to

$$\begin{aligned}\mathbb{E}\left[\|x^{k+1}-x^*\|^2|\mathcal{F}^k\right] &\leq (1-2\mu\alpha_k)\|x^k-x^*\|^2+2\alpha_k\langle x^k-x^*,\nabla F(x^k)-\mathbb{E}[\bar{h}_f^k|\mathcal{F}^k]\rangle+\alpha_k^2\mathbb{E}\left[\|\bar{h}_f^k\|^2|\mathcal{F}^k\right]\\ &\overset{(c)}{\leq}(1-\mu\alpha_k)\|x^k-x^*\|^2+\frac{\alpha_k}{\mu}\left\|\nabla F(x^k)-\mathbb{E}[\bar{h}_f^k|\mathcal{F}^k]\right\|^2+\alpha_k^2\mathbb{E}\left[\|\bar{h}_f^k\|^2|\mathcal{F}^k\right] \end{aligned} \tag{79}$$

where (c) uses the Young's inequality.

The approximation error of $\bar{h}_f^k$ can be bounded by

$$\begin{aligned}&\left\|\nabla F(x^k)-\mathbb{E}[\bar{h}_f^k|\mathcal{F}^k]\right\|^2\\ &\leq 2\left\|\nabla F(x^k)-\overline{\nabla} f(x^k,y^k)\right\|^2+2\mathbb{E}\left[\|\overline{\nabla} f(x^k,y^k)-\mathbb{E}_{\xi^k}[\bar{h}_f^k]\|^2|\mathcal{F}^k\right]\\ &\overset{(a)}{\leq} 2L_f^2\left\|y^k-y^*(x^k)\right\|^2+2\mathbb{E}\left[\|\overline{\nabla} f(x^k,y^k)-\mathbb{E}_{\xi^k}[\bar{h}_f^k]\|^2|\mathcal{F}^k\right]\\ &\overset{(b)}{\leq} 2L_f^2\left\|y^k-y^*(x^k)\right\|^2+2\left\|(H_{yy}^k)^{-1}H_{xy}^k-H_{yy}(x^k,y^k)^{-1}H_{xy}(x^k,y^k)\right\|^2\left\|\nabla_y f(x^k,y^k)\right\|^2\\ &\overset{(c)}{\leq} 2L_f^2\|y^k-y^*(x^k)\|^2+\frac{4C_{g_{xy}}^2C_{f_y}^2}{\mu_g^4}\mathbb{E}[\|H_{yy}^k-H_{yy}(x^k,y^k)\|^2|\mathcal{F}^k]+\frac{4C_{f_y}^2}{\mu_g^2}\mathbb{E}[\|H_{xy}^k-H_{xy}(x^k,y^k)\|^2|\mathcal{F}^k]\end{aligned} \tag{80}$$

where (a) follows from Lemma 5, (b) uses the fact that

$$\mathbb{E}_{\xi^k}[\bar{h}_f^k|\mathcal{F}^k]=\nabla_x f\left(x^k,y^k\right)-(H_{yy}^k)^{-1}H_{xy}^k\nabla_y f\left(x^k,y^k\right) \tag{81}$$

and (c) follows the same steps of (57) and Assumption 3. Plugging (80) into the above completes the proof.

Similar to (34), we first quantify the difference between consecutive Lyapunov functions as

$$\begin{aligned}\mathbb{V}^{k+1}-\mathbb{V}^k=&\underbrace{\|x^{k+1}-x^*\|^2-\|x^k-x^*\|^2}_{\text{Lemma 6}}+\underbrace{\|y^{k+1}-y^*(x^{k+1})\|^2-\|y^k-y^*(x^k)\|^2}_{\text{Lemma 3}}\\ &+\underbrace{\|H_{yy}^{k+1}-\nabla_{yy}^2g(x^{k+1},y^{k+1})\|^2-\|H_{yy}^k-\nabla_{yy}^2g(x^k,y^k)\|^2}_{\text{Lemma 4}}\\ &+\underbrace{\|H_{xy}^{k+1}-\nabla_{xy}^2g(x^{k+1},y^{k+1})\|^2-\|H_{xy}^k-\nabla_{xy}^2g(x^k,y^k)\|^2}_{\text{Lemma 4}}.\end{aligned} \tag{82}$$

Using Lemmas 3-4 and 6 and defining $\tilde{c}_3:=\bar{L}_{g_{xy}}^2+L_{g_{xy}}^2+\bar{L}_{g_{yy}}^2+L_{g_{yy}}^2$, we obtain

$$\begin{aligned}\mathbb{E}[\mathbb{V}^{k+1}]-\mathbb{E}[\mathbb{V}^k]\leq&-\mu\alpha_k\mathbb{E}[\|x^k-x^*\|^2]-\left(\frac{\mu_gL_g\beta_k}{\mu_g+L_g}-\tilde{c}_1\tilde{c}_2\frac{\alpha_k^2}{\beta_k}-\frac{2L_f^2\alpha_k}{\mu}\right)\mathbb{E}[\|y^k-y^*(x^k)\|^2]\\ &-\left(\tau_{k+1}-\tilde{c}_1\tilde{c}_2\frac{\alpha_k^2}{\beta_k}-\frac{4C_{g_{xy}}^2C_{f_y}^2}{\mu_g^4\mu}\alpha_k\right)\mathbb{E}[\|H_{yy}^k-H_{yy}(x^k,y^k)\|^2]\\ &-\left(\tau_{k+1}-\tilde{c}_1\tilde{c}_2\frac{\alpha_k^2}{\beta_k}-\frac{4C_{f_y}^2}{\mu_g^2\mu}\alpha_k\right)\mathbb{E}[\|H_{xy}^k-H_{xy}(x^k,y^k)\|^2]\\ &+\alpha_k^2\mathbb{E}[\|\bar{h}_f^k\|^2]+\left(1+\frac{\mu_gL_g}{\mu_g+L_g}\beta^k\right)\beta_k^2\sigma_{g_y}^2+\tilde{c}_1\tilde{c}_2\frac{\alpha_k^4}{\beta_k}+4\tau_{k+1}^2\sigma_{g_y}^2\\ &+2(1-\tau_{k+1})^2\tilde{c}_3\mathbb{E}[\|x^{k+1}-x^k\|^2]+2(1-\tau_{k+1})^2\tilde{c}_3\mathbb{E}[\|y^{k+1}-y^k\|^2].\end{aligned} \tag{83}$$

Plugging (72) and (61) into (83), we have

$$\begin{aligned}\mathbb{E}[\mathbb{V}^{k+1}]-\mathbb{E}[\mathbb{V}^k] \le& -\mu\alpha_k\mathbb{E}[\|x^k-x^*\|^2]+\underbrace{\left(2+4\tilde{c}_3+8\tilde{c}_3\Big(\frac{C_{g_{xy}}}{\mu_g^2}\Big)^2\right)\left(C_{f_x}^2+\Big(\frac{C_{g_{xy}}}{\mu_g}\Big)^2C_{f_y}^2\right)}_{\tilde{c}_4:=}\alpha_k^2\\ &-\left(\frac{\mu_g L_g\beta_k}{\mu_g+L_g}-\tilde{c}_1\tilde{c}_2\frac{\alpha_k^2}{\beta_k}-\frac{2L_f^2\alpha_k}{\mu}-8\beta_k^2L^2L_g^2\right)\mathbb{E}[\|y^k-y^*(x^k)\|^2]\\ &-\left(\tau_{k+1}-\tilde{c}_1\tilde{c}_2\frac{\alpha_k^2}{\beta_k}-\frac{4C_{g_{xy}}^2C_{f_y}^2}{\mu_g^4\mu}\alpha_k\right)\mathbb{E}[\|H_{yy}^k-H_{yy}(x^k,y^k)\|^2]\\ &-\left(\tau_{k+1}-\tilde{c}_1\tilde{c}_2\frac{\alpha_k^2}{\beta_k}-\frac{4C_{f_y}^2}{\mu_g^2\mu}\alpha_k\right)\mathbb{E}[\|H_{xy}^k-H_{xy}(x^k,y^k)\|^2]\\ &+\left(1+\frac{\mu_g L_g}{\mu_g+L_g}\beta^k\right)\beta_k^2\sigma_{g_y}^2+\tilde{c}_1\tilde{c}_2\frac{\alpha_k^4}{\beta_k}+4\tau_{k+1}^2\sigma_{g_y}^2+8L_g^2\beta_k^2\sigma_{g_y}^2. \end{aligned}\tag{84}$$

We choose the stepsizes α_k,β_k,τ_k as (30) to guarantee that (cf. $c:=\tilde{c}_1\tilde{c}_2$)

$$\begin{aligned}&\text{(a)}\quad \tau_{k+1}-\tilde{c}_1\tilde{c}_2\frac{\alpha_k^2}{\beta_k}-\frac{4C_{g_{xy}}^2C_{f_y}^2}{\mu_g^4\mu}\alpha_k\ge\frac{\beta_k}{4};\quad \text{(b)}\quad \tau_{k+1}-\tilde{c}_1\tilde{c}_2\frac{\alpha_k^2}{\beta_k}-\frac{4C_{f_y}^2}{\mu_g^2\mu}\alpha_k\ge\frac{\beta_k}{4}\\ &\text{(c)}\quad \frac{\mu_g L_g}{\mu_g+L_g}\beta_k-\tilde{c}_1\tilde{c}_2\frac{\alpha_k^2}{\beta_k}-\frac{2L_f^2}{\mu}\alpha_k-8\beta_k^2L^2L_g^2\ge\frac{\mu_g L_g}{4(\mu_g+L_g)}.\end{aligned}\tag{85}$$

Therefore, plugging (85) into (84), we have

$$\begin{aligned}\mathbb{E}[\mathbb{V}^{k+1}]-\mathbb{E}[\mathbb{V}^k]\le& -\mu\alpha_k\mathbb{E}\Big[\|x^k-x^*\|^2\Big]-\frac{\mu_g L_g}{4(\mu_g+L_g)}\beta_k\mathbb{E}\Big[\|y^k-y^*(x^k)\|^2\Big]\\ &-\frac{\beta_k}{4}\mathbb{E}\Big[\|H_{yy}^k-H_{yy}(x^k,y^k)\|^2\Big]-\frac{\beta_k}{4}\mathbb{E}\Big[\|H_{xy}^k-H_{xy}(x^k,y^k)\|^2\Big]+\tilde{c}_6\beta_k^2\\ \le&-\tilde{c}_5\beta_k\mathbb{E}[\mathbb{V}^k]+\tilde{c}_6\beta_k^2\end{aligned}\tag{86}$$

where the first and second inequalities hold since we define

$$\begin{aligned}\tilde{c}_5&:=\min\left\{\frac{\mu\alpha_k}{\beta_k},\frac{\mu_g L_g}{4(\mu_g+L_g)},\frac{1}{4}\right\}=\mathcal{O}(1)\\ \tilde{c}_6&:=\left(1+\frac{\mu_g L_g}{\mu_g+L_g}\beta^k\right)\sigma_{g_y}^2+\frac{\alpha_k^2}{4\beta_k}+4\sigma_{g_y}^2+8L_g^2\sigma_{g_y}^2+\tilde{c}_4=\mathcal{O}(1).\end{aligned}\tag{87}$$

If we choose $\beta_k=\frac{2}{\tilde{c}_5(K_0+k)}$, where K_0 is a sufficiently large constant, then we have

$$\begin{aligned}\mathbb{E}[\mathbb{V}^K]\le&\prod_{k=0}^{K-1}(1-\tilde{c}_5\beta_k)\mathbb{V}^0+\tilde{c}_6\sum_{k=0}^{K-1}\beta_k^2\prod_{j=k+1}^{k-1}(1-\tilde{c}_5\beta_j)\\ \le&\frac{(K_0-2)(K_0-1)}{(K_0+K-2)(K_0+K-1)}\mathbb{V}^0+\frac{\tilde{c}_6}{\tilde{c}_5^2}\sum_{k=0}^{K-1}\frac{4}{(k+K_0)^2}\frac{(k+K_0-1)(k+K_0)}{(K+K_0-2)(K+K_0-1)}\\ \le&\frac{(K_0-1)^2}{(K_0+K-1)^2}\mathbb{V}^0+\frac{4\tilde{c}_6K}{\tilde{c}_5^2(K+K_0-1)^2}\end{aligned}\tag{88}$$

from which the proof is complete.