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

# An Optimal Method for Stochastic Composite Optimization

**Guanghui Lan**

March 06, 2010 (Revised: November 02, 2010)

**Abstract** This paper considers an important class of convex programming (CP) problems, namely, the stochastic composite optimization (SCO), whose objective function is given by the summation of general nonsmooth and smooth stochastic components. Since SCO covers non-smooth, smooth and stochastic CP as certain special cases, a valid lower bound on the rate of convergence for solving these problems is known from the classic complexity theory of convex programming. Note however that the optimization algorithms that can achieve this lower bound had never been developed. In this paper, we show that the simple mirror-descent stochastic approximation method exhibits the best-known rate of convergence for solving these problems. Our major contribution is to introduce the accelerated stochastic approximation (AC-SA) algorithm based on Nesterov's optimal method for smooth CP [32,34], and show that the AC-SA algorithm can achieve the aforementioned lower bound on the rate of convergence for SCO. To the best of our knowledge, it is also the first universally optimal algorithm in the literature for solving non-smooth, smooth and stochastic CP problems. We illustrate the significant advantages of the AC-SA algorithm over existing methods in the context of solving a special but broad class of stochastic programming problems.

**Keywords:** stochastic approximation, convex optimization, stochastic programming, complexity, optimal method, quadratic penalty method, large deviation

**AMS 2000 subject classification:** 62L20, 90C25, 90C15, 68Q25,

The work of the author was partially supported by NSF Grants CMMI-1000347, CCF-0430644 and CCF-0808863, and ONR Grant N000140811104 and N00014-08-1-0033.

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

## 1 Introduction

The basic problem of interest in this paper is the stochastic composite optimization (SCO) given by

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

where $X$ is a convex compact set in the Euclidean space $\mathbb{R}^n$ endowed with inner product $\langle \cdot, \cdot \rangle$ and a norm $\|\cdot\|$ (not necessarily the one associated with inner product), $f : X \to \mathbb{R}$ is a convex function with Lipschitz continuous gradient, that is,

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

($\|\cdot\|_*$ denotes the conjugate norm, see Subsection 1.1), and $h : X \to \mathbb{R}$ is a convex Lipschitz continuous function such that

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

We assume that problem (1) is to be solved by iterative algorithms which acquire the subgradients of $\Psi$ 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 (also independent of search points $x_t$) whose probability distribution $P$ is supported on $\Xi \subseteq \mathbb{R}^d$. The following assumptions are made for the Borel functions $G(x, \xi_t)$.

**A1:** For any $x \in X$, we have

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

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

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

Observe that problem (1) covers several important classes of convex programming problems as certain special cases. For the sake of simplicity, we assume in the following discussion that the domain $X$ is a standard Euclidean ball. Note however that such a simplification is only of local scope for this section only, and that the paper does deal with more general feasible set $X$, allowing non-Euclidean geometry.

*Case I: non-smooth convex optimization* Suppose that the smooth component $f$ in $\Psi$ does not exist, or equivalently $f(x) = 0$ for every $x \in X$, and that there is no noise in the $\mathcal{SO}$, i.e., $\sigma = 0$ in (5). Then, problem (1) becomes the generic non-smooth convex optimization problem that has been well-studied in the Literature. According to Nemirovski and Yudin [31], if the dimension $n$ is sufficiently large, i.e., $n \geq \mathcal{O}(1)N$, where $\mathcal{O}(1)$ denotes an absolute constant, then the rate of convergence for any iterative algorithms to solve nonsmooth convex optimization problems can not be better than $\Psi(\hat{x}_N) - \Psi^* \leq \mathcal{O}(1)(\mathcal{M}/\sqrt{N})$, where $N$ is the number of iterations performed by the algorithm and $\hat{x}_N \in X$ denotes the solution generated by the algorithm after $N$ steps. Moreover, the simple subgradient descent method can achieve, up to a constant factor, the above lower bound. Note that the subgradient descent method is closely related to the gradient projection method of Goldstein and Levitin, Polyak (see [6]). Nemirovski and Yudin [31] also developed the so-called mirror descent algorithm that can be advantageous over the subgradient descent method when $X$ is not an Euclidean ball by using a prox-function (also called Bregman's distance, which was studied by Bregman [7] and many others, see for example, [1,2,15,49] and references therein).

*Case II: smooth convex optimization* Suppose that the non-smooth component $h$ in $\Psi$ does not exist, or equivalently $h(x) = 0$ for every $x \in X$, and that there is no noise in the $\mathcal{SO}$, i.e., $\sigma = 0$ in (5). Then, problem (1) becomes the smooth convex optimization problem. In [31], Nemirovski and Yudin show that, if the dimension $n$ is sufficiently large, i.e., $n \geq \mathcal{O}(1)N$, then the rate of convergence for any iterative algorithms to solve smooth convex optimization problems can not be better than $\Psi(\hat{x}_N) - \Psi^* \leq \mathcal{O}(1)(L/N^2)$. In a seminal work ([32]), Nesterov presented novel smooth convex optimization algorithms whose rate of convergence is bounded by $\Psi(\hat{x}_N) - \Psi^* \leq \mathcal{O}(1)(L/N^2)$. Clearly, Nesterov's methods are optimal, up to a constant factor, for smooth convex optimization when $n \geq \mathcal{O}(1)N$. Nesterov's methods were further studied in [33], [1] and [34] using Bregman's distance and other variants of Nesterov's optimal method can also be found, for example, in [18] and [50].

*Case III: stochastic convex optimization* Suppose that the variance of the $\mathcal{SO}$ is positive, i.e., $\sigma > 0$. Then, problem (1) becomes the stochastic convex optimization problem. There exist two competitive computational approaches for solving stochastic convex optimization based on Monte Carlo sampling techniques, namely, the *Stochastic Approximation* (SA) and the *Sample Average Approximation* (SAA) methods. Both approaches, the SA and SAA methods, have a long history. The classic SA method mimics the gradient descent method and goes back to the pioneering paper by Robbins and Monro [42]. Since then stochastic approximation algorithms became widely used in stochastic optimization (see, e.g., [5,10,11,39,44,17,47] and references therein). An important improvement of the SA method was developed by Polyak [40] and Polyak and Juditsky [41], where longer stepsizes were suggested with consequent averaging of the obtained iterates. In these classical SA-type algorithms, it is assumed that the objective function is twice continuously differentiable and strongly convex. On the other hand, the SAA approach was used by many authors in various contexts under different names. Its basic idea is rather simple: generate a (random) sample $\xi_1, ..., \xi_N$ of size $N$, and approximate the "true" problem (1) by the so-called sample average problem. Recent theoretical studies (cf., [16,45,46]) and numerical experiments (see, e.g., [23,27, 51]) show that the SAA method coupled with a good (deterministic) algorithm could be reasonably efficient for solving certain classes of two stage stochastic programming problems. Current opinion is that the SAA method can efficiently use a specific (say linear) structure of the considered problem, while the SA approach is a crude subgradient method which often performs poorly in practice. Very recently, Nemirovski et. al. [30] demonstrated that a properly modified SA method based on iterate averaging, namely, the mirror-descent SA, can be competitive and even significantly outperform the SAA method for a certain class of stochastic optimization problems. Moreover, this algorithm exhibits an unimprovable rate of convergence $\mathbb{E}[\Psi(\hat{x}_N) - \Psi^*] \leq \mathcal{O}(1)(\mathcal{M} + \sigma)/\sqrt{N}$ for non-smooth stochastic convex optimization (i.e., $f(x) = 0$ for every $x \in X$ in (1) and $\sigma > 0$) even when the dimension $n = 1$ (note that this differs from the two above-mentioned deterministic optimization cases where the lower bounds on the rate of convergence are valid only if $n$ is sufficiently large [31]). Close techniques, based on subgradient averaging, have been proposed in Nesterov [35] and used in [12,14] to solve certain non-smooth stochastic convex optimization problems.

Since SCO covers these subcases described above, we can easily see that the rate of convergence for any iterative algorithms to solve (1) can not be better than

$$\mathbb{E}[\Psi(\hat{x}_N) - \Psi^*] \leq \mathcal{O}(1)\left[\frac{L}{N^2} + \frac{\mathcal{M} + \sigma}{\sqrt{N}}\right], \tag{6}$$

where $\hat{x}_N$ denotes the solution obtained after $N$ 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. However, to the best of our knowledge, none of the existing algorithms achieved this lower bound on the convergence rate. Since the objective function $\Psi$ of (1) is a non-smooth function, we can directly apply the mirror-descent SA [30] to (1) and the resulting rate of convergence (cf. (25)) can not be better than

$$\mathbb{E}[\Psi(\hat{x}_N) - \Psi^*] \leq \mathcal{O}(1)\left[\frac{L + \mathcal{M} + \sigma}{\sqrt{N}}\right]. \tag{7}$$

The best known result so far is given by Juditsky et. al. [13] with the rate of convergence

$$\mathbb{E}[\Psi(\hat{x}_N) - \Psi^*] \leq \mathcal{O}(1)\left[\frac{L}{N} + \frac{\mathcal{M} + \sigma}{\sqrt{N}}\right] \tag{8}$$

by applying an extra-gradient-type algorithm to a variational inequality (v.i.) reformulation of (1). It should be noted that the lower bound on rate of convergence has not been attained even for the deterministic case where $\sigma = 0$. Moreover, with only access to the $\mathcal{SO}$ of the composite function $\Psi$ itself, it is unclear whether the lower bound (6) on the rate of convergence for solving (1) is achievable or not.

To motivate our discussion, we would like to explain why we care about the gap between the convergence rates (7), (8) and the lower bound (6). Suppose that we have an optimal algorithm for solving (1) which achieves the lower bound (6) on the convergence rate. First of all, this algorithm will be a universally optimal method for non-smooth, smooth and stochastic convex optimization. Currently different classes of convex optimization problems are being handled by using different (sub)optimal methods. More specifically, while the mirror descent SA [30] is optimal for deterministic or stochastic non-smooth convex optimization, and Nesterov's method [32] is optimal for deterministic smooth convex optimization, we do not have optimal algorithms for solving general stochastic smooth convex optimization and stochastic composite optimization problems. This is partly due to the difficulty that, although either smooth or nonsmooth optimization has been well-studied separately in the literature, a unified treatment for both of them seems highly non-trivial. Secondly, this optimal algorithm for SCO will allow us to have a very large Lipschitz constant $L$ for problem (1) without affecting the rate of convergence. More specifically, the convergence rates in (7) and (8), will not be affected (up to a constant factor 2), if $L$ is as big as $\mathcal{M} + \sigma$ and $(\mathcal{M} + \sigma)N^{\frac{1}{2}}$, respectively. It can also be easily seen from (6) that the convergence rate of the optimal algorithm will not change (up to a constant factor 2) if $L \leq (\mathcal{M}+\sigma)N^{\frac{3}{2}}$. Clearly, the latter range of $L$ is much bigger than those for the previous two methods and extends much faster as the number of iterations $N$ grows.

Our contribution in this work mainly consists of the following aspects. Firstly, with a novel analysis, it is demonstrated that a slightly modified mirror descent SA algorithm applied to (1) also exhibits the best known so far rate of convergence guaranteed by a more involved stochastic mirror-prox algorithm [13]. Moreover, by substantially generalizing Nesterov's optimal method to non-smooth and stochastic convex optimization, we propose an accelerated SA (AC-SA), which can achieve the theoretically optimal rate of convergence for solving SCO. To the best of our knowledge, it is also the first universally optimal algorithm in the literature for solving non-smooth, smooth and

stochastic convex optimization problems. We also investigate the AC-SA algorithm 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 stochastic oracle is "light-tailed". Thirdly, we illustrate the significant advantages of the AC-SA algorithm over the existing algorithms 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. More specifically, if either the mirror descent or the AC-SA algorithm is applied to solve the quadratic penalization problem, where the violation of the affine constraints is penalized, then the size of the Lagrange multiplier associated with these affine constraints has much smaller effect on the AC-SA algorithm than that on the former one. We has thus shown that the penalty-based approaches are promising for solving our particular problems of interest, while these approaches are generally criticized for their dependence on the unknown size of Lagrange multipliers.

We should distinguish the results obtained in this paper with some related but different development in the literature for solving problems given in the form of (1). In 2005, Nesterov in an important paper [34] presented a first-order method with convergence rate bounded by $\mathcal{O}(1/N)$ to solve a special class of problems given in the form of (1), where the nonsmooth term $h$ is given by

$$h(x) := \sup\{\langle \mathcal{B}y, x\rangle - \phi(y) : y \in Y\},$$

$Y \subseteq \mathbb{R}^m$ is a compact convex set, $\phi : Y \to \mathbb{R}$ is a continuous convex function and $\mathcal{B}$ is a linear operator from $\mathbb{R}^m$ to $\mathbb{R}^n$. Nesterov's approach consists of approximating the nonsmooth function $h$ by a sufficiently close smooth one with Lipschitz continuous gradient and applying the optimal smooth method in [32,34] to the resulting problem with $h$ replaced by its approximation function. In a subsequent paper, Nemirovski [29] proposed an extra-gradient type first-order method for solving a slightly more general class of optimization problems than the one considered by Nesterov [34] and also established an $\mathcal{O}(1/N)$ convergence rate for his method. These first-order methods were further studied in, for example, [1,37,26,8,38,50,28], and successfully used in sparse covariance selection, rank reduction in multivariate linear regression and compressed sensing, etc. (see, for example, [9,24,25,3]). Another line of investigation is also to consider problems given in the form of (1) where the non-smooth component $h$ of $\Psi$ in (1) is sufficiently simple, for example, $h(x) = \|x\|_1$, where $\|\cdot\|_1$ denotes the $l_1$ norm, so that the non-smooth component can be kept as a part of the prox-step (or projection in the Euclidean case) (Nesterov [36], Tseng [50], Lewis and Wright [22]). Consequently, the convergence rate for solving these problems is the same as that of smooth convex optimization, i.e., $\mathcal{O}(1/N^2)$. The above two lines of development differ from ours in the following aspects: (i) those problems considered in [34,29,36] and related references are certain special cases of (1) in the sense that the nonsmooth term $h$ can either be smoothed or kept in the projection. Therefore, it turned out that stronger convergence results can be obtained for those subcases. We, on the other hand, consider a general non-smooth term $h$ in the objective function of (1); (ii) the algorithms developed in [34,36] and related references need to access the smooth and non-smooth component of the composite function $\Psi$ separately. In contrast, our method, in addition to using the structure of the problem, only need to access the composite function itself; iii) in [34, 29,36] and other references mentioned above, only deterministic optimization problems have been considered. We also focus on the situation where the subgradients of $\Psi$ are contaminated by stochastic noise; iv) it should be noted that the development in [34,

36] can be easily incorporated into our method for certain cases where the nonsmooth component $h$ in (1) consists of the aforementioned special structures.

The paper is organized as follows. In Section 2, we present a slightly modified mirror descent SA algorithm applied to (1) and describe its convergence properties. We then discuss in Section 3 the accelerated stochastic approximation algorithm. More specifically, we present the AC-SA algorithm and describe its convergence properties in Subsection 3.1, and 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 $\phi : X \to \mathbb{R}$, its subdifferential $\partial f(\cdot)$ is defined as follows: at a point $x$ from the relative interior of $X$, $\partial\phi$ is comprised of all subgradients $g$ of $\phi$ 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 $x = \lim_{i\to\infty} x_i$, $g = \lim_{i\to\infty} g_i$. Finally, $\partial\phi(x) = \emptyset$ for $x \notin X$. With this definition, it is well-known (see, for example, Ben-Tal and Nemirovksi [4]) that, if a convex function $\phi : X \to \mathbb{R}$ is Lipschitz continuous, with constant $\mathcal{M}$, with respect to a norm $\|\cdot\|$, then the set $\partial\phi(x)$ is nonempty for any $x \in X$ and

$$g \in \partial\phi(x) \Rightarrow |\langle g, d\rangle| \leq \mathcal{M}\|d\|, \ \forall\, d \in \operatorname{lin}(X - X), \tag{9}$$

  in other words,

$$g \in \partial\phi(x) \Rightarrow \|g\|_* \leq \mathcal{M}, \tag{10}$$

  where $\|\cdot\|_*$ denotes the conjugate norm given by $\|g\|_* := \max_{\|d\|\leq 1}\langle g, d\rangle$.
- 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 expectation, $\xi_{[t]}$ being given.

## 2 Modified mirror-descent stochastic approximation

In this section, we present a modified version of the mirror-descent SA method in [30] and demonstrate that it can achieve the best known so far rate of convergence for solving problem (1).

### 2.1 Preliminaries: distance generating function and prox-mapping

Let $\|\cdot\|$ be a (general) norm on $\mathbb{R}^n$ and $\|x\|_* = \sup_{\|y\|\leq 1}\langle y, x\rangle$ be its dual norm. We say that a function $\omega : X \to \mathbb{R}$ is a *distance generating function* modulus $\alpha > 0$ with respect to $\|\cdot\|$, if $\omega$ is convex and continuous on $X$, the set

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

is convex (note that $X^o$ always contains the relative interior of $X$), and restricted to $X^o$, $\omega$ is continuously differentiable and strongly convex with parameter $\alpha$ with respect to $\|\cdot\|$, i.e.,

$$\langle \nabla\omega(x') - \nabla\omega(x), x' - x\rangle \geq \alpha\|x' - x\|^2, \ \ \forall x', x \in X^o. \tag{11}$$

The simplest example of a distance generating function is $\omega(x) = \frac{1}{2}\|x\|_2^2$ (modulus 1 with respect to $\|\cdot\|_2$, $X^o = X$).

Let us define function $V : X^o \times X \to \mathbb{R}_+$ as follows

$$V(x, z) = \omega(z) - [\omega(x) + \langle \nabla\omega(x), z - x\rangle]. \tag{12}$$

In what follows we shall refer to $V(\cdot,\cdot)$ as *prox-function* associated with distance generating function $\omega(x)$. The prox-function $V(\cdot,\cdot)$ is also called the Bregman's distance, which was initially studied by Bregman [7] and later by many others (see [1,2,15,49] and references therein). In particular, Auslender and Teboule [1] give a more geometrically transparent definition for this notion. Note that $V(x,\cdot)$ is nonnegative and is strongly convex modulus $\alpha$ with respect to the norm $\|\cdot\|$. Let us define *prox mapping* $P_x : \mathbb{R}^n \to X^o$, associated with $\omega$ and a point $x \in X^o$, viewed as a parameter, as follows:

$$P_x(y) = \arg\min_{z\in X} \big\{\langle y, z - x\rangle + V(x, z)\big\}. \tag{13}$$

Observe that the minimum in the right hand side of (13) is attained since $\omega$ is continuous on $X$ and $X$ is compact, and all the minimizers belong to $X^o$, whence the minimizer is unique, since $V(x,\cdot)$ is strongly convex on $X^o$. Thus, the prox-mapping is well defined.

The distance generating function $\omega$ also gives rise to the following characteristic entity that will be used frequently in our convergence analysis:

$$D_{\omega,X} := \sqrt{\max_{x\in X}\omega(x) - \min_{x\in X}\omega(x)}. \tag{14}$$

Let $x_1$ be the minimizer of $\omega$ 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 $\omega$ implies that

$$\frac{\alpha}{2}\|x - x_1\|^2 \leq V(x_1, x) \leq \omega(x) - \omega(x_1) \leq D^2_{\omega,X}, \ \ \forall x \in X, \tag{15}$$

and hence

$$\|x - x_1\| \leq \Omega_{\omega,X} := \sqrt{\frac{2}{\alpha}} D_{\omega,X} \text{ and } \|x - x'\| \leq 2\Omega_{\omega,X}, \ \forall x, x' \in X. \tag{16}$$

## 2.2 The algorithm and its convergence properties

The mirror descent SA algorithm, as applied to (1), works with the stochastic oracle of $\Psi$ that satisfies Assumption A1. In some cases, Assumption A1 is augmented by the following "light-tail" assumption.

**A2:** For any $x \in X$, we have

$$\mathbb{E}\left[\exp\{\|G(x,\xi_t) - g(x)\|_*^2/\sigma^2\}\right] \leq \exp\{1\}. \tag{17}$$

It can be easily seen that Assumption A2 implies Assumption A1(b), since by Jensen's inequality,

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

We are now ready to describe the modified mirror descent SA algorithm for solving (1).

**The modified mirror descent SA algorithm:**

0) Let the initial point $x_1$ and the step-sizes $\{\gamma_t\}_{t\ge1}$ be given. Set $t=1$;
1) Call the $\mathcal{SO}$ for computing $G(x_t,\xi_t)$. Set

$$x_{t+1} := P_{x_t}\left(\gamma_t G(x_t,\xi_t)\right), \tag{18}$$

$$x_{t+1}^{av} = \left(\sum_{\tau=1}^{t}\gamma_\tau\right)^{-1}\sum_{\tau=1}^{t}\gamma_\tau x_{\tau+1}. \tag{19}$$

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 $\omega$ over X (see Subsection 2.1). Secondly, observe that the above algorithm only differs from the mirror descent SA algorithm in [30] in the way the averaging step (19) is defined. More specifically, the sequence $\{x_t^{av}\}_{t\ge2}$ is obtained by averaging the iterates $x_t, t\ge2$ with their corresponding weights $\gamma_{t-1}$, while the one in [30] is obtained by taking the average of the whole trajectory $x_t, t\ge1$ with weights $\gamma_t$. It should be mentioned that the above the shifted weights are incorporated mainly for the purpose of facilitating our convergence analysis (more discussions can be found in the comments after the proof of Lemma 3). In fact, if the constant stepsizes are used, i.e., $\gamma_t=\gamma, \forall\, t\ge1$, then the averaging step stated above is exactly the same as the one stated in [30] up to shifting one iterate. The major modification that we made to the standard mirror-descent SA is to define a new stepsize policy which takes into account the structure of problem (1) (c.f. (23) below).

The following result summarizes some convergence properties of the above mirror descent SA algorithm. The proof of this result will be given in Subsection 4.1.

**Theorem 1** *Assume that the step-sizes $\gamma_t$ satisfy $0<\gamma_t\le\alpha/(2L)$, $\forall\, t\ge1$. Let $\{x_{t+1}^{av}\}_{t\ge1}$ be the sequence computed according to (19) by the modified mirror descent SA algorithm. Then we have*

*a) under Assumption A1,*

$$\mathbb{E}\left[\Psi(x_{t+1}^{av})-\Psi^*\right] \le K_0(t),\ \forall\, t\ge1, \tag{20}$$

*where*

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

*$\mathcal{M}$, $\sigma$ and $D_{\omega,X}$ are given in (3), (5) and (14) respectively;*

*b) under Assumptions A1 and A2,*

$$\text{Prob}\left\{\Psi(x_{t+1}^{av}) - \Psi^* > K_0(t) + \Lambda K_1(t)\right\} \le \exp\{-\Lambda^2/3\} + \exp\{-\Lambda\},\ \forall\, \Lambda > 0, t \ge 1, \tag{21}$$

*where*

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

*$\sigma$ and $\Omega_{\omega,X}$ are given in (5) and (14) respectively.*

We now describe a possible strategy for the selection of the stepsizes $\{\gamma_t\}$ for the modified mirror descent SA. 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 the *constant step-size policy* 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 $x_{N+1}^{av} = N^{-1}\sum_{\tau=1}^{N} x_{\tau+1}$ satisfies

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

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

$$\mathbb{E}\left[\Psi(x_{N+1}^{av}) - \Psi^*\right] \le K_0^*(N) := \frac{L\Omega_{\omega,X}^2}{N} + \frac{2\Omega_{\omega,X}\sqrt{4\mathcal{M}^2+\sigma^2}}{\sqrt{N}}, \tag{22}$$

by choosing $\gamma$ as

$$\gamma = \min\left\{\frac{\alpha}{2L}, \sqrt{\frac{\alpha D_{\omega,X}^2}{2N(4\mathcal{M}^2+\sigma^2)}}\right\}. \tag{23}$$

Moreover, with this choice of $\gamma$, we have

$$\begin{aligned} K_1(N) &= \frac{2\Omega_{\omega,X}\sigma}{\sqrt{N}} + \frac{2\gamma\sigma^2}{\alpha} \le \frac{2\Omega_{\omega,X}\sigma}{\sqrt{N}} + \sqrt{\frac{2}{\alpha}}D_{\omega,X}\frac{\sigma^2}{\sqrt{N(4\mathcal{M}^2+\sigma^2)}} \\ &\le \frac{2\Omega_{\omega,X}\sigma}{\sqrt{N}} + \sqrt{\frac{2}{\alpha}}D_{\omega,X}\frac{\sigma}{\sqrt{N}} = \frac{3\Omega_{\omega,X}\sigma}{\sqrt{N}}, \end{aligned}$$

hence, bound (21) implies that

$$\text{Prob}\left\{\Psi(x_{N+1}^{av}) - \Psi^* > K_0^*(N) + \Lambda K_1^*(N)\right\} \le \exp\{-\Lambda^2/3\} + \exp\{-\Lambda\},\ \forall\, \Lambda > 0, \tag{24}$$

where

$$K_1^*(N) := \frac{3\Omega_{\omega,X}\sigma}{\sqrt{N}}.$$

*Remark 1* It is interesting to compare the rate of convergence (22) obtained for the modified mirror descent SA and the one for a direction application of the mirror descent SA in [30]. We first derive the rate of convergence for a direction application of the mirror descent SA in [30] to problem (1). Let $g(x_t) = \mathbb{E}[G(x_t, \xi_t)] = \nabla f(x_t) + h'(x_t)$ for every $t \geq 1$, where $h'(x_t) \in \partial h(x_t)$. Then, in view of assumptions (2), (3), (4) and (5), and relation (10), we have

$$\begin{aligned}
\mathbb{E}[\|G(x_t, \xi_t)\|_*^2] &= \mathbb{E}[\|g(x_t) + G(x_t, \xi_t) - g(x_t)\|_*^2] \\
&\leq 2\|g(x_t)\|_*^2 + 2\mathbb{E}[\|G(x_t, \xi_t) - g(x_t)\|_*^2] \\
&\leq 2\|\nabla f(x_t) + h'(x_t)\|_*^2 + 2\sigma^2 \leq 4\|\nabla f(x_t)\|_*^2 + 4\mathcal{M}^2 + 2\sigma^2 \\
&\leq 4\|\nabla f(x_1) + \nabla f(x_t) - \nabla f(x_1)\|_*^2 + 4\mathcal{M}^2 + 2\sigma^2 \\
&\leq 8\|\nabla f(x_1)\|_*^2 + 8\|\nabla f(x_t) - \nabla f(x_1)\|_*^2 + 4\mathcal{M}^2 + 2\sigma^2 \\
&\leq 8\|\nabla f(x_1)\|_*^2 + 8L^2\|x_t - x_1\|^2 + 4\mathcal{M}^2 + 2\sigma^2 \\
&\leq 8\|\nabla f(x_1)\|_*^2 + 8L^2\Omega_{\omega,X}^2 + 4\mathcal{M}^2 + 2\sigma^2,
\end{aligned}$$

which, in view of relations (2.46) or (2.56) of [30], implies that the rate of convergence for a direct application of the mirror descent SA algorithm to (1) is bounded by

$$\mathcal{O}(1)\left[\frac{\Omega_{\omega,X}(\|\nabla f(x_1)\|_* + L\Omega_{\omega,X} + \mathcal{M} + \sigma)}{\sqrt{N}}\right]. \tag{25}$$

Clearly, the above bound is always worse than the one given in (22). In particular, in the range

$$L \leq \frac{\sqrt{N(4\mathcal{M}^2 + \sigma^2)}}{\Omega_{\omega,X}}, \tag{26}$$

the first component in (22) (for abbreviation, the $L$-component) merely does not affect the error estimate (22). Note that the range stated in (26) extends as $N$ increases, meaning that, if $N$ is large, 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 (25).

Indeed, such an improvement comes from the new stepsize we just defined in (23). Observe that if the first term in the definition of $\gamma$ in (23) dominates, i.e., when $L$ is large, then $\gamma$ in (23) does not depend on $N$. It follows from this fact that the rate of convergence of the modified mirror descent SA algorithm will be $\mathcal{O}(1)1/N$, similarly to the gradient descent method for smooth convex optimization. However, in the original mirror-descent SA algorithm, the stepsize given by (c.f., (2.47) of [30])

$$\gamma = \sqrt{\frac{\alpha D_{\omega,X}^2}{2N(4\mathcal{M}^2 + \sigma^2)}} \tag{27}$$

does not consider the possibility of the existence of a dominant smooth component in the objective of (1). As a result, the rate of convergence of the the original mirror-descent SA algorithm with stepsize (27) will always be $\mathcal{O}(1)1/\sqrt{N}$. ∎

It should be noted that the mirror descent SA is a direct descendant of the mirror descent algorithm [31]. It is well-known that algorithms of this type are not optimal for smooth convex optimization and hence can not be optimal for SCO. On the other hand, as discussed in Section 1, Nesterov's methods [32,33] and its variants were shown to be optimal for solving smooth convex optimization problems. We will investigate a possible extension of Nesterov's method for solving problem (1) in the next section.

## 3 Accelerated stochastic approximation

In this section, we provide a substantial generalization of Nesterov's methods ([32,33]) to solve non-smooth and stochastic convex optimization. As a result, we develop a new SA-type algorithm, referred to as the accelerated SA (AC-SA) method, which can achieve the theoretically optimal rate of convergence for solving (1). More specifically, we will state the algorithm and its convergence results in Subsection 3.1 and illustrate its significant advantages over the existing SA algorithms 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

$$x_{t+1} = P_{x_t}(\gamma_t G(x_t^{md}, \xi_t)), \tag{28}$$
$$x_{t+1}^{ag} = \beta_t^{-1} x_{t+1} + (1-\beta_t^{-1}) x_t^{ag}, \tag{29}$$

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, similarly to the mirror descent SA algorithm, we assume that the initial point $x_1$ is the minimizer of $\omega$ over X. 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 mirror descent SA algorithm, that is, each iteration of the above algorithm requires only one call to the $\mathcal{SO}$ and one solution of the subproblem (28). Thirdly, while Nesterov's optimal method and its variants [32–34, 1, 18, 50] were designed for solving deterministic smooth convex optimization problems, the AC-SA algorithm described above is capable of solving non-smooth and stochastic convex optimization problems as well.

*Remark 2* Note that the above AC-SA algorithm is a little simpler than the original Nesterov's method in [32–34], whose comprehension seems a bit difficult. Here we would like to provide some intuitive explanation of the updates of the AC-SA algorithm, driven from the point view of the proximal-type algorithms. For the sake of simplicity, let us suppose that the first-order information is exact, i.e., $\sigma = 0$. Observe that the points $x_t^{md}$, $t \geq 1$, are used to construct certain model functions of $\Psi(\cdot)$ in (1), namely,

$$m_t(x) := f(x_t^{md}) + \langle g(x_t^{md}), x - x_t^{md} \rangle.$$

The search points $x_t$, $t \geq 1$, can be viewed as the prox-centers and the stepsizes $\gamma_t$, $t \geq 1$, control the instability of the model $m_t(x)$ so that we will not move too far away from the current prox-center while taking step (28). The search points $x_t^{ag}$, $t \geq 1$, are used to evaluate the objective values. Since

$$f(x_{t+1}^{ag}) \leq \beta_t^{-1} f(x_{t+1}) + (1 - \beta_t^{-1}) f(x_t^{ag}),$$

the function value of $x_{t+1}^{ag}$ might be smaller than that of $x_{t+1}$. An immediate improvement of the algorithm would be to take $x_{t+1}^{ag}$ as the one with the smallest objective value among the following three points, $x_t^{ag}$, $x_{t+1}$ and $\beta_t^{-1} x_{t+1} + (1 - \beta_t^{-1}) x_t^{ag}$, provided that these function values can be easily computed. The essence of Nesterov's methods, as well as the AC-SA algorithm, is to coordinate the building of the model function $m_t(x)$, the selection of the prox-center $x_t$ and the evaluation of the objective value through a careful selection of $\beta_t$ and $\gamma_t$, $t \geq 1$. ∎

The following theorem states the main convergence results of the AC-SA algorithm applied to stochastic composite optimization. The proof of this result will be given in Subsection 4.2.

**Theorem 2** *Assume that the stepsizes $\beta_t \in [1, \infty)$ and $\gamma_t \in \mathbb{R}_+$ are chosen such that $\beta_1 = 1$ and the following conditions hold*

$$0 < (\beta_{t+1} - 1)\gamma_{t+1} \leq \beta_t \gamma_t \text{ and } 2L\gamma_t \leq \alpha \beta_t, \ \forall t \geq 1. \tag{30}$$

*Let $\{x_{t+1}^{ag}\}_{t \geq 1}$ be the sequence computed according to (29) by the AC-SA algorithm. Then we have*

*a) under Assumption A1,*

$$\mathbb{E}[\Psi(x_{t+1}^{ag}) - \Psi^*] \leq \hat{K}_0(t), \ \forall\, t \geq 1, \tag{31}$$

*where*

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

*$\mathcal{M}$, $\sigma$ and $D_{\omega,X}$ are given in (3), (5) and (14) respectively;*

*b) under Assumptions A1 and A2,*

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

*where*

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

*$\sigma$ and $\Omega_{\omega,X}$ are given in (5) and (14) respectively.*

*Remark 3* It is interesting to note the similarity between the results stated in Theorem 1 for the modified mirror-descent SA and those obtained in Theorem 2 for the accelerated SA. Comparing (20) with (31) (resp., (21) with (32)), we can easily that the only difference exists in the factors of $K_0(t)$ and $\hat{K}_0(t)$ (resp., $K_1(t)$ and $\hat{K}_1(t)$). More specifically, the factor $1/\sum_{\tau=1}^{t}\gamma_t$ in $K_0(t)$ and $K_1(t)$ is replaced by $1/[(\beta_{t+1}-1)\gamma_{t+1}]$ in $\hat{K}_0(t)$ and $\hat{K}_1(t)$ while all other terms are the same. ∎

We now discuss the determination of the stepsizes $\beta_t$ and $\gamma_t$ in the accelerated SA 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 (30) is given by:

$$\beta_t = \frac{t+1}{2} \text{ and } \gamma_t = \frac{t+1}{2}\gamma \tag{33}$$

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

**Corollary 1** *Suppose that the stepsizes $\beta_t$ and $\gamma_t$ in the AC-SA algorithm are set to*

$$\beta_t = \frac{t+1}{2}, \quad \gamma_t = \frac{t+1}{2}\min\left\{\frac{\alpha}{2L}, \frac{\sqrt{6\alpha}D_{\omega,X}}{(N+2)^{\frac{3}{2}}(4\mathcal{M}^2+\sigma^2)^{\frac{1}{2}}}\right\}, \ \forall t \geq 1, \tag{34}$$

*where $N$ is a fixed in advance number of iterations. Then, we have under Assumption A1,*

$$\mathbb{E}[\Psi(x_{N+1}^{ag}) - \Psi^*] \leq \hat{K}_0^*(N) := \frac{4L\Omega_{\omega,X}^2}{N(N+2)} + \frac{4\Omega_{\omega,X}\sqrt{4\mathcal{M}^2+\sigma^2}}{\sqrt{N}}, \tag{35}$$

*if in addition, Assumption A2 holds, then*

$$\text{Prob}\left\{\Psi(x_{N+1}^{ag}) - \Psi^* > \hat{K}_0^*(N) + \Lambda\hat{K}_1^*(N)\right\} \leq \exp\{-\Lambda^2/3\} + \exp\{-\Lambda\}, \ \forall\, \Lambda > 0, \tag{36}$$

*where*

$$\hat{K}_1^*(N) := \frac{6\Omega_{\omega,X}\sigma}{\sqrt{N}}.$$

*Proof.* Clearly, the stepsizes $\{\beta_t\}_{t\geq 1}$ and $\{\gamma_t\}_{t\geq 1}$ stated in (34) satisfy the conditions $\beta_1 = 1$, $\beta_t > 1, \forall t \geq 2$, and (30). Denoting

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

we then conclude from Theorem 2 that, under Assumption A1,

$$\mathbb{E}[\Psi(x_{N+1}^{ag}) - \Psi^*] \leq \mathcal{T}_0 := \frac{4D_{\omega,X}^2}{N(N+2)\gamma^*} + \frac{8\gamma^*(4\mathcal{M}^2+\sigma^2)}{\alpha N(N+2)}\sum_{\tau=1}^{N}\left(\frac{\tau+1}{2}\right)^2, \tag{37}$$

and that, under Assumptions A1 and A2,

$$\text{Prob}\left\{\Psi(x_{N+1}^{ag}) - \Psi^* > \mathcal{T}_0 + \Lambda\mathcal{T}_1)\right\} \leq \exp\{-\Lambda^2/3\} + \exp\{-\Lambda\}, \ \forall\, \Lambda > 0, \tag{38}$$

where

$$\mathcal{T}_1 := \frac{8\Omega_{\omega,X}\sigma}{N(N+2)}\sqrt{\sum_{\tau=1}^{N}\left(\frac{\tau+1}{2}\right)^2} + \frac{8\gamma^*\sigma^2}{N(N+2)\alpha}\sum_{\tau=1}^{N}\left(\frac{\tau+1}{2}\right)^2.$$

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

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

and

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

Our claim immediately follows by substituting the above bounds of $\mathcal{T}_0$ and $\mathcal{T}_1$ into (37) and (38). ∎

We now make a few observations regarding the results obtained in Theorem 2 and Corollary 1. Firstly, it is interesting to compare bounds (35) and (22) obtained for the AC-SA algorithm and the mirror descent 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 (35)) does not affect the error estimate. Specifically, within the range

$$L \leq \frac{\sqrt{4\mathcal{M}^2+\sigma^2}N^{\frac{3}{2}}}{\Omega_{\omega,X}}, \tag{39}$$

which extends much faster than (26) 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 1 still hold when the Lipschitz constant $L = 0$. More specifically, we consider the case where $f(x) = 0$ for any $x \in X$. In this case, the stepsizes $\{\beta_t\}_{t\geq 1}$ and $\{\gamma_t\}_{t\geq 1}$ in (34) become

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

and the error estimate (35) reduces to

$$\mathbb{E}[h(x_{N+1}^{ag}) - h^*] \leq \frac{4\Omega_{\omega,X}\sqrt{4\mathcal{M}^2+\sigma^2}}{\sqrt{N}},$$

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

$$x_{N+1}^{ag} = \frac{2}{N+1}x_{N+1} + \frac{N-1}{N+1}x_N^{ag} = \frac{\sum_{t=1}^{N}(tx_{t+1})}{\sum_{t=1}^{N}t}.$$

Hence, in contrast to the usual *constant stepsize* or *decreasing stepsize* policy (see [30]), the stepsizes $\gamma_t$ in step (28) and the weights for taking the average in step (29) are

increasing with the increment of $t$. It is worth noting that the aforementioned increasing stepsize policy applies directly to the search points $x_t$, $t \geq 2$, while a previously defined increasing stepsize policy in [35] is used for aggregating the subgradients that have been computed.

Finally, note that if there is no stochastic error for the computed subgradient of $\Psi$, i.e., $\sigma = 0$, then it follows from (35) that

$$\Psi(x_{N+1}^{ag}) - \Psi^* \leq \frac{4L\Omega_{\omega,X}^2}{N(N+2)} + \frac{8\Omega_{\omega,X}\mathcal{M}}{\sqrt{N}},$$

which implies 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 literature of deterministic convex optimization.

### 3.2 Application to stochastic programming

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

Consider the problem of

$$\begin{array}{rl} \tilde{h}^* := \min\limits_{x} & \left\{\tilde{h}(x) := \mathbb{E}[\tilde{H}(x,\xi)]\right\} \\ \text{s.t.} & \mathcal{A}x - b = 0, \quad x \in X, \end{array} \tag{40}$$

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

$$\mathbb{E}[\tilde{H}(x,\xi)] = \textstyle\int_{\Xi} \tilde{H}(x,\xi)dP(\xi) \tag{41}$$

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, (40) becomes a convex programming problem. We also make the following assumptions:

**A3:**

a) It is possible to generate an iid sample $\xi_1, \xi_2, ...,$ of realizations of random vector $\xi$.
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 $x$. Observe that, under certain regularity assumptions [48,30], $\tilde{\mathcal{H}}(x,\xi)$ can be taken as a measurable subgradient of the function $\tilde{H}(x,\xi)$ in (40).
c) There is a constant $\mathcal{M} > 0$ such that

$$\forall x \in X : \mathbb{E}\left[\exp\{\|\tilde{\mathcal{H}}(x,\xi)\|_*^2/\mathcal{M}^2\}\right] \leq \exp\{1\}. \tag{42}$$

For the case where the feasible region consists only of the simple convex set $X$, or equivalently $\mathcal{A} \equiv 0$, Nemirovski et. al. demonstrated in [30] that the mirror descent SA algorithm can substantially outperform the sampling averaging approximation (Shapiro [45]), a widely used approach for stochastic programming in practice. When $\mathcal{A}$ is not identically 0, the mirror descent SA algorithm can still be applied directly to problem (40) 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. Moreover, the selection of the norm $\|\cdot\|$ and the distance generating function $\omega$ will be problem dependent. In other words, it is not clear what is the optimal way for choosing these parameter settings. More discussions about the parameter settings when the domain is relatively simple were given in [30, 21].

One alternative approach to alleviate this difficulty is to apply the quadratic penalty approach: instead of solving (40), we solve certain penalization problem of (40) obtained by penalizing the violation of the constraint $\mathcal{A}x - b = 0$. In particular, given a penalty parameter $\rho > 0$, we solve

$$\tilde{\Psi}^* = \tilde{\Psi}^*_\rho := \inf_{x \in X} \left\{ \tilde{\Psi}_\rho(x) := \tilde{f}_\rho(x) + \tilde{h}(x) \right\}, \tag{43}$$

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 $\mathbb{R}^m$. Define the operator norm $\|\mathcal{A}\| := \max\{\|\mathcal{A}x\|_* : \|x\| \le 1\}$. It can be easily seen that $\nabla\tilde{f}_\rho(x) = \rho\mathcal{A}^*(\mathcal{A}x - b)$ and hence that

$$\|\nabla\tilde{f}_\rho(x) - \nabla\tilde{f}_\rho(x')\|_* = \rho\|\mathcal{A}^*\mathcal{A}(x - x')\|_* \le \rho\|\mathcal{A}^*\|\|\mathcal{A}\|\|x - x'\| = \rho\|\mathcal{A}\|^2\|x - x'\|, \ \forall\, x, x' \in X, \tag{44}$$

where the last equality follows from the fact that $\|\mathcal{A}\| = \|\mathcal{A}^*\|$. Moreover, in view of Assumption A3 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 $\mathbb{E}[\|\tilde{\mathcal{H}}(x, \xi_t)\|_*^2] \le \mathcal{M}^2$ and hence that $\|\tilde{h}'(x)\|_* = \|\mathbb{E}[\tilde{\mathcal{H}}(x, \xi_t)]\|_* \le \mathcal{M}$, which together with the fact $\tilde{h}(x) - \tilde{h}(x') \le \langle\tilde{h}'(x), x - x'\rangle, \forall x, x' \in X$ due to the convexity of $\tilde{h}$, clearly imply that

$$|\tilde{h}(x) - \tilde{h}(x')| \le \mathcal{M}\|x - x'\|, \ \ \forall x, x' \in X. \tag{45}$$

Therefore, the penalization problem (43) is given in the form of (1) and hence can be approximately solved by either the modified mirror descent SA or the AC-SA algorithm developed in this paper.

It is well-known that the near-optimal solutions of the penalization problem (43) also yield near-optimal solutions of (40) if the penalty parameter $\rho$ is sufficiently large. In this paper, we are interested in obtaining one particular type of near-optimal solutions of (40) defined in the following way. First note that $x^*$ is an optimal solution of (40) if, and only if, $x^* \in X$, $\mathcal{A}x^* - b = 0$ and $\tilde{h}(x^*) \le \tilde{h}^*$. This observation leads us to our definition of a near optimal solution $\tilde{x} \in X$ of (40), 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 [19].

**Definition:** Let $\epsilon_p, \epsilon_o > 0$ be given, $\tilde{x} \in X$ is called an $(\epsilon_p, \epsilon_o)$-*primal solution* for (40) if

$$\|\mathcal{A}\tilde{x} - b\| \le \epsilon_p \text{ and } \tilde{h}(\tilde{x}) - \tilde{h}^* \le \epsilon_0. \tag{46}$$

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 (40)

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

is nonempty. It was observed in [19] 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 (40). It is worth noting that some other types of near-optimal solutions of (40), for example, the primal-dual near-optimal solutions defined in [19], can also be obtained by applying the quadratic penalty approach.

We are now ready to state the iteration-complexity bounds for the modified mirror descent SA and the AC-SA algorithm, applied to the penalization problem (43), to compute an $(\epsilon_p, \epsilon_o)$-primal solution of (40).

**Theorem 3** *Let $y^*$ be an arbitrary Lagrange multiplier for (40). Also let the confidence level $\eta \in (0, 1)$ and the accuracy tolerance $(\epsilon_p, \epsilon_o) \in \mathbb{R}_{++} \times \mathbb{R}_{++}$ be given. If*

$$\rho = \rho(\nu) := \left( \frac{\sqrt{\epsilon_o + 4\epsilon_p \nu} + \sqrt{\epsilon_o}}{\sqrt{2}\epsilon_p} \right)^2 \tag{47}$$

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

*a) the modified mirror descent SA algorithm applied to (43) finds an $(\epsilon_p, \epsilon_o)$-primal solution of (40) in at most*

$$N_{md}(\nu) := \left\lceil \max\left\{ 2R(\nu)^2, (8\sqrt{2} + 12\lambda)^2 S \right\} \right\rceil \tag{48}$$

*iterations;*

*b) the AC-SA algorithm applied to (43) finds an $(\epsilon_p, \epsilon_o)$-primal solution of (40) in at most*

$$N_{ac}(\nu) := \left\lceil \max\left\{ 2\sqrt{2}R(\nu), (16\sqrt{2} + 24\lambda)^2 S \right\} \right\rceil \tag{49}$$

*iterations,*

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

$$R(\nu) := \frac{\sqrt{\rho(\nu)}\|\mathcal{A}\|\Omega}{\sqrt{\epsilon_o}}, \quad S := \left( \frac{\Omega\mathcal{M}}{\epsilon_o} \right)^2, \tag{50}$$

*$\Omega$ and $\mathcal{M}$ are given by (14) and (42) respectively.*

We now make a few observations regarding Theorem 3. First, note that the choice of $\rho$ given by (47) requires $\nu \geq \|y^*\|$ and that the iteration-complexity bounds $N_{md}(\nu)$ and $N_{ac}(\nu)$ obtained in Theorem 3 are non-decreasing with respect to $\nu$. Second, since the quantity $\|y^*\|$ is not known a priori, it is necessary to guess the value of $\nu$. Note however that the influence of $\nu$, whence $\|y^*\|$, on the bound $N_{ac}(\nu)$ is much weaker than that on the bound $N_{md}(\nu)$. For example, assume that $\epsilon_p = \epsilon_o = \epsilon$. By some

straightforward computation, it can be easily seen that the value of $N_{ac}(\nu)$ does not change when

$$\|y^*\| \leq \nu \leq \frac{1}{4}\left[\left(\frac{(16\sqrt{2}+24\lambda)^2\Omega M^2}{\|A\|\epsilon}-1\right)^2-1\right],$$

while the range of $\nu$ that does not affect $N_{md}(\nu)$ is given by

$$\|y^*\| \leq \nu \leq \frac{1}{4}\left[\left(\frac{(8\sqrt{2}+12\lambda)M}{\|\mathcal{A}\|}-1\right)^2-1\right].$$

In other words, the AC-SA algorithm allows a big range for $\nu$ (or $\|y^*\|$), as high as $\mathcal{O}(1/\epsilon^2)$, without affecting the effort to find good approximate solutions of (40), while the corresponding one for the modified mirror descent SA algorithm is much smaller, roughly in $\mathcal{O}(1)$. As a consequence, when $\epsilon \downarrow 0$, the size of Lagrange multiplier associated with the affine constraints does not affect, asymptotically, the rate of convergence for the AC-SA algorithm applied to (40). Finally, even if $\nu$ does affect the bounds $N_{ac}(\nu)$ or $N_{md}(\nu)$ (i.e., $\nu$ sits outside the ranges described above), the first bound is in $\mathcal{O}(R(\nu))$ while the latter one is in $\mathcal{O}(R(\nu)^2)$.

## 4 Convergence analysis

The goal of this section is to prove the main results of this paper, namely, Theorems 1, 2, and 3.

### 4.1 Convergence analysis for the modified mirror descent SA

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

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

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 $\omega$. It can be viewed as a Bregman version of "growth formula" for strongly convex functions and is based on a Pythagora like formula for Bregman distances; It is worth noting that the result described below does not assume the strong-convexity of the function $\omega$.

**Lemma 1** *Let $X$ be a convex set in $\mathcal{E}$ and $p,\omega : X \to \mathbb{R}$ 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 (12), it is easy to verify the following known *three points identity*:

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

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 $\Psi$ and $f$.

**Lemma 2** *Let the functions $\Psi : X \to \mathbb{R}$ and $f : X \to \mathbb{R}$ be defined in (1). We have*

$$0 \leq f(y) - f(x) - \langle \nabla f(x), y - x \rangle \leq \tfrac{L}{2} \|y - x\|^2 \tag{51}$$

$$0 \leq \Psi(y) - \Psi(x) - \langle \Psi'(x), y - x \rangle \leq \tfrac{L}{2} \|y - x\|^2 + 2\mathcal{M}\|y - x\| \tag{52}$$

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

*Proof.* The first inequalities in both relations (51) and (52) follow immediately from the convexity of $f$ and $\Psi$ respectively. The second inequality in (51) is well-known (see Theorem 2.1.5 of [33] for a proof). This inequality, together with the fact $h(y) - h(x) \leq \mathcal{M}\|y - x\|$ 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

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

where the last inequality follows from (9) with $g = h'(x)$ and $d = x - y$. ∎

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

$$bu - \frac{au^2}{2} \leq \frac{b^2}{2a}, \ \forall a > 0. \tag{53}$$

**Lemma 3** *Assume that the stepsizes $\gamma_\tau$ satisfy $L\gamma_\tau < \alpha$, $\tau \geq 1$. Let $x_1, \cdots, x_\tau \in X$ be given and $(x_{\tau+1}, x_{\tau+1}^{av}) \in X \times X$ be a pair computed according (18) and (19). 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*

$$\gamma_\tau[\Psi(x_{\tau+1}) - \Psi(x)] + V(x_{\tau+1}, x) \leq V(x_\tau, x) + \Delta_\tau(x), \ \forall\, x \in X, \tag{54}$$

*where*

$$\Delta_\tau(x) := \gamma_\tau\langle\delta_\tau, x - x_\tau\rangle + \frac{(2\mathcal{M} + \|\delta_\tau\|_*)^2\gamma_\tau^2}{2(\alpha - L\,\gamma_\tau)}. \tag{55}$$

*Proof.* Denoting $d_\tau := x_{\tau+1} - x_\tau$, due to the strong-convexity of $\omega$, we have $\alpha\|d_\tau\|^2/2 \leq V(x_\tau, x_{\tau+1})$, which together with (52), then imply that

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

where the last inequality follows from (53) with $u = \|d_\tau\|$, $b = (2\mathcal{M} + \|\delta_\tau\|_*)\gamma_\tau$, and $a = \alpha - L\gamma_\tau$.

Moreover, it follows from the identity (18), (13), and Lemma 1 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

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

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 (54). ∎

Note that, while relation (2.36) in [30] bounds the optimality gap of the current iterate $x_t$ in the mirror-descent SA, Lemma 3 provides an estimate on the optimality gap of the next iterate $x_{t+1}$. This is why we need to incorporate the shifted weights in the mirror-descent SA algorithm when a smooth component appears in the objective function. It is possible to derive a relation similar to (2.36) of [30] for unconstrained smooth CP problems, see Lemma 23 of [20]. However, we still do not know how to obtain such a formula here for constrained smooth CP problems.

Now let us state the following well-known result for the martingale-difference sequence. The proof of this result can be found, for example, in Lemma 2 of [21].

**Lemma 4** *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 (54) from $\tau = 1$ to $t$, we have

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

where the last inequality follows from (15), which, in view of the fact that

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

then implies that

$$\left(\sum_{\tau=1}^{t}\gamma_\tau\right)\left[\Psi(x_{t+1}^{av}) - \Psi^*\right] \leq D_{\omega,X}^2 + \sum_{\tau=1}^{t}\Delta_\tau(\bar{x}). \tag{56}$$

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

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

we then conclude from (56) that

$$\begin{aligned}\left(\sum_{\tau=1}^{t}\gamma_\tau\right)\left[\Psi(x_{t+1}^{av}) - \Psi^*\right] &\leq D_{\omega,X}^2 + \sum_{\tau=1}^{t}\left[\zeta_\tau + \frac{\gamma_\tau^2}{\alpha - L\gamma_\tau}\left(4\mathcal{M}^2 + \|\delta_\tau\|_*^2\right)\right] \\ &\leq D_{\omega,X}^2 + \sum_{\tau=1}^{t}\left[\zeta_\tau + \frac{2\gamma_\tau^2}{\alpha}\left(4\mathcal{M}^2 + \|\delta_\tau\|_*^2\right)\right], \end{aligned}\tag{57}$$

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

Note that the pair $(x_t, x_t^{av})$ 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 (57) and noting that under Assumption A1, $\mathbb{E}[\|\delta_\tau\|_*^2] \leq \sigma^2$, and

$$\mathbb{E}_{|\xi_{[\tau-1]}}[\zeta_\tau] = 0, \tag{58}$$

we obtain

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

which clearly implies part a).

We now show part b) holds. Clearly, by (58), $\{\zeta_\tau\}_{t\ge1}$ is a martingale-difference sequence. Moreover, it follows from (16) and (17) that

$$\mathbb{E}_{|\xi_{[\tau-1]}}\left[\exp\{\zeta_\tau^2/(2\gamma_\tau\Omega_{\omega,X}\sigma)^2\}\right]\le\mathbb{E}_{|\xi_{[\tau-1]}}\left[\exp\{(2\gamma_\tau\Omega_{\omega,X}\|\delta_\tau\|_*)^2/(2\gamma_\tau\Omega_{\omega,X}\sigma)^2\}\right]\le\exp(1),$$

The previous two observations, in view of Lemma 4, then imply that

$$\forall\Lambda\ge0:\mathrm{Prob}\left\{\sum_{\tau=1}^{t}\zeta_\tau>2\Lambda\Omega_{\omega,X}\sigma\sqrt{\sum_{\tau=1}^{t}\gamma_\tau^2}\right\}\le\exp\{-\Lambda^2/3\}.\tag{59}$$

Now observe that under Assumption A2,

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

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

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

whence, taking expectations,

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

It then follows from Markov's inequality that

$$\forall\Lambda\ge0:\mathrm{Prob}\left\{\sum_{\tau=1}^{t}\gamma_\tau^2\|\delta_\tau\|_*^2>(1+\Lambda)\sigma^2\sum_{\tau=1}^{t}\gamma_\tau^2\right\}\le\exp\{-\Lambda\}.\tag{60}$$

Combining (57), (59), and (60), and rearranging the terms, we obtain (21). ∎

### 4.2 Convergence analysis for the accelerated SA

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 A1.

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

**Lemma 5** *Assume that the stepsizes $\beta_\tau$ and $\gamma_\tau$ satisfy $\beta_\tau \geq 1$ and $L\gamma_\tau < \alpha\beta_\tau$ for all $\tau \geq 1$. 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 (28) and (29). Then, for every $x \in X$, we have*

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

*where*

$$\hat{\Delta}_\tau = \hat{\Delta}_\tau(x) := \gamma_\tau\langle \delta_\tau, x - x_\tau\rangle + \frac{(2\mathcal{M} + \|\delta_\tau\|_*)^2 \beta_\tau \gamma_\tau^2}{2(\alpha\beta_\tau - L\gamma_\tau)}. \tag{61}$$

*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 (52) and the relation $\alpha\|d_\tau\|^2/2 \leq V(x_\tau, x_{\tau+1})$ then imply that

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

Noting that

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

we conclude from the previous conclusion that

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

where the last inequality follows from (53) with $u = \|d_\tau\|$, $b = (2\mathcal{M} + \|\delta_\tau\|_*)\gamma_\tau$, and $a = (\alpha\beta_\tau - L\gamma_\tau)/\beta_\tau$.

Moreover, it follows from the identity (28), (13), and Lemma 1 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

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

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

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

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 $\beta_\tau \ge 1$, (30), and Lemma 5 with $x = \bar{x}$ that, for any $t \ge 1$,

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

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 (15) and the fact $V(x_{t+1},\bar{x}) \ge 0$.

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

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

where the last inequality follows from (30), we then conclude from the previous observation that

$$(\beta_{t+1}-1)\gamma_{t+1}[\Psi(x_{t+1}^{ag}) - \Psi(\bar{x})] \le D_{\omega,X}^2 + \sum_{\tau=1}^t \left[\zeta_\tau + \frac{2}{\alpha}(4\mathcal{M}^2 + \|\delta_\tau\|_*^2)\gamma_\tau^2\right]. \quad (62)$$

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 (62) and noting that under Assumption A1, $\mathbb{E}[\|\delta_\tau\|_*^2] \le \sigma^2$ and $\mathbb{E}_{|\xi_{[\tau-1]}}[\zeta_\tau] = 0$, we obtain

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

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 3.

**Lemma 6** *Assume that $\rho > 0$ in (43). If $\tilde{x} \in X$ is an approximate solution of (43) satisfying*

$$\tilde{\Psi}_\rho(\tilde{x}) - \tilde{\Psi}_\rho^* \leq \delta, \tag{63}$$

*then*

$$\|\mathcal{A}\tilde{x} - b\| \leq \frac{2}{\rho}\|y^*\| + \sqrt{\frac{2\delta}{\rho}} \tag{64}$$

$$\tilde{h}(\tilde{x}) - \tilde{h}^* \leq \delta, \tag{65}$$

*where $y^*$ is an arbitrary Lagrange multiplier associated with (40).*

*Proof.* Let $v(u) := \inf\{\tilde{h}(x) : \mathcal{A}x - b = u,\ x \in X\}$ be the value function associated with (40). It is well-known (c.f. Theorem 29.1 of [43]) that our assumptions imply that $v$ is a convex function such that $-y^* \in \partial v(0)$. Hence,

$$v(u) - v(0) \geq \langle -y^*, u\rangle \ \forall u \in \mathbb{R}^m.$$

Letting $u := \mathcal{A}\tilde{x} - b$, we conclude from the above observation, the facts that $v(u) \leq \tilde{h}(\tilde{x})$ and $v(0) \geq \tilde{\Psi}_\rho^*$, and assumption (63), that

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

which clearly implies (64). Moreover, the fact that $\tilde{h}^* = v(0) \geq \tilde{\Psi}_\rho^*$ implies that

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

■

We are now ready to prove Theorem 3.

**Proof of Theorem 3:** Let $\rho = \rho(\nu)$ for some $\nu \geq \|y^*\|$ and assume that $\tilde{x} \in X$ satisfies (63) with $\delta = \epsilon_o$. Let $\rho_* := \rho(\|y^*\|)$ and observe that $\rho_* \leq \rho(\nu)$ for every $\nu \geq \|y^*\|$. It follows from the previous observation and Lemma 6 that $\tilde{h}(\tilde{x}) - \tilde{h}^* \leq \epsilon_o$ and

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

$$= \frac{1}{\sqrt{\rho_*}}\left(\frac{\sqrt{\epsilon_o + 4\epsilon_p\|y^*\|} - \sqrt{\epsilon_o}}{\sqrt{2}} + \sqrt{2\epsilon_o}\right) \ =\ \frac{\sqrt{\epsilon_o} + \sqrt{\epsilon_o + 4\epsilon_p\|y^*\|}}{\sqrt{2\rho_*}} = \epsilon_p,$$

and hence that $\tilde{x}$ is an $(\epsilon_p, \epsilon_o)$-primal solution of (40).

Clearly, by (44), we have $L = \rho\|\mathcal{A}\|^2$. Observe that the gradient for the smooth component $\tilde{f}_\rho$ in $\tilde{\Psi}_\rho$ (see (43)) can be computed exactly and hence that the error of approximating the subgradient of $\tilde{\Psi}_\rho$ exists only in the non-smooth component $\tilde{h}$. For

any given point $x \in X$, let $\tilde{\mathcal{H}}(x, \xi_t)$ be the output from the stochastic oracle of $\tilde{h}$ and $\tilde{h}'(x) = \mathbb{E}[\tilde{\mathcal{H}}(x, \xi_t)]$. It follows from (9), Jensen's inequality, and Assumption A3.c) that

$$\begin{aligned}
&\mathbb{E}\left[\exp\left\{\|\tilde{\mathcal{H}}(x,\xi) - \tilde{h}'(x)\|_*^2/(4\mathcal{M}^2)\right\}\right] \le \mathbb{E}\left[\exp\left\{\left(2\|\tilde{\mathcal{H}}(x,\xi)\|_*^2 + 2\|\tilde{h}'(x)\|_*^2\right)/(4\mathcal{M}^2)\right\}\right] \\
&\le \mathbb{E}\left[\exp\left\{\left(2\|\tilde{\mathcal{H}}(x,\xi)\|_*^2 + 2\mathcal{M}^2\right)/(4\mathcal{M}^2)\right\}\right] \le \exp(1/2)\,\mathbb{E}\left[\exp\left\{\|\tilde{\mathcal{H}}(x,\xi)\|_*^2/(2\mathcal{M}^2)\right\}\right] \\
&\le \exp(1/2)\left(\mathbb{E}\left[\exp\left\{\|\tilde{\mathcal{H}}(x,\xi)\|_*^2/\mathcal{M}^2\right\}\right]\right)^{\frac{1}{2}} \le \exp(1),
\end{aligned}$$

which then implies that Assumption A2 holds with $\sigma = 2\mathcal{M}$. The previous observations together with (22) and (24) then imply that

$$K_0^*(N_{md}) + \lambda K_1^*(N_{md}) = \frac{\rho\|\mathcal{A}\|^2\Omega^2}{N_{md}} + \frac{4\sqrt{2}+3\lambda}{\sqrt{N_{md}}}\Omega\mathcal{M} \le \frac{\epsilon_o}{2} + \frac{\epsilon_o}{2} \le \epsilon_o.$$

The previous conclusion, in view of the definition of $\lambda$ and (24), clearly imply the claim in part a). Part b) follows similarly from (36) and the definition of $\lambda$, by noting that

$$\begin{aligned}
\hat{K}_0^*(N_{ac}) + \lambda\hat{K}_1^*(N_{ac}) &= \frac{4\rho\|\mathcal{A}\|^2\Omega^2}{N_{ac}(N_{ac}+2)} + \frac{8\sqrt{2}+12\lambda}{\sqrt{N_{ac}}}\Omega\mathcal{M} \le \frac{4\rho\|\mathcal{A}\|^2\Omega^2}{N_{ac}^2} + \frac{8\sqrt{2}+12\lambda}{\sqrt{N_{ac}}}\Omega\mathcal{M} \\
&\le \frac{\epsilon_o}{2} + \frac{\epsilon_o}{2} \le \epsilon_o.
\end{aligned}$$

■

## 5 Concluding remarks

In this paper, two subgradient-type methods for solving the stochastic composite problem (1), namely: the modified mirror descent SA and AC-SA algorithm, are analyzed and compared. From the theoretical view of perspective, our contributions are to close the theoretical gap between the lower and upper bounds on the rate of convergence for solving this class of problems and to develop the first universally optimal method in convex optimization through a unified convergence analysis for non-smooth, smooth and stochastic convex optimization. Some important properties on the penalty-based approaches are also discovered for the first time in the literature with the development of the AC-SA algorithm. These results will greatly extend the application of the mirror descent SA-type algorithms to convex programming problems with more structured objective functions or more complicated feasible regions. Moreover, we believe that the AC-SA algorithm developed here possesses significant potential of application, in those traditional areas for stochastic approximation, e.g., statistical learning, Markov-decision process and 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 Professor Alex Shapiro who has carefully proofread this paper. The author thanks the anonymous referees for their very useful suggestions.

## References


1. A. Auslender and M. Teboulle. Interior gradient and proximal methods for convex and conic optimization. *SIAM Journal on Optimization*, 16:697–725, 2006.
2. H.H. Bauschke, J.M. Borwein, and P.L. Combettes. Bregman monotone optimization algorithms. *SIAM Journal on Controal and Optimization*, 42:596–636, 2003.
3. S. Becker, J. Bobin, and E. Candes. Nesta: A fast and accurate first-order method for sparse recovery. Manuscript, California Institute of Technology, 2009.
4. A. Ben-Tal and A.Nemirovski. Non-euclidean restricted memory level method for large-scale convex optimization. *Mathematical Programming*, 102:407–456, 2005.
5. A. Benveniste, M. Métivier, and P. Priouret. *Algorithmes adaptatifs et approximations stochastiques*. Masson, 1987. English translation: *Adaptive Algorithms and Stochastic Approximations*, Springer Verlag (1993).
6. D. Bertsekas. *Nonlinear Programming*. Athena Scientific, New York, second edition, 1999.
7. L.M. Bregman. The relaxation method of finding the common point convex sets and its application to the solution of problems in convex programming. *USSR Comput. Math. Phys.*, 7:200–217, 1967.
8. A. d'Aspremont. Smooth optimization with approximate gradient. *SIAM Journal on Optimization*, 19:1171–1183, 2008.
9. A. d'Aspremont, O. Banerjee, and L. El Ghaoue. First-order methods for sparse covariance selection. *SIAM Journal on Matrix Analysis and its Applications*, 30:56–66, 2008.
10. Y. Ermoliev. Stochastic quasigradient methods and their application to system optimization. *Stochastics*, 9:1–36, 1983.
11. A. Gaivoronski. Nonstationary stochastic programming problems. *Kybernetika*, 4:89–92, 1978.
12. A. Juditsky, A. Nazin, A. B. Tsybakov, and N. Vayatis. Recursive aggregation of estimators via the mirror descent algorithm with average. *Problems of Information Transmission*, 41:n.4, 2005.
13. A. Juditsky, A. Nemirovski, and C. Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. Manuscript, Georgia Institute of Technology, Atlanta, GA, 2008. submitted to SIAM Journal on Control and Optimization.
14. A. Juditsky, P. Rigollet, and A. B. Tsybakov. Learning by mirror averaging. *Annals of Statistics*, 36:2183–2206, 2008.
15. K.C. Kiwiel. Proximal minimization methods with generalized bregman functions. *SIAM Journal on Controal and Optimization*, 35:1142–1168, 1997.
16. A. J. Kleywegt, A. Shapiro, and T. Homem de Mello. The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization*, 12:479–502, 2001.
17. H. J. Kushner and G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications*, volume 35 of *Applications of Mathematics*. Springer-Verlag, New York, 2003.
18. G. Lan, Z. Lu, and R. D. C. Monteiro. Primal-dual first-order methods with $\mathcal{O}(1/\epsilon)$ iteration-complexity for cone programming. *Mathematical Programming*, 2009. to appear.
19. G. Lan and R. D. C. Monteiro. Iteration-complexity of first-order penalty methods for convex programming. Manuscript, School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA, June 2008.
20. G. Lan and R. D. C. Monteiro. Iteration-complexity of first-order augmented lagrangian methods for convex programming. Manuscript, School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA, May 2009.
21. G. Lan, A. Nemirovski, and A. Shapiro. Validation analysis of robust stochastic approximation method. *submitted to Mathematical Programming*, 2008. *E-print available at:* `http://www.optimization-online.org`.
22. A.S. Lewis and S.J. Wright. A proximal method for composite minimization. Manuscript, Cornell University, Ithaca, NY, 2009.
23. J. Linderoth, A. Shapiro, and S. Wright. The empirical behavior of sampling methods for stochastic programming. *Annals of Operations Research*, 142:215–241, 2006.
24. Z. Lu. Smooth optimization approach for sparse covariance selection. *SIAM Journal on Optimization*, 19:1807–1827, 2009.
25. Z. Lu, R.D.C. Monteiro, and M. Yuan. Convex optimization methods for dimension reduction and coefficient estimation in multivariate linear regression. Manuscript, School of ISyE, Georgia Tech, Atlanta, GA, 30332, USA, January 2008.

26. Z. Lu, A. Nemirovski, and R. D. C. Monteiro. Large-scale semidefinite programming via saddle point mirror-prox algorithm. *Mathematical programming*, 109:211–237, 2007.
27. W. K. Mak, D.P. Morton, and R.K. Wood. Monte carlo bounding techniques for determining solution quality in stochastic programs. *Operations Research Letters*, 24:47–56, 1999.
28. R.D.C. Monteiro and B.F. Svaiter. On the complexity of the hybrid proximal extragradient method for the iterates and the ergodic mean. Manuscript, School of ISyE, Georgia Tech, Atlanta, GA, 30332, USA, March 2009.
29. A. Nemirovski. Prox-method with rate of convergence $o(1/t)$ for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15:229–251, 2004.
30. A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19:1574–1609, 2009.
31. A. Nemirovski and D. Yudin. *Problem complexity and method efficiency in optimization*. Wiley-Interscience Series in Discrete Mathematics. John Wiley, XV, 1983.
32. 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. translated as Soviet Math. Docl.
33. Y. E. Nesterov. *Introductory Lectures on Convex Optimization: a basic course*. Kluwer Academic Publishers, Massachusetts, 2004.
34. Y. E. Nesterov. Smooth minimization of nonsmooth functions. *Mathematical Programming*, 103:127–152, 2005.
35. Y. E. Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120:221–259, 2006.
36. 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.
37. Y. E. Nesterov. Smoothing technique and its applications in semidefinite optimization. *Mathematical Programming*, 110:245–259, 2007.
38. J. Peña. Nash equilibria computation via smoothing techniques. *Optima*, 78:12–13, 2008.
39. G.C. Pflug. Optimization of stochastic models. In *The Interface Between Simulation and Optimization*. Kluwer, Boston, 1996.
40. B.T. Polyak. New stochastic approximation type procedures. *Automat. i Telemekh.*, 7:98–107, 1990.
41. B.T. Polyak and A.B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM J. Control and Optimization*, 30:838–855, 1992.
42. H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 1951.
43. R. T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, NJ, 1970.
44. A. Ruszczyński and W. Sysk. A method of aggregate stochastic subgradients with on-line stepsize rules for convex stochastic programming problems. *Mathematical Programming Study*, 28:113–131, 1986.
45. A. Shapiro. Monte carlo sampling methods. In A. Ruszczyński and A. Shapiro, editors, *Stochastic Programming*. North-Holland Publishing Company, Amsterdam, 2003.
46. A. Shapiro and A. Nemirovski. On complexity of stochastic programming problems. In V. Jeyakumar and A.M. Rubinov, editors, *Continuous Optimization: Current Trends and Applications*, pages 111–144. Springer, 2005.
47. J.C. Spall. *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*. John Wiley, Hoboken, NJ, 2003.
48. V. Strassen. The existence of probability measures with given marginals. *Annals of Mathematical Statistics*, 30:423–439, 1965.
49. M. Teboulle. Convergence of proximal-like algorithms. *SIAM Journal on Optimization*, 7:1069–1083, 1997.
50. P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. Manuscript, University of Washington, Seattle, May 2008.
51. B. Verweij, S. Ahmed, A. J. Kleywegt, G. Nemhauser, and A. Shapiro. The sample average approximation method applied to stochastic routing problems: a computational study. *Computational Optimization and Applications*, 24:289–333, 2003.