Open access • Journal Article • DOI:10.1137/110848876

# Optimal Stochastic Approximation Algorithms for Strongly Convex Stochastic Composite Optimization, II: Shrinking Procedures and Optimal Algorithms — Source link

Saeed Ghadimi, Guanghui Lan

**Institutions:** University of Florida

**Published on:** 22 Oct 2013 - Siam Journal on Optimization (Society for Industrial and Applied Mathematics)

**Topics:** Stochastic approximation, Convex optimization, Convex function and Rate of convergence

# Optimal Stochastic Approximation Algorithms for Strongly Convex Stochastic Composite Optimization I: a Generic Algorithmic Framework *†

Saeed Ghadimi ‡ Guanghui Lan §

July 1, 2010 (Revised: September 20, 2011, June 15, 2012)

**Abstract**

In this paper we present a generic algorithmic framework, namely, the accelerated stochastic approximation (AC-SA) algorithm, for solving strongly convex stochastic composite optimization (SCO) problems. While the classical stochastic approximation (SA) algorithms are asymptotically optimal for solving differentiable and strongly convex problems, the AC-SA algorithm, when employed with proper stepsize policies, can achieve optimal or nearly optimal rates of convergence for solving different classes of SCO problems during a given number of iterations. Moreover, we investigate these AC-SA algorithms in more detail, such as, establishing the large-deviation results associated with the convergence rates and introducing efficient validation procedure to check the accuracy of the generated solutions.

**Keywords:** stochastic approximation, convex optimization, stochastic programming, complexity, large deviation

## 1 Introduction

Convex programming (CP) under noisy first-order information has attracted considerable interest during the past decades for its applications in a broad spectrum of disciplines including statistical estimation, signal processing and operations research, etc. In the classical setting, we consider the problem of $\min_{x\in X} \Psi(x)$, where $X$ is a closed convex set and $\Psi : X \to \mathbb{R}$ is a strongly convex and differentiable function for which only noisy gradient information is available. In 1951, Robbins and Monro in their pioneering paper [34] proposed the stochastic approximation (SA) algorithm for solving this type of problems. This approach, referred to as the *classical SA*, mimics the simplest gradient descent method by using noisy gradient information in place of the exact gradients. Since then SA algorithms became widely used in stochastic optimization (see, e.g., [6, 10, 11, 21, 31, 35, 40]

*The manuscript is available on www.optimization-online.org.

†Both authors were partially supported by NSF AWARD CMMI-1000347.

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

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

and references therein) and, due to especially low demand for computer memory, in signal processing (cf., [6] and references therein). An important improvement of the SA method was developed by Polyak [32] and Polyak and Juditsky [33], where longer stepsizes were suggested together with the averaging of the obtained iterates. The analysis of these SA methods (goes back to the works [8] and [36]) focused on obtaining the asymptotically optimal rate of convergence $\mathbb{E}[\Psi(x_t) - \Psi_*] = O(t^{-1})$ (here $x_t$ is $t$-th iterate and $\Psi_*$ is the minimal value of $\Psi(x)$ over $x \in X$). However, it is difficult to implement "asymptotically optimal" stepsize policy, especially in the beginning, so that the algorithms often perform poorly in practice (e.g., [40, Section 4.5.3.]).

In last few years, there has been a revival of interest in SA methods for stochastic convex optimization and their applications (e.g. [15, 18, 23, 25, 29, 37, 39]). These developments, motivated by complexity theory in continuous optimization [26], concerned the convergence properties of SA-type methods during a finite number of iterations. For example, Nemirovski et al. [25] presented a properly modified SA approach, namely, mirror descent SA, for minimizing general non-smooth convex functions. They demonstrated that the mirror descent SA exhibits an optimal $\mathcal{O}(1/\epsilon^2)$ expected iteration-complexity for solving these problems and that the constant factor associated with this iteration-complexity bound is also essentially unimprovable. The mirror descent SA was shown in [23, 25] to be competitive to the widely-accepted sample average approximation approach (see, e.g., [20, 38]) and even significantly outperform it for solving a class of stochastic programming problems. For example, when X in (1.1) is a simplex of large dimension, the mirror descent SA builds approximate solutions $10 - 40$ times faster than an SAA based algorithm while keeping similar solution quality. Similar techniques, based on subgradient averaging, have been proposed in [15, 18, 29]. While these techniques dealt with general non-smooth CP problems, Lan [22] presented the first unified optimal method for smooth, nonsmooth and stochastic optimization, which explicitly takes into account the smoothness of the objective function. However, note that none of these techniques could achieve the optimal expected iteration-complexity for minimizing differentiable and strongly convex functions (a.k.a. the classical setting of SA).

In this paper, we study a class of strongly convex stochastic composite optimization (SCO) problems given by

$$\Psi^* := \min_{x \in X} \{\Psi(x) := f(x) + \mathcal{X}(x)\}, \tag{1.1}$$

where $X$ is a closed convex set in Euclidean space $\mathcal{E}$, $\mathcal{X}(x)$ is a simple convex function with known structure (e.g., $\mathcal{X}(x) = 0$ or $\mathcal{X}(x) = \|x\|_1$), and $f : X \to \mathbb{R}$ is a general convex function such that for some $L \geq 0$, $M \geq 0$ and $\mu \geq 0$,

$$\frac{\mu}{2}\|y - x\|^2 \leq f(y) - f(x) - \langle f'(x), y - x\rangle \leq \frac{L}{2}\|y - x\|^2 + M\|y - x\|, \quad \forall x, y \in X, \tag{1.2}$$

where $f'(x) \in \partial f(x)$ and $\partial f(x)$ denotes the subdifferential of $f$ at $x$. Moreover, we assume that the first-order information of $f$ is available via subsequent calls to a stochastic oracle ($\mathcal{SO}$). More specifically, at the $t$-th call, $x_t \in X$ being the input, the $\mathcal{SO}$ outputs the quantity $F(x_t, \xi_t)$ and a vector $G(x_t, \xi_t)$, where $\{\xi_t\}_{t \geq 1}$ is a sequence of independently and identically distributed (i.i.d) random variables such that $\mathbb{E}[F(x, \xi_t)] = f(x)$ and $\mathbb{E}[G(x, \xi_t)] \equiv g(x) \in \partial f(x)$ for any $x \in X$. The following assumption is made throughout the paper.

**A1:** For any $x \in X$ and $t \geq 1$, we have $\mathbb{E}\left[\|G(x, \xi_t) - g(x)\|_*^2\right] \leq \sigma^2$.

Since the parameters $L, M, \mu, \sigma$ can be zero, problem (1.1) described above covers a wide range of CP problems. In particular, if $f$ is a general Lipschitz continuous function with constant $M_f$, then relation (1.2) holds with $L = 0, \mu = 0$ and $M = 2M_f$. If $f$ is a strongly convex smooth function in $\mathcal{C}^{1,1}_{L/\mu}$ (e.g., [28]), then (1.2) is satisfied with $M = 0$. Clearly, relation (1.2) also holds if $f$ is given as the summation of smooth and nonsmooth convex functions. Moreover, problem (1.1) covers different classes of deterministic CP problems if $\sigma = 0$ (in particular, letting $\sigma = 0$, $M = 0$, $\mu = 0$, problem (1.1) becomes a class of composite CP problems studied by Nesterov [30] and later Tseng [43], Lewis and Wright [24]). To subsume all these different possible combinations, we refer to the aforementioned class of CP problems as $\mathcal{F}_{X,\mathcal{X}}(L, M, \sigma, \mu)$.

Since strong convexity has been extensively studied, we can have a long list of application problems of the form $\mathcal{F}_{X,\mathcal{X}}(L, M, \sigma, \mu)$. In particular, most applications of the classical SA fall as one of its special cases, namely, $\mathcal{F}_{X,0}(L, 0, \sigma, \mu)$. To motivate our discussion, let us mention a few concrete examples in statistical learning which help to represent massive data in a compact way [13]. Consider a set of observed data $S = \{(u_i, v_i)\}_{i=1}^m$, drawn at random from an unknown distribution $\mathcal{D}$ on $U \times V$. We would like to find a linear form $\mathcal{V}(u) = \langle x, u\rangle$ to describe the relation between $u_i$ and $v_i$. To this end, we can solve different problems of the form (1.1). In particular, let $\rho, \upsilon > 0$ be user-defined parameters, we can use the following different types of learning models

- Ridge regression: $f(x) = \mathbb{E}[(\langle x, u\rangle - v)^2] + \rho\|x\|_2^2$, $\mathcal{X}(x) = 0$ and $X = \mathbb{R}^d$;
- support vector machine [5]: $f(x) = \mathbb{E}[\max\{0, v\langle x, u\rangle] + \rho\|x\|_2^2$, $\mathcal{X}(x) = 0$ and $X = \mathbb{R}^d$;
- elastic net regression [46]: $f(x) = \mathbb{E}[(\langle x, u\rangle - v)^2] + \rho\|x\|_2^2$, $\mathcal{X}(x) = \upsilon\|x\|_1$ and $X = \mathbb{R}^d$,

where the expectations are taken w.r.t. $u$ and $v$. Observe that the above problems are unconstrained problems. It is well-known (see Section 1.1.1 of Judistky and Nemirovski [16]) that the convergence rate of first-order methods heavily depends on the starting point or the size of feasible set. For this reason, one may prefer to reformulating the above unconstrained problems into constrained ones. For example, in the ridge regression problem, the $l_2$ norm regularization term can be stated as a constraint, i.e., $\|x\|_2 \le b$ for some $b > 0$. Some other constrained problems fitting in our setting are as follows.

- Lasso regression [42]: $f(x) = \mathbb{E}[(\langle x, u\rangle - v)^2]$, $\mathcal{X}(x) = 0$ and $X = \{x \in \mathbb{R}^d : \|x\|_1 \le b\}$ for some $b > 0$;
- Metric learning [44]: $f(x) = \mathbb{E}[|\mathrm{tr}(xuu^T) - v|]$, $\mathcal{X}(x) = 0$ and $X = \{x \in \mathbb{R}^{d\times d} : x \succeq 0, \mathrm{tr}(x) \le b\}$, for some $b > 0$.

Note that, although the latter set of examples have bounded feasible sets, their objective funtions are not necessarily strongly convex.

If $\mu > 0$ in (1.2), then, by the classic complexity theory for convex programming (see, e.g., Theorems 5.3.1 and 7.2.6 of [26], Theorem 2.1.13 of [28], [45] and [17]), to find *an $\epsilon$-solution of (1.1)*, i.e., a point $\bar{x} \in X$ s.t. $\mathbb{E}[\Psi(\bar{x}) - \Psi^*] \le \epsilon$, the number of calls (or iterations) to $\mathcal{SO}$ cannot be smaller than [1]

$$\mathcal{O}(1)\left(\sqrt{\frac{L}{\mu}}\log\frac{L\|x_0-x^*\|^2}{\epsilon}+\frac{(M+\sigma)^2}{\mu\epsilon}\right), \tag{1.3}$$

where $x_0$ denotes an initial point, $x^*$ is the optimal solution of problem (1.1) and $\mathcal{O}(1)$ represents an absolute constant. However, it is not clear if such a lower complexity bound is achievable or not. As shown in [25], the iteration complexity for the classical SA for solving $\mathcal{F}_{X,0}(L,0,\sigma,\mu)$ is given by

$$\mathcal{O}(1)\left(\frac{L}{\epsilon}\max\left\{\frac{\bar{G}^2}{\mu^2},\|x_0-x^*\|^2\right\}\right), \tag{1.4}$$

where $\bar{G}^2 := \sup_{x\in X}\mathbb{E}[G(x,\xi)]^2$. Note that, in our setting, $M=0$ and $\bar{G}^2$ is in the order of $\sigma^2+L^2\max_{x\in X}\|x-x^*\|^2$ (see Remark 1 of [22]). Clearly, bound (1.4) is substantially worse than (1.3) in terms of the dependence on $L$, $\mu$ and $\|x_0-x^*\|$, although both of them are of $\mathcal{O}(1/\epsilon)$. As a result, the classical SA method of this type is very sensitive to these problem parameters and the selection of initial points.

Our contribution in this paper mainly consists of the following aspects. Firstly, by properly modifying the well-known Nesterov's optimal smooth method [27, 28], we develop a generic accelerated stochastic approximation (AC-SA) algorithmic framework, which can be specialized to yield optimal or nearly optimal methods for different classes of SCO problems in $\mathcal{F}_{X,\mathcal{X}}(L,M,\sigma,\mu)$. In particular, we study and compare two different variants of optimal AC-SA algorithms for solving SCO problems without assuming strong convexity, i.e., $\mathcal{F}_{X,\mathcal{X}}(L,M,\sigma,0)$. Moreover, we present one AC-SA algorithm for solving strongly convex SCO problems, which exhibits a nearly optimal iteration-complexity

$$\mathcal{O}(1)\left(\sqrt{\frac{L\|x_0-x^*\|^2}{\epsilon}}+\frac{(M+\sigma)^2}{\mu\epsilon}\right). \tag{1.5}$$

Note that the above complexity bound is significantly better than (1.4) in terms of the dependence on $L$, $\mu$, $\sigma$ and $\|x_0-x^*\|$ (note $M=0$ for this comparison). In particular, it is worth mentioning that the dependence on the selection of the initial point, i.e., $\|x_0-x^*\|$, has been considerably reduced. In fact, if the second term in (1.5) majorizes the first one, then the quantity $\|x_0-x^*\|$ does not affect the complexity bound (up to a factor of 2). It is also worth noting that this algorithm employs a very simple stepsize policy without requiring any information about $\sigma$ and the initial point $x_0$ or the size of the feasible set.

Secondly, while the aforementioned complexity results evaluate, on average, the performance of the AC-SA algorithms over many different runs, we also study the large-deviations of these complexity results, in order to estimate the performance of a single run of these algorithms. More specifically, under certain "light-tail" assumptions on the $\mathcal{SO}$, we investigate the iteration-complexity of finding an $(\epsilon,\Lambda)$-*solution* of (1.1), i.e., a point $\bar{x}\in X$ s.t. $\text{Prob}\{\Psi(\bar{x})-\Psi^*>\epsilon\}\le\Lambda$, for a given confidence level $\Lambda\in(0,1)$.

[1]While the other terms in (1.3) come from deterministic CP, we briefly discuss how the critical term $\sigma^2/(\mu\epsilon)$ in (1.3) is derived. Consider the problem of $\min_x\left\{\Psi(x)=\mu(x-\alpha)^2\right\}$ with unknown $\alpha$. Also suppose that the stochastic gradient is given by $2\mu(x-\alpha-\xi/\mu)$ with $\xi\sim\mathbb{N}(0,\sigma^2)$. Under this setting, our optimization problem is equivalent to the estimation of the unknown mean $\alpha$ from the observations of $\zeta=\alpha+\xi/\mu\sim\mathbb{N}(\alpha,\sigma^2/\mu^2)$, and the residual is $\mu$ times the expected squared error of recovery of the mean $\alpha$. By standard statistical reasons, when the initial range for $\alpha$ is larger than $\sigma/\mu$, to make this expected squared error smaller than $\delta^2\equiv\epsilon/\mu$, or equivalently, $\mathbb{E}[\Psi(\bar{x})-\Psi_*]\le\epsilon$, the number of observations we need is at least $N=\mathcal{O}(1)\left((\sigma^2/\mu^2)/\delta^2\right)=\mathcal{O}(1)\left(\sigma^2/(\mu\epsilon)\right)$.

Finally, one crucial problem in most SA-type methods is how to check the accuracy of the generated solutions. For this purpose, we show that, with little additional computational effort, the developed AC-SA algorithms can also output a sequence of lower bounds on $\Psi^*$. These lower bounds, when coupled with certain stochastic upper bounds, can provide online accuracy certificates for the generated solutions. In particular, we demonstrate that the gap between these upper and lower bounds converges to 0 as rapidly as the sequence of objective values of the generated solutions converges to $\Psi^*$.

The paper is organized as follows. We start by reviewing some basic concepts, namely, the distance-generating functions and prox-functions in Section 2. In Section 3, we present a generic AC-SA algorithmic framework for solving $\mathcal{F}_{X,\mathcal{X}}(L, M, \sigma, \mu)$ and establish its convergence properties. We then demonstrate in Section 4 that the generic AC-SA algorithm, if employed with suitable stepsize policies, can achieve optimal or nearly optimal expected rates of convergence for solving different classes of problems in $\mathcal{F}_{X,\mathcal{X}}(L, M, \sigma, \mu)$. In Section 5, we discuss the computation of certain stochastic lower and upper bounds of $\Psi^*$ during the execution of the AC-SA algorithms. Some concluding remarks are given in Section 6.

In the companion paper [12], we show that the lower complexity bound (1.3) is actually achievable provided that the value of $\sigma$ and a bound on $\Psi(x_0) - \Psi^*$ are given. Also, observing that a single-run of aforementioned nearly optimal algorithms for strongly convex SCO has significantly worse theoretical convergence performance than the average one over many runs, we develop ways to improve the former iteration-complexity results so that they become comparable to the latter ones. Some numerical results demonstrating the effectiveness of different variants of the AC-SA algorithm are also presented in [12].

## 1.1 Notation and terminology

- $\mathcal{E}$ is endowed with inner product $\langle \cdot, \cdot \rangle$ and an arbitrary norm $\|\cdot\|$ (not necessarily the one induced by the inner product $\langle \cdot, \cdot \rangle$).

- For a convex lower semicontinuous function $\phi : X \to \mathbb{R}$, 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 $\phi$ 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 $x = \lim\limits_{i\to\infty} x_i$, $g = \lim\limits_{i\to\infty} g_i$. Finally, $\partial\phi(x) = \emptyset$ for $x \notin X$. With this definition, it is well-known (see, for example, Ben-Tal and Nemirovksi [3]) that, if a convex function $\phi : X \to \mathbb{R}$ 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
$$g \in \partial\phi(x) \Rightarrow |\langle g, d\rangle| \le M\|d\|, \ \forall\, d \in \operatorname{lin}(X - X), \tag{1.6}$$
in other words,
$$g \in \partial\phi(x) \Rightarrow \|g\|_* \le M, \tag{1.7}$$
where $\|\cdot\|_*$ denotes the conjugate norm given by $\|g\|_* := \max_{\|d\|\le 1}\langle g, d\rangle$.

- 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 with $\xi_{[t]}$ being given.

## 2 Preliminary: distance generating function and prox-function

In this section, we review the concept of prox-function (i.e., proximity control function). By using the generalized prox-function in place of the usual Euclidean distance function, the developed algorithms will be capable of adjusting to the different geometry of the feasible set.

We say that a function $\omega : X \to \mathbb{R}$ is a *distance generating function* with modulus $\nu > 0$ with respect to $\|\cdot\|$, if $\omega$ is continuously differentiable and strongly convex with parameter $\nu$ with respect to $\|\cdot\|$, i.e.,

$$\langle x - z, \nabla\omega(x) - \nabla\omega(z)\rangle \geq \nu\|x - z\|^2, \ \ \forall x, z \in X. \tag{2.1}$$

The *prox-function* associated with $\omega$ is given by

$$V(x, z) \equiv V_\omega(x, z) = \omega(z) - [\omega(x) + \langle\nabla\omega(x), z - x\rangle]. \tag{2.2}$$

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, 19, 41] and references therein). In this paper, we assume that the prox-function $V(x, z)$ is chosen such that the solution of

$$P_\omega(g, x) := \arg\min_{z\in X} \{\langle g, z\rangle + V(x, z) + \mathcal{X}(z)\} \tag{2.3}$$

is easily computable for any $g \in \mathcal{E}^*$ and $x \in X$. We point out below a few examples where such an assumption is satisfied.

- If $X$ is relatively simple, e.g., Euclidean ball, simplex or $l_1$ ball, and $\mathcal{X}(x) = 0$, then by properly choosing the distance generating function $\omega(\cdot)$, we can obtain closed form solutions of problem (2.3). This is the standard setting used in the regular SA methods [25, 16].

- If the problem is unconstrained, i.e., $X = \mathcal{E}$, and $\mathcal{X}(x)$ is relatively simple, we can derive closed form solutions of (2.3) for some interesting cases. For example, if $\mathcal{X}(x) = \|x\|_1$ and $\omega(x) = \|x\|_2^2/2$, then an explicit solution of (2.3) is readily given by its first-order optimality condition. A similar example is given by $\mathcal{X}(x) = \sum_{i=1}^d \sigma_i(x)$ and $\omega(x) = \mathrm{tr}(x^T x)/2$, where $\sigma_i(x)$, $i = 1, \ldots, d$, denote the singular values of $x \in \mathbb{R}^{d\times d}$.

- If $X$ is relatively simple and $\mathcal{X}(x)$ is nontrivial, we can still compute closed form solutions of (2.3) for some interesting special cases, e.g., when $X$ is the standard simplex, $\omega(x) = \sum_{i=1}^d x_i \log x_i$ and $\mathcal{X}(x) = \sum_{i=1}^d x_i$. However, in general, slightly more computational effort than regular SA methods is needed to solve problem (2.3). For example, we can apply a simple bisection procedure to solve the Lagrangian dual of (2.3) if the Lagrangian relaxation of (2.3) has an explicit solution.

If there exists a constant $\mathcal{Q}$ such that $V(x, z) \leq \frac{\mathcal{Q}}{2}\|x - z\|^2$ for any $x, z \in X$, then we say that the prox-function $V(\cdot, \cdot)$ is growing quadratically. Moreover, the smallest constant $\mathcal{Q}$ satisfying the previous relation is called the *quadratic growth constant* of $V(\cdot, \cdot)$. For example, if $X = \mathbb{R}^n$ and $\omega(x) = \|x\|_2^2/2$, then we have $V(x, z) = \|x - z\|_2^2/2$ and $\mathcal{Q} = 1$. Another example is given by

**Example 1** *Let $X = \{x \in \mathbb{R}^n : \|x\|_1 \leq 1\}$, where $\|x\|_1 = \sum_{i=1}^n |x_i|$. We can take*

$$\omega(x) = \frac{1}{2}\|x\|_p^2 = \frac{1}{2}\left(\sum_{i=1}^n |x_i|^p\right)^{\frac{2}{p}}$$

*for some $p > 1$. In particular, if $p = 1 + 1/\ln n$, then $\omega(x)$ is strongly convex with modulus $\mu = 1/(e^2 \ln n)$ and the quadratic growth constant $\mathcal{Q}$ equals 1 for the prox-function $V(x,z)$ (see, e.g. [4, 9, 16, 26]).*

Without loss of generality, we assume that $\mathcal{Q} = 1$ for the prox-function $V(x,z)$ if it grows quadratically, i.e.,

$$V(x,z) \leq \frac{1}{2}\|x - z\|^2, \quad \forall x, z \in X. \tag{2.4}$$

Indeed, if $\mathcal{Q} \neq 1$, we can multiply the corresponding distance generating function $\omega$ by $1/\mathcal{Q}$ and the resulting prox-function will satisfy (2.4).

# 3 A generic accelerated stochastic approximation algorithm

This section contains two subsections. In Subsection 3.1, we present a generic accelerated stochastic approximation (AC-SA) algorithm and discuss its convergence properties. Subsection 3.2 is dedicated to the convergence analysis of this algorithm.

## 3.1 The algorithm and its convergence properties

In this subsection, we present the generic AC-SA algorithm for solving $\mathcal{F}_{X,\mathcal{X}}(L, M, \sigma, \mu)$ and establish its convergence properties. This algorithm maintains the updating of three intertwined sequences, namely, $\{x_t\}, \{x_t^{ag}\}$ and $\{x_t^{md}\}$. The AC-SA algorithm is obtained by replacing the gradients with stochastic (sub)gradients in Nesterov's method for smooth CP [27, 28]. This is in spirit similar to the relation of SA and gradient descent method. However, the generalization of Nesterov's smooth method to nonsmooth and stochastic CP seems to be more involved, partly because of the intercorrelation of the aforementioned three sequences.

**A generic AC-SA algorithm**

Input: $x_0 \in X$, prox-function $V(x,z)$, stepsize parameters $\{\alpha_t\}_{t \geq 1}$ and $\{\gamma_t\}_{t \geq 1}$ s.t. $\alpha_1 = 1$, $\alpha_t \in (0,1)$ for any $t \geq 2$, and $\gamma_t > 0$ for any $t \geq 1$.

0) Set the initial points $x_0^{ag} = x_0$ and $t = 1$;

1) Set

$$x_t^{md} = \frac{(1-\alpha_t)(\mu + \gamma_t)}{\gamma_t + (1 - \alpha_t^2)\mu} x_{t-1}^{ag} + \frac{\alpha_t[(1-\alpha_t)\mu + \gamma_t]}{\gamma_t + (1-\alpha_t^2)\mu} x_{t-1}; \tag{3.1}$$

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

$$x_t = \arg\min_{x \in X} \left\{ \alpha_t[\langle G_t, x\rangle + \mathcal{X}(x) + \mu V(x_t^{md}, x)] + [(1-\alpha_t)\mu + \gamma_t]\, V(x_{t-1}, x) \right\}, \tag{3.2}$$

$$x_t^{ag} = \alpha_t x_t + (1-\alpha_t) x_{t-1}^{ag}; \tag{3.3}$$

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

We now make a few comments about the above algorithmic framework. Firstly, note that, in view of the definition of $V(x,z)$ in (2.2), problems (2.3) and (3.2) are given in the same form. In particular, their objective functions are composed of three terms: a linear function of $x$, the simple convex function $\mathcal{X}(x)$ and the strongly convex function $\omega(x)$ multiplied by a positive scalar. Hence, the assumption that the solution of (2.3) is easily computable guarantees that the solution of (3.2) is also easily computable.

Secondly, observe that the points $x_t^{md}$, $t \geq 1$, are used to construct certain model functions of $\Psi(\cdot)$ in (1.1), namely,

$$l_\Psi(x_t^{md}, x) := f(x_t^{md}) + \langle f'(x_t^{md}), x - x_t^{md} \rangle + \mu V(x_t^{md}, x) + \mathcal{X}(x). \tag{3.4}$$

If $\mu = 0$ or $V(z,x)$ grows quadratically (c.f., (2.4)), then by (1.1), (1.2) and (3.4), we have

$$l_\Psi(z,x) \leq f(x) + \mathcal{X}(x) = \Psi(x), \quad \forall\, z, x \in X. \tag{3.5}$$

The search points $x_t$, $t \geq 1$, are used as prox-centers to control the proximity and the stepsizes $\gamma_t$, $t \geq 1$, control the instability of the model, so that we will not move too far away from the current prox-center when taking step (3.2). The search points $x_t^{ag}$, $t \geq 1$, are used to evaluate the objective values. Since

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

the function value of $x_t^{ag}$ might be smaller than that of $x_t$. An immediate improvement of the algorithm would be to take $x_t^{ag}$ as the one with the smallest objective value among the following three points, $x_{t-1}^{ag}$, $x_t$ and $\alpha_t x_t + (1-\alpha_t) x_{t-1}^{ag}$, 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 $l_\Psi(z,x)$, the selection of the prox-center $x_t$ and the evaluation of the objective value through a careful selection of the stepsize parameters $\alpha_t$ and $\gamma_t$, $t \geq 1$.

In some cases, Assumption A1 for the $\mathcal{SO}$ is augmented by the following "light-tail" assumption.

**A2:** For any $x \in X$ and $t \geq 1$, we have $\mathbb{E}\left[\exp\{\|G(x,\xi_t) - g(x)\|_*^2/\sigma^2\}\right] \leq \exp\{1\}$.

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

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

Theorem 1 below summarizes the main convergence properties of the generic AC-SA algorithm. The proof of this result is given in Section 3.2.

**Theorem 1** *Consider the generic AC-SA algorithm for $\mathcal{F}_{X,\mathcal{X}}(L,M,\sigma,\mu)$ and suppose that condition (2.4) holds whenever $\mu > 0$. Also assume that $\{\alpha_t\}_{t\geq 1}$ and $\{\gamma_t\}_{t\geq 1}$ are chosen such that*

$$\nu(\mu + \gamma_t) > L\alpha_t^2, \tag{3.6}$$

$$\gamma_1/\Gamma_1 = \gamma_2/\Gamma_2 = \ldots, \tag{3.7}$$

*where*

$$\Gamma_t := \begin{cases} 1, & t = 1, \\ (1-\alpha_t)\Gamma_{t-1}, & t \geq 2. \end{cases} \tag{3.8}$$

*Then,*

*a) under Assumption A1, we have*

$$\mathbb{E}[\Psi(x_t^{ag}) - \Psi^*] \le \mathcal{B}_e(t) := \Gamma_t \gamma_1 V(x_0, x^*) + \Gamma_t \sum_{\tau=1}^{t} \frac{2\alpha_\tau^2 (M^2 + \sigma^2)}{\Gamma_\tau [\nu(\mu + \gamma_\tau) - L\alpha_\tau^2]}, \tag{3.9}$$

*for any $t \ge 1$, where $x^*$ is an arbitrary optimal solution of (1.1).*

*b) under Assumption A2, we have*

$$\text{Prob}\left\{\Psi(x_t^{ag}) - \Psi^* \ge \mathcal{B}_e(t) + \lambda \, \mathcal{B}_p(t)\right\} \le \exp\{-\lambda^2/3\} + \exp\{-\lambda\}, \tag{3.10}$$

*for any $\lambda > 0$ and $t \ge 1$, where*

$$\mathcal{B}_p(t) \; := \; \sigma \Gamma_t R_X(x^*) \left( \sum_{\tau=1}^{t} \frac{\alpha_\tau^2}{\Gamma_\tau^2} \right)^{\frac{1}{2}} + \Gamma_t \sum_{\tau=1}^{t} \frac{2\alpha_\tau^2 \sigma^2}{\Gamma_\tau [\nu(\mu + \gamma_\tau) - L\alpha_\tau^2]}, \tag{3.11}$$

$$R_X(x^*) \; := \; \max_{x \in X} \|x - x^*\|. \tag{3.12}$$

*c) If $X$ is compact and $\{\alpha_t\}_{t \ge 1}$ and $\{\gamma_t\}_{t \ge 1}$ are chosen such that relation (3.6) and*

$$\gamma_1/\Gamma_1 \le \gamma_2/\Gamma_2 \le \ldots, \tag{3.13}$$

*then Parts a) and b) still hold by simply replacing the first term in the definition of $\mathcal{B}_e(t)$ with $\gamma_t \bar{V}(x^*)$, where $\bar{V}(x) := \max_{u \in X} V(u, x)$.*

A few remarks about the results obtained in Theorem 1 are in place. First, Theorem 1.a) states a general error bound on $\mathbb{E}[\Psi(x_t^{ag}) - \Psi^*]$, which can be applied to any subclasses of CP problems in $\mathcal{F}_{X,\mathcal{X}}(L, M, \sigma, \mu)$. However, we cannot assess the quality of these bounds, since the stepsize parameters $\{\alpha_t\}$ and $\{\gamma_t\}$ are not specified yet. We will show how these parameters are specified for solving some special classes of SCO problems in Section 4. Second, while Theorem 1.a) evaluates the efficiency of the generic AC-SA algorithm on average over many runs of the algorithm, Theorem 1.b) estimates the quality of the solutions obtained by a single run of the generic AC-SA algorithm. It should be noted, however, that the bound $\mathcal{B}_p(t)$ defined in (3.11) can be significantly larger than the bound $\mathcal{B}_e(t)$ defined in (3.9) (See Subsection 4.2). Finally, observe that one can rarely compute the error measure of $\Psi(x_t^{ag}) - \Psi^*$, since $\Psi^*$ is usually unknown in practice. As a consequence, to check the solution accuracy seems to be a difficult problem for the AC-SA algorithm. In Section 5, we will show that, with little additional computational effort, one can compute online lower and upper bounds of $\Psi^*$ to check the accuracy of the solutions generated by the generic AC-SA algorithm.

### 3.2 Convergence analysis

Our main goal in this subsection is to prove the convergence results of the AC-SA algorithm described in Theorem 1. We first establish two technical results. Lemma 2 states some properties of the projection step (3.2) and Lemma 3 describes some properties about the composite function $\Psi$. This is followed by two intermediate results, i.e., Propositions 4 and 5, which summarize some important recursions of the AC-SA algorithm. Theorem 1 then follows directly from Proposition 5.

The first technical result below characterizes the solution of the projection step (3.2). It is worth noting that the function $\omega$ is not necessarily strongly convex.

**Lemma 2** *Let the convex function $p : X \to \mathbb{R}$, the points $\tilde{x}, \tilde{y} \in X$ and the scalars $\mu_1, \mu_2 \geq 0$ be given. Let $\omega : X \to \mathbb{R}$ be a differentiable convex function and $V(x, z)$ be defined in (2.2). If*

$$u^* \in \text{Argmin}\{p(u) + \mu_1 V(\tilde{x}, u) + \mu_2 V(\tilde{y}, u) : u \in X\},$$

*then for any $u \in X$, we have*

$$p(u^*) + \mu_1 V(\tilde{x}, u^*) + \mu_2 V(\tilde{y}, u^*) \leq p(u) + \mu_1 V(\tilde{x}, u) + \mu_2 V(\tilde{y}, u) - (\mu_1 + \mu_2) V(u^*, u).$$

*Proof.* The definition of $u^*$ and the fact $V(\tilde{x}, \cdot)$ is a differentiable convex function imply that, for some $p'(u^*) \in \partial p(u^*)$, we have

$$\langle p'(u^*) + \mu_1 \nabla V(\tilde{x}, u^*) + \mu_2 \nabla V(\tilde{y}, u^*), u - u^* \rangle \geq 0, \quad \forall u \in X,$$

where $\nabla V(\tilde{x}, u^*)$ denotes the gradient of $V(\tilde{x}, \cdot)$ at $u^*$. Using the definition of $V(x, z)$ in (2.2), it is easy to verify that

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

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

$$\begin{aligned}
p(u) + \mu_1 V(\tilde{x}, u) + \mu_2 V(\tilde{y}, u) &= p(u) + \mu_1 [V(\tilde{x}, u^*) + \langle \nabla V(\tilde{x}, u^*), u - u^* \rangle + V(u^*, u)] \\
&\quad + \mu_2 [V(\tilde{y}, u^*) + \langle \nabla V(\tilde{y}, u^*), u - u^* \rangle + V(u^*, u)] \\
&\geq p(u^*) + \mu_1 V(\tilde{x}, u^*) + \mu_2 V(\tilde{y}, u^*) \\
&\quad + \langle p'(u^*) + \mu_1 \nabla V(\tilde{x}, u^*) + \mu_2 \nabla V(\tilde{y}, u^*), u - u^* \rangle \\
&\quad + (\mu_1 + \mu_2) V(u^*, u) \\
&\geq [p(u^*) + \mu_1 V(\tilde{x}, u^*) + \mu_2 V(\tilde{y}, u^*)] + (\mu_1 + \mu_2) V(u^*, u).
\end{aligned}$$

■

The following result describes some important properties of the composite function $\Psi$.

**Lemma 3** *Let $x_t^{ag} := (1 - \alpha_t) x_{t-1}^{ag} + \alpha_t x_t$ for some $\alpha_t \in [0, 1]$ and $(x_{t-1}^{ag}, x_t) \in X \times X$. We have*

$$\Psi(x_t^{ag}) \leq (1 - \alpha_t) \Psi(x_{t-1}^{ag}) + \alpha_t [f(z) + \langle f'(z), x_t - z \rangle + \mathcal{X}(x_t)] + \frac{L}{2} \|x_t^{ag} - z\|^2 + M \|x_t^{ag} - z\|,$$

*for any $z \in X$.*

*Proof.* First observe that by the definition of $x_t^{ag}$ and the convexity of $f$, we have

$$\begin{aligned}
f(z) + \langle f'(z), x_t^{ag} - z \rangle &= f(z) + \langle f'(z), \alpha_t x_t + (1 - \alpha_t) x_{t-1}^{ag} - z \rangle \\
&= (1 - \alpha_t)[f(z) + \langle f'(z), x_{t-1}^{ag} - z \rangle] + \alpha_t [f(z) + \langle f'(z), x_t - z \rangle] \\
&\leq (1 - \alpha_t) f(x_{t-1}^{ag}) + \alpha_t [f(z) + \langle f'(z), x_t - z \rangle].
\end{aligned}$$

Using this observation, (1.1), (1.2), (3.3), the definition of $x_t^{ag}$ and the convexity of $\mathcal{X}(x)$, we have

$$\begin{aligned}
\Psi(x_t^{ag}) &= f(x_t^{ag}) + \mathcal{X}(x_t^{ag}) \leq f(z) + \langle f'(z), x_t^{ag} - z \rangle + \frac{L}{2} \|x_t^{ag} - z\|^2 + M \|x_t^{ag} - z\| + \mathcal{X}(x_t^{ag}) \\
&\leq (1 - \alpha_t) f(x_{t-1}^{ag}) + \alpha_t [f(z) + \langle f'(z), x_t - z \rangle] + \frac{L}{2} \|x_t^{ag} - z\|^2 + M \|x_t^{ag} - z\| \\
&\quad + (1 - \alpha_t) \mathcal{X}(x_{t-1}^{ag}) + \alpha_t \mathcal{X}(x_t) \\
&= (1 - \alpha_t) \Psi(x_{t-1}^{ag}) + \alpha_t [f(z) + \langle f'(z), x_t - z \rangle + \mathcal{X}(x_t)] + \frac{L}{2} \|x_t^{ag} - z\|^2 + M \|x_t^{ag} - z\|.
\end{aligned}$$

■

In the sequel, we use $\delta_t$, $t \geq 1$, to denote the error for the computation of the subgradient of $f$, i.e.,

$$\delta_t \equiv G(x_t^{md}, \xi_t) - f'(x_t^{md}), \quad \forall\, t \geq 1, \tag{3.14}$$

where $f'(x_t^{md})$ represents an arbitrary element of $\partial f(x_t^{md})$ wherever it appears.

The following proposition establishes a basic recursion for the generic AC-SA algorithm.

**Proposition 4** *Let $(x_{t-1}, x_{t-1}^{ag}) \in X \times X$ be given. Also let $(x_t^{md}, x_t, x_t^{ag}) \in X \times X \times X$ be computed according to (3.1), (3.2) and (3.3). If condition (3.6) holds, then for any $x \in X$, we have*

$$\begin{array}{rcl}\Psi(x_t^{ag}) + \mu V(x_t, x) & \leq & (1-\alpha_t)[\Psi(x_{t-1}^{ag}) + \mu V(x_{t-1}, x)] + \alpha_t l_\Psi(x_t^{md}, x) \\ & & + \gamma_t[V(x_{t-1}, x) - V(x_t, x)] + \Delta_t(x),\end{array} \tag{3.15}$$

*where $l_\Psi(z, x)$ is defined in (3.4),*

$$\Delta_t(x) \;:=\; \alpha_t \langle \delta_t, x - x_{t-1}^+ \rangle + \frac{\alpha_t^2 (M + \|\delta_t\|_*)^2}{\nu(\mu+\gamma_t) - L\alpha_t^2}, \tag{3.16}$$

$$x_{t-1}^+ \;:=\; \frac{\alpha_t \mu}{\mu + \gamma_t} x_t^{md} + \frac{(1-\alpha_t)\mu + \gamma_t}{\mu + \gamma_t} x_{t-1}. \tag{3.17}$$

*Proof.* We first establish some basic relations among the search points $x_t^{ag}$, $x_t^{md}$ and $x_t$. Denote $d_t := x_t^{ag} - x_t^{md}$. It follows from (3.1) and (3.3) that

$$\begin{array}{rcl}d_t & = & \alpha_t x_t + (1-\alpha_t) x_{t-1}^{ag} - x_t^{md} \\ & = & \alpha_t \left( x_t - \dfrac{\alpha_t \mu}{\mu+\gamma_t} x_t^{md} - \dfrac{(1-\alpha_t)\mu + \gamma_t}{\mu + \gamma_t} x_{t-1} \right) = \alpha_t (x_t - x_{t-1}^+),\end{array} \tag{3.18}$$

which, in view of the convexity of $\|\cdot\|^2$ and the strong-convexity of $\omega$, implies that

$$\begin{array}{rcl}\dfrac{\nu(\mu+\gamma_t)}{2\alpha_t^2}\|d_t\|^2 & \leq & \dfrac{\nu(\mu+\gamma_t)}{2(\mu+\gamma_t)}\left[\alpha_t \mu \|x_t - x_t^{md}\|^2 + [(1-\alpha_t)\mu + \gamma_t]\,\|x_t - x_{t-1}\|^2\right] \\ & \leq & \alpha_t \mu V(x_t^{md}, x_t) + [(1-\alpha_t)\mu + \gamma_t] V(x_{t-1}, x_t).\end{array} \tag{3.19}$$

Using the above result and Lemma 3 (with $z = x_t^{md}$), we have

$$\begin{array}{rcl}\Psi(x_t^{ag}) & \leq & (1-\alpha_t)\Psi(x_{t-1}^{ag}) + \alpha_t[f(x_t^{md}) + \langle f'(x_t^{md}), x_t - x_t^{md}\rangle + \mathcal{X}(x_t)] + \dfrac{L}{2}\|d_t\|^2 + M\|d_t\| \\ & = & (1-\alpha_t)\Psi(x_{t-1}^{ag}) + \alpha_t[f(x_t^{md}) + \langle f'(x_t^{md}), x_t - x_t^{md}\rangle + \mathcal{X}(x_t)] + \\ & & \dfrac{\nu(\mu+\gamma_t)}{2\alpha_t^2}\|d_t\|^2 - \dfrac{\nu(\mu+\gamma_t) - L\alpha_t^2}{2\alpha_t^2}\|d_t\|^2 + M\|d_t\| \\ & \leq & (1-\alpha_t)\Psi(x_{t-1}^{ag}) + \alpha_t\left[f(x_t^{md}) + \langle f'(x_t^{md}), x_t - x_t^{md}\rangle + \mathcal{X}(x_t) + \mu V(x_t^{md}, x_t)\right] + \\ & & [(1-\alpha_t)\mu + \gamma_t] V(x_{t-1}, x_t) - \dfrac{\nu(\mu+\gamma_t) - L\alpha_t^2}{2\alpha_t^2}\|d_t\|^2 + M\|d_t\|.\end{array} \tag{3.20}$$

Now let us apply the results regarding the projection step in (3.2). Specifically, by using Lemma 2 with $p(u) = \alpha_t [\langle G_t, u\rangle + \mathcal{X}(u)]$, $\mu_1 = \alpha_t \mu$, $\mu_2 = (1-\alpha_t)\mu + \gamma_t$, $\tilde{x} = x_t^{md}$ and $\tilde{y} = x_{t-1}$, we have

$$\begin{aligned}
&\alpha_t[f(x_t^{md}) + \langle G_t, x_t - x_t^{md}\rangle + \mathcal{X}(x_t) + \mu V(x_t^{md}, x_t)] + [(1-\alpha_t)\mu + \gamma_t]V(x_{t-1}, x_t)\\
&\le \alpha_t[f(x_t^{md}) + \langle G_t, x - x_t^{md}\rangle + \mathcal{X}(x) + \mu V(x_t^{md}, x)] + [(1-\alpha_t)\mu + \gamma_t]V(x_{t-1}, x)\\
&\quad -(\mu+\gamma_t)V(x_t, x)\\
&= \alpha_t l_\Psi(x_t^{md}, x) + \alpha_t\langle \delta_t, x - x_t^{md}\rangle + [(1-\alpha_t)\mu + \gamma_t]V(x_{t-1}, x) - (\mu+\gamma_t)V(x_t, x), \quad (3.21)
\end{aligned}$$

for any $x \in X$, where the last equality follows from (3.4) and (3.14). Combining (3.20) and (3.21), and using the fact that $f'(x_t^{md}) = G_t - \delta_t$ due to (3.14), we obtain

$$\begin{aligned}
\Psi(x_t^{ag}) \;\le\; & (1-\alpha_t)[\Psi(x_{t-1}^{ag}) + \mu V(x_{t-1}, x)] + \alpha_t l_\Psi(x_t^{md}, x) + \gamma_t[V(x_{t-1}, x) - V(x_t, x)]\\
& - \mu V(x_t, x) \underbrace{- \frac{\nu(\mu+\gamma_t) - L\alpha_t^2}{2\alpha_t^2}\|d_t\|^2 + M\|d_t\| + \alpha_t\langle \delta_t, x - x_t\rangle}_{U_t}, \quad \forall x \in X.
\end{aligned}$$

It remains to show that the term $U_t$ defined above is bounded by $\Delta_t(x)$ in (3.16). Indeed, we have, by the last identity in (3.18),

$$\begin{aligned}
U_t &= -\frac{\nu(\mu+\gamma_t) - L\alpha_t^2}{2\alpha_t^2}\|d_t\|^2 + M\|d_t\| - \langle \delta_t, d_t\rangle + \langle \delta_t, d_t + \alpha_t(x - x_t)\rangle\\
&= -\frac{\nu(\mu+\gamma_t) - L\alpha_t^2}{2\alpha_t^2}\|d_t\|^2 + M\|d_t\| - \langle \delta_t, d_t\rangle + \alpha_t\langle \delta_t, x - x_{t-1}^+\rangle\\
&\le -\frac{\nu(\mu+\gamma_t) - L\alpha_t^2}{2\alpha_t^2}\|d_t\|^2 + (M + \|\delta_t\|_*)\|d_t\| + \alpha_t\langle \delta_t, x - x_{t-1}^+\rangle\\
&\le \frac{\alpha_t^2}{\nu(\mu+\gamma_t) - L\alpha_t^2}(M + \|\delta_t\|_*)^2 + \alpha_t\langle \delta_t, x - x_{t-1}^+\rangle = \Delta_t(x), \qquad (3.22)
\end{aligned}$$

where the last inequality follows from the maximization of a simple concave quadratic function w.r.t. $\|d_t\|$. ■

Proposition 5 below follows from Proposition 4 by taking summation of the relations in (3.15). This result will also be used later in Section 5 for validation analysis.

**Proposition 5** *Let $\{x_t^{ag}\}_{t\ge1}$ be computed by the generic AC-SA algorithm for solving $\mathcal{F}_{X,\mathcal{X}}(L, M, \sigma, \mu)$. Also assume that $\{\alpha_t\}_{t\ge1}$ and $\{\gamma_t\}_{t\ge1}$ are chosen such that relation (3.6) holds. We have*

$$\Psi(x_t^{ag}) + \mu V(x_t, x) - \Gamma_t \sum_{\tau=1}^t \left[\frac{\alpha_\tau}{\Gamma_\tau} l_\Psi(x_\tau^{md}, x)\right] \le \Gamma_t \sum_{\tau=1}^t \frac{\gamma_\tau}{\Gamma_\tau}[V(x_{\tau-1}, x) - V(x_\tau, x)] + \Gamma_t \sum_{\tau=1}^t \frac{\Delta_\tau(x)}{\Gamma_\tau}, \tag{3.23}$$

*for any $x \in X$ and any $t \ge 1$, where $l_\Psi(z, x)$ and $\Delta_\tau(x)$ are defined in (3.4) and (3.16), respectively.*

*Proof.* Dividing both sides of relation (3.15) by $\Gamma_t$, and using the definition of $\Gamma_t$ in (3.8) and the fact that $\alpha_1 = 1$, we have

$$\begin{aligned}
\frac{1}{\Gamma_t}[\Psi(x_t^{ag}) + \mu V(x_t, x)] \;\le\; & \frac{1}{\Gamma_{t-1}}[\Psi(x_{t-1}^{ag}) + \mu V(x_{t-1}, x)] + \frac{\alpha_t}{\Gamma_t} l_\Psi(x_t^{md}, x)\\
& + \frac{\gamma_t}{\Gamma_t}[V(x_{t-1}, x) - V(x_t, x)] + \frac{\Delta_t(x)}{\Gamma_t}, \ \forall t \ge 2,
\end{aligned}$$

and

$$\frac{1}{\Gamma_1}\left[\Psi(x_1^{ag}) + \mu V(x_1, x)\right] \quad \leq \quad \frac{\alpha_1}{\Gamma_1} l_\Psi(x_1^{md}, x) + \frac{\gamma_1}{\Gamma_1}[V(x_0, x) - V(x_1, x)] + \frac{\Delta_1(x)}{\Gamma_1}.$$

Summing up the above inequalities, we have

$$\frac{1}{\Gamma_t}\left[\Psi(x_t^{ag}) + \mu V(x_t, x)\right] \leq \sum_{\tau=1}^{t} \frac{\alpha_\tau}{\Gamma_\tau} l_\Psi(x_t^{md}, x) + \sum_{\tau=1}^{t} \frac{\gamma_\tau}{\Gamma_\tau}[V(x_{\tau-1}, x) - V(x_\tau, x)] + \sum_{\tau=1}^{t} \frac{\Delta_\tau(x)}{\Gamma_\tau},$$

which clearly implies (3.23). ■

To prove Theorem 1, we also need the following well-known result for the martingale-difference. A proof of this result can be found, for example, in [23].

**Lemma 6** *Let $\xi_1, \xi_2, ...$ be a sequence of i.i.d 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 for any $\Lambda \geq 0$,* $\operatorname{Prob}\left\{\sum_{t=1}^N \zeta_t > \Lambda \left(\sum_{t=1}^N \sigma_t^2\right)^{\frac{1}{2}}\right\} \leq \exp\{-\Lambda^2/3\}$.

We are now ready to prove Theorem 1.

**Proof of Theorem 1:** We first show Part a). Observe that by the definition of $\Gamma_t$ in (3.8) and the fact that $\alpha_1 = 1$, we have

$$\sum_{\tau=1}^{t} \frac{\alpha_\tau}{\Gamma_\tau} = \frac{\alpha_1}{\Gamma_1} + \sum_{\tau=2}^{t} \frac{1}{\Gamma_\tau}\left(1 - \frac{\Gamma_\tau}{\Gamma_{\tau-1}}\right) = \frac{1}{\Gamma_1} + \sum_{\tau=2}^{t}\left(\frac{1}{\Gamma_\tau} - \frac{1}{\Gamma_{\tau-1}}\right) = \frac{1}{\Gamma_t}. \tag{3.24}$$

Using the previous observation and (3.5), we obtain

$$\Gamma_t \sum_{\tau=1}^{t}\left[\frac{\alpha_\tau}{\Gamma_\tau} l_\Psi(x_\tau^{md}, x)\right] \leq \Gamma_t \sum_{\tau=1}^{t}\left[\frac{\alpha_\tau}{\Gamma_\tau}\Psi(x)\right] = \Psi(x), \quad \forall x \in X. \tag{3.25}$$

Moreover, it follows from the condition (3.7) that

$$\Gamma_t \sum_{\tau=1}^{t} \frac{\gamma_\tau}{\Gamma_\tau}[V(x_{\tau-1}, x) - V(x_\tau, x)] = \Gamma_t \frac{\gamma_1}{\Gamma_1}[V(x_0, x) - V(x_t, x)] \leq \Gamma_t \gamma_1 V(x_0, x), \tag{3.26}$$

where the last inequality follows from the facts that $\Gamma_1 = 1$ and that $V(x_t, x) \geq 0$. Using the fact that $V(x_t, x) \geq 0$ due to (2.2) and replacing the above two bounds into (3.23), we have

$$\Psi(x_t^{ag}) - \Psi(x) \leq \Psi(x_t^{ag}) + \mu V(x_t, x) - \Psi(x) \leq \Gamma_t \gamma_1 V(x_0, x) + \Gamma_t \sum_{\tau=1}^{t} \frac{\Delta_\tau(x)}{\Gamma_\tau}, \quad \forall x \in X, \tag{3.27}$$

where $\Delta_\tau(x)$ is defined in (3.16). Observe that the triple $(x_t^{md}, x_{t-1}, x_{t-1}^{ag})$ 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 (3.27) and noting that under Assumption A1, $\mathbb{E}[\|\delta_\tau\|_*^2] \leq \sigma^2$, and

$$\mathbb{E}_{|\xi_{[\tau-1]}}[\langle \delta_\tau, x^* - x_{t-1}^+ \rangle] = 0, \tag{3.28}$$

we have

$$\begin{aligned}\mathbb{E}\left[\Psi(x_t^{ag}) - \Psi^*\right] &\le \Gamma_t \gamma_1 V(x_0, x^*) + \Gamma_t \sum_{\tau=1}^{t} \frac{\alpha_\tau^2 \mathbb{E}\left[(M + \|\delta_\tau\|_*)^2\right]}{\Gamma_\tau\left[\nu(\mu+\gamma_\tau) - L\alpha_\tau^2\right]} \\ &\le \Gamma_t \gamma_1 V(x_0, x^*) + \Gamma_t \sum_{\tau=1}^{t} \frac{2\alpha_\tau^2 (M^2 + \sigma^2)}{\Gamma_\tau\left[\nu(\mu+\gamma_\tau) - L\alpha_\tau^2\right]}.\end{aligned}$$

To show part b), let us denote $\zeta_\tau := \Gamma_\tau^{-1}\alpha_\tau \langle \delta_\tau, x^* - x_{\tau-1}^+\rangle$. Clearly, from the definition of $R_X(x^*)$ given by (3.12), we have $\|x^* - x_{\tau-1}^+\| \le R_X(x^*)$, which together with Assumption A2 imply that

$$\begin{aligned}\mathbb{E}_{|\xi_{[\tau-1]}}\left[\exp\{\zeta_\tau^2/[\Gamma_\tau^{-1}\alpha_\tau \sigma R_X(x^*)]^2\}\right] &\le \mathbb{E}_{|\xi_{[\tau-1]}}\left[\exp\{(\|\delta_\tau\|_* \|x^* - x_{\tau-1}^+\|)^2/[\sigma R_X(x^*)]^2\}\right] \\ &\le \mathbb{E}_{|\xi_{[\tau-1]}}\left[\exp\{(\|\delta_\tau\|_*)^2/\sigma^2\}\right] \le \exp(1).\end{aligned}$$

Moreover, it follows from (3.28) that $\{\zeta_\tau\}_{\tau \ge 1}$ is a martingale-difference. Using the previous two observations and Lemma 6, we have

$$\forall \lambda \ge 0 : \text{Prob}\left\{\sum_{\tau=1}^{t} \zeta_\tau > \lambda\, \sigma R_X(x^*) \left[\sum_{\tau=1}^{t} (\Gamma_\tau^{-1}\alpha_\tau)^2\right]^{\frac{1}{2}}\right\} \le \exp\{-\lambda^2/3\}. \tag{3.29}$$

Also observe that under Assumption A2, $\mathbb{E}_{|\xi_{[\tau-1]}}\left[\exp\{\|\delta_\tau\|_*^2/\sigma^2\}\right] \le \exp\{1\}$. Setting

$$\pi_\tau^2 = \frac{\alpha_\tau^2}{\Gamma_\tau[\nu(\mu+\gamma_\tau) - L\alpha_\tau^2]} \quad \text{and} \quad \theta_\tau = \frac{\pi_\tau^2}{\sum_{\tau=1}^{t} \pi_\tau^2},$$

we have

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

whence, taking expectations,

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

It then follows from Markov's inequality that

$$\forall \lambda \ge 0 : \text{Prob}\left\{\sum_{\tau=1}^{t} \pi_\tau^2 \|\delta_\tau\|_*^2 > (1+\lambda)\sigma^2 \sum_{\tau=1}^{t} \pi_\tau^2\right\} \le \exp\{-\lambda\}. \tag{3.30}$$

Combining (3.27), (3.29), and (3.30), and rearranging the terms, we obtain (3.10).

Finally, observing that by the condition (3.13), the fact that $V(u,x) \ge 0$ and the definition of $\bar{V}(x)$,

$$\begin{aligned}\Gamma_t \sum_{\tau=1}^{t} \frac{\gamma_\tau}{\Gamma_\tau}[V(x_{\tau-1}, x) - V(x_\tau, x)] &\le \Gamma_t \left[\frac{\gamma_1}{\Gamma_1}\bar{V}(x) + \sum_{\tau=2}^{t}\left(\frac{\gamma_\tau}{\Gamma_\tau} - \frac{\gamma_{\tau-1}}{\Gamma_{\tau-1}}\right)\bar{V}(x) - \frac{\gamma_t}{\Gamma_t} V(x_t, x)\right] \\ &\le \gamma_t \bar{V}(x) - \gamma_t V(x_t, x) \le \gamma_t \bar{V}(x),\end{aligned} \tag{3.31}$$

we can show part c) similarly to part a) and part b) by replacing the bound in (3.26) with the one given above. ∎

# 4 Optimal and nearly optimal algorithms for SCO

Our goal in this section is to specialize the generic AC-SA algorithm to obtain optimal or nearly optimal algorithms for solving different types of SCO problems.

## 4.1 Optimal AC-SA algorithms for problems without strong convexity

In this subsection, we consider problem (1.1), but now the objective function $f$ is not necessarily strongly convex. We present the AC-SA algorithms for solving these problems by setting $\mu = 0$ and properly choosing the stepsize parameters $\{\alpha_t\}_{t\geq 1}$ and $\{\gamma_t\}_{t\geq 1}$ in the generic AC-SA algorithmic framework.

Observe that, if $\mu$ is set to 0 in the generic AC-SA algorithm, then the identities (3.1) and (3.2), respectively, reduce to

$$
\begin{aligned}
x_t^{md} &= (1-\alpha_t) x_{t-1}^{ag} + \alpha_t x_{t-1}, && (4.1)\\
x_t &= \arg\min_{x\in X} \left\{ \alpha_t \left[ \langle G_t, x\rangle + \mathcal{X}(x) \right] + \gamma_t V(x_{t-1}, x) \right\}. && (4.2)
\end{aligned}
$$

We will study and compare two AC-SA algorithms for $\mathcal{F}_{X,\mathcal{X}}(L, M, \sigma, 0)$, each of them employed with a different *stepsize policy to choose* $\{\alpha_t\}_{t\geq 1}$ *and* $\{\gamma_t\}_{t\geq 1}$.

The first stepsize policy and its associated convergence results stated below are similar to those introduced in [22]. This result follows as an immediate consequence of Theorem 1.

**Proposition 7** *Let* $\{x_t^{ag}\}_{t\geq 1}$ *be computed by the AC-SA algorithm for* $\mathcal{F}_{X,\mathcal{X}}(L, M, \sigma, 0)$ *with*

$$
\alpha_t = \frac{2}{t+1} \quad \textit{and} \quad \gamma_t = \frac{4\gamma}{\nu t(t+1)}, \quad \forall\, t \geq 1, \tag{4.3}
$$

*for some* $\gamma \geq 2L$*. Then, under Assumption A1, we have* $\mathbb{E}[\Psi(x_t^{ag}) - \Psi^*] \leq \mathcal{C}_{e,1}(t)$, $\forall t \geq 1$, *where*

$$
\mathcal{C}_{e,1}(t) \equiv \mathcal{C}_{e,1}(x_0, \gamma, t) := \frac{4\gamma V(x_0, x^*)}{\nu t(t+1)} + \frac{4(M^2+\sigma^2)(t+2)}{3\gamma}. \tag{4.4}
$$

*If in addition, Assumption A2 holds, then,* $\forall \lambda > 0, \forall t \geq 1$,

$$
\operatorname{Prob}\left\{ \Psi(x_t^{ag}) - \Psi^* > \mathcal{C}_{e,1}(t) + \lambda\, \mathcal{C}_{p,1}(t) \right\} \leq \exp\{-\lambda^2/3\} + \exp\{-\lambda\}, \tag{4.5}
$$

*where*

$$
\mathcal{C}_{p,1}(t) \equiv \mathcal{C}_{p,1}(\gamma, t) := \frac{2\sigma R_X(x^*)}{\sqrt{3t}} + \frac{4\sigma^2(t+2)}{3\gamma}. \tag{4.6}
$$

*Proof.* Clearly, by the definition of $\Gamma_t$ in (3.8), the stepsize policy (4.3), and the facts that $\gamma \geq 2L$ and $\mu = 0$, we have

$$
\Gamma_t = \frac{2}{t(t+1)}, \quad \frac{\gamma_t}{\Gamma_t} = \gamma_1 = \frac{2\gamma}{\nu}, \quad \nu(\mu+\gamma_t) - L\alpha_t^2 \geq \nu\gamma_t - \frac{\gamma\alpha_t^2}{2} \geq \frac{2\gamma}{(t+1)^2}, \quad \forall t \geq 1, \tag{4.7}
$$

and hence the specification of $\alpha_t$ and $\gamma_t$ in (4.3) satisfies conditions (3.6) and (3.7). It can also be easily seen from the previous result and (4.3) that

$$\begin{aligned}\sum_{\tau=1}^{t}\frac{\alpha_\tau^2}{\Gamma_\tau(\nu\gamma_\tau - L\alpha_\tau^2)} &\le \sum_{\tau=1}^{t}\frac{\tau(\tau+1)}{\gamma} = \frac{1}{3\gamma}t(t+1)(t+2), &(4.8)\\ \sum_{\tau=1}^{t}(\Gamma_\tau^{-1}\alpha_\tau)^2 &= \sum_{\tau=1}^{t}\tau^2 = \frac{t(t+1)(2t+1)}{6} \le \frac{t(t+1)^2}{3}, &(4.9)\end{aligned}$$

Now let $\mathcal{B}_e(t)$ and $\mathcal{B}_p(t)$ be defined in (3.9) and (3.11) respectively. By (4.7), (4.8) and (4.9), we have

$$\begin{aligned}\mathcal{B}_e(t) &\le \Gamma_t\left[\gamma_1 V(x_0,x^*) + \frac{2(M^2+\sigma^2)}{3\gamma}t(t+1)(t+2)\right] = \mathcal{C}_{e,1}(t),\\ \mathcal{B}_p(t) &\le \Gamma_t\left[\sigma R_X(x^*)\left(\frac{t(t+1)^2}{3}\right)^{\frac{1}{2}} + \frac{2\sigma^2}{3\gamma}t(t+1)(t+2)\right] = \mathcal{C}_{p,1}(t),\end{aligned}$$

which, in view of Theorem 1, clearly imply our results. ■

We now briefly discuss how to derive the optimal rate of convergence for solving $\mathcal{F}_{X,\mathcal{X}}(L,M,\sigma,0)$. Given a fixed in advance number of iterations $N$, let us suppose that the stepsize parameters $\{\alpha_t\}_{t=1}^N$ and $\{\gamma_t\}_{t=1}^N$ are set to (4.3) with

$$\gamma = \gamma_N^* = \max\left\{2L, \left[\frac{\nu(M^2+\sigma^2)N(N+1)(N+2)}{3V(x_0,x^*)}\right]^{\frac{1}{2}}\right\}. \tag{4.10}$$

Note that $\gamma_N^*$ in (4.10) is obtained by minimizing $\mathcal{C}_{e,1}(N)$ (c.f. (4.4)) with respect to $\gamma$ over the interval $[2L,+\infty)$. Then, it can be shown from (4.4) and (4.6) that

$$\begin{aligned}\mathcal{C}_{e,1}(x_0,\gamma_N^*,N) &\le \frac{8LV(x_0,x^*)}{\nu N(N+1)} + \frac{8\sqrt{(M^2+\sigma^2)V(x_0,x^*)}}{\sqrt{\nu(N+1)}} =: \mathcal{C}_{e,1}^*(N), &(4.11)\\ \mathcal{C}_{p,1}(\gamma_N^*,N) &\le \frac{2\sigma R_X(x^*)}{\sqrt{3N}} + \frac{4\sigma\sqrt{V(x_0,x^*)}}{\sqrt{\nu(N+1)}} =: \mathcal{C}_{p,1}^*(N). &(4.12)\end{aligned}$$

Indeed, let

$$\bar\gamma := \left[\frac{\nu(M^2+\sigma^2)N(N+1)(N+2)}{3V(x_0,x^*)}\right]^{\frac{1}{2}}.$$

According to the relation (4.10), we have $\bar\gamma \le \gamma_N^* \le 2L+\bar\gamma$. Using these facts and (4.4) we obtain

$$\begin{aligned}\mathcal{C}_{e,1}(x_0,\gamma_N^*,N) &\le \frac{4(2L+\bar\gamma)V(x_0,x^*)}{\nu N(N+1)} + \frac{4(M^2+\sigma^2)(N+2)}{3\bar\gamma}\\ &= \frac{8LV(x_0,x^*)}{\nu N(N+1)} + 4\left[\frac{\bar\gamma V(x_0,x^*)}{\nu N(N+1)} + \frac{(M^2+\sigma^2)(N+2)}{3\bar\gamma}\right]\\ &= \frac{8LV(x_0,x^*)}{\nu N(N+1)} + \frac{8\sqrt{(M^2+\sigma^2)(N+2)V(x_0,x^*)}}{\sqrt{3\nu N(N+1)}}.\end{aligned}$$

Noting that $N+2 \le 3N$ in the second term of the above equation, we obtain (4.11). Also by (4.6), we have

$$\mathcal{C}_{p,1}(N) \le \frac{2\sigma R_X(x^*)}{\sqrt{3N}} + \frac{4\sigma^2(N+2)}{3\bar{\gamma}},$$

which leads to (4.12).

Hence, by Proposition 7, we have, under Assumption A1, $\mathbb{E}[\Psi(x_N^{ag}) - \Psi^*] \le \mathcal{C}_{e,1}^*(N)$, which gives us an optimal expected rate of convergence for solving $\mathcal{F}_{X,\mathcal{X}}(L, M, \sigma, 0)$. Moreover, if Assumption A2 holds, then $\mathrm{Prob}\{\Psi(x_N^{ag}) - \Psi^* \ge \mathcal{C}_{e,1}^*(N) + \lambda \mathcal{C}_{p,1}^*(N)\} \le \exp(-\lambda^2/3) + \exp(-\lambda)$. It is worth noting that both $\mathcal{C}_{p,1}^*$ and $\mathcal{C}_{e,1}^*$ are in the same order of magnitude, i.e., $\mathcal{O}(1/\sqrt{N}\,)$. Observe that we need to estimate a bound on $V(x_0, x^*)$ to implement this stepsize policy since $V(x_0, x^*)$ is usually unknown.

One possible drawback of the stepsize policy (4.3) with $\gamma = \gamma_N^*$ is the need of fixing $N$ in advance. In Proposition 8, we propose an alternative stepsize policy which does not require to fix the number of iterations $N$. Note that, to apply this stepsize policy properly, we need to assume that all the iterates $\{x_t\}_{t\ge1}$ stay in a bounded set. A similar stepsize policy was recently developed by Hu et al. [14]. However, their work focused on unconstrained CP problems, for which the boundedness of $\{x_t\}_{t\ge1}$ and hence the convergence of their algorithm cannot be guaranteed, theoretically speaking.

**Proposition 8** *Assume that $X$ is compact. Let $\{x_t^{ag}\}_{t\ge1}$ be computed by the AC-SA algorithm for $\mathcal{F}_{X,\mathcal{X}}(L, M, \sigma, 0)$ with*

$$\alpha_t = \frac{2}{t+1} \quad \text{and} \quad \gamma_t = \frac{4L}{\nu t(t+1)} + \frac{2\gamma}{\nu\sqrt{t}}, \quad \forall\, t \ge 1, \tag{4.13}$$

*for some $\gamma > 0$. Then, under Assumption A1, we have $\mathbb{E}[\Psi(x_N^{ag}) - \Psi^*] \le \mathcal{C}_{e,2}(t)$, $\forall t \ge 1$, where*

$$\mathcal{C}_{e,2}(t) \equiv \mathcal{C}_{e,2}(\gamma, t) := \frac{4L\bar{V}(x^*)}{\nu t(t+1)} + \frac{2\gamma \bar{V}(x^*)}{\nu\sqrt{t}} + \frac{8\sqrt{2}}{3\gamma\sqrt{t}}(M^2 + \sigma^2), \tag{4.14}$$

*and $\bar{V}(\cdot)$ is defined in Theorem 1.c). If in addition, Assumption A2 holds, then, $\forall \lambda > 0$, $\forall t \ge 1$,*

$$\mathrm{Prob}\left\{\Psi(x_t^{ag}) - \Psi^* > \mathcal{C}_{e,2}(t) + \lambda\, \mathcal{C}_{p,2}(t)\right\} \le \exp\{-\lambda^2/3\} + \exp\{-\lambda\}, \tag{4.15}$$

*where*

$$\mathcal{C}_{p,2}(t) \equiv \mathcal{C}_{p,2}(\gamma, t) := \frac{2\sigma R_X(x^*)}{\sqrt{3t}} + \frac{8\sqrt{2}\sigma^2}{3\gamma\sqrt{t}}. \tag{4.16}$$

*Proof.* Clearly, by the definition of $\Gamma_t$ in (3.8), the stepsize policy (4.13) and the fact that $\mu = 0$, we have

$$\Gamma_t = \frac{2}{t(t+1)}, \quad \frac{\gamma_t}{\Gamma_t} = \frac{2L}{\nu} + \frac{\gamma}{\nu}(t+1)\sqrt{t}, \quad \nu(\mu + \gamma_t) - L\alpha_t^2 = \nu\gamma_t - L\alpha_t^2 \ge \frac{2\gamma}{\sqrt{t}}, \tag{4.17}$$

and hence the specification of $\alpha_t$ and $\gamma_t$ in (4.13) satisfies conditions (3.6) and (3.13). It can also be easily seen from the previous observations and (4.13) that

$$\sum_{\tau=1}^{t} (\Gamma_\tau^{-1}\alpha_\tau)^2 = \sum_{\tau=1}^{t} \tau^2 = \frac{t(t+1)(2t+1)}{6} \le \frac{t(t+1)^2}{3}, \tag{4.18}$$

$$\sum_{\tau=1}^{t}\frac{\alpha_\tau^2}{\Gamma_\tau(\nu\gamma_\tau - L\alpha_\tau^2)} \le \sum_{\tau=1}^{t}\frac{\sqrt{\tau}}{\gamma} \le \frac{1}{\gamma}\int_1^{t+1}\sqrt{x}dx \le \frac{2}{3\gamma}(t+1)^{\frac{3}{2}} \le \frac{2\sqrt{2}}{3\gamma}(t+1)\sqrt{t}. \tag{4.19}$$

Now let $\mathcal{B}_e'(t)$ be obtained by replacing the first term in the definition of $\mathcal{B}_e(t)$ in (3.9) with $\gamma_t \bar{V}(x^*)$ and $\mathcal{B}_p(t)$ be defined in (3.11). By (4.13), (4.17), (4.18) and (4.19), we have

$$\begin{aligned}
\mathcal{B}_e'(t) &\le \gamma_t \bar{V}(x^*) + \Gamma_t \frac{4\sqrt{2}(M^2+\sigma^2)}{3\gamma}(t+1)\sqrt{t} = \mathcal{C}_{e,2}(t), \\
\mathcal{B}_p(t) &\le \Gamma_t\left[\sigma R_X(x^*)\left(\frac{t(t+1)^2}{3}\right)^{\frac{1}{2}} + \frac{4\sqrt{2}\sigma^2}{3\gamma}(t+1)\sqrt{t}\right] = \mathcal{C}_{p,2}(t),
\end{aligned}$$

which, in view of Theorem 1.c), then clearly imply our results. ■

Clearly, if we set $\gamma$ in the stepsize policy (4.13) as

$$\gamma = \tilde{\gamma}^* := \left[\frac{4\sqrt{2}(M^2+\sigma^2)\nu}{3\bar{V}(x^*)}\right]^{\frac{1}{2}},$$

then by (4.14), we have

$$\mathbb{E}[\Psi(x_N^{ag}) - \Psi^*] \le \frac{4L\bar{V}(x^*)}{\nu N(N+1)} + \frac{8}{\sqrt{\nu N}}\left[\frac{\sqrt{2}\bar{V}(x^*)}{3}(M^2+\sigma^2)\right]^{\frac{1}{2}} =: \mathcal{C}_{e,2}^*,$$

which also gives an optimal expected rate of convergence for $\mathcal{F}_{X,\mathcal{X}}(L, M, \sigma, 0)$. As discussed before, one obvious advantage of the stepsize policy (4.13) with $\gamma = \tilde{\gamma}^*$ over the one in (4.3) with $\gamma = \gamma_N^*$ is that the former one does not require the knowledge of $N$. Hence, it allows possibly earlier termination of the AC-SA algorithm, especially when coupled with the validation procedure in Subsection 5. Note however, that the convergence rate $\mathcal{C}_{e,1}^*$ depends on $V(x_0, x^*)$, which can be significantly smaller than $\bar{V}(x^*)$ in $\mathcal{C}_{e,2}^*$ given a good starting point $x_0 \in X$.

## 4.2 Nearly optimal AC-SA algorithms for strongly convex problems

The objective of this subsection is to present an AC-SA algorithm for solving $\mathcal{F}_{X,\mathcal{X}}(L, M, \sigma, \mu)$ with $\mu > 0$. We demonstrate the generic AC-SA algorithm in Subsection 3, if employed with a suitable stepsize policy, can achieve a nearly optimal expected rate of convergence for $\mathcal{F}_{X,\mathcal{X}}(L, M, \sigma, \mu)$ and the optimal expected rate of convergence for $\mathcal{F}_{X,\mathcal{X}}(0, M, \sigma, \mu)$. Also, throughout this subsection we assume that the prox-function $V(x, z)$ grows quadratically with constant 1.

We start by presenting the AC-SA algorithm for $\mathcal{F}_{X,\mathcal{X}}(L, M, \sigma, \mu)$ with a simple stepsize policy and discussing its convergence properties. It is worth noting that this stepsize policy does not depend on $\mu$, $\sigma$, $M$ and $V(x_0, x^*)$, and hence it is quite convenient for implementation.

**Proposition 9** *Let $\{x_t^{ag}\}_{t\ge1}$ be computed by the AC-SA algorithm for $\mathcal{F}_{X,\mathcal{X}}(L, M, \sigma, \mu)$ with*

$$\alpha_t = \frac{2}{t+1} \quad \textit{and} \quad \gamma_t = \frac{4L}{\nu t(t+1)}, \quad \forall\, t \ge 1. \tag{4.20}$$

*If $\mu > 0$ and condition (2.4) holds, then under Assumption A1, we have*

$$\mathbb{E}[\Psi(x_t^{ag}) - \Psi^*] \le \mathcal{D}_e(t) := \frac{4LV(x_0, x^*)}{\nu t(t+1)} + \frac{8(M^2+\sigma^2)}{\nu\mu(t+1)}, \quad \forall\, t \ge 1. \tag{4.21}$$

*If in addition, Assumption A2 holds, then, $\forall \lambda > 0, \forall t \ge 1$,*

$$\text{Prob}\left\{\Psi(x_t^{ag}) - \Psi^* \ge \mathcal{D}_e(t) + \lambda \mathcal{D}_p(t)\right\} \le \exp\{-\lambda^2/3\} + \exp\{-\lambda\}, \tag{4.22}$$

*where*

$$\mathcal{D}_p(t) := \frac{2\sigma R_X(x^*)}{\sqrt{3t}} + \frac{8\sigma^2}{\nu\mu(t+1)}. \tag{4.23}$$

*Proof.* Clearly, by the definition of $\Gamma_t$ in (3.8) and the stepsize policy (4.20), we have

$$\Gamma_t = \frac{2}{t(t+1)}, \quad \frac{\gamma_t}{\Gamma_t} = \frac{2L}{\nu}, \quad \nu(\mu+\gamma_t) - L\alpha_t^2 \ge \nu\mu, \tag{4.24}$$

and hence that the specification of $\alpha_t$ and $\gamma_t$ in (4.20) satisfies conditions (3.6) and (3.7). It can also be easily seen from the previous results and (4.20) that (4.9) holds and that

$$\sum_{\tau=1}^{t} \frac{\alpha_\tau^2}{\Gamma_\tau\left[\nu(\mu+\gamma_\tau) - L\alpha_\tau^2\right]} \le \sum_{\tau=1}^{t} \frac{\alpha_\tau^2}{\nu\mu\Gamma_\tau} \le \frac{2t}{\nu\mu}, \tag{4.25}$$

where the last inequality is due to the fact that $\alpha_\tau^2/\Gamma_\tau \le 2$ by (4.20) and (4.24). Let $\mathcal{B}_e(t)$ and $\mathcal{B}_p(t)$ be defined in (3.9) and (3.11), respectively. By (4.9), (4.24) and (4.25), we have

$$\begin{aligned}
\mathcal{B}_e(t) &\le \Gamma_t\left[\gamma_1 V(x_0, x^*) + \frac{4t(M^2+\sigma^2)}{\nu\mu}\right] = \mathcal{D}_e(t), \\
\mathcal{B}_p(t) &\le \Gamma_t\left[\sigma R_X(x^*)\left(\frac{t(t+1)^2}{3}\right)^{\frac{1}{2}} + \frac{4t\sigma^2}{\nu\mu}\right] = \mathcal{D}_p(t),
\end{aligned}$$

which, in view of Theorem 1, clearly imply our results. ∎

We now make a few remarks about the results obtained in Proposition 9. First, in view of (1.3), the AC-SA algorithm with the stepsize policy (4.20) achieves the optimal rate of convergence for solving $\mathcal{F}_{X,\mathcal{X}}(0, M, \sigma, \mu)$, i.e., for those problems without a smooth component. It is also nearly optimal for solving $\mathcal{F}_{X,\mathcal{X}}(L, M, \sigma, \mu)$, in the sense that the second term $8(M^2+\sigma^2)/[\nu\mu(t+1)]$ of $\mathcal{D}_e(t)$ in (4.21) is unimprovable. The first term of $\mathcal{D}_e(t)$ (for abbreviation, $L$-component) depends on the product of $L$ and $V(x_0, x^*)$, which can be as big as $LV(x_0, x^*) \le 2(t+1)(M^2+\sigma^2)/\mu$ without affecting the rate of convergence (up to a constant factor 2). Note that in comparison with (1.3), it seems that it is possible to improve the $L$-component of $\mathcal{D}_e(t)$. However, unless both $M$ and $\sigma$ are zero, such an improvement can be hardly achievable, without increasing the second term of $\mathcal{D}_e(t)$, within the generic AC-SA algorithmic framework.

Second, observe that the bounds $\mathcal{D}_e(t)$ and $\mathcal{D}_p(t)$, defined in (4.21) and (4.23) respectively, are not in the same order of magnitude, that is, $\mathcal{D}_e(t) = \mathcal{O}(1/t)$ and $\mathcal{D}_p(t) = \mathcal{O}(1/\sqrt{t}\,)$. We now discuss some consequences of this fact. By (4.21) and Markov's inequality, under Assumption A1, we have

$\text{Prob}\left\{\Psi(x_t^{ag}) - \Psi^* \geq \lambda \mathcal{D}_e(t)\right\} \leq 1/\lambda$ for any $\lambda > 0$ and $t \geq 1$. Hence, for a given confidence level $\Lambda \in (0,1)$, one can easily see that the number of iterations for finding an $(\epsilon, \Lambda)$-solution $\bar{x} \in X$ such that $\text{Prob}\left\{\Psi(\bar{x}) - \Psi^* < \epsilon\right\} \geq 1 - \Lambda$ can be bounded by

$$\mathcal{O}\left\{\frac{1}{\Lambda}\left(\sqrt{\frac{LV(x_0, x^*)}{\nu\epsilon}} + \frac{M^2 + \sigma^2}{\nu\mu\epsilon}\right)\right\}. \tag{4.26}$$

Moreover, if Assumption A2 holds, then by setting the value of $\lambda$ in (4.22) such that $\exp(-\lambda^2/3) + \exp(-\lambda) \leq \Lambda$ and using definitions of $\mathcal{D}_e$ and $\mathcal{D}_p$ in (4.21) and (4.23), we conclude that the number of iterations for finding an $(\epsilon, \Lambda)$-solution of (1.1) can be bounded by

$$\mathcal{O}\left\{\sqrt{\frac{LV(x_0, x^*)}{\nu\epsilon}} + \frac{M^2 + \sigma^2}{\nu\mu\epsilon} + \frac{\sigma^2}{\nu\mu\epsilon}\log\frac{1}{\Lambda} + \left(\frac{\sigma R_X(x^*)}{\epsilon}\log\frac{1}{\Lambda}\right)^2\right\}. \tag{4.27}$$

Note that the above iteration-complexity bound has a significantly worse dependence on $\epsilon$ than the one in (4.26), although it depends only logarithmically on $1/\Lambda$.

# 5 Validation analysis for the AC-SA algorithms

One critical problem associated with SA-type methods is that it is difficult to check the accuracy of the generated solutions. In this subsection, we show that one can compute, with little additional computational effort, certain stochastic lower bounds of the optimal value of (1.1) during the execution of the AC-SA algorithms. These stochastic lower bounds, when grouped with certain stochastic upper bounds on the optimal value, can provide online accuracy certificates for the generated solutions.

We start by discussing the accuracy certificates for the generic AC-SA algorithm in Subsection 3. Let $l_\Psi(z, x)$ be defined in (3.4) and denote

$$\text{lb}_t := \min_{x \in X}\left\{\underline{\Psi}_t(x) := \Gamma_t \sum_{\tau=1}^{t}\left[\frac{\alpha_\tau}{\Gamma_\tau} l_\Psi(x_\tau^{md}, x)\right]\right\}. \tag{5.1}$$

By (3.25), the function $\underline{\Psi}_t(\cdot)$ underestimates $\Psi(\cdot)$ everywhere on $X$. Note however that $\text{lb}_t$ is unobservable since $\underline{\Psi}_t(\cdot)$ is not known exactly. Along with $\text{lb}_t$, let us define

$$\tilde{\text{lb}}_t = \min_{x \in X}\left\{\tilde{\underline{\Psi}}_t(x) := \Gamma_t \sum_{\tau=1}^{t}\frac{\alpha_\tau}{\Gamma_\tau}\tilde{l}_\Psi(x_\tau^{md}, \xi_\tau, x)\right\}, \tag{5.2}$$

where

$$\tilde{l}_\Psi(z, \xi, x) := F(z, \xi) + \langle G(z, \xi), x - z\rangle + \mu V(z, x) + \mathcal{X}(x).$$

In view of the assumption that problem (2.3) is easy to solve, the bound $\tilde{\text{lb}}_t$ is easily computable. Moreover, since $x_t^{md}$ is a function of $\xi_{[t-1]}$, and $\xi_t$ is independent of $\xi_{[t-1]}$, we have that

$$\begin{aligned} \mathbb{E}\big[\tilde{\text{lb}}_t\big] &= \mathbb{E}\left[\mathbb{E}_{\xi_{[t-1]}}\left[\min_{x \in X}\left(\Gamma_t \sum_{\tau=1}^{t}\tilde{l}_\Psi(x_\tau^{md}, \xi_\tau, x)\right)\right]\right] \leq \mathbb{E}\left[\min_{x \in X}\mathbb{E}_{\xi_{[t-1]}}\left[\left(\Gamma_t \sum_{\tau=1}^{t}\tilde{l}_\Psi(x_\tau^{md}, \xi_\tau, x)\right)\right]\right] \\ &= \mathbb{E}\left[\min_{x \in X}\underline{\Psi}_t(x)\right] = \mathbb{E}\big[\text{lb}_t\big] \leq \Psi^*. \end{aligned} \tag{5.3}$$

That is, on average, $\tilde{\mathrm{lb}}_t$ gives a lower bound for the optimal value of (1.1). In order to see how good the lower bound $\tilde{\mathrm{lb}}_t$ is, we estimate the expectations and probabilities of the corresponding errors in Theorem 10. To establish the large-deviation results for $\tilde{\mathrm{lb}}_t$, we also need the following assumption for the $\mathcal{SO}$.

**A3:** For any $x \in X$ and $t \geq 1$, we have $\mathbb{E}\left[\exp\{\|F(x,\xi_t) - f(x)\|_*^2/Q^2\}\right] \leq \exp\{1\}$ for some $Q > 0$.

Note that while Assumption A2 describes certain "light-tail" assumption about the stochastic gradients $G(x,\xi)$, Assumption A3 imposes a similar restriction on the function values $F(x,\xi)$. Such an additional assumption is needed to establish the large deviation properties for the derived stochastic online lower and upper bounds on $\Psi^*$, both of which involve the estimation of function values, i.e., $F(x_t^{md},\xi_t)$ in (5.2) and $F(x_t^{ag},\xi_t)$ in (5.12). On the other hand, we do not need to use the estimation of function values in the AC-SA algorithms discussed in Sections 3 and 4.

**Theorem 10** *Consider the generic AC-SA algorithm for solving $\mathcal{F}_{X,\mathcal{X}}(L,M,\sigma,\mu)$ and suppose that condition (2.4) holds whenever $\mu > 0$. Also assume that $\{\alpha_t\}_{t\geq 1}$ and $\{\gamma_t\}_{t\geq 1}$ are chosen such that relations (3.6) and (3.7) hold. Let $\tilde{\mathrm{lb}}_t$ be defined in (5.2). Then,*

*a) under Assumption A1, we have, for any $t \geq 2$,*

$$\mathbb{E}[\Psi(x_t^{ag}) - \tilde{\mathrm{lb}}_t] \leq \tilde{\mathcal{B}}_e(t) := \Gamma_t\gamma_1 \max_{x\in X} V(x_0,x) + \Gamma_t \sum_{\tau=1}^{t} \frac{2\alpha_\tau^2(M^2+\sigma^2)}{\Gamma_\tau[\nu(\mu+\gamma_\tau) - L\alpha_\tau^2]}; \tag{5.4}$$

*b) if Assumptions A2 and A3 hold, then for any $t \geq 1$ and $\lambda > 0$,*

$$\mathrm{Prob}\left\{\Psi(x_t^{ag}) - \tilde{\mathrm{lb}}_t > \tilde{\mathcal{B}}_e(t) + \lambda\tilde{\mathcal{B}}_p(t)\right\} \leq 2\exp(-\lambda^2/3) + \exp(-\lambda), \tag{5.5}$$

*where*

$$\tilde{\mathcal{B}}_p(t) := Q\Gamma_t \left(\sum_{\tau=1}^{t} \frac{\alpha_\tau^2}{\Gamma_\tau^2}\right)^{\frac{1}{2}} + 2\sigma\Gamma_t R_X(x^*) \left(\sum_{\tau=1}^{t} \frac{\alpha_\tau^2}{\Gamma_\tau^2}\right)^{\frac{1}{2}} + \sigma^2\Gamma_t \sum_{\tau=1}^{t} \frac{2\alpha_\tau^2}{\Gamma_\tau[\nu(\mu+\gamma_\tau) - L\alpha_\tau^2]}, \tag{5.6}$$

*and $R_X(x^*)$ is defined in (3.12);*

*c) If $\{\alpha_t\}_{t\geq 1}$ and $\{\gamma_t\}_{t\geq 1}$ are chosen such that (3.6) and (3.13) (rather than (3.7)) hold, then Parts a) and b) still hold by simply replacing the first term in the definition of $\tilde{\mathcal{B}}_e(t)$ with $\gamma_t \max_{x\in X} \bar{V}(x)$, where $\bar{V}(x)$ is defined in Theorem 1.c).*

*Proof.* Let $\zeta_t := F(x_t^{md},\xi_t) - f(x_t^{md})$, $t \geq 1$, and $\delta_t$ be defined in (3.14). Noting that by the definitions (3.23) and (3.26), relation (5.2), and the fact that $V(x_t, x) \geq 0$ due to (2.2), we have

$$\begin{aligned}
\Psi(x_t^{ag}) - \underline{\tilde{\Psi}}_t(x) &= \Psi(x_t^{ag}) - \Gamma_t \sum_{\tau=1}^{t} \frac{\alpha_\tau}{\Gamma_\tau} \left[ l_\Psi(x_\tau^{md}, x) + \zeta_\tau + \langle \delta_\tau, x - x_\tau^{md} \rangle \right] \\
&\leq \Gamma_t \sum_{\tau=1}^{t} \frac{\gamma_\tau}{\Gamma_\tau} [V(x_{\tau-1}, x) - V(x_\tau, x)] + \Gamma_t \sum_{\tau=1}^{t} \frac{1}{\Gamma_\tau} \left[ \Delta_\tau(x) - \alpha_\tau \left( \zeta_\tau + \langle \delta_\tau, x - x_\tau^{md} \rangle \right) \right] \\
&\leq \Gamma_t \gamma_1 V(x_0, x) + \Gamma_t \sum_{\tau=1}^{t} \frac{1}{\Gamma_\tau} \left[ \Delta_\tau(x) - \alpha_\tau \left( \zeta_\tau + \langle \delta_\tau, x - x_\tau^{md} \rangle \right) \right] \\
&= \Gamma_t \gamma_1 V(x_0, x) + \Gamma_t \sum_{\tau=1}^{t} \frac{1}{\Gamma_\tau} \left[ \alpha_\tau \langle \delta_\tau, x_\tau^{md} - x_{\tau-1}^{+} \rangle + \frac{\alpha_\tau^2 (M + \|\delta_\tau\|_*)^2}{\nu(\mu + \gamma_\tau) - L\alpha_\tau^2} - \alpha_\tau \zeta_\tau \right], \quad (5.7)
\end{aligned}$$

where the last identity follows from (3.16). Note that $x_t^{md}$ and $x_{t-1}^{+}$ are functions of $\xi_{[t-1]} = (\xi_1, ..., \xi_{t-1})$ and that $\xi_t$ is independent of $\xi_{[t-1]}$. Using arguments similar to the ones in the proof of (3.9) and (3.10), we can show (5.4) and (5.5).

We now show part c). If (3.13) holds, in view of Proposition 5 and relation (3.31), the inequality (5.7) holds with the first term $\Gamma_t \gamma_1 V(x_0, x)$ in the right-hand-side replaced by $\gamma_t \bar{V}(x)$. The rest of the proof is exactly the same as Parts a) and b). ■

We now add a few comments about the results obtained in Theorem 10. First, note that relations (5.4) and (5.5) tells us how the gap between $\Psi(x_t^{ag})$ and $\tilde{\text{lb}}_t$ converges to zero. By comparing these two relations with (3.9) and (3.10), we can easily see that both $\Psi(x_t^{ag}) - \tilde{\text{lb}}_t$ and $\Psi(x_t^{ag}) - \Psi^*$ converge to zero in the same order of magnitude.

Second, it is possible to develop validation analysis results for the specialized AC-SA algorithms in Subsection 4. In particular, Proposition 11 below discusses the lower bounds $\tilde{\text{lb}}_t^*$ for the nearly optimal AC-SA algorithm for $\mathcal{F}_{X,\mathcal{X}}(L, M, \sigma, \mu)$. The proof of this result is similar to that of Proposition 9 and hence the details are skipped.

**Proposition 11** *Let $x_t^{ag}$ be computed by the AC-SA algorithm for $\mathcal{F}_{X,\mathcal{X}}(L, M, \sigma, \mu)$ with stepsize policy (4.20). Also let $\tilde{\text{lb}}_t$ be defined as in (5.2). If $\mu > 0$ and condition (2.4) holds, then under Assumption A1, we have, $\forall t \geq 1$,*

$$\mathbb{E}[\Psi(x_t^{ag}) - \tilde{\text{lb}}_t] \leq \tilde{\mathcal{D}}_e(t) := \frac{4L \max_{x \in X} V(x_0, x)}{\nu t(t+1)} + \frac{8(M^2 + \sigma^2)}{\nu\mu(t+1)}. \quad (5.8)$$

*If Assumptions A2 and A3 hold, then, $\forall \lambda > 0, \forall t \geq 1$,*

$$\text{Prob}\left\{ \Psi(x_t^{ag}) - \tilde{\text{lb}}_t > \tilde{\mathcal{D}}_e(t) + \lambda \tilde{\mathcal{D}}_p(t) \right\} \leq 2\exp(-\lambda^2/3) + \exp(-\lambda), \quad (5.9)$$

*where*

$$\tilde{\mathcal{D}}_p(t) := \frac{2Q}{(t+1)^{\frac{1}{2}}} + \frac{4\sigma R_X(x^*)}{(t+1)^{\frac{1}{2}}} + \frac{8\sigma^2}{\nu\mu(t+1)}, \quad (5.10)$$

*$R_X(x^*)$ is defined in (3.12), and $Q$ is from Assumption A3.*

Theorem 10 presents a way to assess the quality of the solutions $x_t^{ag}$, $t \ge 1$, by computing the gap between $\Psi(x_t^{ag})$ and $\tilde{\text{lb}}_t$ (c.f. (5.2)). While $\tilde{\text{lb}}_t$ can be computed easily, the estimation of of $\Psi(x_t^{ag})$ can be time consuming, requiring a large number of samples for $\xi$. In the remaining part of this section, we will briefly discuss how to enhance these lower bounds with efficiently computable upper bounds on the optimal value $\Psi^*$ so that one can assess the quality of the generated solutions in an online manner. More specifically, for any $t \ge 1$, let us denote

$$\beta_t := \sum_{\tau=\lceil t/2 \rceil}^{t} \tau,$$

$$\text{ub}_t := \beta_t^{-1} \sum_{\tau=\lceil t/2 \rceil}^{t} \tau \Psi(x_\tau^{ag}) \quad \text{and} \quad \bar{x}_t^{ag} := \beta_t^{-1} \sum_{\tau=\lceil t/2 \rceil}^{t} \tau x_\tau^{ag}. \tag{5.11}$$

Clearly, we have $\text{ub}_t \ge \Psi(\bar{x}_t^{ag}) \ge \Psi^*$ due to the convexity of $\Psi$. Also let us define

$$\bar{\text{ub}}_t = \beta_t^{-1} \sum_{\tau=\lceil t/2 \rceil}^{t} \tau \{F(x_\tau^{ag}, \xi_\tau) + \mathcal{X}(x_\tau^{ag})\}, \quad \forall\, t \ge 1. \tag{5.12}$$

Since $\mathbb{E}_{\xi_\tau}[F(x_\tau^{ag}, \xi_\tau)] = f(x_\tau^{ag})$, we have $\mathbb{E}[\bar{\text{ub}}_t] = \text{ub}_t \ge \Psi^*$. That is, $\bar{\text{ub}}_t$, $t \ge 1$, on average, provide online upper bounds on $\Psi^*$. Accordingly, we define the new online lower bounds as

$$\bar{\text{lb}}_t = \beta_t^{-1} \sum_{\tau=\lceil t/2 \rceil}^{t} \tau \, \tilde{\text{lb}}_\tau, \quad \forall\, t \ge 1, \tag{5.13}$$

where $\tilde{\text{lb}}_\tau$ is defined in (5.2).

To bound the gap between these lower and upper bounds, let $\tilde{\mathcal{B}}_e(\tau)$ be defined in (5.4) and suppose that $\tilde{\mathcal{B}}_e(t) = \mathcal{O}(t^{-q})$ for some $q \in [1/2, 1]$. In view of Theorem 10.a), (5.11) and (5.13), we have

$$\begin{aligned}
\mathbb{E}[\bar{\text{ub}}_t - \bar{\text{lb}}_t] &= \beta_t^{-1} \sum_{\tau=\lceil t/2 \rceil}^{t} \tau[\Psi(x_\tau^{ag}) - \tilde{\text{lb}}_\tau] \le \beta_t^{-1} \sum_{\tau=\lceil t/2 \rceil}^{t} [\tau \tilde{\mathcal{B}}_e(\tau)] \\
&= \mathcal{O}\left( \beta_t^{-1} \sum_{\tau=\lceil t/2 \rceil}^{t} \tau^{1-q} \right) = \mathcal{O}(t^{-q}), \quad t \ge 3,
\end{aligned}$$

where the last identity follows from the facts that $\sum_{\tau=\lceil t/2 \rceil}^{t} \tau^{1-q} = \mathcal{O}(t^{2-q})$ and that

$$\beta_t \ge \frac{1}{2}\left[ t(t+1) - \left(\frac{t}{2}+1\right)\left(\frac{t}{2}+2\right)\right] \ge \frac{1}{8}\left(3t^2 - 2t - 8\right).$$

Therefore, the gap between the online upper bound $\bar{\text{ub}}_t$ and lower bound $\bar{\text{lb}}_t$ converges to 0 in the same order of magnitude as the one between $\Psi(x_t^{ag})$ and $\tilde{\text{lb}}_t$. It should be mentioned that the stochastic upper bound $\bar{\text{ub}}_t$, on average, overestimates the value of $\Psi(\bar{x}_t^{ag})$ (c.f. (5.11)), indicating that one can also use $\bar{x}_t^{ag}$, $t \ge 1$, as the output of the AC-SA algorithm.

## 6 Concluding remarks

In this paper, we presented a generic AC-SA algorithmic framework for solving strongly convex SCO problems. We showed that it can yield optimal algorithms for solving SCO problems without possessing strong convexity and a nearly optimal algorithm for solving strongly convex SCO problems. It is worth noting that the latter algorithm is also optimal for nonsmooth strongly convex problems. Moreover, large-deviation results associated with the convergence rates for these AC-SA algorithms are studied and certain stochastic lower and upper bounds of the optimal value are presented to provide accuracy certificates for the generated solutions.

However, there remain a few interesting questions that have not been answered, e.g., "Whether the lower complexity bound stated in (1.3) for minimizing strongly convex SCO problems can be achievable?" and "Whether the large-deviation results associated with the convergence results for the nearly optimal AC-SA algorithm in Subsection 4.2 can be improvable?" We will study these problems in [12], a companion work of the current paper.

**Acknowledgement:** The authors are very grateful to the associate editor and two anonymous referees for their very useful suggestions for improving the quality and exposition of the paper.

## 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] A. Ben-Tal and A.Nemirovski. Non-euclidean restricted memory level method for large-scale convex optimization. *Mathematical Programming*, 102:407–456, 2005.

[4] A. Ben-Tal, T. Margalit, and A.Nemirovski. The ordered subsets mirror descent optimization method with applications to tomography. *SIAM Journal on Optimization*, 12:79–108, 2001.

[5] K.P. Bennett and O.L. Mangasarian. Robust linear programming discrimination of two linearly inseparable sets. *Optimization Methods and Software*, 1:23–24, 1992.

[6] A. Benveniste, M. Métivier, and P. Priouret. *Algorithmes adaptatifs et approximations stochastiques*. Masson, 1987. English translation: *Adaptive Algorithms and Stochastic Approximations*, Springer Verlag (1993).

[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] K.L. Chung. On a stochastic approximation method. *Annals of Mathematical Statistics*, pages 463–483, 1954.

[9] C. D. Dang and G. Lan. On the convergence properties of non-euclidean extragradient methods for variational inequalities with generalized monotone operators. Manuscript, Department of

Industrial and Systems Engineering, University of Florida, Gainesville, FL 32611, USA, April 2012. Available on http://www.optimization-online.org/.

[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] S. Ghadimi and G. Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, II: shrinking procedures and optimal algorithms. Manuscript, Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL 32611, USA, 2010. Submitted to *SIAM Journal on Optimization*.

[13] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer Series in Statistics. Springer, 2009.

[14] C. Hu, J. T. Kwok, and W. Pan. Accelerated gradient methods for stochastic optimization and online learning. In *NIPS*, 2009.

[15] 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.

[16] A. Juditsky and A. Nemirovski. *First-Order Methods for Nonsmooth Convex Large-Scale Optimization, I: General Purpose Methods* . in Optimization for Machine Learning, Eds: S. Sra, S. Nowozin and S.J. Wright. MIT press, 2011.

[17] A. Juditsky and Y. Nesterov. Primal-dual subgradient methods for minimizing uniformly convex functions. Manuscript.

[18] A. Juditsky, P. Rigollet, and A. B. Tsybakov. Learning by mirror averaging. *Annals of Statistics*, 36:2183–2206, 2008.

[19] K.C. Kiwiel. Proximal minimization methods with generalized bregman functions. *SIAM Journal on Controal and Optimization*, 35:1142–1168, 1997.

[20] 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.

[21] H. J. Kushner and G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications*, volume 35 of *Applications of Mathematics*. Springer-Verlag, New York, 2003.

[22] G. Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 2010. Forthcoming, online first, DOI: 10.1007/s10107-010-0434-y.

[23] G. Lan, A. Nemirovski, and A. Shapiro. Validation analysis of mirror descent stochastic approximation method. *Mathematical Programming*, 2011. Forthcoming, online first.

[24] A.S. Lewis and S.J. Wright. A proximal method for composite minimization. Manuscript, Cornell University, Ithaca, NY, 2009.

[25] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19:1574–1609, 2009.

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

[27] 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.

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

[29] Y. E. Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120:221–259, 2006.

[30] 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.

[31] G.C. Pflug. Optimization of stochastic models. In *The Interface Between Simulation and Optimization*. Kluwer, Boston, 1996.

[32] B.T. Polyak. New stochastic approximation type procedures. *Automat. i Telemekh.*, 7:98–107, 1990.

[33] B.T. Polyak and A.B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM J. Control and Optimization*, 30:838–855, 1992.

[34] H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 1951.

[35] 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.

[36] J. Sacks. Asymptotic distribution of stochastic approximation. *Annals of Mathematical Statistics*, 29:373–409, 1958.

[37] S. Shalev-Shwartz, Y. Singer, and N. Srebro. Pegasos: Primal estimated sub-gradient solver for svm. In *In ICML*, pages 807–814, 2007.

[38] A. Shapiro. Monte carlo sampling methods. In A. Ruszczyński and A. Shapiro, editors, *Stochastic Programming*. North-Holland Publishing Company, Amsterdam, 2003.

[39] S. Smale and Y. Yao. Online learning algorithms. *Found. Comp. Math*, 6:145–170, 2005.

[40] J.C. Spall. *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*. John Wiley, Hoboken, NJ, 2003.

[41] M. Teboulle. Convergence of proximal-like algorithms. *SIAM Journal on Optimization*, 7:1069–1083, 1997.

[42] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 58:267–288, 1996.

[43] P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. Manuscript, University of Washington, Seattle, May 2008.

[44] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell. Distance metric learning, with application to clustering with side-information. In *Advances in Neural Information Processing Systems 15*, pages 505–512. MIT Press, 2002.

[45] D.B. Yudin and A.S. Nemirovskii. Computational complexity of strictly convex programming. *Eknomika i Matematicheskie Metody*, 3:550–569, 1977.

[46] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320, 2005.