Submitted 10/20; Revised 3/22; Published 8/22

On Acceleration for Convex Composite Minimization with Noise-Corrupted Gradients and Approximate Proximal Mapping

Qiang Zhou[*] ZHOUQIANG@U.NUS.EDU
School of Cyber Science and Engineering
Southeast University
Nanjing, Jiangsu 211189, China
Purple Mountain Laboratories
Nanjing, Jiangsu 211111, China

Sinno Jialin Pan✉ SINNOPAN@NTU.EDU.SG
School of Computer Science and Engineering
Nanyang Technological University
Singapore 639798

Editor: Julien Mairal

Abstract

The accelerated proximal methods (APM) have become one of the most important optimization tools for large-scale convex composite minimization problems, due to their wide range of applications and the optimal convergence rate in first-order algorithms. However, most existing theoretical results of APM are obtained by assuming that the gradient oracle is exact and the proximal mapping must be exactly solved, which may not hold in practice. This work presents a theoretical study of APM by allowing to use inexact gradient oracle and approximate proximal mapping. Specifically, we analyze inexact APM by improving the approximate duality gap technique (ADGT) which was originally designed for convergence analysis for first-order methods with both exact gradient oracle and proximal mapping. Our approach has several advantages: 1) we provide a unified convergence analysis that allows both inexact gradient oracle and approximate proximal mapping; 2) our proof is generic that naturally recovers the convergence rates of both accelerated and non-accelerated proximal methods, on top of which the advantages and the disadvantages of acceleration can be easily derived; 3) we derive the same convergence bound as previous methods in terms of inexact gradient oracle, but a tighter convergence bound in terms of approximate proximal mapping.

Keywords: Convex Composite Minimization, Accelerated Proximal Methods, Noisy Gradients, Approximate Proximal Mapping, Bregman Divergence

[*] Most of this work was performed while the first author worked as a postdoc at NTU, Singapore.

1. Introduction

For large-scale convex optimization problems, as it is prohibitive to compute the Hessian matrix, first-order algorithms have become the most important technique due to much cheaper per-iteration cost for evaluating first-order gradients. For smooth and convex functions, non-accelerated first-order algorithms converge with a rate of $O(1/k)$, where k is the number of times for querying first-order gradients. This result, however, is less desirable and not optimal (Nesterov, 2013; Bubeck, 2015). To improve it, Nesterov (1983) presented the accelerated gradient descent (AGD) method to achieve a convergence rate of $O(1/k^2)$ for smooth and convex functions, which is optimal for first-order algorithms (Nemirovsky and Yudin, 1983). Since then, accelerated first-order algorithms have been extensively studied (Nesterov, 2005; Tseng, 2008; Beck and Teboulle, 2009; Bubeck et al., 2015; Bubeck, 2015; Nesterov, 2013; Allen-Zhu and Orecchia, 2017; Xu et al., 2018; Yao et al., 2017; Zhou et al., 2020; Ye et al., 2020).

In the literature, convergence analysis is mainly obtained by assuming that the gradient oracle is exact (noiseless). That means there exists a black-box that can return an exact first-order gradient for any given point (Nesterov, 2013; Bubeck, 2015). However, in many applications, only an approximate or a noise-corrupted (i.e., inexact) gradient is available. For example, gradients may only be approximately computed when applying the smoothing technique to non-smooth functions (Nesterov, 2005), which depends on solving another auxiliary problem that might not be easily solved. In the case of inexact gradient, it has been empirically observed that non-accelerated first-order algorithms (e.g., gradient descent (GD)) significantly outperform AGD (Hardt, 2014). In other words, empirically, non-accelerated first-order algorithms are more robust with inexact gradient oracle than their accelerated counterparts. Therefore, it is important to theoretically study the robustness of accelerated first-order algorithms with inexact gradient oracle.

To this end, several models with inexact gradient oracle have been introduced to study the robustness of accelerated first-order algorithms. Existing works can be classified into two categories: deterministic perturbation and stochastic noise. In the first category, the inexact gradient oracle is generally defined by extending existing properties of exact gradient to the case of inexact gradient (d'Aspremont, 2008; Devolder et al., 2014). In the latter category, the true gradient is assumed to be corrupted by a stochastic noise (Lan, 2012; Ghadimi and Lan, 2012, 2013; Zhang et al., 2014; Dvurechensky and Gasnikov, 2016; Jain et al., 2018; Wangni et al., 2018; Kulunchakov and Mairal, 2019a; Aybat et al., 2019; Wang and Zhang, 2019; Kulunchakov and Mairal, 2019b, 2020; Assran and Rabbat, 2020). Recently, Cohen et al. (2018) presented a theoretical study for the robustness of accelerated algorithms by using a more general model of noise-corrupted gradient oracle. Although their analysis is based on stochastic noise, it can also recover the results based on deterministic perturbation models presented in (d'Aspremont, 2008; Devolder et al., 2014). However, their results are limited to convex and smooth objectives which are less desirable, as the objectives of many machine learning problems are convex composite, (Tibshirani, 1996; Bach et al., 2012; Bassily et al., 2014; Parikh and Boyd, 2014; Tan et al., 2015a,b) instead of smooth (Hastie et al., 2009).

To overcome the aforementioned limitation, we present a study of the robustness of accelerated proximal methods (APM) for convex composite objectives. Specifically, we consider the minimization problem in the form of

$$\min_{\mathbf{x}\in\mathcal{X}} P(\mathbf{x}) \overset{\text{def}}{=} f(\mathbf{x}) + h(\mathbf{x}), \tag{1}$$

where $f(\mathbf{x})$ is convex and L-smooth w.r.t. $\|\cdot\|$ (refer to Definition 1), and $h(\mathbf{x})$ is convex but non-smooth. Many problems in machine learning can be cast by (1) (Bach et al., 2012). Typically, $f(\cdot)$ defines a convex loss function for training examples, and $h(\cdot)$ regularizes the model to promote a specified structure.

To solve (1), each iteration of proximal method (PM) (Parikh and Boyd, 2014; Nemirovski, 2004) first takes gradient descent on the smooth function $f(\mathbf{x})$ and then performs *proximal mapping* on the non-smooth function $h(\mathbf{x})$.

$$\mathbf{x}_{i+1} = \text{Prox}_{a_i h}\big(\mathbf{x}_i - a_i \nabla f(\mathbf{x}_i)\big), \tag{2}$$

where a_i is the step size at the i-th iteration and $\text{Prox}_{a_i h}(\cdot)$ is the proximal mapping of $h(\cdot)$

$$\text{Prox}_{a_i h}(\widehat{\mathbf{x}}) = \operatorname*{argmin}_{\mathbf{x}\in\mathcal{X}} \left\{ \frac{1}{2a_i} \|\mathbf{x} - \widehat{\mathbf{x}}\|_2^2 + h(\mathbf{x}) \right\}. \tag{3}$$

Unlike AGD+ (Cohen et al., 2018), APM for convex composite minimization may not have an analytical solution at every iteration due to the proximal mapping of non-smooth $h(\mathbf{x})$. Therefore, we simultaneously study the robustness of APM with inexact gradient oracle and approximate proximal mapping.

The framework we adopt is based on the approximate duality gap technique (ADGT) introduced by Diakonikolas and Orecchia (2019). The ADGT constructs an estimate iteratively for the duality gap of the optimal solution, which can be easily tracked and should be improved as the algorithm converges. Using ADGT to prove the convergence of APM is conceptually clear. Typically, ADGT may be used to analyze the convergence of existing first-order methods, but also useful in designing new first-order algorithms with tight convergence bound. Our construction is, however, different to the original one since ours does not assume the lower bound problem is exactly solved, and thus some important properties of ADGT do not hold. We note that ADGT have been used before for analyzing acceleration with noisy gradient oracle for convex and smooth objectives (Cohen et al., 2018), but not for accelerated proximal methods with approximate proximal mapping for convex composite minimization. Specifically, we are interested in presenting a unified convergence analysis of APM with both noise-corrupted gradients and approximate proximal mapping.

In the following, we discuss the main differences between our work and the most related works, and summarize our contributions.

- We present a unified analysis method that covers both non-accelerated and accelerated proximal methods for convex composite minimization. In particular, the convergence rates of PM and APM can be obtained by a common convergence proof with different choices of parameters. This naturally provides a comparison between PM and APM, which helps us to deeply understand the advantages as well as the disadvantages of APM over PM. In addition, our analysis allows the norm to be an arbitrary norm instead of only the Euclidean ℓ_2 norm. Thus, our approach works for the generalized proximal mapping (i.e., Bregman divergence) while the analysis of (Schmidt et al., 2011; Kulunchakov and Mairal, 2019a) only works for standard proximal mapping (i.e., squared Euclidean distance).

- Our approach allows both inexact gradient oracle and approximate generalized proximal mapping. Unlike AGD+ (Cohen et al., 2018), APM for convex composite minimization may not have an analytical solution for the generalized proximal mapping, i.e., (11), due to either the complicated $h(\mathbf{x})$ or general Bergman divergence. Therefore, it is important to study the convergence rate of APM by allowing the generalized proximal mapping to be solved approximately in expectation.

- To address the challenge of approximate proximal mapping, we present a different method to construct the approximate duality gap that makes the convergence bound simpler and tighter (Cohen et al., 2018; Diakonikolas and Orecchia, 2019, 2018; Jain et al., 2018). In particular, our method is different from AGD+ (Cohen et al., 2018) even when the objective is smooth, i.e., $h(\mathbf{x}) = 0$. Taking Algorithm 1 for example, we define a different formulation for updating $\mathbf{v}_i$ i.e., (8). It only requires $\widetilde{\nabla} f(\mathbf{x}_i)$ in our method while AGD+ takes all previous $\widetilde{\nabla} f(\mathbf{x}_j)$ from $j = 1$ to i (Cohen et al., 2018). Specifically, if we choose to directly apply the idea of AGD+ (Cohen et al., 2018) to the case of convex composite minimization (1), the update of $\mathbf{v}_i$, i.e., (8) in Algorithm 1, becomes

$$\mathbf{v}_i \approx \operatorname*{argmin}_{\mathbf{v}\in\mathcal{X}} \left\{ \frac{1}{A_i}\sum_{j=1}^{i} a_j \big\langle \widetilde{\nabla} f(\mathbf{x}_j), \mathbf{v} - \mathbf{x}_j \big\rangle + \frac{1}{A_i} D_\psi(\mathbf{v}, \mathbf{x}_0) + h(\mathbf{v}) \right\}. \tag{4}$$

The comparison between (8) and (4) implies that our Algorithm 1 is different from AGD+ (Cohen et al., 2018) even when the objective is smooth, i.e., $h(\mathbf{x}) = 0$. Next, we show that the upper bound of $A_k\mathbb{E}[G_k]$ obtained by our method (8) is better than that of AGD+ (4). Applying Lemma 4, $A_k\mathbb{E}[G_k]$ obtained by our method is

$$A_k\mathbb{E}[G_k] \le D_\psi(\mathbf{x}^\star, \mathbf{x}_0) + \sum_{i=1}^{k} a_i \big\langle \boldsymbol{\eta}_i, \mathbf{x}^\star - \mathbf{v}_i \big\rangle + \overbrace{\underbrace{\sum_{i=1}^{k} a_i \big\langle \mathbf{w}_i, \mathbf{x}^\star - \mathbf{v}_i \big\rangle}_{\textcircled{1}} + \sum_{i=1}^{k} a_i \varepsilon_i}^{\textcircled{2}}. \tag{5}$$

In contrast, $A_k\mathbb{E}[G_k]$ obtained by using (4) is

$$A_k\mathbb{E}[G_k] \le D_\psi(\mathbf{x}^\star, \mathbf{x}_0) + \sum_{i=1}^{k} a_i \langle \boldsymbol{\eta}_i, \mathbf{x}^\star - \mathbf{v}_i \rangle + \overbrace{\underbrace{\sum_{i=1}^{k-1} A_i \big\langle \mathbf{w}_i, \mathbf{v}_{i+1} - \mathbf{v}_i \big\rangle}_{\textcircled{3}} + 2\sum_{i=1}^{k} A_i \varepsilon_i}^{\textcircled{4}}. \tag{6}$$

Compared with AGD+ (Cohen et al., 2018), our method has at least two advantages. 1) Comparing ① in (5) and ③ in (6), we observe that our bound (5) is simpler than (6). Unlike (Cohen et al., 2018), both $\mathbf{v}_i$ and $\mathbf{v}_{i+1}$ are not an exact solution of proximal in our case, thus our method avoids to handle two inexact solutions together, which makes it easier to analyze. 2) More importantly, our bound is also tighter than (6). Specifically, ② in (5) is the sum of $a_i\varepsilon_i$ over i, while ④ in (6) is the sum of $A_i\varepsilon_i$ over i. For accelerated methods, we have $a_i \sim O(i)$ and $A_i \sim O(i^2)$ (refer to Remark 2 for details). Thus, our bound (5) is tighter than (6).

- Our analysis achieves the same convergence bound as existing accelerated stochastic gradient methods (Kulunchakov and Mairal, 2020, 2019a; Aybat et al., 2019; Ghadimi and Lan, 2013) in terms of inexact gradient oracle, but a tighter convergence bound than (Kulunchakov and Mairal, 2019a; Schmidt et al., 2011) in terms of approximate proximal mapping (see Tables 1 and 2 for details). To the best of our knowledge, our work is the nevertheless the first to achieve such a tight bound.
- We analyze the effect of non-smooth regularization $h(\mathbf{x})$ to the robustness of APM with inexact gradient oracle by leveraging the equivalence between convex composite minimization and constrained smooth optimization. Our analysis suggests that APM is more robust with inexact gradient oracle when a stronger regularization is used because it leads to a smaller feasible set (see Proposition 3 and Section 7.1).

1.1 Related Work

In the literature, several works have been proposed to study the effect of approximate proximal mapping or linear oracle in first-order methods. Lin et al. (2017, 2015) present a generic framework to accelerate first-order methods for convex (Lin et al., 2017) and non-convex objectives (Paquette et al., 2018). Under the same inexactness setting (Definition 5), our results are better than theirs. For example, to obtain $O(1/k^2)$ convergence rate for convex objectives, (Lin et al., 2017, Proposition 6) suggests that the error is required to decrease faster than $O(1/k^4)$, while our method only requires faster than $O(1/k^3)$ as shown in (44) of Corollary 2. For strongly convex objectives, we also achieves better convergence bound than (Lin et al., 2017) in terms of approximate proximal mapping.

To guarantee the convergence, Lan and Zhou (2016) prove that the error of inexact linear oracle of conditional gradient descent is required to decrease as the iteration (see Lan and Zhou, 2016, 2.30). It is worth noting the conditional gradient can only handle a special case of (1) in which the non-smooth $h(\mathbf{x})$ is an indicator function a convex set. By using a stronger inexactness, Ben-Tal and Nemirovskii (2001); Kamzolov et al. (2020a); Stonyakin et al. (2020) show that the error accumulation due to approximate proximal mapping can be removed. Taking (Ben-Tal and Nemirovskii, 2001) for example (as they use the same inexactness), it requires the approximate solution to satisfy $\Psi_i(\mathbf{v}_i) - \Psi_i(\mathbf{v}_i^\star) \le \varepsilon - D_\psi(\mathbf{v}_i, \mathbf{v}_i^\star)$ (see Ben-Tal and Nemirovskii, 2001, Lemma 5.5.1) where $D_\psi(\mathbf{v}_i, \mathbf{v}_i^\star) \ge \frac{\gamma}{2}\|\mathbf{v}_i - \mathbf{v}_i^\star\|^2$. In contrast, our method only needs to satisfy $\Psi_i(\mathbf{v}_i) - \Psi_i(\mathbf{v}_i^\star) \le \varepsilon$ as shown in Definition 5. Kamzolov et al. (2020b) considers the same gradient noise as our work, however, they assume the objective function must be μ-strongly convex and L-smooth. Recently, Assran and Rabbat (2020) also studied the convergence of Nesterov's accelerated gradient with inexact gradient oracle where they assume the objectives not only smooth and strongly convex but also twice continuously differentiable. In contrast, we study both generally and strongly convex objectives, with smooth $f(\mathbf{x})$ and non-smooth $h(\mathbf{x})$ objective functions. In addition, Davis et al. (2018) also studied the convergence of non-accelerated first-order proximal methods with inexact gradient oracle.

This paper is organized as follows. Section 2 introduces the notation and preliminaries. Section 3 presents the APM with inexact gradient oracle and approximate proximal mapping for convex composite minimization; Section 4 is devoted to convergence analysis of inexact APM; Section 5 introduces the extension to the case of strongly convex objectives and Section 6 presents the extension to bounded variance models; Finally, we demonstrate our analysis by experiments in Section 7 and conclude the paper in Section 8.

2. Notation and Preliminaries

Throughout this paper, we use lower-case and upper-case boldface characters (e.g., $\mathbf{x}$ and $\mathbf{X}$) to denote vectors and matrices, respectively. Let $\mathbf{0}$ be a vector or matrix with all its entries equal to 0. We assume that the feasible region $\mathcal{X} \in \mathbb{R}^n$ considered in problem (1) is a closed convex set. We assume that there is an arbitrary but fixed norm $\|\cdot\|$ associated with $\mathcal{X}$. Then, all statements about functions properties are described with respect to the norm. For generic norm $\|\cdot\|$, its dual norm $\|\cdot\|_*$ is defined as $\|\mathbf{y}\|_* = \sup_{\mathbf{x}}\{\langle \mathbf{x}, \mathbf{y}\rangle \mid \|\mathbf{x}\| \leq 1\}$, where $\langle\cdot,\cdot\rangle$ denotes the inner product.

Definition 1 (L-Smooth) *We say a function $f : \mathcal{X} \to \mathbb{R}$ is L-smooth with respect to $\|\cdot\|$, if it is differentiable and satisfies*

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle + \frac{L}{2}\|\mathbf{y}-\mathbf{x}\|^2, \forall \mathbf{x}, \mathbf{y} \in \mathcal{X}.$$

Definition 2 (γ-Strongly Convex) *We say a function $f : \mathcal{X} \to \mathbb{R}$ is γ-strongly convex with respect to $\|\cdot\|$, if it satisfies*

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle + \frac{\gamma}{2}\|\mathbf{y}-\mathbf{x}\|^2, \forall \mathbf{x}, \mathbf{y} \in \mathcal{X}.$$

Definition 3 (Convex Conjugate) *For $f(\mathbf{x})$, its convex conjugate $f^*(\mathbf{y})$ is defined as*

$$f^*(\mathbf{y}) \overset{\text{def}}{=} \sup_{\mathbf{x}} \big\{\langle \mathbf{x}, \mathbf{y}\rangle - f(\mathbf{x})\big\}.$$

Lemma 1 (Hiriart-Urruty and Lemaréchal, 1993, Theorem 4.22) *If $f(\mathbf{x})$ is closed and γ-strongly convex with respect to $\|\cdot\|$, then $f^*(\mathbf{y})$ is $\frac{1}{\gamma}$-smooth with respect to the dual norm $\|\cdot\|_*$ and $\nabla f^*(\mathbf{y}) = \operatorname{argmax}_{\mathbf{x}} \big\{\langle \mathbf{x}, \mathbf{y}\rangle - f(\mathbf{x})\big\}$.*

3. Inexact APM for Convex Composite Minimization

This section presents the accelerated proximal methods (APM) with inexact gradient oracle and approximate proximal mapping to address the convex composite problem (1).

3.1 Generalized Proximal Mapping

This section is dedicated to the concept of generalized proximal mapping. To this end, we first introduce the classical Bregman divergence, which plays a key role for defining the generalized proximal mapping.

Definition 4 (Bregman Divergence) *(Bregman, 1967; Censor and Zenios, 1998) Let $\psi : \mathcal{X} \to \mathbb{R}$ be a continuously differentiable function and γ-strongly convex with respect to $\|\cdot\|$. The Bregman divergence is defined as*

$$D_{\psi}(\mathbf{x}, \mathbf{y}) \overset{\text{def}}{=} \psi(\mathbf{x}) - \psi(\mathbf{y}) - \langle \nabla\psi(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle, \forall \mathbf{x}, \mathbf{y} \in \mathcal{X}.$$

The Bregman divergence $D_{\psi}(\mathbf{x}, \mathbf{y})$ is essentially the difference between $\psi(\mathbf{x})$ and its first-order approximation provided by $\mathbf{y}$. It includes many well-know examples.

- Squared Euclidean distance: let $\psi(\mathbf{x}) \overset{\text{def}}{=} \frac{1}{2}\|\mathbf{x}\|_2^2$, then $D_{\psi}(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|_2^2$.
- Squared Mahalanobis distance: let $\psi(\mathbf{x}) \overset{\text{def}}{=} \frac{1}{2}\mathbf{x}^\top \mathbf{M}\mathbf{x}$ where $\mathbf{M} \succeq 0$ is a positive semi-definite matrix, then $D_{\psi}(\mathbf{x}, \mathbf{y}) = \frac{1}{2}(\mathbf{x} - \mathbf{y})^\top \mathbf{M}(\mathbf{x} - \mathbf{y})$.
- Kullback-Leibler divergence: let $\Omega \overset{\text{def}}{=} \{\mathbf{x} \in \mathbb{R}_+^n : \sum_i x_i = 1\}$ and $\psi(\mathbf{x}) \overset{\text{def}}{=} \sum_i x_i \log x_i$, then $D_{\psi}(\mathbf{x}, \mathbf{y}) = \sum_i x_i \log \frac{x_i}{y_i}$ for $\mathbf{x}, \mathbf{y} \in \Omega$.

Given Definition 4, we can generalize the proximal mapping (3) from squared Euclidean distance to Bregman divergence (Parikh and Boyd, 2014). Specifically, the proximal method with generalized proximal mapping for (1) is defined as

$$\mathbf{x}_{i+1} \overset{\text{def}}{=} \operatorname*{argmin}_{\mathbf{x} \in \mathcal{X}} \Big\{ \big\langle \nabla f(\mathbf{x}_i), \mathbf{x} - \mathbf{x}_i \big\rangle + \frac{1}{a_i} D_{\psi}(\mathbf{x}, \mathbf{x}_i) + h(\mathbf{x}) \Big\}, \tag{7}$$

where a_i is the step-size at ith iteration. The Bregman divergence used in generalized proximal mapping provides many advantages. For example, it can be considered as preconditioning, that allows us to use a more accurate model of $f(\mathbf{x})$ around $\mathbf{x}_i$ (Liu et al., 2019; Wang et al., 2019). By choosing an approximate $D_{\psi}(\cdot, \cdot)$, it is expected

$$f(\mathbf{x}) \approx f(\mathbf{x}_i) + \big\langle \nabla f(\mathbf{x}_i), \mathbf{x} - \mathbf{x}_i \big\rangle + \frac{1}{a_i} D_{\psi}(\mathbf{x}, \mathbf{x}_i).$$

By doing so, this can improve the convergence of optimization algorithm, especially for an ill-conditioned $f(\mathbf{x})$. For example, Liu et al. (2019) demonstrate that the Bregman divergence $D_{\psi}(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\mathbf{x}^\top \mathbf{M}\mathbf{y}$ can significantly improve the performance over SVRG (Johnson and Zhang, 2013) and Katyusha X (Allen-Zhu, 2018), where $\mathbf{M} \succ 0$ is a fixed preconditioner.

Algorithm 1 Inexact APM for Convex Composite Minimization (1)

```
Input: starting point $\mathbf{x}_0$
$A_0 = 0$ and $\mathbf{y}_0 = \mathbf{v}_0 = \mathbf{x}_0$
for $i = 1$ to $k$ do
    Set $A_i := A_{i-1} + a_i$
    Set $\mathbf{x}_i := \frac{A_{i-1}}{A_i}\mathbf{y}_{i-1} + \frac{a_i}{A_i}\mathbf{v}_{i-1}$
    Solve
        $\mathbf{v}_i \approx \operatorname*{argmin}_{\mathbf{v}\in\mathcal{X}} \left\{ \langle \widetilde{\nabla} f(\mathbf{x}_i), \mathbf{v} \rangle + \frac{1}{a_i} D_\psi(\mathbf{v}, \mathbf{v}_{i-1}) + h(\mathbf{v}) \right\}$    (8)
    such that $\mathbb{E}\left[\Psi_i(\mathbf{v}_i) - \Psi_i(\mathbf{v}_i^\star)\right] \leq \varepsilon_i$.
    Set $\mathbf{y}_i := \frac{A_{i-1}}{A_i}\mathbf{y}_{i-1} + \frac{a_i}{A_i}\mathbf{v}_i$
end for
Output: $\mathbf{y}_k$
```

3.2 APM with Inexact Gradient Oracle and Approximate Proximal Mapping

The gradient oracle in generalized proximal mapping (7) is assume exact. To study the robustness of APM with inexact gradient oracle, following (Cohen et al., 2018), we assume the true gradient $\nabla f(\mathbf{x}_i)$ in (7) is corrupted by an additive noise $\boldsymbol{\eta}_i$:

$$\widetilde{\nabla} f(\mathbf{x}_i) = \nabla f(\mathbf{x}_i) + \boldsymbol{\eta}_i, \tag{9}$$

where $\boldsymbol{\eta}_i$ can be either a deterministic or random variable. In Section 6, we show the generalization of our approach to the bounded variance noise models from (Lan, 2012; Ghadimi and Lan, 2012). Since $\boldsymbol{\eta}_i$ is allowed to be a random variable, it can be either the error when the gradient is only estimated from a stochastic subset (Lan, 2012; Ghadimi and Lan, 2012, 2013; Atchadé et al., 2017; Krichene and Bartlett, 2017; Jain et al., 2018; Kulunchakov and Mairal, 2019a, 2020) or the intentionally added Gaussian noise of the gradient in differential private empirical risk minimization (Bassily et al., 2014).

By replacing the noise-corrupted gradient $\widetilde{\nabla} f(\mathbf{x}_i)$, the generalized proximal mapping (7) becomes

$$\mathbf{x}_{i+1} \overset{\text{def}}{=} \operatorname*{argmin}_{\mathbf{x}\in\mathcal{X}} \left\{ \langle \widetilde{\nabla} f(\mathbf{x}_i), \mathbf{x} - \mathbf{x}_i \rangle + \frac{1}{a_i} D_\psi(\mathbf{x}, \mathbf{x}_i) + h(\mathbf{x}) \right\}. \tag{10}$$

In order to obtain accelerated convergence rate, we introduce another sequence variables $\{\mathbf{v}_i\}_{i\geq 0}$ and use it to perform extrapolation for $\mathbf{x}_i$ (Nesterov, 1983, 2013). By doing so, (10) becomes

$$\mathbf{v}_i^\star \overset{\text{def}}{=} \operatorname*{argmin}_{\mathbf{v}\in\mathcal{X}} \left\{ \overbrace{\langle \widetilde{\nabla} f(\mathbf{x}_i), \mathbf{v} \rangle + \frac{1}{a_i} D_\psi(\mathbf{v}, \mathbf{v}_{i-1}) + h(\mathbf{v})}^{\Psi_i(\mathbf{v})} \right\}, \tag{11}$$

where $\mathbf{v}_i^\star$ is the optimal solution to the generalized proximal mapping (11). Since $\Psi_i(\mathbf{v})$ is strongly convex, the generalized proximal mapping (11) has a unique optimum. Applying Lemma 1, the optimum is given in the following proposition.

Proposition 1 *Let* $\psi_i(\mathbf{v}) \overset{\text{def}}{=} \psi(\mathbf{v}) + a_i h(\mathbf{v})$ *and* $\mathbf{z}_i \overset{\text{def}}{=} \nabla\psi(\mathbf{v}_{i-1}) - a_i \widetilde{\nabla} f(\mathbf{x}_i)$*, then* $\mathbf{v}_i^\star = \nabla\psi_i^*(\mathbf{z}_i)$*.*

Proof The proof is straightforward by applying Lemma 1. ■

However, solving (11) exactly may be expensive and impractical due to the following two reasons:

- The $h(\mathbf{v})$ may have a complicated form, for example, general overlapping group sparsity (Huang et al., 2011; Mairal et al., 2011; Schmidt et al., 2011), OSCAR(Bondell and Reich, 2008), total variation (Beck and Teboulle, 2009), etc.
- The Bregman divergence leads to expensive computation for solving (11), for example, one needs to compute inverse for the preconditioner matrix (Liu et al., 2019).

In practice, an approximate solution is obtained by employing some iterative algorithm to solve the generalized proximal mapping (11) up to a prescribed accuracy. Algorithm 1 presents the APM with noise-corrupted gradient and approximate proximal mapping for convex composite minimization (1). It only requires the generalized proximal mapping to be solved approximately *in expectation* up to a certain precision so that $\mathbf{v}_i$ is an ε_i-optimal solution to (11) by Definition 5.

Definition 5 *For strongly convex* $\Psi_i(\mathbf{v})$ *and a non-negative scalar* ε_i*,* $\mathbf{v}_i$ *is said to be an* ε_i*-optimal solution to* $\min_{\mathbf{v}\in\mathcal{X}} \Psi_i(\mathbf{v})$ *in expectation if* $\mathbb{E}\big[\Psi_i(\mathbf{v}_i) - \inf\limits_{\mathbf{v}\in\mathcal{X}} \Psi_i(\mathbf{v})\big] \le \varepsilon_i$*.*

4. Convergence Analysis

In this section, we present convergence analysis for Algorithm 1. For convenience, we assume that $\psi(\mathbf{x})$ is ξ-smooth w.r.t. $\|\cdot\|$. Our analysis approach builds on the approximate duality gap technique (ADGT) (Diakonikolas and Orecchia, 2019). In Section 4.1, we define an approximate duality gap for the optimal solution of (1). Then, in Section 4.2, we present a generic convergence result without making any assumption on the gradient noise $\boldsymbol{\eta}_k$. Finally, we present specific convergence results for bounded and unbounded $\mathcal{X}$ in Sections 4.3 and 4.4, respectively.

4.1 Approximate Duality Gap

The key idea of ADGT is to first construct an upper bound U_k and a lower bound L_k to the optimal function value $P(\mathbf{x}^\star)$ where $\mathbf{x}^\star = \operatorname{argmin}_{\mathbf{x}\in\mathcal{X}} P(\mathbf{x})$ is the minimizer of $P(\mathbf{x})$ over $\mathcal{X}$, then use them to define an approximate optimality gap $G_k \overset{\text{def}}{=} U_k - L_k$. Consequently, the convergence of Algorithm 1 can then be proved by showing G_k is converging. Specifically, we analyze the evolution of A_kG_k where $A_k = \sum_{i=1}^{k} a_i$. Note that A_k is monotonically increasing, thus G_k is decreasing if A_kG_k is non-increasing.

We now describe the choices of the upper bound U_k and the lower bound L_k. It is worth noting that their choices are critical for convergence analysis since ADGT proves the convergence of optimization algorithms by tracking the evolution of G_k. It is naturally to choose $U_k \overset{\text{def}}{=} P(\mathbf{y}_k)$ as the upper bound. To construct a lower bound L_k to $P(\mathbf{x}^\star)$, we can apply the convexity of $P(\mathbf{x})$. By the convexity of $P(\mathbf{x})$,

$$P(\mathbf{x}^\star) \geq \big(f(\mathbf{x}_k) + \big\langle \nabla f(\mathbf{x}_k), \mathbf{x}^\star - \mathbf{x}_k \big\rangle + h(\mathbf{x}^\star)\big) + \frac{1}{a_k} D_\psi(\mathbf{x}^\star, \mathbf{v}_{k-1}) - \frac{1}{a_k} D_\psi(\mathbf{x}^\star, \mathbf{v}_{k-1}). \tag{12}$$

Note that only inexact gradient oracle $\widetilde{\nabla} f(\mathbf{x}_k)$ is available, thus we substitute $\nabla f(\mathbf{x}_k) = \widetilde{\nabla} f(\mathbf{x}_k) - \boldsymbol{\eta}_k$ into (12),

$$P(\mathbf{x}^\star) \geq \big\langle \widetilde{\nabla} f(\mathbf{x}_k), \mathbf{x}^\star - \mathbf{x}_k \big\rangle + \frac{1}{a_k} D_\psi(\mathbf{x}^\star, \mathbf{v}_{k-1}) + h(\mathbf{x}^\star) + f(\mathbf{x}_k) - \langle \boldsymbol{\eta}_k, \mathbf{x}^\star - \mathbf{x}_k \rangle - \frac{1}{a_k} D_\psi(\mathbf{x}^\star, \mathbf{v}_{k-1}).$$

Then, a lower bound L_k for $P(\mathbf{x}^\star)$ can be obtained by minimizing the right-hand side of the above inequality. Formally, the L_k is defined as

$$L_k \overset{\text{def}}{=} \overbrace{\min_{\mathbf{v} \in \mathcal{X}} \left\{ \big\langle \widetilde{\nabla} f(\mathbf{x}_k), \mathbf{v} - \mathbf{x}_k \big\rangle + \frac{1}{a_k} D_\psi(\mathbf{v}, \mathbf{v}_{k-1}) + h(\mathbf{v}) \right\}}^{\text{❀}} + f(\mathbf{x}_k) - \langle \boldsymbol{\eta}_k, \mathbf{x}^\star - \mathbf{x}_k \rangle - \frac{1}{a_k} D_\psi(\mathbf{x}^\star, \mathbf{v}_{k-1}), \tag{13}$$

where ❀ is essentially same as line 6 in Algorithm 1. Thus, the updating of $\mathbf{v}_k$ is equivalent to construct the lower bound L_k.

Given U_k and L_k, it is straightforward to show $G_k = U_k - L_k \geq P(\mathbf{y}_k) - P(\mathbf{x}^\star)$. We will prove the convergence of G_k by showing A_kG_k is non-increasing. To this end, we define

$$E_k \overset{\text{def}}{=} A_kG_k - A_{k-1}G_{k-1}, \forall k \geq 1. \tag{14}$$

As described in (Diakonikolas and Orecchia, 2019), if we treat Algorithm 1 as a discretization of a underlying continuous-time dynamical system with Lyapunov function A_kG_k, then E_k is the discretization error at iteration k. If $E_i, \forall i \leq k$ are bounded, (14) leads to

$$G_k = \frac{A_1G_1 + \sum_{i=2}^{k} E_i}{A_k}.$$

Thus, it suffices to bound the sum of E_i.

Relation with existing works. Note that the lower bound L_k defined in (13) bear similarities with the approximate duality gap technique introduced by Diakonikolas and Orecchia (2019). However, the lower bound L_k in (Diakonikolas and Orecchia, 2019; Cohen et al., 2018; Diakonikolas and Orecchia, 2018) is constructed by a linear combination of all $\mathbf{x}_i, \forall i \leq k$ seen so far. In contrast, our lower bound in (13) is constructed by only using the latest $\mathbf{x}_k$. This construction for lower bound leads to several advantages when the lower bound problem ❀ in (13) only be approximately solved. As we will see later, our method simplifies the analysis but also leads to a tighter error bound.

4.2 Generic Convergence Result

We start by analyzing the optimality condition of approximate solution $\mathbf{v}_i$ to (8). To this end, we first extend the definition of ε-subdifferential (Bertsekas et al., 2003) to include the stochastic case.

Definition 6 (ε-subdifferential in expectation) *Given a convex function f and a non-negative scalar ε, the ε-subdifferential in expectation of f at $\mathbf{x}$ is defined as*

$$\partial_\varepsilon f(\mathbf{x}) \overset{\text{def}}{=} \big\{\mathbf{w} \mid \varepsilon \geq \mathbb{E}\big[f(\mathbf{x}) + \langle \mathbf{w}, \mathbf{y} - \mathbf{x}\rangle - f(\mathbf{y})\big], \forall \mathbf{y} \in \mathcal{X}\big\}. \tag{15}$$

Remark 1 *If both f and $\mathbf{x}$ are deterministic, it reduces to the standard ε-subdifferential (Bertsekas et al., 2003).*

Since $\mathbf{v}_i$ is an ε_i-optimal solution to (8) in expectation, the next lemma provides its ε_i-subdifferential in expectation, that will play a key role for convergence analysis.

Lemma 2 *If $\mathbf{v}_i$ is a ε_i-optimal solution to (8) in expectation and ψ is ξ-smooth w.r.t. $\|\cdot\|$, then there exists $\mathbf{w}_i$ with $\mathbb{E}[\|\mathbf{w}_i\|_*^2] \leq 2\xi\varepsilon_i/a_i$ such that*

$$\frac{1}{a_i}\big(\nabla\psi(\mathbf{v}_{i-1}) - \nabla\psi(\mathbf{v}_i)\big) - \widetilde{\nabla} f(\mathbf{x}_i) - \mathbf{w}_i \in \partial_{\varepsilon_i} h(\mathbf{v}_i).$$

Proof For convenience, we define

$$\Phi_i(\mathbf{v}) \overset{\text{def}}{=} \big\langle \widetilde{\nabla} f(\mathbf{x}_i), \mathbf{v}\big\rangle + \frac{1}{a_i} D_\psi(\mathbf{v}, \mathbf{v}_{i-1}).$$

Then, $\Psi_i(\mathbf{v})$ defined in (11) can be rewritten as $\Psi_i(\mathbf{v}) = \Phi_i(\mathbf{v}) + h(\mathbf{v})$. By Definition 6, the ε_i-subdifferential in expectation of $\Phi_i(\mathbf{v})$ at $\mathbf{v}_i$ is

$$\partial_{\varepsilon_i}\Phi_i(\mathbf{v}_i) = \big\{\mathbf{w} \mid \varepsilon_i \geq \mathbb{E}\big[\Phi_i(\mathbf{v}_i) + \langle \mathbf{w}, \mathbf{v} - \mathbf{v}_i\rangle - \Phi_i(\mathbf{v})\big], \forall \mathbf{v} \in \mathcal{X}\big\}.$$

It is equivalent to

$$\partial_{\varepsilon_i}\Phi_i(\mathbf{v}_i) = \left\{\mathbf{w} \mid \varepsilon_i \geq \mathbb{E}\left[\max_{\mathbf{v}\in\mathcal{X}}\left\{\frac{1}{a_i}D_\psi(\mathbf{v}_i, \mathbf{v}_{i-1}) + \langle \mathbf{w} - \widetilde{\nabla} f(\mathbf{x}_i), \mathbf{v} - \mathbf{v}_i\rangle - \frac{1}{a_i}D_\psi(\mathbf{v}, \mathbf{v}_{i-1})\right\}\right]\right\}.$$

It can be rewritten as

$$\partial_{\varepsilon_i}\Phi_i(\mathbf{v}_i) = \left\{\mathbf{w} \mid \varepsilon_i \geq \mathbb{E}\left[\max_{\mathbf{v}\in\mathcal{X}}\left\{\frac{1}{a_i}\big(\psi(\mathbf{v}_i) - \psi(\mathbf{v})\big) + \left\langle \mathbf{w} - \widetilde{\nabla} f(\mathbf{x}_i) + \frac{1}{a_i}\nabla\psi(\mathbf{v}_{i-1}), \mathbf{v} - \mathbf{v}_i\right\rangle\right\}\right]\right\}.$$

Since $\psi(\mathbf{v})$ is ξ-smooth w.r.t to $\|\cdot\|$, it implies

$$\frac{1}{a_i}\big(\psi(\mathbf{v}_i) - \psi(\mathbf{v})\big) \geq -\frac{\xi}{2a_i}\|\mathbf{v} - \mathbf{v}_i\|^2 - \frac{1}{a_i}\big\langle \nabla\psi(\mathbf{v}_i), \mathbf{v} - \mathbf{v}_i\big\rangle.$$

Applying this to the maximization term, we come up with

$$\partial_{\varepsilon_i}\Phi_i(\mathbf{v}_i) \subseteq \left\{ \mathbf{w} \;\middle|\; \varepsilon_i \geq \mathbb{E}\Bigg[\max_{\mathbf{v}\in\mathcal{X}} \bigg\{ \Big\langle \mathbf{w} - \widetilde{\nabla} f(\mathbf{x}_i) - \frac{1}{a_i}\big(\nabla\psi(\mathbf{v}_i) - \nabla\psi(\mathbf{v}_{i-1})\big), \mathbf{v} - \mathbf{v}_i \Big\rangle - \frac{\xi}{2a_i}\|\mathbf{v} - \mathbf{v}_i\|^2 \bigg\} \Bigg] \right\}.$$

Solving the maximization problem, it becomes

$$\partial_{\varepsilon_i}\Phi_i(\mathbf{v}_i) \subseteq \left\{ \mathbf{w} \;\middle|\; \frac{2\xi\varepsilon_i}{a_i} \geq \mathbb{E}\left[\left\| \mathbf{w} - \widetilde{\nabla} f(\mathbf{x}_i) - \frac{1}{a_i}\big(\nabla\psi(\mathbf{v}_i) - \nabla\psi(\mathbf{v}_{i-1})\big) \right\|_*^2 \right] \right\}.$$

Thus, for any $\mathbf{z} \in \partial_{\varepsilon_i}\Phi_i(\mathbf{v}_i)$, it must can be expressed as the form

$$\mathbf{z} = \widetilde{\nabla} f(\mathbf{x}_i) + \frac{1}{a_i}\big(\nabla\psi(\mathbf{v}_i) - \nabla\psi(\mathbf{v}_{i-1})\big) + \mathbf{w}_i \quad \text{with} \quad \mathbb{E}\left[\|\mathbf{w}_i\|_*^2\right] \leq \frac{2\xi\varepsilon_i}{a_i}. \tag{16}$$

By Definition 5, $\mathbf{v}_i$ is an ε_i-optimal solution of $\Psi_i(\mathbf{v})$ in expectation if and only if

$$\mathbb{E}\left[\Psi_i(\mathbf{v}_i) - \inf_{\mathbf{v}\in\mathcal{X}} \Psi_i(\mathbf{v})\right] \leq \varepsilon_i.$$

Invoking Definition 6, this is equivalent to $\mathbf{0}$ belongings to the ε_i-subgradient in expectation of $\Psi_i(\mathbf{v}_i)$. Combining this with (Bertsekas et al., 2003, Proposition 4.3.1), we come up with

$$\mathbf{0} \in \partial_{\varepsilon_i}\Psi_i(\mathbf{v}_i) \subset \partial_{\varepsilon_i}\Phi_i(\mathbf{v}_i) + \partial_{\varepsilon_i} h(\mathbf{v}_i).$$

Therefore, there must exists some $\mathbf{z}$ such that $\mathbf{z} \in \partial_{\varepsilon_i}\Phi_i(\mathbf{v}_i)$ and $-\mathbf{z} \in \partial_{\varepsilon_i} h(\mathbf{v}_i)$. Invoking (16), there must exist $\mathbf{w}_i$ with $\mathbb{E}[\|\mathbf{w}_i\|_*^2] \leq 2\xi\varepsilon_i/a_i$ such that

$$\frac{1}{a_i}\big(\nabla\psi(\mathbf{v}_{i-1}) - \nabla\psi(\mathbf{v}_i)\big) - \widetilde{\nabla} f(\mathbf{x}_i) - \mathbf{w}_i \in \partial_{\varepsilon_i} h(\mathbf{v}_i).$$

This completes the proof. ■

Given Lemma 2, the difference between $\Psi_i(\mathbf{v}_i)$ and $\Psi_i(\mathbf{v})$ can be bounded for any $\mathbf{v}$.

Lemma 3 *If $\mathbf{v}_i$ is a ε_i-optimal solution to* (8) *in expectation and ψ is ξ-smooth w.r.t. $\|\cdot\|$, then there exists $\mathbf{w}_i$ with $\mathbb{E}[\|\mathbf{w}_i\|_*^2] \leq 2\xi\varepsilon_i/a_i$ such that, $\forall \mathbf{v} \in \mathcal{X}$,*

$$\mathbb{E}\left[\Psi_i(\mathbf{v}_i) + \frac{1}{a_i}D_\psi(\mathbf{v}, \mathbf{v}_i) - \langle \mathbf{w}_i, \mathbf{v} - \mathbf{v}_i \rangle - \Psi_i(\mathbf{v})\right] \leq \varepsilon_i.$$

Proof By the convexity of $h(\mathbf{v})$, Definition 6 implies

$$\mathbb{E}\left[h(\mathbf{v}_i) + \langle \mathbf{w}, \mathbf{v} - \mathbf{v}_i \rangle\right] - \varepsilon_i \leq h(\mathbf{v}), \forall \mathbf{w} \in \partial_{\varepsilon_i} h(\mathbf{v}_i), \mathbf{v} \in \mathcal{X}.$$

Applying Lemma 2, there exits $\mathbf{w}_i$ with $\mathbb{E}[\|\mathbf{w}_i\|_*^2] \leq 2\xi\varepsilon_i/a_i$ such that

$$\begin{aligned}
h(\mathbf{v}) &\geq \mathbb{E}\left[h(\mathbf{v}_i) + \left\langle \frac{1}{a_i}\big(\nabla\psi(\mathbf{v}_{i-1}) - \nabla\psi(\mathbf{v}_i)\big) - \widetilde{\nabla} f(\mathbf{x}_i) - \mathbf{w}_i, \mathbf{v} - \mathbf{v}_i \right\rangle\right] - \varepsilon_i \\
&= \mathbb{E}\left[h(\mathbf{v}_i) + \frac{1}{a_i}\big\langle \nabla\psi(\mathbf{v}_{i-1}), \mathbf{v} - \mathbf{v}_i\big\rangle - \frac{1}{a_i}\big\langle \nabla\psi(\mathbf{v}_i), \mathbf{v} - \mathbf{v}_i\big\rangle - \big\langle \widetilde{\nabla} f(\mathbf{x}_i) + \mathbf{w}_i, \mathbf{v} - \mathbf{v}_i\big\rangle\right] - \varepsilon_i \\
&= \mathbb{E}\Big[h(\mathbf{v}_i) + \frac{1}{a_i}\big\langle \nabla\psi(\mathbf{v}_{i-1}), \mathbf{v} - \mathbf{v}_{i-1}\big\rangle - \frac{1}{a_i}\big\langle \nabla\psi(\mathbf{v}_{i-1}), \mathbf{v}_i - \mathbf{v}_{i-1}\big\rangle - \frac{1}{a_i}\big\langle \nabla\psi(\mathbf{v}_i), \mathbf{v} - \mathbf{v}_i\big\rangle \\
&\qquad - \big\langle \widetilde{\nabla} f(\mathbf{x}_i) + \mathbf{w}_i, \mathbf{v} - \mathbf{v}_i\big\rangle\Big] - \varepsilon_i \\
&= \mathbb{E}\left[h(\mathbf{v}_i) - \frac{1}{a_i}D_\psi(\mathbf{v}, \mathbf{v}_{i-1}) + \frac{1}{a_i}D_\psi(\mathbf{v}_i, \mathbf{v}_{i-1}) + \frac{1}{a_i}D_\psi(\mathbf{v}, \mathbf{v}_i) - \big\langle \widetilde{\nabla} f(\mathbf{x}_i) + \mathbf{w}_i, \mathbf{v} - \mathbf{v}_i\big\rangle\right] - \varepsilon_i.
\end{aligned}$$

By reorganizing both sides, it becomes

$$\begin{aligned}
&\mathbb{E}\left[h(\mathbf{v}_i) + \big\langle \widetilde{\nabla} f(\mathbf{x}_i), \mathbf{v}_i\big\rangle + \frac{1}{a_i}D_\psi(\mathbf{v}_i, \mathbf{v}_{i-1}) + \frac{1}{a_i}D_\psi(\mathbf{v}, \mathbf{v}_i) - \big\langle \mathbf{w}_i, \mathbf{v} - \mathbf{v}_i\big\rangle - \varepsilon_i\right] \\
&\leq \mathbb{E}\left[h(\mathbf{v}) + \big\langle \widetilde{\nabla} f(\mathbf{x}_i), \mathbf{v}\big\rangle + \frac{1}{a_i}D_\psi(\mathbf{v}, \mathbf{v}_{i-1})\right].
\end{aligned}$$

By the definition of $\Psi_i(\cdot)$, it can be written as

$$\mathbb{E}\left[\Psi_i(\mathbf{v}_i) + \frac{1}{a_i}D_\psi(\mathbf{v}, \mathbf{v}_i) - \big\langle \mathbf{w}_i, \mathbf{v} - \mathbf{v}_i\big\rangle - \Psi_i(\mathbf{v})\right] \leq \varepsilon_i.$$

This completes the proof. ■

Given Lemma 3, we are able to bound E_k for convex $f(\mathbf{x})$ in the following lemma.

Lemma 4 *Let $\mathbf{x}_k, \mathbf{y}_k, \mathbf{v}_k$ evolve according to Algorithm 1 where $\mathbf{v}_k$ is a ε_k-optimal solution in expectation to (11). We define $E_k^e \overset{\text{def}}{=} A_k\big(f(\mathbf{y}_k) - f(\mathbf{x}_k) - \langle \nabla f(\mathbf{x}_k), \mathbf{y}_k - \mathbf{x}_k\rangle\big) - D_\psi(\mathbf{v}_k, \mathbf{v}_{k-1})$. Then there exists $\mathbf{w}_k$ with $\mathbb{E}[\|\mathbf{w}_k\|_*^2] \leq 2\xi\varepsilon_k/a_k$ such that,*

$$\mathbb{E}\left[E_k\right] \leq \mathbb{E}\left[E_k^e\right] + \mathbb{E}\left[E_k^\eta\right] + \mathbb{E}\left[E_k^\varepsilon\right] + \mathbb{E}\left[D_\psi(\mathbf{x}^\star, \mathbf{v}_{k-1}) - D_\psi(\mathbf{x}^\star, \mathbf{v}_k)\right], \forall k \geq 1, \tag{17}$$

where $E_k^\eta \overset{\text{def}}{=} a_k\langle \boldsymbol{\eta}_k, \mathbf{x}^\star - \mathbf{v}_k\rangle$ and $E_k^\varepsilon \overset{\text{def}}{=} a_k\big(\langle \mathbf{w}_k, \mathbf{x}^\star - \mathbf{v}_k\rangle + \varepsilon_k\big)$.

Proof From the definition of E_k,

$$E_k = A_kG_k - A_{k-1}G_{k-1} = A_kP(\mathbf{y}_k) - A_{k-1}P(\mathbf{y}_{k-1}) - a_kP(\mathbf{x}^\star).$$

Substituting $P(\mathbf{y}_k)$, it becomes

$$\begin{aligned}
E_k &= A_k\big(f(\mathbf{y}_k) + h(\mathbf{y}_k)\big) - A_{k-1}P(\mathbf{y}_{k-1}) - a_kP(\mathbf{x}^\star) \\
&= A_k\big(f(\mathbf{y}_k) - f(\mathbf{x}_k) - \langle \nabla f(\mathbf{x}_k), \mathbf{y}_k - \mathbf{x}_k\rangle\big) + A_k\big(f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{y}_k - \mathbf{x}_k\rangle + h(\mathbf{y}_k)\big)
\end{aligned}$$

$$- A_{k-1}P(\mathbf{y}_{k-1}) - a_kP(\mathbf{x}^\star). \tag{18}$$

By the definition of $\mathbf{y}_k$ and convexity of $h(\cdot)$,

$$\begin{aligned}
&A_k\big(f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{y}_k - \mathbf{x}_k\rangle + h(\mathbf{y}_k)\big)\\
&= A_k\bigg\{ f(\mathbf{x}_k) + \bigg\langle \nabla f(\mathbf{x}_k), \frac{A_{k-1}}{A_k}\mathbf{y}_{k-1} + \frac{a_k}{A_k}\mathbf{v}_k - \mathbf{x}_k \bigg\rangle + h\bigg(\frac{A_{k-1}}{A_k}\mathbf{y}_{k-1} + \frac{a_k}{A_k}\mathbf{v}_k\bigg)\bigg\}\\
&\le A_k\bigg\{ f(\mathbf{x}_k) + \frac{A_{k-1}}{A_k}\langle \nabla f(\mathbf{x}_k), \mathbf{y}_{k-1} - \mathbf{x}_k\rangle + \frac{a_k}{A_k}\langle \nabla f(\mathbf{x}_k), \mathbf{v}_k - \mathbf{x}_k\rangle + \frac{A_{k-1}}{A_k}h(\mathbf{y}_{k-1}) + \frac{a_k}{A_k}h(\mathbf{v}_k)\bigg\}\\
&\le A_{k-1}P(\mathbf{y}_{k-1}) + a_k\big(f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{v}_k - \mathbf{x}_k\rangle + h(\mathbf{v}_k)\big).
\end{aligned} \tag{19}$$

On the other hand, $a_kP(\mathbf{x}^\star) = a_k\big(f(\mathbf{x}^\star) + h(\mathbf{x}^\star)\big)$ can be lower bounded as

$$\begin{aligned}
a_kP(\mathbf{x}^\star) &\ge a_k\big(f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{x}^\star - \mathbf{x}_k\rangle + h(\mathbf{x}^\star)\big) + D_\psi(\mathbf{x}^\star, \mathbf{v}_{k-1}) - D_\psi(\mathbf{x}^\star, \mathbf{v}_{k-1})\\
&= a_k\Big(\big\langle \widetilde{\nabla} f(\mathbf{x}_k), \mathbf{x}^\star - \mathbf{x}_k\big\rangle + \frac{1}{a_k}D_\psi(\mathbf{x}^\star, \mathbf{v}_{k-1}) + h(\mathbf{x}^\star)\Big)\\
&\quad + a_k\Big(f(\mathbf{x}_k) - \langle \boldsymbol{\eta}_k, \mathbf{x}^\star - \mathbf{x}_k\rangle - \frac{1}{a_k}D_\psi(\mathbf{x}^\star, \mathbf{v}_{k-1})\Big).
\end{aligned}$$

Taking expectation for the left-hand side, we come up with

$$\begin{aligned}
a_kP(\mathbf{x}^\star) &\ge a_k\mathbb{E}\left[\left(\Big\langle \widetilde{\nabla} f(\mathbf{x}_k), \mathbf{x}^\star - \mathbf{x}_k\Big\rangle + \frac{1}{a_k}D_\psi(\mathbf{x}^\star, \mathbf{v}_{k-1}) + h(\mathbf{x}^\star)\right)\right]\\
&\quad + a_k\mathbb{E}\left[\left(f(\mathbf{x}_k) - \langle \boldsymbol{\eta}_k, \mathbf{x}^\star - \mathbf{x}_k\rangle - \frac{1}{a_k}D_\psi(\mathbf{x}^\star, \mathbf{v}_{k-1})\right)\right].
\end{aligned}$$

Applying Lemma 3 with $\mathbf{v} = \mathbf{x}^\star$, there exists $\mathbf{w}_k$ with $\mathbb{E}[\|\mathbf{w}_k\|_*^2] \le 2\xi\varepsilon_k/a_k$ such that

$$\begin{aligned}
&a_kP(\mathbf{x}^\star)\\
&\ge a_k\mathbb{E}\left[\left(\Big\langle \widetilde{\nabla} f(\mathbf{x}_k), \mathbf{v}_k - \mathbf{x}_k\Big\rangle + \frac{1}{a_k}D_\psi(\mathbf{v}_k, \mathbf{v}_{k-1}) + h(\mathbf{v}_k) + \frac{1}{a_k}D_\psi(\mathbf{x}^\star, \mathbf{v}_k) - \big\langle \mathbf{w}_k, \mathbf{x}^\star - \mathbf{v}_k\big\rangle - \varepsilon_k\right)\right]\\
&\quad + a_k\mathbb{E}\left[\left(f(\mathbf{x}_k) - \langle \boldsymbol{\eta}_k, \mathbf{x}^\star - \mathbf{x}_k\rangle - \frac{1}{a_k}D_\psi(\mathbf{x}^\star, \mathbf{v}_{k-1})\right)\right]\\
&= a_k\mathbb{E}\left[(f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{v}_k - \mathbf{x}_k\rangle + h(\mathbf{v}_k)) + D_\psi(\mathbf{v}_k, \mathbf{v}_{k-1}) + D_\psi(\mathbf{x}^\star, \mathbf{v}_k) - D_\psi(\mathbf{x}^\star, \mathbf{v}_{k-1})\right]\\
&\quad - a_k\mathbb{E}\left[(\langle \boldsymbol{\eta}_k + \mathbf{w}_k, \mathbf{x}^\star - \mathbf{v}_k\rangle + \varepsilon_k)\right].
\end{aligned} \tag{20}$$

Substituting (19) and (20) into (18), we obtain

$$\begin{aligned}
\mathbb{E}[E_k] \le A_k\mathbb{E}\big[\big(f(\mathbf{y}_k) - f(\mathbf{x}_k) - \langle \nabla f(\mathbf{x}_k), \mathbf{y}_k - \mathbf{x}_k\rangle\big) - D_\psi(\mathbf{v}_k, \mathbf{v}_{k-1})\big]&\\
+ \mathbb{E}\big[D_\psi(\mathbf{x}^\star, \mathbf{v}_{k-1}) - D_\psi(\mathbf{x}^\star, \mathbf{v}_k)\big] + a_k\mathbb{E}\big[\big(\langle \boldsymbol{\eta}_k + \mathbf{w}_k, \mathbf{x}^\star - \mathbf{v}_k\rangle + \varepsilon_k\big)\big]&.
\end{aligned}$$

By using the definitions of E_k^e, E_k^η and E_k^ε, it becomes

$$\mathbb{E}\left[E_k\right] \le \mathbb{E}\left[E_k^e\right] + \mathbb{E}\left[E_k^\eta\right] + \mathbb{E}\left[E_k^\varepsilon\right] + \mathbb{E}\left[D_\psi(\mathbf{x}^\star, \mathbf{v}_{k-1}) - D_\psi(\mathbf{x}^\star, \mathbf{v}_k)\right].$$

This completes the proof. ■

Note that E_k^η is due to the inexact gradient oracle and E_k^ε is incurred from approximate proximal mapping. Before proving the main results, we first introduce a proposition to bound E_k^e that is useful for later analysis.

Proposition 2 *Assume f is L-smooth w.r.t. $\|\cdot\|$ and $\frac{a_k^2}{A_k} \leq \frac{\gamma}{L}, \forall k \geq 1$. Let $\mathbf{x}_k, \mathbf{y}_k, \mathbf{v}_k$ evolve according to Algorithm 1 where $\mathbf{v}_k$ is a ε_k-optimal solution in expectation to* (11)*. Then it holds that $E_k^e \leq 0, \forall k \geq 1$.*

Proof The proof is provided in Appendix A.1. ∎

Then, without making further assumption on $\boldsymbol{\eta}_k$, we have the following general convergence result, which is a direct consequence of the Lemma 4 and Proposition 2.

Theorem 1 *Assume f is L-smooth w.r.t. $\|\cdot\|$ and $\frac{a_k^2}{A_k} \leq \frac{\gamma}{L}, \forall k \geq 1$. Let $\mathbf{x}_k, \mathbf{y}_k, \mathbf{v}_k$ evolve according to Algorithm 1 where $\mathbf{v}_k$ is a ε_k-optimal solution in expectation to* (11)*. Then,*

$$\mathbb{E}\left[G_k\right] \leq \frac{1}{A_k}\left(D_\psi(\mathbf{x}^\star, \mathbf{x}_0) + \sum_{i=1}^{k} \mathbb{E}\left[E_i^\eta + E_i^\varepsilon\right]\right), \tag{21}$$

where $E_i^\eta \overset{\text{def}}{=} a_i\langle \boldsymbol{\eta}_i, \mathbf{x}^\star - \mathbf{v}_i\rangle$ and $E_i^\varepsilon \overset{\text{def}}{=} a_i\big(\langle \mathbf{w}_i, \mathbf{x}^\star - \mathbf{v}_i\rangle + \varepsilon_i\big)$.

Proof It can be proved by applying Lemma 4. Applying Proposition 2, (17) becomes

$$\mathbb{E}[E_i] = \mathbb{E}[A_i G_i - A_{i-1}G_{i-1}] \leq \mathbb{E}\left[D_\psi(\mathbf{x}^\star, \mathbf{v}_{i-1}) - D_\psi(\mathbf{x}^\star, \mathbf{v}_i) + E_i^\eta + E_i^\varepsilon\right]. \tag{22}$$

Substituting (22) into G_k, we obtain

$$\mathbb{E}\left[A_k G_k + D_\psi(\mathbf{x}^\star, \mathbf{v}_k)\right] \leq D_\psi(\mathbf{x}^\star, \mathbf{v}_0) + \sum_{i=1}^{k}\left(\mathbb{E}\big[E_i^\eta\big] + \mathbb{E}\big[E_i^\varepsilon\big]\right). \tag{23}$$

By noting $D_\psi(\mathbf{x}^\star, \mathbf{v}_k) \geq 0$, it implies

$$A_k \mathbb{E}[G_k] \leq D_\psi(\mathbf{x}^\star, \mathbf{v}_0) + \sum_{i=1}^{k}\left(\mathbb{E}\big[E_i^\eta\big] + \mathbb{E}\big[E_i^\varepsilon\big]\right). \tag{24}$$

This completes the proof. ∎

Theorem 1 allows us to recover convergence rates both for PM and APM with different choices of a_k. When the gradient oracle is exact (noiseless) and the proximal mapping is exactly solved, we have $\boldsymbol{\eta}_k = \mathbf{0}$ and $\varepsilon_k = 0, \forall k \geq 1$. The following theorem gives the convergence rate of Algorithm 1 in this ideal case.

Theorem 2 *Assume f is L-smooth and $\frac{a_k^2}{A_k} \leq \frac{\gamma}{L}, \forall k \geq 1$. Let $\mathbf{x}_k, \mathbf{y}_k, \mathbf{v}_k$ evolve according to Algorithm 1, where $\mathbf{v}_k$ is an exact solution (i.e., $\varepsilon_k = 0, \forall k \geq 1$) to* (11)*. If the gradient oracle is noiseless (i.e., $\boldsymbol{\eta}_k = \mathbf{0}, \forall k \geq 1$), then*

$$G_k \leq \frac{D_\psi(\mathbf{x}^\star, \mathbf{x}_0)}{A_k}, \forall k \geq 1.$$

Proof It can be proved by applying Theorem 1. Note that $E_i^{\eta} = 0, \forall i \geq 1$ holds since $\boldsymbol{\eta}_i = \mathbf{0}$. In addition, it is straightforward to show $E_i^{\varepsilon} = 0$ as $\mathbf{v}_i$ is exact solution to (11) that leads to $\varepsilon_i = 0$ and $\mathbf{w}_i = \mathbf{0}, \forall i \geq 1$. Then, Theorem 2 can be proved by substituting $E_i^{\eta} = 0$ and $E_i^{\varepsilon} = 0$ into Theorem 1. This completes the proof. ■

Remark 2 *In fact, we present a unified analysis method that covers both non-accelerated and accelerated proximal methods for convex composite minimization* (1)*. In particular, the convergence rates of* PM *and* APM *can be obtained by a common convergence proof with different choices of parameters a_k. On the one hand, it is easy to see that $a_k = \frac{\gamma(k+1)}{2L}, \forall k \geq 1$ satisfies the condition $\frac{a_k^2}{A_k} \leq \frac{\gamma}{L}$. In this case, we have $A_k = \frac{\gamma k(k+3)}{4L}$. Then, we come up with $G_k \leq \frac{4LD_{\psi}(\mathbf{x}^{\star},\mathbf{x}_0)}{\gamma k(k+3)}$. This is essentially the optimal convergence rate $O(\frac{1}{k^2})$ of accelerated first-order algorithms for convex and smooth objectives (Nesterov, 1983). On the other hand, setting a_k to be a constant $a_k = \frac{\gamma}{L}$ is also valid for $\frac{a_k^2}{A_k} \leq \frac{\gamma}{L}$. In this case, we obtain $A_k = \frac{k\gamma}{L}$ and $G_k \leq \frac{LD_{\psi}(\mathbf{x}^{\star},\mathbf{x}_0)}{\gamma k}$. Indeed, it recovers non-accelerated first-order algorithms for convex and smooth objectives with convergence rate $O(\frac{1}{k})$.*

Next, we show the convergence of APM with inexact gradient oracle and approximate proximal mapping. By controlling the gradient noise $\boldsymbol{\eta}_k$, we will apply Theorem 1 to obtain specific convergence rates.

4.3 Convergence for Bounded $\mathcal{X}$

We first study the case when the domain $\mathcal{X}$ in (1) is bounded. By defining $R_{\mathbf{x}^{\star}} \overset{\text{def}}{=} \sup_{\mathbf{x}\in\mathcal{X}} \|\mathbf{x} - \mathbf{x}^{\star}\|$, the domain $\mathcal{X}$ is said to be bounded if $R_{\mathbf{x}^{\star}}$ is bounded.

Theorem 3 *Consider problem* (1) *where $\mathcal{X}$ is bounded. Assume f is L-smooth and $\frac{a_k^2}{A_k} \leq \frac{\gamma}{L}, \forall k \geq 1$. Let $\mathbf{x}_k, \mathbf{y}_k, \mathbf{v}_k$ evolve according to Algorithm 1, where $\mathbf{v}_k$ is a ε_k-optimal solution in expectation to* (11)*. If $\{\boldsymbol{\eta}_i\}_{i\geq 1}$ are independent random variables, then $\forall k \geq 1$:*

$$\mathbb{E}\big[G_k\big] \leq \frac{1}{A_k}\left(D_{\psi}(\mathbf{x}^{\star},\mathbf{x}_0) + R_{\mathbf{x}^{\star}}\sum_{i=1}^{k} a_i\mathbb{E}\big[\|\boldsymbol{\eta}_i\|_*\big] + \sum_{i=1}^{k}\Big(R_{\mathbf{x}^{\star}}\sqrt{2\xi a_i\varepsilon_i} + a_i\varepsilon_i\Big)\right).$$

Proof It can be proved by applying Theorem 1. Specifically, $\mathbb{E}\big[E_i^{\eta}\big]$ can be relaxed as

$$\mathbb{E}\big[E_i^{\eta}\big] = \mathbb{E}\big[a_i\langle\boldsymbol{\eta}_i, \mathbf{x}^{\star} - \mathbf{v}_i\rangle\big] \leq a_i\mathbb{E}\big[\|\boldsymbol{\eta}_i\|_*\|\mathbf{x}^{\star} - \mathbf{v}_i\|\big] \leq R_{\mathbf{x}^{\star}}a_i\mathbb{E}\big[\|\boldsymbol{\eta}_i\|_*\big], \quad (25)$$

where the first inequality follows by applying dual norm inequality $\langle\mathbf{x},\mathbf{y}\rangle \leq \|\mathbf{x}\|_*\|\mathbf{y}\|$. On the other hand, $\mathbb{E}\big[E_i^{\varepsilon}\big]$ can be relaxed as

$$\mathbb{E}\big[E_i^{\varepsilon}\big] = \mathbb{E}\big[a_i\langle\mathbf{w}_i, \mathbf{x}^{\star} - \mathbf{v}_i\rangle\big] + a_i\varepsilon_i \leq \mathbb{E}\big[a_i\|\mathbf{w}_i\|_*\|\mathbf{x}^{\star} - \mathbf{v}_i\|\big] + a_i\varepsilon_i \leq R_{\mathbf{x}^{\star}}\sqrt{2\xi a_i\varepsilon_i} + a_i\varepsilon_i, \quad (26)$$

where the last inequality follows from $\mathbb{E}[\|\mathbf{w}_i\|_*] \leq \sqrt{\mathbb{E}[\|\mathbf{w}_i\|_*^2]} \leq \sqrt{2\xi\varepsilon_i/a_i}$. Substituting (25) and (26) into (24), $A_k\mathbb{E}[G_k]$ becomes

$$A_k\mathbb{E}\big[G_k\big] \leq D_\psi(\mathbf{x}^\star, \mathbf{x}_0) + R_{\mathbf{x}^\star}\sum_{i=1}^k a_i\mathbb{E}\big[\|\boldsymbol{\eta}_i\|_*\big] + \sum_{i=1}^k \big(R_{\mathbf{x}^\star}\sqrt{2\xi a_i\varepsilon_i} + a_i\varepsilon_i\big).$$

Therefore, $\mathbb{E}[G_k]$ is upper bounded as

$$\mathbb{E}\big[G_k\big] \leq \frac{1}{A_k}\left(D_\psi(\mathbf{x}^\star, \mathbf{x}_0) + R_{\mathbf{x}^\star}\sum_{i=1}^k a_i\mathbb{E}\big[\|\boldsymbol{\eta}_i\|_*\big] + \sum_{i=1}^k \Big(R_{\mathbf{x}^\star}\sqrt{2\xi a_i\varepsilon_i} + a_i\varepsilon_i\Big)\right).$$

This completes the proof. ■

Remark 3 *In fact, the same bound on* $\mathbb{E}\big[G_k\big]$ *as Theorem 3 holds even if* $\{\boldsymbol{\eta}_i\}_{i\geq 1}$ *are not independent but* $\mathbb{E}\big[\|\boldsymbol{\eta}_i\|_*\big] \leq \sigma, \mathbb{E}\big[\|\boldsymbol{\eta}_i\|_*^2\big] \leq \delta, \forall i$. *Under this condition, we have*

$$\mathbb{E}\big[G_k\big] \leq \frac{1}{A_k}\left(D_\psi(\mathbf{x}^\star, \mathbf{x}_0) + \sigma R_{\mathbf{x}^\star}A_k + \sum_{i=1}^k \Big(R_{\mathbf{x}^\star}\sqrt{2\xi a_i\varepsilon_i} + a_i\varepsilon_i\Big)\right), \forall k \geq 1.$$

For smooth objectives, Cohen et al. (2018) demonstrate that the presence of domain boundary makes accelerated methods more robust with inexact gradient oracle, as the boundary of the feasible set limits the variance. Unlike (Cohen et al., 2018), the convex composite minimization problem (1) can be rewritten as an equivalent constrained optimization problem even it is unconstrained. Next, we show that the stronger regularization makes accelerated methods more robust with inexact gradient oracle when the non-smooth $h(\mathbf{x})$ is a regularization. Without loss of generality, we assume that $h(\mathbf{x})$ can be written as $h(\mathbf{x}) \overset{\text{def}}{=} \lambda g(\mathbf{x})$ where λ is a regularization parameter and $g(\mathbf{x})$ is a regularization, e.g., $g(\mathbf{x}) = \|\mathbf{x}\|_1$. By doing so, (1) becomes

$$\min_{\mathbf{x}\in\mathcal{X}} P(\mathbf{x}) \overset{\text{def}}{=} f(\mathbf{x}) + \lambda g(\mathbf{x}). \tag{27}$$

The larger value of λ leads to the stronger regularization. Let $\mathbf{x}_\lambda^\star$ be the optimal solution of (27) when the value of regularization parameter is λ. The next proposition discusses the effect of regularization parameter λ to the robustness of Algorithm 1 with inexact gradient oracle.

Proposition 3 *We consider the case of* $\mathcal{X} = \mathbb{R}^n$, *i.e.,* (27) *is an unconstrained optimization problem. The problem* (27) *is equivalent to*

$$\min_{\mathbf{x}} f(\mathbf{x}) \quad \text{s.t. } g(\mathbf{x}) \leq C_\lambda \quad \textit{where} \quad C_\lambda \overset{\text{def}}{=} g(\mathbf{x}_\lambda^\star). \tag{28}$$

Let $\widehat{\mathbf{x}}_\lambda^\star$ *be the optimal solution to* (28) *when the constraint parameter is* C_λ. *It holds that* $f(\widehat{\mathbf{x}}_\lambda^\star) = f(\mathbf{x}_\lambda^\star)$ *and* $g(\widehat{\mathbf{x}}_\lambda^\star) = g(\mathbf{x}_\lambda^\star)$. *As the increase of* λ, *Algorithm 1 becomes more robust with inexact gradient oracle because* C_λ *is decreased.*

Proof We first prove the equivalence between (27) and (28). The optimality condition of $\widehat{\mathbf{x}}_\lambda^\star$ to (28) implies

$$f(\widehat{\mathbf{x}}_\lambda^\star) \leq f(\mathbf{x}_\lambda^\star) \quad \text{and} \quad g(\widehat{\mathbf{x}}_\lambda^\star) \leq g(\mathbf{x}_\lambda^\star). \tag{29}$$

Combining them together, we obtain

$$f(\widehat{\mathbf{x}}_\lambda^\star) + \lambda g(\widehat{\mathbf{x}}_\lambda^\star) \leq f(\mathbf{x}_\lambda^\star) + \lambda g(\mathbf{x}_\lambda^\star). \tag{30}$$

On the other hand, the optimality condition of $\mathbf{x}_\lambda^\star$ to (27) implies

$$f(\mathbf{x}_\lambda^\star) + \lambda g(\mathbf{x}_\lambda^\star) \leq f(\widehat{\mathbf{x}}_\lambda^\star) + \lambda g(\widehat{\mathbf{x}}_\lambda^\star). \tag{31}$$

Combining (30) and (31), we come up with

$$f(\mathbf{x}_\lambda^\star) + \lambda g(\mathbf{x}_\lambda^\star) = f(\widehat{\mathbf{x}}_\lambda^\star) + \lambda g(\widehat{\mathbf{x}}_\lambda^\star).$$

Invoking (29), we come up with $f(\mathbf{x}_\lambda^\star) = f(\widehat{\mathbf{x}}_\lambda^\star)$ and $g(\mathbf{x}_\lambda^\star) = g(\widehat{\mathbf{x}}_\lambda^\star)$. Thus, the problem (27) is equivalent to (28).

Next, we prove the effect of λ to the robustness of Algorithm 1. For any $\lambda_1 \leq \lambda_2$, we show that $C_{\lambda_1} \geq C_{\lambda_2}$. Let $\widehat{\mathbf{x}}_{\lambda_1}^\star$ and $\widehat{\mathbf{x}}_{\lambda_2}^\star$ be the optimal solutions of (28) with different constraint parameters C_{λ_1} and C_{λ_2}, respectively. By the equivalence between (27) and (28), $\widehat{\mathbf{x}}_{\lambda_1}^\star$ and $\widehat{\mathbf{x}}_{\lambda_2}^\star$ are also the optimal solutions of (27) when the regularization parameter are λ_1 and λ_2, respectively. Their optimality conditions imply

$$\begin{aligned} f(\widehat{\mathbf{x}}_{\lambda_1}^\star) + \lambda_1 g(\widehat{\mathbf{x}}_{\lambda_1}^\star) &\leq f(\widehat{\mathbf{x}}_{\lambda_2}^\star) + \lambda_1 g(\widehat{\mathbf{x}}_{\lambda_2}^\star), \\ f(\widehat{\mathbf{x}}_{\lambda_2}^\star) + \lambda_2 g(\widehat{\mathbf{x}}_{\lambda_2}^\star) &\leq f(\widehat{\mathbf{x}}_{\lambda_1}^\star) + \lambda_2 g(\widehat{\mathbf{x}}_{\lambda_1}^\star). \end{aligned}$$

Summing up the above two inequalities, we obtain

$$(\lambda_1 - \lambda_2)\big(g(\widehat{\mathbf{x}}_{\lambda_1}^\star) - g(\widehat{\mathbf{x}}_{\lambda_2}^\star)\big) \leq 0.$$

Since $\lambda_1 \leq \lambda_2$, it implies $g(\widehat{\mathbf{x}}_{\lambda_1}^\star) \geq g(\widehat{\mathbf{x}}_{\lambda_2}^\star)$, thus $C_{\lambda_1} \geq C_{\lambda_2}$. We define $\mathcal{X}_{\lambda_1} \overset{\text{def}}{=} \{\mathbf{x} \mid g(\mathbf{x}) \leq C_{\lambda_1}\}$ and $\mathcal{X}_{\lambda_2} \overset{\text{def}}{=} \{\mathbf{x} \mid g(\mathbf{x}) \leq C_{\lambda_2}\}$. It is straightforward to show $\mathcal{X}_{\lambda_2} \subseteq \mathcal{X}_{\lambda_1}$. By the definition of $R_{\mathbf{x}^\star}$, we have $R_{\mathbf{x}_{\lambda_2}^\star} \leq R_{\mathbf{x}_{\lambda_1}^\star}$ due to $\mathcal{X}_{\lambda_2} \subseteq \mathcal{X}_{\lambda_1}$. Invoking Theorem 2, $\mathbb{E}\big[G_k\big]$ and $\mathrm{Var}\big[G_k\big]$ increase linearly and quadratically with respect to $R_{\mathbf{x}^\star}$, respectively. Thus, Algorithm 1 becomes more robust with inexact gradient oracle as the increase of λ because $R_{\mathbf{x}^\star}$ becomes smaller. ■

4.4 Convergence for Unbounded $\mathcal{X}$

The results presented in Theorem 3 only hold for the case when $\mathcal{X}$ is bounded. However, in many machine learning problems, the domain $\mathcal{X}$ is unbounded, e.g., $\mathcal{X} = \mathbb{R}^n$. To study the convergence rate of Algorithm 1 for the case of unbounded domain, we present theoretical results in the following theorem by assuming that the noise samples $\{\boldsymbol{\eta}_i\}_{i\geq 1}$ are zero-mean and independent.

Theorem 4 *Assume f is L-smooth and $\frac{a_k^2}{A_k} \leq \frac{\gamma}{L}, \forall k \geq 1$. Let $\mathbf{x}_k, \mathbf{y}_k, \mathbf{v}_k$ evolve according to Algorithm 1, where $\mathbf{v}_k$ is a ε_k-optimal to (11) in expectation and $\{\boldsymbol{\eta}_i\}_{i\geq 1}$ are zero-mean and independent random variables. Then $\forall k \geq 1$:*

$$\mathbb{E}[G_k] \leq \frac{1}{A_k}\left(\frac{3}{2}D_\psi(\mathbf{x}^\star, \mathbf{x}_0) + \sum_{i=1}^{k}\frac{3a_i^2}{\gamma}\mathbb{E}\big[\|\boldsymbol{\eta}_i\|_*^2\big] + \left(\frac{9}{4} + 4\sqrt{\frac{\xi}{\gamma}}\right)\left(\sum_{i=1}^{k}\sqrt{a_i\varepsilon_i}\right)^2\right). \tag{32}$$

Proof It can be proved by following the proof of Theorem 1. Applying $\mathbb{E}\big[D_\psi(\mathbf{x}^\star, \mathbf{v}_k)\big] \geq \frac{\gamma}{2}\mathbb{E}[\|\mathbf{x}^\star - \mathbf{v}_k\|^2]$ to (23), it becomes

$$A_k\mathbb{E}[G_k] + \frac{\gamma}{2}\mathbb{E}[\|\mathbf{x}^\star - \mathbf{v}_k\|^2] \leq D_\psi(\mathbf{x}^\star, \mathbf{v}_0) + \sum_{i=1}^{k}\mathbb{E}\big[E_i^\eta\big] + \sum_{i=1}^{k}\mathbb{E}\big[E_i^\varepsilon\big]. \tag{33}$$

Bounding $\mathbb{E}\big[E_i^\eta\big]$:
We define $\widehat{\mathbf{v}}_i^\star \overset{\text{def}}{=} \nabla\psi_i^*\big(\nabla\psi(\mathbf{v}_{i-1}) - a_i\nabla f(\mathbf{x}_i)\big) = \nabla\psi_i^*\big(\mathbf{z}_i + a_i\boldsymbol{\eta}_i\big)$ that is the optimal solution to (11) when both gradient oracle and proximal mapping are exact. Then, $\mathbb{E}\big[E_i^\eta\big]$ can be written as

$$\mathbb{E}\big[E_i^\eta\big] = \mathbb{E}\big[a_i\langle\boldsymbol{\eta}_i, \mathbf{x}^\star - \widehat{\mathbf{v}}_i^\star\rangle\big] + \mathbb{E}\big[a_i\langle\boldsymbol{\eta}_i, \widehat{\mathbf{v}}_i^\star - \mathbf{v}_i^\star\rangle\big] + \mathbb{E}\big[a_i\langle\boldsymbol{\eta}_i, \mathbf{v}_i^\star - \mathbf{v}_i\rangle\big]. \tag{34}$$

By the fact that $\widehat{\mathbf{v}}_i^\star$ is independent of $\boldsymbol{\eta}_i$ and $\mathbb{E}[\boldsymbol{\eta}_i] = \mathbf{0}$, we note that

$$\mathbb{E}\big[a_i\langle\boldsymbol{\eta}_i, \mathbf{x}^\star - \widehat{\mathbf{v}}_i^\star\rangle\big] = 0. \tag{35}$$

Regarding $\mathbb{E}\big[a_i\langle\boldsymbol{\eta}_i, \widehat{\mathbf{v}}_i^\star - \mathbf{v}_i^\star\rangle\big]$, it can be relaxed as

$$\mathbb{E}\big[a_i\langle\boldsymbol{\eta}_i, \widehat{\mathbf{v}}_i^\star - \mathbf{v}_i^\star\rangle\big] \leq a_i\mathbb{E}\big[\|\boldsymbol{\eta}_i\|_*\|\widehat{\mathbf{v}}_i^\star - \mathbf{v}_i^\star\|\big] \leq a_i\mathbb{E}\left[\|\boldsymbol{\eta}_i\|_*\frac{1}{\gamma}\|\mathbf{z}_i + a_i\boldsymbol{\eta}_i - \mathbf{z}_i\|_*\right] \leq \frac{a_i^2}{\gamma}\mathbb{E}\big[\|\boldsymbol{\eta}_i\|_*^2\big], \tag{36}$$

where the second inequality follows from ψ is γ-strongly convex w.r.t. $\|\cdot\|$ and Lemma 1. The strong convexity of $\Psi_i(\mathbf{v})$ and definition of $\mathbf{v}_i$ lead to

$$\varepsilon_i \geq \mathbb{E}\big[\Psi_i(\mathbf{v}_i) - \Psi_i(\mathbf{v}_i^\star)\big] \geq \frac{\gamma}{2a_i}\mathbb{E}\big[\|\mathbf{v}_i - \mathbf{v}_i^\star\|^2\big] \geq \frac{\gamma}{2a_i}\left(\mathbb{E}\big[\|\mathbf{v}_i - \mathbf{v}_i^\star\|\big]\right)^2.$$

Thus, $\mathbb{E}\big[a_i\langle\boldsymbol{\eta}_i, \mathbf{v}_i^\star - \mathbf{v}_i\rangle\big]$ can be bounded as

$$\mathbb{E}\big[a_i\langle\boldsymbol{\eta}_i, \mathbf{v}_i^\star - \mathbf{v}_i\rangle\big] \leq a_i\mathbb{E}\big[\|\boldsymbol{\eta}_i\|_*\|\mathbf{v}_i^\star - \mathbf{v}_i\|\big] \leq \sqrt{\frac{2a_i^3\varepsilon_i}{\gamma}}\mathbb{E}\big[\|\boldsymbol{\eta}_i\|_*\big]. \tag{37}$$

Bounding $\mathbb{E}\big[E_i^\varepsilon\big]$:
It can be bounded as

$$\mathbb{E}\big[a_i\big(\langle\mathbf{w}_i, \mathbf{x}^\star - \mathbf{v}_i\rangle + \varepsilon_i\big)\big] \leq a_i\varepsilon_i + a_i\mathbb{E}\big[\|\mathbf{w}_i\|_*\|\mathbf{x}^\star - \mathbf{v}_i\|\big] \leq a_i\varepsilon_i + \sqrt{2\xi a_i\varepsilon_i}\mathbb{E}\big[\|\mathbf{x}^\star - \mathbf{v}_i\|\big]. \tag{38}$$

where the last inequality follows from $\|\mathbf{w}_i\|_*^2 \le 2\xi\varepsilon_i/a_i$.
Now we are ready to prove (32). Substituting (35), (36) and (37) into (34), we obtain

$$\mathbb{E}\big[E_i^\eta\big] \le \frac{a_i^2}{\gamma}\mathbb{E}\big[\|\boldsymbol{\eta}_i\|_*^2\big] + \sqrt{\frac{2a_i^3\varepsilon_i}{\gamma}}\mathbb{E}\big[\|\boldsymbol{\eta}_i\|_*\big]. \tag{39}$$

Substituting (38) and (39) into (33),

$$\begin{aligned}&A_k\mathbb{E}[G_k] + \frac{\gamma}{2}\mathbb{E}\big[\|\mathbf{x}^\star - \mathbf{v}_k\|^2\big] \qquad (40)\\ &\le D_\psi(\mathbf{x}^\star, \mathbf{v}_0) + \sum_{i=1}^k\left(\frac{a_i^2}{\gamma}\mathbb{E}\big[\|\boldsymbol{\eta}_i\|_*^2\big] + \sqrt{\frac{2a_i^3\varepsilon_i}{\gamma}}\mathbb{E}\big[\|\boldsymbol{\eta}_i\|_*\big] + a_i\varepsilon_i\right) + \sum_{i=1}^k\sqrt{2\xi a_i\varepsilon_i}\mathbb{E}\big[\|\mathbf{x}^\star - \mathbf{v}_i\|\big].\end{aligned}$$

Note that

$$\sqrt{\frac{2a_i^3\varepsilon_i}{\gamma}}\mathbb{E}\big[\|\boldsymbol{\eta}_i\|_*\big] = \sqrt{\frac{2a_i^2}{\gamma}}\mathbb{E}\big[\|\boldsymbol{\eta}_i\|_*\big]\sqrt{a_i\varepsilon_i} \le \frac{a_i^2}{\gamma}\mathbb{E}\big[\|\boldsymbol{\eta}_i\|_*^2\big] + \frac{1}{2}a_i\varepsilon_i,$$

where the last inequality follows from $ab \le \frac{1}{2}(a^2+b^2)$ and $(\mathbb{E}[X])^2 \le \mathbb{E}[X^2]$. Substituting this result into (40),

$$\begin{aligned}&A_k\mathbb{E}[G_k] + \frac{\gamma}{2}\mathbb{E}\big[\|\mathbf{x}^\star - \mathbf{v}_k\|^2\big]\\ &\le D_\psi(\mathbf{x}^\star, \mathbf{v}_0) + \sum_{i=1}^k\left(\frac{2a_i^2}{\gamma}\mathbb{E}\big[\|\boldsymbol{\eta}_i\|_*^2\big] + \frac{3}{2}a_i\varepsilon_i\right) + \sum_{i=1}^k\sqrt{2\xi a_i\varepsilon_i}\mathbb{E}\big[\|\mathbf{x}^\star - \mathbf{v}_i\|\big]. \qquad (41)\end{aligned}$$

Since $\mathbb{E}[G_k] \ge 0$, it implies

$$\frac{\gamma}{2}\mathbb{E}\big[\|\mathbf{x}^\star - \mathbf{v}_k\|^2\big] \le D_\psi(\mathbf{x}^\star, \mathbf{v}_0) + \sum_{i=1}^k\left(\frac{2a_i^2}{\gamma}\mathbb{E}\big[\|\boldsymbol{\eta}_i\|_*^2\big] + \frac{3}{2}a_i\varepsilon_i\right) + \sum_{i=1}^k\sqrt{2\xi a_i\varepsilon_i}\mathbb{E}\big[\|\mathbf{x}^\star - \mathbf{v}_i\|\big].$$

Applying Lemma 6 with $S_k \overset{\text{def}}{=} D_\psi(\mathbf{x}^\star, \mathbf{v}_0) + \sum_{i=1}^k\left(\frac{2a_i^2}{\gamma}\mathbb{E}\big[\|\boldsymbol{\eta}_i\|_*^2\big] + \frac{3}{2}a_i\varepsilon_i\right)$, $\vartheta_i \overset{\text{def}}{=} 2\sqrt{\frac{\xi}{\gamma}a_i\varepsilon_i}$ and $u_i \overset{\text{def}}{=} \sqrt{\frac{\gamma}{2}}\mathbb{E}\big[\|\mathbf{x}^\star - \mathbf{v}_i\|\big]$, we obtain

$$\begin{aligned}&D_\psi(\mathbf{x}^\star, \mathbf{v}_0) + \sum_{i=1}^k\left(\frac{2a_i^2}{\gamma}\mathbb{E}\big[\|\boldsymbol{\eta}_i\|_*^2\big] + \frac{3}{2}a_i\varepsilon_i\right) + \sum_{i=1}^k\sqrt{2\xi a_i\varepsilon_i}\mathbb{E}\big[\|\mathbf{x}^\star - \mathbf{v}_i\|\big]\\ &\le \frac{3}{2}\left(D_\psi(\mathbf{x}^\star, \mathbf{v}_0) + \sum_{i=1}^k\left(\frac{2a_i^2}{\gamma}\mathbb{E}\big[\|\boldsymbol{\eta}_i\|_*^2\big] + \frac{3}{2}a_i\varepsilon_i\right) + \left(\sum_{i=1}^k 2\sqrt{\frac{\xi}{\gamma}a_i\varepsilon_i}\right)^2\right)\\ &\le \frac{3}{2}D_\psi(\mathbf{x}^\star, \mathbf{v}_0) + \sum_{i=1}^k\frac{3a_i^2}{\gamma}\mathbb{E}\big[\|\boldsymbol{\eta}_i\|_*^2\big] + \left(\frac{9}{4} + 4\sqrt{\frac{\xi}{\gamma}}\right)\left(\sum_{i=1}^k\sqrt{a_i\varepsilon_i}\right)^2.\end{aligned}$$

Substituting it into (41), we come up with

$$\mathbb{E}[G_k] \le \frac{1}{A_k}\left(\frac{3}{2}D_\psi(\mathbf{x}^\star, \mathbf{x}_0) + \sum_{i=1}^k\frac{3a_i^2}{\gamma}\mathbb{E}\big[\|\boldsymbol{\eta}_i\|_*^2\big] + \left(\frac{9}{4} + 4\sqrt{\frac{\xi}{\gamma}}\right)\left(\sum_{i=1}^k\sqrt{a_i\varepsilon_i}\right)^2\right).$$

This completes the proof. ■

Next, we present several corollaries of Theorem 4 to recover both accelerated and non-accelerated proximal methods for various cases and compare our results with related works. We generally assume the noisy gradient oracle and approximate proximal mapping satisfy $\mathbb{E}\big[\|\boldsymbol{\eta}_i\|_2^2\big] \leq \sigma^2, \varepsilon_i \leq \varepsilon, \forall i \geq 1$. Convergence results for exact gradient oracle and proximal mapping can be also recovered naturally from the corollaries. We start by applying Theorem 4 with step-size $a_i = \frac{\gamma(i+1)}{2L}$ and $a_i = \frac{\gamma}{L}$, which present convergence rates of accelerated and non-accelerated proximal methods for convex composite minimization problems, respectively.

Corollary 1 *Consider the same setting as Theorem 4 and assume* $\mathbb{E}\big[\|\boldsymbol{\eta}_i\|_*^2\big] \leq \sigma^2, \varepsilon_i \leq \varepsilon, \forall i \geq 1$, *where* σ^2 *and* ε *are constants. For APM with* $a_i = \frac{\gamma(i+1)}{2L}, \forall i \geq 1$, *we have*

$$\mathbb{E}[G_k] \leq \frac{6LD_\psi(\mathbf{x}^\star, \mathbf{x}_0)}{\gamma k(k+3)} + \frac{(2k+3)\sigma^2}{L} + \left(6 + \frac{32}{3}\sqrt{\frac{\xi}{\gamma}}\right)(k+2)\varepsilon, \forall k \geq 1. \tag{42}$$

For PM with $a_i = \frac{\gamma}{L}, \forall i \geq 1$, *we have*

$$\mathbb{E}[G_k] \leq \frac{3LD_\psi(\mathbf{x}^\star, \mathbf{x}_0)}{2\gamma k} + \frac{3\sigma^2}{L} + \left(\frac{9}{4} + 4\sqrt{\frac{\xi}{\gamma}}\right)k\varepsilon, \forall k \geq 1. \tag{43}$$

Consider the special case in which both gradient oracle and proximal mapping are exact (i.e., $\boldsymbol{\eta}_i = \mathbf{0}, \varepsilon_i = 0, \forall i \geq 1$), (42) and (43) recover the $O(1/k^2)$ (accelerated) and $O(1/k)$ (non-accelerated) convergence rates for convex composite objectives, respectively.

Remark 4 *The Corollary 1 suggests both noisy gradient oracle and approximate proximal mapping lead to error accumulation*[1] *for accelerated proximal methods. In contrast, the noisy gradient oracle does not lead to error accumulation for non-accelerated proximal methods. Thus, the acceleration comes at the expense of being less robust to noisy gradient oracle and approximate proximal mapping.*

For unbound composite minimization, Theorem 4 suggests that the acceleration leads to error accumulation for APM with noisy gradient oracle and proximal mapping as we have $a_k \sim O(k)$ and $A_k \sim O(k^2)$. Thus, the APM may fail to attain the optimal convergence rate, even produce divergent result. Nevertheless, the error accumulation of APM can be avoided if we postulate the magnitude of gradient noise (i.e., $\{\mathbb{E}\big[\|\boldsymbol{\eta}_i\|_*^2\big]\}_{i\geq 1}$) and error of approximate proximal mapping (i.e., $\{\epsilon_i\}_{i\geq 1}$) vanishes with the number of iterations. This can be achieved if the estimates of the gradient and the proximal mapping improve over iterations (Atchadé et al., 2017; Devolder et al., 2014).

[1] Following (Devolder et al., 2014; Cohen et al., 2018), the bound on the error incurred due to noisy gradient oracle or approximate proximal mapping does not accumulate if it is not increased as the number of iterations, otherwise there is error accumulation.

Table 1: Comparisons of convergence bound on gradient noise variance and proximal mapping error for convex composite minimization by assuming $\mathbb{E}\big[\|\boldsymbol{\eta}_i\|_2^2\big] \leq \sigma^2, \varepsilon_i \leq \varepsilon, \forall i \geq 1$.

	Method	σ^2	ε
Accelerated	(Schmidt et al., 2011, Proposition 2)	$O(k^2)$	$O(k^2)$
	(Kulunchakov and Mairal, 2019a, Proposition 4)	$O(k)$	$O(k^2)$
	Ours (59)	$O(k)$	$O(k)$
Non-accelerated	(Schmidt et al., 2011, Proposition 1)	$O(k)$	$O(k)$
	Ours (60)	$O(1)$	$O(k)$

Corollary 2 *Under the same setting as Theorem 4, we assume that* $\mathbb{E}\big[\|\boldsymbol{\eta}_i\|_*^2\big] \leq \sigma^2(i+1)^{-p}$ *and* $\epsilon_i \leq \varepsilon(i+1)^{-q}$ *hold for some* $\delta, \varepsilon, p, q$*. Set* $a_i = \frac{\gamma(i+1)}{2L}$ *for* APM*. For noisy gradient oracle and inexact proximal mapping, if* $p > 3$ *and* $q > 3$*,* $\forall k \geq 1$*:*

$$\mathbb{E}\big[G_k\big] \leq \frac{1}{k(k+3)}\left(\frac{6LD_\psi(\mathbf{x}^\star, \mathbf{x}_0)}{\gamma} + \frac{3\sigma^2}{L(p-3)} + \frac{18\gamma + 32\sqrt{\gamma\xi}}{\gamma(q-3)^2}\right). \tag{44}$$

Set $a_i = \frac{\gamma}{L}$ *for* PM*. For noisy gradient oracle and proximal mapping, if* $p > 1$ *and* $q > 2$*,* $\forall k \geq 1$*:*

$$\mathbb{E}\big[G_k\big] \leq \frac{1}{k}\left(\frac{3LD_\psi(\mathbf{x}^\star, \mathbf{x}_0)}{2\gamma} + \frac{3\sigma^2}{L(p-1)} + \frac{9\gamma + 16\sqrt{\gamma\xi}}{\gamma(q-2)^2}\varepsilon\right). \tag{45}$$

Remark 5 *For* APM *with approximate proximal mapping, our work suggests that the optimal convergence rate can be preserved if* $\{\epsilon_i\}_{i\geq 1}$ *vanishes faster than* $O(1/k^3)$ *while it requires at least* $O(1/k^4)$ *(Kulunchakov and Mairal, 2019a; Schmidt et al., 2011). This is consistent with the comparison presented in Section 4.4.1.*

4.4.1 Comparisons with Existing Works with Convexity

To understand the novelty of our results, we provide detailed comparisons with them under the same setting as Corollary 1, i.e., $\mathbb{E}\big[\|\boldsymbol{\eta}_i\|_2^2\big] \leq \sigma^2, \varepsilon_i \leq \varepsilon, \forall i \geq 1$, where σ^2 and ε are constants.

Acceleration with noisy gradient oracle and inexact proximal mapping has also been studied in (Kulunchakov and Mairal, 2019a; Schmidt et al., 2011). Following the setting of (Kulunchakov and Mairal, 2019a), we set $\kappa = L$ for convex objectives (i.e., $\mu = 0$). As shown in (11) of (Kulunchakov and Mairal, 2019a), we have $\delta_j = \sigma^2/L$. Then, the Proposition 4 of (Kulunchakov and Mairal, 2019a) implies

$$\mathbb{E}\big[P(\mathbf{x}_k) - P(\mathbf{x}^\star)\big] \leq \frac{2e^{1+\gamma}}{(k+1)^2}\left(L\|\mathbf{x}_0 - \mathbf{x}^\star\|^2 + \sum_{j=1}^{k}(j+1)^2\frac{\sigma^2}{L} + \sum_{j=1}^{k}\frac{(j+1)^{3+\gamma}\varepsilon}{\gamma}\right).$$

It is easy to see that $(j+1)^{3+\gamma} > (j+1)^3$ for any $\gamma \in (0,1]$. For convenience, we approximate $(j+1)^{3+\gamma}$ by $(j+1)^3$. Consequently, the error bound of (Kulunchakov and Mairal, 2019a) is at least

$$\mathbb{E}\left[P(\mathbf{x}_k) - P(\mathbf{x}^\star)\right] \leq \frac{2e^{1+\gamma}L\|\mathbf{x}_0 - \mathbf{x}^\star\|^2}{(k+1)^2} + \frac{e^{1+\gamma}(2k+6)\sigma^2}{3L} + \frac{e^{1+\gamma}}{2\gamma}(k+2)^2\varepsilon. \tag{46}$$

Similarly, the Proposition 2 of (Schmidt et al., 2011) implies

$$P(\mathbf{x}_k) - P(\mathbf{x}^\star) \leq \frac{2L}{(k+1)^2}\left(\|\mathbf{x}_0 - \mathbf{x}^\star\|^2 + 2\sum_{i=1}^{k} i\left(\frac{\sigma}{L} + \sqrt{\frac{2\varepsilon}{L}}\right) + \sqrt{2\sum_{i=1}^{k}\frac{i^2\varepsilon}{L}}\right)^2.$$

It can be rewritten as

$$P(\mathbf{x}_k) - P(\mathbf{x}^\star) \leq \frac{6L\|\mathbf{x}_0 - \mathbf{x}^\star\|^2}{(k+1)^2} + \frac{6k^2\sigma^2}{L} + \frac{4}{3L}(3k^2 + k)\varepsilon. \tag{47}$$

In addition, Schmidt et al. (2011) also studied the convergence rate of non-accelerated proximal methods. Specifically, the Proposition 1 of (Schmidt et al., 2011) implies

$$P(\mathbf{x}_k) - P(\mathbf{x}^\star) \leq \frac{3L\|\mathbf{x}_0 - \mathbf{x}^\star\|_2^2}{2k} + \frac{3k\sigma^2}{2L} + 12k\varepsilon. \tag{48}$$

Table 1 summarizes the detailed comparisons of (42) and (43) with (46), (47) and (48). From the comparison, we obtain the following conclusions:

- For the case of acceleration, our result is better than (Schmidt et al., 2011) in terms of both noisy gradient variance and approximate proximal mapping error and tighter than (Kulunchakov and Mairal, 2019a) in term of approximate proximal mapping error.

- For the case of without acceleration, our result is better than (Schmidt et al., 2011) in terms of noisy gradient variance.

Acceleration with noisy gradient oracle but exact proximal mapping has also been studied in (Kulunchakov and Mairal, 2020; Cohen et al., 2018; Ghadimi and Lan, 2012). Next, we show that our method also achieves the optimal sublinear convergence rate for finite horizon as aforementioned works.

Algorithm 2 Inexact APM for Strongly Convex Composite Minimization (1)

```
Input: starting point $\mathbf{x}_0$, strongly convex parameter $\mu$
$A_0 = 1$ and $\mathbf{y}_0 = \mathbf{v}_0 = \mathbf{x}_0$
for $i = 1$ to $k$ do
    Set $A_i \coloneqq A_{i-1} + a_i$
    Set $\mathbf{x}_i \coloneqq \frac{A_i}{A_i+a_i}\mathbf{y}_{i-1} + \frac{a_i}{A_i+a_i}\mathbf{v}_{i-1}$
    Solve
        $\mathbf{v}_i \approx \operatorname*{argmin}_{\mathbf{v}} \left\{ \left\langle \widetilde{\nabla} f(\mathbf{x}_i), \mathbf{v} \right\rangle + \frac{\mu}{2}\|\mathbf{v} - \mathbf{x}_i\|^2 + \frac{\mu A_{i-1}}{a_i} D_\psi(\mathbf{v}, \mathbf{v}_{i-1}) + h(\mathbf{v}) \right\} \quad (51)$
    such that $\Psi_i(\mathbf{v}_i) - \Psi_i(\mathbf{v}_i^\star) \leq \epsilon_i$.
    Set $\mathbf{y}_i \coloneqq \frac{A_{i-1}}{A_i}\mathbf{y}_{i-1} + \frac{a_i}{A_i}\mathbf{v}_i$
end for
Output: $\mathbf{y}_k$
```

Corollary 3 *Under the same setting as Corollary 1 but the proximal mapping is exact, i.e., $\varepsilon = 0$. Consider a fixed budget K of iterations of Algorithm 1. If $a_i = \frac{(i+1)\zeta}{2}$ where $\zeta = \min\left(\frac{\gamma}{L}, \sqrt{\frac{3\gamma D_\psi(\mathbf{x}^\star, \mathbf{x}_0)}{2\sigma^2(K+1)^3}}\right)$, then*

$$\mathbb{E}[G_K] \leq \frac{3LD_\psi(\mathbf{x}^\star, \mathbf{x}_0)}{\gamma(K+1)^2} + \sigma\sqrt{\frac{6D_\psi(\mathbf{x}^\star, \mathbf{x}_0)}{\gamma(K+1)}}. \tag{49}$$

If $a_i = \min\left(\frac{\gamma}{L}, \frac{1}{\sigma}\sqrt{\frac{\gamma D_\psi(\mathbf{x}^\star, \mathbf{x}_0)}{2K}}\right)$, then

$$\mathbb{E}[G_K] \leq \frac{3LD_\psi(\mathbf{x}^\star, \mathbf{x}_0)}{2\gamma K} + 3\sigma\sqrt{\frac{D_\psi(\mathbf{x}^\star, \mathbf{x}_0)}{2\gamma K}}. \tag{50}$$

Similar to (Kulunchakov and Mairal, 2020), (49) and (50) show that our analysis is able to recover the optimal noise-dependency for accelerated (Ghadimi and Lan, 2013) and non-accelerated (Nemirovski et al., 2009) stochastic first-order optimization, respectively.

5. APM for Strongly Convex Composite Minimization

In this section, we extend the study to μ-strongly convex $f(\mathbf{x})$. In this case, we slightly modify Algorithm 1 by considering the strong convexity of $f(\mathbf{x})$. Following the customary for smooth and strongly convex minimization (Nesterov, 2013; Bubeck, 2015; Cohen et al., 2018), we assume that $\|\cdot\| = \|\cdot\|_2$ so that $f(\mathbf{x})$ is L-smooth and μ-strongly w.r.t. the ℓ_2 norm in this setting. We take $\psi(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_2^2$ (i.e., $\gamma = 1, \xi = 1$) for simplicity as in (Cohen et al., 2018). Throughout this section, we only consider the case of $\mathcal{X}$ is unbounded, i.e., $\mathcal{X} = \mathbb{R}^n$.

Next, we apply the same idea presented in Section 4.1 to construct the approximate duality gap for μ-strongly convex $f(\mathbf{x})$. Same as before, we choose $U_k = P(\mathbf{y}_k)$ as the upper bound to $P(\mathbf{x}^\star)$ where $\mathbf{y}_k$ is the current solution. To construct a lower bound L_k to $P(\mathbf{x}^\star)$, we apply the μ-strong convexity of $P(\mathbf{x})$.

$$\begin{aligned} P(\mathbf{x}^\star) \geq\ & f(\mathbf{x}_k) + \big\langle \nabla f(\mathbf{x}_k), \mathbf{x}^\star - \mathbf{x}_k \big\rangle + \frac{\mu}{2}\|\mathbf{x}^\star - \mathbf{x}_k\|^2 + h(\mathbf{x}^\star) + \frac{\mu A_{k-1}}{a_k} D_\psi(\mathbf{x}^\star, \mathbf{v}_{k-1}) \\ & - \frac{\mu A_{k-1}}{a_k} D_\psi(\mathbf{x}^\star, \mathbf{v}_{k-1}). \end{aligned} \tag{52}$$

Substituting $\nabla f(\mathbf{x}_k) = \widetilde{\nabla} f(\mathbf{x}_k) - \boldsymbol{\eta}_k$ into (52), we obtain

$$\begin{aligned} P(\mathbf{x}^\star) \geq\ & \big\langle \widetilde{\nabla} f(\mathbf{x}_k), \mathbf{x}^\star - \mathbf{x}_k \big\rangle + \frac{\mu}{2}\|\mathbf{x}^\star - \mathbf{x}_k\|^2 + \frac{\mu A_{k-1}}{a_k} D_\psi(\mathbf{x}^\star, \mathbf{v}_{k-1}) + h(\mathbf{x}^\star) \\ & + f(\mathbf{x}_k) - \langle \boldsymbol{\eta}_k, \mathbf{x}^\star - \mathbf{x}_k \rangle - \frac{\mu A_{k-1}}{a_k} D_\psi(\mathbf{x}^\star, \mathbf{v}_{k-1}). \end{aligned}$$

Minimizing the right-hand side with respect to $\mathbf{x}^\star$, we obtain the lower bound L_k as following

$$\begin{aligned} L_k \overset{\text{def}}{=}\ & \min_{\mathbf{v}} \left\{ \big\langle \widetilde{\nabla} f(\mathbf{x}_k), \mathbf{v} - \mathbf{x}_k \big\rangle + \frac{\mu}{2}\|\mathbf{v} - \mathbf{x}_k\|^2 + \frac{\mu A_{k-1}}{a_k} D_\psi(\mathbf{v}, \mathbf{v}_{k-1}) + h(\mathbf{v}) \right\} \\ & + f(\mathbf{x}_k) - \langle \boldsymbol{\eta}_k, \mathbf{x}^\star - \mathbf{x}_k \rangle - \frac{\mu A_{k-1}}{a_k} D_\psi(\mathbf{x}^\star, \mathbf{v}_{k-1}). \end{aligned} \tag{53}$$

In view of (53), we define the following minimization problem for μ-strongly $f(\mathbf{x})$:

$$\mathbf{v}_i^\star \overset{\text{def}}{=} \operatorname*{argmin}_{\mathbf{v}} \Big\{ \overbrace{\big\langle \widetilde{\nabla} f(\mathbf{x}_i), \mathbf{v} \big\rangle + \tfrac{\mu}{2}\|\mathbf{v} - \mathbf{x}_i\|^2 + \tfrac{\mu A_{i-1}}{a_i} D_\psi(\mathbf{v}, \mathbf{v}_{i-1}) + h(\mathbf{v})}^{\Psi_i(\mathbf{v})} \Big\}. \tag{54}$$

In addition, the update of $\mathbf{x}_i$ in Algorithm 1 is changed as

$$\mathbf{x}_i := \frac{A_i}{A_i + a_i}\mathbf{y}_{i-1} + \frac{a_i}{A_i + a_i}\mathbf{v}_{i-1}.$$

For initialization, we set $\mathbf{y}_0 = \mathbf{v}_0 = \mathbf{x}_0$ and $A_0 = 1$. Algorithm 2 presents the detailed inexact APM for μ-strongly convex $f(\mathbf{x})$.

Similar to Section 4.1, we prove the convergence of Algorithm 2 by showing the approximate duality gap $G_k = U_k - L_k$ is converged. Note that $E_k = A_k G_k - A_{k-1} G_{k-1}, \forall k \geq 1$, thus

$$A_k G_k = A_0 G_0 + \sum_{i=1}^{k} E_i, \tag{55}$$

where initial gap G_0 is given by $G_0 = P(\mathbf{x}_0) - P(\mathbf{x}^\star)$ as $\mathbf{y}_0$ is initialized to be $\mathbf{x}_0$. In Section 5.1, we present a generic convergence result for G_k. Then, we apply it to obtain various specific convergence results in Section 5.2.

5.1 Generic Convergence Result

In view of (55), the main convergence argument is to bound $E_k, \forall k \geq 1$. To bound it, we need to first establish the upper bound of approximate $\mathbf{v}_k$ that is presented in Lemma 8 of Appendix B. By applying it, we can bound $E_k, \forall k \geq 1$, for μ-strongly convex $f(\mathbf{x})$ as follows.

Lemma 5 *Let $\mathbf{x}_k, \mathbf{y}_k, \mathbf{v}_k$ evolve according to Algorithm 2 where $\mathbf{v}_k$ is a ε_k-optimal solution to* (51) *in expectation. If $f(\mathbf{x})$ is μ-strongly convex and $0 < \frac{a_k}{A_k} \leq \sqrt{\frac{\mu}{L}}$, then there exists $\mathbf{w}_k$ with $\mathbb{E}[\|\mathbf{w}_k\|_2^2] \leq \frac{2\mu A_k \varepsilon_k}{a_k}$ such that,*

$$\mathbb{E}[E_k] \leq \mathbb{E}\left[E_k^{\eta} + E_k^{\varepsilon} + \frac{\mu A_{k-1}}{2}\|\mathbf{x}^\star - \mathbf{v}_{k-1}\|_2^2 - \frac{\mu A_k}{2}\|\mathbf{x}^\star - \mathbf{v}_k\|_2^2\right], \forall k \geq 1, \tag{56}$$

where $E_k^{\eta} \overset{\text{def}}{=} a_k\langle \boldsymbol{\eta}_k, \mathbf{x}^\star - \mathbf{v}_k\rangle$ and $E_k^{\varepsilon} \overset{\text{def}}{=} a_k\big(\langle \mathbf{w}_k, \mathbf{x}^\star - \mathbf{v}_k\rangle + \varepsilon_k\big)$.

Proof The detailed proof is provided in Appendix B.1. ■

Same as before, E_k^{η} comes from inexact gradient oracle and E_k^{ε} is incurred from approximate proximal mapping. Given Lemma 5, G_k can be bounded by applying $A_k G_k = A_0 G_0 + \sum_{i=1}^k E_i$. We define

$$\Delta_k \overset{\text{def}}{=} G_k + \frac{\mu}{2}\|\mathbf{x}^\star - \mathbf{v}_k\|^2 \quad \text{and} \quad \Theta_k \overset{\text{def}}{=} \prod_{i=1}^k (1 - \theta_i),$$

where $\theta_i = \frac{a_i}{A_i}, \forall i \geq 1$. In the next theorem, we prove a generic convergence result for the Δ_k.

Theorem 5 *Let $\mathbf{x}_k, \mathbf{y}_k, \mathbf{v}_k$ evolve according to Algorithm 2, where $\mathbf{v}_k$ is a ε_k-optimal solution to* (51) *in expectation. If $f(\mathbf{x})$ is μ-strongly convex and $\Delta_0 = P(\mathbf{x}_0) - P(\mathbf{x}^\star) + \frac{\mu}{2}\|\mathbf{x}_0 - \mathbf{x}^\star\|_2^2, A_0 = 1, 0 < \theta_i = \frac{a_i}{A_i} \leq \sqrt{\mu/L}, \forall i \geq 1$, then there exists $\mathbf{w}_i$ with $\mathbb{E}\left[\|\mathbf{w}_i\|_2^2\right] \leq 2\mu A_i \varepsilon_i / a_i$ such that,*

$$\mathbb{E}\left[\Delta_k\right] \leq \Theta_k \mathbb{E}\left[\Delta_0 + \sum_{i=1}^k \tfrac{\theta_i}{\Theta_i}\big(\langle \mathbf{w}_i + \boldsymbol{\eta}_i, \mathbf{x}^\star - \mathbf{v}_i\rangle + \varepsilon_i\big)\right], \forall k \geq 1. \tag{57}$$

Proof It can be proved by applying Lemma 5. The detailed proof is provided in Appendix B.2. ■

In the next section, we apply Theorem 5 and bound these error terms in (57) to present various specific convergence results.

5.2 Specific Convergence Results

Here, we show that both accelerated and non-accelerated convergence rates can be recovered from Theorem 5 with different choices of $\frac{a_i}{A_i}$. The next theorem presents the convergence result for the case of $\frac{a_i}{A_i} = \beta$ is fixed for all iterations.

Theorem 6 *Under the same setting as Theorem 5, let* $\Delta_0 = P(\mathbf{x}_0) - P(\mathbf{x}^\star) + \frac{\mu}{2}\|\mathbf{x}_0 - \mathbf{x}^\star\|_2^2$ *and* $\{\boldsymbol{\eta}_i\}_{i\geq 1}$ *are zero-mean and independent random variables. If* $\theta_i = \frac{a_i}{A_i} = \beta$ *and* $\beta \leq \sqrt{\mu/L}$*, then*

$$\mathbb{E}[G_k] \leq \big(1-\beta\big)^k \left(\frac{3}{2}\Delta_0 + \frac{3\beta^2}{\mu}\widetilde{G}_k^{\eta} + \frac{25\beta}{4}\widetilde{G}_k^{\varepsilon}\right), \tag{58}$$

where

$$\widetilde{G}_k^{\eta} \overset{\text{def}}{=} \sum_{i=1}^{k} \big(1-\beta\big)^{-i}\mathbb{E}\left[\|\boldsymbol{\eta}_i\|_2^2\right] \quad \textit{and} \quad \widetilde{G}_k^{\varepsilon} \overset{\text{def}}{=} \left(\sum_{i=1}^{k} \big(1-\beta\big)^{-i/2}\sqrt{\varepsilon_i}\right)^2.$$

Proof The detailed proof is provided in Appendix B.3. ■

Remark 6 *In fact, same as the case of general convex, Theorem 6 presents a unified convergence results for both accelerated and non-accelerated first-order proximal methods for* μ*-strongly convex composite minimization. Specifically, if both the gradient oracle and proximal mapping are exact, Theorem 6 reduces to*

$$\mathbb{E}[G_k] \leq \frac{3}{2}\Delta_0\big(1-\beta\big)^k.$$

It is easy to see that both $\frac{a_i}{A_i} = \beta = \sqrt{\mu/L}$ *and* $\frac{a_i}{A_i} = \beta = \mu/L$ *satisfy the condition presented in Theorem 6, that lead to the convergence results of accelerated and non-accelerated first-order proximal methods, respectively.*

Next, we present several corollaries of Theorem 6 to recover both accelerated and non-accelerated proximal methods for various cases and compare our results with related works. Same as before, we assume the noisy gradient oracle and approximate proximal mapping satisfy $\mathbb{E}\big[\|\boldsymbol{\eta}_i\|_2^2\big] \leq \sigma^2, \varepsilon_i \leq \varepsilon, \forall i \geq 1$. We start by applying Theorem 6 with $\frac{a_i}{A_i} = \beta = \sqrt{\mu/L}$ and $\frac{a_i}{A_i} = \beta = \mu/L$, which present accelerated and non-accelerated proximal methods for strongly convex objectives, respectively.

Corollary 4 *Consider the same setting as Theorem 6, where* f *is* μ*-strongly convex,* $\Delta_0 = P(\mathbf{x}_0) - P(\mathbf{x}^\star) + \frac{\mu}{2}\|\mathbf{x}_0 - \mathbf{x}^\star\|_2^2$*. Assume* $\mathbb{E}\big[\|\boldsymbol{\eta}_i\|_2^2\big] \leq \sigma^2$ *and* $\varepsilon_i \leq \varepsilon, \forall i \geq 1$*, where* σ^2 *and* ε *are constants. If we set* $\beta = \sqrt{\mu/L}$*, then* (58) *recovers the accelerated convergence rate of first-order proximal methods*

$$\mathbb{E}[G_k] \leq \frac{3}{2}\left(1 - \sqrt{\frac{\mu}{L}}\right)^k \Delta_0 + \frac{3\sigma^2}{\sqrt{\mu L}} + 25\sqrt{\frac{L}{\mu}}\varepsilon. \tag{59}$$

Table 2: Comparisons of error bounds on gradient noise variance and proximal mapping error for strongly convex composite minimization by assuming $\mathbb{E}\big[\|\boldsymbol{\eta}_i\|_2^2\big] \le \sigma^2, \varepsilon_i \le \varepsilon, \forall i \ge 1$.

	Method	σ^2	ε
Accelerated	(Schmidt et al., 2011, Proposition 4)	$O\left(\frac{L}{\mu^2}\right)$	$O\left(\left(\frac{L}{\mu}\right)^2\right)$
	(Kulunchakov and Mairal, 2019a, Proposition 4)	$O\left(\frac{1}{\sqrt{\mu L}}\right)$	$O\left(\frac{L}{\mu}\right)$
	Ours (59)	$O\left(\frac{1}{\sqrt{\mu L}}\right)$	$O\left(\sqrt{\frac{L}{\mu}}\right)$
Non-accelerated	Ours (60)	$O\left(\frac{1}{L}\right)$	$O\left(\frac{L}{\mu}\right)$

If we set $\beta = \mu/L$, then (58) *recovers the non-accelerated convergence rate of first-order proximal methods*

$$\mathbb{E}[G_k] \le \frac{3}{2}\left(1 - \frac{\mu}{L}\right)^k \Delta_0 + \frac{3\sigma^2}{L} + \frac{25L}{\mu}\varepsilon. \tag{60}$$

5.2.1 Comparisons with Existing Works with Strong Convexity

Same as Section 4.4.1, we also compare our results with related works under the same setting as Corollary 4, i.e., $\mathbb{E}\big[\|\boldsymbol{\eta}_i\|_2^2\big] \le \sigma^2, \varepsilon_i \le \varepsilon, \forall i \ge 1$, where σ^2 and ε are constants.

Acceleration with noisy gradient oracle and inexact proximal mapping has also been studied in (Kulunchakov and Mairal, 2019a; Schmidt et al., 2011). Following the setting of (Kulunchakov and Mairal, 2019a), we set $\kappa = L - \mu$ for μ-strongly convex objectives. As shown in (11) of (Kulunchakov and Mairal, 2019a), we have $\delta_j = \sigma^2/L$. Then, the Proposition 4 of (Kulunchakov and Mairal, 2019a) implies

$$\mathbb{E}\big[P(\mathbf{x}_k) - P(\mathbf{x}^\star)\big] \le \left(1 - \frac{1}{2}\sqrt{\frac{\mu}{L}}\right)^k \left(2\big(P(\mathbf{x}_0) - P(\mathbf{x}^\star)\big) + 4\sum_{j=1}^{k}\left(1 - \frac{1}{2}\sqrt{\frac{\mu}{L}}\right)^{-j}\left(\frac{\sigma^2}{L} + \sqrt{\frac{L}{\mu}}\varepsilon\right)\right).$$

Relaxing the right-hand side, it becomes

$$\mathbb{E}\big[P(\mathbf{x}_k) - P(\mathbf{x}^\star)\big] \le 2\left(1 - \frac{1}{2}\sqrt{\frac{\mu}{L}}\right)^k \big(P(\mathbf{x}_0) - P(\mathbf{x}^\star)\big) + \frac{8\sigma^2}{\sqrt{\mu L}} + \frac{8L}{\mu}\varepsilon. \tag{61}$$

Similarly, the Proposition 4 of (Schmidt et al., 2011) implies

$$P(\mathbf{x}_k) - P(\mathbf{x}^\star)$$

$$\leq \left(1-\sqrt{\frac{\mu}{L}}\right)^k \left(\sqrt{2\left(P(\mathbf{x}_0)-P(\mathbf{x}^\star)\right)}+\sqrt{\frac{2}{\mu}}\sum_{i=1}^k \left(1-\sqrt{\frac{\mu}{L}}\right)^{-\frac{1}{2}}\left(\sigma+\sqrt{2L\varepsilon}\right)+\sqrt{\sum_{i=1}^k \left(1-\sqrt{\frac{\mu}{L}}\right)^i \varepsilon}\right)^2 .$$

It can be further relaxed as

$$P(\mathbf{x}_k)-P(\mathbf{x}^\star) \leq 6\left(1-\sqrt{\frac{\mu}{L}}\right)^k \left(P(\mathbf{x}_0)-P(\mathbf{x}^\star)\right)+\frac{24L}{\mu^2}\sigma^2+\left(\frac{96L^2}{\mu^2}+\frac{24L}{\mu}\varepsilon\right). \quad (62)$$

Same as the convex case, Schmidt et al. (2011) also studied the convergence rate of non-accelerated proximal methods, i.e., the Proposition 3 of (Schmidt et al., 2011). However, they presented the convergence rate of $\|\mathbf{x}_k-\mathbf{x}^\star\|$ instead of $(P(\mathbf{x}_k)-P(\mathbf{x}^\star))$. In the case of inexact proximal mapping, it is not straightforward to obtain the convergence rate of $(P(\mathbf{x}_k)-P(\mathbf{x}^\star))$ from that of $\|\mathbf{x}_k-\mathbf{x}^\star\|$. Thus, we do not present the comparison with the Proposition 3 of (Schmidt et al., 2011).

Table 2 summarizes the detailed comparisons of (59) and (60) with (61) and (62). From the comparison, we observe that our result is better than (Schmidt et al., 2011) in terms of both noisy gradient variance and approximate proximal mapping error and tighter than (Kulunchakov and Mairal, 2019a) in term of approximate proximal mapping error. In particular, the advantages of our convergence bound is more significant when the problem is badly conditioned, i.e., $L \gg \mu$.

Acceleration with noisy gradient oracle but exact proximal mapping has also been studied in (Kulunchakov and Mairal, 2020; Cohen et al., 2018; Ghadimi and Lan, 2013). Next, we show that our method also achieves the optimal complexity similar to aforementioned works. We first derive a specific convergence result for this case by applying Theorem 5 with $\varepsilon_i = 0$ and $\mathbf{w}_i = \mathbf{0}, \forall i$.

Theorem 7 *Under the same setting as Theorem 5 but the proximal mapping is exact, i.e.,* $\varepsilon_i = 0$ *and* $\mathbf{w}_i = \mathbf{0}$. *If* $\theta_i = \frac{a_i}{A_i} \leq \sqrt{\mu/L}$, *then*

$$\mathbb{E}\left[\Delta_k\right] \leq \Theta_k\left(\Delta_0+\frac{1}{\mu}\sum_{i=1}^k \frac{\theta_i^2}{\Theta_i}\mathbb{E}\left[\|\boldsymbol{\eta}_i\|_2^2\right]\right), \forall k \geq 1. \quad (63)$$

Assume $\mathbb{E}\left[\|\boldsymbol{\eta}_i\|_2^2\right] \leq \sigma^2, \forall i \geq 1$, *where* σ^2 *is a constant. If we set* $\theta_i = \sqrt{\mu/L}, \forall i \geq 1$, *then* (63) *recovers the accelerated convergence rate of first-order proximal methods*

$$\mathbb{E}\left[\Delta_k\right] \leq \left(1-\sqrt{\frac{\mu}{L}}\right)^k \Delta_0+\frac{\sigma^2}{\sqrt{\mu L}}. \quad (64)$$

If we set $\theta_i = \mu/L, \forall i \geq 1$, *then* (63) *recovers the non-accelerated convergence rate of first-order proximal methods*

$$\mathbb{E}\left[\Delta_k\right] \leq \left(1-\frac{\mu}{L}\right)^k \Delta_0+\frac{\sigma^2}{L}. \quad (65)$$

The detailed proof of Theorem 7 is provided in Appendix B.3.

Next, we show that the worst-case complexity can be improved by a restart mechanism with decreasing θ_i, then we obtain an algorithm with optimal complexity similar to (Kulunchakov and Mairal, 2019a, 2020; Ghadimi and Lan, 2013). Suppose we aim to obtain a solution $\mathbf{y}_k$ such that $\mathbb{E}\big[P(\mathbf{y}_k) - P(\mathbf{x}^\star) + \frac{\mu}{2}\|\mathbf{y}_k - \mathbf{x}^\star\|_2^2\big] \le \epsilon$ where $\epsilon \le 2\sigma^2/\sqrt{\mu L}$ and $\epsilon \le 2\sigma^2/L$ for accelerated and non-accelerated proximal methods, respectively. The detailed procedure is presented in Algorithm 3.

Algorithm 3 Inexact APM with Restart

```
1: Input: starting point $\mathbf{x}_0$, strongly convex parameter $\mu$
2: Stage 1: Use $\mathbf{x}_0$ as the initialization. Run the Algorithm 2 with $\theta_i = \sqrt{\frac{\mu}{L}}$ and $\theta_i = \frac{\mu}{L}, \forall i \ge 1$ for accelerated and non-accelerated proximal methods, respectively. Stop the procedure until it obtains a solution $\widehat{\mathbf{y}}_{\widehat{k}}$ such that $\mathbb{E}\big[P(\widehat{\mathbf{y}}_{\widehat{k}}) - P(\mathbf{x}^\star) + \frac{\mu}{2}\|\widehat{\mathbf{y}}_{\widehat{k}} - \mathbf{x}^\star\|_2^2\big] \le \delta$ where $\delta \le 2\sigma^2/\sqrt{\mu L}$ and $\delta \le 2\sigma^2/L$ for accelerated and non-accelerated proximal methods, respectively.
3: Stage 2: Restart the Algorithm 2 by using $\widehat{\mathbf{y}}_{\widehat{k}}$ as the initialization and $\theta_i = \min\big(\sqrt{\frac{\mu}{L}}, \frac{2}{i+2}\big)$ and $\theta_i = \min\big(\frac{\mu}{L}, \frac{2}{i+2}\big)$ for accelerated and non-accelerated proximal methods, respectively, to obtain a solution $\mathbf{y}_k$ such that $\mathbb{E}\big[P(\mathbf{y}_k) - P(\mathbf{x}^\star) + \frac{\mu}{2}\|\mathbf{y}_k - \mathbf{x}^\star\|_2^2\big] \le \epsilon$.
```

Corollary 5 *Under the same setting as Theorem 7 and assume $\mathbb{E}\big[\|\boldsymbol{\eta}_i\|_2^2\big] \le \sigma^2, \forall i \ge 1$, where σ^2 is a constant. For accelerated proximal method, the complexity of the restart mechanism presented in Algorithm 3 to achieve $\mathbb{E}\big[P(\mathbf{y}_k) - P(\mathbf{x}^\star) + \frac{\mu}{2}\|\mathbf{y}_k - \mathbf{x}^\star\|_2^2\big] \le \epsilon$ is*

$$O\left(\sqrt{\frac{L}{\mu}}\log\left(\frac{2\Delta_0}{\epsilon}\right)\right) + O\left(\frac{\sigma^2}{\mu\epsilon}\right). \tag{66}$$

For non-accelerated proximal method, the complexity of the restart mechanism presented in Algorithm 3 to achieve $\mathbb{E}\big[P(\mathbf{y}_k) - P(\mathbf{x}^\star) + \frac{\mu}{2}\|\mathbf{y}_k - \mathbf{x}^\star\|_2^2\big] \le \epsilon$ is

$$O\left(\frac{L}{\mu}\log\left(\frac{2\Delta_0}{\epsilon}\right)\right) + O\left(\frac{\sigma^2}{\mu\epsilon}\right). \tag{67}$$

6. Extension to Bounded Variance Models

In this section, we show that how to generalize our approach to the bounded variance noise models from (Lan, 2012; Ghadimi and Lan, 2012). In such a model, we assume $\widetilde{\nabla} f(\mathbf{x}_i) = G(\mathbf{x}_i, \boldsymbol{\xi}_i)$ is an unbiased estimate of the gradient $\nabla f(\mathbf{x}_i)$ for all $i \ge 1$ and its variance is bounded by σ^2, where $\{\boldsymbol{\xi}_i\}_{i\ge 1}$'s are i.i.d. randoms vectors. The definition implies $\boldsymbol{\eta}_i = G(\mathbf{x}_i, \boldsymbol{\xi}_i) - \nabla f(\mathbf{x}_i)$, $\mathbb{E}[\boldsymbol{\eta}_i] = \mathbf{0}$ and $\mathbb{E}[\|\boldsymbol{\eta}_i\|_*^2] \le \sigma^2$ for all $i \ge 1$. Let $\mathcal{F}_k$ denote the natural filtration up to (and including) iteration k.

For simplicity, we assume the proximal mapping is exact that implies $\mathbf{v}_k = \mathbf{v}_k^\star$ and $\mathbb{E}[E_k^\varepsilon] = 0$. For convex composite minimization, we have the following convergence result.

Theorem 8 *Assume f is L-smooth and $\frac{a_k^2}{A_k} \leq \frac{\gamma}{L}, \forall k \geq 1$. Let $\mathbf{x}_k, \mathbf{y}_k, \mathbf{v}_k$ evolve according to Algorithm 1, where $\mathbf{v}_k$ is the exact solution to (11). In addition, $\widetilde{\nabla} f(\mathbf{x}_i) = G(\mathbf{x}_i, \boldsymbol{\xi}_i)$ is an unbiased estimate of the gradient $\nabla f(\mathbf{x}_i)$ for all $i \geq 1$ and its variance is upper bounded by σ^2, where $\{\boldsymbol{\xi}_i\}_{i\geq 1}$'s are i.i.d. randoms vectors. Then $\forall k \geq 1$:*

$$\mathbb{E}\left[G_k\right] \leq \frac{1}{A_k}\left(D_\psi(\mathbf{x}^\star, \mathbf{x}_0) + \sum_{i=1}^{k} \frac{a_i^2 \sigma^2}{\gamma}\right). \tag{68}$$

Proof We define $\widehat{\mathbf{v}}_k^\star \overset{\text{def}}{=} \nabla \psi_k^*\big(\nabla\psi(\mathbf{v}_{k-1}) - a_k \nabla f(\mathbf{x}_k)\big) = \nabla\psi_k^*\big(\mathbf{z}_k + a_k \boldsymbol{\eta}_k\big)$ that is the optimal solution to (11) when both gradient oracle and proximal mapping are exact. Note that $\widehat{\mathbf{v}}_k^\star$ is measurable w.r.t. $\mathcal{F}_{k-1}$ as $\{\mathbf{x}_i\}_{i=1}^k$ and $\{\boldsymbol{\xi}_i\}_{i=1}^{k-1}$ are measurable w.r.t. $\mathcal{F}_{k-1}$. It follows that

$$\mathbb{E}\big[E_k^\eta | \mathcal{F}_{k-1}\big] = \mathbb{E}\big[a_k \langle \boldsymbol{\eta}_k, \mathbf{x}^\star - \widehat{\mathbf{v}}_k^\star\rangle | \mathcal{F}_{k-1}\big] + \mathbb{E}\big[a_k \langle \boldsymbol{\eta}_k, \widehat{\mathbf{v}}_k^\star - \mathbf{v}_k^\star\rangle | \mathcal{F}_{k-1}\big] + \mathbb{E}\big[a_k \langle \boldsymbol{\eta}_k, \mathbf{v}_k^\star - \mathbf{v}_k\rangle | \mathcal{F}_{k-1}\big]. \tag{69}$$

Note $\mathbb{E}\big[a_k\langle \boldsymbol{\eta}_k, \mathbf{x}^\star - \widehat{\mathbf{v}}_k^\star\rangle | \mathcal{F}_{k-1}\big] = 0$ as $\widehat{\mathbf{v}}_k^\star$ is independent of $\boldsymbol{\eta}_k$ and $\mathbb{E}[\boldsymbol{\eta}_k] = \mathbf{0}$. In addition, it holds $\mathbb{E}\big[a_k \langle \boldsymbol{\eta}_k, \mathbf{v}_k^\star - \mathbf{v}_k\rangle | \mathcal{F}_{k-1}\big] = 0$ as the proximal mapping is exact, i.e., $\mathbf{v}_k^\star = \mathbf{v}_k$. Plugging them into (69), we obtain

$$\mathbb{E}\big[E_k^\eta | \mathcal{F}_{k-1}\big] = \mathbb{E}\big[a_k \langle \boldsymbol{\eta}_k, \widehat{\mathbf{v}}_k^\star - \mathbf{v}_k^\star\rangle | \mathcal{F}_{k-1}\big] \leq \frac{a_k^2}{\gamma}\mathbb{E}\big[\|\boldsymbol{\eta}_k\|_*^2\big] \leq \frac{a_k^2 \sigma^2}{\gamma}, \tag{70}$$

where the first inequality is obtained by applying the same arguments as the proof of (36). In order to bound $A_k G_k$, we next show that $\Gamma_k \overset{\text{def}}{=} A_k G_k + D_\psi(\mathbf{x}^\star, \mathbf{v}_k) - \sum_{i=1}^k \frac{a_i^2\sigma^2}{\gamma}$ is a supermartingale. Specifically,

$$\begin{aligned} \mathbb{E}\left[\Gamma_k - \Gamma_{k-1} | \mathcal{F}_{k-1}\right] &= \mathbb{E}\left[A_k G_k + D_\psi(\mathbf{x}^\star, \mathbf{v}_k) - A_{k-1}G_{k-1} - D_\psi(\mathbf{x}^\star, \mathbf{v}_{k-1}) - \frac{a_k^2\sigma^2}{\gamma}\Big| \mathcal{F}_{k-1}\right] \\ &= \mathbb{E}\left[E_k + D_\psi(\mathbf{x}^\star, \mathbf{v}_k) - D_\psi(\mathbf{x}^\star, \mathbf{v}_{k-1}) - \frac{a_k^2\sigma^2}{\gamma}\Big|\mathcal{F}_{k-1}\right] \\ &\leq \mathbb{E}\left[E_k^e + E_k^\eta + E_k^\varepsilon - \frac{a_k^2\sigma^2}{\gamma}\Big|\mathcal{F}_{k-1}\right] \leq \mathbb{E}\left[E_k^\eta - \frac{a_k^2\sigma^2}{\gamma}\Big|\mathcal{F}_{k-1}\right], \end{aligned}$$

where the second equality is obtained by applying the definition of E_k from (14) and the first inequality is obtained by plugging (17). The second inequality follows from $E_k^e \leq 0$ due to Proposition 2 and $E_k^\varepsilon = 0$ as the proximal mapping is exact. Combining it with (70), we obtain

$$\mathbb{E}\left[\Gamma_k - \Gamma_{k-1}|\mathcal{F}_{k-1}\right] \leq 0.$$

This shows Γ_k is a supermartingale. Hence, we can conclude that $\mathbb{E}[\Gamma_k] \leq \mathbb{E}[\Gamma_1]$.

$$\mathbb{E}\left[A_k G_k + D_\psi(\mathbf{x}^\star, \mathbf{v}_k) - \sum_{i=1}^k \frac{a_i^2\sigma^2}{\gamma}\right] \leq \mathbb{E}\left[A_1 G_1 + D_\psi(\mathbf{x}^\star, \mathbf{v}_1) - \frac{a_1^2\sigma^2}{\gamma}\right].$$

It can be rewritten as

$$\mathbb{E}\left[A_k G_k + D_\psi(\mathbf{x}^\star, \mathbf{v}_k)\right] \leq \mathbb{E}\left[A_1 G_1 - \left(D_\psi(\mathbf{x}^\star, \mathbf{v}_1) - \frac{a_1^2\sigma^2}{\gamma}\right)\right] + \sum_{i=1}^{k} \frac{a_i^2\sigma^2}{\gamma}. \tag{71}$$

Next, we show the upper bound of $\mathbb{E}[A_1 G_1]$. In view of (14) and $A_0 = 0$, we obtain $E_1 \overset{\text{def}}{=} A_1 G_1 - A_0 G_0 = A_1 G_1$. Combining this with (17),

$$\mathbb{E}[A_1 G_1] = \mathbb{E}[E_1] \leq \mathbb{E}\left[E_1^\eta\right] + \mathbb{E}\left[D_\psi(\mathbf{x}^\star, \mathbf{v}_0) - D_\psi(\mathbf{x}^\star, \mathbf{v}_1)\right],$$

where the inequality follows from $E_1^e \leq 0$ and $E_1^\varepsilon = 0$ due to Proposition 2 and the proximal mapping is exact, respectively. It is easy to show $\mathbb{E}\left[E_1^\eta\right] \leq \frac{a_1^2\sigma^2}{\gamma}$ by applying the same argument as (70). Thus,

$$\mathbb{E}[A_1 G_1] \leq \mathbb{E}\left[D_\psi(\mathbf{x}^\star, \mathbf{v}_0) + \left(\frac{a_1^2\sigma^2}{\gamma} - D_\psi(\mathbf{x}^\star, \mathbf{v}_1)\right)\right].$$

Substituting it into (71), we come up with

$$\mathbb{E}\left[G_k\right] \leq \frac{1}{A_k}\left(D_\psi(\mathbf{x}^\star, \mathbf{x}_0) + \sum_{i=1}^{k} \frac{a_i^2\sigma^2}{\gamma}\right).$$

This completes the proof for (68). ■

By choosing different values for a_i, Theorem 8 naturally recover both accelerated and non-accelerated convergence rates of first-order proximal methods.

Corollary 6 *Consider the same setting as Theorem 8. For APM with $a_i = \frac{\gamma(i+1)}{2L}, \forall i \geq 1$, we have*

$$\mathbb{E}[G_k] \leq \frac{4LD_\psi(\mathbf{x}^\star, \mathbf{x}_0)}{\gamma k(k+3)} + \frac{(2k+3)\sigma^2}{3L}. \tag{72}$$

For PM with $a_i = \frac{\gamma}{L}, \forall i \geq 1$, we have

$$\mathbb{E}[G_k] \leq \frac{LD_\psi(\mathbf{x}^\star, \mathbf{x}_0)}{\gamma k} + \frac{\sigma^2}{L}. \tag{73}$$

Similarly, we can also have the following convergence result strongly convex composite minimization.

Theorem 9 *Let $\mathbf{x}_k, \mathbf{y}_k, \mathbf{v}_k$ evolve according to Algorithm 2, where $\mathbf{v}_k$ is the exact solution to (51). In addition, $\widetilde{\nabla} f(\mathbf{x}_i) = G(\mathbf{x}_i, \boldsymbol{\xi}_i)$ is an unbiased estimate of the gradient $\nabla f(\mathbf{x}_i)$ for all $i \geq 1$ and its variance is bounded by σ^2, where $\{\boldsymbol{\xi}_i\}_{i\geq 1}$'s are i.i.d. randoms vectors. If $f(\mathbf{x})$ is μ-strongly convex, then $\forall \geq 1$:*

$$\mathbb{E}\left[G_k\right] \leq \frac{1}{A_k}\left(P(\mathbf{x}_0) - P(\mathbf{x}^\star) + \frac{\mu}{2}\|\mathbf{x}_0 - \mathbf{x}^\star\|_2^2 + \frac{\sigma^2}{\mu}\sum_{i=1}^{k} \frac{a_i^2}{A_i}\right). \tag{74}$$

Proof Following the proof of Theorem 8 until (70), it becomes

$$\begin{aligned}\mathbb{E}\big[E_k^{\eta}|\mathcal{F}_{k-1}\big] = \mathbb{E}\left[a_k\langle \boldsymbol{\eta}_k, \widehat{\mathbf{v}}_k^{\star} - \mathbf{v}_k^{\star}\rangle|\mathcal{F}_{k-1}\right] &\leq a_k\mathbb{E}\left[\|\boldsymbol{\eta}_k\|_2\|\widehat{\mathbf{v}}_k^{\star} - \mathbf{v}_k^{\star}\|_2|\mathcal{F}_{k-1}\right] \\ &\leq a_k\mathbb{E}\left[\|\boldsymbol{\eta}_i\|_2\tfrac{1}{\mu A_k}\|\mathbf{z}_k + a_k\boldsymbol{\eta}_k - \mathbf{z}_k\|_2\right] \\ &\leq \frac{a_k^2}{\mu A_k}\mathbb{E}\left[\|\boldsymbol{\eta}_k\|_2^2\right],\end{aligned}$$

where the second inequality follows from $\psi_k(\mathbf{v})$ is (μA_k)-strongly convex. Thus, we obtain

$$\mathbb{E}\big[E_k^{\eta}|\mathcal{F}_{k-1}\big] \leq \frac{a_k^2\sigma^2}{\mu A_k}. \tag{75}$$

In order to bound A_kG_k, we next show that $\Gamma_k \overset{\text{def}}{=} A_kG_k + \frac{\mu A_k}{2}\|\mathbf{x}^{\star} - \mathbf{v}_k\|_2^2 - \sum_{i=1}^{k}\frac{a_i^2\sigma^2}{\mu A_i}$ is a supermartingale. Specifically,

$$\begin{aligned}&\mathbb{E}\left[\Gamma_k - \Gamma_{k-1}|\mathcal{F}_{k-1}\right] \\ &= \mathbb{E}\left[A_kG_k + \frac{\mu A_k}{2}\|\mathbf{x}^{\star} - \mathbf{v}_k\|_2^2 - A_{k-1}G_{k-1} - \frac{\mu A_{k-1}}{2}\|\mathbf{x}^{\star} - \mathbf{v}_{k-1}\|_2^2 - \frac{a_k^2\sigma^2}{\mu A_k}\Big|\mathcal{F}_{k-1}\right] \\ &= \mathbb{E}\left[E_k + \frac{\mu A_k}{2}\|\mathbf{x}^{\star} - \mathbf{v}_k\|_2^2 - \frac{\mu A_{k-1}}{2}\|\mathbf{x}^{\star} - \mathbf{v}_{k-1}\|_2^2 - \frac{a_k^2\sigma^2}{\mu A_k}\Big|\mathcal{F}_{k-1}\right] \\ &\leq \mathbb{E}\left[E_k^{\eta} + E_k^{\varepsilon} - \frac{a_k^2\sigma^2}{\mu A_k}\Big|\mathcal{F}_{k-1}\right] \leq \mathbb{E}\left[E_k^{\eta} - \frac{a_k^2\sigma^2}{\mu A_k}\Big|\mathcal{F}_{k-1}\right],\end{aligned}$$

where the second equality is obtained by applying the definition of E_k from (14) and the first inequality is obtained by plugging (56). The second inequality follows from $E_k^{\varepsilon} = 0$ as the proximal mapping is exact. Combining it with (75), we obtain

$$\mathbb{E}\left[\Gamma_k - \Gamma_{k-1}|\mathcal{F}_{k-1}\right] \leq 0. \tag{76}$$

This shows Γ_k is a supermartingale. Hence, we can conclude that $\mathbb{E}[\Gamma_k] \leq \mathbb{E}[\Gamma_1]$.

$$\mathbb{E}\left[A_kG_k + \frac{\mu A_k}{2}\|\mathbf{x}^{\star} - \mathbf{v}_k\|_2^2 - \sum_{i=1}^{k}\frac{a_i^2\sigma^2}{\mu A_i}\right] \leq \mathbb{E}\left[A_1G_1 + \frac{\mu A_1}{2}\|\mathbf{x}^{\star} - \mathbf{v}_1\|_2^2 - \frac{a_1^2\sigma^2}{\mu A_1}\right].$$

It can be rewritten as

$$\mathbb{E}\left[A_kG_k + \frac{\mu A_k}{2}\|\mathbf{x}^{\star} - \mathbf{v}_k\|_2^2\right] \leq \mathbb{E}\left[A_1G_1 - \left(\frac{a_1^2\sigma^2}{\mu A_1} - \frac{\mu A_1}{2}\|\mathbf{x}^{\star} - \mathbf{v}_1\|_2^2\right)\right] + \sum_{i=1}^{k}\frac{a_i^2\sigma^2}{\mu A_i}. \tag{77}$$

Next, we show the upper bound of $\mathbb{E}[A_1G_1]$. Combining (14) with (17),

$$\mathbb{E}[A_1G_1 - A_0G_0] = \mathbb{E}[E_1] \leq \mathbb{E}\left[E_1^{\eta}\right] + \mathbb{E}\left[\frac{\mu A_0}{2}\|\mathbf{x}^{\star} - \mathbf{x}_0\|_2^2 - \frac{\mu A_1}{2}\|\mathbf{x}^{\star} - \mathbf{v}_1\|_2^2\right],$$

where the inequality follows from $E_1^{\varepsilon} = 0$ as the proximal mapping is exact. It is easy to show $\mathbb{E}\left[E_1^{\eta}\right] \leq \frac{a_1^2\sigma^2}{\mu A_1}$ by applying the same argument as (75). Thus,

$$\mathbb{E}[A_1G_1] \leq \mathbb{E}\left[\frac{a_1^2\sigma^2}{\mu A_1} - \frac{\mu A_1}{2}\|\mathbf{x}^{\star} - \mathbf{v}_1\|_2^2\right] + A_0G_0 + \frac{\mu A_0}{2}\|\mathbf{x}^{\star} - \mathbf{x}_0\|_2^2.$$

Substituting it into (77), we come up with

$$\mathbb{E}\left[G_k\right] \leq \frac{1}{A_k}\left(A_0 G_0 + \frac{\mu A_0}{2}\|\mathbf{x}^\star - \mathbf{x}_0\|_2^2 + \sum_{i=1}^{k}\frac{a_i^2\sigma^2}{\mu A_i}\right).$$

Substituting $A_0 = 1$, it can be rewritten as

$$\mathbb{E}\left[G_k\right] \leq \frac{1}{A_k}\left(P(\mathbf{x}_0) - P(\mathbf{x}^\star) + \frac{\mu}{2}\|\mathbf{x}_0 - \mathbf{x}^\star\|_2^2 + \frac{\sigma^2}{\mu}\sum_{i=1}^{k}\frac{a_i^2}{A_i}\right).$$

This completes the proof. ∎

By choosing different values for $\frac{a_i}{A_i}$, Theorem 9 can recover both accelerated and non-accelerated convergence rates of first-order proximal methods for μ-strongly convex objectives.

Corollary 7 *Consider the same setting as Theorem 9. If $f(\mathbf{x})$ is μ-strongly convex and $0 < \theta_i = \frac{a_i}{A_i} = \sqrt{\mu/L}, \forall i \geq 1$, then $\forall \geq 1$:*

$$\mathbb{E}\left[G_k\right] \leq \left(1 - \sqrt{\frac{\mu}{L}}\right)^k \left(P(\mathbf{x}_0) - P(\mathbf{x}^\star) + \frac{\mu}{2}\|\mathbf{x}_0 - \mathbf{x}^\star\|_2^2\right) + \frac{\sigma^2}{\sqrt{\mu L}}. \tag{78}$$

If $f(\mathbf{x})$ is μ-strongly convex and $0 < \theta_i = \frac{a_i}{A_i} = \mu/L, \forall i \geq 1$, then $\forall \geq 1$:

$$\mathbb{E}\left[G_k\right] \leq \left(1 - \frac{\mu}{L}\right)^k \left(P(\mathbf{x}_0) - P(\mathbf{x}^\star) + \frac{\mu}{2}\|\mathbf{x}_0 - \mathbf{x}^\star\|_2^2\right) + \frac{\sigma^2}{L}. \tag{79}$$

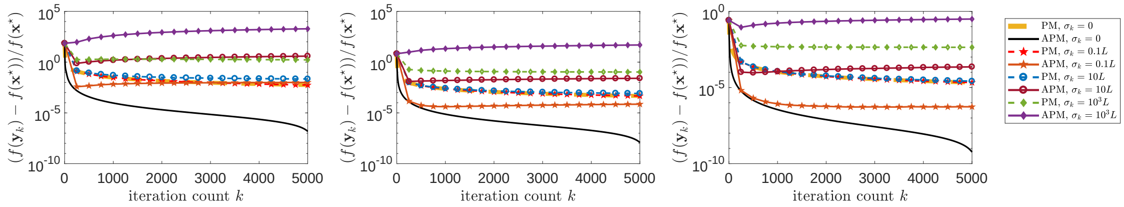
Same as Corollary 5, we can apply the restart mechanism to improve the worst-case complexity by using decreasing step-sizes a_i, i.e., the value of $\theta_i = a_i/A_i, \forall i \geq 1$ is decreasing. We skip the details as it is similar to Corollary 5.

7. Experiments

To demonstrate our theoretical results, we consider using the APM with noise-corrupted gradient and approximate proximal mapping to solve the Octagonal Shrinkage and Clustering Algorithm for Regression (OSCAR) (Bondell and Reich, 2008):

$$\min_{\mathbf{x}\in\mathbb{R}^n}\{\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \lambda_1\|\mathbf{x}\|_1 + \lambda_2\sum_{i<j}\max(|x_i|, |x_j|)\},$$

where $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{b} \in \mathbb{R}^m$. We use a synthetic dataset generated by the regression model $\mathbf{b} = \mathbf{A}\mathbf{x}^\star + \boldsymbol{\epsilon}$ with $m = 3000, n = 5000$. The model $\mathbf{x}^\star$ is generated by Matlab code: `repmat([zeros(85, 1); 3 * ones(10, 1); -3 * ones(5, 1)]), 50, 1)`. Each row of



(a) $\lambda_1 = 0.02, \lambda_2 = 0.04$ (b) $\lambda_1 = 0.2, \lambda_2 = 0.4$ (c) $\lambda_1 = 2, \lambda_2 = 4$

Figure 1: Results of PM and APM with inexact gradient oracle but exact proximal mapping.

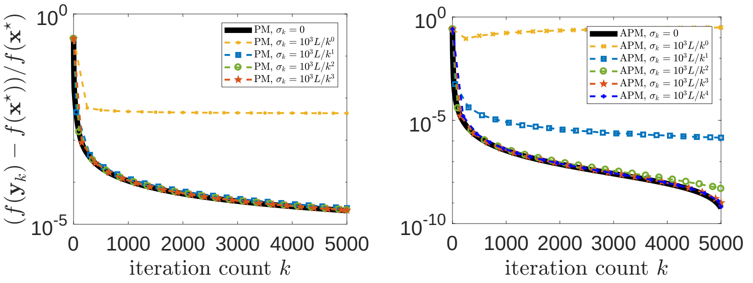
$\mathbf{A}$ is generated as $\mathbf{a}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ where $\Sigma_{ij} = 0.7^{|i-j|}$. Then b_i is obtained by $b_i = \langle \mathbf{a}_i, \mathbf{x}^\star \rangle + \epsilon_i$ where $\epsilon_i \sim \mathcal{N}(0, 1)$. We set $\psi(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_2^2$ in all experiments. In view of Remark 2, we set $a_k = \frac{1}{L}$ and $a_k = \frac{k+1}{2L}$ for Algorithm 1 to recover non-accelerated (PM) and accelerated (APM) proximal methods, respectively. To simulate inexact gradient oracle, we generate additive gradient noise $\boldsymbol{\eta}_k \sim \mathcal{N}(\mathbf{0}, \sigma_k \mathbf{I})$, where $\mathbf{I}$ is the identical matrix. The proximal mapping (8) for OSCAR is solved by the method proposed in (Zhong and Kwok, 2011) and the precision is controlled by ε_k, i.e., $\mathbf{v}_k$ is a ε_k-optimal solution to (8). To study the performance of Algorithm 1 for different values of λ_1 and λ_2, we test three pairs: $(0.02, 0.04)$, $(0.2, 0.4)$ and $(2, 4)$. We initialize $\mathbf{x}_0$ as a vector of zeros. For all experiments, we report the mean result of 20 random trials.

7.1 Results of Inexact Gradient Oracle

We first demonstrate the behavior of APM by only considering inexact gradient oracle while the proximal mapping is exact. For L-smooth f, we consider three kinds of gradient noise: $\boldsymbol{\eta}_k \sim \mathcal{N}(\mathbf{0}, 0.1L\mathbf{I}), \mathcal{N}(\mathbf{0}, 10L\mathbf{I})$ and $\mathcal{N}(\mathbf{0}, 10^3L\mathbf{I})$.

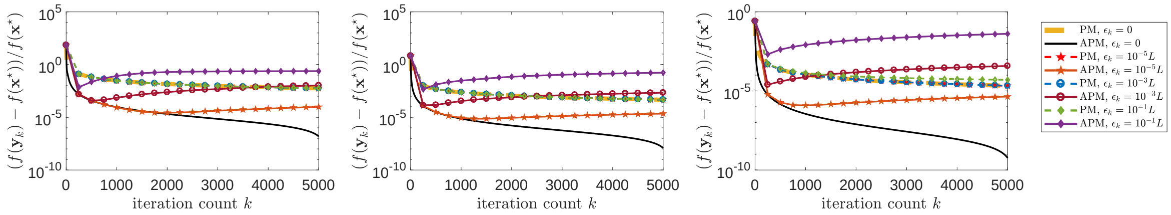
Figure 1 shows the results of PM and APM for three different groups of (λ_1, λ_2). In each group, the comparisons of PM and APM suggests that APM is less robust with inexact gradient oracle. Thus, the faster convergence rate of APM comes at the expense of being less robust to gradient noise as the noise becomes larger. Taking Figure 1 (b) for instance, APM performs significantly better than PM for exact gradient oracle as they converge to optimal solution as the rates of $O(1/k^2)$ and $O(1/k)$, respectively. When gradient is slightly corrupted (i.e., $\sigma_k = 0.1L$), APM shows significant performance degradation but still converges faster than PM. However, as the gradient noise becomes larger, APM fails to achieve the optimal convergence rate and it becomes worse than PM. In particular, for $\sigma_k = 10^3L$, APM even fails to decrease the objective value while PM still converges.

The comparison of Figure 1 (a), 1 (b) and 1 (c) suggests that APM becomes more robust to gradient noise as the increase of regularization parameter. Specifically, APM starts to perform worse than PM when $\sigma_k \geq 0.1L$ for regularization parameter $(0.02, 0.04)$. In contrast, APM still converges to smaller objective value than PM for $(0.2, 0.4)$ and



(a) (b)

Figure 2: Performance of PM and APM with inexact gradient oracle but exact proximal mapping for OSCAR with $\lambda_1 = 2$ and $\lambda_2 = 4$. The gradient oracle becomes more and more accurate as iteration count by setting $\sigma_k = 10^3L/k^p$.



(a) $\lambda_1 = 0.02, \lambda_2 = 0.04$ (b) $\lambda_1 = 0.2, \lambda_2 = 0.4$ (c) $\lambda_1 = 2, \lambda_2 = 4$

Figure 3: Results of PM and APM with approximate proximal mapping but exact gradient oracle for OSCAR.

$(2, 4)$. More importantly, the performance gap between APM and PM is more significant for $(2, 4)$ than that of $(0.2, 0.4)$. This can be interpreted by Proposition 3. By rewriting (1) as an equivalent constrained problem, larger λ_1 and λ_2 leads to smaller feasible set that can decrease the effect of gradient noise as stated in Theorem 3.

Figure 2 shows the performance of PM and APM if the magnitude of noise vanishes with the number of iterations. Specifically, the noise variance σ_k is set as $10^3L/k^p$ with $p = 1, 2, 3, 4$. For PM, it achieves $O(1/k)$ when $p > 1$. In contrast, it requires $p > 3$ for APM to achieve $O(1/k^2)$ convergence rate. This is consistent with Corollary 2 that suggests Algorithm 1 can preserve the optimal convergence rate of APM if $p > 3$ and proximal mapping is exact.

7.2 Results of Approximate Proximal Mapping

Next, we show the result of APM by using approximate proximal mapping but exact gradient oracle. Specifically, the approximate proximal mapping finds a $\mathbf{v}_k$ such that it is a

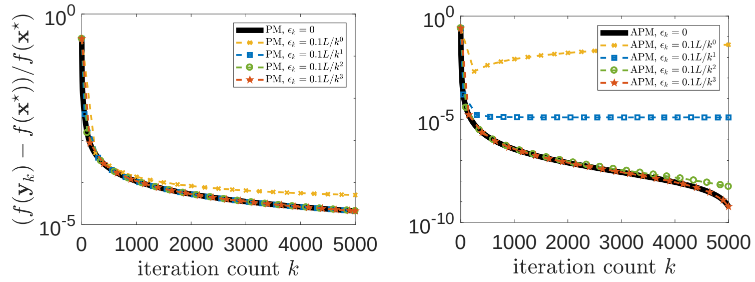
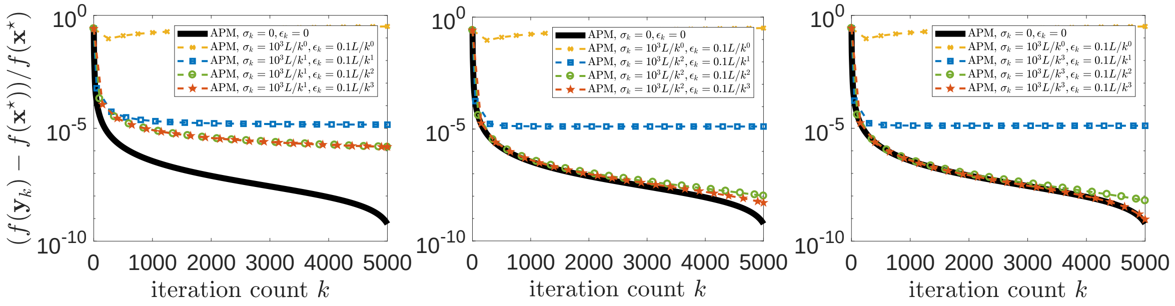


Figure 4: Performance of PM and APM with approximate proximal mapping but exact gradient oracle for OSCAR with $\lambda_1 = 2$ and $\lambda_2 = 4$. The proximal mapping becomes more and more accurate as iteration count by setting $\epsilon_k = 10^{-1}L/k^q$.



(a) (b) (c)

Figure 5: Performance of PM and APM with inexact gradient oracle and approximate proximal mapping simultaneously for OSCAR with $\lambda_1 = 2$ and $\lambda_2 = 4$. The gradient oralce and proximal mapping become more and more accurate as iteration count by setting $\sigma_k = 10^3 L/k^p$ and $\epsilon_k = 10^{-1}L/k^q$.

ε_k-optimal solution of (8) as Definition 5. For L-smooth f, we test three different precision for approximate proximal mapping: $\varepsilon_k = 10^{-5}L, 10^{-3}L$ and $10^{-1}L$.

Figure 3 demonstrates the performance of PM and APM for three different groups of (λ_1, λ_2). For fixed (λ_1, λ_2), the comparison of PM and APM implies that the performance of APM is also more sensitive with approximate proximal mapping than PM. Similar to the case of inexact gradient oracle, the faster convergence of APM over PM comes at the expense of being more sensitive with approximate proximal mapping. Thus, the APM fails to achieve the optimal convergence rate $O(1/k^2)$ as the approximate proximal mapping becomes less accurate. As shown in Figure 3, APM performs worse than PM and even fails to converge in the case of $\varepsilon_k = 10^{-3}L$ and $\varepsilon_k = 10^{-1}L$. In contrast, PM only shows slight performance degradation in most cases.

Comparing Figure 3 (a) with 3 (b) and 3 (c), we observe that the APM becomes less sensitive to the precision of approximate proximal mapping as the decrease of regularization parameter. Taking $\varepsilon_k = 10^{-5}L$ for example, APM performs better than PM for all three groups of (λ_1, λ_2). However, the performance gap is obviously larger for smaller (λ_1, λ_2). It is mainly because solving (11) is easier for smaller regularization parameter. In particular, it has an analytical solution if the regularization parameter is zero.

Figure 4 shows the performance of PM and APM if the magnitude of error incurred by approximate proximal mapping vanishes with the number of iterations. Specifically, the error ϵ_k is set as $0.1L/k^q$ with $q = 1, 2, 3$. For PM, it achieves $O(1/k)$ if $q > 2$ and gradient oracle is exact. In contrast, APM requires $q > 3$ for the same case. This is consistent with Corollary 2 that the optimal convergence rate of APM can be preserved if $q > 3$ and gradient oracle is exact. As discussed in Remark 5, it is better than existing results for approximate proximal mapping defined by Definition 5.

7.3 Results of Inexact Gradient Oracle and Approximate Proximal Mapping

Next, we demonstrate the performance of APM by simultaneously considering inexact gradient oracle and approximate proximal mapping. Suggested by Corollary 2, the noise variance σ_k is set as $10^3L/k^p$ with $p = 1, 2, 3$ and the error ε_k is set as $0.1L/k^q$ with $q = 1, 2, 3$. Figure 5 (a) to (c) show the performance of APM for $\sigma_k = 10^3L/k^1, 10^3L/k^2$ and $10^3L/k^3$, respectively. It should be compared with the case of inexact gradient oracle but exact proximal mapping in Figure 2. Specifically, in that case, the optimal convergence rate of APM can be preserved when σ_k decreases at a rate faster than $O(1/k^3)$. In contrast, if the proximal mapping is also inexact, it requires σ_k decreases also at a rate faster than $O(1/k^3)$.

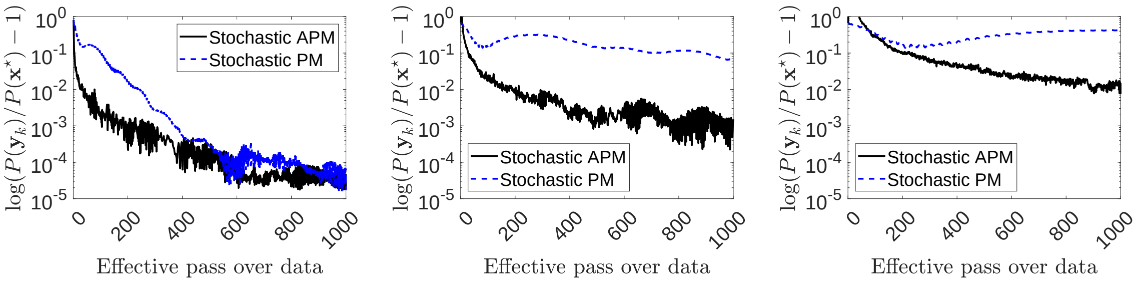
7.4 Results of Bounded Variance Models

In this section, we perform experiments to verify the performance of our method for bounded variance models presented in Section 6. Specifically, we consider the inexact gradient oracle $\widetilde{\nabla} f(\mathbf{x})$ is obtained by a mini-batch stochastic gradient. Following classical works in optimization methods for machine learning, e.g., (Kulunchakov and Mairal, 2020, 2019a; Schmidt et al., 2017), we consider the logistic regression problem. Give a training dataset by $\{(\mathbf{a}_i, b_i)\}_{i=1}^n$ where $\mathbf{a}_i \in \mathbb{R}^p$ and $b_i \in \{-1, 1\}$, the optimization formulation is

$$\min_{\mathbf{x}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^n \log\left(1 + \exp\left(-b_i\langle \mathbf{a}_i, \mathbf{x}\rangle\right)\right) + \frac{\lambda}{2}\|\mathbf{x}\|^2. \tag{80}$$

Note that the logistic loss function is convex and L-smooth where $L = 0.25$. The regularization is λ-strongly convex due to the squared ℓ_2-norm.

We run the Algorithm 3 on the `alpha` dataset that is from the Pascal Large Scale Learning Challenge website and it includes $n = 500,000$ samples in dimension $p = 500$. We compute $\widetilde{\nabla} f(\mathbf{x})$ by a mini-batch stochastic gradient with batch-size $m = 100$. Specifically, we consider three regularization parameters: $\lambda = 1/10n, \lambda = 1/100n$ and $\lambda = 1/1000n$



(a) $\lambda = 1/10n$ (b) $\lambda = 1/100n$ (c) $\lambda = 1/1000n$

Figure 6: Result of PM and APM with mini-batch stochastic gradient oracle on the `alpha` dataset for $\lambda = 1/10n, \lambda = 1/100n$ and $\lambda = 1/1000n$.

where n is number of training samples. We run each experiment five times with a different random initialization and report the average result of the five experiments in each figure. Figure 6 shows the results of APM and PM with mini-batch stochastic gradient oracle.

Comparing the results of APM and PM for each λ, we conclude that acceleration with decreased step size is effective even the gradient oracle is inexact. In addition, the results for different values of λ imply acceleration is more effective when the problem is badly condition, i.e., smaller λ. Specifically, the result of PM is similar to that of APM if $\lambda = 1/10n$, while there is significantly performance gap between them when $\lambda = 1/100n$ and $\lambda = 1/1000n$.

8. Conclusion

In this work, we have presented a study on APM with inexact gradient oracle and approximate proximal mapping. Our method is generic that naturally recover the convergence rates of both accelerated and non-accelerated first-order proximal methods. Our analysis achieves same convergence bound as previous works in terms of inexact gradient oracle, but a tighter convergence bound in terms of approximate proximal mapping that is more significant when the problem is badly conditioned. Numerical results on several datasets clearly corroborate our analysis.

Acknowledgments

Qiang Zhou is supported by National Science Foundation of China under Grant 62106045 and Southeast University Startup Fund (the Fundamental Research Funds for the Central Universities) under Grant 2242021R10096. Sinno J. Pan thanks the support of the 2020 Microsoft Research Asia collaborative research grant.

Appendix A. Omitted Proofs for Section 4

Following is a useful lemma on non-negative random sequences that will be used in proof of Theorem 4, that is inspired by (Schmidt et al., 2011; Lin et al., 2015).

Lemma 6 *We consider three non-negative sequences $\{S_k\}_{k\geq 0}$, $\{\vartheta_k\}_{k\geq 0}$ and $\{u_k\}_{k\geq 0}$ where $\{S_k\}_{k\geq 0}$ is increased and $\{u_k\}_{k\geq 0}$ is random. If $S_0 \geq \mathbb{E}[u_0^2]$ and $\forall k \geq 0$:*

$$\mathbb{E}[u_k^2] \leq S_k + \sum_{i=1}^{k} \vartheta_i \mathbb{E}[u_i]. \tag{81}$$

Then, $\forall k \geq 0$:

$$S_k + \sum_{i=1}^{k} \vartheta_i \mathbb{E}[u_i] \leq \frac{3}{2}\left(S_k + \left(\sum_{i=1}^{k} \vartheta_i\right)^2\right). \tag{82}$$

Proof It can be proved by using (Schmidt et al., 2011, Lemma 1) and the fact of $\mathbb{E}[u_k^2] \geq (\mathbb{E}[u_k])^2$. We first prove following inequality.

$$\mathbb{E}[u_k] \leq \frac{1}{2}\sum_{i=1}^{k} \vartheta_i + \left(S_k + \left(\frac{1}{2}\sum_{i=1}^{k} \vartheta_i\right)^2\right)^{1/2}. \tag{83}$$

First, it is straightforward to show (83) holds for $k = 0$. Then, we assume (83) hold for $k-1$ and prove it also true for k by induction. By the fact that $\mathbb{E}[u_k^2] \geq (\mathbb{E}[u_k])^2$, the inequality (81) implies

$$(\mathbb{E}[u_k])^2 \leq S_k + \sum_{i=1}^{k} \vartheta_i \mathbb{E}[u_i].$$

Applying (Schmidt et al., 2011, Lemma 1), we can obtain

$$\mathbb{E}[u_k] \leq \frac{1}{2}\sum_{i=1}^{k} \vartheta_i + \left(S_k + \left(\frac{1}{2}\sum_{i=1}^{k} \vartheta_i\right)^2\right)^{1/2}.$$

This completes the proof of (83). Relaxing the right-hand side of the above inequality, we obtain

$$\mathbb{E}[u_k] \leq \sqrt{S_k} + \sum_{i=1}^{k} \vartheta_i.$$

For any $i \leq k$, we have

$$\mathbb{E}[u_i] \leq \sqrt{S_i} + \sum_{j=1}^{i} \vartheta_j \leq \sqrt{S_k} + \sum_{i=1}^{k} \vartheta_i.$$

Substituting the upper bound of $\mathbb{E}[u_i]$ into (81), we obtain

$$S_k + \sum_{i=1}^{k} \vartheta_i \mathbb{E}[u_i] \le S_k + \sum_{i=1}^{k} \vartheta_i \left(\sqrt{S_k} + \sum_{i=1}^{k} \vartheta_i \right) \le \frac{3}{2} \left(S_k + \left(\sum_{i=1}^{k} \vartheta_i \right)^2 \right).$$

This completes the proof. ■

A.1 Proof of Proposition 2

Proof By the fact that f and ψ are L-smooth and γ-strongly convex w.r.t. $\|\cdot\|$, respectively, E_k^e can be upper bounded as

$$E_k^e = A_k\big(f(\mathbf{y}_k) - f(\mathbf{x}_k) - \big\langle \nabla f(\mathbf{x}_k), \mathbf{y}_k - \mathbf{x}_k \big\rangle\big) - D_\psi(\mathbf{v}_k, \mathbf{v}_{k-1}) \le A_k \frac{L}{2}\|\mathbf{y}_k - \mathbf{x}_k\|^2 - \frac{\gamma}{2}\|\mathbf{v}_k - \mathbf{v}_{k-1}\|^2.$$

Substituting $\mathbf{y}_k - \mathbf{x}_k = \frac{a_k}{A_k}(\mathbf{v}_k - \mathbf{v}_{k-1})$, it becomes

$$E_k^e \le A_k \frac{L}{2}\frac{a_k^2}{A_k^2}\|\mathbf{v}_k - \mathbf{v}_{k-1}\|^2 - \frac{\gamma}{2}\|\mathbf{v}_k - \mathbf{v}_{k-1}\|^2 \Rightarrow E_k^e = \frac{L}{2}\Big(\frac{a_k^2}{A_k} - \frac{\gamma}{L}\Big)\|\mathbf{v}_k - \mathbf{v}_{k-1}\|^2 \le 0.$$

This completes the proof. ■

A.2 Proof of Corollary 1

Proof It is straightforward to prove it by applying Theorem 4. Substituting $\mathbb{E}\big[\|\boldsymbol{\eta}_i\|_*^2\big] \le \sigma^2, \varepsilon_i \le \varepsilon, \forall i \ge 1$ into (32), we obtain

$$\mathbb{E}[G_k] \le \frac{1}{A_k}\left(\frac{3}{2} D_\psi(\mathbf{x}^\star, \mathbf{v}_0) + \frac{3\sigma^2}{\gamma}\sum_{i=1}^{k} a_i^2 + \left(\frac{9}{4} + 4\sqrt{\frac{\xi}{\gamma}}\right)\left(\sum_{i=1}^{k}\sqrt{a_i}\right)^2 \varepsilon \right), \forall k \ge 1. \quad (84)$$

For APM, we have $a_i = \frac{\gamma(i+1)}{2L}, \forall i \ge 1$ and $A_k = \frac{\gamma k(k+3)}{4L}$. Substituting these into (84), we come up with

$$\mathbb{E}[G_k] \le \frac{6LD_\psi(\mathbf{x}^\star, \mathbf{x}_0)}{\gamma k(k+3)} + \frac{(2k+3)\sigma^2}{L} + \left(6 + \frac{32}{3}\sqrt{\frac{\xi}{\gamma}}\right)(k+2)\varepsilon, \forall k \ge 1,$$

where we use the inequality $\sum_{i=1}^{k}\sqrt{i} \le \frac{2}{3}(k+1)^{3/2}$. This completes the proof of (42).
For PM, we have $a_i = \frac{\gamma}{L}, \forall i \ge 1$ and $A_k = \frac{\gamma k}{L}$. Substituting these into (84), we come up with

$$\mathbb{E}[G_k] \le \frac{3LD_\psi(\mathbf{x}^\star, \mathbf{x}_0)}{2\gamma k} + \frac{3\sigma^2}{L} + \left(\frac{9}{4} + 4\sqrt{\frac{\xi}{\gamma}}\right) k\varepsilon, \forall k \ge 1.$$

This completes the proof of (43). ■

A.3 Proof of Corollary 2

Proof It can be proved by applying Theorem 4.
Proof of (44): Note that we have $a_i = \frac{\gamma(i+1)}{2L}$ and $A_k = \frac{\gamma k(k+3)}{4L}$.
If $\mathbb{E}\big[\|\boldsymbol{\eta}_i\|_*^2\big] \leq \sigma(i+1)^{-p}$ with $p > 3$,

$$\sum_{i=1}^{k} \frac{3a_i^2}{\gamma}\mathbb{E}\big[\|\boldsymbol{\eta}_i\|_*^2\big] \leq \frac{3\gamma\sigma^2}{4L^2}\sum_{i=1}^{k}(i+1)^{(2-p)}.$$

Relaxing the summation to integral leads to

$$\sum_{i=1}^{k} \frac{3a_i^2}{\gamma}\mathbb{E}\big[\|\boldsymbol{\eta}_i\|_*^2\big] \leq \frac{3\gamma\sigma^2}{4L^2}\int_1^\infty x^{(2-p)}\,\mathrm{d}x = \frac{3\gamma\sigma^2}{4(3-p)L^2}x^{(3-p)}\Big|_1^\infty = \frac{3\gamma\sigma^2}{4(p-3)L^2}. \tag{85}$$

If $\epsilon_i \leq \varepsilon(i+1)^{-q}$ with $q > 3$,

$$\left(\frac{9}{4} + 4\sqrt{\frac{\xi}{\gamma}}\right)\left(\sum_{i=1}^{k}\sqrt{a_i\varepsilon_i}\right)^2 \leq \frac{9\gamma + 16\sqrt{\gamma\xi}}{8L}\left(\sum_{i=2}^{k+1} i^{\frac{1-q}{2}}\right)^2 \varepsilon$$

Relaxing the summation to integral leads to

$$\left(\frac{9}{4} + 4\sqrt{\frac{\xi}{\gamma}}\right)\left(\sum_{i=1}^{k}\sqrt{a_i\varepsilon_i}\right)^2 \leq \frac{9\gamma + 16\sqrt{\gamma\xi}}{8L}\varepsilon\left(\int_1^\infty x^{\frac{1-q}{2}}\,\mathrm{d}x\right)^2 = \frac{9\gamma + 16\sqrt{\gamma\xi}}{2L(q-3)^2}\varepsilon. \tag{86}$$

Substituting (85) and (86) into (32), we obtain

$$\mathbb{E}\big[G_k\big] \leq \frac{1}{k(k+3)}\left(\frac{6LD_\psi(\mathbf{x}^\star, \mathbf{x}_0)}{\gamma} + \frac{3\sigma^2}{L(p-3)} + \frac{18\gamma + 32\sqrt{\gamma\xi}}{\gamma(q-3)^2}\right).$$

This completes the proof for (44).
Proof of (45): Note that we have $a_i = \frac{\gamma}{L}$ and $A_k = \frac{\gamma k}{L}$.
If $\mathbb{E}\big[\|\boldsymbol{\eta}_i\|_*^2\big] \leq \sigma^2(i+1)^{-p}$ with $p > 1$,

$$\sum_{i=1}^{k} \frac{3a_i^2}{\gamma}\mathbb{E}\big[\|\boldsymbol{\eta}_i\|_*^2\big] \leq \frac{3\gamma\sigma^2}{L^2}\sum_{i=2}^{k+1} i^{-p}.$$

Relaxing the summation to integral leads to

$$\sum_{i=1}^{k} \frac{3a_i^2}{\gamma}\mathbb{E}\big[\|\boldsymbol{\eta}_i\|_*^2\big] \leq \frac{3\gamma\sigma^2}{L^2}\int_1^\infty x^{-p}\,\mathrm{d}x = \frac{3\gamma\sigma^2}{(1-p)L^2}x^{(1-p)}\Big|_1^\infty = \frac{3\gamma\sigma^2}{L^2(p-1)}. \tag{87}$$

If $\varepsilon_i \leq \varepsilon(i+1)^{-q}$ with $q > 2$,

$$\left(\frac{9}{4} + 4\sqrt{\frac{\xi}{\gamma}}\right)\left(\sum_{i=1}^{k}\sqrt{a_i\varepsilon_i}\right)^2 \leq \frac{9\gamma + 16\sqrt{\gamma\xi}}{4L}\left(\sum_{i=2}^{k+1} i^{-\frac{q}{2}}\right)^2 \varepsilon.$$

Relaxing the summation to integral leads to

$$\left(\frac{9}{4}+4\sqrt{\frac{\xi}{\gamma}}\right)\left(\sum_{i=1}^{k}\sqrt{a_i\varepsilon_i}\right)^2 \leq \frac{9\gamma+16\sqrt{\gamma\xi}}{4L}\left(\int_1^\infty x^{-\frac{q}{2}}\,\mathrm{d}x\right)^2\varepsilon = \frac{9\gamma+16\sqrt{\gamma\xi}}{L(q-2)^2}. \tag{88}$$

Substituting (87) and (88) into (32), we obtain

$$\mathbb{E}[G_k] \leq \frac{1}{k}\left(\frac{3LD_\psi(\mathbf{x}^\star,\mathbf{x}_0)}{2\gamma}+\frac{3\sigma^2}{L(p-1)}+\frac{9\gamma+16\sqrt{\gamma\xi}}{\gamma(q-2)^2}\varepsilon\right).$$

This completes the proof for (45). ■

A.4 Proof for Cororllary 3

Proof Substituting the value of $a_i = \frac{\zeta(i+1)}{2}, \mathbb{E}\left[\|\boldsymbol{\eta}_i\|_2^2\right] \leq \sigma^2$ and $\varepsilon_i = 0$ into (32), we obtain

$$\mathbb{E}[G_K] \leq \frac{3D_\psi(\mathbf{x}^\star,\mathbf{x}_0)}{\zeta K(K+3)}+\frac{\zeta\sigma^2(K+1)(K+2)(2K+3)}{2\gamma K(K+3)}.$$

Relaxing the right-hand side, we obtain

$$\mathbb{E}[G_K] \leq \frac{3D_\psi(\mathbf{x}^\star,\mathbf{x}_0)}{\zeta(K+1)^2}+\frac{2\zeta\sigma^2}{\gamma}(K+1).$$

Optimizing the right-hand size upper bound with respect to ζ, we obtain

$$\mathbb{E}[G_K] \leq \frac{3LD_\psi(\mathbf{x}^\star,\mathbf{x}_0)}{\gamma(K+1)^2}+\sigma\sqrt{\frac{6D_\psi(\mathbf{x}^\star,\mathbf{x}_0)}{\gamma(K+1)}}.$$

This completes the proof of (49).
Substituting the value of $a_i = \zeta, \mathbb{E}\left[\|\boldsymbol{\eta}_i\|_2^2\right] \leq \sigma^2$ and $\varepsilon_i = 0$ into (32), we obtain

$$\mathbb{E}[G_K] \leq \frac{3D_\psi(\mathbf{x}^\star,\mathbf{x}_0)}{2K\zeta}+\frac{3\sigma^2\zeta}{\gamma}.$$

Plugging $\zeta = \min\left(\frac{\gamma}{L}, \frac{1}{\sigma}\sqrt{\frac{\gamma D_\psi(\mathbf{x}^\star,\mathbf{x}_0)}{2K}}\right)$, we come up with

$$\mathbb{E}[G_K] \leq \frac{3LD_\psi(\mathbf{x}^\star,\mathbf{x}_0)}{2\gamma K}+3\sigma\sqrt{\frac{D_\psi(\mathbf{x}^\star,\mathbf{x}_0)}{2\gamma K}}.$$

This completes the proof of (50). ■

Appendix B. Omitted Proofs for Section 5

Before presenting proofs for main results, we first introduce two useful lemmas that allows us to bound the ε-subgradient of $h(\mathbf{v})$ at $\mathbf{v}_i$.

Lemma 7 *If $\mathbf{v}_i$ is an ε_i-optimal solution to* (51) *in expectation, then there exists $\mathbf{w}_i$ with $\mathbb{E}[\|\mathbf{w}_i\|_2^2] \leq 2\mu A_i \varepsilon_i / a_i$ such that*

$$\frac{\mu\big(a_i\mathbf{x}_i + A_{i-1}\mathbf{v}_{i-1} - A_i\mathbf{v}_i\big)}{a_i} - \widetilde{\nabla} f(\mathbf{x}_i) - \mathbf{w}_i \in \partial_{\varepsilon_i} h(\mathbf{v}_i).$$

Proof For convenience, we define

$$\Phi_i(\mathbf{v}) \overset{\text{def}}{=} \big\langle \widetilde{\nabla} f(\mathbf{x}_i), \mathbf{v} \big\rangle + \tfrac{\mu}{2}\|\mathbf{v} - \mathbf{x}_i\|_2^2 + \tfrac{\mu A_{i-1}}{a_i} D_\psi(\mathbf{v}, \mathbf{v}_{i-1}).$$

Then, $\Psi_i(\mathbf{v})$ can be rewritten as $\Psi_i(\mathbf{v}) = \Phi_i(\mathbf{v}) + h(\mathbf{v})$. By Definition 6, the ε_i-subdifferential in expectation of $\Phi_i(\mathbf{v})$ at $\mathbf{v}_i$ is

$$\partial_{\varepsilon_i} \Phi_i(\mathbf{v}_i) = \big\{ \mathbf{w} \bigm| \varepsilon_i \geq \mathbb{E}\left[\Phi_i(\mathbf{v}_i) + \langle \mathbf{w}, \mathbf{v} - \mathbf{v}_i \rangle - \Phi_i(\mathbf{v})\right], \forall \mathbf{v} \big\}.$$

It can be equivalent to

$$\partial_{\varepsilon_i} \Phi_i(\mathbf{v}_i) = \Big\{ \mathbf{w} \bigm| \varepsilon_i \geq \mathbb{E}\Big[\max_{\mathbf{v}} \left\{\Phi_i(\mathbf{v}_i) + \langle \mathbf{w}, \mathbf{v} - \mathbf{v}_i \rangle - \Phi_i(\mathbf{v})\right\}\Big] \Big\}.$$

Noting $D_\psi(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|_2^2$ and substituting $\Phi_i(\mathbf{v})$, it becomes

$$\partial_{\varepsilon_i} \Phi_i(\mathbf{v}_i) = \left\{ \mathbf{w} \Bigm| \mathbb{E}\left[\left\| \mathbf{w} - \widetilde{\nabla} f(\mathbf{x}_i) - \frac{\mu\big(A_i\mathbf{v}_i - a_i\mathbf{x}_i - A_{i-1}\mathbf{v}_{i-1}\big)}{a_i} \right\|_2^2\right] \leq \frac{2\mu A_i \varepsilon_i}{a_i} \right\}.$$

Thus, for any $\mathbf{z} \in \partial_{\varepsilon_i} \Phi_i(\mathbf{v}_i)$, it must can be expressed as the form

$$\mathbf{z} = \widetilde{\nabla} f(\mathbf{x}_i) + \frac{\mu\big(A_i\mathbf{v}_i - a_i\mathbf{x}_i - A_{i-1}\mathbf{v}_{i-1}\big)}{a_i} + \mathbf{w}_i \quad \text{with} \quad \mathbb{E}\left[\|\mathbf{w}_i\|_2^2\right] \leq \frac{2\mu A_i \varepsilon_i}{a_i}. \tag{89}$$

By Definition 5, $\mathbf{v}_i$ is an ε_i-optimal solution of $\Psi_i(\mathbf{v})$ in expectation if and only if

$$\mathbb{E}\Big[\Psi_i(\mathbf{v}_i) - \inf_{\mathbf{v}} \Psi_i(\mathbf{v})\Big] \leq \varepsilon_i.$$

Invoking Definition 6, this is equivalent to $\mathbf{0}$ belongings to the ε_i-subgradient in expectation of $\Psi_i(\mathbf{v}_i)$. Combining this with (Bertsekas et al., 2003, Proposition 4.3.1), we come up with

$$\mathbf{0} \in \partial_{\varepsilon_i} \Psi_i(\mathbf{v}_i) \subset \partial_{\varepsilon_i} \Phi_i(\mathbf{v}_i) + \partial_{\varepsilon_i} h(\mathbf{v}_i).$$

Therefore, there must exists some $\mathbf{z}$ such that $\mathbf{z} \in \partial_{\varepsilon_i}\Phi_i(\mathbf{v}_i)$ and $-\mathbf{z} \in \partial_{\varepsilon_i}h(\mathbf{v}_i)$. Invoking (89), there must exist $\mathbf{w}_i$ with $\mathbb{E}[\|\mathbf{w}_i\|_2^2] \leq 2\mu A_i\varepsilon_i/a_i$ such that

$$\frac{\mu\big(a_i\mathbf{x}_i + A_{i-1}\mathbf{v}_{i-1} - A_i\mathbf{v}_i\big)}{a_i} - \widetilde{\nabla} f(\mathbf{x}_i) - \mathbf{w}_i \in \partial_{\varepsilon_i}h(\mathbf{v}_i).$$

This completes the proof. ■

Lemma 8 *If $\mathbf{v}_i$ is an ε_i-optimal solution to* (51) *in expectation, then there exists $\mathbf{w}_i$ with $\mathbb{E}[\|\mathbf{w}_i\|_2^2] \leq 2\mu A_i\varepsilon_i/a_i$ such that, $\forall \mathbf{v} \in \mathcal{X}$,*

$$\mathbb{E}\left[\Psi_i(\mathbf{v}_i) + \frac{\mu A_i}{2a_i}\|\mathbf{v} - \mathbf{v}_i\|_2^2 - \big\langle \mathbf{w}_i, \mathbf{v} - \mathbf{v}_i\big\rangle - \Psi_i(\mathbf{v})\right] \leq \varepsilon_i.$$

Proof By the convexity of $\Phi_i(\mathbf{v})$, Definition 6 implies

$$\mathbb{E}\left[h(\mathbf{v}_i) - \big\langle \mathbf{w}, \mathbf{v} - \mathbf{v}_i\big\rangle\right] - \varepsilon_i \leq h(\mathbf{v}), \forall \mathbf{w} \in \partial_{\varepsilon_i}h(\mathbf{v}_i), \mathbf{v} \in \mathcal{X}.$$

Applying Lemma 7, there exits $\mathbf{w}_i$ with $\mathbb{E}[\|\mathbf{w}_i\|_2^2] \leq 2\mu A_i\varepsilon_i/a_i$ such that

$$h(\mathbf{v}) \geq \mathbb{E}\left[h(\mathbf{v}_i) + \left\langle \frac{\mu\big(a_i\mathbf{x}_i + A_{i-1}\mathbf{v}_{i-1} - A_i\mathbf{v}_i\big)}{a_i} - \widetilde{\nabla} f(\mathbf{x}_i) - \mathbf{w}_i, \mathbf{v} - \mathbf{v}_i\right\rangle\right] - \varepsilon_i.$$

It can be rewritten as

$$h(\mathbf{v}) \geq \mathbb{E}\left[h(\mathbf{v}_i) + \mu\big\langle \mathbf{x}_i - \mathbf{v}_i, \mathbf{v} - \mathbf{v}_i\big\rangle + \frac{\mu A_{i-1}}{a_i}\big\langle \mathbf{v}_{i-1} - \mathbf{v}_i, \mathbf{v} - \mathbf{v}_i\big\rangle - \big\langle \mathbf{w}_i + \widetilde{\nabla} f(\mathbf{x}_i), \mathbf{v} - \mathbf{v}_i\big\rangle\right] - \varepsilon_i. \tag{90}$$

Note that

$$\mu\big\langle \mathbf{x}_i - \mathbf{v}_i, \mathbf{v} - \mathbf{v}_i\big\rangle = \frac{\mu}{2}\|\mathbf{v} - \mathbf{v}_i\|_2^2 + \frac{\mu}{2}\|\mathbf{v}_i - \mathbf{x}_i\|_2^2 - \frac{\mu}{2}\|\mathbf{v} - \mathbf{x}_i\|_2^2,$$
$$\frac{\mu A_{i-1}}{a_i}\big\langle \mathbf{v}_{i-1} - \mathbf{v}_i, \mathbf{v} - \mathbf{v}_i\big\rangle = \frac{\mu A_{i-1}}{2a_i}\|\mathbf{v} - \mathbf{v}_i\|_2^2 + \frac{\mu A_{i-1}}{2a_i}\|\mathbf{v}_i - \mathbf{v}_{i-1}\|_2^2 - \frac{\mu A_{i-1}}{2a_i}\|\mathbf{v} - \mathbf{v}_{i-1}\|_2^2.$$

Substituting them into (90), we obtain

$$\begin{aligned} h(\mathbf{v}) \geq \mathbb{E}\Bigg[&h(\mathbf{v}_i) + \frac{\mu}{2}\|\mathbf{v} - \mathbf{v}_i\|_2^2 + \frac{\mu}{2}\|\mathbf{v}_i - \mathbf{x}_i\|_2^2 - \frac{\mu}{2}\|\mathbf{v} - \mathbf{x}_i\|_2^2 + \frac{\mu A_{i-1}}{2a_i}\|\mathbf{v} - \mathbf{v}_i\|_2^2 \\ &+ \frac{\mu A_{i-1}}{2a_i}\|\mathbf{v}_i - \mathbf{v}_{i-1}\|_2^2 - \frac{\mu A_{i-1}}{2a_i}\|\mathbf{v} - \mathbf{v}_{i-1}\|_2^2 - \big\langle \mathbf{w}_i + \widetilde{\nabla} f(\mathbf{x}_i), \mathbf{v} - \mathbf{v}_i\big\rangle\Bigg] - \varepsilon_i. \end{aligned}$$

By rearranging both sides, we obtain

$$\mathbb{E}\left[\Psi_i(\mathbf{v}_i) + \frac{\mu A_i}{2a_i}\|\mathbf{v} - \mathbf{v}_i\|_2^2 - \big\langle \mathbf{w}_i, \mathbf{v} - \mathbf{v}_i\big\rangle - \Psi_i(\mathbf{v})\right] \leq \varepsilon_i.$$

This completes the proof. ■

B.1 Proof for Lemma 5

Proof From the definition of E_k,

$$E_k = A_k P(\mathbf{y}_k) - A_{k-1}P(\mathbf{y}_{k-1}) - a_k P(\mathbf{x}^\star).$$

Substituting $P(\mathbf{y}_k)$, it becomes

$$E_k = A_k f(\mathbf{y}_k) - A_{k-1}f(\mathbf{y}_{k-1}) + A_k h\Big(\frac{A_{k-1}}{A_k}\mathbf{y}_{k-1} + \frac{a_k}{A_k}\mathbf{v}_k\Big) - A_{k-1}h(\mathbf{y}_{k-1}) - a_k P(\mathbf{x}^\star).$$

By convexity of $h(\mathbf{x})$, E_k is upper bounded as

$$E_k \leq A_k f(\mathbf{y}_k) - A_{k-1}f(\mathbf{y}_{k-1}) + a_k h(\mathbf{v}_k) - a_k P(\mathbf{x}^\star).$$

Taking expectations for both sides, we obtain

$$\mathbb{E}[E_k] \leq \mathbb{E}\left[A_k f(\mathbf{y}_k) - A_{k-1}f(\mathbf{y}_{k-1}) + a_k h(\mathbf{v}_k) - a_k P(\mathbf{x}^\star)\right]. \tag{91}$$

By strong convexity of $P(\mathbf{x})$, $a_k P(\mathbf{x}^\star)$ can be lower bounded by using $\mathbf{v}_k$.

$$\begin{aligned}
a_k P(\mathbf{x}^\star) \geq\ & \mathbb{E}\left[a_k\Big(f(\mathbf{x}_k) + \big\langle \nabla f(\mathbf{x}_k), \mathbf{x}^\star - \mathbf{x}_k\big\rangle + \frac{\mu}{2}\|\mathbf{x}^\star - \mathbf{x}_k\|_2^2 + h(\mathbf{x}^\star)\Big)\right] \\
&+ \mathbb{E}\left[\mu A_{k-1}\Big(D_\psi(\mathbf{x}^\star, \mathbf{v}_{k-1}) - D_\psi(\mathbf{x}^\star, \mathbf{v}_{k-1})\Big)\right] \\
=\ & \mathbb{E}\left[a_k\Big(\big\langle \widetilde{\nabla} f(\mathbf{x}_k), \mathbf{x}^\star - \mathbf{x}_k\big\rangle + h(\mathbf{x}^\star) + \frac{\mu}{2}\|\mathbf{x}^\star - \mathbf{x}_k\|_2^2 + \frac{\mu A_{k-1}}{a_k}D_\psi(\mathbf{x}^\star, \mathbf{v}_{k-1})\Big)\right] \\
&+ \mathbb{E}\left[a_k\Big(f(\mathbf{x}_k) - \langle \boldsymbol{\eta}_k, \mathbf{x}^\star - \mathbf{x}_k\rangle - \frac{\mu A_{k-1}}{a_k}D_\psi(\mathbf{x}^\star, \mathbf{v}_{k-1})\Big)\right].
\end{aligned}$$

Note that $D_\psi(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|_2^2$ in this case as $\psi(\cdot) = \frac{1}{2}\|\mathbf{x}\|_2^2$. The above inequality becomes

$$\begin{aligned}
a_k P(\mathbf{x}^\star) \geq\ & \mathbb{E}\left[a_k\Big(\big\langle \widetilde{\nabla} f(\mathbf{x}_k), \mathbf{x}^\star - \mathbf{x}_k\big\rangle + h(\mathbf{x}^\star) + \frac{\mu}{2}\|\mathbf{x}^\star - \mathbf{x}_k\|_2^2 + \frac{\mu A_{k-1}}{2a_k}\|\mathbf{x}^\star - \mathbf{v}_{k-1}\|_2^2\Big)\right] \\
&+ \mathbb{E}\left[a_k\Big(f(\mathbf{x}_k) - \langle \boldsymbol{\eta}_k, \mathbf{x}^\star - \mathbf{x}_k\rangle - \frac{\mu A_{k-1}}{2a_k}\|\mathbf{x}^\star - \mathbf{v}_{k-1}\|_2^2\Big)\right].
\end{aligned}$$

Applying Lemma 8 with $\mathbf{v} = \mathbf{x}^\star$, there exists $\mathbf{w}_k$ with $\mathbb{E}[\|\mathbf{w}_k\|_2^2] \leq 2\mu A_k \varepsilon_k / a_k$ such that

$$\begin{aligned}
a_k P(\mathbf{x}^\star) \geq\ & \mathbb{E}\Bigg[a_k\Big(\big\langle \widetilde{\nabla} f(\mathbf{x}_k), \mathbf{v}_k - \mathbf{x}_k\big\rangle + h(\mathbf{v}_k) + \frac{\mu}{2}\|\mathbf{v}_k - \mathbf{x}_k\|_2^2 + \frac{\mu A_{k-1}}{2a_k}\|\mathbf{v}_k - \mathbf{v}_{k-1}\|_2^2 \\
&\qquad + \frac{\mu A_k}{2a_k}\|\mathbf{x}^\star - \mathbf{v}_k\|_2^2\Big)\Bigg] \\
&+ \mathbb{E}\left[a_k\Big(f(\mathbf{x}_k) - \langle \boldsymbol{\eta}_k, \mathbf{x}^\star - \mathbf{x}_k\rangle - \frac{\mu A_{k-1}}{2a_k}\|\mathbf{x}^\star - \mathbf{v}_{k-1}\|_2^2\Big) - a_k\big(\langle \mathbf{w}_k, \mathbf{x}^\star - \mathbf{v}_k\rangle + \varepsilon_k\big)\right].
\end{aligned}$$

It can be rewritten as

$$a_k P(\mathbf{x}^\star)\ \geq \mathbb{E}\left[a_k f(\mathbf{x}_k) + a_k h(\mathbf{v}_k) + a_k\big\langle \nabla f(\mathbf{x}_k), \mathbf{v}_k - \mathbf{x}_k\big\rangle + \frac{\mu a_k}{2}\|\mathbf{v}_k - \mathbf{x}_k\|_2^2 + \frac{\mu A_{k-1}}{2}\|\mathbf{v}_k - \mathbf{v}_{k-1}\|_2^2\right]$$

$$+\mathbb{E}\left[\frac{\mu A_k}{2}\|\mathbf{x}^\star-\mathbf{v}_k\|_2^2-\frac{\mu A_{k-1}}{2}\|\mathbf{x}^\star-\mathbf{v}_{k-1}\|_2^2-a_k\big(\big\langle\mathbf{w}_k+\boldsymbol{\eta}_k,\mathbf{x}^\star-\mathbf{v}_k\big\rangle+\varepsilon_k\big)\right]. \quad (92)$$

Denote $\mathbf{u}_k=\frac{a_k}{A_k}\mathbf{x}_k+\frac{A_{k-1}}{A_k}\mathbf{v}_{k-1}$. By Jensen's inequality,

$$\frac{\mu a_k}{2}\|\mathbf{v}_k-\mathbf{x}_k\|_2^2+\frac{\mu A_{k-1}}{2}\|\mathbf{v}_k-\mathbf{v}_{k-1}\|_2^2=\frac{\mu A_k}{2}\Big(\frac{a_k}{A_k}\|\mathbf{v}_k-\mathbf{x}_k\|_2^2+\frac{A_{k-1}}{A_k}\|\mathbf{v}_k-\mathbf{v}_{k-1}\|_2^2\Big)\geq\frac{\mu A_k}{2}\|\mathbf{v}_k-\mathbf{u}_k\|_2^2.$$

Using $\frac{a_k}{A_k}\leq\sqrt{\frac{\mu}{L}}$, it becomes

$$\frac{\mu a_k}{2}\|\mathbf{v}_k-\mathbf{x}_k\|_2^2+\frac{\mu A_{k-1}}{2}\|\mathbf{v}_k-\mathbf{v}_{k-1}\|_2^2\geq\frac{LA_k}{2}\frac{\mu}{L}\|\mathbf{v}_k-\mathbf{u}_k\|_2^2\geq\frac{LA_k}{2}\Big\|\frac{a_k}{A_k}\big(\mathbf{v}_k-\mathbf{u}_k\big)\Big\|_2^2.$$

In addition,

$$\begin{aligned}a_k\big\langle\nabla f(\mathbf{x}_k),\mathbf{v}_k-\mathbf{x}_k\big\rangle&=a_k\big\langle\nabla f(\mathbf{x}_k),\mathbf{v}_k-\mathbf{u}_k\big\rangle+a_k\Big\langle\nabla f(\mathbf{x}_k),\frac{A_{k-1}}{A_k}\mathbf{v}_{k-1}+\frac{a_k}{A_k}\mathbf{x}_k-\mathbf{x}_k\Big\rangle\\&=a_k\big\langle\nabla f(\mathbf{x}_k),\mathbf{v}_k-\mathbf{u}_k\big\rangle+A_{k-1}\Big\langle\nabla f(\mathbf{x}_k),\frac{a_k}{A_k}\big(\mathbf{v}_{k-1}-\mathbf{x}_k\big)\Big\rangle. \quad (93)\end{aligned}$$

Combining them with L-smoothness of $f(\mathbf{x})$,

$$\begin{aligned}&a_k\big\langle\nabla f(\mathbf{x}_k),\mathbf{v}_k-\mathbf{u}_k\big\rangle+\frac{\mu a_k}{2}\|\mathbf{v}_k-\mathbf{x}_k\|_2^2+\frac{\mu A_{k-1}}{2}\|\mathbf{v}_k-\mathbf{v}_{k-1}\|_2^2\\&\geq a_k\big\langle\nabla f(\mathbf{x}_k),\mathbf{v}_k-\mathbf{u}_k\big\rangle+\frac{LA_k}{2}\Big\|\frac{a_k}{A_k}\big(\mathbf{v}_k-\mathbf{u}_k\big)\Big\|_2^2\\&=A_k\Big\{\Big\langle\nabla f(\mathbf{x}_k),\mathbf{x}_k+\frac{a_k}{A_k}\big(\mathbf{v}_k-\mathbf{u}_k\big)-\mathbf{x}_k\Big\rangle+\frac{L}{2}\Big\|\mathbf{x}_k+\frac{a_k}{A_k}\big(\mathbf{v}_k-\mathbf{u}_k\big)-\mathbf{x}_k\Big\|_2^2\Big\}.\end{aligned}$$

Since f is L-smooth w.r.t. $\|\cdot\|$, it becomes

$$a_k\big\langle\nabla f(\mathbf{x}_k),\mathbf{v}_k-\mathbf{u}_k\big\rangle+\frac{\mu a_k}{2}\|\mathbf{v}_k-\mathbf{x}_k\|_2^2+\frac{\mu A_{k-1}}{2}\|\mathbf{v}_k-\mathbf{v}_{k-1}\|_2^2\geq A_kf\Big(\mathbf{x}_k+\frac{a_k}{A_k}\big(\mathbf{v}_k-\mathbf{u}_k\big)\Big)-A_kf(\mathbf{x}_k).$$

Note that $\mathbf{y}_k=\mathbf{x}_k+\frac{a_k}{A_k}\big(\mathbf{v}_k-\mathbf{u}_k\big)$, we come up with

$$a_k\big\langle\nabla f(\mathbf{x}_k),\mathbf{v}_k-\mathbf{u}_k\big\rangle+\frac{\mu a_k}{2}\|\mathbf{v}_k-\mathbf{x}_k\|_2^2+\frac{\mu A_{k-1}}{2}\|\mathbf{v}_k-\mathbf{v}_{k-1}\|_2^2\geq A_kf(\mathbf{y}_k)-A_kf(\mathbf{x}_k). \quad (94)$$

By definition of $\mathbf{u}_k$, the second term of (93) can be written as

$$A_{k-1}\Big\langle\nabla f(\mathbf{x}_k),\frac{a_k}{A_k}\big(\mathbf{v}_{k-1}-\mathbf{x}_k\big)\Big\rangle=A_{k-1}\big\langle\nabla f(\mathbf{x}_k),\mathbf{x}_k-\mathbf{y}_{k-1}\big\rangle. \quad (95)$$

Combining (94) and (95), we obtain

$$\begin{aligned}&a_k\big\langle\nabla f(\mathbf{x}_k),\mathbf{v}_k-\mathbf{x}_k\big\rangle+\frac{\mu a_k}{2}\|\mathbf{v}_k-\mathbf{x}_k\|_2^2+\frac{\mu A_{k-1}}{2}\|\mathbf{v}_k-\mathbf{v}_{k-1}\|_2^2\\&\geq A_k\big(f(\mathbf{y}_k)-f(\mathbf{x}_k)\big)+A_{k-1}\big\langle\nabla f(\mathbf{x}_k),\mathbf{x}_k-\mathbf{y}_{k-1}\big\rangle.\end{aligned}$$

Substituting the above inequality into (92),

$$\begin{aligned}
a_k P(\mathbf{x}^\star) \geq\ & \mathbb{E}\left[a_k f(\mathbf{x}_k) + a_k\big\langle \nabla f(\mathbf{x}_k), \mathbf{v}_k - \mathbf{x}_k\big\rangle + a_k h(\mathbf{v}_k) + \frac{\mu a_k}{2}\|\mathbf{v}_k - \mathbf{x}_k\|_2^2 + \frac{\mu A_{k-1}}{2}\|\mathbf{v}_k - \mathbf{v}_{k-1}\|_2^2\right] \\
& + \mathbb{E}\left[\frac{\mu A_k}{2}\|\mathbf{x}^\star - \mathbf{v}_k\|_2^2 - \frac{\mu A_{k-1}}{2}\|\mathbf{x}^\star - \mathbf{v}_{k-1}\|_2^2 - a_k\big(\big\langle \mathbf{w}_k + \boldsymbol{\eta}_k, \mathbf{x}^\star - \mathbf{v}_k\big\rangle + \varepsilon_k\big)\right] \\
\geq\ & \mathbb{E}\left[a_k f(\mathbf{x}_k) + A_k f(\mathbf{y}_k) - A_k f(\mathbf{x}_k) + A_{k-1}\big\langle \nabla f(\mathbf{x}_k), \mathbf{x}_k - \mathbf{y}_{k-1}\big\rangle + a_k h(\mathbf{v}_k)\right] \\
& + \mathbb{E}\left[\frac{\mu A_k}{2}\|\mathbf{x}^\star - \mathbf{v}_k\|_2^2 - \frac{\mu A_{k-1}}{2}\|\mathbf{x}^\star - \mathbf{v}_{k-1}\|_2^2 - a_k\big(\big\langle \mathbf{w}_k + \boldsymbol{\eta}_k, \mathbf{x}^\star - \mathbf{v}_k\big\rangle + \varepsilon_k\big)\right] \\
=\ & \mathbb{E}\left[A_k f(\mathbf{y}_k) - A_{k-1}\Big(f(\mathbf{x}_k) + \big\langle \nabla f(\mathbf{x}_k), \mathbf{y}_{k-1} - \mathbf{x}_k\big\rangle\Big) + a_k h(\mathbf{v}_k)\right] \\
& + \mathbb{E}\left[\frac{\mu A_k}{2}\|\mathbf{x}^\star - \mathbf{v}_k\|_2^2 - \frac{\mu A_{k-1}}{2}\|\mathbf{x}^\star - \mathbf{v}_{k-1}\|_2^2 - a_k\big(\big\langle \mathbf{w}_k + \boldsymbol{\eta}_k, \mathbf{x}^\star - \mathbf{v}_k\big\rangle + \varepsilon_k\big)\right].
\end{aligned}$$

Applying the convexity of $f(\mathbf{x})$, we obtain

$$\begin{aligned}
a_k P(\mathbf{x}^\star) \geq\ & \mathbb{E}\left[A_k f(\mathbf{y}_k) - A_{k-1} f(\mathbf{y}_{k-1}) + a_k h(\mathbf{v}_k) + \frac{\mu A_k}{2}\|\mathbf{x}^\star - \mathbf{v}_k\|_2^2 - \frac{\mu A_{k-1}}{2}\|\mathbf{x}^\star - \mathbf{v}_{k-1}\|_2^2\right] \\
& - \mathbb{E}\left[a_k\left(\big\langle \mathbf{w}_k + \boldsymbol{\eta}_k, \mathbf{x}^\star - \mathbf{v}_k\big\rangle + \varepsilon_k\right)\right]. \qquad (96)
\end{aligned}$$

Substituting (96) into (91),

$$\mathbb{E}[E_k] \leq \mathbb{E}\left[E_k^\eta + E_k^\varepsilon + \frac{\mu A_{k-1}}{2}\|\mathbf{x}^\star - \mathbf{v}_{k-1}\|_2^2 - \frac{\mu A_k}{2}\|\mathbf{x}^\star - \mathbf{v}_k\|_2^2\right].$$

This completes the proof. ■

B.2 Proof of Theorem 5

Proof Applying Lemma 5, we obtain

$$\sum_{i=1}^k \mathbb{E}[E_i] \leq \sum_{i=1}^k \left[E_k^\eta + E_k^\varepsilon\right] + \frac{\mu A_0}{2}\|\mathbf{x}^\star - \mathbf{v}_0\|^2 - \frac{\mu A_k}{2}\mathbb{E}\left[\|\mathbf{x}^\star - \mathbf{v}_k\|^2\right].$$

Substituting this upper bound into $A_k G_k - A_0 G_0 = \sum_{i=1}^k E_i$, we obtain

$$A_k\mathbb{E}\left[G_k + \frac{\mu}{2}\|\mathbf{x}^\star - \mathbf{v}_k\|^2\right] \leq A_0\mathbb{E}\left[G_0 + \frac{\mu}{2}\|\mathbf{x}^\star - \mathbf{v}_0\|^2\right] + \sum_{i=1}^k a_i\mathbb{E}\left[\big\langle \boldsymbol{\eta}_i + \mathbf{w}_i, \mathbf{x}^\star - \mathbf{v}_i\big\rangle + \varepsilon_i\right].$$

Using the definition of Δ_k, it becomes

$$\mathbb{E}\left[\Delta_k\right] \leq \frac{A_0}{A_k}\Delta_0 + \sum_{i=1}^k \frac{a_i}{A_k}\mathbb{E}\left[\big\langle \boldsymbol{\eta}_i + \mathbf{w}_i, \mathbf{x}^\star - \mathbf{v}_i\big\rangle + \varepsilon_i\right]. \qquad (97)$$

Substituting θ_i, $\frac{A_0}{A_k}$ can be written as

$$\frac{A_0}{A_k} = \frac{A_{k-1}}{A_k} \times \cdots \times \frac{A_1}{A_2} \times \frac{A_0}{A_1} = \frac{A_k - a_k}{A_k} \times \cdots \times \frac{A_2 - a_2}{A_2} \times \frac{A_1 - a_1}{A_1} = \prod_{i=1}^{k} (1-\theta_i) = \Theta_k. \quad (98)$$

Similarly, $\frac{a_i}{A_k}$ can be written as

$$\frac{a_i}{A_k} = \frac{A_{k-1}}{A_k} \times \cdots \times \frac{A_i}{A_{i+1}} \times \frac{a_i}{A_i} = \theta_i \prod_{j=i+1}^{k} (1-\theta_j) = \frac{\Theta_k}{\Theta_i}\theta_i. \quad (99)$$

Substituting (98) and (99) into (97), we obtain

$$\mathbb{E}\left[\Delta_k\right] \leq \Theta_k \Delta_0 + \sum_{i=1}^{k} \frac{\Theta_k}{\Theta_i} \theta_i \mathbb{E}\left[\langle \boldsymbol{\eta}_i + \mathbf{w}_i, \mathbf{x}^\star - \mathbf{v}_i \rangle + \varepsilon_i\right].$$

It can be rewritten as

$$\mathbb{E}\left[\Delta_k\right] \leq \Theta_k \mathbb{E}\left[\Delta_0 + \sum_{i=1}^{k} \frac{\theta_i}{\Theta_i} \left(\langle \mathbf{w}_i + \boldsymbol{\eta}_i, \mathbf{x}^\star - \mathbf{v}_i \rangle + \varepsilon_i\right)\right].$$

This completes the proof. ■

B.3 Proof of Theorem 6

Proof We first introduce some notations. We define $\psi_i(\mathbf{v}) \overset{\text{def}}{=} \frac{\mu A_i}{2}\|\mathbf{v}\|_2^2 + a_i h(\mathbf{v})$ and $\mathbf{z}_i \overset{\text{def}}{=} \mu(a_i \mathbf{x}_i + A_{i-1}\mathbf{v}_{i-1}) - a_i \widetilde{\nabla} f(\mathbf{x}_i)$. Same as Proposition 1, it holds that $\mathbf{v}_i^\star = \nabla \psi_k^*(\mathbf{z}_i)$. Applying $\theta_i = \beta, \forall i \geq 1$, (57) becomes

$$\begin{aligned} &\mathbb{E}[G_k] + \frac{\mu}{2}\mathbb{E}\big[\|\mathbf{x}^\star - \mathbf{v}_k\|_2^2\big] \\ &\leq \big(1-\beta\big)^k \left(\Delta_0 + \beta \sum_{i=1}^{k} \big(1-\beta\big)^{-i} \mathbb{E}\big[\langle \boldsymbol{\eta}_i, \mathbf{x}^\star - \mathbf{v}_i \rangle\big] + \beta \sum_{i=1}^{k} \big(1-\beta\big)^{-i} \big(\mathbb{E}\big[\langle \mathbf{w}_i, \mathbf{x}^\star - \mathbf{v}_i \rangle\big] + \varepsilon_i\big)\right). \end{aligned} \quad (100)$$

Bounding $\mathbb{E}[\langle \boldsymbol{\eta}_i, \mathbf{x}^\star - \mathbf{v}_i \rangle]$: We define $\widehat{\mathbf{v}}_i^\star \overset{\text{def}}{=} \nabla \psi_i^*(\mu(a_i \mathbf{x}_i + A_{i-1}\mathbf{v}_{i-1}) - a_i \nabla f(\mathbf{x}_i)) = \nabla \psi_i^*(\mathbf{z}_i + a_i \boldsymbol{\eta}_i)$ that is optimal solution to (54) when both gradient oracle and proximal mapping are exact. Then, $\mathbb{E}\big[\langle \boldsymbol{\eta}_i, \mathbf{x}^\star - \mathbf{v}_i \rangle\big]$ can be written as

$$\mathbb{E}\big[\langle \boldsymbol{\eta}_i, \mathbf{x}^\star - \mathbf{v}_i \rangle\big] = \mathbb{E}\big[\langle \boldsymbol{\eta}_i, \mathbf{x}^\star - \widehat{\mathbf{v}}_i^\star \rangle\big] + \mathbb{E}\big[\langle \boldsymbol{\eta}_i, \widehat{\mathbf{v}}_i^\star - \mathbf{v}_i^\star \rangle\big] + \mathbb{E}\big[\langle \boldsymbol{\eta}_i, \mathbf{v}_i^\star - \mathbf{v}_i \rangle\big]. \quad (101)$$

Since $\widehat{\mathbf{v}}_i^\star$ is independent of $\boldsymbol{\eta}_i$ and $\mathbb{E}[\boldsymbol{\eta}_i] = \mathbf{0}$, we obtain

$$\mathbb{E}\big[\langle \boldsymbol{\eta}_i, \mathbf{x}^\star - \widehat{\mathbf{v}}_i^\star \rangle\big] = 0. \quad (102)$$

It is easy to see that $\psi_i(\mathbf{v})$ is (μA_i)-strongly convex. Applying Lemma 1, we obtain

$$\mathbb{E}\big[\langle \boldsymbol{\eta}_i, \widehat{\mathbf{v}}_i^\star - \mathbf{v}_i^\star\rangle\big] \le \mathbb{E}\big[\|\boldsymbol{\eta}_i\|_2\|\widehat{\mathbf{v}}_i^\star - \mathbf{v}_i^\star\|_2\big] \le \mathbb{E}\Big[\|\boldsymbol{\eta}_i\|_2\tfrac{1}{\mu A_i}\|\mathbf{z}_i + a_i\boldsymbol{\eta}_i - \mathbf{z}_i\|_2\Big] \le \tfrac{a_i}{\mu A_i}\mathbb{E}\big[\|\boldsymbol{\eta}_i\|_2^2\big]. \tag{103}$$

Substituting $a_i/A_i = \beta$ into (103), we come up with

$$\mathbb{E}\big[\langle \boldsymbol{\eta}_i, \widehat{\mathbf{v}}_i^\star - \mathbf{v}_i^\star\rangle\big] \le \frac{\beta}{\mu}\mathbb{E}\big[\|\boldsymbol{\eta}_i\|_2^2\big]. \tag{104}$$

Regarding $\mathbb{E}\big[\langle \boldsymbol{\eta}_i, \mathbf{v}_i^\star - \mathbf{v}_i\rangle\big]$, we have $\mathbb{E}\big[\langle \boldsymbol{\eta}_i, \mathbf{v}_i^\star - \mathbf{v}_i\rangle\big] = 0$ if the proximal mapping is exactly solved as $\mathbf{v}_i^\star = \mathbf{v}_i$. Otherwise, the strong convexity of $\Psi_i(\cdot)$ and definition of $\mathbf{v}_i$ lead to

$$\varepsilon_i \ge \mathbb{E}\big[\Psi_i(\mathbf{v}_i) - \Psi_i(\mathbf{v}_i^\star)\big] \ge \tfrac{\mu A_i}{2a_i}\mathbb{E}\big[\|\mathbf{v}_i - \mathbf{v}_i^\star\|_2^2\big] \Rightarrow \mathbb{E}\big[\|\mathbf{v}_i - \mathbf{v}_i^\star\|_2\big] \le \sqrt{\frac{2a_i\varepsilon_i}{\mu A_i}}.$$

Thus, $\mathbb{E}[\langle \boldsymbol{\eta}_i, \mathbf{v}_i^\star - \mathbf{v}_i\rangle]$ can be bounded as

$$\mathbb{E}\big[\langle \boldsymbol{\eta}_i, \mathbf{v}_i^\star - \mathbf{v}_i\rangle\big] \le \mathbb{E}\big[\|\boldsymbol{\eta}_i\|_2\|\mathbf{v}_i^\star - \mathbf{v}_i\|_2\big] \le \sqrt{\frac{2a_i\varepsilon_i}{\mu A_i}}\mathbb{E}\big[\|\boldsymbol{\eta}_i\|_2\big] \le \sqrt{\frac{2\beta\varepsilon_i}{\mu}}\mathbb{E}\big[\|\boldsymbol{\eta}_i\|_2\big]. \tag{105}$$

Substituting (102), (104) and (105) into (101), we obtain

$$\mathbb{E}\big[\langle \boldsymbol{\eta}_i, \mathbf{x}^\star - \mathbf{v}_i\rangle\big] \le \frac{\beta}{\mu}\mathbb{E}\big[\|\boldsymbol{\eta}_i\|_2^2\big] + \sqrt{\frac{2\beta\varepsilon_i}{\mu}}\mathbb{E}\big[\|\boldsymbol{\eta}_i\|_2\big]. \tag{106}$$

Bounding $\mathbb{E}[\langle \mathbf{w}_i, \mathbf{x}^\star - \mathbf{v}_i\rangle]$: Applying $\mathbb{E}\left[\|\mathbf{w}_i\|_2^2\right] \le 2\mu A_i\varepsilon_i/a_i$, we come up with

$$\mathbb{E}\big[\langle \mathbf{w}_i, \mathbf{x}^\star - \mathbf{v}_i\rangle\big] \le \mathbb{E}\big[\|\mathbf{w}_i\|_2\|\mathbf{x}^\star - \mathbf{v}_i\|_2\big] \le \sqrt{\frac{2\mu\varepsilon_i}{\beta}}\mathbb{E}\big[\|\mathbf{x}^\star - \mathbf{v}_i\|_2\big]. \tag{107}$$

We are now ready to prove (58). Substituting $\beta = \sqrt{\frac{\mu}{L}}$, (106) and (107) into (100),

$$\begin{aligned}
&\mathbb{E}[G_k] + \frac{\mu}{2}\mathbb{E}\big[\|\mathbf{x}^\star - \mathbf{v}_k\|_2^2\big] \\
&\le \big(1-\beta\big)^k\Bigg(\Delta_0 + \sum_{i=1}^k \big(1-\beta\big)^{-i}\left(\frac{\beta^2}{\mu}\mathbb{E}\big[\|\boldsymbol{\eta}_i\|_2^2\big] + \sqrt{\frac{2\beta^3\varepsilon_i}{\mu}}\mathbb{E}\big[\|\boldsymbol{\eta}_i\|_2\big] + \beta\varepsilon_i\right) \\
&\qquad\qquad + \sum_{i=1}^k \big(1-\beta\big)^{-i}\sqrt{2\mu\beta\varepsilon_i}\mathbb{E}\big[\|\mathbf{x}^\star - \mathbf{v}_i\|_2\big]\Bigg).
\end{aligned} \tag{108}$$

Note that

$$\sqrt{\frac{2\beta^3\varepsilon_i}{\mu}}\mathbb{E}\big[\|\boldsymbol{\eta}_i\|_2\big] = \sqrt{\frac{2\beta^2}{\mu}}\mathbb{E}\big[\|\boldsymbol{\eta}_i\|_2\big]\sqrt{\beta\varepsilon_i} \le \frac{\beta^2}{\mu}\mathbb{E}\big[\|\boldsymbol{\eta}_i\|_2^2\big] + \frac{1}{2}\beta\varepsilon_i,$$

where the last inequality follows from $ab \le \frac{1}{2}(a^2+b^2)$ and $\mathbb{E}[X^2] \le (\mathbb{E}[X])^2$. Substituting this result to (108), we obtain

$$\mathbb{E}[G_k] + \frac{\mu}{2}\mathbb{E}\big[\|\mathbf{x}^\star - \mathbf{v}_k\|_2^2\big]$$

$$\leq \big(1-\beta\big)^k \left(\Delta_0 + \sum_{i=1}^k \big(1-\beta\big)^{-i}\left(\frac{2\beta^2}{\mu}\mathbb{E}\big[\|\boldsymbol{\eta}_i\|_2^2\big] + \frac{3}{2}\beta\varepsilon_i\right) + \sum_{i=1}^k \big(1-\beta\big)^{-i}\sqrt{2\mu\beta\varepsilon_i}\mathbb{E}\big[\|\mathbf{x}^\star - \mathbf{v}_i\|_2\big]\right). \tag{109}$$

Since $\mathbb{E}[G_k] \geq 0$, it implies

$$\frac{\mu}{2}\mathbb{E}\big[\|\mathbf{x}^\star - \mathbf{v}_k\|_2^2\big]$$
$$\leq \big(1-\beta\big)^k \left(\Delta_0 + \sum_{i=1}^k \big(1-\beta\big)^{-i}\left(\frac{2\beta^2}{\mu}\mathbb{E}\big[\|\boldsymbol{\eta}_i\|_2^2\big] + \frac{3}{2}\beta\varepsilon_i\right) + \sum_{i=1}^k \big(1-\beta\big)^{-i}\sqrt{2\mu\beta\varepsilon_i}\mathbb{E}\big[\|\mathbf{x}^\star - \mathbf{v}_i\|_2\big]\right).$$

Diving both sides by $\big(1-\beta\big)^k$, it becomes

$$\frac{\mu}{2}\big(1-\beta\big)^{-k}\mathbb{E}\big[\|\mathbf{x}^\star - \mathbf{v}_k\|_2^2\big]$$
$$\leq \Delta_0 + \sum_{i=1}^k \big(1-\beta\big)^{-i}\left(\frac{2\beta^2}{\mu}\mathbb{E}\big[\|\boldsymbol{\eta}_i\|_2^2\big] + \frac{3}{2}\beta\varepsilon_i\right) + \sum_{i=1}^k \big(1-\beta\big)^{-i}\sqrt{2\mu\beta\varepsilon_i}\mathbb{E}\big[\|\mathbf{x}^\star - \mathbf{v}_i\|_2\big].$$

Applying Lemma 6 with $S_k \overset{\text{def}}{=} \Delta_0 + \sum_{i=1}^k \big(1-\beta\big)^{-i}\left(\frac{2\beta^2}{\mu}\mathbb{E}\big[\|\boldsymbol{\eta}_i\|_2^2\big] + \frac{3}{2}\beta\varepsilon_i\right), \vartheta_i \overset{\text{def}}{=} 2\big(1-\beta\big)^{-i/2}\sqrt{\beta\varepsilon_i}$ and $u_i \overset{\text{def}}{=} \sqrt{\frac{\mu}{2}}\big(1-\beta\big)^{-i/2}\mathbb{E}\big[\|\mathbf{x}^\star - \mathbf{v}_i\|_2\big]$, we obtain

$$\Delta_0 + \sum_{i=1}^k \big(1-\beta\big)^{-i}\left(\frac{2\beta^2}{\mu}\mathbb{E}\big[\|\boldsymbol{\eta}_i\|_2^2\big] + \frac{3}{2}\beta\varepsilon_i\right) + \sum_{i=1}^k \big(1-\beta\big)^{-i}\sqrt{2\mu\beta\varepsilon_i}\mathbb{E}\big[\|\mathbf{x}^\star - \mathbf{v}_i\|_2\big]$$
$$\leq \frac{3}{2}\left(\Delta_0 + \sum_{i=1}^k \big(1-\beta\big)^{-i}\left(\frac{2\beta^2}{\mu}\mathbb{E}\big[\|\boldsymbol{\eta}_i\|_2^2\big] + \frac{3}{2}\beta\varepsilon_i\right) + \left(\sum_{i=1}^k 2\big(1-\beta\big)^{-i/2}\sqrt{\beta\varepsilon_i}\right)^2\right)$$
$$\leq \frac{3}{2}\Delta_0 + \frac{3\beta^2}{\mu}\sum_{i=1}^k \big(1-\beta\big)^{-i}\mathbb{E}\big[\|\boldsymbol{\eta}_i\|_2^2\big] + \frac{25\beta}{4}\left(\sum_{i=1}^k \big(1-\beta\big)^{-i/2}\sqrt{\varepsilon_i}\right)^2.$$

Substituting it into (109),

$$\mathbb{E}[G_k] \leq \big(1-\beta\big)^k \left(\frac{3}{2}\Delta_0 + \frac{3\beta^2}{\mu}\sum_{i=1}^k \big(1-\beta\big)^{-i}\mathbb{E}\big[\|\boldsymbol{\eta}_i\|_2^2\big] + \frac{25\beta}{4}\left(\sum_{i=1}^k \big(1-\beta\big)^{-i/2}\sqrt{\varepsilon_i}\right)^2\right).$$

This completes the proof of (58). ■

B.4 Proof of Corollary 4

Proof It is straightforward to prove it by applying Theorem 6. Substituting $\mathbb{E}\big[\|\boldsymbol{\eta}_i\|_2^2\big] \leq \sigma^2, \varepsilon_i \leq \varepsilon, \forall i \geq 1$ into (58), we obtain

$$\mathbb{E}[G_k] \leq \frac{3}{2}\big(1-\beta\big)^k\Delta_0 + \frac{3\beta^2\sigma^2}{\mu}\sum_{i=1}^k \big(1-\beta\big)^{k-i} + \frac{25\beta\varepsilon}{4}\left(\sum_{i=1}^k \big(1-\beta\big)^{(k-i)/2}\right)^2. \tag{110}$$

Substituting $\beta = \sqrt{\frac{\mu}{L}}$ into (110), we come up with

$$\mathbb{E}[G_k] \leq \frac{3}{2}\left(1 - \sqrt{\frac{\mu}{L}}\right)^k \Delta_0 + \frac{3\sigma^2}{\sqrt{\mu L}} + 25\sqrt{\frac{L}{\mu}}\varepsilon,$$

where the inequality follows from $\sqrt{1-x} \leq 1 - \frac{x}{2}$. This completes the proof of (59). Substituting $\beta = \frac{\mu}{L}$ into (110), we come up with

$$\mathbb{E}[G_k] \leq \frac{3}{2}\left(1 - \frac{\mu}{L}\right)^k \Delta_0 + \frac{3\sigma^2}{L} + \frac{25L}{\mu}\varepsilon.$$

This completes the proof of (60). ■

B.5 Proof for Theorem 7

Proof Plugging $\varepsilon_i = 0$ and $\mathbf{w}_i = \mathbf{0}$ into (57), it becomes

$$\mathbb{E}\left[\Delta_k\right] \leq \Theta_k \mathbb{E}\left[\Delta_0 + \sum_{i=1}^{k} \tfrac{\theta_i}{\Theta_i}\langle \boldsymbol{\eta}_i, \mathbf{x}^\star - \mathbf{v}_i\rangle\right], \forall k \geq 1. \tag{111}$$

Then, we follow the proof of Theorem 6 to bound $\mathbb{E}\big[\langle \boldsymbol{\eta}_i, \mathbf{x}^\star - \mathbf{v}_i\rangle\big]$. In our case, we have $\mathbb{E}\big[\langle \boldsymbol{\eta}_i, \mathbf{v}_i^\star - \mathbf{v}_i\rangle\big] = 0$ as the proximal mapping is exactly solved as $\mathbf{v}_i^\star = \mathbf{v}_i$. Substituting $\mathbb{E}\big[\langle \boldsymbol{\eta}_i, \mathbf{v}_i^\star - \mathbf{v}_i\rangle\big] = 0$, (102) and (103) into (101), we obtain

$$\mathbb{E}\big[\langle \boldsymbol{\eta}_i, \mathbf{x}^\star - \mathbf{v}_i\rangle\big] \leq \frac{\theta_i}{\mu}\mathbb{E}\big[\|\boldsymbol{\eta}_i\|_2^2\big]. \tag{112}$$

Plugging (112) into (111), we obtain

$$\mathbb{E}\left[\Delta_k\right] \leq \Theta_k \left(\Delta_0 + \frac{1}{\mu}\sum_{i=1}^{k} \tfrac{\theta_i^2}{\Theta_i}\mathbb{E}\big[\|\boldsymbol{\eta}_i\|_2^2\big]\right), \forall k \geq 1.$$

This completes the proof of (63). Then, it is straightforward to prove (64) and (65) by replacing $\theta_i = \sqrt{\mu/L}$ and $\theta_i = \mu/L$, respectively. ■

B.6 Proof of Corollary 5

Proof We first prove the case of accelerated proximal method. Given the linear convergence rate (64), the number of iterations of the first stage strategy is as following

$$\left(1 - \sqrt{\frac{\mu}{L}}\right)^{\widehat{k}} \Delta_0 \leq \frac{\sigma^2}{\sqrt{\mu L}} \Rightarrow \widehat{k} \geq \sqrt{\frac{L}{\mu}}\log\left(\frac{2\Delta_0}{\epsilon}\right).$$

For the second stage, we apply Theorem 7 for $k \geq \widetilde{k} = \left\lceil 2\sqrt{\frac{L}{\mu}} - 2 \right\rceil$,

$$\begin{aligned}
\mathbb{E}\left[\Delta_k\right] &\leq \Theta_k \left(\Delta_0 + \frac{\sigma^2}{\mu} \sum_{i=1}^{k} \frac{\theta_i^2}{\Theta_i} \right) \\
&= \Theta_k \left(\mathbb{E}\left[P(\widehat{\mathbf{y}}_{\widehat{k}}) - P(\mathbf{x}^\star) + \frac{\mu}{2} \|\widehat{\mathbf{y}}_{\widehat{k}} - \mathbf{x}^\star\|_2^2 \right] + \frac{\sigma^2}{\mu} \sum_{i=1}^{\widetilde{k}-1} \frac{\theta_i^2}{\Theta_i} \right) + \Theta_k \frac{\sigma^2}{\mu} \sum_{i=\widetilde{k}}^{k} \frac{\theta_i^2}{\Theta_i} \\
&\leq \Theta_k \left(\frac{2\sigma^2}{\sqrt{\mu L}} + \frac{\sigma^2}{\sqrt{\mu L}} \sum_{i=1}^{\widetilde{k}-1} \frac{\theta_i}{\Theta_i} \right) + \Theta_k \frac{\sigma^2}{\mu} \sum_{i=\widetilde{k}}^{k} \frac{\theta_i^2}{\Theta_i} \\
&= \frac{\widetilde{k}(\widetilde{k}+1)}{(k+1)(k+2)} \left(\Theta_{\widetilde{k}-1} \frac{2\sigma^2}{\sqrt{\mu L}} + \frac{\sigma^2}{\sqrt{\mu L}} \Theta_{\widetilde{k}-1} \sum_{i=1}^{\widetilde{k}-1} \frac{\theta_i}{\Theta_i} \right) + \frac{4\sigma^2}{\mu(k+1)(k+2)} \sum_{i=\widetilde{k}}^{k} \frac{i+1}{i+2} \\
&= \frac{\widetilde{k}(\widetilde{k}+1)}{(k+1)(k+2)} \left(\Theta_{\widetilde{k}-1} \frac{2\sigma^2}{\sqrt{\mu L}} + \frac{\sigma^2}{\sqrt{\mu L}} \left(1 - \Theta_{\widetilde{k}-1}\right) \right) + \frac{4\sigma^2}{\mu(k+1)(k+2)} \sum_{i=\widetilde{k}}^{k} \frac{i+1}{i+2} \\
&\leq \frac{\widetilde{k}(\widetilde{k}+1)}{(k+1)(k+2)} \frac{2\sigma^2}{\sqrt{\mu L}} + \frac{4\sigma^2}{\mu(k+1)(k+2)} \sum_{i=\widetilde{k}}^{k} \frac{i+1}{i+2} \\
&\leq \frac{\widetilde{k}}{(k+1)(k+2)} \frac{4\sigma^2}{\mu} + \frac{4\sigma^2}{\mu(k+2)} \\
&\leq \frac{8\sigma^2}{\mu(k+2)},
\end{aligned}$$

where the second and third equations follow from (Kulunchakov and Mairal, 2020, Lemma 26) and (Kulunchakov and Mairal, 2020, Lemma 27), respectively. The last two inequalities are obtained by applying the definition of $\widetilde{k}$. Thus, the complexity for the second stage is $O(\sigma^2/(\mu\epsilon))$. Combining two stages together, we obtain the complexity shown in (66).
For the case of non-accelerated proximal method, it can be easily proved by following the above proof. ■

B.7 Proof of Corollary 6

Proof It is straightforward to prove it by applying Theorem 6. Substituting $\mathbb{E}\left[\|\boldsymbol{\eta}_i\|_2^2\right] \leq \sigma^2, \varepsilon_i \leq \varepsilon, \forall i \geq 1$ into (58), we obtain

$$\mathbb{E}[G_k] \leq \frac{3}{2}\beta^k \Delta_0 + \frac{3\sigma^2}{L} \sum_{i=1}^{k} \beta^{k-i} + \frac{25}{4} \sqrt{\frac{\mu}{L}} \varepsilon \left(\sum_{i=1}^{k} \beta^{(k-i)/2} \right)^2 .$$

Applying $\beta = 1 - \sqrt{\frac{\mu}{L}}$, we come up with

$$\mathbb{E}[G_k] \leq \frac{3}{2}\left(1 - \sqrt{\tfrac{\mu}{L}}\right)^k \Delta_0 + \frac{3\sigma^2}{\sqrt{\mu L}} + 25\sqrt{\frac{L}{\mu}}\varepsilon,$$

where the inequality follows from $\sqrt{1-x} \leq 1 - \frac{x}{2}$. This completes the proof. ■

B.8 Proof of Corollary 7

Proof Substituting the results of (98) and (99) into (74), we obtain

$$\mathbb{E}\left[G_k\right] \leq \Theta_k \left(\Delta_0 + \sqrt{\frac{1}{\mu L}}\sigma^2 \sum_{i=1}^{k} \frac{\theta_i}{\Theta_i}\right).$$

Substituting $\theta_i = \sqrt{\mu/L}$ and $\Theta_i = (1 - \sqrt{\mu/L})^i$, it becomes

$$\mathbb{E}\left[G_k\right] \leq \left(1 - \sqrt{\frac{\mu}{L}}\right)^k \left(P(\mathbf{x}_0) - P(\mathbf{x}^\star) + \frac{\mu}{2}\|\mathbf{x}_0 - \mathbf{x}^\star\|_2^2\right) + \frac{\sigma^2}{\sqrt{\mu L}}.$$

This completes the proof of (78). It is straightforward to prove (79) by following the above proof with $\theta_i = \mu/L$. ■

References

Zeyuan Allen-Zhu. Katyusha X: practical momentum method for stochastic sum-of-nonconvex optimization. In *Proceedings of International Conference on Machine Learning (ICML)*, pages 179–185, 2018.

Zeyuan Allen-Zhu and Lorenzo Orecchia. Linear coupling: An ultimate unification of gradient and mirror descent. In *Proceedings of ACM Conference on Innovations in Theoretical Computer Science Conference (ITCS)*, pages 3:1–3:22, 2017.

Mahmoud Assran and Mike Rabbat. On the convergence of nesterov's accelerated gradient method in stochastic settings. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 119, pages 410–420, 2020.

Yves F. Atchadé, Gersende Fort, and Eric Moulines. On perturbed proximal gradient algorithms. *Journal of Machine Learning Research (JMLR)*, 18:10:1–10:33, 2017.

Necdet Serhat Aybat, Alireza Fallah, Mert Gürbüzbalaban, and Asuman E. Ozdaglar. A universally optimal multistage accelerated stochastic gradient method. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 8523–8534, 2019.

Francis R. Bach, Rodolphe Jenatton, Julien Mairal, and Guillaume Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends in Machine Learning*, 4(1):1–106, 2012.

Raef Bassily, Adam D. Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *Proceedings of the Annual Symposium on Foundations of Computer Science (FOCS)*, pages 464–473, 2014.

Amir Beck and Marc Teboulle. Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. *IEEE Transactions on Image Processing (TIP)*, 18(11):2419–2434, 2009.

Aharon Ben-Tal and Arkadii Nemirovskii. *Lectures on Modern Convex Optimization - Analysis, Algorithms, and Engineering Applications.* MPS-SIAM series on optimization. SIAM, 2001.

Dimitri P. Bertsekas, Angelia Nedi, and Asuman E. Ozdaglar. *Convex Analysis and Optimization.* Athena Scientific, 2003.

Howard Bondell and Brian Reich. Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with oscar. *Biometrics*, 64(1):115–123, 2008.

Lev M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200–217, 1967.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357, 2015.

Sébastien Bubeck, Yin Tat Lee, and Mohit Singh. A geometric alternative to nesterov's accelerated gradient descent. *CoRR*, abs/1506.08187, 2015.

Yair Censor and Stavros Zenios. *Parallel Optimization: Theory, Algorithms, and Applications.* Oxford University Press, 1998.

Michael Cohen, Jelena Diakonikolas, and Lorenzo Orecchia. On acceleration with noise-corrupted gradients. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1018–1027, 2018.

Alexandre d'Aspremont. Smooth optimization with approximate gradient. *SIAM Journal on Optimization (STOPT)*, 19(3):1171–1183, 2008.

Damek Davis, Dmitriy Drusvyatskiy, and Kellie J. MacPhee. Stochastic model-based minimization under high-order growth. *CoRR*, abs/1807.00255, 2018. URL `http://arxiv.org/abs/1807.00255`.

Olivier Devolder, François Glineur, and Yurii Nesterov. First-order methods of smooth convex optimization with inexact oracle. *Mathematical Programming*, 146(1-2):37–75, 2014.

Jelena Diakonikolas and Lorenzo Orecchia. Accelerated extra-gradient descent: A novel accelerated first-order method. In *Proceedings of the Innovations in Theoretical Computer Science Conference (ITCS)*, volume 94, pages 23:1–23:19, 2018.

Jelena Diakonikolas and Lorenzo Orecchia. The approximate duality gap technique: A unified theory of first-order methods. *SIAM Journal on Optimization (SIOPT)*, 29(1): 660–689, 2019.

Pavel Dvurechensky and Alexander Gasnikov. Stochastic intermediate gradient method for convex problems with stochastic inexact oracle. *J. Optimization Theory and Applications*, 171(1):121–145, 2016.

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

Saeed Ghadimi and Guanghui Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, II: Shrinking procedures and optimal algorithms. *SIAM Journal on Optimization (STOPT)*, 23(4):2061–2089, 2013.

Moritz Hardt. Robustness versus acceleration. `http://blog.mrtz.org/2014/08/18/robustness-versus-acceleration.html`, 2014.

Trevor Hastie, Robert Tibshirani, and Jerome H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, 2009.

Junzhou Huang, Tong Zhang, and Dimitris N. Metaxas. Learning with structured sparsity. *Journal of Machine Learning Research (JMLR)*, 12:3371–3412, 2011.

Prateek Jain, Sham M. Kakade, Rahul Kidambi, raneeth Netrapalli, and Aaron Sidford. Accelerating stochastic gradient descent for least squares regression. In *Proceedings of the Annual Conference On Learning Theory (COLT)*, pages 545–604, 2018.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems (NIPS)*, pages 315–323, 2013.

Dmitry Kamzolov, Pavel Dvurechensky, and Alexander Gasnikov. Universal intermediate gradient method for convex problems with inexact oracle. *Optimization Methods and Software*, (2):1–28, 2020a.

Dmitry Kamzolov, Pavel Dvurechensky, and Alexander Gasnikov. Robustness of accelerated first-order algorithms for strongly convex optimization problems. *IEEE Transactions on Automatic Control*, 2020b.

Walid Krichene and Peter L. Bartlett. Acceleration and averaging in stochastic descent dynamics. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 6796–6806, 2017.

Andrei Kulunchakov and Julien Mairal. A generic acceleration framework for stochastic composite optimization. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 12556–12567, 2019a.

Andrei Kulunchakov and Julien Mairal. Estimate sequences for variance-reduced stochastic composite optimization. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 3541–3550, 2019b.

Andrei Kulunchakov and Julien Mairal. Estimate sequences for stochastic composite optimization: Variance reduction, acceleration, and robustness to noise. *Journal of Machine Learning Research (JMLR)*, 21:155:1–155:52, 2020.

Guanghui Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1-2):365–397, 2012.

Guanghui Lan and Yi Zhou. Conditional gradient sliding for convex optimization. *SIAM Journal on Optimization (SIOPT)*, 26(2):1379–1409, 2016.

Hongzhou Lin, Julien Mairal, and Zaïd Harchaoui. A universal catalyst for first-order optimization. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 3384–3392, 2015.

Hongzhou Lin, Julien Mairal, and Zaïd Harchaoui. Catalyst acceleration for first-order convex optimization: from theory to practice. *Journal of Machine Learning Research (JMLR)*, 18:212:1–212:54, 2017.

Yanli Liu, Fei Feng, and Wotao Yin. Acceleration of SVRG and katyusha X by inexact preconditioning. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 4003–4012, 2019.

Julien Mairal, Rodolphe Jenatton, Guillaume Obozinski, and Francis R. Bach. Convex and network flow optimization for structured sparsity. *Journal of Machine Learning Research (JMLR)*, 12:2681–2720, 2011.

Arkadi Nemirovski. Prox-method with rate of convergence o(1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization (SIOPT)*, 15(1):229–251, 2004.

Arkadi Nemirovski, Anatoli B. Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization (SIOPT)*, 19(4):1574–1609, 2009.

Arkadii Nemirovsky and David Yudin. *Problem Complexity and Method Efficiency in Optimization.* Wiley Interscience Series in Discrete Mathematics, 1983.

Yurii Nesterov. A method for solving the convex programming problem with convergence rate $O(1/\mathrm{k\hat{}2})$. *Doklady AN SSSR (translated as Soviet Mathematics Doklady)*, 269:543–547, 1983.

Yurii Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 140(1):127–152, 2005.

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer Science & Business Media, 2013.

Courtney Paquette, Hongzhou Lin, Dmitriy Drusvyatskiy, Julien Mairal, and Zaïd Harchaoui. Catalyst for gradient-based nonconvex optimization. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 613–622, 2018.

Neal Parikh and Stephen P. Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):127–239, 2014.

Mark Schmidt, Nicolas Le Roux, and Francis R. Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1-2):83–112, 2017.

Mark W. Schmidt, Nicolas Le Roux, and Francis R. Bach. Convergence rates of inexact proximal-gradient methods for convex optimization. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 1458–1466, 2011.

Fedor Stonyakin, Alexander Tyurin, Alexander Gasnikov, Pavel Dvurechensky, Artem Agafonov, Darina Dvinskikh, Dmitry Pasechnyuk, Sergei Artamonov, and Victorya Piskunova. Inexact relative smoothness and strong convexity for optimization and variational inequalities by inexact model. *arXiv preprint arXiv:2001.09013*, 2020.

Mingkui Tan, Ivor W. Tsang, and Li Wang. Matching pursuit LASSO part I: sparse recovery over big dictionary. *IEEE Transactions on Signal Processing (TSP)*, 63(3):727–741, 2015a.

Mingkui Tan, Ivor W. Tsang, and Li Wang. Matching pursuit LASSO part II: applications and sparse recovery over batch signals. *IEEE Transactions on Signal Processing (TSP)*, 63(3):742–753, 2015b.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B*, 58(1):267–288, 1996.

Paul Tseng. On accelerated proximal gradient methods for convex-concave optimization. *Submitted to SIAM Journal on Optimization (SIOPT)*, 2008.

Jialei Wang and Tong Zhang. Utilizing second order information in minibatch stochastic variance reduced proximal iterations. *Journal of Machine Learning Research (JMLR)*, 20:42:1–42:56, 2019.

Xiaoyu Wang, Xiao Wang, and Ya-Xiang Yuan. Stochastic proximal quasi-Newton methods for non-convex composite optimization. *Optimization Methods and Software*, 34(5):922–948, 2019.

Jianqiao Wangni, Jialei Wang, Ji Liu, and Tong Zhang. Gradient sparsification for communication-efficient distributed optimization. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 1306–1316, 2018.

Pan Xu, Tianhao Wang, and Quanquan Gu. Continuous and discrete-time accelerated stochastic mirror descent for strongly convex functions. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 5488–5497, 2018.

Quanming Yao, James T. Kwok, Fei Gao, Wei Chen, and Tie-Yan Liu. Efficient inexact proximal gradient algorithm for nonconvex problems. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3308–3314, 2017.

Haishan Ye, Luo Luo, and Zhihua Zhang. Nesterov's acceleration for approximate newton. *Journal of Machine Learning Research (JMLR)*, 21:142:1–142:37, 2020.

Weizhong Zhang, Lijun Zhang, Yao Hu, Rong Jin, Deng Cai, and Xiaofei He. Sparse learning for stochastic composite optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 893–900, 2014.

Wenliang Zhong and James T. Kwok. Efficient sparse modeling with automatic feature grouping. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 9–16, 2011.

Dongruo Zhou, Yuan Cao, and Quanquan Gu. Accelerated factored gradient descent for low-rank matrix factorization. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 4430–4440, 2020.