Open access • Journal Article • DOI:10.1007/S10107-010-0434-Y

An optimal method for stochastic composite optimization — Source link

Guanghui Lan

Institutions: University of Florida

Published on: 01 Jun 2012 - Mathematical Programming (Springer-Verlag)

Topics: Random coordinate descent, Stochastic optimization, Stochastic programming, Stochastic approximation and Convex optimization

Efficient Methods for Stochastic Composite Optimization

Guanghui Lan

School of Industrial and Systems Engineering
Georgia Institute of Technology, Atlanta, GA 30332-0205
Email: glan@isye.gatech.edu

June 21, 2008

Abstract

This paper considers an important class of convex programming problems whose objective function Ψ is given by the summation of a smooth and non-smooth component. Further, it is assumed that the only information available for the numerical scheme to solve these problems is the subgradient of Ψ contaminated by stochastic noise. Our contribution mainly consists of the following aspects. Firstly, with a novel analysis, it is demonstrated that the simple robust mirror-descent stochastic approximation method applied to the aforementioned problems exhibits the best known so far rate of convergence guaranteed by a more involved stochastic mirror-prox algorithm. Moreover, by incorporating some ideas of the optimal method for smooth minimization, we propose an accelerated scheme, which can achieve, uniformly in dimension, the theoretically optimal rate of convergence for solving this class of problems. Finally, the significant advantages of the accelerated scheme over the existing algorithms are illustrated in the context of solving a class of stochastic programming problems whose feasible region is a simple compact convex set intersected with an affine manifold.

Keywords: stochastic approximation, convex optimization, stochastic programming, complexity, optimal method, quadratic penalty method, large deviation
AMS 2000 subject classification: Primary: 62L20, 90C25, 90C15, 68Q25; Secondary:49M37, 60F10
OR/MS subject classification: Primary: programming, nondifferentible, stochastic; Secondary: statistic, sampling

1 Introduction

The basic problem of interest in this paper is

$$\Psi^* := \min_{x \in X} \{\Psi(x) := f(x) + h(x)\}, \tag{1}$$

where X is a convex compact set in Euclidean space $\mathcal{E}$ with inner product $\langle \cdot, \cdot \rangle$, $f : X \to \Re$ is a convex function with Lipschitz continuous gradient, that is,

$$\|\nabla f(x) - \nabla f(x')\|_* \leq L\|x - x'\|, \ \forall x, x' \in X, \tag{2}$$

($\|\cdot\|$ is a given norm in $\mathcal{E}$, $\|\cdot\|_*$ denotes its conjugate norm, i.e., $\|y\|_* := \max_{\|z\|\leq 1}\langle y, z\rangle$), and $h : X \to \Re$ is a convex Lipschitz continuous function such that

$$|h(x) - h(x')| \leq M\|x - x'\|, \ \forall x, x' \in X. \tag{3}$$

We are interested in the situation where problem (1) is solved by an iterative algorithm which acquires the subgradients of Ψ via subsequent calls to a stochastic oracle ($\mathcal{SO}$). Specifically, at iteration t of the algorithm, $x_t \in X$ being the input, the $\mathcal{SO}$ outputs a vector $G(x_t, \xi_t)$, where $\{\xi_t\}_{t\geq 1}$ is a sequence of i.i.d. random variables which are also independent of the search point x_t. The following assumptions are made for the Borel functions $G(x, \xi_t)$.

Assumption I: For any $x \in X$, we have

$$\text{a)} \quad \mathbb{E}[G(x, \xi_t)] \equiv g(x) \in \partial\Psi(x) \tag{4}$$

$$\text{b)} \quad \mathbb{E}\left[\|G(x, \xi_t) - g(x)\|_*^2\right] \leq Q^2, \tag{5}$$

where $\partial\Psi(x)$ denotes the subdifferential of Ψ at x (see Subsection 1.1).

Let us consider the normalized problem of (1) where X is the unit Euclidean ball in $\mathcal{E} = \Re^n$. According to the classical complexity theorem for convex programming by Nemirovski and Yudin [7], if the dimension n is sufficiently large, the rate of convergence for any iterative algorithms for solving (1) can not be better than

$$\mathcal{O}(1)\left[\frac{L}{N^2} + \frac{M + Q}{\sqrt{N}}\right], \tag{6}$$

where N is the number of iterations performed by the algorithm. This means that, for any algorithms solving problem (1), one can always point out a "bad" problem instance satisfying (2), (3), (4), and (5), such that the expected error of the solution generated at the N-step of the algorithm will be, up to a constant factor, greater than the lower bound stated above. Moreover, to the best of our knowledge, none of the existing algorithms, uniformly in the dimension, achieved this lower bound. Somewhat surprisingly, this optimal rate of convergence has not been attained even for the deterministic case where $Q = 0$. The best known result so far is given by Juditsky et al. [3] with the rate of convergence

$$\mathcal{O}(1)\left[\frac{L}{N} + \frac{M + Q}{\sqrt{N}}\right] \tag{7}$$

by applying an extra-gradient-type algorithm to a variational inequality (v.i.) reformulation of (1).

In this paper, we focus on the subgradient-type methods applied directly to the original problem (1). We show that, with a novel analysis, the simple robust stochastic approximation method (RM-SA) developed in Juditsky et al. [2] also exhibits the guaranteed rate of convergence in (7). Moreover, by incorporating some ideas of the optimal method for smooth minimization (Nesterov [8, 9], Lan et al. [6]), we present an accelerated stochastic approximation (AC-SA) algorithm, which can achieve, uniformly in dimension, the optimal rate of convergence given by (6). We investigate this accelerated scheme in more details, for example, derive the exponential bounds for the large deviations of the resulting solution inaccuracy from the expected one, provided the noise from the $\mathcal{SO}$ is "light-tailed". The significant advantages of the AC-SA method over the existing algorithms are illustrated in the context of solving a class of stochastic programming problems whose feasible region is a simple compact convex set intersected with an affine manifold.

The paper is organized as follows. In Section 2, we give a brief introduction to the RM-SA algorithm applied to (1) and present the new convergence result. Section 3 discusses the

accelerated stochastic approximation method. More specifically, we present the AC-SA algorithm and its convergence properties in Subsection 3.1, and outline an application to demonstrate the advantages of this algorithm in Subsection 3.2. Section 4 is devoted to proving the main results of this paper. Finally, some concluding remarks are made in Section 5.

1.1 Notation and terminology

- For a convex lower semicontinuous function $\phi : X \to \Re$, its subdifferential $\partial\phi(\cdot)$ is defined as follows: at a point x from the relative interior of X, $\partial\phi$ is comprised of all subgradients g of ϕ at x which are in the linear span of $X - X$. For a point $x \in X \backslash \operatorname{rint} X$, the set $\partial\phi(x)$ consists of all vectors g, if any, such that there exists $x_i \in \operatorname{rint} X$ and $g_i \in \partial\phi(x_i)$, $i = 1, 2, \cdots$, with $x = \lim\limits_{i\to\infty} x_i$, $g = \lim\limits_{i\to\infty} g_i$. Finally, $\partial\phi(x) = \emptyset$ for $x \notin X$. Note that with this definition (see, for example, Ben-Tal and Nemirovski [1]), if a convex function $\phi : X \to \Re$ is Lipschitz continuous, with constant M, with respect to a norm $\|\cdot\|$, then the set $\partial\phi(x)$ is nonempty for any $x \in X$ and

$$g \in \partial\phi(x) \Rightarrow |\langle g, d\rangle| \leq M\|d\|, \ \forall\, d \in \operatorname{lin}(X - X). \tag{8}$$

- For the random process $\xi_1, \xi_2, ...$, we set $\xi_{[t]} := (\xi_1, ..., \xi_t)$, and denote by $\mathbb{E}_{|\xi_{[t]}}$ the conditional, $\xi_{[t]}$ being given, expectation.

2 Robust mirror-descent stochastic approximation

In this section, we give a brief introduction to the robust mirror-descent stochastic approximation method and present a new rate of convergence for this algorithm applied to (1).

2.1 Preliminaries: distance generating function and prox-mapping

We say that a function $\omega : X \to \mathbb{R}$ is a *distance generating function* modulus $\alpha > 0$, with respect to the norm $\|\cdot\|$, if ω is convex and continuous on X, the set

$$X^o := \big\{x \in X : \text{there exists } p \in \mathbb{R}^n \text{ such that } x \in \arg\min_{u\in X}[p^T u + \omega(u)]\big\}$$

is convex, and $\omega(\cdot)$ restricted to X^o is continuously differentiable and strongly convex with parameter α with respect to $\|\cdot\|$, i.e.,

$$\langle \nabla\omega(x) - \nabla\omega(x'), x - x'\rangle \geq \alpha\|x - x'\|^2, \ \ \forall x, x' \in X^o. \tag{9}$$

Note that set X^o always contains the relative interior of set X.

The *prox-function* $V : X^o \times X \to \mathbb{R}_+$ and the *prox-mapping* $P_x : \mathbb{R}^n \to X^o$ associated with $\omega(x)$ are defined as

$$V(x, z) := \omega(z) - \omega(x) - \langle \nabla\omega(x), z - x\rangle, \tag{10}$$

$$P_x(y) := \arg\min_{z\in X}\big\{y^T(z - x) + V(x, z)\big\}. \tag{11}$$

The distance generating function ω also gives rise to the following two characteristic entities that will be used frequently in our convergence analysis:

$$D_{\omega,X} := \sqrt{\max_{x\in X}\omega(x) - \min_{x\in X}\omega(x)}, \quad \Omega = \Omega_{\omega,X} := \sqrt{\frac{2}{\alpha}}D_{\omega,X}. \tag{12}$$

Let x_1 be the minimizer of ω over X. Observe that $x_1 \in X^o$, whence $\nabla w(x_1)$ is well defined and satisfies $\langle \nabla\omega(x_1), x - x_1\rangle \geq 0$ for all $x \in X$, which combined with the strong convexity of ω implies that

$$\frac{\alpha}{2}\|x - x_1\|^2 \leq V(x_1, x) \leq \omega(x) - \omega(x_1) \leq D^2_{\omega,X}, \ \ \forall x \in X, \tag{13}$$

and hence

$$\|x - x_1\| \leq \Omega, \ \ \forall x \in X, \text{ and } \|x - x'\| \leq \|x - x_1\| + \|x' - x_1\| \leq 2\Omega, \ \forall x, x' \in X. \tag{14}$$

2.2 Robust mirror-descent SA algorithm and its convergence properties

The RM-SA algorithm, as applied to (1), works with the stochastic oracle of Ψ that satisfies Assumption I. In some cases, Assumption I is augmented by the following "light-tail" assumption.

Assumption II: For any $x \in X$, we have

$$\mathbb{E}\left[\exp\{\|G(x,\xi_t) - g(x)\|_*^2/Q^2\}\right] \ \ \leq \ \ \exp\{1\}. \tag{15}$$

It can be easily seen that Assumption II implies Assumption I.b), since by Jensen's inequality,

$$\exp\left(\mathbb{E}[\|G(x,\xi_t) - g(x)\|_*^2/Q^2]\right) \ \ \leq \ \ \mathbb{E}\left[\exp\{\|G(x,\xi_t) - g(x)\|_*^2/Q^2\}\right] \leq \exp\{1\}.$$

We are now ready to state the RM-SA algorithm applied to (1).

The RM-SA algorithm:

0) Let the initial point x_1 and the step-sizes $\{\gamma_t\}_{t\geq 1}$ be given. Set $t = 1$;

1) Call the $\mathcal{SO}$ for computing $G(x_t, \xi_t)$. Set

$$x_{t+1} \ := \ P_{x_t}\left(\gamma_t G(x_t, \xi_t)\right), \tag{16}$$

$$x^{ag}_{t+1} \ = \ \left(\sum_{\tau=1}^{t}\gamma_\tau\right)^{-1}\sum_{\tau=1}^{t}\gamma_\tau x_{\tau+1}. \tag{17}$$

2) Set $t \leftarrow t + 1$ and go to Step 1.

end

We now make a few comments about the above algorithm. Firstly, without loss of generality, we will assume from now on that the initial point x_1 is given by the minimizer of ω over X (see Subsection 2.1). Secondly, observe that the above algorithm only slightly differs from the RM-SA algorithm of Juditsky et al. [2] in the way the averaging step (17) is defined. More specifically, the sequence $\{x^{ag}_t\}_{t\geq 2}$ is obtained by averaging the iterates $x_t, t \geq 2$ with their corresponding weights

γ_{t-1}, while the one in [2] is obtained by taking the average of the whole trajectory $x_t, t \geq 1$ with weights γ_t. Note however, if the constant step-sizes are used, i.e., $\gamma_t = \gamma, \forall\, t \geq 1$, then the RM-SA algorithm stated above is exactly the same as the one stated in [2] up to shifting one iterate in the averaging step. Thirdly, in view of Proposition 2.1 in [2], the rate of convergence of the RM-SA algorithm, as applied to problem (1), is given by

$$\mathcal{O}(1)\left[\frac{\Omega(LD_X + M + Q)}{\sqrt{N}}\right], \tag{18}$$

where N is the number of iterations performed by the algorithm, D_X is the $\|\cdot\|$-diameter of X, i.e., $D_X := \max_{x,x' \in X} \|x - x'\|$, L, M, and Q are defined in (2), (3), and (5) respectively. The goal of this section is to demonstrate that, with certain appropriately chosen step-sizes γ_t and a different convergence analysis, a stronger rate of convergence can be derived for the RM-SA algorithm applied to (1).

We start with stating a general convergence result of the above RM-SA algorithm without specifying the step-sizes γ_t.

Theorem 1 *Assume that the step-sizes γ_t satisfy $0 < \gamma_t \leq \alpha/(2L)$, $\forall\, t \geq 1$. Let $\{x_{t+1}^{ag}\}_{t \geq 1}$ be the sequence computed according to (17) by the RM-SA algorithm. Then we have*

a) under Assumption I,

$$\mathbb{E}\left[\Psi(x_{t+1}^{ag}) - \Psi^*\right] \leq K_0(t),\ \forall\, t \geq 1, \tag{19}$$

where

$$K_0(t) := \left(\sum_{\tau=1}^{t} \gamma_\tau\right)^{-1} \left[D_{\omega,X}^2 + \frac{2}{\alpha}(4M^2 + Q^2)\sum_{\tau=1}^{t} \gamma_\tau^2\right],$$

M, Q, and $D_{\omega,X}$ are given in (3), (5), and (12) respectively;

b) under Assumptions I and II,

$$\text{Prob}\left\{\Psi(x_{t+1}^{ag}) - \Psi^* > K_0(t) + \Lambda K_1(t)\right\} \leq \exp\{-\Lambda^2/3\} + \exp\{-\Lambda\},\ \forall\, \Lambda > 0, t \geq 1, \tag{20}$$

where

$$K_1(t) := \left(\sum_{\tau=1}^{t} \gamma_\tau\right)^{-1} \left[2\Omega Q\sqrt{\sum_{\tau=1}^{t} \gamma_\tau^2} + \frac{2}{\alpha}Q^2\sum_{\tau=1}^{t} \gamma_\tau^2\right],$$

Q and Ω are given in (5) and (12) respectively.

For the sake of simplicity, let us suppose that the number of iterations for the above algorithm is fixed in advance, say equal to N, and that a *constant step-size strategy* is applied, i.e., $\gamma_t = \gamma, t = 1, \cdots, N$, for some $\gamma < \alpha/(2L)$ (note that the assumption of constant step-sizes does not hurt the efficiency estimate). We then conclude from Theorem 1 that the obtained solution $x_{N+1}^{ag} = N^{-1}\sum_{\tau=1}^{N} x_{\tau+1}$ satisfies

$$\mathbb{E}\left[\Psi(x_{N+1}^{ag}) - \Psi^*\right] \leq \frac{D_{\omega,X}^2}{N\gamma} + \frac{2\gamma}{\alpha}\left(4M^2 + Q^2\right).$$

Minimizing the right-hand-side of the above inequality with respect to γ over the interval $(0, \alpha/(2L)]$, we conclude that

$$\mathbb{E}\left[\Psi(x_{N+1}^{ag}) - \Psi^*\right] \leq K_0^*(N) := \frac{L\Omega^2}{N} + \frac{2\Omega\sqrt{4M^2+Q^2}}{\sqrt{N}}, \tag{21}$$

by choosing γ as

$$\gamma = \min\left\{\frac{\alpha}{2L}, \sqrt{\frac{\alpha D_{\omega,X}^2}{2N(4M^2+Q^2)}}\right\}.$$

Moreover, with this choice of γ, we have

$$\begin{aligned} K_1(N) &= \frac{2\Omega Q}{\sqrt{N}} + \frac{2\gamma Q^2}{\alpha} \leq \frac{2\Omega Q}{\sqrt{N}} + \sqrt{\frac{2}{\alpha}} D_{\omega,X} \frac{Q^2}{\sqrt{N(4M^2+Q^2)}} \\ &\leq \frac{2\Omega Q}{\sqrt{N}} + \sqrt{\frac{2}{\alpha}} D_{\omega,X} \frac{Q}{\sqrt{N}} = \frac{3\Omega Q}{\sqrt{N}}, \end{aligned}$$

hence, bound (20) implies that

$$\text{Prob}\left\{\Psi(x_{N+1}^{ag}) - \Psi^* > K_0^*(N) + \Lambda K_1^*(N)\right\} \leq \exp\{-\Lambda^2/3\} + \exp\{-\Lambda\}, \, \forall\, \Lambda > 0, \tag{22}$$

where

$$K_1^*(N) := \frac{3\Omega Q}{\sqrt{N}}.$$

It is interesting to compare bounds (21) and (18) derived for the RM-SA algorithm. Clearly, if $D_{\omega,X} \leq \sqrt{\alpha N} D_X$ (which is typically the case in applications, see [2] and [3]), the latter one is always worse than the first one. Moreover, in the range

$$L \leq \frac{\sqrt{N(4M^2+Q^2)}}{\Omega}, \tag{23}$$

the first component in (21) (for abbreviation, the L-component) merely does not affect the error estimate (21). Note that the range stated in (23) extends as N increases, meaning that, if N is large and $Q = \mathcal{O}(1)M$, the presence of the smooth component f in the objective function of (1) does not affect the complexity of finding good approximate solutions. In contrast, this phenomenon does not appear in the error estimate (18) derived for the RM-SA algorithm in [2] which employs certain simple step-size strategies without taking into account the structure of the objective function Ψ.

3 The accelerated stochastic approximation method

Motivated by Nesterov's optimal method ([8, 9]) and its variants ([6]) for smooth minimization, we present in this section an accelerated stochastic approximation method, which achieves the theoretically optimal rate of convergence for solving (1). Specifically, we state the algorithm and its convergence results in Subsection 3.1 and outline an application to illustrate its advantages over the RM-SA algorithm in Subsection 3.2.

3.1 The algorithm and its main convergence properties

The AC-SA algorithm for solving problem (1) is comprised of the updating of three sequences: $\{x_t\}_{t\geq 1}$, $\{x_t^{ag}\}_{t\geq 1}$, and $\{x_t^{md}\}_{t\geq 1}$. Here, we use the superscript "ag" (which stands for "aggregated") in the sequence obtained by taking a convex combination of all the previous iterates x_t, and the superscript "md" (which stands for "middle") in the sequence obtained by taking a convex combination of the current iterate x_t with the current aggregated iterate x_t^{ag}. The algorithm is stated as follows.

The AC-SA algorithm:

0) Let the initial points $x_1^{ag} = x_1$, and the step-sizes $\{\beta_t\}_{t\geq 1}$ and $\{\gamma_t\}_{t\geq 1}$ be given. Set $t = 1$.

1) Set $x_t^{md} = \beta_t^{-1} x_t + (1 - \beta_t^{-1}) x_t^{ag}$,

2) Call the $\mathcal{SO}$ for computing $G(x_t^{md}, \xi_t)$. Set

$$x_{t+1} = P_{x_t}(\gamma_t G(x_t^{md}, \xi_t)), \tag{24}$$

$$x_{t+1}^{ag} = \beta_t^{-1} x_{t+1} + (1 - \beta_t^{-1}) x_t^{ag}, \tag{25}$$

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

end

We now make a few comments regarding the AC-SA algorithm described above. Firstly, similar to the RM-SA algorithm, we assume that the initial point x_1 is the minimizer of ω over X (see Subsection 2.1). Secondly, it is worth noting that the major computation cost in each iteration of the AC-SA algorithm is exactly the same as the one of the RM-SA algorithm, that is, each iteration of the above algorithm requires only one call to the $\mathcal{SO}$ and one solution of the subproblem (24). Thirdly, observe that the AC-SA algorithm reduces to a variant of Nesterov's optimal method ([8, 9]) if the non-smooth component $h(\cdot)$ in the objective function $\Psi(\cdot)$ does not appear and that there is no noise in the computed gradient, i.e., $Q = 0$ in (5).

The following theorem states the main convergence results of the AC-SA algorithm described above.

Theorem 2 *Assume that the step-sizes $\beta_t \in [1, \infty)$ and $\gamma_t \in \Re_+$ are chosen such that $\beta_1 = 1$ and the following conditions hold*

$$0 < (\beta_{t+1} - 1)\gamma_{t+1} \leq \beta_t \gamma_t \text{ and } 2L\gamma_t \leq \alpha\beta_t \ \forall t \geq 1. \tag{26}$$

Let $\{x_{t+1}^{ag}\}_{t\geq 1}$ be the sequence computed according to (25) by the AC-SA algorithm. Then we have

a) under Assumption I,

$$\mathbb{E}[\Psi(x_{t+1}^{ag}) - \Psi^*] \leq \hat{K}_0(t), \ \forall\, t \geq 1, \tag{27}$$

where

$$\hat{K}_0(t) := \frac{1}{(\beta_{t+1} - 1)\gamma_{t+1}} \left[D_{\omega,X}^2 + \frac{2}{\alpha}(4M^2 + Q^2) \sum_{\tau=1}^{t} \gamma_\tau^2 \right],$$

M, Q, and $D_{\omega,X}$ are given in (3), (5), and (12) respectively;

b) under Assumptions I and II,

$$\text{Prob}\left\{\Psi(x_{t+1}^{ag}) - \Psi^* > \hat{K}_0(t) + \Lambda\hat{K}_1(t)\right\} \leq \exp\{-\Lambda^2/3\} + \exp\{-\Lambda\},\ \forall\,\Lambda > 0, t \geq 1, \quad (28)$$

where

$$\hat{K}_1(t) := \frac{1}{(\beta_{t+1}-1)\gamma_{t+1}}\left[2\Omega Q\sqrt{\sum_{\tau=1}^{t}\gamma_\tau^2} + \frac{2}{\alpha}Q^2\sum_{\tau=1}^{t}\gamma_\tau^2\right],$$

Q and Ω are given in (5) and (12) respectively.

We now discuss the determination of the step-sizes β_t and γ_t in the accelerated scheme so as to achieve the optimal rate of convergence for solving (1). Observing that a pair of sequences $\{\beta_t\}_{t\geq 1}$ and $\{\gamma_t\}_{t\geq 1}$ satisfying condition (26) is given by:

$$\beta_t = \frac{t+1}{2} \text{ and } \gamma_t = \frac{t+1}{2}\gamma \quad (29)$$

for any $0 < \gamma \leq \alpha/(2L)$, we obtain the following corollary of Theorem 2 by appropriately choosing this parameter γ, .

Corollary 3 *Suppose that the step-sizes β_t and γ_t in the AC-SA algorithm are set to*

$$\beta_t = \frac{t+1}{2},\quad \gamma_t = \frac{t+1}{2}\min\left\{\frac{\alpha}{2L}, \frac{\sqrt{6\alpha}D_{\omega,X}}{(N+2)^{\frac{3}{2}}(4M^2+Q^2)^{\frac{1}{2}}}\right\},\ \forall t \geq 1, \quad (30)$$

where N is a fixed in advance number of iterations. Then, we have under Assumption I,

$$\mathbb{E}[\Psi(x_{N+1}^{ag}) - \Psi^*] \leq \hat{K}_0^*(N) := \frac{4L\Omega^2}{N(N+2)} + \frac{4\Omega\sqrt{4M^2+Q^2}}{\sqrt{N}}, \quad (31)$$

if in addition, Assumption II holds, then

$$\text{Prob}\left\{\Psi(x_{N+1}^{ag}) - \Psi^* > \hat{K}_0^*(N) + \Lambda\hat{K}_1^*(N)\right\} \leq \exp\{-\Lambda^2/3\} + \exp\{-\Lambda\},\ \forall\,\Lambda > 0, \quad (32)$$

where

$$\hat{K}_1^*(N) := \frac{10\Omega Q}{\sqrt{N}}.$$

Proof. Clearly, the step-sizes $\{\beta_t\}_{t\geq 1}$ and $\{\gamma_t\}_{t\geq 1}$ stated in (30) satisfy the conditions $\beta_1 = 1$, $\beta_t > 1, \forall t \geq 2$, and (26). Denoting

$$\gamma^* := \min\left\{\frac{\alpha}{2L}, \frac{\sqrt{6\alpha}D_{\omega,X}}{(N+2)^{\frac{3}{2}}(4M^2+Q^2)^{\frac{1}{2}}}\right\},$$

we then conclude from Theorem 2 that, under Assumption I,

$$\mathbb{E}[\Psi(x_{N+1}^{ag}) - \Psi^*] \leq \mathcal{T}_0 := \frac{4D_{\omega,X}^2}{N(N+2)\gamma^*} + \frac{8\gamma^*(4M^2+Q^2)}{\alpha N(N+2)}\sum_{\tau=1}^{N}\left(\frac{\tau+1}{2}\right)^2, \quad (33)$$

and that, under Assumptions I and II,

$$\text{Prob}\left\{\Psi(x_{N+1}^{ag}) - \Psi^* > \mathcal{T}_0 + \Lambda\mathcal{T}_1)\right\} \leq \exp\{-\Lambda^2/3\} + \exp\{-\Lambda\}, \, \forall\, \Lambda > 0, \tag{34}$$

where

$$\mathcal{T}_1 := \frac{8\Omega Q}{N(N+2)}\sqrt{\sum_{\tau=1}^{N}\left(\frac{\tau+1}{2}\right)^2} + \frac{8\gamma^* Q^2}{N(N+2)\alpha}\sum_{\tau=1}^{N}\left(\frac{\tau+1}{2}\right)^2.$$

Moreover, using the simple observations $\sum_{\tau=1}^{N}(\tau+1)^2 \leq \int_1^{N+1}(u+1)^2 du \leq (N+2)^3/3$, $N+2 \leq 3N$ due to $N \geq 1$, and the definition of γ^*, we obtain

$$\begin{aligned}
\mathcal{T}_0 &\leq \frac{4D_{\omega,X}^2}{N(N+2)\gamma^*} + \frac{2\gamma^*(4M^2+Q^2)(N+2)^2}{3\alpha N} \leq \frac{8LD_{\omega,X}^2}{N(N+2)\alpha} + \frac{8D_{\omega,X}(4M^2+Q^2)^{\frac{1}{2}}(N+2)^{\frac{1}{2}}}{\sqrt{6\alpha}N} \\
&\leq \frac{8LD_{\omega,X}^2}{N(N+2)\alpha} + \frac{8D_{\omega,X}(4M^2+Q^2)^{\frac{1}{2}}}{\sqrt{2\alpha N}} = \frac{4L\Omega^2}{N(N+2)} + \frac{4\Omega\sqrt{4M^2+Q^2}}{\sqrt{N}} = \hat{K}_0^*(N),
\end{aligned}$$

and

$$\begin{aligned}
\mathcal{T}_1 &\leq \frac{8\Omega Q}{\sqrt{3}N}(N+2)^{\frac{1}{2}} + \frac{2\gamma^* Q^2}{3N\alpha}(N+2)^2 \leq \frac{8\Omega Q}{\sqrt{N}} + \frac{2Q^2(N+2)^{\frac{1}{2}}}{3N\sqrt{\alpha}}\frac{\sqrt{6}D_{\omega,X}}{\sqrt{4M^2+Q^2}} \\
&\leq \frac{8\Omega Q}{\sqrt{N}} + \frac{2\sqrt{2}D_{\omega,X}Q}{\sqrt{\alpha N}} = \frac{10\Omega Q}{\sqrt{N}} = \hat{K}_1^*(N).
\end{aligned}$$

Our claim immediately follows by substituting the above bounds of $\mathcal{T}_0$ and $\mathcal{T}_1$ into (33) and (34). ■

We now make a few observations regarding the results obtained in Theorem 2 and Corollary 3. Firstly, it is interesting to compare bounds (31) and (21) obtained for the AC-SA algorithm and the RM-SA algorithm respectively. Clearly, the first one is always better than the latter one up to a constant factor provided that $L > 0$. Moreover, the AC-SA algorithm substantially enlarges the range of L in which the L-component (the first component in (31)) does not affect the error estimate. Specifically, in the range

$$L \leq \frac{\sqrt{4M^2+Q^2}N^{\frac{3}{2}}}{\Omega}, \tag{35}$$

which extends much faster than (23) as N increases, the L-component does not change the order of magnitude for the rate of convergence associated with the AC-SA algorithm.

Secondly, observe that the results obtained in Theorem 2 and Corollary 3 still hold when the Lipschitz constant $L = 0$. More specifically, we consider the case where $f(\cdot) \equiv 0$. In this case, the step-sizes $\{\beta_t\}_{t\geq 1}$ and $\{\gamma_t\}_{t\geq 1}$ in (30) become

$$\beta_t = \frac{t+1}{2}, \quad \gamma_t = \frac{\sqrt{6\alpha}D_{\omega,X}(t+1)}{2(N+2)^{\frac{3}{2}}(4M^2+Q^2)^{\frac{1}{2}}}, \; 1 \leq t \leq N+1,$$

and the error estimate (31) reduces to

$$\mathbb{E}[h(x_{N+1}^{ag}) - h^*] \leq \frac{4\Omega\sqrt{4M^2+Q^2}}{\sqrt{N}},$$

where $h^* := \min_{x\in X} h(x)$. Note also that one alternative characterization of x^{ag}_{N+1} is given by

$$\begin{aligned} x^{ag}_{N+1} &= \frac{2}{N+1}x_{N+1} + \frac{N-1}{N+1}x^{ag}_N = \frac{2}{N+1}x_{N+1} + \frac{2(N-1)}{N(N+1)}x_N + \frac{(N-2)(N-1)}{N(N+1)}x^{ag}_{N-1} \\ &= \frac{2}{N+1}x_{N+1} + \frac{2(N-1)}{N(N+1)}x_N + \frac{2(N-2)}{N(N+1)}x_{N-1} + \cdots + \frac{2}{N(N+1)}x_2 \\ &= \frac{\sum_{t=1}^N (t x_{t+1})}{\sum_{t=1}^N t}. \end{aligned}$$

Hence, in contrast to the usual *constant step-size* or *decreasing step-size* strategy (see [2]), the step-sizes γ_t in step (24) and the weights for taking the average in step (25) are increasing with the increment of t. To the best of our knowledge, this is the first time that an *increasing step-size* strategy is introduced in the literature of stochastic approximation or subgradient methods.

Finally, note that if there is no stochastic error for the computed subgradient of Ψ, i.e., $Q = 0$, then bound (31) reads

$$\Psi(x^{ag}_{N+1}) - \Psi^* \le \frac{4L\Omega^2}{N(N+2)} + \frac{8\Omega M}{\sqrt{N}},$$

which basically says that the impact of the smooth component on the efficiency estimate vanishes very quickly as N grows. This result also seems to be new in the area of deterministic convex optimization.

3.2 An illustrative application

The goal of this subsection is to demonstrate the significant advantages of the AC-SA algorithm over the existing algorithms, for example, the RM-SA algorithm, when applied for solving certain class of stochastic programming problems.

Consider the problem of

$$\begin{aligned} \tilde{h}^* := \min_x \quad & \{\tilde{h}(x) := \mathbb{E}[\tilde{H}(x,\xi)]\} \\ \text{s.t.} \quad & \mathcal{A}x - b = 0, \quad x \in X, \end{aligned} \tag{36}$$

where $X \subset \mathbb{R}^n$ is a nonempty compact convex set, $\mathcal{A} : \Re^n \to \Re^m$ is a linear operator, $b \in \Re^m$ is given, ξ is a random vector whose probability distribution P is supported on set $\Xi \subseteq \mathbb{R}^d$ and $\tilde{H} : X \times \Xi \to \mathbb{R}$. We assume that for every $\xi \in \Xi$ the function $\tilde{H}(\cdot,\xi)$ is convex on X, and that the expectation

$$\mathbb{E}[\tilde{H}(x,\xi)] = \int_\Xi \tilde{H}(x,\xi) dP(\xi) \tag{37}$$

is well defined and finite valued for every $x \in X$. It follows that function $\tilde{h}(\cdot)$ is convex and finite valued on X. Moreover, we assume that $\tilde{h}(\cdot)$ is continuous on X. Of course, continuity of $\tilde{h}(\cdot)$ follows from convexity if $\tilde{h}(\cdot)$ is finite valued and convex on a neighborhood of X. With these assumptions, (36) becomes a convex programming problem. We also make the following assumptions:

Assumption III:

a) It is possible to generate an iid sample $\xi_1, \xi_2, ...,$ of realizations of random vector ξ.

b) We have access to a "black box" subroutine (a stochastic oracle). At i-th call, $x \in X$ being the input, the oracle returns a *stochastic subgradient* – a vector $\tilde{\mathcal{H}}(x, \xi_i)$ such that for every $x \in X$, the vector $\mathbb{E}[\tilde{\mathcal{H}}(x, \xi)]$ is well defined and is a subgradient of $\tilde{h}(\cdot)$ at x.

c) There is a constant $M > 0$ such that

$$\forall x \in X : \mathbb{E}\left[\exp\{\|\tilde{\mathcal{H}}(x, \xi)\|_*^2/M^2\}\right] \leq \exp\{1\}. \tag{38}$$

For the case where the feasible region consists only of the simple convex set X, or equivalently $\mathcal{A} \equiv 0$, Juditsky et. al. demonstrated in [2] that the RM-SA algorithm can substantially outperform the sampling averaging approximation (Shapiro [10]), a widely used approach for stochastic programming in practice. When $\mathcal{A}$ is not identically 0, the RM-SA algorithm can still be applied directly to problem (36) but this approach would require the computation of the prox-mapping onto the feasible region $X \cap \{x : \mathcal{A}x - b = 0\}$, which can be very expensive for many practical problems.

One alternative approach to alleviate this difficulty is to apply the quadratic penalty approach: instead of solving (36), we solve certain penalization problem of (36) obtained by penalizing the violation of the constraint $\mathcal{A}x - b = 0$. In particular, given a penalty parameter $\rho > 0$, we solve

$$\tilde{\Psi}^* = \tilde{\Psi}_\rho^* := \inf_{x \in X} \left\{ \tilde{\Psi}_\rho(x) := \tilde{f}_\rho(x) + \tilde{h}(x) \right\}, \tag{39}$$

where $\tilde{f}_\rho(x) := \rho\|\mathcal{A}x - b\|^2/2$ and $\|\cdot\|$ denotes the norm induced by the inner product $\langle \cdot, \cdot \rangle$ in $\Re^m$. Define the operator norm $\|\mathcal{A}\| := \max\{\|\mathcal{A}x\|_* : \|x\| \leq 1\}$. It can be easily seen that $\nabla \tilde{f}_\rho(x) = \rho\mathcal{A}^*(\mathcal{A}x - b)$ and hence that

$$\|\nabla \tilde{f}_\rho(x) - \nabla \tilde{f}_\rho(x')\|_* = \rho\|\mathcal{A}^*\mathcal{A}(x - x')\|_* \leq \rho\|\mathcal{A}^*\|\|\mathcal{A}\|\|x - x'\| = \rho\|\mathcal{A}\|^2\|x - x'\|, \forall\, x, x' \in X, \tag{40}$$

where the last equality follows from the fact that $\|\mathcal{A}\| = \|\mathcal{A}^*\|$. Moreover, in view of Assumption III and Jensen's inequality, for any $x \in X$, there exists $\tilde{h}'(x) := \mathbb{E}[\tilde{\mathcal{H}}(x, \xi_t)] \in \partial\tilde{h}(x)$ such that $\mathbb{E}[\|\tilde{\mathcal{H}}(x, \xi_t)\|_*^2] \leq M^2$ and hence that $\|\tilde{h}'(x)\|_* = \|\mathbb{E}[\tilde{\mathcal{H}}(x, \xi_t)]\|_* \leq M$, which together with the fact $\tilde{h}(x) - \tilde{h}(x') \leq \langle \tilde{h}'(x), x - x' \rangle, \forall x, x' \in X$ due to the convexity of $\tilde{h}$, clearly imply that

$$\|\tilde{h}(x) - \tilde{h}(x')\| \leq M\|x - x'\|, \ \ \forall x, x' \in X. \tag{41}$$

Therefore, the penalization problem (39) is given in the form of (1), and can be approximately solved by either the RM-SA or the AC-SA algorithm developed in this paper.

It is well-known that the near-optimal solutions of the penalization problem (39) also yield near-optimal solutions of (36) if the penalty parameter ρ is sufficiently large. In this paper, we are interested in obtaining one particular type of near-optimal solutions of (36) defined in the following way. First note that x^* is an optimal solution of (36) if, and only if, $x^* \in X$, $\mathcal{A}x^* - b = 0$ and $\tilde{h}(x^*) \leq \tilde{h}^*$. This observation leads us to our definition of a near optimal solution $\tilde{x} \in X$ of (36), which essentially requires the primal infeasibility measure $\|\mathcal{A}\tilde{x} - b\|_2$ and the primal optimality gap $[\tilde{h}(\tilde{x}) - \tilde{h}^*]^+$ to be both small (Lan and Monteiro [4]).

Definition: Let $\epsilon_p, \epsilon_o > 0$ be given, $\tilde{x} \in X$ is called an (ϵ_p, ϵ_o)-*primal solution* for (36) if

$$\|\mathcal{A}\tilde{x} - b\| \leq \epsilon_p \text{ and } \tilde{h}(\tilde{x}) - \tilde{h}^* \leq \epsilon_0. \tag{42}$$

One drawback of the above notion of near optimality of $\tilde{x}$ is that it says nothing about the size of $[\tilde{h}(\tilde{x}) - \tilde{h}^*]^-$. Assume that the set of Lagrange multiplier for (36)

$$Y^* := \{y \in \Re^m : \tilde{h}^* = \inf\{\tilde{h}(x) + \langle \mathcal{A}x - b, y\rangle : x \in X\}$$

is nonempty. It was observed in [4] that this quantity can be bounded as $[\tilde{h}(\tilde{x}) - \tilde{h}^*]^- \leq \epsilon_p \|y^*\|$, where $y^* \in Y^*$ is an arbitrary Lagrange multiplier for (36). It is worth noting that some other types of near-optimal solutions of (36), for example, the primal-dual near-optimal solutions defined in [4], can also be obtained by applying the quadratic penalty approach.

We are now ready to state the iteration-complexity bounds for the RM-SA and the AC-SA algorithm, applied to the penalization problem (39), to compute an (ϵ_p, ϵ_o)-primal solution of (36).

Theorem 4 *Let y^* be an arbitrary Lagrange multiplier for (36). Also let the confidence level $\eta \in (0,1)$ and the accuracy tolerance $(\epsilon_p, \epsilon_o) \in \Re_{++} \times \Re_{++}$ be given. If*

$$\rho = \rho(t) := \left(\frac{\sqrt{\epsilon_o + 4\epsilon_p t} + \sqrt{\epsilon_o}}{\sqrt{2}\epsilon_p} \right)^2 \tag{43}$$

for some $t \geq \|y^*\|$, then, with probability greater than $1 - \eta$,

a) the RM-SA algorithm applied to (39) finds an (ϵ_p, ϵ_o)-primal solution of (36) in at most

$$N_{rm}(t) := \left\lceil \max\left\{ 2R(t)^2, (4\sqrt{5} + 6\lambda)^2 S \right\} \right\rceil \tag{44}$$

iterations;

b) the AC-SA algorithm applied to (39) finds an (ϵ_p, ϵ_o)-primal solution of (36) in at most

$$N_{ac}(t) := \left\lceil \max\left\{ \sqrt{2}R(t), (8\sqrt{5} + 20\lambda)^2 S \right\} \right\rceil \tag{45}$$

iterations,

where λ satisfies $\exp(-\lambda^2/3) + \exp(-\lambda) \leq \eta$ (clearly $\lambda = \mathcal{O}(1)\log 1/\eta$),

$$R(t) := \frac{\sqrt{\rho(t)}\|\mathcal{A}\|\Omega}{\sqrt{\epsilon_o}}, \quad S := \left(\frac{\Omega M}{\epsilon_o} \right)^2, \tag{46}$$

Ω and M are given by (12) and (38) respectively.

We now make a few observations regarding Theorem 4. First, the choice of ρ given by (43) requires that $t \geq \|y^*\|$ and that the iteration-complexity bounds $N_{rm}(t)$ and $N_{ac}(t)$ obtained in Theorem 4 are non-decreasing with respect to t. Second, since the quantity $\|y^*\|$ is not known a priori, it is necessary to guess the value of t. Note however that the influence of t, whence $\|y^*\|$, on the bound $N_{ac}(t)$ is much weaker than that on the bound $N_{rm}(t)$. For example, assume that

$\epsilon_p = \epsilon_o = \epsilon$. By some straightforward computation, it can be easily seen that the value of $N_{ac}(t)$ does not change when

$$\|y^*\| \leq t \leq \frac{1}{4}\left[\left(\frac{(8\sqrt{5}+20\lambda)^2\Omega M^2}{\|A\|\epsilon} - 1\right)^2 - 1\right],$$

while the range of t that does not affect $N_{rm}(t)$ is given by

$$\|y^*\| \leq t \leq \frac{1}{4}\left[\left(\frac{(4\sqrt{5}+6\lambda)M}{\|\mathcal{A}\|} - 1\right)^2 - 1\right].$$

In other words, the AC-SA algorithm allows a big range for t (or $\|y^*\|$), as high as $\mathcal{O}(1/\epsilon^2)$, without affecting the effort to find good approximate solutions of (36), while the corresponding one for the RM-SA algorithm is much smaller, roughly in $\mathcal{O}(1)$. Moreover, even if t does affect the bounds $N_{ac}(t)$ or $N_{rm}(t)$ (i.e., t sits outside the ranges described above), the first bound is in $\mathcal{O}(R(t))$ while the latter one is in $\mathcal{O}(R(t)^2)$.

4 Convergence analysis

The goal of this section is to prove the main results of this paper, namely, Theorems 1, 2, and 4.

4.1 Convergence analysis for RM-SA algorithm

This subsection is to devoted to the proof of Theorem 1. Before proving this result, we establish a few technical results from which Theorem 1 immediately follows.

Let $p(u)$ be a convex function over a convex set $X \in \mathcal{E}$. Assume that $\hat{u}$ is an optimal solution of the problem $\min\{p(u) + \|u - \tilde{x}\|^2 : u \in X\}$ for some $\tilde{x} \in X$. Due to the well-known fact that the sum of a convex and a strongly convex function is also strongly convex, one can easily see that

$$p(u) + \|u - \tilde{x}\|^2 \geq \min\{p(u) + \|u - \tilde{x}\|^2 : u \in X\} + \|u - \hat{u}\|^2.$$

The next lemma generalizes this result to the case where the function $\|u - \tilde{x}\|^2$ is replaced with the prox-function $V(\tilde{x}, u)$ associated with a convex function ω. It is worth noting that the result described below does not assume the strong-convexity of the function ω.

Lemma 5 *Let X be a convex set in $\mathcal{E}$ and $p, \omega : X \to \Re$ be differentiable convex functions. Assume that $\hat{u}$ is an optimal solution of $\min\{p(u) + V(\tilde{x}, u) : u \in X\}$. Then,*

$$\min\{p(u) + V(\tilde{x}, u) : u \in X\} \leq p(u) + V(\tilde{x}, u) - V(\hat{u}, u), \quad \forall u \in X.$$

Proof. The definition of $\hat{u}$ and the fact that $p(\cdot) + V(\tilde{x}, \cdot)$ is a differentiable convex function imply that

$$\langle \nabla p(\hat{u}) + \nabla V(\tilde{x}, \hat{u}), u - \hat{u} \rangle \geq 0, \quad \forall u \in X,$$

where $\nabla V(\tilde{x}, \hat{u})$ denotes the gradient of $V(\tilde{x}, \cdot)$ at $\hat{u}$. Using the definition of the prox-function (10), it is easy to verify that

$$V(\tilde{x}, u) = V(\tilde{x}, \hat{u}) + \langle \nabla V(\tilde{x}, \hat{u}), u - \hat{u} \rangle + V(\hat{u}, u), \quad \forall u \in X.$$

Using the above two relations and the assumption that p is convex, we then conclude that

$$\begin{aligned} p(u) + V(\tilde{x}, u) &= p(u) + V(\tilde{x}, \hat{u}) + \langle \nabla V(\tilde{x}, \hat{u}), u - \hat{u} \rangle + V(\hat{u}, u)] \\ &\geq p(\hat{u}) + V(\tilde{x}, \hat{u}) + \langle \nabla p(\hat{u}) + \nabla V(\tilde{x}, \hat{u}), u - \hat{u} \rangle + V(\hat{u}, u) \\ &\geq p(\hat{u}) + V(\tilde{x}, \hat{u}) + V(\hat{u}, u), \end{aligned}$$

and hence that the lemma holds. ∎

The following lemma summarizes some properties of the objective function Ψ and f.

Lemma 6 *Let the functions $\Psi : X \to \Re$ and $f : X \to \Re$ be defined in (1). We have*

$$0 \leq f(y) - f(x) - \langle \nabla f(x), y - x \rangle \leq \tfrac{L}{2} \|y - x\|^2 \tag{47}$$

$$0 \leq \Psi(y) - \Psi(x) - \langle \Psi'(x), y - x \rangle \leq \tfrac{L}{2} \|y - x\|^2 + 2M\|y - x\| \tag{48}$$

for any $x, y \in X$, where $\Psi'(x) \in \partial \Psi(x)$.

Proof. The first inequalities in both relations (47) and (48) follow immediately from the convexity of f and Ψ respectively. The second inequality in (47) is well-known (see Theorem 2.1.5 of [9] for a proof). This inequality, together with the fact $h(y) - h(x) \leq M\|y - x\|$ due to the Lipschitz-continuity of h and the identity $\Psi'(x) = \nabla f(x) + h'(x)$ for some $h'(x) \in \partial h(x)$, then imply that

$$\begin{aligned} \Psi(y) &= f(y) + h(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2 + h(x) + M\|y - x\| \\ &= \Psi(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2 + M\|y - x\| \\ &= \Psi(x) + \langle \Psi'(x), y - x \rangle + \frac{L}{2}\|y - x\|^2 + M\|y - x\| - \langle h'(x), y - x \rangle \\ &\leq \Psi(x) + \langle \Psi'(x), y - x \rangle + \frac{L}{2}\|y - x\|^2 + 2M\|y - x\|, \end{aligned}$$

where the last inequality follows from (8) with $g = h'(x)$ and $d = x - y$. ∎

The following lemma establishes an important recursion for the RM-SA algorithm. Before stating this result, we mention the following simple inequality that will be used more than once in this section:

$$bu - \frac{au^2}{2} \leq \frac{b^2}{2a}, \ \forall a > 0. \tag{49}$$

Lemma 7 *Assume that the step-sizes γ_τ satisfy $L\gamma_\tau < \alpha$, $\tau \geq 1$. Let $x_1, \cdots, x_\tau \in X$ be given and $(x_{\tau+1}, x^{ag}_{\tau+1}) \in X \times X$ be a pair computed according (16) and (17). Also let $\delta_\tau := G(x_\tau, \xi_\tau) - g(x_\tau)$, where $g(x_\tau) = \mathbb{E}[G(x_\tau, \xi_\tau)] \in \partial \Psi(x_\tau)$. Then, we have*

$$\gamma_\tau[\Psi(x_{\tau+1}) - \Psi(x)] + V(x_{\tau+1}, x) \leq V(x_\tau, x) + \Delta_\tau(x), \ \forall\, x \in X, \tag{50}$$

where

$$\Delta_\tau(x) := \gamma_\tau \langle \delta_\tau, x - x_\tau \rangle + \frac{(2M + \|\delta_\tau\|_*)^2 \gamma_\tau^2}{2(\alpha - L\,\gamma_\tau)}. \tag{51}$$

Proof. Denoting $d_\tau := x_{\tau+1} - x_\tau$, due to the strong-convexity of ω, we have $\alpha\|d_\tau\|^2/2 \leq V(x_\tau, x_{\tau+1})$, which together with (48), then imply that

$$\begin{aligned}
\gamma_\tau\Psi(x_{\tau+1}) &\leq \gamma_\tau[\Psi(x_\tau) + \langle g(x_\tau), d_\tau\rangle + \frac{L}{2}\|d_\tau\|^2 + 2M\|d_\tau\|] \\
&= \gamma_\tau[\Psi(x_\tau) + \langle g(x_\tau), d_\tau\rangle] + \frac{\alpha}{2}\|d_\tau\|^2 - \frac{\alpha - L\gamma_\tau}{2}\|d_\tau\|^2 + 2M\gamma_\tau\|d_\tau\| \\
&\leq \gamma_\tau[\Psi(x_\tau) + \langle g(x_\tau), d_\tau\rangle] + V(x_\tau, x_{\tau+1}) - \frac{\alpha - L\gamma_\tau}{2}\|d_\tau\|^2 + 2M\gamma_\tau\|d_\tau\| \\
&= \gamma_\tau[\Psi(x_\tau) + \langle G(x_\tau, \xi_\tau), d_\tau\rangle] - \gamma_\tau\langle\delta_\tau, d_\tau\rangle + V(x_\tau, x_{\tau+1}) - \frac{\alpha - L\gamma_\tau}{2}\|d_\tau\|^2 + 2M\gamma_\tau\|d_\tau\| \\
&\leq \gamma_\tau[\Psi(x_\tau) + \langle G(x_\tau, \xi_\tau), d_\tau\rangle] + V(x_\tau, x_{\tau+1}) - \frac{\alpha - L\gamma_\tau}{2}\|d_\tau\|^2 + (2M + \|\delta_\tau\|_*)\gamma_\tau\|d_\tau\| \\
&\leq \gamma_\tau[\Psi(x_\tau) + \langle G(x_\tau, \xi_\tau), d_\tau\rangle] + V(x_\tau, x_{\tau+1}) + \frac{(2M + \|\delta_\tau\|_*)^2\gamma_\tau^2}{2(\alpha - L\gamma_\tau)},
\end{aligned}$$

where the last inequality follows from (49) with $u = \|d_\tau\|$, $b = (2M + \|\delta\|_*)\gamma_\tau$, and $a = \alpha - L\gamma_\tau$.

Moreover, it follows from the identity (16), (11), and Lemma 5 with $\tilde{x} = x_\tau$, $\hat{u} = x_{\tau+1}$, and $p(\cdot) \equiv \gamma_\tau\langle G(x_\tau, \xi_\tau), \cdot - x_\tau\rangle$ that

$$\begin{aligned}
&\gamma_\tau\Psi(x_\tau) + [\gamma_\tau\langle G(x_\tau, \xi_\tau), x_{\tau+1} - x_\tau\rangle + V(x_\tau, x_{\tau+1})] \\
&\quad\leq \gamma_\tau\Psi(x_\tau) + [\gamma_\tau\langle G(x_\tau, \xi_\tau), x - x_\tau\rangle + V(x_\tau, x) - V(x_{\tau+1}, x)] \\
&\quad= \gamma_\tau[\Psi(x_\tau) + \langle g(x_\tau), x - x_\tau\rangle] + \gamma_\tau\langle\delta_\tau, x - x_\tau\rangle + V(x_\tau, x) - V(x_{\tau+1}, x) \\
&\quad\leq \gamma_\tau\Psi(x) + \gamma_\tau\langle\delta_\tau, x - x_\tau\rangle + V(x_\tau, x) - V(x_{\tau+1}, x),
\end{aligned}$$

where the last inequality follows from the convexity of $\Psi(\cdot)$ and the fact $g(x_\tau) \in \partial\Psi(x_\tau)$.

Combining the above two conclusions and rearranging the terms, we obtain (50). ∎

Now let us state the following well-known large-deviation result for the martingale sequence (see for example, Lemma 6 of [5], for a proof).

Lemma 8 *Let $\xi_1, \xi_2, ...$ be a sequence of iid random variables, and $\zeta_t = \zeta_t(\xi_{[t]})$ be deterministic Borel functions of $\xi_{[t]}$ such that $\mathbb{E}_{|\xi_{[t-1]}}[\zeta_t] = 0$ a.s. and $\mathbb{E}_{|\xi_{[t-1]}}[\exp\{\zeta_t^2/\sigma_t^2\}] \leq \exp\{1\}$ a.s., where $\sigma_t > 0$ are deterministic. Then*

$$\forall\Lambda \geq 0 : \text{Prob}\left\{\sum_{t=1}^N \zeta_t > \Lambda\sqrt{\sum_{t=1}^N \sigma_t^2}\right\} \leq \exp\{-\Lambda^2/3\}.$$

We are now ready to prove Theorem 1.

Proof of Theorem 1: Let $\bar{x}$ be an optimal solution of (1). Summing up (50) from $\tau = 1$ to t, we have

$$\begin{aligned}
\sum_{\tau=1}^t [\gamma_\tau(\Psi(x_{\tau+1}) - \Psi^*)] &\leq V(x_1, \bar{x}) - V(x_{t+1}, \bar{x}) + \sum_{\tau=1}^t \Delta_\tau(\bar{x}) \\
&\leq V(x_1, \bar{x}) + \sum_{\tau=1}^t \Delta_\tau(\bar{x}) \leq D_{\omega,X}^2 + \sum_{\tau=1}^t \Delta_\tau(\bar{x}),
\end{aligned}$$

where the last inequality follows from (13), which, in view of the fact

$$\Psi(x_{t+1}^{ag}) \leq (\sum_{\tau=1}^{t} \gamma_\tau)^{-1} \sum_{\tau=1}^{t} \gamma_\tau \Psi(x_{\tau+1}),$$

then implies that

$$\left(\sum_{\tau=1}^{t} \gamma_\tau\right) \left[\Psi(x_{t+1}^{ag}) - \Psi^*\right] \leq D_{\omega,X}^2 + \sum_{\tau=1}^{t} \Delta_\tau(\bar{x}). \tag{52}$$

Denoting $\zeta_\tau := \gamma_\tau \langle \delta_\tau, \bar{x} - x_\tau \rangle$ and observing that

$$\Delta_\tau(\bar{x}) = \zeta_\tau + \frac{(2M + \|\delta_\tau\|_*)^2 \gamma_\tau^2}{2(\alpha - L\,\gamma_\tau)} \leq \zeta_\tau + \frac{\gamma_\tau^2}{\alpha - L\gamma_\tau} \left(4M^2 + \|\delta_\tau\|_*^2\right),$$

we then conclude from (52) that

$$\begin{aligned} \left(\sum_{\tau=1}^{t} \gamma_\tau\right) \left[\Psi(x_{t+1}^{ag}) - \Psi^*\right] &\leq D_{\omega,X}^2 + \sum_{\tau=1}^{t} \left[\zeta_\tau + \frac{\gamma_\tau^2}{\alpha - L\gamma_\tau} \left(4M^2 + \|\delta_\tau\|_*^2\right)\right] \\ &\leq D_{\omega,X}^2 + \sum_{\tau=1}^{t} \left[\zeta_\tau + \frac{2\gamma_\tau^2}{\alpha} \left(4M^2 + \|\delta_\tau\|_*^2\right)\right], \end{aligned} \tag{53}$$

where the last inequality follows from the assumption that $\gamma_t \leq \alpha/(2L)$.

Note that the pair (x_t, x_t^{ag}) is a function of the history $\xi_{[t-1]} := (\xi_1, ..., \xi_{t-1})$ of the generated random process and hence is random. Taking expectations of both sides of (53) and noting that under assumption I, $\mathbb{E}[\|\delta_\tau\|_*^2] \leq Q^2$, and

$$\mathbb{E}_{|\xi_{[\tau-1]}}[\zeta_\tau] = 0, \tag{54}$$

we obtain

$$\left(\sum_{\tau=1}^{t} \gamma_\tau\right) \mathbb{E}\left[\Psi(x_{t+1}^{ag}) - \Psi^*\right] \leq D_{\omega,X}^2 + \frac{2}{\alpha}(4M^2 + Q^2) \sum_{\tau=1}^{t} \gamma_\tau^2,$$

which clearly implies part a).

We now show part b) holds. Clearly, by (54), $\{\zeta_\tau\}_{t\geq 1}$ is a marting-gale sequences. Moreover, it follows from (14) and (15) that

$$\mathbb{E}_{|\xi_{[\tau-1]}} \left[\exp\{\zeta_\tau^2/(2\gamma_\tau \Omega Q)^2\}\right] \leq \mathbb{E}_{|\xi_{[\tau-1]}} \left[\exp\{(2\gamma_\tau \Omega \|\delta_\tau\|_*)^2/(2\gamma_\tau \Omega Q)^2\}\right] \leq 1,$$

The previous two observations, in view of Lemma 8, then imply that

$$\forall \Lambda \geq 0 : \text{Prob}\left\{\sum_{\tau=1}^{t} \zeta_\tau > 2\Lambda \Omega Q \sqrt{\sum_{\tau=1}^{t} \gamma_\tau^2}\right\} \leq \exp\{-\Lambda^2/3\}. \tag{55}$$

Now observe that under Assumption II,

$$\mathbb{E}_{|\xi_{\tau-1}} \left[\exp\{\|\delta_\tau\|_*^2/Q^2\}\right] \leq \exp\{1\}.$$

Setting $\theta_\tau = \gamma_\tau^2/\sum_{\tau=1}^t \gamma_\tau^2$, we have

$$\exp\left\{\sum_{\tau=1}^t \theta_t(\|\delta_\tau\|_*^2/Q^2)\right\} \le \sum_{\tau=1}^t \theta_\tau \exp\{\|\delta_\tau\|_*^2/Q^2\},$$

whence, taking expectations,

$$\mathbb{E}\left[\exp\left\{\sum_{\tau=1}^t \gamma_\tau^2\|\delta_\tau\|_*^2/\left(Q^2\sum_{\tau=1}^t \gamma_\tau^2\right)\right\}\right] \le \exp\{1\}.$$

It then follows from Markov's inequality that

$$\forall \Lambda \ge 0 : \text{Prob}\left\{\sum_{\tau=1}^t \gamma_\tau^2\|\delta_\tau\|_*^2 > (1+\Lambda)Q^2\sum_{\tau=1}^t \gamma_\tau^2\right\} \le \exp\{-\Lambda\}. \tag{56}$$

Combining (53), (55), and (56), and rearranging the terms, we obtain (28). ∎

4.2 Convergence analysis for the AC-SA algorithm

The goal of this subsection is to prove Theorem 2.

In the sequel, with a little abuse of the notation, we use the following entity to denote the error for the computed subgradient at each iteration t of the AC-SA algorithm:

$$\delta_t := G(x_t^{md}, \xi_t) - g(x_t^{md}),$$

where $g'(x_t^{md}) = \mathbb{E}[G(x_t^{md}, \xi_t)] \in \partial\Psi(x_t^{md})$ under Assumption I.

The following lemma establishes an important recursion for the AC-SA algorithm.

Lemma 9 *Assume that the step-sizes β_τ and γ_τ satisfy $\beta_\tau \ge 1$ and $L\gamma_\tau < \alpha\beta_\tau$ for all $\tau \ge 1$. Let $(x_\tau, x_\tau^{ag}) \in X \times X$ be given and set $x_\tau^{md} \equiv \beta_\tau^{-1}x_\tau + (1-\beta_\tau^{-1})x_\tau^{ag}$. Also let $(x_{\tau+1}, x_{\tau+1}^{ag}) \in X \times X$ be a pair computed according to (24) and (25). Then, for every $x \in X$, we have*

$$\beta_\tau\gamma_\tau[\Psi(x_{\tau+1}^{ag}) - \Psi(x)] + V(x_{\tau+1}, x) \le (\beta_\tau - 1)\gamma_\tau[\Psi(x_\tau^{ag}) - \Psi(x)] + V(x_\tau, x) + \hat{\Delta}_\tau,$$

where

$$\hat{\Delta}_\tau = \hat{\Delta}_\tau(x) \;\; := \;\; \gamma_\tau\langle\delta_\tau, x - x_\tau\rangle + \frac{(2M + \|\delta\|_*)^2\beta_\tau\gamma_\tau^2}{2(\alpha\beta_\tau - L\gamma_\tau)}. \tag{57}$$

Proof. Denoting $d_\tau := x_{\tau+1} - x_\tau$, it can be easily seen that

$$x_{\tau+1}^{ag} - x_\tau^{md} = \beta_\tau^{-1}x_{\tau+1} + (1-\beta_\tau^{-1})x_\tau^{ag} - x_\tau^{md} = \beta_\tau^{-1}(x_{\tau+1} - x_\tau) = \beta_\tau^{-1}d_\tau.$$

The above observation together with (48) and the relation $\alpha\|d_\tau\|^2/2 \le V(x_\tau, x_{\tau+1})$ then imply that

$$\begin{aligned}
&\beta_\tau\gamma_\tau\Psi(x_{\tau+1}^{ag}) \le \beta_\tau\gamma_\tau[\Psi(x_\tau^{md}) + \langle g(x_\tau^{md}), x_{\tau+1}^{ag} - x_\tau^{md}\rangle + \tfrac{L}{2}\|x_{\tau+1}^{ag} - x_\tau^{md}\|^2 + 2M\|x_{\tau+1}^{ag} - x_\tau^{md}\|] \\
&= \beta_\tau\gamma_\tau[\Psi(x_\tau^{md}) + \langle g(x_\tau^{md}), x_{\tau+1}^{ag} - x_\tau^{md}\rangle] + \tfrac{L\gamma_\tau}{2\beta_\tau}\|d_\tau\|^2 + 2M\gamma_\tau\|d_\tau\| \\
&\le \beta_\tau\gamma_\tau[\Psi(x_\tau^{md}) + \langle g(x_\tau^{md}), x_{\tau+1}^{ag} - x_\tau^{md}\rangle] + V(x_\tau, x_{\tau+1}) - \tfrac{\alpha\beta_\tau - L\gamma_\tau}{2\beta_\tau}\|d_\tau\|^2 + 2M\gamma_\tau\|d_\tau\|.
\end{aligned}$$

Noting that

$$\begin{aligned}
&\beta_\tau\gamma_\tau[\Psi(x_\tau^{md}) + \langle g(x_\tau^{md}), x_{\tau+1}^{ag} - x_\tau^{md}\rangle] = \beta_\tau\gamma_\tau[\Psi(x_\tau^{md}) + \langle g(x_\tau^{md}), (1-\beta_\tau^{-1})x_\tau^{ag} + \beta_\tau^{-1}x_{\tau+1} - x_\tau^{md}\rangle]\\
&= (\beta_\tau - 1)\gamma_\tau[\Psi(x_\tau^{md}) + \langle g(x_\tau^{md}), x_\tau^{ag} - x_\tau^{md}\rangle] + \gamma_\tau[\Psi(x_\tau^{md}) + \langle g(x_\tau^{md}), x_{\tau+1} - x_\tau^{md}\rangle]\\
&\le (\beta_\tau - 1)\gamma_\tau\Psi(x_\tau^{ag}) + \gamma_\tau[\Psi(x_\tau^{md}) + \langle g(x_\tau^{md}), x_{\tau+1} - x_\tau^{md}\rangle]\\
&= (\beta_\tau - 1)\gamma_\tau\Psi(x_\tau^{ag}) + \gamma_\tau[\Psi(x_\tau^{md}) + \langle G(x_\tau^{md}, \xi_\tau), x_{\tau+1} - x_\tau^{md}\rangle - \langle \delta_\tau, x_{\tau+1} - x_\tau^{md}\rangle]\\
&= (\beta_\tau - 1)\gamma_\tau\Psi(x_\tau^{ag}) + \gamma_\tau[\Psi(x_\tau^{md}) + \langle G(x_\tau^{md}, \xi_\tau), x_{\tau+1} - x_\tau^{md}\rangle - \langle \delta_\tau, x_\tau - x_\tau^{md}\rangle - \langle \delta_\tau, d_\tau\rangle]\\
&\le (\beta_\tau - 1)\gamma_\tau\Psi(x_\tau^{ag}) + \gamma_\tau[\Psi(x_\tau^{md}) + \langle G(x_\tau^{md}, \xi_\tau), x_{\tau+1} - x_\tau^{md}\rangle - \langle \delta_\tau, x_\tau - x_\tau^{md}\rangle + \|\delta_\tau\|_*\|d_\tau\|],
\end{aligned}$$

we conclude from the previous conclusion that

$$\begin{aligned}
&\beta_\tau\gamma_\tau\Psi(x_{\tau+1}^{ag}) \le (\beta_\tau - 1)\gamma_\tau\Psi(x_\tau^{ag}) + \gamma_\tau[\Psi(x_\tau^{md}) + \langle G(x_\tau^{md}, \xi_\tau), x_{\tau+1} - x_\tau^{md}\rangle] + V(x_\tau, x_{\tau+1})\\
&-\gamma_\tau\langle \delta_\tau, x_\tau - x_\tau^{md}\rangle - \tfrac{\alpha\beta_\tau - L\gamma_\tau}{2\beta_\tau}\|d_\tau\|^2 + (2M + \|\delta\|_*)\gamma_\tau\|d_\tau\|\\
&\le (\beta_\tau - 1)\gamma_\tau\Psi(x_\tau^{ag}) + \gamma_\tau[\Psi(x_\tau^{md}) + \langle G(x_\tau^{md}, \xi_\tau), x_{\tau+1} - x_\tau^{md}\rangle] + V(x_\tau, x_{\tau+1})\\
&-\gamma_\tau\langle \delta_\tau, x_\tau - x_\tau^{md}\rangle + \tfrac{(2M+\|\delta\|_*)^2\beta_\tau\gamma_\tau^2}{2(\alpha\beta_\tau - L\gamma_\tau)},
\end{aligned}$$

where the last inequality follows from (49) with $u = \|d_\tau\|$, $b = (2M + \|\delta\|_*)\gamma_\tau$, and $a = (\alpha\beta_\tau - L\gamma_\tau)/\beta_\tau$.

Moreover, it follows from the identity (24), (11), and Lemma 5 with $\tilde{x} = x_\tau$, $\hat{u} = x_{\tau+1}$, and $p(\cdot) \equiv \gamma_\tau\langle G(x_\tau^{md}, \xi_\tau), \cdot - x_\tau^{md}\rangle$ that

$$\begin{aligned}
&\gamma_\tau\Psi(x_\tau^{md}) + [\gamma_\tau\langle G(x_\tau^{md}, \xi_\tau), x_{\tau+1} - x_\tau^{md}\rangle + V(x_\tau, x_{\tau+1})]\\
&\quad\le\quad \gamma_\tau\Psi(x_\tau^{md}) + [\gamma_\tau\langle G(x_\tau^{md}, \xi_\tau), x - x_\tau^{md}\rangle + V(x_\tau, x) - V(x_{\tau+1}, x)]\\
&\quad=\quad \gamma_\tau[\Psi(x_\tau^{md}) + \langle g(x_\tau^{md}), x - x_\tau^{md}\rangle] + \gamma_\tau\langle \delta_\tau, x - x_\tau^{md}\rangle + V(x_\tau, x) - V(x_{\tau+1}, x)\\
&\quad\le\quad \gamma_\tau\Psi(x) + \gamma_\tau\langle \delta_\tau, x - x_\tau^{md}\rangle + V(x_\tau, x) - V(x_{\tau+1}, x),
\end{aligned}$$

where the last inequality follows from the convexity of $\Psi(\cdot)$ and the fact $g(x_\tau^{md}) \in \partial\Psi(x_\tau^{md})$.

Combining the previous two conclusions, we obtain

$$\beta_\tau\gamma_\tau\Psi(x_{\tau+1}^{ag}) \le (\beta_\tau - 1)\gamma_\tau\Psi(x_\tau^{ag}) + \gamma_\tau\Psi(x) + V(x_\tau, x) - V(x_{\tau+1}, x) + \gamma_\tau\langle \delta_\tau, x - x_\tau\rangle + \tfrac{(2M+\|\delta\|_*)^2\beta_\tau\gamma_\tau^2}{2(\alpha\beta_\tau - L\gamma_\tau)}$$

Our claim immediately follows from the above inequality by subtracting $\beta_\tau\gamma_\tau\Psi(x)$ from both sides and rearranging the terms. ∎

We are now ready to prove Theorem 2.

Proof of Theorem 2: Let $\bar{x}$ be an optimal solution of (1). It follows from the fact that $\Psi(x) \ge \Psi(\bar{x}) = \Psi^*$, $\forall x \in X$, the fact $\beta_\tau \ge 1$, (26), and Lemma 9 with $x = \bar{x}$ that, for any $t \ge 1$,

$$\begin{aligned}
(\beta_{t+1} - 1)\gamma_{t+1}[\Psi(x_{t+1}^{ag}) - \Psi^*] &\le \beta_t\gamma_t[\Psi(x_{t+1}^{ag}) - \Psi^*]\\
&\le\quad (\beta_t - 1)\gamma_t[\Psi(x_t^{ag}) - \Psi^*] + V(x_t, \bar{x}) - V(x_{t+1}, \bar{x}) + \hat{\Delta}_t(\bar{x}),
\end{aligned}$$

from which it follows inductively that

$$\begin{aligned}
(\beta_{t+1} - 1)\gamma_{t+1}[\Psi(x_{t+1}^{ag}) - \Psi^*] &\le (\beta_1 - 1)\gamma_1[\Psi(x_1^{ag}) - \Psi^*] + V(x_1, \bar{x}) - V(x_{t+1}, \bar{x}) + \sum_{\tau=1}^{t}\hat{\Delta}_\tau(\bar{x})\\
&= V(x_1, \bar{x}) - V(x_{t+1}, \bar{x}) + \sum_{\tau=1}^{t}\hat{\Delta}_\tau(\bar{x}) \le D_{\omega,X}^2 + \sum_{\tau=1}^{t}\hat{\Delta}_\tau(\bar{x}),
\end{aligned}$$

where the first equality follows from the assumption $\beta_1 = 1$ and the last inequality follows from (13) and the fact $V(x_{t+1}, \bar{x}) \geq 0$.

Denoting $\zeta_\tau := \gamma_\tau \langle \delta_\tau, \bar{x} - x_\tau \rangle$ and observing that

$$\begin{aligned} \hat{\Delta}_\tau(\bar{x}) &= \zeta_\tau + \frac{(2M + \|\delta\|_*)^2 \beta_\tau \gamma_\tau^2}{2(\alpha\beta_\tau - L\gamma_\tau)} \leq \zeta_\tau + \frac{\beta_\tau \gamma_\tau^2}{\alpha\beta_\tau - L\gamma_\tau}(4M^2 + \|\delta_\tau\|_*^2) \\ &\leq \zeta_\tau + \frac{2}{\alpha}(4M^2 + \|\delta_\tau\|_*^2)\gamma_\tau^2, \end{aligned}$$

where the last inequality follows from (26), we then conclude from the previous observation that

$$(\beta_{t+1} - 1)\gamma_{t+1}[\Psi(x_{t+1}^{ag}) - \Psi(\bar{x})] \leq D_{\omega,X}^2 + \sum_{\tau=1}^{t} \left[\zeta_\tau + \frac{2}{\alpha}(4M^2 + \|\delta_\tau\|_*^2)\gamma_\tau^2 \right]. \tag{58}$$

Note that the triple $(x_t, x_t^{ag}, x_t^{md})$ is a function of the history $\xi_{[t-1]} := (\xi_1, ..., \xi_{t-1})$ of the generated random process and hence is random. Taking expectations of both sides of (58) and noting that under assumption I, $\mathbb{E}[\|\delta_\tau\|_*^2] \leq Q^2$ and $\mathbb{E}_{|\xi_{[\tau-1]}}[\zeta_\tau] = 0$, we obtain

$$(\beta_{t+1} - 1)\gamma_{t+1}\mathbb{E}[\Psi(x_{t+1}^{ag}) - \Psi^*] \leq D_{\omega,X}^2 + \frac{2}{\alpha}(4M^2 + Q^2)\sum_{\tau=1}^{t} \gamma_\tau^2,$$

which clearly implies part a).

The proof of part b) is similar to the one of Theorem 1.b), and hence the details are skipped. ■

4.3 Convergence analysis for quadratic penalty method

The goal of this subsection is to prove Theorem 4.

Lemma 10 *If $\tilde{x} \in X$ is an approximate solution of (39) satisfying*

$$\tilde{\Psi}_\rho(\tilde{x}) - \tilde{\Psi}^* \leq \delta, \tag{59}$$

then

$$\|\mathcal{A}\tilde{x} - b\| \leq \frac{2}{\rho}\|y^*\| + \sqrt{\frac{2\delta}{\rho}} \tag{60}$$

$$\tilde{h}(\tilde{x}) - \tilde{h}^* \leq \delta, \tag{61}$$

where y^* is an arbitrary Lagrange multiplier associated with (36).

Proof. Denote $v(u) := \inf\{\tilde{h}(x) : \mathcal{A}x - b = u,\ x \in X\}$. It is well-known that our assumptions imply that v is a convex function such that $-y^* \in \partial v(0)$. Hence,

$$v(u) - v(0) \geq (-y^*)^T u,\ \forall u \in \Re^m.$$

Letting $u := \mathcal{A}\tilde{x} - b$, we conclude from the above observation, the facts that $v(u) \leq \tilde{h}(\tilde{x})$ and $v(0) \geq \tilde{\Psi}^*$, and assumption (59), that

$$\begin{aligned} -\|y^*\|\|u\| + \rho\|u\|^2/2 &\leq (-y^*)^T u + \rho\|u\|^2/2 \\ &\leq v(u) - v(0) + \rho\|u\|^2/2 \leq \tilde{h}(\tilde{x}) + \rho\|u\|^2/2 - v(0) \\ &\leq \tilde{h}(\tilde{x}) + \rho\|u\|^2/2 - \phi^* = \tilde{\Psi}_\rho(\tilde{x}) - \tilde{\Phi}^* \leq \delta, \end{aligned}$$

which clearly implies (60). Moreover, the fact that $\tilde{h}^* = v(0) \geq \tilde{\Phi}^*$ implies that

$$\tilde{h}(\tilde{x}) - \tilde{h}^* \leq \tilde{h}(\tilde{x}) + \rho\|u\|^2/2 - \Phi^* = \tilde{\Psi}_\rho(\tilde{x}) - \tilde{\Phi}^* \leq \delta.$$

■

We are now ready to prove Theorem 4.

Proof of Theorem 4: Let $\tilde{x} \in X$ satisfies (59) with $\delta = \epsilon_o$. Let $\rho_* := \rho(\|y^*\|)$ and observe that $\rho_* \leq \rho(t)$ for every $t \geq \|y^*\|$. It follows from the previous observation and Lemma 10 that $\tilde{h}(\tilde{x}) - \tilde{h}^* \leq \epsilon_o$ and

$$\begin{aligned} \|A\tilde{x} - b\| &\leq \frac{2}{\rho(t)}\|y^*\| + \sqrt{\frac{2\epsilon_o}{\rho(t)}} \leq \frac{2}{\rho_*}\|y^*\| + \sqrt{\frac{2\epsilon_o}{\rho_*}} = \frac{1}{\sqrt{\rho_*}}\left(\frac{2\sqrt{2}\epsilon_p\|y^*\|}{\sqrt{\epsilon_o + 4\epsilon_p\|y^*\|} + \sqrt{\epsilon_o}} + \sqrt{2\epsilon_o}\right) \\ &= \frac{1}{\sqrt{\rho_*}}\left(\frac{\sqrt{\epsilon_o + 4\epsilon_p\|y^*\|} - \sqrt{\epsilon_o}}{\sqrt{2}} + \sqrt{2\epsilon_o}\right) = \frac{\sqrt{\epsilon_o} + \sqrt{\epsilon_o + 4\epsilon_p\|y^*\|}}{\sqrt{2\rho_*}} = \epsilon_p, \end{aligned}$$

and hence that $\tilde{x}$ is an (ϵ_p, ϵ_o)-primal solution of (36).

Using Assumption III.c), (40), and (41), we have $L = \rho\|\mathcal{A}\|^2$ and $Q = M$, which together with (21) and (22) then imply that

$$K_0^*(N_{rm}) + \lambda K_1^*(N_{rm}) = \frac{\rho\|\mathcal{A}\|^2\Omega^2}{N_{rm}} + \frac{2\sqrt{5} + 3\lambda}{\sqrt{N_{rm}}}\Omega M \leq \frac{\epsilon_o}{2} + \frac{\epsilon_o}{2} \leq \epsilon_o.$$

The previous conclusion, in view of the definition of λ and (22), clearly imply the claim in part a). Part b) follows similarly from (32) and the definition of λ, by noting that

$$\begin{aligned} \hat{K}_0^*(N_{ac}) + \lambda\hat{K}_1^*(N_{ac}) &= \frac{4\rho\|\mathcal{A}\|^2\Omega^2}{N_{ac}(N_{ac}+2)} + \frac{4\sqrt{5} + 10\lambda}{\sqrt{N_{ac}}}\Omega M \leq \frac{4\rho\|\mathcal{A}\|^2\Omega^2}{N_{ac}^2} + \frac{4\sqrt{5} + 10\lambda}{\sqrt{N_{ac}}}\Omega M \\ &\leq \frac{\epsilon_o}{2} + \frac{\epsilon_o}{2} \leq \epsilon_o. \end{aligned}$$

■

5 Concluding remarks

In this paper, two subgradient-type methods for solving the stochastic composite problem (1), namely: the RM-SA and AC-SA algorithm, are analyzed and compared. From the theoretical view of perspective, our contribution is to close the gap between the lower and upper bounds on the rate of convergence for solving this class of problems. As shown in the illustrative example, we believe that the AC-SA algorithm developed herein possesses significant potential of application, for

example, in stochastic programming, large-scale convex programming, as well as in those traditional areas for stochastic approximation, e.g., statistics, digital signal processing.

Acknowledgement: The author would like to express the sincerest appreciation to Professor Arkadi Nemirovski, for the motivating discussion and very insightful comments on some results of this paper. The author is also very grateful to Professors Renato Monteiro and Alex Shapiro for their patient guidance during the Ph.D study, which paved the way for conducting this research.

References

[1] Ben-Tal, A. and Nemirovski, A., Non-euclidean restricted memory level method for large-scale convex optimization, *Mathematical Programming*, **102**, 407-456 (2005).

[2] Juditsky, A., Lan, G., Nemirovski, A., and Shapiro, A. (2007), Stochastic approximation approach to stochastic programming, submitted to *SIAM Journal on Optimization E-print available at:* `http://www.optimization-online.org`.

[3] Juditsky, A., Nemirovski, A., and Tauvel, C. (2008), Solving variational inequalities with stochastic mirror-prox algorithm, submitted to *SIAM Journal on Control and Optimization*

[4] Lan, G. and Monteiro, R. (2008), Iteration-complexity of first-order penalty methods for convex programming, submitted to *Mathematical Programming*, *E-print available at:* `http://www.optimization-online.org`.

[5] Lan, G., Nemirovski, A., and Shapiro, A. (2008), Validation analysis of robust stochastic approximation method, submitted to *Mathematical Programming*, *E-print available at:* `http://www.optimization-online.org`.

[6] Lan, G., Lu, Z., and Monteiro, R. (2006), Primal-dual first-order methods with $\mathcal{O}(1/\epsilon)$ iteration-complexity for Cone Programing, submitted to *Mathematical Programming*, *E-print available at :* `http://www.optimization-online.org`.

[7] Nemirovski, A., Yudin, D., *Problem complexity and method efficiency in optimization*, Wiley-Interscience Series in Discrete Mathematics, John Wiley, XV, 1983.

[8] Nesterov, Y. , A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$, *Doklady AN SSSR*, **269**, 543-547 (1983).

[9] Nesterov, Y., *Introductory Lectures on Convex Optimization: a basic course*, Kluwer Academic Publishers, Massachusetts, 2004.

[10] Shapiro, A., Monte Carlo sampling methods, in: Ruszczyński, A. and Shapiro, A., (Eds.), *Stochastic Programming*, Handbook in OR & MS, Vol. 10, North-Holland Publishing Company, Amsterdam, 2003.