A Single-Timescale Method for Stochastic Bilevel Optimization

Tianyi Chen\textsuperscript{*} \hspace{1cm} Yuejiao Sun\textsuperscript{†} \hspace{1cm} Quan Xiao\textsuperscript{*} \hspace{1cm} Wotao Yin\textsuperscript{†}

\textsuperscript{*}Rensselaer Polytechnic Institute \hspace{2cm} \textsuperscript{†}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 \( \epsilon \)-stationary point of the bilevel problem, STABLE requires \( \mathcal{O}(\epsilon^{-2}) \) samples in total; and to achieve an \( \epsilon \)-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

\[
\begin{align*}
\min_{x \in \mathcal{X}} & \quad F(x) := \mathbb{E}_\xi [f(x, y^*(x); \xi)] \quad \text{(upper)} \\
\text{s.t.} & \quad y^*(x) \in \arg\min_{y \in \mathbb{R}^d_y} \mathbb{E}_\phi [g(x, y; \phi)] \quad \text{(lower)}
\end{align*}
\]

where \( f \) and \( g \) are differentiable functions; \( \xi \) and \( \phi \) 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.
Table 1: Sample complexity of several state-of-the-art algorithms (BSA in Ghadimi and Wang 2018, TTSA in Hong et al. 2019, stocBiO in Ji et al. 2020) to achieve an $\epsilon$-stationary point of $F(x)$ in the nonconvex setting and an $\epsilon$-optimal solution of $F(x)$ in the strongly convex setting; the notation $\tilde{O}(\cdot)$ hides logarithmic terms of $\epsilon^{-1}$.

<table>
<thead>
<tr>
<th></th>
<th>STABLE</th>
<th>BSA</th>
<th>TTSA</th>
<th>stocBiO</th>
</tr>
</thead>
<tbody>
<tr>
<td>batch size</td>
<td>$\tilde{O}(1)$</td>
<td>$\tilde{O}(1)$</td>
<td>$\tilde{O}(1)$</td>
<td>$\tilde{O}(\epsilon^{-1})$</td>
</tr>
<tr>
<td># of loops</td>
<td>Single</td>
<td>Double</td>
<td>Single</td>
<td>Double</td>
</tr>
<tr>
<td># of samples (nonconvex)</td>
<td>$\tilde{O}(\epsilon^{-2})$ in $\xi$</td>
<td>$\tilde{O}(\epsilon^{-2})$ in $\phi$</td>
<td>$\tilde{O}(\epsilon^{-2.5})$ in $\xi$</td>
<td>$\tilde{O}(\epsilon^{-2})$ in $\xi$</td>
</tr>
<tr>
<td># of samples (strongly convex)</td>
<td>$\tilde{O}(\epsilon^{-1})$ in $\xi$</td>
<td>$\tilde{O}(\epsilon^{-1})$ in $\phi$</td>
<td>$\tilde{O}(\epsilon^{-1.5})$ in $\xi$</td>
<td>$\tilde{O}(\epsilon^{-1.5})$ in $\phi$</td>
</tr>
<tr>
<td>complexity of $y$ update</td>
<td>$O(d_y^3)$</td>
<td>$O(d_y^3)$</td>
<td>$O(d_y^3)$</td>
<td>$O(d_y^3)$</td>
</tr>
</tbody>
</table>

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 $\epsilon$-stationary point, they only need the sample complexity $O(\epsilon^{-2})$ 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. Due to the nature of two-timescale update, it incurs the sub-optimal sample complexity $O(\epsilon^{-2.5})$. 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 $O(\epsilon^{-1.5})$ 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 (16) 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 $O(\epsilon^{-2})$ 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 $O(\epsilon^{-2})$. 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 Panagiotidis, 2018; Luo et al., 2020; Rafla 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 $\epsilon$-stationary point of (1), STABLE only requires $O(\epsilon^{-2})$ samples in total. In the strongly convex case, to achieve $\epsilon$-optimal solution of (1), STABLE only requires $O(\epsilon^{-1})$ 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 $O(d^3)$ 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 $O(d^2)$ 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 $\ell_2$ 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 $X$ as

$$\min_{x \in \mathbb{R}^d} f(x) := f(x, y^*(x))$$

s.t. $y^*(x) \in \arg \min_{y \in \mathbb{R}^d} g(x, y)$

(2)

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

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

$$\nabla^2_{xy}g(x, y) := \begin{bmatrix}
\frac{\partial^2}{\partial x_1 \partial y_1} g(x, y) & \cdots & \frac{\partial^2}{\partial x_1 \partial y_d} g(x, y) \\
\vdots & \ddots & \vdots \\
\frac{\partial^2}{\partial x_d \partial y_1} g(x, y) & \cdots & \frac{\partial^2}{\partial x_d \partial y_d} g(x, y)
\end{bmatrix}.$$
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 \( \sigma^2_{x_k}, \sigma^2_{y_k}, \sigma^2_{g_{xy}}, \sigma^2_{g_{yy}} \), respectively. Moreover, their moments are bounded by

\[
\mathbb{E}[\|\nabla_x f(x,y;\xi)\|^p] \leq C_{f_x}^p, \quad p = 2, 4
\]
\[
\mathbb{E}[\|\nabla_y f(x,y;\xi)\|^p] \leq C_{f_y}^p, \quad p = 2, 4
\]
\[
\mathbb{E}[\|\nabla^2_{xy} g(x,y;\phi)\|^2] \leq C_{g_{xy}},
\]
\[
\mathbb{E}[\|\nabla^2_{yy} g(x,y;\phi)\|^2] \leq C_{g_{yy}}^2.
\] (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

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

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., 2010). 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 \( F(x) \) 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

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

As we will show in Lemma 5 of Appendix, Assumptions 1-3 ensure that \( VF(\cdot) \), \( \nabla_x f(x,\cdot) \), 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

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

With \( \alpha_k \) and \( \beta_k \) denoting two sequences of stepsizes, the TTSA recursion is given by

\[
y^{k+1} = y^k - \beta_k h^k_y, \quad x^{k+1} = P_X\left(x^k - \alpha_k \left( \nabla_x f(x^k,y^k;\xi^k) - h^k_{xy}(\phi^k)\nabla_{yy}^{-1} \nabla_y f(x^k,y^k;\xi^k)\right)\right) (7a)
\]

where \( \nabla_{yy}^{-1} \) 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 \( O(1/\sqrt{k}) \) 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

1: initialize: $x^0$, $y^0$, $H_{xy}^0$, $H_{yy}^0$, stepsizes $\{\alpha_k, \beta_k\}$.
2: for $k = 0, 1, \ldots, K-1$ do
3: compute $h_{xy}^k(\phi^k)$ and $h_{yy}^k(\phi^k)$
4: update $H_{xy}^k$ via (9a)
5: compute $h_{xy}^k(\phi^k)$ and $h_{yy}^k(\phi^k)$
6: update $H_{yy}^k$ via (9b)
7: compute $\nabla_x f(x^k, y^k; \xi^k)$, $\nabla_y f(x^k, y^k; \xi^k)$
8: > randomly select datum $\xi^k$
9: update $x^k$ and $y^k$ via (8)
10: 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

\[
x^{k+1} = P_X \left( x^k - \alpha_k \nabla_x f(x^k, y^k; \xi^k) - H_{xy}^k (H_{yy}^k)^{-1} \nabla_y f(x^k, y^k; \xi^k) \right)
\]

(8a)

\[
y^{k+1} = y^k - \beta_k h_y^k - (H_{yy}^k)^{-1} H_{xy}^k (x^{k+1} - x^k).
\]

(8b)

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

\[
H_{xy}^k = \mathcal{P} \left( (1 - \tau_k) \left( H_{xy}^{k-1} - h_{xy}^k(\phi^k) \right) + h_{xy}^k(\phi^k) \right)
\]

(9a)

\[
H_{yy}^k = \mathcal{P} \left( (1 - \tau_k) \left( H_{yy}^{k-1} - h_{yy}^k(\phi^k) \right) + h_{yy}^k(\phi^k) \right)
\]

(9b)

where $\mathcal{P}$ is the projection to set $\{ X : \|x\| \leq C_{gsx} \}$ and $\mathcal{P}$ is the projection to set $\{ X : x \succeq \mu_y I \}$.

Compared with (7) and other 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 $\nabla_x f(x^k)$, which is responsible for its improved convergence. Note that we use three stepsizes $\alpha_k, \beta_k$ and $\tau_k$ in (8), we call our method a single-timescale algorithm because the upper- and lower-level variables use the same order of step sizes 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

\[
\dot{x}(t) = -\alpha T(x(t), y(x(t)))
\]

(10)

by choosing an operator $T$. 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 $T(x, y) = \nabla_x f(x, y^*)$ and instead use $y$ to approximate $y^*(x)$. Here note that we have dropped $(t)$ for conciseness. Hence, define the operator as

\[
T(x, y) := \nabla_x f(x, y) - \nabla^2_{xy} g(x, y) \nabla^2_{yy} g(x, y) \nabla_x f(x, y).
\]

(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

\[
(C1) \quad \dot{V} < 0;
\]

\[
(C2) \quad \dot{V} = 0 \quad \text{if and only if} \quad \nabla F(x) = 0 \quad \text{and} \quad y = y^*(x).
\]

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

\[
\dot{x} = -\alpha \nabla_x f(x, y) - \alpha \nabla^2_{xy} g(x, y) \nabla^2_{yy} g(x, y) \nabla_y f(x, y)
\]

\[
\dot{y} = -\beta \nabla_y g(x, y) - [\nabla^2_{yy} g(x, y)]^{-1} \nabla_{yy} g(x, y) \dot{x}
\]

(12)

and choose the constants $\alpha$ and $\beta$ appropriately, then there exists a Lyapunov function $V$ of the $x$- and $y$-dynamics that satisfies (C1) and (C2).
Let us try to mitigate the term \( \| y(x) - y^*(x) \|^2 \) by defining the following new Lyapunov function:

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

which implies that

\[
\dot{V} = -\alpha \langle T(x, y^*(x)), T(x, y) \rangle + \langle y - y^*(x), \dot{y} - \nabla y^*(x) \dot{x} \rangle \\
\leq -\frac{\alpha}{2} \| T(x, y^*(x)) \|^2 + \frac{\alpha L_f^2}{2} \| y - y^*(x) \|^2 \\
+ \langle y - y^*(x), \dot{y} - \nabla y^*(x) \dot{x} \rangle \\
\leq -\frac{\alpha}{2} \| T(x, y^*(x)) \|^2 - \left( \beta - \frac{\alpha L_f^2}{2} \right) \| y - y^*(x) \|^2 \\
+ \langle y - y^*(x), \dot{y} + \beta \dot{y} - y^*(x) \rangle - \nabla y^*(x) \dot{x}
\]

where \( \beta > 0 \) is a fixed constant. The first two terms in the RHS of (16) are non-positive given that \( \alpha \geq 0 \) and \( \beta \geq \alpha L_f^2 / 2 \), 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

\[
\dot{y} \approx -\beta (y - y^*(x)) + \nabla y^*(x) \dot{x}
\]

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))

\[
\nabla_x y(x) := -\left( \nabla^2_{yy} g(x, y) \right)^{-1} \nabla^2_{yy} g(x, y).
\]

These choices lead to the \( y \)-dynamics:

\[
\dot{y} = -\beta \nabla_y g(x, y) + \nabla x y^*(x) \dot{x}.
\]  

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

\[
\langle y - y^*(x), \dot{y} - \nabla y^*(x) \dot{x} \rangle \\
= -\langle y - y^*(x), \beta \nabla y g(x, y) \rangle \\
- \nabla_x y^*(x) \dot{x}.
\]

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

\[
\langle y - y^*(x), \nabla_y g(x, y) - \nabla_y g(x, y^*(x)) \rangle \geq \mu_g \| y - y^*(x) \|^2
\]

where \( \nabla_y g(x, y^*(x)) = 0 \) as \( y^*(x) \) minimizes \( g(x, y) \). Therefore, plugging (21) into (20), we have

\[
\langle y - y^*(x), \dot{y} - \nabla y^*(x) \dot{x} \rangle \\
\leq -\langle y - y^*(x), \nabla_x y^*(x) - \nabla x y^*(x) \dot{x} \rangle - \beta \mu_g \| y - y^*(x) \|^2 \\
\leq \mu \| y - y^*(x) \|^2 - \beta \mu_g \| y - y^*(x) \|^2 \\
\leq \alpha B_x L_y \| y - y^*(x) \|^2 - \beta \mu_g \| y - y^*(x) \|^2
\]

where the second inequality uses the Cauchy-Schwarz inequality, and the last inequality follows the bound...
With the insights gained from the continuous-time Lyapunov function analysis, we can set \( \alpha \) 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

\[
\dot{V} \leq -\frac{\alpha}{2} \| T(x, y^*(x)) \|^2 - \left( \beta \mu_g - \frac{\alpha L_y^2}{2} - \alpha B_x L_y \right) \| y - y^*(x) \|^2.
\]

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

With the insights gained from the continuous-time Lyapunov function analysis, we define the following Lyapunov function

\[
\Psi^k := \Phi_{1/\rho}(x^k) + \| y^k - y^*(x^k) \|^2 + \| H_{yy} - \nabla^2_{yy} g(x^k, y^k) \|^2 + \| H_{xy} - \nabla^2_{xy} g(x^k, y^k) \|^2
\]

which mimics the continuous-time Lyapunov function for the deterministic problem. Similar to the ODE analysis, we need to quantify the difference between two Lyapunov functions \( \Psi^{k+1} - \Psi^k \). We will first analyze the descent of the Moreau Envelope 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

\[
\mathbb{E}[\Phi_{1/\rho}(x^{k+1})] - \mathbb{E}[\Phi_{1/\rho}(x^k)] \\
\leq \frac{(\mu_F + \rho)\rho_k}{4} \| \hat{x}(x^k) - x^k \|^2 + 2\rho_k^2 \left( C_{F_x}^2 + \frac{C_{g_{yy}}^2 C_{F_y}^2}{\mu_g^2} \right) + \frac{2L_y^2 \rho_k}{\mu_F + \rho} \| y^k - y^*(x^k) \|^2 + \frac{4C_{F_x}^2 C_{g_{xy}} \rho_k}{(\mu_F + \rho)\mu_g^2} \mathbb{E}[\| H_{yy}^k - \nabla^2_{yy}(x^k, y^k) \|^2] + \frac{4C_{F_y}^2 \rho_k}{(\mu_F + \rho)\mu_g^2} \mathbb{E}[\| H_{xy}^k - \nabla^2_{xy}(x^k, y^k) \|^2]
\]

where \( L_f, L_F \) are defined in Lemma 5 of Appendix, and \( C_{g_{yy}} \) is the projection radius of \( \Phi_{1/\rho} \).

Lemma 2 implies that the descent of the Moreau Envelope of the upper-level objective functions depends on the error of the lower-level variable \( y^k \), and the estimation errors of \( H_{yy}^k \) and \( H_{xy}^k \). 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, we will utilize the near-stationarity condition (25) as a tool to quantify the convergence of STABLE when \( F(x) \) is non-convex.
and setting $\rho > |\mu_F|$, if we choose the stepsizes as

$$\beta_k \leq \min \left\{ \frac{1}{\sqrt{K}} \frac{\mu_g / L_g}{32(\mu_g + L_g)c} \right\}$$  \hspace{1cm} (28a)

$$\alpha_k \leq \min \left\{ \beta_k, \frac{(c + 4\rho C_{g_s}^2 C_{f_y}^2)/(\mu_g + \rho g^2) - 1}{\sqrt{K}} \right\}$$

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

$$\frac{1}{K} \sum_{k=1}^{K} \mathbb{E} \left[ \|z(x^k) - x^k\|^2 \right] = \mathcal{O}\left(\frac{1}{\sqrt{K}}\right)$$

and

$$\mathbb{E} \left[ \|y^K - y^*(x^K)\|^2 \right] = \mathcal{O}\left(\frac{1}{K}\right)$$  \hspace{1cm} (29)

where $y^*(x^K)$ is the minimizer of the problem $\{1\}$, 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 $\mathcal{O}(K^{-1/2})$. Since each iteration of STABLE only uses two samples (see Algorithm 1), the sample complexity to achieve an $\epsilon$-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 TSST 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 batch-size or 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 $\mu$-strongly convex in $x$, that is, $\nabla_x^2 F(x) \succeq \mu I$.

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 Assump-

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

$$\beta_k \leq \min \left\{ \frac{1}{\sqrt{K}} \frac{\mu_g / L_g}{32(\mu_g + L_g)c} \right\}$$  \hspace{1cm} (30a)

$$\alpha_k \leq \min \left\{ \beta_k, \frac{(c + 4\rho C_{g_s}^2 C_{f_y}^2)/(\mu_g + \rho g^2) - 1}{\sqrt{K}} \right\}$$

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

$$\frac{1}{K} \sum_{k=1}^{K} \mathbb{E} \left[ \|z(x^k) - x^k\|^2 \right] = \mathcal{O}\left(\frac{1}{\sqrt{K}}\right)$$

and

$$\mathbb{E} \left[ \|y^K - y^*(x^K)\|^2 \right] = \mathcal{O}\left(\frac{1}{K}\right)$$  \hspace{1cm} (31)

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

Theorem 3 implies that to achieve an $\epsilon$-optimal solution for both the lower-level and upper-level problems, the sample complexity of STABLE is $\mathcal{O}(\epsilon^{-2})$. 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 TSST’s sample complexity $\mathcal{O}(\epsilon^{-2.5})$ for an $\epsilon$-optimal upper-level solution and $\mathcal{O}(\epsilon^{-1.5})$ for an $\epsilon$-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 TSST 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.
We test all the algorithms in a hyper-parameter optimization task which aims to find the optimal hyperparameter $x \in \mathbb{R}^d$ (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 $\xi$, and $\mathcal{D}_{\text{val}}$ and $\mathcal{D}_{\text{tra}}$ denote, respectively, the training and validation datasets. Specifically, we aim to solve
\[
\min_{x \in \mathbb{R}^d} \mathbb{E}_{\xi \sim \mathcal{D}_{\text{val}}} [\ell(y^*(x); \xi)]
\]
\[
\text{s.t. } y^*(x) = \arg \min_{y \in \mathbb{R}^d} \mathbb{E}_{\phi \sim \mathcal{D}_{\text{tra}}} [\ell(y; \phi)] + \sum_{i=1}^d x_i y_i^2. \quad (32)
\]
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 Reusseelaer-IBM AI Research Collaboration (http://airc.rpi.edu), part of the IBM AI Horizons Network (http://ibm.biz/AIHorizons) and NSF 2134168.

References


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


Tianyi Lin, Chi Jin, and Michael Jordan. On gradient descent ascent for nonconvex-concave minimax probl-
Tianyi Chen, Yuejiao Sun, Quan Xiao, Wotao Yin


A 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

\[
V^k := \Phi_{1/\rho}(x^k) + \|y^k - y^*(x^k)\|^2 + \|H^k_{yy} - \nabla^2_{yy}g(x^k, y^k)\|^2 + \|H^k_{xy} - \nabla^2_{xy}g(x^k, y^k)\|^2
\]  

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

\[
V^{k+1} - V^k = \Phi_{1/\rho}(x^{k+1}) - \Phi_{1/\rho}(x^k) + \|y^{k+1} - y^*(x^{k+1})\|^2 - \|y^k - y^*(x^k)\|^2
\]

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 Envelope 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 \(\beta_k \leq \frac{2}{\mu_y + \bar{L}_y}\), then \(y^{k+1}\) satisfies

\[
\begin{align*}
E\left[\|y^*(x^{k+1}) - y^{k+1}\|^2 | \mathcal{F}^k \right] & \leq \left(1 - \frac{\mu_y L_y \beta_k}{\mu_y + \bar{L}_y} + \frac{\sigma^2_{g_y}}{\sigma^2_{g_y}}\right) \|y^k - y^*(x^k)\|^2 + \left(1 + \frac{\mu_y L_y \beta_k}{\mu_y + \bar{L}_y} + \sigma^2_{g_y}\right) \|y^k - y^*(x^k)\|^2 \\
& + \frac{\sigma^2_{g_y}}{\beta_k} + E\left[\|H^k_{yy} - \nabla^2_{yy}g(x^k, y^k)\|^2 | \mathcal{F}^k \right] \frac{\sigma^2_{g_y}}{\beta_k} + E\left[\|H^k_{xy} - \nabla^2_{xy}g(x^k, y^k)\|^2 | \mathcal{F}^k \right] \frac{\sigma^2_{g_y}}{\beta_k}.
\end{align*}
\]  

(35)

Roughly speaking, Lemma 3 implies that if the stepsizes \(\alpha_k^2\) and \(\beta_k^2\) and the estimation errors of \(H^k_{yy}\) and \(H^k_{xy}\) 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 \(H^k_{yy}\) and \(H^k_{xy}\), we will next build upon the results in (Chen et al., 2020) (Lemma 2) to analyze the estimation errors.

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

\[
E\left[\|H^k_{xy} - \nabla^2_{xy}g(x^k, y^k)\|^2 | \mathcal{F}^k \right] \leq (1 - \tau_k)^2 \left\|H^k_{xy} - \nabla^2_{xy}g(x^{k-1}, y^{k-1})\right\|^2 + 2\sigma^2_{g_y} \|x^k - x^{k-1}\|^2 + 2(1 - \tau_k)^2 (\bar{L}^2_{g_{xy}} + L^2_{g_{xy}}) \|y^k - y^{k-1}\|^2
\]

\[
+ 2(1 - \tau_k)^2 (\bar{L}^2_{g_{xy}} + L^2_{g_{xy}}) \|x^k - x^{k-1}\|^2
\]

(36)
where the constants \(L_{g_{xy}}, L_{g_{yy}}, \bar{L}_{g_{xy}}, \bar{L}_{g_{yy}}, \bar{\sigma}_{g_{xy}}, \bar{\sigma}_{g_{yy}}\) are defined in Assumptions 1 and 3. And likewise, the mean square error of \(H_{xy}^k\) satisfies
\[
\mathbb{E}_k \left[ \left\| H_{xy}^k - \nabla^2_{yy} g(x^k, y^k) \right\|_2^2 \right] \leq (1 - \tau_k)^2 \left\| H_{yy}^{k-1} - \nabla^2_{yy} g(x^{k-1}, y^{k-1}) \right\|_2^2 + 2\tau_k^2 \sigma^2_{g_{yy}} + 2(1 - \tau_k)^2 (\bar{L}_{g_{yy}}^2 + \bar{L}_{g_{yy}}^2) \| x^k - x^{k-1} \|_2^2 + 2(1 - \tau_k)^2 (\bar{L}_{g_{yy}}^2 + \bar{L}_{g_{yy}}^2) \| y^k - y^{k-1} \|_2^2.
\]

Intuitively, the update of \(x^k\) is bounded and so is the update of \(y^k\), and thus \(\| x^k - x^{k-1} \|_2^2 = \mathcal{O}(\alpha^2_k - 1)\) and \(\| y^k - y^{k-1} \|_2^2 = \mathcal{O}(\beta^2_k - 1)\). Plugging them into the RHS of Lemma 4, it suggests that if the stepsizes \(\alpha_k, \beta_k, \tau_k\) 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
\[
\mathbb{E}_k [\| y^{k+1} \| - \mathbb{E}_k [\| y^k \|] - c_1 \mathbb{E}_k [\| y^k - y^* (x^k) \|] - c_2 \mathbb{E}_k [\| \hat{x} - x^k \| - x^k] + c_3
\]
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
\[
\| \nabla_x f(x, y^*(x)) - \nabla_x f(x, y) \| \leq L_f \| y^*(x) - y \| \tag{39a}
\]
\[
\| \nabla F(x_1) - \nabla F(x_2) \| \leq L_F \| x_1 - x_2 \| \tag{39b}
\]
\[
\| y^*(x_1) - y^*(x_2) \| \leq L_y \| x_1 - x_2 \| \tag{39c}
\]

and the constants \(L_f, L_y, L_F\) are defined as
\[
L_f := L_{f x} + \frac{C_{g_{xy}} L_{f y}}{\mu_y} + \frac{C_{f_y}}{\mu_y} \left( L_{f_{xy}} + \frac{C_{g_{xy}} L_{g_{yy}}}{\mu_g} \right), \quad L_y := \frac{C_{g_{xy}}}{\mu_y}
\]
\[
L_F := L_{f x} + \frac{C_{g_{xy}} (L_{f y} + L_f)}{\mu_y} + \frac{C_{f_y}}{\mu_y} \left( L_{f_{xy}} + \frac{C_{g_{xy}} L_{g_{yy}}}{\mu_g} \right)
\]
where the constants are defined in Assumptions 1–3.

### C Proof of Proposition 1

**Proof:** Define the Jacobian matrix
\[
\nabla_{xy}(x) = \begin{bmatrix}
\frac{\partial}{\partial x_1} y_1(x) & \cdots & \frac{\partial}{\partial x_d} y_1(x) \\
\vdots & \ddots & \vdots \\
\frac{\partial}{\partial x_1} y_d(x) & \cdots & \frac{\partial}{\partial x_d} y_d(x)
\end{bmatrix}.
\]

By the chain rule, it follows that
\[
\nabla F(x) := \nabla_x f(x, y^*(x)) + \nabla_{xy} y^*(x)^\top \nabla_y f(x, y^*(x)).
\]

The minimizer \(y^*(x)\) satisfies
\[
\nabla_y g(x, y^*(x)) = 0, \quad \text{thus} \quad \nabla_x (\nabla_y g(x, y^*(x))) = 0.
\]

By the chain rule again, it follows that
\[
\nabla^2_{xy} g(x, y^*(x)) + \nabla_{xy} y^*(x)^\top \nabla^2_{yy} g(x, y^*(x)) = 0.
\]

By Assumption 2, \(\nabla^2_{yy} g(x, y^*(x))\) is invertible, so
\[
\nabla_{xy} y^*(x)^\top := -\nabla^2_{xy} g(x, y^*(x)) \left[ \nabla^2_{yy} g(x, y^*(x)) \right]^{-1}.
\]

By substituting (42) into (40), we arrive at (44).
D Proof of Lemma 2
Proof: Now we turn to analyze the update of $x$. For convenience, we define the update in (8a) as
\[ x^{k+1} = \mathcal{P}_\mathcal{X} (x^k - \alpha_k \hat{h}_f^k) \quad \text{with} \quad \hat{h}_f^k := \nabla_x f(x^k, y^k; \xi^k) - H_{xy}^k (H_{yy}^k)^{-1} \nabla_y f(x^k, y^k; \xi^k) \] (43)
and let $\hat{x}^k$ and $\hat{x}$ denote $\hat{x}(x^k)$ and $\hat{x}(x)$.
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
\[ F(x) + \frac{\rho}{2} \|x - \hat{x}\|^2 - F(\hat{x}) - \frac{\rho}{2} \|\hat{x} - x\|^2 = F(x) - F(\hat{x}) - \frac{\rho}{2} \|\hat{x} - x\|^2 \geq 0. \]
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 $\mu_F$ such that $\mu_F + \rho > 0$ and using the definition of Moreau Envelop, we have that
\[
\Phi_{1/\rho}(x^{k+1}) = F(\hat{x}^{k+1}) + \frac{\rho}{2} \|x^{k+1} - \hat{x}^{k+1}\|^2 \leq F(\hat{x}^k) + \frac{\rho}{2} \|x^k - \hat{x}^k\|^2 \\
= F(\hat{x}^k) + \frac{\rho}{2} \|x^k - \hat{x}^k\|^2 + \frac{\rho}{2} \|x^{k+1} - x^k\|^2 + \rho \langle x^k - \hat{x}^k, x^{k+1} - x^k \rangle \\
= \Phi_{1/\rho}(x^k) + \frac{\rho}{2} \|x^{k+1} - x^k\|^2 + \rho \langle x^k - \hat{x}^k, x^{k+1} - x^k \rangle - \rho \|x^k - x^{k+1}\|^2 \\
\leq \Phi_{1/\rho}(x^k) + \rho \alpha_k \langle \hat{x}^k - x^k, \hat{h}_f^k \rangle + \rho \alpha_k \langle \hat{h}_f^k, x^k - x^{k+1} \rangle \] (44)
where the fourth inequality is due to $\langle x^k - \alpha_k \hat{h}_f^k - x^{k+1}, \hat{x}^k - x^{k+1} \rangle \leq 0$ using the definition of $\mathcal{P}_\mathcal{X}$. Then taking the conditional expectation of both sides in (44), we have that
\[
\mathbb{E} \left[ \Phi_{1/\rho}(x^{k+1}) | \mathcal{F}^k \right] \leq \Phi_{1/\rho}(x^k) + 2 \alpha_k \mathbb{E} \left[ \|\hat{h}_f^k\|^2 | \mathcal{F}^k \right] + \mathbb{E} \left[ \langle \hat{x}^k - x^k, \hat{h}_f^k \rangle | \mathcal{F}^k \right] + \mathbb{E} \left[ \langle \hat{x}^k - x^k, \nabla F(x^k) \rangle | \mathcal{F}^k \right] \\
\leq \mathbb{E} \left[ \langle \hat{x}^k - x^k, \hat{h}_f^k \rangle | \mathcal{F}^k \right] + \mathbb{E} \left[ \|\nabla F(x^k)\|^2 | \mathcal{F}^k \right] \] (45)
where the second inequality comes from (61). Then we bound the third term in (45) and get that
\[
\mathbb{E} \left[ \langle \hat{x}^k - x^k, \hat{h}_f^k \rangle | \mathcal{F}^k \right] \leq \mathbb{E} \left[ \|\hat{x}^k - x^k, \hat{h}_f^k \| \nabla_x f(x^k, y^k) + \nabla_x f(x^k, y^k) - \nabla F(x^k) + \nabla F(x^k) \| \mathcal{F}^k \right] \\
\leq \mathbb{E} \left[ \|\hat{x}^k - x^k \| \nabla_x f(x^k, y^k) \| \mathcal{F}^k \right] + \mathbb{E} \left[ \|\hat{x}^k - x^k \| \nabla_x f(x^k, y^k) \| \mathcal{F}^k \right] \\
\leq \frac{\gamma_k}{4} \|\hat{x}^k - x^k\|^2 + \frac{1}{\gamma_k} \|\nabla_x f(x^k, y^k)\|^2 \mathbb{E} \left[ \|H_{yy}^k\| - H_{yy}^k \| \mathcal{F}^k \right] \\
+ \frac{\gamma_k}{4} \|\hat{x}^k - x^k\|^2 + \frac{1}{\gamma_k} \|\nabla_x f(x^k, y^k)\|^2 \mathbb{E} \left[ \|H_{yy}^k\| - H_{yy}^k \| \mathcal{F}^k \right] \\
\leq \frac{\gamma_k}{4} \|\hat{x}^k - x^k\|^2 + \frac{1}{\gamma_k} \|\nabla_x f(x^k, y^k)\|^2 \mathbb{E} \left[ \|H_{yy}^k\| - H_{yy}^k \| \mathcal{F}^k \right] \\
\leq \frac{L_2^2}{\gamma_k} \|y^g - y^g(x^k)\|^2 - \frac{\mu_F + \rho}{2} \|\hat{x}^k - x^k\|^2 \\
\] where the third inequality uses Young’s inequality with parameter $\gamma_k$, (61) in (Hong et al., 2020) and the fact that
\[
\mathbb{E} \left[ \hat{h}_f^k \|\mathcal{F}^k \right] = \nabla_x f(x^k, y^k) - (H_{yy}^k)^{-1} H_{xy}^k \nabla_y f(x^k, y^k); \] (46)
We first bound the first approximation error in the RHS of (50) by

\[ \mathbb{E} \left[ (\tilde{x}^k - x^k, \tilde{h}_g^k) | \mathcal{F}_k \right] \leq -\frac{\mu_L + \rho}{4} \| \tilde{x}^k - x^k \|^2 + 4C_L^2 C_{\alpha \
\eta}^2 \mathbb{E} \| H_{yy}^k - H_{yy}(x^k, y^k) \|^2 | \mathcal{F}_k \] 

+ \frac{4C_L^2}{(\mu_L + \rho) \mu_g^2} \mathbb{E} \| H_{xy}^k - H_{xy}(x^k, y^k) \|^2 | \mathcal{F}_k \] 

+ \frac{2L_f^2}{\mu_L + \rho} \| y^k - y^*(x^k) \|^2 \] (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

\[ \mathbb{E} \left[ \| y^{k+1} - y^*(x^{k+1}) \|^2 | \mathcal{F}_k \right] \]

\[ = \mathbb{E} \left[ \| y^k - \beta_k h_g^k - y^*(x^k) - y^*(x^{k+1}) - (H_{yy}^k)^{-1}(H_{xy}^k)^\top (x^{k+1} - x^k) \|^2 | \mathcal{F}_k \right] \]

\[ \leq (1 + \varepsilon) \mathbb{E} \| y^k - \beta_k h_g^k - y^*(x^k) \|^2 | \mathcal{F}_k \]

\[ + (1 + \varepsilon^{-1}) \mathbb{E} \| y^*(x^k) - y^*(x^{k+1}) - (H_{yy}^k)^{-1}(H_{xy}^k)^\top (x^{k+1} - x^k) \|^2 | \mathcal{F}_k \]. (48)

The upper bound of \( I_1 \) can be derived as

\[ I_1 = \| y^k - y^*(x^k) \|^2 - 2\beta_k \mathbb{E} [(y^k - y^*(x^k), h_g^k) | \mathcal{F}_k] + \beta_k^2 \mathbb{E} \| h_g^k \|^2 | \mathcal{F}_k \]

\[ \leq (a) \| y^k - y^*(x^k) \|^2 - 2\beta_k (y^k - y^*(x^k) - \nabla_y g(x^k, y^k)) + \beta_k^2 \| \nabla_y g(x^k, y^k) \|^2 + \beta_k^2 \sigma_g^2 \]

\[ \leq (b) \left( 1 - \frac{2\mu_g L_g}{\mu_g + L_g} \beta_k \right) \| y^k - y^*(x^k) \|^2 + \beta_k \left( \beta_k - \frac{2}{\mu_g + L_g} \right) \| \nabla_y g(x^k, y^k) \|^2 + \beta_k^2 \sigma_g^2 \]

\[ \leq (c) \left( 1 - \frac{2\mu_g L_g}{\mu_g + L_g} \beta_k \right) \| y^k - y^*(x^k) \|^2 + \beta_k \sigma_g^2 \] (49)

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

The upper bound of \( I_2 \) can be derived as

\[ I_2 = \mathbb{E} \left[ \| y^*(x^k) - y^*(x^{k+1}) - (H_{yy}^k)^{-1}(H_{xy}^k)^\top (x^{k+1} - x^k) \|^2 | \mathcal{F}_k \right] \]

\[ \leq 3\mathbb{E} \left[ \| y^*(x^{k+1}) - y^*(x^k) - \nabla_x y^*(x^k)(x^{k+1} - x^k) \|^2 | \mathcal{F}_k \right] \]

\[ + 3\mathbb{E} \left[ \| \nabla_y y^*(x^k) - H_{yy}(x^k, y^k)^{-1} H_{xy}(x^k, y^k)^\top (x^{k+1} - x^k) \|^2 | \mathcal{F}_k \right] \]

\[ + 3\mathbb{E} \left[ \| (H_{yy}(x^k, y^k)^{-1} H_{xy}(x^k, y^k)^\top - (H_{yy}^k)^{-1}(H_{xy}^k)^\top) (x^{k+1} - x^k) \|^2 | \mathcal{F}_k \right]. \] (50)

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

\[ \| y^*(x^{k+1}) - y^*(x^k) - \nabla_x y^*(x^k)(x^{k+1} - x^k) \|^2 \]

\[ = \left\| \int_0^1 \nabla_x y^*(x^k + t(x^{k+1} - x^k))(x^{k+1} - x^k) dt - \nabla_x y^*(x^k)(x^{k+1} - x^k) \right\|^2 \]

\[ \leq \int_0^1 \| \nabla_x y^*(x^k + t(x^{k+1} - x^k)) - \nabla_x y^*(x^k) \|^2 \| x^{k+1} - x^k \| dt \leq \frac{I^2}{2} \| x^{k+1} - x^k \|^4 \] (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
\[ \mathbb{E} \left[ \left\| \nabla_x y^*(x^k) - H_{yy}(x^k, y^k)^{-1} H_{xy}(x^k, y^k) \right\| \right] \leq \mathbb{E} \left[ \left\| \nabla_x y^*(x^k) - H_{yy}(x^k, y^k)^{-1} H_{xy}(x^k, y^k) \right\|^2 \right] \leq 2 \mathbb{E} \left[ \left\| \nabla_x y^*(x^k) - H_{yy}(x^k, y^k)^{-1} H_{xy}(x^k, y^k) \right\|^2 \right] \leq 2 C_{g_{xy}}^2 \mathbb{E} \left[ \left\| H_{yy}(x^k, y^k)^{-1} H_{xy}(x^k, y^k) \right\|^2 \right] + \frac{2}{\mu_g} \mathbb{E} \left[ \left\| H_{yy}(x^k, y^k) - H_{xy}(x^k, y^k) \right\|^2 \right]. \]
where the second inequality follows from $\|H^k_{xy}\| \leq C_{g_{xy}}$ and $H^k_{yy} \leq \mu_g I$.

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

$$I_2 \leq \frac{3L_y^2}{2} E[\|x^{k+1} - x^k\|^4 |\mathcal{F}^k] + \frac{6C^2_{g_{xy}}}{\mu_g^2} \mu_g \|H_{yy}(x^k, y^*(x^k)) - H_{yy}(x^k, y^k)\|^2 E[\|x^{k+1} - x^k\|^2 |\mathcal{F}^k]
+ \frac{6}{\mu_g} E[\|H_{xy}(x^k, y^*(x^k)) - H_{xy}(x^k, y^k)\|^2 E[\|x^{k+1} - x^k\|^2 |\mathcal{F}^k]
+ \frac{6C^2_{g_{yx}}}{\mu_g^2} E[\|H^k_{xy} - H_{xy}(x^k, y^k)\|^2 E[\|x^{k+1} - x^k\|^2 |\mathcal{F}^k]
+ \frac{6}{\mu_g} E[\|H^k_{xy} - H_{xy}(x^k, y^k)\|^2 E[\|x^{k+1} - x^k\|^2 |\mathcal{F}^k]].$$

(58)

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

$$I_2 \leq \frac{3L_y^2}{2} E[\|x^{k+1} - x^k\|^4 |\mathcal{F}^k] + \frac{2}{\mu_g} \left( C_{g_{xy}} L_{g_{yy}} + L_{g_{yx}} \right) \|y^k - y^*(x^k)\|^2 E[\|x^{k+1} - x^k\|^2 |\mathcal{F}^k]
+ \frac{2C_{g_{yx}}}{\mu_g} E[\|H^k_{xy} - H_{xy}(x^k, y^k)\|^2 E[\|x^{k+1} - x^k\|^2 |\mathcal{F}^k]
+ \frac{2}{\mu_g} E[\|H^k_{xy} - H_{xy}(x^k, y^k)\|^2 E[\|x^{k+1} - x^k\|^2 |\mathcal{F}^k]].$$

(59)

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

$$\|\tilde{h}_f^k\| = \left\| \nabla_x f(x^k, y^k; \xi^k) - (H^k_{yy})^{-1}H^k_{xy} \nabla_y f(x^k, y^k; \xi^k) \right\|
\leq \|\nabla_x f(x^k, y^k; \xi^k)\| + \|(H^k_{yy})^{-1}H^k_{xy} \nabla_y f(x^k, y^k; \xi^k)\|
\leq \|\nabla_x f(x^k, y^k; \xi^k)\| + \frac{C_{g_{xy}}}{\mu_g} \|\nabla_y f(x^k, y^k; \xi^k)\|$$

(60)

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

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

$$E[\|\tilde{h}_f^k\|^p |\mathcal{F}^k, H^k_{xy}, H^k_{yy}] \leq 2^{p-1} E \left[ \|\nabla_x f(x^k, y^k; \xi^k)\|^p |\mathcal{F}^k, H^k_{xy}, H^k_{yy} \right]
+ 2^{p-1} \left( \frac{C_{g_{xy}}}{\mu_g} \right)^p E \left[ \|\nabla_y f(x^k, y^k; \xi^k)\|^p |\mathcal{F}^k, H^k_{xy}, H^k_{yy} \right]
\leq 2^{p-1} \left( C_{f_x}^p + \left( \frac{C_{g_{xy}}}{\mu_g} \right)^p C_{f_y}^p \right)$$

(61)

where the last inequality from Assumption 3. And thus

$$E \left[ \|x^{k+1} - x^k\|^p |\mathcal{F}^k, H^k_{xy}, H^k_{yy} \right] \leq 2^{p-1} \left( C_{f_x}^p + \left( \frac{C_{g_{xy}}}{\mu_g} \right)^p C_{f_y}^p \right) \alpha_k^p.$$
Plugging (62) into (59), we have

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

Now let us define the constants as

\[
\hat{c}_1 := \max \left\{ 12L_y^2 \left( C_{f_x}^4 + \left( \frac{C_{g_{yx}}}{\mu_y} \right)^4 C_{f_y}^4 \right), \frac{12}{\mu_g} \left( \frac{C_{g_{yx}}^2 L_{g_{yy}}}{\mu_g^2} + L_{g_{yy}} \right) \left( C_{f_x}^2 + \left( \frac{C_{g_{yx}}}{\mu_y} \right)^2 C_{f_y}^2 \right) \right\} \\
\hat{c}_2 := \frac{2}{\mu_g + L_g} + \frac{\mu_g + L_g}{\mu_g L_g}, \quad c := \hat{c}_1 \hat{c}_2.
\]

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

\[
\mathbb{E} \left[ \|y^{k+1} - y^*(x^{k+1})\|^2 |F^k\right] \\
\leq \left( 1 - \frac{\mu_g L_g}{\mu_g + L_g} \beta_k \right) \|y^k - y^*(x^k)\|^2 + \left( 1 + \frac{\mu_g L_g}{\mu_g + L_g} \beta_k \right) \beta_k \sigma_g^2 + \hat{c}_1 \hat{c}_2 \frac{\alpha_k}{\beta_k} \\
+ \hat{c}_1 \hat{c}_2 \mathbb{E} \left[ \|H_{yy}^k - H_{yy}(x^k, y^k)\|^2 |F^k\right] \frac{\alpha_k^2}{\beta_k} \\
+ \hat{c}_1 \hat{c}_2 \mathbb{E} \left[ \|H_{xy}^k - H_{xy}(x^k, y^k)\|^2 |F^k\right] \frac{\alpha_k^2}{\beta_k} \\
\] (64)

where we have used the fact that

\[
\left( 1 + \frac{\mu_g L_g}{\mu_g + L_g} \beta_k \right) \left( 1 - \frac{2\mu_g L_g}{\mu_g + L_g} \beta_k \right) \leq 1 - \frac{\mu_g L_g}{\mu_g + L_g} \beta_k \\
\left( 1 + \frac{\mu_g L_g}{\mu_g + L_g} \beta_k \right)^{-1} \leq \frac{1}{\beta_k} \left( \frac{2}{\mu_g + L_g} + \frac{\mu_g + L_g}{\mu_g L_g} \right) = \frac{\hat{c}_2}{\beta_k}
\]

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

\section*{F Proof of Lemma 4}

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

\[
h_{yy}^k(\phi) := \nabla^2_{yy} g(x^k, y^k; \phi), \quad h_{xy}^k(\phi) := \nabla^2_{xy} g(x^k, y^k; \phi).
\] (65)

For notational brevity in the analysis, we define

\[
H_{xy}(x, y) := \nabla^2_{xy} g(x, y), \quad H_{yy}(x, y) := \nabla^2_{yy} g(x, y).
\] (66)
and rewrite the update of (9) as

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

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

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

\[
E\left[ \| H_{xy}^k - H_{xy}(x^k, y^k) \|^2 | \mathcal{F}_k \right] \leq E\left[ \| \hat{H}_{xy}^k - H_{xy}(x^k, y^k) \|^2 | \mathcal{F}_k \right]
= \left\| E\left[ \hat{H}_{xy}^k - H_{xy}(x^k, y^k) | \mathcal{F}_k \right] \right\|^2 + \sum_{i,j} \text{Var} \left[ (\hat{H}_{xy}^k - H_{xy}(x^k, y^k))_{i,j} | \mathcal{F}_k \right] \quad (68)
\]

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

We first analyze the bias term in (68) by

\[
E\left[ \hat{H}_{xy}^k - H_{xy}(x^k, y^k) | \mathcal{F}_k \right]
\]

\[
E\left[ (1 - \tau_k) \left( H_{xy}^{k-1} + h_{xy}^{k-1}(\phi^k) - h_{xy}^{k-1}(\phi^k) \right) + \tau_k h_{xy}^k(\phi^k) - H_{xy}(x^k, y^k) | \mathcal{F}_k \right]
= (1 - \tau_k) \left( H_{xy}^{k-1} + H_{xy}(x^k, y^k) - H_{xy}(x^{k-1}, y^{k-1}) \right) + \tau_k H_{xy}(x^k, y^k) - H_{xy}(x^k, y^k)
= (1 - \tau_k) \left( H_{xy}^{k-1} - H_{xy}(x^{k-1}, y^{k-1}) \right). \quad (69)
\]

The variance term in (68) follows

\[
\sum_{i,j} \text{Var} \left[ (\hat{H}_{xy}^k - H_{xy}(x^k, y^k))_{i,j} | \mathcal{F}_k \right] = \sum_{i,j} \text{Var} \left[ (\hat{H}_{xy}^k)_{i,j} | \mathcal{F}_k \right]
\]

\[
(67a) \sum_{i,j} \text{Var} \left[ (1 - \tau_k)(h_{xy}^k(\phi^k) - h_{xy}^{k-1}(\phi^k))_{i,j} + \tau_k (h_{xy}^k(\phi^k))_{i,j} | \mathcal{F}_k \right]
\leq 2(1 - \tau_k)^2 \sum_{i,j} \text{Var} \left[ (h_{xy}^k(\phi^k) - h_{xy}^{k-1}(\phi^k))_{i,j} | \mathcal{F}_k \right] + 2\tau_k^2 \sum_{i,j} \text{Var} \left[ (h_{xy}^k(\phi^k))_{i,j} | \mathcal{F}_k \right]
\]

\[
(\text{a}) \leq 2(1 - \tau_k)^2 \mathbb{E} \left[ \| h_{xy}^k(\phi^k) - h_{xy}^{k-1}(\phi^k) \|^2 | \mathcal{F}_k \right] + 2\tau_k^2 \sum_{i,j} \text{Var} \left[ (h_{xy}^k(\phi^k))_{i,j} | \mathcal{F}_k \right]
\]

\[
(\text{b}) \leq 2(1 - \tau_k)^2 \left( \frac{L_{gxy}^2}{\sigma_{gxy}^2} + 2 \| x^k - x^{k-1} \|^2 + \| y^k - y^{k-1} \|^2 \right) + 2\tau_k^2 \sigma_{gxy}^2 \quad (70)
\]

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

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

\[
\mathbb{E}[\| H_{xy}^k - H_{xy}(x^k, y^k) \|^2 | \mathcal{F}_k] \leq (1 - \tau_k)^2 \| H_{xy}^{k-1} - H_{xy}(x^{k-1}, y^{k-1}) \|^2 + 2\tau_k^2 \sigma_{gxy}^2
+ 2(1 - \tau_k)^2 \left( \frac{L_{gxy}^2}{\sigma_{gxy}^2} + 2 \| x^k - x^{k-1} \|^2 + \| y^k - y^{k-1} \|^2 \right).
\]

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

\[
\mathbb{E}[\| H_{yy}^k - H_{yy}(x^k, y^k) \|^2 | \mathcal{F}_k] \leq (1 - \tau_k)^2 \| H_{yy}^{k-1} - H_{yy}(x^{k-1}, y^{k-1}) \|^2 + 2\tau_k^2 \sigma_{gxy}^2
+ 2(1 - \tau_k)^2 \left( \frac{L_{gxy}^2}{\sigma_{gxy}^2} + 2 \| x^k - x^{k-1} \|^2 + \| y^k - y^{k-1} \|^2 \right).
\]

The proof is then complete.
Proof: Using Lemmas 2 and 4, we respectively bound the four difference terms in (31) and obtain

\[
\begin{align*}
\mathbb{E}[\bar{V}^{k+1}] - \mathbb{E}[\bar{V}^k] &\leq -\frac{(\mu + \rho)\alpha_k}{4} \mathbb{E}[\|\tilde{x}(x^k) - x^k\|^2] - \left(\frac{\mu_g L_g}{\mu_g + L_g} \beta_k - \tilde{c}_1 \tilde{c}_2 \frac{\alpha_k^2}{\beta_k} - \frac{2L_f^2 \rho \alpha_k}{\mu + \rho}\right) \mathbb{E}[\|y^k - y^*(x^k)\|^2] \\
&\quad - \left(\tau_k + 1 - \tilde{c}_1 \tilde{c}_2 \frac{\alpha_k^2}{\beta_k} - \frac{4\rho \alpha_k C_g^2}{(\mu + \rho)\mu_g^2}\right) \mathbb{E}[\|H_{xy}^k - H_{xy}(x, y^k)\|^2] \\
&\quad + 4(1 - \tau_k) \tilde{c}_3 \alpha_k^2 \left(C_{f_x}^2 + \frac{C_{g_{xy}}}{\mu_g} \right)^2 \mathbb{E}[\|H_{xy}^k - H_{xy}(x, y^k)\|^2] \\
&\quad + 2\rho \tilde{c}_2 \left(C_{f_x}^2 + \frac{C_{g_{xy}}}{\mu_g} \right)^2 \mathbb{E}[\|H_{xy}^k - H_{xy}(x, y^k)\|^2] \\
&\quad + 4(1 - \tau_k) \tilde{c}_3 \alpha_k^2 \left(C_{f_x}^2 + \frac{C_{g_{xy}}}{\mu_g} \right)^2 \mathbb{E}[\|H_{xy}^k - H_{xy}(x, y^k)\|^2] \\
&\quad + 2\rho \tilde{c}_2 \left(C_{f_x}^2 + \frac{C_{g_{xy}}}{\mu_g} \right)^2 \mathbb{E}[\|H_{xy}^k - H_{xy}(x, y^k)\|^2] \\
&\quad + 4(1 - \tau_k) \tilde{c}_3 \alpha_k^2 \left(C_{f_x}^2 + \frac{C_{g_{xy}}}{\mu_g} \right)^2 \mathbb{E}[\|H_{xy}^k - H_{xy}(x, y^k)\|^2].
\end{align*}
\]

where the constant is defined as \( \tilde{c}_3 := \tilde{L}_{g_{xy}}^2 + \tilde{L}_{g_{zy}}^2 + \tilde{L}_{g_{zy}}^2 + L_{g_{zy}}^2 \).

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

\[
\mathbb{E}[\|y^{k+1} - y^k\|^2] = \mathbb{E}[\|\beta h_k - (H_{xy}^{k+1})^{-1} H_{xy}^k (x^{k+1} - x^k)\|^2]
\]

\[
\leq \frac{\beta_k^2 \mathbb{E}[\|h_k\|^2]}{2} + 2 \mathbb{E}[\|H_{xy}^k\|^{-2} \|H_{xy}^k\|^2 \|x^{k+1} - x^k\|^2]
\]

\[
\leq 2\beta_k^2 \mathbb{E}[\|\nabla_y g(x^k, y^k)\|^2] + 2\beta_k^2 \sigma_{g_{yx}}^2 + 2 \mathbb{E}[\|H_{xy}^k\|^{-2} \|H_{xy}^k\|^2 \|x^{k+1} - x^k\|^2]
\]

\[
\leq 2\beta_k^2 \mathbb{E}[\|\nabla_y g(x^k, y^k)\|^2] + 2\beta_k^2 \sigma_{g_{yx}}^2 + 2 \mathbb{E}[\|x^{k+1} - x^k\|^2]
\]

\[
\leq 4\beta_k^2 L_y^2 \mathbb{E}[\|y^k - y^*(x^k)\|^2] + 2\beta_k^2 \sigma_{g_{yx}}^2 + 2 \left(C_{g_{xy}}^{\mu_g} \right)^2 \mathbb{E}[\|x^{k+1} - x^k\|^2]
\]

where (a) follows from \( \mathbb{E}[X^2] = \mathbb{E}[X] + \mathbb{E}[X^2] \) and Assumption 3, (b) uses the upper and lower projections of \( H_{xy}^k \) and \( H_{xy}^k \) in (31), 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 (42) to bound \( \mathbb{E}[\|x^{k+1} - x^k\|^2] \) and using (71)–(72), we have

\[
\begin{align*}
\mathbb{E}[\bar{V}^{k+1}] - \mathbb{E}[\bar{V}^k] &\leq -\frac{(\mu + \rho)\alpha_k}{4} \mathbb{E}[\|\tilde{x}(x^k) - x^k\|^2] + \alpha_k^2 \left(2 \rho + 4\tilde{c}_3 + 8\tilde{c}_3 \frac{C_{g_{xy}}}{\mu_g} \right)^2 \left(C_{f_x}^2 + \frac{C_{g_{xy}}}{\mu_g} \right)^2 \\
&\quad - \left(\frac{\mu_g L_g}{\mu_g + L_g} \beta_k - \tilde{c}_1 \tilde{c}_2 \frac{\alpha_k^2}{\beta_k} - \frac{2L_f^2 \rho \alpha_k}{\mu + \rho} - 8\tilde{c}_3 \beta_k^2 L_g^2\right) \mathbb{E}[\|y^k - y^*(x^k)\|^2] \\
&\quad - \frac{1}{\sqrt{K}} \tilde{c}_1 \tilde{c}_2 \frac{\alpha_k^2}{\beta_k} - \frac{4\rho C_{g_{xy}}^2 C_{f_x}^2 \alpha_k}{(\mu + \rho)\mu_g^4} \mathbb{E}[\|H_{xy}^k - H_{xy}(x, y^k)\|^2] \\
&\quad - \frac{1}{\sqrt{K}} \tilde{c}_1 \tilde{c}_2 \frac{\alpha_k^2}{\beta_k} - \frac{4\rho C_{f_x}^2 \alpha_k}{(\mu + \rho)\mu_g^2} \mathbb{E}[\|H_{xy}^k - H_{xy}(x, y^k)\|^2] \\
&\quad + \left(1 + \frac{\mu_g L_g}{\mu_g + L_g} \beta_k\right) \beta_k^2 \sigma_{g_{yx}}^2 + \tilde{c}_1 \tilde{c}_2 \frac{\alpha_k^4}{\beta_k} + \frac{4\sigma_{g_{yx}}^2}{\rho} + 4\tilde{c}_3 \beta_k^2 \sigma_{g_{yx}}^2.
\end{align*}
\]
Choosing the stepsize $\alpha_k$ as (28), it will lead to (cf. $c := \tilde{c}_1\tilde{c}_2$)

\[
\frac{1}{\sqrt{K}} - \tilde{c}_1\tilde{c}_2\frac{\alpha_k^2}{\beta_k} - 4\rho C_{g,x}^2 C_{f,x}^2 \frac{\alpha_k}{(\mu_F + \rho)\mu_g^4} \geq \frac{1}{\sqrt{K}} - \tilde{c}_1\tilde{c}_2\alpha_k - \frac{4\rho C_{g,x}^2 C_{f,x}^2 \alpha_k}{(\mu_F + \rho)\mu_g^4} \geq 0 \tag{74a}
\]

\[
\frac{1}{\sqrt{K}} - \tilde{c}_1\tilde{c}_2\frac{\alpha_k^2}{\beta_k} - 4\rho C_{f,x}^2 \alpha_k \geq \frac{1}{\sqrt{K}} - \tilde{c}_1\tilde{c}_2\alpha_k - \frac{4\rho C_{f,x}^2 \alpha_k}{(\mu_F + \rho)\mu_g^4} \geq 0 \tag{74b}
\]

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 $\beta_k$ as (28) will lead to

\[
\frac{\mu_g L_g - \beta_k}{\mu_g + L_g} - \tilde{c}_1\tilde{c}_2\frac{\alpha_k^2}{\beta_k} - \frac{2L_l^2 \rho \alpha_k}{\mu_F + \rho} - 8\tilde{c}_3\beta_k^2 L_g \geq \frac{\mu_g L_g - \beta_k}{\mu_g + L_g} - \tilde{c}_1\tilde{c}_2\alpha_k - \frac{2L_l^2 \rho}{\mu_F + \rho} \alpha_k - 8\tilde{c}_3\beta_k^2 L_g \tag{74c}
\]

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

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

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

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

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

\[
\mathbb{E}[[x^{k+1} - x^*]^2] \leq (1 - \mu\alpha_k)\mathbb{E}[[x^k - x^*]^2] + \frac{2L_l^2}{\mu} \alpha_k \mathbb{E}[[y^k - y^*(x^k)]^2] + \alpha_k^2 \mathbb{E}[[\tilde{h}_f^k]^2] + \frac{4C_{g,x}^2 C_{f,x}^2}{\mu_g \mu} \alpha_k \mathbb{E}[[H_{yy} - \nabla_{yy}g(x^k, y^k)]^2] + \frac{4C_{f,x}^2}{\mu_g \mu} \alpha_k \mathbb{E}[[H_{xy} - \nabla_{xy}g(x^k, y^k)]^2] \tag{76}
\]

where $L_f, L_F$ are defined in Lemma 3 and $C_{g,x}$ is the projection radius of $H_{xy}$ in (3a).

**Proof:** We start with

\[
\mathbb{E}[[x^{k+1} - x^*]^2] \leq \mathbb{E}[[x^k - \alpha_k\tilde{h}^k_f - x^*]^2]
\]

\[
= ||x^k - x^*||^2 - 2\alpha_k\langle x^k - x^*, \mathbb{E}[\tilde{h}^k_f] \rangle + \alpha_k^2 \mathbb{E}[[\tilde{h}^k_f]^2]
\]

\[
= ||x^k - x^*||^2 - 2\alpha_k\langle x^k - x^*, \nabla F(x^k) \rangle
\]

\[
+ 2\alpha_k\langle x^k - x^* , \nabla F(x^k) - \mathbb{E}[\tilde{h}^k_f] \rangle + \alpha_k^2 \mathbb{E}[[\tilde{h}^k_f]^2] \tag{77}
\]

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

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

\[
-\langle x^k - x^*, \nabla F(x^k) - \nabla F(x^*) \rangle \leq -\mu||x^k - x^*||^2 \tag{78}
\]
where (c) uses the Young’s inequality.

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

\[
\begin{align*}
\|\nabla F(x^k) - \mathbb{E}[\tilde{h}_f^k|\mathcal{F}_k]\|^2 \\
&\leq 2\|\nabla F(x^k) - \nabla f(x^k, y^k)\|^2 + 2\mathbb{E}\left[\|\nabla f(x^k, y^k) - \mathbb{E}_{\xi^k}[\tilde{h}_f^k]\|^2|\mathcal{F}_k\right] \\
&\leq 2L_y^2\|y^k - y^*(x^k)\|^2 + 2\mathbb{E}\left[\|\nabla f(x^k, y^k) - \mathbb{E}_{\xi^k}[\tilde{h}_f^k]\|^2|\mathcal{F}_k\right] \\
&\leq 2L_y^2\|y^k - y^*(x^k)\|^2 + 2\left((H_{yy}^k)^{-1}H_{xy}^k - H_{yy}(x^k, y^k)^{-1}H_{xy}(x^k, y^k)\right)\|\nabla y f(x^k, y^k)\|^2 \\
&\leq 2L_y^2\|y^k - y^*(x^k)\|^2 + \left(\frac{4C_g^2C_f^2}{\mu_g}\right)^2\mathbb{E}[\|H_{yy}^k - H_{yy}(x^k, y^k)\|^2] + \left(\frac{4C_f^2}{\mu_g}\right)^2\mathbb{E}[\|H_{xy}^k - H_{xy}(x^k, y^k)\|^2]
\end{align*}
\]  

(80)

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

\[
\mathbb{E}_{\xi^k}[\tilde{h}_f^k|\mathcal{F}_k] = \nabla_x f(x^k, y^k) - (H_{yy}^k)^{-1}H_{xy}^k \nabla_y f(x^k, y^k)
\]  

(81)

and (c) follows the same steps of (34) and Assumption 3. Plugging (80) into the above completes the proof.

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

\[
\begin{align*}
\mathcal{V}^{k+1} - \mathcal{V}^k &= \|x^{k+1} - x^*\|^2 - \|x^k - x^*\|^2 + \|y^{k+1} - y^*(x^{k+1})\|^2 - \|y^k - y^*(x^k)\|^2 \\
&= \|H_{yy}^{k+1} - \nabla_y^2 g(x^{k+1}, y^{k+1})\|^2 - \|H_{yy}^k - \nabla_y^2 g(x^k, y^k)\|^2
\end{align*}
\]  

(82)

Using Lemmas 5 and 6 and defining $\hat{c}_3 := \hat{L}_{gys}^2 + L_{gys}^2 + \hat{L}_{gys}^2 + L_{gys}^2$, we obtain

\[
\begin{align*}
\mathbb{E}[\mathcal{V}^{k+1}] - \mathbb{E}[\mathcal{V}^k] &\leq -\mu_k \mathbb{E}[\|x^k - x^*\|^2] - \left(\frac{\mu_y L_y^3 \beta_k}{\mu_y + L_y} - \hat{c}_1 \hat{c}_2 \frac{\alpha_k^2}{\beta_k} - \frac{2L_y^2 \alpha_k}{\mu}\right) \mathbb{E}[\|y^k - y^*(x^k)\|^2] \\
&- \left(\tau_{k+1} - \hat{c}_1 \hat{c}_2 \frac{\alpha_k^2}{\beta_k} - \frac{4C_g^2C_f^2}{\mu_g\mu} \alpha_k\right) \mathbb{E}[\|H_{yy}^k - H_{yy}(x^k, y^k)\|^2] \\
&- \left(\tau_{k+1} - \hat{c}_1 \hat{c}_2 \frac{\alpha_k^2}{\beta_k} - \frac{4C_f^2}{\mu_g\mu} \alpha_k\right) \mathbb{E}[\|H_{xy}^k - H_{xy}(x^k, y^k)\|^2] \\
&+ \alpha_k^2 \mathbb{E}[\|\tilde{h}_f^k\|^2] + \left(1 + \frac{H_y^k L_y}{\mu_y + L_y} \beta_k\right) \beta_k^2 \sigma_g^2 + \hat{c}_1 \hat{c}_2 \frac{\alpha_k^4}{\beta_k} + 4\tau_{k+1}^2 \sigma_g^2 \\
&+ 2(1 - \tau_{k+1})^2 \hat{c}_3 \mathbb{E}[\|x^{k+1} - x^k\|^2] + 2(1 - \tau_{k+1})^2 \hat{c}_3 \mathbb{E}[\|y^{k+1} - y^k\|^2].
\end{align*}
\]  

(83)
Plugging (72) and (81) into (83), we have

\[
E[V^{k+1}] - E[V^k] \leq -\mu \alpha_k E[\|x^k - x^*\|^2] + \left(2 + 4\tilde{c}_3 + 8\tilde{c}_3 \left(\frac{C_{g,v}}{\mu^3}\right)^2\right) \left(C_{f,v}^2 + \left(\frac{C_{g,v}}{\mu}\right)^2 C_{f,v}^2\right) \alpha_k^2
\]

\[
- \left(\frac{\mu_g L_g \beta_k}{\mu_g + L_g} - \tilde{c}_1 \tilde{c}_2 \frac{\alpha_k^2}{\beta_k} - \frac{2L_f^2 \alpha_k}{\mu} - 8\beta_k^2 L_g^2\right) E[\|y^k - y^*(x^k)\|^2]
\]

\[
- \left(\frac{\tau_{k+1} - \tilde{c}_1 \tilde{c}_2 \alpha_k}{\beta_k} - \frac{4C_{g,v}^2 C_{f,v}^2}{\mu^3 \beta_k} \alpha_k\right) E[\|H_{yy}^k - H_{yy}(x^k, y^k)\|^2]
\]

\[
- \left(\frac{\tau_{k+1} - \tilde{c}_1 \tilde{c}_2 \alpha_k}{\beta_k} - \frac{4C_{f,v}^2}{\mu^3 \beta_k} \alpha_k\right) E[\|H_{xy}^k - H_{xy}(x^k, y^k)\|^2]
\]

\[
+ \left(1 + \frac{\mu_g L_g}{\mu_g + L_g} \beta_k \right) \beta_k \sigma_{g,v}^2 + \tilde{c}_1 \tilde{c}_2 \frac{\alpha_k^4}{\beta_k} + 4\tau_{k+1} \sigma_{g,v}^2 + 8L_g^2 \beta_k^2 \sigma_{g,v}^2.
\] (84)

We choose the stepsizes \(\alpha_k, \beta_k, \tau_k\) as (30) to guarantee that (cf. \(c := \tilde{c}_1 \tilde{c}_2\))

\[
(a) \quad \tau_{k+1} - \tilde{c}_1 \tilde{c}_2 \frac{\alpha_k^2}{\beta_k} - \frac{4C_{g,v}^2 C_{f,v}^2}{\mu^3 \beta_k} \alpha_k \geq \frac{\beta_k}{4}; \quad (b) \quad \tau_{k+1} - \tilde{c}_1 \tilde{c}_2 \frac{\alpha_k^2}{\beta_k} - \frac{4C_{f,v}^2}{\mu^3 \beta_k} \alpha_k \geq \frac{\beta_k}{4}
\]

\[
(c) \quad \frac{\mu_g L_g}{\mu_g + L_g} \beta_k - \tilde{c}_1 \tilde{c}_2 \frac{\alpha_k^2}{\beta_k} - \frac{2L_f^2}{\mu} \alpha_k - 8\beta_k^2 L_g^2 \geq \frac{\mu_g L_g}{4(\mu_g + L_g)}.
\] (85)

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

\[
E[V^{k+1}] - E[V^k] \leq -\mu \alpha_k E[\|x^k - x^*\|^2] - \frac{\mu_g L_g}{4(\mu_g + L_g)} \beta_k E[\|y^k - y^*(x^k)\|^2]
\]

\[
- \frac{\beta_k}{4} E[\|H_{yy}^k - H_{yy}(x^k, y^k)\|^2] - \frac{\beta_k}{4} E[\|H_{xy}^k - H_{xy}(x^k, y^k)\|^2] + \tilde{c}_5 \beta_k^2
\]

\[
\leq -\tilde{c}_5 \beta_k E[V^k] + \tilde{c}_6 \beta_k^2
\] (86)

where the first and second inequalities hold since we define

\[
\tilde{c}_5 := \min \left\{ \frac{\mu \alpha_k}{\beta_k}, \frac{\mu_g L_g}{4(\mu_g + L_g)}, \frac{1}{4} \right\} = O(1)
\]

\[
\tilde{c}_6 := \left(1 + \frac{\mu_g L_g}{\mu_g + L_g} \beta_k \right) \sigma_{g,v}^2 + \frac{\alpha_k^4}{4\beta_k} + 4\tau_{k+1} \sigma_{g,v}^2 + 8L_g^2 \beta_k^2 \sigma_{g,v}^2 + \tilde{c}_4 = O(1).
\] (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

\[
E[V^K] \leq \prod_{k=0}^{K-1} (1 - \tilde{c}_5 \beta_k) V^0 + \tilde{c}_6 \sum_{k=0}^{K-1} \beta_k^{k-1} \prod_{j=k+1}^{K-1} (1 - \tilde{c}_5 \beta_j)
\]

\[
\leq \frac{(K_0 - 2)(K_0 - 1)}{(K_0 + K - 2)(K_0 + K - 1)} V^0 + \tilde{c}_6 \sum_{k=0}^{K-1} \frac{4}{(k + K_0 - 1)(k + K_0)} \leq \frac{(K_0 - 1)^2}{(K_0 + K - 1)^2} V^0 + \frac{4\tilde{c}_6 K}{\tilde{c}_5^2(K + K_0 - 1)^2}
\] (88)

from which the proof is complete.