ALGORITHMS FOR STOCHASTIC OPTIMIZATION WITH EXPECTATION CONSTRAINTS *

GUANGHUI LAN † AND ZHIQIANG ZHOU ‡

Abstract. This paper considers the problem of minimizing an expectation function over a closed convex set, coupled with an expectation constraint on either decision variables or problem parameters. We first present a new stochastic approximation (SA) type algorithm, namely the cooperative SA (CSA), to handle problems with the expectation constraint on devision variables. We show that this algorithm exhibits the optimal $\mathcal{O}(1/\sqrt{N})$ rate of convergence, in terms of both optimality gap and constraint violation, when the objective and constraint functions are generally convex, where N denotes the number of iterations. Moreover, we show that this rate of convergence can be improved to $\mathcal{O}(1/N)$ if the objective and constraint functions are strongly convex. We then present a variant of CSA, namely the cooperative stochastic parameter approximation (CSPA) algorithm, to deal with the situation when the expectation constraint is defined over problem parameters and show that it exhibits similar optimal rate of convergence to CSA. It is worth noting that CSA and CSPA are primal methods which do not require the iterations on the dual space and/or the estimation on the size of the dual variables. To the best of our knowledge, this is the first time that such optimal SA methods for solving expectation constrained stochastic optimization are presented in the literature.

Keywords: convex programming, stochastic optimization, complexity, subgradient method

AMS 2000 subject classification: 90C25, 90C06, 90C22, 49M37

1. Introduction.

In this paper, we study two related stochastic programming (SP) problems with expectation constraints. The first one is a classical SP problem with the expectation constraint over the decision variables, formally defined as

$$\begin{aligned} \min\ & f(x) := \mathbb{E}[F(x,\zeta)] \\ \text{s.t. } & g(x) := \mathbb{E}[G(x,\xi)] \le 0, \\ & x \in X, \end{aligned} \tag{1.1}$$

where $X \subseteq \mathbb{R}^n$ is a convex compact set, ζ and ξ are random vectors supported on $\mathcal{P} \subseteq \mathbb{R}^p$ and $\mathcal{Q} \subseteq \mathbb{R}^q$, respectively, $F(x,\zeta) : X \times \mathcal{P} \mapsto \mathbb{R}$ and $G(x,\xi) : X \times \Xi \mapsto \mathbb{R}$ are closed convex functions w.r.t. x for a.e. $\zeta \in \mathcal{P}$ and $\xi \in \mathcal{Q}$. Moreover, we assume that ζ and ξ are independent of x. Under these assumptions, (1.1) is a convex optimization problem.

Problem (1.1) has many applications in operations research, finance and data analysis. One motivating example is SP with the conditional value at risk (CVaR) constraint. In an important work [31], Rockafellar and Uryasev shows that a class of asset allocation problem can be modeled as

$$\begin{array}{ll} \min_{x,\tau} & -\mu^T x \\ \text{s.t.} & \tau + \frac{1}{\beta}\mathbb{E}\{[-\xi^T x - \tau]_+\} \le 0, \\ & \sum_{i=1}^n x_i = 1, x \ge 0, \end{array} \tag{1.2}$$

where ξ denotes the random return with mean $\mu = \mathbb{E}[\xi]$. Expectation constraints also play an important role in providing tight convex approximation to chance constrained problems (e.g., Nemirovksi and Shapiro [24]). Some other important applications of (1.1) can be found in semi-supervised learning (see, e.g., [6]). For example, one can use the objective function to define the fidelity of the model for the labelled data, while using the constraint to enforce some other properties of the model for the unlabelled data (e.g., proximity for data with similar features).

While problem (1.1) covers a wide class of problems with expectation constraints over the decision variables, in practice we often encounter the situation where the expectation constraint is defined over the problem

---

*The results of this paper were first presented at the Annual INFORMS meeting in Oct, 2015, https://informs.emeetingsonline.com/emeetings/formbuilder/clustersessiondtl.asp?csnno=24236&mmnno=272&ppnno=91687

†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).

parameters. Under these circumstances our goal is to find a pair of parameters x^* and decision variables $y^*(x^*)$ such that

$$y^*(x^*) \in \mathrm{Argmin}_{y \in Y} \{\phi(x^*, y) := \mathbb{E}[\Phi(x^*, y, \zeta)]\}, \tag{1.3}$$

$$x^* \in \{x \in X | g(x) := \mathbb{E}[G(x, \xi)] \leq 0\}. \tag{1.4}$$

Here $\Phi(x, y, \zeta)$ is convex w.r.t. y for a.e. $\zeta \in \mathcal{P}$ but possibly nonconvex w.r.t. (x, y) jointly, and $G(\cdot, \xi)$ is convex w.r.t. x for a.e. $\xi \in \mathcal{Q}$. Moreover, we assume that ζ and ξ are independent of x and y, respectively, while ζ is not necessarily independent of x^*. Note that (1.3)-(1.4) defines a pair of optimization and feasibility problems coupled through the following ways: a) the solution to (1.4) defines an admissible parameter of (1.3); b) the random variables ζ and ξ can be dependent, e.g., $\zeta = \xi$; and c) ζ can be a random variable with probability distribution parameterized by x^*.

Problem (1.3)-(1.4) also has many applications, especially in data analysis. One such example is to learn a classifier w with a certain metric $\bar{A}$ using the support vector machine model:

$$\min_{w} \mathbb{E}[l(w; (\bar{A}^{\frac{1}{2}} u, v))] + \frac{\lambda}{2}\|w\|^2, \tag{1.5}$$

$$\bar{A} \in \left\{A \succeq 0 | \mathbb{E}[|\mathrm{Tr}(A(u_i - v_j)(u_i - v_j)^T) - b_{ij}|] \leq 0, \mathrm{Tr}(A) \leq C\right\}, \tag{1.6}$$

where $l(w; (\theta, y)) = \max\{0, 1 - y\langle w, \theta\rangle\}$ denotes the hinge loss function, $u, u_i, u_j \in \mathbb{R}^n$, $v, v_i, v_j \in \{+1, -1\}$, and $b_{ij} \in \mathbb{R}$ are the random variables satisfying certain probability distributions, and $\lambda, C > 0$ are certain given parameters. In this problem, (1.5) is used to learn the classifer w by using the metric $\bar{A}$ satisfying certain requirements in (1.6), including the low rank (or nuclear norm) assumption. Problem (1.3)-(1.4) can also be used in some data-driven applications, where one can use (1.4) to specify the parameters for the probabilistic models associated with the random variable ζ, as well as some other applications for multi-objective stochastic optimization.

In spite of its wide applicability, the study on efficient solution methods for expectation constrained optimization is still limited. For the sake of simplicity, suppose for now that ζ is given as a deterministic vector and hence that the objective functions f and ϕ in (1.1) and (1.3) are easily computable. One popular method to solve stochastic optimization problems is called the sample average approximation (SAA) approach ([35, 18, 38]). To apply SAA for (1.1) and (1.4), we first generate a random sample $\xi_i, i = 1, \ldots, N$, for some $N \geq 1$ and then approximate g by $\tilde{g}(x) = \frac{1}{N}\sum_{i=1}^N G(x, \xi_i)$. The main issues associated with the SAA for solving (1.1) include: i) the deterministic SAA problem might not be feasible; ii) the resulting deterministic SAA problem is often difficult to solve especially when N is large, requiring going through the whole sample $\{\xi_1, \ldots, \xi_N\}$ at each iteration; and ii) it is not applicable to the on-line setting where one needs to update the decision variable whenever a new piece of sample ξ_i, $i = 1, \ldots N$, is collected.

A different approach to solve stochastic optimization problems is called stochastic approximation (SA), which was initially proposed in a seminal paper by Robbins and Monro[30] in 1951 for solving strongly convex SP problems. This algorithm mimics the gradient descent method by using the stochastic gradient $F'(x, \xi_i)$ rather than the original gradient $f'(x)$ for minimizing $f(x)$ in (1.1) over a simple convex set X (see also [4, 10, 11, 26, 32, 36]). An important improvement of this algorithm was developed by Polyak and Juditsky([28],[29]) through using longer steps and then averaging the obtained iterates. Their method was shown to be more robust with respect to the choice of stepsize than classic SA method for solving strongly convex SP problems. More recently, Nemirovski et al. [23] presented a modified SA method, namely, the mirror descent SA method, and demonstrated its superior numerical performance for solving a general class of nonsmooth convex SP problems. The SA algorithms have been intensively studied over the past few years (see, e.g., [19, 12, 13, 9, 39, 14, 22, 33]). It should be noted, however, that none of these SA algorithms are applicable to expectation constrained problems, since each iteration of these algorithms requires the projection over the feasible set $\{x \in X | g(x) \leq 0\}$, which is computationally prohibitive as g is given in the form of expectation.

In this paper, we intend to develop efficient solution methods for solving expectation constrained problems by properly addressing the aforementioned issues associated with existing SA methods. Our contribution mainly exists in the following several aspects. Firstly, inspired by Polayk's subgradient method for constrained optimization [27, 25], we present a new SA algorithm, namely the cooperative SA (CSA) method for solving the SP problem with expectation constraint in (1.1). At the k-th iteration, CSA performs a projected subgradient step along either $F'(x_k, \zeta_k)$ or $G'(x_k, \xi_k)$ over the set X, depending on whether the stochastic condition $G(x_k, \xi_k) \le \eta_k$ is satisfied or not. We introduce an index set $\mathcal{B} := \{1 \le k \le N : G(x_k, \xi_k) \le \eta_k\}$ in order to compute the output solution as a weighted average of the iterates in $\mathcal{B}$. By carefully bounding $|\mathcal{B}|$, we show that the number of iterations performed by the CSA algorithm to find an ϵ-solution of (1.1), i.e., a point $\bar{x} \in X$ s.t. $\mathbb{E}[f(\bar{x}) - f^*] \le \epsilon$ and $\mathbb{E}[g(\bar{x})] \le \epsilon$, can be bounded by $\mathcal{O}(1/\epsilon^2)$. Moreover, when both f and g are strongly convex, by using a different set of algorithmic parameters we show that the complexity of the CSA method can be significantly improved to $\mathcal{O}(1/\epsilon)$. It it is worth mentioning that this result is new even for solving deterministic strongly convex problems with functional constraints. We also established the large-deviation properties for the CSA method under certain light-tail assumptions.

Secondly, we develop a variant of CSA, namely the cooperative stochastic parameter approximation (C-SPA) method for solving the SP problem with expectation constraints on problem parameters in (1.3)-(1.4). At the k-the iteration, CSPA performs either a projected subgradient step along $\Phi'(x_k, y_k, \zeta_k)$ over the set Y, or along $G'(x_k, \xi_k)$ over the set X, depending on whether the stochastic condition $G(x_k, \xi_k) \le \eta_k$ is satisfied or not. We then define the output solution as a randomly selected iterate from the index set $\mathcal{B}$ (rather than the weighted average as in CSA) mainly due to the possible non-convexity of Φ w.r.t. (x, y). By carefully bounding $|\mathcal{B}|$, we show that the number of iterations performed by the CSPA algorithm to find an ϵ-solution of (1.3)-(1.4), i.e., a pair of solution $(\bar{x}, \bar{y})$ s.t. $\mathbb{E}[g(\bar{x})] \le \epsilon$ and $\mathbb{E}[\phi(\bar{x}, \bar{y}) - \phi(\bar{x}, y^*(\bar{x})] \le \epsilon$, can be bounded by $\mathcal{O}(1/\epsilon^2)$. Moreover, this bound can be significantly improved to $\mathcal{O}(1/\epsilon)$ if G and Φ are strongly convex w.r.t. x and y, respectively. We also derive the large-deviation properties for the CSPA method and design a two-phase procedure to improve such properties of CSPA under certain light-tail assumptions.

To the best of our knowledge, all the aforementioned algorithmic developments are new in the stochastic optimization literature. It is also worth mentioning a few alternative or related methods to solve (1.1) and (1.3)-(1.4). First, without efficient methods to directly solve (1.1), current practice resorts to reformulate it as $\min_{x \in X} \lambda f(x) + (1 - \lambda) g(x)$ for some $\lambda \in (0, 1)$. However, one then has to face the difficulty of properly specifying λ, since an optimal selection would depend on the unknown dual multiplier. As a consequence, we cannot assess the quality of the solutions obtained by solving this reformulated problem. Second, one alternative approach to solve (1.1) is the penalty-based or primal-dual approach. However these methods would require either the estimation of the optimal dual variables or iterations performed on the dual space (see [7], [23] and [20]). Moreover, the rate of convergence of these methods for functional constrained problems has not been well-understood other than conic constraints even for the deterministic setting. Third, in [17] (and see references therein), Jiang and Shanbhag developed a coupled SA method to solve a stochastic optimization problem with parameters given by another optimization problem, and hence is not applicable to problem (1.3)-(1.4). Moreover, each iteration of their method requires two stochastic subgradient projection steps and hence is more expensive than that of CSPA.

The remaining part of this paper is organized as follows. In Section 2, we present the CSA algorithm and establish its convergence properties under general convexity and strong convexity assumptions. Then in Section 3, we develop a variant of the CSA algorithm, namely the CSPA for solving SP problems with the expectation constraint over problem parameters and discuss its convergence properties. We then present some numerical results for these new SA methods in section 4. Finally some concluding remarks are added in Section 5.

2. Expectation constraints over decision variables. In this section we present the cooperative SA (CSA) algorithm for solving convex stochastic optimization problems with the expectation constraint over decision variables. More specifically, we first briefly review the distance generating function and prox-mapping in Subsection 2.1. We then describe the CSA algorithm and discuss its convergence properties for solving general convex problems in Subsection 2.2, and show how to improve the convergence of this algorithm by imposing strong convexity assumptions in Subsection 2.3. Finally, we establish some large deviation properties for the CSA method in Subsection 2.4.

2.1. Preliminary: prox-mapping. Recall that a function $\omega_X : X \mapsto R$ is a distance generating function with parameter α, if ω_X is continuously differentiable and strongly convex with parameter α with respect to $\|\cdot\|$. Without loss of generality, we assume throughout this paper that $\alpha = 1$, because we can always rescale $\omega_X(x)$ to $\bar{\omega}_X(x) = \omega_X(x)/\alpha$. Therefore, we have

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

The prox-function associated with ω is given by

$$V_X(z, x) = \omega_X(x) - \omega_X(z) - \langle\nabla\omega_X(z), x - z\rangle.$$

$V_X(\cdot,\cdot)$ is also called the Bregman's distance, which was initially studied by Bregman [5] and later by many others (see [1],[2] and [37]). In this paper we assume the prox-function $V_X(x, z)$ is chosen such that, for a given $x \in X$, the prox-mapping $P_{x,X} : \mathbb{R}^n \mapsto \mathbb{R}^n$ defined as

$$P_{x,X}(\phi) := \text{argmin}_{z\in X}\{\langle\phi, z\rangle + V_X(x, z)\} \tag{2.1}$$

is easily computed.

It can be seen from the strong convexity of $\omega(\cdot,\cdot)$ that

$$V_X(x, z) \geq \frac{1}{2}\|x - z\|^2, \forall x, z \in X. \tag{2.2}$$

Whenever the set X is bounded, the distance generating function ω_X also gives rise to the diameter of X that will be used frequently in our convergence analysis:

$$D_X \equiv D_{X,\omega_X} := \sqrt{\max_{x,z\in X} V_X(x, z)}. \tag{2.3}$$

The following lemma follows from the optimality condition of (2.1) and the definition of the prox-function (see the proof in [23]).

LEMMA 1. *For every $u, x \in X$, and $y \in \mathbb{R}^n$, we have*

$$V_X(P_{x,X}(y), u) \leq V_X(x, u) + y^T(u - x) + \frac{1}{2}\|y\|_*^2,$$

where the $\|\cdot\|_*$ denotes the conjugate of $\|\cdot\|$, i.e., $\|y\|_* = \max\{\langle x, y\rangle \,|\, \|x\| \leq 1\}$.

2.2. The CSA method for general convex problems. In this section, we consider the constrained optimization problem in (1.1). We assume the expectation functions $f(x)$ and $g(x)$, in addition to being well-defined and finite-valued for every $x \in X$, are continuous and convex on X. Throughout this section, we make the following assumption for the stochastic subgradients.

ASSUMPTION 1. *For any $x \in X$, a.e. $\zeta \in \mathcal{P}$ and a.e. $\xi \in \mathcal{Q}$,*

$$\mathbb{E}[\|F'(x,\zeta)\|_*^2] \leq M_F^2 \quad \text{and} \quad \mathbb{E}[\|G'(x,\xi)\|_*^2] \leq M_G^2,$$

where $F'(x,\zeta) \in \partial_x F(x,\zeta)$ and $G'(x,\xi) \in \partial_x G(x,\xi)$.

The CSA method can be viewed as a stochastic counterpart of Polayk's subgradient method, which was originally designed for solving deterministic nonsmooth convex optimization problems (see [27] and a more recent generalization in [3]). At each iterate x_k, $k \geq 0$, depending on whether $g(x_k) \leq \eta_k$ for some tolerance $\eta_k > 0$, it moves either along the subgradient direction $f'(x_k)$ or $g'(x_k)$, with an appropriately chosen stepsize γ_k which also depends on $\|f'(x_k)\|$ and $\|g'(x_k)\|$. However, Polayk's subgradient method cannot be applied to solve (1.1) because we do not have access to exact information about f', g' and g. The CSA method differs from Polyak's subgradient method in the following three aspects. Firstly, the search direction h_k is defined in a stochastic manner: we first check if the solution x_k we computed at iteration k violates the condition $G(x_k, \xi_k) \le \eta_k$ for some $\eta_k \ge 0$ and a random realization ξ_k of ξ. If so, we set the $h_k = G'(x_k, \xi_k)$ in order to control the violation of expectation constraint. Otherwise, we set $h_k = F'(x_k, \zeta_k)$. Secondly, for some $1 \le s \le N$, we partition the indices $I = \{s, ..., N\}$ into two subsets: $\mathcal{B} = \{s \le k \le N | G(x_k, \xi_k) \le \eta_k\}$ and $\mathcal{N} = I \setminus \mathcal{B}$, and define the output $\bar{x}_{N,s}$ as an ergodic mean of x_k over $\mathcal{B}$. This differs from the Polyak's subgradient method that defines the output solution as the best $x_k, k \in \mathcal{B}$, with the smallest objective value. Thirdly, while the original Polayk's subgradient method were developed only for general nonsmooth problems, we show that the CSA method also exhibits an optimal rate of convergence for solving strongly convex problems by properly choosing $\{\gamma_k\}$ and $\{\eta_k\}$.

Algorithm 1 The cooperative SA algorithm

Input: initial point $x_1 \in X$, stepsizes $\{\gamma_k\}$ and tolerances $\{\eta_k\}$.

for $k = 1, 2, \dots, N$

Set

$$h_k = \begin{cases} F'(x_k, \zeta_k), & \text{if } G(x_k, \xi_k) \le \eta_k; \\ G'(x_k, \xi_k), & \text{otherwise.} \end{cases} \tag{2.4}$$

$$x_{k+1} = P_{x_k, X}(\gamma_k h_k). \tag{2.5}$$

end for

Output: Set $\mathcal{B} = \{s \le k \le N | G(x_k, \xi_k) \le \eta_k\}$ for some $1 \le s \le N$, and define the output

$$\bar{x}_{N,s} = (\textstyle\sum_{k \in \mathcal{B}} \gamma_k)^{-1} (\textstyle\sum_{k \in \mathcal{B}} \gamma_k x_k), \tag{2.6}$$

Notice that each iteration of the CSA algorithm only requires at most one realization ζ_k and ξ_k. Hence, this algorithm can make progress in an on-line fashion as more and more samples of ζ_k and ξ_k are collected. Our goal in the remaining part of this section is to establish the rate of convergence associated with this algorithm, in terms of both the distance to the optimal value and the violation of constraints. It should also be noted that Algorithm 1 is conceptional only as we have not specified a few algorithmic parameters (e.g. $\{\gamma_k\}$ and $\{\eta_k\}$). We will come back to this issue after establishing some general properties about this method.

The following result establishes a simple but important recursion about the CSA method for stochastic optimization with expectation constraints.

PROPOSITION 2. *For any* $1 \le s \le N$*, we have*

$$\sum_{k \in \mathcal{N}} \gamma_k (\eta_k - G(x, \xi_k)) + \sum_{k \in \mathcal{B}} \gamma_k \langle F'(x_k, \zeta_k), x_k - x \rangle \le V(x_s, x) + \frac{1}{2} \sum_{k \in \mathcal{B}} \gamma_k^2 \|F'(x_k, \zeta_k)\|_*^2 + \frac{1}{2} \sum_{k \in \mathcal{N}} \gamma_k^2 \|G'(x_k, \xi_k)\|_*^2, \tag{2.7}$$

for all $x \in X$*.*

Proof. For any $s \le k \le N$, using Lemma 1, we have

$$V(x_{k+1}, x) \le V(x_k, x) + \gamma_k \langle h_k, x - x_k \rangle + \frac{1}{2} \gamma_k^2 \|h_k\|_*^2. \tag{2.8}$$

Observe that if $k \in \mathcal{B}$, we have $h_k = F'(x_k, \zeta_k)$, and

$$\langle h_k, x_k - x \rangle = \langle F'(x_k, \zeta_k), x_k - x \rangle.$$

Moreover, if $k \in \mathcal{N}$, we have $h_k = G'(x_k, \xi_k)$ and

$$\langle h_k, x_k - x \rangle = \langle G'(x_k, \xi_k), x_k - x \rangle \ge G(x_k, \xi_k) - G(x, \xi_k) \ge \eta_k - G(x, \xi_k).$$

Summing up the inequalities in (2.8) from $k = s$ to N and using the previous two observations, we obtain

$$\begin{aligned} V(x_{k+1}, x) &\le V(x_s, x) - \textstyle\sum_{k=s}^N \gamma_k \langle h_k, x_k - x\rangle + \frac{1}{2}\sum_{k=s}^N \gamma_k^2 \|h_k\|_*^2 \\ &\le V(x_s, x) - \left[\textstyle\sum_{k\in\mathcal{N}} \gamma_k \langle G'(x_k, \xi_k), x_k - x\rangle + \sum_{k\in\mathcal{B}} \gamma_k \langle F'(x_k, \zeta_k), x_k - x\rangle\right] + \frac{1}{2}\sum_{k=s}^N \gamma_k^2 \|h_k\|_*^2 \\ &\le V(x_s, x) - \left[\textstyle\sum_{k\in\mathcal{N}} \gamma_k (\eta_k - G(x, \xi_k)) + \sum_{k\in\mathcal{B}} \gamma_k \langle F'(x_k, \zeta_k), x_k - x\rangle\right] \\ &\quad + \tfrac{1}{2}\textstyle\sum_{k\in\mathcal{B}} \gamma_k^2 \|F'(x_k, \zeta_k)\|_*^2 + \frac{1}{2}\sum_{k\in\mathcal{N}} \gamma_k^2 \|G'(x_k, \xi_k)\|_*^2. \end{aligned} \tag{2.9}$$

Rearranging the terms in above inequality, we obtain (2.7) ∎

Using Proposition 2, we present below a sufficient condition under which the output solution $\bar{x}_{N,s}$ is well-defined.

Lemma 3. *Let x^* be an optimal solution of (1.1). If*

$$\tfrac{N-s+1}{2} \min_{k\in\mathcal{N}} \gamma_k \eta_k > D_X^2 + \tfrac{1}{2}\textstyle\sum_{k\in\mathcal{B}} \gamma_k^2 M_F^2 + \frac{1}{2}\sum_{k\in\mathcal{N}} \gamma_k^2 M_G^2, \tag{2.10}$$

then $\mathcal{B} \ne \emptyset$, i.e., $\bar{x}_{N,s}$ is well-defined. Moreover, we have one of the following two statements holds,

a) $|\mathcal{B}| \ge (N - s + 1)/2$,

b) $\sum_{k\in\mathcal{B}} \gamma_k \langle f'(x_k), x_k - x^*\rangle \le 0$.

Proof. Taking expectation w.r.t. ξ_k and ζ_k on both sides of (2.7) and fixing $x = x^*$, we have

$$\begin{aligned} \sum_{k\in\mathcal{N}} \gamma_k[\eta_k - g(x^*)] + \sum_{k\in\mathcal{B}} \gamma_k \langle f'(x_k), x_k - x^*\rangle &\le V(x_s, x^*) + \frac{1}{2}\sum_{k\in\mathcal{B}} \gamma_k^2 M_F^2 + \frac{1}{2}\sum_{k\in\mathcal{N}} \gamma_k^2 M_G^2 \\ &\le D_X^2 + \frac{1}{2}\sum_{k\in\mathcal{B}} \gamma_k^2 M_F^2 + \frac{1}{2}\sum_{k\in\mathcal{N}} \gamma_k^2 M_G^2. \end{aligned} \tag{2.11}$$

If $\sum_{k\in\mathcal{B}} \gamma_k \langle f'(x_k), x_k - x^*\rangle \le 0$, part b) holds. If $\sum_{k\in\mathcal{B}} \gamma_k \langle f'(x_k), x_k - x^*\rangle \ge 0$, we have

$$\sum_{k\in\mathcal{N}} \gamma_k[\eta_k - g(x^*)] \le V(x_s, x^*) + \frac{1}{2}\sum_{k\in\mathcal{B}} \gamma_k^2 M_F^2 + \frac{1}{2}\sum_{k\in\mathcal{N}} \gamma_k^2 M_G^2,$$

which, in view of $g(x^*) \le 0$, implies that

$$\sum_{k\in\mathcal{N}} \gamma_k \eta_k \le V(x_s, x^*) + \frac{1}{2}\sum_{k\in\mathcal{B}} \gamma_k^2 M_F^2 + \frac{1}{2}\sum_{k\in\mathcal{N}} \gamma_k^2 M_G^2. \tag{2.12}$$

Suppose that $|\mathcal{B}| < (N - s + 1)/2$, i.e., $|\mathcal{N}| \ge (N - s + 1)/2$. Then,

$$\textstyle\sum_{k\in\mathcal{N}} \gamma_k \eta_k \ge \frac{N-s+1}{2} \min_{k\in\mathcal{N}} \gamma_k \eta_k > V(x_s, x^*) + \frac{1}{2}\sum_{k\in\mathcal{B}} \gamma_k^2 M_F^2 + \frac{1}{2}\sum_{k\in\mathcal{N}} \gamma_k^2 M_G^2,$$

which contradicts with (2.12). Hence, part a) holds. ∎

Now we are ready to establish the main convergence properties of the CSA method.

Theorem 4. *Suppose that $\{\gamma_k\}$ and $\{\eta_k\}$ in the CSA algorithm are chosen such that (2.10) holds. Then for any $1 \le s \le N$, we have*

$$\mathbb{E}[f(\bar{x}_{N,s}) - f(x^*)] \le \tfrac{2D_X^2 + \max\{M_F^2, M_G^2\}\sum_{s\le k\le N} \gamma_k^2}{(N-s+1)\min_{s\le k\le N} \gamma_k}, \tag{2.13}$$

$$\mathbb{E}[g(\bar{x}_{N,s})] \le \max_{s\le k\le N} \eta_k. \tag{2.14}$$

Proof. We first show (2.13). By Lemma 3, if Lemma 3 part (b) holds, dividing both sides of (**??**) by $\sum_{k\in\mathcal{B}} \gamma_k$ and taking expectation, we have

$$\mathbb{E}[f(\bar{x}_{N,s}) - f(x^*)] \le 0. \tag{2.15}$$

If $|\mathcal{B}| \geq (N-s+1)/2$, we have $\sum_{k\in\mathcal{B}} \gamma_k \geq |\mathcal{B}| \min_{k\in\mathcal{B}} \gamma_k \geq \frac{N-s+1}{2} \min_{k\in\mathcal{B}} \gamma_k$. It follows from the definition of $\bar{x}_{N,s}$ in (2.6), the convexity of $f(\cdot)$ and (2.11) that

$$\begin{array}{rcl}\sum_{k\in\mathcal{N}} \gamma_k\eta_k + \sum_{k\in\mathcal{B}} \gamma_k E[f(\bar{x}_{N,s}) - f(x^*)] & \leq & \sum_{k\in\mathcal{N}} \gamma_k\eta_k + \sum_{k\in\mathcal{B}} E[\gamma_k(f(x_k) - f(x^*))] \\ & \leq & D_X^2 + \frac{1}{2}\sum_{k\in\mathcal{B}} \gamma_k^2 M_F^2 + \frac{1}{2}\sum_{k\in\mathcal{N}} \gamma_k^2 M_G^2,\end{array}$$

which implies that

$$|\mathcal{N}| \min_{k\in\mathcal{N}} \gamma_k\eta_k + \left(\sum_{k\in\mathcal{B}} \gamma_k\right) E[f(\bar{x}_{N,s}) - f(x^*)] \leq D_X^2 + \frac{1}{2}\sum_{k\in\mathcal{B}} \gamma_k^2 M_F^2 + \tfrac{1}{2}\sum_{k\in\mathcal{N}} \gamma_k^2 M_G^2. \tag{2.16}$$

Using this bound and the fact $\gamma_k\eta_k \geq 0$ in (2.16), we have

$$\begin{aligned}\mathbb{E}[f(\bar{x}_{N,s}) - f(x^*)] &\leq \frac{2D_X^2 + \sum_{k\in\mathcal{B}} \gamma_k^2 M_F^2 + \sum_{k\in\mathcal{N}} \gamma_k^2 M_G^2}{(N-s+1)\min_{k\in I} \gamma_k} \\ &\leq \frac{2D_X^2 + \max\{M_F^2, M_G^2\} \sum_{s\leq k\leq N} \gamma_k^2}{(N-s+1)\min_{k\in\mathcal{B}} \gamma_k}.\end{aligned} \tag{2.17}$$

Combining these two inequalities (2.15) and (2.17), we have (2.13). Now we show that (2.14) holds. For any $k \in \mathcal{B}$, we have $G(x_k, \xi_k) \leq \eta_k$. Taking expectations w.r.t. ξ_k on both sides, we have $\mathbb{E}[g(x_k)] \leq \eta_k$, which, in view of the definition of $\bar{x}_{N,s}$ in (2.6) and the convexity of $g(\cdot)$, then implies that

$$\mathbb{E}[g(\bar{x}_{N,s})] \leq \frac{\mathbb{E}[\sum_{k\in\mathcal{B}} \gamma_k g(x_k)]}{\sum_{k\in\mathcal{B}} \gamma_k} \leq \frac{\sum_{k\in\mathcal{B}} \gamma_k\eta_k}{\sum_{k\in\mathcal{B}} \gamma_k} \leq \max_{s\leq k\leq N} \eta_k. \tag{2.18}$$

∎

Below we provide a few specific selections of $\{\gamma_k\}$, $\{\eta_k\}$ and s that lead to the optimal rate of convergence for the CSA method. In particular, we will present a constant and variable stepsize policy, respectively, in Corollaries 5 and 6.

Corollary 5. *If s=1,$\gamma_k = \frac{D_X}{\sqrt{N}(M_F+M_G)}$ and $\eta_k = \frac{4(M_F+M_G)D_X}{\sqrt{N}}$, $k = 1, ...N$, then*

$$\begin{aligned}\mathbb{E}[f(\bar{x}_{N,s}) - f(x^*)] &\leq \tfrac{4D_X(M_F+M_G)}{\sqrt{N}}, \\ \mathbb{E}[g(\bar{x}_{N,s})] &\leq \tfrac{4D_X(M_F+M_G)}{\sqrt{N}}.\end{aligned}$$

Proof. First, observe that condition (2.10) holds by using the facts that

$$\begin{aligned}&\tfrac{N-s+1}{2} \min_{k\in\mathcal{N}} \gamma_k\eta_k = \tfrac{N}{2}\tfrac{4D_X^2}{N} = 2D_X^2, \\ &D_X^2 + \tfrac{1}{2}\sum_{k\in\mathcal{B}} \gamma_k^2 M_F^2 + \tfrac{1}{2}\sum_{k\in\mathcal{N}} \gamma_k^2 M_G^2 \\ &\leq D_X^2 + \tfrac{1}{2}\sum_{k\in\mathcal{B}} \tfrac{D_X^2 M_F^2}{N(M_F+M_G)^2} + \tfrac{1}{2}\sum_{k\in\mathcal{N}} \tfrac{D_X^2 M_G^2}{N(M_F+M_G)^2} \\ &\leq D_X^2 + \tfrac{1}{2}\sum_{k=1}^N \tfrac{D_X^2}{N} \leq 2D_X^2.\end{aligned}$$

It then follows from Lemma 3 and Theorem 4 that

$$\begin{aligned}\mathbb{E}[f(\bar{x}_{N,s}) - f(x^*)] &\leq \frac{2D_X(M_F+M_G) + \sum_{k\in\mathcal{B}} \frac{D_X M_F^2}{N(M_F+M_G)} + \sum_{k\in\mathcal{N}} \frac{D_X M_G^2}{N(M_F+M_G)}}{\sqrt{N}} \leq \tfrac{4D_X(M_F+M_G)}{\sqrt{N}}, \\ \mathbb{E}[g(\bar{x}_{N,s})] &\leq \max_{s\leq k\leq N} \eta_k = \tfrac{4D_X(M_F+M_G)}{\sqrt{N}}.\end{aligned}$$

∎

Corollary 6. *If $s = \frac{N}{2}$, $\gamma_k = \frac{D_X}{\sqrt{k}(M_F+M_G)}$ and $\eta_k = \frac{4D_X(M_F+M_G)}{\sqrt{k}}$, $k = 1, 2, ..., N$, then*

$$\begin{aligned}\mathbb{E}[f(\bar{x}_{N,s}) - f(x^*)] &\leq \tfrac{4D_X(1+\frac{1}{2}\log 2)(M_F+M_G)}{\sqrt{N}}, \\ \mathbb{E}[g(\bar{x}_{N,s})] &\leq \tfrac{4\sqrt{2}D_X(M_F+M_G)}{\sqrt{N}}.\end{aligned}$$

Proof. The proof is similar to that of corollary 4 and hence the details are skipped. ∎

In view of Corollaries 5 and 6, the CSA algorithm achieves an $\mathcal{O}(1/\sqrt{N})$ rate of convergence for solving problem (1.1). This convergence rate seems to be unimprovable as it matches the optimal rate of convergence for deterministic convex optimization problems with functional constraints [25]. However, to the best of our knowledge, no such complexity bounds have been obtained before for solving stochastic optimization problems with functional expectation constraints.

2.3. Strongly convex objective and strongly convex constraints. In this subsection, we are interested in establishing the convergence of the CSA algorithm applied to strongly convex problems. More specifically, we assume that the objective function F and constraint function G are both strongly convex w.r.t. x, i.e., $\exists \mu_F > 0$ and $\mu_G > 0$ s.t.

$$\begin{aligned}
F(x_1,\zeta) &\geq F(x_2,\zeta) + \langle F'(x_2,\zeta), x_1 - x_2\rangle + \tfrac{\mu_F}{2}\|x_1 - x_2\|^2, \forall x_1, x_2 \in X,\\
G(x_1,\xi) &\geq G(x_2,\xi) + \langle G'(x_2,\xi), x_1 - x_2\rangle + \tfrac{\mu_G}{2}\|x_1 - x_2\|^2, \forall x_1, x_2 \in X.
\end{aligned}$$

In order to estimate the convergent rate of the CSA algorithm for solving strongly convex problems, we need to assume that the prox-function $V_X(\cdot,\cdot)$ satisfies a quadratic growth condition

$$V_X(z,x) \leq \frac{Q}{2}\|z - x\|^2, \forall z, x \in X. \tag{2.19}$$

Moreover, letting γ_k be the stepsizes used in the CSA method, and denoting

$$a_k = \begin{cases} \frac{\mu_F \gamma_k}{Q}, & k \in \mathcal{B}, \\ \frac{\mu_G \gamma_k}{Q}, & k \in \mathcal{N}, \end{cases} \quad \text{and } A_k = \begin{cases} 1, & k = 1, \\ (1 - a_k)A_{k-1}, & k \geq 2. \end{cases}$$

we modify the output in Algorithm 1 to

$$\bar{x}_{N,s} = \frac{\sum_{k\in\mathcal{B}} \rho_k x_k}{\sum_{k\in\mathcal{B}} \rho_k}, \tag{2.20}$$

where $\rho_k = \gamma_k / A_k$. The following simple result will be used in the convergence analysis of the CSA method.

Lemma 7. *If $a_k \in (0,1]$, $k = 0,1,2,\ldots$, $A_k > 0, \forall k \geq 1$, and $\{\Delta_k\}$ satisfies*

$$\Delta_{k+1} \leq (1 - a_k)\Delta_k + B_k, \forall k \geq 1,$$

then we have

$$\tfrac{\Delta_{k+1}}{A_k} \leq (1 - a_1)\Delta_1 + \textstyle\sum_{i=1}^{k} \tfrac{B_i}{A_i}.$$

Below we provide an important recursion about CSA applied to strongly convex problems. This result differs from Proposition 2 for the general convex case in that we use different weight ρ_k rather than γ_k.

Proposition 8. *For any $1 \leq s \leq N$, we have*

$$\begin{aligned}
\textstyle\sum_{k\in\mathcal{N}} \rho_k(\eta_k - G(x,\xi_k)) + \sum_{k\in\mathcal{B}} \rho_k[F(x_k,\zeta_k) - F(x,\zeta_k)] \leq (1 - a_s)D_X^2 \\
+ \tfrac{1}{2}\textstyle\sum_{k\in\mathcal{B}} \rho_k\gamma_k\|F'(x_k,\zeta_k)\|_*^2 + \tfrac{1}{2}\sum_{k\in\mathcal{N}} \rho_k\gamma_k\|G'(x_k,\xi_k)\|_*^2.
\end{aligned} \tag{2.21}$$

Proof. Consider the iteration k, $\forall s \leq k \leq N$. If $k \in \mathcal{B}$, by Lemma 1 and the strong convexity of $F(x,\zeta)$, we have

$$\begin{aligned}
V(x_{k+1},x) &\leq V(x_k,x) - \gamma_k\langle h_k, x_k - x\rangle + \tfrac{1}{2}\gamma_k^2\|F'(x_k,\zeta_k)\|_*^2 \\
&= V(x_k,x) - \gamma_k\langle F'(x_k,\zeta_k), x_k - x\rangle + \tfrac{1}{2}\gamma_k^2\|F'(x_k,\zeta_k)\|_*^2 \\
&\leq V(x_k,x) - \gamma_k\left[F(x_k,\zeta_k) - F(x,\zeta_k) + \tfrac{\mu_F}{2}\|x_k - x\|_*^2\right] + \tfrac{1}{2}\gamma_k^2\|F'(x_k,\xi_k)\|_*^2 \\
&\leq \left(1 - \tfrac{\mu_F\gamma_k}{Q}\right)V(x_k,x) - \gamma_k[F(x_k,\zeta_k) - F(x,\zeta_k)] + \tfrac{1}{2}\gamma_k^2\|F'(x_k,\zeta_k)\|_*^2.
\end{aligned}$$

Similarly for $k \in \mathcal{N}$, using Lemma 1 and the strong convexity of $G(x,\xi)$, we have

$$\begin{aligned} V(x_{k+1},x) &\le V(x_k,x) - \gamma_k \langle h_k, x_k - x\rangle + \tfrac{1}{2}\gamma_k^2 \|G'(x_k,\xi_k)\|_*^2 \\ &= V(x_k,x) - \gamma_k \langle G'(x_k,\xi_k), x_k - x\rangle + \tfrac{1}{2}\gamma_k^2 \|G'(x_k,\xi_k)\|_*^2 \\ &\le V(x_k,x) - \gamma_k \left[(G(x_k,\xi_k) - G(x,\xi_k)) + \tfrac{\mu_G}{2}\|x_k - x\|_*^2 \right] + \tfrac{1}{2}\gamma_k^2 \|G'(x_k,\xi_k)\|_*^2 \\ &\le \left(1 - \tfrac{\mu_G \gamma_k}{Q}\right) V(x_k,x) - \gamma_k(\eta_k - G(x,\xi_k)) + \tfrac{1}{2}\gamma_k^2 \|G'(x_k,\xi_k)\|_*^2. \end{aligned}$$

Summing up these inequalities for $s \le k \le N$ and using Lemma 7, we have

$$\begin{aligned} \tfrac{V(x_{N+1},x)}{A_N} \le (1-a_s)\, V(x_s,x) &- \left[\textstyle\sum_{k\in\mathcal{N}} \tfrac{\gamma_k}{A_k}(\eta_k - G(x,\xi_k)) + \textstyle\sum_{k\in\mathcal{B}} \tfrac{\gamma_k}{A_k}[F(x_k,\zeta_k) - F(x,\zeta_k)] \right] \\ &+ \tfrac{1}{2}\textstyle\sum_{k\in\mathcal{N}} \tfrac{\gamma_k^2}{A_k}\|G'(x_k,\xi_k)\|_*^2 + \tfrac{1}{2}\textstyle\sum_{k\in\mathcal{B}} \tfrac{\gamma_k^2}{A_k}\|F'(x_k,\zeta_k)\|_*^2, \end{aligned}$$

Using the fact $V(x_{N+1},x)/A_N \ge 0$ and the definition of ρ_k, and rearranging the terms in the above inequality, we obtain (2.21). ∎

Lemma 9 below provides a sufficient condition which guarantees $\bar{x}_{N,s}$ to be well-defined.

LEMMA 9. *Let x^* be the optimal solution of (1.1). If*

$$\tfrac{N-s+1}{2} \min_{k\in\mathcal{N}} \rho_k \eta_k > (1-a_s) D_X^2 + \tfrac{1}{2}\textstyle\sum_{k\in\mathcal{N}} \rho_k \gamma_k M_G^2 + \tfrac{1}{2}\textstyle\sum_{k\in\mathcal{B}} \rho_k \gamma_k M_F^2, \tag{2.22}$$

then $\mathcal{B} \ne \emptyset$ and hence $\bar{x}_{N,s}$ is well-defined. Moreover, we have one of the following two statements holds,

a) $|\mathcal{B}| \ge (N-s+1)/2$,

b) $\sum_{k\in\mathcal{B}} \rho_k [f(x_k) - f(x^*)] \le 0$.

Proof. The proof of this result is similar to that of Lemma 2 and hence the details are skipped. ∎

With the help of Proposition 8, we are ready to establish the main convergence properties of the CSA method for solving strongly convex problems.

THEOREM 10. *Suppose that $\{\gamma_k\}$ and $\{\eta_k\}$ in the CSA algorithm are chosen such that (2.22) holds. Then for any $1 \le s \le N$, we have*

$$\mathbb{E}[f(\bar{x}_{N,s}) - f(x^*)] \le ((N-s+1)\min_{s\le k\le N} \rho_k)^{-1} \left(2(1-a_s)D_X^2 + \textstyle\sum_{k\in\mathcal{B}} \rho_k\gamma_k M_F^2 + \textstyle\sum_{k\in\mathcal{N}} \rho_k \gamma_k M_G^2\right), \tag{2.23}$$

$$\mathbb{E}[g(\bar{x}_{N,s})] \le \max_{k\in\mathcal{B}} \eta_k. \tag{2.24}$$

Proof. Taking expectation w.r.t. $\zeta_i, \xi_i, 1 \le i \le k$, on both sides of (2.21) (with $x = x^*$) and using Assumption 1, we have

$$\textstyle\sum_{k\in\mathcal{N}} \rho_k(\eta_k - g(x^*)) + \sum_{k\in\mathcal{B}} \rho_k \mathbb{E}[f(x_k) - f(x^*)] \le (1-a_s)D_X^2 + \tfrac{1}{2}\sum_{k\in\mathcal{B}} \rho_k\gamma_k M_F^2 + \tfrac{1}{2}\sum_{k\in\mathcal{N}} \rho_k\gamma_k M_G^2.$$

(2.23) then immediately follows from the above inequality, (2.20), the convexity of f and the fact that $g(x^*) \le 0$. Moreover, (2.24) follows similarly to (2.18). ∎

Below we provide a stepsize policy of s, γ_k and η_k in order to achieve the optimal rate of convergence for solving strongly convex problems.

COROLLARY 11. *Let $s = \frac{N}{2}$, $\gamma_k = \begin{cases} \frac{2Q}{\mu_F(k+1)}, & \text{if } k\in\mathcal{B}; \\ \frac{2Q}{\mu_G(k+1)}, & \text{if } k \in \mathcal{N}, \end{cases}$, $\eta_k = \frac{2\mu_G Q}{k}\left(\frac{2D_X^2}{k} + \max\left\{\frac{M_F^2}{\mu_F^2}, \frac{M_G^2}{\mu_G^2}\right\}\right)$, then we have*

$$\begin{aligned} \mathbb{E}[f(\bar{x}_{N,s}) - f(x^*)] &\le \tfrac{8\mu_F D_X^2}{N^2 Q} + \tfrac{4\mu_F Q}{N}\max\{\tfrac{M_F^2}{\mu_F^2}, \tfrac{M_G^2}{\mu_G^2}\}, \\ \mathbb{E}[g(\bar{x}_{N,s})] &\le \tfrac{16\mu_G Q D_X^2}{N^2} + \tfrac{4\mu_G Q}{N}\max\left\{\tfrac{M_F^2}{\mu_F^2}, \tfrac{M_G^2}{\mu_G^2}\right\}. \end{aligned}$$

Proof. Based on our selection of s, γ_k, η_k and the definition of a_k, A_k and ρ_k, we have

$$a_k = \tfrac{2}{k+1},\ A_k = \prod_{i=2}^{k}(1-a_i) = \tfrac{2}{k(k+1)},\ \rho_k = \begin{cases} \frac{kQ}{\mu_F}, & \text{if } k \in \mathcal{B}; \\ \frac{kQ}{\mu_G}, & \text{if } k \in \mathcal{N}, \end{cases}$$

For $\forall s \le k \le N$, by the definition of s, γ_k and η_k, we have

$$\begin{aligned}&(1-a_s)V(x_s,x) + \tfrac{1}{2}\textstyle\sum_{k\in\mathcal{N}}\rho_k\gamma_k M_G^2 + \tfrac{1}{2}\sum_{k\in\mathcal{B}}\rho_k\gamma_k M_F^2 \\ &\le D_X^2 + \tfrac{1}{2}\textstyle\sum_{k\in\mathcal{B}}\frac{\gamma_k^2}{A_k}M_F^2 + \tfrac{1}{2}\sum_{k\in\mathcal{N}}\frac{\gamma_k^2}{A_k}M_G^2 \le D_X^2 + Q^2(|\mathcal{B}|\frac{M_F^2}{\mu_F^2} + |\mathcal{N}|\frac{M_G^2}{\mu_G^2}) \le D_X^2 + \frac{Q^2N}{2}\max\left\{\frac{M_F^2}{\mu_F^2}, \frac{M_G^2}{\mu_G^2}\right\},\end{aligned}$$

$$\tfrac{N-s+1}{2}\min_{k\in\mathcal{N}}\rho_k\eta_k = \tfrac{N}{4}\min_{k\in\mathcal{N}}\tfrac{kQ}{\mu_G}\tfrac{2\mu_G Q}{k}\left(\tfrac{2D_X^2}{k} + \max\left\{\tfrac{M_F^2}{\mu_F^2}, \tfrac{M_G^2}{\mu_G^2}\right\}\right) \ge D_X^2 + \tfrac{Q^2N}{2}\max\left\{\tfrac{M_F^2}{\mu_F^2}, \tfrac{M_G^2}{\mu_G^2}\right\}.$$

Combining the above two inequalities, we can easily see that condition (2.22) holds. It then follows from Theorem 10 that

$$\begin{aligned}\mathbb{E}[f(\bar{x}_{N,s}) - f(x^*)] &\le ((N-s+1)\min_{s\le k\le N}\rho_k)^{-1}\left(2(1-a_s)D_X^2 + \textstyle\sum_{k\in\mathcal{B}}\rho_k\gamma_k M_F^2 + \sum_{k\in\mathcal{N}}\rho_k\gamma_k M_G^2\right) \\ &\le \tfrac{8\mu_F D_X^2}{N^2Q} + \tfrac{4\mu_F Q}{N}\max\{\tfrac{M_F^2}{\mu_F^2}, \tfrac{M_G^2}{\mu_G^2}\}, \\ \mathbb{E}[g(\bar{x}_{N,s})] &\le \max_{k\in\mathcal{B}}\eta_k \le \tfrac{16\mu_G Q D_X^2}{N^2} + \tfrac{4\mu_G Q}{N}\max\left\{\tfrac{M_F^2}{\mu_F^2}, \tfrac{M_G^2}{\mu_G^2}\right\}.\end{aligned}$$

■

In view of Corollary 11, the CSA algorithm can achieve the optimal rate of convergence for strongly convex optimization with strongly convex constraints. To the best of our knowledge, this is the first time such a complexity result is obtained in the literature and this result is new also for the deterministic setting.

2.4. Probability of large deviation. In the previous two subsections, we established the expected convergence properties over many runs of the CSA algorithm. In this subsection, we are interested in the large deviation properties for a single run of this method. For the sake of simplicity, we focus only on the general convex case, i.e., f and g are not necessarily strongly convex.

First note that by Corollary 5 and the Markov's inequality, we have

$$\begin{aligned}\text{Prob}\left(f(\bar{x}_{N,s}) - f(x^*) > \lambda_1 \frac{4D_X(1+\frac{1}{2}\log 2)(M_F+M_G)}{\sqrt{N}}\right) < \tfrac{1}{\lambda_1}, \forall\lambda_1 \ge 0; \\ \text{Prob}\left(g(\bar{x}_{N,s}) > \lambda_2 \frac{16\log 2 D_X(M_F+M_G)}{\sqrt{N}}\right) < \tfrac{1}{\lambda_2}, \forall\lambda_2 \ge 0.\end{aligned}$$

It then follows that in order to find a solution $\bar{x}_{N,s} \in X$ such that

$$\text{Prob}\left(f(\bar{x}_{N,s}) - f(x^*) \le \epsilon \text{ and } g(\bar{x}_{N,s}) \le \epsilon\right) > 1 - \Lambda,$$

the number of iteration performed by the CSA method can be bounded by

$$\mathcal{O}\left\{\tfrac{1}{\epsilon^2\Lambda^2}\right\}. \tag{2.25}$$

We will show that this result can be significantly improved if Assumption A1 is augmented by the following "light-tail" assumption.

Assumption 2. *For and $x \in X$,*

$$\begin{aligned}\mathbb{E}[\exp\{\|F'(x,\zeta)\|_*^2/M_F^2\}] &\le \exp\{1\}, \\ \mathbb{E}[\exp\{\|G'(x,\xi)\|_*^2/M_G^2\}] &\le \exp\{1\}, \\ \mathbb{E}[\exp\{(G(x,\xi) - g(x))^2/\sigma^2\}] &\le \exp\{1\}.\end{aligned}$$

We first present the following Bernstein inequality that will be used to establish the large-deviation properties of the CSA method (e.g. see [23]). Note that in the sequel, we denote $\xi_{[k]} := \{\xi_1, \ldots, \xi_k\}$.

LEMMA 12. *Let $\xi_1, \xi_2, \ldots$ be a sequence of i.i.d. random variables, and $\zeta_t = \zeta(\xi_{[t]})$ be deterministic Borel functions of $\xi_{[t]}$ such that $\mathbb{E}[\zeta_t] = 0$ a.s. and $\mathbb{E}[\exp\{\zeta_t^2/\sigma_t^2\}] \le \exp\{1\}$ a.s., where $\sigma_t > 0$ are deterministic. Then*

$$\forall \lambda \ge 0 : \text{Prob}\left\{\textstyle\sum_{t=1}^N \zeta_t > \lambda \sqrt{\sum_{t=1}^N \sigma_t^2}\right\} \le \exp\{-\lambda^2/3\}.$$

Now we are ready to establish the large deviation properties of the CSA algorithm.

THEOREM 13. *Under Assumption 2, $\forall \lambda \ge 0$,*

$$\text{Prob}\{f(\bar{x}_{N,s}) - f(x^*) \ge K_0 + \lambda K_1\} \le 2\exp\{-\lambda\} + 2\exp\{-\frac{\lambda^2}{3}\}, \tag{2.26}$$

$$\text{Prob}\left\{g(\bar{x}_{N,s}) \ge \left(\textstyle\sum_{k\in\mathcal{B}} \gamma_k\right)^{-1} \left(\sum_{k\in\mathcal{B}} \gamma_k \eta_k + \lambda\sigma\sqrt{\sum_{k\in\mathcal{B}} \gamma_k^2}\right)\right\} \le \exp\{-\lambda^2/3\}, \tag{2.27}$$

where $K_0 = \frac{\frac{1}{2}D_X^2 + M_F^2 \sum_{k\in\mathcal{B}} \gamma_k^2 + M_G^2 \sum_{k\in\mathcal{N}} \gamma_k^2}{\sum_{k\in\mathcal{B}} \gamma_k}$ *and*

$$K_1 = \frac{M_F^2 \sum_{k\in\mathcal{B}} \gamma_k^2 + M_G^2 \sum_{k\in\mathcal{N}} \gamma_k^2 + \sigma\sqrt{\sum_{k\in\mathcal{N}} \gamma_k^2} + M_F D_X \sqrt{\sum_{k\in\mathcal{B}} \gamma_k^2}}{\sum_{k\in\mathcal{B}} \gamma_k}.$$

Proof. Let $G(x, \xi_k) = g(x) + \delta_k$ and $F'(x_k, \zeta_k) = f'(x_k) + \Delta_k$. It follows from the inequality (2.7) (with $x = x^*$) and the fact $g(x^*) \le 0$ that

$$\begin{aligned}\textstyle\sum_{k\in\mathcal{N}} \gamma_k \eta_k + (\sum_{k\in\mathcal{B}} \gamma_k)(f(\bar{x}_{N,s}) - f(x^*)) \le D_X^2 + \sum_{k\in\mathcal{B}} \gamma_k^2 \|F'(x_k, \zeta_k)\|_*^2 \\ + \sum_{k\in\mathcal{N}} \gamma_k^2 \|G'(x_k, \xi_k)\|_*^2 + \sum_{k\in\mathcal{N}} \gamma_k \delta_k - \sum_{k\in\mathcal{B}} \gamma_k \langle \Delta_k, x_k - x^*\rangle.\end{aligned} \tag{2.28}$$

Now we provide probabilistic bounds for $\sum_{k\in\mathcal{B}} \gamma_k^2 \|F'(x_k, \zeta_k)\|_*^2$, $\sum_{k\in\mathcal{N}} \gamma_k^2 \|G'(x_k, \xi_k)\|_*^2$, $\sum_{k\in\mathcal{N}} \gamma_k \delta_k$ and $\sum_{k\in\mathcal{B}} \gamma_k \langle \Delta_k, x_k - x^*\rangle$. First, setting $\theta_k = \gamma_k^2 / \sum_{k\in\mathcal{B}} \gamma_k^2$, using the fact that $\mathbb{E}[\exp\{\|F'(x_k, \zeta_k)\|_*^2 / M_F^2\}] \le exp\{1\}$ and Jensens inequality, we have

$$\exp\{\textstyle\sum_{k\in\mathcal{B}} \theta_k (\|F'(x_k, \zeta_k)\|_*^2 / M_F^2)\} \le \sum_{k\in\mathcal{B}} \theta_k \exp\{\|F'(x_k, \zeta_k)\|_*^2 / M_F^2\},$$

and hence that

$$\mathbb{E}[\exp\{\textstyle\sum_{k\in\mathcal{B}} \gamma_k^2 \|F'(x_k, \zeta_k)\|_*^2 / M_F^2 \sum_{k\in\mathcal{B}} \gamma_k^2\}] \le \exp\{1\}.$$

It then follows from Markov's inequality that $\forall \lambda \ge 0$,

$$\begin{aligned}&\text{Prob}(\textstyle\sum_{k\in\mathcal{B}} \gamma_k^2 \|F'(x_k, \zeta_k)\|_*^2 > (1 + \lambda) M_F^2 \sum_{k\in\mathcal{B}} \gamma_k^2) \\ &= \text{Prob}\left(\exp\left\{\frac{\sum_{k\in\mathcal{B}} \gamma_k^2 \|F'(x_k, \zeta_k)\|_*^2}{M_F^2 \sum_{k\in\mathcal{B}} \gamma_k^2}\right\} > \exp(1 + \lambda)\right) \\ &\le \tfrac{\exp\{1\}}{\exp\{1+\lambda\}} \le \exp\{-\lambda\}.\end{aligned} \tag{2.29}$$

Similarly, we have

$$\text{Prob}\left(\textstyle\sum_{k\in\mathcal{N}} \gamma_k^2 \|G'(x_k, \xi_k)\|_*^2 > (1 + \lambda) M_G^2 \sum_{k\in\mathcal{N}} \gamma_k^2\right) \le \exp\{-\lambda\}. \tag{2.30}$$

Second, for $\sum_{k\in\mathcal{N}} \gamma_k \delta_k$, setting $\iota_k = \gamma_k / \sum_{k\in\mathcal{B}} \gamma_k$, and noting that $\mathbb{E}[\delta_k] = 0$ and $\mathbb{E}[\exp\{\delta_k^2/\sigma^2\}] \le \exp\{1\}$, we obtain $\mathbb{E}[\iota_k \delta_k] = 0$, $\mathbb{E}[\exp\{\iota_k^2 \delta_k^2 / \zeta_k^2 \sigma^2\}] \le \exp\{1\}$. By lemma 12, we have

$$\text{Prob}\left\{\textstyle\sum_{k\in\mathcal{N}} \gamma_k \delta_k > \lambda\sigma\sqrt{\sum_{k\in\mathcal{N}} \gamma_k^2}\right\} \le \exp\{-\lambda^2/3\}. \tag{2.31}$$

Lastly, let us consider $\sum_{k\in\mathcal{B}}\gamma_k\langle\Delta_k, x_k - x^*\rangle$. Setting $\beta_k = \gamma_k\langle\Delta_k, x_k - x^*\rangle$ and noting that $\mathbb{E}[\|\Delta_k\|_*^2] \le (2M_F)^2$, we have

$$\mathbb{E}[\exp\{\beta_k^2/(2M_F\gamma_k D_X)^2\}] \le \exp\{1\},$$

which, in view of Lemma 12, implies that

$$\text{Prob}\left\{\sum_{k\in\mathcal{B}}\beta_k > 2\lambda M_F D_X\sqrt{\sum_{k\in\mathcal{B}}\gamma_k^2}\right\} \le \exp\{-\lambda^2/3\}. \tag{2.32}$$

Combining (2.29),(2.30), (2.31) and (2.32), and rearranging the terms we get (2.26).

Let us show that (2.27) holds. Clearly, by the convexity of $g(\cdot)$ and definition of $\bar{x}_{N,s}$, we have

$$g(\bar{x}_{N,s}) = g(\sum_{k\in\mathcal{B}}\iota_k x_k) \le \sum_{k\in\mathcal{B}}\iota_k g(x_k) = \sum_{k\in\mathcal{B}}\iota_k[G(x_k,\xi_k) - \delta_k].$$

Also,

$$\text{Prob}\{\sum_{k\in\mathcal{B}}\iota_k\delta_k > \lambda\sigma\sqrt{\sum_{k\in\mathcal{B}}\iota_k^2}\} \le \exp\{-\lambda^2/3\}.$$

Hence, we have (2.33) holds.

■

Applying the stepsize strategy in Corollary 5 to Theorem 12, then it follows that the number of iterations performed by the CSA method can be bounded by

$$\mathcal{O}\left\{\frac{1}{\epsilon^2}(\log\frac{1}{\Lambda})^2\right\}.$$

We can see that the above result significantly improves the one in (2.25).

3. Expectation constraints over problem parameters. In this section, we are interested in solving a class of parameterized stochastic optimization problems whose parameters are defined by expectation constraints as described in (1.3)-(1.4), under the assumption that such a pair of solutions satisfying (1.3)-(1.4) exists. Our goal in this section is to present a variant of the CSA algorithm to approximately solve problem (1.3)-(1.4) and establish its convergence properties. More specifically, we discuss this variant of the CSA algorithm when applied to the parameterized stochastic optimization problem in (1.3)-(1.4) and then consider a modified problem by imposing certain strong convexity assumptions to the function $\Phi(x,y,\zeta)$ w.r.t. y and $G(x,\xi)$ w.r.t. x in Subsections 3.1 and 3.2, respectively. In Subsection 3.3, we discuss some large deviation properties for the variant of the CSA method for the problem defined by (1.3)-(1.4).

3.1. Stochastic optimization with parameter feasibility constraints. Given tolerance $\eta > 0$ and target accuracy $\epsilon > 0$, we will present a variant of the CSA algorithm, namely cooperative stochastic parameter approximation (CSPA), to find a pair of approximate solutions $(\bar{x},\bar{y}) \in X\times Y$ s.t. $\mathbb{E}[g(\bar{x})] \le \eta$ and $\mathbb{E}[\phi(\bar{x},\bar{y}) - \phi(\bar{x},y)] \le \epsilon$, $\forall y \in Y$, in this subsection. Before we describe the CSPA method, we need slightly modify Assumption 1.

Assumption 3. *For any $x \in X$ and $y \in Y$,*

$$\mathbb{E}[\|\Phi'(x,y,\zeta)\|_*^2] \le M_\Phi^2 \quad \textit{and} \quad \mathbb{E}[\|G'(x,\xi)\|_*^2] \le M_G^2,$$

where $\Phi'(x,y,\zeta) \in \partial_y\Phi(x,y,\zeta)$ and $G'(x,\xi) \in \partial_x G(x,\xi)$.

We assume that the distance generating functions $\omega_X : X \mapsto \mathbb{R}$ and $\omega_Y : Y \mapsto \mathbb{R}$ are strongly convex with modulus 1 w.r.t. given norms in $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively, and that their associated prox-mappings $P_{x,X}$ and $P_{y,Y}$ (see (2.1)) are easily computable.

We make the following modifications to the CSA method in Section 2.1 in order to apply it to solve problem (1.3)-(1.4). Firstly, we still check the solution (x_k, y_k) to see whether x_k violates the condition $G(x_k,\xi_k) \le \eta_k$. If so, we set the search direction as $G'(x_k,\xi_k)$ to update x_k, while keeping y_k intact. Otherwise, we only update y_k along the direction $\Phi'(x_k, y_k, \zeta_k)$. Secondly, we define the output as a randomly selected (x_k, y_k) according to a certain probability distribution instead of the ergodic mean of $\{(x_k, y_k)\}$. Since we are solving a coupled optimization and feasibility problem, each iteration of our algorithm only updates either y_k or x_k and requires the computation of either Φ' or G' depending on whether $G(x_k, \xi_k) \le \eta_k$. This differs from the SA method used in Jiang and Shanbhag [17] that requires two projection steps and the computation of two subgradients at each iteration to solve a different parameterized stochastic optimization problem.

Algorithm 2 The cooperative stochastic parameter approximation method

Input: initial point (x_0, y_0), stepsize $\{\gamma_k\}$, tolerance $\{\eta_k\}$, number of iterations N.

for k=1,2,...,N

if $G(x_k, \xi_k) \le \eta_k$

$$x_{k+1} = x_k, y_{k+1} = P_{y_k,Y}(\gamma_k \Phi'(x_k, y_k, \zeta_k)); \tag{3.1}$$

else

$$x_{k+1} = P_{x_k,X}(\gamma_k G'(x_k, \xi_k)), y_{k+1} = y_k. \tag{3.2}$$

end if

end for

Output: Set $\mathcal{B} := \{s \le k \le N | G(x_k, \xi_k) \le \eta_k\}$ for some $1 \le s \le N$, and define the output (x_R, y_R), where R is randomly chosen according to

$$\text{Prob}\{R = k\} = \frac{\gamma_k}{\sum_{k\in\mathcal{B}} \gamma_k}, k \in \mathcal{B}. \tag{3.3}$$

With a little abuse of notation, we still use $\mathcal{B}$ to represent the set $\{s \le k \le N | G(x_k, \xi_k) \le \eta_k\}$, $I = \{s, \ldots, N\}$, and $\mathcal{N} = I \setminus \mathcal{B}$. The following result mimics Proposition 2.

Proposition 14. *Under Assumption 3, for any $1 \le s \le N$, we have*

$$\sum_{k\in\mathcal{B}} \gamma_k \langle \nabla_y \Phi(x_k, y_k, \zeta_k), y_k - y\rangle \le D_Y^2 + \tfrac{1}{2} \sum_{k\in\mathcal{B}} \gamma_k^2 \|\Phi'(x_k, y_k, \zeta_k)\|_*^2, \ \forall y \in Y, \tag{3.4}$$

$$\sum_{k\in\mathcal{N}} \gamma_k [\eta_k - G(x, \xi_k)] \le D_X^2 + \tfrac{1}{2} \sum_{k\in\mathcal{N}} \gamma_k^2 \|G'(x_k, \xi_k)\|_*^2, \ \forall x \in X, \tag{3.5}$$

where $D_X \equiv D_{X,w_x}$ and $D_Y \equiv D_{Y,w_y}$ are defined as in (2.3).

Proof. By Lemma 1, if $k \in \mathcal{B}$,

$$V(y_{k+1}, y) \le V(y_k, y) + \gamma_k \langle \Phi'(x_k, y_k, \zeta_k), y - y_k\rangle + \tfrac{1}{2}\gamma_k^2 \|\Phi'(x_k, y_k, \zeta_k)\|_*^2.$$

Also note that $V(y_{k+1}, y) = V(y_k, y)$ for $k \in \mathcal{N}$. Summing up these relations for $k \in \mathcal{B} \cup \mathcal{N}$ and using the fact that $V(y_s, y) \le D_Y^2$, we have

$$\begin{aligned} V(y_{N+1}, y) &\le V(y_s, y) + \tfrac{1}{2} \textstyle\sum_{k\in\mathcal{B}} \gamma_k^2 \|\Phi'(x_k, y_k, \zeta_k)\|_*^2 - \sum_{k\in\mathcal{B}} \gamma_k \langle \Phi'(x_k, y_k, \zeta_k), y_k - y\rangle \\ &\le D_Y^2 + \tfrac{1}{2} \textstyle\sum_{k\in\mathcal{B}} \gamma_k^2 \|\Phi'(x_k, y_k, \zeta_k)\|_*^2 - \sum_{k\in\mathcal{B}} \gamma_k \langle \Phi'(x_k, y_k, \zeta_k), y_k - y\rangle. \end{aligned} \tag{3.6}$$

Similarly for $k \in \mathcal{N}$, we have

$$\begin{aligned} V(x_{k+1}, x) &\le V(x_k, x) + \gamma_k \langle G'(x_k, \xi_k), x - x_k\rangle + \tfrac{1}{2}\gamma_k^2 \|G'(x_k, \xi_k)\|_*^2 \\ &\le V(x_k, x) - \gamma_k (\eta_k - G(x, \xi_k)) + \tfrac{1}{2}\gamma_k^2 \|G'(x_k, \xi_k)\|_*^2, \end{aligned}$$

where the last inequality follows from the convexity of $G(x, \xi)$ and the fact that $G(x_k, \xi_k) \ge \eta_k$. Also observe that $V(x_{k+1}, x) = V(x_k, x)$, $\forall k \in \mathcal{B}$. Summing up these relations for $k \in \mathcal{N} \cup \mathcal{B}$ and using the fact that $V(x_s, x) \le D_X^2$, we obtain

$$\begin{aligned} V(x_{N+1}, x) &\le V(x_s, x) + \tfrac{1}{2} \textstyle\sum_{k\in\mathcal{N}} \gamma_k^2 \|G'(x_k, \xi_k)\|_*^2 - \sum_{k\in\mathcal{N}} (\eta_k - G(x, \xi_k)) \\ &\le D_X^2 + \tfrac{1}{2} \textstyle\sum_{k\in\mathcal{N}} \gamma_k^2 \|G'(x_k, \xi_k)\|_*^2 - \sum_{k\in\mathcal{N}} (\eta_k - G(x, \xi_k)). \end{aligned} \tag{3.7}$$

Using the facts $V(y_{N+1}, y) \ge 0$ and $V(x_{N+1}, x) \ge 0$, and rearranging the terms in (3.6) and (3.7), we then obtain (3.4) and (3.5), respectively. ∎

The following result provides a sufficient condition under which (x_R, y_R) is well-defined.

LEMMA 15. *Under Assumption 3, if*

$$\tfrac{N-s+1}{2} \min_{k\in\mathcal{N}} \gamma_k \eta_k > D_X^2 + \tfrac{M_G^2}{2} \textstyle\sum_{k\in\mathcal{N}} \gamma_k^2, \tag{3.8}$$

then $\mathcal{B} \neq \emptyset$, i.e., (x_R, y_R) is well-defined. Futhermore, we have $|\mathcal{B}| \ge \frac{N-s+1}{2}$.

Proof. Under Assumption 3, taking expectation w.r.t. ξ_k on both sides of (3.5) and fixing $x = x^*$, we have

$$\sum_{k\in\mathcal{N}} \gamma_k(\eta_k - g(x^*)) \le D_X^2 + \frac{M_G^2}{2} \sum_{k\in\mathcal{N}} \gamma_k^2. \tag{3.9}$$

For contradiction, suppose that $|\mathcal{B}| < \frac{N-s+1}{2}$, i.e., $|\mathcal{N}| \ge \frac{N-s+1}{2}$. The above relation, in view of $g(x^*) \le 0$, implies that

$$\frac{N-s+1}{2} \min_{k\in\mathcal{N}} \gamma_k \eta_k \le \sum_{k\in\mathcal{N}} \gamma_k \eta_k \le D_X^2 + \frac{M_G^2}{2} \sum_{k\in\mathcal{N}} \gamma_k^2,$$

which contradicts with (3.8). ∎

Theorem 16 below summarizes the main convergence properties for Algorithm 2 applied to problem (1.3)-(1.4).

THEOREM 16. *Under Assumption 3, if the stepsizes $\{\gamma_k\}$ and tolerances $\{\eta_k\}$ are chosen such that (3.8) holds, we have*

$$\mathbb{E}[\phi(x_R, y_R) - \phi(x_R, y^*(x_R))] \le \frac{2D_Y^2 + M_\Phi^2 \sum_{k\in\mathcal{B}} \gamma_k^2}{2\sum_{k\in\mathcal{B}} \gamma_k}, \tag{3.10}$$

$$\mathbb{E}[g(x_R)] \le \left(\textstyle\sum_{k\in\mathcal{B}} \gamma_k\right)^{-1} \textstyle\sum_{k\in\mathcal{B}} \gamma_k \eta_k, \tag{3.11}$$

where the expectation is taken w.r.t. R and $\xi_1, \ldots, \xi_N$.

Proof. We first show that (3.10) holds. Using the convexity of $\Phi(x_k, y_k, \zeta_k)$ w.r.t. y_k and then taking expectation w.r.t. ζ_k on both sides of (3.4) (with $y = y^*(x_k)$), we have

$$\sum_{k\in\mathcal{B}} \gamma_k \mathbb{E}_{\zeta_k} \left[\phi(x_k, y_k) - \phi(x_k, y^*(x_k)) | \zeta_{[k-1]}\right] \le D_Y^2 + \frac{M_\Phi^2}{2} \sum_{k\in\mathcal{B}} \gamma_k^2.$$

Dividing $\sum_{k\in\mathcal{B}} \gamma_k$ on both sides of the above inequality and using the distribution of R in (3.3), we obtain (3.10).

Then we need to show (3.11) holds. Let $\delta_k = G(x_k, \xi_k) - g(x_k)$ and hence $\mathbb{E}[\delta_k | \xi_{[k-1]}] = 0$. By the definition of set $\mathcal{B}$, we have $G(x_k, \xi_k) \le \eta_k \ \forall k \in \mathcal{B}$ and

$$\sum_{k\in\mathcal{B}} \gamma_k g(x_k) = \sum_{k\in\mathcal{B}} \gamma_k [G(x_k, \xi_k) - \delta_k] \le \sum_{k\in\mathcal{B}} \gamma_k \eta_k - \sum_{k\in\mathcal{B}} \gamma_k \delta_k, \tag{3.12}$$

which, in view of the fact that $\{\delta_k\}$ is a martingale-difference sequence, implies that

$$\sum_{k\in\mathcal{B}} \gamma_k \mathbb{E}_{\xi_{[N]}}[g(x_k)] \le \sum_{k\in\mathcal{B}} \gamma_k \eta_k.$$

Using the distribution of R in (3.3), we obtain (3.11). ∎

It is worth noting that the result in (3.10) still holds if we replace $y^*(x_R)$ by any fixed $y \in Y$, i.e., $\mathbb{E}[\phi(x_R, y_R) - \phi(x_R, y)] \le (2D_Y^2 + M_\Phi^2 \sum_{k\in\mathcal{B}} \gamma_k^2)/(2\sum_{k\in\mathcal{B}}\gamma_k)$, $\forall y \in Y$. Below we provide a special selection of s, $\{\gamma_k\}$ and $\{\eta_k\}$.

Corollary 17. *Let $s = \frac{N}{2}+1$, $\gamma_k = \frac{D_X}{M_G\sqrt{k}}$ and $\eta_k = \frac{8M_G D_X}{\sqrt{k}}$ for $k = 1,\dots,N$. Then we have*

$$\mathbb{E}[\phi(x_R, y_R) - \phi(x_R, y^*(x_R))] \le \tfrac{8M_\Phi D_Y}{\sqrt{N}} \max\{\nu, \tfrac{1}{\nu}\},$$
$$\mathbb{E}[g(x_R)] \le \tfrac{32\log 2 M_G D_X}{\sqrt{N}},$$

where $\nu := (M_G D_Y)/(M_\Phi D_X)$.

Proof. Similarly to Corollary 5, we can show that (3.8) holds. It then follows from Lemma 15 and Theorem 16 that

$$\textstyle\sum_{k\in\mathcal{B}} \gamma_k = \sum_{k\in\mathcal{B}} \frac{D_X}{M_G\sqrt{k}} \ge \frac{D_X}{M_G}\frac{N}{4}\frac{1}{\sqrt{N}} = \frac{D_X\sqrt{N}}{4M_G}.$$

$$\begin{aligned}\mathbb{E}[\phi(x_R, y_R) - \phi(x_R, y^*)] &\le \tfrac{4M_G}{D_X\sqrt{N}}\left[D_Y^2 + \tfrac{1}{2}\textstyle\sum_{k\in\mathcal{B}} \frac{D_X^2 M_\Phi^2}{M_G^2 k}\right] \\ &\le \tfrac{4M_G}{D_X\sqrt{N}}\left[D_Y^2 + \tfrac{1}{2}\textstyle\sum_{k=N/2}^{N} \frac{D_X^2 M_\Phi^2}{M_G^2 k}\right] \\ &\le \tfrac{4M_G}{D_X\sqrt{N}}[D_Y^2 + \tfrac{\log 2}{2} D_X^2 \tfrac{M_\Phi^2}{M_G^2}] \le \tfrac{8M_\Phi D_Y}{\sqrt{N}}\max\{\nu, \tfrac{1}{\nu}\},\end{aligned}$$

$$\mathbb{E}[g(x_R)] \le \tfrac{4M_G}{D_X\sqrt{N}} \textstyle\sum_{k\in\mathcal{B}} \frac{8D_X^2}{k} \le \frac{32\log 2 M_G D_X}{\sqrt{N}}.$$

∎

By Corollary (17), the CSPA method applied to (1.3)-(1.4) can achieve an $\mathcal{O}(1/\sqrt{N})$ rate of convergence.

3.2. CSPA with strong convexity assumptions. In this subsection, we modify problem (1.3)-(1.4) by imposing certain strongly convexity assumptions to Φ and G with respect to y and x, respectively, i.e., $\exists \mu_\Phi, \mu_G > 0$, s.t.

$$\Phi(x, y_1, \zeta) \ge \Phi(x, y_2, \zeta) + \langle \Phi'(x, y_2, \zeta), y_1 - y_2\rangle + \tfrac{\mu_\Phi}{2}\|y_1 - y_2\|^2, \quad \forall y_1, y_2 \in Y. \tag{3.13}$$
$$G(x_1, \xi) \ge G(x_2, \xi) + \langle G'(x_2, \xi), x_1 - x_2\rangle + \tfrac{\mu_G}{2}\|x_1 - x_2\|^2, \quad \forall x_1, x_2 \in X. \tag{3.14}$$

We also assume that the pair of solutions (x^*, y^*) exists for problem (1.3)-(1.4). Our main goal in this subsection is to estimate the convergence properties of the CSPA algorithm under these new assumptions.

We need to modify the probability distribution (3.3) used in the CSPA algorithm as follows. Given the stepsize γ_k, modulus μ_G and μ_Φ, and growth parameter Q (see (2.19)), let us define

$$a_k := \begin{cases} (\mu_\Phi\gamma_k)/Q, & k \in \mathcal{B}; \\ (\mu_G\gamma_k)/Q, & k \in \mathcal{N}, \end{cases} \quad \text{and } A_k := \begin{cases} \prod_{i\in\mathcal{B}, i\le k}(1-a_i), & k \in \mathcal{B}; \\ \prod_{i\in\mathcal{N}, i\le k}(1-a_i), & k \in \mathcal{N}. \end{cases} \tag{3.15}$$

Also, let us denote $\rho_k := \gamma_k/A_k$. Then the probability distribution of R is modified to

$$\text{Prob}\{R = k\} = \tfrac{\rho_k}{\sum_{k\in\mathcal{B}}\rho_k}, k \in \mathcal{B}. \tag{3.16}$$

The following result shows some simple but important properties for the modified CSPA method applied to problem (1.3)-(1.4).

Proposition 18. *For any $s \le k \le m$, we have*

$$\textstyle\sum_{k\in\mathcal{B}} \rho_k[\Phi(x_k, y_k, \zeta_k) - \Phi(x_k, y, \zeta_k)] \le (1 - \frac{\mu_\Phi\gamma_s}{Q})V_Y(y_s, y) + \frac{1}{2}\sum_{k\in\mathcal{B}} \rho_k\gamma_k\|\Phi'(x_k, y_k, \zeta_k)\|_*^2, \quad \forall y \in Y \tag{3.17}$$
$$\textstyle\sum_{k\in\mathcal{N}} \rho_k\left[\eta_k - G(x, \xi_k)\right] \le \left(1 - \frac{\mu_G\gamma_s}{Q}\right) V_X(x_s, x) + \frac{1}{2}\sum_{k\in\mathcal{N}} \gamma_k\rho_k\|G'(x_k, \xi_k)\|_*^2, \quad \forall x \in X. \tag{3.18}$$

Proof. Using Lemma 1 and the strong convexity of Φ w.r.t. y, for $k \in B$, we have

$$\begin{aligned} V_Y(y_{k+1}, y) &\le V_Y(y_k, y) - \gamma_k \langle \Phi'(x_k, y_k, \zeta_k), y_k - y\rangle + \tfrac{1}{2}\gamma_k^2 \|\Phi'(x_k, y_k, \zeta_k)\|_*^2 \\ &\le V_Y(y_k, y) - \gamma_k \left[\Phi(x_k, y_k, \zeta_k) - \Phi(x_k, y, \zeta_k) + \tfrac{\mu_\Phi}{2}\|y_k - y\|^2\right] + \tfrac{1}{2}\gamma_k^2 \|\Phi'(x_k, y_k, \zeta_k)\|_*^2 \\ &\le \left(1 - \tfrac{\mu_\Phi \gamma_k}{Q}\right) V_Y(y_k, y) - \gamma_k[\Phi(x_k, y_k, \zeta_k) - \Phi(x_k, y, \zeta_k)] + \tfrac{1}{2}\gamma_k^2 \|\Phi'(x_k, y_k, \zeta_k)\|_*^2. \end{aligned}$$

Also note that $V_Y(y_{k+1}, y) = V_Y(y_k, y)$ for all $k \in \mathcal{N}$. Summing up these relations for all $k \in \mathcal{B} \cup \mathcal{N}$ and using Lemma 7, we obtain

$$\tfrac{V_Y(y_N, y)}{A_{N+1}} \le \left(1 - \tfrac{\mu_\Phi \gamma_s}{Q}\right) V_Y(y_s, y) - \textstyle\sum_{k\in\mathcal{B}} \tfrac{\gamma_k}{A_k}[\Phi(x_k, y_k, \zeta_k) - \Phi(x_k, y, \zeta_k)] + \tfrac{1}{2}\sum_{k\in\mathcal{B}} \tfrac{\gamma_k^2}{A_k}\|\Phi'(x_k, y_k, \zeta_k)\|_*^2. \quad (3.19)$$

Similarly for $k \in \mathcal{N}$, we have

$$\begin{aligned} V_X(x_{k+1}, x) &\le V_X(x_k, x) - \gamma_k \langle G'(x_k, \xi_k), x_k - x\rangle + \tfrac{1}{2}\gamma_k^2 \|G'(x_k, \xi_k)\|_*^2 \\ &\le V_X(x_k, x) - \gamma_k \left[G(x_k, \xi_k) - G(x, \xi_k) + \tfrac{\mu_G}{2}\|x_k - x\|^2\right] + \tfrac{1}{2}\gamma_k^2 \|G'(x_k, \xi_k)\|_*^2 \\ &\le \left(1 - \tfrac{\mu_G \gamma_k}{Q}\right) V_X(x_k, x) - \gamma_k[\eta_k - G(x, \xi_k)] + \tfrac{1}{2}\gamma_k^2 \|G'(x_k, \xi_k)\|_*^2, \end{aligned}$$

where the last inequality follows from the fact that $G(x_k, \xi_k) > \eta_k$ for $k \in \mathcal{N}$. Note that $V_X(x_{k+1}, x) = V_X(x_k, x)$ for all $k \in \mathcal{B}$. Summing up these relations for all $k \in \mathcal{N} \cup \mathcal{B}$ and using Lemma 7, we have

$$\tfrac{V_X(x_{N+1}, x)}{A_N} \le \left(1 - \tfrac{\mu_G \gamma_s}{Q}\right) V_X(x_s, x) - \textstyle\sum_{k\in\mathcal{N}} \tfrac{\gamma_k}{A_k}[\eta_k - G(x, \xi_k)] + \tfrac{1}{2}\sum_{k\in\mathcal{N}} \tfrac{\gamma_k^2}{A_k}\|G'(x_k, \xi_k)\|_*^2. \quad (3.20)$$

Using the facts that $V_Y(y_{N+1}, y)/A_N \ge 0$ and $V_X(x_{N+1}, x)/A_N \ge 0$, and rearranging the terms in (3.19) and (3.20), we obtain (3.17) and (3.18), respectively.

■

Lemma 19 below provides a sufficient condition which guarantees that the output solution (x_R, y_R) is well-defined. We will provide a lower bound of $|\mathcal{B}|$ later in Corollary 21 because to do so requires a specific selection of stepsizes $\{\gamma_k\}$ and tolerances $\{\eta_k\}$.

Lemma 19. *If*

$$\sum_{k=s}^{N} \rho_k \eta_k > (1 - \tfrac{\mu_G \gamma_s}{Q}) V_X(x_s, x^*) + \tfrac{M_G^2}{2} \sum_{k\in\mathcal{N}} \rho_k \gamma_k, \quad (3.21)$$

then $\mathcal{B} \neq \emptyset$, i.e., (x_R, y_R) is well-defined.

Proof. Taking expectation w.r.t. ξ_k on both sides of (3.18) and fixing $x = x^*$, we have

$$\textstyle\sum_{k\in\mathcal{N}} \rho_k[\eta_k - g(x^*)] \le \left(1 - \tfrac{\mu_G \gamma_s}{Q}\right) V_X(x_s, x^*) + \tfrac{M_G^2}{2} \sum_{k\in\mathcal{N}} \gamma_k \rho_k,$$

which, in view of the fact that $g(x^*) \le 0$, implies that

$$\textstyle\sum_{k\in\mathcal{N}} \rho_k \eta_k \le \left(1 - \tfrac{\mu_G \gamma_s}{Q}\right) V_X(x_s, x^*) + \tfrac{M_G^2}{2} \sum_{k\in\mathcal{N}} \tfrac{\gamma_k^2}{A_k}. \quad (3.22)$$

For contradiction, suppose that $\mathcal{B} = \emptyset$, i.e., $|\mathcal{N}| = N - s + 1$, then we have

$$\textstyle\sum_{k\in\mathcal{N}} \rho_k \eta_k = \sum_{k=s}^{N} \rho_k \eta_k > (1 - \tfrac{\mu_G \gamma_s}{Q}) V_X(x_s, x) + \tfrac{M_G^2}{2} \sum_{k\in\mathcal{N}} \rho_k \gamma_k,$$

which contradicts with (3.22). ■

Now let us establish the rate of convergence of the modified CSPA method for problem (1.3)-(1.4).

THEOREM 20. *Suppose that* $\{\rho_k\}$ *and* $\{\eta_k\}$ *are chosen such that* (3.21) *holds. Then*

$$\mathbb{E}[\phi(x_R, y_R) - \phi(x_R, y^*(x_R))] \le \left(\sum_{k\in B}\rho_k\right)^{-1}\left((1-\tfrac{\mu_\Phi\gamma_s}{Q})D_Y^2 + \tfrac{M_\Phi^2}{2}\sum_{k\in B}\rho_k\gamma_k\right). \tag{3.23}$$

$$\mathbb{E}[g(x_R)] \le \left(\sum_{k\in B}\rho_k\right)^{-1}\left(\sum_{k\in B}\rho_k\eta_k\right). \tag{3.24}$$

Proof. We first show that (3.23) holds. Using (3.17) with $y = y^*(x_k)$ and taking expectation w.r.t. ξ_k on the both sides of (3.17), we obtain

$$\sum_{k\in\mathcal{B}}\rho_k[\phi(x_k, y_k) - \phi(x_k, y^*(x_k))] \le (1-\tfrac{\mu_\Phi\gamma_s}{Q})V_Y(y_s, y^*(x_k)) + \tfrac{M_\Phi^2}{2}\sum_{k\in\mathcal{B}}\rho_k\gamma_k.$$

Dividing $\sum_{k\in\mathcal{B}}\rho_k$ from both sides of the above inequality, and using the definition of the probability distribution of R in (3.16) and the fact that $V(y_s, y^*(x_k)) \le D_Y^2$, we obtain (3.23). We can show (3.24) by using an argument similar to the one used in the proof of (2.24), and hence the details are skipped. ∎

Now we provide a specific selection of $\{\gamma_k\}$ and $\{\eta_k\}$ that not only satisfies (3.21), but also guarantees $|\mathcal{B}| \ge (N-s+1)/2$. While the selection of η_k only depends on iteration index k, i.e.,

$$\eta_k = \frac{8QM_G^2}{k\mu_G}, \tag{3.25}$$

the selection of γ_k depends on the particular position of iteration index k in set $\mathcal{B}$ or $\mathcal{N}$. More specifically, let $\tau_{\mathcal{B}(k)}$ and $\tau_{\mathcal{N}(k)}$ be the position of index k in set $\mathcal{B}$ and set $\mathcal{N}$, respectively (for example, $\mathcal{B} = \{1,3,5,9,10\}$ and $\mathcal{N} = \{2,4,6,7,8\}$. If $k = 9$, then $\tau_{B(k)} = 4$). We define γ_k as

$$\gamma_k = \begin{cases} \frac{2Q}{\mu_\Phi(\tau_{\mathcal{B}(k)}+1)}, & k\in\mathcal{B}; \\ \frac{2Q}{\mu_G(\tau_{\mathcal{N}(k)}+1)}, & k\in\mathcal{N}. \end{cases} \tag{3.26}$$

Such a selection of γ_k can be conveniently implemented by using two separate counters in each iteration to represent $\tau_{\mathcal{B}(k)}$ and $\tau_{\mathcal{N}(k)}$.

COROLLARY 21. *Let* $s = 1$, η_k *and* γ_k *be given in* (3.25) *and* (3.26), *respectively. Then we have*

a) $|\mathcal{B}| \ge N/2$;

b) $\mathbb{E}[\phi(x_R, y_R) - \phi(x_R, y^*)] \le \frac{8(QM_\Phi)^2}{(N+2)\mu_\Phi}$ *and* $\mathbb{E}[g(x_R)] \le \frac{(8QM_G)^2}{(N+2)\mu_G}$.

Proof. For any $k\in\mathcal{B}$, we have $\gamma_k = 2Q/(\mu_\Phi(\tau_{B(k)}+1))$, which, in views of the definition of a_k, A_k and ρ_k, then implies that

$$a_k = \tfrac{2}{\tau_{B(k)}+1},\ A_k = \tfrac{2}{\tau_{B(k)}(\tau_{B(k)}+1)} \text{ and } \rho_k = \tfrac{Q\tau_{B(k)}}{\mu_\Phi},\ \forall k\in\mathcal{B}. \tag{3.27}$$

Similarly for any $k\in\mathcal{N}$, we have $\gamma_k = 2Q/(\mu_G(\tau_{N(k)}+1))$, and hence

$$a_k = \tfrac{2}{\tau_{N(k)}+1},\ A_k = \tfrac{2}{\tau_{N(k)}(\tau_{N(k)}+1)} \text{ and } \rho_k = \tfrac{Q\tau_{N(k)}}{\mu_G},\ \forall k\in\mathcal{N}. \tag{3.28}$$

Let us show part a) holds. For contradiction, suppose $|\mathcal{B}| < N/2$, i.e., $|\mathcal{N}| \ge N/2$. Observe that by (3.25) and (3.28),

$$\begin{aligned}\sum_{\tau_{N(k)}=1}^{N/2}\rho_k\eta_k &= \sum_{\tau_{N(k)}=1}^{N/2}\tfrac{8\tau_{N(k)}Q^2M_G^2}{k\mu_G^2} = \tfrac{8Q^2M_G^2}{\mu_G^2}\sum_{j=1}^{N/2}\tfrac{j}{k} \\ &\ge \tfrac{8Q^2M_G^2}{\mu_G^2}\sum_{j=1}^{N/2}\tfrac{j}{N/2+j} = \tfrac{8Q^2M_G^2}{\mu_G^2}\left(\tfrac{N}{2} - \tfrac{N}{2}\sum_{j=1}^{N/2}\tfrac{1}{N/2+j}\right) \\ &= \tfrac{4NQ^2M_G^2}{\mu_G^2}\left(1 - \sum_{j=N/2+1}^{N}\tfrac{1}{j}\right) > \tfrac{NQ^2M_G^2}{\mu_G^2},\end{aligned} \tag{3.29}$$

and

$$(1-\frac{\mu_G\gamma_1}{Q})D_X^2+\frac{1}{2}\sum_{k\in N}\rho_k\gamma_k M_G^2=\frac{Q^2M_G^2}{\mu_G^2}\sum_{\tau_{N(k)}=1}^{|\mathcal{N}|}\frac{\tau_{N(k)}}{\tau_{N(k)}+1}\le\frac{NQ^2M_G^2}{\mu_G^2}. \tag{3.30}$$

Combining (3.29) and (3.30), we have

$$\sum_{k\in\mathcal{N}}\rho_k\eta_k=\sum_{\tau_{N(k)}=1}^{|\mathcal{N}|}\rho_k\eta_k\ge\sum_{\tau_{N(k)}=1}^{N/2}\rho_k\eta_k>(1-\frac{\mu_G\gamma_s}{Q})V_X(x_s,x)+\frac{M_G^2}{2}\sum_{k\in\mathcal{N}}\rho_k\gamma_k,$$

which contradicts with (3.22). Therefore, we have $|\mathcal{B}|\ge N/2$. Using the fact that $\sum_{k=1}^{N}\rho_k\eta_k\ge\sum_{\tau_{N(k)}=1}^{N/2}\rho_k\eta_k$, we also obtain (3.21). We now show part b) holds. It follows from part a) and (3.27) that

$$\sum_{k\in B}\rho_k\ge\sum_{\tau_{B(k)}=1}^{N/2}\rho_k=\frac{8\mu_\Phi}{(N+2)N}.$$

We then conclude from Theorem 20, the above inequality, and the relations in (3.27) and (3.28) that

$$\mathbb{E}[\phi(x_R,y_R)-\phi(x_R,y^*(x_R))]\le\left(\sum_{k\in B}\rho_k\right)^{-1}\left[(1-\frac{\mu_\Phi\gamma_s}{Q})D_Y^2+\frac{M_\Phi^2}{2}\sum_{k\in B}\rho_k\gamma_k\right]\le\frac{8(QM_\Phi)^2}{(N+2)\mu_\Phi},$$
$$\mathbb{E}[g(x_R)]\le\frac{8\mu_\Phi}{(N+2)N}\sum_{k\in B}\frac{8Q^2M_G^2\tau_{B(k)}}{k\mu_G\mu_\Phi}\le\frac{(8QM_G)^2}{(N+2)\mu_G}.$$

■

Clearly, comparing to the result in (17), the above modified CSPA method can improve the rate of convergence to $\mathcal{O}(1/N)$ under the strong convexity assumptions.

3.3. Probability of large deviation. In this subsection, we study the large deviation properties for a single run of the CSPA algorithm applied to problem (1.3)-(1.4), and we focus on the basic setting without strong convexity assumptions. Using the stepsize strategy in Corollary 17 and Markov's inequality, we have $\forall\lambda\ge 0$,

$$\text{Prob}\left(\phi(x_R,y_R)-\phi(x_R,y^*(x_R))\ge\lambda\frac{8M_\Phi D_Y}{\sqrt{N}}\max\{\nu,\tfrac{1}{\nu}\}\right)\le\frac{1}{\lambda},$$
$$\text{Prob}\left(g(x_R)\ge\lambda\frac{32\log 2D_X M_G}{\sqrt{N}}\right)\le\frac{1}{\lambda}.$$

It then follows that the number of iterations performed by CSPA for finding an (ϵ,Λ)-solution, i.e., a pair of solutions $(x_R,y_R)\in X\times Y$ s.t.

$$\text{Prob}(\phi(x_R,y_R)-\phi(x_R,y^*(x_R))\le\epsilon \text{ and } g(x_R)\le\epsilon)\ge 1-\Lambda, \tag{3.31}$$

after disregarding a few constant factors, can be bounded by

$$\mathcal{O}\left\{\frac{1}{\epsilon^2\Lambda^2}\right\}. \tag{3.32}$$

This complexity bound seems rather pessimistic in terms of its dependence on Λ^2. In order to improve the large deviation properties and hence the reliability of the CSPA method, we will present a two-phase CSPA method by introducing a post-optimization phase to select the output solution from a short list of solutions grouped into T subsets $P_1,\ldots,P_T$, generated by several runs of the CSPA method with warm-starting. Then in the post-optimization phase, we define $u^{(j)}=(x^{(j)},y^{(j)})\ \forall j=1,\ldots,T$ as the one with the smallest value of $\bar{g}(x),\forall x\in P_j$, and the best solution $\bar{u}$ as the one with the smallest value of $\bar{\phi}(u^{(j)}), j=1,...,S$. It is different from the two-phase procedure for nonconvex stochastic programming by Ghadimi and Lan[15], where one does not need to take into account the violation of constraints $g(x)$. This two-phase CSPA algorithm is described in Algorithm 3.

In Algorithm 3, the number of samples of ξ is given by $T^2\times N$ and $T^2\times S$, respectively, for the optimization and post-optimization phase. Theorem 22 estimates the bounds on the total number of samples required for

Algorithm 3 A two-phase CSPA algorithm

Input: initial point $u_1 \equiv (x_1, y_1)$, number of runs T^2, and sample size S.

Optimization phase

for $t = 1, 2, \ldots, T^2$

Call the CSPA method with initial point $\widehat{u}_{t-1}$ where $\widehat{u}_0 = u_1$ and $\widehat{u}_t \equiv (x_{R_t}, y_{R_t})$ are the outputs of the t-th run of the CSPA algorithm, iteration limit N, stepsize $\{\gamma_k\}$, tolerance $\{\eta_k\}$ and the probability mass function P_R in (3.3).

end for

Post-optimization phase

Divide all the candidate solutions $\{\widehat{u}_1, \widehat{u}_2, \ldots, \widehat{u}_{T^2}\}$ into T sets: $P_1 := \{\widehat{u}_1, \ldots, \widehat{u}_T\}$, $P_2 := \{\widehat{u}_{T+1}, \ldots, \widehat{u}_{2T}\}$,$\ldots$,$P_T := \{\widehat{u}_{(T-1)T+1}, \ldots, \widehat{u}_{T^2}\}$.

1) Choose a solution $u^{(j)} \equiv (x^{(j)}, y^{(j)})$ from P_j such that

$$\bar{g}(x^{(j)}) = \min_{u \equiv (x,y) \in P_j} \bar{g}(x), \forall j = 1, \ldots, T,$$

where $\bar{g}(x) := \sum_{k=1}^{S} G(x, \xi_k)/S$, and then create a new set $\bar{P} = \{u^{(1)}, u^{(2)}, \ldots, u^{(T)}\}$.

2) Choose a solution $\bar{u} \equiv (\bar{x}, \bar{y})$ from $\bar{P}$ such that

$$\bar{\phi}(\bar{u}) = \min_{u \equiv (x,y) \in \bar{P}} \bar{\phi}(x, y), \text{ where } \bar{\phi}(x, y) = \tfrac{1}{S} \textstyle\sum_{k=1}^{S} \Phi(x, y, \zeta_k).$$

---

this two-phase CSPA method.

THEOREM 22. *If s, $\{\gamma_k\}$ and $\{\eta_k\}$ are set according to Corollary 17, then*

$$\text{Prob}\left(\phi(\bar{x}, \bar{y}) - \phi(\bar{x}, y^*(\bar{x})) \leq \tfrac{16 M_\Phi D_Y}{\sqrt{N}} \max\{\nu, \tfrac{1}{\nu}\} + \tfrac{2\lambda_1 \sigma}{\sqrt{S}}\right) \geq 1 - \left(2^{-T} + \tfrac{T+1}{\lambda_1}\right), \tag{3.33}$$

$$\text{Prob}\left(g(\bar{x}) \leq \tfrac{64 \log 2 D_X M_G}{\sqrt{N}} + \tfrac{2\lambda_2 \sigma}{\sqrt{S}}\right) \geq 1 - \left(2^{-T} + \tfrac{T+1}{\lambda_2}\right). \tag{3.34}$$

Proof. Denote $\epsilon_{j,k} := \Phi(u^{(j)}, \zeta_k) - \phi(u^{(j)})$, $\delta_{t,k} := G(x_t, \xi_k) - g(x_t)$, $\bar{u} := (\bar{x}, \bar{y})$ and $u^* := (\bar{x}, y^*(\bar{x}))$. We first show that (3.33) holds. By the definition of $\bar{\phi}$ and $\bar{u}$, we have

$$\begin{aligned}
|\bar{\phi}(\bar{u}) - \phi(u^*)| &= \min_j |\phi(u^{(j)}) - \phi(u^*) + \bar{\phi}(u^{(j)}) - \phi(u^{(j)})| \\
&\leq \min_j \{|\phi(u^{(j)}) - \phi(u^*)| + |\bar{\phi}(u^{(j)}) - \phi(u^{(j)})|\} \\
&\leq \min_j |\phi(u^{(j)}) - \phi(u^*)| + \max_j |\bar{\phi}(u^{(j)}) - \phi(u^{(j)})|.
\end{aligned}$$

Using the above observation and the triangle inequality, we have

$$\begin{aligned}
|\phi(\bar{u}) - \phi(u^*)| &\leq |\bar{\phi}(\bar{u}) - \phi(u^*)| + |\phi(\bar{u}) - \bar{\phi}(\bar{u})| \\
&\leq \min_j |\phi(u^{(j)}) - \phi(u^*)| + \max_j |\bar{\phi}(u^{(j)}) - \phi(u^{(j)})| + |\phi(\bar{u}) - \bar{\phi}(\bar{u})|.
\end{aligned}$$

We now provide certain probabilistic upper bounds to the three terms in the right hand side of the above inequality. Firstly, using the fact that $\bar{\phi}(u^{(j)}) - \phi(u^{(j)}) = \frac{1}{S} \sum_{k=1}^{S} \epsilon_{j,k}$, and $\mathbb{E}[|\epsilon_{j,k}|^2] \leq \sigma^2$, we have

$$\mathbb{E}[|\bar{\phi}(u^{(j)}) - \phi(u^{(j)})|^2] = \tfrac{1}{S^2} \mathbb{E}[\textstyle\sum_{k=1}^{S} \epsilon_{j,k}^2] \leq \tfrac{\sigma^2}{S},$$

which, by Markov's inequality, implies that

$$\text{Prob}\left(|\bar{\phi}(u^{(j)}) - \phi(u^{(j)})| \ge \tfrac{\lambda\sigma}{\sqrt{S}}\right) \le \tfrac{1}{\lambda},\ \forall j = 1, \ldots, T, \tag{3.35}$$

and that

$$\text{Prob}\left(|\phi(\bar{u}) - \bar{\phi}(\bar{u})| \ge \tfrac{\lambda\sigma}{\sqrt{S}}\right) \le \tfrac{1}{\lambda}. \tag{3.36}$$

Using this observation, we have

$$\text{Prob}\left(\max_j |\bar{\phi}(u^{(j)}) - \phi(u^{(j)})| \ge \tfrac{\lambda\sigma}{\sqrt{S}}\right) \le \textstyle\sum_j \text{Prob}\left(|\bar{\phi}(u^{(j)}) - \phi(u^{(j)})| \ge \tfrac{\lambda\sigma}{\sqrt{S}}\right) \le \tfrac{T}{\lambda}. \tag{3.37}$$

Secondly, from Corollory 17, we have

$$\mathbb{E}[\phi(u^{(j)}) - \phi(u^*)] \le \tfrac{8M_\Phi D_Y}{\sqrt{N}} \max\{\nu, \tfrac{1}{\nu}\}.$$

Let E_t be the event that $|\phi(u^{(t)}) - \phi(u^*)| \ge \frac{16M_\Phi D_Y}{\sqrt{N}} \max\{\nu, \frac{1}{\nu}\}$. Using Markov's inequality(taking $\lambda = 2$), we have

$$\text{Prob}\left(E_t | \cap_{i=1}^{t-1} E_i\right) \le \tfrac{1}{2},\ t = 1, \ldots, T,$$

which implies that

$$\text{Prob}\left(\min_j |\phi(u^{(j)}) - \phi(u^*)| \ge \tfrac{16M_\Phi D_Y}{\sqrt{N}} \max\{\nu, \tfrac{1}{\nu}\}\right) = \text{Prob}(\cap_{i=1}^{T} E_i) = \prod_{i=1}^{T} \text{Prob}(E_i | \cap_{k=1}^{i-1} E_k) \le 2^{-T}. \tag{3.38}$$

Combining (3.36),(3.37) and(3.38), we have

$$\text{Prob}\left(\phi(\bar{u}) - \phi(u^*) \le \tfrac{16M_\Phi D_Y}{\sqrt{N}} \max\{\nu, \tfrac{1}{\nu}\} + \tfrac{2\lambda\sigma}{\sqrt{S}}\right) \ge 1 - \left(2^{-T} + \tfrac{T+1}{\lambda}\right).$$

Then, we show (3.34) holds. Observe that

$$\begin{aligned} |\bar{g}(x^{(j)})| &= \min_{(x,y)\in P_j} |g(x) + \bar{g}(x) - g(x)| \\ &\le \min_{(x,y)\in P_j} \{|g(x)| + |\bar{g}(x) - g(x)|\} \\ &\le \min_{(x,y)\in P_j} |g(x)| + \max_{(x,y)\in P_j} |\bar{g}(x) - g(x)|. \end{aligned}$$

By the triangle inequality and the above observation, we have

$$\begin{aligned} &|g(x^{(j)})| \le |\bar{g}(x^{(j)})| + |g(x^{(j)}) - \bar{g}(x^{(j)})| \\ &\le \min_{(x,y)\in P_j} |g(x)| + \max_{(x,y)\in P_j} |\bar{g}(x) - g(x)| + |g(x^{(j)}) - \bar{g}(x^{(j)})|. \end{aligned}$$

Using the above inequality and the argument similar to the proof for (3.33), we can show (3.34). ■

Let $\epsilon > 0$ and $\Lambda \in (0, 1)$ be given. If parameters (T, N, S) are set to

$$T = \log \tfrac{2}{\Lambda},\ N = \tfrac{32^2 D_Y^2 M_\Phi^2 \max\{\nu^2, 1/\nu^2\}}{\epsilon^2} \text{ and } S = \tfrac{64(T+1)^2\sigma^2}{\epsilon^2\Lambda^2}, \tag{3.39}$$

then the total number of required samples, given by $T^2(N + S)$, for finding an (ϵ, Λ)-solution of problem (1.3)-(1.4), after disregarding a few constant factors, can be bounded by

$$\mathcal{O}\left\{\tfrac{1}{\epsilon^2}(\log \tfrac{1}{\Lambda})^2 + \tfrac{1}{\epsilon^2\Lambda^2}(\log \tfrac{1}{\Lambda})^4 + \tfrac{1}{\epsilon^2\Lambda^2}(\log \tfrac{1}{\Lambda})^3 + \tfrac{1}{\epsilon^2\Lambda^2}(\log \tfrac{1}{\Lambda})^2\right\}. \tag{3.40}$$

Although this bound is still in the order of $\mathcal{O}(1/(\epsilon^2\Lambda^2))$, the number of samples required for the optimization phase, given by $T^2 \times N$, has been significantly reduced to $\mathcal{O}((\log \frac{1}{\Lambda})^2/\epsilon^2)$. We will show the bound obtained in (3.40) can be further improved under the following light-tail assumption.

ASSUMPTION 4.

$$\mathbb{E}[\exp\{(\Phi(x,y,\zeta)-\phi(x,y))^2/\sigma^2\}] \le \exp\{1\},$$
$$\mathbb{E}[\exp\{(G(x,\xi)-g(x))^2/\sigma^2\}] \le \exp\{1\}.$$

THEOREM 23. *Under Assumption 4, if s, $\{\gamma_k\}$ and $\{\eta_k\}$ are set according to Corollary 17, then the following statements hold.*

$$\text{Prob}\left(\phi(\bar{x},\bar{y})-\phi(\bar{x},y^*(\bar{x})) \le \tfrac{16M_\Phi D_Y}{\sqrt{N}}\max\{\nu,\tfrac{1}{\nu}\}+\tfrac{2\lambda\sigma}{\sqrt{S}}\right) \ge 1-\left(2^{-T}+(1+T)\exp\{\tfrac{-\lambda^2}{3}\}\right), \tag{3.41}$$

$$\text{Prob}\left(g(\bar{x}) \le \tfrac{64\log 2D_X M_G}{\sqrt{N}}+\tfrac{2\lambda\sigma}{\sqrt{S}}\right) \ge 1-\left(2^{-T}+(1+T)\exp\{\tfrac{-\lambda^2}{3}\}\right). \tag{3.42}$$

Proof. The proof is similar to the one for Theorem 22 except that we use Lemma 12 to bound $\text{Prob}\left(|\bar{\phi}(u^{(j)})-\phi(u^{(j)})| \ge (\lambda\sigma)/\sqrt{S}\right)$. Hence the details are skipped. ∎

Let $\epsilon > 0$ and $\Lambda \in (0,1)$ be given. In view of Corollary 17, if parameters (T,N,S) are set to

$$T=\log\tfrac{2}{\Lambda},\ \ N=\tfrac{32^2M_\Phi^2D_Y^2\max\{\nu^2,1/\nu^2\}}{\epsilon^2}\ \text{and}\ S=\tfrac{48\sigma^2(\log(1+T)/\Lambda)^2}{\epsilon^2},$$

then the total number of required samples of ξ, given by $T^2(N+S)$, for finding an (ϵ,Λ)-solution defined in (3.31) can be bounded by

$$\mathcal{O}\left\{\frac{1}{\epsilon^2}(\log\frac{1}{\Lambda})^2\right\} \tag{3.43}$$

after disregarding a few constant factors. In comparison with the results in (3.32) and (3.40), this bound is considerably smaller in terms of its dependence on Λ.

4. Numerical Experiment. In this section, we present some numerical results of our computational experiments for solving two problems: an asset allocation problem with conditional value at risk (CVaR) constraint and a parameterized classification problem. More specifically, we report the numerical results obtained from the CSA and CSPA method applied to these two problems in Subsection 4.1 and 4.2, respectively.

4.1. Asset allocation problem. Our goal of this subsection is to examine the performance of the CSA method applied to the CVaR constrained problem in (1.2). Apparently, there is one problem associated with applying the CSA algorithm to this model – the feasible region X is unbounded. Lan, Nemirovski and Shapiro (see [21] Section 4.2) show that τ can be restricted to

$$\left[\underline{\mu}+\sqrt{\tfrac{\beta}{1-\beta}}\sigma, \bar{\mu}+\sqrt{\tfrac{1-\beta}{\beta}}\sigma\right],$$

where $\underline{\mu} := \min_{y\in Y}\{-\bar{\xi}^T y\}$ and $\bar{\mu} := \max_{y\in Y}\{-\bar{\xi}^T y\}$.

In this experiment, we consider four instances. The first three instances are randomly generated according to the factor model in Goldfarb and Iyengar (see Section 7 of [16]) with different number of stocks (500, 1000 and 2000), while the last instance consists of the 95 stocks from $S\&P100$ (excluding SBC, ATI, GS, LU and VIA-B) obtained from [38], the mean $\bar{\xi}$ and covariance Σ are estimated by the historical monthly data from 1996 to 2002. The reliability level $\beta = 0.05$ and the number of samples used to evaluate the solution is $n = 50,000$. For SAA algorithm, the deterministic SAA problem to (1.2) is defined by

$$\begin{array}{ll} \min_{x,\tau} & -\mu^T x \\ \text{s.t.} & \tau+\frac{1}{\beta N}\sum_{i=1}^N[-\xi_i^T x-\tau]_+ \le 0, \\ & \sum_{i=1}^n x_i = 1, x \ge 0, \end{array} \tag{4.1}$$

We implemented the SAA approach by using Polyak's subgradient method for solving convex programming problems with functional constraints (see [27, 3]). The main reasons why we did not use the linear programming (LP) method to (4.1) include: 1) problem (4.1) might be infeasible for some instances; and 2) we tried the LP method with CVX toolbox for an instance with 500 stocks and the CPU time is thousands times larger than that of the CSA method. In our experiment, we adjust the stepsize strategy by multiplying γ_k and η_k with some scaling parameters c_g and c_e, respectively. These parameters are chosen as a result of pilot runs of our algorithm (see [21] for more details). We have found that the "best parameters" in Table 4.1 slightly outperforms other parameter settings we have considered.

TABLE 4.1
The stepsize factor

		best c_g	best c_e
Number of stocks	500	0.5	0.005
	1000	0.5	0.05
	2000	0.5	0.05

Notations in Tables 4.2-4.5.

N: the sample size(the number of steps in SA, and the size of the sample used to build the stochastic average in SAA).
Obj.: the objective function value of our solution, i.e. the loss of the portfolio.
Cons.: the constraint function value of our solution.
CPU: the processing time in seconds for each method.

TABLE 4.2
Random Sample with 500 Assets

		N=500	N=1000	N=2000	N=5000
CSA	Obj.	-4.698	-4.658	-4.867	-4.856
	Cons.	2.009e+00	1.754e+00	2.411e+00	1.646e+00
	CPU	1.524e-01	2.706e-01	5.505e-01	1.390e+00
SAA	Obj.	-4.977	-4.977	-4.984	-4.984
	Cons.	1.442e+01	9.178e+00	6.807e+00	4.538e+00
	CPU	2.767e+00	1.205e+01	4.481e+01	2.668e+02

TABLE 4.3
Random Sample with 1000 Assets

		N=500	N=1000	N=2000	N=5000
CSA	Obj.	-4.531	-4.665	-4.634	-4.883
	Cons.	1.559e+01	1.549e+01	1.567e+01	1.786e+01
	CPU	2.180e-01	4.319e-01	9.025e-01	2.111e+00
SAA	Obj.	-4.994	-4.994	-4.994	-4.994384
	Cons.	3.530e+01	3.701e+01	2.979e+01	1.908e+01
	CPU	5.460e+00	2.345e+01	1.106e+02	6.935e+02

TABLE 4.4
Random Sample with 2000 Assets

		N=500	N=1000	N=2000	N=5000
CSA	Obj.	-4.266	-4.188	-4.598	-4.374
	Cons.	2.221e+02	1.422e+02	9.315e+01	6.167e+01
	CPU	3.745e-01	7.499e-01	1.527e+00	3.704e+00
SAA	Obj.	-4.982	-4.987	-4.987	-4.994
	Cons.	2.393e+02	2.387e+02	1.712e+02	1.124e+02
	CPU	1.545e+01	7.755e+01	3.079e+02	2.054e+03

The following conclusions can be made from the numerical results. First, as far as the quality of solutions is concerned, the CSA method is at least as good as SAA method and it may outperform SAA for some

TABLE 4.5
Comparing the CSA and SAA for the CVaR model

		N=500	N=1000	N=2000	N=5000	N=10000
CSA	Obj.	-3.531	-3.537	-3.542	-3.548	-3.560
	Cons.	3.382e+00	2.188e-01	1.106e-01	2.724e-01	-7.102e-01
	CPU	8.315e-02	1.422e-01	2.778e-01	7.251e-01	1.415e+00
SAA	Obj.	-3.530	-3.541	-3.541	-3.544	-3.559
	Cons.	3.385e+00	7.163e-01	6.989e-01	6.988e-01	7.061e-01
	CPU	3.155e+00	1.221e+01	4.834e+01	3.799e+02	1.462e+03

instances. Second, the CSA method can significantly reduce the processing time than SAA method for all the instances.

4.2. Classification and metric learning problem. In this subsection, our goal is to examine the efficiency of the CSPA algorithm applied to a classification problem with the metric as parameter. In this experiment, we use the expectation of hinge loss function, described in [34], as objective function, and formulate the constraint with the loss function of metric learning problem in [8], see formal definition in (1.5)-(1.6). For each i, j, we are given samples $u_i, u_j \in \mathbb{R}^d$ and a measure $b_{ij} \geq 0$ of the similarity between the samples u_i and u_j ($b_{ij} = 0$ means u_i and u_j are the same). The goal is to learn a metric A such that $\langle (u_i - u_j), A(u_i - u_j)\rangle \approx b_{ij}$, and to do classification among all the samples u projected by the learned metric A.

For solving this class of problems in machine learning, one widely accepted approach is to learn the metric in the first step and then solve the classification problem with the obtained optimal metric. However, this approach is not applicable to the online setting since once the dataset is updated with new samples, this approach has to go through all the samples to update A and ω. On the other hand, the CSPA algorithm optimizes the metric A and classifier ω simultaneously, and only needs to take one new sample in each iteration.

In this experiment, our goal is to test the solution quality of the CSPA algorithm with respect to the number of iterations. More specifically, we consider 2 instances of this problem with different dimension ($d = 100$ and 200, respectively). Since we are dealing with the online setting, our sample size for training A and ω is increasing with the number of iterations. The size for the sample used to estimate the parameters and the one used to evaluate the quality of solution (or testing sample) are set to 100 and $10,000$, respectively. Also, the parameters T and S in the 2-phase CSPA are given by (3.39). Within each trial, we test the objective and constraint value of the output solution over training sample and testing sample, respectively.

TABLE 4.6
Convergence with a small number of samples

Iteration Number	$\phi(\omega, A)$	g(A)
1	1.927	33.205
20	0.322	33.205
200	0.313	32.620

In Table 4.6, $\phi(\omega, A)$ and $g(A)$, respectively, denote the estimated objective and constraint values when the testing sample size $n = 500$ and $d = 100$. Table 4.6 shows that the CSPA method converges very fast even when iteration number is less than 200. Table 4.7 and Table 4.8 shows the CSPA method decreases the objective value and constraint value as the sample size (number of iterations N) increases. These experiments demonstrate that we can improve both the metric and the classifier simultaneously by using the CSPA method as more and more data are collected.

Notations in Table 4.7 and 4.8.

Obj. Train: The objective function value using training sample at the output solution.
Cons. Train: The constraint function value using training sample at the output solution.
Obj. Test: The objective function value using testing sample at the output solution.
Cons. Test: The constraint function value using testing sample at the output solution.

TABLE 4.7
d = 100

N	Obj. Train	Cons. Train	Obj. Test	Cons. Test
200	0.316	33.484	0.329	33.407
500	0.311	23.636	0.314	24.316
1000	0.294	16.722	0.303	16.936
2000	0.289	10.970	0.301	11.006
5000	0.288	5.306	0.287	5.222

TABLE 4.8
d = 200

N	Obj. Train	Cons. Train	Obj. Test	Cons. Test
200	1.837	3.709	1.925	3.734
500	1.109	3.937	1.026	3.749
1000	1.361	3.494	1.342	3.498
2000	0.811	3.458	0.824	3.415
5000	0.756	3.390	0.753	3.433

5. Conclusions. In this paper, we present a new stochastic approximation type method, the CSA method, for solving the stochastic convex optimization problems with functional expectation constraints. Moreover, we show that a variant of CSA method, the CSPA method, is applicable to a class of parameterized stochastic problem in (1.3)-(1.4). We show that these methods exhibit theoretically optimal rate of convergence for solving a few different classes of expectation constrained stochastic optimization problems and demonstrated their effectiveness through some preliminary numerical experiments.

REFERENCE

[1] Alfred Auslender and Marc Teboulle. Interior gradient and proximal methods for convex and conic optimization. *SIAM Journal on Optimization*, 16(3):697–725, 2006.

[2] H.H. Bauschke, J.M. Borwein, and P.L. Combettes. Bregman monotone optimization algorithms. *SIAM Journal on Controal and Optimization*, 42:596–636, 2003.

[3] Amir Beck, Aharon Ben-Tal, Nili Guttmann-Beck, and Luba Tetruashvili. The comirror algorithm for solving nonsmooth constrained convex problems. *Operations Research Letters*, 38(6):493–498, 2010.

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

[5] Lev M Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics*, 7(3):200–217, 1967.

[6] O. Chapelle, B. Scholkopf, and A. Zien. *Semi-supervised learning*. MIT Press, Cambridge, Mass., 2006.

[7] 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.

[8] J. C. Duchi, P. L. Bartlett, and M. J. Wainwright. Randomized smoothing for stochastic optimization. *SIAM Journal on Optimization*, 22:674–701, 2012.

[9] J. C. Duchi, S. Shalev-shwartz, Y. Singer, and A. Tewari. Composite objective mirror descent. *In Proceedings of the Twenty Third Annual Conference on Computational Learning Theory*, 2010.

[10] Y. Ermoliev. Stochastic quasigradient methods and their application to system optimization. *Stochastics*, 9:1–36, 1983.

[11] A. Gaivoronski. Nonstationary stochastic programming problems. *Kybernetika*, 4:89–92, 1978.

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

[13] 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.

[14] S. Ghadimi and G. Lan. Accelerated gradient methods for nonconvex nonlinear and stochastic optimization. Technical report, Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL 32611, USA, June 2013.

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

[16] Donald Goldfarb and Garud Iyengar. Robust portfolio selection problems. *Mathematics of Operations Research*, 28(1):1–38, 2003.

[17] Hao Jiang and Uday V Shanbhag. On the solution of stochastic optimization and variational problems in imperfect information regimes. *arXiv preprint arXiv:1402.1457*, 2014.

[18] A. J. Kleywegt, A. Shapiro, and T. Homem de Mello. The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization*, 12:479–502, 2001.

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

[20] G. Lan, Z. Lu, and R. D. C. Monteiro. Primal-dual first-order methods with $\mathcal{O}(1/\epsilon)$ iteration-complexity for cone programming. *Mathematical Programming*, 126:1–29, 2011.

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

[22] A. Nedic. On stochastic subgradient mirror-descent algorithm with weighted averaging. Technical report, 2012.

[23] Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

[24] Arkadi Nemirovski and Alexander Shapiro. Convex approximations of chance constrained programs. *SIAM Journal on Optimization*, 17(4):969–996, 2006.

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

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

[27] Boris T Polyak. A general method of solving extremum problems. *Doklady Akademii Nauk SSSR*, 174(1):33, 1967.

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

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

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

[31] R.T. Rockafellar and S.P. Uryasev. Optimization of conditional value-at-risk. *The Journal of Risk*, 2:21–41, 2000.

[32] A. Ruszczyński and W. Sysk. A method of aggregate stochastic subgradients with on-line stepsize rules for convex stochastic programming problems. *Mathematical Programming Study*, 28:113–131, 1986.

[33] M. Schmidt, N. L. Roux, and F. Bach. Minimizing finite sums with the stochastic average gradient. Technical report, September 2013.

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

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

[36] James C Spall. *Introduction to stochastic search and optimization: estimation, simulation, and control*, volume 65. John Wiley & Sons, 2005.

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

[38] Wei Wang and Shabbir Ahmed. Sample average approximation of expected value constrained stochastic programs. *Operations Research Letters*, 36(5):515–519, 2008.

[39] L. Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *J. Mach. Learn. Res.*, pages 2543–2596, 2010.