On the Convergence and Complexity of Proximal Gradient and Accelerated Proximal Gradient Methods under Adaptive Gradient Estimation

Raghu Bollapragada[†‡] Shagun Gupta[†]

July 19, 2025

Abstract

In this paper, we propose a proximal gradient method and an accelerated proximal gradient method for solving composite optimization problems, where the objective function is the sum of a smooth and a convex, possibly nonsmooth, function. We consider settings where the smooth component is either a finite-sum function or an expectation of a stochastic function, making it computationally expensive or impractical to evaluate its gradient. To address this, we utilize gradient estimates within the proximal gradient framework. Our methods dynamically adjust the accuracy of these estimates, increasing it as the iterates approach a solution, thereby enabling high-precision solutions with minimal computational cost. We analyze the methods when the smooth component is nonconvex, convex, or strongly convex, using a biased gradient estimate. In all cases, the methods achieve the optimal iteration complexity for first-order methods. When the gradient estimate is unbiased, we further refine the analysis to show that the methods simultaneously achieve optimal iteration complexity and optimal complexity in terms of the number of stochastic gradient evaluations. Finally, we validate our theoretical results through numerical experiments.

1 Introduction

In this paper, we consider composite optimization problems of the form

$$\min_{x\in\mathbb{R}^d} \ \phi(x) = f(x) + h(x), \tag{1.1}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is a continuously differentiable function and $h : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is a closed, convex, proper, and possibly nonsmooth function. We focus on settings where $h(x)$ admits a simple structure that enables efficient computation of the proximal operator,

$$\text{prox}_{\alpha,h}(y) = \underset{x\in\mathbb{R}^d}{\text{argmin}} \ h(x) + \frac{1}{2\alpha}\|x - y\|^2, \quad \text{with } \alpha > 0, \tag{1.2}$$

[†]Operations Research and Industrial Engineering Program, University of Texas at Austin. (raghu.bollapragada@utexas.edu,shagungupta@utexas.edu)

[‡]Corresponding author.

which allows for the use of proximal gradient methods [5, 14] to efficiently solve (1.1). Examples of such $h(x)$ include the l_1-norm penalty and the indicator function of a set to enforce simple convex constraints such as box constraints, norm-ball constraints, or boundary conditions. This class of problems has been extensively studied due to its wide range of applications, including image processing [15, 18], data science [13, 22], and inverse problems [2, 41]. When $f(x)$ is a convex function, Nesterov's acceleration [30] can be applied to proximal gradient methods to achieve improved convergence rates, as shown in [3, 29, 36, 37].

We analyze problems where the smooth component $f(x)$ takes one of the following forms:

$$f(x) = \frac{1}{N}\sum_{i=1}^{N} F(x, \xi_i) \qquad (1.3), \qquad \text{or} \qquad f(x) = \mathbb{E}[F(x, \xi)], \qquad (1.4)$$

where (1.3) is a finite-sum function resulting in a finite-sum problem over the dataset $\mathcal{S} = \{\xi_1, \xi_2, \ldots, \xi_N\}$ with $F : \mathbb{R}^d \times \mathcal{S} \to \mathbb{R}$, and (1.4) is an expectation function resulting in an expectation problem over the random variable ξ with the associated probability space $(\Xi, \Omega, \mathcal{P})$, $F : \mathbb{R}^d \times \Xi \to \mathbb{R}$ and $\mathbb{E}[\cdot]$ denotes the expectation with respect to $\mathcal{P}$. Since computing the exact gradient of $f(x)$ in these settings is computationally prohibitive, proximal gradient methods rely on gradient estimates [4, 17, 19, 23, 24, 32, 34, 43]. The main computational costs in such methods are: (1) the number of proximal operator evaluations (i.e., iterations) and, (2) the number of stochastic gradient evaluations. Most existing methods use unbiased gradient estimates of fixed accuracy [17, 19, 24, 34], typically achieving optimal performance with respect to only one of the two costs. Adaptive gradient estimation methods, by contrast, dynamically adjust the accuracy of the gradient estimate based on the quality of the current iterate. By using low-accuracy estimates in the early stages and gradually increasing the accuracy of the estimates, they can achieve high-accuracy solutions with minimal computational effort. To this end, we propose proximal gradient methods with adaptive gradient estimation to optimize both cost metrics. Furthermore, in many practical scenarios, the bias in the gradient estimate is intrinsic and cannot be fully eliminated, such as federated learning with non-IID data [25] and Bayesian optimization using surrogate models [38].

As a result, we propose and analyze a proximal gradient method and an accelerated proximal gradient method for the finite-sum problem (1.3) and the expectation problem (1.4) that adaptively control the accuracy of the estimate and allow for biased gradient estimates. These methods achieve the optimal iteration complexity for first-order methods when the objective function is nonconvex, convex and strongly convex while using biased gradient estimates. When the gradient estimate is unbiased, the methods additionally attain the optimal complexity for the number of stochastic gradient evaluations in all three settings. A summary of these complexity results is provided in Table 1.

1.1 Literature Review

Proximal gradient methods for deterministic composite optimization problems are well studied; see [5, 14] and references therein. Accelerated variants for deterministic convex problems, such as Nesterov's acceleration [29, 37], the fast iterative shrinkage-thresholding algorithm (FISTA) [3], and its backtracking extension [36], are also well established in the literature. Several works [7, 35, 37, 39] have analyzed proximal and accelerated proximal gradient methods with inexact gradients, where gradient accuracy is controlled through a predetermined deterministic sequence, in settings similar to the finite-sum problem (1.3). In contrast, our work employs Nesterov's acceleration and

adaptively adjusts the accuracy of the gradient estimates based on the current iterate for both the finite-sum (1.3) and expectation (1.4) problems.

Many proximal and accelerated proximal gradient methods have been proposed that employ unbiased stochastic gradient estimates for the finite-sum problem (1.3) and the expectation problem (1.4). In [19,23,24], the authors proposed accelerated methods that achieve the optimal complexity in terms of the number of stochastic gradient evaluations for the expectation problem (1.4) when the objective function is convex or strongly convex. In [17], accelerated proximal gradient methods using Nesterov's acceleration were analyzed, establishing optimal complexity results for the number of proximal operator evaluations and the number of stochastic gradient evaluations for the expectation problem when the objective function is nonconvex. In [20, 26, 34], the authors employ accelerated proximal gradient methods with variance reduction techniques and achieve the optimal complexity in terms of stochastic gradient evaluations (in expectation) for the finite-sum problem (1.3). In [32], the authors extend the FISTA based method [36] to the expectation problem (1.4) while allowing for biased gradient estimates. The method replaced the backtracking line search in [36] with a step search mechanism to adaptively determine the step size, and assumed control over the error in the gradient estimate in probability, unlike the other works discussed above, which control the expected error in the gradient estimate. Although this method establishes stronger convergence guarantees than convergence in expectation under its assumptions, the analysis is limited to general convex objectives and results in suboptimal complexity for the number of stochastic gradient evaluations.

In [4, 43], the authors employed unbiased gradient estimates in proximal gradient methods and used adaptive sampling strategies to control the accuracy of the gradient estimate for the expectation problem (1.4). They established theoretical convergence guarantees for various objective functions but did not provide complexity results for the number of stochastic gradient evaluations. We adopt similar conditions to [43] to control the accuracy of the gradient estimate, while allowing for biased estimates and extending the approach to accelerated proximal gradient methods, achieving optimal iteration complexity for first-order methods [9, 31]. When unbiased gradient estimates are available via sample average approximations, our conditions guide sample size selection in a way similar to [4, 43], and also yield the optimal complexity for the number of stochastic gradient evaluations for the expectation problem (1.4) for first-order methods [1, 16, 17, 24]. These results are summarized in Table 1. Similar complexity guarantees for the finite-sum problem (1.3) can also be established following the same procedure and are omitted for brevity.

Lastly, while our methods assume exact solutions to the proximal operator, several works have explored algorithms with inexact proximal updates [7,37]. Additionally, approaches that go beyond the structured nonsmoothness in problem (1.1) have been investigated in [42] for constrained settings, and in [21,27,44] where smoothing techniques are employed to develop zeroth-order methods.

1.2 Contributions

We summarize our main contributions as follows:

1. We propose a proximal gradient method and an accelerated proximal gradient method for both the finite-sum problem (1.3) and the expectation problem (1.4). These methods employ gradient estimates whose accuracy is adaptively controlled using deterministic and stochastic generalizations of the well-known "norm condition" [6].

2. We show that the proposed methods achieve the optimal complexity in terms of the number of proximal operator evaluations for first-order methods, even when using biased gradient

Table 1: Summary of the best-established complexity results in this paper.

$f(x)$	Proximal Gradient		Accelerated Proximal Gradient	
	Proximal Operators	Gradients (Unbiased)	Proximal Operators	Gradients (Unbiased)
Nonconvex	$\mathcal{O}\left(\frac{1}{\epsilon}\right)$	$\mathcal{O}\left(\frac{1}{\epsilon^2}\right)$	-	-
Convex	$\mathcal{O}\left(\frac{1}{\epsilon}\right)$	$\mathcal{O}\left(\frac{1}{\epsilon^2}\right)$	$\mathcal{O}\left(\frac{1}{\sqrt{\epsilon}}\right)$	$\mathcal{O}\left(\frac{1}{\epsilon^{1.5}}\right)$
Strongly Convex	$\mathcal{O}\left(\kappa \log \frac{1}{\epsilon}\right)$	$\mathcal{O}\left(\frac{\kappa}{\epsilon}\right)$	$\mathcal{O}\left(\sqrt{\kappa} \log \frac{1}{\epsilon}\right)$	$\mathcal{O}\left(\frac{\sqrt{\kappa}}{\epsilon}\right)$

Note: The solution accuracy ϵ depends on the nature of the objective function: for nonconvex functions, it refers to the squared norm of the gradient; for convex and strongly convex functions, it corresponds to the optimality gap in function value. Here, κ denotes the condition number.

estimates, for nonconvex, convex, and strongly convex objective functions in both the finite-sum and expectation problems; see Table 1.

3. We further show that, when the gradient estimates are unbiased, the methods simultaneously achieve the optimal complexity for the number of stochastic gradient evaluations for the expectation problem (1.4) for first-order methods, across nonconvex, convex and strongly convex objective functions; see Table 1.

1.3 Paper Organization

The paper is organized as follows. In Section 2, we describe the proposed algorithm, the adaptive conditions used to control the accuracy of the gradient estimate, and the preliminary assumptions. In Section 3, we present the theoretical analysis, covering the nonconvex case in Subsection 3.1, the convex case in Subsection 3.2, and the strongly convex case in Subsection 3.3. We illustrate the empirical performance of the proposed algorithm in Section 4 and provide concluding remarks in Section 5.

1.4 Notation

Let $\mathbb{R}$ denote the set of real numbers and $\mathbb{R}^d$ denote the set of d dimensional real vectors. Unless otherwise specified, $|\cdot|$ denotes the Euclidean norm of a vector, and $|\cdot|$ denotes either the absolute value of a real number or the cardinality of a set, depending on context. The ceiling function is denoted by $\lceil \cdot \rceil$. Expectation and variance with respect to the distribution $\mathcal{P}$ are denoted as $\mathbb{E}[\cdot]$ and $\mathrm{Var}[\cdot]$, respectively. We denote the optimal value for problem (1.1) as ϕ^*.

2 Proposed Algorithm

In this section, we present the preliminary assumptions and describe the proposed (accelerated) proximal gradient method, along with the conditions used to adaptively control the accuracy of the gradient estimates. For the case of unbiased gradient estimates, we also provide a sequence of sample average approximations that satisfy these conditions.

We begin with the following assumption about the objective function.

Assumption 2.1. *The function $f : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable and has L-Lipschitz continuous gradients (i.e., f is L-smooth). The function $h : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is closed, convex, and proper.*

A well-known result for functions with Lipschitz continuous gradients (see [31]) is

$$f(a) \leq f(b) + \nabla f(a)^T (b - a) + \tfrac{L}{2}\|b - a\|^2 \quad \forall a, b \in \mathbb{R}^d. \tag{2.1}$$

Under Assumption 2.1, since $h(x)$ is a convex function, computing the proximal operator (1.2) is well-defined. Specifically, it corresponds to the unique solution of a strongly convex optimization problem.

We now describe the proposed algorithm. At each iteration $k \geq 0$, the algorithm maintains two iterates: the decision variable x_k and an auxiliary variable y_k. It computes g_k, an estimate of $\nabla f(y_k)$, which may be biased. The accuracy of this estimate is controlled through adaptive conditions described later in this section. The next iterate x_{k+1} is obtained by taking a step from y_k in the direction of g_k, followed by evaluating the proximal operator (1.2), i.e. $x_{k+1} = \text{prox}_{\alpha_k, h}(y_k - \alpha_k g_k)$, where $\alpha_k > 0$ is the step size. Under Assumption 2.1, this is equivalent to

$$x_{k+1} = \underset{x \in \mathbb{R}^d}{\text{argmin}} \; f(y_k) + g_k^T (x - y_k) + \tfrac{1}{2\alpha_k}\|x - y_k\|^2 + h(x), \quad \text{with } \alpha_k \in \left(0, \tfrac{1}{L}\right]. \tag{2.2}$$

The auxiliary variable is then updated in one of two ways. Under the proximal gradient method (referred to as **Option I**), it is set as $y_{k+1} = x_{k+1}$ and under the accelerated proximal gradient method (**Option II**), it is updated using the rule $y_{k+1} = x_{k+1} + \beta_{k+1}(x_{k+1} - x_k)$, where $\{\beta_k\}$ is a user-defined sequence. The complete procedure is summarized in Algorithm 2.1 and the following remark.

Algorithm 2.1 (Accelerated) Proximal Gradient Method with Adaptive Gradient Estimation

Inputs : Initial iterate x_0, initial auxiliary iterate $y_0 = x_0$, step size sequence $\{\alpha_k\}$, and acceleration parameter sequence $\{\beta_k\}$.

1: **for** k = 0, 1, 2, ... **do**
2: Compute a gradient estimate g_k of $\nabla f(y_k)$
3: Proximal step: $x_{k+1} = \text{prox}_{\alpha_k, h}(y_k - \alpha_k g_k)$
4: **Option I:** Set $y_{k+1} = x_{k+1}$
5: **Option II:** Set $y_{k+1} = x_{k+1} + \beta_{k+1}(x_{k+1} - x_k)$
6: **end for**

Remark 2.1. *We make the following remarks regarding Algorithm 2.1.*

- ***Gradient Estimate (Line 2):*** *At each iteration, the algorithm computes a possibly biased gradient estimate g_k for $\nabla f(y_k)$. The accuracy of this estimate is adaptively controlled to ensure sufficient progress and is described below.*

- ***Proximal Step (Line 3):*** *The proximal step computes the proximal operator defined in (1.2). We assume that the function $h(x)$ is simple enough such that the proximal operator can be evaluated efficiently.*
- ***Acceleration Option (Lines 4-5):*** *The iterate update y_{k+1} follows either* ***Option I****, corresponding to the standard proximal gradient method, or* ***Option II****, which incorporates acceleration based on [29, 37], where $\{\beta_k\}$ is a predetermined user-defined sequence.*

We now introduce some quantities to describe the conditions controlling the accuracy of the gradient estimate. The reduced gradient at iteration $k \geq 0$ is defined as

$$R_{\alpha_k}(y_k) = \tfrac{1}{\alpha_k}\left(y_k - x_{k+1}\right). \tag{2.3}$$

The true step computed using $\nabla f(y_k)$ and the corresponding true reduced gradient at iteration $k \geq 0$ are defined as

$$\hat{x}_{k+1} = \operatorname{prox}_{\alpha_k,h}(y_k - \alpha_k \nabla f(y_k)) \quad \text{and} \quad R^{true}_{\alpha_k}(y_k) = \tfrac{1}{\alpha_k}\left(y_k - \hat{x}_{k+1}\right), \tag{2.4}$$

respectively. If $\left\|R^{true}_{\alpha_k}(y_k)\right\| = 0$, then y_k is a stationary point for $\phi(x)$ [4]. The error in the reduced gradient can be bounded in terms of the gradient estimation error using the contraction property of the proximal operator as follows:

$$\begin{aligned}\|R_\alpha(y_k) - R^{true}_\alpha(y_k)\| &= \tfrac{1}{\alpha_k}\|\operatorname{prox}_{\alpha_k,h}(y_k - \alpha_k g_k) - \operatorname{prox}_{\alpha_k,h}(y_k - \alpha_k \nabla f(y_k))\| \\ &\leq \|g_k - \nabla f(y_k)\|. \end{aligned} \tag{2.5}$$

We also define a nested sequence of σ-algebras $\{\mathcal{G}_k\}$ where $\mathcal{G}_0 = \{x_0\}$ and $\mathcal{G}_k = \{x_0, g_0, g_1, \ldots, g_{k-1}\}$ $\forall k \geq 1$. Hence, both x_k and y_k are specified under $\mathcal{G}_k$. We denote the conditional expectation given $\mathcal{G}_k$ as $\mathbb{E}_k[\cdot] = \mathbb{E}[\cdot|\mathcal{G}_k]$ and the total expectation, (i.e., the expectation given the initial conditions) as $\mathbb{E}[\cdot] = \mathbb{E}[\cdot|\mathcal{G}_0]$.

We next introduce the conditions controlling the accuracy of the gradient estimate that guarantee fast convergence, analogous to deterministic methods. We adapt the conditions proposed in [43] for proximal gradient methods, originally based on the well-known *norm condition* [6, 8, 11, 33] for smooth optimization, to our (accelerated) proximal setting, and further introduce relaxations to improve efficiency. The proposed conditions are presented below.

Condition 2.1. *The gradient estimate g_k in Algorithm 2.1, $\forall k \geq 0$, is chosen such that:*

1. *For the finite-sum problem (1.3): With constants $\eta_k \in [0, 1)$ and $\iota_0, \delta_k \geq 0$,*

$$\|g_k - \nabla f(y_k)\| \leq \tfrac{\eta_k}{2}\|R_{\alpha_k}(y_k)\| + \iota_0\delta_k.$$

2. *For the expectation problem (1.4): With constants $\tilde{\eta}_k \in [0, 1)$ and $\tilde{\iota}_0, \tilde{\delta}_k \geq 0$,*

$$\mathbb{E}_k\left[\|g_k - \nabla f(y_k)\|^2\right] \leq \tfrac{\tilde{\eta}_k^2}{4}\left\|\mathbb{E}_k\left[R_{\alpha_k}(y_k)\right]\right\|^2 + \tilde{\iota}_0^2\tilde{\delta}_k^2.$$

Condition 2.1, while utilizing sampled quantities on the right-hand side of the inequality, controls the accuracy of the gradient estimate using the true reduced gradient (2.4) as the optimality measure, as shown in Lemma B.1 using (2.5). To relax restrictions on the gradient estimation error,

we introduce the sequences δ_k and $\tilde{\delta}_k$ in addition to the optimality measures on the right-hand side of Condition 2.1; these are appropriately chosen to ensure good performance of the proposed algorithms. We show that Algorithm 2.1 achieves the optimal complexity for the number of proximal operator evaluations when employing a biased gradient estimate that satisfies Condition 2.1. When the gradient estimate is unbiased, we construct sample average approximations, where Condition 2.1 governs the sample size to ensure fast convergence. To determine the complexity of the number of stochastic gradient evaluations in this case, we introduce an additional assumption regarding the variance (or error) of the stochastic gradients. We then present a lemma that constructs a sequence of unbiased gradient estimates satisfying Condition 2.1.

Assumption 2.2. *For any run of Algorithm 2.1, $\exists \sigma \geq 0$ such that $\forall k \geq 0$,*

1. *For the finite-sum problem (1.3): $\|\nabla F(y_k, \xi_i) - \nabla f(y_k)\|^2 \leq \sigma^2$, $\forall i = 1, 2, \ldots, N$.*
2. *For the expectation problem (1.4): $\mathrm{Var}\left[\nabla F(y_k, \xi) | \mathcal{G}_k\right] \leq \sigma^2$.*

Assumption 2.2 is frequently employed to characterize the complexity of the number of stochastic gradient evaluations, as done for finite-sum problems in [20, 26, 34] and for expectation problems in [4, 16, 17, 24]. Since Algorithm 2.1 performs a proximal step over the closed and proper function $h(x)$ each iteration, it is reasonable to expect a bounded deviation of the components gradients for the finite-sum problem (1.3) and the variance to remain bounded for the expectation problem (1.4), over the set of iterates. We now present a sequence of unbiased gradient estimates in the form of sample average approximations that satisfy Condition 2.1 for the expectation problem (1.4).

Lemma 2.2. *Suppose Assumptions 2.1 and 2.2 hold in Algorithm 2.1 for the expectation problem (1.4). Let $g_k = \frac{1}{|S_k|} \sum_{\xi \in S_k} \nabla F(y_k, \xi)$, where S_k is a set of i.i.d. samples from $\mathcal{P}$, independent of $\mathcal{G}_k$, $\forall k \geq 0$. Then, Condition 2.1 is satisfied $\forall k \geq 0$ if*

$$|S_k| = \left\lceil \frac{\sigma^2}{\frac{\tilde{\eta}_k^2}{4} \left\| \mathbb{E}_k \left[R_{\alpha_k}(y_k) \right] \right\|^2 + \iota_0^2 \tilde{\delta}_k^2} \right\rceil. \tag{2.6}$$

Proof. In iteration $k \geq 0$, from Assumption 2.2 and the definition of g_k, the gradient error can be bounded as,

$$\mathbb{E}_k \left[\|g_k - \nabla f(y_k)\|^2 \right] = \frac{\mathrm{Var}[\nabla F(y_k, \xi) | \mathcal{G}_k]}{|S_k|} \leq \frac{\sigma^2}{|S_k|}.$$

From (2.6), we have $|S_k| \geq \frac{\sigma^2}{\frac{\tilde{\eta}_k^2}{4} \left\| \mathbb{E}_k \left[R_{\alpha_k}(y_k) \right] \right\|^2 + \iota_0^2 \tilde{\delta}_k^2}$. Therefore, the gradient error is bounded as,

$$\mathbb{E}_k \left[\|g_k - \nabla f(y_k)\|^2 \right] \leq \frac{\sigma^2}{|S_k|} \leq \frac{\tilde{\eta}_k^2}{4} \left\| \mathbb{E}_k \left[R_{\alpha_k}(y_k) \right] \right\|^2 + \iota_0^2 \tilde{\delta}_k^2,$$

satisfying Condition 2.1. □

Similar to Lemma 2.2, for the finite-sum problem (1.3) under Assumption 2.2, at iteration $k \geq 0$, if $g_k = \frac{1}{|S_k|} \sum_{\xi \in S_k} \nabla F(y_k, \xi)$, where $S_k \subseteq \mathcal{S}$, Condition 2.1 is satisfied if

$$|S_k| = \left\lceil \frac{N}{\left(1 + \frac{\eta_k \left\| R_{\alpha_k}(y_k) \right\| + 2\iota_0 \delta_k}{2\sigma} \right)} \right\rceil,$$

as shown in Lemma B.2. While we only present the complexity of the number of stochastic gradient evaluations for the expectation problem (1.4) using unbiased gradient estimates, results for the finite-sum problem (1.3) can be established through a similar procedure.

3 Theoretical Analysis

In this section, we present the theoretical results for Algorithm 2.1. We begin with some preliminary results, followed by the analysis for three classes of objective functions: nonconvex (Subsection 3.1), general convex (Subsection 3.2), and strongly convex (Subsection 3.3). For each class, we consider both the finite-sum problem (1.3) and the expectation problem (1.4), and derive the iteration complexity, i.e., the number of proximal operator evaluations required to obtain an $\epsilon > 0$ accurate solution when using biased gradient estimates. We then refine these results for the expectation problem when the gradient estimate is unbiased and provide the complexity in terms of the number of stochastic gradient evaluations. All results naturally extend to smooth unconstrained optimization as a special case of composite optimization with $h(x) = 0$.

We begin with a technical descent lemma for problem (1.1) under Assumption 2.1.

Lemma 3.1. *Suppose Assumption 2.1 holds. Then, $\forall x \in \mathbb{R}^d$ and $\forall k \geq 0$, the iterates generated by Algorithm 2.1 satisfy,*

$$\begin{aligned}\phi(x_{k+1}) \leq f(y_k) + \nabla f(y_k)^T(x - y_k) + (g_k - \nabla f(y_k))^T(x - y_k) + \tfrac{1}{2\alpha_k}\|x - y_k\|^2 + h(x) \\ + (\nabla f(y_k) - g_k)^T(x_{k+1} - y_k) - \left(\tfrac{1}{2\alpha_k} - \tfrac{L}{2}\right)\|x_{k+1} - y_k\|^2.\end{aligned}$$

Proof. From (2.1),

$$\begin{aligned}\phi(x_{k+1}) &\leq f(y_k) + \nabla f(y_k)^T(x_{k+1} - y_k) + \tfrac{L}{2}\|x_{k+1} - y_k\|^2 + h(x_{k+1}) \\ &= f(y_k) + g_k^T(x_{k+1} - y_k) + \tfrac{1}{2\alpha_k}\|x_{k+1} - y_k\|^2 + h(x_{k+1}) \\ &\quad + (\nabla f(y_k) - g_k)^T(x_{k+1} - y_k) - \left(\tfrac{1}{2\alpha_k} - \tfrac{L}{2}\right)\|x_{k+1} - y_k\|^2 \\ &\leq f(y_k) + g_k^T(x - y_k) + \tfrac{1}{2\alpha_k}\|x - y_k\|^2 + h(x) \\ &\quad + (\nabla f(y_k) - g_k)^T(x_{k+1} - y_k) - \left(\tfrac{1}{2\alpha_k} - \tfrac{L}{2}\right)\|x_{k+1} - y_k\|^2,\end{aligned}$$

where the last inequality follows $\forall x \in \mathbb{R}^d$ from (2.2). Adding and subtracting $\nabla f(y_k)^T(x - y_k)$ on the right-hand side completes the proof. □

Next, we present a collection of inequalities for the gradient error under Condition 2.1, which will be frequently used in the subsequent analysis.

Lemma 3.2. *Suppose Assumption 2.1 holds and the gradient estimate g_k satisfies Condition 2.1. Then, $\forall k \geq 0$ in Algorithm 2.1:*

1. *For the finite-sum problem* (1.3)*:*

$$\|g_k - \nabla f(y_k)\|^2 \leq \tfrac{\eta_k^2}{2} \|R_{\alpha_k}(y_k)\|^2 + 2\iota_0^2\delta_k^2, \tag{3.1}$$

$$(\nabla f(y_k) - g_k)^T(x_{k+1} - y_k) \leq \tfrac{\alpha_k\eta_k}{2}\|R_{\alpha_k}(y_k)\|^2 + \alpha_k\iota_0\delta_k\|R_{\alpha_k}(y_k)\|, \tag{3.2}$$

$$(\nabla f(y_k) - g_k)^T(x_{k+1} - y_k) \leq \tfrac{\alpha_k\left(\eta_k+\iota_0^2\right)}{2}\|R_{\alpha_k}(y_k)\|^2 + \tfrac{\alpha_k\delta_k^2}{2}, \tag{3.3}$$

$$\|R_{\alpha_k}^{true}(y_k)\|^2 \leq 2\left(1 + \tfrac{\eta_k}{2}\right)^2 \|R_{\alpha_k}(y_k)\|^2 + 2\iota_0^2\delta_k^2. \tag{3.4}$$

2. *For the expectation problem* (1.4)*:*

$$\mathbb{E}_k[\|g_k - \nabla f(y_k)\|] \leq \tfrac{\tilde{\eta}_k}{2}\sqrt{\mathbb{E}_k\left[\|R_{\alpha_k}(y_k)\|^2\right]} + \tilde{\iota}_0\tilde{\delta}_k, \tag{3.5}$$

$$\mathbb{E}_k[(\nabla f(y_k) - g_k)^T(x_{k+1} - y_k)] \leq \tfrac{\alpha_k\tilde{\eta}_k}{2}\mathbb{E}_k\left[\|R_{\alpha_k}(y_k)\|^2\right] + \alpha_k\tilde{\iota}_0\tilde{\delta}_k\sqrt{\mathbb{E}_k\left[\|R_{\alpha_k}(y_k)\|^2\right]}, \tag{3.6}$$

$$\mathbb{E}_k[(\nabla f(y_k) - g_k)^T(x_{k+1} - y_k)] \leq \tfrac{\alpha_k\left(\tilde{\eta}_k+\tilde{\iota}_0^2\right)}{2}\mathbb{E}_k\left[\|R_{\alpha_k}(y_k)\|^2\right] + \tfrac{\alpha_k\tilde{\delta}_k^2}{2}, \tag{3.7}$$

$$\left\|R_{\alpha_k}^{true}(y_k)\right\|^2 \leq 2\left(1 + \tfrac{\tilde{\eta}_k^2}{4}\right)\|\mathbb{E}_k\left[R_{\alpha_k}(y_k)\right]\|^2 + 2\tilde{\iota}_0^2\tilde{\delta}_k^2. \tag{3.8}$$

Proof. For the finite-sum problem (1.3), from Condition 2.1,

$$\|g_k - \nabla f(y_k)\|^2 \leq \left(\tfrac{\eta_k}{2}\|R_{\alpha_k}(y_k)\| + \iota_0\delta_k\right)^2 \leq \tfrac{\eta_k^2}{2}\|R_{\alpha_k}(y_k)\|^2 + 2\iota_0^2\delta_k^2,$$

where the last inequality follows from the identity $(a+b)^2 \leq 2a^2 + 2b^2$, completing the proof of (3.1). For (3.2), by the Cauchy-Schwartz inequality,

$$\begin{aligned}(\nabla f(y_k) - g_k)^T(x_{k+1} - y_k) &\leq \|\nabla f(y_k) - g_k\|\|x_{k+1} - y_k\| = \alpha_k\|\nabla f(y_k) - g_k\|\|R_{\alpha_k}(y_k)\| \\ &\leq \tfrac{\alpha_k\eta_k}{2}\|R_{\alpha_k}(y_k)\|^2 + \alpha_k\iota_0\delta_k\|R_{\alpha_k}(y_k)\|,\end{aligned}$$

where the second inequality follows from Condition 2.1, completing the proof. Applying the identity $ab \leq \frac{a^2+b^2}{2}$ to the right-hand side of the above inequality yields (3.3). Finally for (3.4),

$$\|R_{\alpha_k}^{true}(y_k)\| \leq \|R_{\alpha_k}(y_k)\| + \|R_{\alpha_k}^{true}(y_k) - R_{\alpha_k}(y_k)\| \leq \|R_{\alpha_k}(y_k)\| + \|\nabla f(y_k) - g_k\|,$$

where the second inequality follows from (2.5). Substituting Condition 2.1 in the above inequality, then squaring both sides and applying the identity $(a+b)^2 \leq 2a^2 + 2b^2$, completes the proof of (3.4).

For the expectation problem (1.4), from Condition 2.1,

$$\begin{aligned}(\mathbb{E}_k[\|g_k - \nabla f(y_k)\|])^2 \leq \mathbb{E}_k\left[\|g_k - \nabla f(y_k)\|^2\right] &\leq \tfrac{\tilde{\eta}_k^2}{4}\|\mathbb{E}_k\left[R_{\alpha_k}(y_k)\right]\|^2 + \tilde{\iota}_0^2\tilde{\delta}_k^2 \\ &\leq \tfrac{\tilde{\eta}_k^2}{4}\mathbb{E}_k\left[\|R_{\alpha_k}(y_k)\|^2\right] + \tilde{\iota}_0^2\tilde{\delta}_k^2 \leq \left(\tfrac{\tilde{\eta}_k}{2}\sqrt{\mathbb{E}_k\left[\|R_{\alpha_k}(y_k)\|^2\right]} + \tilde{\iota}_0\tilde{\delta}_k\right)^2,\end{aligned}$$

yielding (3.5). For (3.6), from the Cauchy-Schwartz inequality,

$$\begin{aligned}\mathbb{E}_k\left[(\nabla f(y_k) - g_k)^T(x_{k+1} - y_k)\right] &\leq \sqrt{\mathbb{E}_k\left[\|\nabla f(y_k) - g_k\|^2\right]}\sqrt{\mathbb{E}_k\left[\|x_{k+1} - y_k\|^2\right]} \\ &= \alpha_k\sqrt{\mathbb{E}_k\left[\|\nabla f(y_k) - g_k\|^2\right]}\sqrt{\mathbb{E}_k\left[\|R_{\alpha_k}(y_k)\|^2\right]} \\ &\leq \alpha_k\left(\tfrac{\tilde{\eta}_k}{2}\sqrt{\mathbb{E}_k\left[\|R_{\alpha_k}(y_k)\|^2\right]} + \tilde{\iota}_0\tilde{\delta}_k\right)\sqrt{\mathbb{E}_k\left[\|R_{\alpha_k}(y_k)\|^2\right]},\end{aligned}$$

where the equality follows from (2.3) and the second inequality follows from Condition 2.1, completing the proof. Applying the identity $ab \leq \frac{a^2+b^2}{2}$ to (3.6) yields (3.7). Finally for (3.8), using the identity $(a+b)^2 \leq 2a^2 + 2b^2$,

$$\begin{aligned}\left\|R_{\alpha_k}^{true}(y_k)\right\|^2 &\leq 2\left\|\mathbb{E}_k\left[R_{\alpha_k}(y_k)\right]\right\|^2 + 2\left\|\mathbb{E}_k\left[R_{\alpha_k}^{true}(y_k) - R_{\alpha_k}(y_k)\right]\right\|^2 \\ &\leq 2\left\|\mathbb{E}_k\left[R_{\alpha_k}(y_k)\right]\right\|^2 + 2\mathbb{E}_k\left[\left\|R_{\alpha_k}^{true}(y_k) - R_{\alpha_k}(y_k)\right\|^2\right] \\ &\leq 2\left\|\mathbb{E}_k\left[R_{\alpha_k}(y_k)\right]\right\|^2 + 2\mathbb{E}_k\left[\left\|\nabla f(y_k) - g_k\right\|^2\right],\end{aligned}$$

where the second inequality follows from Jenson's inequality, the third inequality follows from (2.5), and further using Condition 2.1 completes the proof. □

3.1 Nonconvex Objective Function

In this section, we present the theoretical analysis for Algorithm 2.1 when the smooth function $f(x)$, and thus the composite function $\phi(x)$, is nonconvex. The analysis is limited to **Option I** in Algorithm 2.1, as **Option II** does not yield any improvements for nonconvex objective functions up to constant factors, as noted in [17]. We first establish the convergence of Algorithm 2.1 when using a biased gradient estimate, followed by the complexity of number of proximal operator evaluations. We then establish the complexity of the number of stochastic gradient evaluation for the expectation problem when using an unbiased gradient estimate.

Theorem 3.3. *Suppose Assumption 2.1 holds and the gradient estimate g_k satisfies Condition 2.1. Then, for Algorithm 2.1 with **Option I**:*

1. *For the finite-sum problem (1.3), if the parameters in Condition 2.1 are chosen such that $\{\eta_k\} = \eta \in [0,1)$, $\iota_0 \in \left[0, \sqrt{\frac{1-\eta}{2}}\right)$ and $\sum_{k=0}^{\infty} \delta_k^2 < \infty$, and the step size is chosen such that $\{\alpha_k\} = \alpha \leq \frac{1-\eta}{2L}$, then $\{x_k\}$ converges to a stationary point with $\min_{k=0,\ldots,K-1} \|R_{\alpha_k}^{true}(x_k)\|^2 = \mathcal{O}\left(\frac{1}{K}\right)$, $\forall K \geq 1$.*
2. *For the expectation problem (1.4), if the parameters in Condition 2.1 are chosen such that $\{\tilde{\eta}_k\} = \tilde{\eta} \in [0,1)$, $\tilde{\iota}_0 \in \left[0, \sqrt{\frac{1-\tilde{\eta}}{2}}\right)$ and $\sum_{k=0}^{\infty} \tilde{\delta}_k^2 < \infty$, and the step size is chosen such that $\{\alpha_k\} = \alpha \leq \frac{1-\tilde{\eta}}{2L}$, then $\{x_k\}$ converges to a stationary point in expectation with $\min_{k=0,\ldots,K-1} \mathbb{E}\left[\left\|R_{\alpha_k}^{true}(x_k)\right\|^2\right] = \mathcal{O}\left(\frac{1}{K}\right)$, $\forall K \geq 1$.*

Proof. With $y_k = x_k$ under **Option I**, consider the result from Lemma 3.1. Substituting $x = x_k$ and using (2.3) yields,

$$\phi(x_{k+1}) \leq \phi(x_k) + (\nabla f(x_k) - g_k)^T(x_{k+1} - x_k) - \alpha_k^2\left(\tfrac{1}{2\alpha_k} - \tfrac{L}{2}\right)\|R_{\alpha_k}(x_k)\|^2. \tag{3.9}$$

For the finite-sum problem (1.3), under the defined parameters, substituting (3.3) from Lemma 3.2 into (3.9) yields,

$$\begin{aligned}\phi(x_{k+1}) &\leq \phi(x_k) + \tfrac{\alpha\left(\eta+\iota_0^2\right)}{2}\|R_\alpha(x_k)\|^2 + \tfrac{\alpha\delta_k^2}{2} - \alpha^2\left(\tfrac{1}{2\alpha} - \tfrac{L}{2}\right)\|R_\alpha(x_k)\|^2 \\ &= \phi(x_k) - \alpha\left[\tfrac{1}{2} - \tfrac{\alpha L}{2} - \tfrac{\eta}{2} - \tfrac{\iota_0^2}{2}\right]\|R_\alpha(x_k)\|^2 + \tfrac{\alpha\delta_k^2}{2}.\end{aligned}$$

Rearranging the above inequality and substituting $c(\eta,\alpha) = \frac{1}{2} - \frac{\alpha L}{2} - \frac{\eta}{2} - \frac{\iota_0^2}{2}$, we get,

$$\|R_\alpha(x_k)\|^2 \leq \tfrac{\phi(x_k)-\phi(x_{k+1})}{\alpha c(\alpha,\eta)} + \tfrac{\delta_k^2}{2c(\alpha,\eta)}.$$

Under the defined parameters, $c(\eta,\alpha) \geq \frac{1}{2} - \frac{(1-\eta)}{4} - \frac{\eta}{2} - \frac{\iota_0^2}{2} = \frac{1-\eta}{4} - \frac{\iota_0^2}{2} > \frac{1-\eta}{4} - \frac{1-\eta}{4} = 0$. Thus, substituting the above bound into (3.4) from Lemma 3.2 yields,

$$\|R_\alpha^{true}(x_k)\|^2 \leq 2\left(1+\tfrac{\eta}{2}\right)^2 \left[\tfrac{\phi(x_k)-\phi(x_{k+1})}{\alpha c(\alpha,\eta)} + \tfrac{\delta_k^2}{2c(\alpha,\eta)}\right] + 2\iota_0^2\delta_k^2.$$

The telescoping sum of the above inequality for $k = 0,\ldots,K-1$ yields,

$$\sum_{k=0}^{K-1} \|R_\alpha^{true}(x_k)\|^2 \leq 2\left(1+\tfrac{\eta}{2}\right)^2 \left[\tfrac{\phi(x_0)-\phi(x_K)}{\alpha c(\alpha,\eta)} + \sum_{k=0}^{K-1} \tfrac{\delta_k^2}{2c(\alpha,\eta)}\right] + 2\iota_0^2 \sum_{k=0}^{K-1} \delta_k^2.$$

Rearranging the above inequality and using $\phi(x_K) \geq \phi^*$, we get,

$$\min_{k=0,\ldots,K-1} \|R_\alpha^{true}(x_k)\|^2 \leq \tfrac{1}{K}\left\{2\left(1+\tfrac{\eta}{2}\right)^2 \left[\tfrac{\phi(x_0)-\phi^*}{\alpha c(\alpha,\eta)}\right] + \left[\tfrac{(2+\eta)^2}{4c(\alpha,\eta)} + 2\iota_0^2\right] \sum_{k=0}^{K-1} \delta_k^2\right\},$$

where all terms within the curly brackets on the right-hand side are bounded due to the condition $\sum_{k=0}^{\infty} \delta_k^2 < \infty$, completing the proof for the finite-sum problem (1.3).

For the expectation problem (1.4), under the defined parameters, taking a conditional expectation of (3.9) given $\mathcal{G}_k$ and substituting (3.7) from Lemma 3.2 yields,

$$\begin{aligned}\mathbb{E}_k\left[\phi(x_{k+1})\right] &\leq \phi(x_k) + \tfrac{\alpha\left(\tilde{\eta}+\tilde{\iota}_0^2\right)}{2}\mathbb{E}_k\left[\|R_\alpha(x_k)\|^2\right] + \tfrac{\alpha\tilde{\delta}_k^2}{2} - \alpha^2\left(\tfrac{1}{2\alpha} - \tfrac{L}{2}\right)\mathbb{E}_k\left[\|R_\alpha(x_k)\|^2\right] \\ &= \phi(x_k) + \tfrac{\alpha\tilde{\delta}_k^2}{2} - \alpha\left[\tfrac{1}{2} - \tfrac{\alpha L}{2} - \tfrac{\tilde{\eta}}{2} - \tfrac{\tilde{\iota}_0^2}{2}\right]\mathbb{E}_k\left[\|R_\alpha(x_k)\|^2\right].\end{aligned}$$

Rearranging the above inequality and substituting $\tilde{c}(\alpha,\tilde{\eta}) = \frac{1}{2} - \frac{\alpha L}{2} - \frac{\tilde{\eta}}{2} - \frac{\tilde{\iota}_0^2}{2}$, we get,

$$\mathbb{E}_k\left[\|R_\alpha(x_k)\|^2\right] \leq \tfrac{\phi(x_k)-\mathbb{E}_k[\phi(x_{k+1})]}{\alpha\tilde{c}(\alpha,\tilde{\eta})} + \tfrac{\tilde{\delta}_k^2}{2\tilde{c}(\alpha,\tilde{\eta})}.$$

Under the defined parameters, $\tilde{c}(\alpha,\tilde{\eta}) > 0$. Since $\|\mathbb{E}_k[R_{\alpha_k}(x_k)]\|^2 \leq \mathbb{E}_k\left[\|R_{\alpha_k}(x_k)\|^2\right]$ by Jenson's inequality, substituting the above bound into (3.8) from Lemma 3.2 yields,

$$\left\|R_\alpha^{true}(x_k)\right\|^2 \leq 2\left(1+\tfrac{\tilde{\eta}^2}{4}\right)\left[\tfrac{\phi(x_k)-\mathbb{E}_k[\phi(x_{k+1})]}{\alpha\tilde{c}(\alpha,\tilde{\eta})} + \tfrac{\tilde{\delta}_k^2}{2\tilde{c}(\alpha,\tilde{\eta})}\right] + 2\tilde{\iota}_0^2\tilde{\delta}_k^2.$$

Taking the total expectation of the above inequality and summing telescopically over $k = 0,\ldots,K-1$ yields,

$$\sum_{k=0}^{K-1} \mathbb{E}\left[\left\|R_\alpha^{true}(x_k)\right\|^2\right] \leq 2\left(1+\tfrac{\tilde{\eta}^2}{4}\right)\left[\tfrac{\phi(x_0)-\mathbb{E}[\phi(x_K)]}{\alpha\tilde{c}(\alpha,\tilde{\eta})} + \sum_{k=0}^{K-1} \tfrac{\tilde{\delta}_k^2}{2\tilde{c}(\alpha,\tilde{\eta})}\right] + \sum_{k=0}^{K-1} 2\tilde{\iota}_0^2\tilde{\delta}_k^2.$$

Rearranging the above inequality and using $\phi(x_K) \geq \phi^*$, we get,

$$\min_{k=0,\ldots,K-1} \mathbb{E}\left[\left\|R_\alpha^{true}(x_k)\right\|^2\right] \leq \tfrac{1}{K}\left\{2\left(1+\tfrac{\tilde{\eta}^2}{4}\right)\left[\tfrac{\phi(x_0)-\phi^*}{\alpha\tilde{c}(\alpha,\tilde{\eta})}\right] + \left[\tfrac{4+\tilde{\eta}^2}{4\tilde{c}(\alpha,\tilde{\eta})} + 2\tilde{\iota}_0^2\right]\sum_{k=0}^{K-1} \tilde{\delta}_k^2\right\}.$$

where all terms within the curly brackets on the right-hand side are bounded due to the condition $\sum_{k=0}^{\infty} \tilde{\delta}_k^2 < \infty$, completing the proof for the expectation problem (1.4). □

Theorem 3.3 establishes a sublinear convergence rate for Algorithm 2.1 with **Option I** for nonconvex objective functions, when using a possibly biased gradient estimate satisfying Condition 2.1. Thus, an $\epsilon > 0$ accurate solution, i.e., $\left\|R^{true}_{\alpha_k}(x_k)\right\|^2 \leq \epsilon$ for the finite-sum problem (1.3) and $\mathbb{E}\left[\left\|R^{true}_{\alpha_k}(x_k)\right\|^2\right] \leq \epsilon$ for the expectation problem (1.4), can be achieved in $\mathcal{O}\left(\frac{1}{\epsilon}\right)$ iterations (proximal operator evaluations), matching the complexity bounds for deterministic first-order methods [9, 10]. We now present the complexity for the number of stochastic gradient evaluations for Algorithm 2.1 for the expectation problem (1.4) with a nonconvex objective function when the gradient estimate is unbiased.

Theorem 3.4. *Suppose Assumptions 2.1 and 2.2 hold, and Condition 2.1 is satisfied for the expectation problem (1.4) via the unbiased gradient estimate in Lemma 2.2. Then, for Algorithm 2.1 with **Option I**, if the parameters in Condition 2.1 are chosen such that* $\{\tilde{\eta}_k\} = \tilde{\eta} \in (0,1)$, $\tilde{\iota}_0 \in \left[0, \sqrt{\frac{1-\tilde{\eta}}{2}}\right)$, $\tilde{\delta}_k = \frac{1}{(k+1)^{1+\nu}}$ $\forall k \geq 0$ *with* $\nu > 0$, *and the step size is chosen such that* $\{\alpha_k\} = \alpha \leq \frac{1-\tilde{\eta}}{2L}$, *a solution satisfying* $\min\left\{\mathbb{E}\left[\left\|R^{true}_{\alpha_k}(x_k)\right\|^2\right], \left\|R^{true}_{\alpha_k}(x_k)\right\|^2\right\} \leq \epsilon$ *with* $\epsilon > 0$ *is achieved in* $\mathcal{O}\left(\epsilon^{-(2+\nu)}\right)$ *stochastic gradient evaluations. If* $\{\tilde{\delta}_k\} = 0$, *this improves to* $\mathcal{O}\left(\epsilon^{-2}\right)$.

Proof. Let $K_\epsilon \geq 1$ be the first iteration that achieves the desired solution accuracy. Hence, $\left\|R^{true}_{\alpha_k}(x_k)\right\|^2 > \epsilon$ $\forall k \leq K_\epsilon - 1$ and from (3.8), $\left\|\mathbb{E}_k\left[R_{\alpha_k}(x_k)\right]\right\|^2 > \left(1 + \frac{\tilde{\eta}^2}{4}\right)^{-1}\left[\frac{\epsilon}{2} - \tilde{\iota}_0^2\tilde{\delta}_k^2\right]$ $\forall k \leq K_\epsilon - 1$. Thus, the total number of stochastic gradient evaluations can be bounded using Lemma 2.2 as,

$$\sum_{k=0}^{K_\epsilon-1}|S_k| = \sum_{k=0}^{K_\epsilon-1}\left\lceil \frac{\sigma^2}{\frac{\tilde{\eta}_k^2}{4}\left\|\mathbb{E}_k\left[R_{\alpha_k}(x_k)\right]\right\|^2 + \tilde{\iota}_0^2\tilde{\delta}_k^2}\right\rceil \leq \sum_{k=0}^{K_\epsilon-1}\frac{2\sigma^2(4+\tilde{\eta}^2)}{\tilde{\eta}^2\epsilon + 8\tilde{\iota}_0^2\tilde{\delta}_k^2} + 1 \leq \sum_{k=0}^{K_\epsilon-1}\frac{2\sigma^2(4+\tilde{\eta}^2)}{\tilde{\eta}^2\epsilon} + \frac{\sigma^2(4+\tilde{\eta}^2)}{4\tilde{\iota}_0^2\tilde{\delta}_k^2} + K_\epsilon.$$

For Algorithm 2.1 with **Option I**, K_ϵ is at most $\mathcal{O}\left(\frac{1}{\epsilon}\right)$ from Theorem 3.3, yielding,

$$\sum_{k=0}^{K_\epsilon-1}|S_k| \leq \frac{2\sigma^2(4+\tilde{\eta}^2)}{\tilde{\eta}^2\epsilon}K_\epsilon + \frac{\sigma^2(4+\tilde{\eta}^2)}{4\tilde{\iota}_0^2}K_\epsilon^{2+\nu} + K_\epsilon = \mathcal{O}\left(\frac{1}{\epsilon^{2+\nu}}\right).$$

Following the same procedure, if $\{\tilde{\delta}_k\} = 0$, $\sum_{k=0}^{K_\epsilon-1}|S_k| \leq \frac{2\sigma^2(4+\tilde{\eta}^2)}{\tilde{\eta}^2\epsilon}K_\epsilon + K_\epsilon = \mathcal{O}\left(\frac{1}{\epsilon^2}\right)$. □

Theorem 3.4 matches the optimal complexity for the number of stochastic gradients for the expectation problem (1.4) over nonconvex objective functions [16]. We conclude this section with a corollary to Theorem 3.4, using a definition of an ϵ-accurate solution similar to that in [17], under the parameter setting $\{\tilde{\eta}_k\} = 0$.

Corollary 3.5. *Suppose the conditions in Theorem 3.4 hold. If the parameters in Condition 2.1 are chosen as* $\{\tilde{\eta}_k\} = 0$, $\tilde{\iota}_0 \in \left(0, \frac{1}{\sqrt{2}}\right)$ *and* $\tilde{\delta}_k = \frac{1}{(k+1)^{1+\nu}}$ $\forall k \geq 0$ *with* $\nu > 0$, *and the step size is chosen such that* $\{\alpha_k\} = \alpha \leq \frac{1}{2L}$, *a solution satisfying* $\mathbb{E}\left[\left\|R^{true}_{\alpha_k}(x_k)\right\|^2\right] \leq \epsilon$ *with* $\epsilon > 0$ *is achieved in* $\mathcal{O}\left(\epsilon^{-(2+\nu)}\right)$ *stochastic gradient evaluations.*

Proof. The proof follows from the same procedure as Theorem 3.4. □

The parameter settings in Corollary 3.5 reduce Condition 2.1 to maintaining a predetermined sequence of gradient estimation errors, similar to [37].

3.2 General Convex Objective Function

In this section, we provide the theoretical analysis of Algorithm 2.1 when the smooth function $f(x)$, and thus the composite function $\phi(x)$, is convex. We begin by stating the basic assumptions and definitions, along with some mathematical identities that will be used throughout the analysis.

Assumption 3.1. *The function $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth and convex, and the function $h : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is closed, convex, and proper.*

Under Assumption 3.1, let $x^* \in \mathbb{R}^d$ denote an optimal solution where the optimal value ϕ^* is attained. Since $f(x)$ is differentiable and convex, from [31],

$$f(b) \geq f(a) + \nabla f(a)^T(b-a) \quad \forall a, b \in \mathbb{R}^d. \tag{3.10}$$

For Algorithm 2.1 with **Option II**, under Assumption 3.1, we define the sequence $\{\beta_k\}$ and two additional sequences $\{\theta_k\}$ and $\{v_k\}$ for the analysis as,

$$\beta_k = \tfrac{k-1}{k+2} \quad \forall k \geq 1, \quad \theta_k = \tfrac{2}{k+1} \quad \forall k \geq 0, \text{ and } \quad v_k = x_{k-1} + \tfrac{1}{\theta_k}(x_k - x_{k-1}) \quad \forall k \geq 1, \tag{3.11}$$

with $v_0 = x_0$. Under these definitions, y_k can be expressed as,

$$\begin{aligned} y_k &= x_k + \tfrac{k-1}{k+2}(x_k - x_{k-1}) = \left(1 - \tfrac{2}{k+2}\right) x_k + \tfrac{2}{k+2}\left(x_{k-1} + \tfrac{k+1}{2}(x_k - x_{k-1})\right) \\ &= (1-\theta_{k+1})x_k + \theta_{k+1} v_k \quad \forall k \geq 0. \end{aligned} \tag{3.12}$$

From (3.11) and (3.12), an alternative update form for $\{v_k\}$ can be derived as,

$$\begin{aligned} v_{k+1} &= x_k + \tfrac{1}{\theta_{k+1}}(x_{k+1} - x_k) = v_k - \tfrac{1}{\theta_{k+1}}\left((1-\theta_{k+1})x_k + \theta_{k+1} v_k - x_{k+1}\right) \\ &= v_k - \tfrac{1}{\theta_{k+1}}(y_k - x_{k+1}) \quad \forall k \geq 0. \end{aligned} \tag{3.13}$$

Finally, we also introduce a useful identity for $\{\theta_k\}$ as,

$$\tfrac{(1-\theta_k)}{\theta_k^2} = \tfrac{(k-1)(k+1)}{4} = \tfrac{k^2-1}{4} \leq \tfrac{k^2}{4} = \tfrac{1}{\theta_{k-1}^2} \quad \forall k \geq 1. \tag{3.14}$$

We now establish a general descent lemma under Assumption 3.1.

Lemma 3.6. *Suppose Assumption 3.1 holds. Then, $\forall z \in \mathbb{R}^d$ and $\forall k \geq 0$, the iterates generated by Algorithm 2.1 satisfy,*

$$\begin{aligned} \phi(x_{k+1}) \leq \phi(z) &+ (g_k - \nabla f(y_k))^T(z - y_k) + (\nabla f(y_k) - g_k)^T(x_{k+1} - y_k) \\ &+ \left[\tfrac{L}{2} - \tfrac{1}{\alpha_k}\right]\|x_{k+1} - y_k\|^2 + \tfrac{1}{\alpha_k}(y_k - x_{k+1})^T(y_k - z). \end{aligned}$$

Proof. From (2.1),

$$\begin{aligned} \phi(x_{k+1}) &\leq f(y_k) + \nabla f(y_k)^T(x_{k+1} - y_k) + \tfrac{L}{2}\|x_{k+1} - y_k\|^2 + h(x_{k+1}) \\ &\leq f(z) - \nabla f(y_k)^T(z - y_k) + \nabla f(y_k)^T(x_{k+1} - y_k) + \tfrac{L}{2}\|x_{k+1} - y_k\|^2 + h(x_{k+1}) \\ &\leq f(z) - \nabla f(y_k)^T(z - y_k) + \nabla f(y_k)^T(x_{k+1} - y_k) + \tfrac{L}{2}\|x_{k+1} - y_k\|^2 \\ &\quad + h(z) - \left(\tfrac{y_k - x_{k+1}}{\alpha_k} - g_k\right)^T(z - x_{k+1}) \\ &= \phi(z) - \nabla f(y_k)^T(z - y_k) + \nabla f(y_k)^T(x_{k+1} - y_k) + \tfrac{L}{2}\|x_{k+1} - y_k\|^2 \\ &\quad - \left(\tfrac{y_k - x_{k+1}}{\alpha_k} - g_k\right)^T(z - y_k + y_k - x_{k+1}), \end{aligned}$$

where the second inequality follows from (3.10) $\forall z \in \mathbb{R}^d$, and the third inequality follows from the convexity of $h(x)$ and the definition (2.2) for x_{k+1} which yields $0 \in g_k + \partial h(x_{k+1}) + \frac{x_{k+1}-y_k}{\alpha_k}$. Rearranging the terms in the last equality completes the proof. □

Using Lemma 3.6, we establish recursive bounds on the optimality gap in function value for Algorithm 2.1 with **Option I** and **Option II**, under Assumption 3.1.

Lemma 3.7. *Suppose Assumption 3.1 holds.*

1. *For Algorithm 2.1 with **Option I**, $\forall k \geq 0$, the iterates satisfy,*

$$\begin{aligned}\phi(x_{k+1}) - \phi^* \leq (g_k - \nabla f(x_k))^T(x^* - x_k) + (\nabla f(x_k) - g_k)^T(x_{k+1} - x_k) \\ + \tfrac{1}{2\alpha_k}\left[\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 - (1 - \alpha_k L)\|x_k - x_{k+1}\|^2\right].\end{aligned}$$

2. *For Algorithm 2.1 with **Option II**, $\forall k \geq 0$, the iterates satisfy,*

$$\begin{aligned}\phi(x_{k+1}) - \phi^* \leq (1-\theta_{k+1})(\phi(x_k) - \phi^*) + \theta_{k+1}(g_k - \nabla f(y_k))^T(x^* - v_k) \\ + (\nabla f(y_k) - g_k)^T(x_{k+1} - y_k) \\ + \tfrac{\theta_{k+1}^2}{2\alpha_k}\left[\|v_k - x^*\|^2 - \|v_{k+1} - x^*\|^2 - (1 - \alpha_k L)\|v_k - v_{k+1}\|^2\right].\end{aligned}$$

Proof. For Algorithm 2.1 with **Option I**, consider the result from Lemma 3.6 with $y_k = x_k$ and $z = x^*$,

$$\begin{aligned}\phi(x_{k+1}) - \phi^* \leq (g_k - \nabla f(x_k))^T(x^* - x_k) + (\nabla f(x_k) - g_k)^T(x_{k+1} - x_k) \\ + \left[\tfrac{L}{2} - \tfrac{1}{\alpha_k}\right]\|x_{k+1} - x_k\|^2 + \tfrac{1}{\alpha_k}(x_k - x_{k+1})^T(x_k - x^*).\end{aligned}$$

Applying Lemma A.1 with $a_1 = x_k - x_{k+1}$, $a_2 = x_k - x^*$ and $c = \frac{\alpha_k L}{2}$ completes the proof for **Option I**.

For Algorithm 2.1 with **Option II**, consider the result from Lemma 3.6 with $z = \theta_{k+1}x^* + (1 - \theta_{k+1})x_k$. Since $\theta_{k+1} \in [0, 1]$ $\forall k \geq 0$, from Assumption 3.1,

$$\begin{aligned}\phi(x_{k+1}) - \phi^* &\leq (1-\theta_{k+1})(\phi(x_k) - \phi^*) + (g_k - \nabla f(y_k))^T(\theta_{k+1}x^* + (1-\theta_{k+1})x_k - y_k) \\ &\quad + (\nabla f(y_k) - g_k)^T(x_{k+1} - y_k) + \left[\tfrac{L}{2} - \tfrac{1}{\alpha_k}\right]\|x_{k+1} - y_k\|^2 \\ &\quad + \tfrac{1}{\alpha_k}(y_k - x_{k+1})^T(y_k - \theta_{k+1}x^* - (1-\theta_{k+1})x_k) \\ &= (1-\theta_{k+1})(\phi(x_k) - \phi^*) + \theta_{k+1}(g_k - \nabla f(y_k))^T(x^* - v_k) + (\nabla f(y_k) - g_k)^T(x_{k+1} - y_k) \\ &\quad + \theta_{k+1}^2\left[\tfrac{L}{2} - \tfrac{1}{\alpha_k}\right]\|v_{k+1} - v_k\|^2 + \tfrac{\theta_{k+1}^2}{\alpha_k}(v_k - v_{k+1})^T(v_k - x^*),\end{aligned}$$

where the equality follows from (3.12) and (3.13). Applying Lemma A.1 with $a_1 = v_k - v_{k+1}$, $a_2 = v_k - x^*$ and $c = \frac{\alpha_k L}{2}$ completes the proof for **Option II**. □

We now present the convergence of Algorithm 2.1 under Assumption 3.1 when using a biased gradient estimate, followed by the complexity of number of proximal operator evaluations.

Theorem 3.8. *Suppose Assumption 3.1 holds and the gradient estimate g_k satisfies Condition 2.1.*

1. *For Algorithm 2.1 with* ***Option I****:*

   (a) *For the finite-sum problem (1.3), if the parameters in Condition 2.1 are chosen such that* $\{\eta_k\} \searrow 0$, $\{\eta_k\} \leq \eta < \frac{1}{2}$, $\sum_{k=0}^{\infty} \eta_k^2 < \infty$, $\iota_0^2 \in \left[0, \frac{1}{2} - \eta\right)$, $\sum_{k=0}^{\infty} \delta_k < \infty$ *and* $\sum_{k=0}^{\infty} \delta_k^2 < \infty$, *and the step size is chosen such that* $\{\alpha_k\} = \alpha \leq \frac{1-2(\eta_k+\iota_0^2)}{2L}$ $\forall k \geq 0$, *then* $\{\phi(x_k)\}$ *converges to the optimal value with* $\min_{k=0,\ldots,K-1} \phi(x_k) - \phi^* = \mathcal{O}\left(\frac{1}{K}\right)$ $\forall K \geq 1$.

   (b) *For the expectation problem (1.3), if the parameters in Condition 2.1 are chosen such that* $\{\tilde{\eta}_k\} \searrow 0$, $\{\tilde{\eta}_k\} \leq \tilde{\eta} < \frac{1}{2}$, $\sum_{k=0}^{\infty} \tilde{\eta}_k^2 < \infty$, $\tilde{\iota}_0^2 \in \left[0, \frac{1}{2} - \tilde{\eta}\right)$, $\sum_{k=0}^{\infty} \tilde{\delta}_k < \infty$ *and* $\sum_{k=0}^{\infty} \tilde{\delta}_k^2 < \infty$, *and the step size is chosen such that* $\{\alpha_k\} = \alpha \leq \frac{1-2(\tilde{\eta}_k+\tilde{\iota}_0^2)}{2L}$ $\forall k \geq 0$, *then* $\{\phi(x_k)\}$ *converges to the optimal value in expectation with* $\min_{k=0,\ldots,K-1} \mathbb{E}\left[\phi(x_k) - \phi^*\right] = \mathcal{O}\left(\frac{1}{K}\right)$ $\forall K \geq 1$.

2. *For Algorithm 2.1 with* ***Option II****:*

   (a) *For the finite-sum problem (1.3), if the parameters in Condition 2.1 are chosen such that* $\eta_k = \hat{\eta} t_k$ *and* $\delta_k = \hat{\delta} u_k$ $\forall k \geq 0$, *where* $\{\eta_k\} \leq \eta < \frac{1}{2}$, $\{t_k\} \searrow 0$, $\sum_{k=0}^{\infty} t_k^2 < \infty$, $\sum_{k=0}^{\infty} k t_k^2 < \infty$, $\sum_{k=0}^{\infty} (k+2)^2 u_k < \infty$, $\sum_{k=0}^{\infty} (k+2)^2 u_k^2 < \infty$, $\iota_0^2 \in \left[0, \frac{1}{2} - \eta\right)$ *and* $\hat{\eta}, \hat{\delta} \geq 0$ *are sufficiently small, and the step size is chosen such that* $\{\alpha_k\} = \alpha \leq \frac{1-2(\eta_k+\iota_0^2)}{2L}$ $\forall k \geq 0$, *then* $\{\phi(x_k)\}$ *converges to the optimal value with* $\phi(x_k) - \phi^* = \mathcal{O}\left(\frac{1}{(k+1)^2}\right)$ $\forall k \geq 0$.

   (b) *For the expectation problem (1.4), if the parameters in Condition 2.1 are chosen such that* $\tilde{\eta}_k = \hat{\eta} \tilde{t}_k$ *and* $\tilde{\delta}_k = \hat{\delta} \tilde{u}_k$ $\forall k \geq 0$, *where* $\{\tilde{\eta}_k\} \leq \tilde{\eta} < \frac{1}{2}$, $\{\tilde{t}_k\} \searrow 0$, $\sum_{k=0}^{\infty} \tilde{t}_k^2 < \infty$, $\sum_{k=0}^{\infty} k \tilde{t}_k^2 < \infty$, $\sum_{k=0}^{\infty} (k+2)^2 \tilde{u}_k < \infty$, $\sum_{k=0}^{\infty} (k+2)^2 \tilde{u}_k^2 < \infty$, $\tilde{\iota}_0^2 \in \left[0, \frac{1}{2} - \tilde{\eta}\right)$ *and* $\hat{\eta}, \hat{\delta} \geq 0$ *are sufficiently small, and the step size is chosen such that* $\{\alpha_k\} = \alpha \leq \frac{1-2(\tilde{\eta}_k+\tilde{\iota}_0^2)}{2L}$ $\forall k \geq 0$, *then* $\{\phi(x_k)\}$ *converges to the optimal value in expectation with* $\mathbb{E}\left[\phi(x_k) - \phi^*\right] = \mathcal{O}\left(\frac{1}{(k+1)^2}\right)$ $\forall k \geq 0$.

Proof. For Algorithm 2.1 with **Option I**, applying Cauchy-Schwarz inequality to the result in Lemma 3.7, we get,

$$\begin{aligned} \phi(x_{k+1}) - \phi^* \leq \|g_k - \nabla f(x_k)\| \|x_k - x^*\| + (\nabla f(x_k) - g_k)^T (x_{k+1} - x_k) \\ + \tfrac{1}{2\alpha_k} \left[\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 - (1 - \alpha_k L) \|x_k - x_{k+1}\|^2\right]. \end{aligned} \tag{3.15}$$

For the finite-sum problem (1.3) for Algorithm 2.1 with **Option I**, substituting Condition 2.1 and (3.2) from Lemma 3.2 into (3.15),

$$\begin{aligned} \phi(x_{k+1}) - \phi^* &\leq \left(\tfrac{\eta_k}{2\alpha_k} \|x_k - x_{k+1}\| + \iota_0 \delta_k\right) \|x_k - x^*\| + \left(\tfrac{\eta_k}{2\alpha_k} \|x_k - x_{k+1}\|^2 + \iota_0 \delta_k \|x_k - x_{k+1}\|\right) \\ &\quad + \tfrac{1}{2\alpha_k} \left[\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 - (1 - \alpha_k L) \|x_k - x_{k+1}\|^2\right] \\ &\leq \tfrac{\eta_k^2}{4\alpha_k} \|x_k - x^*\|^2 + \tfrac{1}{4\alpha_k} \|x_k - x_{k+1}\|^2 + \iota_0 \delta_k \|x_k - x^*\| + \tfrac{\alpha_k \delta_k^2}{2} + \tfrac{\iota_0^2 \|x_k - x_{k+1}\|^2}{2\alpha_k} \\ &\quad + \tfrac{1}{2\alpha_k} \left[\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 - (1 - \alpha_k L - \eta_k) \|x_k - x_{k+1}\|^2\right] \\ &= \tfrac{\eta_k^2}{4\alpha_k} \|x_k - x^*\|^2 + \iota_0 \delta_k \|x_k - x^*\| + \tfrac{\alpha_k \delta_k^2}{2} \\ &\quad + \tfrac{1}{2\alpha_k} \left[\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 - \left(\tfrac{1}{2} - \alpha_k L - \eta_k - \iota_0^2\right) \|x_k - x_{k+1}\|^2\right], \end{aligned}$$

where the second inequality follows from the identity $ab \leq \frac{a^2+b^2}{2}$ and young's inequality as $\iota_0\delta_k\|x_k - x_{k+1}\| \leq \frac{\alpha_k\delta_k^2}{2} + \frac{\iota_0^2\|x_k - x_{k+1}\|^2}{2\alpha_k}$ with $\alpha_k > 0$. Under the defined parameters, the constant $\frac{1}{2} - \alpha_k L - \eta_k - \iota_0^2 \geq 0$, and hence the last term in the upper bound can be omitted. Applying young's inequality again, $\|x_k - x^*\| \leq \frac{\|x_k - x^*\|^2}{2\alpha_k} + \frac{\alpha_k}{2}$, we get,

$$\begin{aligned}\phi(x_{k+1}) - \phi^* &\leq \left(\frac{\eta_k^2}{4\alpha_k} + \frac{\iota_0\delta_k}{2\alpha_k}\right)\|x_k - x^*\|^2 + \frac{\alpha_k(\iota_0\delta_k + \delta_k^2)}{2} + \frac{1}{2\alpha_k}\left[\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2\right] \\ &\leq \left(\frac{\eta_k^2}{4\alpha} + \frac{\iota_0\delta_k}{2\alpha}\right)\|x_k - x^*\|^2 + \frac{(\iota_0\delta_k + \delta_k^2)}{2L} + \frac{1}{2\alpha}\left[\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2\right], \quad (3.16)\end{aligned}$$

where the second inequality follows from $\alpha = \{\alpha_k\} \leq \frac{1}{L}$. In the above bound, since $\phi(x_{k+1}) - \phi^* \geq 0$, one can create a recursive relation,

$$\|x_{k+1} - x^*\|^2 \leq \left(\frac{\eta_k^2}{2} + \iota_0\delta_k + 1\right)\|x_k - x^*\|^2 + \frac{\alpha}{L}\left(\iota_0\delta_k + \delta_k^2\right).$$

From Lemma A.3, with $T_k = \|x_k - x^*\|^2$, $a_k = \frac{\eta_k^2}{2} + \iota_0\delta_k$ and $s_k = \frac{\alpha}{L}\left(\iota_0\delta_k + \delta_k^2\right)$, under the defined parameters, the sequence $\left\{\|x_k - x^*\|^2\right\}$ is bounded. Taking a telescopic sum of (3.16) for $k = 0, \ldots, K-1$ yields,

$$\sum_{k=0}^{K-1}\phi(x_{k+1}) - \phi^* \leq \sum_{k=0}^{K-1}\left(\frac{\eta_k^2}{4\alpha} + \frac{\iota_0\delta_k}{2\alpha}\right)\|x_k - x^*\|^2 + \sum_{k=0}^{K-1}\frac{(\iota_0\delta_k + \delta_k^2)}{2L} + \frac{1}{2\alpha}\|x_0 - x^*\|^2.$$

Rearranging the terms in the above inequality,

$$\min_{k=0,\ldots,K-1}\phi(x_{k+1}) - \phi^* \leq \frac{1}{K}\left\{\sum_{k=0}^{\infty}\left(\frac{\eta_k^2}{4\alpha} + \frac{\iota_0\delta_k}{2\alpha}\right)\|x_k - x^*\|^2 + \sum_{k=0}^{\infty}\frac{(\iota_0\delta_k + \delta_k^2)}{2L} + \frac{1}{2\alpha}\|x_0 - x^*\|^2\right\},$$

where all terms within the curly brackets on the right-hand side are bounded, completing the proof for the finite-sum problem (1.3) for Algorithm 2.1 with **Option I**.

For the expectation problem (1.4) for Algorithm 2.1 with **Option I**, consider the conditional expectation of (3.15) given $\mathcal{G}_k$. Substituting (3.5) and (3.6) from Lemma 3.2 yields,

$$\begin{aligned}\mathbb{E}_k\left[\phi(x_{k+1}) - \phi^*\right] &\leq \left(\frac{\tilde{\eta}_k}{2\alpha_k}\sqrt{\mathbb{E}_k\left[\|x_k - x_{k+1}\|^2\right]} + \tilde{\iota}_0\tilde{\delta}_k\right)\|x_k - x^*\| \\ &\quad + \left(\frac{\tilde{\eta}_k}{2\alpha_k}\mathbb{E}_k\left[\|x_k - x_{k+1}\|^2\right] + \tilde{\iota}_0\tilde{\delta}_k\sqrt{\mathbb{E}_k\left[\|x_k - x_{k+1}\|^2\right]}\right) \\ &\quad + \frac{1}{2\alpha_k}\mathbb{E}_k\left[\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 - (1 - \alpha_k L)\|x_k - x_{k+1}\|^2\right] \\ &\leq \frac{\tilde{\eta}_k^2}{4\alpha_k}\|x_k - x^*\|^2 + \tilde{\iota}_0\tilde{\delta}_k\|x_k - x^*\| + \frac{\alpha_k\tilde{\delta}_k^2}{2} \\ &\quad + \frac{1}{2\alpha_k}\mathbb{E}_k\left[\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 - \left(\frac{1}{2} - \alpha_k L - \tilde{\eta}_k - \tilde{\iota}_0^2\right)\|x_k - x_{k+1}\|^2\right],\end{aligned}$$

where the second inequality follows from young's inequality. Under the defined parameters, the constant $\frac{1}{2} - \alpha_k L - \tilde{\eta}_k - \tilde{\iota}_0^2 \geq 0$, and hence the last term in the upper bound can be omitted. Taking the total expectation of the reduced bound and applying young's inequality yields,

$$\begin{aligned}\mathbb{E}[\phi(x_{k+1}) - \phi^*] &\leq \left(\frac{\tilde{\eta}_k^2}{4\alpha_k} + \frac{\tilde{\iota}_0\tilde{\delta}_k}{2\alpha_k}\right)\mathbb{E}\left[\|x_k - x^*\|^2\right] + \frac{\alpha_k(\tilde{\iota}_0\tilde{\delta}_k + \tilde{\delta}_k^2)}{2} + \frac{1}{2\alpha_k}\mathbb{E}\left[\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|\right] \\ &\leq \left(\frac{\tilde{\eta}_k^2}{\alpha} + \frac{\tilde{\iota}_0\tilde{\delta}_k}{2\alpha}\right)\mathbb{E}\left[\|x_k - x^*\|^2\right] + \frac{(\tilde{\iota}_0\tilde{\delta}_k + \tilde{\delta}_k^2)}{2L} + \frac{1}{2\alpha}\mathbb{E}\left[\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|\right],\end{aligned}$$

where the second inequality follows from $\alpha = \{\alpha_k\} \leq \frac{1}{L}$. Similar to the finite-sum problem, one can create a recursive relation and use Lemma A.3 to show that $\left\{\mathbb{E}\left[\|x_k - x^*\|^2\right]\right\}$ is a bounded sequence. Taking a telescopic sum of the bound for $k = 0, \ldots, K-1$ yields,

$$\sum_{k=0}^{K-1} \mathbb{E}[\phi(x_{k+1}) - \phi^*] \leq \sum_{k=0}^{K-1} \left(\frac{\tilde{\eta}_k^2}{4\alpha} + \frac{\tilde{\iota}_0\tilde{\delta}_k}{2\alpha}\right) \mathbb{E}\left[\|x_k - x^*\|^2\right] + \sum_{k=0}^{K-1} \frac{(\tilde{\iota}_0\tilde{\delta}_k + \tilde{\delta}_k^2)}{2L} + \frac{1}{2\alpha}\mathbb{E}\left[\|x_0 - x^*\|^2\right].$$

Rearranging the terms in the above inequality,

$$\min_{k=0,\ldots,K-1} \mathbb{E}[\phi(x_{k+1}) - \phi^*] \leq \frac{1}{K}\left\{\sum_{k=0}^{\infty} \left(\frac{\tilde{\eta}_k^2}{4\alpha} + \frac{\tilde{\iota}_0\tilde{\delta}_k}{2\alpha}\right) \mathbb{E}\left[\|x_k - x^*\|^2\right] + \sum_{k=0}^{\infty} \frac{(\tilde{\iota}_0\tilde{\delta}_k + \tilde{\delta}_k^2)}{2L} + \frac{1}{2\alpha}\mathbb{E}\left[\|x_0 - x^*\|^2\right]\right\},$$

where all terms within the curly brackets on the right-hand side are bounded, completing the proof for the expectation problem (1.4) for Algorithm 2.1 with **Option I**.

For Algorithm 2.1 with **Option II**, applying Cauchy-Schwartz inequality to the result in Lemma 3.7 yields,

$$\begin{aligned}\phi(x_{k+1}) - \phi^* \leq{}& (1-\theta_{k+1})(\phi(x_k) - \phi^*) + \theta_{k+1}\|g_k - \nabla f(y_k)\|\|v_k - x^*\| \\ &+ (\nabla f(y_k) - g_k)^T(x_{k+1} - y_k) - \frac{\theta_{k+1}^2}{2\alpha_k}(1 - \alpha_k L)\|v_k - v_{k+1}\|^2 \\ &+ \frac{\theta_{k+1}^2}{2\alpha_k}\left[\|v_k - x^*\|^2 - \|v_{k+1} - x^*\|^2\right].\end{aligned} \tag{3.17}$$

For the finite-sum problem (1.3) for Algorithm 2.1 with **Option II**, substituting Condition 2.1, (3.2) from Lemma 3.2, and (3.13) into (3.15) yields,

$$\begin{aligned}\phi(x_{k+1}) - \phi^* \leq{}& (1-\theta_{k+1})(\phi(x_k) - \phi^*) + \theta_{k+1}\left(\frac{\theta_{k+1}\eta_k}{2\alpha_k}\|v_k - v_{k+1}\| + \iota_0\delta_k\right)\|v_k - x^*\| \\ &+ \left(\frac{\eta_k\theta_{k+1}^2}{2\alpha_k}\|v_k - v_{k+1}\|^2 + \theta_{k+1}\iota_0\delta_k\|v_k - v_{k+1}\|\right) - \frac{\theta_{k+1}^2}{2\alpha_k}(1 - \alpha_k L)\|v_k - v_{k+1}\|^2 \\ &+ \frac{\theta_{k+1}^2}{2\alpha_k}\left[\|v_k - x^*\|^2 - \|v_{k+1} - x^*\|^2\right] \\ \leq{}& (1-\theta_{k+1})(\phi(x_k) - \phi^*) + \theta_{k+1}\iota_0\delta_k\|v_k - x^*\| + \frac{\theta_{k+1}^2\eta_k^2}{4\alpha_k}\|v_k - x^*\|^2 + \frac{\alpha_k\delta_k^2}{2} \\ &- \frac{\theta_{k+1}^2}{2\alpha_k}\left(\frac{1}{2} - \alpha_k L - \eta_k - \iota_0^2\right)\|v_k - v_{k+1}\|^2 + \frac{\theta_{k+1}^2}{2\alpha_k}\left[\|v_k - x^*\|^2 - \|v_{k+1} - x^*\|^2\right],\end{aligned}$$

where the second inequality follows from young's inequality. Under the defined parameters, the constant $\frac{1}{2} - \alpha_k L - \eta_k - \iota_0^2 \geq 0$, and hence the fifth term in the upper bound can be ignored. Applying young's inequality to the reduced bound yields,

$$\begin{aligned}\phi(x_{k+1}) - \phi^* \leq{}& (1-\theta_{k+1})(\phi(x_k) - \phi^*) + \frac{\theta_{k+1}^2}{2\alpha_k}\left[\iota_0\delta_k + \frac{\eta_k^2}{2}\right]\|v_k - x^*\|^2 + \frac{\alpha_k(\delta_k^2 + \iota_0\delta_k)}{2} \\ &+ \frac{\theta_{k+1}^2}{2\alpha_k}\left[\|v_k - x^*\|^2 - \|v_{k+1} - x^*\|^2\right].\end{aligned}$$

Dividing the above inequality by θ_{k+1}^2 and applying (3.14), we get,

$$\begin{aligned}\frac{\phi(x_{k+1}) - \phi^*}{\theta_{k+1}^2} \leq{}& \frac{1}{\theta_k^2}(\phi(x_k) - \phi^*) + \frac{1}{2\alpha_k}\left[\iota_0\delta_k + \frac{\eta_k^2}{2}\right]\|v_k - x^*\|^2 + \frac{\alpha_k(\delta_k^2 + \iota_0\delta_k)}{2\theta_{k+1}^2} \\ &+ \frac{1}{2\alpha_k}\left[\|v_k - x^*\|^2 - \|v_{k+1} - x^*\|^2\right] \\ \leq{}& \frac{1}{\theta_k^2}(\phi(x_k) - \phi^*) + \frac{1}{2\alpha}\left[\iota_0\delta_k + \frac{\eta_k^2}{2}\right]\|v_k - x^*\|^2 + \frac{\delta_k^2 + \iota_0\delta_k}{2\theta_{k+1}^2 L} \\ &+ \frac{1}{2\alpha}\left[\|v_k - x^*\|^2 - \|v_{k+1} - x^*\|^2\right],\end{aligned}$$

where the second inequality follows from $\alpha = \{\alpha_k\} \leq \frac{1}{L}$. By Lemma A.4, with $R_k = \frac{1}{\theta_k^2}(\phi(x_k) - \phi^*)$, $T_k = \frac{1}{2\alpha}\|v_k - x^*\|^2$, $a_k = \iota_0\delta_k + \frac{\eta_k^2}{2}$, and $s_k = \frac{\delta_k^2 + \iota_0\delta_k}{2\theta_{k+1}^2 L}$, the sequence $\{R_k\}$ is bounded under the specified parameters provided that $\hat{\eta}$ and $\hat{\delta}$ are chosen to be sufficiently small. Thus, $\exists C \geq 0$ such that, $\phi(x_k) - \phi^* \leq C\theta_k^2 = \frac{4C}{(k+1)^2}$, completing the proof for the finite-sum problem (1.3) for Algorithm 2.1 with **Option II**.

For the expectation problem (1.4) for Algorithm 2.1 with **Option II**, consider the conditional expectation of (3.17) given $\mathcal{G}_k$. Substituting (3.5) and (3.6) from Lemma 3.2, and (3.13) yields,

$$\begin{aligned}
\mathbb{E}_k\left[\phi(x_{k+1}) - \phi^*\right] &\leq (1-\theta_{k+1})(\phi(x_k) - \phi^*) + \theta_{k+1}\left(\tfrac{\theta_{k+1}\tilde{\eta}_k}{2\alpha_k}\sqrt{\mathbb{E}_k\left[\|v_k - v_{k+1}\|^2\right]} + \tilde{\iota}_0\tilde{\delta}_k\right)\|v_k - x^*\| \\
&\quad + \tfrac{\theta_{k+1}^2\tilde{\eta}_k}{2\alpha_k}\mathbb{E}_k\left[\|v_k - v_{k+1}\|^2\right] + \theta_{k+1}\tilde{\iota}_0\tilde{\delta}_k\sqrt{\mathbb{E}_k\left[\|v_k - v_{k+1}\|^2\right]} \\
&\quad - \tfrac{\theta_{k+1}^2}{2\alpha_k}(1 - \alpha_k L)\,\mathbb{E}_k\left[\|v_k - v_{k+1}\|^2\right] + \tfrac{\theta_{k+1}^2}{2\alpha_k}\mathbb{E}_k\left[\|v_k - x^*\|^2 - \|v_{k+1} - x^*\|^2\right] \\
&\leq (1-\theta_{k+1})(\phi(x_k) - \phi^*) + \theta_{k+1}\tilde{\iota}_0\tilde{\delta}_k\|v_k - x^*\| + \tfrac{\theta_{k+1}^2\tilde{\eta}_k^2}{4\alpha_k}\|v_k - x^*\|^2 + \tfrac{\alpha_k\tilde{\delta}_k^2}{2} \\
&\quad - \tfrac{\theta_{k+1}^2}{2\alpha_k}\left(\tfrac{1}{2} - \alpha_k L - \tilde{\eta}_k - \tilde{\iota}_0^2\right)\mathbb{E}_k\left[\|v_k - v_{k+1}\|^2\right] + \tfrac{\theta_{k+1}^2}{2\alpha_k}\mathbb{E}_k\left[\|v_k - x^*\|^2 - \|v_{k+1} - x^*\|^2\right],
\end{aligned}$$

where the second inequality follows from young's inequality. Under the defined parameters, the constant $\frac{1}{2} - \alpha_k L - \tilde{\eta}_k - \tilde{\iota}_0^2 \geq 0$, and hence the fifth term in the upper bound can be ignored. Applying young's inequality to the reduced bound yields,

$$\begin{aligned}
\mathbb{E}_k\left[\phi(x_{k+1}) - \phi^*\right] &\leq (1-\theta_{k+1})(\phi(x_k) - \phi^*) + \tfrac{\theta_{k+1}^2}{2\alpha_k}\left[\tilde{\iota}_0\tilde{\delta}_k + \tfrac{\tilde{\eta}_k^2}{2}\right]\|v_k - x^*\|^2 + \tfrac{\alpha_k(\tilde{\delta}_k^2 + \tilde{\iota}_0\tilde{\delta}_k)}{2} \\
&\quad + \tfrac{\theta_{k+1}^2}{2\alpha_k}\mathbb{E}_k\left[\|v_k - x^*\|^2 - \|v_{k+1} - x^*\|^2\right].
\end{aligned}$$

Taking the total expectation of the above bound, dividing by θ_{k+1}^2 and using (3.14), we get,

$$\begin{aligned}
\mathbb{E}\left[\tfrac{\phi(x_{k+1}) - \phi^*}{\theta_{k+1}^2}\right] &\leq \tfrac{1}{\theta_k^2}\mathbb{E}\left[\phi(x_k) - \phi^*\right] + \tfrac{1}{2\alpha_k}\left[\tilde{\iota}_0\tilde{\delta}_k + \tfrac{\tilde{\eta}_k^2}{2}\right]\mathbb{E}\left[\|v_k - x^*\|^2\right] + \tfrac{\alpha_k(\tilde{\delta}_k^2 + \tilde{\iota}_0\tilde{\delta}_k)}{2\theta_{k+1}^2} \\
&\quad + \tfrac{1}{2\alpha_k}\mathbb{E}\left[\|v_k - x^*\|^2 - \|v_{k+1} - x^*\|^2\right] \\
&\leq \tfrac{1}{\theta_k^2}\mathbb{E}\left[\phi(x_k) - \phi^*\right] + \tfrac{1}{2\alpha}\left[\tilde{\iota}_0\tilde{\delta}_k + \tfrac{\tilde{\eta}_k^2}{2}\right]\mathbb{E}\left[\|v_k - x^*\|^2\right] + \tfrac{\tilde{\delta}_k^2 + \tilde{\iota}_0\tilde{\delta}_k}{2\theta_{k+1}^2 L} \\
&\quad + \tfrac{1}{2\alpha}\mathbb{E}\left[\|v_k - x^*\|^2 - \|v_{k+1} - x^*\|^2\right],
\end{aligned}$$

where the final inequality results from $\alpha = \{\alpha_k\} \leq \frac{1}{L}$. By Lemma A.4, with $R_k = \frac{1}{\theta_k^2}\mathbb{E}\left[\phi(x_k) - \phi^*\right]$, $T_k = \frac{1}{2\alpha}\mathbb{E}\left[\|v_k - x^*\|^2\right]$, $a_k = \tilde{\iota}_0\tilde{\delta}_k + \frac{\tilde{\eta}_k^2}{2}$, and $s_k = \frac{\tilde{\delta}_k^2 + \tilde{\iota}_0\tilde{\delta}_k}{2\theta_{k+1}^2 L}$, the sequence $\{R_k\}$ is bounded under the specified parameters provided that $\hat{\eta}$ and $\hat{\delta}$ are chosen to be sufficiently small. Thus, $\exists C \geq 0$ such that, $\mathbb{E}\left[\phi(x_k) - \phi^*\right] \leq C\theta_k^2 = \frac{4C}{(k+1)^2}$, completing the proof for the expectation problem (1.4) for Algorithm 2.1 with **Option II**. □

Theorem 3.8 establishes a sublinear rate of convergence for Algorithm 2.1 with **Option I** and **Option II** for a general convex objective function when using a biased gradient estimate satisfying Condition 2.1. Thus, an $\epsilon > 0$ accurate solution, i.e., $\phi(x_k) - \phi^* \leq \epsilon$ for the finite-sum problem (1.3) and $\mathbb{E}\left[\phi(x_k) - \phi^*\right] \leq \epsilon$ for the expectation problem (1.4), which are stronger definitions than those defined in Subsection 3.1, can be achieved in $\mathcal{O}\left(\frac{1}{\epsilon}\right)$ iterations (proximal operator evaluations) with

Option I and in $\mathcal{O}\left(\frac{1}{\sqrt{\epsilon}}\right)$ iterations (proximal operator evaluations) with **Option II**, matching the results for deterministic first-order methods [31].

When gradient estimate is unbiased, the required conditions on the parameters can be simplified for the expectation problem (1.4) as shown in the following corollary.

Corollary 3.9. *Suppose the conditions in Theorem 3.8 hold for the expectation problem (1.4) and the gradient estimate is unbiased, i.e.* $\mathbb{E}_k[g_k] = \nabla f(y_k)$.

1. *For Algorithm 2.1 with* ***Option I****, if the parameters in Condition 2.1 are chosen such that* $\{\tilde{\eta}_k\} = \tilde{\eta} \in [0, 1)$, $\tilde{\iota}_0^2 \in [0, 1 - \tilde{\eta})$, *and* $\sum_{k=0}^{\infty} \tilde{\delta}_k^2 < \infty$, *and the step size is chosen such that* $\{\alpha_k\} = \alpha \leq \frac{1-\tilde{\eta}-\tilde{\iota}_0^2}{L}$, *then* $\{\phi(x_k)\}$ *converges to the optimal value in expectation with* $\min_{k=0,\ldots,K-1} \mathbb{E}[\phi(x_k) - \phi^*] = \mathcal{O}\left(\frac{1}{K}\right)$ $\forall K \geq 1$.

2. *For Algorithm 2.1 with* ***Option II****, if the parameters in Condition 2.1 are chosen such that* $\{\tilde{\eta}_k\} = \tilde{\eta} \in [0, 1)$, $\tilde{\iota}_0^2 \in [0, 1 - \tilde{\eta})$, *and* $\sum_{k=0}^{\infty} \frac{\tilde{\delta}_k^2}{\theta_{k+1}^2} < \infty$, *and the step size is chosen such that* $\{\alpha_k\} = \alpha \leq \frac{1-\tilde{\eta}-\tilde{\iota}_0^2}{L}$, *then* $\{\phi(x_k)\}$ *converges to the optimal value in expectation with* $\mathbb{E}[\phi(x_K) - \phi^*] = \mathcal{O}\left(\frac{1}{K^2}\right)$ $\forall K \geq 1$.

Proof. For Algorithm 2.1 with **Option I**, consider the conditional expectation of the result in Lemma 3.7 given $\mathcal{G}_k$ under the defined parameters. From the assumption of an unbiased gradient estimate and substituting (3.7) from Lemma 3.2, the bound reduces to,

$$\begin{aligned}
\mathbb{E}_k\left[\phi(x_{k+1}) - \phi^*\right] &\leq \tfrac{\left(\tilde{\eta}+\tilde{\iota}_0^2\right)}{2\alpha}\mathbb{E}_k\left[\|x_k - x_{k+1}\|^2\right] + \tfrac{\alpha\tilde{\delta}_k^2}{2} \\
&\quad + \tfrac{1}{2\alpha}\left[\mathbb{E}_k\left[\|x_k - x^*\|^2\right] - \mathbb{E}_k\left[\|x_{k+1} - x^*\|^2\right] - (1 - \alpha L)\,\mathbb{E}_k\left[\|x_k - x_{k+1}\|^2\right]\right] \\
&= \tfrac{1}{2\alpha}\left[\mathbb{E}_k\left[\|x_k - x^*\|^2\right] - \mathbb{E}_k\left[\|x_{k+1} - x^*\|^2\right]\right] + \tfrac{\alpha\tilde{\delta}_k^2}{2} \\
&\quad - \tfrac{1}{2\alpha}\left[1 - \alpha L - \tilde{\eta} - \tilde{\iota}_0^2\right]\mathbb{E}_k\left[\|x_k - x_{k+1}\|^2\right],
\end{aligned}$$

where the constant $1 - \alpha L - \tilde{\eta} - \tilde{\iota}_0^2 \geq 0$ under the defined parameters. Taking a telescopic sum for $k = 0, \ldots, K-1$ of the total expectation of the reduced bound yields,

$$\sum_{k=0}^{K-1} \mathbb{E}\left[\phi(x_{k+1}) - \phi^*\right] \leq \tfrac{1}{2\alpha}\left[\mathbb{E}\left[\|x_0 - x^*\|^2\right] - \mathbb{E}\left[\|x_K - x^*\|^2\right]\right] + \sum_{k=0}^{K-1} \tfrac{\alpha\tilde{\delta}_k^2}{2}.$$

Rearranging the terms of the above inequality, we get

$$\min_{k=0,\ldots,K-1} \mathbb{E}\left[\phi(x_{k+1}) - \phi^*\right] \leq \tfrac{1}{K}\left\{\tfrac{1}{2\alpha}\mathbb{E}\left[\|x_0 - x^*\|^2\right] + \tfrac{\alpha}{2}\sum_{k=0}^{K-1} \tilde{\delta}_k^2\right\},$$

where all terms within the curly brackets on the right-hand side are bounded due to the condition $\sum_{k=0}^{\infty} \tilde{\delta}_k^2 < \infty$, completing the proof for **Option I**.

For Algorithm 2.1 with **Option II**, consider the conditional expectation of the result in Lemma 3.7 given $\mathcal{G}_k$ under the defined parameters. From the assumption of an unbiased gradient estimate and

substituting (3.7) from Lemma 3.2, the bound reduces to,

$$\begin{aligned}\mathbb{E}_k\left[\phi(x_{k+1})-\phi^*\right] &\leq (1-\theta_{k+1})(\phi(x_k)-\phi^*)+\tfrac{(\tilde{\eta}+\tilde{\iota}_0^2)}{2\alpha}\mathbb{E}_k\left[\|x_{k+1}-y_k\|^2\right]+\tfrac{\alpha\tilde{\delta}_k^2}{2}\\ &\quad+\tfrac{\theta_{k+1}^2}{2\alpha}\mathbb{E}_k\left[\|v_k-x^*\|^2-\|v_{k+1}-x^*\|^2-(1-\alpha L)\|v_k-v_{k+1}\|^2\right]\\ &=(1-\theta_{k+1})(\phi(x_k)-\phi^*)+\tfrac{\theta_{k+1}^2}{2\alpha}\mathbb{E}_k\left[\|v_k-x^*\|^2-\|v_{k+1}-x^*\|^2\right]+\tfrac{\alpha\tilde{\delta}_k^2}{2}\\ &\quad-\tfrac{\theta_{k+1}^2}{2\alpha}\left[1-\alpha L-\tilde{\eta}-\tilde{\iota}_0^2\right]\mathbb{E}_k\left[\|v_k-v_{k+1}\|^2\right]\end{aligned}$$

where the equality follows from (3.13) and the constant $1-\alpha L-\tilde{\eta}-\tilde{\iota}_0^2\geq 0$ under the defined parameters. Dividing the total expectation of the reduced bound by θ_{k+1}^2 and applying (3.14) yields,

$$\tfrac{1}{\theta_{k+1}^2}\mathbb{E}\left[\phi(x_{k+1})-\phi^*\right]\leq\tfrac{1}{\theta_k^2}\mathbb{E}\left[\phi(x_k)-\phi^*\right]+\tfrac{1}{2\alpha}\mathbb{E}\left[\|v_k-x^*\|^2-\|v_{k+1}-x^*\|^2\right]+\tfrac{\alpha\tilde{\delta}_k^2}{2\theta_{k+1}^2}.$$

Taking a telescoping sum of the above for $k=0,\ldots,K-1$ yields,

$$\tfrac{1}{\theta_K^2}\mathbb{E}\left[\phi(x_K)-\phi^*\right]\leq\tfrac{1}{\theta_0^2}\left[\phi(x_0)-\phi^*\right]+\tfrac{1}{2\alpha}\mathbb{E}\left[\|v_0-x^*\|^2-\|v_K-x^*\|^2\right]+\sum_{k=0}^{K-1}\tfrac{\alpha\tilde{\delta}_k^2}{2\theta_{k+1}^2}$$

Multiplying the above inequality by θ_K^2 yields,

$$\mathbb{E}\left[\phi(x_K)-\phi^*\right]\leq\tfrac{4}{(K+1)^2}\left\{\tfrac{1}{4}\left[\phi(x_0)-\phi^*\right]+\tfrac{1}{2\alpha}\mathbb{E}\left[\|v_0-x^*\|^2-\|v_K-x^*\|^2\right]+\sum_{k=0}^{K-1}\tfrac{\alpha\tilde{\delta}_k^2}{2\theta_{k+1}^2}\right\},$$

where all terms within the curly brackets on the right-hand side are bounded due to the defined $\tilde{\delta}_k$, completing the proof for **Option II**. □

Compared to Theorem 3.8, the parameter settings in Corollary 3.9 for Algorithm 2.1 and Condition 2.1 are less restrictive, as the use of an unbiased gradient estimate simplifies the analysis by reducing the number of error terms involved.

We now present the complexity for number of stochastic gradient evaluations for Algorithm 2.1 when an unbiased gradient estimate is used for the expectation problem (1.4) under Assumption 3.1.

Theorem 3.10. *Suppose Assumptions 2.2 and 3.1 hold, and Condition 2.1 is satisfied for the expectation problem (1.4) via the unbiased gradient estimate in Lemma 2.2. Then, to achieve a solution satisfying* $\min\left\{\mathbb{E}\left[\phi(x_k)-\phi^*\right],\left\|R_{\alpha_k}^{true}(y_k)\right\|^2\right\}\leq\epsilon$ *with* $\epsilon>0$*, the number of stochastic gradient evaluations required is as follows:*

1. *For Algorithm 2.1 with* ***Option I****, if the parameters in Condition 2.1 are chosen such that* $\{\tilde{\eta}_k\}=\tilde{\eta}\in(0,1)$*,* $\tilde{\iota}_0^2\in[0,1-\tilde{\eta})$*, and* $\tilde{\delta}_k^2=\frac{1}{(k+1)^{1+\nu}}\ \forall k\geq 0$*, where* $\nu>0$*, and the step size is chosen such that* $\{\alpha_k\}=\alpha\leq\frac{1-\tilde{\eta}-\tilde{\iota}_0^2}{L}$*, then the number of stochastic gradient evaluations required is* $\mathcal{O}\left(\epsilon^{-(2+\nu)}\right)$*. If* $\{\tilde{\delta}_k\}=0$*, this improves to* $\mathcal{O}\left(\epsilon^{-2}\right)$*.*

2. *For Algorithm 2.1 with* ***Option II****, if the parameters in Condition 2.1 are chosen such that* $\{\tilde{\eta}_k\}=\tilde{\eta}\in(0,1)$*,* $\tilde{\iota}_0^2\in[0,1-\tilde{\eta})$*, and* $\tilde{\delta}_k^2=\frac{1}{(k+1)^{3+2\nu}}\ \forall k\geq 0$*, where* $\nu>0$*, and the step size*

is chosen such that $\{\alpha_k\} = \alpha \leq \frac{1-\tilde{\eta}-\tilde{\iota}_0^2}{L}$*, then the number of stochastic gradient evaluations required is* $\mathcal{O}\left(\epsilon^{-(2+\nu)}\right)$*. If* $\{\tilde{\delta}_k\} = 0$*, this improves to* $\mathcal{O}\left(\epsilon^{-\frac{3}{2}}\right)$*.*

Proof. Let $K_\epsilon \geq 1$ be the first iteration that achieves the desired solution accuracy. Hence, $\left\|R_{\alpha_k}^{true}(y_k)\right\|^2 > \epsilon \; \forall k \leq K_\epsilon - 1$ and from (3.8), $\|\mathbb{E}_k\left[R_{\alpha_k}(y_k)\right]\|^2 > \left(1+\frac{\tilde{\eta}^2}{4}\right)^{-1}\left[\frac{\epsilon}{2} - \tilde{\iota}_0^2\tilde{\delta}_k^2\right] \; \forall k \leq K_\epsilon - 1$. Thus, the total number of stochastic gradient evaluations can be bounded using Lemma 2.2 as,

$$\sum_{k=0}^{K_\epsilon-1}|S_k| = \sum_{k=0}^{K_\epsilon-1}\left\lceil \frac{\sigma^2}{\frac{\tilde{\eta}_k^2}{4}\left\|\mathbb{E}_k\left[R_{\alpha_k}(y_k)\right]\right\|^2 + \tilde{\iota}_0^2\tilde{\delta}_k^2}\right\rceil \leq \sum_{k=0}^{K_\epsilon-1}\frac{2\sigma^2(4+\tilde{\eta}^2)}{\tilde{\eta}^2\epsilon + 8\tilde{\iota}_0^2\tilde{\delta}_k^2} + 1 \leq \sum_{k=0}^{K_\epsilon-1}\frac{2\sigma^2(4+\tilde{\eta}^2)}{\tilde{\eta}^2\epsilon} + \frac{\sigma^2(4+\tilde{\eta}^2)}{4\tilde{\iota}_0^2\tilde{\delta}_k^2} + K_\epsilon.$$

For Algorithm 2.1 with **Option I**, K_ϵ is at most $\mathcal{O}\left(\frac{1}{\epsilon}\right)$ from Corollary 3.9, yielding

$$\sum_{k=0}^{K_\epsilon-1}|S_k| \leq \frac{2\sigma^2(4+\tilde{\eta}^2)}{\tilde{\eta}^2\epsilon}K_\epsilon + \frac{\sigma^2(4+\tilde{\eta}^2)}{4\tilde{\iota}_0^2}K_\epsilon^{2+\nu} + K_\epsilon = \mathcal{O}\left(\frac{1}{\epsilon^{2+\nu}}\right).$$

Following the same procedure, if $\{\tilde{\delta}_k\} = 0$, $\sum_{k=0}^{K_\epsilon-1}|S_k| \leq \frac{2\sigma^2(4+\tilde{\eta}^2)}{\tilde{\eta}^2\epsilon}K_\epsilon + K_\epsilon = \mathcal{O}\left(\frac{1}{\epsilon^2}\right)$.

For Algorithm 2.1 with **Option II**, K_ϵ is at most $\mathcal{O}\left(\frac{1}{\sqrt{\epsilon}}\right)$ from Corollary 3.9, yielding

$$\sum_{k=0}^{K_\epsilon-1}|S_k| \leq \frac{2\sigma^2(4+\tilde{\eta}^2)}{\tilde{\eta}^2\epsilon}K_\epsilon + \frac{\sigma^2(4+\tilde{\eta}^2)}{4\tilde{\iota}_0^2}K_\epsilon^{4+2\nu} + K_\epsilon = \mathcal{O}\left(\frac{1}{\epsilon^{2+\nu}}\right).$$

Following the same procedure, if $\{\tilde{\delta}_k\} = 0$, $\sum_{k=0}^{K_\epsilon-1}|S_k| \leq \frac{2\sigma^2(4+\tilde{\eta}^2)}{\tilde{\eta}^2\epsilon}K_\epsilon + K_\epsilon = \mathcal{O}\left(\frac{1}{\epsilon^{3/2}}\right)$.

□

Theorem 3.10 matches the optimal complexity for the number of stochastic gradient evaluations for the expectation problem (1.4) with a general convex objective function [24]. We conclude this section with a corollary to Theorem 3.10, similar to Corollary 3.5, using a definition of an ϵ-accurate solution similar to that in [24], under the parameter setting $\{\tilde{\eta}_k\} = 0$.

Corollary 3.11. *Suppose the conditions in Theorem 3.10 hold. Then, to achieve a solution satisfying* $\mathbb{E}\left[\phi(x_k) - \phi^*\right] \leq \epsilon$ *with* $\epsilon > 0$*:*

1. *For Algorithm 2.1 with* ***Option I****, if the parameters in Condition 2.1 are chosen as* $\{\tilde{\eta}_k\} = 0$*,* $\tilde{\iota}_0^2 \in (0,1)$*, and* $\tilde{\delta}_k^2 = \frac{1}{(k+1)^{1+\nu}} \; \forall k \geq 0$*, where* $\nu > 0$*, and the step size is chosen such that* $\{\alpha_k\} = \alpha \leq \frac{1-\tilde{\iota}_0^2}{L}$*, then the number of stochastic gradient evaluations required is* $\mathcal{O}\left(\epsilon^{-(2+\nu)}\right)$*.*
2. *For Algorithm 2.1 with* ***Option II****, if the parameters in Condition 2.1 are chosen as* $\{\tilde{\eta}_k\} = 0$*,* $\tilde{\iota}_0^2 \in (0,1)$*, and* $\tilde{\delta}_k^2 = \frac{1}{(k+1)^{3+2\nu}} \; \forall k \geq 0$*, where* $\nu > 0$*, and the step size is chosen such that* $\{\alpha_k\} = \alpha \leq \frac{1-\tilde{\iota}_0^2}{L}$*, then the number of stochastic gradient evaluations required is* $\mathcal{O}\left(\epsilon^{-(2+\nu)}\right)$*.*

Proof. The proof follows from the same procedure as Theorem 3.10. □

The conditions in Corollary 3.11 reduce Condition 2.1 to using a predetermined error sequence for the gradient estimates, similar to [37].

3.3 Strongly Convex Objective Function

In this section, we present the theoretical analysis of Algorithm 2.1 when the smooth function $f(x)$, and thus the composite function $\phi(x)$, is strongly convex. We begin by stating the basic assumptions and definitions, along with some mathematical identities that will be used throughout the analysis.

Assumption 3.2. *The function $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth and μ-strongly convex, i.e.,*

$$f(\gamma a + (1-\gamma)b) \leq \gamma f(a) + (1-\gamma)f(b) - \tfrac{\mu}{2}\gamma(1-\gamma)\|a-b\|^2 \quad \forall a, b \in \mathbb{R}^d,\ \forall \gamma \in [0,1],$$

and the function $h : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is closed, convex, and proper.

Under Assumption 3.2, let $x^* \in \mathbb{R}^d$ be the unique optimal solution. Since $f(x)$ is differentiable and strongly convex, from [31],

$$f(b) \geq f(a) + \nabla f(a)^T(b-a) + \tfrac{\mu}{2}\|b-a\|^2 \quad \forall a, b \in \mathbb{R}^d. \tag{3.18}$$

For Algorithm 2.1 with **Option II**, under Assumption 3.2, we define the sequence $\{\beta_k\}$ and two additional sequences $\{\theta_k\}$ and $\{v_k\}$ for the analysis as,

$$\beta_k = \tfrac{1-\theta_k}{1+\theta_k} \quad \text{and} \quad \theta_k = \sqrt{\mu\alpha_k} \quad \forall k \geq 0, \text{ and} \quad v_k = x_{k-1} + \tfrac{1}{\theta_k}(x_k - x_{k-1}) \quad \forall k \geq 1, \tag{3.19}$$

with $v_0 = x_0$. Under these definitions, y_k can be expressed as,

$$\begin{aligned} y_k &= x_k + \tfrac{1-\theta_k}{1+\theta_k}(x_k - x_{k-1}) = \tfrac{1}{1+\theta_k}(x_k + \theta_k x_{k-1} + x_k - x_{k-1}) \\ &= \tfrac{x_k + \theta_k v_k}{1+\theta_k} \quad \forall k \geq 0. \end{aligned} \tag{3.20}$$

When a constant step size is employed in Algorithm 2.1, i.e., $\{\alpha_k\} = \alpha$ and $\{\theta_k\} = \theta = \sqrt{\mu\alpha}$, then using (3.19) and (3.20), the update form for $\{v_k\}$ can be expressed as,

$$\begin{aligned} v_{k+1} &= x_k + \tfrac{1}{\theta}(x_{k+1} - x_k) = v_k\left(1 + \tfrac{\theta^2}{1+\theta} - \theta - \tfrac{1}{1+\theta}\right) + x_k\left(\tfrac{\theta}{1+\theta} - \tfrac{1}{\theta(1+\theta)}\right) + x_{k+1}\left(\tfrac{1}{\theta}\right) \\ &= v_k + \theta\left(\tfrac{x_k + \theta v_k}{1+\theta} - v_k\right) - \tfrac{1}{\theta}\left(\tfrac{x_k + \theta v_k}{1+\theta} - x_{k+1}\right) \\ &= v_k + \theta(y_k - v_k) - \tfrac{1}{\theta}(y_k - x_{k+1}) \quad \forall k \geq 0. \end{aligned} \tag{3.21}$$

We now establish a descent lemma that further refines Lemma 3.1 under Assumption 3.2.

Lemma 3.12. *Suppose Assumption 3.2 holds. Then, $\forall x \in \mathbb{R}^d$ and $\forall k \geq 0$, the iterates generated by Algorithm 2.1 satisfy,*

$$\begin{aligned} \phi(x_{k+1}) \leq \phi(x) &+ \left[\tfrac{1}{2\alpha_k} - \tfrac{\mu}{2}\right]\|x - y_k\|^2 + (g_k - \nabla f(y_k))^T(x - y_k) \\ &+ (\nabla f(y_k) - g_k)^T(x_{k+1} - y_k) - \left(\tfrac{1}{2\alpha_k} - \tfrac{L}{2}\right)\|x_{k+1} - y_k\|^2. \end{aligned}$$

Proof. From (3.18), the first two terms in the result from Lemma 3.1 can be bounded as,

$$\begin{aligned} \phi(x_{k+1}) \leq f(x) &- \tfrac{\mu}{2}\|x - y_k\|^2 + (g_k - \nabla f(y_k))^T(x - y_k) + \tfrac{1}{2\alpha_k}\|x - y_k\|^2 + h(x) \\ &+ (\nabla f(y_k) - g_k)^T(x_{k+1} - y_k) - \left(\tfrac{1}{2\alpha_k} - \tfrac{L}{2}\right)\|x_{k+1} - y_k\|^2, \end{aligned}$$

completing the proof. □

While the result in Lemma 3.12 is sufficient for analyzing Algorithm 2.1 with **Option I**, a different Lyapunov function is required to analyze Algorithm 2.1 with **Option II** for a strongly convex objective function. We first establish a recursion for the distance between $\{v_k\}$ and the optimal solution x^* for Algorithm 2.1 with **Option II**.

Lemma 3.13. *Suppose Assumption 3.2 holds. Then, $\forall k \geq 0$, the iterates generated by Algorithm 2.1 with **Option II** and $\{\alpha_k\} = \alpha \leq \frac{1}{L}$ satisfy,*

$$\begin{aligned}\tfrac{\mu}{2}\|v_{k+1} - x^*\|^2 = \tfrac{\mu}{2}(1-\theta)\|v_k - x^*\|^2 - \tfrac{\mu}{2}\theta(1-\theta)\|v_k - y_k\|^2 + \tfrac{\mu\theta}{2}\|y_k - x^*\|^2 \\ + \tfrac{1}{2\alpha}\|y_k - x_{k+1}\|^2 + \tfrac{1}{\alpha}(y_k - x_{k+1})^T\left[\theta x^* + (1-\theta)x_k - y_k\right],\end{aligned}$$

where $\theta = \sqrt{\mu\alpha}$.

Proof. From (3.21),

$$v_{k+1} - x^* = v_k - x^* + \theta(y_k - v_k) - \tfrac{1}{\theta}(y_k - x_{k+1}) = (1-\theta)(v_k - x^*) + \theta(y_k - x^*) - \tfrac{1}{\theta}(y_k - x_{k+1}).$$

Taking the euclidean norm of the above and squaring both sides yields,

$$\begin{aligned}\|v_{k+1} - x^*\|^2 = (1-\theta)^2\|v_k - x^*\|^2 + \theta^2\|y_k - x^*\|^2 + \tfrac{1}{\theta^2}\|y_k - x_{k+1}\|^2 \\ + 2\theta(1-\theta)(v_k - x^*)^T(y_k - x^*) - \tfrac{2}{\theta}(y_k - x_{k+1})^T\left[(1-\theta)(v_k - x^*) + \theta(y_k - x^*)\right].\end{aligned}$$

Multiplying the above equality by $\frac{\mu}{2} = \frac{\theta^2}{2\alpha}$ and using $(1-\theta)^2 = 1 - \theta - \theta(1-\theta)$, we get,

$$\begin{aligned}\tfrac{\mu}{2}\|v_{k+1} - x^*\|^2 = \tfrac{\mu}{2}(1-\theta)\|v_k - x^*\|^2 - \tfrac{\mu}{2}\theta(1-\theta)\|v_k - x^*\|^2 + \tfrac{\mu\theta^2}{2}\|y_k - x^*\|^2 + \tfrac{1}{2\alpha}\|y_k - x_{k+1}\|^2 \\ + \mu\theta(1-\theta)(v_k - x^*)^T(y_k - x^*) - \tfrac{\theta}{\alpha}(y_k - x_{k+1})^T\left[(1-\theta)(v_k - x^*) + \theta(y_k - x^*)\right].\end{aligned} \tag{3.22}$$

The second, third and fifth terms on the right-hand side of (3.22) can be simplified together as,

$$\begin{aligned}&-\tfrac{\mu}{2}\theta(1-\theta)\|v_k - x^*\|^2 + \tfrac{\mu\theta^2}{2}\|y_k - x^*\|^2 + \mu\theta(1-\theta)(v_k - x^*)^T(y_k - x^*) \\ &= -\tfrac{\mu}{2}\theta(1-\theta)\|v_k - x^*\|^2 + \mu\theta(1-\theta)(v_k - x^*)^T(y_k - x^*) - \tfrac{\mu\theta(1-\theta)}{2}\|y_k - x^*\|^2 + \tfrac{\mu\theta}{2}\|y_k - x^*\|^2 \\ &= -\tfrac{\mu}{2}\theta(1-\theta)\|v_k - y_k\|^2 + \tfrac{\mu\theta}{2}\|y_k - x^*\|^2,\end{aligned}$$

where the second equality follows from $\theta^2 = \theta - \theta(1-\theta)$. The last term on the right-hand side of (3.22) can be simplified as,

$$\begin{aligned}&\tfrac{1}{\alpha}(y_k - x_{k+1})^T\left[\theta(1-\theta)(v_k - x^*) + \theta^2(y_k - x^*)\right] \\ &\quad= \tfrac{1}{\alpha}(y_k - x_{k+1})^T\left[\theta(1-\theta)\left(\tfrac{y_k(1+\theta) - x_k}{\theta} - x^*\right) + \theta^2(y_k - x^*)\right] \\ &\quad= \tfrac{1}{\alpha}(y_k - x_{k+1})^T\left[y_k - (1-\theta)x_k - \theta x^*\right],\end{aligned}$$

where the first equality follows from (3.20). Substituting these simplified expressions into (3.22) completes the proof. □

Using Lemma 3.13, we now establish the Lyapunov function and the recursive relation to analyze Algorithm 2.1 with **Option II** under Assumption 3.2.

Lemma 3.14. *Suppose Assumption 3.2 holds. Then, $\forall k \geq 0$, the iterates generated by Algorithm 2.1 with* ***Option II*** *and $\{\alpha_k\} = \alpha \leq \frac{1}{L}$ satisfy,*

$$\begin{aligned}
&\phi(x_{k+1}) - \phi^* + \tfrac{\mu}{2}\|v_{k+1} - x^*\|^2 \\
&\quad \leq (1-\theta)\left[\phi(x_k) - \phi^* + \tfrac{\mu}{2}\|v_k - x^*\|^2\right] - \tfrac{\mu}{2}\theta(1-\theta)\|v_k - y_k\|^2 - \left[\tfrac{1}{2\alpha} - \tfrac{L}{2}\right]\|x_{k+1} - y_k\|^2 \\
&\qquad + (g_k - \nabla f(y_k))^T(\theta x^* + (1-\theta)x_k - y_k) + (g_k - \nabla f(y_k))^T(y_k - x_{k+1}),
\end{aligned}$$

where $\theta = \sqrt{\mu\alpha}$.

Proof. From (2.1),

$$\begin{aligned}
\phi(x_{k+1}) &\leq f(y_k) + \nabla f(y_k)^T(x_{k+1} - y_k) + \tfrac{L}{2}\|x_{k+1} - y_k\|^2 + h(x_{k+1}) \\
&\leq f(y_k) + \nabla f(y_k)^T(x_{k+1} - y_k) + \tfrac{L}{2}\|x_{k+1} - y_k\|^2 + h(x) - \left(\tfrac{y_k - x_{k+1}}{\alpha_k} - g_k\right)^T(x - x_{k+1}) \\
&\leq f(x) - \nabla f(y_k)^T(x - y_k) - \tfrac{\mu}{2}\|x - y_k\|^2 + \nabla f(y_k)^T(x_{k+1} - y_k) + \tfrac{L}{2}\|x_{k+1} - y_k\|^2 \\
&\quad + h(x) - \left(\tfrac{y_k - x_{k+1}}{\alpha_k} - g_k\right)^T(x - x_{k+1}) \\
&= \phi(x) - \tfrac{\mu}{2}\|x - y_k\|^2 + \tfrac{L}{2}\|x_{k+1} - y_k\|^2 - \left(\tfrac{y_k - x_{k+1}}{\alpha_k}\right)^T(x - x_{k+1}) \\
&\quad - (\nabla f(y_k) - g_k)^T(x - x_{k+1})
\end{aligned}$$

where the second inequality follows, $\forall x \in \mathbb{R}^d$, from the convexity of $h(x)$ and the definition (2.2) for x_{k+1} which yields $0 \in g_k + \partial h(x_{k+1}) + \frac{x_{k+1} - y_k}{\alpha_k}$, and the third inequality follows from (3.18). Substituting $x = \theta x^* + (1-\theta)x_k$ where $\theta \in [0,1]$, from Assumption 3.2,

$$\begin{aligned}
\phi(x_{k+1}) - \phi^* \leq (1-\theta)\left[\phi(x_k) - \phi^*\right] - \tfrac{\mu}{2}\left[\theta(1-\theta)\|x_k - x^*\|^2 + \|x - y_k\|^2\right] \\
+ \tfrac{L}{2}\|x_{k+1} - y_k\|^2 - \left(\tfrac{y_k - x_{k+1}}{\alpha_k}\right)^T(x - x_{k+1}) - (\nabla f(y_k) - g_k)^T(x - x_{k+1}).
\end{aligned} \tag{3.23}$$

The second term on the right-hand side of (3.23) can be simplified as,

$$\begin{aligned}
&\theta(1-\theta)\|x_k - x^*\|^2 + \|x - y_k\|^2 \\
&\quad = \theta(1-\theta)\|x_k - x^*\|^2 + \|\theta x^* + (1-\theta)x_k - y_k\|^2 \\
&\quad = \theta(1-\theta)\|x_k - x^*\|^2 + (1-\theta)^2\|x_k - x^*\|^2 + \|y_k - x^*\|^2 + 2(1-\theta)(x_k - x^*)^T(x^* - y_k) \\
&\quad = (1-\theta)\|x_k - x^*\|^2 + \|y_k - x^*\|^2 + 2(1-\theta)(x_k - x^*)^T(x^* - y_k) \\
&\quad \geq (1-\theta)\|x_k - x^*\|^2 + \|y_k - x^*\|^2 - (1-\theta)\|x_k - x^*\|^2 - (1-\theta)\|y_k - x^*\|^2 \\
&\quad = \theta\|y_k - x^*\|^2,
\end{aligned}$$

where the inequality follows from the identity $2a^Tb \geq -\|a\|^2 - \|b\|^2$, $\forall a, b \in \mathbb{R}^d$. Substituting this bound into (3.23) and adding the result from Lemma 3.13, we get,

$$\begin{aligned}
&\phi(x_{k+1}) - \phi^* + \tfrac{\mu}{2}\|v_{k+1} - x^*\|^2 \\
&\quad \leq (1-\theta)\left[\phi(x_k) - \phi^* + \tfrac{\mu}{2}\|v_k - x^*\|^2\right] - \tfrac{\mu}{2}\theta(1-\theta)\|v_k - y_k\|^2 + \tfrac{L}{2}\|x_{k+1} - y_k\|^2 \\
&\qquad - \left(\tfrac{y_k - x_{k+1}}{\alpha_k}\right)^T(\theta x^* + (1-\theta)x_k - x_{k+1} - y_k + y_k) \\
&\qquad - (\nabla f(y_k) - g_k)^T(\theta x^* + (1-\theta)x_k - x_{k+1} - y_k + y_k) \\
&\qquad + \tfrac{1}{2\alpha}\|y_k - x_{k+1}\|^2 + \tfrac{1}{\alpha}(y_k - x_{k+1})^T\left[\theta x^* + (1-\theta)x_k - y_k\right],
\end{aligned}$$

where simplifying the expression completes the proof. □

We now present the convergence of Algorithm 2.1 under Assumption 3.2 when using a biased gradient estimate, and then discuss the corresponding complexity of number of proximal operator evaluations.

Theorem 3.15. *Suppose Assumption 3.2 holds and the gradient estimate g_k satisfies Condition 2.1.*

1. *For Algorithm 2.1 with* ***Option I****:*

   (a) *For the finite-sum problem (1.3), if the parameters in Condition 2.1 are chosen such that $\{\eta_k\} = \eta \in \left[0, \frac{1}{2}\right)$, $\iota_0 < \infty$, and $\delta_k = \delta^k \ \forall k \geq 0$, where $\delta \in [0, 1)$, and the step size is chosen such that $\{\alpha_k\} = \alpha \leq \frac{1-4\eta^2}{2L}$, then $\{x_k\}$ converges to x^* at a linear rate as,*

$$\phi(x_k) - \phi^* \leq \max\left\{1 - \tfrac{\mu\alpha}{3}, \delta^2\right\}^{k+1} \max\left\{\phi(x_0) - \phi^*, \tfrac{12\iota_0^2}{\mu}\right\}.$$

   (b) *For the expectation problem (1.4), if the parameters in Condition 2.1 are chosen such that $\{\tilde{\eta}_k\} = \tilde{\eta} \in \left[0, \frac{1}{\sqrt{2}}\right)$, $\tilde{\iota}_0 < \infty$, and $\tilde{\delta}_k = \tilde{\delta}^k \ \forall k \geq 0$, where $\tilde{\delta} \in [0, 1)$, and the step size is chosen such that $\{\alpha_k\} = \alpha \leq \frac{1-2\tilde{\eta}^2}{2L}$, then $\{x_k\}$ converges to x^* in expectation at a linear rate as,*

$$\mathbb{E}[\phi(x_k) - \phi^*] \leq \max\left\{1 - \tfrac{\mu\alpha}{3}, \tilde{\delta}^2\right\}^{k+1} \max\left\{\phi(x_0) - \phi^*, \tfrac{6\tilde{\iota}_0^2}{\mu}\right\}.$$

2. *For Algorithm 2.1 with* ***Option II****:*

   (a) *For the finite-sum problem (1.3), if the parameters in Condition 2.1 are chosen such that $\{\eta_k\} = \eta \leq \sqrt{\frac{\hat{c}}{4(L+\hat{c})}}$ where $\hat{c} = \frac{\mu}{4}(1 - \sqrt{\frac{\mu}{L}})$, $\iota_0 < \infty$, and $\delta_k = \delta^k \ \forall k \geq 0$, where $\delta \in [0, 1)$, and the step size is chosen as $\{\alpha_k\} = \alpha = \frac{1}{2(L+\hat{c})}$, then $\{x_k\}$ converges to x^* at a linear rate as,*

$$\phi(x_k) - \phi^* \leq \max\left\{1 - \tfrac{\sqrt{\alpha\mu}}{4}, \delta^2\right\}^{k+1} \max\left\{\phi(x_0) - \phi^* + \tfrac{\mu}{2}\|x_0 - x^*\|^2, \tfrac{8\iota_0^2}{\hat{c}\sqrt{\mu\alpha}}\right\}.$$

   (b) *For the expectation problem (1.4), if the parameters in Condition 2.1 are chosen such that $\{\tilde{\eta}_k\} = \tilde{\eta} \leq \sqrt{\frac{\hat{c}}{2(L+\hat{c})}}$ where $\hat{c} = \frac{\mu}{4}(1 - \sqrt{\frac{\mu}{L}})$, $\iota_0 < \infty$, and $\tilde{\delta}_k = \tilde{\delta}^k \ \forall k \geq 0$, where $\delta \in [0, 1)$, and the step size is chosen as $\{\alpha_k\} = \alpha = \frac{1}{2(L+\hat{c})}$, then $\{x_k\}$ converges to x^* in expectation at a linear rate as,*

$$\mathbb{E}\left[\phi(x_k) - \phi^*\right] \leq \max\left\{1 - \tfrac{\sqrt{\alpha\mu}}{4}, \tilde{\delta}^2\right\}^{k+1} \max\left\{\phi(x_0) - \phi^* + \tfrac{\mu}{2}\|x_0 - x^*\|^2, \tfrac{4\tilde{\iota}_0^2}{\hat{c}\sqrt{\mu\alpha}}\right\}.$$

Proof. For Algorithm 2.1 with **Option I**, consider the result in Lemma 3.12 with $y_k = x_k$. Applying young's inequality with constants $c_1, c_2 > 0$ as $(g_k - \nabla f(y_k))^T(x - y_k) \leq \frac{\alpha_k}{2c_1}\|g_k - \nabla f(x_k)\|^2 + \frac{c_1}{2\alpha_k}\|x -$

$x_k\|^2$ and $(\nabla f(y_k) - g_k)^T(x_{k+1} - y_k) \leq \frac{\alpha_k}{2c_2}\|\nabla f(x_k) - g_k\|^2 + \frac{c_2}{2\alpha_k}\|x_{k+1} - x_k\|^2$ yields,

$$\begin{aligned}\phi(x_{k+1}) &\leq \phi(x) + \left[\tfrac{1}{2\alpha_k} - \tfrac{\mu}{2}\right]\|x - x_k\|^2 + \tfrac{\alpha_k}{2c_1}\|g_k - \nabla f(x_k)\|^2 + \tfrac{c_1}{2\alpha_k}\|x - x_k\|^2 \\ &\quad + \tfrac{\alpha_k}{2c_2}\|\nabla f(x_k) - g_k\|^2 + \tfrac{c_2}{2\alpha_k}\|x_{k+1} - x_k\|^2 - \left(\tfrac{1}{2\alpha_k} - \tfrac{L}{2}\right)\|x_{k+1} - x_k\|^2 \\ &= \phi(x) + \tfrac{1}{2\alpha_k}\left[1 - \mu\alpha_k + c_1\right]\|x - x_k\|^2 + \tfrac{\alpha_k}{2}\left[\tfrac{1}{c_1} + \tfrac{1}{c_2}\right]\|g_k - \nabla f(x_k)\|^2 \\ &\quad - \tfrac{1}{2\alpha_k}\left(1 - \alpha_k L - c_2\right)\|x_{k+1} - x_k\|^2.\end{aligned}$$

Substituting $x = \nu_k x^* + (1-\nu_k)x_k$, where $\nu_k \in [0,1]$, setting $c_1 = c_2 = \frac{1}{2}$, and using Assumption 3.2 in the above inequality, we get,

$$\begin{aligned}\phi(x_{k+1}) - \phi^* &\leq (1-\nu_k)(\phi(x_k) - \phi^*) + \left[-\tfrac{\mu}{2}\nu_k(1-\nu_k) + \tfrac{\nu_k^2}{4\alpha_k}\left(3 - 2\mu\alpha_k\right)\right]\|x_k - x^*\|^2 \\ &\quad + 2\alpha_k\|g_k - \nabla f(x_k)\|^2 - \tfrac{1}{4\alpha_k}\left(1 - 2\alpha_k L\right)\|x_{k+1} - x_k\|^2.\end{aligned}$$

Substituting $\nu_k = \frac{2\mu\alpha_k}{3}$ reduces the bound to

$$\phi(x_{k+1}) - \phi^* \leq \left(1 - \tfrac{2\mu\alpha_k}{3}\right)(\phi(x_k) - \phi^*) + 2\alpha_k\|g_k - \nabla f(x_k)\|^2 - \tfrac{1}{4\alpha_k}\left(1 - 2\alpha_k L\right)\|x_{k+1} - x_k\|^2. \quad (3.24)$$

For the finite-sum problem (1.3) for Algorithm 2.1 with **Option I**, under the defined parameters, substituting (3.1) from Lemma 3.2 into (3.24) yields,

$$\begin{aligned}\phi(x_{k+1}) - \phi^* &\leq \left(1 - \tfrac{2\mu\alpha}{3}\right)(\phi(x_k) - \phi^*) + 2\alpha\left(\tfrac{\eta^2}{2\alpha^2}\|x_{k+1} - x_k\|^2 + 2\iota_0^2\delta^{2k}\right) \\ &\quad - \tfrac{1}{4\alpha}\left(1 - 2\alpha L\right)\|x_{k+1} - x_k\|^2 \\ &= \left(1 - \tfrac{2\mu\alpha}{3}\right)(\phi(x_k) - \phi^*) + 4\alpha\iota_0^2\delta^{2k} - \tfrac{1}{4\alpha}\left(1 - 2\alpha L - 4\eta^2\right)\|x_{k+1} - x_k\|^2,\end{aligned}$$

where the constant $1 - 2\alpha L - 4\eta^2 \geq 0$ under the defined parameters. Hence, the bound reduces to,

$$\phi(x_{k+1}) - \phi^* \leq \left(1 - \tfrac{2\mu\alpha}{3}\right)(\phi(x_k) - \phi^*) + 4\alpha\iota_0^2\delta^{2k},$$

where applying Lemma A.2 with $\omega = \frac{\mu\alpha}{3}$ completes the proof for the finite-sum problem (1.3) for Algorithm 2.1 with **Option I**.

For the expectation problem (1.4) for Algorithm 2.1 with **Option I**, consider the conditional expectation of (3.24) given $\mathcal{G}_k$. From Condition 2.1, under the defined parameters, we get,

$$\begin{aligned}\mathbb{E}_k\left[\phi(x_{k+1}) - \phi^*\right] &\leq \left(1 - \tfrac{2\mu\alpha}{3}\right)(\phi(x_k) - \phi^*) + 2\alpha\left(\tfrac{\tilde{\eta}^2}{4}\mathbb{E}_k\left[\|x_{k+1} - x_k\|^2\right] + \tilde{\iota}_0^2\tilde{\delta}^{2k}\right) \\ &\quad - \tfrac{1}{4\alpha}\left(1 - 2\alpha L\right)\mathbb{E}_k\left[\|x_{k+1} - x_k\|^2\right] \\ &\leq \left(1 - \tfrac{2\mu\alpha}{3}\right)(\phi(x_k) - \phi^*) + 2\alpha\tilde{\iota}_0^2\tilde{\delta}^{2k} - \tfrac{1}{4\alpha}\left(1 - 2\alpha L - 2\tilde{\eta}^2\right)\mathbb{E}_k\left[\|x_{k+1} - x_k\|^2\right],\end{aligned}$$

where the constant $1 - 2\alpha L - 2\tilde{\eta}^2 \geq 0$ under the defined parameters. Hence, the total expectation of the above bound yields,

$$\mathbb{E}\left[\phi(x_{k+1}) - \phi^*\right] \leq \left(1 - \tfrac{2\mu\alpha}{3}\right)\mathbb{E}\left[\phi(x_k) - \phi^*\right] + 2\alpha\tilde{\iota}_0^2\tilde{\delta}^{2k},$$

where applying Lemma A.2 with $\omega = \frac{\mu\alpha}{3}$ completes the proof for the expectation problem (1.4) for Algorithm 2.1 with **Option I**.

For Algorithm 2.1 with **Option II**, consider the result in Lemma 3.14. Substituting v_k from (3.20) yields,

$$\begin{aligned}
&\phi(x_{k+1}) - \phi^* + \tfrac{\mu}{2}\|v_{k+1} - x^*\|^2 \\
&\quad \leq (1-\theta)\left[\phi(x_k) - \phi^* + \tfrac{\mu}{2}\|v_k - x^*\|^2\right] - \tfrac{1}{2\alpha}\left[1 - \alpha L\right]\|x_{k+1} - y_k\|^2 \\
&\qquad - \tfrac{\mu}{2}\theta(1-\theta)\left\|\tfrac{(1+\theta)y_k - x_k}{\theta} - y_k\right\|^2 + (g_k - \nabla f(y_k))^T(y_k - x_{k+1}) \\
&\qquad + (g_k - \nabla f(y_k))^T(\theta x^* + y_k(1+\theta) - \theta v_k - \theta x_k - y_k).
\end{aligned} \tag{3.25}$$

The last three terms on the right-hand side of (3.25) can be simplified using young's inequality with constants $c_1, c_2 > 0$ as,

$$\begin{aligned}
&-\tfrac{\mu}{2\theta}(1-\theta)\|y_k - x_k\|^2 + (g_k - \nabla f(y_k))^T(y_k - x_{k+1}) + \theta(g_k - \nabla f(y_k))^T(y_k - x_k + x^* - v_k) \\
&\quad \leq -\tfrac{\mu}{2\theta}(1-\theta)\|y_k - x_k\|^2 + \tfrac{\alpha}{2c_1}\|g_k - \nabla f(y_k)\|^2 + \tfrac{c_1}{2\alpha}\|y_k - x_{k+1}\|^2 \\
&\qquad + \tfrac{\alpha}{2c_2}\|g_k - \nabla f(y_k)\|^2 + \tfrac{c_2\theta^2}{2\alpha}\|y_k - x_k + x^* - v_k\|^2 \\
&\quad \leq -\left[\tfrac{\mu}{2\theta}(1-\theta) - \tfrac{c_2\theta^2}{\alpha}\right]\|y_k - x_k\|^2 + \tfrac{\alpha}{2}\left[\tfrac{1}{c_1} + \tfrac{1}{c_2}\right]\|g_k - \nabla f(y_k)\|^2 \\
&\qquad + \tfrac{c_1}{2\alpha}\|y_k - x_{k+1}\|^2 + \tfrac{c_2\theta^2}{\alpha}\|v_k - x^*\|^2,
\end{aligned}$$

where the second inequality follows from the identity $\|a+b\|^2 \leq 2(\|a\|^2 + \|b\|^2)\ \forall a, b \in \mathbb{R}^d$. Setting $c_2 = \frac{\theta(1-\theta)}{4}$ yields the constant $\frac{\mu}{2\theta}(1-\theta) - \frac{c_2\theta^2}{\alpha} = \frac{\mu}{2\theta}(1-\theta) - \frac{\theta(1-\theta)\mu}{4} = \frac{\mu(1-\theta)}{2}\left[\frac{1}{\theta} - \frac{\theta}{2}\right] \geq 0$, since $\theta \in (0,1)$. Therefore, the corresponding term can be dropped, and (3.24) reduces to

$$\begin{aligned}
&\phi(x_{k+1}) - \phi^* + \tfrac{\mu}{2}\|v_{k+1} - x^*\|^2 \\
&\quad \leq (1-\theta)\left[\phi(x_k) - \phi^*\right] + \left[\tfrac{\mu(1-\theta)}{2} + \tfrac{c_2\theta^2}{\alpha}\right]\|v_k - x^*\|^2 \\
&\qquad - \tfrac{1}{2\alpha}\left[1 - \alpha L - c_1\right]\|x_{k+1} - y_k\|^2 + \tfrac{\alpha}{2}\left[\tfrac{1}{c_1} + \tfrac{1}{c_2}\right]\|g_k - \nabla f(y_k)\|^2 \\
&\quad \leq \left(1 - \tfrac{\theta}{2}\right)\left[\phi(x_k) - \phi^* + \tfrac{\mu}{2}\|v_k - x^*\|^2\right] - \tfrac{1}{2\alpha}\left[1 - \alpha L - c_1\right]\|x_{k+1} - y_k\|^2 \\
&\qquad + \tfrac{\alpha}{2}\left[\tfrac{1}{c_1} + \tfrac{1}{c_2}\right]\|g_k - \nabla f(y_k)\|^2,
\end{aligned} \tag{3.26}$$

where the second inequality follows from $\frac{\mu(1-\theta)}{2} + \frac{c_2\theta^2}{\alpha} = \frac{\mu(1-\theta)}{2} + \frac{\theta(1-\theta)\mu}{4} = \frac{\mu}{2}(1-\theta)\left(1 + \frac{\theta}{2}\right) \leq \frac{\mu}{2}\left(1 - \frac{\theta}{2}\right)$ since $\theta \in (0,1)$ and $\phi(x_k) - \phi^* \geq 0$.

For the finite-sum problem (1.3) for Algorithm 2.1 with **Option II**, substituting (3.1) into (3.26) under the defined parameters yields,

$$\begin{aligned}
&\phi(x_{k+1}) - \phi^* + \tfrac{\mu}{2}\|v_{k+1} - x^*\|^2 \\
&\quad \leq \left(1 - \tfrac{\theta}{2}\right)\left[\phi(x_k) - \phi^* + \tfrac{\mu}{2}\|v_k - x^*\|^2\right] - \tfrac{1}{2\alpha}\left[1 - \alpha L - c_1\right]\|x_{k+1} - y_k\|^2 \\
&\qquad + \tfrac{\alpha}{2}\left[\tfrac{1}{c_1} + \tfrac{1}{c_2}\right]\left[\tfrac{\eta^2}{2}\left\|\tfrac{x_{k+1} - y_k}{\alpha}\right\|^2 + 2\iota_0^2\delta^{2k}\right] \\
&\quad \leq \left(1 - \tfrac{\theta}{2}\right)\left[\phi(x_k) - \phi^* + \tfrac{\mu}{2}\|v_k - x^*\|^2\right] + \tfrac{\alpha}{2}\left[\tfrac{1}{c_1} + \tfrac{1}{c_2}\right]\left[2\iota_0^2\delta^{2k}\right] \\
&\qquad - \tfrac{1}{2\alpha}\left[1 - \alpha L - c_1 - \tfrac{\eta^2}{2}\left(\tfrac{1}{c_1} + \tfrac{1}{c_2}\right)\right]\|x_{k+1} - y_k\|^2.
\end{aligned}$$

With $c_1 = \alpha\hat{c}$ and previously defined $c_2 = \frac{\sqrt{\alpha\mu}(1-\sqrt{\alpha\mu})}{4} \geq \frac{\alpha\mu}{4}\left(1-\sqrt{\frac{\mu}{L}}\right) = \alpha\hat{c}$ under the defined parameters (since $\alpha \leq \frac{1}{L}$ and $\sqrt{\alpha\mu} \leq 1$), the constant multiplying the last term can be upper bounded as,

$$1 - \alpha L - c_1 - \frac{\eta^2}{2}\left(\frac{1}{c_1} + \frac{1}{c_2}\right) \geq 1 - \alpha(L+\hat{c}) - \frac{\eta^2}{\alpha\hat{c}} \geq 1 - \frac{(L+\hat{c})}{2(L+\hat{c})} - \frac{2(L+\hat{c})}{\hat{c}}\frac{\hat{c}}{4(L+\hat{c})} = 0.$$

Thus, the last term in the upper bound can be omitted, yielding,

$$\phi(x_{k+1}) - \phi^* + \tfrac{\mu}{2}\|v_{k+1} - x^*\|^2 \leq \left(1-\tfrac{\theta}{2}\right)\left[\phi(x_k) - \phi^* + \tfrac{\mu}{2}\|v_k - x^*\|^2\right] + \tfrac{2\iota_0^2\delta^{2k}}{\hat{c}},$$

where applying Lemma A.2 with $\omega = \frac{\theta}{4}$ completes the proof for the finite-sum problem (1.3) for Algorithm 2.1 with **Option II**.

For the expectation problem (1.4) for Algorithm 2.1 with **Option II**, consider the conditional expectation of (3.26) given $\mathcal{G}_k$. From Condition 2.1 under the defined parameters, we get,

$$\begin{aligned}
&\mathbb{E}_k\left[\phi(x_{k+1}) - \phi^* + \tfrac{\mu}{2}\|v_{k+1} - x^*\|^2\right] \\
&\quad\leq \left(1-\tfrac{\theta}{2}\right)\left[\phi(x_k) - \phi^* + \tfrac{\mu}{2}\|v_k - x^*\|^2\right] - \tfrac{1}{2\alpha}\left[1 - \alpha L - c_1\right]\mathbb{E}_k\left[\|x_{k+1} - y_k\|^2\right] \\
&\qquad + \tfrac{\alpha}{2}\left[\tfrac{1}{c_1} + \tfrac{1}{c_2}\right]\left[\tfrac{\tilde{\eta}^2}{4\alpha^2}\mathbb{E}_k\left[\|x_{k+1} - y_k\|^2\right] + \tilde{\iota}_0^2\tilde{\delta}^{2k}\right] \\
&\quad\leq \left(1-\tfrac{\theta}{2}\right)\left[\phi(x_k) - \phi^* + \tfrac{\mu}{2}\|v_k - x^*\|^2\right] + \tfrac{\alpha}{2}\left[\tfrac{1}{c_1} + \tfrac{1}{c_2}\right]\left[\tilde{\iota}_0^2\tilde{\delta}^{2k}\right] \\
&\qquad - \tfrac{1}{2\alpha}\left[1 - \alpha L - c_1 - \tfrac{\tilde{\eta}^2}{4}\left(\tfrac{1}{c_1} + \tfrac{1}{c_2}\right)\right]\mathbb{E}_k\left[\|x_{k+1} - y_k\|^2\right].
\end{aligned}$$

With $c_1 = \alpha\hat{c}$ and previously defined $c_2 = \frac{\sqrt{\alpha\mu}(1-\sqrt{\alpha\mu})}{4} \geq \frac{\alpha\mu}{4}\left(1-\sqrt{\frac{\mu}{L}}\right) = \alpha\hat{c}$ under the defined parameters, the constant for the last term can be upper bounded as,

$$1 - \alpha L - c_1 - \frac{\tilde{\eta}^2}{4}\left(\frac{1}{c_1} + \frac{1}{c_2}\right) \geq 1 - \alpha(L+\hat{c}) - \frac{\tilde{\eta}^2}{2\alpha\hat{c}} \geq 1 - \frac{(L+\hat{c})}{2(L+\hat{c})} - \frac{2(L+\hat{c})}{2\hat{c}}\frac{\hat{c}}{2(L+\hat{c})} = 0.$$

Thus, the last term in the upper bound can be omitted, yielding,

$$\mathbb{E}\left[\phi(x_{k+1}) - \phi^* + \tfrac{\mu}{2}\|v_{k+1} - x^*\|^2\right] \leq \left(1-\tfrac{\theta}{2}\right)\mathbb{E}\left[\phi(x_k) - \phi^* + \tfrac{\mu}{2}\|v_k - x^*\|^2\right] + \tfrac{\tilde{\iota}_0^2\tilde{\delta}^{2k}}{\hat{c}},$$

where applying Lemma A.2 with $\omega = \frac{\theta}{4}$ completes the proof for the expectation problem (1.4) for Algorithm 2.1 with **Option II**. □

Theorem 3.15 establishes linear rate of convergence for Algorithm 2.1 with **Option I** and **Option II** for strongly convex objective functions when using a biased gradient estimate satisfying Condition 2.1. Thus, an $\epsilon > 0$ accurate solution, with the same definition as in Subsection 3.2, can be achieved in $\mathcal{O}\left(\kappa\log\frac{1}{\epsilon}\right)$ iterations (proximal operator evaluations) with **Option I** and in $\mathcal{O}\left(\sqrt{\kappa}\log\frac{1}{\epsilon}\right)$ iterations (proximal operator evaluations) with **Option II**, where $\kappa = \frac{L}{\mu}$ is the condition number, matching the results for deterministic first-order methods [31]. We note that the presented analysis is significantly simpler than that of [37], where the authors analyze accelerated proximal gradient methods with predetermined deterministic errors in the gradient estimate and the solution to the proximal operator. In [37], the analysis establishes that the sequences $\{x_k\}$ and $\{v_k\}$ remain within a finite distance of x^* to accommodate the errors resulting from using a biased gradient estimate. In contrast, the adaptive nature of the error in our gradient estimates allows us to incorporate this error directly into the Lyapunov function, thereby simplifying the analysis.

The parameter settings required in Theorem 3.15 can be simplified when the gradient estimate for the expectation problem (1.4) is unbiased, as shown in the following corollary.

Corollary 3.16. *Suppose the conditions in Theorem 3.15 hold for the expectation problem (1.4) and the gradient estimate is unbiased, i.e.* $\mathbb{E}_k[g_k] = \nabla f(y_k)$.

1. *For Algorithm 2.1 with* ***Option I****, if the parameters in Condition 2.1 are chosen such that* $\{\tilde{\eta}_k\} = \tilde{\eta} \in [0,1)$ *and* $\tilde{\delta}_k = \tilde{\delta}^k \ \forall k \geq 0$*, where* $\tilde{\delta} \in [0,1)$*, and the step size is chosen such that* $\{\alpha_k\} = \alpha \leq \frac{2-\tilde{\eta}^2}{2L}$*, then* $\{x_k\}$ *converges to* x^* *in expectation at a linear rate as,*

$$\mathbb{E}[\phi(x_k) - \phi^*] \leq \max\left\{1 - \tfrac{\mu\alpha}{2}, \tilde{\delta}^2\right\}^{k+1} \max\left\{\phi(x_0) - \phi^*, \tfrac{2\tilde{\iota}_0^2}{\mu}\right\}.$$

2. *For Algorithm 2.1 with* ***Option II****, if the parameters in Condition 2.1 are chosen such that* $\{\tilde{\eta}_k\} = \tilde{\eta} \in [0,1)$*,* $\tilde{\iota}_0^2 \in [0, \tilde{\eta})$*,* $\tilde{\delta}_k = \tilde{\delta}^k \ \forall k \geq 0$*, where* $\tilde{\delta} \in [0,1)$*, and the step size is chosen such that* $\{\alpha_k\} = \alpha \leq \frac{1-\tilde{\eta}-\tilde{\iota}_0^2}{L}$*, then* $\{x_k\}$ *converges to* x^* *in expectation at a linear rate as,*

$$\mathbb{E}\left[\phi(x_k) - \phi(x^*)\right] \leq \max\left\{1 - \tfrac{\sqrt{\mu\alpha}}{2}, \tilde{\delta}^2\right\}^{k+1} \max\left\{\phi(x_0) - \phi^* + \tfrac{\mu}{2}\|x_0 - x^*\|^2, \sqrt{\tfrac{\alpha}{\mu}}\right\}.$$

Proof. For Algorithm 2.1 with **Option I**, consider the result in Lemma 3.12 with $y_k = x_k$. Under the defined parameters and $x = \mu\alpha x^* + (1-\mu\alpha)x_k$ where $\mu\alpha \leq 1$, using Assumption 3.2, we get,

$$\begin{aligned}
\phi(x_{k+1}) - \phi^* &\leq (1-\mu\alpha)(\phi(x_k) - \phi^*) - \tfrac{\mu^2\alpha(1-\mu\alpha)}{2}\|x_k - x^*\|^2 + \tfrac{\mu^2\alpha}{2}(1-\mu\alpha)\|x_k - x^*\|^2 \\
&\quad + \mu\alpha(g_k - \nabla f(x_k))^T(x^* - x_k) + (\nabla f(x_k) - g_k)^T(x_{k+1} - x_k) - \tfrac{1}{2\alpha}(1-\alpha L)\|x_{k+1} - x_k\|^2 \\
&= (1-\mu\alpha)(\phi(x_k) - \phi^*) + \mu\alpha(g_k - \nabla f(x_k))^T(x^* - x_k) \\
&\quad + (\nabla f(x_k) - g_k)^T(x_{k+1} - x_k) - \tfrac{1}{2\alpha}(1-\alpha L)\|x_{k+1} - x_k\|^2.
\end{aligned}$$

Taking conditional expectation of the above inequality given $\mathcal{G}_k$, the second term on the right-hand side is zero due to the assumption of an unbiased gradient estimate, yielding,

$$\begin{aligned}
\mathbb{E}_k\left[\phi(x_{k+1}) - \phi^*\right] &\leq (1-\mu\alpha)(\phi(x_k) - \phi^*) + \mathbb{E}_k\left[(\nabla f(x_k) - g_k)^T(x_{k+1} - x_k)\right] \\
&\quad - \tfrac{1}{2\alpha}(1-\alpha L)\,\mathbb{E}_k\left[\|x_{k+1} - x_k\|^2\right] \\
&\leq (1-\mu\alpha)(\phi(x_k) - \phi^*) + \tfrac{\tilde{\eta}^2}{4\alpha}\mathbb{E}_k\left[\|x_{k+1} - x_k\|^2\right] + \alpha\tilde{\iota}_0^2\tilde{\delta}^{2k} \\
&\quad - \tfrac{1}{2\alpha}(1-\alpha L)\,\mathbb{E}_k\left[\|x_{k+1} - x_k\|^2\right] \\
&= (1-\mu\alpha)(\phi(x_k) - \phi^*) + \alpha\tilde{\iota}_0^2\tilde{\delta}^{2k} - \tfrac{1}{2\alpha}\left(1 - \alpha L - \tfrac{\tilde{\eta}^2}{2}\right)\mathbb{E}_k\left[\|x_{k+1} - x_k\|^2\right],
\end{aligned}$$

where the second inequality follows from Cauchy-Schwartz inequality and Condition 2.1 as,

$$\begin{aligned}
&\mathbb{E}_k[(\nabla f(x_k) - g_k)^T(x_{k+1} - x_k)] \\
&\quad = \mathbb{E}_k[(\nabla f(x_k) - g_k)^T(x_{k+1} - x_k - \hat{x}_{k+1} + \hat{x}_{k+1})] = \mathbb{E}_k[(\nabla f(x_k) - g_k)^T(x_{k+1} - \hat{x}_{k+1})] \\
&\quad \leq \mathbb{E}_k[\|\nabla f(x_k) - g_k\|\|x_{k+1} - \hat{x}_{k+1}\|] = \alpha\mathbb{E}_k[\|\nabla f(x_k) - g_k\|\|R_\alpha^{true}(x_k) - R_\alpha(x_k)\|]] \\
&\quad \leq \alpha\mathbb{E}_k[\|\nabla f(x_k) - g_k\|^2] \leq \tfrac{\tilde{\eta}^2}{4\alpha}\mathbb{E}_k\left[\|x_{k+1} - x_k\|^2\right] + \alpha\tilde{\iota}_0^2\tilde{\delta}^{2k}.
\end{aligned}$$

Under the defined parameters, $1 - \alpha L - \frac{\tilde{\eta}^2}{2} \geq 0$. Hence, the total expectation of the bound yields,

$$\mathbb{E}\left[\phi(x_{k+1}) - \phi^*\right] \leq (1-\mu\alpha)\mathbb{E}\left[\phi(x_k) - \phi^*\right] + \alpha\tilde{\iota}_0^2\tilde{\delta}^{2k},$$

where applying Lemma A.2 with $\omega = \frac{\mu\alpha}{2}$ completes the proof for **Option I**.

For Algorithm 2.1 with **Option II**, consider the result in Lemma 3.14 while simplifying the upper bound by ignoring the negative term $-\frac{\mu}{2}\theta(1-\theta)\|v_k - y_k\|^2$. Taking conditional expectation given $\mathcal{G}_k$, from the assumption of an unbiased gradient estimate, the bound reduces to,

$$\begin{aligned}
&\mathbb{E}_k\left[\phi(x_{k+1}) - \phi(x^*) + \tfrac{\mu}{2}\|v_{k+1} - x^*\|^2\right] \\
&\quad\leq (1-\theta)\left[\phi(x_k) - \phi(x^*) + \tfrac{\mu}{2}\|v_k - x^*\|^2\right] - \tfrac{1}{2\alpha}(1-\alpha L)\,\mathbb{E}_k\left[\|x_{k+1} - y_k\|^2\right] \\
&\qquad + \mathbb{E}_k\left[(g_k - \nabla f(y_k))^T(y_k - x_{k+1})\right] \\
&\quad\leq (1-\theta)\left[\phi(x_k) - \phi(x^*) + \tfrac{\mu}{2}\|v_k - x^*\|^2\right] - \tfrac{1}{2\alpha}(1-\alpha L)\,\mathbb{E}_k\left[\|x_{k+1} - y_k\|^2\right] \\
&\qquad + \tfrac{\alpha(\tilde{\eta}+\tilde{\iota}_0^2)}{2}\mathbb{E}_k\left[\|R_\alpha(y_k)\|^2\right] + \tfrac{\alpha\tilde{\delta}^{2k}}{2} \\
&\quad= (1-\theta)\left[\phi(x_k) - \phi(x^*) + \tfrac{\mu}{2}\|v_k - x^*\|^2\right] + \tfrac{\alpha\tilde{\delta}^{2k}}{2} - \tfrac{\alpha}{2}(1 - \alpha L - \tilde{\eta} - \tilde{\iota}_0^2)\mathbb{E}_k\left[\|R_\alpha(y_k)\|^2\right],
\end{aligned}$$

where the second inequality follows from (3.7) from Lemma 3.2. Under the defined parameters, $1 - \alpha L - \tilde{\eta}_k - \tilde{\iota}_0^2 \geq 0$. Hence, the total expectation of the bound yields,

$$\mathbb{E}\left[\phi(x_{k+1}) - \phi(x^*) + \tfrac{\mu}{2}\|v_{k+1} - x^*\|^2\right] \leq (1-\theta)\mathbb{E}\left[\phi(x_k) - \phi(x^*) + \tfrac{\mu}{2}\|v_k - x^*\|^2\right] + \tfrac{\alpha\tilde{\delta}^{2k}}{2},$$

where applying Lemma A.2 with $\omega = \frac{\theta}{2}$ completes the proof for **Option II**. □

The parameter settings in Corollary 3.16 are less restrictive compared to those in Theorem 3.15, due to the unbiased nature of the gradient approximation, particularly with respect to the step size in **Option II**.

We now present the complexity for the number of stochastic gradient evaluations for Algorithm 2.1 when an unbiased gradient estimate is used for the expectation problem (1.4) under Assumption 3.2. Unlike in Subsection 3.1 and Subsection 3.2, the parameter setting $\{\eta_k\} = 0$ is incorporated directly into the next theorem, as the optimal complexity for stochastic gradient evaluations is achieved under this parameter setting, along with a definition of an ϵ-accurate solution similar to that in [16].

Theorem 3.17. *Suppose Assumptions 2.2 and 3.2 hold, and Condition 2.1 is satisfied for the expectation problem (1.4) via the unbiased gradient estimate in Lemma 2.2. Then, to achieve a solution satisfying* $\min\left\{\mathbb{E}\left[\phi(x_k) - \phi^*\right], \left\|R_{\alpha_k}^{true}(y_k)\right\|^2\right\} \leq \epsilon$ *with* $\epsilon > 0$*, the number of stochastic gradient evaluations required is as follows:*

1. *For Algorithm 2.1 with* ***Option I****, if the parameters in Condition 2.1 are chosen such that* $\{\tilde{\eta}_k\} = \tilde{\eta} \in [0,1)$ *and* $\tilde{\delta}_k = \tilde{\delta}^k\ \forall k \geq 0$*, where* $\tilde{\delta}^2 = 1 - \frac{\mu\alpha}{2}$*,* $\iota_0 > 0$*, and the step size is chosen such that* $\{\alpha_k\} = \alpha \leq \frac{2-\tilde{\eta}^2}{2L}$*, then the number of stochastic gradient evaluations required is* $\mathcal{O}\left(\frac{\kappa}{\epsilon}\log\left(\frac{1}{\epsilon}\right)\right)$*, where* $\kappa = \frac{L}{\mu}$*. If* $\{\tilde{\eta}_k\} = 0$*, then the number of stochastic gradient evaluations reduces to* $\mathcal{O}\left(\frac{\kappa}{\epsilon}\right)$ *to achieve a solution satisfying* $\mathbb{E}\left[\phi(x_k) - \phi^*\right] \leq \epsilon$ *with* $\epsilon > 0$*.*
2. *For Algorithm 2.1 with* ***Option II****, if the parameters in Condition 2.1 are chosen such that* $\{\tilde{\eta}_k\} = \tilde{\eta} \in [0,1)$*,* $\tilde{\iota}_0^2 \in (0, \tilde{\eta})$*,* $\tilde{\delta}_k = \tilde{\delta}^k\ \forall k \geq 0$*, where* $\tilde{\delta}^2 = 1 - \frac{\sqrt{\mu\alpha}}{2}$*, and the step size is chosen such that* $\{\alpha_k\} = \alpha \leq \frac{1-\tilde{\eta}-\tilde{\iota}_0^2}{L}$*, then the number of stochastic gradient evaluations required is* $\mathcal{O}\left(\frac{\sqrt{\kappa}}{\epsilon}\log\left(\frac{1}{\epsilon}\right)\right)$*, where* $\kappa = \frac{L}{\mu}$*. If* $\{\tilde{\eta}_k\} = 0$*, then the number of stochastic gradient evaluations reduces to* $\mathcal{O}\left(\frac{\sqrt{\kappa}}{\epsilon}\right)$ *to achieve a solution satisfying* $\mathbb{E}\left[\phi(x_k) - \phi^*\right] \leq \epsilon$ *with* $\epsilon > 0$*.*

Proof. Let $K_\epsilon \geq 1$ be the first iteration that achieves the desired solution accuracy. Hence, $\left\|R^{true}_{\alpha_k}(y_k)\right\|^2 > \epsilon \; \forall k \leq K_\epsilon - 1$ and from (3.8), $\left\|\mathbb{E}_k\left[R_{\alpha_k}(y_k)\right]\right\|^2 > \left(1+\frac{\tilde{\eta}^2}{4}\right)^{-1}\left[\frac{\epsilon}{2} - \tilde{\iota}_0^2\tilde{\delta}_k^2\right] \; \forall k \leq K_\epsilon - 1$. Thus, the total number of stochastic gradient evaluations can be bounded using Lemma 2.2 as,

$$\sum_{k=0}^{K_\epsilon-1}|S_k| = \sum_{k=0}^{K_\epsilon-1}\left\lceil \frac{\sigma^2}{\frac{\tilde{\eta}_k^2}{4}\left\|\mathbb{E}_k\left[R_{\alpha_k}(y_k)\right]\right\|^2 + \tilde{\iota}_0^2\tilde{\delta}_k^2}\right\rceil \leq \sum_{k=0}^{K_\epsilon-1}\frac{2\sigma^2(4+\tilde{\eta}^2)}{\tilde{\eta}^2\epsilon + 8\tilde{\iota}_0^2\tilde{\delta}_k^2} + 1 \leq \sum_{k=0}^{K_\epsilon-1}\frac{2\sigma^2(4+\tilde{\eta}^2)}{\tilde{\eta}^2\epsilon} + \frac{\sigma^2(4+\tilde{\eta}^2)}{4\tilde{\iota}_0^2\tilde{\delta}_k^2} + K_\epsilon.$$

For Algorithm 2.1 with **Option I**, K_ϵ is at most $\mathcal{O}\left(\log_{\frac{1}{\tilde{\delta}^2}}\left(\frac{1}{\epsilon}\right)\right)$ from Corollary 3.16, yielding,

$$\sum_{k=0}^{K_\epsilon-1}|S_k| \leq \frac{2\sigma^2(4+\tilde{\eta}^2)}{\tilde{\eta}^2\epsilon}K_\epsilon + \frac{\sigma^2(4+\tilde{\eta}^2)}{4\tilde{\iota}_0^2}\frac{\frac{1}{\tilde{\delta}^{2(K_\epsilon+1)}}-1}{\frac{1}{\tilde{\delta}^2}-1} + K_\epsilon = \mathcal{O}\left(\tfrac{\kappa}{\epsilon}\log\left(\tfrac{1}{\epsilon}\right)\right).$$

Following the same procedure, if $\{\tilde{\eta}_k\} = 0$, a solution satisfying $\mathbb{E}\left[\phi(x_k) - \phi^*\right] \leq \epsilon$ with $\epsilon > 0$ is achieved in $\tilde{K}_\epsilon = \mathcal{O}\left(\log_{\frac{1}{\tilde{\delta}^2}}\left(\frac{1}{\epsilon}\right)\right)$ iterations from Corollary 3.16, and the total number of stochastic gradient evaluations is $\sum_{k=0}^{\tilde{K}_\epsilon-1}|S_k| \leq \frac{\sigma^2}{\tilde{\iota}_0^2}\frac{\frac{1}{\tilde{\delta}^{2(K_\epsilon+1)}}-1}{\frac{1}{\tilde{\delta}^2}-1} + \tilde{K}_\epsilon = \mathcal{O}\left(\frac{\kappa}{\epsilon}\right)$.

For Algorithm 2.1 with **Option II**, K_ϵ is at most $\mathcal{O}\left(\log_{\frac{1}{\tilde{\delta}^2}}\left(\frac{1}{\epsilon}\right)\right)$ from Corollary 3.16, yielding,

$$\sum_{k=0}^{K_\epsilon-1}|S_k| \leq \frac{2\sigma^2(4+\tilde{\eta}^2)}{\tilde{\eta}^2\epsilon}K_\epsilon + \frac{\sigma^2(4+\tilde{\eta}^2)}{4\tilde{\iota}_0^2}\frac{\frac{1}{\tilde{\delta}^{2(K_\epsilon+1)}}-1}{\frac{1}{\tilde{\delta}^2}-1} + K_\epsilon = \mathcal{O}\left(\tfrac{\sqrt{\kappa}}{\epsilon}\log\left(\tfrac{1}{\epsilon}\right)\right).$$

Following the same procedure, if $\{\tilde{\eta}_k\} = 0$, a solution satisfying $\mathbb{E}\left[\phi(x_k) - \phi^*\right] \leq \epsilon$ with $\epsilon > 0$ is achieved in $\tilde{K}_\epsilon = \mathcal{O}\left(\log_{\frac{1}{\tilde{\delta}^2}}\left(\frac{1}{\epsilon}\right)\right)$ iterations from Corollary 3.16, and the total number of stochastic gradient evaluations is $\sum_{k=0}^{\tilde{K}_\epsilon-1}|S_k| \leq \frac{\sigma^2}{\tilde{\iota}_0^2}\frac{\frac{1}{\tilde{\delta}^{2(K_\epsilon+1)}}-1}{\frac{1}{\tilde{\delta}^2}-1} + \tilde{K}_\epsilon = \mathcal{O}\left(\frac{\sqrt{\kappa}}{\epsilon}\right)$.

□

Theorem 3.17 matches the optimal complexity for the number of stochastic gradient evaluations for the expectation problem (1.4) with a strongly convex objective function [16]. Unlike the cases of nonconvex and general convex objective functions, a predefined sequence of decreasing gradient errors results in the optimal complexity for strongly convex objective functions.

4 Numerical Experiments

In this section, we illustrate the performance of Algorithm 2.1 on a synthetic strongly convex quadratic problem of the form

$$\min_{x\in\mathbb{R}^{10}} \; \phi(x) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{2}x^TQ_ix + b_i^Tx + \mathcal{I}_{[\|x\|^2\leq 1]}, \tag{4.1}$$

where $Q_i \in \mathbb{R}^{10\times 10}$ is positive definite ($Q_i \succ 0$) and $b_i \in \mathbb{R}^{10}$ are samples from a finite dataset with $N = 10^5$, generated via the process outlined in [28] with condition number $\kappa \approx 10^4$. The problem is constrained to the convex feasible region described by $\|x\|^2 \leq 1$, using the indicator function of a set as

$$\mathcal{I}_{[\|x\|^2\leq 1]} = \begin{cases} 0 & \text{if } \|x\|^2 \leq 1, \\ \infty & \text{otherwise.} \end{cases}$$

Thus, the objective function is strongly convex[1].

The gradient estimate g_k used in Algorithm 2.1 is a sample average approximation with sample set $S_k \subseteq \mathcal{S}$ $\forall k \geq 0$. We present results for five different sample selection strategies (S_k). The "Deterministic" label corresponds to using the true problem gradient, i.e., the full dataset. The "Stochastic" label corresponds to using 256 samples every iteration, i.e., $|S_k| = 256$ $\forall k \geq 0$. The "Geometric" label corresponds to starting from 32 samples and increasing the number of samples by 5% each iteration, i.e., $|S_{k+1}| = \lceil 1.05|S_k| \rceil$ $\forall k \geq 0$, equivalent to setting $\{\tilde{\eta}_k\} = 0$ in Condition 2.1. The "Adaptive" label corresponds to using Condition 2.1 to control the accuracy of the gradient estimate, with $\{\tilde{\eta}_k\} = \tilde{\eta} = 0.1$, $\iota_0 = 0$, and the sample size selected using sampled estimates as described in [43]. The "Stochastic", "Geometric" and "Adaptive" strategies approximate problem (4.1) as an expectation problem over a uniform distribution, as done in [20, 34, 43], and use an unbiased gradient estimate by sampling each iteration with replacement independent of the current iterate. The "Adaptive-biased" label corresponds to using Condition 2.1 to control the accuracy of the gradient estimate by determining the sample size with the same parameters, while introducing bias by maintaining $S_k \subseteq S_{k+1}$ $\forall k \geq 0$. We evaluate both **Option I** (Proximal Gradient) and **Option II** (Accelerated Proximal Gradient) for each sample selection strategy, with results labeled using "-I" and "-II", respectively. For **Option II**, we set $\beta_k = \beta = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} \approx 0.98$. The step size $\{\alpha_k\} = \alpha$ was tuned for each result over the set $\{10^{-i} | i = 0, 1, \ldots, 6\}$. The performance was measured by the optimality gap in function value, evaluated against both the number of proximal operator evaluations and the number of stochastic gradient evaluations.

When Condition 2.1 is used to control the accuracy of the gradient estimate, the method achieves one of the most efficient performances in terms of gradient evaluations, while also outperforming constant sample size strategies in terms of proximal operator evaluations. Finally, although introducing bias in the gradient estimate degrades performance, the algorithm still converges and performs well under **Option II**.

The results are summarized in Fig. 1. First, for all sample selection strategies, **Option II** yields better performance with respect to number of proximal operator evaluations during the initial phase. Second, a stochastic gradient estimate is initially more efficient than the deterministic approach with respect to the number of gradient evaluations, but less efficient in terms of number of proximal operator evaluations. When Condition 2.1 is used to control the accuracy of the gradient estimate, the method achieves the most efficient performance in terms of gradient evaluations, while also outperforming constant sample size strategies in terms of proximal operator evaluations. Finally, although introducing bias in the gradient estimate deteriorates performance, the algorithm still converges and performs well under **Option II**.

[1] Additional numerical experiments for l_1-regularized binary classification logistic regression problems are provided in Appendix C.

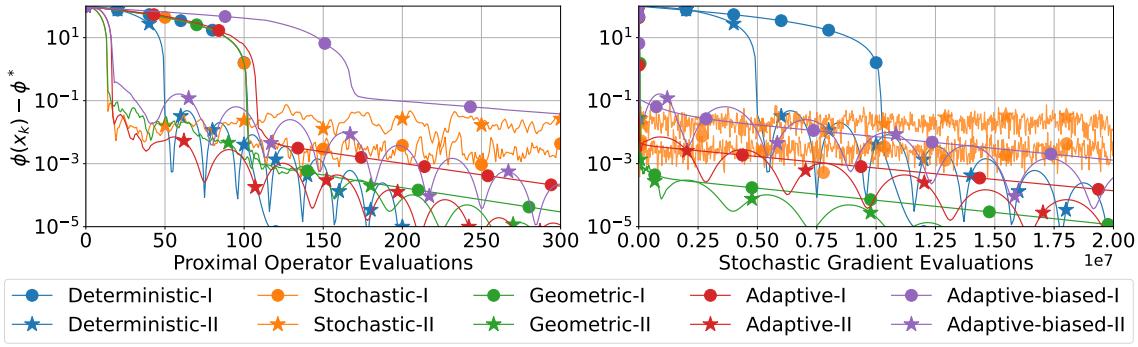


Figure 1: Optimality Gap ($\phi(x_k) - \phi^*$) with respect to the number of proximal operator evaluations and the number of stochastic gradient evaluations of Algorithm 2.1, evaluated under **Option I** (Proximal Gradient) and **Option II** (Accelerated Proximal Gradient), using "Deterministic", "Stochastic", "Geometric", "Adaptive" and "Adaptive-biased" sampling strategies on the strongly convex quadratic problem (4.1).

5 Final Remarks

In this paper, we have proposed a proximal gradient method and an accelerated proximal gradient method for composite optimization problems, where the smooth component is either a finite-sum function or an expectation of a stochastic function. The methods employed possibly biased estimates for the gradient of the smooth component, with the accuracy of these estimates adaptively adjusted via extensions of generalized "norm" conditions tailored to the composite optimization setting to achieve computational efficiency. For nonconvex, convex, and strongly convex objective functions, the methods achieved the optimal iteration complexity even with biased gradient estimates. When the gradient estimate is unbiased, we refined the analysis which allowed for less restrictive parameter settings. In this case, the methods simultaneously achieved the optimal complexity for both the number of proximal operator evaluations and the number of stochastic gradient evaluations for nonconvex, convex, and strongly convex objective functions. Finally, we conducted preliminary numerical experiments that validated our theoretical results.

References

[1] Yossi Arjevani, Yair Carmon, John C Duchi, Dylan J Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199(1):165–214, 2023.

[2] Heinz H Bauschke, Regina S Burachik, Patrick L Combettes, Veit Elser, D Russell Luke, and Henry Wolkowicz. *Fixed-point algorithms for inverse problems in science and engineering*, volume 49. Springer Science & Business Media, 2011.

[3] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

[4] Florian Beiser, Brendan Keith, Simon Urbainczyk, and Barbara Wohlmuth. Adaptive sampling strategies for risk-averse stochastic optimization with constraints. *IMA Journal of Numerical Analysis*, 43(6):3729–3765, 2023.

[5] Dimitri Bertsekas. *Convex optimization algorithms*. Athena Scientific, 2015.

[6] Raghu Bollapragada, Richard Byrd, and Jorge Nocedal. Adaptive sampling strategies for stochastic optimization. *SIAM Journal on Optimization*, 28(4):3312–3343, 2018.

[7] Silvia Bonettini, Simone Rebegoldi, and Valeria Ruggiero. Inertial variable metric techniques for the inexact forward–backward algorithm. *SIAM Journal on Scientific Computing*, 40(5):A3180–A3210, 2018.

[8] Richard H Byrd, Gillian M Chin, Jorge Nocedal, and Yuchen Wu. Sample size selection in optimization methods for machine learning. *Mathematical programming*, 134(1):127–155, 2012.

[9] Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points i. *Mathematical Programming*, 184(1):71–120, 2020.

[10] Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points ii: first-order methods. *Mathematical Programming*, 185(1):315–355, 2021.

[11] Richard G Carter. On the global convergence of trust region algorithms using inexact gradient information. *SIAM Journal on Numerical Analysis*, 28(1):251–265, 1991.

[12] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[13] Xi Chen, Seyoung Kim, Qihang Lin, Jaime G Carbonell, and Eric P Xing. Graph-structured multi-task regression and an efficient optimization method for general fused lasso. *arXiv preprint arXiv:1005.3579*, 2010.

[14] John C Duchi, Shai Shalev-Shwartz, Yoram Singer, and Ambuj Tewari. Composite objective mirror descent. In *Colt*, volume 10, pages 14–26. Citeseer, 2010.

[15] Jalal M Fadili and Gabriel Peyré. Total variation projection with first order schemes. *IEEE Transactions on Image Processing*, 20(3):657–669, 2010.

[16] Saeed Ghadimi and Guanghui Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization i: A generic algorithmic framework. *SIAM Journal on Optimization*, 22(4):1469–1492, 2012.

[17] Saeed Ghadimi and Guanghui Lan. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Mathematical Programming*, 156(1):59–99, 2016.

[18] Per Christian Hansen, James G Nagy, and Dianne P O'leary. *Deblurring images: matrices, spectra, and filtering*. SIAM, 2006.

[19] Chonghai Hu, Weike Pan, and James Kwok. Accelerated gradient methods for stochastic optimization and online learning. *Advances in Neural Information Processing Systems*, 22, 2009.

[20] Sashank J Reddi, Suvrit Sra, Barnabas Poczos, and Alexander J Smola. Proximal stochastic methods for nonsmooth nonconvex finite-sum optimization. *Advances in neural information processing systems*, 29, 2016.

[21] Afrooz Jalilzadeh, Uday Shanbhag, Jose Blanchet, and Peter W Glynn. Smoothed variable sample-size accelerated proximal methods for nonsmooth stochastic convex programs. *Stochastic Systems*, 12(4):373–410, 2022.

[22] Rodolphe Jenatton, Julien Mairal, Guillaume Obozinski, and Francis R Bach. Proximal methods for sparse hierarchical dictionary learning. In *ICML*, volume 1, page 2. Citeseer, 2010.

[23] Andrei Kulunchakov and Julien Mairal. A generic acceleration framework for stochastic composite optimization. *Advances in Neural Information Processing Systems*, 32, 2019.

[24] Guanghui Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1):365–397, 2012.

[25] Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. *arXiv preprint arXiv:1907.02189*, 2019.

[26] Zhize Li and Jian Li. A simple proximal stochastic gradient method for nonsmooth nonconvex optimization. *Advances in neural information processing systems*, 31, 2018.

[27] Luke Marrinan, Uday V Shanbhag, and Farzad Yousefian. Zeroth-order gradient and quasi-newton methods for nonsmooth nonconvex stochastic optimization. *arXiv preprint arXiv:2401.08665*, 2023.

[28] Aryan Mokhtari, Qing Ling, and Alejandro Ribeiro. Network newton distributed optimization methods. *IEEE Transactions on Signal Processing*, 65(1):146–161, 2016.

[29] Yu Nesterov. Gradient methods for minimizing composite functions. *Mathematical programming*, 140(1):125–161, 2013.

[30] Yurii Nesterov. A method for unconstrained convex minimization problem with the rate of convergence o (1/k2). In *Dokl. Akad. Nauk. SSSR*, volume 269, page 543, 1983.

[31] Yurii Nesterov et al. *Lectures on convex optimization*, volume 137. Springer, 2018.

[32] Lam M Nguyen, Katya Scheinberg, and Trang H Tran. Stochastic ista/fista adaptive step search algorithms for convex composite optimization. *arXiv preprint arXiv:2402.15646*, 2024.

[33] Thomas O'Leary-Roseberry and Raghu Bollapragada. Fast unconstrained optimization via hessian averaging and adaptive gradient sampling methods. *arXiv preprint arXiv:2408.07268*, 2024.

[34] Nhan H Pham, Lam M Nguyen, Dzung T Phan, and Quoc Tran-Dinh. Proxsarah: An efficient algorithmic framework for stochastic composite nonconvex optimization. *Journal of Machine Learning Research*, 21(110):1–48, 2020.

[35] Simone Rebegoldi and Luca Calatroni. Scaled, inexact, and adaptive generalized fista for strongly convex optimization. *SIAM Journal on Optimization*, 32(3):2428–2459, 2022.

[36] Katya Scheinberg, Donald Goldfarb, and Xi Bai. Fast first-order methods for composite convex optimization with backtracking. *Foundations of Computational Mathematics*, 14:389–417, 2014.

[37] Mark Schmidt, Nicolas Roux, and Francis Bach. Convergence rates of inexact proximal-gradient methods for convex optimization. *Advances in neural information processing systems*, 24, 2011.

[38] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems*, 25, 2012.

[39] Tao Sun, Roberto Barrio, Hao Jiang, and Lizhi Cheng. Convergence rates of accelerated proximal gradient algorithms under independent noise. *Numerical Algorithms*, 81:631–654, 2019.

[40] William F Trench. Conditional convergence of infinite products. *The American mathematical monthly*, 106(7):646–651, 1999.

[41] Joel A Tropp and Stephen J Wright. Computational methods for sparse solution of linear inverse problems. *Proceedings of the IEEE*, 98(6):948–958, 2010.

[42] Jingyi Wang and Cosmin G. Petra. A sequential quadratic programming algorithm for non-smooth problems with upper-c^2 objective. *SIAM Journal on Optimization*, 33(3):2379–2405, 2023.

[43] Yuchen Xie, Raghu Bollapragada, Richard Byrd, and Jorge Nocedal. Constrained and composite optimization via adaptive sampling methods. *IMA Journal of Numerical Analysis*, 44(2):680–709, 2024.

[44] Qinzi Zhang, Hoang Tran, and Ashok Cutkosky. Private zeroth-order nonsmooth nonconvex optimization. *arXiv preprint arXiv:2406.19579*, 2024.

A Technical Results

In this section, we present some technical results that have been used in the paper.

Lemma A.1. *Given $a_1, a_2 \in \mathbb{R}^d$ and $c \in \mathbb{R}$,*

$$a_1^T a_2 + (c-1)\|a_1\|^2 = \tfrac{1}{2}(\|a_2\|^2 - \|a_1 - a_2\|^2 + (2c-1)\|a_1\|^2).$$

Proof. The proof follows as,

$$a_1^T a_2 + (c-1)\|a_1\|^2 = \tfrac{1}{2}(2a_1^T a_2 - \|a_1\|^2 + (2c-1)\|a_1\|^2) = \tfrac{1}{2}(\|a_2\|^2 - \|a_1 - a_2\|^2 + (2c-1)\|a_1\|^2).$$

□

Lemma A.2. *Given a non-negative sequence $\{T_k\}$ such that $T_{k+1} \leq \rho_1 T_k + a\rho_2^k$ where $\rho_1, \rho_2 \in [0, 1)$ and $0 \leq a < \infty$, the sequence $\{T_k\} \to 0$ at a linear rate as,*

$$T_k \leq \max\{\rho_1 + \omega, \rho_2\}^{k+1} \max\left\{T_0, \tfrac{a}{\omega}\right\},$$

where $\omega > 0$ such that $\rho_1 + \omega < 1$.

Proof. The proof follows by induction. For T_0, the result trivially holds. Then, if the result holds for T_k,

$$\begin{aligned} T_{k+1} \leq \rho_1 T_k + a\rho_2^k &\leq \rho_1 \max\{\rho_1+\omega, \rho_2\}^k \max\left\{T_0, \tfrac{a}{\omega}\right\} + a\rho_2^k \\ &\leq \max\{\rho_1+\omega, \rho_2\}^k \max\left\{T_0, \tfrac{a}{\omega}\right\} (\rho_1+\omega) \\ &\leq \max\{\rho_1+\omega, \rho_2\}^{k+1} \max\left\{T_0, \tfrac{a}{\omega}\right\}, \end{aligned}$$

thus completing the proof. □

Lemma A.3. *Given non-negative sequences $\{T_k\}$, $\{a_k\}$ and $\{s_k\}$ such that $T_{k+1} \leq (1+a_k)T_k + s_k$, $T_0 < \infty$, $\sum_{k=0}^{\infty} a_k < \infty$, $\sum_{k=0}^{\infty} s_k < \infty$, then the sequence $\{T_k\}$ is bounded.*

Proof. Unrolling the recursion,

$$T_{k+1} \leq T_0 \prod_{i=0}^{k}(1+a_i) + \sum_{i=0}^{k} s_i \prod_{j=i+1}^{k}(1+a_i) \leq T_0 \prod_{i=0}^{\infty}(1+a_i) + \sum_{i=0}^{\infty} s_i \prod_{j=0}^{\infty}(1+a_j),$$

where the second inequality holds due to all the additive terms being non-negative and the product terms being at least one. From [40, Theorem 1], since $\sum_{k=0}^{\infty} a_k < \infty$ with $\{a_k\} \geq 0$, the infinite product $\prod_{i=0}^{\infty}(1+a_i) < \infty$. Thus, $T_{k+1} \leq \prod_{i=0}^{\infty}(1+a_i)\left[T_0 + \sum_{i=0}^{\infty} s_i\right] < \infty$. □

Lemma A.4. *Given non-negative sequences $\{R_k\}$, $\{T_k\}$, $\{a_k\}$ and $\{s_k\}$ such that $R_{k+1} + T_{k+1} \leq R_k + (1+a_k)T_k + s_k$, $T_0 < \infty$, $R_0 < \infty$, $\sum_{k=0}^{\infty} s_k < \infty$, and $a_k = \hat{\rho}\rho_k$ $\forall k \geq 0$ where $\sum_{k=0}^{\infty} \rho_k < \infty$, $\sum_{k=0}^{\infty} k\rho_k < \infty$ and $\hat{\rho}$ can be controlled to be sufficiently small, then the sequence $\{R_k\}$ is bounded.*

Proof. Let $\sum_{k=0}^{\infty} s_k < \bar{s}$, $\sum_{k=0}^{\infty} a_k < \bar{a}$. From [40, Theorem 1], as $\sum_{k=0}^{\infty} a_k < \infty$ with $\{a_k\} \geq 0$, $\exists \hat{a} > 0$ such that $\prod_{i=0}^{\infty}(1+a_i) < \hat{a}$. We now unroll the recursion for T_k as,

$$\begin{aligned} T_{k+1} &\leq (1+a_k)T_k + s_k + R_k - R_{k+1} \leq T_0 \prod_{i=0}^{k}(1+a_i) + \sum_{i=0}^{k}(s_i + R_i - R_{i+1}) \prod_{j=i+1}^{k}(1+a_j) \\ &\leq T_0 \prod_{i=0}^{k}(1+a_i) + \sum_{i=0}^{k}(s_i+R_i) \prod_{j=i+1}^{k}(1+a_j) \leq \prod_{i=0}^{\infty}(1+a_i)\left[T_0 + \sum_{i=0}^{k}(s_i+R_i)\right] \\ &\leq \hat{a}\left[T_0 + \sum_{i=0}^{k}(s_i+R_i)\right]. \end{aligned} \tag{A.1}$$

We unroll the recursion for R_k using a telescopic sum as,

$$R_{k+1} - R_0 \leq T_0 - T_{k+1} + \sum_{i=0}^{k}(a_i T_i + s_i).$$

Further unrolling the above inequality and using (A.1) yields,

$$\begin{aligned}
R_{k+1} &\leq R_0 + T_0 + \sum_{i=0}^{k} (a_i T_i + s_i) \\
&= R_0 + T_0 + \hat{a} T_0 \sum_{i=0}^{k} a_i + \hat{a} \sum_{i=0}^{k} \left(a_i \sum_{j=0}^{i-1} s_j \right) + \hat{a} \sum_{i=0}^{k} \left(a_i \sum_{j=0}^{i-1} R_j \right) + \sum_{i=0}^{k} s_i \\
&\leq R_0 + T_0 + \hat{a} T_0 \sum_{i=0}^{k} a_i + \hat{a} \left(\sum_{i=0}^{k} s_i \right) \left(\sum_{i=0}^{k} a_i \right) + \hat{a} \sum_{i=0}^{k} \left(a_i \sum_{j=0}^{i-1} R_j \right) + \sum_{i=0}^{k} s_i \\
&\leq R_0 + T_0 + \hat{a}\bar{a} T_0 + \hat{a}\bar{a}\bar{s} + \hat{a} \sum_{i=0}^{k} \left(a_i \sum_{j=0}^{i-1} R_j \right) + \bar{s} = \hat{R} + \hat{a} \sum_{i=0}^{k} \left(a_i \sum_{j=0}^{i-1} R_j \right),
\end{aligned}$$

where $\hat{R} = R_0 + T_0 + \hat{a}\bar{a}T_0 + \hat{a}\bar{a}\bar{s} + \bar{s}$. Let $\sum_{k=0}^{\infty} k\rho_k = \bar{\rho}$ and we define a constant $C = \frac{\hat{R}}{1-\hat{a}\hat{\rho}\bar{\rho}} > 0$ for sufficiently small $\hat{\rho}$ such that $1 - \hat{a}\hat{\rho}\bar{\rho} \in (0, 1)$. We show via induction that $\{R_k\} \leq C$. First, $R_0 \leq \hat{R} \leq C$ as $1 - \hat{a}\hat{\rho}\bar{\rho} \in (0, 1)$. Then, if the induction holds for $k \geq 0$,

$$R_{k+1} \leq \hat{R} + \hat{a} \sum_{i=0}^{k} \left(a_i \sum_{j=0}^{i-1} R_j \right) \leq \hat{R} + \hat{a} \sum_{i=0}^{k} (a_i i C) \leq \hat{R} + \hat{a}\hat{\rho}\bar{\rho} C = \hat{R} + \hat{a}\hat{\rho}\bar{\rho} \frac{\hat{R}}{1-\hat{a}\hat{\rho}\bar{\rho}} = C,$$

completing the proof. □

B Additional Proofs

In this section, we present proofs that have been omitted from the paper for brevity.

Lemma B.1. *Suppose Assumption 2.1 holds and Condition 2.1 is satisfied in Algorithm 2.1. Then,*

1. *For the finite-sum problem (1.3):*

$$\left(1 - \tfrac{\eta_k}{2}\right) \|g_k - \nabla f(y_k)\| \leq \tfrac{\eta_k}{2} \left\| R^{true}_{\alpha_k}(y_k) \right\| + \iota_0 \delta_k, \quad \forall k \geq 0.$$

2. *For the expectation problem (1.4):*

$$\left(1 - \tfrac{\tilde{\eta}_k^2}{2}\right) \mathbb{E}_k \left[\|g_k - \nabla f(y_k)\|^2 \right] \leq \tfrac{\tilde{\eta}_k^2}{2} \mathbb{E}_k \left[\left\| R^{true}_{\alpha_k}(y_k) \right\|^2 \right] + \tilde{\iota}_0^2 \tilde{\delta}_k^2, \quad \forall k \geq 0.$$

Proof. For the finite-sum problem (1.3), using Condition 2.1, we get,

$$\begin{aligned}
\|g_k - \nabla f(y_k)\| &\leq \tfrac{\eta_k}{2} \left\| R_{\alpha_k}(y_k) - R^{true}_{\alpha_k}(y_k) + R^{true}_{\alpha_k}(y_k) \right\| + \iota_0 \delta_k \\
&\leq \tfrac{\eta_k}{2} \left\| R_{\alpha_k}(y_k) - R^{true}_{\alpha_k}(y_k) \right\| + \tfrac{\eta_k}{2} \left\| R^{true}_{\alpha_k}(y_k) \right\| + \iota_0 \delta_k \\
&\leq \tfrac{\eta_k}{2} \|g_k - \nabla f(y_k)\| + \tfrac{\eta_k}{2} \left\| R^{true}_{\alpha_k}(y_k) \right\| + \iota_0 \delta_k,
\end{aligned}$$

where the final inequality follows form (2.5). Rearranging the final inequality yields the desired result for the finite-sum problem (1.3).

For the expectation problem (1.4), from Condition 2.1 and using Jensen's inequality, we get,

$$\begin{aligned}\mathbb{E}_k\left[\|g_k - \nabla f(y_k)\|^2\right] &\leq \tfrac{\tilde{\eta}_k^2}{4}\mathbb{E}_k\left[\left\|R_{\alpha_k}(y_k) - R_{\alpha_k}^{true}(y_k) + R_{\alpha_k}^{true}(y_k)\right\|^2\right] + \tilde{\iota}_0^2\tilde{\delta}_k^2 \\ &\leq \tfrac{\tilde{\eta}_k^2}{2}\mathbb{E}_k\left[\left\|R_{\alpha_k}(y_k) - R_{\alpha_k}^{true}(y_k)\right\|^2\right] + \tfrac{\tilde{\eta}_k^2}{2}\mathbb{E}_k\left[\left\|R_{\alpha_k}^{true}(y_k)\right\|^2\right] + \tilde{\iota}_0^2\tilde{\delta}_k^2 \\ &\leq \tfrac{\tilde{\eta}_k^2}{2}\mathbb{E}_k\left[\|g_k - \nabla f(y_k)\|^2\right] + \tfrac{\tilde{\eta}_k^2}{2}\mathbb{E}_k\left[\left\|R_{\alpha_k}^{true}(y_k)\right\|^2\right] + \tilde{\iota}_0^2\tilde{\delta}_k^2,\end{aligned}$$

where the second inequality follows from the identity $(a+b)^2 \leq 2a^2 + 2b^2$ and the final inequality follows from (2.5). Rearranging the final inequality yields the desired result for the expectation problem (1.4). □

Lemma B.2. *Suppose Assumptions 2.1 and 2.2 hold in Algorithm 2.1 for the finite-sum problem (1.3). Let $g_k = \frac{1}{|S_k|}\sum_{\xi\in S_k}\nabla F(y_k,\xi)$, where $S_k \subseteq \mathcal{S}$ $\forall k \geq 0$. Then, Condition 2.1 is satisfied $\forall k \geq 0$ if*

$$|S_k| = \left\lceil \frac{N}{\left(1+\frac{\eta_k\left\|R_{\alpha_k}(y_k)\right\|+2\iota_0\delta_k}{2\sigma}\right)} \right\rceil.$$

Proof. In iteration $k \geq 0$, from the definition of g_k, the gradient error can be bounded as,

$$\begin{aligned}\|g_k - \nabla f(y_k)\| &= \tfrac{1}{|S_k|}\left\|\sum_{\xi\in S_k}\left(\nabla F(y_k,\xi) - \nabla f(y_k)\right)\right\| \\ &= \tfrac{1}{|S_k|}\left\|\sum_{\xi\in \mathcal{S}/S_k}\left(\nabla F(y_k,\xi) - \nabla f(y_k)\right)\right\| \\ &\leq \tfrac{1}{|S_k|}\sum_{\xi\in\mathcal{S}/S_k}\|\nabla F(y_k,\xi) - \nabla f(y_k)\| \\ &\leq \tfrac{1}{|S_k|}\sum_{\xi\in\mathcal{S}/S_k}\sigma = \tfrac{N-|S_k|}{|S_k|}\sigma,\end{aligned}$$

where the second equality following from the definition of the finite-sum problem (1.3) and the last inequality follows from Assumption 2.2. From the defined sample size, we have $|S_k| \geq \frac{N}{\left(1+\frac{\eta_k\left\|R_{\alpha_k}(y_k)\right\|+2\iota_0\delta_k}{2\sigma}\right)}$. Therefore, the gradient error can be bounded as,

$$\|g_k - \nabla f(y_k)\| \leq \tfrac{N}{|S_k|}\sigma - \sigma \leq \left(1 + \frac{\eta_k\left\|R_{\alpha_k}(y_k)\right\|+2\iota_0\delta_k}{2\sigma}\right)\sigma - \sigma = \frac{\eta_k\left\|R_{\alpha_k}(y_k)\right\|+2\iota_0\delta_k}{2},$$

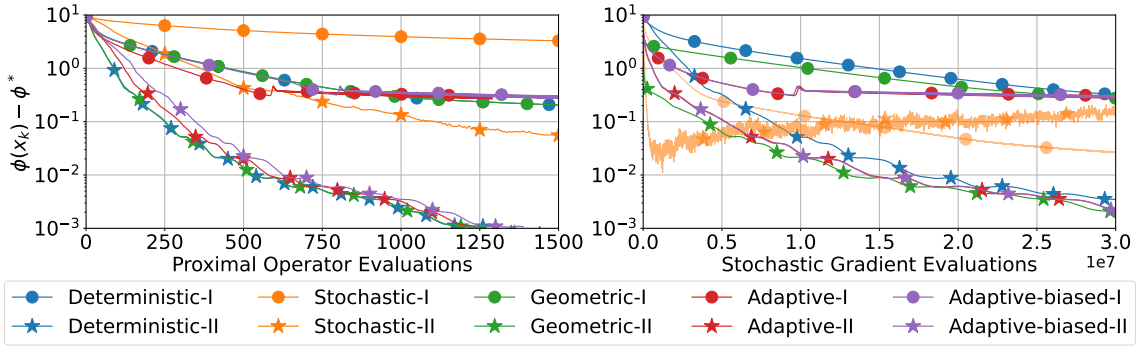
satisfying Condition 2.1. □

C Additional Numerical Experiments

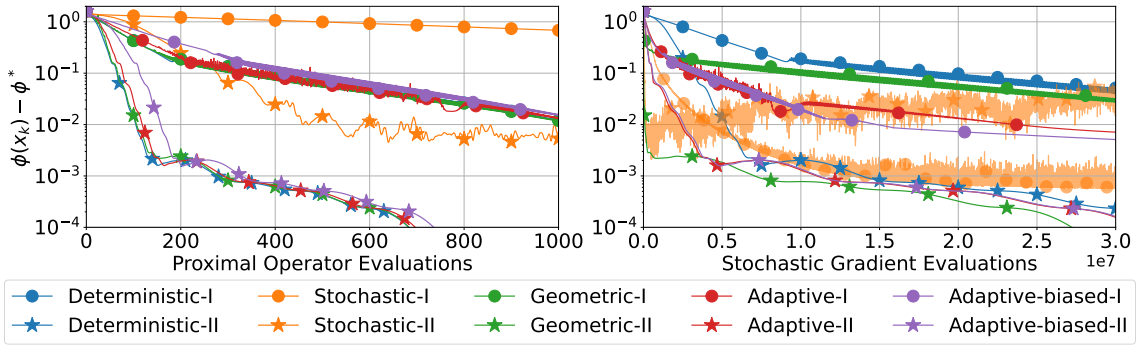
In this section, we illustrate the performance of Algorithm 2.1 on l_1-regularized binary classification logistic regression problems of the form

$$\min_{x\in\mathbb{R}^d} \phi(x) = \frac{1}{N}\sum_{i=1}^{N} \log(1+e^{-b_i A_i x}) + \frac{1}{N}\|x\|_1, \tag{C.1}$$

where $a_i \in \mathbb{R}^d$ is the feature vector (including one for the bias term) and $b_i \in \{0,1\}$ is the label for each datapoint $i \in \{1,2,\ldots,N\}$. Experiments were performed on the a9a dataset ($d = 123$, $N = 32,561$) and the ijcnn dataset ($d = 23$, $N = 49,990$) [12]. The sequence $\{\beta_k\}$ for **Option II** is chosen as described in Subsection 3.2, since the logistic regression binary cross-entropy loss is convex not strongly convex. All other implementation details are the same as in Section 4. The results are summarized in Fig. 2.



(a) a9a dataset ($d = 123$, $N = 32,561$, [12])



(b) ijcnn dataset ($d = 23$, $N = 49,990$, [12])

Figure 2: Optimality Gap ($\phi(x_k) - \phi^*$) with respect to the number of proximal operator evaluations and the number of stochastic gradient evaluations of Algorithm 2.1, evaluated under **Option I** (Proximal Gradient) and **Option II** (Accelerated Proximal Gradient), using "Deterministic", "Stochastic", "Geometric", "Adaptive" and "Adaptive-biased" sampling strategies for l_1-regularized binary classification logistic regression (C.1) on; (a) a9a dataset ($d = 123$, $N = 32,561$, [12]) and (b) ijcnn dataset ($d = 23$, $N = 49,990$).