Noname manuscript No.
(will be inserted by the editor)

Accelerated Gradient Methods for Nonconvex Nonlinear and Stochastic Programming

Saeed Ghadimi · Guanghui Lan

the date of receipt and acceptance should be inserted later

Abstract In this paper, we generalize the well-known Nesterov's accelerated gradient (AG) method, originally designed for convex smooth optimization, to solve nonconvex and possibly stochastic optimization problems. We demonstrate that by properly specifying the stepsize policy, the AG method exhibits the best known rate of convergence for solving general nonconvex smooth optimization problems by using first-order information, similarly to the gradient descent method. We then consider an important class of composite optimization problems and show that the AG method can solve them uniformly, i.e., by using the same aggressive stepsize policy as in the convex case, even if the problem turns out to be nonconvex. More specifically, the AG method exhibits an optimal rate of convergence if the composite problem is convex, and improves the best known rate of convergence if the problem is nonconvex. Based on the AG method, we also present new nonconvex stochastic approximation methods and show that they can improve a few existing rates of convergence for nonconvex stochastic optimization. To the best of our knowledge, this is the first time that the convergence of the AG method has been established for solving nonconvex nonlinear programming in the literature.

Keywords: nonconvex optimization, stochastic programming, accelerated gradient, complexity
AMS 2000 subject classification: 62L20, 90C25, 90C15, 68Q25,

1 Introduction

In 1983, Nesterov in a celebrated work [23] presented the accelerated gradient (AG) method for solving a class of convex programming (CP) problems given by

$$\Psi^* = \min_{x \in \mathbb{R}^n} \Psi(x). \tag{1.1}$$

Here $\Psi(\cdot)$ is a convex function with Lipschitz continuous gradients, i.e., $\exists\, L_\Psi > 0$ such that (s.t.)

$$\|\nabla\Psi(y) - \nabla\Psi(x)\| \leq L_\Psi \|y - x\| \quad \forall x, y \in \mathbb{R}^n. \tag{1.2}$$

Nesterov shows that the number of iterations performed by this algorithm to find a solution $\bar{x}$ s.t. $\Psi(\bar{x}) - \Psi^* \leq \epsilon$ can be bounded by $\mathcal{O}(1/\sqrt{\epsilon})$, which significantly improves the $\mathcal{O}(1/\epsilon)$ complexity bound possessed by the gradient descent method. Moreover, in view of the classic complexity theory for convex optimization by Nemirovski and Yudin [22], the above $\mathcal{O}(1/\sqrt{\epsilon})$ iteration complexity bound is not improvable for smooth convex optimization when n is sufficiently large.

This research was partially supported by NSF grants CMMI-1000347, CMMI-1254446, DMS-1319050, and ONR grant N00014-13-1-0036.

Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL 32611, (email: sghadimi@ufl.edu).

Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL 32611, (email: glan@ise.ufl.edu).

Address(es) of author(s) should be given

Nesterov's AG method has attracted much interest recently due to the increasing need to solve large-scale CP problems by using fast first-order methods. In particular, Nesterov in an important work [25] shows that by using the AG method and a novel smoothing scheme, one can improve the complexity for solving a broad class of saddle-point problems from $\mathcal{O}(1/\epsilon^2)$ to $\mathcal{O}(1/\epsilon)$. The AG method has also been generalized by Nesterov [26], Beck and Teboulle [3], and Tseng [31] to solve an emerging class of composite CP problems whose objective function is given by the summation of a smooth component and another relatively simple nonsmooth component (e.g., the l_1 norm). Lan [14] further shows that the AG method, when employed with proper stepsize policies, is optimal for solving not only smooth CP problems, but also general (not necessarily simple) nonsmooth and stochastic CP problems. More recently, some key elements of the AG method, e.g., the multi-step acceleration scheme, have been adapted to significantly improve the convergence properties of a few other first-order methods (e.g., level methods [15]). However, to the best of our knowledge, all the aforementioned developments require explicitly the convexity assumption about Ψ. Otherwise, if Ψ in (1.1) is not necessarily convex, it is unclear whether the AG method still converges or not.

This paper aims to generalize the AG method, originally designed for smooth convex optimization, to solve more general nonlinear programming (NLP) (possibly nonconvex and stochastic) problems, and thus to present a unified treatment and analysis for convex, nonconvex and stochastic optimization. While this paper focuses on the theoretical development of the AG method, our study has also been motivated by the following more practical considerations in solving nonlinear programming problems. First, many general nonlinear objective functions are locally convex. A unified treatment for both convex and nonconvex problems will help us to make use of such local convex properties. In particular, we intend to understand whether one can apply the well-known aggressive stepsize policy in the AG method under a more general setting to benefit from such local convexity. Second, many nonlinear objective functions arising from sparse optimization (e.g., [5,8]) and machine learning (e.g., [7,19]) usually consist of both convex and nonconvex components, corresponding to the data fidelity and sparsity regularization terms respectively. One interesting question is whether one can design more efficient algorithms for solving these nonconvex composite problems by utilizing their convexity structure. Third, the convexity of some objective functions represented by a black-box procedure is usually unknown, e.g., in simulation-based optimization [1,9,2,18]. A unified treatment and analysis can thus help us to deal with such structural ambiguity. Fourth, in some cases, the objective functions are nonconvex with respect to (w.r.t.) a few decision variables jointly, but convex w.r.t. each one of them separately. Many machine learning/imaging processing problems are given in this form (e.g., [19]). Current practice is to first run an NLP solver to find a stationary point, and then a CP solver after one variable (e.g., dictionary in [19]) is fixed. A more powerful, unified treatment for both convex and nonconvex problems is desirable to better handle these types of problems.

Our contribution mainly lies in the following three aspects. First, we consider the classic NLP problem given in the form of (1.1), where $\Psi(\cdot)$ is a smooth (possibly nonconvex) function satisfying (1.2) (denoted by $\Psi \in \mathcal{C}^{1,1}_{L_\Psi}(\mathbb{R}^n)$). In addition, we assume that $\Psi(\cdot)$ is bounded from below. We demonstrate that the AG method, when employed with a certain stepsize policy, can find an ϵ-solution of (1.1), i.e., a point $\bar{x}$ such that $\|\nabla\Psi(\bar{x})\|^2 \le \epsilon$, in at most $\mathcal{O}(1/\epsilon)$ iterations, which is the best-known complexity bound possessed by first-order methods to solve general NLP problems (e.g., the gradient descent method [24,4] and the trust region method [29]). Note that if Ψ is convex and a more aggressive stepsize policy is applied in the AG method, then the aforementioned complexity bound can be improved to $\mathcal{O}(1/\epsilon^{1/3})$.

Second, we consider a class of composite problems given by

$$\min_{x\in\mathbb{R}^n} \Psi(x) + \mathcal{X}(x), \quad \Psi(x) := f(x) + h(x), \tag{1.3}$$

where $f \in \mathcal{C}^{1,1}_{L_f}(\mathbb{R}^n)$ is possibly nonconvex, $h \in \mathcal{C}^{1,1}_{L_h}(\mathbb{R}^n)$ is convex, and $\mathcal{X}$ is a simple convex (possibly non-smooth) function with bounded domain (e.g., $\mathcal{X}(x) = \mathcal{I}_X(x)$ with $\mathcal{I}_X(\cdot)$ being the indicator function of a convex compact set $X \subset \mathbb{R}^n$). Clearly, we have $\Psi \in \mathcal{C}^{1,1}_{L_\Psi}(\mathbb{R}^n)$ with $L_\Psi = L_f + L_h$. Since $\mathcal{X}$ is possibly non-differentiable, we need to employ a different termination criterion based on the gradient mapping $\mathcal{G}(\cdot,\cdot,\cdot)$ (see (2.38)) to analyze the complexity of the AG method. Observe, however, that if $\mathcal{X}(x) = 0$, then we have $\mathcal{G}(x, \nabla\Psi(x), c) = \nabla\Psi(x)$ for any $c > 0$. We show that the same aggressive stepsize policy as the AG method for the convex problems can be applied for solving problem (1.3) no matter if $\Psi(\cdot)$ is convex or not. More specifically, the AG method exhibits an optimal rate of convergence in terms of functional optimality gap if $\Psi(\cdot)$ turns out to be convex. In addition, we show that one can find a solution $\bar{x} \in \mathbb{R}^n$ s.t. $\|\mathcal{G}(x, \nabla\Psi(x), c)\|^2 \le \epsilon$ in at most

$$\mathcal{O}\left\{\left(\frac{L_\Psi^2}{\epsilon}\right)^{1/3} + \frac{L_\Psi L_f}{\epsilon}\right\}$$

iterations. The above complexity bound improves the one established in [13] for the projected gradient method applied to problem (1.3) in terms of their dependence on the Lipschtiz constant L_h. In addition, it is significantly better than the latter bound when L_f is small enough (see Section 2.2 for more details).

Third, we consider stochastic NLP problems in the form of (1.1) or (1.3), where only noisy first-order information about Ψ is available via subsequent calls to a stochastic oracle ($\mathcal{SO}$). More specifically, at the k-th call, $x_k \in \mathbb{R}^n$ being the input, the $\mathcal{SO}$ outputs a stochastic gradient $G(x_k, \xi_k)$, where $\{\xi_k\}_{k\geq 1}$ are random vectors whose distributions P_k are supported on $\Xi_k \subseteq \mathbb{R}^d$. The following assumptions are also made for the stochastic gradient $G(x_k, \xi_k)$.

Assumption 1 *For any $x \in \mathbb{R}^n$ and $k \geq 1$, we have*

$$a)\ \ \mathbb{E}[G(x, \xi_k)] = \nabla\Psi(x), \tag{1.4}$$

$$b)\ \ \mathbb{E}\left[\|G(x, \xi_k) - \nabla\Psi(x)\|^2\right] \leq \sigma^2. \tag{1.5}$$

Currently, the randomized stochastic gradient (RSG) method initially studied by Ghadimi and Lan [12] and later improved in [13,6] seems to be the only available stochastic approximation (SA) algorithm for solving the aforementioned general stochastic NLP problems, while other SA methods (see, e.g., [28,21,30,27,14,12,10]) require the convexity assumption about Ψ. However, the RSG method and its variants are only nearly optimal for solving convex SP problems. Based on the AG method, we present a randomized stochastic AG (RSAG) method for solving general stochastic NLP problems and show that if $\Psi(\cdot)$ is convex, then the RSAG exhibits an optimal rate of convergence in terms of functional optimality gap, similarly to the accelerated SA method in [14]. In this case, the complexity bound in (1.6) in terms of the residual of gradients can be improved to

$$\mathcal{O}\left(\frac{L_\Psi^{\frac{2}{3}}}{\epsilon^{\frac{1}{3}}} + \frac{L_\Psi^{\frac{2}{3}}\sigma^2}{\epsilon^{\frac{4}{3}}}\right).$$

Moreover, if $\Psi(\cdot)$ is nonconvex, then the RSAG method can find an ϵ-solution of (1.1), i.e., a point $\bar{x}$ s.t. $\mathbb{E}[\|\nabla\Psi(\bar{x})\|^2] \leq \epsilon$ in at most

$$\mathcal{O}\left(\frac{L_\Psi}{\epsilon} + \frac{L_\Psi\sigma^2}{\epsilon^2}\right) \tag{1.6}$$

calls to the $\mathcal{SO}$. We also generalize these complexity analyses to a class of nonconvex stochastic composite optimization problems by introducing a mini-batch approach into the RSAG method and improve a few complexity results presented in [13] for solving these stochastic composite optimization problems.

This paper is organized as follows. In Section 2, we present the AG algorithm and establish its convergence properties for solving problems (1.1) and (1.3). We then generalize the AG method for solving stochastic nonlinear and composite optimization problems in Section 3. Some brief concluding remarks are given in Section 4.

2 The accelerated gradient algorithm

Our goal in this section is to show that the AG method, which is originally designed for smooth convex optimization, also converges for solving nonconvex optimization problems after incorporating some proper modification. More specifically, we first present an AG method for solving a general class of nonlinear optimization problems in Subsection 2.1 and then describe the AG method for solving a special class of nonconvex composite optimization problems in Subsection 2.2.

2.1 Minimization of smooth functions

In this subsection, we assume that $\Psi(\cdot)$ is a differentiable nonconvex function, bounded from below and its gradient satisfies in (1.2). It then follows that (see, e.g., [24])

$$|\Psi(y) - \Psi(x) - \langle\nabla\Psi(x), y - x\rangle| \leq \frac{L_\Psi}{2}\|y - x\|^2 \quad \forall x, y \in \mathbb{R}^n. \tag{2.1}$$

While the gradient descent method converges for solving the above class of nonconvex optimization problems, it does not achieve the optimal rate of convergence, in terms of the functional optimality gap, when $\Psi(\cdot)$ is convex. On the other hand, the original AG method in [23] is optimal for solving convex optimization problems, but does not necessarily converge for solving nonconvex optimization problems. Below, we present a modified AG method and show that by properly specifying the stepsize policy, it not only achieves the optimal rate of convergence for convex optimization, but also exhibits the best-known rate of convergence as shown in [24,4] for solving general smooth NLP problems by using first-order methods.

Algorithm 1 The accelerated gradient (AG) algorithm

Input: $x_0 \in \mathbb{R}^n$, $\{\alpha_k\}$ s.t. $\alpha_1 = 1$ and $\alpha_k \in (0,1)$ for any $k \geq 2$, $\{\beta_k > 0\}$, and $\{\lambda_k > 0\}$.
0. Set the initial points $x_0^{ag} = x_0$ and $k = 1$.
1. Set

$$x_k^{md} = (1-\alpha_k)x_{k-1}^{ag} + \alpha_k x_{k-1}. \tag{2.2}$$

2. Compute $\nabla\Psi(x_k^{md})$ and set

$$x_k = x_{k-1} - \lambda_k \nabla\Psi(x_k^{md}), \tag{2.3}$$
$$x_k^{ag} = x_k^{md} - \beta_k \nabla\Psi(x_k^{md}). \tag{2.4}$$

3. Set $k \leftarrow k+1$ and go to step 1.

Note that, if $\beta_k = \alpha_k\lambda_k \ \ \forall k \geq 1$, then we have $x_k^{ag} = \alpha_k x_k + (1-\alpha_k)x_{k-1}^{ag}$. In this case, the above AG method is equivalent to one of the simplest variants of the well-known Nesterov's method (see, e.g., [24]). On the other hand, if $\beta_k = \lambda_k, \ \ k = 1,2,\ldots$, then it can be shown by induction that $x_k^{md} = x_{k-1}$ and $x_k^{ag} = x_k$. In this case, Algorithm 1 reduces to the gradient descent method. We will show in this subsection that the above AG method actually converges for different selections of $\{\alpha_k\}$, $\{\beta_k\}$, and $\{\lambda_k\}$ in both convex and nonconvex case.

To establish the convergence of the above AG method, we need the following simple technical result (see Lemma 3 of [15] for a slightly more general result).

Lemma 1 *Let $\{\alpha_k\}$ be the stepsizes in the AG method and the sequence $\{\theta_k\}$ satisfies*

$$\theta_k \leq (1-\alpha_k)\theta_{k-1} + \eta_k, \quad k = 1,2,\ldots, \tag{2.5}$$

where

$$\Gamma_k := \begin{cases} 1, & k = 1, \\ (1-\alpha_k)\Gamma_{k-1}, & k \geq 2. \end{cases} \tag{2.6}$$

Then we have $\theta_k \leq \Gamma_k \sum_{i=1}^k (\eta_i/\Gamma_i)$ for any $k \geq 1$.

Proof. Noting that $\alpha_1 = 1$ and $\alpha_k \in (0,1)$ for any $k \geq 2$. These observations together with (2.6) then imply that $\Gamma_k > 0$ for any $k \geq 1$. Dividing both sides of (2.5) by Γ_k, we obtain

$$\frac{\theta_1}{\Gamma_1} \leq \frac{(1-\alpha_1)\theta_0}{\Gamma_1} + \frac{\eta_1}{\Gamma_1} = \frac{\eta_1}{\Gamma_1}$$

and

$$\frac{\theta_i}{\Gamma_i} \leq \frac{(1-\alpha_i)\theta_{i-1}}{\Gamma_i} + \frac{\eta_i}{\Gamma_i} = \frac{\theta_{i-1}}{\Gamma_{i-1}} + \frac{\eta_i}{\Gamma_i}, \quad \forall i \geq 2.$$

The result then immediately follows by summing up the above inequalities and rearranging the terms. ∎

We are now ready to describe the main convergence properties of the AG method.

Theorem 1 *Let $\{x_k^{md}, x_k^{ag}\}_{k\geq 1}$ be computed by Algorithm 1 and Γ_k be defined in (2.6).*

a) If $\{\alpha_k\}$, $\{\beta_k\}$, and $\{\lambda_k\}$ are chosen such that

$$C_k := 1 - L_\Psi \lambda_k - \frac{L_\Psi(\lambda_k - \beta_k)^2}{2\alpha_k \Gamma_k \lambda_k}\left(\sum_{\tau=k}^{N} \Gamma_\tau\right) > 0, \tag{2.7}$$

then for any $N \geq 1$, we have

$$\min_{k=1,\ldots,N} \|\nabla\Psi(x_k^{md})\|^2 \leq \frac{\Psi(x_0) - \Psi^*}{\sum_{k=1}^{N} \lambda_k C_k}. \tag{2.8}$$

b) Suppose that $\Psi(\cdot)$ is convex and that an optimal solution x^* exists for problem (1.1). If $\{\alpha_k\}$, $\{\beta_k\}$, and $\{\lambda_k\}$ are chosen such that

$$\alpha_k \lambda_k \leq \beta_k < \frac{1}{L_\Psi}, \tag{2.9}$$

$$\frac{\alpha_1}{\lambda_1 \Gamma_1} \geq \frac{\alpha_2}{\lambda_2 \Gamma_2} \geq \cdots, \tag{2.10}$$

then for any $N \geq 1$, we have

$$\min_{k=1,\ldots,N} \|\nabla\Psi(x_k^{md})\|^2 \leq \frac{\|x_0 - x^*\|^2}{\lambda_1 \sum_{k=1}^{N} \Gamma_k^{-1} \beta_k (1 - L_\Psi \beta_k)}, \tag{2.11}$$

$$\Psi(x_N^{ag}) - \Psi(x^*) \leq \frac{\Gamma_N \|x_0 - x^*\|^2}{2\lambda_1}. \tag{2.12}$$

Proof. We first show part a). Denote $\Delta_k := \nabla\Psi(x_{k-1}) - \nabla\Psi(x_k^{md})$. By (1.2) and (2.2), we have

$$\|\Delta_k\| = \|\nabla\Psi(x_{k-1}) - \nabla\Psi(x_k^{md})\| \leq L_\Psi \|x_{k-1} - x_k^{md}\| = L_\Psi(1 - \alpha_k)\|x_{k-1}^{ag} - x_{k-1}\|. \tag{2.13}$$

Also by (2.1) and (2.3), we have

$$\begin{aligned}
\Psi(x_k) &\leq \Psi(x_{k-1}) + \langle \nabla\Psi(x_{k-1}), x_k - x_{k-1}\rangle + \frac{L_\Psi}{2}\|x_k - x_{k-1}\|^2 \\
&= \Psi(x_{k-1}) + \langle \Delta_k + \nabla\Psi(x_k^{md}), -\lambda_k \nabla\Psi(x_k^{md})\rangle + \frac{L_\Psi \lambda_k^2}{2}\|\nabla\Psi(x_k^{md})\|^2 \\
&= \Psi(x_{k-1}) - \lambda_k\left(1 - \frac{L_\Psi \lambda_k}{2}\right)\|\nabla\Psi(x_k^{md})\|^2 - \lambda_k \langle \Delta_k, \nabla\Psi(x_k^{md})\rangle \\
&\leq \Psi(x_{k-1}) - \lambda_k\left(1 - \frac{L_\Psi \lambda_k}{2}\right)\|\nabla\Psi(x_k^{md})\|^2 + \lambda_k \|\Delta_k\| \cdot \|\nabla\Psi(x_k^{md})\|,
\end{aligned} \tag{2.14}$$

where the last inequality follows from the Cauchy-Schwarz inequality. Combining the previous two inequalities, we obtain

$$\begin{aligned}
\Psi(x_k) &\leq \Psi(x_{k-1}) - \lambda_k\left(1 - \frac{L_\Psi \lambda_k}{2}\right)\|\nabla\Psi(x_k^{md})\|^2 + L_\Psi(1 - \alpha_k)\lambda_k \|\nabla\Psi(x_k^{md})\| \cdot \|x_{k-1}^{ag} - x_{k-1}\| \\
&\leq \Psi(x_{k-1}) - \lambda_k\left(1 - \frac{L_\Psi \lambda_k}{2}\right)\|\nabla\Psi(x_k^{md})\|^2 + \frac{L_\Psi \lambda_k^2}{2}\|\nabla\Psi(x_k^{md})\|^2 + \frac{L_\Psi(1 - \alpha_k)^2}{2}\|x_{k-1}^{ag} - x_{k-1}\|^2 \\
&= \Psi(x_{k-1}) - \lambda_k(1 - L_\Psi \lambda_k)\|\nabla\Psi(x_k^{md})\|^2 + \frac{L_\Psi(1 - \alpha_k)^2}{2}\|x_{k-1}^{ag} - x_{k-1}\|^2,
\end{aligned} \tag{2.15}$$

where the second inequality follows from the fact that $ab \leq (a^2 + b^2)/2$. Now, by (2.2), (2.3), and (2.4), we have

$$\begin{aligned}
x_k^{ag} - x_k &= (1 - \alpha_k)x_{k-1}^{ag} + \alpha_k x_{k-1} - \beta_k \nabla\Psi(x_k^{md}) - [x_{k-1} - \lambda_k \nabla\Psi(x_k^{md})] \\
&= (1 - \alpha_k)(x_{k-1}^{ag} - x_{k-1}) + (\lambda_k - \beta_k)\nabla\Psi(x_k^{md}),
\end{aligned}$$

which, in the view of Lemma 1, implies that

$$x_k^{ag} - x_k = \Gamma_k \sum_{\tau=1}^{k} \left(\frac{\lambda_\tau - \beta_\tau}{\Gamma_\tau} \right) \nabla \Psi(x_\tau^{md}).$$

Using the above identity, the Jensen's inequality for $\|\cdot\|^2$, and the fact that

$$\sum_{\tau=1}^{k} \frac{\alpha_\tau}{\Gamma_\tau} = \frac{\alpha_1}{\Gamma_1} + \sum_{\tau=2}^{k} \frac{1}{\Gamma_\tau} \left(1 - \frac{\Gamma_\tau}{\Gamma_{\tau-1}} \right) = \frac{1}{\Gamma_1} + \sum_{\tau=2}^{k} \left(\frac{1}{\Gamma_\tau} - \frac{1}{\Gamma_{\tau-1}} \right) = \frac{1}{\Gamma_k}, \tag{2.16}$$

we have

$$\begin{aligned} \|x_k^{ag} - x_k\|^2 &= \left\| \Gamma_k \sum_{\tau=1}^{k} \left(\frac{\lambda_\tau - \beta_\tau}{\Gamma_\tau} \right) \nabla \Psi(x_\tau^{md}) \right\|^2 = \left\| \Gamma_k \sum_{\tau=1}^{k} \frac{\alpha_\tau}{\Gamma_\tau} \left[\left(\frac{\lambda_\tau - \beta_\tau}{\alpha_\tau} \right) \nabla \Psi(x_\tau^{md}) \right] \right\|^2 \\ &\leq \Gamma_k \sum_{\tau=1}^{k} \frac{\alpha_\tau}{\Gamma_\tau} \left\| \left(\frac{\lambda_\tau - \beta_\tau}{\alpha_\tau} \right) \nabla \Psi(x_\tau^{md}) \right\|^2 = \Gamma_k \sum_{\tau=1}^{k} \frac{(\lambda_\tau - \beta_\tau)^2}{\Gamma_\tau \alpha_\tau} \|\nabla \Psi(x_\tau^{md})\|^2. \end{aligned} \tag{2.17}$$

Replacing the above bound in (2.15), we obtain

$$\begin{aligned} \Psi(x_k) &\leq \Psi(x_{k-1}) - \lambda_k(1 - L_\Psi \lambda_k) \|\nabla \Psi(x_k^{md})\|^2 + \frac{L_\Psi \Gamma_{k-1}(1-\alpha_k)^2}{2} \sum_{\tau=1}^{k-1} \frac{(\lambda_\tau - \beta_\tau)^2}{\Gamma_\tau \alpha_\tau} \|\nabla \Psi(x_\tau^{md})\|^2 \\ &\leq \Psi(x_{k-1}) - \lambda_k(1 - L_\Psi \lambda_k) \|\nabla \Psi(x_k^{md})\|^2 + \frac{L_\Psi \Gamma_k}{2} \sum_{\tau=1}^{k} \frac{(\lambda_\tau - \beta_\tau)^2}{\Gamma_\tau \alpha_\tau} \|\nabla \Psi(x_\tau^{md})\|^2 \end{aligned} \tag{2.18}$$

for any $k \geq 1$, where the last inequality follows from the definition of Γ_k in (2.6) and the fact that $\alpha_k \in (0, 1]$ for all $k \geq 1$. Summing up the above inequalities and using the definition of C_k in (2.7), we have

$$\begin{aligned} \Psi(x_N) &\leq \Psi(x_0) - \sum_{k=1}^{N} \lambda_k(1 - L_\Psi \lambda_k) \|\nabla \Psi(x_k^{md})\|^2 + \frac{L_\Psi}{2} \sum_{k=1}^{N} \Gamma_k \sum_{\tau=1}^{k} \frac{(\lambda_\tau - \beta_\tau)^2}{\Gamma_\tau \alpha_\tau} \|\nabla \Psi(x_\tau^{md})\|^2 \\ &= \Psi(x_0) - \sum_{k=1}^{N} \lambda_k(1 - L_\Psi \lambda_k) \|\nabla \Psi(x_k^{md})\|^2 + \frac{L_\Psi}{2} \sum_{k=1}^{N} \frac{(\lambda_k - \beta_k)^2}{\Gamma_k \alpha_k} \left(\sum_{\tau=k}^{N} \Gamma_\tau \right) \|\nabla \Psi(x_k^{md})\|^2 \\ &= \Psi(x_0) - \sum_{k=1}^{N} \lambda_k C_k \|\nabla \Psi(x_k^{md})\|^2. \end{aligned} \tag{2.19}$$

Re-arranging the terms in the above inequality and noting that $\Psi(x_N) \geq \Psi^*$, we obtain

$$\min_{k=1,\ldots,N} \|\nabla \Psi(x_k^{md})\|^2 \left(\sum_{k=1}^{N} \lambda_k C_k \right) \leq \sum_{k=1}^{N} \lambda_k C_k \|\nabla \Psi(x_k^{md})\|^2 \leq \Psi(x_0) - \Psi^*,$$

which, in view of the assumption that $C_k > 0$, clearly implies (2.8).

We now show part b). First, note that by (2.4), we have

$$\begin{aligned} \Psi(x_k^{ag}) &\leq \Psi(x_k^{md}) + \langle \nabla \Psi(x_k^{md}), x_k^{ag} - x_k^{md} \rangle + \frac{L_\Psi}{2} \|x_k^{ag} - x_k^{md}\|^2 \\ &= \Psi(x_k^{md}) - \beta_k \|\nabla \Psi(x_k^{md})\|^2 + \frac{L_\Psi \beta_k^2}{2} \|\nabla \Psi(x_k^{md})\|^2. \end{aligned} \tag{2.20}$$

Also by the convexity of $\Psi(\cdot)$ and (2.2),

$$\begin{aligned}\Psi(x_k^{md}) - [(1-\alpha_k)\Psi(x_{k-1}^{ag}) + \alpha_k\Psi(x)] &= \alpha_k\left[\Psi(x_k^{md}) - \Psi(x)\right] + (1-\alpha_k)\left[\Psi(x_k^{md}) - \Psi(x_{k-1}^{ag})\right] \\ &\le \alpha_k\langle\nabla\Psi(x_k^{md}), x_k^{md} - x\rangle + (1-\alpha_k)\langle\nabla\Psi(x_k^{md}), x_k^{md} - x_{k-1}^{ag}\rangle \\ &= \langle\nabla\Psi(x_k^{md}), \alpha_k(x_k^{md} - x) + (1-\alpha_k)(x_k^{md} - x_{k-1}^{ag})\rangle \\ &= \alpha_k\langle\nabla\Psi(x_k^{md}), x_{k-1} - x\rangle. \end{aligned} \tag{2.21}$$

It also follows from (2.3) that

$$\begin{aligned}\|x_{k-1} - x\|^2 - 2\lambda_k\langle\nabla\Psi(x_k^{md}), x_{k-1} - x\rangle + \lambda_k^2\|\nabla\Psi(x_k^{md})\|^2 \\ = \|x_{k-1} - \lambda_k\nabla\Psi(x_k^{md}) - x\|^2 = \|x_k - x\|^2,\end{aligned}$$

and hence that

$$\alpha_k\langle\nabla\Psi(x_k^{md}), x_{k-1} - x\rangle = \frac{\alpha_k}{2\lambda_k}\left[\|x_{k-1} - x\|^2 - \|x_k - x\|^2\right] + \frac{\alpha_k\lambda_k}{2}\|\nabla\Psi(x_k^{md})\|^2. \tag{2.22}$$

Combining (2.20), (2.21), and (2.22), we obtain

$$\begin{aligned}\Psi(x_k^{ag}) &\le (1-\alpha_k)\Psi(x_{k-1}^{ag}) + \alpha_k\Psi(x) + \frac{\alpha_k}{2\lambda_k}\left[\|x_{k-1} - x\|^2 - \|x_k - x\|^2\right] \\ &\quad - \beta_k\left(1 - \frac{L_\Psi\beta_k}{2} - \frac{\alpha_k\lambda_k}{2\beta_k}\right)\|\nabla\Psi(x_k^{md})\|^2 \\ &\le (1-\alpha_k)\Psi(x_{k-1}^{ag}) + \alpha_k\Psi(x) + \frac{\alpha_k}{2\lambda_k}\left[\|x_{k-1} - x\|^2 - \|x_k - x\|^2\right] \\ &\quad - \frac{\beta_k}{2}\left(1 - L_\Psi\beta_k\right)\|\nabla\Psi(x_k^{md})\|^2, \end{aligned} \tag{2.23}$$

where the last inequality follows from the assumption in (2.9). Subtracting $\Psi(x)$ from both sides of the above inequality and using Lemma 1, we conclude that

$$\begin{aligned}\frac{\Psi(x_N^{ag}) - \Psi(x)}{\Gamma_N} &\le \sum_{k=1}^N \frac{\alpha_k}{2\lambda_k\Gamma_k}\left[\|x_{k-1} - x\|^2 - \|x_k - x\|^2\right] - \sum_{k=1}^N \frac{\beta_k}{2\Gamma_k}\left(1 - L_\Psi\beta_k\right)\|\nabla\Psi(x_k^{md})\|^2 \\ &\le \frac{\|x_0 - x\|^2}{2\lambda_1} - \sum_{k=1}^N \frac{\beta_k}{2\Gamma_k}\left(1 - L_\Psi\beta_k\right)\|\nabla\Psi(x_k^{md})\|^2 \quad \forall x \in \mathbb{R}^n, \end{aligned} \tag{2.24}$$

where the second inequality follows from the simple relation that

$$\sum_{k=1}^N \frac{\alpha_k}{\lambda_k\Gamma_k}\left[\|x_{k-1} - x\|^2 - \|x_k - x\|^2\right] \le \frac{\alpha_1\|x_0 - x\|^2}{\lambda_1\Gamma_1} = \frac{\|x_0 - x\|^2}{\lambda_1} \tag{2.25}$$

due to (2.10) and the fact that $\alpha_1 = \Gamma_1 = 1$. Hence, (2.12) immediately follows from the above inequality and the assumption in (2.9). Moreover, fixing $x = x^*$, re-arranging the terms in (2.24), and noting the fact that $\Psi(x_N^{ag}) \ge \Psi(x^*)$, we obtain

$$\begin{aligned}\min_{k=1,\ldots,N}\|\nabla\Psi(x_k^{md})\|^2\sum_{k=1}^N \frac{\beta_k}{2\Gamma_k}\left(1 - L_\Psi\beta_k\right) &\le \sum_{k=1}^N \frac{\beta_k}{2\Gamma_k}\left(1 - L_\Psi\beta_k\right)\|\nabla\Psi(x_k^{md})\|^2 \\ &\le \frac{\|x^* - x_0\|^2}{2\lambda_1}, \end{aligned}$$

which together with (2.9), clearly imply (2.11). ∎

We add a few observations about Theorem 1. First, in view of (2.23), it is possible to use a different assumption than the one in (2.9) on the stepsize policies for the convex case. In particular, we only need

$$2 - L_\Psi \beta_k - \frac{\alpha_k \lambda_k}{\beta_k} > 0 \tag{2.26}$$

to show the convergence of the AG method for minimizing smooth convex problems. However, since the condition given by (2.9) is required for minimizing composite problems in Subsections 2.2 and 3.2, we state this assumption for the sake of simplicity. Second, there are various options for selecting $\{\alpha_k\}$, $\{\beta_k\}$, and $\{\lambda_k\}$ to guarantee the convergence of the AG algorithm. Below we provide some of these selections for solving both convex and nonconvex problems.

Corollary 1 *Suppose that $\{\alpha_k\}$ and $\{\beta_k\}$ in the AG method are set to*

$$\alpha_k = \frac{2}{k+1} \quad \text{and} \quad \beta_k = \frac{1}{2L_\Psi}. \tag{2.27}$$

a) If $\{\lambda_k\}$ satisifies

$$\lambda_k \in \left[\beta_k, \left(1 + \frac{\alpha_k}{4}\right)\beta_k\right] \quad \forall k \geq 1, \tag{2.28}$$

then for any $N \geq 1$, we have

$$\min_{k=1,\ldots,N} \|\nabla \Psi(x_k^{md})\|^2 \leq \frac{6L_\Psi[\Psi(x_0) - \Psi^*]}{N}. \tag{2.29}$$

b) Assume that $\Psi(\cdot)$ is convex and that an optimal solution x^* exists for problem (1.1). If $\{\lambda_k\}$ satisfies

$$\lambda_k = \frac{k\,\beta_k}{2} \quad \forall k \geq 1, \tag{2.30}$$

then for any $N \geq 1$, we have

$$\min_{k=1,\ldots,N} \|\nabla \Psi(x_k^{md})\|^2 \leq \frac{96L_\Psi^2 \|x_0 - x^*\|^2}{N^2(N+1)}, \tag{2.31}$$

$$\Psi(x_N^{ag}) - \Psi(x^*) \leq \frac{4L_\Psi \|x_0 - x^*\|^2}{N(N+1)}. \tag{2.32}$$

Proof. We first show part a). Note that by (2.6) and (2.27), we have

$$\Gamma_k = \frac{2}{k(k+1)}, \tag{2.33}$$

which implies that

$$\sum_{\tau=k}^{N} \Gamma_\tau = \sum_{\tau=k}^{N} \frac{2}{\tau(\tau+1)} = 2\sum_{\tau=k}^{N} \left(\frac{1}{\tau} - \frac{1}{\tau+1}\right) \leq \frac{2}{k}. \tag{2.34}$$

It can also be easily seen from (2.28) that $0 \leq \lambda_k - \beta_k \leq \alpha_k \beta_k/4$. Using these observations, (2.27), and (2.28), we have

$$\begin{aligned} C_k &= 1 - L_\Psi \left[\lambda_k + \frac{(\lambda_k - \beta_k)^2}{2\alpha_k \Gamma_k \lambda_k}\left(\sum_{\tau=k}^{N} \Gamma_\tau\right)\right] \\ &\geq 1 - L_\Psi \left[\left(1 + \frac{\alpha_k}{4}\right)\beta_k + \frac{\alpha_k^2 \beta_k^2}{16} \frac{1}{k\alpha_k \Gamma_k \beta_k}\right] \\ &= 1 - \beta_k L_\Psi \left(1 + \frac{\alpha_k}{4} + \frac{1}{16}\right) \\ &\geq 1 - \beta_k L_\Psi \frac{21}{16} = \frac{11}{32}, \\ \lambda_k C_k &\geq \frac{11\beta_k}{32} = \frac{11}{64L_\Psi} \geq \frac{1}{6L_\Psi}. \end{aligned} \tag{2.35}$$

Combining the above relation with (2.8), we obtain (2.29).

We now show part b). Observe that by (2.27) and (2.30), we have

$$\alpha_k\lambda_k = \frac{k}{k+1}\beta_k < \beta_k,$$
$$\frac{\alpha_1}{\lambda_1\Gamma_1} = \frac{\alpha_2}{\lambda_2\Gamma_2} = \ldots = 4L_\Psi,$$

which implies that conditions (2.9) and (2.10) hold. Moreover, we have

$$\sum_{k=1}^{N}\Gamma_k^{-1}\beta_k(1-L_\Psi\beta_k) = \frac{1}{4L_\Psi}\sum_{k=1}^{N}\Gamma_k^{-1} \geq \frac{1}{8L_\Psi}\sum_{k=1}^{N}k^2 = \frac{1}{48L_\Psi}N(N+1)(2N+1) \geq \frac{N^2(N+1)}{24L_\Psi}. \tag{2.36}$$

Using (2.33) and the above bounds in (2.11) and (2.12), we obtain (2.31) and (2.32). ∎

We now add a few remarks about the results obtained in Corollary 1. First, the rate of convergence in (2.29) for the AG method is in the same order of magnitude as that for the gradient descent method ([24]). It is also worth noting that by choosing $\lambda_k = \beta_k$ in (2.28), the rate of convergence for the AG method is just changed up to a constant factor. However, in this case, the AG method is reduced to the gradient descent method as mentioned earlier in this subsection. Second, if the problem is convex, by choosing more aggressive stepsize $\{\lambda_k\}$ in (2.30), the AG method exhibits the optimal rate of convergence in (2.32). Moreover, with such a selection of $\{\lambda_k\}$, the AG method can find a solution $\bar{x}$ such that $\|\nabla\Psi(\bar{x})\|^2 \leq \epsilon$ in at most $\mathcal{O}(1/\epsilon^{\frac{1}{3}})$ iterations according to (2.31). The latter result has also been established in [20, Proposition 5.2] for an accelerated hybrid proximal extra-gradient method when applied to convex problems.

Observe that $\{\lambda_k\}$ in (2.28) for general nonconvex problems is in the order of $\mathcal{O}(1/L_\Psi)$, while the one in (2.30) for convex problems are more aggressive (in $\mathcal{O}(k/L_\Psi)$). An interesting question is whether we can apply the same stepsize policy in (2.30) for solving general NLP problems no matter they are convex or not. We will discuss such a uniform treatment for both convex and nonconvex optimization for solve a certain class of composite problems in next subsection.

2.2 Minimization of nonconvex composite functions

In this subsection, we consider a special class of NLP problems given in the form of (1.3). Our goal in this subsection is to show that we can employ a more aggressive stepsize policy in the AG method, similar to the one used in the convex case (see Theorem 1.b) and Corollary 1.b)), to solve these composite problems, even if $\Psi(\cdot)$ is possibly nonconvex.

Throughout this subsection, we make the following assumption about the convex (possibly non-differentiable) component $\mathcal{X}(\cdot)$ in (1.3).

Assumption 2 *There exists a constant M such that $\|\mathcal{P}(x,y,c)\| \leq M$ for any $c \in (0,+\infty)$ and $x, y \in \mathbb{R}^n$, where $\mathcal{P}(x,y,c)$ is given by*

$$\mathcal{P}(x,y,c) := \mathrm{argmin}_{u\in\mathbb{R}^n}\left\{\langle y,u\rangle + \frac{1}{2c}\|u-x\|^2 + \mathcal{X}(u)\right\}. \tag{2.37}$$

Next result shows a certain class of functions $\mathcal{X}(\cdot)$ which assures that Assumption 2 is satisfied.

Lemma 2 *If $\mathcal{X}(\cdot)$ is a proper closed convex function with bounded domain, then Assumption 2 is satisfied.*

Proof. Denote $X \equiv \mathrm{dom}(\mathcal{X}) := \{u|\mathcal{X}(u) < +\infty\}$. Note that by assumption, X is nonempty and bounded. Also observe that (2.37) is equivalent to

$$\mathcal{P}(x,y,c) = \mathrm{argmin}_{u\in\mathbb{R}^n}\left\{w(u) := \frac{1}{2c}\|u-x+cy\|^2 + \mathcal{X}(u)\right\}.$$

For any $u \notin X$, we have $\mathcal{X}(u) = +\infty$, which together with the fact that $c > 0$ then imply that $w(u) = +\infty$ for any $u \notin X$. Hence, $\mathcal{P}(x,y,c) \in X$ and the result follows immediately. ∎

Based on the above result, we can give the following examples. Let $X \subseteq \mathbb{R}^n$ be a given convex compact set. It can be easily seen that Assumption 2 holds if $\mathcal{X}(x) = \mathcal{I}_X(x)$. Here $\mathcal{I}_X$ is the indicator function of X given by

$$\mathcal{I}_X(x) = \begin{cases} 0 & x \in X, \\ +\infty & x \notin X. \end{cases}$$

Another important example is given by $\mathcal{X}(x) = \mathcal{I}_X(x) + \|x\|_1$, where $\|\cdot\|_1$ denotes the l_1 norm.

Observe that $\mathcal{P}(x, y, c)$ in (2.37) also gives rise to an important quantity that will be used frequently in our convergence analysis, i.e.,

$$\mathcal{G}(x, y, c) := \frac{1}{c}[x - \mathcal{P}(x, y, c)]. \tag{2.38}$$

In particular, if $y = \nabla\Psi(x)$, then $\mathcal{G}(x, y, c)$ is called the gradient mapping at x, which has been used as a termination criterion for solving constrained or composite NLP problems (see, e.g., [22,24,16,13,17]). It can be easily seen that $\mathcal{G}(x, \nabla\Psi(x), c) = \nabla\Psi(x)$ for any $c > 0$ when $\mathcal{X}(\cdot) = 0$. For more general $\mathcal{X}(\cdot)$, the following result shows that as the size of $\mathcal{G}(x, \nabla\Psi(x), c)$ vanishes, $\mathcal{P}(x, \nabla\Psi(x), c)$ approaches to a stationary point of problem (1.3).

Lemma 3 *Let $x \in \mathbb{R}^n$ be given and denote $g \equiv \nabla\Psi(x)$. If $\|\mathcal{G}(x, g, c)\| \le \epsilon$ for some $c > 0$, then*

$$-\nabla\Psi(\mathcal{P}(x, g, c)) \in \partial\mathcal{X}(\mathcal{P}(x, g, c)) + \mathcal{B}(\epsilon(cL_\Psi + 1)),$$

where $\partial\mathcal{X}(\cdot)$ denotes the subdifferential of $\mathcal{X}(\cdot)$ and $\mathcal{B}(r) := \{x \in \mathbb{R}^n : \|x\| \le r\}$.

Proof. By the optimality condition of (2.37), we have $-\nabla\Psi(x) - \frac{1}{c}(\mathcal{P}(x, g, c) - x) \in \partial\mathcal{X}(\mathcal{P}(x, g, c))$, which implies that

$$-\nabla\Psi(\mathcal{P}(x, g, c)) + \left[\nabla\Psi(\mathcal{P}(x, g, c)) - \nabla\Psi(x) - \frac{1}{c}(\mathcal{P}(x, g, c) - x)\right] \in \partial\mathcal{X}(\mathcal{P}(x, g, c)). \tag{2.39}$$

Our conclusion immediately follows from the above relation and the simple fact that

$$\begin{aligned}\|\nabla\Psi(\mathcal{P}(x, g, c)) - \nabla\Psi(x) - \frac{1}{c}(\mathcal{P}(x, g, c) - x\| &\le L_\Psi\|\mathcal{P}(x, g, c) - x\| + \frac{1}{c}\|\mathcal{P}(x, g, c) - x\| \\ &= (cL_\Psi + 1)\|\mathcal{G}(x, g, c)\|.\end{aligned}$$

■

The following result shows that $\mathcal{G}(x, \cdot, c)$ is Lipschitz continuous (see, e.g., Proposition 1 of [13]).

Lemma 4 *For any $y_1, y_2 \in \mathbb{R}^n$, we have $\|\mathcal{G}(x, y_1, c) - \mathcal{G}(x, y_2, c)\| \le \|y_1 - y_2\|$.*

We are now ready to describe the AG algorithm for solving problem (1.3), which differs from Algorithm 1 only in Step 2.

Algorithm 2 The AG method for composite optimization

Replace (2.3) and (2.4) in Step 2 of the Algorithm 1, respectively, by

$$x_k = \mathcal{P}(x_{k-1}, \nabla\Psi(x_k^{md}), \lambda_k), \tag{2.40}$$

$$x_k^{ag} = \mathcal{P}(x_k^{md}, \nabla\Psi(x_k^{md}), \beta_k). \tag{2.41}$$

A few remarks about Algorithm 2 are in place. First, observe that the subproblems (2.40) and (2.41) are given in the form of (2.37) and hence that under Assumption 2, the search points x_k and x_k^{ag} $\forall k \ge 1$, will stay in a bounded set. Second, we need to assume that $\mathcal{X}(\cdot)$ is simple enough so that the subproblems (2.40) and (2.41) are easily computable. Third, in view of (2.38) and (2.41), we have

$$\mathcal{G}(x_k^{md}, \nabla\Psi(x_k^{md}), \beta_k) = \frac{1}{\beta_k}(x_k^{md} - x_k^{ag}). \tag{2.42}$$

We will use $\|\mathcal{G}(x_k^{md}, \nabla\Psi(x_k^{md}), \beta_k)\|$ as a termination criterion in the above AG method for composite optimization.

Before establishing the convergence of the above AG method, we first state a technical result which shows that the relation in (2.1) can be enhanced for composite functions.

Lemma 5 *Let $\Psi(\cdot)$ be defined in (1.3). For any $x, y \in \mathbb{R}^n$, we have*

$$-\frac{L_f}{2}\|y-x\|^2 \le \Psi(y) - \Psi(x) - \langle \nabla\Psi(x), y-x\rangle \le \frac{L_\Psi}{2}\|y-x\|^2. \tag{2.43}$$

Proof. We only need to show the first relation since the secone one follows from (2.1). Indeed,

$$\begin{aligned}
\Psi(y) - \Psi(x) &= \int_0^1 \langle \nabla\Psi(x+t(y-x)), y-x\rangle dt \\
&= \int_0^1 \langle \nabla f(x+t(y-x)), y-x\rangle dt + \int_0^1 \langle \nabla h(x+t(y-x)), y-x\rangle dt \\
&= \langle \nabla f(x), y-x\rangle + \int_0^1 \langle \nabla f(x+t(y-x)) - \nabla f(x), y-x\rangle dt \\
&\quad + \langle \nabla h(x), y-x\rangle + \int_0^1 \langle \nabla h(x+t(y-x)), y-x\rangle dt \\
&\ge \langle \nabla f(x), y-x\rangle + \int_0^1 \langle \nabla f(x+t(y-x)) - \nabla f(x), y-x\rangle dt + \langle \nabla h(x), y-x\rangle \\
&\ge \langle \nabla\Psi(x), y-x\rangle - \frac{L_f}{2}\|y-x\|^2 \quad \forall x, y \in \mathbb{R}^n,
\end{aligned}$$

where the first inequality follows from the fact that $\langle \nabla h(x+t(y-x), y-x\rangle \ge 0$ due to the convexity of h, and the last inequality follows from the fact that

$$\langle \nabla f(x+t(y-x)) - \nabla f(x), y-x\rangle \ge -\|f(x+t(y-x)) - \nabla f(x)\|\|y-x\| \ge -L_f t\|y-x\|^2.$$

■

We are now ready to describe the main convergence properties of Algorithm 2 for solving problem (1.3).

Theorem 2 *Suppose that Assumption 2 holds and that $\{\alpha_k\}$, $\{\beta_k\}$, and $\{\lambda_k\}$ in Algorithm 2 are chosen such that (2.9) and (2.10) hold. Also assume that an optimal solution x^* exists for problem (1.3). Then for any $N \ge 1$, we have*

$$\min_{k=1,\ldots,N} \|\mathcal{G}(x_k^{md}, \nabla\Psi(x_k^{md}), \beta_k)\|^2 \le 2\left[\sum_{k=1}^N \Gamma_k^{-1}\beta_k(1-L_\Psi\beta_k)\right]^{-1}\left[\frac{\|x_0-x^*\|^2}{2\lambda_1} + \frac{L_f}{\Gamma_N}(\|x^*\|^2 + 2M^2)\right], \tag{2.44}$$

where $\mathcal{G}(\cdot,\cdot,\cdot)$ is defined in (2.38). If, in addition, $L_f = 0$, then we have

$$\Phi(x_N^{ag}) - \Phi(x^*) \le \frac{\Gamma_N\|x_0-x^*\|^2}{2\lambda_1}, \tag{2.45}$$

where $\Phi(x) \equiv \Psi(x) + \mathcal{X}(x)$.

Proof. By the assumption that $\Psi \in \mathcal{C}_{L_\Psi}^{1,1}(\mathbb{R}^n)$, we have

$$\Psi(x_k^{ag}) \le \Psi(x_k^{md}) + \langle \nabla\Psi(x_k^{md}), x_k^{ag} - x_k^{md}\rangle + \frac{L_\Psi}{2}\|x_k^{ag} - x_k^{md}\|^2. \tag{2.46}$$

Also by Lemma 5, we have

$$\begin{aligned}
&\Psi(x_k^{md}) - [(1-\alpha_k)\Psi(x_{k-1}^{ag}) + \alpha_k\Psi(x)] = \alpha_k[\Psi(x_k^{md}) - \Psi(x)] + (1-\alpha_k)[\Psi(x_k^{md}) - \Psi(x_{k-1}^{ag}] \\
&\quad \le \alpha_k\left[\langle \nabla\Psi(x_k^{md}), x_k^{md} - x\rangle + \frac{L_f}{2}\|x_k^{md} - x\|^2\right] \\
&\qquad + (1-\alpha_k)\left[\langle \nabla\Psi(x_k^{md}), x_k^{md} - x_{k-1}^{ag}\rangle + \frac{L_f}{2}\|x_k^{md} - x_{k-1}^{ag}\|^2\right] \\
&\quad = \langle \nabla\Psi(x_k^{md}), x_k^{md} - \alpha_k x - (1-\alpha_k)x_{k-1}^{ag}\rangle + \frac{L_f\alpha_k}{2}\|x_k^{md} - x\|^2 + \frac{L_f(1-\alpha_k)}{2}\|x_k^{md} - x_{k-1}^{ag}\|^2 \\
&\quad \le \langle \nabla\Psi(x_k^{md}), x_k^{md} - \alpha_k x - (1-\alpha_k)x_{k-1}^{ag}\rangle + \frac{L_f\alpha_k}{2}\|x_k^{md} - x\|^2 + \frac{L_f\alpha_k^2(1-\alpha_k)}{2}\|x_{k-1}^{ag} - x_{k-1}\|^2,
\end{aligned} \tag{2.47}$$

where the last inequality follows from the fact that $x_k^{md} - x_{k-1}^{ag} = \alpha_k(x_{k-1}^{ag} - x_{k-1})$ due to (2.2). Now, by Lemma 2 of [11] for the solutions of subproblems (2.40) and (2.41), we have

$$\langle \nabla\Psi(x_k^{md}), x_k - x\rangle + \mathcal{X}(x_k) \le \mathcal{X}(x) + \frac{1}{2\lambda_k}\left[\|x_{k-1}-x\|^2 - \|x_k - x\|^2 - \|x_k - x_{k-1}\|^2\right], \tag{2.48}$$

$$\langle \nabla\Psi(x_k^{md}), x_k^{ag} - x\rangle + \mathcal{X}(x_k^{ag}) \le \mathcal{X}(x) + \frac{1}{2\beta_k}\left[\|x_k^{md}-x\|^2 - \|x_k^{ag} - x\|^2 - \|x_k^{ag} - x_k^{md}\|^2\right] \tag{2.49}$$

for any $x \in \mathbb{R}^n$. Letting $x = \alpha_k x_k + (1-\alpha_k)x_{k-1}^{ag}$ in (2.49), we have

$$\begin{aligned}
&\langle \nabla\Psi(x_k^{md}), x_k^{ag} - \alpha_k x_k - (1-\alpha_k)x_{k-1}^{ag}\rangle + \mathcal{X}(x_k^{ag}) \\
&\quad\le \mathcal{X}(\alpha_k x_k + (1-\alpha_k)x_{k-1}^{ag}) + \frac{1}{2\beta_k}\left[\|x_k^{md} - \alpha_k x_k - (1-\alpha_k)x_{k-1}^{ag}\|^2 - \|x_k^{ag} - x_k^{md}\|^2\right] \\
&\quad\le \alpha_k\mathcal{X}(x_k) + (1-\alpha_k)\mathcal{X}(x_{k-1}^{ag}) + \frac{1}{2\beta_k}\left[\alpha_k^2\|x_k - x_{k-1}\|^2 - \|x_k^{ag} - x_k^{md}\|^2\right],
\end{aligned}$$

where the last inequality follows from the convexity of $\mathcal{X}$ and (2.2). Summing up the above inequality with (2.48) (with both sides multiplied by α_k), we obtain

$$\begin{aligned}
&\langle \nabla\Psi(x_k^{md}), x_k^{ag} - \alpha_k x - (1-\alpha_k)x_{k-1}^{ag}\rangle + \mathcal{X}(x_k^{ag}) \le (1-\alpha_k)\mathcal{X}(x_{k-1}^{ag}) + \alpha_k\mathcal{X}(x) \\
&+\frac{\alpha_k}{2\lambda_k}\left[\|x_{k-1}-x\|^2 - \|x_k - x\|^2\right] + \frac{\alpha_k(\lambda_k\alpha_k - \beta_k)}{2\beta_k\lambda_k}\|x_k - x_{k-1}\|^2 - \frac{1}{2\beta_k}\|x_k^{ag} - x_k^{md}\|^2 \\
&\le (1-\alpha_k)\mathcal{X}(x_{k-1}^{ag}) + \alpha_k\mathcal{X}(x) + \frac{\alpha_k}{2\lambda_k}\left[\|x_{k-1}-x\|^2 - \|x_k - x\|^2\right] - \frac{1}{2\beta_k}\|x_k^{ag} - x_k^{md}\|^2,
\end{aligned} \tag{2.50}$$

where the last inequality follows from the assumption that $\alpha_k\lambda_k \le \beta_k$. Combining (2.46), (2.47), and (2.50), and using the definition $\Phi(x) \equiv \Psi(x) + \mathcal{X}(x)$, we have

$$\begin{aligned}
\Phi(x_k^{ag}) &\le (1-\alpha_k)\Phi(x_{k-1}^{ag}) + \alpha_k\Phi(x) - \frac{1}{2}\left(\frac{1}{\beta_k} - L_\Psi\right)\|x_k^{ag} - x_k^{md}\|^2 \\
&\quad + \frac{\alpha_k}{2\lambda_k}\left[\|x_{k-1}-x\|^2 - \|x_k - x\|^2\right] + \frac{L_f\alpha_k}{2}\|x_k^{md} - x\|^2 + \frac{L_f\alpha_k^2(1-\alpha_k)}{2}\|x_{k-1}^{ag} - x_{k-1}\|^2.
\end{aligned} \tag{2.51}$$

Subtracting $\Phi(x)$ from both sides of the above inequality, re-arranging the terms, and using Lemma 1 and relation (2.25), we obtain

$$\frac{\Phi(x_N^{ag}) - \Phi(x)}{\Gamma_N} + \sum_{k=1}^N \frac{1 - L_\Psi\beta_k}{2\beta_k\Gamma_k}\|x_k^{ag} - x_k^{md}\|^2 \le \frac{\|x_0 - x\|^2}{2\lambda_1} + \frac{L_f}{2}\sum_{k=1}^N \frac{\alpha_k}{\Gamma_k}[\|x_k^{md} - x\|^2 + \alpha_k(1-\alpha_k)\|x_{k-1}^{ag} - x_{k-1}\|^2].$$

Now letting $x = x^*$ in the above inequality, and observing that by Assumption 2 and (2.2),

$$\begin{aligned}
&\|x_k^{md} - x^*\|^2 + \alpha_k(1-\alpha_k)\|x_{k-1}^{ag} - x_{k-1}\|^2 \\
&\le 2[\|x^*\|^2 + \|x_k^{md}\|^2 + \alpha_k(1-\alpha_k)(\|x_{k-1}^{ag}\|^2 + \|x_{k-1}\|^2)] \\
&\le 2[\|x^*\|^2 + (1-\alpha_k)\|x_{k-1}^{ag}\|^2 + \alpha_k\|x_{k-1}\|^2 + \alpha_k(1-\alpha_k)(\|x_{k-1}^{ag}\|^2 + \|x_{k-1}\|^2)] \\
&\le 2[\|x^*\|^2 + \|x_{k-1}^{ag}\|^2 + \|x_{k-1}\|^2] \le 2(\|x^*\|^2 + 2M^2),
\end{aligned} \tag{2.52}$$

we obtain

$$\begin{aligned}
\frac{\Phi(x_N^{ag}) - \Phi(x^*)}{\Gamma_N} + \sum_{k=1}^N \frac{1 - L_\Psi\beta_k}{2\beta_k\Gamma_k}\|x_k^{ag} - x_k^{md}\|^2 &\le \frac{\|x_0 - x\|^2}{2\lambda_1} + L_f\sum_{k=1}^N \frac{\alpha_k}{\Gamma_k}(\|x^*\|^2 + 2M^2) \\
&\le \frac{\|x_0 - x\|^2}{2\lambda_1} + \frac{L_f}{\Gamma_N}(\|x^*\|^2 + 2M^2),
\end{aligned} \tag{2.53}$$

where the last inequality follows from (2.16). The above relation, in view of (2.9) and the assumption $L_f = 0$, then clearly implies (2.45). Moreover, it follows from the above relation, (2.42), and the fact $\Phi(x_N^{ag}) - \Phi(x^*) \geq 0$ that

$$\sum_{k=1}^{N} \frac{\beta_k(1 - L_\Psi \beta_k)}{2\Gamma_k} \|\mathcal{G}(x_k^{md}, \nabla\Psi(x_k^{md}), \beta_k)\|^2 = \sum_{k=1}^{N} \frac{1 - L_\Psi \beta_k}{2\beta_k \Gamma_k} \|x_k^{ag} - x_k^{md}\|^2$$
$$\leq \frac{\|x_0 - x^*\|^2}{2\lambda_1} + \frac{L_f}{\Gamma_N}(\|x^*\|^2 + 2M^2),$$

which, in view of (2.9), then clearly implies (2.44). ∎

As shown in Theorem 2, we can have a uniform treatment for both convex and nonconvex composite problems. More specifically, we allow the same stepsize policies in Theorem 1.b) to be used for both convex and nonconvex composite optimization. In the next result, we specialize the results obtained in Theorem 2 for a particular selection of $\{\alpha_k\}$, $\{\beta_k\}$, and $\{\lambda_k\}$.

Corollary 2 *Suppose that Assumption 2 holds and that $\{\alpha_k\}$, $\{\beta_k\}$, and $\{\lambda_k\}$ in Algorithm 2 are set to (2.27) and (2.30). Also assume that an optimal solution x^* exists for problem (1.3). Then for any $N \geq 1$, we have*

$$\min_{k=1,\ldots,N} \|\mathcal{G}(x_k^{md}, \nabla\Psi(x_k^{md}), \beta_k)\|^2 \leq 24 L_\Psi \left[\frac{4L_\Psi \|x_0 - x^*\|^2}{N^2(N+1)} + \frac{L_f}{N}(\|x^*\|^2 + 2M^2) \right]. \tag{2.54}$$

If, in addition, $L_f = 0$, then we have

$$\Phi(x_N^{ag}) - \Phi(x^*) \leq \frac{4L_\Psi \|x_0 - x^*\|^2}{N(N+1)}. \tag{2.55}$$

Proof. The results directly follow by plugging the value of Γ_k in (2.33), the value of λ_1 in (2.30), and the bound (2.36) into (2.44) and (2.45), respectively. ∎

Clearly, it follows from (2.54) that after running the AG method for at most $N = \mathcal{O}(L_\Psi^{\frac{2}{3}}/\epsilon^{\frac{1}{3}} + L_\Psi L_f/\epsilon)$ iterations, we have $-\nabla\Psi(x_N^{ag}) \in \partial\mathcal{X}(x_N^{ag}) + \mathcal{B}(\epsilon)$. Using the fact that $L_\Psi = L_f + L_h$, we can easily see that if either the smooth convex term $h(\cdot)$ or the nonconvex term $f(\cdot)$ becomes zero, then the previous complexity bound reduces to $\mathcal{O}(L_f^2/\epsilon)$ or $\mathcal{O}(L_h^2/\epsilon^{\frac{1}{3}})$, respectively.

It is interesting to compare the rate of convergence obtained in (2.54) with the one obtained in [13] for the projected gradient method applied to problem (1.3) . More specifically, let $\{p_k\}$ and $\{\nu_k\}$, respectively, denote the iterates and stepsizes in the projected gradient method. Also assume that the component $\mathcal{X}(\cdot)$ in (1.3) is Lipschitz continuous with Lipschtiz constant $L_\mathcal{X}$. Then, by Corollary 1 of [13], we have

$$\min_{k=1,\ldots,N} \|\mathcal{G}(p_k, \nabla\Psi(p_k), \nu_k\|^2 \leq \frac{L_\Psi[\Phi(p_0) - \Phi(x^*)]}{N}$$
$$\leq \frac{L_\Psi}{N}\left(\|\nabla\Psi(x^*)\| + L_\mathcal{X}\right)(\|x^*\| + M) + \frac{L_\Psi^2}{N}(\|x^*\|^2 + M^2), \tag{2.56}$$

where the last inequality follows from

$$\Phi(p_0) - \Phi(x^*) = \Psi(p_0) - \Psi(x^*) + \mathcal{X}(p_0) - \mathcal{X}(x^*)$$
$$\leq \langle \nabla\Psi(x^*), p_0 - x^* \rangle + \frac{L_\Psi}{2}\|p_0 - x^*\|^2 + L_\mathcal{X}\|p_0 - x^*\|$$
$$\leq \left(\|\nabla\Psi(x^*)\| + L_\mathcal{X}\right)\|p_0 - x^*\| + \frac{L_\Psi}{2}\|p_0 - x^*\|^2$$
$$\leq \left(\|\nabla\Psi(x^*)\| + L_\mathcal{X}\right)(\|x^*\| + M) + L_\Psi(\|x^*\|^2 + M^2).$$

Comparing (2.54) with (2.56), we can make the following observations. First, the bound in (2.54) does not depend on $L_\mathcal{X}$ while the one in (2.56) may depend on $L_\mathcal{X}$. Second, if the second terms in both (2.54) and (2.56) are the dominating ones, then the rate of convergence of the AG method is bounded by $\mathcal{O}(L_\Psi L_f/N)$, which is better than the $\mathcal{O}(L_\Psi^2/N)$

rate of convergence possessed by the projected gradient method, in terms of their dependence on the Lipschitz constant L_h. Third, consider the case when $L_f = \mathcal{O}(L_h/N^2)$. By (2.54), we have

$$\min_{k=1,\ldots,N} \|\mathcal{G}(x_k^{md}, \nabla\Psi(x_k^{md}), \beta_k)\|^2 \leq \frac{96L_\Psi^2\|x_0 - x^*\|^2}{N^3}\left(1 + \frac{L_f N^2(\|x^*\|^2 + 2M^2))}{4(L_f + L_h)\|x_0 - x^*\|^2}\right),$$

which implies that the rate of convergence of the AG method is bounded by

$$\mathcal{O}\left(\frac{L_h^2}{N^3}\left[\|x_0 - x^*\|^2 + \|x^*\|^2 + M^2\right]\right).$$

The previous bound is significantly better than the $\mathcal{O}(L_h^2/N)$ rate of convergence possessed by the projected gradient method for this particular case. Finally, it should be noted, however, that the projected gradient method in [13] can be used to solve more general problems as it does not require the domain of $\mathcal{X}$ to be bounded. Instead, it only requires the objective function $\Phi(x)$ to be bounded from below.

3 The stochastic accelerated gradient method

Our goal in this section is to present a stochastic counterpart of the AG algorithm for solving stochastic optimization problems. More specifically, we discuss the convergence of this algorithm for solving general smooth (possibly nonconvex) SP problems in Subsection 3.1, and for a special class of composite SP problems in Subsection 3.2.

3.1 Minimization of stochastic smooth functions

In this subsection, we consider problem (1.1), where $\Psi \in \mathcal{C}_{L_\Psi}^{1,1}(\mathbb{R}^n)$ is bounded from below. Moreover, we assume that the first-order information of $\Psi(\cdot)$ is obtained by the $\mathcal{SO}$, which satisfies Assumption 1. It should also be mentioned that in the standard setting for SP, the random vectors ξ_k, $k = 1, 2, \ldots$, are independent of each other (see, e.g., [22,21]). However, our assumption here is slightly weaker, since we do not need to require ξ_k, $k = 1, 2, \ldots$, to be independent.

While Nesterov's method has been generalized by Lan [14] to achieve the optimal rate of convergence for solving both smooth and nonsmooth convex SP problem, it is unclear whether it converges for nonconvex SP problems. On the other hand, although the RSG method ([12]) converges for nonconvex SP problems, it cannot achieve the optimal rate of convergence when applied to convex SP problems. Below, we present a new SA-type algorithm, namely, the randomized stochastic AG (RSAG) method which not only converges for nonconvex SP problems, but also achieves an optimal rate of convergence when applied to convex SP problems by properly specifying the stepsize policies.

The RSAG method is obtained by replacing the exact gradients in Algorithm 1 with the stochastic ones and incorporating a randomized termination criterion for nonconvex SP first studied in [12]. This algorithm is formally described as follows.

Algorithm 3 The randomized stochastic AG (RSAG) algorithm

Input: $x_0 \in \mathbb{R}^n$, $\{\alpha_k\}$ s.t. $\alpha_1 = 1$ and $\alpha_k \in (0,1)$ for any $k \geq 2$, $\{\beta_k > 0\}$ and $\{\lambda_k > 0\}$, iteration limit $N \geq 1$, and probability mass function $P_R(\cdot)$ s.t.

$$\text{Prob}\{R = k\} = p_k, \quad k = 1, \ldots, N. \tag{3.1}$$

0. Set $x_0^{ag} = x_0$ and $k = 1$. Let R be a random variable with probability mass function P_R.
1. Set x_k^{md} to (2.2).
2. Call the $\mathcal{SO}$ for computing $G(x_k^{md}, \xi_k)$ and set

$$x_k = x_{k-1} - \lambda_k G(x_k^{md}, \xi_k), \tag{3.2}$$

$$x_k^{ag} = x_k^{md} - \beta_k G(x_k^{md}, \xi_k). \tag{3.3}$$

3. If $k = R$, **terminate** the algorithm. Otherwise, set $k = k + 1$ and go to step 1.

We now add a few remarks about the above RSAG algorithm. First, similar to our discussion in the previous section, if $\alpha_k = 1, \quad \beta_k = \lambda_k \quad \forall k \geq 1$, then the above algorithm reduces to the classical SA algorithm. Moreover, if $\beta_k = \lambda_k \quad \forall k \geq 1$, the above algorithm reduces to the accelerated SA method in [14]. Second, we have used a random number R to terminate the above RSAG method for solving general (not necessarily convex) NLP problems. Equivalently, one can run the RSAG method for N iterations and then randomly select the search points (x_R^{md}, x_R^{ag}) as the output of Algorithm 3 from the trajectory (x_k^{md}, x_k^{ag}), $k = 1, \ldots, N$. Note, however, that the remaining $N - R$ iterations will be surplus.

We are now ready to describe the main convergence properties of the RSAG algorithm applied to problem (1.1) under the stochastic setting.

Theorem 3 *Let $\{x_k^{md}, x_k^{ag}\}_{k\geq 1}$ be computed by Algorithm 3 and Γ_k be defined in (2.6). Also suppose that Assumption 1 holds.*

a) If $\{\alpha_k\}$, $\{\beta_k\}$, $\{\lambda_k\}$, and $\{p_k\}$ are chosen such that (2.7) holds and

$$p_k = \frac{\lambda_k C_k}{\sum_{k=1}^N \lambda_k C_k}, \quad k = 1, \ldots, N, \tag{3.4}$$

where C_k is defined in (2.7), then for any $N \geq 1$, we have

$$\mathbb{E}[\|\nabla \Psi(x_R^{md})\|^2] \leq \frac{1}{\sum_{k=1}^N \lambda_k C_k} \left[\Psi(x_0) - \Psi^* + \frac{L_\Psi \sigma^2}{2} \sum_{k=1}^N \lambda_k^2 \left(1 + \frac{(\lambda_k - \beta_k)^2}{\alpha_k \Gamma_k \lambda_k^2} \sum_{\tau=k}^N \Gamma_\tau \right) \right], \tag{3.5}$$

where the expectation is taken with respect to R and $\xi_{[N]} := (\xi_1, ..., \xi_N)$.

b) Suppose that $\Psi(\cdot)$ is convex and that an optimal solution x^* exists for problem (1.1). If $\{\alpha_k\}$, $\{\beta_k\}$, $\{\lambda_k\}$, and $\{p_k\}$ are chosen such that (2.10) holds,

$$\alpha_k \lambda_k \leq L_\Psi \beta_k^2, \quad \beta_k < 1/L_\Psi, \tag{3.6}$$

and

$$p_k = \frac{\Gamma_k^{-1} \beta_k (1 - L_\Psi \beta_k)}{\sum_{k=1}^N \Gamma_k^{-1} \beta_k (1 - L_\Psi \beta_k)} \tag{3.7}$$

for all $k = 1, ..., N$, then for any $N \geq 1$, we have

$$\mathbb{E}[\|\nabla \Psi(x_R^{md})\|^2] \leq \frac{(2\lambda_1)^{-1} \|x_0 - x^*\|^2 + L_\Psi \sigma^2 \sum_{k=1}^N \Gamma_k^{-1} \beta_k^2}{\sum_{k=1}^N \Gamma_k^{-1} \beta_k (1 - L_\Psi \beta_k)}, \tag{3.8}$$

$$\mathbb{E}[\Psi(x_R^{ag}) - \Psi(x^*)] \leq \frac{\sum_{k=1}^N \beta_k (1 - L_\Psi \beta_k) \left[(2\lambda_1)^{-1} \|x_0 - x^*\|^2 + L_\Psi \sigma^2 \sum_{j=1}^k \Gamma_j^{-1} \beta_j^2 \right]}{\sum_{k=1}^N \Gamma_k^{-1} \beta_k (1 - L_\Psi \beta_k)}. \tag{3.9}$$

Proof. We first show part a). Denote $\delta_k := G(x_k^{md}, \xi_k) - \nabla \Psi(x_k^{md})$ and $\Delta_k := \nabla \Psi(x_{k-1}) - \nabla \Psi(x_k^{md})$. By (2.1) and (3.2), we have

$$\begin{aligned}
\Psi(x_k) &\leq \Psi(x_{k-1}) + \langle \nabla \Psi(x_{k-1}), x_k - x_{k-1} \rangle + \frac{L_\Psi}{2} \|x_k - x_{k-1}\|^2 \\
&= \Psi(x_{k-1}) + \langle \Delta_k + \nabla \Psi(x_k^{md}), -\lambda_k [\nabla \Psi(x_k^{md}) + \delta_k] \rangle + \frac{L_\Psi \lambda_k^2}{2} \|\nabla \Psi(x_k^{md}) + \delta_k\|^2 \\
&= \Psi(x_{k-1}) + \langle \Delta_k + \nabla \Psi(x_k^{md}), -\lambda_k \nabla \Psi(x_k^{md}) \rangle - \lambda_k \langle \nabla \Psi(x_{k-1}), \delta_k \rangle + \frac{L_\Psi \lambda_k^2}{2} \|\nabla \Psi(x_k^{md}) + \delta_k\|^2 \\
&\leq \Psi(x_{k-1}) - \lambda_k \left(1 - \frac{L_\Psi \lambda_k}{2} \right) \|\nabla \Psi(x_k^{md})\|^2 + \lambda_k \|\Delta_k\| \, \|\nabla \Psi(x_k^{md})\| + \frac{L_\Psi \lambda_k^2}{2} \|\delta_k\|^2 \\
&\quad - \lambda_k \langle \nabla \Psi(x_{k-1}) - L_\Psi \lambda_k \nabla \Psi(x_k^{md}), \delta_k \rangle,
\end{aligned}$$

which, in view of (2.13) and the fact that $ab \le (a^2+b^2)/2$, then implies that

$$\begin{aligned}\Psi(x_k) &\le \Psi(x_{k-1}) - \lambda_k\left(1-\frac{L_\Psi\lambda_k}{2}\right)\|\nabla\Psi(x_k^{md})\|^2 + \lambda_k L_\Psi(1-\alpha_k)\|x_{k-1}^{ag}-x_{k-1}\|\,\|\nabla\Psi(x_k^{md})\| \\ &\quad + \frac{L_\Psi\lambda_k^2}{2}\|\delta_k\|^2 - \lambda_k\langle\nabla\Psi(x_{k-1}) - L_\Psi\lambda_k\nabla\Psi(x_k^{md}),\delta_k\rangle \\ &\le \Psi(x_{k-1}) - \lambda_k(1-L_\Psi\lambda_k)\|\nabla\Psi(x_k^{md})\|^2 + \frac{L_\Psi(1-\alpha_k)^2}{2}\|x_{k-1}^{ag}-x_{k-1}\|^2 + \frac{L_\Psi\lambda_k^2}{2}\|\delta_k\|^2 \\ &\quad -\lambda_k\langle\nabla\Psi(x_{k-1}) - L_\Psi\lambda_k\nabla\Psi(x_k^{md}),\delta_k\rangle.\end{aligned}$$

Noting that similar to (2.17), we have

$$\begin{aligned}\|x_{k-1}^{ag}-x_{k-1}\|^2 &\le \Gamma_{k-1}\sum_{\tau=1}^{k-1}\frac{(\lambda_\tau-\beta_\tau)^2}{\Gamma_\tau\alpha_\tau}\|\nabla\Psi(x_\tau^{md})+\delta_k\|^2 \\ &= \Gamma_{k-1}\sum_{\tau=1}^{k-1}\frac{(\lambda_\tau-\beta_\tau)^2}{\Gamma_\tau\alpha_\tau}\left[\|\nabla\Psi(x_\tau^{md})\|^2+\|\delta_\tau\|^2+2\langle\nabla\Psi(x_\tau^{md}),\delta_\tau\rangle\right].\end{aligned}$$

Combining the previous two inequalities and using the fact that $\Gamma_{k-1}(1-\alpha_k)^2 \le \Gamma_k$, we obtain

$$\begin{aligned}\Psi(x_k) &\le \Psi(x_{k-1}) - \lambda_k(1-L_\Psi\lambda_k)\|\nabla\Psi(x_k^{md})\|^2 + \frac{L_\Psi\lambda_k^2}{2}\|\delta_k\|^2 - \lambda_k\langle\nabla\Psi(x_{k-1}) - L_\Psi\lambda_k\nabla\Psi(x_k^{md}),\delta_k\rangle \\ &\quad + \frac{L_\Psi\Gamma_k}{2}\sum_{\tau=1}^{k}\frac{(\lambda_\tau-\beta_\tau)^2}{\Gamma_\tau\alpha_\tau}\left[\|\nabla\Psi(x_\tau^{md})\|^2+\|\delta_\tau\|^2+2\langle\nabla\Psi(x_\tau^{md}),\delta_\tau\rangle\right].\end{aligned}$$

Summing up the above inequalities, we obtain

$$\begin{aligned}\Psi(x_N) &\le \Psi(x_0) - \sum_{k=1}^N\lambda_k(1-L_\Psi\lambda_k)\|\nabla\Psi(x_k^{md})\|^2 - \sum_{k=1}^N\lambda_k\langle\nabla\Psi(x_{k-1}) - L_\Psi\lambda_k\nabla\Psi(x_k^{md}),\delta_k\rangle \\ &\quad + \sum_{k=1}^N\frac{L_\Psi\lambda_k^2}{2}\|\delta_k\|^2 - \frac{L_\Psi}{2}\sum_{k=1}^N\Gamma_k\sum_{\tau=1}^k\frac{(\lambda_\tau-\beta_\tau)^2}{\Gamma_\tau\alpha_\tau}\left[\|\nabla\Psi(x_\tau^{md})\|^2+\|\delta_\tau\|^2+2\langle\nabla\Psi(x_\tau^{md}),\delta_\tau\rangle\right] \\ &= \Psi(x_0) - \sum_{k=1}^N\lambda_k C_k\|\nabla\Psi(x_k^{md})\|^2 + \frac{L_\Psi}{2}\sum_{k=1}^N\lambda_k^2\left(1+\frac{(\lambda_k-\beta_k)^2}{\alpha_k\Gamma_k\lambda_k^2}\sum_{\tau=k}^N\Gamma_\tau\right)\|\delta_k\|^2 - \sum_{k=1}^N b_k,\end{aligned}$$

where $b_k = \langle\lambda_k\nabla\Psi(x_{k-1}) - \left[L_\Psi\lambda_k^2 + \frac{L_\Psi(\lambda_k-\beta_k)^2}{\Gamma_k\alpha_k}\left(\sum_{\tau=k}^N\Gamma_\tau\right)\right]\nabla\Psi(x_k^{md}),\delta_k\rangle$. Taking expectation w.r.t. $\xi_{[N]}$ on both sides of the above inequality and noting that under Assumption 1, $\mathbb{E}[\|\delta_k\|^2] \le \sigma^2$ and $\{b_k\}$ is a martingale difference, we have

$$\sum_{k=1}^N\lambda_k C_k\mathbb{E}_{\xi_{[N]}}[\|\nabla\Psi(x_k^{md})\|^2] \le \Psi(x_0) - \Psi(x_N) + \frac{L_\Psi\sigma^2}{2}\sum_{k=1}^N\lambda_k^2\left(1+\frac{(\lambda_k-\beta_k)^2}{\alpha_k\Gamma_k\lambda_k^2}\sum_{\tau=k}^N\Gamma_\tau\right).$$

Dividing both sides of the above relation by $\sum_{k=1}^N\lambda_k C_k$, and using the facts that $\Psi(x_N) \ge \Psi^*$ and

$$\mathbb{E}[\|\nabla\Psi(x_R^{md})\|^2] = \mathbb{E}_{R,\xi_{[N]}}[\|\nabla\Psi(x_R^{md})\|^2] = \frac{\sum_{k=1}^N\lambda_k C_k\mathbb{E}_{\xi_{[N]}}[\|\nabla\Psi(x_k^{md})\|^2]}{\sum_{k=1}^N\lambda_k C_k},$$

we obtain (3.5).

We now show part b). By (2.1), (3.3), and (2.21), we have

$$\begin{aligned}\Psi(x_k^{ag}) &\le \Psi(x_k^{md}) + \langle \nabla\Psi(x_k^{md}), x_k^{ag} - x_k^{md}\rangle + \frac{L_\Psi}{2}\|x_k^{ag} - x_k^{md}\|^2 \\ &= \Psi(x_k^{md}) - \beta_k\|\nabla\Psi(x_k^{md})\|^2 + \beta\langle \nabla\Psi(x_k^{md}), \delta_k\rangle + \frac{L_\Psi\beta_k^2}{2}\|\nabla\Psi(x_k^{md}) + \delta_k\|^2 \\ &\le (1-\alpha_k)\Psi(x_{k-1}^{ag}) + \alpha_k\Psi(x) + \alpha_k\langle \nabla\Psi(x_k^{md}), x_{k-1} - x\rangle \\ &\quad -\beta_k\|\nabla\Psi(x_k^{md})\|^2 + \beta_k\langle \nabla\Psi(x_k^{md}), \delta_k\rangle + \frac{L_\Psi\beta_k^2}{2}\|\nabla\Psi(x_k^{md}) + \delta_k\|^2. \end{aligned} \tag{3.10}$$

Similar to (2.22), we have

$$\alpha_k\langle \nabla\Psi(x_k^{md}) + \delta_k, x_{k-1} - x\rangle = \frac{\alpha_k}{2\lambda_k}\left[\|x_{k-1} - x\|^2 - \|x_k - x\|^2\right] + \frac{\alpha_k\lambda_k}{2}\|\nabla\Psi(x_k^{md}) + \delta_k\|^2.$$

Combining the above two inequalities and using the fact that

$$\|\nabla\Psi(x_k^{md}) + \delta_k\|^2 = \|\nabla\Psi(x_k^{md})\|^2 + \|\delta_k\|^2 + 2\langle \nabla\Psi(x_k^{md}), \delta_k\rangle,$$

we obtain

$$\begin{aligned}\Psi(x_k^{ag}) &\le (1-\alpha_k)\Psi(x_{k-1}^{ag}) + \alpha_k\Psi(x) + \frac{\alpha_k}{2\lambda_k}\left[\|x_{k-1} - x\|^2 - \|x_k - x\|^2\right] \\ &\quad - \beta_k\left(1 - \frac{L_\Psi\beta_k}{2} - \frac{\alpha_k\lambda_k}{2\beta_k}\right)\|\nabla\Psi(x_k^{md})\|^2 + \left(\frac{L_\Psi\beta_k^2 + \alpha_k\lambda_k}{2}\right)\|\delta_k\|^2 \\ &\quad + \langle \delta_k, (\beta_k + L_\Psi\beta_k^2 + \alpha_k\lambda_k)\nabla\Psi(x_k^{md}) + \alpha_k(x - x_{k-1})\rangle.\end{aligned}$$

Subtracting $\Psi(x)$ from both sides of the above inequality, and using Lemma 1 and (2.25), we have

$$\begin{aligned}\frac{\Psi(x_N^{ag}) - \Psi(x)}{\Gamma_N} &\le \frac{\|x_0 - x\|^2}{2\lambda_1} - \sum_{k=1}^N \frac{\beta_k}{2\Gamma_k}\left(2 - L_\Psi\beta_k - \frac{\alpha_k\lambda_k}{\beta_k}\right)\|\nabla\Psi(x_k^{md})\|^2 \\ &\quad + \sum_{k=1}^N\left(\frac{L_\Psi\beta_k^2 + \alpha_k\lambda_k}{2\Gamma_k}\right)\|\delta_k\|^2 + \sum_{k=1}^N b_k' \quad \forall x \in \mathbb{R}^n,\end{aligned}$$

where $b_k' = \Gamma_k^{-1}\langle \delta_k, (\beta_k + L_\Psi\beta_k^2 + \alpha_k\lambda_k)\nabla\Psi(x_k^{md}) + \alpha_k(x - x_{k-1})\rangle$. The above inequality together with the first relation in (3.6) then imply that

$$\begin{aligned}\frac{\Psi(x_N^{ag}) - \Psi(x)}{\Gamma_N} &\le \frac{\|x_0 - x\|^2}{2\lambda_1} - \sum_{k=1}^N \frac{\beta_k}{\Gamma_k}\left(1 - L_\Psi\beta_k\right)\|\nabla\Psi(x_k^{md})\|^2 \\ &\quad + \sum_{k=1}^N \frac{L_\Psi\beta_k^2}{\Gamma_k}\|\delta_k\|^2 + \sum_{k=1}^N b_k' \quad \forall x \in \mathbb{R}^n.\end{aligned}$$

Taking expectation (with respect to $\xi_{[N]}$) on both sides of the above relation, and noting that under Assumption 1, $\mathbb{E}[\|\delta_k\|^2] \le \sigma^2$ and $\{b_k'\}$ is a martingale difference, we obtain, $\forall x \in \mathbb{R}^n$,

$$\frac{1}{\Gamma_N}\mathbb{E}_{\xi_{[N]}}[\Psi(x_N^{ag}) - \Psi(x)] \le \frac{\|x_0 - x\|^2}{2\lambda_1} - \sum_{k=1}^N \frac{\beta_k}{\Gamma_k}\left(1 - L_\Psi\beta_k\right)\mathbb{E}_{\xi_{[N]}}[\|\nabla\Psi(x_k^{md})\|^2] + \sigma^2\sum_{k=1}^N \frac{L_\Psi\beta_k^2}{\Gamma_k}. \tag{3.11}$$

Now, fixing $x = x^*$ and noting that $\Psi(x_N^{ag}) \ge \Psi(x^*)$, we have

$$\sum_{k=1}^N \frac{\beta_k}{\Gamma_k}\left(1 - L_\Psi\beta_k\right)\mathbb{E}_{\xi_{[N]}}[\|\nabla\Psi(x_k^{md})\|^2] \le \frac{\|x_0 - x^*\|^2}{2\lambda_1} + \sigma^2\sum_{k=1}^N \frac{L_\Psi\beta_k^2}{\Gamma_k},$$

which, in view of the definition of x_R^{md}, then implies (3.8). It also follows from (3.11) and (3.6) that, for any $N \geq 1$,

$$\mathbb{E}_{\xi_{[N]}}[\Psi(x_N^{ag}) - \Psi(x^*)] \leq \Gamma_N \left(\frac{\|x_0 - x\|^2}{2\lambda_1} + \sigma^2 \sum_{k=1}^{N} \frac{L_\Psi \beta_k^2}{\Gamma_k} \right),$$

which, in view of the definition of x_R^{ag}, then implies that

$$\begin{aligned} \mathbb{E}[\Psi(x_R^{ag}) - \Psi(x^*)] &= \sum_{k=1}^{N} \frac{\Gamma_k^{-1}\beta_k(1 - L_\Psi\beta_k)}{\sum_{k=1}^{N} \Gamma_k^{-1}\beta_k(1 - L_\Psi\beta_k)} \mathbb{E}_{\xi_{[N]}}[\Psi(x_k^{ag}) - \Psi(x^*)] \\ &\leq \frac{\sum_{k=1}^{N} \beta_k(1 - L_\Psi\beta_k)\left[(2\lambda_1)^{-1}\|x_0 - x\|^2 + L_\Psi\sigma^2 \sum_{j=1}^{k} \Gamma_j^{-1}\beta_j^2\right]}{\sum_{k=1}^{N} \Gamma_k^{-1}\beta_k(1 - L_\Psi\beta_k)}. \end{aligned}$$

■

We now add a few remarks about the results obtained in Theorem 3. First, note that similar to the deterministic case, we can use the assumption in (2.26) instead of the one in (3.6). Second, the expectations in (3.5), (3.8), and (3.9) are taken with respect to one more random variable R in addition to ξ coming from the $\mathcal{SO}$. Specifically, the output of the Algorithm 3 is chosen randomly from the generated trajectory $\{(x_1^{md}, x_1^{ag}), \ldots, (x_N^{md}, x_N^{ag})\}$ according to (3.1), as mentioned earlier in this subsection. Third, the probabilities $\{p_k\}$ depend on the choice of $\{\alpha_k\}$, $\{\beta_k\}$, and $\{\lambda_k\}$.

Below, we specialize the results obtained in Theorem 3 for some particular selections of $\{\alpha_k\}$, $\{\beta_k\}$, and $\{\lambda_k\}$.

Corollary 3 *The following statements hold for Algorithm 3 when applied to problem (1.1) under Assumption 1.*

a) If $\{\alpha_k\}$ and $\{\lambda_k\}$ in the RSAG method are set to (2.27) and (2.28), respectively, $\{p_k\}$ is set to (3.4), $\{\beta_k\}$ is set to

$$\beta_k = \min\left\{ \frac{8}{21L_\Psi}, \frac{\tilde{D}}{\sigma\sqrt{N}} \right\}, \quad k \geq 1 \tag{3.12}$$

for some $\tilde{D} > 0$, and an iteration limit $N \geq 1$ is given, then we have

$$\mathbb{E}[\|\nabla\Psi(x_R^{md})\|^2] \leq \frac{21L_\Psi[\Psi(x_0) - \Psi^*]}{4N} + \frac{2\sigma}{\sqrt{N}}\left(\frac{\Psi(x_0) - \Psi^*}{\tilde{D}} + L_\Psi\tilde{D} \right) =: \mathcal{U}_N. \tag{3.13}$$

b) Assume that $\Psi(\cdot)$ is convex and that an optimal solution x^* exists for problem (1.1). If $\{\alpha_k\}$ is set to (2.27), $\{p_k\}$ is set to (3.7), $\{\beta_k\}$ and $\{\lambda_k\}$ are set to

$$\beta_k = \min\left\{ \frac{1}{2L_\Psi}, \left(\frac{\tilde{D}^2}{L_\Psi^2\sigma^2N^3} \right)^{\frac{1}{4}} \right\} \tag{3.14}$$

$$\text{and} \quad \lambda_k = \frac{kL_\Psi\beta_k^2}{2}, \quad k \geq 1, \tag{3.15}$$

for some $\tilde{D} > 0$, and an iteration limit $N \geq 1$ is given, then we have

$$\mathbb{E}[\|\nabla\Psi(x_R^{md})\|^2] \leq \frac{96L_\Psi^2\|x_0 - x^*\|^2}{N^3} + \frac{L_\Psi^{\frac{1}{2}}\sigma^{\frac{3}{2}}}{N^{\frac{3}{4}}}\left(\frac{12\|x_0 - x^*\|^2}{\tilde{D}^{\frac{3}{2}}} + 2\tilde{D}^{\frac{1}{2}} \right), \tag{3.16}$$

$$\mathbb{E}[\Psi(x_R^{ag}) - \Psi(x^*)] \leq \frac{48L_\Psi\|x_0 - x^*\|^2}{N^2} + \frac{12\sigma}{\sqrt{N}}\left(\frac{\|x_0 - x^*\|^2}{\tilde{D}} + \tilde{D} \right). \tag{3.17}$$

Proof. We first show part a). It follows from (2.28), (2.35), and (3.12) that

$$C_k \geq 1 - \frac{21}{16}L_\Psi\beta_k \geq \frac{1}{2} > 0 \quad \text{and} \quad \lambda_k C_k \geq \frac{\beta_k}{2}.$$

Also by (2.28), (2.33), (2.34), and (3.12), we have

$$\lambda_k^2\left[1+\frac{(\lambda_k-\beta_k)^2}{\alpha_k\Gamma_k\lambda_k^2}\left(\sum_{\tau=k}^N\Gamma_\tau\right)\right]\le\lambda_k^2\left[1+\frac{1}{\alpha_k\Gamma_k\lambda_k^2}\left(\frac{\alpha_k\beta_k}{4}\right)^2\frac{2}{k}\right]=\lambda_k^2+\frac{\beta_k^2}{8}$$
$$\le\left[\left(1+\frac{\alpha_k}{4}\right)^2+\frac{1}{8}\right]\beta_k^2\le 2\beta_k^2$$

for any $k\ge 1$. These observations together with (3.5) then imply that

$$\mathbb{E}[\|\nabla\Psi(x_R^{md})\|^2]\le\frac{2}{\sum_{k=1}^N\beta_k}\left(\Psi(x_0)-\Psi^*+L_\Psi\sigma^2\sum_{k=1}^N\beta_k^2\right)$$
$$\le\frac{2[\Psi(x_0)-\Psi^*]}{N\beta_1}+2L_\Psi\sigma^2\beta_1$$
$$\le\frac{2[\Psi(x_0)-\Psi^*]}{N}\left\{\frac{21L_\Psi}{8}+\frac{\sigma\sqrt{N}}{\tilde D}\right\}+\frac{2L_\Psi\tilde D\sigma}{\sqrt{N}},$$

which implies (3.12).

We now show part b). It can be easily checked that (2.10) and (3.6) hold in view of (3.14) and (3.15). By (2.33) and (3.14), we have

$$\sum_{k=1}^N\Gamma_k^{-1}\beta_k(1-L_\Psi\beta_k)\ge\frac{1}{2}\sum_{k=1}^N\Gamma_k^{-1}\beta_k=\frac{\beta_1}{2}\sum_{k=1}^N\Gamma_k^{-1},\tag{3.18}$$
$$\sum_{k=1}^N\Gamma_k^{-1}\ge\sum_{k=1}^N\frac{k^2}{2}=\frac{1}{12}N(N+1)(2N+1)\ge\frac{1}{6}N^3.\tag{3.19}$$

Using these observations, (2.33), (3.8), (3.14), and (3.15), we have

$$\mathbb{E}[\|\nabla\Psi(x_R^{md})\|^2]\le\frac{2}{\beta_1\sum_{k=1}^N\Gamma_k^{-1}}\left(\frac{\|x_0-x^*\|^2}{L_\Psi\beta_1^2}+L_\Psi\sigma^2\beta_1^2\sum_{k=1}^N\Gamma_k^{-1}\right)$$
$$=\frac{2\|x_0-x^*\|^2}{L_\Psi\beta_1^3\sum_{k=1}^N\Gamma_k^{-1}}+2L_\Psi\sigma^2\beta_1\le\frac{12\|x_0-x^*\|^2}{L_\Psi N^3\beta_1^3}+2L_\Psi\sigma^2\beta_1$$
$$\le\frac{96L_\Psi^2\|x_0-x^*\|^2}{N^3}+\frac{L_\Psi^{\frac{1}{2}}\sigma^{\frac{3}{2}}}{N^{\frac{3}{4}}}\left(\frac{12\|x_0-x^*\|^2}{\tilde D^{\frac{3}{2}}}+2\tilde D^{\frac{1}{2}}\right).$$

Also observe that by (2.33) and (3.14), we have

$$1-L_\Psi\beta_k\le 1\ \text{ and }\ \sum_{j=1}^k\Gamma_j^{-1}=\frac{1}{2}\sum_{j=1}^k j(j+1)\le\sum_{j=1}^k j^2\le k^3$$

for any $k\ge 1$. Using these observations, (3.9), (3.14), (3.18), and (3.19), we obtain

$$\mathbb{E}[\Psi(x_R^{ag})-\Psi(x^*)]\le\frac{2}{\sum_{k=1}^N\Gamma_k^{-1}}\left[N(2\lambda_1)^{-1}\|x_0-x^*\|^2+L_\Psi\sigma^2\beta_1^2\sum_{k=1}^N k^3\right]$$
$$\le\frac{12\|x_0-x^*\|^2}{N^2L_\Psi\beta_1^2}+\frac{12L_\Psi\sigma^2\beta_1^2}{N^3}\sum_{k=1}^N k^3$$
$$\le\frac{12\|x_0-x^*\|^2}{N^2L_\Psi\beta_1^2}+12L_\Psi\sigma^2\beta_1^2N$$
$$\le\frac{48L_\Psi\|x_0-x^*\|^2}{N^2}+\frac{12\sigma}{N^{\frac{1}{2}}}\left(\frac{\|x_0-x^*\|^2}{\tilde D}+\tilde D\right).$$

■

We now add a few remarks about the results obtained in Corollary 3. First, note that, the stepsizes $\{\beta_k\}$ in the above corollary depend on the parameter $\tilde{D}$. While the RSAG method converges for any $\tilde{D} > 0$, by minimizing the RHS of (3.13) and (3.17), the optimal choices of $\tilde{D}$ would be $\sqrt{[\Psi(x_0^{ag}) - \Psi(x^*)]/L_\Psi}$ and $\|x_0 - x^*\|$, respectively, for solving nonconvex and convex smooth SP problems. With such selections for $\tilde{D}$, the bounds in (3.13), (3.16), and (3.17), respectively, reduce to

$$\mathbb{E}[\|\nabla\Psi(x_R^{md})\|^2] \le \frac{21L_\Psi[\Psi(x_0) - \Psi^*]}{4N} + \frac{4\sigma[L_\Psi(\Psi(x_0) - \Psi^*)]^{\frac{1}{2}}}{\sqrt{N}}, \tag{3.20}$$

$$\mathbb{E}[\|\nabla\Psi(x_R^{md})\|^2] \le \frac{96L_\Psi^2\|x_0 - x^*\|^2}{N^3} + \frac{14(L_\Psi\|x_0 - x^*\|)^{\frac{1}{2}}\sigma^{\frac{3}{2}}}{N^{\frac{3}{4}}}, \tag{3.21}$$

and

$$\mathbb{E}[\Psi(x_R^{ag}) - \Psi(x^*)] \le \frac{48L_\Psi\|x_0 - x^*\|^2}{N^2} + \frac{24\|x_0 - x^*\|\sigma}{\sqrt{N}}. \tag{3.22}$$

Second, the rate of convergence of the RSAG algorithm in (3.13) for general nonconvex problems is the same as that of the RSG method [12] for smooth nonconvex SP problems. However, if the problem is convex, then the complexity of the RSAG algorithm will be significantly better than the latter algorithm. More specifically, in view of (3.22), the RSAG is an optimal method for smooth stochastic optimization [14], while the rate of convergence of the RSG method is only nearly optimal. Moreover, in view of (3.16), if $\Psi(\cdot)$ is convex, then the number of iterations performed by the RSAG algorithm to find an ϵ-solution of (1.1), i.e., a point $\bar{x}$ such that $\mathbb{E}[\|\nabla\Psi(\bar{x})\|^2] \le \epsilon$, can be bounded by

$$\mathcal{O}\left\{\left(\frac{1}{\epsilon^{\frac{1}{3}}} + \frac{\sigma^2}{\epsilon^{\frac{4}{3}}}\right)(L_\Psi\|x_0 - x^*\|)^{\frac{2}{3}}\right\}.$$

To the best of our knowledge, this complexity result seems to be new in the literature.

In addition to the aforementioned expected complexity results of the RSAG method, we can establish their associated large deviation properties. For example, by Markov's inequality and (3.13), we have

$$\text{Prob}\left\{\|\nabla\Psi(x_R^{md})\|^2 \ge \lambda\mathcal{U}_N\right\} \le \frac{1}{\lambda} \quad \forall\lambda > 0, \tag{3.23}$$

which implies that the total number of calls to the $\mathcal{SO}$ performed by the RSAG method for finding an (ϵ, Λ)-*solution* of problem (1.1), i.e., a point $\bar{x}$ satisfying $\text{Prob}\{\|\nabla\Psi(\bar{x})\|^2 \le \epsilon\} \ge 1 - \Lambda$ for some $\epsilon > 0$ and $\Lambda \in (0, 1)$, after disregarding a few constant factors, can be bounded by

$$\mathcal{O}\left\{\frac{1}{\Lambda\epsilon} + \frac{\sigma^2}{\Lambda^2\epsilon^2}\right\}. \tag{3.24}$$

To improve the dependence of the above bound on the confidence level Λ, we can design a variant of the RSAG method which has two phases: optimization and post-optimization phase. The optimization phase consists of independent runs of the RSAG method to generate a list of candidate solutions and the post-optimization phase then selects a solution from the generated candidate solutions in the optimization phase (see [12, Subsection 2.2] for more details).

3.2 Minimization of nonconvex stochastic composite functions

In this subsection, we consider the stochastic composite problem (1.3), which satisfies both Assumptions 1 and 2. Our goal is to show that under the above assumptions, we can choose the same aggressive stepsize policy in the RSAG method no matter if the objective function $\Psi(\cdot)$ in (1.3) is convex or not.

We will modify the RSAG method in Algorithm 3 by replacing the stochastic gradient $\nabla\Psi(x_k^{md}, \xi_k)$ with

$$\bar{G}_k = \frac{1}{m_k}\sum_{i=1}^{m_k} G(x_k^{md}, \xi_{k,i}) \tag{3.25}$$

for some $m_k \geq 1$, where $G(x_k^{md}, \xi_{k,i}), i = 1, \ldots, m_k$ are the stochastic gradients returned by the i-th call to the $\mathcal{SO}$ at iteration k. Such a mini-batch approach has been used for nonconvex stochastic composite optimization in [13,6]. The modified RSAG algorithm is formally described as follows.

Algorithm 4 The RSAG algorithm for stochastic composite optimization

Replace (3.2) and (3.3), respectively, in Step 2 of Algorithm 3 by

$$x_k = \mathcal{P}(x_{k-1}, \bar{G}_k, \lambda_k), \tag{3.26}$$
$$x_k^{ag} = \mathcal{P}(x_k^{md}, \bar{G}_k, \beta_k), \tag{3.27}$$

where $\bar{G}_k$ is defined in (3.25) for some $m_k \geq 1$.

A few remarks about the above RSAG algorithm are in place. First, note that by calling the $\mathcal{SO}$ multiple times at each iteration, we can obtain a better estimator for $\nabla\Psi(x_k^{md})$ than the one obtained by using one call to the $\mathcal{SO}$ as in Algorithm 3. More specifically, under Assumption 1, we have

$$\begin{aligned} \mathbb{E}[\bar{G}_k] &= \frac{1}{m_k}\sum_{i=1}^{m_k}\mathbb{E}[G(x_k^{md}, \xi_{k,i})] = \nabla\Psi(x_k^{md}), \\ \mathbb{E}[\|\bar{G}_k - \nabla\Psi(x_k^{md})\|^2] &= \frac{1}{m_k^2}\mathbb{E}\left[\|\sum_{i=1}^{m_k}[G(x_k^{md}, \xi_{k,i}) - \nabla\Psi(x_k^{md})]\|^2\right] \leq \frac{\sigma^2}{m_k}, \end{aligned} \tag{3.28}$$

where the last inequality follows from [13, p.11]. Thus, by increasing m_k, we can decrease the error existing in the estimation of $\nabla\Psi(x_k^{md})$. We will discuss the appropriate choice of m_k later in this subsection. Second, since we do not have access to $\nabla\Psi(x_k^{md})$, we cannot compute the exact gradient mapping, i.e., $\mathcal{G}(x_k^{md}, \nabla\Psi(x_k^{md}), \beta_k)$ as the one used in Subsection 2.2 for composite optimization. However, by (2.38) and (3.26), we can compute an approximate stochastic gradient mapping given by $\mathcal{G}(x_k^{md}, \bar{G}_k, \beta_k)$. Indeed, by Lemma 4 and (3.28) , we have

$$\mathbb{E}[\|\mathcal{G}(x_k^{md}, \nabla\Psi(x_k^{md}), \beta_k) - \mathcal{G}(x_k^{md}, \bar{G}_k, \beta_k)\|^2] \leq \mathbb{E}[\|\bar{G}_k - \nabla\Psi(x_k^{md})\|^2] \leq \frac{\sigma^2}{m_k}. \tag{3.29}$$

We are ready to describe the main convergence properties of Algorithm 4 for solving nonconvex stochastic composite problems.

Theorem 4 *Suppose that $\{\alpha_k\}$, $\{\beta_k\}$, $\{\lambda_k\}$, and $\{p_k\}$ in Algorithm 4 satisfy (2.9), (2.10), and (3.7). Then under Assumptions 1 and 2, we have*

$$\begin{aligned} \mathbb{E}[\|\mathcal{G}(x_R^{md}, \nabla\Psi(x_R^{md}), \beta_R)\|^2] \leq 8\left[\sum_{k=1}^{N}\Gamma_k^{-1}\beta_k(1 - L_\Psi\beta_k)\right]^{-1}\Bigg[\frac{\|x_0 - x^*\|^2}{2\lambda_1} + \frac{L_f}{\Gamma_N}(\|x^*\|^2 + 2M^2) \\ + \sigma^2\sum_{k=1}^{N}\frac{\beta_k\left(4 + (1 - L_\Psi\beta_k)^2\right)}{4\Gamma_k(1 - L_\Psi\beta_k)m_k}\Bigg], \end{aligned} \tag{3.30}$$

where the expectation is taken with respect to R and $\xi_{k,i}$, $k = 1, .., N$, $i = 1, ..., m_k$. If, in addition, $L_f = 0$, then we have

$$\begin{aligned} \mathbb{E}[\Phi(x_R^{ag}) - \Phi(x^*)] \leq \left[\sum_{k=1}^{N}\Gamma_k^{-1}\beta_k(1 - L_\Psi\beta_k)\right]^{-1}\Bigg[\sum_{k=1}^{N}\beta_k(1 - L_\Psi\beta_k)\Bigg(\frac{\|x_0 - x^*\|^2}{2\lambda_1} \\ + \sigma^2\sum_{j=1}^{k}\frac{\beta_j(4 + (1 - L_\Psi\beta_j)^2)}{4\Gamma_j(1 - L_\Psi\beta_j)m_j}\Bigg)\Bigg], \end{aligned} \tag{3.31}$$

where $\Phi(x) \equiv \Psi(x) + \mathcal{X}(x)$.

Proof. Denoting $\bar{\delta}_k \equiv \bar{G}_k - \nabla\Psi(x_k^{md})$ and $\bar{\delta}_{[k]} \equiv \{\bar{\delta}_1, \ldots, \bar{\delta}_k\}$ for any $k \geq 1$, and using Lemma 2 of [11] for the solutions of subproblems (3.26) and (3.27), we have

$$\langle \nabla\Psi(x_k^{md}) + \bar{\delta}_k, x_k - x\rangle + \mathcal{X}(x_k) \leq \mathcal{X}(x) + \frac{1}{2\lambda_k}\left[\|x_{k-1} - x\|^2 - \|x_k - x\|^2 - \|x_k - x_{k-1}\|^2\right], \tag{3.32}$$

$$\langle \nabla\Psi(x_k^{md}) + \bar{\delta}_k, x_k^{ag} - x\rangle + \mathcal{X}(x_k^{ag}) \leq \mathcal{X}(x) + \frac{1}{2\beta_k}\left[\|x_k^{md} - x\|^2 - \|x_k^{ag} - x\|^2 - \|x_k^{ag} - x_k^{md}\|^2\right] \tag{3.33}$$

for any $x \in \mathbb{R}^n$. Letting $x = \alpha_k x_k + (1-\alpha_k)x_{k-1}^{ag}$ in (3.33), we have

$$\begin{aligned}
&\langle \nabla\Psi(x_k^{md}) + \bar{\delta}_k, x_k^{ag} - \alpha_k x_k - (1-\alpha_k)x_{k-1}^{ag}\rangle + \mathcal{X}(x_k^{ag}) \\
&\quad \leq \mathcal{X}(\alpha_k x_k + (1-\alpha_k)x_{k-1}^{ag}) + \frac{1}{2\beta_k}\left[\|x_k^{md} - \alpha_k x_k - (1-\alpha_k)x_{k-1}^{ag}\|^2 - \|x_k^{ag} - x_k^{md}\|^2\right] \\
&\quad \leq \alpha_k \mathcal{X}(x_k) + (1-\alpha_k)\mathcal{X}(x_{k-1}^{ag}) + \frac{1}{2\beta_k}\left[\alpha_k^2\|x_k - x_{k-1}\|^2 - \|x_k^{ag} - x_k^{md}\|^2\right],
\end{aligned}$$

where the last inequality follows from the convexity of $\mathcal{X}$ and (2.2). Summing up the above inequality with (3.32) (with both sides multiplied by α_k), we obtain

$$\begin{aligned}
&\langle \nabla\Psi(x_k^{md}) + \bar{\delta}_k, x_k^{ag} - \alpha_k x - (1-\alpha_k)x_{k-1}^{ag}\rangle + \mathcal{X}(x_k^{ag}) \leq (1-\alpha_k)\mathcal{X}(x_{k-1}^{ag}) + \alpha_k\mathcal{X}(x) \\
&+\frac{\alpha_k}{2\lambda_k}\left[\|x_{k-1} - x\|^2 - \|x_k - x\|^2\right] + \frac{\alpha_k(\lambda_k\alpha_k - \beta_k)}{2\beta_k\lambda_k}\|x_k - x_{k-1}\|^2 - \frac{1}{2\beta_k}\|x_k^{ag} - x_k^{md}\|^2 \\
&\leq (1-\alpha_k)\mathcal{X}(x_{k-1}^{ag}) + \alpha_k\mathcal{X}(x) + \frac{\alpha_k}{2\lambda_k}\left[\|x_{k-1} - x\|^2 - \|x_k - x\|^2\right] - \frac{1}{2\beta_k}\|x_k^{ag} - x_k^{md}\|^2,
\end{aligned} \tag{3.34}$$

where the last inequality follows from the assumption that $\alpha_k\lambda_k \leq \beta_k$. Combining the above relation with (2.46) and (2.47), and using the definition $\Phi(x) \equiv \Psi(x) + \mathcal{X}(x)$, we have

$$\begin{aligned}
\Phi(x_k^{ag}) &\leq (1-\alpha_k)\Phi(x_{k-1}^{ag}) + \alpha_k\Phi(x) - \frac{1}{2}\left(\frac{1}{\beta_k} - L_\Psi\right)\|x_k^{ag} - x_k^{md}\|^2 + \langle \bar{\delta}_k, \alpha_k(x - x_{k-1}) + x_k^{md} - x_k^{ag}\rangle \\
&\quad + \frac{\alpha_k}{2\lambda_k}\left[\|x_{k-1} - x\|^2 - \|x_k - x\|^2\right] + \frac{L_f\alpha_k}{2}\|x_k^{md} - x\|^2 + \frac{L_f\alpha_k^2(1-\alpha_k)}{2}\|x_{k-1}^{ag} - x_{k-1}\|^2 \\
&\leq (1-\alpha_k)\Phi(x_{k-1}^{ag}) + \alpha_k\Phi(x) + \langle \bar{\delta}_k, \alpha_k(x - x_{k-1})\rangle - \frac{1}{4}\left(\frac{1}{\beta_k} - L_\Psi\right)\|x_k^{ag} - x_k^{md}\|^2 + \frac{\beta_k\|\bar{\delta}_k\|^2}{1 - L_\Psi\beta_k} \\
&\quad + \frac{\alpha_k}{2\lambda_k}\left[\|x_{k-1} - x\|^2 - \|x_k - x\|^2\right] + \frac{L_f\alpha_k}{2}\|x_k^{md} - x\|^2 + \frac{L_f\alpha_k^2(1-\alpha_k)}{2}\|x_{k-1}^{ag} - x_{k-1}\|^2,
\end{aligned}$$

where the last inequality follows from the Young's inequality. Subtracting $\Phi(x)$ from both sides of the above inequality, re-arranging the terms, and using Lemma 1 and (2.25), we obtain

$$\begin{aligned}
&\frac{\Phi(x_N^{ag}) - \Phi(x)}{\Gamma_N} + \sum_{k=1}^N \frac{1 - L_\Psi\beta_k}{4\beta_k\Gamma_k}\|x_k^{ag} - x_k^{md}\|^2 \leq \frac{\|x_0 - x\|^2}{2\lambda_1} + \sum_{k=1}^N \frac{\alpha_k}{\Gamma_k}\langle \bar{\delta}_k, x - x_{k-1}\rangle \\
&+ \frac{L_f}{2}\sum_{k=1}^N \frac{\alpha_k}{\Gamma_k}[\|x_k^{md} - x\|^2 + \alpha_k(1-\alpha_k)\|x_{k-1}^{ag} - x_{k-1}\|^2] + \sum_{k=1}^N \frac{\beta_k\|\bar{\delta}_k\|^2}{\Gamma_k(1 - L_\Psi\beta_k)} \quad \forall x \in \mathbb{R}^n.
\end{aligned}$$

Letting $x = x^*$ in the above inequality, and using (2.16) and (2.52), we have

$$\begin{aligned}
\frac{\Phi(x_N^{ag}) - \Phi(x^*)}{\Gamma_N} + \sum_{k=1}^N \frac{1 - L_\Psi\beta_k}{4\beta_k\Gamma_k}\|x_k^{ag} - x_k^{md}\|^2 &\leq \frac{\|x_0 - x^*\|^2}{2\lambda_1} + \sum_{k=1}^N \frac{\alpha_k}{\Gamma_k}\langle \bar{\delta}_k, x^* - x_{k-1}\rangle \\
&\quad + \frac{L_f}{\Gamma_N}(\|x^*\|^2 + 2M^2) + \sum_{k=1}^N \frac{\beta_k\|\bar{\delta}_k\|^2}{\Gamma_k(1 - L_\Psi\beta_k)}.
\end{aligned}$$

Taking expectation from both sides of the above inequality, noting that under Assumption 1, $\mathbb{E}[\langle \bar{\delta}_k, x^* - x_{k-1}\rangle | \bar{\delta}_{[k-1]}] = 0$, and using (3.28) and the definition of the gradient mapping in (2.38), we conclude

$$\frac{\mathbb{E}_{\bar{\delta}_{[N]}}[\Phi(x_N^{ag}) - \Phi(x^*)]}{\Gamma_N} + \sum_{k=1}^{N} \frac{\beta_k\,[1 - L_\Psi \beta_k]}{4\Gamma_k} \mathbb{E}_{\bar{\delta}_{[N]}}[\|\mathcal{G}(x_k^{md}, \bar{G}_k, \beta_k)\|^2$$
$$\leq \frac{\|x_0 - x^*\|^2}{2\lambda_1} + \frac{L_f}{\Gamma_N}(\|x^*\|^2 + 2M^2) + \sigma^2 \sum_{k=1}^{N} \frac{\beta_k}{\Gamma_k(1 - L_\Psi \beta_k) m_k},$$

which, together with the fact that $\mathbb{E}_{\bar{\delta}_{[N]}}[\|\mathcal{G}(x_k^{md}, \nabla\Psi(x_k^{md}), \beta_k)\|^2] \leq 2(\mathbb{E}_{\bar{\delta}_{[N]}}[\|\mathcal{G}(x_k^{md}, \bar{G}_k, \beta_k)\|^2] + \sigma^2/m_k)$ due to (3.29), then imply that

$$\frac{\mathbb{E}_{\bar{\delta}_{[N]}}[\Phi(x_N^{ag}) - \Phi(x)]}{\Gamma_N} + \sum_{k=1}^{N} \frac{\beta_k(1 - L_\Psi \beta_k)}{8\Gamma_k} \mathbb{E}_{\bar{\delta}_{[N]}}[\|\mathcal{G}(x_k^{md}, \nabla\Psi(x_k^{md}), \beta_k)\|^2$$
$$\leq \frac{\|x_0 - x^*\|^2}{2\lambda_1} + \frac{L_f}{\Gamma_N}(\|x^*\|^2 + 2M^2) + \sigma^2 \left(\sum_{k=1}^{N} \frac{\beta_k}{\Gamma_k(1 - L_\Psi \beta_k) m_k} + \sum_{k=1}^{N} \frac{\beta_k(1 - L_\Psi \beta_k)}{4\Gamma_k m_k} \right)$$
$$= \frac{\|x_0 - x^*\|^2}{2\lambda_1} + \frac{L_f}{\Gamma_N}(\|x^*\|^2 + 2M^2) + \sigma^2 \sum_{k=1}^{N} \frac{\beta_k \left[4 + (1 - L_\Psi \beta_k)^2\right]}{4\Gamma_k(1 - L_\Psi \beta_k) m_k}. \tag{3.35}$$

Since the above relation is similar to the relation (3.11), the rest of proof is also similar to the last part of the proof for Theorem 3 and hence the details are skipped. ∎

Theorem 4 shows that by using the RSAG method in Algorithm 4, we can have a unified treatment and analysis for stochastic composite problem (1.3), no matter it is convex or not. In the next result, we specialize the results obtained in Theorem 4 for some particular selections of $\{\alpha_k\}$, $\{\beta_k\}$, and $\{\lambda_k\}$.

Corollary 4 *Suppose that the stepsizes $\{\alpha_k\}$, $\{\beta_k\}$, and $\{\lambda_k\}$ in Algorithm 4 are set to (2.27) and (2.30), respectively, and $\{p_k\}$ is set to (3.7). Also assume that an optimal solution x^* exists for problem (1.3). Then under Assumptions 1 and 2, for any $N \geq 1$, we have*

$$\mathbb{E}[\|\mathcal{G}(x_R^{md}, \nabla\Psi(x_R^{md}), \beta_R)\|^2] \leq 96 L_\Psi \left[\frac{4L_\Psi \|x_0 - x^*\|^2}{N^2(N+1)} + \frac{L_f}{N}(\|x^*\|^2 + 2M^2) + \frac{3\sigma^2}{L_\Psi N^3} \sum_{k=1}^{N} \frac{k^2}{m_k} \right]. \tag{3.36}$$

If, in addition, $L_f = 0$, then for any $N \geq 1$, we have

$$\mathbb{E}[\Phi(x_R^{ag}) - \Phi(x^*)] \leq \frac{12 L_\Psi \|x_0 - x^*\|^2}{N(N+1)} + \frac{7\sigma^2}{L_\Psi N^3} \sum_{k=1}^{N} \sum_{j=1}^{k} \frac{j^2}{m_j}. \tag{3.37}$$

Proof. Similar to Corollary 1.b), we can easily show that (2.9) and (2.10) hold. By (3.30), (2.27), (2.30), (2.33), and (2.36), we have

$$\mathbb{E}[\|\mathcal{G}(x_R^{md}, \nabla\Psi(x_R^{md}), \beta_R)\|^2] \leq \frac{192 L_\Psi}{N^2(N+1)} \left[2L_\Psi \|x_0 - x^*\|^2 + \frac{N(N+1)L_f}{2}(\|x^*\|^2 + 2M^2) \right.$$
$$\left. + \sigma^2 \sum_{k=1}^{N} \frac{17k(k+1)}{32 L_\Psi m_k} \right],$$

which clearly implies (3.36). By (3.31), (2.27), (2.30), (2.33), and (2.36), we have

$$\mathbb{E}[\Phi(x_R^{ag}) - \Phi(x^*)] \leq \frac{24 L_\Psi}{N^2(N+1)} \left[\frac{N}{2} \|x_0 - x^*\|^2 + \frac{\sigma^2}{4L_\Psi} \sum_{k=1}^{N} \sum_{j=1}^{k} \frac{17j(j+1)}{32 L_\Psi m_j} \right],$$

which implies (3.37). ∎

Note that all the bounds in the above corollary depend on $\{m_k\}$ and they may not converge to zero for all values of $\{m_k\}$. In particular, if $\{m_k\}$ is set to a positive integer constant, then the last terms in (3.36) and (3.37), unlike the other terms, will not vanish as the algorithm advances. On the other hand, if $\{m_k\}$ is very big, then each iteration of Algorithm 4 will be expensive due to the computation of stochastic gradients. Next result provides an appropriate selection of $\{m_k\}$.

Corollary 5 *Suppose that the stepsizes $\{\alpha_k\}$, $\{\beta_k\}$, and $\{\lambda_k\}$ in Algorithm 4 are set to (2.27) and (2.30), respectively, and $\{p_k\}$ is set to (3.7). Also assume that an optimal solution x^* exists for problem (1.3), an iteration limit $N \geq 1$ is given, and*

$$m_k = \left\lceil \frac{\sigma^2}{L_\Psi \tilde{D}^2} \min\left\{ \frac{k}{L_f}, \frac{k^2 N}{L_\Psi} \right\} \right\rceil, \quad k = 1, 2, \ldots, N \tag{3.38}$$

for some parameter $\tilde{D}$. Then under Assumptions 1 and 2, we have

$$\mathbb{E}[\|\mathcal{G}(x_R^{md}, \nabla\Psi(x_R^{md}), \beta_R)\|^2] \leq 96 L_\Psi \left[\frac{4L_\Psi(\|x_0 - x^*\|^2 + \tilde{D}^2)}{N^3} + \frac{L_f(\|x^*\|^2 + 2M^2 + 3\tilde{D}^2)}{N} \right]. \tag{3.39}$$

If, in addition, $L_f = 0$, then

$$\mathbb{E}[\Phi(x_R^{ag}) - \Phi(x^*)] \leq \frac{L_\Psi}{N^2} \left(12\|x_0 - x^*\|^2 + 7\tilde{D}^2 \right). \tag{3.40}$$

Proof. By (3.38), we have

$$\frac{\sigma^2}{L_\Psi N^3} \sum_{k=1}^{N} \frac{k^2}{m_k} \leq \frac{\tilde{D}^2}{N^3} \sum_{k=1}^{N} k^2 \max\left\{ \frac{L_f}{k}, \frac{L_\Psi}{k^2 N} \right\} \leq \frac{\tilde{D}^2}{N^3} \sum_{k=1}^{N} k^2 \left\{ \frac{L_f}{k} + \frac{L_\Psi}{k^2 N} \right\} \leq \frac{L_f \tilde{D}^2}{N} + \frac{L_\Psi \tilde{D}^2}{N^3},$$

which together with (3.36) imply (3.39). If $L_f = 0$, then due to (3.38), we have

$$m_k = \left\lceil \frac{\sigma^2 k^2 N}{L_\Psi^2 \tilde{D}^2} \right\rceil, \quad k = 1, 2, \ldots, N. \tag{3.41}$$

Using this observation, we have

$$\frac{\sigma^2}{L_\Psi N^3} \sum_{k=1}^{N} \sum_{j=1}^{k} \frac{j^2}{m_j} \leq \frac{L_\Psi \tilde{D}^2}{N^2},$$

which, in view of (3.37), then implies (3.40). ∎

We now add a few remarks about the results obtained in Corollary 5. First, we conclude from (3.39) and Lemma 3 that by running Algorithm 4 for at most

$$\mathcal{O}\left\{ \left[\frac{L_\Psi^2(\|x_0 - x^*\|^2 + \tilde{D}^2)}{\epsilon} \right]^{\frac{1}{3}} + \frac{L_f L_\Psi(M^2 + \|x^*\|^2 + \tilde{D}^2)}{\epsilon} \right\}$$

iterations, we have $-\nabla\Psi(x_R^{ag}) \in \partial\mathcal{X}(x_R^{ag}) + \mathcal{B}(\epsilon)$. Also at the k-th iteration of this algorithm, the $\mathcal{SO}$ is called m_k times and hence the total number of calls to the $\mathcal{SO}$ equals to $\sum_{k=1}^{N} m_k$. Now, observe that by (3.38), we have

$$\sum_{k=1}^{N} m_k \leq \sum_{k=1}^{N} \left(1 + \frac{k\sigma^2}{L_f L_\Psi \tilde{D}^2} \right) \leq N + \frac{\sigma^2 N^2}{L_f L_\Psi \tilde{D}^2}. \tag{3.42}$$

Using these two observations, we conclude that the total number of calls to the $\mathcal{SO}$ performed by Algorithm 4 to find an ϵ-stationary point of problem (1.3) i.e., a point $\bar{x}$ satisfying $-\nabla\Psi(\bar{x}) \in \partial\mathcal{X}(\bar{x}) + \mathcal{B}(\epsilon)$ for some $\epsilon > 0$, can be bounded by

$$\mathcal{O}\left\{\left[\frac{L_\Psi^2(\|x_0 - x^*\|^2 + \tilde{D}^2)}{\epsilon}\right]^{\frac{1}{3}} + \frac{L_f L_\Psi(M^2 + \|x^*\|^2 + \tilde{D}^2)}{\epsilon} + \left[\frac{L_\Psi^{\frac{1}{2}}(\|x_0 - x^*\|^2 + \tilde{D}^2)\sigma^3}{L_f^{\frac{3}{2}}\tilde{D}^3\epsilon}\right]^{\frac{2}{3}} + \frac{L_f L_\Psi(M^2 + \|x^*\|^2 + \tilde{D}^2)^2\sigma^2}{\tilde{D}^2\epsilon^2}\right\}. \tag{3.43}$$

Second, note that there are various choices for the parameter $\tilde{D}$ in the definition of m_k. While Algorithm 4 converges for any $\tilde{D}$, an optimal choice would be $\sqrt{\|x^*\|^2 + M^2}$ for solving composite nonconvex SP problems, if the last term in (3.43) is the dominating one. Third, due to (3.40) and (3.41), it can be easily shown that when $L_f = 0$, Algorithm 4 possesses an optimal complexity for solving convex SP problems which is similar to the one obtained in the Subsection 3.1 for smooth problems. Fourth, note that the definition of $\{m_k\}$ in Corollary 5 depends on the iteration limit N. In particular, due to (3.38), we may call the $\mathcal{SO}$ many times (depending on N) even at the beginning of Algorithm 4. In the next result, we specify a different choice for $\{m_k\}$ which is independent of N. However, the following result is slightly weaker than the one in (3.39) when $L_f = 0$.

Corollary 6 *Suppose that the stepsizes $\{\alpha_k\}$, $\{\beta_k\}$, and $\{\lambda_k\}$ in Algorithm 4 are set to (2.27) and (2.30), respectively, and $\{p_k\}$ is set to (3.7). Also assume that an optimal solution x^* exists for problem (1.3), and*

$$m_k = \left\lceil \frac{\sigma^2 k}{L_\Psi \tilde{D}^2} \right\rceil, \quad k = 1, 2, \ldots \tag{3.44}$$

for some parameter $\tilde{D}$. Then under Assumptions 1 and 2, for any $N \geq 1$, we have

$$\mathbb{E}[\|\mathcal{G}(x_R^{md}, \nabla\Psi(x_R^{md}), \beta_R)\|^2] \leq 96 L_\Psi \left[\frac{4L_\Psi\|x_0 - x^*\|^2}{N^3} + \frac{L_f(\|x^*\|^2 + 2M^2) + 3\tilde{D}^2}{N}\right]. \tag{3.45}$$

Proof. Observe that by (3.44), we have

$$\frac{\sigma^2}{L_\Psi N^3}\sum_{k=1}^{N}\frac{k^2}{m_k} \leq \frac{\tilde{D}^2}{N^3}\sum_{k=1}^{N} k \leq \frac{\tilde{D}^2}{N}.$$

Using this observation and (3.36), we obtain (3.45). ∎

Using Markov's inequality, (3.42), (3.44), and (3.45), we conclude that the total number of calls to the $\mathcal{SO}$ performed by Algorithm 4 for finding an (ϵ, Λ)-*solution* of problem (1.3), i.e., a point $\bar{x}$ satisfying $\text{Prob}\{\|\mathcal{G}(\bar{x}, \nabla\Psi(\bar{x}), c)\|^2 \leq \epsilon\} \geq 1 - \Lambda$ for any $c > 0$, some $\epsilon > 0$ and $\Lambda \in (0, 1)$, can be bounded by (3.24) after disregarding a few constant factors. We can also design a two-phase method for improving the dependence of this bound on the confidence level Λ (see [13, Subsection 4.2] for more details).

4 Concluding remarks

In this paper, we present a generalization of Nesterov's AG method for solving general nonlinear (possibly nonconvex and stochastic) optimization problems. We show that the AG method employed with proper stepsize policy possesses the best known rate of convergence for solving smooth nonconvex problems, similar to the gradient descent method. We also show that this algorithm allows us to have a uniform treatment for solving a certain class of composite optimization problems no matter it is convex or not. In particular, we show that the AG method exhibits an optimal rate of convergence when the composite problem is convex and improves the best known rate of convergence if it is

nonconvex. Based on the AG method, we present a randomized stochastic AG method and show that it can improve a few existing rate of convergence results for solving nonconvex stochastic optimization problems. To the best of our knowledge, this is the first time that Nesterov's method has been generalized and analyzed for solving nonconvex optimization problems in the literature.

References

1. S. Andradóttir. A review of simulation optimization techniques. *Proceedings of the 1998 Winter Simulation Conference*, pages 151–158.
2. S. Asmussen and P. W. Glynn. *Stochastic Simulation: Algorithm and Analysis*. Springer, New York, USA, 2000.
3. A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sciences*, 2:183–202, 2009.
4. C. Cartis, N. I. M. Gould, and Ph. L. Toint. On the complexity of steepest descent, newton's and regularized newton's methods for nonconvex unconstrained optimization. *SIAM Journal on Optimization*, 20(6):2833–2852, 2010.
5. X. Chen, D. Ge, Z. Wang, and Y. Ye. Complexity of unconstrained $l_2 - l_p$ minimization. *Mathematical Programming*, 2012. DOI 10.1007/s10107-012-0613-0.
6. C. D. Dang and G. Lan. Stochastic block mirror descent methods for nonsmooth and stochastic optimization. Manuscript, Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL 32611, USA, August 2013.
7. J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96:13481360, 2001.
8. M. Feng, J. E. Mitchell, J.-S. Pang, X. Shen, , and A. W˙ Technical report.
9. M. Fu. Optimization for simulation: Theory vs. practice. *INFORMS Journal on Computing*, 14:192–215, 2002.
10. S. Ghadimi and G. Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, II: shrinking procedures and optimal algorithms. Technical report, 2010. *SIAM Journal on Optimization* (to appear).
11. S. Ghadimi and G. Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, I: a generic algorithmic framework. *SIAM Journal on Optimization*, 22:1469–1492, 2012.
12. S. Ghadimi and G. Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. Technical report, Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL 32611, USA, June 2012. *SIAM Journal on Optimization* (to appear).
13. S. Ghadimi, G. Lan, and H. Zhang. Mini-batch stochastic approximation methods for constrained nonconvex stochastic programming. Manuscript, Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL 32611, USA, August 2013.
14. G. Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1):365–397, 2012.
15. G. Lan. Bundle-level type methods uniformly optimal for smooth and non-smooth convex optimization. Manuscript, Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL 32611, USA, January 2013. *Mathematical Programming* (to appear).
16. G. Lan and R. D. C. Monteiro. Iteration-complexity of first-order penalty methods for convex programming. *Mathematical Programming*, 138:115–139, 2013.
17. G. Lan and R. D. C. Monteiro. Iteration-complexity of first-order augmented lagrangian methods for convex programming. Technical report, Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL 32611, USA, September 2013. *Mathematical Programming* (Under second-round review).
18. A. M. Law. *Simulation Modeling and Analysis*. McGraw Hill, New York, 2007.
19. J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online dictionary learning for sparse coding. In *In ICML*, pages 689–696, 2009.
20. R.D.C. Monteiro and B.F. Svaiter. An accelerated hybrid proximal extragradient method for convex optimization and its implications to second-order methods. Manuscript, School of ISyE, Georgia Tech, Atlanta, GA, 30332, USA, May 2011.
21. A. S. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19:1574–1609, 2009.
22. A. S. Nemirovski and D. Yudin. *Problem complexity and method efficiency in optimization*. Wiley-Interscience Series in Discrete Mathematics. John Wiley, XV, 1983.
23. Y. E. Nesterov. A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$. *Doklady AN SSSR*, 269:543–547, 1983.
24. Y. E. Nesterov. *Introductory Lectures on Convex Optimization: a basic course*. Kluwer Academic Publishers, Massachusetts, 2004.
25. Y. E. Nesterov. Smooth minimization of nonsmooth functions. *Mathematical Programming*, 103:127–152, 2005.
26. Y. E. Nesterov. Gradient methods for minimizing composite objective functions. Technical report, Center for Operations Research and Econometrics (CORE), Catholic University of Louvain, September 2007.
27. B.T. Polyak and A.B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM J. Control and Optimization*, 30:838–855, 1992.
28. H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 1951.
29. A. Sartenaer S. Gratton and Ph. L. Toint. Recursive trust-region methods for multiscale nonlinear optimization. *SIAM Journal on Optimization*, 19:414–444, 2008.
30. J.C. Spall. *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*. John Wiley, Hoboken, NJ, 2003.
31. P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. Manuscript, University of Washington, Seattle, May 2008.