Dynamic Stochastic Approximation for Multi-stage Stochastic Optimization

Guanghui Lan [*] Zhiqiang Zhou [†]

July 11, 2017

Abstract

In this paper, we consider multi-stage stochastic optimization problems with convex objectives and conic constraints at each stage. We present a new stochastic first-order method, namely the dynamic stochastic approximation (DSA) algorithm, for solving these types of stochastic optimization problems. We show that DSA can achieve an optimal $\mathcal{O}(1/\epsilon^4)$ rate of convergence in terms of the total number of required scenarios when applied to a three-stage stochastic optimization problem. We further show that this rate of convergence can be improved to $\mathcal{O}(1/\epsilon^2)$ when the objective function is strongly convex. We also discuss variants of DSA for solving more general multi-stage stochastic optimization problems with the number of stages $T > 3$. The developed DSA algorithms only need to go through the scenario tree once in order to compute an ϵ-solution of the multi-stage stochastic optimization problem. To the best of our knowledge, this is the first time that stochastic approximation type methods are generalized for multi-stage stochastic optimization with $T \geq 3$.

1 Introduction

Multi-stage stochastic optimization aims at optimal decision-making over multiple periods of time, where the decision in the current period has to take into account what will happen in the future. Optimal decisions over a certain time horizon is of paramount importance to different applications areas including finance, logistics, robotics and clinic trials etc. In this paper, we are interested in solving a class of multi-stage stochastic optimization problems given by

$$\begin{array}{lll} \min\{\ h^1(x^1,c^1) + & \mathbb{E}_{\xi^2}\{\min\{\ h^2(x^2,c^2) \quad + \ldots + & \mathbb{E}_{\xi^T}[\min\{\ h^T(x^T,c^T)\}]\}\}\} \\ \text{s.t. } A^1x^1 - b^1 \in K^1 & \text{s.t. } A^2x^2 - b^2 - B^2x^1 \in K^2, & \text{s.t. } A^Tx^T - b^T - B^Tx^{T-1} \in K^T, \\ \quad x^1 \in X^1, & \quad x^2 \in X^2, & \quad x^T \in X^T. \end{array} \tag{1.1}$$

Here T denotes the number of stages, $h_t(\cdot, c^t)$ are relatively simple convex functions, K^t are closed convex cones, $\xi^t := (A^t, b^t, B^t, c^t)$, $t = 2, \ldots, T$, are the random vectors at stage t, and $\mathbb{E}_{\xi^t}$ denote the conditional expectation with respect to ξ^t given $(\xi^2, \ldots, \xi^{t-1})$. By defining value functions, we can write problem (1.1) equivalently as

$$\begin{array}{ll} \min & \left\{h^1(x^1,c^1) + v^2(x^1)\right\} \\ \text{s.t.} & A^1x^1 - b^1 \in K^1, \\ & x^1 \in X^1, \end{array} \tag{1.2}$$

[*]H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, 30332. (email: george.lan@isye.gatech.edu).

[†]H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, 30332. (email: zzhoubrian@gatech.edu).

where the value factions v^t are recursively defined by

$$\begin{array}{lcl} v^t(x^{t-1}) & := & \mathbb{E}_{\xi^t}[V^t(x^{t-1},\xi^t)], \quad t=2,\ldots,T-1, \\ V^t(x^{t-1},\xi^t) & := & \min \ \left\{h^t(x^t,c^t)+v^{t+1}(x^t)\right\} \\ & & \quad \text{s.t.} \ \ A^tx^t-b^t-B^tx^{t-1}\in K^t, \\ & & \qquad\quad x^t\in X^t, \end{array} \tag{1.3}$$

and

$$\begin{array}{lcl} v^T(x^{T-1}) & := & \mathbb{E}_{\xi^T}[V^T(x^{T-1},\xi^T)], \\ V^T(x^{T-1},\xi^T) & := & \min \ h^T(x^T,c^T) \\ & & \quad \text{s.t.} \ \ A^Tx^T-b^T-B^Tx^{T-1}\in K^T, \\ & & \qquad\quad x^T\in X^T. \end{array} \tag{1.4}$$

In particular, if h^t are affine, $K^t=\{0\}$ and X^t are polyhedral, then problem (1.1) reduces to the well-known multi-stage stochastic linear programming problem (see, e.g., [2]). The incorporation of the nonlinear (but convex) objective function $h^t(x^t,c^t)$ and conic constraints $A^tx^t-b^t-B^tx^{t-1}\in K^t$ allows us to model a much wider class of problems. Moreover, if $T=2$, then problem (1.1) is often referred to as a two-stage (or static) stochastic programming problem.

In spite of its wide applicability, multi-stage stochastic optimization remains highly challenging to solve. Many existing methods for multi-stage stochastic optimization are based on sample average approximation (see Nemirovski and Shapiro [30] and Shapiro [31]). In this approach, one first generate a deterministic counterpart of (1.1) by replacing the expectations with (conditional) sample averages. In particular, if the number of stages $T=3$, the total number of samples (a.k.a. scenarios) cannot be smaller than $\mathcal{O}(1/\epsilon^4)$ in general. Once after a deterministic approximation of (1.1) is generated, one can then develop decomposition methods to solve it to certain accurary. The most popular decomposition methods consists of stage-based and scenario-based decomposition method. One widely-used stage-based method is the stochastic dual dynamic programming (SDDP) algorithm, which is essentially an approximate cutting plane method, first presented by Pereira and Pinto [24] and later studied by Shapiro [32], Donohue and Birge [5], and Hindsberger [12] etc. The progressive hedging algorithm by Rockafellar and Wets [28] is a well-known scenario-based decomposition method, which basically applies an augmented Lagrangian method to penalize the violation of the non-anticipativity constraints. All these methods assume that the scenario tree has been generated and will go through the scenario tree many times. Usually there are no performance guarantees provided regarding their rate of convergence, i.e., the number of times one needs to go through the scenario tree. In SDDP, one also needs to assume that random vectors are stage-wise independent.

Recently, a different approach called stochastic approximation (SA) has attracted much attention for solving static stochastic optimization problems given in the form of

$$\min_{x\in X}\left\{f(x):=\mathbb{E}_\xi[F(x,\xi)]\right\}, \tag{1.5}$$

where X is a closed convex set, ξ denotes the random vecctor and $F(\cdot,\xi)$ is a closed convex function. Observe that when $T=2$, problem (1.1) can be cast in the form of (1.5) and hence one can apply the aforementioned SA methods to solve these two-stage stochastic optimization problems (see [19, 14]). The basic SA algorithm, initially proposed by Robbins and Monro [27], mimics the simple projected gradient descent method by replacing exact gradient with its unbiased estimator. Important improvements for the SA methods have been made by Nemirovski and Yudin [20] and later by Polayk and Juditsky ([25, 26]). During the past few years, significant progress has been made in SA methods [19, 13, 6, 7, 8, 17, 10]. In particular, Nemirovski et. al. [19] presented a properly modified SA approach, namely, mirror descent SA for solving general nonsmooth convex SP problems. Lan [13] introduced an accelerated SA method, based on Nesterov's accelerated gradient method [21], for solving smooth, nonsmooth and stochastic optimization in a uniform manner. Novel nonconvex SA methods and their accelerated versions have also been studied in [8, 10, 9]. All these SA algorithms only need to access one single ξ_k at each iteration, and hence do not require much memory. It has been shown in [19, 14] that SA methods can significantly outperform the SAA approach for solving static (or two-stage) stochastic programming problems. However, it remains unclear whether these SA methods can be generalized for multi-stage stochastic optimization problems with $T \geq 3$.

In this paper, we intend to shed some light on this problem by developing a dynamic stochastic approximation (DSA) method for multi-stage stochastic optimization. The basic idea of the DSA method is to design an inexact primal-dual SA method for solving the t-th stage optimization problem in order to compute an approximate stochastic subgradient for its associated value functions v^t. In the pursuit of this idea, we manage to resolve the following difficulties. First, the first-order information for the value function v^{t+1} used to solve the t-stage subproblem is not only stochastic, but also biased. We need to control the bias associated with such first-order information. Second, in order to establish the convergence of stochastic optimization subroutines for solving the t-stage problem, we need to guarantee that the variance of approximate stochastic subgradients and hence the dual multipliers associated with the $t+1$-stage problem are bounded. Third, we need to make sure that the approximation errors do not accumulate quickly as the number of stages T increases. By properly addressing these issues, we were able to show that the DSA method can achieve an optimal $\mathcal{O}(1/\epsilon^4)$ rate of convergence in terms of the number of random samples when applied to a three-stage stochastic optimization problem. We further show that this rate of convergence can be improved to $\mathcal{O}(1/\epsilon^2)$ when the objective function is strongly convex. Moreover, we discuss variants of the DSA method which exhibit optimal rate of convergence for solving more general multi-stage stochastic optimization problems with $T > 3$. The developed DSA algorithms only need to go through the scenario tree once in order to compute an ϵ-solution of the multi-stage stochastic optimization problem. As a result, the required memory for DSA increases only linearly with respect to T. To the best of our knowledge, this is the first time that stochastic approximation type methods are generalized to and their complexities are established for multi-stage stochastic optimization.

This paper is organized as follows. In Section 2, we introduce the basic scheme of the DSA algorithm and establish its main convergence properties for solving three-stage stochastic optimization problems. In Section 3, we apply the DSA to the problem (2.1) under the strongly convex assumption on the objective function at each stage. and we develop variants of the DSA method for solving more general form of (1.1) with $T > 3$ in Section 4. Finally, some concluding remarks are made in Section 5.

1.1 Notation and terminology

For a closed convex set X, a function $\omega_X : X \mapsto R$ is called a distance generating function with parameter α, if ω_X is continuously differentiable and strongly convex with parameter α with respect to $\|\cdot\|$. Therefore, we have

$$\langle x_1 - x, \nabla \omega_X(x_1) - \nabla \omega_X(x) \rangle \geq \|x_1 - x\|^2, \forall x_1, x \in X.$$

The prox-function associated with ω_X is given by

$$P_X(x, x_1) = \omega_X(x_1) - \omega_X(x) - \langle \nabla \omega_X(x), x_1 - x \rangle, \forall x_1, x \in X.$$

It can be easily seen that

$$P_X(x, x_1) \geq \tfrac{\alpha_X}{2} \|x_1 - x\|^2, \ \forall x_1, x \in X. \tag{1.6}$$

If X is bounded, we define the diameter of the set X as

$$\Omega_X^2 := \max_{x_1, x \in X} P_X(x, x_1). \tag{1.7}$$

For a given closed convex cone K_*, we choose the distance generating function $\omega_{K_*}(y) = \|y\|_2^2/2$. For simplicity, we often skip the subscript of $\|\cdot\|_2$ whenever we apply it to an unbounded set (such as a cone).

For a given closed convex set $X \subseteq \mathbb{R}^n$ and a closed convex function $V : X \to \mathbb{R}$, $g(x)$ is called an ϵ-subgradient of V at $x \in X$ if

$$V(x_1) \geq V(x) + \langle g(x), x_1 - x\rangle - \epsilon \ \ \forall x_1 \in X. \tag{1.8}$$

The collection of all such ϵ-subgradients of V at x is called the ϵ-subdeifferential of V at x, denoted by $\partial_\epsilon V(x)$.

Assume that V is Lipschitz continuous in an ϵ-neighborhood of X, i.e.,

$$|V(x_1) - V(x)| \leq M_0\|x_1 - x\|, \ \forall x_1, x \in X_\epsilon := \{p \in \mathbb{R}^n : p = r + x, x \in X, \|r\| \leq \epsilon\}. \tag{1.9}$$

We can show that

$$\|g(x)\|_* \leq M_0 + 1 \ \ \forall x \in X. \tag{1.10}$$

Indeed, if $\|\cdot\| = \|\cdot\|_2$, the result follows immediately by setting $d = \epsilon g(x)/\|g(x)\|_2$ and $x_1 = x+d$ in (1.8). Otherwise, we need to choose d properly s.t. $\|d\| = \epsilon$ and $\langle g(x), d\rangle = \epsilon\|g(x)\|_*$. It should be noted, however, that if V is Lipschitz continuous over X (rather than X_ϵ), then one cannot guarantee the boundedness of an ϵ-subgradient of V.

2 Three-stage problems with generally convex objectives

Our goal in this section is to introduce the basic scheme of the DSA algorithm and discuss its convergence properties. For the sake of simplicity, we will focus on three-stage stochastic optimization problems with simple convex objective functions in this section. Extensions to strongly convex cases and more general form of multi-stage stochastic optimization problems will be studied in later sections.

2.1 Value functions and stochastic ϵ-subgradients

Consider the following three-stage stochastic programming problem:

$$\begin{array}{lll}\min\{\ h^1(x^1, c^1) + & \mathbb{E}_{\xi^2}\{\min\{\ h^2(x^2, c^2) & +\mathbb{E}_{\xi^3}[\min\{\ h^3(x^3, c^3)\}]\}\}\} \\ \quad \text{s.t. } A^1x^1 - b^1 \in K^1 & \quad \text{s.t. } A^2x^2 - b^2 - B^2x^1 \in K^2, & \quad \text{s.t. } A^3x^3 - b^3 - B^3x^2 \in K^3, \\ \qquad x^1 \in X^1, & \qquad x^2 \in X^2, & \qquad x^3 \in X^3,\end{array} \tag{2.1}$$

where $X^t \subseteq \mathbb{R}^{n_t}$, $t = 1, 2, 3$, are compact convex sets for some $n_t > 0$, $h^t : X^t \to \mathbb{R}$ are relatively simple convex functions, A^t denote the linear mappings from $\mathbb{R}^{n_t}$ to $\mathbb{R}^{m_t}$ for some $m_t > 0$, and $K^t \subseteq \mathbb{R}^{m_t}$ are closed convex cones. Note that the first expectation in (2.1) is taken w.r.t. the random vector $\xi^2 \equiv (A^2, b^2, B^2, c^2)$ and the second one is the conditional expectation with respect to $\xi^3 \equiv (A^3, b^3, B^2, c^3)$ given ξ^2. As an example, if $h^t(x^t, c^t) = \langle c^t, x^t\rangle$, $K^t = \{0\}$ and X^t are polyhedronal, then (2.1) reduces to a well-known three-stage stochastic linear programming problem.

We can write problem (2.1) in a more compact form by using value functions as discussed in Section 1. More specifically, let $V^3(x^2, \xi^3)$ be the stochastic value function at the third stage and $v^3(x^2)$ be the corresponding (expected) value function:

$$\begin{array}{lll}V^3(x^2, \xi^3) & := & \min\ h^3(x^3, c^3) \\ & & \quad \text{s.t. } \ A^3x^3 - b^3 - B^3x^2 \in K^3, \\ & & \qquad\quad x^3 \in X^3. \\ v^3(x^2) & := & \mathbb{E}_{\xi^3}[V^3(x^2, \xi^3)].\end{array} \tag{2.2}$$

We can then define the stochastic value function $V^2(x^1, \xi^2)$ and its corresponding (expected) value function as

$$\begin{array}{lll} V^2(x^1,\xi^2) & := & \min \ \left\{h^2(x^2,c^2) + v^3(x^2)\right\} \\ & & \quad \text{s.t.} \ \ A^2x^2 - b^2 - B^2x^1 \in K^2, \\ & & \qquad\quad x^2 \in X^2. \\ v^2(x^1) & := & \mathbb{E}_{\xi^2}[V^2(x^1,\xi^2)]. \end{array} \tag{2.3}$$

Problem (2.1) can then be formulated equivalently as

$$\begin{array}{ll} \min & \left\{h^1(x^1,c^1) + v^2(x^1)\right\} \\ \text{s.t.} & A^1x^1 - b^1 \in K^1, \\ & x^1 \in X^1. \end{array} \tag{2.4}$$

Throughout this paper, we assume that the expected value functions $v^2(x^1)$ and $v^3(x^2)$ are well-defined and finite-valued for any $x^1 \in X^1$ and $x^2 \in X^2$, respectively.

In order to solve problem (2.4), we need to understand how to compute first-order information about the value functions v^2 and v^3. Since both v^2 and v^3 are given in the form of expectation, the exact first-order information is hard to compute. We resort to the computation of a stochastic ϵ-subgradient of these value functions defined as follows.

Definition 1 *$G(u,\xi)$ is called a stochastic ϵ-subgradient of the value function $v(u) = \mathbb{E}_\xi[V(u,\xi)]$ if $G(u,\xi)$ is an unbiased estimator of an ϵ-subgradient of v, i.e.,*

$$\mathbb{E}_\xi[G(u,\xi)] = g(u) \quad \textit{and} \quad g(u) \in \partial_\epsilon v(u). \tag{2.5}$$

In order to compute a stochastic ϵ-subgradient of v^2 (resp., v^3), we need to compute an approximate subgradient of the corresponding stochastic value function $V^2(x^1,\xi^2)$ (resp., $V^3(x^2,\xi^3)$). To this end, we further assume that strong Lagrange duality holds for the optimization problems defined in (2.3) (resp.,(2.2)) almost surely. In other words, these problems can be formulated as saddle point problems:

$$V^2(x^1,\xi^2) = \max_{y^2\in K^2_*}\min_{x^2\in X^2} \langle b^2 + B^2x^1 - A^2x^2, y^2\rangle + h^2(x^2,c^2) + v^3(x^2), \tag{2.6}$$

$$V^3(x^2,\xi^3) = \max_{y^3\in K^3_*}\min_{x^3\in X^3} \langle b^3 + B^3x^2 - A^3x^3, y^3\rangle + h^3(x^3,c^3), \tag{2.7}$$

where K^2_* and K^3_* are corresponding dual cones to K^2 and K^3, respectively.

Observe that both (2.6) and (2.7) can be viewed as special cases of the following more generic saddle point problem

$$V(u,\xi) \equiv V(u,(A,b,B,c)) := \max_{y\in K_*}\min_{x\in X}\langle b + Bu - Ax, y\rangle + h(x,c) + \tilde{v}(x), \tag{2.8}$$

where $A : \mathbb{R}^n \to m$ and $B : \mathbb{R}^{n_0} \to m$ denote the linear mappings. For example, (2.7) is a special case of (2.8) with $u = x^2$, $y = y^3$, $K_* = k^3_*$, $b = b^3$, $B = B^3$, $A = A^3$, $h = h^3$ and $\tilde{v} = 0$. Let

$$(x_*, y_*) \in Z \equiv X \times K_*$$

be a pair of optimal solutions of the saddle point problem (2.6), i.e.,

$$V(u,\xi) = \langle y_*, b + Bu - Ax_*\rangle + h(x_*,c) + \tilde{v}(x_*) = h(x_*,c) + \tilde{v}(x_*), \tag{2.9}$$

where the second identity follows from the complementary slackness of Lagrange duality. It is worth noting that the first stage problem can also be viewed as a special case of (2.8), since (2.4) is equivalent to

$$\max_{y\in K^1_*}\min_{x^1\in X^1} \ \left\{\langle b^1 - A^1x^1, y^1\rangle + h^1(x^1,c^1) + v^2(x^1)\right\}. \tag{2.10}$$

Below we provide a different characterization of an ϵ-subgradient of V than the one in (1.8).

Lemma 1 *Let $\bar{z} := (\bar{x}, \bar{y}) \in Z$ and $u \in \mathbb{R}^{n_0}$ be given. If*

$$Q(\bar{z}; x, y_*) := \langle y_*, b + Bu - A\bar{x}\rangle + h(\bar{x}, c) + \tilde{v}(\bar{x}) \\ - \langle \bar{y}, b + Bu - Ax\rangle - h(x, c) - \tilde{v}(x) \le \epsilon, \ \forall x \in X, \tag{2.11}$$

then $B^T\bar{y}$ is an ϵ-subgradient of $V(u, \xi)$ at u.

Proof. For simplicity, let us denote $V(u) \equiv V(u, \xi)$. For any $u_1 \in \text{dom}V$, we denote (x_1^*, y_1^*) as a pair of primal-dual solution of (2.8) (with $u = u_1$). Hence,

$$V(u_1) := \langle y_1^*, b + Bu_1 - Ax_1^*\rangle + h(x_1^*, c) + \tilde{v}(x_1^*). \tag{2.12}$$

It follows from the definition of V in (2.8) and (2.11) that

$$\begin{aligned} V(u) &= \langle y_*, b + Bu - Ax_*\rangle + h(x_*, c) + \tilde{v}(x_*) \\ &\le \langle y_*, b + Bu - A\bar{x}\rangle + h(\bar{x}, c) + \tilde{v}(\bar{x}) \\ &\le \langle \bar{y}, b + Bu - Ax_1^*\rangle + h(x_1^*, c) + \tilde{v}(x_1^*) + \epsilon \end{aligned} \tag{2.13}$$

Observe that by

$$\begin{aligned} \langle \bar{y}, b + Bu - Ax_1^*\rangle &= \langle \bar{y}, B(u - u_1)\rangle + \langle \bar{y}, b + Bu_1 - Ax_1^*\rangle \\ &\le \langle \bar{y}, B(u - u_1)\rangle + \langle y_1^*, b + Bu_1 - Ax_1^*\rangle, \end{aligned}$$

where the last inequality follows from the assumption that (x_1^*, y_1^*) is a pair of optimal solution of (2.8) with $u = u_1$. Combining these two observations and using (2.12), we have

$$V(u) \le \langle B^T\bar{y}, u - u_1\rangle + V(u_1) + \epsilon,$$

which, in view of (1.8), implies that $B^T\bar{y}$ is an ϵ-subgradient of $V(u)$. ■

In view of Lemma 1, in order to compute a stochastic subgradient of $v(u) = \mathbb{E}[V(u, \xi)]$ at a given point u, we first generate a random realization ξ and then try to find a pair of solutions $(\bar{x}, \bar{y})$ satisfying (2.11). We can then use $B^T\bar{y}$ as a stochastic ϵ-subgradient of v at u. However, when applied to the value function v^2 of the second stage, the difficulty exists in that the function $\tilde{v}$, being the value function v^3 of the third stage, is also given in the form of expectation. We will discuss how to address these issues in more details in the next subsection.

2.2 The DSA algorithm

Our goal in this subsection is to describe the basic scheme of our dynamic stochastic approximation algorithm applied to problem (2.4).

Our algorithm relies on the following three key primal-dual steps, referred to as stochastic primal-dual transformation (SPDT), applied to the generic saddle point problem in (2.8) at every stage.

$(p_+, d_+, \tilde{d}) = \text{SPDT}(p, d, d\_, \tilde{v}', u, \xi, h, X, K_*, \theta, \tau, \eta)$:

$$\tilde{d} = \theta(d - d\_) + d. \tag{2.14}$$

$$p_+ = \text{argmin}_{x \in X} \langle b + Bu - Ax, \tilde{d}\rangle + h(x, c) + \langle \tilde{v}', x\rangle + \tau P_X(p, x). \tag{2.15}$$

$$d_+ = \text{argmin}_{y \in K_*} \langle -b - Bu + Ap_+, y\rangle + \tfrac{\eta}{2}\|y - d\|^2. \tag{2.16}$$

In the above primal-dual tranformation, the input $(p, d, d\_)$ denotes the current primal solution, dual solution, and the previous dual solution, respectively. Moreover, the input $\tilde{v}'$ denotes a stochastic ϵ-subgradient for $\tilde{v}$ at the current search point p. The parameters (u, ξ, h, X, K_*) describes the problem in (2.8) and (θ, τ, η) are certain algorithmic parameters to be specified.

Given these input parameters, the relation in (2.14) defines a dual extrapolation (or prediction) step to estimate the dual variable $\tilde{d}$ for the next iterate. Based on this estimate, (2.15) performs a primal prox-mapping to compute p_+, and then (2.16) updates in the dual space to compute d_+ by using the updated p_+. We assume that the above SPDT operator can be performed very fast or even has explicit expressions.

In order to solve problem (2.4), we will combine the above primal-dual transformation applied to all the three stages together with the scenario generation for the random variables ξ^2 and ξ^3 in the second and third stage, as well as certain averaging steps in both the primal and dual spaces. We are now ready to describe the basic scheme of the DSA algorithm.

Algorithm 1 The basic DSA algorithm for three-stage problems

Input: initial points (z_0^1, z_0^2, z_0^3).
$\xi^1 = (A^1, 0, b^1, c^1)$.
for $i = 1, 2, \ldots, N_1$ **do**
  Generate an random realization of $\xi_i^2 = (A_i^2, B_i^2, b_i^2, c_i^2)$.
  for $j = 1, 2, \ldots, N_2$ **do**
    Generate an random realization of $\xi_j^3 = (A_j^3, B_j^3, b_j^3, c_j^3)$ (conditional on ξ_i^2).
    for $k = 1, 2, \ldots, N_3$ **do**
      $(x_k^3, y_k^3, \tilde{y}_k^3) = \text{SPDT}(x_{k-1}^3, y_{k-1}^3, y_{k-2}^3, 0, x_{j-1}^2, \xi_j^3, h^3, X^3, K_*^3, \theta_k^3, \tau_k^3, \eta_k^3)$.
    end for
    $(\bar{x}_j^3, \bar{y}_j^3) = \sum_{k=1}^{N_3} w_k^3 (x_k^3, y_k^3) / \sum_{k=1}^{N_3} w_k^3$.
    $(x_j^2, y_j^2, \tilde{y}_j^2) = \text{SPDT}(x_{j-1}^2, y_{j-1}^2, y_{j-2}^2, (B_j^3)^T \bar{y}_j^3, x_{i-1}^1, \xi_i^2, h^2, X^2, K_*^2, \theta_j^2, \tau_j^2, \eta_j^2)$.
  end for
  $(\bar{x}_i^2, \bar{y}_i^2) = \sum_{j=1}^{N_2} w_j^2 (x_j^2, y_j^2) / \sum_{j=1}^{N_2} w_j^2$.
  $(x_i^1, y_i^1, \tilde{y}_i^1) = \text{SPDT}(x_{i-1}^1, y_{i-1}^1, y_{i-2}^1, (B_i^2)^T \bar{y}_i^2, 0, \xi^1, h^1, X^1, K_*^1, \theta_i^1, \tau_i^1, \eta_i^1)$.
end for
Output: $(\bar{x}^1, \bar{y}^1) = \sum_{i=1}^{N_1} w_i^1 (x_i^1, y_i^1) / \sum_{i=1}^{N_1} w_i^1$.

This algorithm consists of three loops. The innermost (third) loop runs N_3 steps of SPDT in order to compute an approximate stochastic subgradient ($(B_j^3)^T \bar{y}_j^3$) of the value function v^3 of the third stage. The second loop consists of N_2 SPDTs applied to the saddle point formulation of the second-stage problem, which requires the output from the third loop. The outer loop applies N_1 SPDTs to the saddle point formulation of the first-stage optimization problem in (2.4), using the stochastic approximate subgradients ($(B_i^2)^T \bar{y}_i^2$) for v^2 computed by the second loop. In this algorithm, we need to generate N_1 and $N_1 \times N_2$ realizations for the random vectors ξ^2 and ξ^3, respectively. Observe that the DSA algorithm described above is conceptual only since we have not specified any algorithmic parameters yet. We will come back to this issue after establishing some general convergence properties about this method in the next two subsections.

2.3 Basic tools: inexact primal-dual stochastic approximation

In this subsection, we provide some basic tools for the convergence analysis of the DSA method.

Our analysis will be centered around an inexact primal-dual stochastic approximation (I-PDSA) method, which consists of iterative applications of the SPDTs defined in (2.14), (2.15) and (2.16) to solve the generic saddle point problem in (2.8).

Algorithm 2 formally describes the I-PDSA method for solving (2.8), which evolves from the primal-dual method by Chambolle and Pork in [3]. The primal-dual method in [3] is an efficient and simple method for solving saddle point problems, which can be viewed as a refined version of the primal-dual hybrid gradient method by Arrow et al. [1]. However, its design and analysis is more closely related to a few recent important works which established the $\mathcal{O}(1/k)$ rate of convergence for solving bilinear saddle point problems (e.g., [22, 18, 16, 11]). In particular, it is equivalent to a linearized version of the alternative direction method of multipliers. The first stochastic version of the primal-dual method was studied by Chen, Lan and Ouyang [4] together with an accelerated scheme. Lan and Zhou [15] revealed some inherent relationship between Nesterov's accelerated gradient method and the primal-dual method, and presented an optimal randomized incremental gradient method. In this section, we provide a thorough analysis for an inexact version of stochastic primal-dual method that differ from the previous studies of stochastic primal-dual method in the following two aspects. First, we need to deal with stochastic and biased subgradient information about the value function $\tilde{v}$. Second, we will investigate how to guarantee the boundedness of output dual multiplier $\bar{y}_N$ in order to generate an approximate stochastic subgradient of V with bounded variance.

Algorithm 2 Inexact primal-dual stochastic approximation

$\zeta = (A, B, b, c)$.
for $k = 1, 2, \cdots, N$ **do**
Let G_{k-1} be a stochastic, independent of x_{k-1}, $\bar{\epsilon}$-subgradient of $\tilde{v}$, i.e.,

$$g(x_{k-1}) \equiv \mathbb{E}[G_{k-1}] \in \partial_{\bar{\epsilon}}\tilde{v}(x_{k-1}). \tag{2.17}$$

$(x_k, y_k, \tilde{y}_k) = \mathrm{SPDT}(x_{k-1}, y_{k-1}, y_{k-2}, G_{k-1}, u, \zeta, h, X, K_*, \theta_k, \tau_k, \eta_k)$.
end for
Output: $\bar{z}_N \equiv (\bar{x}_N, \bar{y}_N) = \sum_{k=1}^N w_k(x_k, y_k)/\sum_{k=1}^N w_k$.

Throughout this section, we assume that there exists $M > 0$ such that

$$\mathbb{E}[\|G_k\|_*^2] \le M^2 \quad \forall k \ge 1. \tag{2.18}$$

This assumption, in view of (2.17) and Jensen's inequality, then implies that $\|g(x_k)\|_* \le M$. For notational convenience, we assume that the Lipschitz constant of the function $\tilde{v}$ is also bounded by M. Indeed, by definition, any exact subgradient can be viewed as an $\bar{\epsilon}$-subgradient. Hence, the size of subgradient (and the Lipschtiz constant of $\tilde{v}$) can also be bounded by M. Later in this section (see Corollary 7), we will discuss different ways to ensure that the assumption in (2.18) holds.

Below we discuss some convergence properties for Algorithm 2. More specifically, we will first establish in Proposition 2 the relation between (x_{k-1}, y_{k-1}) and (x_k, y_k) after running one step of SPDT, and then discuss in Theorems 3 and 5 the convergence properties of Algorithm 2 applied to problem (2.8). Using these results, we will establish the convergence of the DSA method for solving problem (2.1) in Section 2.4.

Proposition 2 *Let Q be defined in (2.11). For any $1 \le k \le N$ and $(x, y) \in X \times K_*$, we have*

$$\begin{aligned}
&Q(z_k, z) + \langle A(x_k - x), y_k - y_{k-1}\rangle - \theta_k\langle A(x_{k-1} - x), y_{k-1} - y_{k-2}\rangle \\
&\le \tau_k[P_X(x_{k-1}, x) - P_X(x_k, x)] + \tfrac{\eta_k}{2}(\|y - y_{k-1}\|^2 - \|y - y_k\|^2) - \tfrac{\alpha_X\tau_k}{2}\|x_k - x_{k-1}\|^2 \\
&\quad - \tfrac{\eta_k}{2}\|y_{k-1} - y_k\|^2 + \langle \Delta_{k-1}, x_{k-1} - x\rangle + (M + \|G_{k-1}\|_*)\|x_k - x_{k-1}\| + \bar{\epsilon} \\
&\quad + \theta_k\langle A(x_k - x_{k-1}), y_{k-1} - y_{k-2}\rangle,
\end{aligned} \tag{2.19}$$

where

$$\Delta_k := g(x_k) - G_k. \tag{2.20}$$

Proof. By the Lipschitz continuity of $\tilde{v}$ and the definition of an $\bar{\epsilon}$-subgradient, we have

$$\begin{aligned}
\tilde{v}(x_k) &\le \tilde{v}(x_{k-1}) + M\|x_k - x_{k-1}\| \\
&\le \tilde{v}(x) + \langle g(x_{k-1}), x_{k-1} - x\rangle + M\|x_k - x_{k-1}\| + \bar{\epsilon}.
\end{aligned}$$

Moreover, by (2.20), we have

$$\begin{aligned}\langle g(x_{k-1}), x_{k-1} - x\rangle &= \langle G_{k-1}, x_{k-1} - x\rangle + \langle \Delta_{k-1}, x_{k-1} - x\rangle \\ &= \langle G_{k-1}, x_k - x\rangle + \langle G_{k-1}, x_{k-1} - x_k\rangle + \langle \Delta_{k-1}, x_{k-1} - x\rangle \\ &\le \langle G_{k-1}, x_k - x\rangle + \|G_{k-1}\|_* \|x_k - x_{k-1}\| + \langle \Delta_{k-1}, x_{k-1} - x\rangle.\end{aligned}$$

Combining the above two inequalities, we obtain

$$\tilde{v}(x_k) - \tilde{v}(x) \le \langle G_{k-1}, x_k - x\rangle + \langle \Delta_{k-1}, x_{k-1} - x\rangle + (M + \|G_{k-1}\|_*)\|x_k - x_{k-1}\| + \bar{\epsilon}. \qquad (2.21)$$

Also observe that by the optimality conditions of (2.15) and (2.16) (with input $p = x_{k-1}, d = y_{k-1}, d_{\_} = y_{k-2}, \tilde{v}' = G_{k-1}, u = u, \xi = (A, B, b, c), h = h, X = X, K_* = K_*, \theta = \theta_k, \tau = \tau_k, \eta = \eta_k$, output $(p_+, d_+, \tilde{d}) = (x_k, y_k, \tilde{y}_k)$ (see, e.g., Lemma 2 of [13]), we have

$$\begin{aligned}&\langle -A(x_k - x), \tilde{y}_k\rangle + h(x_k, c) - h(x, c) + \langle G_{k-1}, x_k - x\rangle \\ &\qquad \le \tau_k[P_X(x_{k-1}, x) - P_X(x_k, x) - P_X(x_{k-1}, x_k)], \forall x \in X, \qquad (2.22)\\ &\langle -b - Bu + Ax_k, y_k - y\rangle \le \tfrac{\eta_k}{2}[\|y_{k-1} - y\|^2 - \|y_k - y\|^2 - \|y_{k-1} - y_k\|^2], \forall y \in K_*. \qquad (2.23)\end{aligned}$$

Using the definition of Q in (2.11) and the relations (2.21), (2.22) and (2.23), we have

$$\begin{aligned}&Q(z_k, z) + \langle A(x_k - x), y_k - \tilde{y}_k\rangle \le \tau_k[P_X(x_{k-1}, x) - P_X(x_k, x)] + \tfrac{\eta_k}{2}[\|y_{k-1} - y\|^2 - \|y_k - y\|^2] \\ &\quad - \tau_k P_X(x_{k-1}, x_k) - \tfrac{\eta_k}{2}\|y_{k-1} - y_k\|^2 + \langle \Delta_{k-1}, x_{k-1} - x\rangle + (M + \|G_{k-1}\|_*)\|x_k - x_{k-1}\| + \bar{\epsilon}.\end{aligned}$$

Also note that by the definition of $\tilde{y}_k$ (i.e., $\tilde{d}$ in (2.14)), we have $\tilde{y}_k = \theta_k(y_{k-1} - y_{k-2}) + y_{k-1}$ and hence

$$\begin{aligned}\langle A(x_k - x), y_k - \tilde{y}_k\rangle &= \langle A(x_k - x), y_k - y_{k-1}\rangle - \theta_k\langle A(x_k - x), y_{k-1} - y_{k-2}\rangle \\ &= \langle A(x_k - x), y_k - y_{k-1}\rangle - \theta_k\langle A(x_{k-1} - x), y_{k-1} - y_{k-2}\rangle \\ &\quad - \theta_k\langle A(x_k - x_{k-1}), y_{k-1} - y_{k-2}\rangle.\end{aligned}$$

Our result then immediately follows from the above two relations and the strong convexity of P_X (see (1.6)). ■

We are now ready to establish some important convergence properties for the iterative applications of SPDTs stated in Algorithm 2.

Theorem 3 *If the parameters $\{\theta_k\}$, $\{w_k\}$, $\{\tau_k\}$ and $\{\eta_k\}$ in Algorithm 2 satisfy*

$$\begin{aligned}w_k\theta_k &= w_{k-1}, 1 \le k \le N, && (a)\\ w_k\tau_k &\ge w_{k+1}\tau_{k+1}, 1 \le k \le N-1, && (b)\\ w_k\eta_k &\ge w_{k+1}\eta_{k+1}, 1 \le k \le N-1, && (c)\\ w_k\tau_k\eta_{k-1}\alpha_X &\ge 2w_{k-1}\|A\|^2, 1 \le k \le N-1, && (d)\\ \tau_N\eta_N\alpha_X &\ge 2\|A\|^2, && (e)\end{aligned} \qquad (2.24)$$

then we have

$$Q(\bar{z}_N, z) \le \frac{1}{\sum_{k=1}^N w_k}\left(w_1\tau_1 P_X(x_0, x) + \tfrac{w_1\eta_1}{2}\|y_0 - y\|^2 - \tfrac{w_N\eta_N}{2}\|y_N - y\|^2 + \textstyle\sum_{k=1}^N \Lambda_k\right) \qquad (2.25)$$

for any $z \in Z$, where

$$\Lambda_k := w_k\left[(M + \|G_{k-1}\|_*)^2/(\alpha_X\tau_k) + \langle \Delta_k, x_{k-1} - x\rangle + \bar{\epsilon}\right]. \qquad (2.26)$$

Proof. Multiplying the both sides of (2.19) by w_k for each $k \ge 1$, summing them up over $1 \le k \le N$ and using the relations in (2.24).a), (2.24).b) and (2.24).c), we have

$$\begin{aligned}
&\textstyle\sum_{k=1}^N w_k Q(z_k, z) \\
\le& w_1\tau_1 P_X(x_0, x) + \tfrac{w_1\eta_1}{2}\|y_0 - y\|^2 - \tfrac{w_N\eta_N}{2}\|y_N - y\|^2 + \textstyle\sum_{k=1}^N w_k\bar{\epsilon} \\
&- w_N\tau_N P_X(x_N, x) - w_N\langle A(x_N - x), y_N - y_{N-1}\rangle - \tfrac{w_N\eta_N}{2}\|y_N - y_{N-1}\|^2 \\
&- \textstyle\sum_{k=1}^N [\tfrac{\alpha_X w_k\tau_k}{4}\|x_k - x_{k-1}\|^2 + \tfrac{w_{k-1}\eta_{k-1}}{2}\|y_{k-1} - y_{k-2}\|^2 \\
&+ w_{k-1}\langle A(x_k - x_{k-1}), y_{k-1} - y_{k-2}\rangle] - \textstyle\sum_{k=1}^N \tfrac{\alpha_X w_k\tau_k}{4}\|x_k - x_{k-1}\|^2 \\
&+ \textstyle\sum_{k=1}^N w_k(M + \|G_{k-1}\|_*)\|x_k - x_{k-1}\| + \sum_{k=1}^N w_k\langle \Delta_k, x_{k-1} - x\rangle.
\end{aligned} \tag{2.27}$$

Now, by the Cauchy-Schwarz inequality and the strong convexity of P_X and (2.24).e),

$$\begin{aligned}
&- \tau_N P_X(x_N, x) - \langle A(x_N - x), y_N - y_{N-1}\rangle - \tfrac{\eta_N}{2}\|y_N - y_{N-1}\|^2 \\
\le& - \tfrac{\alpha_X\tau_N}{2}\|x - x_N\|^2 + \|A\|\|x_N - x\|\|y_N - y_{N-1}\| - \tfrac{\eta_N}{2}\|y_N - y_{N-1}\|^2 \le 0.
\end{aligned}$$

Similarly, by the Cauchy-Schwarz inequality and (2.24).d), we have

$$\begin{aligned}
&- \textstyle\sum_{k=1}^N [\tfrac{\alpha_X w_k\tau_k}{4}\|x_k - x_{k-1}\|^2 + \tfrac{w_{k-1}\eta_{k-1}}{2}\|y_{k-1} - y_{k-2}\|^2 \\
&+ w_{k-1}\langle A(x_k - x_{k-1}), y_{k-1} - y_{k-2}\rangle] \le 0.
\end{aligned}$$

Moreover, using the fact that $-at^2/2 + b \le b^2/(2a)$, we can easily see that

$$- \textstyle\sum_{k=1}^N \left[\tfrac{\alpha_X\tau_k}{4}\|x_k - x_{k-1}\|^2 + (M + \|G_{k-1}\|_*)\|x_k - x_{k-1}\|\right] \le \sum_{k=1}^N \tfrac{(M+\|G_{k-1}\|_*)^2}{\tau_k\alpha_X}.$$

Using the above three inequalities in (2.27), we have

$$\begin{aligned}
\textstyle\sum_{k=1}^N w_k Q(z_k, z) \le& w_1\tau_1 P_X(x_0, x) + \tfrac{w_1\eta_1}{2}\|y_0 - y\|^2 - \tfrac{w_N\eta_N}{2}\|y_N - y\|^2 \\
&+ \textstyle\sum_{k=1}^N w_k \left(\tfrac{(M+\|G_{k-1}\|_*)^2}{\alpha_X\tau_k} + \langle \Delta_k, x_{k-1} - x\rangle + \bar{\epsilon}\right).
\end{aligned}$$

Dividing both sides of above inequality by $\sum_{k=1}^N w_k$, and using the convexity of Q and the definition of $\bar{z}_N$, we obtain (2.25). ■

We also need the following technical result for the analysis of Algorithm 2.

Lemma 4 *Let $x_0^v \equiv x_0$ and*

$$x_k^v := \text{argmin}_{x\in X}\{\langle \Delta_{k-1}, x\rangle + \tau_k P_X(x_{k-1}^v, x)\} \tag{2.28}$$

for any $k \ge 1$. Then for any $x \in X$,

$$\textstyle\sum_{k=1}^N w_k\langle \Delta_{k-1}, x_{k-1}^v - x\rangle \le \sum_{k=1}^N w_k\tau_k[P_X(x_{k-1}, x) - P_X(x_k, x)] + \sum_{k=1}^N \tfrac{w_k\|\Delta_{k-1}\|_*^2}{2\alpha_X\tau_k}. \tag{2.29}$$

Proof. It follows from the definition of x_k^v in (2.28) and Lemma 2.1 of [19] that

$$\tau_k P_X(x_k^v, x) \le \tau_k P_X(x_{k-1}^v, x) - \langle \Delta_{k-1}, x_{k-1}^v - x\rangle + \tfrac{\|\Delta_{k-1}\|_*^2}{2\alpha_X\tau_k},$$

for all $k \ge 1$. Multiplying w_k on both sides of the above inequality and summing them up from $k = 1$ to N, we obtain (2.29). ■

Theorem 5 below provides certain bounds for the following two gap functions:

$$\text{gap}_*(\bar{z}) \equiv \text{gap}_*(\bar{z}, X) := \max\{Q(\bar{z}; x, y_*) : x \in X\}, \tag{2.30}$$

$$\text{gap}_\delta(\bar{z}) \equiv \text{gap}_\delta(\bar{z}, X, K_*) := \max\{Q(\bar{z}, x, y) + \langle \delta, y\rangle : (x, y) \in X \times K_*\}. \tag{2.31}$$

The gap function in (2.30) will be used to measure the error associated with an approximate subgradient, while the perturbed gap function in (2.31) will be used to measure both functional optimality gap and infeasibility of the conic constraint. In particular, we will apply the first gap function to the second and third stage, and the latter one to the first stage.

Theorem 5 *Suppose the parameters $\{\theta_k\}$, $\{w_k\}$, $\{\tau_k\}$ and $\{\eta_k\}$ in Algorithm 2 satisfy (2.24).*

a) *For any $N \geq 1$, we have*

$$\mathbb{E}[\mathrm{gap}_*(\bar{z}_N)] \leq (\textstyle\sum_{k=1}^N w_k)^{-1} \left[2w_1\tau_1\Omega_X^2 + \frac{w_1\eta_1}{2}\|y_* - y_0\|^2 + \sum_{k=1}^N \frac{6w_k M^2}{\alpha_X \tau_k}\right] + \bar{\epsilon}. \quad (2.32)$$

b) *If, in addition, $w_1\eta_1 = \ldots = w_N\eta_N$, then*

$$\mathbb{E}[\mathrm{gap}_\delta(\bar{z}_N)] \leq (\textstyle\sum_{k=1}^N w_k)^{-1} \left[2w_1\tau_1\Omega_X^2 + \frac{w_1\eta_1}{2}\|y_0\|^2 + \sum_{k=1}^N \frac{6w_k M^2}{\alpha_X \tau_k}\right] + \bar{\epsilon}, \quad (2.33)$$

$$\mathbb{E}[\|\delta\|] \leq \frac{w_1\eta_1}{\sum_{k=1}^N w_k}\left[2\|y_* - y_0\| + 2\sqrt{\frac{\tau_1}{\eta_1}}\Omega_X + \sqrt{\frac{2}{w_1\eta_1}\sum_{k=1}^N w_k\left(\frac{6M^2}{\alpha_X\tau_k} + \bar{\epsilon}\right)}\right], \quad (2.34)$$

$$\mathbb{E}[\|y_* - \bar{y}_N\|^2] \leq \|y_* - y_0\|^2 + (\textstyle\sum_{k=1}^N w_k)^{-1}\sum_{k=1}^N \frac{2}{\eta_k}\left[2w_1\tau_1\Omega_X^2 + \sum_{i=1}^k w_i(\frac{6M^2}{\tau_i} + \bar{\epsilon})\right], \quad (2.35)$$

where $\delta := (\sum_{k=1}^N w_k)^{-1}[w_1\eta_1(y_0 - y_N)]$.

Proof. We first prove part (a). Letting $y = y_*$ in (2.25) and using the definition of Ω_X in (1.7), we have

$$Q(\bar{z}_N; x, y_*) \leq (\textstyle\sum_{k=1}^N w_k)^{-1}\left[w_1\tau_1\Omega_X^2 + \frac{w_1\eta_1}{2}\|y_* - y_0\|^2 - \frac{w_N\eta_N}{2}\|y_* - y_N\|^2 + \sum_{k=1}^N \Lambda_k\right]. \quad (2.36)$$

Maximizing w.r.t. $x \in X$ on both sides of (2.37), and then taking expectation w.r.t. $\xi_1, \ldots, \xi_N$, we have

$$\mathbb{E}[\mathrm{gap}_*(\bar{z}_N)] \leq (\textstyle\sum_{k=1}^N w_k)^{-1}\left[w_1\tau_1\Omega_X^2 + \frac{w_1\eta_1}{2}\|y_* - y_0\|^2 + \mathbb{E}[\sum_{k=1}^N \Lambda_k]\right] \quad (2.37)$$

Now it follows from (2.26) and (2.29) that

$$\begin{aligned}\textstyle\sum_{k=1}^N \Lambda_k &= \textstyle\sum_{k=1}^N w_k\left(\frac{(M+\|G_{k-1}\|_*)^2}{\tau_k\alpha_X} + \bar{\epsilon} + \langle\Delta_{k-1}, x_{k-1} - x_{k-1}^v\rangle + \langle\Delta_{k-1}, x_{k-1}^v - x\rangle\right)\\ &\leq \textstyle\sum_{k=1}^N w_k\left(\frac{2M^2+2\|G_{k-1}\|_*^2}{\tau_k\alpha_X} + \bar{\epsilon} + \langle\Delta_{k-1}, x_{k-1} - x_{k-1}^v\rangle\right) + w_1\tau_1\Omega_X^2 + \sum_{k=1}^N \frac{w_k\|\Delta_{k-1}\|_*^2}{2\alpha_X\tau_k}.\end{aligned}$$

Note that the random noises Δ_k are independent of x_{k-1} and $\mathbb{E}[\Delta_k] = 0$, hence $\mathbb{E}[\langle\Delta_k, x_{k-1} - x_k^v\rangle] = 0$. Moreover, using the relations that $\mathbb{E}[\|G_{k-1}\|_*^2 \leq M^2$, $\|g(x_{k-1})\| \leq M$ and the triangle inequality, we have

$$\mathbb{E}[\|\Delta_{k-1}\|_*^2] = \mathbb{E}[\|G_{k-1} - g(x_{k-1})\|_*^2] \leq \mathbb{E}[(\|G_{k-1}\|_* + \|g(x_{k-1})\|_*)^2] \leq 4M^2. \quad (2.38)$$

Therefore,

$$\mathbb{E}[\textstyle\sum_{k=1}^N \Lambda_k] \leq w_1\tau_1\Omega_X^2 + \sum_{k=1}^N w_k\left(\frac{6M^2}{\alpha_X\tau_k} + \bar{\epsilon}\right). \quad (2.39)$$

The result (2.32) then follows by using the above relation in (2.37).

We now show part (b) holds. Adding $\langle\delta, y\rangle$ to both sides of (2.25) and using the fact that $w_1\eta_1 = w_N\eta_N$, we have

$$\begin{aligned}Q(\bar{z}_N, z) + \langle\delta, y\rangle &\leq (\textstyle\sum_{k=1}^N w_k)^{-1}[w_1\tau_1 P_X(x_0, x) + w_1\eta_1\left(\frac{1}{2}\|y_0 - y\|^2 - \frac{1}{2}\|y_N - y\|^2 + \langle y_0 - y_N, y\rangle\right)\\ &\quad + \textstyle\sum_{k=1}^N \Lambda_k]\\ &\leq (\textstyle\sum_{k=1}^N w_k)^{-1}[w_1\tau_1 P_X(x_0, x) + \frac{w_1\eta_1}{2}\|y_0\|^2 + \sum_{k=1}^N \Lambda_k].\end{aligned}$$

Maximizing both sides of the above inequality w.r.t. $(x, y) \in X \times K_*$, taking expectation w.r.t $\xi_1, \ldots, \xi_N$ and using (2.31), we obtain

$$\mathbb{E}[\mathrm{gap}_\delta(\bar{z}_N)] \leq (\textstyle\sum_{k=1}^N w_k)^{-1}\left[w_1\tau_1\Omega_X^2 + \frac{w_1\eta_1}{2}\|y_0\|^2 + \mathbb{E}[\sum_{k=1}^N \Lambda_k]\right].$$

The result in (2.33) then follows from the above inequality and (2.39). Now fixing $x = x_*$ in (2.36) and using the fact $Q(\bar{z}_N; x_*, y_*) \geq 0$, we have

$$\tfrac{w_N \eta_N}{2} \|y_* - y_N\|^2 \leq w_1 \tau_1 \Omega_X^2 + \tfrac{w_1 \eta_1}{2} \|y_* - y_0\|^2 + \textstyle\sum_{k=1}^N \Lambda_k.$$

Taking expectation w.r.t. $\xi_1, \ldots, \xi_N$ on both sides of this inequality and using (2.39), we conclude

$$\tfrac{w_N \eta_N}{2} \mathbb{E}[\|y_* - y_N\|^2] \leq 2 w_1 \tau_1 \Omega_X^2 + \tfrac{w_1 \eta_1}{2} \|y_* - y_0\|^2 + \textstyle\sum_{k=1}^N w_k \left(\frac{6M^2}{\alpha_X \tau_k} + \bar{\epsilon} \right), \tag{2.40}$$

which implies that

$$\mathbb{E}[\|y_* - y_N\|] \leq 2 \sqrt{\tfrac{\tau_1}{\eta_1}} \Omega_X + \|y_* - y_0\| + \sqrt{\tfrac{2}{w_1 \eta_1} \textstyle\sum_{k=1}^N w_k \left(\frac{6M^2}{\alpha_X \tau_k} + \bar{\epsilon} \right)}.$$

Using the above inequality and the fact that $\|\delta\| \leq (\sum_{k=1}^N w_k)^{-1} [w_1 \eta_1 (\|y_0 - y_*\| + \|y_* - y_N\|)$, we obtain (2.34). Observe that (2.40) holds for any y_k, $k = 1, \ldots, N$, and hence that

$$\tfrac{w_k \eta_k}{2} \mathbb{E}[\|y_* - y_k\|^2] \leq 2 w_1 \tau_1 \Omega_X^2 + \tfrac{w_1 \eta_1}{2} \|y_* - y_0\|^2 + \textstyle\sum_{i=1}^k w_i \left(\frac{6M^2}{\alpha_X \tau_i} + \bar{\epsilon} \right).$$

Using the above inequality, the convexity of $\|\cdot\|^2$ and the fact that $\bar{y}_N = \sum_{k=1}^N (w_k x_k) / \sum_{k=1}^N w_k$, we conclude that

$$\begin{aligned} \mathbb{E}[\|y_* - \bar{y}_N\|^2] &\leq (\textstyle\sum_{k=1}^N w_k)^{-1} \sum_{k=1}^N \left[\frac{4 w_1 \tau_1 \Omega_X^2}{\eta_k} + \frac{w_1 \eta_1}{\eta_k} \|y_* - y_0\|^2 + \frac{2}{\eta_k} \sum_{i=1}^k w_i (\frac{6M^2}{\tau_i} + \bar{\epsilon}) \right] \\ &= \|y_* - y_0\|^2 + (\textstyle\sum_{k=1}^N w_k)^{-1} \sum_{k=1}^N \left[\frac{4 w_1 \tau_1 \Omega_X^2}{\eta_k} + \frac{2}{\eta_k} \sum_{i=1}^k w_i (\frac{6M^2}{\tau_i} + \bar{\epsilon}) \right], \end{aligned}$$

where the second identity follows from the fact that $w_k \eta_k = w_1 \eta_1$. ∎

Below we provide two different parameter settings for $\{w_k\}, \{\tau_k\}$ and $\{\eta_k\}$ satisfying (2.24). While the first one in Corollary 6 leads to slightly better rate of convergence, the second one in Corollary 6 can guarantee the boundedness of the dual solution in expectation. We will discuss how to use these results when analyzing the convergence of the DSA algorithm.

Corollary 6 *If*

$$w_k = w = 1, \tau_k = \tau = \max\{\tfrac{M\sqrt{3N}}{\Omega_X \sqrt{\alpha_X}}, \tfrac{\sqrt{2}\|A\|}{\sqrt{\alpha_X}}\} \text{ and } \eta_k = \eta = \tfrac{\sqrt{2}\|A\|}{\sqrt{\alpha_X}}, \forall 1 \leq k \leq N, \tag{2.41}$$

then

$$\mathbb{E}[\text{gap}_*(\bar{z}_N)] \leq \tfrac{\sqrt{2}\|A\|(2\Omega_X^2 + \|y_* - y_0\|^2)}{\sqrt{\alpha_X} N} + \tfrac{4\sqrt{3} M \Omega_X}{\sqrt{\alpha_X N}} + \bar{\epsilon}, \tag{2.42}$$

$$\mathbb{E}[\text{gap}_\delta(\bar{z}_N)] \leq \tfrac{\sqrt{2}\|A\|(2\Omega_X^2 + \|y_0\|^2)}{\sqrt{\alpha_X} N} + \tfrac{4\sqrt{3} M \Omega_X}{\sqrt{\alpha_X N}} + \bar{\epsilon}, \tag{2.43}$$

$$\mathbb{E}[\|\delta\|] \leq \tfrac{2\sqrt{2\alpha_X}\|A\|\|y_* - y_0\| + 4\Omega_X \|A\|}{\alpha_X N} + \tfrac{2M(\sqrt{6}\|A\| + \sqrt{3\alpha_X})}{\alpha_X \sqrt{N}} + \sqrt{\tfrac{3\|A\|\bar{\epsilon}}{N\sqrt{\alpha_X}}}, \tag{2.44}$$

$$\mathbb{E}[\|y_* - \bar{y}_N\|^2] \leq \|y_* - y_0\|^2 + 4\Omega_X^2 + \tfrac{2\sqrt{6N} M \Omega_X}{\|A\|} + \tfrac{3\alpha_X (N+1) M^2}{\|A\|^2} + \tfrac{(N+1)\bar{\epsilon}}{2}. \tag{2.45}$$

Proof. We can easily check that the parameter setting in (2.41) satisfies (2.24). It follows from (2.32) and (2.41) that

$$\mathbb{E}[\text{gap}_*(\bar{z}_N)] \leq \tfrac{1}{N} \left[2\tau \Omega_X^2 + \tfrac{\eta}{2} \|y_* - y_0\|^2 + \tfrac{6NM^2}{\alpha_X \tau} \right] + \bar{\epsilon} \leq \tfrac{\sqrt{2}\|A\|(2\Omega_X^2 + \|y_* - y_0\|^2)}{\sqrt{\alpha_X} N} + \tfrac{4\sqrt{3} M \Omega_X}{\sqrt{\alpha_X N}} + \bar{\epsilon}.$$

Moreover, we have $w_1 \eta_1 = w_N \eta_N$. Hence, by (2.33) and (2.41),

$$\mathbb{E}[\text{gap}_\delta(\bar{z}_N)] \leq \tfrac{1}{N} \left[2\tau \Omega_X^2 + \tfrac{\eta}{2} \|y_0\|^2 + \tfrac{6NM^2}{\alpha_X \tau} \right] + \bar{\epsilon} \leq \tfrac{\sqrt{2}\|A\|(2\Omega_X^2 + \|y_0\|^2)}{\sqrt{\alpha_X} N} + \tfrac{4\sqrt{3} M \Omega_X}{\sqrt{\alpha_X N}} + \bar{\epsilon}.$$

Also by (2.34) and (2.41),

$$\mathbb{E}[\|\delta\|] \le \tfrac{\eta}{N}\left[2\|y_* - y_0\| + 2\sqrt{\tfrac{\tau}{\eta}}\Omega_X + \sqrt{\tfrac{2N}{\eta}\left(\tfrac{6M^2}{\alpha_X \tau} + \bar\epsilon\right)}\right]$$
$$\le \tfrac{2\sqrt{2}\|A\|\|y_*-y_0\|}{N\sqrt{\alpha_X}} + \tfrac{2\Omega_X}{N}\left(\tfrac{2\|A\|}{\alpha_X} + \tfrac{\sqrt{6N}\|A\|M}{\Omega_X\alpha_X}\right) + \tfrac{2\sqrt{M}}{\sqrt{\alpha_X N}} + \tfrac{\sqrt{2\bar\epsilon}}{\sqrt{N}}\sqrt{\tfrac{\sqrt{2}\|A\|}{\sqrt{\alpha_X}}},$$

which implies (2.44). Finally, by (2.34) and (2.41),

$$\mathbb{E}[\|y_* - \bar{y}_N\|^2] \le \|y_* - y_0\|^2 + \tfrac{1}{N}\left[\textstyle\sum_{k=1}^N \tfrac{4\tau_k}{\eta_k}\Omega_X^2 + \sum_{k=1}^N \tfrac{2}{\eta_k}\sum_{i=1}^k\left(\tfrac{6M^2}{\tau_i} + \bar\epsilon\right)\right]$$
$$\le \|y_* - y_0\|^2 + 4\Omega_X^2 + \tfrac{2\sqrt{6N}M\Omega_X}{\|A\|} + \tfrac{3\alpha_X(N+1)M^2}{\|A\|^2} + \tfrac{(N+1)\bar\epsilon}{2}.$$

■

In view of (2.45), if $M > 0$ or N is not properly chosen, $\mathbb{E}[\|y_* - \bar{y}_N\|^2]$ might be unbounded. In the following corollary, we slightly modify the selection of τ and η in (2.41) in order to guarantee the boundedness of $\mathbb{E}[\|y_* - \bar{y}_N\|^2]$ even when $M > 0$.

Corollary 7 *If*

$$w_k = w = 1, \tau_k = \tau = \max\{\tfrac{M\sqrt{3N}}{\Omega_X\sqrt{\alpha_X}}, \tfrac{\sqrt{2}\|A\|}{\sqrt{\alpha_X N}}\} \text{ and } \eta_k = \eta = \tfrac{\sqrt{2N}\|A\|}{\sqrt{\alpha_X}}, \forall 1 \le k \le N, \tag{2.46}$$

then

$$\mathbb{E}[\mathrm{gap}_*(\bar{z}_N)] \le \tfrac{2\sqrt{2}\|A\|\Omega_X^2}{N\sqrt{\alpha_X N}} + \tfrac{\|A\|\|y_*-y_0\|^2 + 4\sqrt{3}M\Omega_X}{\sqrt{\alpha_X N}} + \bar\epsilon, \tag{2.47}$$

$$\mathbb{E}[\mathrm{gap}_\delta(\bar{z}_N)] \le \tfrac{2\sqrt{2}\|A\|\Omega_X^2}{N\sqrt{\alpha_X N}} + \tfrac{\|A\|\|y_0\|^2 + 4\sqrt{3}M\Omega_X}{\sqrt{\alpha_X N}} + \bar\epsilon, \tag{2.48}$$

$$\mathbb{E}[\|\delta\|] \le \tfrac{2\sqrt{2}\|A\|\|y_*-y_0\| + 4\sqrt{M\|A\|\Omega_X}}{\sqrt{\alpha_X N}} + \tfrac{2\sqrt{6}\|A\|M}{\alpha_X} + \tfrac{4\Omega_X^2\|A\|^2}{N\alpha_X} + \sqrt{\tfrac{3\|A\|\bar\epsilon}{\sqrt{\alpha_X N}}}, \tag{2.49}$$

$$\mathbb{E}[\|y_* - \bar{y}_N\|^2] \le \|y_* - y_0\|^2 + \tfrac{2\Omega_X^2}{N} + \tfrac{\sqrt{6}(1+\alpha_X)M\Omega_X}{\|A\|} + \tfrac{\sqrt{\alpha_X N}\bar\epsilon}{\sqrt{2}\|A\|}. \tag{2.50}$$

Proof. The proofs of (2.47)-(2.50) are similar to Corollary 6 and hence the details are skipped. ■

Note that by using the parameter setting (2.46), we still obtain the optimal rate of convergence in terms of the dependence on N, with a slightly worse dependence on $\|A\|$ and $\|y_*\|$ than the one obtained by using the parameter setting in (2.41). However, using the setting (2.46), we can bound $\mathbb{E}[\|\bar{y}_N - y_*\|^2]$ as long as $N = \mathcal{O}(1/\bar\epsilon^2)$, while this statement does not necessarily hold for the parameter setting in (2.41).

We now state a technical result regarding the functional optimality gap and primal infeasibility, which slightly generalize Proposition 2.1 of [23] to conic programming.

Lemma 8 *If there exist random vectors $\delta \in \mathbb{R}^m$ and $\bar{z} \equiv (\bar{x}, \bar{y}) \in Z$ such that*

$$\mathbb{E}[\mathrm{gap}_\delta(\bar{z})] \le \epsilon_o, \tag{2.51}$$

then

$$\mathbb{E}[h(\bar{x}, c) + \tilde{v}(\bar{x}) - (h(x^*, c) + \tilde{v}(x^*))] \le \epsilon_0,$$
$$A\bar{x} - b - \delta \in K \text{ a.s.},$$

where x^* is an optimal solution of problem (2.8).

Proof. Letting $x = x^*$ and $y = 0$ in the definition of (2.31), we can easily see that

$$h(\bar{x}, c) + \tilde{v}(\bar{x}) - (h(x^*, c) + \tilde{v}(x^*)) \leq \text{gap}_\delta(\bar{z}).$$

Moreover, in view of (2.11) and (2.31), we must have $A\bar{x} - Bu - b - \delta \in K$ almost surely. Otherwise, $\mathbb{E}[\text{gap}_\delta(\bar{z})]$ would be unbounded as y runs throughout K^* in the definition of $\text{gap}_\delta(\bar{z})$. ■

In the next result, we will provide a bound on the optimal dual variable y_*. By doing so, we show that the complexity of Algorithm 2 only depends on the parameters for the primal problem along with the smallest nonzero eigenvalue of A and the initial point y_0, even though the algorithm is primal-dual type method.

Lemma 9 *Let (x_*, y_*) be an optimal solution to problem (2.8). If the subgradients of the objective function $v_h(x) := h(x, c) + \tilde{v}(\cdot)$ are bounded, i.e., $\|v_h'(x)\|_2 \leq M_h$ for any $x \in X$, then there exists y_* s.t.*

$$\|y_*\| \leq \tfrac{M_h}{\sigma_{min}(A)}, \tag{2.52}$$

where $\sigma_{min}(A)$ denotes the smallest nonzero singular value of A.

Proof. We consider two cases. Case 1: $A^T y_* = 0$, i.e., y_* belongs to the null space of A. Since for any $\lambda \geq 0$, λy_* is still an optimal dual solution to problem (2.8), we have (2.52) holds. Case 2: $A^T y_* \neq 0$. By the definition of the saddle point, we have

$$\langle b + Bu - Ax_*, y_* \rangle + h(x_*, c) + \tilde{v}(x^*) \leq \langle b + Bu - Ax, y_* \rangle + h(x, c) + \tilde{v}(x), \ \forall x \in X,$$

which implies

$$h(x_*, c) + \tilde{v}(x^*) + \langle A^T y_*, x - x^* \rangle \leq h(x, c) + \tilde{v}(x), \ \forall x \in X. \tag{2.53}$$

Hence $A^T y_*$ is a subgradient of v_h at the point x^*. Without loss of generality, we assume that y^* belongs to the column space of A^T (i.e., y^* is perpendicular to the eigenspace associated with eigenvalue 0). Otherwise we can show that the projection of y^* onto the column space of A^T will also satisfy (2.53). Using this observation, we have

$$\|A^T y_*\|_2^2 = (y_*)^T A A^T y_* = (y_*)^T U^T \Lambda U y_* \geq \sigma_{min}(AA^T)\|Uy_*\|^2 = \sigma_{min}^2(A)\|y_*\|^2,$$

where U is an orthonormal matrix whose rows consist of the eigenvectors of AA^T and Λ is the diagonal matrix whose elements are the corresponding eigenvalues. Our result then follows from the above inequality and the fact that $\|A^T y_*\|_2 \leq M_h$. ■

2.4 Convergence analysis for DSA

Our goal in this subsection is to establish the complexity of the DSA algorithm for solving problem 2.4.

The basic idea is to apply the results we obtained in the previous section regarding the iterative applications of SPDTs to the three loops stated in the DSA algorithm. More specifically, using these results we will show how to generate stochastic ϵ-subgradients for the value functions v^2 and v^3 in the middle and innermost loops, respectively, and how to compute a nearly optimal solution for problem 2.4 in the outer loop of the DSA algorithm .

In order to apply these results to the saddle-point reformulation for the second and first stage problems (see (2.6) and (2.10), we need to make sure that the condition in (2.18) holds for the value functions, v^3 and v^2 respectively, associated with the optimization problems in their subsequent stages. For this purpose, we assume that the less aggressive algorithmic parameter setting in (2.46) is applied to solve the second stage saddle point problems in (2.6), while a more aggressive parameter setting in (2.41) is used to solve the first stage and last stage saddle point problems in (2.10) and (2.7), respectively. Moreover, we need the boundedness of the operators B^2 and B^3:

$$\|B^2\| \le \mathcal{B}_2 \quad \text{and} \quad \|B^3\| \le \mathcal{B}_3 \tag{2.54}$$

in order to guarantee that the generated stochastic subgradients for the value functions v^2 and v^3 have bounded variance.

For notational convenience, we use $\Omega_i \equiv \Omega_{X^i}$ and $\alpha_i \equiv \alpha_{X^i}$, $i = 1, 2, 3$, to denote the diameter and strongly convex modulus associated with the distance generating function for the feasible set X^i (see (1.7)). Lemma 10 shows some convergence properties for the innermost loop of the DSA algorithm.

Lemma 10 *If the parameters $\{w_k^3\}$, $\{\tau_k^3\}$ and $\{\eta_k^3\}$ are set to* (2.41) *(with $M = 0$ and $A = A_j^3$) and*

$$N_3 \equiv N_{3,j} := \tfrac{3\sqrt{2}\|A_j^3\|[2(\Omega_3)^2 + \|y_{*,j}^3 - y_0^3\|^2]}{\sqrt{\alpha_3}\epsilon}, \tag{2.55}$$

then $B_j^3 \bar{y}_j^3$ is a stochastic $(\epsilon/3)$-subgradient of the value function v^3 at x_{j-1}^2. Moreover, there exists a constant $M_3 \ge 0$ such that $\|v^3(x_1) - v^3(x_2)\| \le M_3\|x_1 - x_2\|, \forall x_1, x_2 \in X^3$ and

$$\mathbb{E}[\|B_j^3 \bar{y}_j^3\|_*^2] \le M_3. \tag{2.56}$$

Proof. The innermost loop of the DSA algorithm is equivalent to the application of Algorithm 2 to the last stage saddle point problem in (2.7). Note that for this problem, we do not have any subsequent stages and hence $\tilde{v} = 0$. In other words, the subgradients of $\tilde{v}$ are exact. In view of Corollary 6 (with $M = 0$ and $\bar{\epsilon} = 0$) and the definition of N_3 in (2.55), we have

$$\text{gap}_*(\bar{z}_j^3) \le \tfrac{\sqrt{2}\|A_j^3\|[2(\Omega_3)^2 + \|y_*^3 - y_0^3\|^2]}{\sqrt{\alpha_3}N_3} \le \tfrac{\epsilon}{3}.$$

This observation, in view of Lemma 1, then implies that $B_j^3 \bar{y}_j^3$ is a stochastic $(\epsilon/3)$-subgradient of v^3 at x_{j-1}^2. Moreover, it follows from (2.45) (with $M = 0$ and $\bar{\epsilon} = 0$) that

$$\|y_{*,j}^3 - \bar{y}_j^3\|^2 \le \|y_{*,j}^3 - y_0^3\|^2 + 4(\Omega_3)^2 + \tfrac{(N_3+1)\epsilon}{2}.$$

This inequality, in view of the selection of N_3 in (2.55), the assumption that $y_{*,j}^3$ is well-defined, and (2.54), then implies the latter part of our result. ■

Lemma 11 describes some convergence properties for the middle loop of the DSA algorithm.

Lemma 11 *Assume that the parameters for the innermost loop are set according to Lemma 10. If the parameters $\{w_j^2\}$, $\{\tau_j^2\}$ and $\{\eta_j^2\}$ for the middle loop are set to* (2.46) *(with $M = M_3$ and $A = A_i^2$) and*

$$N_2 \equiv N_{2,i} := \left(\tfrac{12\sqrt{2}\|A_i^2\|\Omega_2}{\sqrt{\alpha_2}\epsilon}\right)^{\frac{2}{3}} + \left[\tfrac{6\left(\|A_i^2\|\|y_{*,i}^2 - y_0^2\|^2 + 4\sqrt{3}M_3\Omega_2\right)}{\sqrt{\alpha_2}\epsilon}\right]^2, \tag{2.57}$$

then $B_i^2 \bar{y}_i^2$ is a stochastic $(2\epsilon/3)$-subgradient of the value function v^2 at x_{i-1}^1. Moreover, there exists a constant $M_2 \ge 0$ such that $\|v^2(x_1) - v^2(x_2)\| \le M_2\|x_1 - x_2\|, \forall x_1, x_2 \in X^2$ and

$$\mathbb{E}[\|B_i^2 \bar{y}_i^2\|_*^2] \le M_2. \tag{2.58}$$

Proof. The middle loop of the DSA algorithm is equivalent to the application of Algorithm 2 to the second stage saddle point problem in (2.6). Note that for this problem, we have $\tilde{v} = v^3$. Moreover, by Lemma 10, the stochastic subgradients of v^3 are computed by the innermost loop with tolerance $\bar{\epsilon} = \epsilon/3$. In view of Corollary 7 (with $M = M_3$ and $\bar{\epsilon} = \epsilon/3$) and the definition of N_2 in (2.57), we have

$$\text{gap}_*(\bar{z}_i^2) \leq \tfrac{2\sqrt{2}\|A_i^2\|\Omega_2}{N_2\sqrt{\alpha_2 N_2}} + \tfrac{\|A_i^2\|\|y_{*,i}^2 - y_0^2\|^2 + 4\sqrt{3}M_3\Omega_2}{\sqrt{\alpha_2 N_2}} + \bar{\epsilon} \leq \tfrac{2\epsilon}{3}.$$

This observation, in view of Lemma 1, then implies that $B_i^2 \bar{y}_i^2$ is a stochastic $(2\epsilon/3)$-subgradient v^3 at x_{j-1}^2. Moreover, it follows from (2.50) (with $M = M_3$ and $\bar{\epsilon} = \epsilon/3$) that

$$\|y_{*,i}^2 - \bar{y}_i^2\|^2 \leq \|y_{*,i}^2 - y_0^2\|^2 + \tfrac{2\Omega_2^2}{N_2} + \tfrac{\sqrt{6}(1+\alpha_2)M_3\Omega_2}{\|A_i^2\|} + \tfrac{\sqrt{\alpha_2 N_2}\epsilon}{3\sqrt{2}\|A_i^2\|}.$$

This inequality, in view of the selection of N_2 in (2.57), the assumption that $y_{*,i}^2$ is well-defined, and (2.54), then implies the latter part of our result. ∎

We are now ready to establish the main convergence properties of the DSA algorithm applied to a three-stage stochastic optimization problem.

Theorem 12 *Suppose that the parameters for the innermost and middle loop in the DSA algorithm are set according to Lemma 10 and Lemma 11, respectively. If the parameters $\{w_i\}$, $\{\tau_i\}$ and $\{\eta_i\}$ for the outer loop are set to (2.41) (with $M = M_2$ and $A = A^1$) and*

$$N_1 := \max\left\{ \tfrac{6\sqrt{2}\|A^1\|[2(\Omega_1)^2 + \|y_0^1\|^2]}{\sqrt{\alpha_1}\epsilon} + \left(\tfrac{24\sqrt{3}M_2\Omega_1}{\sqrt{\alpha_1}\epsilon}\right)^2, \right.$$
$$\left. \tfrac{6\|A^1\|(\sqrt{2\alpha_1}\|y_*^1 - y_0^1\| + 2\Omega_1 + 3\sqrt{\alpha_1})}{\alpha_1\epsilon} + \left(\tfrac{6\sqrt{3}M_2(\sqrt{2}\|A^1\| + \sqrt{\alpha_1})}{\alpha_1\epsilon}\right)^2 \right\}, \tag{2.59}$$

then we will find a solution $\bar{x}^1 \in X^1$ and a vector $\delta \in \mathbb{R}^{m^1}$ s.t.

$$\begin{aligned}
&\mathbb{E}[h(\bar{x}^1, c) + v^2(\bar{x}^1) - (h(x^*, c) + v^2(x^*))] \leq \epsilon,\\
&A\bar{x}^1 - b - \delta \in K^1, a.s.,\\
&\mathbb{E}[\|\delta\|] \leq \epsilon,
\end{aligned}$$

where x^* denotes the optimal solution of problem 2.4.

Proof. The outer loop of the DSA algorithm is equivalent to the application of Algorithm 2 to the first stage saddle point problem in (2.10). Note that for this problem, we have $\tilde{v} = v^2$. Moreover, by Lemma 11, the stochastic subgradients of v^2 are computed by the middle loop with tolerance $\bar{\epsilon} = 2\epsilon/3$. In view of Corollary 6 (with $M = M_2$ and $\bar{\epsilon} = 2\epsilon/3$) and the definition of N_1 in (2.59), we conclude that there exist $\delta \in \mathbb{R}^{m^1}$ s.t.

$$\begin{aligned}
\mathbb{E}[\text{gap}_\delta(\bar{z}_N^1)] &\leq \tfrac{\sqrt{2}\|A^1\|(2\Omega_1^2 + \|y_0^1\|^2)}{\sqrt{\alpha_1}N_1} + \tfrac{4\sqrt{3}M_2\Omega_1}{\sqrt{\alpha_1 N_1}} + \tfrac{2\epsilon}{3} \leq \epsilon,\\
\mathbb{E}[\|\delta\|] &\leq \tfrac{2\sqrt{2\alpha_1}\|A^1\|\|y_*^1 - y_0^1\| + 4\Omega_1\|A^1\|}{\alpha_1 N_1} + \tfrac{2M_2(\sqrt{6}\|A^1\| + \sqrt{3\alpha_1})}{\alpha_1\sqrt{N_1}} + \sqrt{\tfrac{2\|A^1\|\epsilon}{N_1\sqrt{\alpha_1}}} \leq \epsilon,
\end{aligned}$$

which together with Lemma 8 then imply our result. ∎

We now add a few remarks about the convergence of the DSA algorithm. Firstly, it follows from Lemma 11 and Theorem 12 that the number of random samples ξ_2 and ξ_3 are given by

$$N_1 = \mathcal{O}(1/\epsilon^2) \quad \text{and} \quad N_1 \times N_2 = \mathcal{O}(1/\epsilon^4), \tag{2.60}$$

respectively. Secondly, it turns out that the convergence of the DSA algorithm relies on the dual variable y_*^1, $y_{*,i}^2$, and $y_{*,j}^3$. We can use Lemma 9 as a tool to estimate the size of the dual variables and some tools from random matrix theory [29] to estimate the smallest singular values in case these quantities are not easily computable.

3 Three-stage problems with strongly convex objectives

In this section, we show that the complexity of the DSA algorithm can be significantly improved if the objective functions h^i, $i = 1, 2, 3$, are strongly convex. We will first refine the convergence properties of Algorithm 2 under the strong convexity assumption about $h(x, c)$ and then use these results to improve the complexity results of the DSA algorithm.

3.1 Basic tools: inexact primal-dual stochastic approximation under strong convexity

Our goal in this subsection is to study the convergence properties of Algorithm 2 applied to problem (2.8) under the assumption that $h(x, c)$ is strongly convex, i.e., $\exists \mu_h > 0$ s.t.

$$h(x_1, c) - h(x_2, c) - \langle h'(x_2, c), x_1 - x_2 \rangle \geq \mu_h P_X(x_2, x_1),\ \forall x_1, x_2 \in X. \tag{3.1}$$

Proposition 13 below shows the relation between (x_{k-1}, y_{k-1}) and (x_k, y_k) after running one step of SPDT when the assumption about h in (3.18) is satisfied.

Proposition 13 *Let Q and Δ_k be defined in (2.11) and (2.20), respectively. For any $1 \leq k \leq N$ and $(x, y) \in X \times K_*$, we have*

$$\begin{aligned}
&Q(z_k, z) + \langle A(x_k - x), y_k - y_{k-1} \rangle - \theta_k \langle A(x_{k-1} - x), y_{k-1} - y_{k-2} \rangle \\
&\leq \tau_k P_X(x_{k-1}, x) - (\tau_k + \mu_h) P_X(x_k, x) + \tfrac{\eta_k}{2}[\|y_{k-1} - y\|^2 - \|y_k - y\|^2] \\
&\quad - \tfrac{\alpha_X \tau_k}{2}\|x_k - x_{k-1}\|^2 - \tfrac{\eta_k}{2}\|y_{k-1} - y_k\|^2 + \bar{\epsilon} + (M + \|G_{k-1}\|_*)\|x_k - x_{k-1}\| \\
&\quad + \theta_k \langle A(x_k - x_{k-1}), y_{k-1} - y_{k-2} \rangle + \langle \Delta_{k-1}, x_{k-1} - x \rangle,
\end{aligned} \tag{3.2}$$

Proof. Since h is strongly convex, we can rewrite (2.22) as

$$\begin{aligned}
&\langle -A_k(x_k - x), \tilde{y}_k \rangle + h(x_k, c_k) - h(x, c_k) + \langle G(x_{k-1}, \xi_k), x_k - x \rangle \\
&\qquad \leq \tau_k P_X(x_{k-1}, x) - (\tau_k + \mu_h) P_X(x_k, x) - \tau_k P_X(x_{k-1}, x_k).
\end{aligned}$$

It then follows from (2.11), (2.21), (2.23) and the above inequality that

$$\begin{aligned}
&Q(z_k, z) + \langle A(x_k - x), y_k - \tilde{y}_k \rangle \leq \tau_k P_X(x_{k-1}, x) - (\tau_k + \mu_h) P_X(x_k, x) - \tau_k P_X(x_{k-1}, x_k) \\
&+ \tfrac{\eta_k}{2}[\|y_{k-1} - y\|^2 - \|y_k - y\|^2 - \|y_{k-1} - y_k\|^2] + (M + \|G_{k-1}\|_*)\|x_k - x_{k-1}\| + \langle \Delta_{k-1}, x_{k-1} - x \rangle + \bar{\epsilon}.
\end{aligned}$$

Similarly to the proof of (2), using the above relation, the definition of $\tilde{y}_k$ in (2.14) and the strong convexity of P in (1.6), we have (3.2). ■

With the help of Proposition 13, we can provide bounds of two gap functions $\text{gap}_*(\bar{z}_N)$ and $\text{gap}_*\delta(\bar{z}_N)$ under the strong convexity assumption of h.

Theorem 14 *Suppose that the parameters $\{\theta_k\}$, $\{w_k\}$, $\{\tau_k\}$ and $\{\eta_k\}$ satisfy (2.24) with (2.24).b) replaced by*

$$w_k(\mu_h + \tau_k) \geq w_{k+1}\tau_{k+1},\ k = 1, \ldots, N-1. \tag{3.3}$$

a) *For $N \geq 1$, we have*

$$\mathbb{E}[\text{gap}_*(\bar{z}_N)] \leq (\textstyle\sum_{k=1}^N w_k)^{-1}[2w_1\tau_1\Omega_X^2 + \frac{w_1\eta_1}{2}\|y_0 - y_*\|^2 + \sum_{k=1}^N \frac{6M^2 w_k}{\alpha_X \tau_k}] + \bar{\epsilon}. \tag{3.4}$$

b) *If, in addition, $w_1\eta_1 = \ldots = w_N\eta_N$, then*

$$\mathbb{E}[\text{gap}_\delta(\bar{z}_N)] \leq (\textstyle\sum_{k=1}^N w_k)^{-1}[2w_1\tau_1\Omega_X^2 + \frac{w_1\eta_1}{2}\|y_0\|^2 + \sum_{k=1}^N \frac{6M^2 w_k}{2\alpha_X \tau_k}] + \bar{\epsilon}, \tag{3.5}$$

$$\mathbb{E}[\|\delta\|] \leq \frac{w_1\eta_1}{\sum_{k=1}^N w_k}\left[2\|y_* - y_0\| + 2\sqrt{\tfrac{\tau_1}{\eta_1}}\Omega_X + \sqrt{\tfrac{2}{w_1\eta_1}\textstyle\sum_{k=1}^N w_k\left(\frac{6M^2}{\alpha_X\tau_k} + \bar{\epsilon}\right)}\right], \tag{3.6}$$

$$\mathbb{E}[\|y_* - \bar{y}_N\|^2] \leq \|y_* - y_0\|^2 + (\textstyle\sum_{k=1}^N w_k)^{-1}\sum_{k=1}^N \frac{2}{\eta_k}\left[2w_1\tau_1\Omega_X^2 + \sum_{i=1}^k w_i(\frac{6M^2}{\tau_i} + \bar{\epsilon})\right],$$

where $\delta = (\sum_{k=1}^N w_k)^{-1}[w_1\eta_1(y_0 - y_N)]$.

Proof. We first show part a) holds. Multiplying both sides of (3.2) by w_k for every $k \geq 1$, summing up the resulting inequalities over $1 \leq k \leq N$, and using the relations in (2.24) and (3.3), we have

$$
\begin{aligned}
&\textstyle\sum_{k=1}^{N} w_k Q(z_k, z) \\
&\leq \textstyle\sum_{k=1}^{N} [w_k \tau_k P_X(x_{k-1}, x) - w_k(\tau_k + \mu_h) P_X(x_k, x)] - \sum_{k=1}^{N} \frac{\alpha_X w_k \tau_k}{2} \|x_k - x_{k-1}\|^2 \\
&\quad + \textstyle\sum_{k=1}^{N} [\frac{w_k \eta_k}{2} \|y_{k-1} - y\|^2 - \frac{w_k \eta_k}{2} \|y_k - y\|^2] - \sum_{k=1}^{N} \frac{w_k \eta_k}{2} \|y_{k-1} - y_k\|^2 \\
&\quad + \textstyle\sum_{k=1}^{N} w_{k-1} \langle A(x_{k-1} - x_k), y_{k-1} - y_{k-2} \rangle + \sum_{k=1}^{N} w_k \bar{\epsilon} + w_N \langle A(x - x_N), y_N - y_{N-1} \rangle \\
&\quad + \textstyle\sum_{k=1}^{N} w_k (M + \|G_{k-1}\|_*) \|x_k - x_{k-1}\| + \sum_{k=1}^{N} w_k \langle \Delta_{k-1}, x_{k-1} - x \rangle \\
&\leq w_1 \tau_1 P_X(x_0, x) + \tfrac{w_1 \eta_1}{2} \|y_0 - y\|^2 - \tfrac{w_N \eta_N}{2} \|y_N - y\|^2 \\
&\quad + \textstyle\sum_{k=1}^{N} w_k \bar{\epsilon} + \sum_{k=1}^{N} \frac{(M + \|G_{k-1}\|_*)^2 w_k}{\alpha_X \tau_k} + \sum_{k=1}^{N} w_k \langle \Delta_{k-1}, x_{k-1} - x \rangle \\
&\quad - w_N (\tau_N + \mu_h) P_X(x_N, x) + w_N \langle A(x - x_N), y_N - y_{N-1} \rangle - \tfrac{w_N \eta_N}{2} \|y_N - y_{N-1}\|^2 \\
&\leq w_1 \tau_1 P_X(x_0, x) + \tfrac{w_1 \eta_1}{2} \|y_0 - y\|^2 - \tfrac{w_N \eta_N}{2} \|y_N - y\|^2 \\
&\quad + \textstyle\sum_{k=1}^{N} w_k \bar{\epsilon} + \sum_{k=1}^{N} \frac{(M + \|G_{k-1}\|_*)^2 w_k}{\alpha_X \tau_k} + \sum_{k=1}^{N} w_k \langle \Delta_{k-1}, x_{k-1} - x \rangle,
\end{aligned}
$$

where the last two inequalities follows from similar techniques in the proof of Theorem 3. Dividing both sides of the above inequality, and using the convexity of Q and the definition of $\bar{z}_N$, we have

$$
\begin{aligned}
\max_{z \in X \times K_*} Q(\bar{z}_N, z) \leq (\textstyle\sum_{k=1}^{N} w_k)^{-1} [& w_1 \tau_1 \Omega_X^2 + \tfrac{w_1 \eta_1}{2} \|y_0 - y\|^2 - \tfrac{w_N \eta_N}{2} \|y_N - y\|^2 \\
& + \textstyle\sum_{k=1}^{N} w_k \bar{\epsilon} + \sum_{k=1}^{N} \frac{(M + \|G_{k-1}\|_*)^2 w_k}{\alpha_X \tau_k} + \sum_{k=1}^{N} w_k \langle \Delta_{k-1}, x_{k-1} - x \rangle]
\end{aligned} \tag{3.7}
$$

which, in view of (2.29) and (2.30), then implies

$$
\begin{aligned}
\text{gap}_*(\bar{z}_N) \leq (\textstyle\sum_{k=1}^{N} w_k)^{-1} [& 2 w_1 \tau_1 \Omega_X^2 + \tfrac{w_1 \eta_1}{2} \|y_0 - y_*\|^2 - \tfrac{w_N \eta_N}{2} \|y_N - y_*\|^2 \\
& + \textstyle\sum_{k=1}^{N} w_k \epsilon + \sum_{k=1}^{N} \frac{[\|\Delta_k\|_*^2 + 2(M + \|G_{k-1}\|_*)^2] w_k}{2 \alpha_X \tau_k} + \sum_{k=1}^{N} w_k \langle \Delta_{k-1}, x_{k-1} - x_{k-1}^v \rangle].
\end{aligned}
$$

Taking expectation w.r.t. ξ_k on both sides of above inequality, and using (2.38) and the fact that $x_{k-1} - x_{k-1}^v$ is independent of Δ_{k-1}, we have

$$
\mathbb{E}[\text{gap}_*(\bar{z}_N)] \leq (\textstyle\sum_{k=1}^{N} w_k)^{-1} [2 w_1 \tau_1 \Omega_X^2 + \tfrac{w_1 \eta_1}{2} \|y_0 - y_*\|^2 + \sum_{k=1}^{N} \frac{6 M^2 w_k}{\alpha_X \tau_k}] + \bar{\epsilon}.
$$

The proof of part b) is similar to the one for Theorem 5.b) and hence the details are skipped. ■

In the following two corollaries, we provide two different parameter settings for the selection of $\{w_k\}$, $\{\tau_k\}$ and $\{\eta_k\}$, both of which can guarantee the convergence of Algorithm 2 in terms of the gap functions $\mathbb{E}[\text{gap}_*(\bar{z}_N)]$ and $\mathbb{E}[\text{gap}_\delta(\bar{z}_N)]$. Moreover, the first one in Corollary 15 if $M = 0$ and N is properly chosen in order to ensure the boundedness of $\mathbb{E}[\|y_* - \bar{y}_N\|^2]$ while the other one in Corollary 16 can relax such assumptions.

Corollary 15 *If*

$$
w_k = k, \;\; \tau_k = \tfrac{k-1}{2} \mu_h \;\; \textit{and} \;\; \eta_k = \tfrac{4\|A\|^2}{k \alpha_X \mu_h}, \tag{3.8}
$$

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

$$
\mathbb{E}[\text{gap}_*(\bar{z}_N)] \leq \tfrac{8\|A\|^2 \|y_0 - y_*\|^2}{\alpha_X \mu_h (N+1) N} + \tfrac{24 M^2}{\alpha_X \mu_h (N+1)} + \bar{\epsilon}, \tag{3.9}
$$

$$
\mathbb{E}[\text{gap}_\delta(\bar{z}_N)] \leq \tfrac{8\|A\|^2 \|y_0\|^2}{\alpha_X \mu_h (N+1) N} + \tfrac{24 M^2}{\alpha_X \mu_h (N+1)} + \bar{\epsilon}, \tag{3.10}
$$

$$
\mathbb{E}[\|\delta\|] \leq \tfrac{16\|A\|^2 \|y_* - y_0\|}{N(N+1) \alpha_X \mu_h} + \tfrac{8\sqrt{6}\|A\| M}{\alpha_X \mu_h N^{3/2}} + \tfrac{4\|A\|\sqrt{\bar{\epsilon}}}{(N+1)\sqrt{\alpha_X \mu_h}}, \tag{3.11}
$$

$$
\mathbb{E}[\|y_* - \bar{y}_N\|^2] \leq \|y_* - y_0\|^2 + \tfrac{12 M^2 \alpha_X N}{\|A\|^2} + \tfrac{N(N+1)\alpha_X \mu_h}{2\|A\|^2} \bar{\epsilon}. \tag{3.12}
$$

Proof. Clearly, the parameters w_k, τ_k and η_k in (3.8) satisfy (2.24) with (2.24).b) replaced by (3.3). It then follows from Theorem 14 and (3.8) that

$$
\begin{array}{lll}
\mathbb{E}[\text{gap}_*(\bar{z}_N)] & \leq & \frac{2}{N(N+1)}\left[\frac{4\|A\|^2\|y_*-y_0\|^2}{\alpha_X\mu_h}+\frac{12M^2N}{\alpha_X\mu_h}\right]+\bar{\epsilon} \\
 & \leq & \frac{8\|A\|^2\|y_0-y_*\|^2}{\alpha_X\mu_h(N+1)N}+\frac{24M^2}{\alpha_X\mu_h(N+1)}+\bar{\epsilon}, \\
\mathbb{E}[\text{gap}_\delta(\bar{z}_N)] & \leq & \frac{8\|A\|^2\|y_0\|^2}{\alpha_X\mu_h(N+1)N}+\frac{24M^2}{\alpha_X\mu_h(N+1)}+\bar{\epsilon}, \\
\mathbb{E}[\|\delta\|] & \leq & \frac{8\|A\|^2}{\alpha_X\mu_h N(N+1)}\left[2\|y_*-y_0\|+\sqrt{\frac{\alpha_X\mu_h}{2\|A\|^2}\left(\frac{6M^2}{\alpha_X}2N+\frac{N(N+1)}{2}\bar{\epsilon}\right)}\right] \\
 & \leq & \frac{16\|A\|^2\|y_*-y_0\|}{N(N+1)\alpha_X\mu_h}+\frac{8\sqrt{6}\|A\|M}{\alpha_X\mu_h N^{3/2}}+\frac{4\|A\|\sqrt{\bar{\epsilon}}}{(N+1)\sqrt{\alpha_X\mu_h}}, \\
\mathbb{E}[\|y_*-\bar{y}_N\|^2] & \leq & \|y_*-y_0\|^2+\frac{2}{N(N+1)}\sum_{k=1}^N\frac{k\alpha_X\mu_h}{2\|A\|^2}\left(2N\frac{12M^2}{\mu_h}+\frac{N(N+1)}{2}\bar{\epsilon}\right) \\
 & = & \|y_*-y_0\|^2+\frac{12M^2\alpha_X N}{\|A\|^2}+\frac{N(N+1)\alpha_X\mu_h}{2\|A\|^2}\bar{\epsilon}.
\end{array}
$$

■

Corollary 16 *If*

$$w_k=k,\ \tau_k=\tfrac{k-1}{2}\mu_h \ \ \text{and} \ \ \eta_k=\tfrac{4\|A\|^2N}{k\alpha_X\mu_h}, \tag{3.13}$$

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

$$\mathbb{E}[\text{gap}_*(\bar{z}_N)]\leq\tfrac{8\|A\|^2\|y_0-y_*\|^2+24M^2}{\alpha_X\mu_h(N+1)}+\bar{\epsilon}, \tag{3.14}$$

$$\mathbb{E}[\text{gap}_\delta(\bar{z}_N)]\leq\tfrac{8\|A\|^2\|y_0\|^2+24M^2}{\alpha_X\mu_h(N+1)}+\bar{\epsilon}, \tag{3.15}$$

$$\mathbb{E}[\|\delta\|]\leq\tfrac{16\|A\|^2\|y_*-y_0\|}{(N+1)\alpha_X\mu_h+16\sqrt{3}\|A\|M}+\tfrac{4\|A\|\sqrt{\bar{\epsilon}}}{\sqrt{(N+1)\alpha_X\mu_h}}, \tag{3.16}$$

$$\mathbb{E}[\|y_*-\bar{y}_N\|^2]\leq\|y_*-y_0\|^2+\tfrac{24M^2\alpha_X}{\|A\|^2}+\tfrac{(N+1)\alpha_X\mu_h}{2\|A\|^2}\bar{\epsilon}. \tag{3.17}$$

Proof. The proofs of (3.14)-(3.17) are similar to Corollary 15 and hence the details are skipped. ■

3.2 Convergence analysis for DSA under strong convexity

Our goal in this subsection is to establish the complexity of the DSA algorithm for solving problem 2.4 under the strong convex assumption about h^i, $i=1,2,3$, i.e., $\exists\mu_i>0$ s.t.

$$h^i(x_1,c)-h^i(x_2,c)-\langle (h^i)'(x_2,c),x_1-x_2\rangle\geq\mu_i P_{X^i}(x_2,x_1),\ \forall x_1,x_2\in X^i. \tag{3.18}$$

We describe some convergence properties for the innermost and middle loop of the DSA algorithm under the strong convexity assumptions in (3.18) in Lemma 17 and 18, respectively. The proofs for these results are similar to those for Lemma 10 and 11.

Lemma 17 below describes the convergence properties for the innermost loop of the DSA algorithm.

Lemma 17 *If the parameters* $\{w_k^3\}$, $\{\tau_k^3\}$ *and* $\{\eta_k^3\}$ *are set to* (3.8) *(with* $M=0$ *and* $A=A_j^3$*) and*

$$N_3\equiv N_{3,j}:=\tfrac{2\sqrt{6}\|A_j^3\|\|y_{*,j}^3-y_0^3\|}{\sqrt{\alpha_3\mu_3\epsilon}}, \tag{3.19}$$

then $B_j^3\bar{y}_j^3$ *is a stochastic* $(\epsilon/3)$*-subgradient of the value function* v^3 *at* x_{j-1}^2*. Moreover, there exists a constant* $M_3\geq 0$ *such that* $\|v^3(x_1)-v^3(x_2)\|\leq M_3\|x_1-x_2\|,\forall x_1,x_2\in X^3$ *and*

$$\mathbb{E}[\|B_j^3\bar{y}_j^3\|_*^2]\leq M_3. \tag{3.20}$$

Proof. In view of Corollary 15 (with $M = 0$ and $\bar\epsilon = 0$) and the definition of N_3 in (3.19), we have

$$\text{gap}_*(\bar z_j^3) \le \tfrac{8\|A_j^3\|^2\|y_0^3 - y_*^3\|^2}{\alpha_3\mu_3(N_3+1)N_3} \le \tfrac{\epsilon}{3}.$$

This observation, in view of Lemma 1, then implies that $B_j^3\bar y_j^3$ is a stochastic $(\epsilon/3)$-subgradient of v^3 at x_{j-1}^2. Moreover, it follows from (3.12) (with $M = 0$ and $\bar\epsilon = 0$) that $\|y_{*,j}^3 - \bar y_j^3\| \le \|y_{*,j}^3 - y_0^3\|$. This inequality, in view of the selection of N_3 in (3.19), the assumption that $y_{*,j}^3$ is well-defined, and (2.54), then implies the latter part of our result. ∎

Lemma 17 below describes the convergence properties for the middle loop of the DSA algorithm.

Lemma 18 *Assume that the parameters for the innermost loop are set according to Lemma 17. If the parameters $\{w_j^2\}$, $\{\tau_j^2\}$ and $\{\eta_j^2\}$ for the middle loop are set to* (3.13) *(with $M = M_3$ and $A = A_i^2$) and*

$$N_2 \equiv N_{2,i} := \tfrac{24\|A_i^2\|^2\|y_0^2 - y_{*,i}^2\|^2 + 72M_3^2}{\alpha_2\mu_2\epsilon}, \tag{3.21}$$

then $B_i^2\bar y_i^2$ is a stochastic $(2\epsilon/3)$-subgradient of the value function v^2 at x_{i-1}^1. Moreover, there exists a constant $M_2 \ge 0$ such that $\|v^2(x_1) - v^2(x_2)\| \le M_2\|x_1 - x_2\|, \forall x_1, x_2 \in X^2$ and

$$\mathbb{E}[\|B_i^2\bar y_i^2\|_*^2] \le M_2. \tag{3.22}$$

Proof. By Lemma 17, the stochastic subgradients of v^3 are computed by the innermost loop with tolerance $\bar\epsilon = \epsilon/3$. In view of Corollary 16 (with $M = M_3$ and $\bar\epsilon = \epsilon/3$) and the definition of N_2 in (3.21), we have

$$\text{gap}_*(\bar z_i^2) \le \tfrac{8\|A_i^2\|^2\|y_0^2 - y_{*,i}^2\|^2 + 24M_3^2}{\alpha_2\mu_2(N_2+1)} + \bar\epsilon \le \tfrac{2\epsilon}{3}.$$

This observation, in view of Lemma 1, then implies that $B_i^2\bar y_i^2$ is a stochastic $(2\epsilon/3)$-subgradient v^3 at x_{j-1}^2. Moreover, it follows from (3.17) (with $M = M_3$ and $\bar\epsilon = \epsilon/3$) that

$$\|y_{*,i}^2 - \bar y_i^2\|^2 \le \|y_{*,i}^2 - y_0^2\|^2 + \tfrac{24M_3^2\alpha_2}{\|A_i^2\|^2} + \tfrac{(N_2+1)\alpha_2\mu_2}{6\|A_i^2\|^2}\epsilon.$$

This inequality, in view of the selection of N_2 in (3.21), the assumption that $y_{*,i}^2$ is well-defined, and (2.54), then implies the latter part of our result. ∎

We are now ready to state the main convergence properties of the DSA algorithm for solving strongly convex three-stage stochastic optimization problems.

Theorem 19 *Suppose that the parameters for the innermost and middle loop in the DSA algorithm are set according to Lemma 17 and Lemma 18, respectively. If the parameters $\{w_i\}$, $\{\tau_i\}$ and $\{\eta_i\}$ for the outer loop are set to (3.8) (with $M = M_2$ and $A = A^1$) and*

$$N_1 := \max\left\{ \tfrac{4\sqrt{3}\|A^1\|\|y_0^1\|}{\sqrt{\alpha_1\mu_1\epsilon}} + \tfrac{4(6M_2)^2}{\alpha_1\mu_1\epsilon}, \; \tfrac{4\sqrt{3}\|A^1\|(\sqrt{\|y_*^1 - y_0^1\|} + \sqrt{2})}{\sqrt{\alpha_1\mu_1\epsilon}} + \left(\tfrac{24\sqrt{6}\|A^1\|M_2}{\alpha_1\mu_1\epsilon}\right)^{2/3} \right\}, \tag{3.23}$$

then we will find a solution $\bar x^1 \in X^1$ and a vector $\delta \in \mathbb{R}^{m^1}$ s.t.

$$\begin{aligned} &\mathbb{E}[h(\bar x^1, c) + v^2(\bar x^1) - (h(x^*, c) + v^2(x^*))] \le \epsilon, \\ &A\bar x^1 - b - \delta \in K^1, a.s., \\ &\mathbb{E}[\|\delta\|] \le \epsilon, \end{aligned}$$

where x^* denotes the optimal solution of problem 2.4.

Proof. By Lemma 18, the stochastic subgradients of v^2 are computed by the middle loop with tolerance $\bar{\epsilon} = 2\epsilon/3$. In view of Corollary 15 (with $M = M_2$ and $\bar{\epsilon} = 2\epsilon/3$) and the definition of N_1 in (3.23), we conclude that there exist $\delta \in \mathbb{R}^{m^1}$ s.t.

$$\begin{aligned} \mathbb{E}[\mathrm{gap}_\delta(\bar{z}^1_N)] &\le \tfrac{8\|A^1\|^2\|y^1_0\|^2}{\alpha_1\mu_1(N_1+1)N_1} + \tfrac{24M_2^2}{\alpha_1\mu_1(N_1+1)} + \tfrac{2\epsilon}{3} \le \epsilon, \\ \mathbb{E}[\|\delta\|] &\le \tfrac{16\|A^1\|^2\|y^1_*-y^1_0\|}{N_1(N_1+1)\alpha_1\mu_1} + \tfrac{8\sqrt{6}\|A^1\|M_2}{\alpha_1\mu_1 N_1^{3/2}} + \tfrac{4\|A^1\|\sqrt{2\epsilon}}{(N_1+1)\sqrt{3\alpha_1\mu_1}} \le \epsilon, \end{aligned}$$

which together with Lemma 8 then imply our result. ∎

In view of Lemma 18 and Theorem 19, the number of random samples ξ_2 and ξ_3 will be bounded by N_1 and $N_1 \times N_2$, i.e., $\mathcal{O}(1/\epsilon)$ and $\mathcal{O}(1/\epsilon^2)$, respectively.

4 DSA for general multi-stage stochastic optimization

In this section, we consider a multi-stage stochastic optimization problem given by

$$\begin{aligned} \min \;& \left\{h^1(x^1, c^1) + v^2(x^1)\right\} \\ \text{s.t.} \;& A^1x^1 - b^1 \in K^1, \\ & x^1 \in X^1, \end{aligned} \tag{4.1}$$

where the value factions v^t, $t = 2, \ldots, T$, are recursively defined by

$$\begin{aligned} v^t(x^{t-1}) &:= F^{t-1}(x^{t-1}, p^{t-1}) + \mathbb{E}_{\xi^t}[V^t(x^{t-1}, \xi^t)], \quad t = 2, \ldots, T-1, \\ V^t(x^{t-1}, \xi^t) &:= \min \left\{h^t(x^t, c^t) + v^{t+1}(x^t)\right\} \\ &\qquad \text{s.t. } A^tx^t - b^t - B^tx^{t-1} \in K^t, \\ &\qquad\quad\; x^t \in X^t, \end{aligned} \tag{4.2}$$

and

$$\begin{aligned} v^T(x^{T-1}) &:= \mathbb{E}_{\xi^T}[V^T(x^{T-1}, \xi^T)], \\ V^T(x^{T-1}, \xi^T) &:= \min \; h^T(x^T, c^T) \\ &\qquad \text{s.t. } A^Tx^T - b^T - B^Tx^{T-1} \in K^T, \\ &\qquad\quad\; x^T \in X^T. \end{aligned} \tag{4.3}$$

Here $\xi^t := (A^t, b^t, B^t, c^t, p^t)$ are random variables, $h_t(\cdot, c^t)$ are relatively simple functions, $F_t(\cdot, p^t)$ are general (not necessarily simple) Lipschitz continuous convex functions and K^t are convex cones, $\forall t = 1, \ldots, T$. We also assume that one can compute the subgradient $F'(x^t, p^t)$ of function $F^t(x^t, p^t)$ at any point $x^t \in X^t$ for a given parameter p^t.

Problem (4.1) is more general than problem (2.1) (or equivalently problem (2.4)) in the following sense. First, we are dealing with a more complicated multi-stage stochastic optimization problem where the number of stages T (4.1) can be greater than three. Second, the value function $v^t(x^{t-1})$ in (4.2) is defined as the summation of $F^{t-1}(x^{t-1}, p^{t-1})$ and $\mathbb{E}_{\xi^t}[V^t(x^{t-1}, \xi^t)]$, where F^{t-1} is not necessarily simple. We intend to generalize the DSA algorithm in Sections 2 and 3 for solving problem (4.1). More specifically, we show how to compute a stochastic ϵ-subgradient of v^{t+1} at x^t, $t = 1, \ldots, T-2$, in a recursive manner until we obtain the ϵ-subgradient of v^T at x^{T-1}.

We are now ready to formally state the DSA algorithm for solving the multi-stage stochastic optimization problem in (4.1). Observe that the following notations will be used in the algorithm:

- N_t is the number of iterations for stage t subproblem and k_t is the corresponding index, i.e., $k_t = 1, \ldots, N_t$.
- $\xi^t_{k_{t-1}} = (A^t_{k_{t-1}}, b^t_{k_{t-1}}, B^t_{k_{t-1}}, c^t_{k_{t-1}}, p^t_{k_{t-1}})$ is the k_{t-1} th random scenarios in stage t subproblem, $(x^t_{k_t}, y^t_{k_t})$ are the k_t th iterates in stage t subproblem.

- For simplicity, we denote $\xi^t_{k_{t-1}}$ as ξ^t_k, $(x^t_{k_t}, y^t_{k_t})$ as (x^t_k, y^t_k).

Algorithm 3 DSA for multi-stage stochastic programs

Input: initial points $\{x^t_0\}$, $k_t = 1, \forall t$, iteration number N_t and stepsize strategy $\{w_k\}$.
Start with procedure DSA$(1, 0)$.
procedure: DSA(t, u)
for $k_t = 1, \ldots, N_t$ **do**
  if $t < T$ **then**
    Generate random scenarios ξ^{t+1}_k.
    $(\bar{x}^{t+1}, \bar{y}^{t+1}) = \text{DSA}(t+1, x^t_k)$ and $G(x^t_{k-1}, \xi^{t+1}_k) = (B^{t+1}_k)^T \bar{y}^{t+1}$.
  else
    $G(x^T_{k-1}, \xi^{T+1}_k) = 0$.
  end if
  $(x^t_k, y^t_k) = \text{SPDT}(x^t_{k-1}, y^t_{k-1}, y^t_{k-2}, G(x^t_{k-1}, \xi^{t+1}_k), u, \xi^t_{k-1}, h^t, X^t, K^t_*, \theta^t_k, \tau^t_k, \eta^t_k)$.
end for
return: $\bar{z}^t = \sum_{k=1}^{N_t} w_k z^t_k / \sum_{k=1}^{N_t} w_k$.

In order to show the convergence of the above DSA algorithm, we need the following assumption on the boundedness of the operators B^t:

$$\|B^t\| \le \mathcal{B}_t, \ \forall t = 2, \cdots, T. \tag{4.4}$$

Lemma 20 below establishes some convergence properties of the DSA algorithm for solving the last stage problem.

Lemma 20 *Suppose that the algorithmic parameters in the DSA algorithm applied to problem 4.1 are chosen as follows.*

a) *For a general convex problem,* $\{w^T_k\}$, $\{\tau^T_k\}$ *and* $\{\eta^T_k\}$ *are set to* (2.41) *(with* $M = 0$ *and* $A = A^T_k$*) and*

$$N_T \equiv N_{T,k} := \tfrac{T\sqrt{2}\|A^T_k\|[2(\Omega_T)^2 + \|y^T_{*,k} - y^T_0\|^2]}{\sqrt{\alpha_T}\epsilon}. \tag{4.5}$$

b) *Under the strongly convex assumption* (3.18), $\{w^T_k\}$, $\{\tau^T_k\}$ *and* $\{\eta^T_k\}$ *are set to* (3.8) *(with* $M = 0$ *and* $A = A^T_k$*) and*

$$N_T \equiv N_{T,k} := \tfrac{\sqrt{8T}\|A^T_k\|\|y^T_{*,k} - y^T_0\|}{\sqrt{\alpha_T \mu_T \epsilon}}. \tag{4.6}$$

Then $B^T_k \bar{y}^T_k$ *is a stochastic* (ϵ/T)*-subgradient of the value function* v^T *at* x^{T-1}_{k-1}*. Moreover, there exists a constant* $M_T \ge 0$ *such that* $\|v^T(x_1) - v^T(x_2)\| \le M_T\|x_1 - x_2\|, \forall x_1, x_2 \in X^T$ *and*

$$\mathbb{E}[\|B^T_k \bar{y}^T_k\|^2_*] \le M_T. \tag{4.7}$$

Proof. The innermost loop of the DSA algorithm is equivalent to the application of Algorithm 2 to the last stage saddle point problem in (2.7). Note that for this problem, we do not have any subsequent stages and hence $\tilde{v} = 0$. In other words, the subgradients of $\tilde{v}$ are exact. To show part a), in view of Corollary 6 (with $M = 0$ and $\bar{\epsilon} = 0$) and the definition of N_T in (4.5), we have

$$\text{gap}_*(\bar{z}^T_k) \le \tfrac{\sqrt{2}\|A^T_k\|[2(\Omega_T)^2 + \|y^T_* - y^T_0\|^2]}{\sqrt{\alpha_T} N_T} \le \tfrac{\epsilon}{T}.$$

This observation, in view of Lemma 1, then implies that $B^T_k \bar{y}^T_k$ is a stochastic (ϵ/T)-subgradient of v^T at x^{T-1}_{j-1}. Moreover, it follows from (2.45) (with $M = 0$ and $\bar{\epsilon} = 0$) that

$$\|y^T_{*,k} - \bar{y}^T_k\|^2 \le \|y^T_{*,k} - y^T_0\|^2 + 4(\Omega_T)^2 + \tfrac{(N_T+1)\epsilon}{2}.$$

This inequality, in view of the selection of N_T in (4.5), the assumption that $y^T_{*,k}$ is well-defined, and (4.4), then implies the latter part of our result. Similarly, the result in (4.6) follows from Corollary 15 (with $M = 0$ and $\bar\epsilon = 0$) and the definition of N_T in (4.6). ■

We show in Lemma 21 some convergence properties of the middle loops of the DSA algorithm.

Lemma 21 *Assume that the parameters for the innermost loop are set according to Lemma 20. Moreover, suppose that the algorithmic parameters for the middle loops are chosen as follows.*

a) *For general convex problem, the parameters* $\{w^t_k\}$, $\{\tau^t_k\}$ *and* $\{\eta^t_k\}$ *for the middle loops* ($t = 2, \ldots, T-1$) *are set to* (2.46) *(with* $M = M_{t+1}$ *and* $A = A^t_k$*) and*

$$N_t \equiv N_{t,k} := \left(\tfrac{4\sqrt{2}T\|A^t_k\|\Omega_t}{\sqrt{\alpha_t}\epsilon}\right)^{\frac{2}{3}} + \left[\tfrac{2T\left(\|A^t_k\|\|y^t_{*,k}-y^t_0\|^2+4\sqrt{3}M_{t+1}\Omega_t\right)}{\sqrt{\alpha_t}\epsilon}\right]^2. \tag{4.8}$$

b) *Under strongly convex assumption* (3.18)*, the parameters* $\{w^t_k\}$, $\{\tau^t_k\}$ *and* $\{\eta^t_k\}$ *for the middle loops* ($t = 2, \ldots, T-1$) *are set to* (3.13) *(with* $M = M_{t+1}$ *and* $A = A^t_k$*) and*

$$N_t \equiv N_{t,k} := \tfrac{8T\|A^t_k\|^2\|y^t_0-y^t_{*,k}\|^2+24TM^2_{t+1}}{\alpha_t\mu_t\epsilon}. \tag{4.9}$$

Then $B^t_k\bar{y}^t_k$ *is a stochastic* $((T+1-t)\epsilon/T)$*-subgradient of the value function* v^t *at* x^{t-1}_{k-1}*. Moreover, there exists a constant* $M_t \geq 0$ *such that* $\|v^t(x_1) - v^t(x_2)\| \leq M_t\|x_1 - x_2\|, \forall x_1, x_2 \in X^t$ *and*

$$\mathbb{E}[\|B^t_k\bar{y}^t_k\|^2_*] \leq M_t. \tag{4.10}$$

Proof. The middle loops ($t = 2, \ldots, T-1$) of the DSA algorithm applied to multistage stochastic optimization is equivalent to the application of Algorithm 2 to the second stage saddle point problem in (2.6). Note that for this problem, we have $\tilde v = v^{t+1}$. Moreover, by Lemma 20, the stochastic subgradients of v^T are computed by the innermost loop with tolerance $\bar\epsilon = \epsilon/T$. To show part a), in view of Corollary 7 (with $M = M_{t+1}$ and $\bar\epsilon = (T-t)\epsilon/T$) and the definition of N_t in (4.8), we have

$$\text{gap}_*(\bar z^t_k) \leq \tfrac{2\sqrt{2}\|A^t_k\|\Omega_t}{N_t\sqrt{\alpha_t N_t}} + \tfrac{\|A^t_k\|\|y^t_{*,k}-y^t_0\|^2+4\sqrt{3}M_{t+1}\Omega_t}{\sqrt{\alpha_t N_t}} + \bar\epsilon \leq \tfrac{(T+1-t)\epsilon}{T}.$$

This observation, in view of Lemma 1, then implies that $B^t_k\bar y^t_k$ is a stochastic $((T+1-t)\epsilon/T)$-subgradient v^t at x^{t-1}_{k-1}. Moreover, it follows from (2.50) (with $M = M_{t+1}$ and $\bar\epsilon = (T-t)\epsilon/T$) that

$$\|y^t_{*,k} - \bar y^t_k\|^2 \leq \|y^t_{*,k} - y^t_0\|^2 + \tfrac{2\Omega^2_t}{N_t} + \tfrac{\sqrt{6}(1+\alpha_t)M_{t+1}\Omega_t}{\|A^t_k\|} + \tfrac{\sqrt{\alpha_t N_t}\epsilon}{3\sqrt{2}\|A^t_k\|}.$$

This inequality, in view of the selection of N_t in (4.8), the assumption that $y^t_{*,k}$ is well-defined, and (4.4), then implies the latter part of our result. Similarly, in view of Corollary 16, we have part b). ■

We are now ready to establish the main convergence properties of the DSA algorithm for solving general multi-stage stochastic optimization problems with $T \geq 3$.

Theorem 22 *Suppose that the parameters for the inner loops in the DSA algorithm are set according to Lemma 20 and Lemma 21. Moreover, assume that the algorithmic parameters in the outer loop of the DSA algorithm are chosen as follows.*

a) *For general convex problem, the parameters* $\{w_k\}$, $\{\tau_k\}$ *and* $\{\eta_k\}$ *for the outer loop are set to (2.41) (with* $M = M_2$ *and* $A = A^1$*) and*

$$N_1 := \max\left\{\tfrac{2\sqrt{2}T\|A^1\|[2(\Omega_1)^2+\|y^1_0\|^2]}{\sqrt{\alpha_1}\epsilon} + \left(\tfrac{8\sqrt{3}TM_2\Omega_1}{\sqrt{\alpha_1}\epsilon}\right)^2, \right.$$
$$\left. \tfrac{6T\|A^1\|(\sqrt{2\alpha_1}\|y^1_*-y^1_0\|+2\Omega_1)+27(T-1)\sqrt{\alpha_1}\|A^1\|}{\alpha_1 T\epsilon} + \left(\tfrac{6\sqrt{3}M_2(\sqrt{2}\|A^1\|+\sqrt{\alpha_1})}{\alpha_1\epsilon}\right)^2\right\}. \tag{4.11}$$

b) *Under strongly convex assumption* (3.18)*, the parameters* $\{w_k\}$, $\{\tau_k\}$ *and* $\{\eta_k\}$ *for the outer loop are set to (3.8) (with* $M = M_2$ *and* $A = A^1$*) and*

$$N_1 := \max\Big\{\tfrac{4\sqrt{T}\|A^1\|\|y_0^1\|}{\sqrt{\alpha_1\mu_1\epsilon}} + \tfrac{24TM_2^2}{\alpha_1\mu_1\epsilon}, \\ \tfrac{4\sqrt{3}\|A^1\|\sqrt{\|y_*^1-y_0^1\|}}{\sqrt{\alpha_1\mu_1\epsilon}} + \Big(\tfrac{24\sqrt{6}\|A^1\|M_2}{\alpha_1\mu_1\epsilon}\Big)^{2/3} + \tfrac{12\|A^1\|\sqrt{T-1}}{\sqrt{\alpha_1\mu_1T\epsilon}}\Big\}. \tag{4.12}$$

Then we will find a solution $\bar{x}^1 \in X^1$ *and a vector* $\delta \in \mathbb{R}^{m^1}$ *s.t.*

$$\begin{aligned}&\mathbb{E}[h(\bar{x}^1, c) + v^2(\bar{x}^1) - (h(x^*, c) + v^2(x^*))] \le \epsilon,\\ &A\bar{x}^1 - b - \delta \in K^1, a.s.,\\ &\mathbb{E}[\|\delta\|] \le \epsilon,\end{aligned}$$

where x^* *denotes the optimal solution of problem 2.4.*

Proof. The outer loop of the DSA algorithm is equivalent to the application of Algorithm 2 to the first stage saddle point problem in (2.10). Note that for this problem, we have $\tilde{v} = v^2$. Moreover, by Lemma 21, the stochastic subgradients of v^2 are computed by the middle loop with tolerance $\bar{\epsilon} = (T-1)\epsilon/T$. To show part a), in view of Corollary 6 (with $M = M_2$ and $\bar{\epsilon} = (T-1)\epsilon/T$) and the definition of N_1 in (4.11), we conclude that there exist $\delta \in \mathbb{R}^{m^1}$ s.t.

$$\begin{aligned}\mathbb{E}[\text{gap}_\delta(\bar{z}_N^1)] &\le \tfrac{\sqrt{2}\|A^1\|(2\Omega_1^2+\|y_0^1\|^2)}{\sqrt{\alpha_1}N_1} + \tfrac{4\sqrt{3}M_2\Omega_1}{\sqrt{\alpha_1N_1}} + \tfrac{(T-1)\epsilon}{T} \le \epsilon,\\ \mathbb{E}[\|\delta\|] &\le \tfrac{2\sqrt{2\alpha_1}\|A^1\|\|y_*^1-y_0^1\|+4\Omega_1\|A^1\|}{\alpha_1N_1} + \tfrac{2M_2(\sqrt{6}\|A^1\|+\sqrt{3\alpha_1})}{\alpha_1\sqrt{N_1}} + \sqrt{\tfrac{3\|A^1\|(T-1)\epsilon}{N_1T\sqrt{\alpha_1}}} \le \epsilon,\end{aligned}$$

which together with Lemma 8 then imply our result. Similarly, in view of Corollary 15, we have part b). ∎

In view of the results stated in Lemma 20, Lemma 21 and Theorem 22, the total number of scenarios required to find an ϵ-solution of (4.1) is given by $N_2 \times N_3 \times \dots N_T$, and hence will grow exponentially with respect to T, no matter the objective functions are strongly convex or not. These sampling complexity bounds match well with those in [30, 31], implying that multi-stage stochastic optimization problems are essentially intractable for $T \ge 5$ and a moderate target accuracy. Hence, it is reasonable to use the DSA algorithm only for multi-stage stochastic optimization problems with T relatively small and ϵ relatively large. However, it is interesting to point out that the DSA algorithm only needs to go through the scenario tree once and hence its memory requirement increases only linearly with respect to T. Moreover, the development of the complexity bounds of multi-stage stochastic optimization in terms of their dependence on various problem parameters may help us to further explore the structure of the problems and to identify special classes of problems possibly admitting faster solution methods.

5 Conclusion

In this paper, we present a new class of stochastic approximation algorithms, i.e., dynamic stochastic approximation (DSA), for solving multi-stage stochastic optimization problems. This algorithm is developed by reformulating the optimization problem in each stage as a saddle point problem and then recursively applying an inexact primal-dual stochastic approximation algorithm to compute an approximate stochastic subgradient of the previous stage. We establish the convergence of this algorithm by carefully bounding the bias and variance associated with these approximation errors. For a three-stage stochastic optimization problem, we show that the total number of required scenarios to find an ϵ-solution is bounded by $\mathcal{O}(1/\epsilon^4)$ and $\mathcal{O}(1/\epsilon^2)$, respectively, for general convex and strongly convex cases. These bounds are essentially not improvable in terms of their dependence on the target accuracy. We also generalize DSA for solving multi-stage stochastic optimization problems with the number of stages $T > 3$. To the best of our knowledge, this is the first time that stochastic approximation methods have been developed and their complexity is established for multi-stage stochastic optimization.

Observe that this paper focuses on theoretical analysis of the DSA method. The practical performance of this method will depend on the estimation of problem parameters especially those related to the size of subgradients and dual multipliers. It would be interesting to study whether one can estimate these parameters in an online fashion while running these methods, and whether one can further improve the convergence of DSA in terms of its dependence on these problem parameters (e.g., by using accelerated SA methods).

Reference

[1] K. Arrow, L. Hurwicz, and H. Uzawa. *Studies in Linear and Non-linear Programming*. Stanford Mathematical Studies in the Social Sciences. Stanford University Press, 1958.

[2] J.R. Birge and F.V. Louveaux. *Introduction to Stochastic Programming*. Springer, New York, 1997.

[3] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *J. Math. Imaging Vision*, 40:120–145, 2011.

[4] Y. Chen, G. Lan, and Y. Ouyang. Optimal primal-dual methods for a class of saddle point problems. *SIAM Journal on Optimization*, 24(4):1779–1814, 2014.

[5] Christopher J Donohue and John R Birge. The abridged nested decomposition method for multistage stochastic linear programs with relatively complete recourse. *Algorithmic Operations Research*, 1(1), 2006.

[6] S. Ghadimi and G. Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, I: a generic algorithmic framework. *SIAM Journal on Optimization*, 22:1469–1492, 2012.

[7] S. Ghadimi and G. Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, II: shrinking procedures and optimal algorithms. *SIAM Journal on Optimization*, 23:2061–2089, 2013.

[8] S. Ghadimi and G. Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

[9] S. Ghadimi and G. Lan. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Mathematical Programming*, 2015. to appear.

[10] S. Ghadimi, G. Lan, and H. Zhang. Mini-batch stochastic approximation methods for constrained nonconvex stochastic programming. *Mathematical Programming*, 2014. to appear.

[11] Bingsheng He and Xiaoming Yuan. On the o(1/n) convergence rate of the douglas-rachford alternating direction method. *SIAM Journal on Numerical Analysis*, 50(2):700–709, 2012.

[12] Magnus Hindsberger and AB Philpott. Resa: A method for solving multistage stochastic linear programs. *Journal of Applied Operational Research*, 6(1):2–15, 2014.

[13] G. Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1):365–397, 2012.

[14] G. Lan, A. S. Nemirovski, and A. Shapiro. Validation analysis of mirror descent stochastic approximation method. *Mathematical Programming*, 134:425–458, 2012.

[15] Guanghui Lan and Yi Zhou. An optimal randomized incremental gradient method. *Mathematical programming*.

[16] R.D.C. Monteiro and B.F. Svaiter. On the complexity of the hybrid proximal projection method for the iterates and the ergodic mean. *SIAM Journal on Optimization*, 20:2755–2787, 2010.

[17] A. Nedić. On stochastic subgradient mirror-descent algorithm with weighted averaging. 2012.

[18] A. S. 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, 2005.

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

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

[21] 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.

[22] Y. E. Nesterov. Smooth minimization of nonsmooth functions. *Mathematical Programming*, 103:127–152, 2005.

[23] Y. Ouyang, Y. Chen, G. Lan, and E. Pasiliao. An accelerated linearized alternating direction method of multipliers. *SIAM Journal on Imaging Sciences*, 2014. to appear.

[24] Mario VF Pereira and Leontina MVG Pinto. Multi-stage stochastic optimization applied to energy planning. *Mathematical programming*, 52(1-3):359–375, 1991.

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

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

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

[28] R Tyrrell Rockafellar and Roger J-B Wets. Scenarios and policy aggregation in optimization under uncertainty. *Mathematics of operations research*, 16(1):119–147, 1991.

[29] Mark Rudelson and Roman Vershynin. Non-asymptotic theory of random matrices: extreme singular values, 2010.

[30] A. Shapiro and A. Nemirovski. On complexity of stochastic programming problems. *E-print available at: http://www.optimization-online.org*, 2004.

[31] A. Shaprio. On complexity of multistage stochastic programs. *Operations Research Letters*, 34:1–8, 2006.

[32] A. Shaprio. Analysis of stochastic dual dynamic programming method. *European Journal of Operational Research*, 209:63–72, 2011.