On The Complexity of First-Order Methods in Stochastic Bilevel Optimization

Jeongyeol Kwon [1] **Dohyun Kwon** [2 3] **Hanbaek Lyu** [1]

Abstract

We consider the problem of finding stationary points in Bilevel optimization when the lower-level problem is unconstrained and strongly convex. The problem has been extensively studied in recent years; the main technical challenge is to keep track of lower-level solutions $y^*(x)$ in response to the changes in the upper-level variables x. Subsequently, all existing approaches tie their analyses to a genie algorithm that knows lower-level solutions and, therefore, need not query any points far from them. We consider a dual question to such approaches: suppose we have an oracle, which we call y^*-aware, that returns an $O(\epsilon)$-estimate of the lower-level solution, in addition to first-order gradient estimators *locally unbiased* within the $\Theta(\epsilon)$-ball around $y^*(x)$. We study the complexity of finding stationary points with such an y^*-aware oracle: we propose a simple first-order method that converges to an ϵ stationary point using $O(\epsilon^{-6}), O(\epsilon^{-4})$ access to first-order y^*-aware oracles. Our upper bounds also apply to standard unbiased first-order oracles, improving the best-known complexity of first-order methods by $O(\epsilon)$ with minimal assumptions. We then provide the matching $\Omega(\epsilon^{-6})$, $\Omega(\epsilon^{-4})$ lower bounds without and with an additional smoothness assumption on y^*-aware oracles, respectively. Our results imply that any approach that simulates an algorithm with an y^*-aware oracle must suffer the same lower bounds.

[1]Wisconsin Institute for Discovery, Wisconsin, USA [2]Department of Mathematics, University of Seoul, Republic of Korea [3]Center for AI and Natural Sciences, Korea Institute for Advanced Study, Republic of Korea. Correspondence to: Jeongyeol Kwon <jeongyeol.kwon@wisc.edu>, Dohyun Kwon <dh.dohyun.kwon@gmail.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

1. Introduction

Bilevel optimization (Colson et al., 2007) is a fundamental optimization problem that abstracts the core of various critical applications characterized by 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). In essence, Bilevel optimization can be abstractly described as the subsequent minimization problem:

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

where $f, g : \mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \to \mathbb{R}$ are continuously-differentiable functions. The *hyperobjective* $F(x)$ depends on x both directly and indirectly via $y^*(x)$, which is a solution for the lower-level problem of minimizing another function g, which is parametrized by x. Throughout the paper, we assume that the lower-level problem is strongly-convex, *i.e.,* $g(\bar{x}, y)$ is strongly convex in y for all $\bar{x} \in \mathbb{R}^{d_x}$.

Our goal is to find an ϵ-*stationary point* of (P): an x that satisfies $\|\nabla F(x)\| \le \epsilon$. Here the explicit expression of $\nabla F(x)$ can be derived from the implicit function theorem (Krantz & Parks, 2002):

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

Following the standard black-box optimization model (Nemirovskij & Yudin, 1983), we consider the first-order algorithm class that accesses functions through *first-order oracles* that return estimators of first-order derivatives $\hat{\nabla} f(x, y; \zeta), \hat{\nabla} g(x, y; \xi)$ for a given query point (x, y) such that the following holds:

$$\begin{aligned} &\mathbb{E}[\hat{\nabla} f(x, y; \zeta)] = \nabla f(x, y), \\ &\mathbb{E}[\hat{\nabla} g(x, y; \xi)] = \nabla g(x, y), \end{aligned} \tag{2}$$

$$\begin{aligned} &\mathbb{E}[\|\hat{\nabla} f(x, y; \zeta) - \mathbb{E}[\nabla f(x, y; \zeta)]\|^2] \le \sigma_f^2, \\ &\mathbb{E}[\|\hat{\nabla} g(x, y; \xi) - \mathbb{E}[\nabla g(x, y; \xi)]\|^2] \le \sigma_g^2, \end{aligned} \tag{3}$$

$$\mathbb{E}[\|\hat{\nabla} g(x, y^1; \xi) - \hat{\nabla} g(x, y^2; \xi)\|^2] \leq \tilde{l}_{g,1} \|y^1 - y^2\|^2, \quad (4)$$

where $\sigma_f^2, \sigma_g^2 > 0$ are the variance of gradient estimators, $\tilde{l}_{g,1} \in (0, \infty]$ is the stochastic smoothness parameter, and ζ, ξ are independently sampled random variables. The complexity of an algorithm is measured by the worst-case expected number of calls for the first-order oracles until an ϵ-stationary point of (**P**) is found.

$y^*(x)$-Aware Oracles. As we can see in the expression of hypergradients (1), two main challenges distinguish (stochastic) Bilevel optimization from the more standard single-level optimization: at every k^{th} iteration at a query point x^k, we need to estimate **(i)** $y^*(x^k)$, and then estimate **(ii)** $\nabla F(x^k)$, which involves the estimation of Hessian-inverse using y^k. The vast volume of literature has been dedicated to resolving these two issues, starting from double-loop implementations which wait until y^k to be sufficiently close to $y^*(x^k)$ (Ghadimi & Wang, 2018), to recent fully single-loop approaches that updates all variables within $O(1)$-oracle access with incremental improvement of $y^*(x^k)$ estimators (Dagréou et al., 2022; Yang et al., 2023). Notably, all existing approaches require reasonable estimators of $y^*(x)$ for designing an algorithm for Bilevel optimization.

From a practical perspective, while it is standard in Bilevel literature to assume that the inner objective g is strongly convex everywhere, many problems are globally nonconvex, and only *locally strongly convex near* $y^*(x)$. In such cases, a common practice is to obtain a good initialization of $y^*(x)$ via more computationally expensive methods before running faster iterative algorithms to obtain more accurate estimates of lower-level solutions (Jain et al., 2010; Kwon & Caramanis, 2020; Han et al., 2022).

From these motivations, we formulate the following mathematical question: if we can directly obtain a sufficiently good estimator $\hat{y}(x)$ of $y^*(x)$ without additional complexity (*e.g.,* an oracle additionally provides an ϵ-accurate estimate of $y^*(x)$ for a given x), and if we can only query gradients at points (x, y) near $(x, y^*(x))$, what is the fundamental complexity of Bilevel problems? Formally, we consider the following oracle model which we refer to y^*-*aware* oracle:

Definition 1.1 (y^*-Aware Oracle)**.** *An oracle* $\mathrm{O}(\cdot)$ *is* y^**-aware, if there exists* $r \in (0, \infty]$ *such that for every query point* (x, y)*, the following conditions hold.* **(i)** *in addition to stochastic gradients, the oracle also returns* $\hat{y}(x)$ *such that* $\|\hat{y}(x) - y^*(x)\| \leq r/2$*;* **(ii)** *Gradient estimators satisfy* (2)*,* (3) *and* (4) *only if* $\|y - y^*(x)\| \leq r$*; otherwise, the returned gradient estimators can be arbitrary.*

Note that, if we take $r = \infty$ in the above definition, then we recover the usual first-order stochastic gradient oracle that can be queried at any (x, y) with the additional a priori estimator $\hat{y}(x)$ being uninformative. Thus, our oracle model subsumes the models conventionally assumed.

Conceptually, the complexity of any algorithm paired with an y^*-aware oracle can be considered as the lower limit of the problem, unless we can extract significant information from an arbitrary point (x, y) where y is far away from y^*. However, existing approaches view the information obtained at y as meaningful only for the purpose of reaching $y^*(x)$, otherwise containing non-informative biases of size $O(\|y^* - y\|)$. For such approaches, one would expect faster convergence if sufficiently accurate estimates of $y^*(x)$ are provided for free. In this paper, we study lower bounds with such $y^*(x)$-aware oracles, providing a partial answer to the fundamental limits of the problem.

Prior Art. Recent years have witnessed a rapid development of a body of work studying non-asymptotic convergence rates of iterative algorithms to ϵ-stationary points of (**P**) under various assumptions on the stochastic oracles (see Section 1.3 for the detailed overview). A major portion of existing literature assumes access to second-order information of g via Jacobian/Hessian-vector product oracles (which we call *second-order oracles*) as the stationarity measure $\|\nabla F(x)\|$ naturally requires computation of these quantities; see (1). The best-known complexity results with second-order oracles give an $O(\epsilon^{-4})$ upper bound (Ji et al., 2021; Chen et al., 2021), and it can be improved to $O(\epsilon^{-3})$ with variance-reduction when the oracles have additional stochastic smoothness assumptions (Khanduri et al., 2021; Dagréou et al., 2022).

A few recent works have shown that ϵ-stationarity can also be achieved only with first-order oracles (Kwon et al., 2023b; Chen et al., 2023a;b; Lu & Mei, 2023; Yang et al., 2023). With stochastic noises, (Kwon et al., 2023b) proposes an algorithm that finds ϵ-stationary point within $O(\epsilon^{-7})$ access to first-order oracles, and $O(\epsilon^{-5})$ when the gradient estimators are (mean-squared) Lipschitz in expectation. Very recently, (Yang et al., 2023) has shown that an $O(\epsilon^{-3})$ upper-bound is possible only with first-order oracles if we further have an additional *second-order* stochastic smoothness This result matches the best-known rate achieved by second-order baselines under the same condition. We close the remaining gap between first-order methods and second-order methods when we have fewer assumptions on stochastic oracles, *i.e.,* with the standard unbiased, variance-bounded, and possibly (mean-squared) gradient-Lipschitz oracles.

1.1. Overview of Main Results

We first provide high-level ideas on the expected convergence rates of first-order methods. To begin with, suppose for every x, we can directly access $y^*(x)$ and estimators of

Jacobian/Hessian of g without any cost. Then the problem becomes equivalent to finding a stationary point of $F(x)$ with (unbiased) estimators of $\nabla F(x)$. After this reduction to single-level optimization, from the rich literature of nonconvex stochastic optimization (Arjevani et al., 2023), the best achievable complexities are given as $\Theta(\epsilon^{-4})$ and $\Theta(\epsilon^{-3})$ without and with stochastic smoothness, respectively.

When we only have first-order oracles, the complexity can be similarly inferred from the previous bounds and simulating second-order oracles using first-order oracles. Indeed, observe that first-order oracles can simulate Jacobian/Hessian-vector product oracles (with a vector v) through finite differentiation with a precision parameter $\delta > 0$. Namely, $\mathbb{E}\left[\hat{\nabla}^2_{xy} g(x, y^*; \xi)^\top v\right]$ can be estimated as

$$\mathbb{E}\left[\frac{\hat{\nabla}_x g(x, y^* + \delta v; \xi) - \hat{\nabla}_x g(x, y^*; \xi)}{\delta}\right] + O(\delta \|v\|^2).$$

However, without further assumptions, estimators of Jacobian-vector products obtained via the finite differentiation have $O(\delta)$-bias and $O(\delta^{-2})$-amplified variance. Thus, we can use δ at most $O(\epsilon)$ to keep the bias less than ϵ (hence the "oracle-reliability radius" should also satisfy $r = \Omega(\epsilon)$), and require $\Omega(\epsilon^{-2})$ times more oracle access to cancel out the variance amplification to approximately simulate the second-order methods, resulting in total $O(\epsilon^{-6})$ iterations.

When we have stochastic smoothness, *i.e.,* gradient estimators are (mean-squared) Lipschitz as in (4) (with $\tilde{l}_{g,1} < \infty$), then variances of finite-differentiation estimators are still bounded by $O(1)$, and we can obtain the $O(\epsilon^{-4})$ upper bound as if we can access second-order oracles.

Upper Bound. Given the above discussion, we derive upper bounds of $O(\epsilon^{-6})$, $O(\epsilon^{-4})$, without ($\tilde{l}_{g,1} = \infty$) and with ($\tilde{l}_{g,1} < \infty$ in (4)) stochastic smoothness, respectively, with the y^*-aware oracle as long as $r = \Omega(\epsilon)$. In particular, with $r = \infty$, our results improve the best-known upper bounds given in (Kwon et al., 2023b) by the order of $O(\epsilon)$ with minimal assumptions on stochastic oracles. Furthermore, perhaps surprisingly, our result shows that first-order methods are not necessarily worse than second-order methods under nearly the same assumption (as the stochastic smoothness assumption allows us to simulate second-order oracles with no additional cost in ϵ).

Lower Bound. Next, we turn our focus to lower bounds for finding an ϵ-stationary point. For the lower bound, we assume $y^*(x)$-aware oracles with sufficiently good accuracy $r = \Theta(\epsilon)$ and we do not pursue lower bounds for the $r \gg \epsilon$ case in this work (see Conjecture 1). We prove that any black-box algorithms with $y^*(x)$-aware oracles must suffer at least $\Omega(\epsilon^{-6}), \Omega(\epsilon^{-4})$ access to oracles without and with stochastic smoothness, respectively. As an implication, if an algorithm (or analysis), including ours, does not introduce a slowdown by projecting y-coordinates of all query points onto a $\Theta(\epsilon)$-ball around $y^*(x)$, then its iteration complexity cannot be less than the proposed lower bounds.

1.2. Our Approach

Here, we overview our approaches for deriving the claimed upper and lower bounds.

1.2.1. Upper Bound: Penalty Method

In proving the upper bounds, our starting point is to consider an alternative reformulation of (**P**) through the penalty method used in (Kwon et al., 2023b). Specifically, consider a function $\mathcal{L}_\lambda$ with a penalty parameter $\lambda > 0$:

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

In (Kwon et al., 2023b), it was shown that the hyperobjective can be approximated by the following surrogate

$$\mathcal{L}^*_\lambda(x) := \min_y \mathcal{L}_\lambda(x, y) \tag{5}$$

in the sense that $\|\nabla F(x) - \nabla \mathcal{L}^*_\lambda(x)\| \le O(1/\lambda)$, where $\nabla \mathcal{L}^*_\lambda(x)$ is given by

$$\nabla \mathcal{L}^*_\lambda(x) = \nabla_x f(x, y^*_\lambda(x)) + \lambda(\nabla_x g(x, y^*_\lambda(x)) - \nabla_x g(x, y^*(x))), \tag{6}$$

with $y^*_\lambda(x) := \arg\min_y \left(\lambda^{-1} f(x, y) + g(x, y)\right)$. Therefore, we can instead find an ϵ-stationary point of $\mathcal{L}^*_\lambda(x)$, *e.g.,* by running a stochastic gradient descent (SGD) style method on $\mathcal{L}^*_\lambda(x)$ with $\lambda = O(\epsilon^{-1})$.

For now, suppose $y^*_\lambda(x), y^*(x)$ are immediately accessible once x is given. Then, we can construct the unbiased estimator $\hat{\nabla} \mathcal{L}^*_\lambda(x)$ with first-order oracles:

$$\hat{\nabla} \mathcal{L}^*_\lambda(x) = \hat{\nabla} f(x, y^*_\lambda(x); \zeta) + \lambda(\hat{\nabla} g(x, y^*_\lambda(x); \xi^y) - \hat{\nabla} g(x, y^*(x); \xi^z)).$$

However, with the above construction, the variance of $\hat{\nabla} \mathcal{L}^*_\lambda(x)$ is amplified by λ^2 so we need $O(\lambda^2) = O(\epsilon^{-2})$ batch of samples at every iteration to cancel out the amplified variance (similarly to finite-differentation for simulating Jacobian-vector products). Together with the standard sample complexity of SGD (which is $O(\texttt{Var}(\hat{\nabla} \mathcal{L}^*_\lambda)\epsilon^{-4})$), in total $O(\epsilon^{-6})$ oracle access should be sufficient.

When we have stochastic smoothness (4), and if the oracle allows two query points for the same randomness (see Section 2) we can keep the variance controlled by coupling y^*_λ and y^* with the same random variable $\xi^y = \xi^z = \xi$ as

$$\texttt{Var}\left(\hat{\nabla}_x g(x, y^*_\lambda(x); \xi) - \hat{\nabla}_x g(x, y^*(x); \xi)\right) \le \tilde{l}^2_{g,1} \|y^*_\lambda(x) - y^*(x)\|^2.$$

(Kwon et al., 2023b) has shown that $\|y_\lambda^*(x) - y^*(x)\|$ is always bounded by $O(1/\lambda)$, and thus, the variance of $\hat{\nabla}\mathcal{L}_\lambda^*(x)$ is bounded by $O(\sigma^2+\tilde{l}_{g,1}^2)$. Hence, the $O(\epsilon^{-4})$ upper bound can be achieved by running *e.g.,* standard SGD with smooth stochastic oracles.

The key challenge in obtaining upper bounds in both cases is to control the bias. As we cannot directly access $y_\lambda^*(x)$ and $y^*(x)$, we must compute estimators of them (say y and z) by, *e.g.,* running additional gradient steps on y, z in the inner loop before updating x. Then bias comes from the error $\|y - y_\lambda^*(x)\|$ and $\|z - y^*(x)\|$ after the inner-loop iterations, and thus, our analysis requires tight control of overall bias to keep the number of inner-loop iterations optimal.

1.2.2. Lower Bound: Slower Progress

Our lower bound builds on the construction of probabilistic zero-chains by Arjevani et al. (Arjevani et al., 2023) for (single-level) stochastic nonconvex optimization. The key idea of zero-chain is to create a function in a way that, when making a query to an oracle, it discloses only up to one new coordinate. With stochastic oracles, the required number of iterations can be amplified by constructing an oracle that exposes the next new coordinate only with a small probability $p \ll 1$. The result from (Arjevani et al., 2023) constructs such an oracle with $p = O(\epsilon^2)$, achieving the $\Omega(\epsilon^{-4})$ lower bound for single-level optimization.

In Bilevel optimization, obtaining the right lower-bound dependency on ϵ requires care, otherwise, we easily end up with vacuous lower bounds no better than the known lower bounds for single-level optimization. In the noiseless setting, recent work in (Chen et al., 2023a) has achieved $O(\epsilon^{-2})$ upper bound, which is optimal in ϵ (the known lower bound of $\Omega(\epsilon^{-2})$ in deterministic single-level optimization (Carmon et al., 2020) also applies here). This implies that objective functions of the hard-instance cannot have a zero-chain longer than $O(\epsilon^{-2})$. In turn, with variance-bounded stochastic oracles, the $\Omega(\epsilon^{-6})$ lower bound must result from the slowdown of progression in zero-chains by stochastic noises due to the smaller probability of showing the next coordinate.

Our key observation is that we can set the probability p of progression much smaller if the progress in x comes indirectly from g. Specifically, we design $y^*(x)$ such that $y^*(x) = F(x)$ where $F(x)$ is the hard instance given in (Arjevani et al., 2023) (see (17) for the explicit construction), and let f, g such that

$$f(x,y) = y, \quad g(x,y) = (y - F(x))^2. \tag{7}$$

With the oracle model in Definition 1.1 with $r = \Theta(\epsilon)$, the probability of progression can be set $p = O(\epsilon^4)$ over the length $\Omega(\epsilon^{-2})$ zero-chain, and we obtain the desired $\Omega(\epsilon^{-6})$ lower bound for first-order methods with variance-bounded stochastic oracles. The $\Omega(\epsilon^{-4})$ lower bound with additional stochastic smoothness can be shown on the same construction with minor modifications on scaling parameters.

1.3. Related Work

Due to the vast volume of stochastic optimization, we only review the most relevant lines of work.

Upper Bounds for Bilevel Optimization. Bilevel optimization has a long history since its introduction in (Bracken & McGill, 1973). Beyond classical studies on asymptotic landscapes and convergence rates (White & Anandalingam, 1993; Vicente et al., 1994; Colson et al., 2007), initiated by (Ghadimi & Wang, 2018), there has been a surge of interest in developing iterative optimization methods in large-scale problems for solving (**P**). Most convergence analysis have been performed on the standard setting of unconstrained and strongly-convex lower-level optimization with various assumptions on the oracle access to gradients and Hessians (Ghadimi & Wang, 2018; Hong et al., 2020; Ji et al., 2021; Chen et al., 2021; Khanduri et al., 2021; Chen et al., 2022; Sow et al., 2022; Dagréou et al., 2022; Ji et al., 2021; Kwon et al., 2023b; Liu et al., 2021; Sow et al., 2022; Ye et al., 2022; Yang et al., 2023). We mention that there are also a few recent works that study an extension to nonconvex and/or constrained lower-level problems with certain regularity assumptions on the landscape of g around the lower-level solutions, similar to the *local* strong-convexity of the lower-level problems (Kwon et al., 2023a; Chen et al., 2023b; Lu & Mei, 2023; Shen & Chen, 2023; Xiao et al., 2023).

Lower Bounds for Bilevel Optimization. There are relatively few studies on lower bounds for the Bilevel optimization. While the lower bounds for the single-level stochastic optimization (Carmon et al., 2020; Arjevani et al., 2023) also imply the lower bounds of Bilevel optimization, the lower complexity could be much higher than single-level optimization. To our best knowledge, only the work in (Ji & Liang, 2023; Dagréou et al., 2023) has studied the fundamental limits of finding ϵ-stationary points of (**P**). However, (Ji & Liang, 2023) only considers the noiseless setting when the hyperobjective $F(x)$ is convex (note that this is different from assuming f or g is convex). The work in (Dagréou et al., 2023) only considers the case when objective functions are in the form of finite-sum and the second-order oracles are available, and they only provide a lower-bound of single-level finite-sum optimization (Zhou & Gu, 2019).

Stochastic Nonconvex Optimization. There exists a long and rich history of stochastic (single-level) optimization for smooth nonconvex functions. With unbiased first-order gradient oracles, it is well-known that vanilla stochastic gradient descent (SGD) converges to a ϵ-stationary point with at most $O(\epsilon^{-4})$ oracle access (Ghadimi & Lan, 2013),

which is shown to be optimal recently by Arjevani et al. (Arjevani et al., 2023). With smooth stochastic oracles, the rate has been improved to $O(\epsilon^{-3})$ (Fang et al., 2018; Cutkosky & Orabona, 2019), which is also shown to be optimal in (Arjevani et al., 2023). These results are the counterparts to convergence rates of second-order methods for (**P**) with unbiased and smooth Jacobian/Hessian stochastic oracles (Chen et al., 2021; Khanduri et al., 2021). The same rate $O(\epsilon^{-3})$ can also be achieved only with first-order oracles, as shown in (Yang et al., 2023) with assuming higher-order stochastic smoothness.

2. Preliminaries

Throughout the paper, we specify the assumptions on the smoothness of objective functions. The first one is the global smoothness properties:

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

1. *f, g are continuously-differentiable and $l_{f,1}, l_{g,1}$-smooth respectively, jointly in (x, y) over $\mathbb{R}^{d_x \times d_y}$.*
2. *For every $(x, y) \in \mathbb{R}^{d_x \times d_y}$, $\|\nabla_y f(x, y)\| \leq l_{f,0}$.*
3. *For every $\bar{x} \in \mathbb{R}^{d_x}$, $g(\bar{x}, \cdot)$ is μ_g-strongly convex in y.*

In addition to the smoothness of individual objective functions, in Bilevel optimization, we also desire the hyperobjective to be smooth. This can be indirectly assumed by the Lipschitzness of derivatives of g:

Assumption 2. *$\nabla^2_{xy} g, \nabla^2_{yy} g$ are well-defined and $l_{g,2}$-Lipschitz jointly in (x, y) for all $(x, y) \in \mathbb{R}^{d_x \times d_y}$.*

Oracle Classes. We assume that we can access first-order information of objective functions only through stochastic oracles O that are $y^*(x)$-aware for some radius $r = \Omega(\epsilon)$ (see Definition 1.1). Note that if we take $r = \infty$, these oracles become the usual first-order stochastic gradient oracles that can be queried at any (x, y) and $\hat{y}(x)$ is uninformative. Thus our assumption on the oracles includes the conventional ones. We consider that stochastic oracles allow N-simultaneous query: the oracle takes N-simultaneous query points $(\boldsymbol{x}, \boldsymbol{y}) = \{(x^n, y^n)\}_{n=1}^N$ with $N \geq 2$, and the oracle returns $\{(\hat{\nabla} f(x^n, y^n; \zeta), \hat{\nabla} g(x^n, y^n; \xi), \hat{y}^n(x^n))\}_{n=1}^N$, where $\hat{y}^n(x^n)$ is an estimator of $y^*(x^n)$. We denote the oracle class as $\mathcal{O}(N, \sigma^2, \tilde{l}_{g,1}, r)$. We note that $\tilde{l}_{g,1} \geq l_{g,1}$ must always hold.

Additional Notation Throughout the paper, $\|\cdot\|$ denotes the Euclidean norm for vectors and operator norm for matrices, and $\texttt{Var}(\cdot)$ denotes the variance of a random vector. We often denote $a \lesssim b$ when the inequality holds up to some absolute constant. $\mathbb{B}(x, r)$ is a Euclidean ball of radius $r > 0$ around a point x.

3. Upper Bounds

In this section, we prove the $O(\epsilon^{-6})$ and $O(\epsilon^{-4})$ upper bounds without and with stochastic smoothness (4), respectively. We mainly focus on finding a stationary point of the surrogate hyperobjective $\mathcal{L}^*_\lambda(x)$ defined in (5). Thanks to $\|F(x) - \mathcal{L}^*_\lambda(x)\| = O(\lambda^{-1})$ given in Lemma 3.1 from (Kwon et al., 2023b), we get ϵ-stationarity of $F(x)$ with a choice of $\lambda = O(\epsilon^{-1})$. We mention here that our upper bounds do not depend on the 'reliability radius' of the oracle r as long as $r \geq \frac{6 l_{f,0}}{\mu_g \lambda}$, hence the readers may assume the standard oracle model with $r = \infty$.

3.1. $O(\epsilon^{-6})$ Upper Bound

We show that double-loop $\texttt{F}^2\texttt{SA}$ introduced in (Kwon et al., 2023b) can be improved by $O(\epsilon)$ from $O(\epsilon^{-7})$ to $O(\epsilon^{-6})$ with suitable choice of outer-loop batch-size M and inner-loop iterations T.

As in (Kwon et al., 2023b), the main challenge comes from handling the bias raised every iteration by using approximations of $y^*_\lambda(x)$ and $y^*(x)$. Specifically, let y^{k+1} and z^{k+1} be the estimates of $y^*_\lambda(x^k)$ and $y^*(x^k)$ at the k^{th} iteration, respectively. Using these estimates we may approximate $\nabla \mathcal{L}^*_\lambda(x^k)$ by (see (6))

$$G_k := \nabla_x f(x^k, y^{k+1}) + \lambda(\nabla_x g(x^k, y^{k+1}) - \nabla_x g(x^k, z^{k+1})). \quad (8)$$

Comparing $\nabla \mathcal{L}^*_\lambda(x^k)$ and G_k, it is natural to consider

$$\lambda(\|y^{k+1} - y^*_\lambda(x^k)\| + \|z^{k+1} - y^*(x^k)\|) \quad (9)$$

as the order of bias of approximating $\nabla \mathcal{L}^*_\lambda(x^k)$ by G_k. For this bias to be less than ϵ, we need $O(\epsilon/\lambda) = O(\epsilon^2)$ accuracy of y^{k+1} and z^{k+1}, which in turn requesting $T \asymp \epsilon^{-4}$ inner-loop iterations (with SGD on strongly-convex functions) before updating x^k with an unbiased estimate of G_k. A similar idea has been used in (Chen et al., 2023a) where $\tilde{O}(\epsilon^{-3})$ bound in the deterministic setting shown in (Kwon et al., 2023b) has been improved to $\tilde{O}(\epsilon^{-2})$ with a choice of $T \asymp \log(\lambda)$.

In Algorithm 1, we use the following notations:

$$\begin{aligned}
h_y^{k,t} &= \lambda^{-1} \hat{\nabla}_y f(x^k, y^{k,t}; \zeta^{k,t}) + \hat{\nabla}_y g(x^k, y^{k,t}; \xi^{k,t}), \\
h_z^{k,t} &= \hat{\nabla}_y g(x^k, z^{k,t}; \xi^{k,t}), \\
h_x^{k,m} &= \hat{\nabla}_x f(x^k, y^{k+1}; \zeta^x_{k,m}) + \\
&\quad \lambda(\hat{\nabla}_x g(x^k, y^{k+1}; \xi^{k,m}_x) - \hat{\nabla}_x g(x^k, z^{k+1}; \xi^{k,m}_x)).
\end{aligned}$$

In the outer loop, we have $x^{k+1} = x^k - \alpha \hat{G}_k$ where

$$\hat{G}_k := M^{-1} \textstyle\sum_{m=1}^M h_x^{k,m}. \quad (10)$$

Under (2) and (3), $\texttt{Var}(\hat{G}_k) := \mathbb{E}[\|G_k - \hat{G}_k\|^2)] = O(\lambda^2/M)$ as shown in Lemma A.2. Thus, the variance can

Algorithm 1 Penalty Methods with Lower-Level Coupling

Input: total outer-loop iterations: n, batch size: M, step sizes: $\alpha, \{\gamma_t\}_t$, penalty parameter: λ, Oracle reliability radius: r, Trust-region radius for $(y-z)$: r_λ, initializations: x_0, y_0, z_0

for $k = 0...n-1$ **do**
  # Initialize Lower-Level Iteration Variables
  Get $\hat{y}(x^k)$ such that $\|\hat{y}(x^k) - y^*(x^k)\| \leq \frac{r}{2}$.
  $y^{k,0}, z^{k,0} \leftarrow \texttt{Proj}(\hat{y}(x^k), y^k, z^k, r, r_\lambda)$
  # Lower-Level Iteration with Coupling y, z
  for $t = 0, ..., T-1$ **do**
    $\bar{y}^{k,t} \leftarrow y^{k,t} - \gamma_t h_y^{k,t}$, $\bar{z}^{k,t} \leftarrow z^{k,t} - \gamma_t h_z^{k,t}$
    $y^{k,t+1}, z^{k,t+1} \leftarrow \texttt{Proj}(\hat{y}(x^k), \bar{y}^{k,t}, \bar{z}^{k,t}, r, r_\lambda)$
  end for
  $y^{k+1}, z^{k+1} \leftarrow y^{k,T}, z^{k,T}$
  # Upper-Level Iteration with Coupling y, z
  $x^{k+1} \leftarrow x^k - \frac{\alpha}{M} \sum_{m=1}^{M} h_x^{k,m}$
end for

Algorithm 2 `Proj`

Input: y^*-estimator: $\hat{y}$, lower-level variables: y, z, radius parameters: r, r_λ
$y' \leftarrow \Pi_{\mathbb{B}(\hat{y}, \frac{2r}{3})}\{y\}$, $\Delta_y = y' - y$
(If $\tilde{l}_{g,1} = \infty$**):** $z' \leftarrow \Pi_{\mathbb{B}(\hat{y}, \frac{r}{2})}\{z\}$
(Else): $z' \leftarrow \Pi_{\mathbb{B}(y', r_\lambda)}\{z + \Delta_y\}$
Return y', z'

be handled with $M \asymp \epsilon^{-4}$, which suffices to get the $O(\epsilon^{-6})$ upper bound. The proof can be found in Appendix A.2.

Note that Algorithm 1 is a first-order method that has per-sample computational cost bounded by that for computing the gradient estimators $\hat{\nabla}_x f$, $\hat{\nabla}_y f$, $\hat{\nabla}_x g$, and $\hat{\nabla}_y g$.

Theorem 3.1. *Suppose Assumptions 1 and 2 hold and let* $\lambda = \max\left(\frac{\lambda_0}{\epsilon}, \frac{6l_{f,0}}{\mu_g r}\right) \asymp \epsilon^{-1}$, $r_\lambda = \frac{l_{f,0}}{\mu_g \lambda}$ *where* $\lambda_0 := \frac{4l_{f,0}l_{g,1}}{\mu_g^2}\left(l_{f,1} + \frac{2l_{f,0}l_{g,2}}{\mu_g}\right)$. *Under (2) and (3), Algorithm 1 with* $\alpha \ll \frac{1}{l_{g,1}}$, $T \asymp \epsilon^{-4}$, *and* $\gamma_t = \left(\frac{2}{\mu_g} + \frac{\lambda}{l_{f,1}+\lambda l_{g,1}}\right)\frac{1}{1+t}$ *finds an ϵ-stationary point of (P) within $O(\epsilon^{-6})$ oracle calls.*

See (58) for an explicit expression of the bound with implied constants in the $O(\epsilon^{-6})$ bound in the above theorem. It is worth mentioning that compared to the $O(\epsilon^{-7})$ results in (Kwon et al., 2023b), we improve the upper bound to $O(\epsilon^{-6})$, and do not require additional assumption on the Hessian-Lipschitzness of f.

3.2. $O(\epsilon^{-4})$ Upper Bound

If we aim to get $O(\epsilon^{-4})$ upper bound with first-order smooth oracles, $T \asymp \epsilon^{-4}$ inner-loop iterations are too many to achieve the goal. When we have stochastic smoothness (*i.e.,* $\tilde{l}_{g,1} < \infty$), we show that choosing $T \asymp \epsilon^{-2}$ is sufficient to obtain the desired convergence rate. The proof can be found in Appendix A.3.

Theorem 3.2. *Suppose that Assumptions 1-2, (2), (3), and (4) hold. Let the algorithm parameters satisfy the following:*

$$\lambda \geq \max\left(\frac{\lambda_0}{\epsilon}, \frac{6l_{f,0}}{\mu_g r}\right), \quad r_\lambda = \frac{l_{f,0}}{\mu_g \lambda}, \quad \alpha \ll 1. \tag{11}$$

Then, Algorithm 1 with $\lambda \asymp \epsilon^{-1}$, $\gamma \asymp \epsilon^2$, $T \asymp \epsilon^{-2}$, $M \asymp \epsilon^{-2}$, *and* $n \asymp \epsilon^{-2}$ *finds an ϵ-stationary point of* **(P)** *within* $O(\epsilon^{-4})$ *oracle calls.*

To our best knowledge, the best known results under the similar setting achieve the $O(\epsilon^{-5})$ upper bound with the stochastic smoothness of both objective functions f and g jointly in (x, y) (Kwon et al., 2023b). We show that the upper bound can be improved $O(\epsilon^{-4})$ with the only required additional assumption being the stochastic smoothness in y.

Our new observation here is that the estimation for the bias can be tightened by using

$$v^k := y^k - z^k \text{ and } v^*(x) := y_\lambda^*(x) - y^*(x).$$

The following lemma is the key to tighten the bias in terms of y and v:

Lemma 3.3. *Suppose that Assumptions 1-2, (2), (3), (4), and* $r_\lambda = \frac{l_{f,0}}{\mu_g \lambda}$ *hold. Then,*

$$\|\nabla \mathcal{L}_\lambda^*(x^k) - G_k\| \leq l_y \|y^{k+1} - y_\lambda^*(x^k)\| + \lambda l_{g,1}\|v^{k+1} - v^*(x^k)\| + \frac{l_{f,0}l_y}{\mu_g \lambda},$$

where $l_y := l_{f,1} + \frac{l_{g,2}l_{f,0}}{\mu_g}$.

The proof is given in Appendix A. In comparison to (9), the term $\|y^{k+1} - y_\lambda^*(x^k)\|$ on the upper bound does not depend on λ. As a consequence, $T = O(\epsilon^{-2})$ (with $\gamma = O(\epsilon^2)$ is enough to obtain $O(\epsilon)$ accuracy of y^{k+1} and z^{k+1}.

On the other hand, we observe that the variance of stochastic noises in updating $v^k = y^k - z^k$ can be bounded by $O(\|y^k - z^k\|^2)$, which can be bounded by $O(1/\lambda^2)$ (instead of $O(\sigma^2) = O(1)$) with a forced projection step. We observe that the projection makes the distance between v^k and v^* smaller. Let $\bar{v}^{k,t} := \bar{y}^{k,t} - \bar{z}^{k,t}$.

Proposition 3.4. *Suppose that Assumptions 1-2, (2), (3), and (4) hold. Then, for all k, t with* $r_\lambda = \frac{l_{f,0}}{\lambda \mu_g}$:

$$\|v^{k,t+1} - v_k^*\| \leq \|\bar{v}^{k,t} - v_k^*\|. \tag{12}$$

Hence, it suffices to have $O(\epsilon^{-2})$ inner-loop iterations to make the bias in v^k less than $O(\epsilon/\lambda)$, giving $O(\epsilon^{-4})$ upper bound with first-order smooth oracles.

Proposition 3.5. *Under Assumptions 1-2, (2), (3), (4), (11), and $\gamma T > 1/\mu_g$, the iterates of Algorithm 1 satisfy:*

$$\frac{\alpha}{n}\sum_{k=0}^{n-1}\mathbb{E}[\|\nabla\mathcal{L}_\lambda^*(x^k)\|^2 \\ \leq O(n^{-1}) + \frac{\alpha}{n}\sum_{k=0}^{n-1}\texttt{Var}(\hat{G}_k) + O(\lambda^{-2}) + O(\gamma). \quad (13)$$

Lastly, we obtain the tighter estimation of $\texttt{Var}(\hat{G}_k) = O(M^{-1})$ shown in Lemma A.2. Choosing $T \asymp \epsilon^{-2}$, $M \asymp \epsilon^{-2}$, and $n \asymp \epsilon^{-2}$, we conclude the theorem.

4. Lower Bounds

In our lower bound construction, we mainly focus on the iteration complexity dependence on ϵ, and we consider a subset of the function class in which smoothness parameters are absolute constants:

$$\mathcal{F}(1) := \{(f,g) \text{ satisfy Assumptions 1, 2 } | \\ l_{f,0}, l_{f,1}, \mu_g, l_{g,1}, l_{g,2} = O(1)\}. \quad (14)$$

Throughout the section, we consider the smoothness parameters as $O(1)$ quantities and assume that $\epsilon > 0$ is sufficiently small such that ϵ^{-1} dominates any polynomial factors of smoothness parameters. Before we proceed, we describe additional preliminaries for the lower bound.

Algorithm Class. We consider algorithms that access an unknown pair of functions (f,g) via a first-order stochastic oracle O. Following the black-box optimization model (Nemirovskij & Yudin, 1983), we consider an algorithm $\texttt{A} \in \mathcal{A}$ *paired* with $\mathrm{O} \in \mathcal{O}(N, \sigma^2, \tilde{l}_{g,1}, r_\epsilon)$ where $\tilde{l}_{g,1} \geq 100$ and $r_\epsilon := 100\epsilon$. At any iteration step $t \in \mathbb{N}$, $\texttt{A}$ takes the first $(t-1)$ oracle responses and generates the next (randomized) query points:

$$(\boldsymbol{x}^t, \boldsymbol{y}^t) = \texttt{A}\big(\xi_{\mathrm{A}}, (\boldsymbol{x}^0, \boldsymbol{y}^0), \mathrm{O}(\boldsymbol{x}^0, \boldsymbol{y}^0; \zeta^0, \xi^0), ..., \\ (\boldsymbol{x}^{t-1}, \boldsymbol{y}^{t-1}), \mathrm{O}(\boldsymbol{x}^{t-1}, \boldsymbol{y}^{t-1}; \zeta^{t-1}, \xi^{t-1})\big),$$

where ξ_{A} is a random seed generated at the beginning of the optimization procedure. For simplicity, we assume that $\sigma_f = \sigma_g = \sigma$ in this section.

Zero-Respecting Algorithms. An important subclass of randomized algorithm class $\mathcal{A}$ is the zero-respecting algorithm class $\mathcal{A}_{\text{zr}}$ which preserves the support of gradients with respect to x at the queried points:

Definition 4.1. *Suppose $\hat{\nabla}_x f = 0$. An algorithm $\texttt{A}$ is zero-respecting if for all $t \geq 0$ and $n \in [N]$ generated by $\texttt{A}$,*

$$\text{supp}(x^{t,n}) \subseteq \bigcup_{t'<t, n' \in [N]} \text{supp}\left(\hat{\nabla}_x g(x^{t',n'}, y^{t',n'}; \xi^{t'})\right).$$

Zero-respecting algorithm class plays a critical role in dimension-free lower-bound arguments, as we can construct a family of hard instances for all randomized black-box algorithms from one hard example (see Section 4.3).

Probabilistic Zero-Chains. At the core of the lower-bound construction is the construction of a hard example for all zero-respecting algorithms. To study the iteration complexity of a general randomized algorithm class, (Arjevani et al., 2023) introduced the notion of *progress* defined as the following:

$$\texttt{prog}_\alpha(x) := \max\{i \geq 0 | |x_i| > \alpha\}, \ (x_0 \equiv 1). \quad (15)$$

Here, $\alpha \in [0, 1)$ controls the effective threshold of values that would be considered as zero. We also adopt the notion of probabilistic zero-chains from (Arjevani et al., 2023):

Definition 4.2. *A function $g(x,y)$, paired with gradient estimators $\hat{\nabla} g(x,y;\xi)$ with an independent random variable ξ, is a probability-p zero-chain in x if there exists $\alpha > 0$ such that for all (x,y):*

$$\mathbb{P}(\texttt{prog}_0(\hat{\nabla}_x g(x,y;\xi)) > \texttt{prog}_\alpha(x) + 1) = 0, \\ \mathbb{P}(\texttt{prog}_0(\hat{\nabla}_x g(x,y;\xi)) = \texttt{prog}_\alpha(x) + 1) \leq p. \quad (16)$$

4.1. Hard Instance

We start with the base construction from (Carmon et al., 2020) for single-level nonconvex optimization:

$$F(x) = \epsilon^2 \textstyle\sum_{i=1}^{d_x} f_i(x), \\ f_i(x) := \Psi_\epsilon(x_{i-1})\Phi_\epsilon(x_i) - \Psi_\epsilon(-x_{i-1})\Phi_\epsilon(-x_i), \quad (17)$$

where $d_x = \lfloor \epsilon^{-2} \rfloor$, $x_0 \equiv \epsilon$, and $\Psi(x), \Phi(x)$ are a scalar function defined as the following:

$$\Psi(t) := \begin{cases} 0, & \text{if } t \leq 1/2 \\ \exp\left(1 - \frac{1}{(2t-1)^2}\right), & \text{otherwise} \end{cases}, \\ \Phi(t) := \sqrt{e}\int_{-\infty}^{t} e^{-\tau^2/2} d\tau, \quad (18)$$

$\Psi_s(t), \Phi_s(t)$ are short-hands of $\Psi(t/s), \Phi(t/s)$. The above construction satisfies the required smoothness and initial value-gap $F(0) - \inf_x F(x) = O(1)$. We set the hyper-objective to be defined as (17), but the progression of an algorithm is slowed down when the $\nabla F(x)$ is only indirectly accessed via $\hat{\nabla}_x g(x,y;\xi)$ for $\|y - y^*(x)\| \leq r_\epsilon$.

To set up the Bilevel objectives, we let $f(x,y) = y$ and $g(x,y) = (y - F(x))^2$ as in (7). It is straight-forward to see that $y^*(x) = F(x)$ and $f(x, y^*(x)) = F(x)$. Next, we design the expectation of gradient estimators that the oracle returns. To simplify discussion, let $\hat{\nabla} f$ are deterministic,

i.e., $\hat{\nabla} f(x,y;\zeta) = \nabla f(x,y)$ with probability 1. We define the stochastic gradients to be

$$\mathbb{E}[\hat{\nabla}_x g(x,y;\xi)] = \nabla_x \left(r_\epsilon^2 \cdot \phi\left(\frac{y - F(x)}{r_\epsilon} \right)^2 \right), \quad (19)$$

where $\phi(t)$ is a smooth clipping function:

$$\phi(t) := \begin{cases} t, & \text{if } |t| \le 1/2, \\ t - \frac{1}{e}\int_{1/2}^{t} \Psi(\tau)d\tau, & \text{else if } t > 1/2 \\ t + \frac{1}{e}\int_{1/2}^{-t} \Psi(\tau)d\tau, & \text{else} \end{cases}. \quad (20)$$

With the above construction, gradient estimators are unbiased for $|y - y^*(x)| \le r_\epsilon$. Let $\nabla_x g_b(x,y)$ be the short-hand of (19).

Next, we design the coordinate-wise gradient estimators of g w.r.t x as the following. Let $\xi \sim \texttt{Ber}(p)$, and

$$\hat{\nabla}_{x_i} g(x,y;\xi) = \nabla_{x_i} g_b(x,y) \cdot (1 + h_i(x)(\xi/p - 1)),$$

where $h_i(x)$ is a smooth version of the indicator function $\mathbb{1}_{\{i > \texttt{prog}_{\epsilon/4}(x)\}}$ (see details in Appendix, Lemma B.4). Finally, we design stochastic gradients w.r.t y and $\hat{y}$:

$$\hat{\nabla}_y g(x,y;\xi) = 2\left(y - \epsilon^2 \cdot \sum_{i=1}^{d_x} f_i(x)\left(1 + h_i(x)(\xi/p - 1)\right) \right),$$
$$\hat{y}(x) = \epsilon^2 \textstyle\sum_{i=1}^{\texttt{prog}_{\epsilon/2}(x)} f_i(x).$$

The next step is to show the lower complexity bounds of "activating" the last coordinate of x for zero-respecting algorithms. We note that additional noises in $\hat{\nabla}_y g$ and $\hat{y}$ are only required when extending to randomized algorithms, and $\hat{\nabla}_y g$, $\hat{y}$ satisfy the following:

Lemma 4.3. *There exists some absolute constant $c > 0$ such that for $x \in \mathbb{R}^{d_x}, y \in \mathbb{R}$:*

$$\texttt{Var}\left(\hat{\nabla}_y g(x,y;\xi) \right) \lesssim \frac{\epsilon^4}{p} \lesssim \sigma^2,$$
$$\mathbb{E}[\|\hat{\nabla}^2_{yy} g(x,y;\xi)\|^2] \le 2.$$

Furthermore, it holds that $\mathbb{E}[\hat{\nabla}_y g(x,y;\xi)] = \nabla_y g(x,y)$ and $|\hat{y}(x) - F(x)| = O(\epsilon^2) \ll r_\epsilon$ for all $x \in \mathbb{R}^{d_x}$ and $y \in \mathbb{R}$.

4.2. Lower Bounds for Zero-Respecting Algorithms

We start by showing that the construction (f, g) with the oracle $\hat{\nabla} g(x,y;\xi)$ forms a probabilistic zero-chain. The intuition behind the probabilistic zero-chain argument is to amplify the required number of iterations to progress toward the next coordinate from 1 to $O(1/p)$, which gives the overall lower bound of $\Omega(d_x/p)$. Thus, as we set p smaller, the more iterations all zero-respecting algorithms would need to reach the last coordinate of x. The minimum possible value of p is decided by (3) on bounded variances:

$$\texttt{Var}\left(\hat{\nabla}_x g(x,y;\xi) \right) \lesssim \frac{\|\nabla_x g_b(x,y)\|^2_\infty}{p} \lesssim \sigma^2,$$

and the smoothness condition (4):

$$\mathbb{E}[\|\hat{\nabla}^2_{xy} g(x,y;\xi)\|^2] \lesssim \|\nabla F(x)\|^2 + \frac{\|\nabla F(x)\|^2_\infty}{p} \lesssim \tilde{l}^2_{g,1}.$$

The crucial feature of our construction is the upper bound on $\|\nabla_x g_b(x,y)\|_\infty$, which is much smaller than the single-level optimization case:

Lemma 4.4. *There exists some absolute constant $c > 0$ such that for all $x \in \mathbb{R}^{d_x}, y \in \mathbb{R}$,*

$$\|\nabla_x g_b(x,y)\|_\infty \le c r_\epsilon \epsilon = O(\epsilon^2),$$
$$\|\nabla F(x)\|_\infty \le c\epsilon = O(\epsilon).$$

Therefore, (16) holds with $p = \max\left(\epsilon^4/\sigma^2, \epsilon^2/\tilde{l}^2_{g,1}\right)$ and $\alpha = \epsilon/4$.

Note that the scenario only with bounded-variance can be explained by setting $\tilde{l}_{g,1} = \infty$. Thus, with $d_x = O(\epsilon^{-2})$, we obtain the lower bound of $\Omega(\epsilon^{-6}), \Omega(\epsilon^{-4})$ for all zero-respecting algorithms without and with the stochastic smoothness assumption, respectively.

4.3. Lower Bounds for Randomized Algorithms

To convert the lower bound for all zero-respecting algorithms to randomized algorithm classes, our construction can adopt the "randomized coordinate-embedding" argument from (Carmon et al., 2020). We define a class of hard instances for randomized algorithms:

$$f_U(x,y) := y + \frac{1}{10}\|x\|^2,$$
$$g_U(x,y) := (y - F(U^\top \rho(x))^2, \quad (21)$$

where U is a random orthonormal matrix sampled from $\texttt{Ortho}(d, d_x) := \{U \in \mathbb{R}^{d \times d_x} | U^\top U = I\}$ with $d \gg d_x$ (with a slight abuse in notation, the true dimension of x is d instead of d_x), and

$$\rho(x) := x/\sqrt{1 + \|x\|^2/R^2}, \quad (22)$$

where $R = 250\epsilon\sqrt{d_x}$. Then we consider the family of hard instances as the following:

$$\mathcal{F}_{\texttt{hard}} := \{(f_U, g_U) | U \in \texttt{Ortho}(d, d_x), \\ f_U, g_U \text{ defined in (21)}\}. \quad (23)$$

Intuition for the above construction is that for every randomized algorithm $\texttt{A} \in \mathcal{A}$, if we select the low-dimensional

embedding U uniformly randomly from a unit sphere, the sequence of query points generated by A behaves as if it is generated by a zero-respecting algorithm in the embedding space. The corresponding stochastic gradient estimator is now given by

$$\begin{aligned}\hat{\nabla}_x g_U(x,y;\xi) &= J(x)^\top U \hat{\nabla}_x g(U^\top \rho(x), y; \xi),\\ \hat{\nabla}_y g_U(x,y;\xi) &= \hat{\nabla}_y g(U^\top \rho(x), y; \xi),\\ \hat{y}_U(x) &= \hat{y}(U^\top \rho(x)).\end{aligned} \tag{24}$$

where $J(x) \in \mathbb{R}^{d\times d}$ is a Jacobian of $\rho(x)$. As the remaining steps follow the same arguments in (Arjevani et al., 2023), we conclude our lower bound for the randomized algorithms with the oracle model in Definition 1.1:

Theorem 4.5. *Let K be the minimax oracle complexity for the function class $\mathcal{F}(1)$ and oracle class $\mathcal{O}(N, \sigma^2, \tilde{l}_{g,1}^2, r_\epsilon)$ where $\tilde{l}_{g,1} \geq 100$ and $r_\epsilon = 100\epsilon$:*

$$K := \inf_{\mathcal{A}} \sup_{\substack{\mathcal{F}(1)\\ \mathcal{O}(N,\sigma^2,\tilde{l}_{g,1},r_\epsilon)}} \inf\left\{K \in \mathbb{N} | \mathbb{E}[\|\nabla F(x^{K,1})\|] \leq \epsilon\right\}.$$

Then the minimax oracle complexity must be at least:

$$K \gtrsim \min\left(\frac{\sigma^2}{\epsilon^6}, \frac{\tilde{l}_{g,1}^2}{\epsilon^4}\right).$$

Lastly, we leave the following conjecture for the lower bound for future investigation.

Conjecture 1 (Lower bound with globally unbiased oracles)**.** *The current $\Omega(\epsilon^{-6})$ lower bound only holds with $r_\epsilon = \Theta(\epsilon)$. For fully general results, the main challenge is to construct a hard-instance where $g(x,y)$ is strongly-convex in y for all $(x,y) \in \mathbb{R}^{d_x \times d_y}$, while at the same time $\|\nabla_x g(x,y)\|_\infty$ is globally bounded by the same order of $\|\nabla_x g(x, y^*(x))\|_\infty$ for all $y \in \mathbb{R}^{d_y}$. We conjecture that lower bounds remain the same with standard stochastic oracles (i.e., with $r = \infty$).*

5. Concluding Remarks

In this work, we have studied the complexity of first-order methods for Bilevel optimization with y^*-aware oracles. When the oracle gives $r_\epsilon = \Theta(\epsilon)$ estimates of $y^*(x)$ along with $O(r_\epsilon)$-locally reliable gradients, we establish $O(\epsilon^{-6}), O(\epsilon^{-4})$ upper bounds without and with stochastic smoothness, along with the matching $\Omega(\epsilon^{-6}), \Omega(\epsilon^{-4})$ lower bounds. Our upper bound analysis also holds with standard oracles (*i.e.,* with $r = \infty$), improving the best-known results given in (Kwon et al., 2023b). The remaining complexity questions include Conjecture 1, and tight bounds in terms of other smoothness parameters. We conjecture that obtaining such results would require theoretical breakthroughs beyond existing techniques, thereby leaving them as unresolved challenges.

Acknowledgements

JK is partially funded by AFOSR/AFRL grant no. FA9550-18-1-0166. DK is partially supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00252516) and the POSCO Science Fellowship of POSCO TJ Park Foundation. HL is partially supported by NSF grants DMS-2206296 and DMS-2010035.

Impact statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Arjevani, Y., Carmon, Y., Duchi, J. C., Foster, D. J., Srebro, N., and Woodworth, B. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199 (1-2):165–214, 2023.

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

Bottou, L. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010: 19th International Conference on Computational StatisticsParis France, August 22-27, 2010 Keynote, Invited and Contributed Papers*, pp. 177–186. Springer, 2010.

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

Carmon, Y., Duchi, J. C., Hinder, O., and Sidford, A. Lower bounds for finding stationary points i. *Mathematical Programming*, 184(1-2):71–120, 2020.

Chen, L., Ma, Y., and Zhang, J. Near-optimal fully first-order algorithms for finding stationary points in bilevel optimization. *arXiv preprint arXiv:2306.14853*, 2023a.

Chen, L., Xu, J., and Zhang, J. On bilevel optimization without lower-level strong convexity. *arXiv preprint arXiv:2301.00712*, 2023b.

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

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

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

Cutkosky, A. and Orabona, F. Momentum-based variance reduction in non-convex sgd. *Advances in neural information processing systems*, 32, 2019.

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

Dagréou, M., Moreau, T., Vaiter, S., and Ablin, P. A lower bound and a near-optimal algorithm for bilevel empirical risk minimization. *arXiv e-prints*, pp. arXiv–2302, 2023.

Fang, C., Li, C. J., Lin, Z., and Zhang, T. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. *Advances in neural information processing systems*, 31, 2018.

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

Garrigos, G. and Gower, R. M. Handbook of convergence theorems for (stochastic) gradient methods. *arXiv preprint arXiv:2301.11235*, 2023.

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

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

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

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

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

Han, R., Willett, R., and Zhang, A. R. An optimal statistical and computational framework for generalized tensor estimation. *The Annals of Statistics*, 50(1):1–29, 2022.

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

Jain, P., Meka, R., and Dhillon, I. Guaranteed rank minimization via singular value projection. *Advances in Neural Information Processing Systems*, 23, 2010.

Ji, K. and Liang, Y. Lower bounds and accelerated algorithms for bilevel optimization. *J. Mach. Learn. Res.*, 24: 22–1, 2023.

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

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

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

Krantz, S. G. and Parks, H. R. *The implicit function theorem: history, theory, and applications*. Springer Science & Business Media, 2002.

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

Kwon, J. and Caramanis, C. The EM algorithm gives sample-optimality for learning mixtures of well-separated gaussians. In *Conference on Learning Theory*, pp. 2425–2487, 2020.

Kwon, J., Kwon, D., Wright, S., and Nowak, R. On penalty methods for nonconvex bilevel optimization and first-order stochastic approximation. *arXiv preprint arXiv:2309.01753*, 2023a.

Kwon, J., Kwon, D., Wright, S., and Nowak, R. D. A fully first-order method for stochastic bilevel optimization. In *International Conference on Machine Learning*, pp. 18083–18113. PMLR, 2023b.

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

Lu, Z. and Mei, S. First-order penalty methods for bilevel optimization. *arXiv preprint arXiv:2301.01716*, 2023.

Nemirovskij, A. S. and Yudin, D. B. Problem complexity and method efficiency in optimization. 1983.

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

Shen, H. and Chen, T. On penalty-based bilevel gradient descent method. *arXiv preprint arXiv:2302.05185*, 2023.

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

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

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

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

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

Xiao, Q., Shen, H., Yin, W., and Chen, T. Alternating projected sgd for equality-constrained bilevel optimization. In *International Conference on Artificial Intelligence and Statistics*, pp. 987–1023. PMLR, 2023.

Yang, Y., Xiao, P., and Ji, K. Achieving $\mathcal{O}(\epsilon^{-1.5})$ complexity in hessian/jacobian-free stochastic bilevel optimization. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

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

Zhou, D. and Gu, Q. Lower bounds for smooth nonconvex finite-sum optimization. In *International Conference on Machine Learning*, pp. 7574–7583. PMLR, 2019.

Supplementary Materials for "On the Complexity of First-Order Methods in Stochastic Bilevel Optimization"

A. First-order analysis for the upper bound

In this section, we establish the upper bounds on the convergence rate claimed in Theorems 3.1 and 3.2. We use the following notations for filtration: for $k = 0, 1, \cdots, n-1$ and $t = 0, 1, \cdots, T-1$

- $\mathcal{F}_k$: the σ-algebra generated by $x^i, y^i, z^i, y^{i,t}, z^{i,t}$ for all $i = 0, 1, \cdots, k-1$ and $t = 0, 1, \cdots, T-1$.
- $\mathcal{F}_{k,s}$: the σ-algebra generated by $\mathcal{F}_k$ and $y^{k,t}, z^{k,t}$ for all $t = 0, 1, \cdots, s-1$.
- $\mathcal{F}'_k := \mathcal{F}_{k,T}$.

A.1. Preliminaries

Following (Kwon et al., 2023b), we reformulate (**P**) as the following constrained single-level problem:

$$\min_{x \in \mathcal{X}} \quad F(x) := f(x, y^*(x)) \quad \text{s.t.} \quad g(x,y) - g^*(x) \le 0, \tag{**P'**}$$

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

$$\mathcal{L}_\lambda(x,y) = f(x,y) + \lambda(g(x,y) - g^*(x)). \tag{25}$$

For each $x \in \mathbb{R}^{d_x}$, recall that

$$y^*(x) := \arg\min_{y \in \mathbb{R}^{d_y}} g(x,y), \quad y^*_\lambda(x) := \arg\min_{y \in \mathbb{R}^{d_y}} \mathcal{L}_\lambda(x,y), \quad \mathcal{L}^*_\lambda(x) := \mathcal{L}_\lambda(x, y^*_\lambda(x)). \tag{26}$$

Algorithm 1 seeks to optimize $\mathcal{L}^*_\lambda(x)$, instead of the hyperobjective $F(x)$ given in (**P'**). By the first-order optimality condition for $y^*_\lambda(x)$, we have

$$0 = \nabla_y \mathcal{L}_\lambda(x, y^*_\lambda(x)) = \nabla_y f(x, y^*_\lambda(x)) + \lambda \nabla_y g(x, y^*_\lambda(x)). \tag{27}$$

According to Lemma 3.1 in (Kwon et al., 2023b), for any $x \in \mathbb{R}^{d_x}$ and $\lambda \ge 2l_{f,1}/\mu_g$,

$$\nabla \mathcal{L}^*_\lambda(x) = \nabla_x f(x, y^*_\lambda(x)) + \lambda \left(\nabla_x g(x, y^*_\lambda(x)) - \nabla_x g(x, y^*(x)) \right). \tag{28}$$

Hence, in order to estimate $\nabla \mathcal{L}^*_\lambda(x)$, one needs to estimate $y^*_\lambda(x)$ and $y^*(x)$. This are achieved by the inner loop so that $y^{k,T} \approx y^*_\lambda(x^k)$ and $z^{k,T} \approx y^*(x^k)$.

Here, we establish two preliminary lemmas that will be used in the subsequent sections. Lemma A.1 below bounds the bias of the expected gradient estimator G_k in (8) for the gradient $\nabla \mathcal{L}^*_\lambda(x^k)$ of the surrogate hyperobjective. Note that the second part of this lemma is the restatement of Lemma 3.3.

Denote

$$v^*(x) := y^*_\lambda(x) - y^*(x) \text{ and } v^k = y^k - z^k.$$

Also, we often use a short-hand $y^*_{\lambda,k} = y^*_\lambda(x^k), y^*_k = y^*(x^k)$, and $v^*_k = v^*(x^k)$. Recall that $x^{k+1} = x^k - \alpha \hat{G}_k$ where $\hat{G}_k$ is given in (10). We denote

$$G_k := \nabla_x f(x^k, y^{k+1}) + \lambda(\nabla_x g(x^k, y^{k+1}) - \nabla_x g(x^k, z^{k+1})). \tag{29}$$

Lemma A.1. *Suppose that Assumptions 1-2, (2), and (3) hold.*

(i) Then, we have

$$\|\nabla \mathcal{L}^*_\lambda(x^k) - G_k\| \le (l_{f,1} + \lambda l_{g,1}) \left(\|y^{k+1} - y^*_\lambda(x^k)\| + \|z^{k+1} - y^*(x^k)\| \right). \tag{30}$$

(ii) For $r_\lambda = \frac{l_{f,0}}{\mu_g \lambda}$, we have

$$\|\nabla \mathcal{L}_\lambda^*(x^k) - G_k\| \le l_y \|y^{k+1} - y_\lambda^*(x^k)\| + \lambda l_{g,1} \|v^{k+1} - v^*(x^k)\| + \frac{l_{f,0} l_y}{\mu_g \lambda}$$

where $l_y := l_{f,1} + \frac{l_{g,2} l_{f,0}}{\mu_g}$.

Proof. The first part directly follows from the smoothness properties of $f(x,\cdot)$ and $g(x,\cdot)$.

Let us show the second part. Using (28), we directly estimate the difference between $\nabla \mathcal{L}_\lambda^*(x^k)$ and G_k:

$$\begin{aligned}\|\nabla \mathcal{L}_\lambda^*(x^k) - G_k\| \le{}& \|\nabla_x f(x, y_{\lambda,k}^*) - \nabla_x f(x^k, y^{k+1})\| \\ &+ \lambda \|\nabla_x g(x^k, y_{\lambda,k}^*) - \nabla_x g(x^k, y_k^*) - \nabla_x g(x^k, y^{k+1}) + \nabla_x g(x^k, z^{k+1})\|.\end{aligned}$$

Using the regularity of f and g from Assumptions 1 and 2, we have

$$\|\nabla \mathcal{L}_\lambda^*(x^k) - G_k\| \le l_{f,1} \|y^{k+1} - y_{\lambda,k}^*\| + \lambda \|\nabla_{xy}^2 g(x, y_{\lambda,k}^*) v_k^* - \nabla_{xy}^2 g(x, y^{k+1}) v^{k+1}\| + \lambda l_{g,2} \left(\|v_k^*\|^2 + \|v^{k+1}\|^2\right)$$

where $v_k^* = y_{\lambda,k}^* - y_k^*$ and $v^{k+1} = y^{k+1} - z^{k+1}$. Next, note that

$$\nabla_{xy}^2 g(x, y_{\lambda,k}^*) v_k^* - \nabla_{xy}^2 g(x, y^{k+1}) v^{k+1} = (\nabla_{xy}^2 g(x, y_{\lambda,k}^*) - \nabla_{xy}^2 g(x, y^{k+1})) v_k^* + \nabla_{xy}^2 g(x, y^{k+1})(v_k^* - v^{k+1}),$$

which yields

$$\|\nabla_{xy}^2 g(x, y_{\lambda,k}^*) v_k^* - \nabla_{xy}^2 g(x, y^{k+1}) v^{k+1}\| \le l_{g,2} \|y_{\lambda,k}^* - y^{k+1}\| \|v_k^*\| + l_{g,1} \|v_k^* - v^{k+1}\|.$$

Finally, note that $\|v_k^*\|, \|v^k\| < r_\lambda = \frac{l_{f,0}}{\mu_g \lambda}$ for all k due to the projection step, and thus

$$\|\nabla \mathcal{L}_\lambda^*(x^k) - G_k\| \le (l_{f,1} + l_{g,2} \lambda r_\lambda) \|y^{k+1} - y_{\lambda,k}^*\| + \lambda l_{g,1} \|v_k^* - v^{k+1}\| + l_{g,2} \lambda r_\lambda^2.$$

As the last term $l_{g,2} \lambda r_\lambda^2 = l_{g,2} \lambda \left(\frac{l_{f,0}}{\mu_g \lambda}\right)^2 \le \frac{l_{f,0}}{\mu_g \lambda} \frac{l_{g,2} l_{f,0}}{\mu_g} \le \frac{l_{f,0} l_y}{\mu_g \lambda}$, we conclude. □

Next, the following lemma gives a bound on the variance of the stochastic gradient estimator $\hat{G}_k$ in (10).

Lemma A.2. *Suppose that Assumptions 1-2, (2), and (3) hold. Then, the variance of $\hat{G}_k$ is bounded for all k:*

$$\texttt{Var}(\hat{G}_k) := \mathbb{E}[\|\hat{G}_k - G_k\|^2] \le \frac{1}{M} \left(2\sigma_f^2 + 8\lambda^2 \sigma_g^2\right).$$

If we further assume (4) and $r_\lambda = \frac{l_{f,0}}{\mu_g \lambda}$,

$$\texttt{Var}(\hat{G}_k) := \mathbb{E}[\|\hat{G}_k - G_k\|^2] \le \frac{1}{M} \left(2\sigma_f^2 + \frac{8\tilde{l}_{g,1}^2 l_{f,0}^2}{\mu_g^2}\right).$$

Proof. Under (2), and (3), we can bounded the vairance as the following:

$$\begin{aligned}M\mathbb{E}[\|\hat{G}_k - G_k\|^2] \le{}& 2 \underbrace{\mathbb{E}[\|\hat{\nabla}_x f(x^k, y^{k+1}; \zeta_k^x) - \nabla_x f(x^k, y^{k+1})\|^2]}_{\le \sigma_f^2} \\ &+ 4\lambda^2 \underbrace{\mathbb{E}[\|\hat{\nabla}_x g(x^k, y^{k+1}; \zeta_k^x) - \nabla_x g(x^k, y^{k+1})\|^2]}_{\le \sigma_g^2} \\ &+ 4\lambda^2 \underbrace{\mathbb{E}[\|\hat{\nabla}_x g(x^k, z^{k+1}; \zeta_k^x) - \nabla_x g(x^k, z^{k+1})\|^2]}_{\le \sigma_g^2}.\end{aligned}$$

Under (4) and $r_\lambda = \frac{l_{f,0}}{\mu_g \lambda}$, we can use different inequality:

$$\begin{aligned} M\mathbb{E}[\|\hat{G}_k - G_k\|^2] \le 2\underbrace{\mathbb{E}[\|\hat{\nabla}_x f(x^k, y^{k+1}; \zeta_k^x) - \nabla_x f(x^k, y^{k+1})\|^2]}_{\le \sigma_f^2} \\ + 4\lambda^2 \underbrace{\mathbb{E}[\|\hat{\nabla}_x g(x^k, y^{k+1}; \zeta_k^x) - \hat{\nabla}_x g(x^k, z^{k+1}; \zeta_k^x)\|^2]}_{\le \tilde{l}_{g,1}\|y^{k+1} - z^{k+1}\|^2 \le \tilde{l}_{g,1} r_\lambda^2} \\ + 4\lambda^2 \underbrace{\mathbb{E}[\|\nabla_x g(x^k, y^{k+1}) - \nabla_x g(x^k, z^{k+1})\|^2]}_{\le l_{g,1}\|y^{k+1} - z^{k+1}\|^2 \le l_{g,1} r_\lambda^2}. \end{aligned}$$

Nothing that $\tilde{l}_{g,1} \ge l_{g,1}$, we have the lemma. □

A.2. $O(\epsilon^{-6})$ upper bound

In this subsection, we establish the $O(\epsilon^{-6})$ upper bound on the complexity of Algorithm 1 as stated in Theorem 3.1.

We begin with the standard argument that starts with estimating the one-step change in the (surrogate) objective $\mathcal{L}_\lambda^*(x^{k+1}) - \mathcal{L}_\lambda^*(x^k)$.

Proposition A.3. *Under the step size rule $\alpha_k \in (0, \frac{1}{2L})$ for a constant L give in Lemma C.2, we have*

$$\mathbb{E}[\mathcal{L}_\lambda^*(x^{k+1})|\mathcal{F}_k'] - \mathcal{L}_\lambda^*(x^k) \le -\frac{\alpha_k}{2}\|\nabla \mathcal{L}_\lambda^*(x^k)\|^2 - \frac{\alpha_k}{4}\|\hat{G}_k\|^2 + \alpha_k \mathtt{Var}(\hat{G}_k) + \alpha_k \|\nabla \mathcal{L}_\lambda^*(x^k) - G_k\|^2.$$

Proof. The L-smoothness of $\mathcal{L}_\lambda^*(x)$ given in Lemma C.2 and $x^{k+1} = x^k - \alpha_k \hat{G}_k$ yield that

$$\begin{aligned} \mathcal{L}_\lambda^*(x^{k+1}) - \mathcal{L}_\lambda^*(x^k) &\le \langle \nabla \mathcal{L}_\lambda^*(x^k),\, x^{k+1} - x^k \rangle + \frac{L}{2}\|x^{k+1} - x^k\|^2, \\ &= -\alpha_k \langle \nabla \mathcal{L}_\lambda^*(x^k),\, \hat{G}_k \rangle + \frac{\alpha_k^2 L}{2}\|\hat{G}_k\|^2. \\ &= -\frac{\alpha_k}{2}\|\nabla \mathcal{L}_\lambda^*(x^k)\|^2 + \frac{\alpha_k^2 L - \alpha_k}{2}\|\hat{G}_k\|^2 + \frac{\alpha_k}{2}\|\nabla \mathcal{L}_\lambda^*(x^k) - \hat{G}_k\|^2. \end{aligned}$$

Rearranging the right-hand side and using our step size rule $\alpha_k \in (0, \frac{1}{2L})$,

$$\mathcal{L}_\lambda^*(x^{k+1}) - \mathcal{L}_\lambda^*(x^k) \le -\frac{\alpha_k}{2}\|\nabla \mathcal{L}_\lambda^*(x^k)\|^2 - \frac{\alpha_k}{4}\|\hat{G}_k\|^2 + \frac{\alpha_k}{2}\|\nabla \mathcal{L}_\lambda^*(x^k) - \hat{G}_k\|^2. \tag{31}$$

By Young's inequality, the last term on the right-hand side is bounded by

$$\alpha_k \|\nabla \mathcal{L}_\lambda^*(x^k) - G_k\|^2 + \alpha_k \|\hat{G}_k - G_k\|^2. \tag{32}$$

Taking the expectation on both sides, we conclude our claim. □

Next, denote

$$\mathcal{J}_k := \|y^{k,0} - y_\lambda^*(x^k)\|^2 + \|z^{k,0} - y^*(x^k)\|^2. \tag{33}$$

We derive a recursive inequality for the above quantity to obtain the following bound on the weighted sum of the squared norm of the expected gradients.

Proposition A.4. *Under the same setting as in Theorem 3.1, we have*

$$\sum_{k=0}^{n-1} \frac{\alpha_k}{2}\mathbb{E}[\|\nabla \mathcal{L}_\lambda^*(x^k)\|^2] \le \mathbb{E}[\mathcal{L}_\lambda^*(x^0)] - \inf_{x \in \mathbb{R}^{d_x}} \mathbb{E}[\mathcal{L}_\lambda^*(x)] + \sigma_x^2 \sum_{k=0}^{n-1} \alpha_k + (\mathbb{E}[\mathcal{J}_0] + A) \sum_{k=0}^{n-1} \alpha_k \frac{16C^3/\lambda}{\mu_g(1+T)}, \tag{34}$$

where $A = \frac{(\mu_g + 2l_{g,1})^2 \sigma_g^2}{\mu_g^2 l_{g,1}} + O(\lambda^{-1})$.

Proof. From Proposition A.3, we get

$$\sum_{k=0}^{n-1} \frac{\alpha_k}{2} \mathbb{E}[\|\nabla \mathcal{L}_\lambda^*(x^k)\|^2 \le \mathbb{E}[\mathcal{L}_\lambda^*(x^0)] - \inf_{x \in \mathbb{R}^{d_x}} \mathbb{E}[\mathcal{L}_\lambda^*(x)] \tag{35}$$
$$- \frac{1}{4} \underbrace{\sum_{k=0}^{n-1} \left(\alpha_k^{-1} \mathbb{E}[\|x^{k+1} - x^k\|^2] - 2\alpha_k \mathbb{E}[\|\hat{G}_k - \nabla \mathcal{L}_\lambda^*(x^k)\|^2] \right)}_{(*)}.$$

Below, we will show that the sum $(*)$ above is uniformly lower bounded, which is enough to conclude.

Note that G_k is the estimated gradient $\nabla \mathcal{L}_\lambda^*(x^k)$ with the only source of error being approximation errors in $y^{k,T} \approx y_\lambda^*(x^k)$ and $z^{k,T} \approx y^*(x^k)$. The second quantity, $\hat{G}_k$, is the actual estimated gradient we use in Algorithm 1, with independent random noise $\xi_{k,T}^x, \xi_{k,T}^{xy}, \xi_{k,T}^{xz}$ being the additional source of error.

Using (30) in Lemma A.1, and Lemma A.2,

$$\mathbb{E}[\|\hat{G}_k - \nabla \mathcal{L}_\lambda^*(x^k) \,|\, \mathcal{F}_k\|^2] \le 2\sigma_x^2 + 4C^2 \left(\mathbb{E}\left[\|y_\lambda^*(x^k) - y^{k,T}\|^2 \,|\, \mathcal{F}_k\right] + \mathbb{E}\left[\|y^*(x^k) - z^{k,T}\|^2 \,|\, \mathcal{F}_k\right] \right) \tag{36}$$

where $\sigma_x^2 := \frac{\sigma_f^2 + 2\lambda^2 \sigma_g^2}{M}$ and $C = l_{f,1} + \lambda l_{g,1}$. Recall that $y^{k+1} = y^{k,T}$ is obtained by minimizing $\lambda^{-1}\mathcal{L}_\lambda(x, \cdot)$ using PSGD with an unbiased gradient estimator with variance $\lambda^{-2}\sigma_f^2 + \sigma_g^2 = O(1)$ (see Algorithm 1). The objective $\lambda^{-1}\mathcal{L}_\lambda(x, \cdot)$ is $(\mu_g/2)$-strongly convex (by Lemma C.3) and $C/\lambda = O(1)$-smooth where $C := l_{f,1} + \lambda l_{g,1}$. Then, we choose the inner-loop step-size

$$\gamma_t = \left(\frac{2}{\mu_g} + \frac{\lambda}{C} \right) \frac{1}{1 + \sqrt{3} + t}. \tag{37}$$

We then apply Lemma C.1 with $\mu = \mu_g/2$, $L = C/\lambda = l_{g,1} + O(\epsilon)$, $\beta = \frac{2}{\mu_g} + \frac{\lambda}{C} = O(1)$, $\sigma^2 = \lambda^{-2}\sigma_f^2 + \sigma_g^2 = \sigma_g^2 + O(\epsilon^2)$, $\gamma = 1 + \sqrt{3}$ (this value of γ was chosen so that $(\gamma + 1)(1 - \frac{2}{\gamma}) \le 1$) and the constraint set $\mathcal{B}$ to be the radius $2r/3$-ball around $\hat{y}(x)$, which contains $y^*(x)$ by the hypothesis. (When $r = \infty$, $\mathcal{B}$ becomes the whole space.) This gives us

$$\mathbb{E}[\|y^{k,T} - y_\lambda^*(x^k)\|^2 \,|\, \mathcal{F}_k] \le \frac{\max\left\{\tilde{C},\ \|y^{k,0} - y_\lambda^*(x^k)\|^2\right\}}{1 + T}, \tag{38}$$

where $\tilde{C} := \frac{(\mu_g + l_{g,1})^2 \sigma_g^2 (l_{g,1} + 2^{-1})}{2\mu_g^2 l_{g,1}^2} + O(\epsilon)$.

Similarly, since $g(x, \cdot)$ is μ_g-strongly convex and $l_{g,1}$-smooth,

$$\mathbb{E}[\|z^{k,T} - y^*(x^k)\|^2 \,|\, \mathcal{F}_k] \le \frac{\max\left\{\tilde{C},\ \|z^{k,0} - y^*(x^k)\|^2\right\}}{1 + T}. \tag{39}$$

for the same constant $\tilde{C}$ as above. Then bounding the maximum of nonnegative quantities by their sum and combining (36), (38), and (39),

$$\mathbb{E}\left[\|G_k - \nabla \mathcal{L}_\lambda^*(x^k)\|^2 \,\middle|\, \mathcal{F}_k \right] \le \frac{8C^3/\lambda}{\mu_g(1 + T)} \left[2\tilde{C} + \mathcal{J}_k \right], \tag{40}$$

Next, we derive a recursion for $\mathbb{E}[\mathcal{J}_k]$. By Young's inequality, (38), and that y_λ^* is $(4l_{g,1}/\mu_g)$-Lipschitz continuous (see Lemma C.5), we obtain

$$\mathbb{E}[\|y^{k+1,0} - y_\lambda^*(x^{k+1})\|^2 \,|\, \mathcal{F}_k] \tag{41}$$
$$\le 2\mathbb{E}[\|y^{k,T} - y_\lambda^*(x^k)\|^2 \,|\, \mathcal{F}_k] + 2\mathbb{E}[\|y_\lambda^*(x^{k+1}) - y_\lambda^*(x^k)\|^2 \,|\, \mathcal{F}_k] \tag{42}$$
$$\le \frac{8(C/\lambda) \max\left\{\tilde{C},\ \|y^{k,0} - y_\lambda^*(x^k)\|^2\right\}}{\mu_g(1 + T)} + \frac{32 l_{g,1}^2}{\mu_g^2} \mathbb{E}[\|x^{k+1} - x^k\|^2 \,|\, \mathcal{F}_k]. \tag{43}$$

Similarly, we also have

$$\mathbb{E}[\|z^{k+1,0} - y^*(x^{k+1})\|^2 \,|\, \mathcal{F}_k] \tag{44}$$

$$\leq \frac{4l_{g,1}\max\left\{\bar{C},\, \|z^{k,0} - y^*(x^k)\|^2\right\}}{\mu_g(1+T)} + \frac{2l_{g,1}^2}{\mu_g^2}\mathbb{E}[\|x^{k+1} - x^k\|^2 \,|\, \mathcal{F}_k]. \tag{45}$$

Suppose T is large enough so that $\mu_g(1+T) \geq 16C/\lambda$. Then using (38) and (39),

$$\mathbb{E}[\mathcal{J}_{k+1} \,|\, \mathcal{F}_k] \leq \frac{\mathcal{J}_k}{2} + \tilde{C} + \frac{34l_{g,1}^2}{\mu_g^2}\mathbb{E}[\|x^{k+1} - x^k\|^2 \,|\, \mathcal{F}_k], \tag{46}$$

where A is the constant defined in (40). Taking the full expectation and by induction, we get

$$\mathbb{E}[\mathcal{J}_k] \leq 2^{-k}(\mathbb{E}[\mathcal{J}_0] + 2\tilde{C}) + \frac{34l_{g,1}^2}{\mu_g^2}\sum_{i=0}^{k}\left(\frac{1}{2}\right)^{k-1-i}\mathbb{E}[\|x^{i+1} - x^i\|^2]. \tag{47}$$

It follows that, combining (40) and (47), we get

$$\frac{\mu_g}{8C^3/\lambda}\sum_{k=0}^{n-1}\alpha_k\mathbb{E}\left[\|G_k - \nabla\mathcal{L}_\lambda^*(x^k)\|^2\right] \leq \sum_{k=0}^{n-1}\frac{\alpha_k}{(1+T)}\left(2\tilde{C} + \mathbb{E}[\mathcal{J}_k]\right) \tag{48}$$

$$\leq (\mathbb{E}[\mathcal{J}_0] + 2\tilde{C})\sum_{k=0}^{n-1}\frac{2\alpha_k}{(1+T)} + \frac{68l_{g,1}^2}{\mu_g^2}\sum_{k=0}^{n-1}\alpha_k\|x^{k+1} - x^k\|^2. \tag{49}$$

Thus, we deduce for α_k sufficiently small,

$$(*) \geq -(\mathbb{E}[\mathcal{J}_0] + 2\tilde{C})\sum_{k=0}^{n-1}\frac{16\alpha_k^2C^3/\lambda}{\mu_g(1+T)} + \sigma_x^2\sum_{k=0}^{n-1}\alpha_k \tag{50}$$

$$+ \sum_{k=0}^{n-1}\alpha_k^{-1}\mathbb{E}[\|x^{k+1} - x^k\|^2]\left(1 - \frac{4\cdot 8\cdot 68C^3l_{g,1}^2/\lambda}{\mu_g^3}\alpha_k^2\right) \tag{51}$$

$$\geq -(\mathbb{E}[\mathcal{J}_0] + 2\tilde{C})\sum_{k=0}^{n-1}\frac{16\alpha_k^2C^3/\lambda}{\mu_g(1+T)} + \sigma_x^2\sum_{k=0}^{n-1}\alpha_k. \tag{52}$$

Combining the above with (35), we conclude (34). □

Now, we prove the upper bound $O(\epsilon^{-6})$.

Proof of Theorem 3.1: Since $\lambda = O(\epsilon^{-1})$, by Lemma C.5, $\|\nabla\mathcal{L}_\lambda^*(x^k)\| = O(\epsilon)$ implies $\|\nabla F(x^k)\| = O(\epsilon)$. Hence it is enough to show that Algorithm 1 finds an ϵ-stationary point of the surrogate objective $\mathcal{L}_\lambda^*$ within $O(\epsilon^{-6})$ iterations.

For $M \asymp \lambda^4 = \epsilon^{-4}$, we have $\sigma_x^2 = \frac{\sigma_f^2 + 2\lambda^2\sigma_g^2}{M} = O(\sigma_g^2\epsilon^2)$, Then, by Proposition A.4, to obtain ϵ-stationary point, it is enough to have (recall that $C = \Theta(\lambda)$, $A = O(1)$, and $\sigma_x^2 = \Theta(\epsilon^2)$)

$$\frac{C_1}{\sum_{k=0}^{n-1}\alpha_k} + \frac{\sigma_x^2\sum_{k=0}^{n-1}\alpha_k}{\sum_{k=0}^{n-1}\alpha_k} + \frac{C_2\lambda^2\sum_{k=0}^{n-1}\frac{\alpha_k}{T}}{\sum_{k=0}^{n-1}\alpha_k} \leq \epsilon^2, \tag{53}$$

where

$$C_1 = \mathcal{L}_\lambda^*(x^0) - \inf_{x\in\mathbb{R}^{d_x}}\mathcal{L}_\lambda^*(x), \quad C_2 = \frac{l_{g,1}^3}{\mu_g}\left(\mathcal{J}_0 + \frac{(\mu_g + 2l_{g,1})^2\sigma_g^2}{\mu_g^2 l_{g,1}} + O(\lambda^{-1})\right). \tag{54}$$

Recalling $\lambda = \epsilon^{-1}$, note that (53) follows from

$$\frac{C_1}{\sum_{k=0}^{n-1}\alpha_k} + \frac{C_2\epsilon^{-2}}{T} + \frac{\sigma_f^2 + 2\epsilon^{-2}\sigma_g^2}{M} \leq \epsilon^2. \tag{55}$$

Hence it is enough to set each of the three terms in the left-hand side above is at most $\epsilon^2/3$. This is the case when $\alpha_k \equiv \alpha \in (0, \frac{1}{2L})$ where $L = \frac{6l_{g,1}}{\mu_g}\left(l_{f,1} + \frac{l_{g,1}^2}{\mu_g} + \frac{l_{f,0}l_{g,1}l_{g,2}}{\mu_g^2}\right)$ as is in Lem. C.2 and

$$\frac{C_1}{n\alpha} \le \epsilon^2/3, \qquad 3C_2\epsilon^{-2} \le \epsilon^2 T, \qquad 3(\sigma_f^2 + 2\epsilon^{-2}\sigma_g^2) \le M\epsilon^2, \tag{56}$$

so we can set

$$n = 3(C_1/\alpha)\epsilon^{-2}, \qquad T = 3C_2\epsilon^{-4}, \qquad M = 3\epsilon^{-2}(\sigma_f^2 + 2\epsilon^{-2}\sigma_g^2). \tag{57}$$

Thus the total sample complexity is at most

$$n(T+M) \le \frac{9C_1(C_2 + 2\sigma_g^2)}{\alpha}\epsilon^{-6} + \frac{9C_1\sigma_g^2}{\alpha}\epsilon^{-4} = O(\epsilon^{-6}). \tag{58}$$

In particular, is we set $\alpha \ge cL$ for some constant $c > 0$, then the leading term is of order $LC_1(C_2 + 2\sigma_g^2)\epsilon^{-6}$.

□

A.3. $O(\epsilon^{-4})$ upper bound

In this section, we provide the proofs of Theorem 3.2 and Proposition 3.5. Throughout this section, we assume that Assumptions 1-2, (2), (3), and (4) hold and let $\gamma_t \equiv \gamma$ for all t.

As illustrated in Section 3.2, our new key observation in Lemma A.1 is the estimation of $\|\nabla\mathcal{L}_\lambda^*(x^k) - G_k\|$ in terms of $\|y^{k+1} - y_\lambda^*(x^k)\|$ and $\|v^{k+1} - v^*(x^k)\|$. In the subsequent section, we estimate each term. Let

$$\mathcal{I}_k := \|y^k - y_\lambda^*(x^k)\|^2 \text{ and } \mathcal{W}_k := \|v^k - v^*(x^k)\|^2, \tag{59}$$

On the other hand, we have the additional projection step for z in the inner loop of Algorithm 1:

$$z^{k,t+1} \leftarrow \Pi_{\mathbb{B}(y^{k,t+1}, r_\lambda)}\left\{\bar{z}^{k,t} + \Delta_y\right\},$$

where $\Delta_y = y^{k,t+1} - \bar{y}^{k,t+1}$ and therefore $\bar{y}^{k,t} - \bar{z}^{k,t} = y^{k,t+1} - (\bar{z}^{k,t} + \Delta_y)$. In Appendix A.3.2, we verify that with the appropriate choice of r_λ the projection step makes $\|v^{k+1} - v^*(x^k)\|$ smaller, which yields the desired result.

A.3.1. Descent lemma for y

Proposition A.5. *For given $\beta \ge 2$, assume that*

$$\lambda > \frac{l_{f,1}}{l_{g,1}}, \ \ \gamma \in \left(0, \frac{1}{4l_{g,1}}\right) \text{ and } \left(1 - \frac{\mu_g\gamma}{2}\right)^T < \frac{1}{2\beta}. \tag{60}$$

Then, we have

$$\mathbb{E}[\|y^{k+1} - y_\lambda^*(x^k)\|^2 \,|\, \mathcal{F}_k] \le \frac{1}{2\beta}\mathcal{I}_k + \frac{4\gamma}{\mu_g}\mathtt{Var}(\hat{\nabla}_y L_\lambda)\frac{1}{\lambda^2} \tag{61}$$

where

$$\mathtt{Var}(\hat{\nabla}_y L_\lambda) := \sup_{x,y}\mathtt{Var}(\hat{\nabla}_y f(x,y;\zeta) + \lambda\hat{\nabla}_y g(x,y;\xi)). \tag{62}$$

Furthermore,

$$\sum_{k=0}^{n}\mathbb{E}[\mathcal{I}_k] \le 2\mathcal{I}_0 + \frac{8\gamma}{\mu_g}\mathtt{Var}(\hat{\nabla}_y L_\lambda)\frac{n}{\lambda^2} + \frac{64l_{g,1}^2}{\mu_g^2}\sum_{k=0}^{n-1}\mathbb{E}[\|\hat{G}_k\|^2]. \tag{63}$$

Proof. Recall from Lemma C.3 that $\mathcal{L}_\lambda(x,\cdot)$ is $(\lambda\mu_g/2)$-strongly convex. In addition, for given x, $\mathcal{L}_\lambda(x,\cdot)$ is $(l_{f,1}+\lambda l_{g,1})$-smooth in y. Then, Applying Lemma C.1 with the objective function $y \mapsto \lambda^{-1}\mathcal{L}_\lambda(x,y)$, $\mu = \mu_g/2$, $\sigma^2 = \texttt{Var}(\hat{\nabla}_y L_\lambda)$, $L = 4l_{g,1}$ and the constraint set $\mathcal{B}$ to be the radius $2r/3$-ball around $\hat{y}(x)$, which contains $y^*(x)$ by the hypothesis, we get

$$\mathbb{E}[\|y^{k,T} - y_\lambda^*(x^k)\|^2 \,|\, \mathcal{F}_k] \le \left(1 - \frac{\mu_g\gamma}{2}\right)^T \|y_{k,0} - y_\lambda^*(x_k)\|^2 + \frac{4\gamma}{\lambda^2\mu_g}\texttt{Var}(\hat{\nabla}_y L_\lambda). \tag{64}$$

Thanks to our assumptions in (60), we conclude (61).

Next, let us estimate $\mathcal{I}_{k+1}$ in terms of $\mathcal{I}_k$. Young's inequality, the $(4l_{g,1}/\mu_g)$-Lipschitz continuity of y_λ^* in Lemma C.5, and (61) yield

$$\mathbb{E}[\mathcal{I}_{k+1} \,|\, \mathcal{F}_k] \le 2\mathbb{E}[\|y^{k+1} - y_\lambda^*(x^k)\|^2 \,|\, \mathcal{F}_k] + 2\mathbb{E}[\|y_\lambda^*(x^{k+1}) - y_\lambda^*(x^k)\|^2 \,|\, \mathcal{F}_k] \tag{65}$$

$$\le \frac{1}{\beta}\mathcal{I}_k + \frac{4\gamma}{\lambda^2\mu_g}\texttt{Var}(\hat{\nabla}_y L_\lambda) + \frac{32l_{g,1}^2\alpha^2}{\mu_g^2}\mathbb{E}[\|\hat{G}_k\|^2 \,|\, \mathcal{F}_k]. \tag{66}$$

Taking the full expectation and by induction, we get

$$\mathbb{E}[\mathcal{I}_k] \le \beta^{-k}\mathcal{I}_0 + \sum_{i=0}^{k-1}\beta^{-k+1+i}\left(\frac{32l_{g,1}^2\alpha^2}{\mu_g^2}\mathbb{E}[\|\hat{G}_i\|^2] + \frac{4\gamma}{\lambda^2\mu_g}\texttt{Var}(\hat{\nabla}_y L_\lambda)\right), \tag{67}$$

$$\le \beta^{-k}\mathcal{I}_0 + \frac{1}{1-1/\beta}\frac{4\gamma}{\lambda^2\mu_g}\texttt{Var}(\hat{\nabla}_y L_\lambda) + \frac{32l_{g,1}^2\alpha^2}{\mu_g^2}\sum_{i=0}^{k-1}\beta^{-k+1+i}\mathbb{E}[\|\hat{G}_i\|^2]. \tag{68}$$

As $\beta > 2$, we have $\frac{1}{1-1/\beta} \le 2$ and we conclude (63). □

A.3.2. Estimates for v_k^*

Next, we prove Proposition 3.4. Recall that $\bar{z}^{k,t} = z^{k,t} - \gamma_t\hat{\nabla}_y g(x^k, z^{k,t}; \xi_{k,t}^y)$ (before projection to the ball around y^{k+1}), and $\bar{v}^{k,t} := \bar{y}^{k,t} - \bar{z}^{k,t}$. Also recall that $\Delta_y = y^{k,t+1} - \bar{y}^{k,t+1}$ and therefore $\bar{v}^{k,t} = y^{k,t+1} - (\bar{z}^{k,t} + \Delta_y)$. Note that for an appropriate choice of t satisfying $\|u\| \le t \le \|v\|$, the projection of v to a ball of radius t makes the distance to u smaller.

Lemma A.6. *For any $u, v \in \mathbb{R}^{d_y}$ with $\|v\| \ge t$ and $\|u\| \le t$ for some $r > 0$, the following holds:*

$$\left\|\frac{tv}{\|v\|} - u\right\| \le \|v - u\|.$$

Proof of Proposition 3.4: First of all, if for $\bar{v}^{k,t} = \bar{y}^{k,t} - \bar{z}^{k,t}$, $\|\bar{v}^{k,t}\| \le r_\lambda$, then the equality holds in (12). Otherwise, suppose that

$$\|\bar{v}^{k,t}\| > r_\lambda.$$

Recall from Lemma C.4 that for all $x \in \mathbb{R}^{d_x}$,

$$\|v^*(x)\| = \|y_\lambda^*(x) - y^*(x)\| \le \frac{l_{f,0}}{\lambda\mu_g} = r_\lambda. \tag{69}$$

Applying Lemma A.6 with $r = r_\lambda$, we conclude (12). □

Next, we obtain the contraction of $v_{k+1}^* - v_k^*$.

Lemma A.7. *For all k, we have*

$$\|v^*(x^{k+1}) - v^*(x^k)\|^2 = \|v_{k+1}^* - v_k^*\|^2 \lesssim \frac{l_{g,1}^2 l_v^2}{\mu_g^4\lambda^2}\|x^{k+1} - x^k\|^2 + \frac{l_{g,2}^2 l_{f,0}^4}{\mu_g^6\lambda^4}$$

where $l_v := l_{g,1} + \frac{l_{f,0}l_{g,2}}{\mu_g}$.

Proof. We first show that

$$v^*(x) = y_\lambda^*(x) - y^*(x) = -\frac{1}{\lambda}(\nabla_{yy}^2 g(x, y^*(x)))^{-1}\nabla_y f(x, y^*(x)) + O\left(\frac{l_{g,2}l_{f,0}^2}{\mu_g^3\lambda^2}\right). \tag{70}$$

To see this, note that

$$\nabla_y g(x, y^*(x)) = 0, \quad \nabla_y g(x, y_\lambda^*(x)) + \lambda^{-1}\nabla_y f(x, y_\lambda^*(x)) = 0,$$

and thus,

$$\begin{aligned}
-\frac{1}{\lambda}\nabla_y f(x, y_\lambda^*(x)) &= \nabla_y g(x, y_\lambda^*(x)) - \nabla_y g(x, y^*(x)) \\
&= \nabla_{yy}^2 g(x, y^*(x))(y_\lambda^*(x) - y^*(x)) + O(l_{g,2}\|y^*(x) - y_\lambda^*(x)\|^2) \\
&= \nabla_{yy}^2 g(x, y^*(x))(y_\lambda^*(x) - y^*(x)) + O\left(\frac{l_{g,2}l_{f,0}^2}{\mu_g^3\lambda^2}\right).
\end{aligned}$$

Multiplying both sides by $\nabla_{yy}^2 g(x, y^*(x))$ gives (70). Thus, $v_{k+1}^* - v_k^*$ can be expressed as

$$\begin{aligned}
&v_{k+1}^* - v_k^* \\
&= v^*(x^{k+1}) - v^*(x^k) \\
&= \frac{1}{\lambda}\left(\nabla_{yy}^2 g(x^k, y^*(x^k))^{-1}\nabla_y f(x^k, y^*(x^k)) - \nabla_{yy}^2 g(x^{k+1}, y^*(x^{k+1}))^{-1}\nabla_y f(x^{k+1}, y^*(x^{k+1}))\right) + O\left(\frac{l_{g,2}l_{f,0}^2}{\mu_g^3\lambda^2}\right) \\
&= \frac{1}{\lambda}\nabla_{yy}^2 g(x^k, y^*(x^k))^{-1}\left(\nabla_y f(x^k, y^*(x^k)) - \nabla_y f(x^{k+1}, y^*(x^{k+1}))\right) \\
&\quad + \frac{1}{\lambda}\left(\nabla_{yy}^2 g(x^k, y^*(x^k))^{-1} - \nabla_{yy}^2 g(x^{k+1}, y^*(x^{k+1}))^{-1}\right)\nabla_y f(x^k, y^*(x^k)) + O\left(\frac{l_{g,2}l_{f,0}^2}{\mu_g^3\lambda^2}\right) \\
&= \frac{1}{\lambda}\left(\frac{l_{g,1}}{\mu_g} + \frac{l_{g,2}l_{f,0}}{\mu_g^2}\right)\cdot O\left(\|x^k - x^{k+1}\| + \|y^*(x^k) - y^*(x^{k+1})\|\right) + O\left(\frac{l_{g,2}l_{f,0}^2}{\mu_g^3\lambda^2}\right).
\end{aligned}$$

Finally, using that $\|y^*(x^k) - y^*(x^{k+1})\| \le \frac{3l_{g,1}}{\mu_g}\|x^k - x^{k+1}\|$, we have

$$\|v_{k+1}^* - v_k^*\| \lesssim \frac{l_{g,1}}{\lambda\mu_g^2}\underbrace{\left(l_{g,1} + \frac{l_{g,2}l_{f,0}}{\mu_g}\right)}_{l_v}\|x^k - x^{k+1}\| + \frac{l_{g,2}l_{f,0}^2}{\mu_g^3\lambda^2}.$$

Having squares on both sides gives the lemma. □

Lemma A.8. *Suppose that*

$$\gamma \in \left(0, \frac{\mu_g}{4\tilde{l}_{g,1}^2}\right). \tag{71}$$

For all k and $t = 1, 2, \cdots, T-1$, we have

$$\mathbb{E}[\|\bar{v}^{k,t} - v_k^*\|^2|\mathcal{F}_{k,t}] \le (1 - \mu_g\gamma/2)\|v^{k,t} - v_k^*\|^2 + O(\gamma^2)\frac{\tilde{l}_{g,1}^2 l_{f,0}^2}{\lambda^2\mu_g^2}. \tag{72}$$

Proof. In this proof, we only consider the projection step

$$z^{k,t+1} \leftarrow \Pi_{\mathbb{B}(y^{k,t+1}, r_\lambda)}\left\{\bar{z}^{k,t} + \Delta_y\right\}.$$

Other projections can be handled by a small modification of Lemma C.1. By direct computation, we have

$$\begin{aligned}
&\mathbb{E}[\|\bar{v}^{k,t} - v_k^*\|^2|\mathcal{F}_{k,t}] \\
&= \mathbb{E}[\|v_k^* - \bar{v}^{k,t}\|^2 + \|v_{k+1} - \bar{v}^{k,t}\|^2 - 2\langle v_k^* - v^{k,t}, \bar{v}^{k,t} - v^{k,t}\rangle|\mathcal{F}_{k,t}] \\
&\le \|v_k^* - v^{k,t}\|^2 + \gamma^2 \cdot \mathbb{E}[\|\lambda^{-1}\nabla_y f(x^k, y^k; \zeta_{k,t}^y) + \nabla_y g(x^k, y^{k,t}; \xi_{k,t}^y) - \nabla_y g(x^k, z^{k,t}; \xi_{k,t}^y)\|^2|\mathcal{F}_{k,t}] \\
&\quad - 2\gamma\langle v_k^* - v^{k,t}, \lambda^{-1}\nabla_y f(x^k, y^{k,t}) + \nabla_y g(x^k, y^{k,t}) - \nabla_y g(x^k, z^{k,t})\rangle \\
&\le (1 - \mu_g\gamma)\|v_k^* - v^k\|^2 + \frac{\gamma^2}{\lambda^2}(\sigma_f^2 + l_{f,0}^2) + \gamma^2\tilde{l}_{g,1}^2\|y^{k,t} - z^{k,t}\|^2.
\end{aligned}$$

where in the last line, we used co-covercivity of strongly-convex functions $\langle x - y, \nabla g(x) - \nabla g(y)\rangle \ge \mu_g\|x - y\|^2$. Finally, note that

$$\|y^{k,t} - z^{k,t}\|^2 \le 2(\|v^{k,t} - v_k^*\|^2 + \|v_k^*\|^2) \le 2\|v^{k,t} - v_k^*\|^2 + 2r_\lambda^2.$$

Using (71), we conclude (72). □

Lastly, we have the desired estimates for $\|v^{k+1} - v^*(x^k)\|$ and $\mathcal{W}_k$.

Proposition A.9. *For given $\beta > 2$, assume that (71) and*

$$\left(1 - \frac{\mu_g\gamma}{2}\right)^T < \frac{1}{2\beta}. \tag{73}$$

Then,

$$\mathbb{E}[\|v^{k+1} - v^*(x^k)\|^2 \,|\, \mathcal{F}_k] \le \frac{1}{2\beta}\mathbb{E}[\mathcal{W}_k] + O(\gamma)\frac{\tilde{l}_{g,1}^2 l_{f,0}^2}{\lambda^2\mu_g^3}. \tag{74}$$

Furthermore,

$$\frac{\lambda^2}{n}\sum_{k=0}^{n}\mathbb{E}[\mathcal{W}_k] \le \frac{1}{n}2\mathcal{W}_0\lambda^2 + \frac{l_{g,1}^2 l_v^2\alpha^2}{\mu_g^4}\frac{1}{n}\sum_{k=0}^{n-1}\|\hat{G}_k\|^2 + C\frac{\tilde{l}_{g,1}^2 l_{f,0}^2}{\mu_g^3}\gamma + \frac{l_{g,2}^2 l_{f,0}^4}{\mu_g^6}\frac{1}{\lambda^2}. \tag{75}$$

Proof. Proposition 3.4 and Lemma A.8 yield

$$\mathbb{E}[\|v^{k,t+1} - v^*(x^k)\|^2|\mathcal{F}_{k,t}] \le (1 - \mu_g\gamma/2)\|v^{k,t} - v^*(x^k)\|^2 + O(\gamma^2)\cdot\frac{\tilde{l}_{g,1}^2 l_{f,0}^2}{\lambda^2\mu_g^2}. \tag{76}$$

for all t and k. By iterating this for $t = 0, 1, \cdots, T-1$, we have

$$\mathbb{E}[\|v^{k+1} - v^*(x^k)\|^2|\mathcal{F}_k] = \mathbb{E}[\|v^{k,T} - v^*(x^k)\|^2|\mathcal{F}_k] \le (1 - \mu_g\gamma/2)^T\|v^{k,t} - v_k^*\|^2 + O(\gamma)\frac{\tilde{l}_{g,1}^2 l_{f,0}^2}{\lambda^2\mu_g^3}. \tag{77}$$

Using the assumption (73), the above yields the first inequality (74).

On the other hand, using Young's inequality,

$$\mathcal{W}_{k+1} = \|v^{k+1} - v^*(x^{k+1})\|^2 \le 2\|v^{k+1} - v^*(x^k)\|^2 + 2\|v^*(x^{k+1}) - v^*(x^k)\|^2.$$

Applying (74) and Lemma A.7, we obtain that

$$\mathbb{E}[\mathcal{W}_{k+1}|\mathcal{F}_k] \le 2\mathbb{E}[\|v^{k+1} - v^*(x^k)\|^2|\mathcal{F}_k] + \frac{l_{g,1}^2 l_v^2\alpha^2}{\mu_g^4\lambda^2}\|\hat{G}_k\|^2 + \frac{l_{g,2}^2 l_{f,0}^4}{\mu_g^6\lambda^4}, \tag{78}$$

$$\le \frac{1}{\beta}\mathcal{W}_k + \frac{l_{g,1}^2 l_v^2\alpha^2}{\mu_g^4}\|\hat{G}_k\|^2\frac{1}{\lambda^2} + C\frac{\tilde{l}_{g,1}^2 l_{f,0}^2}{\lambda^2\mu_g^3}\frac{\gamma}{\lambda^2} + \frac{l_{g,2}^2 l_{f,0}^4}{\mu_g^6}\frac{1}{\lambda^4}, \tag{79}$$

for some constant $C > 0$. The assumption (73) was used in the last inequality.

Using the parallel argument as in Proposition A.5, we conclude (75). □

A.3.3. Proofs of Theorem 3.2 and Proposition 3.5

The following statement is the detailed version of Proposition 3.5.

Proposition A.10. *Suppose that Assumptions 1-2, (2), (3), and (4). In addition, let the algorithm parameters satisfy (71), (73),*

$$r_\lambda = \frac{l_{f,0}}{\mu_g \lambda}, \quad \alpha \ll 1, \quad r \in [6r_\lambda, \infty] \tag{80}$$

and

$$-\frac{1}{4} + C_1\alpha^2 < 0 \text{ where } C_1 = l_y^2 \frac{64 l_{g,1}^2}{\mu_g^2} + \frac{l_{g,1}^2 l_v^2}{\mu_g^4} \tag{81}$$

hold true. Then,

$$\frac{\alpha}{2}\sum_{k=0}^{n-1}\mathbb{E}[\|\nabla\mathcal{L}_\lambda^*(x^k)\|^2 - \mathcal{L}_\lambda^*(x^0) + \inf_{x\in\mathbb{R}^{d_x}}\mathbb{E}[\mathcal{L}_\lambda^*(x)] \le \alpha l_y^2\mathcal{I}_0 + \alpha l_{g,1}^2\mathcal{W}_0\lambda^2 + \alpha\sum_{k=0}^{n-1}\texttt{Var}(\hat{G}_k) + nO(\lambda^{-2}) + nO(\gamma). \tag{82}$$

Proof. Let $\mathcal{B}_k := \mathbb{E}[\mathcal{L}_\lambda^*(x^{k+1})|\mathcal{F}_k] - \mathcal{L}_\lambda^*(x^k) + \frac{\alpha}{2}\|\nabla\mathcal{L}_\lambda^*(x^k)\|^2$. Recall from Proposition A.3 that

$$\mathcal{B}_k \le -\frac{\alpha}{4}\|\hat{G}_k\|^2 + \alpha\texttt{Var}(\hat{G}_k) + 2\alpha\left(\frac{l_{f,0}l_y}{\mu_g}\right)^2\frac{1}{\lambda^2} + 2\alpha l_y^2\|y^{k+1} - y_\lambda^*(x^k)\|^2 + 2\alpha\lambda^2 l_{g,1}^2\|v^{k+1} - v^*(x^k)\|^2.$$

The upper bounds of the last two terms, $\|y^{k+1} - y_\lambda^*(x^k)\|^2$ and $\|v^{k+1} - v^*(x^k)\|^2$, are given in Proposition A.5, and Proposition A.9, respectively. Using them, we obtain that

$$\begin{aligned}\mathcal{B}_k \le& -\frac{\alpha}{4}\|\hat{G}_k\|^2 + \alpha\texttt{Var}(\hat{G}_k) + 2\alpha\left(\frac{l_{f,0}l_y}{\mu_g}\right)^2\frac{1}{\lambda^2}\\ &+ \alpha l_y^2\mathbb{E}[\mathcal{I}_k]\frac{1}{\beta} + \frac{8\alpha l_y\gamma}{\mu_g}\texttt{Var}(\hat{\nabla}\mathcal{L}_\lambda)\frac{1}{\lambda^2}\\ &+ \alpha l_{g,1}^2\mathbb{E}[\lambda^2\mathcal{W}_k]\frac{1}{\beta} + 2\alpha l_{g,1}^2\frac{\tilde{l}_{g,1}^2 l_{f,0}^2}{\mu_g^3}O(\gamma) + 2\alpha l_{g,1}^2\frac{l_{g,2}^2 l_{f,0}^4}{\mu_g^6}\frac{1}{\lambda^2}\end{aligned}$$

Choosing $\beta = 2$, taking the full expectation and telescoping over $k = 0, \dots, n-1$,

$$\sum_{k=0}^{n-1}\mathcal{B}_k \le \alpha l_y^2\mathcal{I}_0 + \alpha l_{g,1}^2\mathcal{W}_0\lambda^2 + \alpha\left(-\frac{1}{4} + C_1\alpha^2\right)\sum_{k=0}^{n-1}\|\hat{G}_k\|^2 + \alpha\sum_{k=0}^{n-1}\texttt{Var}(\hat{G}_k) \tag{83}$$

$$+\left(2\alpha\left(\frac{l_{f,0}l_y}{\mu_g}\right)^2\frac{1}{\lambda^2} + \frac{8}{\mu_g}\frac{\texttt{Var}(\hat{\nabla}L_\lambda)\gamma}{\lambda^2} + 2\alpha l_{g,1}^2\frac{\tilde{l}_{g,1}^2 l_{f,0}^2}{\mu_g^3}\gamma + 2\alpha l_{g,1}^2\frac{l_{g,2}^2 l_{f,0}^4}{\mu_g^6}\frac{1}{\lambda^2}\right)n \tag{84}$$

for C_1 given in (81).

On the other hand,

$$\frac{\alpha}{2}\sum_{k=0}^{n-1}\mathbb{E}[\|\nabla\mathcal{L}_\lambda^*(x^k)\|^2 - \mathcal{L}_\lambda^*(x^0) + \inf_{x\in\mathbb{R}^{d_x}}\mathbb{E}[\mathcal{L}_\lambda^*(x)] \le \sum_{k=0}^{n-1}\mathcal{B}_k. \tag{85}$$

Thanks to (81), and the fact that $\texttt{Var}(\hat{\nabla}L_\lambda) = O(\lambda^2)$, we conclude. □

Proof of Theorem 3.2: Since $\lambda = O(\epsilon^{-1})$, by Lemma C.5, $\|\nabla \mathcal{L}_\lambda^*(x_k)\| = O(\epsilon)$ implies $\|\nabla F(x_k)\| = O(\epsilon)$. Hence it is enough to show that Algorithm 1 finds an ϵ-stationary point of the surrogate objective $\mathcal{L}_\lambda^*$ within $O(\epsilon^{-4})$ iterations, respectively.

After the projection steps before entering the inner loop, we have $\|y^{k,0} - z^{k,0}\| \le r_\lambda$ for all $k \ge 0$. From Proposition A.10, we have

$$\frac{\alpha}{2}\frac{1}{n}\sum_{k=0}^{n-1}\mathbb{E}[\|\nabla \mathcal{L}_\lambda^*(x^k)\|^2 \le \frac{C}{n} + \frac{\alpha}{n}\sum_{k=0}^{n-1}\texttt{Var}(\hat{G}_k) + O(\lambda^{-2}) + O(\gamma). \tag{86}$$

where $C = \mathcal{L}_\lambda^*(x^0) - \inf_{x\in\mathbb{R}^{d_x}}\mathbb{E}[\mathcal{L}_\lambda^*(x)] + \alpha l_y^2 \mathcal{I}_0 + \alpha l_{g,1}^2 \mathcal{W}_0 \lambda^2$.

For $M \asymp O(\epsilon^{-2})$, Lemma A.2 yields that $\texttt{Var}(\hat{G}_k) = \frac{1}{M}\left(2\sigma_f^2 + \frac{8\tilde{l}_{g,1}^2 l_{f,0}^2}{\mu_g^2}\right) = O(\epsilon^2)$. Choosing $\gamma \asymp O(\epsilon^2)$ and $n \asymp O(\epsilon^{-2})$,

$$\frac{\alpha}{2}\frac{1}{n}\sum_{k=0}^{n-1}\mathbb{E}[\|\nabla \mathcal{L}_\lambda^*(x^k)\|^2 = O(\epsilon^2).$$

For $T \asymp O(\epsilon^{-2})$, the step size rule is satisfied (60). Thus, we conclude that the total complexity is $O(n\cdot(M+T)) = O(\epsilon^{-4})$.

B. Proofs for Lower Bounds

In this section, we establish Theorem 4.5.

B.1. Auxiliary Lemmas for Lower Bounds

Lemma B.1 ((Carmon et al., 2020), Lemma 1)**.** *Suppose $\Psi(t), \Phi(t)$ is defined as in (18):*

$$\Psi(t) = \begin{cases} 0, & t \le 1/2, \\ \exp\left(1 - \frac{1}{(2t-1)^2}\right), & t > 1/2, \end{cases} \qquad \Phi(t) = \sqrt{e}\int_{-\infty}^{t} e^{-\frac{1}{2}\tau^2} d\tau.$$

They satisfy the following properties:

1. *Both Ψ and Φ are infinitely differentiable.*
2. *For all $t \in \mathbb{R}$,*

$$0 \le \Psi(t) \le e,\ 0 \le \Psi'(t) \le \sqrt{54/e},\ |\Psi''(t)| \le 32.5$$
$$0 \le \Phi(t)\sqrt{2\pi e},\ 0 \le \Phi'(t) \le \sqrt{e},\ |\Phi''(t)| \le 1.$$

3. *For all $t \ge 1$ and $|u| \le 1$, $\Psi(t)\Phi'(u) \ge 1$.*

Lemma B.2 ((Carmon et al., 2020), Lemma 1, 2)**.** *If there exists $0 \le i < d_x$ such that $|x_i| \ge \epsilon_x$ and $|x_{i+1}| < \epsilon_x$, then $|\nabla_i F(x)| > \epsilon$. Furthermore, for all $x \in \mathbb{R}^{d_x}$, $\|\nabla F(x)\|_\infty \le 23\epsilon$.*

Lemma B.3. *Suppose $\phi(t)$ is as defined in (20):*

$$\phi(t) = \begin{cases} t + \frac{1}{e}\int_{1/2}^{-t}\Psi(\tau)d\tau, & t < -1/2, \\ t, & -1/2 \le t \le 1/2, \\ t - \frac{1}{e}\int_{1/2}^{t}\Psi(\tau)d\tau, & t > 1/2. \end{cases}$$

It satisfies the following properties:

1. *$\phi(t)$ is infinitely differentiable.*

2. *For all* $t \in \mathbb{R}$,

$$|\phi(t)| < 2,\ 0 \le \phi'(t) \le 1,\ |\phi''(t)| \le \sqrt{54/e^3},\ |\phi'''(t)| \le 32.5/e.$$

Proof. Other properties come immediately from Lemma B.1, we only prove $|\phi(t)| < 2$. To see this, for $t \ge 1/2$,

$$\phi(t) = \frac{1}{2} + \int_{1/2}^{t} 1 - \frac{1}{e}\Psi(\tau)d\tau \le 1 + \int_{1}^{t} \frac{1}{(2\tau-1)^2}d\tau \le 2.$$

where in the first inequality, we used $1 - x \le \exp(-x)$. The same holds for $t \le -1/2$. □

Lemma B.4. *Let* $f_i(x), h_i(x)$ *be defined as the following:*

$$\begin{aligned} f_i(x) &:= \Psi_\epsilon(x_{i-1})\Phi_\epsilon(x_i) - \Psi_\epsilon(-x_{i-1})\Phi_\epsilon(-x_i), \\ h_i(x) &:= \Gamma\left(1 - \left(\textstyle\sum_{j=i}^{d_x} \Gamma^2(|x_j|/\epsilon)\right)^{1/2}\right), \end{aligned} \tag{87}$$

where $\Gamma(t)$ *is given by*

$$\Gamma(t) := \frac{\int_{1/4}^{t} \Lambda(\tau)d\tau}{\int_{1/2}^{1/4} \Lambda(\tau)d\tau}, \quad \textit{where} \quad \Lambda(t) = \begin{cases} 0, & t \le 1/4 \textit{ or } t \ge 1/2, \\ \exp\left(-\frac{1}{100(t-1/4)(1/2-t)}\right), & \textit{otherwise} \end{cases}.$$

Then, $h_i(x)$ *satisfies* $\mathbb{1}_{\{i>\texttt{prog}_{\epsilon/4}(x)\}} \le h_i(x) \le \mathbb{1}_{\{i>\texttt{prog}_{\epsilon/2}(x)\}}$, $O(1)$*-Lipschitzness and the following:*

$$f_i^{(k)}(x)h_i(x) = 0, \qquad \forall i \ne \texttt{prog}_{\epsilon/2}(x) + 1, k \in \mathbb{N}_+.$$

Proof. The construction of $h_i(x)$ is brought from (Arjevani et al., 2023) where the boundedness and $O(1)$-Lipschitzness are guaranteed from the proof of Lemma 4 in (Arjevani et al., 2023). The last property follows from the fact that any k^{th} order derivative of $\Psi(t)$ for $t \le 1/2$ is 0, and thus,

$$\Psi_\epsilon^{(k)}(x_{i-1}) = 0, \qquad \forall i > \texttt{prog}_{\epsilon/2}(x) + 1,$$

and

$$h_i(x) \le \mathbb{1}_{\{i>\texttt{prog}_{\epsilon/2}(x)\}} = 0, \qquad \forall i < \texttt{prog}_{\epsilon/2}(x) + 1.$$

□

Lemma B.5 ((Arjevani et al., 2023), Lemma 15)**.** *For all* $x \in \mathbb{R}^d$, $\rho(x)$ *is* 1*-Lipschitz and* $(3/R)$*-smooth, that is,*

$$\|J(x)\| \le 1,\ \|J(x^1) - J(x^2)\| \le \frac{3}{R}\|x^1 - x^2\|, \qquad \forall x^1, x^2 \in \mathbb{R}^d.$$

The following lemma is the key to converting the zero-chain argument to general randomized algorithms (Arjevani et al., 2023):

Lemma B.6 ((Arjevani et al., 2023), General Version of Lemma 5)**.** *Let* $\mathtt{A} \in \mathcal{A}$ *be a randomized algorithm that accesses functions* $f, g : \mathbb{R}^{d_x \times d_y} \to \mathbb{R}$ *through a stochastic oracle, and generates batched queries with norm-bounded* x*, i.e.,* $\|x^{t,n}\| \le R$ *for all* $t \in \mathbb{N}, n \in [N]$*, with some* $R > 0$*. Let* U *be a random matrix uniformly distributed on* $\texttt{Ortho}(d, d_x)$ *with* $d \gtrsim \frac{R^2 N d_x}{p} \log\left(\frac{N d_x^2}{p\delta}\right)$ *and* $p, \delta > 0$*, and let* u_j *be the* j^{th} *column of* U*. Additionally, let* O_U *be an oracle that takes a batched query* $(\boldsymbol{x}, \boldsymbol{y}) := \{(x^n, y^n)\}_{n=1}^N$*, and returns* $G_U(\boldsymbol{x}, \boldsymbol{y}; \xi)$ *where* ξ *is a random variable and* G_U *is the oracle response to* $(\boldsymbol{x}, \boldsymbol{y})$ *and* ξ *parameterized by* U*. Suppose the following holds for* G_U *with probability* 1*:*

$$G_U(\boldsymbol{x}, \boldsymbol{y}; \xi) = G_{U'}(\boldsymbol{x}, \boldsymbol{y}; \xi), \quad \forall U' \in \texttt{Ortho}(d, d_x) : u'_j = u_j \textit{ for all } j = 1, 2, ..., \max_n \texttt{prog}_{1/4}(U^\top x^n) + 1.$$

and with probability at least $1-p$*:*

$$G_U(\boldsymbol{x},\boldsymbol{y};\xi)=G_{U'}(\boldsymbol{x},\boldsymbol{y};\xi),\quad \forall U'\in \texttt{Ortho}(d,d_x):u'_j=u_j \text{ for all } j=1,2,...,\max_n \texttt{prog}_{1/4}(U^\top x^n).$$

When $\texttt{A}$ *is paired with* O_U*, with probability at least* $1-\delta$*,*

$$\max_{n\in[N]}\texttt{prog}_{1/4}\left(U^\top x^{t,n}\right)<d_x,\qquad \forall t<\frac{d_x-\log(1/\delta)}{2p}.$$

Proof. The original proof in (Arjevani et al., 2023) uses the progress of returned gradient estimators directly as the overall progress of an algorithm:

$$\gamma_x^t=\max_{t'\le t,n\in[N]}\left(\texttt{prog}_0(\hat{\nabla}_x g(U^\top x^{t',n},y^{t',n};\xi^{t'})\right).$$

Our definition is a generalization of the above measure to include other signals $\hat{\nabla}_y g$ and $\hat{y}$ in the oracle response. Specifically, we define

$$\gamma^t=\arg\min_{i\in[d_x]}: G_U(\boldsymbol{x}^{t'},\boldsymbol{y}^{t'};\xi^{t'})=G_{U'}(\boldsymbol{x}^{t'},\boldsymbol{y}^{t'};\xi^{t'}),\quad \forall t'\le t,\ \forall U'\in\texttt{Ortho}(d,d_x):u'_j=u_j \text{ for all } j\in[i].$$

With the above definition, under the event $\mathcal{B}^t:=\{\max_n \texttt{prog}_{1/4}(U^\top x^{t,n})\le\gamma^{t-1}\}$ and filtration $\mathcal{G}^{t-1}$:

$$\mathcal{G}^{t-1}:=\sigma(\xi_{\mathrm{A}},(\boldsymbol{x}^0,\boldsymbol{y}^0),G_U(\boldsymbol{x}^0,\boldsymbol{y}^0;\xi^0),...,(\boldsymbol{x}^{t-1},\boldsymbol{y}^{t-1}),G_U(\boldsymbol{x}^{t-1},\boldsymbol{y}^{t-1};\xi^{t-1})),$$

we can check that

$$\begin{aligned}\mathbb{P}(\gamma^t-\gamma^{t-1}\notin\{0,1\},\mathcal{B}^t|\mathcal{G}^{t-1})=0,\\ \mathbb{P}(\gamma^t-\gamma^{t-1}=1,\mathcal{B}^t|\mathcal{G}^{t-1})\le p.\end{aligned}$$

The rest follows the same steps in (Arjevani et al., 2023), hence we refer the readers to Appendix B.1 in the reference. □

B.2. Proof of Lemma 4.3

We first note that $\mathbb{E}[\hat{\nabla}_y g(x,y;\xi)]=\nabla_y g(x,y)$, and

$$\nabla_y g(x,y)=-2\left(y-\epsilon^2\textstyle\sum_{i=1}^{d_x}f_i(x)\right)=-2\left(y-\epsilon^2\sum_{i=1}^{\texttt{prog}_{\epsilon/2}(x)+1}f_i(x)\right).$$

Furthermore, from Lemma B.4, we can observe that

$$\hat{\nabla}_y g(x,y;\xi)-\nabla_y g(x,y)=2\epsilon^2\cdot f_{\texttt{prog}_{\epsilon/2}(x)+1}(x)h_{\texttt{prog}_{\epsilon/2}(x)+1}(x)(\xi/p-1).$$

Since both $f_i(x)$ and $h_i(x)$ are bounded by $O(1)$ for all i and x, we have $\texttt{Var}(\hat{\nabla}_y g(x,y;\xi))\lesssim\epsilon^4/p$. The remaining properties follow straightforwardly from the construction.

B.3. Proof of Lemma 4.4

We start with writing down the explicit formula for $\nabla_x g_b(x,y)$:

$$\nabla_x g_b(x,y)=-r_\epsilon\phi\left(\frac{y-F(x)}{r_\epsilon}\right)\phi'\left(\frac{y-F(x)}{r_\epsilon}\right)\nabla F(x),$$

Thus,

$$\|\nabla_x g_b(x,y)\|_\infty\lesssim r_\epsilon\|\nabla F(x)\|_\infty\lesssim r_\epsilon\epsilon,$$

where we use Lemma B.2 for the last inequality.

Now we show that probabilitic zero-chain property. Recall that we define

$$\hat{\nabla}_{x_i} g(x, y; \xi) = \nabla_{x_i} g_b(x, y) \cdot (1 + h_i(x)(\xi/p - 1)),$$

and note that from Lemma B.4,

$$\nabla_{x_i} F(x) h_i(x) = O(\epsilon) \cdot \nabla_{x_i} (f_{i-1}(x) + f_i(x)) h_i(x) = 0, \quad \forall i \neq \texttt{prog}_{1/2}(x) + 1.$$

Therefore, the probabilistic zero-chain property is satisfied, and $\texttt{Var}\left(\hat{\nabla}_x g(x, y; \xi)\right) \lesssim \frac{\|\nabla_x g_b(x,y)\|_\infty^2}{p}$ which must be less than $O(\sigma^2)$. We also check the stochastic smoothness:

$$\begin{aligned}
\|\hat{\nabla}_x g(x, y^1; \xi) - \hat{\nabla}_x g(x, y^2; \xi)\| &\lesssim \left(\max_y \|\nabla^2_{xy} g_b(x, y)\| + \max_y \|\nabla^2_{xy} g_b(x, y)\|_\infty (\xi/p)\right) \|y^1 - y^2\| \\
&\lesssim \left(\max_x \|F(x)\| + \max_x \|F(x)\|_\infty (\xi/p)\right) \|y^1 - y^2\| \\
&\lesssim (1 + \epsilon\xi/p) \|y^1 - y^2\|.
\end{aligned}$$

Thus, we have $\mathbb{E}[\|\hat{\nabla}_x g(x, y^1; \xi) - \hat{\nabla}_x g(x, y^2; \xi)\|^2] \lesssim \frac{\epsilon^2}{p} \|y^1 - y^2\|^2$, and this must be less than $\tilde{l}^2_{g,1} \|y^1 - y^2\|^2$. Thus, we conclude that $p = \Omega\left(\max(r_\epsilon^2 \epsilon^2/\sigma^2, \epsilon^2/\tilde{l}^2_{g,1})\right)$.

B.4. Proof of Theorem 4.5

We first reiterate the initial value-gap and smoothness properties of the modified hard instance: for any $U \in \texttt{Ortho}(d, d_x)$, let F_U be the hyperobjective constructed with (f_U, g_U) defined in (21), paired with oracle responses defined in (24). Then,

1. $F_U(0) - F_U^* \leq O(1)$.
2. f_U, g_U satisfies Assumption 1, 2 with $O(1)$ smoothness parameters.
3. $\|\nabla F_U(x)\| > O(\epsilon)$ if $\texttt{prog}_{\epsilon/4}(U^\top x) < d_x$.
4. For all $(x, y) \in \mathbb{R}^{d_x \times d_y}$,
$$\mathbb{E}[\|\hat{\nabla}_x g_U(x, y; \xi) - \mathbb{E}[\hat{\nabla}_x g_U(x, y; \xi)]\|^2] \leq O(r_\epsilon \epsilon)^2/p.$$
5. For all $(x, y) \in \mathbb{R}^{d_x \times d_y}$,
$$\mathbb{E}[\|\hat{\nabla}_x g_U(x, y^1; \xi) - \hat{\nabla}_x g_U(x, y^2; \xi)\|^2] \leq O\left(1 + \frac{\epsilon^2}{p}\right) \cdot \|y^1 - y^2\|^2.$$

The first property immediately follows from the fact that $F(0) = F_U(0)$ and $F_U^* \geq F^*$, and the fact that $F^* = O(\epsilon^2 d_x) = O(1)$. Property 2 is trivial given our construction and Lemma B.1. The third property follows from the proof of Lemma 6 in (Arjevani et al., 2023). For rest two properties, note that

$$\mathbb{E}[\hat{\nabla}_x g_U(x, y; \xi)] = J(x)^\top U \hat{\nabla}_x g_b(U^\top \rho(x), y; \xi),$$

and since $\|J(x)\| = O(1)$ by Lemma B.5, we can similarly show the bounded-variance and smoothness as in Lemma 4.4. Finally, we can invoke Lemma B.6, and we get the theorem.

C. Auxiliary Lemmas

In Lemma C.1 below, we obtain convergence rate of the projected gradient descent (PSGD) algorithm with either fixed or diminishing step sizes. These statements are adapted from the analogous statements for unconstrained SGD for strongly convex objectives (Thm. 5.7 in (Garrigos & Gower, 2023) for fixed step sizes and Thm. 4.7 in (Bottou, 2010) for diminishing step sizes).

Lemma C.1 (Convergence rate of PSGD for strongly convex objectives)**.** *Let $f : \mathbb{R}^p \to \mathbb{R}$ be a L-smooth and μ-strongly convex function for some $\mu, L > 0$. Suppose $f^* := \inf_x f(x) > -\infty$ and denote $x^* := \arg\min_x f(x)$. Furthermore, let $\mathcal{B}$ be a convex set containing x^*. Suppose $G(x, \xi)$ is an unbiased stochastic gradient estimator for f, that is, $\mathbb{E}[G(x, \xi)] = \nabla f(x)$ for all $x \in \mathcal{B}$. Further assume that the variance of the gradient estimation error is bounded: $\mathbb{E}[\|G(x, \xi) - \nabla f(x)\|^2] \le \sigma^2$ for all $x \in \mathcal{B}$. Let $\rho := \frac{2\mu L}{\mu + L}$. Consider the following PSGD iterates:*

$$x_{t+1} \leftarrow \Pi_{\mathcal{B}} \left\{ x_t - \alpha_t G(x_t, \xi_t) \right\}. \tag{88}$$

Then the following hold:

(i) *(fixed step size) Suppose $\alpha_t \equiv \alpha < 2L/(\mu + L)$. Fix $T > \frac{4L \log(L/\mu)}{\mu}$. Then for all $0 \le t \le T$,*

$$\mathbb{E}[\|x^t - x^*\|^2] \le (1 - \mu\alpha)^t \|x^0 - x^*\|^2 + \frac{\alpha\sigma^2}{\mu}. \tag{89}$$

Taking $\alpha = \frac{8 \log T}{\mu T}$, we have $\mathbb{E}[\|x^T - x^*\|^2] \le \frac{1}{T^4}\|x^0 - x^*\|^2 + \frac{8 \log T}{\mu^2 T}\sigma^2$.

(ii) *(diminishing step size) Suppose $\alpha_t = \frac{\beta}{\gamma + t}$, where $\beta > 1/\rho$, $\gamma > 0$ are constants. Then for all $t \ge 0$,*

$$\mathbb{E}[\|x_t - x^*\|^2] \le \frac{\nu}{\gamma + t}, \tag{90}$$

where

$$\nu := \max\left\{ \frac{\beta^2\sigma^2 L}{2(\beta\rho - 1)} + \frac{(\gamma + 1)\beta^2\sigma^2}{\gamma^2},\ (\gamma + 1)\left(1 - \frac{2\beta\mu L}{\gamma(\mu + L)}\right)\|x^0 - x^*\|^2 \right\}. \tag{91}$$

Proof. Let $\bar{x}^t := x^t - \alpha G(x^t; \xi_t)$. Then

$$\begin{aligned} \|\bar{x}^t - x^*\|^2 &= \|\bar{x}^t - x^t\|^2 + \|x^t - x^*\|^2 + 2\langle \bar{x}^t - x^t, x^t - x^* \rangle \\ &= \alpha_t^2 \|G(x^t; \xi_t)\|^2 + \|x^t - x^*\|^2 - 2\alpha_t \langle x^t - x^*, G(x^t; \xi_t) \rangle. \end{aligned}$$

Let $\mathcal{F}_t = \sigma(\xi_1, \ldots, \xi_{t-1})$ denote the σ-algebra generated by the random variables $\xi_1, \ldots, \xi_{t-1}$. Then x_t is measurable w.r.t. $\mathcal{F}_t$, so taking conditional expectation with respect to $\mathcal{F}_t$ and using the co-coercivity of strongly convex functions, we have

$$\mathbb{E}[\|\bar{x}^t - x^*\|^2 \,|\, \mathcal{F}_t] = \alpha_t^2\sigma^2 + \alpha_t^2 \|\nabla f(x^t)\|^2 + \|x^t - x^*\|^2 - 2\alpha_t \underbrace{\langle x^t - x^*, \nabla f(x^t) \rangle}_{\ge \frac{\mu L}{\mu + L}\|x^t - x^*\|^2 + \frac{1}{\mu + L}\|\nabla f(x^t)\|^2} \tag{92}$$

$$\le \left(1 - \frac{2\alpha_t \mu L}{\mu + L}\right)\|x^t - x^*\|^2 + \alpha_t^2\sigma^2, \tag{93}$$

given that $\alpha_t \le 2/(\mu + L)$. Then by the projection lemma and taking the full expectation, we have

$$\mathbb{E}[\|x^{t+1} - x^*\|^2] \le \mathbb{E}[\|\bar{x}^t - x^*\|^2] \le (1 - \alpha_t\rho)\,\mathbb{E}[\|x^t - x^*\|^2] + \alpha_t^2\sigma^2,$$

where we have denoted $\rho := \frac{2\mu L}{\mu + L}$.

To derive **(i)**, suppose $\alpha_t \equiv \alpha < 2L/(\mu + L)$. Then applying (92) recursively, we have

$$\begin{aligned} \mathbb{E}[\|x^t - x^*\|^2] &\le (1 - \mu\alpha)^t \|x^0 - x^*\|^2 + 2\sigma^2 \sum_{j=0}^{t} (1 - \alpha\mu)^j \alpha^2 \\ &\le (1 - \mu\alpha)^t \|x^0 - x^*\|^2 + \frac{\alpha\sigma^2}{\mu}. \end{aligned}$$

Next, we show **(ii)** by induction. We will show that, for all $t \ge 1$, that

$$\mathbb{E}[\|x_t - x^*\|^2] \le \frac{\tilde{\nu}}{\gamma + t}, \tag{94}$$

where

$$\tilde{\nu} := \max\left\{\frac{\beta^2\sigma^2 L}{2(\beta\rho-1)},\, (\gamma+1)\mathbb{E}[\|x_1 - x^*\|^2]\right\}. \tag{95}$$

This is enough to conclude since $\tilde{\nu} \le \nu$, which follows by (92):

$$\mathbb{E}[\|\bar{x}^1 - x^*\|^2] \le \left(1 - \frac{2\beta\mu L}{\gamma(\mu+L)}\right)\|x^0 - x^*\|^2 + \frac{\beta^2}{\gamma^2}\sigma^2. \tag{96}$$

Indeed, (94) holds for $t=1$ by the choice of $\tilde{\nu}$. Denoting $\hat{t} := \gamma + t$, by the induction hypothesis and by the choices of step size α_t and $\tilde{\nu}$,

$$\mathbb{E}[\|x_{t+1} - x^*\|^2] \le \left(1 - \frac{\beta\rho}{\hat{t}}\right)\frac{\tilde{\nu}}{\hat{t}} + \frac{L\sigma^2\beta^2}{2\hat{t}^2} \tag{97}$$

$$\le \left(\frac{\hat{t}-1}{\hat{t}^2}\right)\tilde{\nu} \underbrace{- \left(\frac{\beta\rho - 1}{\hat{t}^2}\right)\tilde{\nu} + \frac{L\sigma^2\beta^2}{2\hat{t}^2}}_{\le 0} \tag{98}$$

$$\le \frac{\hat{t}-1}{(\hat{t}-1)(\hat{t}+1)}\tilde{\nu} \le \frac{\tilde{\nu}}{\gamma+1+t}. \tag{99}$$

This shows the assertion. □

Lemma C.2. *$\mathcal{L}_\lambda^*(x)$ is L-smooth with $L := \frac{6l_{g,1}}{\mu_g}\left(l_{f,1} + \frac{l_{g,1}^2}{\mu_g} + \frac{l_{f,0}l_{g,1}l_{g,2}}{\mu_g^2}\right)$.*

Proof. See Lemma B.8 in (Chen et al., 2023a). □

Lemma C.3. *For $\lambda \ge 2l_{f,1}/\mu_g$, $\mathcal{L}_\lambda(x,\cdot)$ is $(\lambda\mu_g/2)$-strongly convex for each $x \in \mathbb{R}^{d_x}$.*

Proof. Since $f(x,\cdot)$ is $l_{f,1}$-smooth and $g(x,\cdot)$ is μ_g-strongly convex,

$$\mathcal{L}_\lambda(x,y') - \mathcal{L}_\lambda(x,y) \ge \langle \nabla_y \mathcal{L}_\lambda(x,y),\, y'-y\rangle + \frac{\lambda\mu_g - l_{f,1}}{2}\|y'-y\|^2. \tag{100}$$

Hence if $\lambda\mu_g \ge 2l_{f,1}$, then $\mathcal{L}_\lambda(x,\cdot)$ is $(\lambda\mu_g/2)$-strongly convex. □

Lemma C.4. *For $\lambda \ge 2l_{f,1}/\mu_g$, it holds that for each $x \in \mathbb{R}^{d_x}$,*

$$\|y_\lambda^*(x) - y^*(x)\| \le \frac{2l_{f,0}}{\lambda\mu_g} = O(\lambda^{-1}). \tag{101}$$

Proof. Note that if a function $w \mapsto h(w)$ is μ-strongly convex and is minimized at w^*, then for any w,

$$0 \ge h(w^*) - h(w) \ge \langle \nabla h(w),\, w^* - w\rangle + \frac{\mu}{2}\|w^* - w\|^2. \tag{102}$$

Using Cauchy-Schwarz inequality, the above inequality implies

$$\|w^* - w\| \le \frac{2}{\mu}\|\nabla h(w)\|. \tag{103}$$

Apply the above inequality for $h(y) = g(x,y)$ and the first-order optimality condition for $y_\lambda^*(x)$ in (27) to get

$$\|y^*(x) - y_\lambda^*(x)\| \le \frac{2}{\mu_g}\|\nabla_y g(x, y_\lambda^*(x))\| = \frac{2}{\lambda\mu_g}\|\nabla_y f(x, y_\lambda^*(x))\| \le \frac{2l_{f,0}}{\lambda\mu_g}, \tag{104}$$

as desired. □

Lemma C.5. *For $\lambda \geq 2l_{f,1}/\mu_g$, it holds that*

$$|\mathcal{L}_\lambda^*(x) - F(x)| \leq D_0 \lambda^{-1} = O(\lambda^{-1}), \tag{105}$$

where $D_0 := \left(l_{f,1} + \frac{l_{f,1}^2}{\mu_g}\right) \frac{l_{f,1}}{\mu_g}$.

Proof. See Lemma B.3 in (Chen et al., 2023a). □