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 (FSA) method, and study its non-asymptotic convergence properties. Specifically, we show that FSA 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 gener-

ally formulated as the following minimization problem:

$$\min_{x \in X} \quad F(x) := f(x, y^*(x))$$

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

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

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

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.
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; Dağrêou et al., 2022). We are particularly interested in regimes in which such operations are costly and prohibitive (Mehra & Hamm, 2021; Giovannielli 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 \( \epsilon \)-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:

\[
\min_{x \in X, y \in \mathbb{R}^p} \ f(x, y) \quad \text{s.t.} \quad g(x, y) - g^*(x) \leq 0, \quad (P')
\]

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 \( \lambda \) 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 \( \lambda \). 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 \( \epsilon \)-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 \( \lambda_k \) forces the bias quickly. On the other hand, fast growth of \( \lambda_k \) 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 \( \lambda_k \) 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, \( F^2 SA \), for stochastic bilevel optimization. \( F^2 SA \) 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 \( F^2 SA \) in different stochastic regimes. It converges to an \( \epsilon \)-stationary-point of \((P)\) after \( \tilde{O}(\epsilon^{-3.5}) \), \( \tilde{O}(\epsilon^{-2.5}) \), or \( \tilde{O}(\epsilon^{-1.5}) \) iterations if both \( \nabla f \) and \( \nabla g \) contain stochastic noise, if only access to \( \nabla f \) 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 \( \lambda_k \) 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 \( F^2 SA \) 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 \( O(k^{-2/3}) \) 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 \( \lambda_k \) 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 \( \mu_g \)-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 \( \zeta \) and \( \phi \), respectively, where \( E[\nabla f(x, y; \zeta)] = \nabla f(x, y) \) and \( E[\nabla g(x, y; \phi)] = \nabla g(x, y) \). The variances of stochastic gradient estimators are bounded:

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

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** (\( \epsilon \)-stationary point). A point \( x \) is called \( \epsilon \)-stationary if \( \|\nabla F(x)\|^2 \leq \epsilon \), where \( \nabla F \) 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 $\nabla L$:

$$\nabla x L_\lambda(x, y) = \nabla_x f(x, y) + \lambda(\nabla_x g(x, y) - \nabla g^*(x)),$$

$$\nabla y L_\lambda(x, y) = \nabla_y f(x, y) + \lambda \nabla_y g(x, y).$$

Note that

$$\nabla g^*(x) = \nabla_x g(x, y^*(x)) + \nabla_y g^*(x) \nabla_y g(x, y^*(x))$$

$$= \nabla_x g(x, y^*(x)),$$

due to the optimality condition for $g$ at $y^*(x)$. Thus, we could consider optimizing $L_\lambda(x, y)$ by introducing an auxiliary variable $z$ that satisfies $y^*(x)$, and setting up an alternative bilevel formulation (P) with outer-level objective $L_\lambda(x', z)$, 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 $L_\lambda$, which affects the overall step-size design and noise variance.

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

$$L_\lambda^*(x) := \min_y L_\lambda(x, y).$$

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

$$y^*_\lambda(x) := \arg \min_y L_\lambda(x, y).$$

(2)

Since $F(x) = \lim_{\lambda \to \infty} L_\lambda^*(x)$ for every $x \in X$, we could expect that $L_\lambda^*(x)$ 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 $\lambda \geq 2f_{1,1}/\mu_g$, $\nabla L_\lambda^*(x)$ is given by

$$\nabla x L_\lambda(x, y^*_\lambda(x)) = \nabla_x f(x, y^*_\lambda(x)) + \lambda(\nabla_x g(x, y^*_\lambda(x)) - \nabla x g(x, y^*(x))).$$

**Algorithm 1 $\mathbb{F}^2\mathcal{SA}$**

**Input:** step sizes: $\{\alpha_k, \gamma_k\}$, multiplier difference sequence: $\{\delta_k\}$, inner-loop iteration count: $T$, step-size ratio: $\xi$, initializations: $\lambda_0, x_0, y_0, z_0$

1: for $k = 0...K - 1$
2: $z_{k,0} \leftarrow z_k, y_{k,0} \leftarrow y_k$
3: for $t = 0...T - 1$
4: $z_{k,t+1} \leftarrow z_{k,t} - \gamma k h^k_{gyx}$
5: $y_{k,t+1} \leftarrow y_{k,t} - \alpha_k (h^k_{fy} + \lambda_k (h^k_{gyy} - h^k_{gzz}))$
6: end for
7: $z_{k+1} \leftarrow z_{k,T}, y_{k+1} \leftarrow y_{k,T}$
8: $x_{k+1} \leftarrow x_k - \xi \alpha_k (h^k_{fx} + \lambda_k (h^k_{gxy} - h^k_{gzz}))$
9: $\lambda_{k+1} \leftarrow \lambda_k + \delta_k$
10: end for

Furthermore, we have

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

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

Importantly, $\nabla L_\lambda^*(x)$ can be computed only with first-order derivatives of both $f$ and $g$. Thus any first-order method that finds a stationary point of $L_\lambda^*(x)$ 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 $\nabla L_\lambda^*(x)$ 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 $y^*_\lambda(x_k)$ and $y^*(x_k)$, respectively. We gradually increase $\lambda_k$ with $k$, so that the bias in the sequence $\{x_k\}$ converges to 0.

Our Fully First-order Stochastic Approximation ($\mathbb{F}^2\mathcal{SA}$) method is shown in Algorithm 1. We emphasize that the method works with stochastic gradients that are independent unbiased estimators of gradients, i.e.,

$$h^k_{gy} := \nabla_x g(x_k, y_k, t; \phi_{k,t}^x), \quad h^k_{gy} := \nabla_y g(x_k, y_k, t; \phi_{k,t}^y),$$

$$h^k_{gxy} := \nabla_x g(x_k, y_k, t; \phi_{k,t}^x), \quad h^k_{gxx} := \nabla_x g(x_k, z_{k+1}; \phi_{z,z}^x).$$

The algorithm can set $T = 1$ in conjunction with an appropriate choice of $\xi$, 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 $\epsilon$-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 $L_{\lambda_k}$ 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 that...
We now describe the specific rules. Since the step size for
we have stochastic noises in the gradient estimate of $g$. This rate is optimized when
for some $\lambda_k$. We characterize first how fast
move sufficiently fast toward the current target $y_{\lambda,k}^*$ every ite-
ration, dominating the movement of target $y_{\lambda,k}^*$ that results
from updates to $x_k$ and $\lambda_k$ (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:

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

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

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

where $O_{\nu}(1)$ are instance-dependent constants. If $\lambda_k$ 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 $\xi$ sufficiently small. Alternatively, we can set $\xi = 1$ choose $T > 1$ to depend on some instance-specific parameters.

### 3.2. Extension: Integrating Momentum

Given the simple structure of Algorithm 1, we can inte-
grate 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 $T = 1$. 

**Algorithm 2 E3SA**

**Input:** step sizes: $\{\alpha_k, \gamma_k\}$, multiplier difference sequence: $\{\delta_k\}$, momentum-weight sequence $\{\eta_k\}$, step-size ratio: $\xi$, initialization: $\lambda_0, x_0, y_0, z_0$

1: for $k = 0 \ldots K - 1$ do
2: \hspace{1em} $z_{k+1} \leftarrow z_k - \gamma_k h_k$
3: \hspace{1em} $y_{k+1} \leftarrow y_k - \alpha_k (h_k + \lambda_k h_{g})$
4: \hspace{1em} $x_{k+1} \leftarrow x_k - \xi \alpha_k (h_k + \lambda_k (h_{g} - h_{g}))$
5: \hspace{1em} $\lambda_{k+1} \leftarrow \lambda_k + \delta_k$
6: end for

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

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

Other quantities \( \tilde{h}_{fy}, \tilde{h}_{gy}, \tilde{h}_{fz}, \tilde{h}_{gz}, \tilde{h}_{gx} \) 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 Faster Fully First-Order Stochastic Approximation (F³SA); it is described in Algorithm 2, where we simply replace \( h \) with \( \tilde{h} \). Note that we have additional momentum-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 \( y^*_k(x) \).

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

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

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

The smoothness of \( y^*_k(x) \) 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 \( l_{x,0} = \max(1, l_{x,0,0}) \) and \( l_{x,1} = l_{x,0,1} \) where \( \lambda_0 \) 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 \( \lambda_0 \geq 2l_{f1}/\mu_g \) and

\[
\beta_k \leq \gamma_k \leq \min\left(\frac{1}{4l_{f,1}}, \frac{1}{4T\mu_g}\right), \quad \alpha_k \leq \frac{1}{2\xi T^{1/2}},
\]

\[
\frac{\xi}{T} < c_\varepsilon \mu_g \max\left(l_{g,1}l_{x,0}, l_{x,1}\sqrt{M}\right)^{-1} \delta_k \leq \frac{T\mu_g\beta_k \lambda_0^4}{16}
\]

for all \( k \geq 0 \) with a proper absolute constant \( c_\varepsilon > 0 \). Then for any \( K \geq 1 \), the iterates generated by Algorithm 1 satisfy

\[
\sum_{k=0}^{K-1} \xi_k \alpha_k \mathbb{E}[\|\nabla F(x_k)\|^2] \leq O_{\varepsilon}(1) \cdot \sum_k \xi_k \alpha_k \lambda_k^2
\]

\[
+ O_{\varepsilon}(\sigma_f^2) \cdot \sum_k \alpha_k^2 \lambda_k + O_{\varepsilon}(\sigma_g^2) \cdot \sum_k \gamma_k^2 \lambda_k + O_{\varepsilon}(1),
\]

where \( O_{\varepsilon}(1) \) 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 \( V_k \), defined by

\[
V_k := (F(x_k) - F^*) + l_{g,1} \lambda_k \|y_k - y^*_k(x_k)\|^2
\]

\[
+ \frac{\lambda_k l_{g,1}}{2} \|z_k - y^*(x_k)\|^2,
\]

for \( F^* \) is the minimum value of \( F \) and \( y^*_k \) 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 \( y^*_k(x_k) \) since \( y^*_k(x_k) \) 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 \( O_{\varepsilon}(\lambda_k) \).

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:

\[
T = \max\left(32, (c_\varepsilon \mu_g)^{-1} \max\left(l_{g,1}l_{x,0}, \sqrt{M} l_{x,1}\right)\right),
\]

\[
\xi = 1, \quad \alpha_k = \frac{c_a}{(k + k_0)^a}, \quad \gamma_k = \frac{c_a}{(k + k_0)^c},
\]

and for the multiplier increase sequence \( \{\delta_k\} \),

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

with some rate constants \( a, c \in [0, 1] \) and \( a \geq c \). We design the starting value \( \lambda_0 \) of the Lagrange multiplier and the constants as

\[
k_0 \geq \frac{4}{\mu_g} \max\left(\frac{\xi l_{f,1}}{2}, T l_{g,1}, l_{f,1}\right), \quad \lambda_0 \geq \frac{2l_{f,1}}{\mu_g},
\]

\[
f_{\mathcal{F}_1} \leq \frac{1}{2 - \xi T^{-1/2}} f_{\mathcal{F}}.
\]
When we use the momentum-assisting technique, we require

\[ x \text{ w.r.t. } 5 \text{ Assumption 6.} \]

Suppose Assumptions 1-3 and 6 hold. If Theorem 4.3.

\[ (\text{Assumption 5}) \text{ for the analysis, as we no longer make use of} \]

step-size parameters are chosen such that \( \lambda \), \( \mu \), \( \gamma \), \( \sigma \).

We show the following convergence results hold after K iterations of Algorithm 2.

\[ (a) \text{ If stochastic noises are present in both upper-level} \]

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 \( E[\|\nabla F(x_R)\|^2] \approx \frac{\log K}{K^{2/7}} \).

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

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

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 \( O(k^{-2/3}) \). In deterministic settings (no noise), we get a rate of \( O(k^{-2/3}) \). This rate compares to the \( O(k^{-1/2}) \) 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 \( g^*_s \). 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

\[ \beta_k \leq \gamma_k \leq \frac{1}{16l_{g,1}}, \xi \alpha_k \leq \frac{1}{l_{f,1}}, \]

\[ \xi \leq c_\xi \frac{\mu_g}{l_{g,1}l_{g,0}^2}, \delta_k \leq \frac{\mu_g \delta_k}{8}, \]

\[ \gamma_k \leq \frac{1}{2\lambda_0 \mu_g}, \]

\[ \eta_0 = \eta_1 = 1, \delta_k/\gamma_k = o(1), \]

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

\[ \sum_{k=0}^{K-1} \xi \alpha_k E[\|\nabla F(x_k)\|^2] \leq \frac{\log K}{K^{2/7}} \]

\[ + O_\varepsilon(\sigma_f^2) \cdot \sum_k \frac{\eta_{k+1}^2}{\gamma_k} + O_\varepsilon(\sigma_g^2) \cdot \sum_k \frac{\eta_k^2}{\gamma_k^2} + O_\varepsilon(1), \]

where \( O_\varepsilon(1) \) are instance-dependent constants.

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

\[ \alpha_k = c_\alpha = \frac{c_\alpha}{(k + k_0)^2}, \gamma_k = \frac{c_\gamma}{(k + k_0)^2}, \eta_k = (k + 1)^{-2c}, \]

\[ \xi \leq c_\xi \frac{\mu_g}{l_{g,1}l_{g,0}^2}, \delta_k = \frac{\gamma_k}{\alpha_k} - \lambda_k, \lambda_0 \geq \frac{2l_{f,1}}{\mu_g}, \]

\[ k_0 \geq \frac{128}{\mu_g} \max \left( \frac{c_\xi l_{f,1}l_{g,1}}{\mu_g}, \frac{c_\xi l_{g,1}}{\mu_g} \right), \]

\[ c_\gamma = \frac{8}{\mu_g k_0^2}, c_\alpha = \frac{8}{\mu_g \lambda_0^{k_0^2}}, \]

\[ \xi \leq \frac{c_\xi \mu_g}{l_{g,1}l_{g,0}^2}, \delta_k \leq \frac{\mu_g \delta_k}{8}, \]

\[ \gamma_k \leq \frac{1}{2\lambda_0 \mu_g}, \]

\[ \eta_0 = \eta_1 = 1, \delta_k/\gamma_k = o(1), \]

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.

\[ (a) \text{ If stochastic noises are present in both upper-level} \]

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 \( E[\|\nabla F(x_R)\|^2] \approx \frac{\log K}{K^{2/7}} \).

\[ (b) \text{ If stochastic noises are present only in } f \text{ (i.e., } \sigma_f^2 > 0, \sigma_g^2 = 0), \text{ then by setting } a = 1/3 \text{ and } c = 1/4 \text{ in (9)}, \text{ we obtain } E[\|\nabla F(x_R)\|^2] \approx \frac{\log K}{K^{1/2}}. \]

\[ (c) \text{ If we have access to exact information about } f \text{ and } g (i.e., } \sigma_f^2 = \sigma_g^2 = 0), \text{ then by setting } a = 1/3 \text{ and } c = 0 \text{ in (9)}, \text{ we obtain } \|\nabla F(x_K)\|^2 \approx \frac{\log K}{K^{2/7}}. \]

The improvements in rates are different in different stochasticity regimes. For instance, the sample complexity required
When deterministic gradient oracles are available, the autors in (Ye et al., 2022) employed the so called dynamic-barrier method (Gong et al., 2021) to decide the value of $\lambda_k$ 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 $\lambda_k$ 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 $D_{\text{train}} := \{(x_i, \tilde{y}_i)\}_{i=1}^n$ with the label $\tilde{y}_i$ being randomly corrupted with probability $p < 1$. We are also given a small but clean validation set $D_{\text{val}} := \{(x_i, y_i)\}_{i=1}^m$. 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 optimization problem, as follows:

$$\min_{\lambda} \sum_{i=1}^n l(x_i, y_i; w^*)$$

s.t. $w^* \in \arg\min_{w} \sum_{i=1}^n \sigma(\lambda_i)l(\tilde{x}_i, \tilde{y}_i; w) + c||w||^2$.

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 $F^2SA$ approach of Algorithm 1.

We demonstrate the performance of Algorithm 1 ($F^2SA$) 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.\footnote{We report our best results obtained with different hyperparameters for each algorithm.}

Although iteration complexity is worse for first-order methods than SOBO, we observe that $F^2SA$ 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 $F^2SA$ 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


Fully First-Order Stochastic Bilevel Optimization


A. Auxiliary Lemmas

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

A.1. Additional Notation

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Meaning</th>
<th>Less than</th>
</tr>
</thead>
<tbody>
<tr>
<td>$l_{f,0}$</td>
<td>Bound of $|\nabla_x f|, |\nabla_y f|$</td>
<td>$\cdot$</td>
</tr>
<tr>
<td>$l_{f,1}$</td>
<td>Smoothness of $f$</td>
<td>$\cdot$</td>
</tr>
<tr>
<td>$l_{g,0}$</td>
<td>Bound of $|\nabla_x g|$</td>
<td>$\cdot$</td>
</tr>
<tr>
<td>$l_{g,1}$</td>
<td>Smoothness of $g$</td>
<td>$\cdot$</td>
</tr>
<tr>
<td>$\mu_g$</td>
<td>Strong-convexity of $g$</td>
<td>$\cdot$</td>
</tr>
<tr>
<td>$l_{g,2}$</td>
<td>Hessian-continuity of $g$</td>
<td>$\cdot$</td>
</tr>
<tr>
<td>$M_f$</td>
<td>Second-order moment of $\nabla f(x; y; \zeta)$</td>
<td>$l_{f,0}^2 + \sigma_f^2$</td>
</tr>
<tr>
<td>$M_g$</td>
<td>Second-order moment of $\nabla g(x; y; \phi)$</td>
<td>$l_{g,0}^2 + \sigma_g^2$</td>
</tr>
<tr>
<td>$l_{F,1}$</td>
<td>Smoothness of $F(x)$</td>
<td>$l_{*,0} \left( l_{f,1} + \frac{l_{g,1}^2}{\mu_g} + \frac{2l_{f,0}l_{g,1}l_{g,2}}{\mu_g^2} \right)$</td>
</tr>
<tr>
<td>$l_{\lambda,1}$</td>
<td>Smoothness of $y^\lambda_1(x)$ (for $\lambda \geq 2l_{f,1}/\mu_g$ with Assumption 5)</td>
<td>$32(l_{g,2} + \lambda^{-1}\cdot l_{f,2}) \frac{l_{g,1}^2}{\mu_g^2}$</td>
</tr>
<tr>
<td>$l_{*,0}$</td>
<td>$= 1 + \max_{\lambda \geq 2l_{f,1}/\mu_g} l_{\lambda,0}$</td>
<td>$\cdot$</td>
</tr>
<tr>
<td>$l_{*,1}$</td>
<td>$= \max_{\lambda \geq 2l_{f,1}/\mu_g} l_{\lambda,1}$</td>
<td>$\cdot$</td>
</tr>
</tbody>
</table>

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

$$
q_k^x := \nabla_x f(x_k, y^t_k) + \lambda_k (\nabla_x g(x_k, y^t_k) - \nabla_x g(x_k, z_k)),
q_k^y := \nabla_y f(x_k, y^t_k) + \lambda_k \nabla_y g(x_k, y^t_k),
q_k^z := \nabla_y g(x_k, z^t_k).
$$

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:

$$
\|\nabla F(x) - \nabla_x L_\lambda(x, y) + \nabla_{x}^2 g(x, y^*(x))^T \nabla_{y}^2 g(x, y^*(x))^{-1} \nabla_{y} L_\lambda(x, y)\|
\leq 2(l_{g,1}/\mu_g) \|y - y^*(x)\| (l_{f,1} + \lambda \cdot \min(2l_{g,1}, l_{g,2}) \|y - y^*(x)\|).
$$

**Lemma A.3.** Under Assumptions 1, 2 and 5, and $\lambda > 2l_{f,1}/\mu_g$, a function $y^\lambda_1(x)$ is $l_{\lambda,1}$-smooth: for any $x_1, x_2 \in X$, we have

$$
\|\nabla y^\lambda_1(x_1) - \nabla y^\lambda_1(x_2)\| \leq l_{\lambda,1} \|x_1 - x_2\|
$$

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

11
Lemma A.4. For any fixed \( \lambda > 2l_{f,1}/\mu_g \), at every \( k \) iteration conditioned on \( \mathcal{F}_k \), we have
\[
\mathbb{E}[(\|y^*(x_{k+1}) - y^*(x_k)\|^2 | \mathcal{F}_k)] \leq \xi^2 l_{f,0}^2 (\alpha_k^2 \mathbb{E}[\|q_k^g\|^2 | \mathcal{F}_k] + \alpha_k^2 \sigma_f^2 + \beta_k^2 \sigma_g^2).
\]

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
\[
\mathbb{E}[(v_k, y^*(x_{k+1}) - y^*(x_k)) | \mathcal{F}_k] \leq (\xi \alpha_k \eta_k + M \xi^2 l_{f,1}^2 \beta_k^2) \mathbb{E}[\|v_k\|^2 | \mathcal{F}_k] \\
+ \left( \frac{\xi \alpha_k l_{f,0}^2}{4 \eta_k} + \frac{\xi^2 \alpha_k^2}{4} \right) \mathbb{E}[\|q_k^g\|^2 | \mathcal{F}_k] + \frac{\xi^2}{4} (\alpha_k^2 \sigma_f^2 + \beta_k^2 \sigma_g^2),
\]
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
\[
\mathbb{E}[(v_k, y^*_w(x_{k+1}) - y^*_w(x_k)) | \mathcal{F}_k] \leq (\delta_k/\lambda_k + \xi \alpha_k \eta_k + M \xi^2 l_{f,1}^2 \beta_k^2) \mathbb{E}[\|v_k\|^2 | \mathcal{F}_k] \\
+ \left( \frac{\xi \alpha_k l_{f,0}^2}{4 \eta_k} + \frac{\xi^2 \alpha_k^2}{4} \right) \mathbb{E}[\|q_k^g\|^2 | \mathcal{F}_k] + \frac{\xi^2}{4} (\alpha_k^2 \sigma_f^2 + \beta_k^2 \sigma_g^2) + \frac{\delta_k l_{f,0}^2}{\lambda_k \mu_g^2},
\]
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 \( \nabla_{k+1} - \nabla_k \) for the potential function \( \nabla_k \) 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 \( y^*_{\lambda,k} := y^*_w(x_k), y^*_w := y^*(x_k) \), and \( x^* = \arg \min_{x} F(x) \). Recall that \( y^*_w \) and \( y^* \) are given in (2) and (P), respectively. Using the above notation, the potential function given in (4) can be rewritten as
\[
\nabla_k := (F(x_k) - F(x^*)) + \lambda_k l_{g,1} \mathcal{I}_k + \frac{\lambda_k l_{g,1}}{2} \mathcal{J}_k \tag{12}
\]
for each \( k \in \mathbb{N} \). In the following three subsections, we find the upper bound of \( \nabla_{k+1} - \nabla_k \) 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 \( \alpha_k \) is designed to satisfy
\[
(\text{step-size rule}): \quad \alpha_k \leq \frac{1}{2 \xi l_{f,1}}, \tag{13}
\]
which is essential to obtain the negative term \(-\frac{\xi \alpha_k}{4} \|q_k^g\|^2\) 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
\[
(\text{step-size rule}): \quad \frac{\xi}{T} \leq \frac{\mu_g}{96 l_{g,1}}, \tag{14}
\]
The terms, \( \|y_{k+1} - y^*_w\|^2 \) 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 \( k \in \mathbb{N} \)

\[
\mathbb{E}[F(x_{k+1}) - F(x_k)|\mathcal{F}_k] \leq -\frac{\xi \alpha_k}{4} \left( 2\|\nabla F(x_k)\|^2 + \|q_k^2\|^2 \right) + \frac{T \mu_2 \alpha_k \lambda_k^2}{32} \left( 2\|y_{k+1} - y_{\lambda,k}^*\|^2 + \|z_{k+1} - y_k^*\|^2 \right) \\
+ \frac{\xi^2 l_{f,1}}{2} (\alpha_k^2 \sigma_f^2 + \beta_k^2 \sigma_g^2) + \frac{\xi \alpha_k}{2} \cdot 3C \lambda_k \lambda^{-2},
\]

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

Proof. From the smoothness of \( F \),

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

As \( q_k^2 \) satisfies \( \mathbb{E}[x_{k+1} - x_k | \mathcal{F}_k] = \alpha_k q_k^2 \),

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

Note that

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

and thus with (13) we have

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

Next, we bound \( \|\nabla F(x_k) - q_k^2\| \) using the triangle inequality:

\[
\|q_k^2 - \nabla F(x_k)\| = \|q_k^2 - \nabla L_{\lambda,k}^*(x_k) + \nabla L_{\lambda,k}^*(x_k) - \nabla F(x_k)\| \\
\leq \|\nabla f(x_k, y_{k+1}) - \nabla f(x_k, y_{\lambda,k}^*)\| + \lambda_k \|\nabla g(x_k, y_{k+1}) - \nabla g(x_k, y_{\lambda,k}^*)\| \\
+ \lambda_k \|\nabla_x g(x_k, z_{k+1}) - \nabla_x g(x_k, y_{\lambda,k}^*)\| + \|\nabla L_{\lambda,k}^*(x_k) - \nabla F(x_k)\|.
\]

From Lemma 3.1, the term \( \|\nabla L_{\lambda,k}^*(x_k) - \nabla F(x_k)\| \) is bounded by \( C_\lambda/\lambda_k \). Combining with the regularity of \( f \) and \( g \) yields the following:

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

(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

\[
\mathbb{E}[F(x_{k+1}) - F(x_k)|\mathcal{F}_k] \leq -\frac{\xi \alpha_k}{2} \|\nabla F(x_k)\|^2 - \frac{\xi \alpha_k}{4} \|q_k^2\|^2 \\
+ \frac{\xi \alpha_k}{2} \cdot 3C \lambda_k \lambda^{-2} + 3\xi \alpha_k l_{g,1} \lambda_k \|z_{k+1} - y_k^*\|^2 + 6\xi \alpha_k l_{g,1} \lambda_k^2 \|y_{k+1} - y_{\lambda,k}^*\|^2 \]

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 $I_{k+1}$ 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:

\[
(\text{step-size rule}): \quad \frac{\delta_k}{\lambda_k} \leq \frac{T \beta_k \mu_g}{32}, \quad \text{and} \quad 2\xi^2 M I_{*,1}^2 \beta_k^2 < T \beta_k \mu_g / 16. \tag{18}
\]

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

\[
\frac{\xi^2}{T^2} \leq \frac{\mu_g^2}{8} (M I_{*,1}^2)^{-1},
\]

which also holds by (3b) with sufficiently small $c_\xi$.

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

\[
\mathbb{E}[I_{k+1} | F_k] \leq (1 + T \beta_k \mu_g / 4) \mathbb{E}[y_{k+1} - y_{\lambda,k}^*]^2 | F_k] + O \left( \frac{\xi^2 I_{*,0}^2 \alpha_k^2}{\mu_g T \beta_k} \right) \mathbb{E}[\|q_k^i\|^2 | F_k] + O \left( \frac{\delta_k I_{*,0}^2}{\lambda_k^3 \mu_g^2} \right) + O(\xi^2 I_{*,0}^2) \cdot (\alpha_k^2 \sigma_f^2 + \beta_k^2 \sigma_g^2). \tag{19}
\]

where $I_k$ and $q_k^i$ are given in (11) and (10), respectively.

**Proof.** We can start from

\[
\|y_{k+1} - y_{\lambda,k+1}\|^2 = \|y_{k+1} - y_{\lambda,k}^*\|^2 + \|y_{\lambda,k+1} - y_{\lambda,k}^*\|^2 - 2 \langle y_{k+1} - y_{\lambda,k}^*, y_{\lambda,k+1} - y_{\lambda,k}^* \rangle.
\]

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

\[
(ii) : \mathbb{E}[\|y_{\lambda,k+1}^* - y_{\lambda,k}^*\|^2 | F_k] \leq \frac{4\delta_k^2}{\lambda_k^2 \lambda_{k+1}^2 \mu_g^2} + l_{*,0}^2 \mathbb{E}[\|x_{k+1} - x_k\|^2 | F_k]
\]

\[
\leq \frac{4\delta_k^2}{\lambda_k^2 \mu_g^2} + \xi^2 l_{*,0}^2 (\alpha_k^2 \mathbb{E}[\|q_k^i\|^2] + \alpha_k^2 \sigma_f^2 + \beta_k^2 \sigma_g^2).
\]

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

\[
(iii) \leq (2\delta_k / \lambda_k + T \beta_k \mu_g / 8 + 2 M \xi^2 I_{*,1}^2 \beta_k^2) \mathbb{E}[y_{k+1} - y_{\lambda,k}^*]^2 | F_k] + \xi^2 \left( \frac{\alpha_k^2}{2} + \frac{8\delta_k I_{*,0}^2}{\mu_g T \beta_k} \right) \|q_k^i\|^2 + \xi^2 \left( \frac{\alpha_k^2 \sigma_f^2 + \beta_k^2 \sigma_g^2}{2} \right) + \frac{2\delta_k I_{*,0}^2}{\lambda_k^3 \mu_g^2}.
\]

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

\[
\mathbb{E}[I_{k+1} | F_k] \leq (1 + 2\delta_k / \lambda_k + T \beta_k \mu_g / 8 + 2 M \xi^2 I_{*,1}^2 \beta_k^2) \mathbb{E}[y_{k+1} - y_{\lambda,k}^*]^2 ]
\]

\[
+ O \left( \frac{\xi^2 I_{*,0}^2 \alpha_k^2}{\mu_g T \beta_k} \right) \|q_k^i\|^2 + O \left( \frac{\delta_k I_{*,0}^2}{\lambda_k^3 \mu_g^2} \right) + O(\xi^2 I_{*,0}^2) \cdot (\alpha_k^2 \sigma_f^2 + \beta_k^2 \sigma_g^2). \tag{20}
\]

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

Next, we note that $\alpha_k$ and $\beta_k$ are chosen to satisfy

\[
(\text{step size rules}): \quad \alpha_k \leq \frac{1}{8l_{f,1}} \quad \text{and} \quad \beta_k \leq \frac{1}{8l_{g,1}}, \tag{21}
\]

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

---

Fully First-Order Stochastic Bilevel Optimization
Lemma B.3. Under the step-size rule given in (21), it holds that for each $k \in \mathbb{N}$
\[ E[\|y_{k+1} - y^*_k\|^2 | \mathcal{F}_k] \leq (1 - 3T\mu_y\beta_k/4) E[\|y_k - y^*_k\|^2] + T(\alpha^2_k \sigma_f^2 + \beta^2_k \sigma_g^2). \]  

Proof. Since $E[y_{k+1}^{(t+1)} - y_{k}^{(t)} | \mathcal{F}_k] = -\alpha_k \nabla y_{k}^{(t)}$, we have
\[ E[\|y_{k+1} - y^*_k\|^2 | \mathcal{F}_k] = \|y_k - y^*_k\|^2 - 2\alpha_k \langle \nabla y_{k}^{(t)}, y_{k}^{(t)} - y^*_k \rangle + E[\|y_{k+1} - y_{k}^{(t)}\|^2 | \mathcal{F}_k]. \]

As we start from $(k, y_k) \geq 2\mu_y \beta_k$, all $\mathcal{L}_k$ is $(\lambda_k \mu_y / 2)$ strongly convex in $y$, and we have
\[ \max \left( \frac{\lambda_k \mu_y}{2} \|y_{k}^{(t)} - y^*_k\|^2, \frac{1}{I_{f,1} + \lambda_k I_{g,1}} \|\nabla y_{k}^{(t)}\|^2 \right) \leq \langle \nabla y_{k}^{(t)}, y_{k}^{(t)} - y^*_k \rangle. \]

Using $E[\|y_{k+1} - y_{k}^{(t)}\|^2 | \mathcal{F}_k] \leq \alpha^2_k \|\nabla y_{k}^{(t)}\|^2 + \alpha^2_k \sigma_f^2 + \beta^2_k \sigma_g^2$, have
\[ (i) : E[\|y_{k+1} - y^*_k\|^2 | \mathcal{F}_k] \leq (1 - 3\mu_y \beta_k / 4) \|y_{k}^{(t)} - y^*_k\|^2 + (\alpha^2_k \sigma_f^2 + \beta^2_k \sigma_g^2), \]

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

\[ \square \]

B.3. Descent Lemma for $z_k$ towards $y^*_k$

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

\[ \text{(step-size rule):} \quad 2M^2 \xi^2 \lambda^2 \beta^2_k \leq T \mu_y \gamma_k / 16, \quad (23) \]

This condition holds since $\beta_k \leq \gamma_k$, and $\beta_k \leq \frac{1}{4T\mu_y}$ and $\xi^2 \geq 2 \mu_y / 8 (M^2 \xi^2 \beta^2_k)^{-1}$.

Lemma B.4. Under the step-size rule (23), at each $k^{th}$ iteration, the following holds:

\[ E[\mathcal{F}_{k+1} | \mathcal{F}_k] \leq \left( 1 + \frac{3T \gamma_k \mu_y}{8} \right) \cdot E[\|z_{k+1} - y^*_k\|^2 | \mathcal{F}_k] + O \left( \frac{\xi^2 \alpha^2_k l_0^2}{T \mu_y \gamma_k} \right) \|z_k^*\|^2 + O \left( \frac{\xi^2 \alpha^2_k l_0^2}{T \mu_y \gamma_k} \alpha^2_k \sigma_f^2 + \beta^2_k \sigma_g^2 \right). \]  

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

\[ \|z_{k+1} - y^*_k\|^2 = \|\tilde{z}_{k+1} - y^*_k\|^2 + \|\tilde{y}_{k+1} - y^*_k\|^2 - 2 \langle \tilde{z}_{k+1} - y^*_k, \tilde{y}_{k+1} - y^*_k \rangle. \]

Lemma 3.2 implies that

\[ (i) : E[\|\tilde{z}_{k+1} - y^*_k\|^2 | \mathcal{F}_k] \leq \lambda^2 \xi^2 \left( \alpha^2_k \|\nabla z_{k} q_k\|^2 + \alpha^2_k \sigma_f^2 + \beta^2_k \sigma_g^2 \right). \]

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

\[ (ii) : \langle z_{k+1} - y^*_k, \tilde{y}_{k+1} - y^*_k \rangle \leq (T \gamma_k \mu_y / 8 + M \xi^2 l_0^2 \beta^2_k) E[\|z_{k+1} - y^*_k\|^2 | \mathcal{F}_k] + \left( \frac{\xi^2 \alpha^2_k}{4} + \frac{2M \xi^2 \alpha^2_k l_0^2}{T \mu_y \gamma_k} \right) \|z_k^*\|^2 + \frac{\xi^2}{4} \left( \alpha^2_k \sigma_f^2 + \beta^2_k \sigma_g^2 \right). \]

The above bounds and Lemma B.5 imply that

\[ E[\mathcal{F}_{k+1} | \mathcal{F}_k] \leq \left( 1 + \frac{T \gamma_k \mu_y}{4} + 2M \xi^2 l_0^2 \beta^2_k \right) \cdot E[\|z_{k+1} - y^*_k\|^2 | \mathcal{F}_k] + \xi^2 \alpha^2_k \cdot \left( \frac{l_0^2}{T \mu_y \gamma_k} + \frac{1}{2} \right) \|z_k^*\|^2 + \xi^2 \cdot \left( \frac{1}{2} + l_0^2 \right) \left( \alpha^2_k \sigma_f^2 + \beta^2_k \sigma_g^2 \right). \]  

Using (23), we conclude. 

\[ \square \]
Next, $\gamma_k$ is chosen to satisfy the following step-size rules:

\begin{equation}
\text{(step-size rule):} \quad l_{g, 1}\gamma_k \leq 1/4, \quad T\mu_g\gamma_k \leq 1/4, \tag{26}
\end{equation}

which directly comes from (3a).

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

\begin{equation}
\mathbb{E}[\|z_{k+1} - y_k^*\|^2 | J_k] \leq (1 - 3T\mu_g\gamma_k/4)J_k + T\gamma_k^2\sigma_g^2. \tag{27}
\end{equation}

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

$$
\|z_k^{(t+1)} - y_k^*\|^2 = \|z_k^{(t)} - y_k^*\|^2 + \|z_k^{(t+1)} - z_k^{(t)}\|^2 + 2\langle z_k^{(t+1)} - z_k^{(t)}, z_k^{(t)} - y_k^* \rangle
$$

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

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

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

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

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

$$
\mathbb{E}[\|z_k^{(t+1)} - y_k^*\|^2 | J_k] \leq \|z_k^{(t)} - y_k^*\|^2 + \gamma_k^2\sigma_g^2 \cdot \gamma_k(1 - l_{g, 1}\gamma_k)\langle \nabla g_k(z_k^{(t)}), z_k^{(t)} - y_k^*\rangle
$$

$$
\leq (1 - 3T\mu_g\gamma_k/4)\|z_k^{(t)} - y_k^*\|^2 + \gamma_k^2\sigma_g^2.
$$

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

**B.4. Proof of Theorem 4.1**

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

$$
\mathbb{V}_{k+1} - \mathbb{V}_k = F(x_{k+1}) - F(x_k) + \lambda_{k+1}l_{g, 1}\mathbb{I}_{k+1} - l_{g, 1}\mathbb{I}_k
$$

$$
+ \frac{\lambda_{k+1}l_{g, 1}}{2} J_{k+1} - \frac{\lambda_k l_{g, 1}}{2} J_k.
$$

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

$$
\mathbb{E}[\mathbb{V}_{k+1} - \mathbb{V}_k | J_k] \leq -\frac{\xi\alpha_k}{2}\|\nabla F(x_k)\|^2 - \frac{\xi\alpha_k}{4}\mathbb{E}[\|y_k^*\|^2 | J_k] + \frac{\xi\alpha_k}{2} \cdot 3C^2\lambda_k^{-2} + \frac{\xi\alpha_k^2}{2} (\alpha_k^2\sigma_y^2 + \beta_k^2\sigma_g^2)
$$

$$
+ l_{g, 1}\mathbb{E}[\lambda_{k+1}\mathbb{I}_{k+1}] + \left( \frac{\lambda_k T\beta_k\mu_g}{16} \|y_{k+1} - y_{k, k}\|^2 - \lambda_k\mathbb{I}_k | J_k \right)
$$

$$
+ \frac{l_{g, 1}}{2} \mathbb{E}[\lambda_{k+1} J_{k+1}] + \left( \frac{\lambda_k T\gamma_k\mu_g}{32} \|z_{k+1} - y_k^*\|^2 - \lambda_k J_k | J_k \right).
$$

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

$$
(i) \leq \lambda_k \left( 1 + \frac{5T\beta_k\mu_g}{16} + \frac{\delta_k}{\lambda_k} \right) \mathbb{E}[\|y_{k+1} - y_{k, k}\|^2 | J_k] - \lambda_k\mathbb{I}_k
$$
\[
\begin{align*}
&+ O(\xi^2 t_{w,0}^2 \frac{\lambda_k \sigma_f^2}{\mu_g T \beta_k^2}) q_k^2 + O(\xi^2 t_{w,0}^2) \lambda_k \sigma_f^2 + \beta_k^2 \sigma_g^2 + O \left( \frac{l_{f,0}^2}{\mu_g^2} \right) \delta_k \frac{\lambda_k}{\lambda_k^2}. \\
&= (i) \leq \lambda_k \left( 1 + \frac{T \beta_k \mu_g}{2} \right) \mathbb{E}[\|y_{k+1} - y^*_k\|^2 | F_k] - \lambda_k J_k + (iii).
\end{align*}
\]

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

\[
(i) \leq \lambda_k \left( 1 + \frac{T \beta_k \mu_g}{2} \right) \mathbb{E}[\|y_{k+1} - y^*_k\|^2 | F_k] - \lambda_k J_k + (iii).
\]

The estimation of \( \|y_{k+1} - y^*_k\|^2 \) from Lemma B.3 yields that

\[
(i) \leq - \lambda_k T \mu_g \beta_k I_k + O(\xi^2 t_{w,0}^2) \frac{\alpha_k \sigma_f^2}{\mu_g T \beta_k^2} q_k^2 + (iii),
\]

\[
= - \lambda_k T \mu_g \beta_k I_k + O(\xi^2 t_{w,0}^2) \frac{\alpha_k \sigma_f^2}{\mu_g T \beta_k^2} q_k^2 + O(T + \xi^2 t_{w,0}^2) \lambda_k (\alpha_k^2 \sigma_f^2 + \beta_k^2 \sigma_g^2) + O \left( \frac{l_{f,0}^2}{\mu_g^3} \right) \delta_k \frac{\lambda_k}{\lambda_k^2}.
\]

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

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

\[
(ii) \leq \lambda_k \left( 1 + \frac{T \gamma_k \mu_g}{8} + \frac{\lambda_k T \beta_k \mu_g}{32} \right) \mathbb{E}[\|z_{k+1} - y^*_k\|^2 | F_k] - \lambda_k J_k
\]

\[
+ O(\xi^2 t_{w,0}^2) \frac{\lambda_k \alpha_k^2}{T \mu_g T} q_k^2 + O(\xi^2 \lambda_k t_{w,0}^2) (\alpha_k^2 \sigma_f^2 + \beta_k^2 \sigma_g^2) + O(\lambda_k) T \gamma_k^2 \sigma_g^2.
\]

With \( \beta_k \leq \gamma_k \), and thus \( \delta_k / \lambda_k < T \mu_g \gamma_k / 32 \), we have that

\[
(ii) \leq \lambda_k \left( 1 + \frac{T \gamma_k \mu_g}{2} \right) \mathbb{E}[\|z_{k+1} - y^*_k\|^2 | F_k] - \lambda_k J_k + (iv)
\]

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

\[
(ii) \leq - \lambda_k T \mu_g \gamma_k \frac{T \beta_k \mu_g}{4} J_k + O(\xi^2 t_{w,0}^2) \frac{\alpha_k \sigma_f^2}{T \mu_g T} q_k^2 + O(\xi^2 \lambda_k t_{w,0}^2) (\alpha_k^2 \sigma_f^2 + \beta_k^2 \sigma_g^2) + O(\lambda_k) T \gamma_k^2 \sigma_g^2.
\]

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

\[
\mathbb{E}[\mathcal{V}_{k+1} - \mathcal{V}_k | \mathcal{F}_k] \leq - \frac{\xi \alpha_k}{2} || \nabla F(x_k) ||^2 + \frac{\xi \alpha_k}{2} \cdot 3C \lambda_k^{-2} + \frac{\xi^2 l_{f,0}^2}{2} (\alpha_k^2 \sigma_f^2 + \beta_k^2 \sigma_g^2)
\]

\[
- \frac{\xi \alpha_k}{4} \left( 1 - O \left( \frac{\xi l_{g,1} l_{w,0}^2 \beta_k}{\mu_g T \gamma_k} \right) \right) \mathbb{E}[||q_k^2||^2 | F_k]
\]

\[
- \lambda_k l_{g,1} T \mu_g \gamma_k \frac{T \beta_k \mu_g}{4} J_k
\]

\[
+ O(T + \xi^2 t_{w,0}^2) \cdot l_{g,1} \lambda_k (\alpha_k^2 \sigma_f^2 + (\beta_k^2 + \gamma_k^2) \sigma_g^2) + O \left( \frac{l_{g,1} l_{f,0}^2}{\mu_g^3} \right) \delta_k \frac{\lambda_k}{\lambda_k^2}.
\]

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

**Step-size rules:**

\[
\xi l_{g,1} l_{w,0}^2 \beta_k \leq c_1 \mu_g T \gamma_k,
\]

\[
\xi l_{g,1} l_{w,0} \leq c_2 \mu_g T.
\]
for some absolute constants \(c_1, c_2 > 0\), which holds by \(\beta_k \leq \gamma_k\) and (3b) with sufficiently small \(c_\xi > 0\). Once this holds, we can conclude that

\[
\mathbb{E}[V_{k+1} - V_k | F_k] \leq -\frac{\xi \alpha_k \lambda_k}{2} \|\nabla F(x_k)\|^2 - \frac{\lambda_k \mu_k \gamma_k}{4} \|z_k - y_k\|^2 - \frac{\lambda_k \mu_k \gamma_k}{4} \|y_k - y_{\lambda,k}^*\|^2 \\
+ O(\xi C_\lambda^2 \alpha_k \lambda_k^2) + O\left(\frac{l_{g,1}^2 f_0}{\mu_g} \right) \frac{\delta_k}{\lambda_k^2} + O(\xi^2 I_{F,1})(\alpha_k^2 \sigma_\gamma^2 + \beta_k^2 \sigma_\delta^2) \\
+ O(T + \xi^2 I_{\epsilon,0}) \cdot l_{g,1} \lambda_k (\alpha_k^2 \sigma_\gamma^2 + (\beta_k^2 + \gamma_k^2) \sigma_\delta^2).
\]

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 \(\lambda_k = \text{poly}(k)\), 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

\[
\frac{c_\gamma}{2c_\alpha} ((k + k_0 + 1)^{a-c} - (k + k_0)^{a-c}) \leq \frac{(a-c)c_\gamma}{2c_\alpha} (k + k_0)^{1-c+a}.
\]

As long as \(-2c + a \geq 1 - c + a\), or equivalently, \(c \leq 1\) and \(T \geq 32\), it always holds that

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

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

\[
\sum_{k=0}^{K-1} \frac{\mathbb{E}[\|\nabla F(x_k)\|^2]}{(k + k_0)^a} \leq O_T(1) \cdot \sum_k \frac{1}{(k + k_0)^{3a-2c}} + O_T(\sigma_\gamma^2) \cdot \sum_k \frac{1}{(k + k_0)^{a+c}} \\
+ O_T(\sigma_\delta^2) \cdot \sum_k \frac{1}{(k + k_0)^{3c-a}} + O_T(1).
\]

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 deising, we have \(\lambda_k = \gamma_k / (2\alpha_k) = \frac{2\lambda_k}{k_0} (k + k_0)^{b-c} = O(k^b)\). Let \(R\) be a random variable uniformly distributed over \{0, 1, ..., \(K\)\}. Note that the left hand side is larger than

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

We consider three regimes:

**Stochasticity in both upper-level and lower-level objectives: \(\sigma_\gamma^2, \sigma_\delta^2 > 0\).** In this case, we set \(a = 5/7, c = 4/7\), and thus \(\lambda_k = k^{1/7}\). The dominating term is \(\sigma_\gamma^2 \cdot \sum_k (\gamma_k^2 \lambda_k) = \sum_k O(k^{-1}) = O(\log K)\) 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: \(\sigma_\gamma^2 > 0, \sigma_\delta^2 = 0\).** In this case, we can take \(a = 3/5, c = 2/5\). When \(\sigma_\delta^2 = 0\), the dominating term is \(\sigma_\gamma^2 \cdot \sum_k (\gamma_k^2 \lambda_k) = \sum_k O(k^{-1}) = O(\log K)\) 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

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

\[ \tilde{h}_z^k := \nabla_y g(x, z_k; \phi_z^k) + (1 - \eta_k) \left( \tilde{h}_z^{k-1} - \nabla_y g(x_{k-1}, z_{k-1}; \phi_z^k) \right), \]
\[ \tilde{h}_{fy}^k := \nabla_y f(x, y_k; \phi_y^k) + (1 - \eta_k) \left( \tilde{h}_{fy}^{k-1} - \nabla_y f(x_{k-1}, y_{k-1}; \phi_y^k) \right), \]
\[ \tilde{h}_{gy}^k := \nabla_g g(x, y_k; \phi_y^k) + (1 - \eta_k) \left( \tilde{h}_{gy}^{k-1} - \nabla_g g(x_{k-1}, y_{k-1}; \phi_y^k) \right), \]
for the inner variable updates, and

\[ \tilde{h}_{fx}^k := \nabla_x f(x, y_{k+1}; \phi_x^k) + (1 - \eta_k) \left( \tilde{h}_{fx}^{k-1} - \nabla_x f(x_{k-1}, y_k; \phi_y^k) \right), \]
\[ \tilde{h}_{gy}^k := \nabla_g g(x, z_{k+1}; \phi_x^k) + (1 - \eta_k) \left( \tilde{h}_{gy}^{k-1} - \nabla_g g(x_{k-1}, z_k; \phi_x^k) \right). \]

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

\[ \bar{e}_x^k := \tilde{h}_{fx}^k + \lambda_k (\tilde{h}_{gy}^k - \tilde{h}_{gz}^k) - q_x^k, \]
\[ \bar{e}_y^k := \tilde{h}_{fy}^k + \lambda_k (\tilde{h}_{gz}^k) - q_y^k, \]
\[ \bar{e}_z^k := \tilde{h}_z^k - q_z^k, \]

where \( q_x^k, q_y^k, q_z^k \) 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

\[ \mathbb{E}[\|y^*(x_{k+1}) - y^*(x_k)\|^2 | \mathcal{F}_k] \leq 2\xi^2\eta^2_0\alpha_k^2 \left( \mathbb{E}[\|q_x^k\|^2 | \mathcal{F}_k] + \mathbb{E}[\|\bar{e}_x^k\|^2] \right). \]

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

\[ \mathbb{E}[\|y^*_{\lambda_{k+1}}(x_{k+1}) - y^*_{\lambda_k}(x_k)\|^2 | \mathcal{F}_k] \leq 4\xi^2\eta^2_0\alpha_k^2 \left( \mathbb{E}[\|q_x^k\|^2 | \mathcal{F}_k] + \mathbb{E}[\|\bar{e}_x^k\|^2] \right) + \frac{8\delta^2\eta^2_0}{\lambda_k^2\mu_g^2}. \]

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.**

\[ \mathbb{E}[\|\bar{e}_{x,1}^{k+1}\|^2] \leq (1 - \eta_{k+1})^2 (1 + 8\eta_{g,1}\gamma_k^2) \mathbb{E}[\|\bar{e}_{x,1}^k\|^2] + 2\eta_{k+1}^2 \sigma_g^2 + 8\eta_{g,1}^2 (1 - \eta_{k+1})^2 (\xi^2\alpha_k^2 \mathbb{E}[\|\bar{e}_x^k\|^2] + \xi^2\alpha_k^2 \mathbb{E}[\|\bar{e}_y^k\|^2] + \gamma_k^2 \mathbb{E}[\|q_y^k\|^2]), \]
\[ \mathbb{E}[\|\bar{e}_{y,1}^{k+1}\|^2] \leq (1 - \eta_{k+1})^2 (1 + 96\eta_{g,1}\beta_k^2) \mathbb{E}[\|\bar{e}_{y,1}^k\|^2] + 2\eta_{k+1}^2 (\sigma_f^2 + \lambda_{k+1}^2 + \sigma_g^2) + 12\delta_{k+1}^2 \sigma_g^2 + 96\eta_{g,1}^2 (1 - \eta_{k+1})^2 \beta_k^2 (\xi^2\gamma_k^2 + \xi^2\|\bar{e}_x^k\|^2 + \xi^2\|\bar{e}_y^k\|^2 + \gamma_k^2 \mathbb{E}[\|q_y^k\|^2]). \]
Lemma C.4.

\[
E[\|\tilde{e}_{k+1}\|^2] \leq (1 - \eta_{k+1})^2(1 + 240l_g^2, \sigma_g^2)E[\|\tilde{e}_k^2\|^2] + 6\eta_{k+1}^2(\sigma_f^2 + \lambda_{g,k+1}^2\sigma_g^2) + 80\delta_{g,k}^2\sigma_g^2 \\
+ 240l_g^2(1 - \eta_{k+1})^2 \lambda_g^2 (\|e_k\|^2 + \|e_k^2\|^2) + \gamma_k^2(\|e_k\|^2 + \|e_k\|^2) .
\]

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

\[
E[\|z_{k+1} - y_{k+1}\|^2|F_k] \leq (1 + \gamma_k \mu_g / 4)E[\|z_{k+1} - y_{k+1}\|^2|F_k] \\
+ O\left( \frac{\xi^2 \alpha_k^2 l_g^2}{\gamma_k \mu_g} \right) \cdot (E[\|q_k\|^2|F_k] + E[\|\tilde{e}_k\|^2|F_k]).
\]

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

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

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

\[\Box\]

Lemma C.6. If \( \gamma_k \leq 1/(16l_g,1) \), then

\[
E[\|z_{k+1} - y_{k}^*\|^2|F_k] \leq (1 - \gamma_k \mu_g / 2)E[\|z_{k} - y_{k}^*\|^2|F_k] - \frac{\gamma_k}{l_{g,1}} \|q_k\|^2 + O\left( \frac{\gamma_k}{\mu_g} \right) E[\|\tilde{e}_k\|^2|F_k].
\]

Proof. Note that

\[
\|z_{k+1} - y_{k}^*\|^2 = \|z_{k} - y_{k}^*\|^2 + \gamma_k^2 \|\hat{e}_k\|^2 - 2\gamma_k \langle \hat{e}_k, z_{k} - y_{k}^* \rangle \\
\leq \|z_{k} - y_{k}^*\|^2 + 2\gamma_k^2(\|\hat{e}_k\|^2 + \|\tilde{e}_k\|^2) - 2\gamma_k \langle \hat{e}_k, z_{k} - y_{k}^* \rangle - 2\gamma_k \langle \hat{e}_k, z_{k} - y_{k}^* \rangle.
\]

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

\[
\left( \mu_g \|z_{k} - y_{k}^*\|^2, \frac{1}{l_{g,1}} \|q_k\|^2 \right) \leq \langle q_k, z_{k} - y_{k}^* \rangle,
\]

and thus, given \( \gamma_k \leq 1/(16l_g,1) \), we have

\[
E[\|z_{k+1} - y_{k}^*\|^2|F_k] \leq (1 - 3\gamma_k \mu_g / 4)E[\|z_{k} - y_{k}^*\|^2|F_k] - \frac{\gamma_k}{l_{g,1}} \|q_k\|^2 + 2\gamma_k^2 E[\|\tilde{e}_k\|^2|F_k] - 2\gamma_k \langle \hat{e}_k, z_{k} - y_{k}^* \rangle.
\]

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

\[
-2\gamma_k \langle \hat{e}_k, z_{k} - y_{k}^* \rangle \leq \frac{\gamma_k \mu_g}{4} \|z_{k} - y_{k}^*\|^2 + \frac{4\gamma_k}{\mu_g} \|\hat{e}_k\|^2.
\]

Plugging this back, with \( \gamma_k \ll \frac{2\lambda_g}{\mu_g} \), we get the lemma.

\[\Box\]
C.4. Descent Lemma for $y_k$ towards $y_{\lambda,k}^*$

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

\[
\mathbb{E}[\|y_{k+1} - y_{\lambda,k+1}^*\|^2 | \mathcal{F}_k] \leq (1 + \beta_k \mu_g / 4) \mathbb{E}[\|y_{k+1} - y_{\lambda,k}^*\|^2 | \mathcal{F}_k] + O \left( \frac{\xi^2 \alpha_k^2 \beta_0}{\beta_k \mu_g} \right) \cdot \left( \mathbb{E}[\|q_k^e\|^2 | \mathcal{F}_k] + \mathbb{E}[\|e_k^e\|^2 | \mathcal{F}_k] \right) + O \left( \frac{\delta^2 \xi^2}{\lambda_k^2 \mu_g^2} \right).
\]

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

\[
\|y_{k+1} - y_{\lambda,k+1}^*\|^2 = \|y_{k+1} - y_{\lambda,k}^*\|^2 + \|y_{\lambda,k}^* - y_{\lambda,k}^*\|^2 - 2\langle y_{k+1} - y_{\lambda,k}^*, y_{\lambda,k}^* - y_{\lambda,k}^* \rangle
\]

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

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

\[
\mathbb{E}[\|y_{k+1} - y_{\lambda,k}^*\|^2 | \mathcal{F}_k] \leq (1 - \beta_k \mu_g / 2) \mathbb{E}[\|y_{k} - y_{\lambda,k}^*\|^2 | \mathcal{F}_k] - \frac{\alpha_k}{\lambda_k l_{g,1}} \|q_k^y\|^2 + O \left( \frac{\alpha_k^2 \xi^2}{\mu_g \beta_k} \right) \mathbb{E}[\|e_k^y\|^2 | \mathcal{F}_k])
\]

**Proof.** Note that

\[
\|y_{k+1} - y_{\lambda,k}^*\|^2 = \|y_k - y_{\lambda,k}^*\|^2 + 2\alpha_k (\|q_k^y\|^2 + \|e_k^y\|^2) - 2\alpha_k \langle q_k^y, y_k - y_{\lambda,k}^* \rangle - 2\alpha_k \langle e_k^y, y_k - y_{\lambda,k}^* \rangle,
\]

where we used $y_{k+1} - y_k = q_k^y + e_k^y$. 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

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

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

\[
\mathbb{E}[\|y_{k+1} - y_{\lambda,k}^*\|^2 | \mathcal{F}_k] \leq (1 - 3\beta_k \mu_g / 4) \mathbb{E}[\|y_k - y_{\lambda,k}^*\|^2 | \mathcal{F}_k] - \frac{\alpha_k}{\lambda_k l_{g,1}} \|q_k^y\|^2
\]

\[
+ 2\alpha_k \mathbb{E}[\|e_k^y\|^2 | \mathcal{F}_k] - 2\alpha_k \mathbb{E}[\langle e_k^y, y_k - y_{\lambda,k}^* \rangle | \mathcal{F}_k]
\]

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

\[
-2\alpha_k \langle e_k^y, y_k - y_{\lambda,k}^* \rangle \leq -\frac{\beta_k \mu_g}{4} \|y_k - y_{\lambda,k}^*\|^2 + \frac{4\alpha_k^2}{\beta_k \mu_g} \|e_k^y\|^2.
\]

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

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

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

\[
\mathbb{E}[F(x_{k+1}) - F(x_k) | \mathcal{F}_k] \leq -\frac{\xi \alpha_k}{4} \|\nabla F(x_k)\|^2 - \frac{\xi \alpha_k}{4} \mathbb{E}[\|q_k^e\|^2 | \mathcal{F}_k] + 2\xi \alpha_k \cdot \mathbb{E}[\|e_k^e\|^2 | \mathcal{F}_k]
\]

\[
+ \frac{3\xi \lambda_k}{2} \left( 4l_{g,1}^2 - \left( y_{k+1}^* - y_{\lambda,k}^* \right) \|y_{k+1} - y_{\lambda,k}^*\|^2 + l_{g,1}^2 \lambda_k^2 \|z_{k+1} - y_{\lambda,k}^*\|^2 + C_3^2 / \lambda_k^2 \right).
\]
**Proof.** Using the smoothness of $F$,  
\[
F(x_{k+1}) - F(x_k) \leq \langle \nabla F(x_k), x_{k+1} - x_k \rangle + \frac{L_{F,1}}{2} \|x_{k+1} - x_k\|^2.
\]

Note that $x_{k+1} - x_k = \xi \alpha_k (q_k^x + \tilde{e}_k^x)$, and thus  
\[
F(x_{k+1}) - F(x_k) \leq -\xi \alpha_k \langle \nabla_x F(x_k), q_k^x \rangle - \xi \alpha_k \langle \nabla_x F(x_k), \tilde{e}_k^x \rangle + \frac{L_{F,1}}{2} \|x_{k+1} - x_k\|^2 \\
\leq -\frac{\xi \alpha_k}{2} (\|\nabla F(x_k)\|^2 + \|q_k^x\|^2 - \|\nabla F(x_k) - q_k^x\|^2) - \xi \alpha_k \langle \nabla_x F(x_k), \tilde{e}_k^x \rangle + \xi^2 \alpha_k^2 l_{F,1} (\|q_k^x\|^2 + \|\tilde{e}_k^x\|^2).
\]

Using $\|\langle a, b \rangle\| \leq c\|a\|^2 + \frac{1}{c}\|b\|^2$, we have  
\[
-\xi \alpha_k \langle \nabla_x F(x_k), \tilde{e}_k^x \rangle \leq \frac{\xi \alpha_k}{4} \|\nabla_x F(x_k)\|^2 + \xi \alpha_k \|\tilde{e}_k^x\|^2.
\]

Finally, recall (16). Using $(a + b + c)^2 \leq 3(a^2 + b^2 + c^2)$, we have  
\[
\|\nabla F(x_k) - q_k^x\|^2 \leq 3\left(2l_{g,1}^2 \lambda_k^2 \|y_{k+1} - y_{\lambda,k}\|^2 + l_{g,1}^2 \lambda_k^2 \|z_{k+1} - y_{k}^*\|^2 + C_5^2 / \lambda_k^2 \right).
\]

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

**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:
\[
\mathbb{E}[\mathbb{V}_{k+1} - \mathbb{V}_k | \mathcal{F}_k] \leq -\frac{\xi \alpha_k}{4} \|\nabla F(x_k)\|^2 - \frac{\xi \alpha_k}{4} \mathbb{E}[\|q_k^y\|^2 | \mathcal{F}_k] + \frac{\xi \alpha_k}{2} \frac{3C_5^2}{\lambda_k^2} + \frac{l_{g,1} \lambda_k}{2} \left( 1 + \frac{\delta_k}{\lambda_k} \right) \|y_{k+1} - y_{\lambda,k}^*\|^2 + 6 \xi \alpha_k \lambda_k l_{g,1} \|y_{k+1} - y_{\lambda,k}^*\|^2 - \|y_k - y_{\lambda,k}^*\|^2 \\
+ \frac{l_{g,1} \lambda_k}{2} \left( 1 + \frac{\delta_k}{\lambda_k} \right) \|z_{k+1} - y_{\lambda,k}^*\|^2 + 3 \xi \alpha_k \lambda_k l_{g,1} \|z_{k+1} - z_{\lambda,k}^*\|^2 - \|z_k - y_{\lambda,k}^*\|^2 \\
+ \frac{1}{c_\eta l_{g,1}^2} \left( \frac{\mathbb{E}[\|\tilde{e}_{k+1}^x\|^2 | \mathcal{F}_k]}{\gamma_k \lambda_k} - \frac{\mathbb{E}[\|\tilde{e}_{k}^x\|^2 | \mathcal{F}_k]}{\gamma_{k-1} \lambda_k} \right) + 2 \xi \alpha_k \mathbb{E}[\|\tilde{e}_k^x\|^2 | \mathcal{F}_k] \\
+ \frac{1}{c_\eta l_{g,1}^2} \left( \frac{\mathbb{E}[\|\tilde{e}_{k+1}^y\|^2 | \mathcal{F}_k]}{\gamma_k \lambda_k} - \frac{\mathbb{E}[\|\tilde{e}_k^y\|^2 | \mathcal{F}_k]}{\gamma_{k-1} \lambda_k} \right) + \frac{\lambda_k}{c_\eta l_{g,1}^2} \left( \frac{\mathbb{E}[\|\tilde{e}_{k+1}^z\|^2 | \mathcal{F}_k]}{\gamma_k} - \frac{\mathbb{E}[\|\tilde{e}_k^z\|^2 | \mathcal{F}_k]}{\gamma_{k-1}} \right).
\]

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
\[
(i) \leq -\frac{\mu_g \beta_k}{8} \|y_k - y_{\lambda,k}^*\|^2 - \frac{\alpha_k}{\lambda_k l_{g,1}} \|q_k^y\|^2 + O \left( \frac{\xi^2 \alpha_k^2 \lambda_k^2}{\beta_k \mu_g} \mathbb{E}[\|q_k^y\|^2 | \mathcal{F}_k] + \frac{\xi^2 \alpha_k^2 \lambda_k^2}{\beta_k \mu_g} \mathbb{E}[\|\tilde{e}_k^y\|^2 | \mathcal{F}_k] + \frac{\alpha_k^2}{\beta_k \mu_g} \mathbb{E}[\|\tilde{e}_k^y\|^2 | \mathcal{F}_k] \right),
\]
\[
(ii) \leq -\frac{\mu_g \gamma_k}{8} \|z_k - y_{\lambda,k}^*\|^2 - \frac{\gamma_k}{l_{g,1}} \|q_k^y\|^2 + O \left( \frac{\xi^2 \alpha_k^2 \lambda_k^2}{\gamma_k \mu_g} \mathbb{E}[\|q_k^y\|^2 | \mathcal{F}_k] + \frac{\xi^2 \alpha_k^2 \lambda_k^2}{\gamma_k \mu_g} \mathbb{E}[\|\tilde{e}_k^y\|^2 | \mathcal{F}_k] + \frac{\gamma_k}{\mu_g} \mathbb{E}[\|\tilde{e}_k^y\|^2 | \mathcal{F}_k] \right),
\]

22
We can use Lemma C.4 to bound \((iii), (iv)\) and \((v)\). Using the step-size condition given in (8b), we have
\[
\frac{(1 - \eta_{k+1})}{\gamma_k} - \frac{1}{\gamma_{k-1}} \gamma_k - \frac{\gamma_{k-1}}{\gamma_k} - \eta_{k+1} \leq -\eta_{k+1} \frac{1}{2\gamma_k}.
\]
Note that by the same step-size condition, \(\eta_{k+1} \geq O(l_{g,1}^2 l_{k}^2)\), and thus,
\[
(iii) \leq -\eta_{k+1} \frac{1}{2\gamma_k \lambda_k} \mathbb{E}[(\epsilon_k^2)^2 | \mathcal{F}_k] + O (\sigma_f^2) \cdot \frac{\eta_{k+1}}{\lambda_k \gamma_k} + O (\sigma_g^2) \cdot \left( \frac{\eta_{k+1} \lambda_k}{\gamma_k} + \frac{\delta_k^2}{\gamma_k \lambda_k} \right) + O(l_{g,1}^2) \cdot \left( \xi^2 \alpha_k \|q_k^w\|^2 + \alpha_k (\|q_k^w\|^2 + \|q_k^v\|^2) + \gamma_k \lambda_k (\|q_k^w\|^2 + \|q_k^v\|^2) \right).
\]
Similarly, we can use Lemma C.3 and show that
\[
(iv) \leq -\eta_{k+1} \frac{1}{2\gamma_k} \mathbb{E}[(\epsilon_k^2)_{[\mathcal{F}_k]}] + O (\sigma_f^2) \cdot \frac{\eta_{k+1}}{\lambda_k \gamma_k} + O (l_{g,1}^2) \cdot \left( \frac{\xi^2 \alpha_k}{\gamma_k} \|q_k^w\|^2 + \frac{\xi^2 \alpha_k^2}{\gamma_k} \|q_k^v\|^2 + \gamma_k \|q_k^w\|^2 \right).
\]
Plugging inequalities for \((i) - (v)\) back and arranging terms, we get
\[
\mathbb{E}[\nabla V_{k+1} - \nabla V_k | \mathcal{F}_k] \leq -\frac{\xi \alpha_k}{4} \|\nabla F(x_k)\|^2 - \frac{\xi \alpha_k}{4} \mathbb{E}[(\epsilon_k^2)^2 | \mathcal{F}_k] + \frac{3C_k^2}{2} \frac{\xi \alpha_k}{\lambda_k} - l_{g,1} \lambda_k (i) + \frac{l_{g,1} \lambda_k}{2} (ii)
\]
\[
+ \frac{1}{c_{\eta} g^2} (iii) + 2\xi \alpha_k \mathbb{E}[(\epsilon_k^2)^2 | \mathcal{F}_k] + \frac{1}{c_{\eta} l_{g,1}^2} (iv) + \frac{\lambda_k}{c_{\eta} l_{g,1}^2} (v)
\]
\[
\leq -\frac{\xi \alpha_k}{4} \|\nabla F(x_k)\|^2 - \frac{\lambda_k l_{g,1} \mu g y_k}{4} \|y_k - y_k^*\|^2 - \frac{\lambda_k l_{g,1} \mu g k}{4} \|y_k - y_k^*\|^2
\]
\[
- \frac{\xi \alpha_k}{4} \mathbb{E}[q^w_k | \mathcal{F}_k] - \frac{1}{c_{\eta} g^2} (l_{g,1}^2 (1 - O(\xi l_{g,1} l_{g,1}^2 | \mu g - O(\xi c_{g}^{-1}))
\]
\[
- \frac{\alpha_k}{4} \mathbb{E}[q^w_k | \mathcal{F}_k] (1 - O(c_{g}^{-1})) - \frac{\gamma_k \lambda_k}{2} \mathbb{E}[q^w_k | \mathcal{F}_k] (1 - O(c_{g}^{-1}))
\]
\[
+ O \left( C_k^2 \frac{\xi \alpha_k}{\lambda_k} + \frac{l_{g,1}^2 l_{g,1}^2}{\mu g^2} \right) + \text{noise variance terms},
\]
where noise variance terms are
\[
\text{noise terms} = -\frac{\mathbb{E}[q^w_k | \mathcal{F}_k]}{c_{\eta} g^2} \left( \frac{\eta_{k+1}}{2\gamma_k \lambda_k} - O \left( l_{g,1}^2 \xi \alpha_k \right) \right) \left( \eta_{k+1} \xi \alpha_k l_{g,1}^2 \right) - \frac{\mathbb{E}[q^w_k | \mathcal{F}_k]}{l_{g,1}^2 c_{\eta} g^2} \left( \frac{\eta_{k+1}}{2\gamma_k \lambda_k} - O \left( l_{g,1}^2 \xi \alpha_k \right) \right) \left( \eta_{k+1} \xi \alpha_k l_{g,1}^2 \right)
\]
\[
- \frac{\lambda_k \mathbb{E}[q^v_k | \mathcal{F}_k]}{l_{g,1}^2 c_{\eta} g^2} \left( \frac{\eta_{k+1}}{2\gamma_k \lambda_k} - O \left( l_{g,1}^2 \xi \alpha_k \right) \right) \left( \eta_{k+1} \xi \alpha_k l_{g,1}^2 \right)
\]
\[
+ \frac{1}{c_{\eta} l_{g,1}^2} \left( O(\sigma_f^2) \cdot \frac{\eta_{k+1}^2}{\lambda_k \gamma_k} + O(\sigma_g^2) \cdot \left( \frac{\eta_{k+1} \lambda_k}{\gamma_k} + \frac{\delta_k^2}{\gamma_k \lambda_k} \right) \right).
\]
For the all squared terms, with careful design of step-sizes, we can make the coefficient negative. Specifically, we need
\[
\xi l_{g,1} l_{g,1}^2 \ll 1, c_{\eta} \gg 1,
\]
to negate \(q^{(c)} \) 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:

$$\mathbb{E}[\nabla F(x_{k+1}) - \nabla F(x_k)] \leq -\frac{\xi \alpha_k}{4} \|\nabla F(x_k)\|^2 + O(\xi C^2) \cdot \frac{\alpha_k}{\lambda_k} + O(l^2_{f,0} \log 1/\rho) \cdot \frac{\delta_k^2}{\lambda_k}$$

$$+ \frac{1}{c \eta_{g,k}^2} \left( O(\alpha_g^2) \cdot \eta_{k+1}^2 \lambda_k + O(\sigma_g^2) \cdot \left( \frac{\eta_{k+1}^2 \lambda_k}{\gamma_k} + \frac{\delta_k^2}{\gamma_k \lambda_k} \right) \right).$$

**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

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

$$+ O(\sigma_g^2) \cdot \sum_k 1 \cdot \frac{1}{(k + k_0)^{-4c+a}} + O(1).$$

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 $\lambda_k \asymp k^{3/5}$, which yields

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

**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

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

**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

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

**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

$$\nabla \mathcal{L}_\lambda^*(x) = \nabla_x \mathcal{L}_\lambda(x, y^*_\lambda(x)) + \nabla_x y^*_\lambda(x) \nabla_y \mathcal{L}_\lambda(x, y^*_\lambda(x)) = \nabla_x \mathcal{L}_\lambda(x, y^*_\lambda(x)).$$
To compare this to $\nabla F(x)$, we can invoke Lemma A.2 which gives

$$\|\nabla F(x) - \nabla_x L_\lambda(x, y^*_\lambda(x))\| \leq 2(l_{g,1}/\mu_g)\|y^*_\lambda(x) - y^*(x)\|(l_{f,1} + \lambda \cdot \min(2l_{g,1}, l_{g,2})\|y^*(x) - y^*_\lambda(x)\|).$$

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

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

\[ \square \]

D.1.2. PROOF OF LEMMA 3.2

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

$$L_\lambda(x, y) = f(x, y) + \lambda(g(x, y) - g^*(x)),$$

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 $L_\lambda(x, y)$ in $y$.

By the optimality condition at $y^*_\lambda(x_1)$ with $x_1, \lambda_1$, we have

$$\nabla_y f(x_1, y^*_\lambda(x_1)) + \lambda_1 \nabla_y g(x_1, y^*_\lambda(x_1)) = 0,$$

which also implies that $\|g(x_1, y^*_\lambda(x_1))\| \leq l_{f,0}/\lambda_1$. Observe that

$$\nabla_y f(x_2, y^*_\lambda(x_1)) + \lambda_2 \nabla_y g(x_2, y^*_\lambda(x_1)) = (\nabla_y f(x_2, y^*_\lambda(x_1)) - \nabla_y f(x_1, y^*_\lambda(x_1))) + \nabla_y f(x_1, y^*_\lambda(x_1))$$

$$+ \lambda_2(\nabla_y g(x_2, y^*_\lambda(x_1)) - \nabla_y g(x_1, y^*_\lambda(x_1))) + \lambda_2 \nabla_y g(x_1, y^*_\lambda(x_1))$$

$$= (\nabla_y f(x_2, y^*_\lambda(x_1)) - \nabla_y f(x_1, y^*_\lambda(x_1))) + \lambda_2(\nabla_y g(x_2, y^*_\lambda(x_1)) - \nabla_y g(x_1, y^*_\lambda(x_1)))$$

$$+ (\lambda_2 - \lambda_1) \nabla_y g(x_1, y^*_\lambda(x_1)),$$

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

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

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

$$\frac{\lambda_2\mu_g}{2} \|y^*_\lambda(x_1) - y^*_\lambda(x_2)\| \leq \|\nabla_y L_{\lambda_2}(x_2, y^*_\lambda(x_1))\| \leq (l_{f,1} + \lambda_2 l_{g,1})\|x_1 - x_2\| + \frac{\lambda_2 - \lambda_1}{\lambda_1} l_{f,0}.$$

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 $\lambda \geq 2l_{f,1}/\mu_g$,

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

\[ \square \]
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

\[
\nabla F(x) = \nabla_x f(x, y^*(x)) - \nabla_x^2 g(x, y^*(x)) \nabla_{yy}^2 g(x, y^*(x))^{-1} \nabla_y f(x, y^*(x)).
\]

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

\[
\| \nabla F(x_1) - \nabla F(x_2) \| \leq \left( l_{f,1} + \frac{l_{f,0} l_{g,2} + l_{g,1} l_{g,1}}{\mu_g} \right) (\| x_1 - x_2 \| + \| y^*(x_1) - y^*(x_2) \|)
\]

\[
+ l_{g,1} l_{f,0} \| \nabla_{yy}^2 g(x_1, y^*(x_1))^{-1} - \nabla_{yy}^2 g(x_2, y^*(x_2))^{-1} \|.
\]

Thus,

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

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

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 \( y^* = y^*(x) \).

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

\[
\nabla_y \mathcal{L}_\lambda(x, y) = \nabla_y f(x, y) + \lambda \nabla_y g(x, y).
\]

Check that

\[
\nabla F(x) - \nabla_x \mathcal{L}_\lambda(x, y) = \nabla_x f(x, y^*) - \nabla_x f(x, y)
\]

\[
- \nabla_{xx}^2 g(x, y^*) \nabla_{yy}^2 g(x, y^*)^{-1} \nabla_y f(x, y^*) - \lambda (\nabla_x g(x, y) - \nabla_x g(x, y^*)).
\]

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

\[
\nabla_x g(x, y) - \nabla_x g(x, y^*) = \nabla_x g(x, y) - \nabla_x g(x, y^*) - \nabla_{xy} g(x, y^*) (y - y^*)
\]

\[
+ \nabla_{xy} g(x, y^*) (y - y^*).
\]

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

\[
y - y^* = -\nabla_{yy} g(x, y^*)^{-1} (\nabla_y g(x, y) - \nabla_y g(x, y^*) - \nabla_{yy} g(x, y^*) (y - y^*))
\]

\[
+ \frac{1}{\lambda} \nabla_{yy} g(x, y^*)^{-1} (\nabla_y \mathcal{L}_\lambda(x, y) - \nabla_y f(x, y)).
\]

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

\[
\nabla F(x) - \nabla_x \mathcal{L}_\lambda(x, y) = (\nabla_x f(x, y^*) - \nabla_x f(x, y)) - \nabla_{xx}^2 g(x, y^*) \nabla_{yy}^2 g(x, y^*)^{-1} (\nabla_y f(x, y^*) - \nabla_y f(x, y))
\]

\[
- \nabla_{xy}^2 g(x, y^*) \nabla_{yy}^2 g(x, y^*)^{-1} \nabla_y \mathcal{L}_\lambda(x, y)
\]

\[
- \lambda (\nabla_x g(x, y) - \nabla_x g(x, y^*) - \nabla_{xy}^2 g(x, y^*) (y - y^*)
\]

\[
+ \frac{1}{\lambda} \nabla_{yy} g(x, y^*)^{-1} (\nabla_y \mathcal{L}_\lambda(x, y) - \nabla_y f(x, y)).
\]
+ \lambda \nabla^2_{xy} g(x, y^*) \nabla_{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 smoothness of \( \nabla g \) from Assumption 1, we have
\[
\| \nabla_y g(x, y) - \nabla_y g(x, y^*) - \nabla^2_{yy} g(x, y^*)(y - y^*) \| \leq l_{g,2} \| y - y^* \|^2.
\]
When \( \| y - y^* \| \) is too large, the smoothness of \( g \) can be more useful:
\[
\| \nabla_y g(x, y) - \nabla_y g(x, y^*) - \nabla^2_{yy} g(x, y^*)(y - y^*) \| \leq 2l_{g,1} \| y - y^* \|.
\]
Similarly, we have
\[
\| \nabla_x g(x, y) - \nabla_x g(x, y^*) - \nabla^2_{xy} g(x, y^*) \top (y - y^*) \| \leq \min \left( l_{g,2} \| y - y^* \|^2, 2l_{g,1} \| y - y^* \| \right).
\]
On the other hand, by smoothness of \( f \), we also have
\[
\| \nabla_x f(x, y^*) - \nabla_x f(x, y) \| \leq l_{f,1} \| y - y^* \|, \quad \| \nabla_y f(x, y^*) - \nabla_y f(x, y) \| \leq l_{f,1} \| y - y^* \|.
\]
We can conclude that
\[
\| \nabla F(x) - \nabla_x \mathcal{L}(x, y) + \nabla^2_{xy} g(x, y^*) \nabla^2_{yy} g(x, y^*)^{-1} \nabla_y \mathcal{L}(x, y) \|
\leq l_{f,1} (1 + l_{g,1}/\mu_g) \| y - y^* \| + \lambda (1 + l_{g,1}/\mu_g) \| y - y^* \| \min(l_{g,2} \| y - y^* \|, 2l_{g,1}).
\]
We know that \( l_{g,1}/\mu_g \geq 1 \) and thus, we have
\[
\| \nabla F(x) - \nabla_x \mathcal{L}(x, y) + \nabla^2_{xy} g(x, y^*) \nabla^2_{yy} g(x, y^*)^{-1} \nabla_y \mathcal{L}(x, y) \|
\leq 2(l_{g,1}/\mu_g) \| y - y^* \| (l_{f,1} + \lambda \cdot \min(2l_{g,2}, l_{g,2} \| y - y^* \|, 2l_{g,1})),
\]
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
\[
\nabla_y \mathcal{L}(x, y^*_\lambda(x)) = \nabla_y f(x, y^*_\lambda(x)) + \lambda \nabla_y g(x, y^*_\lambda(x)) = 0.
\]
Taking derivative with respect to \( x \), we get
\[
(\nabla^2_{yy} f(x, y^*_\lambda(x)) + \lambda \nabla^2_{yy} g(x, y^*_\lambda(x))) \nabla y^*_\lambda(x) = - (\nabla^2_{xy} g(x, y^*_\lambda(x)) + \lambda \nabla^2_{xy} g(x, y^*_\lambda(x))).
\]
As \( \lambda > 2l_{f,1}/\mu_g \), the left-hand side is positive definite with minimum 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
\[
\frac{\lambda \mu_g}{2} \| \nabla y^*_\lambda(x_1) - \nabla y^*_\lambda(x_2) \| \leq (l_{f,2} + \lambda l_{g,2}) (\| x_1 - x_2 \| + \| y^*_\lambda(x_1) - y^*_\lambda(x_2) \|) \max_{x \in \mathcal{X}} \| \nabla y^*_\lambda(x) \|
+ (l_{f,2} + \lambda l_{g,2}) (\| x_1 - x_2 \| + \| y^*_\lambda(x_1) - y^*_\lambda(x_2) \|)
\leq (l_{f,2} + \lambda l_{g,2}) (1 + l_{\lambda,0})^2 \| x_1 - x_2 \|.
\]
Arranging this, we get
\[
\| \nabla y^*_\lambda(x_1) - \nabla y^*_\lambda(x_2) \| \leq 32 \left( \frac{l_{f,2}}{\lambda} + l_{g,2} \right) \frac{l_{g,1}^2}{\mu_g^3} \| x_1 - x_2 \|.
\]

\( \square \)
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.

\[
E[\|y^*(x_{k+1}) - y^*(x_k)\|^2 | F_k] \leq \frac{l_{\sigma,1}^2}{2} E[\|x_{k+1} - x_k\|^2 | F_k]
\]

\[
\leq \frac{l_{\sigma,1}^2}{4} E\left[\left(\frac{l_{\sigma,1}}{l_{\sigma,1}} + 1\right) \cdot \|x_{k+1} - x_k\|^2 | F_k\right]
\]

\[
\leq \frac{l_{\sigma,1}^2}{4} E[\|v_k\|^2 \cdot E[\|x_{k+1} - x_k\|^2 | F_k'] | F_k]
\]

\[
+ \frac{\xi^2}{4} \left(\alpha_k^2 E[\|q_k\|^2] + \alpha_k^2 \sigma_f^2 + \beta_k^2 \sigma_g^2\right),
\]

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

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

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

\[
E[\|x_{k+1} - x_k\|^2 | F_k'] \leq 2\xi^2 (M_f \alpha_k^2 + 2M_g \beta_k^2) \leq 4M \xi^2 l^2_{\sigma,1} \beta_k^2,
\]

which yields

\[
E[\langle v_k, \nabla_y^*(x_{k+1}) - \nabla_y^*(x_k) (x_{k+1} - x_k) \rangle | F_k] 
\]

\[
\leq M \xi^2 l^2_{\sigma,1} \beta_k^2 E[\|v_k\|^2 | F_k] + \frac{\xi^2}{4} \left(\alpha_k^2 E[\|q_k\|^2 | F_k] + \alpha_k^2 \sigma_f^2 + \beta_k^2 \sigma_g^2\right).
\]

Combining all, we obtain the desired result. \( \square \)
D.2.6. Proof of Lemma A.6

**Proof.** We can start with the following decomposition:

\[
\langle v_k, y_{\lambda_{k+1}}^*(x_{k+1}) - y_{\lambda_k}^*(x_k) \rangle = \langle v_k, y_{\lambda_{k+1}}^*(x_{k+1}) - y_{\lambda_k}^*(x_{k+1}) \rangle + \langle v_k, \nabla y_{\lambda_k}^*(x_k)(x_{k+1} - x_k) \rangle + \langle v_k, y_{\lambda_k}^*(x_k) - y_{\lambda_k}^*(x_k) - \nabla y_{\lambda_k}^*(x_k)(x_{k+1} - x_k) \rangle.
\]

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 \( \langle a, b \rangle \leq c\|a\|^2 + \frac{1}{c}\|b\|^2 \),

\[
\mathbb{E}[\langle v_k, y_{\lambda_{k+1}}^*(x_{k+1}) - y_{\lambda_k}^*(x_{k+1}) \rangle | \mathcal{F}_k] \leq c\mathbb{E}[\|v_k\|^2] + \frac{1}{4c} \mathbb{E}[\|y_{\lambda_{k+1}}^*(x_{k+1}) - y_{\lambda_k}^*(x_{k+1})\|^2]
\]

\[
\leq c\mathbb{E}[\|v_k\|^2] + \frac{1}{c\lambda_k\lambda_{k+1}^2} \frac{l_{f,0}^2}{\mu_g^2},
\]

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

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

Adding this with bounds on other two terms, we get the lemma. \(\square\)

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

\[
\mathbb{E}[\|y^*(x_{k+1}) - y^*(x_k)\|^2] \leq l_{f,0}^2 \mathbb{E}[\|x_{k+1} - x_k\|^2]
\]

\[
\leq \xi^2 \alpha_k^2 \tau_{i,0}^2 \mathbb{E}[\|q_{\hat{k}}^* + \hat{e}_{\hat{k}}^*\|^2] \leq 2\xi^2 \alpha_k^2 \tau_{i,0}^2 (\mathbb{E}[\|q_{\hat{k}}^*\|^2] + \mathbb{E}[\|\hat{e}_{\hat{k}}^*\|^2]).
\]

D.3.2. Proof of Lemma C.2

Using Lemma 3.2, we have

\[
\mathbb{E}[\|y_{\lambda_{k+1}}^*(x_{k+1}) - y_{\lambda_k}^*(x_k)\|^2] \leq \frac{8\delta_k^2}{\lambda_k^2 \lambda_{k+1}^2} + 2l_{f,0}^2 \mathbb{E}[\|x_{k+1} - x_k\|^2]
\]

\[
\leq 4\xi^2 \alpha_k^2 \tau_{i,0}^2 (\mathbb{E}[\|q_{\hat{k}}^*\|^2] + \mathbb{E}[\|\hat{e}_{\hat{k}}^*\|^2]) + \frac{8\delta_k^2}{\lambda_k^2}.
\]

D.3.3. Proof of Lemma C.3

We can start with unfolding the expression for \( \mathbb{E}[\|\hat{e}_{k+1}^*\|^2] \).

\[
\mathbb{E}[\|\hat{e}_{k+1}^*\|^2] = \mathbb{E}[\|\hat{h}_{k+1}^* - q_{k+1}^*\|^2]
\]

\[
= \mathbb{E}[\|\nabla g(x_{k+1}, z_{k+1}; \phi_{k+1}^*) + (1 - \eta_{k+1})(\hat{h}_{k}^* - \nabla g(x_k, z_k; \phi_{k}^*) - q_{k+1})\|^2]
\]

\[
= \mathbb{E}[(1 - \eta_{k+1})\hat{e}_{k}^* + \nabla g(x_{k+1}, z_{k+1}; \phi_{k}^*)
\]

\[
+ (1 - \eta_{k+1})(q_{k}^* - \nabla g(x_k, z_k; \phi_{k}^*) - q_{k+1})\|^2
\]

\[
= (1 - \eta_{k+1})^2 \mathbb{E}[\|\hat{e}_{k}^*\|^2] + \mathbb{E}[\|\eta_k (\nabla g(x_{k+1}, z_{k+1}; \phi_{k}^*) - q_{k+1})
\]

\[
+ (1 - \eta_{k+1})(\nabla g(x_k, z_k; \phi_{k}^*) - q_{k}^* - q_{k+1})\|^2].
\]

In the last equality, we used

\[
\mathbb{E}[\mathbb{E}[\langle \hat{e}_{k}^*, \nabla g(x_{k+1}, z_{k+1}; \phi_{k+1}^*) - q_{k+1} \rangle | \mathcal{F}_{k+1}]] = 0,
\]

\[
\mathbb{E}[\mathbb{E}[\langle \hat{e}_{k}^*, \nabla g(x_{k+1}, z_{k+1}; \phi_{k+1}^*) - q_{k+1} \rangle | \mathcal{F}_{k+1}]] = 0.
\]
Also note that
\[ E[\|\nabla_y g(x_{k+1}, z_{k+1}; \phi_x^{k+1}) - q_k^x\|^2] \leq \sigma_y^2, \]
from the variance boundedness (Assumption 3). We also observe that
\[ E[\nabla_y g(x_{k+1}, z_{k+1}; \phi_x^{k+1}) - \nabla_y g(x_k, z_k; \phi_x^{k+1})]^2 \leq l_{g,1}^2 \left(\|x_{k+1} - x_k\|^2 + \|z_{k+1} - z_k\|^2\right) \]
\[ = l_{g,1}^2 (\xi^2 \alpha_k^2 \|q_k^x + \tilde{e}_k^x\|^2 + \gamma_k \|q_k^x + \tilde{e}_k^x\|^2), \]
due to Assumption 6. The same inequality holds for \( q_{k+1}^x - q_k^x \):
\[ E[\|q_{k+1}^x - q_k^x\|^2] \leq l_{g,1}^2 (\xi^2 \alpha_k^2 \|q_k^x + \tilde{e}_k^x\|^2 + \gamma_k \|q_k^x + \tilde{e}_k^x\|^2). \]
Now we plug these inequalities and using \( \|a + b\|^2 \leq 2(\|a\|^2 + \|b\|^2) \) multiple times, we have
\[ E[\|\tilde{e}_{k+1}^x\|^2] \leq (1 - \eta_k)^2 E[\|\tilde{e}_k^x\|^2] + 2\eta_k^2 \sigma_y^2 + 8l_{g,1}^2 (1 - \eta_k)^2 E[\|q_{k+1}^x (\xi^2 \alpha_k^2 \|q_k^x + \tilde{e}_k^x\|^2 + \gamma_k \|q_k^x + \tilde{e}_k^x\|^2)]. \]
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 \( q_k^y = E[q_k^y (\zeta, \phi)] \). Then we can get a similar bound for \( E[\|\tilde{e}_{k+1}^y\|^2] \):
\[ E[\|\tilde{e}_{k+1}^y\|^2] \leq (1 - \eta_k)^2 E[\|\tilde{e}_k^y\|^2] + 2\eta_k^2 E[\|q_{k+1}^y (\xi^2 \alpha_k^2 \|q_k^y + \tilde{e}_k^y\|^2 + \gamma_k \|q_k^y + \tilde{e}_k^y\|^2)] + 2(1 - \eta_k)^2 E[\|q_{k+1}^y (\xi^2 \alpha_k^2 \|q_k^y + \tilde{e}_k^y\|^2 + \gamma_k \|q_k^y + \tilde{e}_k^y\|^2)] + 2\eta_k^2 (q_k^y - q_{k+1}^y)^2. \]
Using the variance bound similarly, we have
\[ E[\|q_{k+1}^y (\xi^2 \alpha_k^2 \|q_k^y + \tilde{e}_k^y\|^2 + \gamma_k \|q_k^y + \tilde{e}_k^y\|^2)] \leq \sigma_y^2 + \lambda_k \sigma_y^2. \]
Then, we unfold the last term such that
\[ E[\|q_{k+1}^y (\xi^2 \alpha_k^2 \|q_k^y + \tilde{e}_k^y\|^2 + \gamma_k \|q_k^y + \tilde{e}_k^y\|^2)] \]
\[ = E[\|\nabla_y f(x_{k+1}, y_{k+1}; \phi_x^{k+1}) - \nabla_y f(x_k, y_k; \phi_x^{k+1}) + \nabla_y f(x_k, y_k) - \nabla_y f(x_{k+1}, y_{k+1})\|^2] \]
\[ + \lambda_k \nabla_y g(x_{k+1}, y_{k+1}; \phi_x^{k+1}) + \nabla_y g(x_k, y_k) - \nabla_y g(x_{k+1}, y_{k+1})\] \[ + \delta_k (\nabla_y g(x_{k+1}, y_{k+1}; \phi_x^{k+1}) - \nabla_y g(x_k, y_k) - \nabla_y g(x_k, y_k; \phi_x^{k+1}))^2 \]
\[ \leq 12(l_{f,1}^2 + l_{g,1}^2 \lambda_k^2 \|x_{k+1} - x_k\|^2 + \|y_{k+1} - y_k\|^2 + 12 \delta_k^2 \sigma_y^2 \]
\[ \leq 24(l_{f,1}^2 + l_{g,1}^2 \lambda_k^2 \|x_{k+1} - x_k\|^2 + \|y_{k+1} - y_k\|^2) + 12 \delta_k^2 \sigma_y^2. \]
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
\[ E[\|\tilde{e}_{k+1}^y\|^2] \leq (1 - \eta_k)^2 (1 + 96 l_{g,1}^2 \delta_k^2) E[\|\tilde{e}_k^y\|^2] + 2\eta_k^2 (\sigma_y^2 + \lambda_k \sigma_y^2) + 24 \delta_k^2 \sigma_y^2 \]
\[ + 96(1 - \eta_k)^2 l_{g,1}^2 \delta_k^2 \xi^2 \alpha_k^2 \|q_k^y\|^2 + \xi^2 \|\tilde{e}_k^y\|^2 + \|q_k^y\|^2 + \|\tilde{e}_k^y\|^2) + 12 \delta_k^2 \sigma_y^2. \]

D.3.4. Proof of Lemma C.4

Similarly to the case for \( E[\|\tilde{e}_{k+1}^y\|^2] \), let us define \( q_k^x (\zeta, \phi) := \nabla_x f(x_k, y_k; \zeta) + \lambda_k \nabla_x g(x_k, y_k; \phi) - \nabla_x g(x_k, z_k; \phi). \)

We note that \( \xi^2, \phi_x^2 \) are sampled after \( y_{k+1}, z_{k+1} \) is updated but before \( x_k \) is updated. Hence,
\[ E[\|q_k^x (\xi^2 (\zeta^{k+1}, \phi_x^{k+1}) - q_{k+1}^x)|F_{k+1}] = 0, \]
\[ E[\|q_k^x (\xi^2 (\zeta^{k+1}, \phi_x^{k+1}) - q_{k+1}^x)|F_{k+1}] = 0. \]
Thus, following similar procedure, we have
\[ E[\|\tilde{e}_{k+1}^x\|^2] = E[\|q_{k+1}^x (\xi^2 \alpha_k^2 \|q_k^x + \tilde{e}_k^x\|^2 + \gamma_k \|q_k^x + \tilde{e}_k^x\|^2) + (1 - \eta_k) \]

\[ \|q_k^x + \tilde{e}_k^x - q_{k+1}^x (\xi^2 (\zeta^{k+1}, \phi_x^{k+1}) - q_{k+1}^x)|F_{k+1}] = 0. \]

\[ \|q_{k+1}^x (\xi^2 \alpha_k^2 \|q_k^x + \tilde{e}_k^x\|^2 + \gamma_k \|q_k^x + \tilde{e}_k^x\|^2) + (1 - \eta_k) \]

\[ E[\|\tilde{e}_{k+1}^x\|^2] = E[\|q_{k+1}^x \]
\[
\begin{align*}
&\leq (1 - \eta_{k+1})^2 \mathbb{E}[\|\tilde{e}_k^x\|^2] + 2\eta_k^2 \mathbb{E}[\|q^x_{k+1} - q^x_{k+1}\|^2]
+ 2(1 - \eta_{k+1})^2 \mathbb{E}[\|q^x_{k+1} - q^x_{k+1}\| + (q^x_k - q^x_{k+1})^2].
\end{align*}
\]

Note that
\[
\begin{align*}
\mathbb{E}[\|q^x_{k+1} - q^x_{k+1}\| - q^x_{k+1}\|^2]
&= \mathbb{E}[\|\nabla_x f(x_{k+1}, y_{k+2}; x_{k+1}) - \nabla_x f(x_{k+1}, y_{k+2})
+ \lambda_k (\nabla_x g(x_{k+1}, y_{k+2}; x_{k+1}) - \nabla_x g(x_{k+1}, y_{k+2})) + \lambda_k (\nabla_x g(x_{k+1}, z_{k+2}; x_{k+1}) - \nabla_x g(x_{k+1}, z_{k+2}))\|\|^2]
&\leq 3(\sigma^2_f + \lambda_k^2 \sigma^2_g).
\end{align*}
\]

Finally, we have
\[
\begin{align*}
\mathbb{E}[\|q^x_{k+1} - q^x_{k+1}\| - q^x_{k+1}\|^2]
&= \mathbb{E}[\|\nabla_x f(x_{k+1}, y_{k+2}; x_{k+1}) - \nabla_x f(x_{k+1}, y_{k+2})
+ \lambda_k (\nabla_x g(x_{k+1}, y_{k+2}; x_{k+1}) - \nabla_x g(x_{k+1}, y_{k+2})) + \lambda_k (\nabla_x g(x_{k+1}, z_{k+2}; x_{k+1}) - \nabla_x g(x_{k+1}, z_{k+2}))
+ \delta_k (\nabla_y g(x_{k+1}, y_{k+2}; x_{k+1}) - \nabla_y g(x_{k+1}, y_{k+2}))
+ \delta_k (\nabla_y g(x_{k+1}, z_{k+2}; x_{k+1}) - \nabla_y g(x_{k+1}, z_{k+2}))\|\|^2]
\end{align*}
\]

Using Cauchy-Schwarz inequality, we get
\[
\begin{align*}
&\leq 3(\sigma^2_f + \lambda_k^2 \sigma^2_g)(\|x_{k+1} - x_k\|^2 + \|y_{k+2} - y_{k+1}\|^2 + \|z_{k+2} - z_{k+1}\|^2) + 40\delta^2_k \sigma^2_g
&\leq 120\eta^2_k \lambda_k^2 (\|q^x_k\|^2 + \|\tilde{e}^x_k\|^2) + \alpha^2_k (\|q^y_k\|^2 + \|\tilde{e}^y_k\|^2) + \gamma^2_{k+1} (\|q^\tilde{x}_k\|^2 + \|\tilde{e}^\tilde{x}_k\|^2) + 40\delta^2_k \sigma^2_g.
\end{align*}
\]

Combining all, we obtain the result.