A Fully First-Order Method for Stochastic Bilevel Optimization

Jeongyeol Kwon [1] **Dohyun Kwon** [2] **Stephen Wright** [1] **Robert Nowak** [1]

Abstract

We consider stochastic unconstrained bilevel optimization problems when only the first-order gradient oracles are available. While numerous optimization methods have been proposed for tackling bilevel problems, existing methods either tend to require possibly expensive calculations regarding Hessians of lower-level objectives, or lack rigorous finite-time performance guarantees. In this work, we propose a Fully First-order Stochastic Approximation ($\mathtt{F}^2\mathtt{SA}$) method, and study its non-asymptotic convergence properties. Specifically, we show that $\mathtt{F}^2\mathtt{SA}$ converges to an ϵ-stationary solution of the bilevel problem after $\epsilon^{-7/2}, \epsilon^{-5/2}$, and $\epsilon^{-3/2}$ iterations (each iteration using $O(1)$ samples) when stochastic noises are in both level objectives, only in the upper-level objective, and not present (deterministic settings), respectively. We further show that if we employ momentum-assisted gradient estimators, the iteration complexities can be improved to $\epsilon^{-5/2}, \epsilon^{-4/2}$, and $\epsilon^{-3/2}$, respectively. We demonstrate even superior practical performance of the proposed method over existing second-order based approaches on MNIST data-hypercleaning experiments.

1. Introduction

Bilevel optimization (Colson et al., 2007) arises in many important applications that have two-level hierarchical structures, including meta-learning (Rajeswaran et al., 2019), hyper-parameter optimization (Franceschi et al., 2018; Bao et al., 2021), model selection (Kunapuli et al., 2008; Giovannelli et al., 2021), adversarial networks (Goodfellow et al., 2020; Gidel et al., 2018), game theory (Stackelberg et al., 1952) and reinforcement learning (Konda & Tsitsiklis, 1999; Sutton & Barto, 2018). Bilevel optimization can be generally formulated as the following minimization problem:

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

where f and g are continuously differentiable functions and $X \subseteq \mathbb{R}^{d_x}$ is a convex set. The outer objective F depends on x both directly and also indirectly via $y^*(x)$, which is a solution of the lower-level problem of minimizing another function g, which is parametrized by x. Throughout the paper, we assume that $X = \mathbb{R}^{d_x}$ (that is, there are no explicit constraints on x) and that $g(x, y)$ is strongly convex in y, so that $y^*(x)$ is uniquely well-defined for all $x \in X$.

Among various approaches to (**P**), iterative procedures have been predominant due to their simplicity and potential scalability in large-scale applications. Initiated by (Ghadimi & Wang, 2018), a flurry of recent works study efficient iterative procedures and their finite-time performance for solving (**P**), see *e.g.,* (Chen et al., 2021; Hong et al., 2020; Khanduri et al., 2021; Chen et al., 2022; Dagréou et al., 2022; Guo et al., 2021; Sow et al., 2022; Ji et al., 2021; Yang et al., 2021). The underlying idea is based on an algorithm of (stochastic) gradient descent type, applied to F, that is,

$$x_{k+1} = x_k - \alpha_k \nabla F(x_k),$$

with some appropriate step-sizes $\{\alpha_k\}$. Direct application of this approach requires us to compute or estimate the so-called hyper-gradient of F at x, which is

$$\begin{aligned} \nabla F(x) =& \nabla_x f(x, y^*(x)) - \nabla^2_{xy} g(x, y^*(x)) \times \\ & \nabla^2_{yy} g(x, y^*(x))^{-1} \nabla_y f(x, y^*(x)). \end{aligned} \tag{1}$$

There are two major obstacles in computing (1). The first obstacle is that for every given x, we need to search for the optimal solution $y^*(x)$ of the lower problem, which results in updating the lower variable y multiple times before updating x. To tackle this issue, several ideas have been proposed in (Ghadimi & Wang, 2018; Hong et al., 2020; Chen et al., 2021) to effectively track $y^*(x)$ without waiting for too many inner iterations before updating x (we discuss this further in Section 1.2). Following in the spirit of this approach, we show that a single-loop style algorithm can still be implemented using only first-order gradient estimators.

[1]University of Wisconsin-Madison, USA [2]University of Seoul, Korea. Correspondence to: Jeongyeol Kwon <jeongyeol.kwon@wisc.edu>, Dohyun Kwon <dh.dohyun.kwon@gmail.com>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

The second obstacle, which is the main focus of this work, centers around the presence of second-order derivatives of g in (1). Existing approaches mostly require an explicit extraction of second-order information from g with a major focus on estimating the Jacobian and inverse Hessian efficiently with stochastic noises (Ji et al., 2021; Chen et al., 2022; Dagréou et al., 2022). We are particularly interested in regimes in which such operations are costly and prohibitive (Mehra & Hamm, 2021; Giovannelli et al., 2021). Some existing works avoid the second-order computation and only use the first-order information of both upper and lower objectives; see (Giovannelli et al., 2021; Sow et al., 2022; Liu et al., 2021b; Ye et al., 2022). These works either lack a complete finite-time analysis (Giovannelli et al., 2021; Liu et al., 2021b) or are applicable only to deterministic functions (Ye et al., 2022; Sow et al., 2022).

Our goal in this paper is to study a *fully* first-order approach for stochastic bilevel optimization. We propose a gradient-based approach that avoids the estimation of Jacobian and Hessian of g, and finds an ϵ-stationary solution of F using only first-order gradients of f and g. Further, the number of inner iterations remains constant throughout all outer iterations of our algorithm. We provide a finite-time analysis of our method with explicit convergence rates. To our best knowledge, this work is the first to establish non-asymptotic convergence guarantees for stochastic bilevel optimization using only first-order gradient oracles.

1.1. Overview of Main Results

The starting point of our approach is to convert (**P**) to an equivalent constrained single-level version:

$$\min_{x\in X,\ y\in\mathbb{R}^{d_y}} \quad f(x,y) \quad \text{s.t.} \quad g(x,y)-g^*(x)\le 0, \quad \textbf{(P')}$$

where $g^*(x) := g(x, y^*(x))$. The Lagrangian $\mathcal{L}_\lambda$ for (**P'**) with multiplier $\lambda > 0$ is

$$\mathcal{L}_\lambda(x,y) := f(x,y) + \lambda(g(x,y) - g^*(x)).$$

We can minimize $\mathcal{L}_\lambda$ for a given λ by, for example, running (stochastic) gradient descent. As noted in (Ye et al., 2022), the gradient of $\mathcal{L}_\lambda$ can be computed only with gradients of f and g, and thus the entire procedure can be implemented using only with first-order derivatives. In fact, such a reformulation has been attempted in several recent works (e.g., (Liu et al., 2021a; Sow et al., 2022; Ye et al., 2022)). However, the challenge in handling the constrained version (**P'**) is to find an appropriate value of the multiplier λ. Unfortunately, the desired solution $x^* = \arg\min_x F(x)$ can only be obtained at $\lambda = \infty$ (this is a consequence of the fact that the so-called *constraint qualifications* (Wright et al., 1999) are not satisfied for (**P'**)). However, with $\lambda = \infty$, $\mathcal{L}_\lambda(x, y)$ has unbounded smoothness which prevents us from employing gradient-descent style approaches. For these reasons, none of the previously proposed algorithms can obtain a consistent estimator for the original problem $\min_x F(x)$ without access to second derivatives of g.

Nonetheless, we find that (**P'**) is the key to deriving a consistent estimator that converges to an ϵ-stationary point of F in finite time *without* access to second derivatives. The main idea is to start with an initial value $\lambda = \lambda_0 > 0$ and gradually increase it on subsequent iterations: At iteration k, $\lambda_k = O(k^b)$ for some $b \in (0, 1]$. The success of this approach depends crucially on the growth rate captured by the parameter b. On one hand, fast growth of λ_k removes the bias quickly. On the other hand, fast growth of λ_k forces a fast decay of step-sizes due to the growing nonsmoothness of $\mathcal{L}_{\lambda_k}$, which slows down the overall convergence.

Our main technical contribution is to characterize an explicit growth rate of λ_k that optimizes the trade-off between bias and step-sizes, and to provide a non-asymptotic convergence guarantee with explicit rates for the proposed algorithm.

- We propose a fully first-order method, $\texttt{F}^2\texttt{SA}$, for stochastic bilevel optimization. $\texttt{F}^2\texttt{SA}$ is a single-loop style algorithm: For every outer variable update we only update inner variables a constant number of times.
- We characterize explicit convergence rates of $\texttt{F}^2\texttt{SA}$ in different stochastic regimes. It converges to an ϵ-stationary-point of (**P**) after $\tilde{O}(\epsilon^{-3.5})$, $\tilde{O}(\epsilon^{-2.5})$, or $\tilde{O}(\epsilon^{-1.5})$ iterations if both ∇f and ∇g contain stochastic noise, if only access to ∇f is noisy, or if we are in deterministic settings, respectively. These complexities can be improved to $\tilde{O}(\epsilon^{-2.5})$, $\tilde{O}(\epsilon^{-2})$, or $\tilde{O}(\epsilon^{-1.5})$, respectively, if momentum or variance-reduction techniques are employed. The crux of the analysis is to understand the effect of the value of multipliers λ_k on step-sizes, noise variances, and bias.
- We demonstrate the proposed algorithm on a data hyper-cleaning task for MNIST. Even though our theoretical guarantees are not better than existing methods that use second-order information, we illustrate that $\texttt{F}^2\texttt{SA}$ can even outperform such methods in practice.

1.2. Related Work

Bilevel optimization has a long and rich history since its first introduction in (Bracken & McGill, 1973). A number of algorithms have been proposed for bilevel optimization. Classical results include approximation descent (Vicente et al., 1994) and penalty function method (Ishizuka & Aiyoshi, 1992; Anandalingam & White, 1990; White & Anandalingam, 1993) for instance; see (Colson et al., 2007) for a comprehensive overview. These results often deal with several special cases of bilevel-optimization and only provide asymptotic guarantees. Note that the penalty function methods in (Ishizuka & Aiyoshi, 1992; Anandalingam

& White, 1990; White & Anandalingam, 1993) discuss the landscape within the infinitesimal neighborhood of local minimizers, and their results cannot imply practical approaches to find a stationary point non-convex objectives F.

Recently, several papers study gradient-based optimization methods for bilevel optimization and its non-asymptotic analysis. The first non-asymptotic analysis of a double-loop algorithm was given in (Ghadimi & Wang, 2018), where an inner problem finds an approximate solution of $y^*(x)$ given x, which is used to evaluate an approximation of $\nabla F(x)$. Furthermore, (Ghadimi & Wang, 2018) uses the Neuman series approximation to estimate the Hessian inverse when we only have access to the stochastic oracles of second-order derivatives.

The paper (Ghadimi & Wang, 2018) was followed by a flurry of work that improved their result in numerous ways. For instance, (Hong et al., 2020; Chen et al., 2021; 2022; Ji et al., 2021) develop a single-loop style update by properly choosing two step-sizes for the inner and outer iterations, along with the improved sample complexity, *i.e.,* the total number of accesses to first and second-order stochastic oracles. The overall convergence rate is further optimized by using variance-reduction and momentum techniques (Khanduri et al., 2021; Dagréou et al., 2022; Guo et al., 2021; Yang et al., 2021; Huang & Huang, 2021). We do not aim to compete with the convergence rates obtained from this line of work, since all of these method have access to second-order derivatives, even though some computational cost might be saved if good automatic differentiation packages (Margossian, 2019) are available. Rather, we avoid the needs for second-order information altogether, allowing a simple algorithm with low per-iteration complexity for large scale applications.

The results most closely related to ours can be found in (Ye et al., 2022; Sow et al., 2022). (Sow et al., 2022) considers a primal-dual approach for (**P'**), but their main focus is to get a biased solution when g is only convex (not strongly convex), so the lower-level problem may have multiple solutions. Their analysis is restricted to the case in which the overall Lagrangian is strongly-convex in x (which is not usually guaranteed) and they do not provide any guarantees in terms of the true objective F. More recent work in (Ye et al., 2022) is the closest to ours, but they only consider deterministic gradient oracles, and do not provide convergence guarantees in terms of F. Moreover, they prove a convergence guarantee of $O(k^{-1/4})$, whereas we show an improved guarantee of $\tilde{O}(k^{-2/3})$ in the deterministic case.

There are also other lines of work that study a simpler version of the bilevel problem which has no coupling between two variables x and y (*e.g.,* see (Ferris & Mangasarian, 1991; Solodov, 2007; Jiang et al., 2022)). In (Amini & Yousefian, 2019a;b), the Lagrangian formulation is exploited with iteratively increasing multiplier. Note that the nature of single-variable bilevel formulation is different from (**P**) as the former is only interesting when the lower-level problem allows a multiple (convex) solution set. To our best knowledge, the idea of iteratively increasing λ_k with its non-asymptotic guarantee is new in the context of solving (**P**), and has the merit of avoiding (possibly) expensive second-order computation.

2. Preliminaries

We state several assumptions on (**P**) to specify the problem class of interest. We consider (**P**) with the following assumptions on objective functions:

Assumption 1. The functions f and g satisfy the following conditions.

1. f is continuously differentiable and $l_{f,1}$-smooth .
2. g is continuously differentiable and $l_{g,1}$-smooth.
3. For every $\bar{x} \in X$, $\|\nabla_y f(\bar{x}, y)\| \leq l_{f,0}$ for all y.

We focus on well-conditioned bilevel optimization problems, *i.e.,* when $F(x)$ is well-defined, continuous and smooth. The following assumption has been standard for well-conditioned bilevel problems (Ghadimi & Wang, 2018):

Assumption 2. The following holds for g:

1. There exists an $\mu_g > 0$ such that for all $\bar{x} \in X$, $g(\bar{x}, y)$ is μ_g strongly-convex in y.
2. g is two-times continuously differentiable, and $\nabla^2 g$ is $l_{g,2}$-Lipschitz jointly in (x, y).

We assume that we can access first-order information of objective functions only through stochastic gradient oracles:

Assumption 3. We access the gradients of objective functions via unbiased estimators $\nabla f(x, y; \zeta), \nabla g(x, y; \phi)$ depending on random variables ζ and ϕ, respectively, where $\mathbb{E}[\nabla f(x, y; \zeta)] = \nabla f(x, y)$ and $\mathbb{E}[\nabla g(x, y; \phi)] = \nabla g(x, y)$. The variances of stochastic gradient estimators are bounded:

$$\mathbb{E}[\|\nabla f(x, y; \zeta) - \nabla f(x, y)\|^2] \leq \sigma_f^2,$$
$$\mathbb{E}[\|\nabla g(x, y; \phi) - \nabla g(x, y)\|^2] \leq \sigma_g^2.$$

Throughout the paper, we assume that Assumptions 1-3 hold unless specified otherwise. We use the following definition as the optimality criteria for solving (**P**).

Definition 2.1 (ϵ-stationary point)**.** A point x is called ϵ-stationary if $\|\nabla F(x)\|^2 \leq \epsilon$, where ∇F is defined in (1).

Notation. We say $a_k \asymp b_k$ if a_k and b_k decreases (or increases) in the same rate as $k \to \infty$, *i.e.,* $\lim_{k\to\infty} a_k/b_k = \Theta(1)$. Throughout the paper, $\|\cdot\|$ denotes the Euclidean norm on finite dimensional space.

3. Algorithm

In this section, we develop an algorithm that converges to a stationary point of the bilevel problem (*i.e.,* a stationary point of $F(x) = f(x, y^*(x))$) and makes use only of gradients of f and g. Recall the equivalent formulation (**P'**). To see how we can avoid second-order derivatives, we observe the gradient of $\nabla\mathcal{L}_\lambda$:

$$\nabla_x\mathcal{L}_\lambda(x,y) = \nabla_x f(x,y) + \lambda(\nabla_x g(x,y) - \nabla g^*(x)),$$
$$\nabla_y\mathcal{L}_\lambda(x,y) = \nabla_y f(x,y) + \lambda\nabla_y g(x,y).$$

Note that

$$\begin{aligned}\nabla g^*(x) &= \nabla_x g(x, y^*(x)) + \nabla_x y^*(x)\nabla_y g(x, y^*(x)) \\ &= \nabla_x g(x, y^*(x)),\end{aligned}$$

due to the optimality condition for g at $y^*(x)$. Thus, we could consider optimizing $\mathcal{L}_\lambda(x,y)$ by introducing an auxiliary variable z that chases $y^*(x)$, and setting up an alternative bilevel formulation (**P**) with outer-level objective $\mathcal{L}_\lambda(x', z)$, outer variable $x' = (x, y)$, and inner variable z. However, such an approach settles in a different landscape from that of $F(x)$, resulting in a bias. The question is how tightly we can control this bias without compromising too much smoothness of the alternative function $\mathcal{L}_\lambda$, which affects the overall step-size design and noise variance.

To control the bias, we need a better understanding of how the functions $\mathcal{L}_\lambda$ and $F(x)$ are related. Let us introduce an auxiliary function $\mathcal{L}^*_\lambda$ defined as:

$$\mathcal{L}^*_\lambda(x) := \min_y \mathcal{L}_\lambda(x,y).$$

Note that if $\lambda > 2l_{f,1}/\mu_g$, then for every $\bar{x} \in X$, $\mathcal{L}_\lambda(\bar{x}, y)$ is at least $(\lambda\mu_g/2)$ strongly-convex in y, and therefore its minimizer $y^*_\lambda(x)$ is uniquely well-defined:

$$y^*_\lambda(x) := \arg\min_y \mathcal{L}_\lambda(x,y). \tag{2}$$

Since $F(x) = \lim_{\lambda\to\infty}\mathcal{L}^*_\lambda(x)$ for every $x \in X$, we could expect that $\mathcal{L}^*_\lambda(x)$ is a well-defined proxy of $F(x)$ for sufficiently large $\lambda > 0$. The following lemma confirms this intuition.

Lemma 3.1. *For any $x \in X$ and $\lambda \geq 2l_{f,1}/\mu_g$, $\nabla\mathcal{L}^*_\lambda(x)$ is given by*

$$\begin{aligned}\nabla_x\mathcal{L}_\lambda(x, y^*_\lambda(x)) &= \nabla_x f(x, y^*_\lambda(x)) \\ &\quad + \lambda(\nabla_x g(x, y^*_\lambda(x)) - \nabla_x g(x, y^*(x))).\end{aligned}$$

Algorithm 1 $\mathtt{F^2SA}$

Input: step sizes: $\{\alpha_k, \gamma_k\}$, multiplier difference sequence: $\{\delta_k\}$, inner-loop iteration count: T, step-size ratio: ξ, initializations: λ_0, x_0, y_0, z_0

for $k = 0...K-1$ **do**
  $z_{k,0} \leftarrow z_k$, $y_{k,0} \leftarrow y_k$
  for $t = 0...T-1$ **do**
    $z_{k,t+1} \leftarrow z_{k,t} - \gamma_k h^{k,t}_{gz}$
    $y_{k,t+1} \leftarrow y_{k,t} - \alpha_k(h^{k,t}_{fy} + \lambda_k h^{k,t}_{gy})$
  end for
  $z_{k+1} \leftarrow z_{k,T}$, $y_{k+1} \leftarrow y_{k,T}$
  $x_{k+1} \leftarrow x_k - \xi\alpha_k(h^k_{fx} + \lambda_k(h^k_{gxy} - h^k_{gxz}))$
  $\lambda_{k+1} \leftarrow \lambda_k + \delta_k$
end for

Furthermore, we have

$$\|\nabla F(x) - \nabla\mathcal{L}^*_\lambda(x)\| \leq C_\lambda/\lambda,$$

where $C_\lambda := \frac{4l_{f,0}l_{g,1}}{\mu_g^2}\left(l_{f,1} + \frac{2l_{f,0}l_{g,2}}{\mu_g}\right)$.

Importantly, $\nabla\mathcal{L}^*_\lambda(x)$ can be computed only with first-order derivatives of both f and g. Thus any first-order method that finds a stationary point of $\mathcal{L}^*_\lambda(x)$ approximately follows the trajectory of x updated with the exact $\nabla F(x)$, with a bias of $O(1/\lambda)$.

Our strategy is to use $\nabla\mathcal{L}^*_\lambda(x)$ as a proxy to $\nabla F(x)$ for generating a sequence of iterates $\{x_k\}$. Accordingly, we introduce sequences $\{y_k\}$ and $\{z_k\}$ that approximate $y^*_{\lambda_k}(x_k)$ and $y^*(x_k)$, respectively. We gradually increase λ_k with k, so that the bias in the sequence $\{x_k\}$ converges to 0.

Our **F**ully **F**irst-order **S**tochastic **A**pproximation ($\mathtt{F^2SA}$) method is shown in Algorithm 1. We emphasize that the method works with *stochastic* gradients that are independent unbiased estimators of gradients, *i.e.,*

$$h^{k,t}_{gz} := \nabla_y g(x_k, z_{k,t}; \phi^{k,t}_z),\ h^{k,t}_{fy} := \nabla_y f(x_k, y_{k,t}; \zeta^{k,t}_y),$$
$$h^{k,t}_{gy} := \nabla_y g(x_k, y_{k,t}; \phi^{k,t}_y), h^k_{gxy} := \nabla_x g(x_k, y_{k+1}; \phi^k_{xy}),$$
$$h^k_{fx} := \nabla_x f(x_k, y_{k+1}; \zeta^k_x), h^k_{gxz} := \nabla_x g(x_k, z_{k+1}; \phi^k_{xz}).$$

The algorithm can set $T = 1$ in conjunction with an appropriate choice of ξ, allowing a fully single-loop update for all variables.

3.1. Step-Size Design Principle

We describe how we design the step-sizes for Algorithm 1 to achieve convergence to a ϵ-stationary point of F. Several conditions must be satisfied. As will be shown in the analysis, with $(\lambda_k\mu_g/2)$-strong convexity of $\mathcal{L}_{\lambda_k}$ in y, one-step inner iteration of $y_{k,t}$ is a contraction mapping toward $y^*_{\lambda,k}$ with rate $1 - O(\mu_g\beta_k)$. Henceforth, we often use the

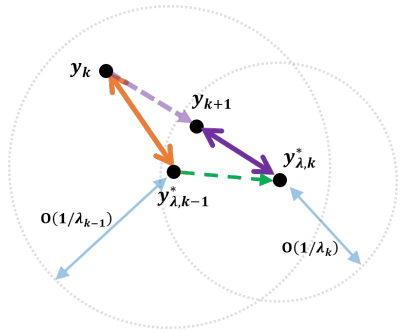


Figure 1: y_k should move faster than $y^*_{\lambda_k}(x_k)$ moves, and stay within $O(1/\lambda_k)$-ball around $y^*_{\lambda_k}(x_k)$.

notation $\beta_k := \alpha_k\lambda_k$, which is the effective step-size for updating y_k. For simplicity, we denote $y^*_{\lambda,k} := y^*_{\lambda_k}(x_k)$ and $y^*_k := y^*(x_k)$.

We now describe the specific rules. Since the step size for x_k is essentially $\xi\alpha_k$, and since we may need to traverse an arbitrary distance from the initial point x_0 to the optimal value of x, we need $\alpha_k = \Omega(1/k)$. On the other hand, since $\beta_k = \alpha_k\lambda_k$ is the effective step size for updating y_k, we need $\beta_k < O(1/l_{g,1}) = O(1)$. Together, these observations imply that the maximum rate of growth for λ_k cannot exceed $O(k)$.

Second, note that $\|x_{k+1} - x_k\|$ is (roughly) proportional to

$$\|\nabla F(x_k)\| + C_\lambda/\lambda_k + \lambda_k\|y_k - y^*_{\lambda,k}\| + \lambda_k\|z_k - y^*_k\|.$$

This rate is optimized when $\|y_k - y^*_{\lambda,k}\| \asymp \|z_k - y^*_k\| \asymp \lambda_k^{-2}$. Thus, the ideal growth rate for λ_k is $\|y_k - y^*_{\lambda,k}\|^{1/2}$ or $\|z_k - y^*_k\|^{1/2}$. We will design the rate of convergence of y_k and z_k to be the same, *i.e.,* $\beta_k \asymp \gamma_k$. For instance, when we have stochastic noises in the gradient estimate of g, *i.e.,* $\sigma_g^2 > 0$, the expected convergence rate of $\|y_k - y^*_{\lambda_k}\|^2$ is $O(\beta_k)$, since the sequence is optimized for strongly convex functions. This suggests $\lambda_k \asymp \beta_k^{-1/4}$ as the ideal rate of growth for λ_k.

The crux of Algorithm 1 is how well y_k (and z_k) can chase $y^*_{\lambda_k}(x_k)$ (resp. $y^*(x_k)$) when x_k and λ_k are changing at every iteration. We characterize first how fast $y^*_\lambda(x)$ moves in relation to the movements of λ and x.

Lemma 3.2. *For any $\lambda_2 \geq \lambda_1 \geq 2l_{f,1}/\mu_g$ and $x_1, x_2 \in X$, we have*

$$\|y^*_{\lambda_1}(x_1) - y^*_{\lambda_2}(x_2)\| \leq \frac{2(\lambda_2 - \lambda_1)}{\lambda_1\lambda_2}\frac{l_{f,0}}{\mu_g} + l_{\lambda,0}\|x_2 - x_1\|,$$

for some $l_{\lambda,0} \leq 3l_{g,1}/\mu_g$.

For Algorithm 1 to converge to a desired point, y_k should move sufficiently fast toward the current target $y^*_{\lambda,k}$ every iteration, dominating the movement of target $y^*_{\lambda,k}$ that results from updates to x_k and λ_k (see Figure 1). At a minimum, the following condition should hold (in expectation):

$$\|y_{k+1} - y^*_{\lambda,k}\| < \|y_k - y^*_{\lambda,k-1}\|.$$

Since $\|y_{k+1} - y^*_{\lambda,k}\|^2$ can be bounded with T-steps of $1 - O(\mu_g\beta_k)$ contractions, starting from y_k, we require

$$(1 - O(T\mu_g\beta_k))\,\|y_k - y^*_{\lambda,k}\|^2 < \|y_k - y^*_{\lambda,k-1}\|^2.$$

Now, applying the bound in Lemma 3.2, the minimal condition is given by:

$$\begin{aligned}\|y^*_{\lambda,k-1} - y^*_{\lambda,k}\| &\leq (l_{f,0}/\mu_g)\cdot(\delta_k/\lambda_k^2) + l_{\lambda,0}\|x_k - x_{k-1}\| \\ &\leq T\mu_g\beta_k\|y_k - y^*_{\lambda,k-1}\|.\end{aligned}$$

Note that $\|y_{k+1} - y^*_{\lambda,k}\|$ should decay faster than λ_k^{-1}. Otherwise, the bias in updating x_k using y_k (to estimate $\nabla\mathcal{L}^*_{\lambda_k}$) is larger than $\lambda_k\|y_{k+1} - y^*_{\lambda,k}\|$, and this amount might blow up. Also, it can be easily seen that $\|x_k - x_{k-1}\| = \Omega(\xi\beta_k\|y_k - y^*_{\lambda,k-1}\|)$. We can thus derive two simple conditions:

$$\frac{\delta_k}{\lambda_k} \leq O_{\mathrm{P}}(1)\cdot\beta_k,\ \frac{\xi}{T} < O_{\mathrm{P}}(1),$$

where $O_{\mathrm{P}}(1)$ are instance-dependent constants. If λ_k grows in some polynomial rate, then $\delta_k/\lambda_k = O(1/k)$ and the first condition is satisfied provided that $\beta_k = \Omega(1/k)$. The second condition indicates the number of inner iterations T required for each outer iteration. We can set $T = 1$ (thus making the algorithm single-loop) by setting ξ sufficiently small. Alternatively, we can set $\xi = 1$ choose $T > 1$ to depend on some instance-specific parameters.

Algorithm 2 $\mathtt{F^3SA}$

Input: step sizes: $\{\alpha_k, \gamma_k\}$, multiplier difference sequence: $\{\delta_k\}$, momentum-weight sequence $\{\eta_k\}$, step-size ratio: ξ, initialization: λ_0, x_0, y_0, z_0

for $k = 0...K-1$ **do**
  $z_{k+1} \leftarrow z_k - \gamma_k\tilde{h}^k_{gz}$
  $y_{k+1} \leftarrow y_k - \alpha_k(\tilde{h}^k_{fy} + \lambda_k\tilde{h}^k_{gy})$
  $x_{k+1} \leftarrow x_k - \xi\alpha_k(\tilde{h}^k_{fx} + \lambda_k(\tilde{h}^k_{gxy} - \tilde{h}^k_{gxz}))$
  $\lambda_{k+1} \leftarrow \lambda_k + \delta_k$
end for

3.2. Extension: Integrating Momentum

Given the simple structure of Algorithm 1, we can integrate variance-reduction techniques to improve the overall convergence rates. One relevant technique is the momentum-assisting technique of (Khanduri et al., 2021) for stochastic bilevel optimization. To simplify the presentation, we consider a fully single-loop variant by setting $T = 1$.

To apply the momentum technique, we only need to replace the simple unbiased gradient estimators h with momentum-assisted gradient estimators $\tilde{h}$. For instance, $\tilde{h}_z^k$ can be defined with a proper momentum weight sequence $\eta_k \in (0,1]$ as follows:

$$\tilde{h}_z^k := \nabla_y g(x_k, z_k; \phi_z^k) + (1-\eta_k)\left(\tilde{h}_z^{k-1} - \nabla_y g(x_{k-1}, z_{k-1}; \phi_z^k)\right).$$

Other quantities $\tilde{h}_{fy}, \tilde{h}_{gy}, \tilde{h}_{fx}, \tilde{h}_{gxy}, \tilde{h}_{gxz}$ are defined similarly, with the same momentum-weight sequence. We defer the full description of those quantities to Appendix C. The version of our algorithm that incorporates momentum is called **F**aster **F**ully **F**irst-order **S**tochastic **A**pproximation ($\mathtt{F}^3\mathtt{SA}$); it is described in Algorithm 2, where we simply replace h with $\tilde{h}$. Note that we have additional moment-weight parameters $\{\eta_k\}$.

4. Main Results

In this section we provide non-asymptotic convergence guarantees of the proposed algorithms. For Algorithm 1, we prove in Theorem 4.1 that the weighted sum of $\|\nabla F(x_k)\|^2$ in expectation is bounded from above. By choosing suitable step sizes, the estimate yields a convergence rate. Dependence on stochastic noises is explicated in Corollaries 4.2. Similar results with better convergence rates and weaker assumptions are proved for Algorithm 2; see Theorem 4.3 and Corollary 4.4.

4.1. Main Result for Algorithm 1

Two mild assumptions are required for exploiting the smoothness of $y_\lambda^*(x)$.

Assumption 4. The gradient with respect to x is bounded for functions f and g:

1. For every $\bar{y}$, $\|\nabla_x f(x,\bar{y})\| \le l_{f,0}$ for all $x \in X$.
2. For every $\bar{y}$, $\|\nabla_x g(x,\bar{y})\| \le l_{g,0}$ for all $x \in X$.

Assumption 5. f is two-times continuously differentiable, and $\nabla^2 f$ is $l_{f,2}$-Lipschitz in (x,y).

The smoothness of $y_\lambda^*(x)$ is used to keep the number of effective inner iterations constant throughout all outer-iterations, as in (Chen et al., 2021).

Before we state our convergence result, let us define some additional notation. We denote the second-moment bound of the x update, $x_{k+1}-x_k$, as $M := \max(l_{f,0}^2+\sigma_f^2, l_{g,0}^2+\sigma_g^2)$. We also denote $l_{*,0} = \max(1, l_{\lambda_0,0})$ and $l_{*,1} = l_{\lambda_0,1}$ where λ_0 is the starting value of Lagrange multiplier.

We are now ready to state our main results for Algorithm 1.

Theorem 4.1. *Suppose that Assumptions 1 - 5 hold, and parameters and step-sizes are chosen such that $\lambda_0 \ge 2l_{f,1}/\mu_g$ and*

$$\beta_k \le \gamma_k \le \min\left(\frac{1}{4l_{g,1}}, \frac{1}{4T\mu_g}\right),\ \alpha_k \le \frac{1}{2\xi l_{F,1}}, \tag{3a}$$

$$\frac{\xi}{T} < c_\xi \mu_g \cdot \max\left(l_{g,1}l_{*,0}^2,\ l_{*,1}\sqrt{M}\right)^{-1},\ \frac{\delta_k}{\lambda_k} \le \frac{T\mu_g\beta_k}{16} \tag{3b}$$

for all $k \ge 0$ with a proper absolute constant $c_\xi > 0$. Then for any $K \ge 1$, the iterates generated by Algorithm 1 satisfy

$$\sum_{k=0}^{K-1} \xi\alpha_k \mathbb{E}[\|\nabla F(x_k)\|^2] \le O_{\mathbb{P}}(1)\cdot\sum_k \xi\alpha_k\lambda_k^{-2} + O_{\mathbb{P}}(\sigma_f^2)\cdot\sum_k \alpha_k^2\lambda_k + O_{\mathbb{P}}(\sigma_g^2)\cdot\sum_k \gamma_k^2\lambda_k + O_{\mathbb{P}}(1),$$

where $O_{\mathbb{P}}(1)$ are instance-dependent constants.

The proof of Theorem 4.1 is given in Appendix B. At a high level, our analysis investigates the decrease in expectation (with k) of the potential function $\mathbb{V}_k$ defined by

$$\mathbb{V}_k := (F(x_k) - F^*) + l_{g,1}\lambda_k\|y_k - y_{\lambda_k}^*(x_k)\|^2 + \frac{\lambda_k l_{g,1}}{2}\|z_k - y^*(x_k)\|^2, \tag{4}$$

where F^* is the minimum value of F and y_λ^* and y^* are given in (2) and (**P**), respectively. That is, in addition to the decrease in values of F and $z_k - y_k^*$ which have been standardized in literature, we track the error between y_k and $y_{\lambda_k}^*(x_k)$ since $y_{\lambda,k}^*$ is the key to compute true $\nabla F(x_k)$ only with gradients. It is also shown in the proof that the right scaling factor for the tracking errors is $O_{\mathbb{P}}(\lambda_k)$.

We now describe how we design step sizes. Note that the conditions (3a) are standard conditions on the step sizes for gradient-based methods with smooth functions. The conditions (3b) arise from the double-loop nature of the problem, as discussed in Section 3.1. In accordance with the step-size design rule (3), we propose the following:

$$T = \max\left(32, (c_\xi\mu_g)^{-1}\max\left(l_{g,1}l_{*,0}^2,\ \sqrt{M}l_{*,1}\right)\right),$$
$$\xi = 1, \alpha_k = \frac{c_\alpha}{(k+k_0)^a},\ \gamma_k = \frac{c_\gamma}{(k+k_0)^c}, \tag{5}$$

and for the multiplier increase sequence $\{\delta_k\}$,

$$\delta_k = \min\left(\frac{T\mu_g}{16}\alpha_k\lambda_k^2,\ \frac{\gamma_k}{2\alpha_k} - \lambda_k\right), \tag{6}$$

with some rate constants $a, c \in [0,1]$ and $a \ge c$. We design the starting value λ_0 of the Lagrange multiplier and the constants as

$$k_0 \ge \frac{4}{\mu_g}\max\left(\frac{\xi l_{F,1}}{2}, Tl_{g,1}, l_{f,1}\right),\ \lambda_0 \ge \frac{2l_{f,1}}{\mu_g},$$

$$c_\gamma = \frac{1}{\mu_g k_0^{1-c}},\ c_\alpha = \frac{1}{2\lambda_0 \mu_g k_0^{1-a}}. \tag{7}$$

These choices simplify the convergence rate analysis, but any set of choices can be used as long as it satisfies (3). With the choices above, we can specify the rate of convergence in three different regimes of stochastic noises.

Corollary 4.2. *Suppose that the conditions of Theorem 4.1 hold, with step-sizes designed as in (5), (6), and (7). Let R be a random variable drawn from a uniform distribution over $\{0, ..., K-1\}$. Then the following convergence results hold after K iterations of Algorithm 1.*

(a) If stochastic noises are present in both upper-level objective f and lower-level objective g (i.e., $\sigma_f^2, \sigma_g^2 > 0$), then by setting $a = 5/7$ and $c = 4/7$ in (5) and (7), we obtain $\mathbb{E}[\|\nabla F(x_R)\|^2] \asymp \frac{\log K}{K^{2/7}}$,

(b) If stochastic noises are present only in f (i.e., $\sigma_f^2 > 0$), $\sigma_g^2 = 0$), then by setting $a = 3/5$ and $c = 2/5$ in (5) and (7), we obtain $\mathbb{E}[\|\nabla F(x_R)\|^2] \asymp \frac{\log K}{K^{2/5}}$,

(c) If we have access to exact information about f and g (i.e., $\sigma_f^2 = \sigma_g^2 = 0$), then by setting $a = 1/3$ and $c = 0$ in (5) and (7), we obtain $\|\nabla F(x_K)\|^2 \asymp \frac{\log K}{K^{2/3}}$,

As these results show, stronger convergence results can be proved when noise is present in fewer places in the problem. If stochastic noise is present only in the upper-level rather than in both levels, the rate can be improved from $O(k^{-2/7})$ to $O(k^{-2/5})$. In deterministic settings (no noise), we get a rate of $O(k^{-2/3})$. This rate compares to the $O(k^{-1})$ rate that can be obtained with second-order based methods.

4.2. Main Result for Algorithm 2

When we use the momentum-assisting technique, we require the stochastic functions to be well-behaved as well.

Assumption 6. Assumption 1 holds for $f(x, y; \zeta)$ and $g(x, y; \phi)$ with probability 1.

One technical benefit of the momentum technique is that now we no longer require the bounded-gradient assumption w.r.t. x (Assumption 4) or the smoothness of Hessian of f (Assumption 5) for the analysis, as we no longer make use of the smoothness of y_λ^*. We show the following convergence result for Algorithm 2.

Theorem 4.3. *Suppose Assumptions 1-3 and 6 hold. If step-size parameters are chosen such that $\lambda_0 \geq 2l_{f,1}/\mu_g$ and*

$$\beta_k \leq \gamma_k \leq \frac{1}{16 l_{g,1}},\ \xi\alpha_k \leq \frac{1}{l_{F,1}},$$
$$\xi \leq c_\xi \frac{\mu_g}{l_{g,1} l_{*,0}^2},\ \frac{\delta_k}{\lambda_k} \leq \frac{\mu_g \beta_k}{8}, \tag{8a}$$

$$\max\left(2\frac{\gamma_{k-1} - \gamma_k}{\gamma_{k-1}}, c_\eta \frac{l_{g,1}^3}{\mu_g}\gamma_k^2\right) \leq \eta_{k+1} \leq 1,$$
$$\eta_0 = \eta_1 = 1,\ \delta_k/\gamma_k = o(1), \tag{8b}$$

with proper absolute constants $c_\xi, c_\eta > 0$, then for any $K \geq 1$, the iterates generated by Algorithm 2 satisfy

$$\sum_{k=0}^{K-1} \xi\alpha_k \mathbb{E}[\|\nabla F(x_k)\|^2] \leq O_P(1) \cdot \sum_k \xi\alpha_k\lambda_k^{-2}$$
$$+ O_P(\sigma_f^2) \cdot \sum_k \frac{\eta_{k+1}^2}{\gamma_k\lambda_k} + O_P(\sigma_g^2) \cdot \sum_k \frac{\eta_{k+1}^2\lambda_k}{\gamma_k} + O_P(1),$$

where $O_P(1)$ are instance-dependent constants.

The proof of Theorem 4.3 appears in Appendix C. We introduce the following step-size design, consistent with (8).

$$\alpha_k = \frac{c_\alpha}{(k+k_0)^a},\ \gamma_k = \frac{c_\gamma}{(k+k_0)^c},\ \eta_k = (k+1)^{-2c} \tag{9a}$$

$$\xi \leq c_\xi \frac{\mu_g}{l_{g,1} l_{*,0}^2},\ \delta_k = \frac{\gamma_k}{\alpha_k} - \lambda_k,\ \lambda_0 \geq \frac{2l_{f,1}}{\mu_g}, \tag{9b}$$

$$k_0 \geq \frac{128}{\mu_g} \max\left(\xi l_{F,1}, l_{g,1}\sqrt{\frac{c_\eta l_{g,1}}{\mu_g}}\right),$$
$$c_\gamma = \frac{8}{\mu_g k_0^{1-c}},\ c_\alpha = \frac{8}{\mu_g \lambda_0 k_0^{1-a}}, \tag{9c}$$

with some rate constants $a, c \in [0, 1]$ and $a \geq c$. As a corollary, we can obtain faster convergence rates for Algorithm 2 than Algorithm 1.

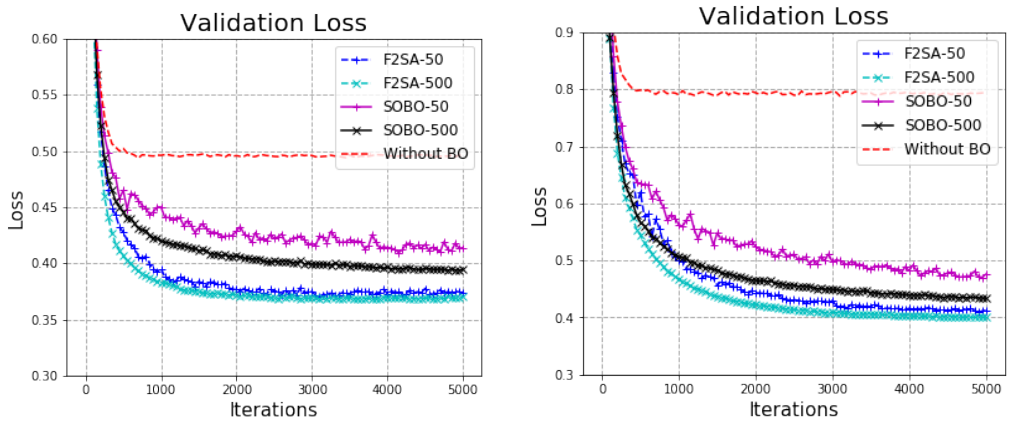
Corollary 4.4. *Suppose the conditions of Theorem 4.3 hold. Suppose that Algorithm 2 is run with step-sizes are designed as in (9). Let R be a random variable drawn from a uniform distribution over $\{0, ..., K-1\}$. Then the following convergence results hold after K iterations of Algorithm 2.*

(a) If stochastic noises are present in both upper-level objective f and lower-level objective g (i.e., $\sigma_f^2, \sigma_g^2 > 0$), then by setting $a = 3/5$ and $c = 2/5$ in (9), we obtain $\mathbb{E}[\|\nabla F(x_R)\|^2] \asymp \frac{\log K}{K^{2/5}}$.

(b) If stochastic noises are present only in f (i.e., $\sigma_f^2 > 0$), $\sigma_g^2 = 0$), then by setting $a = 1/2$ and $c = 1/4$ in (9), we obtain $\mathbb{E}[\|\nabla F(x_R)\|^2] \asymp \frac{\log K}{K^{1/2}}$.

(c) If we have access to exact information about f and g (i.e., $\sigma_f^2 = \sigma_g^2 = 0$), then by setting $a = 1/3$ and $c = 0$ in (9), we obtain $\|\nabla F(x_K)\|^2 \asymp \frac{\log K}{K^{2/3}}$.

The improvements in rates are different in different stochasticity regimes. For instance, the sample complexity required



(a) (b)

Figure 2: Outer objective (validation) loss with label corruption rate: (a) $p = 0.1$, (b) $p = 0.3$.

to achieve ϵ-stationary point is $\tilde{O}_{\mathrm{P}}(\epsilon^{-7/2})$ without momentum and $\tilde{O}_{\mathrm{P}}(\epsilon^{-5/2})$ with momentum — a factor of $O(\epsilon)$ improvement — when stochastic noises are present in both levels. In contrast, when stochastic noises are only in the upper-level objective, then the overall sample complexity is tightened from $\tilde{O}_{\mathrm{P}}(\epsilon^{-5/2})$ to $\tilde{O}_{\mathrm{P}}(\epsilon^{-2})$, an $O(\epsilon^{-0.5})$ improvement. Whether Algorithm 2 achieves the optimal sample complexity for fully first-order methods is an interesting topic for future work.

4.3. Discussion

Because our algorithms do not access second-order derivatives of g, their iteration convergence rate is slower, decreasing from $O(k^{-1/2})$ (e.g., (Chen et al., 2021)) to $O(k^{-2/7})$ for algorithms without momentum and from $O(k^{-2/3})$ (e.g., (Khanduri et al., 2021)) to $O(k^{-2/5})$ for algorithms with momentum. This is not unexpected since we use less information. Our experiments, perhaps surprisingly, do not show a slowdown in the convergence speed. In fact, first-order methods even outperform existing methods that use second-order information of g, as we show in Section 5. We add that in practice, if a bias of $O(1/\lambda^2)$-bias in the solution is not critical to the overall performance, then we can set $\lambda_k := \lambda$ constant at all iterations and choose more aggressive step-sizes, *e.g,* $\alpha_k \asymp k^{-1/2}, \gamma_k \asymp k^{-1/2}$ as in (Chen et al., 2021). Such a strategy yields faster convergence to a certain biased point.

When deterministic gradient oracles are available, the authors in (Ye et al., 2022) employed the so called *dynamic-barrier* method (Gong et al., 2021) to decide the value of λ_k at every iteration, based on $\|\nabla_y g(x_k, z_{k+1}) - \nabla_y g(x_k, y_{k+1})\|$. Such an approach requires precise knowledge of the latter quantity, which is not available in stochastic settings. Our result shows that a simple design of polynomial-rate growth of λ_k is sufficient; an adaptive choice is not needed for good practical performance. Further, the convergence rate reported in (Ye et al., 2022) is $k^{-1/4}$, while our result guarantees $k^{-2/3}$ convergence rate in deterministic settings.

5. Numerical Experiment

We demonstrate the proposed algorithms on a data hyper-cleaning task involving MNIST (Deng, 2012). We are given a noisy training set $\mathcal{D}_{\text{train}} := \{(\tilde{x}_i, \tilde{y}_i)\}_{i=1}^n$ with the label $\tilde{y}_i$ being randomly corrupted with probability $p < 1$. We are also given a small but clean validation set $\mathcal{D}_{\text{val}} := \{(x_i, y_i)\}_{i=1}^m$. The goal is to assign weights to each training data point so that the model trained on the weighted training set yields good performance on the validation set. This task can be formulated as bilevel opimization problem, as follows:

$$\begin{array}{ll} \min\limits_{\lambda} & \sum_{i=1}^m l(x_i, y_i; w^*) \\ \text{s.t.} & w^* \in \arg\min\limits_{w} \sum_{i=1}^n \sigma(\lambda_i) l(\tilde{x}_i, \tilde{y}_i; w) + c\|w\|^2. \end{array}$$

where $\sigma(\cdot)$ is a sigmoid function, $l(x, y; w)$ is a logistic loss function with parameter w and c is a regularization constant. We use $n = 19000$ training samples and $m = 1000$ clean validation samples with regularization parameter $c = 0.01$. We do not include momentum-assisted methods in our discussion, since we do not observe a significant improvement over the $\mathtt{F}^2\mathtt{SA}$ approach of Algorithm 1 .

We demonstrate the performance of Algorithm 1 ($\mathtt{F}^2\mathtt{SA}$) and the second-order based method (SOBO) with batch sizes 50 and 500. We note that several existing second-order methods are in principle the same when momentum or variance-reduction techniques are omitted (Ghadimi & Wang, 2018; Hong et al., 2020; Chen et al., 2021), so we use

the implementation of stocBiO (Ji et al., 2021) as a representative of the other second-order methods. As a baseline, we also add a result from training without bilevel formulation (Without BO), *i.e.,* train on all samples as usual, ignoring the label corruption. Results are shown in Figure 2.[1]

Although iteration complexity is worse for first-order methods than SOBO, we observe that $\mathtt{F^2SA}$ is at least on par with SOBO in this example. It can even give superior performance when the batch size is small. We conjecture that stochastic noises in Hessian become significantly larger than those in gradients, degrading the performance of SOBO. In our experiment, we also observe that the use of a truncated Neumann approximation (Ghadimi & Wang, 2018) for estimating the Hessian-inverse may induce non-negligible bias. In contrast, our fully first-order method $\mathtt{F^2SA}$ is much less sensitive to small batch sizes and free of bias.

6. Conclusion

In this work, we study a fully first-order method for stochastic bilevel optimization and its non-asymptotic convergence behavior. While we focus on well-conditioned bilevel problems, there are already several recent work that considers a more challenging case when the lower-level optimization problem can be non-strongly-convex and non-smooth Liu et al. (2021b;a); Arbel & Mairal (2022). The potential benefit of the first-order method over existing second-order based methods is that it can still be considered to tackle such scenarios, whereas the formula (1) is only available for well-conditioned lower-level problems. We believe it is an important future direction to study a more general class of (**P**) beyond strongly-convex lower-level problems with fully first-order methods. Adding variable-dependent constraints to the lower-level problem would also lead to an interesting extension of fully first-order approaches.

Acknowledgement

This work is partially supported by NSF Awards DMS 2023239 and CCF 2224213, DOE via subcontract 8F-30039 from Argonne National Laboratory, and AFOSR Award FA9550-21-1-0084.

References

Amini, M. and Yousefian, F. An iterative regularized incremental projected subgradient method for a class of bilevel optimization problems. In *2019 American Control Conference (ACC)*, pp. 4069–4074. IEEE, 2019a.

Amini, M. and Yousefian, F. An iterative regularized mirror descent method for ill-posed nondifferentiable stochastic optimization. *arXiv preprint arXiv:1901.09506*, 2019b.

Anandalingam, G. and White, D. A solution method for the linear static stackelberg problem using penalty functions. *IEEE Transactions on automatic control*, 35(10):1170–1173, 1990.

Arbel, M. and Mairal, J. Non-convex bilevel games with critical point selection maps. *arXiv preprint arXiv:2207.04888*, 2022.

Bao, F., Wu, G., Li, C., Zhu, J., and Zhang, B. Stability and generalization of bilevel programming in hyperparameter optimization. *Advances in Neural Information Processing Systems*, 34:4529–4541, 2021.

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

Chen, T., Sun, Y., and Yin, W. Closing the gap: Tighter analysis of alternating stochastic gradient methods for bilevel problems. *Advances in Neural Information Processing Systems*, 34:25294–25307, 2021.

Chen, T., Sun, Y., Xiao, Q., and Yin, W. A single-timescale method for stochastic bilevel optimization. In *International Conference on Artificial Intelligence and Statistics*, pp. 2466–2488. PMLR, 2022.

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

Dagréou, M., Ablin, P., Vaiter, S., and Moreau, T. A framework for bilevel optimization that enables stochastic and global variance reduction algorithms. *arXiv preprint arXiv:2201.13409*, 2022.

Deng, L. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine*, 29(6):141–142, 2012.

Ferris, M. C. and Mangasarian, O. L. Finite perturbation of convex programs. *Applied Mathematics and Optimization*, 23(1):263–273, 1991.

Franceschi, L., Frasconi, P., Salzo, S., Grazzi, R., and Pontil, M. Bilevel programming for hyperparameter optimization and meta-learning. In *International Conference on Machine Learning*, pp. 1568–1577. PMLR, 2018.

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

Gidel, G., Berard, H., Vignoud, G., Vincent, P., and Lacoste-Julien, S. A variational inequality perspective on generative adversarial networks. In *International Conference on Learning Representations*, 2018.

[1] We report our best results obtained with different hyperparameters for each algorithm.

Giovannelli, T., Kent, G., and Vicente, L. N. Bilevel stochastic methods for optimization and machine learning: Bilevel stochastic descent and darts. *arXiv preprint arXiv:2110.00604*, 2021.

Gong, C., Liu, X., and Liu, Q. Automatic and harmless regularization with constrained and lexicographic optimization: A dynamic barrier approach. *Advances in Neural Information Processing Systems*, 34:29630–29642, 2021.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Guo, Z., Hu, Q., Zhang, L., and Yang, T. Randomized stochastic variance-reduced methods for multi-task stochastic bilevel optimization. *arXiv preprint arXiv:2105.02266*, 2021.

Hong, M., Wai, H.-T., Wang, Z., and Yang, Z. A two-timescale framework for bilevel optimization: Complexity analysis and application to actor-critic. *arXiv preprint arXiv:2007.05170*, 2020.

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

Ishizuka, Y. and Aiyoshi, E. Double penalty method for bilevel optimization problems. *Annals of Operations Research*, 34(1):73–88, 1992.

Ji, K., Yang, J., and Liang, Y. Bilevel optimization: Convergence analysis and enhanced design. In *International Conference on Machine Learning*, pp. 4882–4892. PMLR, 2021.

Jiang, R., Abolfazli, N., Mokhtari, A., and Hamedani, E. Y. A conditional gradient-based method for simple bilevel optimization with convex lower-level problem, 2022.

Khanduri, P., Zeng, S., Hong, M., Wai, H.-T., Wang, Z., and Yang, Z. A near-optimal algorithm for stochastic bilevel optimization via double-momentum. *Advances in Neural Information Processing Systems*, 34:30271–30283, 2021.

Konda, V. and Tsitsiklis, J. Actor-critic algorithms. *Advances in neural information processing systems*, 12, 1999.

Kunapuli, G., Bennett, K. P., Hu, J., and Pang, J.-S. Bilevel model selection for support vector machines. *Data mining and mathematical programming*, 45:129–158, 2008.

Liu, R., Liu, X., Yuan, X., Zeng, S., and Zhang, J. A value-function-based interior-point method for non-convex bilevel optimization. In *International Conference on Machine Learning*, pp. 6882–6892. PMLR, 2021a.

Liu, R., Liu, Y., Zeng, S., and Zhang, J. Towards gradient-based bilevel optimization with non-convex followers and beyond. *Advances in Neural Information Processing Systems*, 34:8662–8675, 2021b.

Margossian, C. C. A review of automatic differentiation and its efficient implementation. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, 9(4): e1305, 2019.

Mehra, A. and Hamm, J. Penalty method for inversion-free deep bilevel optimization. In *Asian Conference on Machine Learning*, pp. 347–362. PMLR, 2021.

Nesterov, Y. et al. *Lectures on convex optimization*, volume 137. Springer, 2018.

Rajeswaran, A., Finn, C., Kakade, S. M., and Levine, S. Meta-learning with implicit gradients. *Advances in neural information processing systems*, 32, 2019.

Solodov, M. An explicit descent method for bilevel convex optimization. *Journal of Convex Analysis*, 14(2):227, 2007.

Sow, D., Ji, K., Guan, Z., and Liang, Y. A constrained optimization approach to bilevel optimization with multiple inner minima. *arXiv preprint arXiv:2203.01123*, 2022.

Stackelberg, H. v. et al. Theory of the market economy. 1952.

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press, 2018.

Vicente, L., Savard, G., and Júdice, J. Descent approaches for quadratic bilevel programming. *Journal of Optimization theory and applications*, 81(2):379–399, 1994.

White, D. J. and Anandalingam, G. A penalty function approach for solving bi-level linear programs. *Journal of Global Optimization*, 3(4):397–419, 1993.

Wright, S., Nocedal, J., et al. Numerical optimization. *Springer Science*, 35(67-68):7, 1999.

Yang, J., Ji, K., and Liang, Y. Provably faster algorithms for bilevel optimization. *Advances in Neural Information Processing Systems*, 34:13670–13682, 2021.

Ye, M., Liu, B., Wright, S., Stone, P., and Liu, Q. Bome! bilevel optimization made easy: A simple first-order approach. *arXiv preprint arXiv:2209.08709*, 2022.

A. Auxiliary Lemmas

All deferred proofs in the main text and appendix are directed to Appendix D.

A.1. Additional Notation

Symbol	Meaning	Less than
$l_{f,0}$	Bound of $\|\nabla_x f\|, \|\nabla_y f\|$	$\cdot$
$l_{f,1}$	Smoothness of f	$\cdot$
$l_{g,0}$	Bound of $\|\nabla_x g\|$	$\cdot$
$l_{g,1}$	Smoothness of g	$\cdot$
μ_g	Strong-convexity of g	$\cdot$
$l_{g,2}$	Hessian-continuity of g	$\cdot$
M_f	Second-order moment of $\nabla f(x,y;\zeta)$	$l_{f,0}^2+\sigma_f^2$
M_g	Second-order moment of $\nabla g(x,y;\phi)$	$l_{g,0}^2+\sigma_g^2$
$l_{f,2}$	Hessian-continuity of f (with Assumption 5)	$\cdot$
$l_{F,1}$	Smoothness of $F(x)$	$l_{*,0}\left(l_{f,1}+\frac{l_{g,1}^2}{\mu_g}+\frac{2l_{f,0}l_{g,1}l_{g,2}}{\mu_g^2}\right)$
$l_{\lambda,0}$	Lipschitzness of $y_\lambda^*(x)$ (for all $\lambda \geq 2l_{f,1}/\mu_g$)	$\frac{3l_{g,1}}{\mu_g}$
$l_{\lambda,1}$	Smoothness of $y_\lambda^*(x)$ (for $\lambda \geq 2l_{f,1}/\mu_g$ with Assumption 5)	$32(l_{g,2}+\lambda^{-1}\cdot l_{f,2})\frac{l_{g,1}^2}{\mu_g^3}$
$l_{*,0}$	$=1+\max_{\lambda\geq 2l_{f,1}/\mu_g} l_{\lambda,0}$	$\cdot$
$l_{*,1}$	$=\max_{\lambda\geq 2l_{f,1}/\mu_g} l_{\lambda,1}$	$\cdot$

Table 1: Meaning of Constants

To simplify the representation for the movement of variables, we often use q_k^x, q_k^y and q_k^z defined as

$$\begin{aligned} q_k^x &:= \nabla_x f(x_k, y_{k+1}) + \lambda_k(\nabla_x g(x_k, y_{k+1}) - \nabla_x g(x_k, z_{k+1})), \\ q_{k,t}^y &:= \nabla_y f(x_k, y_{k,t}) + \lambda_k \nabla_y g(x_k, y_{k,t}), \\ q_{k,t}^z &:= \nabla_y g(x_k, z_{k,t}). \end{aligned} \tag{10}$$

The above quantities are the expected movements of $x_k, y_k^{(t)}, z_k^{(t)}$ respectively if there are no stochastic noises in gradient oracles. We also summarize symbols and their meanings for instance-specific constants in Table 1.

A.2. Auxiliary Lemmas

We first state a few lemmas that will be useful in our main proofs.

Lemma A.1. *$F(x) = f(x, y^*(x))$ is $l_{F,1}$-smooth where*

$$l_{F,1} \leq l_{*,0}\left(l_{f,1} + \frac{l_{g,1}^2}{\mu_g} + \frac{2l_{f,0}l_{g,1}l_{g,2}}{\mu_g^2}\right).$$

Lemma A.2. *For any $x, y, \lambda > 0$, the following holds:*

$$\begin{aligned} &\|\nabla F(x) - \nabla_x \mathcal{L}_\lambda(x,y) + \nabla_{xy}^2 g(x, y^*(x))^\top \nabla_{yy}^2 g(x, y^*(x))^{-1} \nabla_y \mathcal{L}_\lambda(x,y)\| \\ &\leq 2(l_{g,1}/\mu_g)\|y - y^*(x)\| \left(l_{f,1} + \lambda \cdot \min(2l_{g,1}, l_{g,2}\|y - y^*(x)\|)\right). \end{aligned}$$

Lemma A.3. *Under Assumptions 1, 2 and 5, and $\lambda > 2l_{f,1}/\mu_g$, a function $y_\lambda^*(x)$ is $l_{\lambda,1}$-smooth: for any $x_1, x_2 \in X$, we have*

$$\|\nabla y_\lambda^*(x_1) - \nabla y_\lambda^*(x_2)\| \leq l_{\lambda,1}\|x_1 - x_2\|$$

where $l_{\lambda,1} \leq 32(l_{g,2} + \lambda^{-1} l_{f,2}) l_{g,1}^2/\mu_g^3$.

Lemma A.4. *For any fixed $\lambda > 2l_{f,1}/\mu_g$, at every k iteration conditioned on $\mathcal{F}_k$, we have*

$$\mathbb{E}[\|y^*(x_{k+1}) - y^*(x_k)\|^2|\mathcal{F}_k] \leq \xi^2 l_{*,0}^2 \left(\alpha_k^2 \mathbb{E}[\|q_k^x\|^2|\mathcal{F}_k] + \alpha_k^2\sigma_f^2 + \beta_k^2\sigma_g^2\right).$$

Lemma A.5. *At every k^{th} iteration, conditioned on $\mathcal{F}_k$, let v_k be a random vector decided before updating x_k. Then for any $\eta_k > 0$, we have*

$$\begin{aligned}\mathbb{E}[\langle v_k, y^*(x_{k+1}) - y^*(x_k)\rangle|\mathcal{F}_k] \leq{}& (\xi\alpha_k\eta_k + M\xi^2 l_{*,1}^2\beta_k^2)\mathbb{E}[\|v_k\|^2|\mathcal{F}_k] \\ &+ \left(\frac{\xi\alpha_k l_{*,0}^2}{4\eta_k} + \frac{\xi^2\alpha_k^2}{4}\right)\mathbb{E}[\|q_k^x\|^2|\mathcal{F}_k] + \frac{\xi^2}{4}(\alpha_k^2\sigma_f^2 + \beta_k^2\sigma_g^2),\end{aligned}$$

where $M := \max\left(l_{f,0}^2 + \sigma_f^2, l_{g,0}^2 + \sigma_g^2\right)$.

Lemma A.6. *Under Assumptions 1-5, at every k^{th} iteration, conditioned on $\mathcal{F}_k$, let v_k be a random vector decided before updating x_k. Then for any $\eta_k > 0$, we have*

$$\begin{aligned}\mathbb{E}[\langle v_k, y_{\lambda_{k+1}}^*(x_{k+1}) - y_{\lambda_k}^*(x_k)\rangle|\mathcal{F}_k] \leq{}& (\delta_k/\lambda_k + \xi\alpha_k\eta_k + M\xi^2 l_{\lambda_k,1}^2\beta_k^2)\mathbb{E}[\|v_k\|^2|\mathcal{F}_k] \\ &+ \left(\frac{\xi\alpha_k l_{*,0}^2}{4\eta_k} + \frac{\xi^2\alpha_k^2}{4}\right)\mathbb{E}[\|q_k^x\|^2|\mathcal{F}_k] + \frac{\xi^2}{4}(\alpha_k^2\sigma_f^2 + \beta_k^2\sigma_g^2) + \frac{\delta_k l_{f,0}^2}{\lambda_k^3\mu_g^2},\end{aligned}$$

where $M := \max\left(l_{f,0}^2 + \sigma_f^2, l_{g,0}^2 + \sigma_g^2\right)$.

B. Main Results for Algorithm 1

In this section, we prove our key estimate, Theorem 4.1. Our aim is to find the upper bound of $\mathbb{V}_{k+1} - \mathbb{V}_k$ for the potential function $\mathbb{V}_k$ given in (4). For x_k and y_k given in Algorithm 1, the following notations will be used:

$$\mathcal{I}_k := \|y_k - y_{\lambda,k}^*\|^2 \text{ and } \mathcal{J}_k := \|z_k - y_k^*\|^2 \tag{11}$$

where $y_{\lambda,k}^* := y_{\lambda_k}^*(x_k)$, $y_k^* := y^*(x_k)$, and $x^* = \arg\min_x F(x)$. Recall that y_λ^* and y^* are given in (2) and (**P**), respectively. Using the above notation, the potential function given in (4) can be rewritten as

$$\mathbb{V}_k := (F(x_k) - F(x^*)) + \lambda_k l_{g,1}\mathcal{I}_k + \frac{\lambda_k l_{g,1}}{2}\mathcal{J}_k \tag{12}$$

for each $k \in \mathbb{N}$. In the following three subsections, we find the upper bound of $\mathbb{V}_{k+1} - \mathbb{V}_k$ in terms of $\mathcal{I}_k$ and $\mathcal{J}_k$. The proof of Theorem 4.1 is given in Section B.4.

B.1. Estimation of $F(x_{k+1}) - F(x_k)$

The step size α_k is designed to satisfy

$$\textbf{(step-size rule):} \qquad \alpha_k \leq \frac{1}{2\xi l_{F,1}}, \tag{13}$$

which is essential to obtain the negative term $-\frac{\xi\alpha_k}{4}\|q_k^x\|^2$ on the right hand side of (15). This negativity plays an important role in the proof of Theorem 4.1 in Section B.4.

On the other hand, we also impose

$$\textbf{(step-size rule):} \qquad \frac{\xi}{T} \leq \frac{\mu_g}{96 l_{g,1}}. \tag{14}$$

The terms, $\|y_{k+1} - y_{\lambda,k}^*\|^2$ and $\|z_{k+1} - y_k^*\|^2$, in the upper bound (15) will be estimated in Lemma B.3 and Lemma B.5, respectively.

Proposition B.1. *Under the step-size rules given in (13), and (14) and $\lambda_k \geq 2l_{f,1}/\mu_g$, it holds that for each $k \in \mathbb{N}$*

$$\mathbb{E}[F(x_{k+1}) - F(x_k)|\mathcal{F}_k] \leq -\frac{\xi\alpha_k}{4}\left(2\|\nabla F(x_k)\|^2 + \|q_k^x\|^2\right) + \frac{T\mu_g\alpha_k\lambda_k^2}{32}\left(2\|y_{k+1} - y_{\lambda,k}^*\|^2 + \|z_{k+1} - y_k^*\|^2\right) + \frac{\xi^2 l_{F,1}}{2}(\alpha_k^2\sigma_f^2 + \beta_k^2\sigma_g^2) + \frac{\xi\alpha_k}{2}\cdot 3C_\lambda^2\lambda_k^{-2}, \tag{15}$$

where q_k^x is given in (10), and $C_\lambda := \frac{4l_{f,0}l_{g,1}}{\mu_g^2}\left(l_{f,1} + \frac{2l_{f,0}l_{g,2}}{\mu_g}\right)$.

Proof. From the smoothness of F,

$$\mathbb{E}[F(x_{k+1}) - F(x_k)|\mathcal{F}_k] \leq \mathbb{E}[\langle\nabla F(x_k), x_{k+1} - x_k\rangle + \frac{l_{F,1}}{2}\|x_{k+1} - x_k\|^2|\mathcal{F}_k].$$

As q_k^x satisfies $\mathbb{E}[x_{k+1} - x_k|\mathcal{F}_k] = \alpha_k q_k^x$,

$$\begin{aligned}\mathbb{E}[F(x_{k+1}) - F(x_k)|\mathcal{F}_k] &= -\xi\alpha_k\langle\nabla_x F(x_k), q_k^x\rangle + \frac{l_{F,1}}{2}\mathbb{E}[\|x_{k+1} - x_k\|^2|\mathcal{F}_k] \\ &= -\frac{\xi\alpha_k}{2}(\|\nabla F(x_k)\|^2 + \|q_k^x\|^2 - \|\nabla F(x_k) - q_k^x\|^2) + \frac{l_{F,1}}{2}\mathbb{E}[\|x_{k+1} - x_k\|^2|\mathcal{F}_k].\end{aligned}$$

Note that

$$\mathbb{E}[\|x_{k+1} - x_k\|^2] \leq \xi^2\alpha_k^2\mathbb{E}[\|q_k^x\|^2 + \xi^2(\alpha_k^2\sigma_f^2 + \beta_k^2\sigma_g^2),$$

and thus with (13) we have

$$\begin{aligned}\mathbb{E}[F(x_{k+1}) - F(x_k)|\mathcal{F}_k] &\leq -\frac{\xi\alpha_k}{2}\|\nabla F(x_k)\|^2 - \frac{\xi\alpha_k}{4}\|q_k^x\|^2 \\ &+ \frac{\xi\alpha_k}{2}\|\nabla F(x_k) - q_k^x\|^2 + \frac{\xi^2 l_{F,1}}{2}(\alpha_k^2\sigma_f^2 + \beta_k^2\sigma_g^2).\end{aligned}$$

Next, we bound $\|\nabla F(x_k) - q_k^x\|$ using the triangle inequality:

$$\begin{aligned}\|q_k^x - \nabla F(x_k)\| &= \|q_k^x - \nabla\mathcal{L}_{\lambda_k}^*(x_k) + \nabla\mathcal{L}_{\lambda_k}^*(x_k) - \nabla F(x_k)\| \\ &\leq \|\nabla_x f(x_k, y_{k+1}) - \nabla_x f(x_k, y_{\lambda,k}^*)\| + \lambda_k\|\nabla_x g(x_k, y_{k+1}) - \nabla_x g(x_k, y_{\lambda,k}^*)\| \\ &\quad + \lambda_k\|\nabla_x g(x_k, z_{k+1}) - \nabla_x g(x_k, y_k^*)\| + \|\nabla\mathcal{L}_{\lambda_k}^*(x_k) - \nabla F(x_k)\|.\end{aligned}$$

From Lemma 3.1, the term $\|\nabla\mathcal{L}_{\lambda_k}^*(x_k) - \nabla F(x_k)\|$ is bounded by C_λ/λ_k. Combining with the regularity of f and g yields the following:

$$\|q_k^x - \nabla F(x_k)\| \leq 2l_{g,1}\lambda_k\|y_{k+1} - y_{\lambda,k}^*\| + l_{g,1}\lambda_k\|z_{k+1} - y_k^*\| + C_\lambda/\lambda_k. \tag{16}$$

Note that $\lambda_k \geq 2l_{f,1}/\mu_g$, and thus $l_{f,1} < l_{g,1}\lambda_k$.

Finally, from Cauchy-Schwartz inequality $(a+b+c)^2 \leq 3(a^2+b^2+c^2)$, we get

$$\begin{aligned}\mathbb{E}[F(x_{k+1}) - F(x_k)|\mathcal{F}_k] &\leq -\frac{\xi\alpha_k}{2}\|\nabla F(x_k)\|^2 - \frac{\xi\alpha_k}{4}\|q_k^x\|^2 \\ &+ \frac{\xi\alpha_k}{2}\cdot 3C_\lambda^2\lambda_k^{-2} + 3\xi\alpha_k l_{g,1}\lambda_k^2\|z_{k+1} - y_k^*\|^2 + 6\xi\alpha_k l_{g,1}\lambda_k^2\|y_{k+1} - y_{\lambda,k}^*\|^2 + \frac{\xi^2 l_{F,1}}{2}(\alpha_k^2\sigma_f^2 + \beta_k^2\sigma_g^2).\end{aligned} \tag{17}$$

The step-size condition (14) concludes our claim. □

B.2. Descent Lemma for y_k towards $y^*_{\lambda,k}$

In this section, the upper bounds of $\mathcal{I}_{k+1}$ and $\|y_{k+1} - y^*_{\lambda,k}\|$ are provided, respectively, in Lemma B.2 and Lemma B.3. The following rule is required to ensure that $\|y_{k+1} - y_{\lambda,k+1}\|^2$ contracts:

$$\textbf{(step-size rule):} \qquad \frac{\delta_k}{\lambda_k} \le \frac{T\beta_k\mu_g}{32}, \text{ and } 2\xi^2 M l^2_{*,1}\beta_k^2 < T\beta_k\mu_g/16. \tag{18}$$

The first condition holds directly from (3b), and the second condition holds since $\beta_k \le \frac{1}{4T\mu_g}$ and also

$$\frac{\xi^2}{T^2} \le \frac{\mu_g^2}{8}(Ml^2_{*,l})^{-1},$$

which also holds by (3b) with sufficiently small c_ξ.

Lemma B.2. *Under the step-size rule (18), it holds that for each $k \in \mathbb{N}$*

$$\begin{aligned}\mathbb{E}[\mathcal{I}_{k+1}|\mathcal{F}_k] &\le (1 + T\beta_k\mu_g/4)\,\mathbb{E}[\|y_{k+1} - y^*_{\lambda,k}\|^2|\mathcal{F}_k] \\ &\quad + O\left(\frac{\xi^2 l^2_{*,0}\alpha_k^2}{\mu_g T\beta_k}\right)\mathbb{E}[\|q_k^x\|^2|\mathcal{F}_k] + O\left(\frac{\delta_k}{\lambda_k^3}\frac{l^2_{f,0}}{\mu_g^2}\right) + O(\xi^2 l^2_{*,0})\cdot(\alpha_k^2\sigma_f^2 + \beta_k^2\sigma_g^2).\end{aligned} \tag{19}$$

where $\mathcal{I}_k$ and q_k^x are given in (11) and (10), respectively.

Proof. We can start from

$$\|y_{k+1} - y^*_{\lambda,k+1}\|^2 = \underbrace{\|y_{k+1} - y^*_{\lambda,k}\|^2}_{(i)} + \underbrace{\|y^*_{\lambda,k+1} - y^*_{\lambda,k}\|^2}_{(ii)} - \underbrace{2\langle y_{k+1} - y^*_{\lambda,k}, y^*_{\lambda,k+1} - y^*_{\lambda,k}\rangle}_{(iii)}.$$

The upper bound of (i) is given in Lemma B.3 below. To bound (ii), we invoke Lemma 3.2 to get

$$\begin{aligned}(ii) : \mathbb{E}[\|y^*_{\lambda,k+1} - y^*_{\lambda,k}\|^2|\mathcal{F}_k] &\le \frac{4\delta_k^2}{\lambda_k^2\lambda_{k+1}^2}\frac{l^2_{f,0}}{\mu_g^2} + l^2_{*,0}\mathbb{E}[\|x_{k+1} - x_k\|^2|\mathcal{F}_k] \\ &\le \frac{4\delta_k^2}{\lambda_k^4}\frac{l^2_{f,0}}{\mu_g^2} + \xi^2 l^2_{*,0}(\alpha_k^2\mathbb{E}[\|q_k^x\|^2] + \alpha_k^2\sigma_f^2 + \beta_k^2\sigma_f^2).\end{aligned}$$

For (iii), recall the smoothness of $y^*_\lambda(x)$ in Lemma A.3, and thus Lemma A.6. By setting $v = y_{k+1} - y^*_{\lambda,k}$ and $\eta_k = T\mu_g\lambda_k/(16\xi)$, and get

$$\begin{aligned}(iii) &\le (2\delta_k/\lambda_k + T\beta_k\mu_g/8 + 2M\xi^2 l^2_{*,1}\beta_k^2)\mathbb{E}[\|y_{k+1} - y^*_{\lambda,k}\|^2|\mathcal{F}_k] \\ &\quad + \xi^2\left(\frac{\alpha_k^2}{2} + \frac{8\alpha_k^2 l^2_{*,0}}{\mu_g T\beta_k}\right)\|q_k^x\|^2 + \frac{\xi^2}{2}(\alpha_k^2\sigma_f^2 + \beta_k^2\sigma_g^2) + \frac{2\delta_k}{\lambda_k^3}\frac{l^2_{f,0}}{\mu_g^3}.\end{aligned}$$

We sum up the $(i), (ii), (iii)$ to conclude

$$\begin{aligned}\mathbb{E}[\mathcal{I}_{k+1}|\mathcal{F}_k] &\le \left(1 + 2\delta_k/\lambda_k + T\beta_k\mu_g/8 + 2M\xi^2 l^2_{*,1}\beta_k^2\right)\mathbb{E}[\|y_{k+1} - y^*_{\lambda,k}\|^2] \\ &\quad + O\left(\frac{\xi^2 l^2_{*,0}\alpha_k^2}{\mu_g T\beta_k}\right)\|q_k^x\|^2 + O\left(\frac{\delta_k}{\lambda_k^3}\frac{l^2_{f,0}}{\mu_g^2}\right) + O(\xi^2 l^2_{*,0})\cdot(\alpha_k^2\sigma_f^2 + \beta_k^2\sigma_g^2).\end{aligned} \tag{20}$$

Lastly, the step-size rule (18) yields our conclusion. □

Next, we note that α_k and β_k are chosen to satisfy

$$\textbf{(step size rules):} \qquad \alpha_k \le \frac{1}{8l_{f,1}} \text{ and } \beta_k \le \frac{1}{8l_{g,1}}, \tag{21}$$

Note that $\beta_k \le \frac{1}{8l_{g,1}}$ is given from the step-size condition (3a), and $\alpha_k \le \frac{1}{8l_{g,1}\lambda_k} \le \frac{1}{8l_{f,1}}$ since $\lambda_k \ge l_{f,1}/\mu_g$.

Lemma B.3. *Under the step-size rule given in (21), it holds that for each $k \in \mathbb{N}$*

$$\mathbb{E}[\|y_{k+1} - y^*_{\lambda,k}\|^2|\mathcal{F}_k] \leq (1 - 3T\mu_g\beta_k/4)\mathcal{I}_k + T(\alpha_k^2\sigma_f^2 + \beta_k^2\sigma_g^2). \tag{22}$$

Proof. Since $\mathbb{E}[y_k^{(t+1)} - y_k^{(t)}|\mathcal{F}_k] = -\alpha_k\nabla_y q_k^{(t)} = -\alpha_k\nabla_y\mathcal{L}_{\lambda_k}(x_k, y_k^{(t)})$, we have

$$\mathbb{E}[\|y_k^{(t+1)} - y^*_{\lambda,k}\|^2|\mathcal{F}_k] = \|y_k^{(t)} - y^*_{\lambda,k}\|^2 - 2\alpha_k\langle\nabla_y q_k^{(t)}, y_k^{(t)} - y^*_{\lambda,k}\rangle + \mathbb{E}[\|y_k^{(t+1)} - y_k^{(t)}\|^2|\mathcal{F}_k].$$

As we start from $\lambda_0 \geq 2\mu_f/\mu_g$, all $\mathcal{L}_k$ is $(\lambda_k\mu_g/2)$-strongly convex in y, and we have

$$\max\left(\frac{\lambda_k\mu_g}{2}\|y_k^{(t)} - y^*_{\lambda,k}\|^2, \frac{1}{l_{f,1} + \lambda_k l_{g,1}}\|\nabla_y q_k^{(t)}\|^2\right) \leq \langle\nabla_y q_k^{(t)}, y_k^{(t)} - y^*_{\lambda,k}\rangle.$$

Using $\mathbb{E}[\|y_k^{(t+1)} - y_k^{(t)}\|^2|\mathcal{F}_k] \leq \alpha_k^2\|\nabla_y q_k^{(t)}\|^2 + \alpha_k^2\sigma_f^2 + \beta_k^2\sigma_g^2$, have

$$(i): \mathbb{E}[\|y_k^{(t+1)} - y^*_{\lambda,k}\|^2|\mathcal{F}_k] \leq (1 - 3\mu_g\beta_k/4)\|y_k^{(t)} - y^*_{\lambda,k}\|^2 + (\alpha_k^2\sigma_f^2 + \beta_k^2\sigma_g^2),$$

where we use $\alpha_k(l_{f,1} + \lambda_k l_{g,1}) = \alpha_k l_{f,1} + \beta_k l_{g,1} \leq 1/4$ if we have (21). Repeating this T times, we get (22). Note that $y_{k+1} = y_k^{(T)}$ and $y_k = y_k^{(0)}$. □

B.3. Descent Lemma for z_k towards y_k^*

Similar to the previous section, we provide the upper bound of $\mathcal{J}_{k+1}$ first and then estimate $\|z_{k+1} - y_k^*\|$ that appears in the upper bound. We work with the following step-size condition:

$$\textbf{(step-size rule):} \qquad 2Ml_{*,1}^2\xi^2\beta_k^2 \leq T\mu_g\gamma_k/16, \tag{23}$$

This condition holds since $\beta_k \leq \gamma_k$, and $\beta_k \leq \frac{1}{4T\mu_g}$ and $\frac{\xi^2}{T^2} \leq \frac{\mu_g^2}{8}(Ml_{*,1}^2)^{-1}$.

Lemma B.4. *Under the step-size rule (23), at each kth iteration, the following holds:*

$$\begin{aligned}\mathbb{E}[\mathcal{J}_{k+1}|\mathcal{F}_k] \leq &\left(1 + \frac{3T\gamma_k\mu_g}{8}\right)\cdot\mathbb{E}[\|z_{k+1} - y_k^*\|^2|\mathcal{F}_k] \\ &+ O\left(\frac{\xi^2\alpha_k^2 l_{*,0}^2}{T\mu_g\gamma_k}\right)\|q_k^x\|^2 + O\left(\xi^2 l_{*,0}^2\right)(\alpha_k^2\sigma_f^2 + \beta_k^2\sigma_g^2).\end{aligned} \tag{24}$$

Proof. We estimate each term in the following simple decomposition.

$$\|z_{k+1} - y^*_{k+1}\|^2 = \underbrace{\|z_{k+1} - y_k^*\|^2}_{(i)} + \underbrace{\|y^*_{k+1} - y_k^*\|^2}_{(ii)} - 2\underbrace{\langle z_{k+1} - y_k^*, y^*_{k+1} - y_k^*\rangle}_{(iii)}.$$

Lemma 3.2 implies that

$$(ii): \mathbb{E}[\|y^*_{k+1} - y_k^*\|^2|\mathcal{F}_k] \leq l_{*,0}^2\xi^2(\alpha_k^2\|\nabla_x q_k\|^2 + \alpha_k^2\sigma_f^2 + \beta_k^2\sigma_g^2).$$

For (iii), we recall Lemma A.5 with $v_k = z_{k+1} - y_k^*$ and $\eta_k = T\mu_g\gamma_k/(8\xi\alpha_k)$, we have

$$\begin{aligned}(iii): \langle z_{k+1} - y_k^*, y^*_{k+1} - y_k^*\rangle \leq &(T\gamma_k\mu_g/8 + M\xi^2 l_{*,1}^2\beta_k^2)\mathbb{E}[\|z_{k+1} - y_k^*\|^2|\mathcal{F}_k] \\ &+ \left(\frac{\xi^2\alpha_k^2}{4} + \frac{2\xi^2\alpha_k^2 l_{*,0}^2}{T\mu_g\gamma_k}\right)\|q_k^x\|^2 + \frac{\xi^2}{4}(\alpha_k^2\sigma_f^2 + \beta_k^2\sigma_g^2).\end{aligned}$$

The above bounds and Lemma B.5 imply that

$$\begin{aligned}\mathbb{E}[\mathcal{J}_{k+1}|\mathcal{F}_k] \leq &\left(1 + \frac{T\gamma_k\mu_g}{4} + 2M\xi^2 l_{*,1}^2\beta_k^2\right)\cdot\mathbb{E}[\|z_{k+1} - y_k^*\|^2|\mathcal{F}_k] \\ &+ \xi^2\alpha_k^2\cdot\left(l_{*,0}^2 + \frac{4l_{*,0}^2}{T\mu_g\gamma_k} + \frac{1}{2}\right)\|q_k^x\|^2 + \xi^2\cdot\left(\frac{1}{2} + l_{*,0}^2\right)(\alpha_k^2\sigma_f^2 + \beta_k^2\sigma_g^2).\end{aligned} \tag{25}$$

Using (23), we conclude. □

Next, γ_k is chosen to satisfy the following step-size rules:

$$\textbf{(step-size rule):} \qquad l_{g,1}\gamma_k \le 1/4, \qquad T\mu_g\gamma_k \le 1/4, \tag{26}$$

which directly comes from (3a).

Lemma B.5. *If (26) holds, then for each $k \in \mathbb{N}$, the following holds:*

$$\mathbb{E}[\|z_{k+1} - y_k^*\|^2|\mathcal{F}_k] \le (1 - 3T\mu_g\gamma_k/4)\mathcal{J}_k + T\gamma_k^2\sigma_g^2. \tag{27}$$

Proof. We analyze one step iteration of the inner loop: for each $t = 0, \cdots, T-1$,

$$\begin{aligned}\|z_k^{(t+1)} - y_k^*\|^2 &= \|z_k^{(t)} - y_k^*\|^2 + \|z_k^{(t+1)} - z_k^{(t)}\|^2 + 2\langle z_k^{(t+1)} - z_k^{(t)}, z_k^{(t)} - y_k^*\rangle \\ &= \|z_k^{(t)} - y_k^*\|^2 + \gamma_k^2\|h_{gz}^{k,t}\|^2 - 2\gamma_k\langle h_{gz}^{k,t}, z_k - y_k^*\rangle.\end{aligned}$$

Here, $z_{k+1} = z_k^{(T)}$ and $z_k = z_k^{(0)}$. Note that $\mathbb{E}[h_{gz}^{k,t}] = \nabla_y g(x_k, z_k^{(t)}) = \nabla_y g_k(z_k^{(t)})$ where $g_k(z_k^{(t)}) := g(x_k, z_k^{(t)})$. Taking expectation,

$$\mathbb{E}[\|z_k^{(t+1)} - y_k^*\|^2|\mathcal{F}_k] \le \|z_k^{(t)} - y_k^*\|^2 + \gamma_k^2\|\nabla g_k(z_k^{(t)})\|^2 + \gamma_k^2\sigma_g^2 - 2\gamma_k\langle\nabla g_k(z_k^{(t)}), z_k^{(t)} - y_k^*\rangle.$$

The strong convexity and smoothness of g_k imply the coercivity and co-coercivity (Nesterov et al., 2018), that is,

$$\max\left(\mu_g\|z_k^{(t)} - y_k^*\|^2, \frac{1}{l_{g,1}}\|\nabla g_k(z_k^{(t)}) - \nabla g_k(y_k^*)\|^2\right) \le \langle\nabla g_k(z_k^{(t)}) - \nabla g_k(y_k^*), z_k^{(t)} - y_k^*\rangle.$$

Note that y_k^* minimizes $g_k(y)$. Use this to cancel out $\gamma_k^2\|\nabla g_k(z_k^{(t)})\|^2$, yielding

$$\begin{aligned}\mathbb{E}[\|z_k^{(t+1)} - y_k^*\|^2|\mathcal{F}_k] &\le \|z_k^{(t)} - y_k^*\|^2 + \gamma_k^2\sigma_g^2 - \gamma_k(1 - l_{g,1}\gamma_k)\langle\nabla g_k(z_k^{(t)}), z_k^{(t)} - y_k^*\rangle \\ &\le (1 - 3\mu_g\gamma_k/4)\|z_k^{(t)} - y_k^*\|^2 + \gamma_k^2\sigma_g^2.\end{aligned}$$

For this to hold, we need a step-size condition (26). We can repeat this relation for T times, and we get (27). □

B.4. Proof of Theorem 4.1

Recall $\mathbb{V}_k$ given in (4). In what follows, we examine

$$\begin{aligned}\mathbb{V}_{k+1} - \mathbb{V}_k = F(x_{k+1}) - F(x_k) + \lambda_{k+1}l_{g,1}\mathcal{I}_{k+1} - \lambda_k l_{g,1}\mathcal{I}_k \\ + \frac{\lambda_{k+1}l_{g,1}}{2}\mathcal{J}_{k+1} - \frac{\lambda_k l_{g,1}}{2}\mathcal{J}_k.\end{aligned}$$

Using the estimate of $F(x_{k+1}) - F(x_k)$ given in Proposition B.1 and rearranging the terms, we have

$$\begin{aligned}\mathbb{E}[\mathbb{V}_{k+1} - \mathbb{V}_k|\mathcal{F}_k] \le &-\frac{\xi\alpha_k}{2}\|\nabla F(x_k)\|^2 - \frac{\xi\alpha_k}{4}\mathbb{E}[\|q_k^x\|^2|\mathcal{F}_k] + \frac{\xi\alpha_k}{2}\cdot 3C_\lambda^2\lambda_k^{-2} + \frac{\xi^2 l_{F,1}}{2}(\alpha_k^2\sigma_f^2 + \beta_k^2\sigma_g^2) \\ &+ l_{g,1}\underbrace{\mathbb{E}[\lambda_{k+1}\mathcal{I}_{k+1} + \frac{\lambda_k T\beta_k\mu_g}{16}\|y_{k+1} - y_{\lambda,k}^*\|^2 - \lambda_k\mathcal{I}_k|\mathcal{F}_k]}_{(i)} \\ &+ \frac{l_{g,1}}{2}\underbrace{\mathbb{E}[\lambda_{k+1}\mathcal{J}_{k+1} + \frac{\lambda_k T\gamma_k\mu_g}{32}\|z_{k+1} - y_k^*\|^2 - \lambda_k\mathcal{J}_k|\mathcal{F}_k]}_{(ii)}\end{aligned}$$

Estimation of (i): From Lemma B.2, and $\lambda_{k+1} = \lambda_k + \delta_k$ yield that

$$(i) \le \lambda_k\left(1 + \frac{5T\beta_k\mu_g}{16} + \frac{\delta_k}{\lambda_k}\right)\mathbb{E}[\|y_{k+1} - y_{\lambda,k}^*\|^2|\mathcal{F}_k] - \lambda_k\mathcal{I}_k$$

$$\underbrace{+ O(\xi^2 l_{\lambda,0}^2)\frac{\lambda_k \alpha_k^2}{\mu_g T \beta_k}\|q_k^x\|^2 + O(\xi^2 l_{*,0}^2)\lambda_k(\alpha_k^2\sigma_f^2 + \beta_k^2\sigma_g^2) + O\left(\frac{l_{f,0}^2}{\mu_g^3}\right)\cdot\frac{\delta_k}{\lambda_k^2}}_{(iii)}.$$

Given the step-size rules (18), we obtain

$$(i) \le \lambda_k\left(1 + \frac{T\beta_k\mu_g}{2}\right)\mathbb{E}[\|y_{k+1} - y_{\lambda,k}^*\|^2|\mathcal{F}_k] - \lambda_k\mathcal{I}_k + (iii).$$

The estimation of $\|y_{k+1} - y_{\lambda,k}^*\|^2$ from Lemma B.3 yields that

$$\begin{aligned}(i) &\le -\frac{\lambda_k T\mu_g\beta_k}{4}\mathcal{I}_k + O(\xi^2 l_{*,0}^2)\frac{\alpha_k}{\mu_g T}\|q_k^x\|^2 + (iii),\\ &= -\frac{\lambda_k T\mu_g\beta_k}{4}\mathcal{I}_k + O(\xi^2 l_{*,0}^2)\frac{\alpha_k}{\mu_g T}\|q_k^x\|^2 + O(T + \xi^2 l_{*,0}^2)\lambda_k(\alpha_k^2\sigma_f^2 + \beta_k^2\sigma_g^2) + O\left(\frac{l_{f,0}^2}{\mu_g^3}\right)\cdot\frac{\delta_k}{\lambda_k^2}.\end{aligned}$$

Here, we use $(1 + a/2)(1 - 3a/4) \le 1 - a/4$ for $a > 0$.

Estimation of (ii)**:** Lemma B.4 yields that

$$\begin{aligned}(ii) \le \lambda_k&\left(1 + \frac{\delta_k}{\lambda_k} + \frac{3T\gamma_k\mu_g}{8} + \frac{\lambda_k T\beta_k\mu_g}{32}\right)\mathbb{E}[\|z_{k+1} - y_k^*\|^2|\mathcal{F}_k] - \lambda_k\mathcal{J}_k\\ &\underbrace{+ O(\xi^2 l_{*,0}^2)\frac{\lambda_{k+1}\alpha_k^2}{T\mu_g\gamma_k}\|q_k^x\|^2 + O(\xi^2\lambda_{k+1}l_{*,0}^2)(\alpha_k^2\sigma_f^2 + \beta_k^2\sigma_g^2)}_{(iv)}.\end{aligned}$$

With $\beta_k \le \gamma_k$, and thus $\delta_k/\lambda_k < T\mu_g\gamma_k/32$, we have that

$$(ii) \le \lambda_k\left(1 + \frac{T\gamma_k\mu_g}{2}\right)\mathbb{E}[\|z_{k+1} - y_k^*\|^2|\mathcal{F}_k] - \lambda_k\mathcal{J}_k + (iv)$$

Similar to the argument for (i) above, Lemma B.5 yields

$$(ii) \le -\frac{\lambda_k T\mu_g\gamma_k}{4}\mathcal{J}_k + O(\xi^2 l_{*,0}^2)\frac{\alpha_k\beta_k}{T\mu_g\gamma_k}\|q_k^x\|^2 + O(\xi^2\lambda_k l_{*,0}^2)(\alpha_k^2\sigma_f^2 + \beta_k^2\sigma_g^2) + O(\lambda_k)T\gamma_k^2\sigma_g^2.$$

Plug the bound for (i) and (ii), after rearranging terms, we get

$$\begin{aligned}\mathbb{E}[\mathbb{V}_{k+1} - \mathbb{V}_k|\mathcal{F}_k] \le &-\frac{\xi\alpha_k}{2}\|\nabla F(x_k)\|^2 + \frac{\xi\alpha_k}{2}\cdot 3C_\lambda^2\lambda_k^{-2} + \frac{\xi^2 l_{F,1}}{2}(\alpha_k^2\sigma_f^2 + \beta_k^2\sigma_g^2)\\ &- \frac{\xi\alpha_k}{4}\left(1 - O\left(\frac{\xi l_{g,1}l_{*,0}^2\beta_k}{\mu_g T\gamma_k}\right) - O\left(\frac{\xi l_{g,1}l_{*,0}^2}{\mu_g T}\right)\right)\mathbb{E}[\|q_k^x\|^2|\mathcal{F}_k]\\ &- \frac{\lambda_k l_{g,1}T\mu_g\beta_k}{4}\mathcal{I}_k - \frac{\lambda_k l_{g,1}T\mu_g\gamma_k}{4}\mathcal{J}_k\\ &+ O(T + \xi^2 l_{*,0}^2)\cdot l_{g,1}\lambda_k(\alpha_k^2\sigma_f^2 + (\beta_k^2 + \gamma_k^2)\sigma_g^2) + O\left(\frac{l_{g,1}l_{f,0}^2}{\mu_g^3}\right)\frac{\delta_k}{\lambda_k^2},\end{aligned}$$

A crucial step here is to ensure that terms driven by $\mathbb{E}[\|q_k^x\|^2]$ is negative. To ensure this, we require

$$\textbf{(step-size rules):}\qquad \begin{aligned}&\xi l_{g,1}l_{*,0}^2\beta_k \le c_1\mu_g T\gamma_k,\\ &\xi l_{g,1}l_{*,0}^2 \le c_2\mu_g T,\end{aligned}$$

for some absolute constants $c_1, c_2 > 0$, which holds by $\beta_k \le \gamma_k$ and (3b) with sufficiently small $c_\xi > 0$. Once this holds, we can conclude that

$$\begin{aligned}\mathbb{E}[\mathbb{V}_{k+1} - \mathbb{V}_k|\mathcal{F}_k] \le &-\frac{\xi\alpha_k}{2}\|\nabla F(x_k)\|^2 - \frac{\lambda_k T\mu_g\gamma_k}{4}\|z_k - y_k^*\|^2 - \frac{\lambda_k T\mu_g\beta_k}{4}\|y_k - y_{\lambda,k}^*\|^2 \\ &+ O(\xi C_\lambda^2)\frac{\alpha_k}{\lambda_k^2} + O\left(\frac{l_{g,1}l_{f,0}^2}{\mu_g^3}\right)\frac{\delta_k}{\lambda_k^2} + O(\xi^2 l_{F,1})(\alpha_k^2\sigma_f^2 + \beta_k^2\sigma_g^2) \\ &+ O(T + \xi^2 l_{*,0}^2)\cdot l_{g,1}\lambda_k(\alpha_k^2\sigma_f^2 + (\beta_k^2 + \gamma_k^2)\sigma_g^2).\end{aligned}$$

We can sum over $k = 0$ to $K - 1$, and leaving only dominating terms, since $\sum_k \delta_k/\lambda_k^2 = O(1)$ (because $\delta_k/\lambda_k = O(1/k)$ and $\lambda_k = \mathrm{poly}(k)$), we have the theorem.

B.5. Proof of Corollary 4.2

We first show that with the step-size design in theorem, $\lambda_k = \gamma_k/(2\alpha_k)$ for all k. To check this, by design, $\lambda_0 = \gamma_0/(2\alpha_0)$ and by mathematical induction,

$$\frac{T\mu_g}{16}\alpha_k\lambda_k^2 = \frac{T}{32}\frac{c_\gamma}{2c_\alpha}(k + k_0)^{-2c+a},$$

and

$$\frac{c_\gamma}{2c_\alpha}((k + k_0 + 1)^{a-c} - (k + k_0)^{a-c}) \le \frac{(a - c)c_\gamma}{2c_\alpha}(k + k_0)^{-1-c+a}.$$

As long as $-2c + a \ge -1 - c + a$, or equivalently, $c \le 1$ and $T \ge 32$, it always holds that

$$\lambda_{k+1} = \frac{c_\gamma}{2c_\alpha}(k + k_0 + 1)^{a-c} = \frac{\gamma_{k+1}}{2\alpha_{k+1}}. \tag{28}$$

Now applying the step-size designs, we obtain the following:

$$\begin{aligned}\sum_{k=0}^{K-1}\frac{\mathbb{E}[\|\nabla F(x_k)\|^2]}{(k + k_0)^a} \le &\, O_{\mathrm{P}}(1)\cdot\sum_k\frac{1}{(k + k_0)^{3a-2c}} + O_{\mathrm{P}}(\sigma_f^2)\cdot\sum_k\frac{1}{(k + k_0)^{a+c}} \\ &+ O_{\mathrm{P}}(\sigma_g^2)\cdot\sum_k\frac{1}{(k + k_0)^{3c-a}} + O_{\mathrm{P}}(1).\end{aligned} \tag{29}$$

We decide the rates $a, c \in [0, 1]$ will be decided differently for different stochasticity. Let $b = a - c$. Note that with the step size deisng, we have $\lambda_k = \gamma_k/(2\alpha_k) = \frac{2\lambda_0}{k_0^{a-c}}(k + k_0)^{a-c} = O(k^b)$. Let R be a random variable uniformly distributed over $\{0, 1, ..., K\}$. Note that the left hand side is larger than

$$\frac{K}{(K + k_0)^a}\sum_{k=1}^{K-1}\frac{1}{K}\mathbb{E}[\|\nabla F(x_k)\|^2] \ge K^{1-a}\cdot\mathbb{E}[\|\nabla F(x_R)\|^2].$$

We consider three regimes:

Stochasticity in both upper-level and lower-level objectives: $\sigma_f^2, \sigma_g^2 > 0$**.** In this case, we set $a = 5/7, c = 4/7$, and thus $\lambda_k = k^{1/7}$. The dominating term is $\sigma_g^2\cdot\sum_k(\gamma_k^2\lambda_k) = \sum_k O(k^{-1}) = O(\log K)$ and $C_\lambda^2\cdot\sum_k(\alpha_k\lambda_k^{-2}) = O(\log K)$. From the left-hand side, we have $K^{1-a} = K^{2/7}$. Therefore,

$$\mathbb{E}[\|\nabla F(x_R)\|^2] = O\left(\frac{\log K}{K^{2/7}}\right).$$

Stochasticity only in the upper-level: $\sigma_f^2 > 0, \sigma_g^2 = 0$**.** In this case, we can take $a = 3/5, c = 2/5$. When $\sigma_g^2 = 0$, the dominating term is $\sigma_f\cdot\sum_k(\alpha_k^2\lambda_k) = \sum_k O(k^{-1}) = O(\log K)$ and $O(C_\lambda^2)\cdot\sum_k(\alpha_k\lambda_k^{-2}) = \sum_k O(k^{-1}) = O(\log K)$. Since $K^{1-a} = O(K^{2/5})$, yielding

$$\mathbb{E}[\|\nabla F(x_R)\|^2] = O\left(\frac{\log K}{K^{2/5}}\right).$$

Deterministic case: $\sigma_f^2 = 0, \sigma_g^2 = 0$. Here, we can take $a = 1/3, c = 0$ with a dominating term $\sum_k (\alpha_k \lambda_k^{-2}) = O(\log K)$. Since there is no stochasticity in the algorithm, we have

$$\|\nabla F(x_K)\|^2 = O\left(\frac{\log K}{K^{2/3}}\right).$$

C. Main Results for Algorithm 2

We start with a few definitions and additional auxiliary lemmas. We first define the momentum-assisted moving direction of variables. They can be recursively defined as

$$\begin{aligned}
\tilde{h}_z^k &:= \nabla_y g(x_k, z_k; \phi_z^k) + (1-\eta_k)\left(\tilde{h}_z^{k-1} - \nabla_y g(x_{k-1}, z_{k-1}; \phi_z^k)\right), \\
\tilde{h}_{fy}^k &:= \nabla_y f(x_k, y_k; \zeta_y^k) + (1-\eta_k)\left(\tilde{h}_{fy}^{k-1} - \nabla_y f(x_{k-1}, y_{k-1}; \zeta_y^k)\right), \\
\tilde{h}_{gy}^k &:= \nabla_y g(x_k, y_k; \phi_y^k) + (1-\eta_k)\left(\tilde{h}_{gy}^{k-1} - \nabla_y g(x_{k-1}, y_{k-1}; \phi_y^k)\right),
\end{aligned}$$

for the inner variable updates, and

$$\begin{aligned}
\tilde{h}_{fx}^k &:= \nabla_x f(x_k, y_{k+1}; \zeta_x^k) + (1-\eta_k)\left(\tilde{h}_{fx}^{k-1} - \nabla_x f(x_{k-1}, y_k; \zeta_x^k)\right), \\
\tilde{h}_{gxy}^k &:= \nabla_x g(x_k, y_{k+1}; \phi_x^k) + (1-\eta_k)\left(\tilde{h}_{gxy}^{k-1} - \nabla_x g(x_{k-1}, y_k; \phi_x^k)\right), \\
\tilde{h}_{gxz}^k &:= \nabla_x g(x_k, z_{k+1}; \phi_x^k) + (1-\eta_k)\left(\tilde{h}_{gxz}^{k-1} - \nabla_x g(x_{k-1}, z_k; \phi_x^k)\right).
\end{aligned}$$

for the outer variable update with some proper choice of η_k. We also define stochastic error terms incurred by random sampling:

$$\begin{aligned}
\tilde{e}_k^x &:= \tilde{h}_{fx}^k + \lambda_k(\tilde{h}_{gxy}^k - \tilde{h}_{gxz}^k) - q_k^x, \\
\tilde{e}_k^y &:= (\tilde{h}_{fy}^k + \lambda_k \tilde{h}_{gy}^k) - q_k^y, \\
\tilde{e}_k^z &:= \tilde{h}_z^k - q_k^z,
\end{aligned} \tag{30}$$

where q_k^x, q_k^y, q_k^z are defined in (10) (we dropped t from subscript since here we consider $T = 1$).

C.1. Additional Auxiliary Lemmas

The following lemmas are analogous of Lemma A.4.

Lemma C.1. *At every k iteration conditioned on $\mathcal{F}_k$, we have*

$$\mathbb{E}[\|y^*(x_{k+1}) - y^*(x_k)\|^2 | \mathcal{F}_k] \le 2\xi^2 l_{*,0}^2 \alpha_k^2 \left(\mathbb{E}[\|q_k^x\|^2 | \mathcal{F}_k] + \mathbb{E}[\|\tilde{e}_k^x\|^2]\right).$$

Lemma C.2. *At every k iteration conditioned on $\mathcal{F}_k$, we have*

$$\mathbb{E}[\|y_{\lambda_{k+1}}^*(x_{k+1}) - y_{\lambda_k}^*(x_k)\|^2 | \mathcal{F}_k] \le 4\xi^2 l_{*,0}^2 \alpha_k^2 \left(\mathbb{E}[\|q_k^x\|^2 | \mathcal{F}_k] + \mathbb{E}[\|\tilde{e}_k^x\|^2]\right) + \frac{8\delta_k^2 l_{f,0}^2}{\lambda_k^4 \mu_g^2}.$$

C.2. Descent Lemma for Noise Variances

A major change in the proof is that now we also track the decrease in stochastic error terms. Specifically, we show the following lemmas.

Lemma C.3.

$$\begin{aligned}
\mathbb{E}[\|\tilde{e}_{k+1}^z\|^2] &\le (1-\eta_{k+1})^2(1 + 8l_{g,1}^2\gamma_k^2)\mathbb{E}[\|\tilde{e}_k^z\|^2] + 2\eta_{k+1}^2\sigma_g^2 \\
&\quad + 8l_{g,1}^2(1-\eta_{k+1})^2\left(\xi^2\alpha_k^2\mathbb{E}[\|q_k^x\|^2] + \xi^2\alpha_k^2\mathbb{E}[\|\tilde{e}_k^x\|^2] + \gamma_k^2\mathbb{E}[\|q_k^z\|^2]\right), \\
\mathbb{E}[\|\tilde{e}_{k+1}^y\|^2] &\le (1-\eta_{k+1})^2(1 + 96l_{g,1}^2\beta_k^2)\mathbb{E}[\|\tilde{e}_k^y\|^2] + 2\eta_{k+1}^2(\sigma_f^2 + \lambda_{k+1}^2\sigma_g^2) + 12\delta_k^2\sigma_g^2 \\
&\quad + 96l_{g,1}^2(1-\eta_{k+1})^2\beta_k^2(\xi^2\|q_k^x\|^2 + \xi^2\|\tilde{e}_k^x\|^2 + \|q_k^y\|^2).
\end{aligned}$$

Lemma C.4.

$$\mathbb{E}[\|\tilde{e}^x_{k+1}\|^2] \le (1-\eta_{k+1})^2(1+240l_{g,1}^2\xi^2\beta_k^2)\mathbb{E}[\|\tilde{e}^x_k\|^2] + 6\eta_{k+1}^2(\sigma_f^2+\lambda_{k+1}^2\sigma_g^2) + 80\delta_k^2\sigma_g^2 \\ + 240l_{g,1}^2(1-\eta_{k+1})^2\lambda_k^2\left(\xi^2\alpha_k^2\|q_k^x\|^2 + \alpha_k^2(\|q_k^y\|^2+\|\tilde{e}_k^y\|^2) + \gamma_k^2(\|q_k^z\|^2+\|\tilde{e}_k^z\|^2)\right).$$

Equipped with these lemmas, we can now proceed as previously in the main proof for Algorithm 1.

C.3. Descent Lemma for z_k towards y_k^*

Lemma C.5. *If $\gamma_k\mu_g < 1/8$, then*

$$\mathbb{E}[\|z_{k+1}-y_{k+1}^*\|^2|\mathcal{F}_k] \le (1+\gamma_k\mu_g/4)\mathbb{E}[\|z_{k+1}-y_k^*\|^2|\mathcal{F}_k] \\ + O\left(\frac{\xi^2\alpha_k^2 l_{*,0}^2}{\gamma_k\mu_g}\right)\cdot(\mathbb{E}[\|q_k^x\|^2|\mathcal{F}_k]+\mathbb{E}[\|\tilde{e}_k^x\|^2|\mathcal{F}_k]).$$

Proof. As before, we can decompose $\|z_{k+1}-y_{k+1}^*\|^2$ as

$$\|z_{k+1}-y_{k+1}^*\|^2 = \|z_{k+1}-y_k^*\|^2 + \|y_{k+1}^*-y_k^*\|^2 - 2\langle z_{k+1}-y_k^*, y_{k+1}^*-y_k^*\rangle \\ \le \|z_{k+1}-y_k^*\|^2 + \left(1+\frac{1}{8\gamma_k\mu_g}\right)\|y_{k+1}^*-y_k^*\|^2 + 4\gamma_k\mu_g\|z_{k+1}-y_k^*\|^2,$$

where we used a general inequality $|\langle a,b\rangle| \le c\|a\|^2 + \frac{1}{4c}\|b\|^2$. We can apply Lemma C.1 for $\|y_{k+1}^*-y_k^*\|^2$, yielding the lemma. □

Lemma C.6. *If $\gamma_k \le 1/(16l_{g,1})$, then*

$$\mathbb{E}[\|z_{k+1}-y_k^*\|^2|\mathcal{F}_k] \le (1-\gamma_k\mu_g/2)\mathbb{E}[\|z_k-y_k^*\|^2|\mathcal{F}_k] - \frac{\gamma_k}{l_{g,1}}\|q_k^z\|^2 + O\left(\frac{\gamma_k}{\mu_g}\right)\mathbb{E}[\|\tilde{e}_k^z\|^2|\mathcal{F}_k]).$$

Proof. Note that

$$\|z_{k+1}-y_k^*\|^2 = \|z_k-y_k^*\|^2 + \gamma_k^2\|\tilde{h}_z^k\|^2 - 2\gamma_k\langle\tilde{h}_z^k, z_k-y_k^*\rangle \\ \le \|z_k-y_k^*\|^2 + 2\gamma_k^2(\|q_k^z\|^2+\|\tilde{e}_k^z\|^2) - 2\gamma_k\langle q_k^z, z_k-y_k^*\rangle - 2\gamma_k\langle\tilde{e}_k^z, z_k-y_k^*\rangle.$$

Since $q_k^z = \nabla_y g(x_k,z_k)$ by definition, by coercivity and co-coercivity of strongly-convex functions, we have

$$\left(\mu_g\|z_k-y_k^*\|^2, \frac{1}{l_{g,1}}\|q_k^z\|^2\right) \le \langle q_k^z, z_k-y_k^*\rangle,$$

and thus, given $\gamma_k \le 1/(16l_{g,1})$, we have

$$\mathbb{E}[\|z_{k+1}-y_k^*\|^2|\mathcal{F}_k] \le (1-3\gamma_k\mu_g/4)\mathbb{E}[\|z_k-y_k^*\|^2|\mathcal{F}_k] - \frac{\gamma_k}{l_{g,1}}\|q_k^z\|^2 + 2\gamma_k^2\mathbb{E}[\|\tilde{e}_k^z\|^2|\mathcal{F}_k] - 2\gamma_k\langle\tilde{e}_k^z, z_k-y_k^*\rangle.$$

Finally, we can use general inequality $|\langle a,b\rangle| \le c\|a\|^2 + \frac{1}{4c}\|b\|^2$ to get

$$-2\gamma_k\langle\tilde{e}_k^z, z_k-y_k^*\rangle \le \frac{\gamma_k\mu_g}{4}\|z_k-y_k^*\|^2 + \frac{4\gamma_k}{\mu_g}\|\tilde{e}_k^z\|^2.$$

Plugging this back, with $\gamma_k^2 \ll \frac{\gamma_k}{\mu_g}$, we get the lemma. □

C.4. Descent Lemma for y_k towards $y^*_{\lambda,k}$

Lemma C.7. *If $\beta_k\mu_g < 1/8$, then*

$$\mathbb{E}[\|y_{k+1} - y^*_{\lambda,k+1}\|^2|\mathcal{F}_k] \le (1+\beta_k\mu_g/4)\mathbb{E}[\|y_{k+1} - y^*_{\lambda,k}\|^2|\mathcal{F}_k] + O\left(\frac{\xi^2\alpha_k^2 l_{*,0}^2}{\beta_k\mu_g}\right)\cdot(\mathbb{E}[\|q_k^x\|^2|\mathcal{F}_k] + \mathbb{E}[\|\tilde{e}_k^x\|^2|\mathcal{F}_k]) + O\left(\frac{\delta_k^2 l_{f,0}^2}{\lambda_k^4\mu_g^2}\right).$$

Proof. As before, we can decompose $\|y_{k+1} - y^*_{\lambda,k+1}\|^2$ as

$$\begin{aligned}\|y_{k+1} - y^*_{\lambda,k+1}\|^2 &= \|y_{k+1} - y^*_{\lambda,k}\|^2 + \|y^*_{k+1} - y^*_{\lambda,k}\|^2 - 2\langle y_{k+1} - y^*_{\lambda,k}, y^*_{\lambda,k+1} - y^*_{\lambda,k}\rangle \\ &\le \|y_{k+1} - y^*_{\lambda,k}\|^2 + \left(1 + \frac{1}{8\beta_k\mu_g}\right)\|y^*_{\lambda,k+1} - y^*_{\lambda,k}\|^2 + 4\beta_k\mu_g\|y_{k+1} - y^*_{\lambda,k}\|^2,\end{aligned}$$

where we used a general inequality $|\langle a,b\rangle| \le c\|a\|^2 + \frac{1}{4c}\|b\|^2$. We can apply Lemma C.2 for $\|y^*_{\lambda,k+1} - y^*_{\lambda,k}\|^2$, since $\beta_k\mu_g \le 1/16$, we get the lemma. □

Lemma C.8. *If $\beta_k \le 1/(16 l_{g,1})$, then*

$$\mathbb{E}[\|y_{k+1} - y^*_{\lambda,k}\|^2|\mathcal{F}_k] \le (1-\beta_k\mu_g/2)\mathbb{E}[\|y_k - y^*_{\lambda,k}\|^2|\mathcal{F}_k] - \frac{\alpha_k}{\lambda_k l_{g,1}}\|q_k^y\|^2 + O\left(\frac{\alpha_k^2}{\mu_g\beta_k}\right)\mathbb{E}[\|\tilde{e}_k^y\|^2|\mathcal{F}_k]).$$

Proof. Note that

$$\|y_{k+1} - y^*_{\lambda,k}\|^2 = \|y_k - y^*_{\lambda,k}\|^2 + 2\alpha_k^2(\|q_k^y\|^2 + \|\tilde{e}_k^y\|^2) - 2\alpha_k\langle q_k^y, y_k - y^*_{\lambda,k}\rangle - 2\alpha_k\langle \tilde{e}_k^y, y_k - y^*_{\lambda,k}\rangle,$$

where we used $y_{k+1} - y_k = q_k^y + \tilde{e}_k^y$. Since $q_k^y = \nabla_y\mathcal{L}_{\lambda_k}$ by definition, again by coercivity and co-coercivity of strongly-convex $\mathcal{L}_{\lambda_k}(x_k,\cdot)$, we have

$$\max\left(\frac{\lambda_k\mu_g}{2}\|y_k - y^*_{\lambda,k}\|^2, \frac{1}{l_{f,1} + \lambda_k l_{g,1}}\|q_k^y\|^2\right) \le \langle q_k^y, y_k - y^*_{\lambda,k}\rangle,$$

and thus, given $\alpha_k\lambda_k \le 1/(16 l_{g,1})$ and $l_{f,1} \le \lambda_k l_{g,1}$, we have

$$\begin{aligned}\mathbb{E}[\|y_{k+1} - y^*_{\lambda,k}\|^2|\mathcal{F}_k] &\le (1 - 3\beta_k\mu_g/4)\mathbb{E}[\|z_k - y_k^*\|^2|\mathcal{F}_k] - \frac{\alpha_k}{\lambda_k l_{g,1}}\|q_k^y\|^2 \\ &\quad + 2\alpha_k^2\mathbb{E}[\|\tilde{e}_k^y\|^2|\mathcal{F}_k] - 2\alpha_k\mathbb{E}[\langle\tilde{e}_k^y, y_k - y^*_{\lambda,k}\rangle|\mathcal{F}_k].\end{aligned}$$

Finally, we can use general inequality $|\langle a,b\rangle| \le c\|a\|^2 + \frac{1}{4c}\|b\|^2$ to get

$$-2\alpha_k\langle\tilde{e}_k^y, y_k - y^*_{\lambda,k}\rangle \le \frac{\beta_k\mu_g}{4}\|y_k - y^*_{\lambda,k}\|^2 + \frac{4\alpha_k^2}{\beta_k\mu_g}\|\tilde{e}_k^y\|^2.$$

Plugging this back, with $\beta_k\mu_g \ll 1$, we get the lemma. □

C.5. Descent Lemma for $F(x_k)$

Lemma C.9. *If $\xi\alpha_k l_{F,1} < 1$, then*

$$\begin{aligned}\mathbb{E}[F(x_{k+1}) - F(x_k)|\mathcal{F}_k] &\le -\frac{\xi\alpha_k}{4}\|\nabla F(x_k)\|^2 - \frac{\xi\alpha_k}{4}\mathbb{E}[\|q_k^x\|^2|\mathcal{F}_k] + 2\xi\alpha_k\cdot\mathbb{E}[\|\tilde{e}_k^x\|^2|\mathcal{F}_k] \\ &\quad + \frac{3\xi\alpha_k}{2}\left(4l_{g,1}^2\lambda_k^2\|y_{k+1} - y^*_{\lambda,k}\|^2 + l_{g,1}^2\lambda_k^2\|z_{k+1} - y_k^*\|^2 + C_\lambda^2/\lambda_k^2\right).\end{aligned}$$

Proof. Using the smoothness of F,

$$F(x_{k+1}) - F(x_k) \le \langle \nabla F(x_k), x_{k+1} - x_k \rangle + \frac{l_{F,1}}{2}\|x_{k+1} - x_k\|^2.$$

Note that $x_{k+1} - x_k = \xi\alpha_k(q_k^x + \tilde{e}_k^x)$, and thus

$$\begin{aligned} F(x_{k+1}) - F(x_k) &\le -\xi\alpha_k\langle \nabla_x F(x_k), q_k^x\rangle - \xi\alpha_k\langle \nabla_x F(x_k), \tilde{e}_k^x\rangle + \frac{l_{F,1}}{2}\|x_{k+1} - x_k\|^2 \\ &\le -\frac{\xi\alpha_k}{2}(\|\nabla F(x_k)\|^2 + \|q_k^x\|^2 - \|\nabla F(x_k) - q_k^x\|^2) - \xi\alpha_k\langle \nabla_x F(x_k), \tilde{e}_k^x\rangle + \xi^2\alpha_k^2 l_{F,1}(\|q_k^x\|^2 + \|\tilde{e}_k^x\|^2). \end{aligned}$$

Using $|\langle a, b\rangle| \le c\|a\|^2 + \frac{1}{4c}\|b\|^2$, we have

$$-\xi\alpha_k\langle \nabla F(x_k), \tilde{e}_k^x\rangle \le \frac{\xi\alpha_k}{4}\|\nabla_x F(x_k)\|^2 + \xi\alpha_k\|\tilde{e}_k^x\|^2.$$

Finally, recall (16). Using $(a+b+c)^2 \le 3(a^2+b^2+c^2)$, we have

$$\|\nabla F(x_k) - q_k^x\|^2 \le 3(4l_{g,1}^2\lambda_k^2\|y_{k+1} - y_{\lambda,k}^*\|^2 + l_{g,1}^2\lambda_k^2\|z_{k+1} - y_k^*\|^2 + C_\lambda^2/\lambda_k^2).$$

Combining all, with $\xi\alpha_k l_{F,1} < 1$, we get the lemma. □

C.6. Decrease in Potential Function

Define the potential function $\mathbb{V}_k$ as the following:

$$\mathbb{V}_k := F(x_k) + l_{g,1}\lambda_k\|y_k - y_{\lambda,k}^*\|^2 + \frac{l_{g,1}\lambda_k}{2}\|z_k - y_k^*\|^2 + \frac{1}{c_\eta l_{g,1}^2\gamma_{k-1}}\left(\frac{\|\tilde{e}_k^x\|^2}{\lambda_k} + \frac{\|\tilde{e}_k^y\|^2}{\lambda_k} + \lambda_k\|\tilde{e}_k^z\|^2\right),$$

with some absolute constant $c_\eta > 0$. We bound the difference in potential function:

$$\begin{aligned} \mathbb{E}[\mathbb{V}_{k+1} - \mathbb{V}_k|\mathcal{F}_k] \le& -\frac{\xi\alpha_k}{4}\|\nabla F(x_k)\|^2 - \frac{\xi\alpha_k}{4}\mathbb{E}[\|q_k^x\|^2|\mathcal{F}_k] + \frac{\xi\alpha_k}{2}\frac{3C_\lambda^2}{\lambda_k^2} \\ &+ l_{g,1}\lambda_k\underbrace{\left(\left(1+\frac{\delta_k}{\lambda_k}\right)\|y_{k+1} - y_{\lambda,k+1}^*\|^2 + 6\xi\alpha_k\lambda_k l_{g,1}\|y_{k+1} - y_{\lambda,k}^*\|^2 - \|y_k - y_{\lambda,k}^*\|^2\right)}_{(i)} \\ &+ \frac{l_{g,1}\lambda_k}{2}\underbrace{\left(\left(1+\frac{\delta_k}{\lambda_k}\right)\|z_{k+1} - y_{k+1}^*\|^2 + 3\xi\alpha_k\lambda_k l_{g,1}\|z_{k+1} - z_k^*\|^2 - \|z_k - y_k^*\|^2\right)}_{(ii)} \\ &+ \frac{1}{c_\eta l_{g,1}^2}\underbrace{\left(\frac{\mathbb{E}[\|\tilde{e}_{k+1}^x\|^2|\mathcal{F}_k]}{\gamma_k\lambda_k} - \frac{\mathbb{E}[\|\tilde{e}_k^x\|^2|\mathcal{F}_k]}{\gamma_{k-1}\lambda_k}\right)}_{(iii)} + 2\xi\alpha_k\mathbb{E}[\|\tilde{e}_k^x\|^2|\mathcal{F}_k] \\ &+ \frac{1}{c_\eta l_{g,1}^2}\underbrace{\left(\frac{\mathbb{E}[\|\tilde{e}_{k+1}^y\|^2|\mathcal{F}_k]}{\gamma_k\lambda_k} - \frac{\mathbb{E}[\|\tilde{e}_k^y\|^2|\mathcal{F}_k]}{\gamma_{k-1}\lambda_k}\right)}_{(iv)} + \frac{\lambda_k}{c_\eta l_{g,1}^2}\underbrace{\left(\frac{\mathbb{E}[\|\tilde{e}_{k+1}^z\|^2|\mathcal{F}_k]}{\gamma_k} - \frac{\mathbb{E}[\|\tilde{e}_k^z\|^2|\mathcal{F}_k]}{\gamma_{k-1}}\right)}_{(v)}. \end{aligned}$$

Using Lemmas C.7, C.8, C.5 and C.6, given that $\delta_k/\lambda_k < \mu_g\beta_k/8$, (i) and (ii) are bounded by

$$(i) \le -\frac{\mu_g\beta_k}{8}\|y_k - y_{\lambda,k}^*\|^2 - \frac{\alpha_k}{\lambda_k l_{g,1}}\|q_k^y\|^2 + O\left(\frac{\xi^2\alpha_k^2 l_{*,0}^2}{\beta_k\mu_g}\mathbb{E}[\|q_k^x\|^2 + \|\tilde{e}_k^x\|^2|\mathcal{F}_k] + \frac{\delta_k^2 l_{f,0}^2}{\lambda_k^4\mu_g^3} + \frac{\alpha_k^2}{\beta_k\mu_g}\mathbb{E}[\|\tilde{e}_k^y\|^2|\mathcal{F}_k]\right),$$

$$(ii) \le -\frac{\mu_g\gamma_k}{8}\|z_k - y_k^*\|^2 - \frac{\gamma_k}{l_{g,1}}\|q_k^z\|^2 + O\left(\frac{\xi^2\alpha_k^2 l_{*,0}^2}{\gamma_k\mu_g}\mathbb{E}[\|q_k^x\|^2 + \|\tilde{e}_k^x\|^2|\mathcal{F}_k] + \frac{\gamma_k}{\mu_g}\mathbb{E}[\|\tilde{e}_k^z\|^2|\mathcal{F}_k]\right),$$

We can use Lemma C.4 to bound (iii), (iv) and (v). Using the step-size condition given in (8b), we have

$$\frac{(1-\eta_{k+1})}{\gamma_k} - \frac{1}{\gamma_{k-1}} = \frac{\frac{\gamma_{k-1}-\gamma_k}{\gamma_{k-1}} - \eta_{k+1}}{\gamma_k} \leq \frac{-\eta_{k+1}}{2\gamma_k}.$$

Note that by the same step-size condition, $\eta_{k+1} \gg O(l_{g,1}^2\gamma_k^2)$, and thus,

$$\begin{aligned}(iii) \leq& -\frac{\eta_{k+1}}{2\gamma_k\lambda_k}\mathbb{E}[\|\tilde{e}_k^x\|^2|\mathcal{F}_k] + O(\sigma_f^2)\cdot\frac{\eta_{k+1}^2}{\lambda_k\gamma_k} + O(\sigma_g^2)\cdot\left(\frac{\eta_{k+1}^2\lambda_k}{\gamma_k} + \frac{\delta_k^2}{\gamma_k\lambda_k}\right)\\ &+ O(l_{g,1}^2)\cdot\left(\xi^2\alpha_k\|q_k^x\|^2 + \alpha_k(\|q_k^y\|^2 + \|\tilde{e}_k^y\|^2) + \gamma_k\lambda_k(\|q_k^z\|^2 + \|\tilde{e}_k^z\|^2)\right).\end{aligned}$$

Similarly, we can use Lemma C.3 and show that

$$\begin{aligned}(iv) \leq& -\frac{\eta_{k+1}}{2\gamma_k\lambda_k}\mathbb{E}[\|\tilde{e}_k^y\|^2|\mathcal{F}_k] + O(\sigma_f^2)\cdot\frac{\eta_{k+1}^2}{\lambda_k\gamma_k} + O(\sigma_g^2)\cdot\left(\frac{\eta_{k+1}^2\lambda_k}{\gamma_k} + \frac{\delta_k^2}{\gamma_k\lambda_k}\right)\\ &+ O(l_{g,1}^2)\alpha_k\cdot\left(\xi^2\|q_k^x\|^2 + \xi^2\|\tilde{e}_k^x\|^2 + \|q_k^y\|^2\right),\\ (v) \leq& -\frac{\eta_{k+1}}{2\gamma_k}\mathbb{E}[\|\tilde{e}_k^z\|^2|\mathcal{F}_k] + O(\sigma_g^2)\cdot\frac{\eta_{k+1}^2}{\gamma_k} + O(l_{g,1}^2)\cdot\left(\frac{\xi^2\alpha_k^2}{\gamma_k}\|q_k^x\|^2 + \frac{\xi^2\alpha_k^2}{\gamma_k}\|\tilde{e}_k^x\|^2 + \gamma_k\|q_k^z\|^2\right).\end{aligned}$$

Plugging inequalities for $(i)-(v)$ back and arranging terms, we get

$$\begin{aligned}\mathbb{E}[\mathbb{V}_{k+1} - \mathbb{V}_k|\mathcal{F}_k] \leq& -\frac{\xi\alpha_k}{4}\|\nabla F(x_k)\|^2 - \frac{\xi\alpha_k}{4}\mathbb{E}[\|q_k^x\|^2|\mathcal{F}_k] + \frac{3C_\lambda^2}{2}\frac{\xi\alpha_k}{\lambda_k^2} - l_{g,1}\lambda_k(i) + \frac{l_{g,1}\lambda_k}{2}(ii)\\ &+ \frac{1}{c_\eta l_{g,1}^2}(iii) + 2\xi\alpha_k\mathbb{E}[\|\tilde{e}_k^x\|^2|\mathcal{F}_k] + \frac{1}{c_\eta l_{g,1}^2}(iv) + \frac{\lambda_k}{c_\eta l_{g,1}^2}(v)\\ \leq& -\frac{\xi\alpha_k}{4}\|\nabla F(x_k)\|^2 - \frac{\lambda_k l_{g,1}\mu_g\beta_k}{4}\|y_k - y_{\lambda,k}^*\|^2 - \frac{\lambda_k l_{g,1}\mu_g\gamma_k}{4}\|z_k - y_k^*\|^2\\ &- \frac{\xi\alpha_k}{4}\mathbb{E}[\|q_k^x\|^2|\mathcal{F}_k]\left(1 - O(\xi l_{g,1}l_{*,0}^2/\mu_g) - O(\xi c_\eta^{-1})\right)\\ &- \alpha_k\mathbb{E}[\|q_k^y\|^2|\mathcal{F}_k]\left(1 - O(c_\eta^{-1})\right) - \frac{\gamma_k\lambda_k}{2}\mathbb{E}[\|q_k^z\|^2|\mathcal{F}_k]\left(1 - O(c_\eta^{-1})\right)\\ &+ O\left(\frac{C_\lambda^2\xi\alpha_k}{\lambda_k^2} + \frac{l_{f,0}^2 l_{g,1}\delta_k^2}{\mu_g^3\lambda_k^3}\right) + \text{noise variance terms},\end{aligned}$$

where noise variance terms are

$$\begin{aligned}\text{noise terms} =& -\frac{\mathbb{E}[\|\tilde{e}_k^x\|^2|\mathcal{F}_k]}{c_\eta l_{g,1}^2}\left(\frac{\eta_{k+1}}{2\gamma_k\lambda_k} - O\left(l_{g,1}^2\xi^2\alpha_k\right) - (c_\eta\xi\alpha_k l_{g,1}^2)\right)\\ &- \frac{\mathbb{E}[\|\tilde{e}_k^y\|^2|\mathcal{F}_k]}{l_{g,1}^2 c_\eta}\left(\frac{\eta_{k+1}}{2\gamma_k\lambda_k} - O(l_{g,1}^2)\alpha_k - O(c_\eta l_{g,1}^3/\mu_g)\alpha_k\right)\\ &- \frac{\lambda_k\mathbb{E}[\|\tilde{e}_k^z\|^2|\mathcal{F}_k]}{l_{g,1}^2 c_\eta}\left(\frac{\eta_{k+1}}{2\gamma_k} - O(l_{g,1}^2)\gamma_k - O(c_\eta l_{g,1}^3/\mu_g)\gamma_k\right)\\ &+ \frac{1}{c_\eta l_{g,1}^2}\left(O(\sigma_f^2)\cdot\frac{\eta_{k+1}^2}{\lambda_k\gamma_k} + O(\sigma_g^2)\cdot\left(\frac{\eta_{k+1}^2\lambda_k}{\gamma_k} + \frac{\delta_k^2}{\gamma_k\lambda_k}\right)\right).\end{aligned}$$

For the all squared terms, with careful design of step-sizes, we can make the coefficient negative. Specifically, we need

$$\xi l_{g,1}l_{*,0}^2/\mu_g \ll 1,\ c_\eta \gg 1,$$

to negate $q_k^{(\cdot)}$ terms, and

$$1 > \eta_{k+1} \gg c_\eta\gamma_k^2(l_{g,1}^3/\mu_g),$$

to suppress noise variance terms, as required in our step-size rules (8). Then, we can simplify the bound for the potential function difference:

$$\mathbb{E}[\mathbb{V}_{k+1} - \mathbb{V}_k|\mathcal{F}_k] \leq -\frac{\xi\alpha_k}{4}\|\nabla F(x_k)\|^2 + O(\xi C_\lambda^2)\cdot\frac{\alpha_k}{\lambda_k^2} + O(l_{f,0}^2 l_{g,1}/\mu_g^3)\cdot\frac{\delta_k^2}{\lambda_k^3} + \frac{1}{c_\eta l_{g,1}^2}\left(O(\sigma_f^2)\cdot\frac{\eta_{k+1}^2}{\lambda_k\gamma_k} + O(\sigma_g^2)\cdot\left(\frac{\eta_{k+1}^2\lambda_k}{\gamma_k} + \frac{\delta_k^2}{\gamma_k\lambda_k}\right)\right).$$

Proof of Theorem 4.3 Summing the above over all $k = 0$ to $K-1$, using $\delta_k/\lambda_k = O(1/k)$ and $1/\lambda_k, \delta_k/\gamma_k = o(1)$, we obtain Theorem 4.3.

C.7. Proof of Corollary 4.4

Using the step-sizes specified in (9), since $\lambda_k = \gamma_k/2\alpha_k \asymp k^{a-c}$, $\delta_k \asymp k^{a-c-1}$. As long as $a - c - 1 < -c$, which is satisfied if $a < 1$, we have $\delta_k/\gamma_k = o(1)$. We can also check that

$$\frac{\delta_k}{\lambda_k} \leq (k + k_0 + 1)^{-1} < \frac{\mu_g\beta_k}{8} = \frac{(k+k_0)^{-c}}{k_0^{1-c}},$$

as long as and $c < 1$. Given the above, we have

$$\sum_{k=0}^{K-1}\frac{\mathbb{E}[\|\nabla F(x_k)\|^2]}{(k+k_0)^a} \leq O_{\mathbb{P}}(1)\cdot\sum_k\frac{1}{(k+k_0)^{3a-2c}} + O_{\mathbb{P}}(\sigma_f^2)\cdot\sum_k\frac{1}{(k+k_0)^{-2c-a}} + O_{\mathbb{P}}(\sigma_g^2)\cdot\sum_k\frac{1}{(k+k_0)^{-4c+a}} + O_{\mathbb{P}}(1).$$

Again, we consider three regimes:

Stochasticity in both upper-level and lower-level objectives: $\sigma_f^2, \sigma_g^2 > 0$. In this case, we set $a = 3/5, c = 2/5$, and thus $\lambda_k \asymp k^{1/5}$, which yields

$$\mathbb{E}[\|\nabla F(x_R)\|^2] \asymp \frac{\log K}{K^{2/5}}.$$

Stochasticity only in the upper-level: $\sigma_f^2 > 0, \sigma_g^2 = 0$. In this case, we can take $a = 2/4, c = 1/4$, and thus $\lambda_k \asymp k^{1/4}$, which yields

$$\mathbb{E}[\|\nabla F(x_R)\|^2] \asymp \frac{\log K}{K^{2/4}}.$$

Deterministic case: $\sigma_f^2 = 0, \sigma_g^2 = 0$. Here, we can take $a = 1/3, c = 0$ and since there is no stochasticity in the algorithm, we have

$$\|\nabla F(x_K)\|^2 \asymp \frac{\log K}{K^{2/3}}.$$

D. Deferred Proofs for Lemmas

D.1. Proofs for Main Lemmas

D.1.1. Proof of Lemma 3.1

Proof. Let $y_\lambda^*(x) := \arg\min_y \mathcal{L}_\lambda(x, y)$. Note that $\nabla_y\mathcal{L}_\lambda(x, y_\lambda^*(x)) = 0$, and thus

$$\nabla\mathcal{L}_\lambda^*(x) = \nabla_x\mathcal{L}_\lambda(x, y_\lambda^*(x)) + \nabla_x y_\lambda^*(x)^\top\nabla_y\mathcal{L}_\lambda(x, y_\lambda^*(x)) = \nabla_x\mathcal{L}_\lambda(x, y_\lambda^*(x)).$$

To compare this to $\nabla F(x)$, we can invoke Lemma A.2 which gives

$$\begin{aligned}&\|\nabla F(x) - \nabla_x \mathcal{L}_\lambda(x, y_\lambda^*(x))\| \\ &\quad \le 2(l_{g,1}/\mu_g)\|y_\lambda^*(x) - y^*(x)\| \left(l_{f,1} + \lambda \cdot \min(2l_{g,1}, l_{g,2}\|y^*(x) - y_\lambda^*(x)\|)\right).\end{aligned}$$

From Lemma 3.2, we use $\|y_\lambda^*(x) - y^*(x)\| \le \frac{2l_{f,0}}{\lambda\mu_g}$, and get

$$\|\nabla F(x) - \nabla_x \mathcal{L}_\lambda(x, y_\lambda^*(x))\| \le \frac{1}{\lambda} \cdot \frac{4l_{f,0}l_{g,1}}{\mu_g^2}\left(l_{f,1} + \frac{2l_{f,0}l_{g,2}}{\mu_g}\right).$$

□

D.1.2. PROOF OF LEMMA 3.2

Proof. Note that $\mathcal{L}_\lambda(x, y)$ is at least $\frac{\lambda\mu_g}{2}$ strongly-convex in y once $\lambda \ge 2l_{f,1}\mu_g$. To see this,

$$\mathcal{L}_\lambda(x, y) = f(x, y) + \lambda(g(x, y) - g^*(x)),$$

which is at least $-l_{f,1} + \lambda\mu_g$-strongly convex in y. If $\lambda > 2l_{f,1}/\mu_g$, this implies at least $\lambda\mu_g/2$ strong-convexity of $\mathcal{L}_\lambda(x, y)$ in y.

By the optimality condition at $y_{\lambda_1}^*(x_1)$ with x_1, λ_1, we have

$$\nabla_y f(x_1, y_{\lambda_1}^*(x_1)) + \lambda_1 \nabla_y g(x_1, y_{\lambda_1}^*(x_1)) = 0,$$

which also implies that $\|g(x_1, y_{\lambda_1}^*(x_1))\| \le l_{f,0}/\lambda_1$. Observe that

$$\begin{aligned}&\nabla_y f(x_2, y_{\lambda_1}^*(x_1)) + \lambda_2 \nabla_y g(x_2, y_{\lambda_1}^*(x_1)) \\ &\quad = (\nabla_y f(x_2, y_{\lambda_1}^*(x_1)) - \nabla_y f(x_1, y_{\lambda_1}^*(x_1))) + \nabla_y f(x_1, y_{\lambda_1}^*(x_1)) \\ &\qquad + \lambda_2(\nabla_y g(x_2, y_{\lambda_1}^*(x_1)) - \nabla_y g(x_1, y_{\lambda_1}^*(x_1))) + \lambda_2 \nabla_y g(x_1, y_{\lambda_1}^*(x_1)) \\ &\quad = (\nabla_y f(x_2, y_{\lambda_1}^*(x_1)) - \nabla_y f(x_1, y_{\lambda_1}^*(x_1))) + \lambda_2(\nabla_y g(x_2, y_{\lambda_1}^*(x_1)) - \nabla_y g(x_1, y_{\lambda_1}^*(x_1))) \\ &\qquad + (\lambda_2 - \lambda_1)\nabla_y g(x_1, y_{\lambda_1}^*(x_1)),\end{aligned}$$

where in the last equality, we applied the optimality condition for $y_{\lambda_1}^*(x_1)$. Then applying the Lipschitzness of $\nabla_y f$ and $\nabla_y g$ in x, we have

$$\|\nabla_y f(x_2, y_{\lambda_1}^*(x_1)) + \lambda_2 \nabla_y g(x_2, y_{\lambda_1}^*(x_1))\| \le l_{f,1}\|x_1 - x_2\| + l_{g,1}\lambda_2\|x_2 - x_1\| + (\lambda_2 - \lambda_1)\frac{l_{f,0}}{\lambda_1}.$$

Since $\mathcal{L}_{\lambda_2}(x_2, y)$ is $\lambda_2\mu_g/2$-strongly convex in y, from the coercivity property of strongly-convex functions, along with the optimality condition with $y_{\lambda_2}^*(x_2)$, we have

$$\frac{\lambda_2\mu_g}{2}\|y_{\lambda_1}^*(x_1) - y_{\lambda_2}^*(x_2)\| \le \|\nabla_y \mathcal{L}_{\lambda_2}(x_2, y_{\lambda_1}^*(x_1))\| \le (l_{f,1} + \lambda_2 l_{g,1})\|x_1 - x_2\| + \frac{\lambda_2 - \lambda_1}{\lambda_1} l_{f,0}.$$

Dividing both sides by $(\lambda_2\mu_g/2)$ concludes the first part of the proof. Note that $y^*(x) = \lim_{\lambda\to\infty} y_\lambda^*(x)$. Thus, for any x and finite $\lambda \ge 2l_{f,1}/\mu_g$,

$$\|y_\lambda^*(x) - y^*(x)\| \le \frac{2l_{f,0}}{\lambda\mu_g}.$$

□

D.2. Proofs for Auxiliary Lemmas

D.2.1. Proof of Lemma A.1

Proof. The proof can be also found in Lemma 2.2 in (Ghadimi & Wang, 2018). We provide the proof for the completeness. Recall that $\nabla F(x)$ is given by

$$\nabla F(x) = \nabla_x f(x, y^*(x)) - \nabla^2_{xy} g(x, y^*(x)) \nabla^2_{yy} g(x, y^*(x))^{-1} \nabla_y f(x, y^*(x)).$$

Using the smoothness of functions and Hessian-continuity of g in assumptions, for any $x_1, x_2 \in X$, we get

$$\begin{aligned}
\|\nabla F(x_1) - \nabla F(x_2)\| &\leq \left(l_{f,1} + \frac{l_{f,0}}{\mu_g} l_{g,2} + \frac{l_{g,1}}{\mu_g} l_{g,1} \right) (\|x_1 - x_2\| + \|y^*(x_1) - y^*(x_2)\|) \\
&\quad + l_{g,1} l_{f,0} \|\nabla^2_{yy} g(x_1, y^*(x_1))^{-1} - \nabla^2_{yy} g(x_2, y^*(x_2))^{-1}\| \\
&\leq \left(l_{f,1} + \frac{l_{f,0}}{\mu_g} l_{g,2} + \frac{l_{g,1}}{\mu_g} \right) l_{*,0} \|x_1 - x_2\| + \frac{l_{g,1} l_{f,0}}{\mu_g^2} l_{g,2} l_{*,0} \|x_1 - x_2\|.
\end{aligned}$$

Thus,

$$\begin{aligned}
l_{F,1} &\leq l_{*,0} \left(l_{f,1} + \frac{l_{f,0} l_{g,2} + l_{g,1}^2}{\mu_g} + \frac{l_{f,0} l_{g,1} l_{g,2}}{\mu_g^2} \right) \\
&\leq l_{*,0} \left(l_{f,1} + \frac{l_{g,1}^2}{\mu_g} + \frac{2 l_{f,0} l_{g,1} l_{g,2}}{\mu_g^2} \right),
\end{aligned}$$

where in the last inequality we used $l_{g,1}/\mu_g \geq 1$. □

D.2.2. Proof of Lemma A.2

We use a short-hand $y^* = y^*(x)$.

$$\begin{aligned}
\nabla_x \mathcal{L}_\lambda(x, y) &= \nabla_x f(x, y) + \lambda(\nabla_x g(x, y) - \nabla_x g(x, y^*)) \\
\nabla_y \mathcal{L}_\lambda(x, y) &= \nabla_y f(x, y) + \lambda \nabla_y g(x, y).
\end{aligned}$$

Check that

$$\begin{aligned}
\nabla F(x) - \nabla_x \mathcal{L}_\lambda(x, y) &= \nabla_x f(x, y^*) - \nabla_x f(x, y) \\
&\quad - \nabla^2_{xy} g(x, y^*) \nabla^2_{yy} g(x, y^*)^{-1} \nabla_y f(x, y^*) - \lambda(\nabla_x g(x, y) - \nabla_x g(x, y^*)).
\end{aligned} \tag{31}$$

We can rearrange terms for $(\nabla_x g(x, y) - \nabla_x g(x, y^*))$ as the following:

$$\begin{aligned}
\nabla_x g(x, y) - \nabla_x g(x, y^*) &= \nabla_x g(x, y) - \nabla_x g(x, y^*) - \nabla_{xy} g(x, y^*)^\top (y - y^*) \\
&\quad + \nabla_{xy} g(x, y^*)^\top (y - y^*).
\end{aligned} \tag{32}$$

Note that from the optimality condition for y^*, $\nabla_y g(x, y^*) = 0$ and from $\nabla_x f(x, y) + \lambda \nabla_y g(x, y) = \nabla_y \mathcal{L}(x, y)$, we can express $y - y^*$ as

$$\begin{aligned}
y - y^* &= -\nabla_{yy} g(x, y^*)^{-1} (\nabla_y g(x, y) - \nabla_y g(x, y^*) - \nabla_{yy} g(x, y^*)(y - y^*)) \\
&\quad + \frac{1}{\lambda} \nabla_{yy} g(x, y^*)^{-1} \left(\nabla_y \mathcal{L}(x, y) - \nabla_y f(x, y) \right).
\end{aligned} \tag{33}$$

Plugging (32) and (33) back to (31), we have

$$\begin{aligned}
\nabla F(x) - \nabla_x \mathcal{L}_\lambda(x, y) &= (\nabla_x f(x, y^*) - \nabla_x f(x, y)) - \nabla^2_{xy} g(x, y^*) \nabla^2_{yy} g(x, y^*)^{-1} (\nabla_y f(x, y^*) - \nabla_y f(x, y)) \\
&\quad - \nabla^2_{xy} g(x, y^*) \nabla^2_{yy} g(x, y^*)^{-1} \nabla_y \mathcal{L}(x, y) \\
&\quad - \lambda(\nabla_x g(x, y) - \nabla_x g(x, y^*) - \nabla^2_{xy} g(x, y^*)^\top (y - y^*))
\end{aligned}$$

$$+ \lambda \nabla^2_{xy} g(x, y^*) \nabla^2_{yy} g(x, y^*)^{-1} (\nabla_y g(x, y) - \nabla_y g(x, y^*) - \nabla^2_{yy} g(x, y^*)(y - y^*)).$$

By the smootheness of ∇g from Assumption 1, we have

$$\|\nabla_y g(x, y) - \nabla_y g(x, y^*) - \nabla^2_{yy} g(x, y^*)(y - y^*)\| \le l_{g,2} \|y - y^*\|^2.$$

When $\|y - y^*\|$ is too large, the smoothness of g can be more useful:

$$\|\nabla_y g(x, y) - \nabla_y g(x, y^*) - \nabla^2_{yy} g(x, y^*)(y - y^*)\| \le 2 l_{g,1} \|y - y^*\|.$$

Similarly, we have

$$\|\nabla_x g(x, y) - \nabla_x g(x, y^*) - \nabla^2_{xy} g(x, y^*)^\top (y - y^*)\| \le \min\left(l_{g,2} \|y - y^*\|^2, 2 l_{g,1} \|y - y^*\|\right).$$

On the other hand, by smootheness of f, we also have

$$\|\nabla_x f(x, y^*) - \nabla_x f(x, y)\| \le l_{f,1} \|y - y^*\|, \ \|\nabla_y f(x, y^*) - \nabla_y f(x, y)\| \le l_{f,1} \|y - y^*\|.$$

We can conclude that

$$\begin{aligned} &\|\nabla F(x) - \nabla_x \mathcal{L}_\lambda(x, y) + \nabla^2_{xy} g(x, y^*) \nabla^2_{yy} g(x, y^*)^{-1} \nabla_y \mathcal{L}(x, y)\| \\ &\quad \le l_{f,1}(1 + l_{g,1}/\mu_g) \|y - y^*\| + \lambda (1 + l_{g,1}/\mu_g) \|y - y^*\| \min(l_{g,2} \|y - y^*\|, 2 l_{g,1}). \end{aligned}$$

We know that $l_{g,1}/\mu_g \ge 1$ and thus, we have

$$\begin{aligned} &\|\nabla F(x) - \nabla_x \mathcal{L}_\lambda(x, y) + \nabla^2_{xy} g(x, y^*) \nabla^2_{yy} g(x, y^*)^{-1} \nabla_y \mathcal{L}(x, y)\| \\ &\quad \le 2 (l_{g,1}/\mu_g) \|y - y^*\| \left(l_{f,1} + \lambda \cdot \min(2 l_{g,1}, l_{g,2} \|y - y^*\|)\right), \end{aligned}$$

yielding the lemma.

D.2.3. Proof of Lemma A.3

Proof. Lipschitzness of $y^*_\lambda(x)$ is immediate from Lemma 3.2. By the optimality condition for $\nabla y^*_\lambda(x)$, we have

$$\nabla_y \mathcal{L}_\lambda(x, y^*_\lambda(x)) = \nabla_y f(x, y^*_\lambda(x)) + \lambda \nabla_y g(x, y^*_\lambda(x)) = 0.$$

Taking derivative with respect to x, we get

$$(\nabla^2_{yy} f(x, y^*_\lambda(x)) + \lambda \nabla^2_{yy} g(x, y^*_\lambda(x))) \nabla y^*_\lambda(x) = -(\nabla^2_{xy} f(x, y^*_\lambda(x)) + \lambda \nabla^2_{xy} g(x, y^*_\lambda(x))).$$

As $\lambda > 2 l_{f,1}/\mu_g$, the left-hand side is positive definite with mininum eigenvalue larger than $\lambda \mu_g / 2$, and we have

$$\nabla y^*_\lambda(x) = -\left(\frac{1}{\lambda} \nabla^2_{yy} f(x, y^*_\lambda(x)) + \nabla^2_{yy} g(x, y^*_\lambda(x))\right)^{-1} \left(\frac{1}{\lambda} \nabla^2_{xy} f(x, y^*_\lambda(x)) + \nabla^2_{xy} g(x, y^*_\lambda(x))\right).$$

To get the smoothness result, we compare this at x_1 and x_2, yielding

$$\begin{aligned} \frac{\lambda \mu_g}{2} \|\nabla y^*_\lambda(x_1) - \nabla y^*_\lambda(x_2)\| &\le (l_{f,2} + \lambda l_{g,2})(\|x_1 - x_2\| + \|y^*_\lambda(x_1) - y^*_\lambda(x_2)\|) \max_{x \in X} \|\nabla y^*_\lambda(x)\| \\ &\quad + (l_{f,2} + \lambda l_{g,2})(\|x_1 - x_2\| + \|y^*_\lambda(x_1) - y^*_\lambda(x_2)\|) \\ &\le (l_{f,2} + \lambda l_{g,2})(1 + l_{\lambda,0})^2 \|x_1 - x_2\|. \end{aligned}$$

Arranging this, we get

$$\|\nabla y^*_\lambda(x_1) - \nabla y^*_\lambda(x_2)\| \le 32 \left(\frac{l_{f,2}}{\lambda} + l_{g,2}\right) \frac{l^2_{g,1}}{\mu^3_g} \|x_1 - x_2\|.$$

□

D.2.4. Proof of Lemma A.4

Proof. This is immediate from Lipschitz continuity in Lemma 3.2 with sending $\lambda_1 = \lambda_2$ to infinity.

$$\begin{aligned}\mathbb{E}[\|y^*(x_{k+1}) - y^*(x_k)\|^2|\mathcal{F}_k] &\leq l_{*,0}^2\mathbb{E}[\|x_{k+1} - x_k\|^2|\mathcal{F}_k] \\ &\leq l_{*,0}^2\xi^2(\alpha_k^2\mathbb{E}[\|q_k^x\|^2|\mathcal{F}_k] + \alpha_k^2\sigma_f^2 + \beta_k^2\sigma_g^2).\end{aligned}$$

□

D.2.5. Proof of Lemma A.5

Proof. We can use the smoothness property of $y^*(x)$ as in (Chen et al., 2021), which is crucial to control the noise variance induced from updating x. We can start with the following:

$$\begin{aligned}\langle v_k, y_{k+1}^* - y_k^*\rangle &= \langle v_k, \nabla y^*(x_k)(x_{k+1} - x_k)\rangle \\ &\quad + \langle v_k, y^*(x_{k+1}) - y^*(x_k) - \nabla y^*(x_k)(x_{k+1} - x_k)\rangle.\end{aligned}$$

On the first term, taking expectation and using $\langle a, b\rangle \leq c\|a\|^2 + \frac{1}{4c}\|b\|^2$,

$$\begin{aligned}\mathbb{E}[\langle v_k, \nabla y^*(x_k)(x_{k+1} - x_k)\rangle|\mathcal{F}_k] &= -\xi\alpha_k\mathbb{E}[\langle v_k, \nabla y^*(x_k)q_k^x\rangle|\mathcal{F}_k] \\ &\leq \xi\alpha_k\eta_k\mathbb{E}[\|v_k\|^2|\mathcal{F}_k] + \frac{\xi\alpha_k}{4\eta_k}\mathbb{E}[\|\nabla y^*(x_k)q_k^x\|^2|\mathcal{F}_k] \\ &\leq \xi\alpha_k\eta_k\mathbb{E}[\|v_k\|^2|\mathcal{F}_k] + \frac{\xi\alpha_k l_{*,0}^2}{4\eta_k}\mathbb{E}[\|q_k^x\|^2|\mathcal{F}_k],\end{aligned}$$

where we used the Lipschitz continuity of $y^*(x)$. For the second term, using smoothness of $y^*(x)$,

$$\begin{aligned}&\mathbb{E}[\langle v_k, y^*(x_{k+1}) - y^*(x_k) - \nabla y^*(x_k)(x_{k+1} - x_k)\rangle|\mathcal{F}_k] \\ &\leq \frac{l_{*,1}}{2}\mathbb{E}[\|v_k\|\|x_{k+1} - x_k\|^2|\mathcal{F}_k] \\ &\leq \frac{l_{*,1}}{4}\mathbb{E}\left[\left(l_{*,1}\|v_k\|^2 + \frac{1}{l_{*,1}}\right)\cdot\|x_{k+1} - x_k\|^2|\mathcal{F}_k\right] \\ &\leq \frac{l_{*,1}^2}{4}\mathbb{E}[\|v_k\|^2\cdot\mathbb{E}\left[\|x_k - x_{k+1}\|^2|\mathcal{F}_k'\right]|\mathcal{F}_k] \\ &\quad + \frac{\xi^2}{4}\left(\alpha_k^2\mathbb{E}[\|q_k^x\|^2] + \alpha_k^2\sigma_f^2 + \beta_k^2\sigma_g^2\right),\end{aligned}$$

where $\mathcal{F}_k'$ is a sigma-algebra generated by stochastic noises up to k^{th} iteration and v_k. Note that

$$\mathbb{E}\left[\|x_k - x_{k+1}\|^2|\mathcal{F}_k'\right] \leq \xi^2\alpha_k^2\mathbb{E}\left[\|q_k^x\|^2|\mathcal{F}_k\right] + \xi^2(\alpha_k^2\sigma_f^2 + \beta_k^2\sigma_g^2),$$

and from boundedness of $\nabla_x f$ and $\nabla_x g$ in Assumption 4, we have $\alpha_k\|q_k^x\| \leq \alpha_k l_{f,0} + 2\beta_k l_{g,0}$. With $M_f = l_{f,0}^2 + \sigma_f^2$, $M_g = l_{g,0}^2 + \sigma_g^2$, and $M = \max(M_f, M_g)$, we get

$$\mathbb{E}\left[\|x_k - x_{k+1}\|^2|\mathcal{F}_k'\right] \leq 2\xi^2(M_f\alpha_k^2 + 2M_g\beta_k^2) \leq 4M\xi^2 l_{*,1}^2\beta_k^2,$$

which yields

$$\begin{aligned}&\mathbb{E}[\langle v_k, y^*(x_{k+1}) - y^*(x_k) - \nabla y^*(x_k)(x_{k+1} - x_k)\rangle|\mathcal{F}_k] \\ &\leq M\xi^2 l_{*,1}^2\beta_k^2\mathbb{E}[\|v_k\|^2|\mathcal{F}_k] + \frac{\xi^2}{4}\left(\alpha_k^2\mathbb{E}[\|q_k^x\|^2|\mathcal{F}_k] + \alpha_k^2\sigma_f^2 + \beta_k^2\sigma_g^2\right).\end{aligned}$$

Combining all, we obtain the desired result. □

D.2.6. PROOF OF LEMMA A.6

Proof. We can start with the following decomposition:

$$\begin{aligned}\langle v_k, y^*_{\lambda_{k+1}}(x_{k+1}) - y^*_{\lambda_k}(x_k)\rangle &= \langle v_k, y^*_{\lambda_{k+1}}(x_{k+1}) - y^*_{\lambda_k}(x_{k+1})\rangle \\ &\quad + \langle v_k, \nabla y^*_{\lambda_k}(x_k)(x_{k+1} - x_k)\rangle \\ &\quad + \langle v_k, y^*_{\lambda_k}(x_k) - y^*_{\lambda_k}(x_k) - \nabla y^*_{\lambda_k}(x_k)(x_{k+1} - x_k)\rangle.\end{aligned}$$

For the second and third terms, we can apply the smoothness of $y_\lambda(x)$ similarly in the proof in D.2.5.

On the first term, taking expectation and using $\langle a, b\rangle \le c\|a\|^2 + \frac{1}{4c}\|b\|^2$,

$$\begin{aligned}\mathbb{E}[\langle v_k, y^*_{\lambda_{k+1}}(x_{k+1}) - y^*_{\lambda_k}(x_{k+1})\rangle|\mathcal{F}_k] &\le c\mathbb{E}[\|v_k\|^2] + \frac{1}{4c}\mathbb{E}[\|y^*_{\lambda_{k+1}}(x_{k+1}) - y^*_{\lambda_k}(x_{k+1})\|^2] \\ &\le c\mathbb{E}[\|v_k\|^2] + \frac{1}{c}\frac{\delta_k^2}{\lambda_k^2\lambda_{k+1}^2}\frac{l_{f,0}^2}{\mu_g^2},\end{aligned}$$

where we applied Lemma 3.2. Take $c = \frac{\delta_k}{\lambda_k}$, getting

$$\mathbb{E}[\langle v_k, y^*_{\lambda_{k+1}}(x_{k+1}) - y^*_{\lambda_k}(x_{k+1})\rangle|\mathcal{F}_k] \le \frac{\delta_k}{\lambda_k}\mathbb{E}[\|v_k\|^2] + \frac{l_{f,0}^2\delta_k}{\mu_g^2\lambda_k^3}.$$

Adding this with bounds on other two terms, we get the lemma. □

D.3. Proofs for Auxiliary Lemmas with Momentum

D.3.1. PROOF OF LEMMA C.1

Due to Lipschitz continuity of $y^*(x)$, we have

$$\begin{aligned}\mathbb{E}[\|y^*(x_{k+1}) - y^*(x_k)\|^2] &\le l_{*,0}^2\mathbb{E}[\|x_{k+1} - x_k\|^2] \\ &\le \xi^2\alpha_k^2 l_{*,0}^2\mathbb{E}[\|q_k^x + \tilde{e}_k^x\|^2] \le 2\xi^2\alpha_k^2 l_{*,0}^2(\mathbb{E}[\|q_k^x\|^2] + \mathbb{E}[\|\tilde{e}_k^x\|^2]).\end{aligned}$$

D.3.2. PROOF OF LEMMA C.2

Using Lemma 3.2, we have

$$\begin{aligned}\mathbb{E}[\|y^*_{\lambda_{k+1}}(x_{k+1}) - y^*_{\lambda_k}(x_k)\|^2] &\le \frac{8\delta_k^2}{\lambda_k^2\lambda_{k+1}^2} + 2l_{\lambda,0}^2\mathbb{E}[\|x_{k+1} - x_k\|^2] \\ &\le 4\xi^2\alpha_k^2 l_{*,0}^2(\mathbb{E}[\|q_k^x\|^2] + \mathbb{E}[\|\tilde{e}_k^x\|^2]) + \frac{8\delta_k^2}{\lambda_k^4}.\end{aligned}$$

D.3.3. PROOF OF LEMMA C.3

We can start with unfolding the expression for $\mathbb{E}[\|\tilde{e}_{k+1}^z\|^2]$.

$$\begin{aligned}\mathbb{E}[\|\tilde{e}_{k+1}^z\|^2] &= \mathbb{E}[\|\tilde{h}_z^{k+1} - q_{k+1}^z\|^2] \\ &= \mathbb{E}[\|\nabla_y g(x_{k+1}, z_{k+1}; \phi_z^{k+1}) + (1-\eta_{k+1})(\tilde{h}_z^k - \nabla_y g(x_k, z_k; \phi_z^{k+1})) - q_{k+1}^z\|^2] \\ &= \mathbb{E}[\|(1-\eta_{k+1})\tilde{e}_k^z + \nabla_y g(x_{k+1}, z_{k+1}; \phi_z^{k+1}) \\ &\qquad + (1-\eta_{k+1})(q_k^z - \nabla_y g(x_k, z_k; \phi_z^{k+1})) - q_{k+1}^z\|^2] \\ &= (1-\eta_{k+1})^2\mathbb{E}[\|\tilde{e}_k^z\|^2] + \mathbb{E}[\|\eta_k(\nabla_y g(x_{k+1}, z_{k+1}; \phi_z^{k+1}) - q_{k+1}^z) \\ &\qquad + (1-\eta_{k+1})(\nabla_y g(x_{k+1}, z_{k+1}; \phi_z^{k+1}) - \nabla_y g(x_k, z_k; \phi_z^{k+1}) + q_k^z - q_{k+1}^z)\|^2].\end{aligned}$$

In the last equality, we used

$$\mathbb{E}[\mathbb{E}[\langle \tilde{e}_k^z, \nabla_y g(x_{k+1}, z_{k+1}; \phi_z^{k+1}) - q_{k+1}^z\rangle|\mathcal{F}_{k+1}]] = 0,$$

$$\mathbb{E}[\mathbb{E}[\langle \tilde{e}_k^z, \nabla_y g(x_k, z_k; \phi_z^{k+1}) - q_k^z \rangle | \mathcal{F}_{k+1}]] = 0.$$

Also note that

$$\mathbb{E}[\|\nabla_y g(x_{k+1}, z_{k+1}; \phi_z^{k+1}) - q_{k+1}^z\|^2] \leq \sigma_g^2,$$

from the variance boundedness (Assumption 3). We also observe that

$$\begin{aligned}\mathbb{E}[\|\nabla_y g(x_{k+1}, z_{k+1}; \phi_z^{k+1}) - \nabla_y g(x_k, z_k; \phi_z^{k+1})\|^2] &\leq l_{g,1}^2(\|x_{k+1} - x_k\|^2 + \|z_{k+1} - z_k\|^2) \\ &= l_{g,1}^2(\xi^2\alpha_k^2\|q_k^x + \tilde{e}_k^x\|^2 + \gamma_k^2\|q_k^z + \tilde{e}_k^z\|^2),\end{aligned}$$

due to Assumption 6. The same inequality holds for $q_{k+1}^z - q_k^z$:

$$\mathbb{E}[\|q_{k+1}^z - q_k^z\|^2] \leq l_{g,1}^2(\xi^2\alpha_k^2\|q_k^x + \tilde{e}_k^x\|^2 + \gamma_k^2\|q_k^z + \tilde{e}_k^z\|^2).$$

Now we plug these inequalities and using $\|a+b\|^2 \leq 2(\|a\|^2 + \|b\|^2)$ multiple times, we have

$$\begin{aligned}\mathbb{E}[\|\tilde{e}_{k+1}^z\|^2] \leq (1-\eta_{k+1})^2(1 + 8l_{g,1}^2\gamma_k^2)\mathbb{E}[\|\tilde{e}_k^z\|^2] + 2\eta_{k+1}^2\sigma_g^2 \\ + 8l_{g,1}^2(1-\eta_{k+1})^2\left(\xi^2\alpha_k^2\mathbb{E}[\|q_k^x\|^2] + \xi^2\alpha_k^2\mathbb{E}[\|\tilde{e}_k^x\|^2] + \gamma_k^2\mathbb{E}[\|q_k^z\|^2]\right).\end{aligned}$$

Similarly, we can repeat similar steps for $\tilde{e}_{k+1}^y$. To simplify the notation, with slight abuse in notation, we let $q_k^y(\zeta, \phi) := \nabla_y f(x_k, y_k; \zeta) + \lambda_k \nabla_y g(x_k, y_k; \phi)$. Note that $q_k^y = \mathbb{E}[q_k^y(\zeta, \phi)]$. Then we can get a similar bound for $\mathbb{E}[\|\tilde{e}_{k+1}^y\|^2]$:

$$\begin{aligned}\mathbb{E}[\|\tilde{e}_{k+1}^y\|^2] \leq (1-\eta_{k+1})^2\mathbb{E}[\|\tilde{e}_k^y\|^2] + 2\eta_{k+1}^2\mathbb{E}[\|q_{k+1}^y(\zeta_y^{k+1}, \phi_y^{k+1}) - q_{k+1}^y\|^2] \\ + 2(1-\eta_{k+1})^2\mathbb{E}[\|(q_{k+1}^y(\zeta_y^{k+1}, \phi_y^{k+1}) - q_k^y(\zeta_y^{k+1}, \phi_y^{k+1})) + (q_k^y - q_{k+1}^y)\|^2].\end{aligned}$$

Using the variance bound similarly, we have

$$\mathbb{E}[\|q_{k+1}^y(\zeta_y^{k+1}, \phi_y^{k+1}) - q_{k+1}^y\|^2] \leq \sigma_f^2 + \lambda_{k+1}^2\sigma_g^2.$$

Then, we unfold the last term such that

$$\begin{aligned}&\mathbb{E}[\|(q_{k+1}^y(\zeta_y^{k+1}, \phi_y^{k+1}) - q_k^y(\zeta_y^{k+1}, \phi_y^{k+1})) + (q_k^y - q_{k+1}^y)\|^2] \\ &= \mathbb{E}[\|(\nabla_y f(x_{k+1}, y_{k+1}; \zeta_y^{k+1}) - \nabla_y f(x_k, y_k; \zeta_y^{k+1}) + \nabla_y f(x_k, y_k) - \nabla_y f(x_{k+1}, y_{k+1})) \\ &\quad + \lambda_k(\nabla_y g(x_{k+1}, y_{k+1}; \phi_y^{k+1}) - \nabla_y g(x_k, y_k; \phi_y^{k+1}) + \nabla_y g(x_k, y_k) - \nabla_y g(x_{k+1}, y_{k+1})) \\ &\quad + \delta_k(\nabla_y g(x_{k+1}, y_{k+1}; \phi_y^{k+1}) - \nabla_y g(x_{k+1}, y_{k+1}) + \nabla_y g(x_k, y_k) - \nabla_y g(x_k, y_k; \phi_y^{k+1}))\|^2] \\ &\leq 12(l_{f,1}^2 + l_{g,1}^2\lambda_k^2)(\|x_{k+1} - x_k\|^2 + \|y_{k+1} - y_k\|^2) + 12\delta_k^2\sigma_g^2 \\ &\leq 24(l_{f,1}^2\alpha_k^2 + l_{g,1}^2\beta_k^2)(\xi^2\|q_k^x\|^2 + \xi^2\|\tilde{e}_k^x\|^2 + \|q_k^y\|^2 + \|\tilde{e}_k^y\|^2) + 12\delta_k^2\sigma_g^2.\end{aligned}$$

We note that we set $\lambda_k \geq 2l_{f,1}/\mu_g$, and thus $l_{f,1} \leq \lambda_k l_{g,1}$. In total, we get

$$\begin{aligned}\mathbb{E}[\|\tilde{e}_{k+1}^y\|^2] \leq (1-\eta_{k+1})^2(1 + 96l_{g,1}^2\beta_k^2)\mathbb{E}[\|\tilde{e}_k^y\|^2] + 2\eta_{k+1}^2(\sigma_f^2 + \lambda_{k+1}^2\sigma_g^2) + 24\delta_k^2\sigma_g^2 \\ + 96(1-\eta_{k+1})^2 l_{g,1}^2\beta_k^2(\xi^2\|q_k^x\|^2 + \xi^2\|\tilde{e}_k^x\|^2 + \|q_k^y\|^2).\end{aligned}$$

D.3.4. Proof of Lemma C.4

Similarly to the case for $\|\tilde{e}_{k+1}^y\|^2$, let us define $q_k^x(\zeta, \phi) := \nabla_x f(x_k, y_{k+1}; \zeta) + \lambda_k(\nabla_x g(x_k, y_{k+1}; \phi) - \nabla_x g(x_k, z_{k+1}; \phi))$. We note that ζ_x^k, ϕ_x^k are sampled after y_{k+1}, z_{k+1} is updated but before x_k is updated. Hence,

$$\begin{aligned}\mathbb{E}[\mathbb{E}[\langle \tilde{e}_k^x, q_{k+1}^x(\zeta_x^{k+1}, \phi_x^{k+1}) - q_{k+1}^x \rangle | \mathcal{F}'_{k+1}]] = 0, \\ \mathbb{E}[\mathbb{E}[\langle \tilde{e}_k^x, q_k^x(\zeta_x^{k+1}, \phi_x^{k+1}) - q_k^x \rangle | \mathcal{F}'_{k+1}]] = 0.\end{aligned}$$

Thus, following similar procedure, we have

$$\mathbb{E}[\|\tilde{e}_{k+1}^x\|^2] = \mathbb{E}[\|q_{k+1}^x(\zeta_x^{k+1}, \phi_x^{k+1}) + (1-\eta_{k+1})(q_k^x + \tilde{e}_k^x - q_k^x(\zeta_x^{k+1}, \phi_x^{k+1})) - q_{k+1}^x\|^2]$$

$$\leq (1-\eta_{k+1})^2\mathbb{E}[\|\tilde{e}_k^x\|^2] + 2\eta_k^2\mathbb{E}[\|q_{k+1}^x(\zeta_x^{k+1},\phi_x^{k+1}) - q_{k+1}^x\|^2]$$
$$+ 2(1-\eta_{k+1})^2\mathbb{E}[\|(q_{k+1}^x(\zeta_x^{k+1},\phi_x^{k+1}) - q_k^x(\zeta_x^{k+1},\phi_x^{k+1})) + (q_k^x - q_{k+1}^x)\|^2].$$

Note that

$$\begin{aligned}
&\mathbb{E}[\|q_{k+1}^x(\zeta_x^{k+1},\phi_x^{k+1}) - q_{k+1}^x\|^2] \\
&\quad = \mathbb{E}[\|(\nabla_x f(x_{k+1},y_{k+2};\zeta_x^{k+1}) - \nabla_x f(x_{k+1},y_{k+2})) \\
&\qquad + \lambda_k(\nabla_x g(x_{k+1},y_{k+2};\phi_x^{k+1}) - \nabla_x g(x_{k+1},y_{k+2})) + \lambda_k(\nabla_x g(x_{k+1},z_{k+2};\phi_x^{k+1}) - \nabla_x g(x_{k+1},z_{k+2}))\|^2] \\
&\quad \leq 3(\sigma_f^2 + \lambda_k^2\sigma_g^2).
\end{aligned}$$

Finally, we have

$$\begin{aligned}
&\mathbb{E}[\|(q_{k+1}^x(\zeta_x^{k+1},\phi_x^{k+1}) - q_k^x(\zeta_x^{k+1},\phi_x^{k+1})) + (q_k^x - q_{k+1}^x)\|^2] \\
&\quad = \mathbb{E}[\|(\nabla_x f(x_{k+1},y_{k+2};\zeta_x^{k+1}) - \nabla_x f(x_k,y_{k+1};\zeta_x^{k+1}) + \nabla_y f(x_k,y_{k+1}) - \nabla_y f(x_{k+1},y_{k+2})) \\
&\qquad + \lambda_k(\nabla_x g(x_{k+1},y_{k+2};\phi_x^{k+1}) - \nabla_x g(x_k,y_{k+1};\phi_x^{k+1}) + \nabla_x g(x_k,y_{k+1}) - \nabla_x g(x_{k+1},y_{k+2})) \\
&\qquad + \lambda_k(\nabla_x g(x_{k+1},z_{k+2};\phi_x^{k+1}) - \nabla_x g(x_k,z_{k+1};\phi_x^{k+1}) + \nabla_x g(x_k,z_{k+1}) - \nabla_x g(x_{k+1},z_{k+2})) \\
&\qquad + \delta_k(\nabla_y g(x_{k+1},y_{k+2};\phi_x^{k+1}) - \nabla_y g(x_{k+1},y_{k+2}) + \nabla_x g(x_k,y_{k+1}) - \nabla_x g(x_k,y_{k+1};\phi_x^{k+1})) \\
&\qquad + \delta_k(\nabla_x g(x_{k+1},z_{k+2};\phi_x^{k+1}) - \nabla_x g(x_{k+1},z_{k+2}) + \nabla_x g(x_k,z_{k+1}) - \nabla_x g(x_k,z_{k+1};\phi_x^{k+1}))\|^2].
\end{aligned}$$

Using Cauchy-Schwartz inequality, we get

$$\begin{aligned}
&\mathbb{E}[\|(q_{k+1}^x(\zeta_x^{k+1},\phi_x^{k+1}) - q_k^x(\zeta_x^{k+1},\phi_x^{k+1})) + (q_k^x - q_{k+1}^x)\|^2] \\
&\quad \leq 30(l_{f,1}^2 + l_{g,1}^2\lambda_k^2)(\|x_{k+1}-x_k\|^2 + \|y_{k+2}-y_{k+1}\|^2 + \|z_{k+2}-z_{k+1}\|^2) + 40\delta_k^2\sigma_g^2 \\
&\quad \leq 120 l_{g,1}^2\lambda_k^2(\xi^2\alpha_k^2(\|q_k^x\|^2 + \|\tilde{e}_k^x\|^2) + \alpha_{k+1}^2(\|q_k^y\|^2 + \|\tilde{e}_k^y\|^2) + \gamma_{k+1}^2(\|q_k^z\|^2 + \|\tilde{e}_k^z\|^2)) + 40\delta_k^2\sigma_g^2.
\end{aligned}$$

Combining all, we obtain the result.