ROBINS-MONRO AUGMENTED LAGRANGIAN METHOD FOR STOCHASTIC CONVEX OPTIMIZATION *

RUI WANG† AND CHAO DING ‡

Abstract. In this paper, we propose a Robbins-Monro augmented Lagrangian method (RMALM) to solve a class of constrained stochastic convex optimization, which can be regarded as a hybrid of the Robbins-Monro type stochastic approximation method and the augmented Lagrangian method of convex optimizations. Under mild conditions, we show that the proposed algorithm exhibits a linear convergence rate. Moreover, instead of verifying a computationally intractable stopping criteria, we show that the RMALM with the increasing subproblem iteration number has a global complexity $\mathcal{O}(1/\varepsilon^{1+q})$ for the ε-solution (i.e., $\mathbb{E}\left(\|x^k - x^*\|^2\right) < \varepsilon$), where q is any positive number. Numerical results on synthetic and real data demonstrate that the proposed algorithm outperforms the existing algorithms.

Key words. stochastic convex optimization, Robbins-Monro augmented Lagrangian method, stochastic approximation, total iteration complexity

MSC codes. 90C15, 90C52, 90C25, 90C06, 65K05

1. Introduction.

In this paper, we consider the following stochastic convex optimization:

$$\begin{aligned} &\min_{x\in X} \ f(x) = f_0(x) + f_1(x), \quad f_1(x) := \mathbb{E}_\xi\left(F(x,\xi)\right), \\ &\text{s.t.} \ \ h(x) := (h_1(x),\ldots,h_M(x))^T \le 0, \end{aligned} \tag{1.1}$$

where $X \subseteq \mathbb{R}^n$ is a nonempty convex and compact set, ξ denotes the random variable whose distribution P is supported on sample space Ω and $\mathbb{E}_\xi$ is the expectation with respect to ξ, the continuously differentiable functions $f_0 : \mathbb{R}^n \mapsto \mathbb{R}$ and $h_j : \mathbb{R}^n \mapsto \mathbb{R}$, $j = 1,\ldots,M$ are convex with respect to x, and the continuously differentiable function $F : \mathbb{R}^n \times \Omega \mapsto \mathbb{R}$ is convex with respect to x for almost sure $\xi \in \Omega$. In addition, we assume that the random variable ξ is independent of x, which implies that (1.1) is a stochastic convex optimization. When ξ is distributed uniformly on a finite set $\{\xi_1,\ldots,\xi_N\}$, the problem (1.1) reduces to the following convex optimization involving the finite-sum objective function:

$$\begin{aligned} &\min_{x\in X} \ f(x) = f_0(x) + \frac{1}{N}\sum_{i=1}^{N} F(x,\xi_i), \\ &\text{s.t.} \ \ h(x) \le 0. \end{aligned} \tag{1.2}$$

The stochastic convex optimization (1.1) appears widely in a variety of applications, including the portfolio optimization [32], the multi-stage stochastic optimization [25], and constrained deep neural networks [7]. Below we give a few concrete examples and more applications of the stochastic convex optimization (1.1) can be found from [2, 35, 36].

Stochastic convex quadratically constrained quadratic program (QCQP).

*This version: August 30, 2022.

Funding: This work was supported in part by the National Key R&D Program of China 2021YFA1000300, 2021YFA1000301, the National Natural Science Foundation of China (No. 12071464), and the Beijing Natural Science Foundation (Z190002).

†Institute of Applied Mathematics, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, P.R. China, School of Mathematical Sciences, University of Chinese Academy of Science, Beijing, P.R. China. (wangrui2020@amss.ac.cn).

‡Institute of Applied Mathematics, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, P.R. China. (dingchao@amss.ac.cn).

Consider the following stochastic convex QCQP:

$$\begin{aligned} &\min_{x\in X}\ f(x)=\mathbb{E}\big(\frac{1}{2}\left\|\xi_H x-\xi_c\right\|^2\big),\\ &\text{s.t. } h_j(x)=\frac{1}{2}x^\top Q_j x+a_j^\top x\le b_j,\ j=1,\dots,M, \end{aligned} \tag{1.3}$$

where $\xi:=(\xi_H,\xi_c)$ with $\xi_H\in\mathbb{R}^{p\times n}$ and $\xi_c\in\mathbb{R}^p$ are random variables, and the symmetric positive semidefinite matrices $Q_j\in\mathbb{R}^{n\times n}$, $a_j\in\mathbb{R}^n$, and $b_j\in\mathbb{R}$, $j=1,\dots,M$ are deterministic. Clearly, this stochastic QCQP is of the form of (1.1).

Two-stage stochastic program. Consider the following two-stage stochastic program:

$$\min_{x\in X}\ f\left(x\right)=f_0\left(x\right)+f_1\left(x\right), \tag{1.4}$$

with $X\subset\mathbb{R}^n$ a nonempty convex and compact set, f_0 is a convex continuously differentiable function, and $f_1\left(x\right)=\mathbb{E}_\xi\left(F\left(x,\xi\right)\right)$ where $\xi:=(\xi_f,\xi_g,\xi_A,\xi_B,\xi_b)$ with the random matrices $\xi_A\in\mathbb{R}^{d\times n}$ and $\xi_B\in\mathbb{R}^{d\times m}$, and random vectors $\xi_f\in\mathbb{R}^p$, $\xi_g\in\mathbb{R}^q$ and $\xi_b\in\mathbb{R}^d$ summarizes all the random variables involved in the second stage:

$$\begin{aligned} F\left(x,\xi\right):=\min_{y\in Y}\quad & f_2\left(x,y,\xi_f\right)\\ \text{s.t.}\quad & \xi_A x+\xi_B y=\xi_b,\quad g\left(x,y,\xi_g\right)\le 0, \end{aligned} \tag{1.5}$$

where $Y\subset\mathbb{R}^m$ is a nonempty convex and compact set, f_2 and g are continuously differentiable functions and jointly convex with respect to the first stage decision variable x and the second stage decision variable y. One of the method for solving the two-stage stochastic program (1.4) is the sample average approximation (SAA) method [18], see [36] for more details. By sampling $\xi_1,\xi_2,\dots,\xi_N$ in the distribution to approximate $\mathbb{E}_\xi\left(F\left(x,\xi\right)\right)$, we can rewrite the two-stage stochastic problem (1.4) as the following convex optimization involving the finite-sum objective function:

$$\begin{aligned} \min_{x\in X,y_1,\dots,y_N\in Y}\ & f_0\left(x\right)+\frac{1}{N}\sum_{i=1}^N f_2\left(x,y_i,\xi_{f_i}\right)\\ \text{s.t.}\quad & \xi_{A_i}x+\xi_{B_i}y_i=\xi_{b_i},\ g\left(x,y_i,\xi_{g_i}\right)\le 0,\quad i=1,2,\dots,N, \end{aligned} \tag{1.6}$$

where y_i represents the second stage decision corresponding to ξ_i. Thus, we may approximately solve the two-stage stochastic problem (1.4) by considering a stochastic convex optimization in the form of (1.2).

Stochastic portfolio optimization. The third motivating example is the portfolio optimization problem involving Conditional Value at Risk (CVaR). In a fundamental work [32], Rockafellar and Uryasev show that a class of asset allocation problems can be modeled as:

$$\text{CVaR}=\min_{a,x\in X}\big\{a+\frac{1}{1-p}E\left(\left[f(x,\xi)-a\right]_+\right)\big\}, \tag{1.7}$$

where f is the loss associated with the decision vector x, to be chosen from a certain subset X of $\mathbb{R}^n$, and the random vector ξ in $\mathbb{R}^m$, $p\in(0,1)$ is a safety (reliability) level chosen by users, a is a threshold of loss f. When the return on a portfolio x is the sum of the returns on the individual instruments in the portfolio, scaled by the proportions x_i. The loss is the negative of return and can be denoted as $f(x,\xi):=-(\xi_1x_1+\cdots+\xi_nx_n)=-\xi^Tx$. Let $m:=E(\xi)$ be the average return for each asset assumed to be known (or estimated). We have $\mathbb{E}\left(\xi^Tx\right)\ge R\Rightarrow m^Tx\ge R$, where R encodes a minimum desired return. The feasible set of portfolios can be written as

$$X=\big\{x\in\mathbb{R}^n\mid\sum_{i=1}^n x_i=1,x\ge 0,-m^Tx\le -R\big\}, \tag{1.8}$$

which is a nonempty convex and compact set. To minimize (1.7) concerning x and a, we approximate the expectation in (1.7) by sampling $\xi_1, \xi_2, \ldots, \xi_N$ with respect to the distribution of ξ, then we obtain the approximate problem:

$$\min_{a,x\in X} \ a + \frac{1}{(1-p)N}\sum_{i=1}^{N}[-\xi_i^T x - a]_+. \tag{1.9}$$

Introducing auxiliary variables y_i for $i = 1, 2, \ldots, N$, it is equivalent to minimizing the problem:

$$\begin{aligned} \min_{a,x\in X,y} \ & a + \frac{1}{(1-p)N}\sum_{i=1}^{N} y_i \\ \text{s.t.} \quad & y_i \geq -\xi_i^T x - a, \quad y_i \geq 0, i = 1, \ldots, N. \end{aligned} \tag{1.10}$$

This problem can be reduced to (1.2).

The stochastic approximation (SA) is an efficient approach to solving the unconstrained stochastic convex optimization. It is firstly proposed by Robbins and Monro [27] in 1951 for strongly convex unconstrained stochastic optimization. Recently, the SA-type algorithms have become popular even beyond the optimization community. This trend can be credited to some extent to the exciting developments in emerging fields such as machine learning [5, 24, 38]. In terms of convergence analysis, for the stochastic convex optimization with easy projection constraint, Nemirovski et al. [23] show that, under the assumption of strong convexity, the SA algorithm exhibits a rate of convergence $\mathcal{O}(1/\varepsilon)$, i.e. after $\mathcal{O}(1/\varepsilon)$ iterations, it holds that $\mathbb{E}\left(\|x^k - x^*\|^2\right) < \varepsilon$, where x^k is the k-th iterate, and x^* is the optimal point.

Recently, Lan and Zhou [19] extend the stochastic approximation idea to the stochastic convex optimization with the single deterministic/stochastic constraint and propose the cooperative stochastic approximation (CSA) method. In [19], Lan and Zhou show that the CSA method has the $\mathcal{O}(1/\varepsilon^2)$ rate of convergence for both optimality gap and constraint violation, where ε denotes the optimality gap and infeasibility. Moreover, when the objective function and constraint are both strongly convex, the rate of convergence of the CSA can be improved to $\mathcal{O}(1/\varepsilon)$. Then, Basu and Nandy [1] extend it to the stochastic convex optimization with multiple constraints. Yu et al. [42] propose an online algorithm for constrained stochastic convex optimization which achieves $\mathcal{O}(1/\sqrt{k})$ expected regret and constraint violation (see [42, Thm 1 and 2]). Moreover, Nemirovski et al. [23] propose the saddle-point mirror stochastic approximation (MSA) to solve the convex-concave stochastic saddle-point problems. It includes the constrained stochastic convex optimization as a special case if certain constraint qualifications hold and achieves an ergodic convergence rate $\mathcal{O}(1/\sqrt{k})$ in terms of the primal-dual gap. The convergence analysis of the above methods mainly focuses on the objective function value and constraint violation. For the iterative sequence convergence property, the penalized stochastic gradient (PSG) method proposed by Xiao [39] owns the convergence rates $\mathcal{O}(1/k^{1/4})$ about $\mathbb{E}(\|\bar{x}^k - x^*\|^2)$ under the restricted strong convexity assumption, where $\bar{x}^k$ is obtained by weighted average. For the general constrained stochastic convex optimization problem with the deterministic/stochastic constraints, Boob et al. [3] propose a primal-dual proximal gradient based method, so-called constraint extrapolation (ConEx) method, in which the linear approximations of the constraint functions are used to define the extrapolation (or acceleration) step. Under the strong convexity assumption on the objective function, we know from [3, Theorem 1] that the ConEx method exhibits a $\mathcal{O}(1/\varepsilon)$ rate of convergence in terms of $\mathbb{E}(\|x^k - x^*\|^2)$ for solving the stochastic convex optimization problem (1.1). In this paper, we propose a Robbins-Monro augmented Lagrangian method for the stochastic convex optimization (1.1), which can be regarded as a hybrid of the stochastic approximation and traditional augmented Lagrangian method. Similar as the ConEx method proposed by [3], our augmented Lagrangian based algorithm exhibits a $\mathcal{O}(1/\varepsilon^{1+q})$ rate of convergence in terms of $\mathbb{E}(\|x^k - x^*\|^2)$ provided only the strong convexity of the objective function, where $q > 0$ is an arbitrarily given number.

The classical augmented Lagrangian method (ALM) proposed by Hestenes [17] and Powell [26] as an algorithm for constrained optimization. Recently, the ALM has been recognized as an efficient method for solving many optimization problems [45, 20, 46], due to its fast linear convergence rate [30, 9, 10]. Furthermore, it also has some combinations for the constrained stochastic convex optimization. Zhang et al. [44] propose a stochastic augmented Lagrangian-type algorithm named stochastic linearized proximal method of multipliers (SLPMM) for stochastic convex optimization, which achieves the complexity of $\mathcal{O}(1/\varepsilon^2)$ for objective reduction and constraint violation. The SLPMM builds small-scale constrained optimization problems by simultaneously sampling the objective function and constraints, then continuously solving the sampled small-scale optimization problem approximates the solution of the stochastic convex optimization. It is worth noting that the same technique has been applied in the other stochastic proximal point algorithm (SPPA) based algorithms. In fact, both Ryu and Boyd [34], and Milzarek et al. [21] propose to construct the small-scale problem for the sampled objective function and solve it sequentially by proximal point algorithm (PPA) to obtain the solution of original problem. However, for the theoretical analysis of SLPMM, authors assume that the sampled subproblem can achieve the exact solution, which may not be satisfied in practice. Xu [40] also proposes an ALM-related method, the primal-dual stochastic gradient method, which alternatively decreases the primal and dual variables in the augmented Lagrangian function. It achieves the $\mathcal{O}(1/\sqrt{k})$ convergence rate of objective reduction and constraint violation for convex case and nearly $\mathcal{O}(\log(k)/k)$ rate for strongly convex case with $\mathbb{E}(\|x^k - x^*\|^2) = \mathcal{O}(\log(k)/k)$.

Traditional inexact ALM controls the accuracy of subproblem solutions by suitable stopping criteria, resulting in the linear or asymptotic superlinear convergence rates [30]. However, verifying the stopping criteria for stochastic convex optimization is usually computationally expensive or even intractable due to the expectations. In order to solve (1.1) and explore the convergence of inexact ALM without stopping criteria, we first analyze the convergence of the ALM with random perturbations called stochastic augmented Lagrangian method (SALM). Then, we propose a Robbins-Monro augmented Lagrangian method (RM-ALM) to solve the constrained stochastic convex optimization (1.1) which is inspired by Robbins-Monro proximal point algorithm (RMPPA) [37]. In the RMPPA, Toulis et al. [37] study the unconstrained stochastic convex optimization by the PPA, which can be viewed as a dual method of the ALM [30]. It enlightens us to solve the constrained stochastic convex optimization by the ALM. However, the RMPPA uses the Robbins-Monro method to solve the PPA subproblem, and its convergence requires the subproblem iteration number to converge to infinity consistently, which is usually unpractical even for most applications. By the adequately chosen step size and subproblem iteration number, we successfully overcome this shortage in the proposed RMALM.

The RMALM can be viewed as a hybrid of the ALM and the Robbins–Monro type method. In the RMALM, we use Robbins–Monro type method to minimize the augmented Lagrangian function at each iteration of ALM, and each inexact solution can be viewed as a perturbation of the exact solution. Instead of verifying a stopping criteria, which is intractable computationally for stochastic optimizations, we show that the RMALM with the increasing subproblem iteration number has a global complexity $\mathcal{O}(1/\varepsilon^{1+q})$ for the ε-solution (i.e., $\mathbb{E}\left(\|x^k - x^*\|^2\right) < \varepsilon$), where q is any positive number. The contributions of this paper exist in the following several aspects.

1) We obtain the almost sure convergence of the stochastic augmented Lagrangian method (SALM). By introducing random noise to the ALM subproblem solution, the convergence properties of the results are analyzed from a probabilistic viewpoint.
2) We design a novel Robbins-Monro augmented Lagrangian method with a convergence rate arbitrarily close to $\mathcal{O}(1/\varepsilon)$. We use a stochastic algorithm to solve the ALM subproblem and avoid the verification of the stopping criteria, which is more practical. Under the strongly convex assumption of the objective function, we obtain the total iteration complexity $\mathcal{O}(1/\varepsilon^{1+q})$, where q is any positive number. It is arbitrarily close to $\mathcal{O}(1/\varepsilon)$. The complexity is shown to be comparable with the

existing related stochastic methods.

3) We show the practical performance of the proposed algorithm by testing it on the stochastic convex QCQP, a two-stage stochastic program, and a stochastic portfolio optimization problem. We compare the proposed method to the CSA method in [19], the MSA method in [23], the PDSG-adp method in [40] and the APriD method in [41]. The numerical results demonstrate that RMALM can decrease the objective value, the constraint violation, and the parameter error faster than other methods in terms of iteration number and running time.

The paper is organized as follows. Some preliminaries and the analysis of the SALM are in Section 2, the algorithm framework and convergence results are in Section 3, the numerical experiments are in Section 4, and the conclusions follow in Section 5.

2. SALM: Stochastic augmented Lagrangian method. Denote $\mathbb{R}^M_+ := \{z \in \mathbb{R}^M \mid z \geq 0\}$ and $\mathbb{R}^M_{++} := \{z \in \mathbb{R}^M \mid z > 0\}$. For any given $x \in \mathbb{R}^n$, $y \in \mathbb{R}^M_+$ and $c > 0$, the augmented Lagrangian function $L(x, y, c)$ of (1.1) takes the following form:

$$L(x,y,c) := f(x) + \frac{c}{2}\left\|\left(h(x) + \frac{y}{c}\right)_+\right\|^2 - \frac{1}{2c}\|y\|^2, \tag{2.1}$$

where $(\cdot)_+$ is the metric projection operator over the non-negative orthant. It is clear that the augmented Lagrangian function L is convex about x. Let $(x^0, y^0) \in \mathbb{R}^n \times \mathbb{R}^M_+$ be a given initial point. For the k-th iteration, the SALM for (1.1) is defined as follows

$$\hat{x}^{k+1} = \arg\min_{x \in X} L(x, y^k, c^k), \tag{2.2a}$$

$$x^{k+1} = \hat{x}^{k+1} - c^k \epsilon^{k+1}, \tag{2.2b}$$

$$y^{k+1} = \max\left\{0, y^k + c^k h(x^{k+1})\right\}, \tag{2.2c}$$

where $h(x) = (h_1(x), \ldots, h_M(x))^T$ and $c^k > 0$ is given. Let $l : \mathbb{R}^n \times \mathbb{R}^M \to [-\infty, \infty]$ be the ordinary Lagrangian function for (1.1), i.e.,

$$l(x,y) = \begin{cases} f(x) + \langle h(x), y\rangle, & \text{if } x \in X \text{ and } y \in \mathbb{R}^M_+, \\ -\infty, & \text{if } x \in X \text{ and } y \notin \mathbb{R}^M_+, \\ \infty, & \text{if } x \notin X. \end{cases} \tag{2.3}$$

Thus, for (1.1), the primal essential objective function $p : \mathbb{R}^n \to (-\infty, \infty]$ and the dual essential objective function $g : \mathbb{R}^M \to [-\infty, \infty)$ take the following forms, respectively:

$$p(x) := \sup_{y \in \mathbb{R}^M} l(x,y) \quad \text{and} \quad g(y) := \inf_{x \in \mathbb{R}^n} l(x,y). \tag{2.4}$$

It is clear from [29] that l is a closed saddle-function. Since the convexity, continuity of f, g and compactness of X, we know from [29, Corollary 37.5.2] that the mapping

$$T_l : (x,y) \to \{(v,u) \mid (v,-u) \in \partial l(x,y)\} \tag{2.5}$$

is a maximal monotone operator in $\mathbb{R}^{n+M}$. Define the inverse of T_l as a mapping $T_l^{-1} : (v,u) \to \{(x,y) \mid (v,u) \in T_l(x,y)\}$. The following definition on the Lipschitz continuity of the inverse of a maximal monotone operator is taken from [30].

DEFINITION 2.1. *For a maximal monotone operator T from a finite dimensional linear vector space $\mathcal{Z}$ to itself, we say that its inverse $T^{-1}(w) := \{z \in \mathcal{Z} \mid w \in T(z)\}$, $w \in \mathcal{Z}$ is Lipschitz continuous at the origin with modulus $a \geq 0$ if there is a unique solution $\hat{z}$ to $z = T^{-1}(0)$, and for some $\tau > 0$ we have $\|z - \hat{z}\| \leq a\|w\|$ whenever $z \in T^{-1}(w)$ and $\|w\| \leq \tau$.*

The dual essential objective function g is a proper closed concave function on $\mathbb{R}^M$. Thus, the mapping $T_g := -\partial g$ is a maximal monotone operator and the dual optimal solutions is given by $T_g^{-1}(0) := \{y \in \mathbb{R}^M \mid 0 \in T_g(y)\}$. It is well-known (cf. e.g., [30]) that the exact

ALM for (1.1) is equivalent with the following dual proximal point algorithm (PPA): for each k,

$$\hat{y}^{k+1} = P^k(y^k) = \left(I + c^k T_g\right)^{-1}(y^k) = \arg\max_{y\in\mathbb{R}^M}\left\{g(y) - \left(1/2c^k\right)\left\|y - y^k\right\|^2\right\}, \tag{2.6}$$

i.e.,

$$\hat{y}^{k+1} := \max\left\{0, y^k + c^k h(\hat{x}^{k+1})\right\}, \tag{2.7}$$

where $\hat{x}^{k+1}$ is given by (2.2a).

Next, we introduce some assumptions on the stochastic convex optimization (1.1), which will be used in the subsequent almost sure convergence analysis of the SALM.

Assumption 2.1. *The convex objective function f of* (1.1) *is strictly convex for all $x \in X$.*

The following assumption is natural and reasonable, since the set X is compact and the function h is continuous.

Assumption 2.2. *The constraint function h in* (1.1) *is Lipschitz continuous over X with parameter $L_h > 0$, i.e.,*

$$\|h(x) - h(y)\| \le L_h \|x - y\| \quad \forall\, x, y \in X.$$

It follows from (2.4) that the dual essential objective function g is given by

$$g(y) = \inf_{x\in\mathbb{R}^n} l(x,y) = \begin{cases} \inf\limits_{x\in X}\left\{f(x) + \langle h(x), y\rangle\right\} & \text{if } y \in \mathbb{R}^M_+, \\ -\infty & \text{otherwise,} \end{cases} \qquad y \in \mathbb{R}^M.$$

It is clear that l is continuously differentiable on $X \times \mathbb{R}^M_+$. Since h_j, $j = 1, \ldots, M$ are convex and X is compact, under Assumption 2.1, we know from the continuity of l that for any $y \in \mathbb{R}^M_+$, $\inf\limits_{x\in X} l(x,y)$ has the unique solution $x(y)$. Moreover, we know from [13, Theorem 9], the optimal solution mapping $x(\cdot) : \mathbb{R}^M_+ \to \mathbb{R}^n$ is continuous. Therefore, it is clear that the corresponding dual essential objective function $g(y) = f(x(y)) + \langle h(x(y)), y\rangle$, $y \in \mathbb{R}^M_+$ is also continuous. The following lemma is on the subdifferential [29] of the concave function g.

Lemma 2.2. *For each k, let $\hat{y}^k$ be given by* (2.7)*. Suppose Assumption* 2.1 *holds. Then, we have that for each k,*

$$x(\hat{y}^k) = \hat{x}^k \quad \textit{and} \quad h(\hat{x}^k) \in \partial g(\hat{y}^k), \tag{2.8}$$

where $\hat{x}^k$ is given by (2.2a)*.*

Proof. Let $y \in \mathbb{R}^M_+$ be arbitrarily given. If $y \in \mathbb{R}^M_{++}$, then since l is continuously differentiable on $X \times \mathbb{R}^M_{++}$, $x(y)$ is the unique solution of $\inf\limits_{x\in X} l(x,y)$ and g is concave , it follows from the Danskin theorem [11] (cf. e.g., [12, Theorem 10.2.1]) and [29, Theorem 25.2] that for any $y \in \mathbb{R}^M_{++}$, g is differentiable at y with the gradient

$$\nabla g(y) = \nabla_y l(x(y), y) = h(x(y)) \quad \text{and} \quad \nabla g(y) = \partial g(y). \tag{2.9}$$

If $y \in \mathbb{R}^M_+ \setminus \mathbb{R}^M_{++}$, then denote U as the set of indexes of y whose elements are equal to 0, i.e., $U := \left\{j \in \{1, \ldots, M\} \mid y_j = 0\right\}$. Define $\{\tilde{y}\} \in \mathbb{R}^M_{++}$ by

$$\tilde{y}_j := \begin{cases} \delta_u & \text{if } j \in U, \\ y_j & \text{otherwise,} \end{cases} \qquad j \in \{1, \ldots, M\},$$

where $\delta_u > 0$ for each $u \in U$. Therefore, by the continuity of $h(x(\cdot))$ over $\mathbb{R}^M_+$, we know that $\lim\limits_{\tilde{y}\to y} h(x(\tilde{y}))$ exists and equals to $h(x(y))$. Thus, by combining (2.9), we obtain that

$$\lim_{\tilde{y}\to y} \nabla g(\tilde{y}) = \lim_{\tilde{y}\to y} h(x(\tilde{y})) = h(x(y)).$$

Together with the concavity and the continuity of g on $\mathbb{R}^M_+$, we obtain that for $\forall y' \in \mathbb{R}^M$

$$\langle h(x(y)), y' - y\rangle = \lim_{\tilde{y}\to y}\langle \nabla g(\tilde{y}), y' - \tilde{y}\rangle \geq \lim_{\tilde{y}\to y} g(y') - g(\tilde{y}) = g(y') - g(y). \tag{2.10}$$

It follows from the definition of subdifferential that $h(x(y)) \in \partial g(y)$.

For each k, from the process of ALM (2.2a), since the augmented Lagrangian function L is convex about x and X is convex, we have that

$$0 \in \nabla f(\hat{x}^k) + c^{k-1}\nabla h(\hat{x}^k)^T\big(h(\hat{x}^k) + \frac{y^{k-1}}{c^{k-1}}\big)_+ + \mathcal{N}_X(\hat{x}^k), \tag{2.11}$$

where $\mathcal{N}_X(\hat{x}^k)$ denotes the normal cone of X at $\hat{x}^k$. Together with (2.7), we obtain that

$$0 \in \nabla f(\hat{x}^k) + \nabla h(\hat{x}^k)^T \hat{y}^k + \mathcal{N}_X(\hat{x}^k),$$

which implies that under the convexity of l about x,

$$\hat{x}^k = \arg\min_{x\in X} f(x) + \big\langle h(x), \hat{y}^k\big\rangle = \arg\min_{x\in X} l(x, \hat{y}^k),$$

thus $\hat{x}^k = x(\hat{y}^k)$ and $h(x(\hat{y}^k)) = h(\hat{x}^k) \in \partial g(\hat{y}^k)$, which leads to (2.8). ∎

In order to study the almost sure convergence of SALM (2.2), we make the following assumption on the random errors ϵ^k.

Assumption 2.3. *There exists a constant $\sigma > 0$ such that for each k,*

$$\mathbb{E}\big(\epsilon^k \mid \mathcal{F}^{k-1}\big) = 0 \quad \textit{and} \quad \mathbb{E}\big(\big\|\epsilon^k\big\|^2 \mid \mathcal{F}^{k-1}\big) \leq \sigma^2,$$

where $\mathcal{F}^{k-1}$ denotes the σ-algebra generated by $\{\epsilon^1, \epsilon^2, \ldots, \epsilon^{k-1}\}$.

In the following theorem, inspired by [37], we obtain the almost sure convergence of the SALM (2.2), in which a supermartingale convergence property [28, Theorem 1] plays a crucial role.

Theorem 2.3. *Suppose Assumption 2.1, Assumption 2.2 and Assumption 2.3 hold. Let $c^k = c^0 k^{-q}$ with $c^0 > 0$ and $q \in (\frac{1}{2}, 1]$. Then, if the dual problem of (1.1) has the unique solution y^*, the iterates y^k of the SALM (2.2) will converge almost surely to y^*.*

Proof. By (2.2), we know that for each k,

$$\begin{aligned}\|y^k - y^*\|^2 &= \|(y^{k-1} + c^{k-1}h(x^k))_+ - y^*\|^2 \leq \|y^{k-1} + c^{k-1}h(x^k) - y^*\|^2\\ &= \|y^{k-1} - y^*\|^2 + 2c^{k-1}\big\langle y^{k-1} - y^*, h(\hat{x}^k)\big\rangle + 2c^{k-1}\big\langle y^{k-1} - y^*, h(x^k) - h(\hat{x}^k)\big\rangle\\ &\quad + \big(c^{k-1}\big)^2\|h(x^k)\|^2. \end{aligned} \tag{2.12}$$

By taking expectations in (2.12) conditional on $\mathcal{F}^{k-1}$, we obtain that

$$\begin{aligned}\mathbb{E}\big(\|y^k - y^*\|^2|\mathcal{F}^{k-1}\big) \leq& \|y^{k-1} - y^*\|^2 + 2c^{k-1}\mathbb{E}(\big\langle y^{k-1} - y^*, h(\hat{x}^k)\big\rangle|\mathcal{F}^{k-1})\\ &+2c^{k-1}\mathbb{E}\big(\big\langle y^{k-1} - y^*, h(x^k) - h(\hat{x}^k)\big\rangle|\mathcal{F}^{k-1}\big) + (c^{k-1})^2\mathbb{E}\big(\|h(x^k)\|^2|\mathcal{F}^{k-1}\big).\end{aligned} \tag{2.13}$$

For each k, denote $R^k := \mathbb{E}\left(\big\langle y^{k-1} - y^*, h(\hat{x}^k)\big\rangle|\mathcal{F}^{k-1}\right)$. Since $\hat{x}^k$ is $\mathcal{F}^{k-1}$-measurable, we have that for each k,

$$\begin{aligned}R^k &= \big\langle y^{k-1} - y^*, h(\hat{x}^k)\big\rangle = \big\langle y^{k-1} - \hat{y}^k, h(\hat{x}^k)\big\rangle + \big\langle \hat{y}^k - y^*, h(\hat{x}^k)\big\rangle\\ &= \big\langle y^{k-1} - (y^{k-1} + c^{k-1}h(\hat{x}^k))_+, h(\hat{x}^k)\big\rangle + \big\langle \hat{y}^k - y^*, h(\hat{x}^k)\big\rangle\\ &\leq \big\langle \hat{y}^k - y^*, h(\hat{x}^k)\big\rangle. \end{aligned} \tag{2.14}$$

By the concavity of g, we know from Lemma 2.2 that for each k,

$$\langle \hat{y}^k - y^*, h(\hat{x}^k)\rangle \leq g(\hat{y}^k) - g(y^*) \leq 0,$$

which implies that $R^k \leq 0$.

On the other hand, by Assumption 2.2, Assumption 2.3 and Jensen's inequality for conditional expectations (cf. e.g., [8]), we know that

$$\begin{aligned}
&\mathbb{E}\big(\langle y^{k-1}-y^*, h(x^k)-h(\hat{x}^k)\rangle|\mathcal{F}^{k-1}\big) \leq \mathbb{E}\big(\|y^{k-1}-y^*\|\|h(x^k)-h(\hat{x}^k)\||\mathcal{F}^{k-1}\big)\\
&\leq (1+\|y^{k-1}-y^*\|^2)L_h\mathbb{E}\big(\|x^k-\hat{x}^k\||\mathcal{F}^{k-1}\big)\\
&\leq c^{k-1}(1+\|y^{k-1}-y^*\|^2)L_h\mathbb{E}\big(\|\epsilon^k\||\mathcal{F}^{k-1}\big)\\
&\leq c^{k-1}(1+\|y^{k-1}-y^*\|^2)L_h\sqrt{\mathbb{E}\big(\|\epsilon^k\|^2|\mathcal{F}^{k-1}\big)}\\
(2.15)\ &\leq c^{k-1}(1+\|y^{k-1}-y^*\|^2)L_h\sigma, \quad a.s.
\end{aligned}$$

It follows from Assumption 2.2 and the boundness of x that for each k,

$$\begin{aligned}
\|h(x^k)\|^2 &= \|h(x^k)-h(x^*)+h(x^*)\|^2 \leq 2\|h(x^k)-h(x^*)\|^2+2\|h(x^*)\|^2\\
&\leq 2L_h^2d^2+2\|h(x^*)\|^2,
\end{aligned} \tag{2.16}$$

where $d := \max\limits_{\forall x,x'\in X}\|x-x'\|$. Then take expectation in (2.16) conditional on $\mathcal{F}^{k-1}$, we conclude that for each k,

$$\mathbb{E}\big(\|h(x^k)\|^2|\mathcal{F}^{k-1}\big) \leq 2L_h^2d^2+2\|h(x^*)\|^2. \tag{2.17}$$

It then follows from (2.13), (2.14), (2.15) and (2.17) that

$$\mathbb{E}\big(\|y^k-y^*\|^2|\mathcal{F}^{k-1}\big) \leq (1+(c^{k-1})^2A_1)\|y^{k-1}-y^*\|^2+2c^{k-1}R^k+(c^{k-1})^2A_2, \quad a.s.,$$

where $A_1 := 2L_h\sigma$ and $A_2 := 2L_h^2d^2+2\|h(x^*)\|^2+2L_h\sigma$ are constants. Since the random variable R^k is non-positive, $\sum c^k=\infty$ and $\sum(c^k)^2<\infty$, by employing [28, Theorem 1], we know that $\left\|y^k-y^*\right\|^2$ converges to some $B\geq 0$ and $\sum c^{k-1}R^k > -\infty$ almost surely.

If $B\neq 0$, we have $\liminf\left\|y^k-y^*\right\|>0$ almost surely. If $\liminf R^k=0$ almost surely, by (2.14) and Lemma 2.2, we have $\liminf\left\langle\hat{y}^k-y^*, h(\hat{x}^k)\right\rangle=0$ almost surely. Thus, there exists a subsequence $\{(\hat{x}^{k_j},\hat{y}^{k_j})\}$ satisfying

$$\lim\left\langle\hat{y}^{k_j}-y^*, h(\hat{x}^{k_j})\right\rangle=0, \quad \text{almost surely.}$$

Since $\hat{y}^k$ satisfies $\hat{y}^k=P^k(y^{k-1})$ and $P^k(y^*)=y^*$ by (2.6) for all k, it then follows from the nonexpansive of P^k [31], we have that for each k,

$$\left\|\hat{y}^k-y^*\right\|^2=\|P^k(y^{k-1})-P^k(y^*)\|^2\leq\|y^{k-1}-y^*\|^2. \tag{2.18}$$

By (2.18) and $\left\|y^k-y^*\right\|^2$ converges to some $B\geq 0$ almost surely, the subsequence $\{\hat{y}^{k_j}\}$ is bounded almost surely. Then, there exists a subsequence $\{(\hat{x}^{k_{j(i)}},\hat{y}^{k_{j(i)}})\}$ and a point $(\hat{x}^*,\hat{y}^*)$ satisfying that $\lim\hat{y}^{k_{j(i)}}=\hat{y}^*$ and $\lim\hat{x}^{k_{j(i)}}=x(\hat{y}^{k_{j(i)}})=\hat{x}^*$ almost surely by Lemma 2.2. It follows the continuity of h that $\langle\hat{y}^*-y^*,h(\hat{x}^*)\rangle=0$. Since y^* is the unique solution for the dual problem of (1.1) and g is concave, we have that g is strictly concave at y^*, i.e. $\langle\hat{y}^*-y^*,h(\hat{x}^*)\rangle<0$ for $\hat{y}^*\neq y^*$, which implies $\hat{y}^*=y^*$. By Assumption 2.2 and Assumption 2.3, we have that

$$\begin{aligned}
\mathbb{E}\big(\|y^k-\hat{y}^k\|^2\big) &= \mathbb{E}\big(\mathbb{E}(\|y^k-\hat{y}^k\|^2\mid\mathcal{F}^{k-1})\big)\\
&\leq \mathbb{E}\big(\mathbb{E}(\|(y^{k-1}+c^{k-1}h(x^k))_+-(y^{k-1}+c^{k-1}h(\hat{x}^k))_+\|^2\mid\mathcal{F}^{k-1})\big)\\
&\leq (c^{k-1})^2\mathbb{E}\big(\mathbb{E}(\|h(x^k)-h(\hat{x}^k)\|^2\mid\mathcal{F}^{k-1})\big)\\
&\leq (c^{k-1})^4L_h^2\mathbb{E}\big(\mathbb{E}(\|\epsilon^k\|^2\mid\mathcal{F}^{k-1})\big)\leq(c^{k-1})^4L_h^2\sigma^2\to 0.
\end{aligned}$$

Let $k=k_{j(i)}$, it follows from Lebesgue's dominated convergence theorem that

$$\lim\mathbb{E}\big(\|y^{k_{j(i)}}-\hat{y}^{k_{j(i)}}\|^2\big)=\mathbb{E}\big(\lim\|y^{k_{j(i)}}-\hat{y}^{k_{j(i)}}\|^2\big)=\mathbb{E}\big(\lim\|y^{k_{j(i)}}-y^*\|^2\big)=0.$$

The above equation implies $\lim \|y^{k_{j(i)}} - y^*\|^2 = 0$ almost surely, which is contradictory to $\liminf \left\|y^k - y^*\right\| > 0$ almost surely. Therefore, we know that $\liminf R^k < 0$ almost surely, which implies that the series $\sum c^{k-1}R^k$ diverges almost surely since $\sum c^k = \infty$. This is contradictory to $\sum c^{k-1}R^k > \infty$ almost surely. Thus, we have $B = 0$. This completes the proof. ∎

3. RMALM: Robbins-Monro augmented Lagrangian method. We know from Theorem 2.3 that under suitable conditions, the general SALM can still be guaranteed to converge almost surely to the original solution of (1.1). In this section, we shall present a practical SALM, so-called the Robbins-Monro augmented Lagrangian method (RMALM) for the stochastic convex optimization (1.1), in which the Robbins-Monro type method [27] is employed to solve the stochastic subproblem (2.2a), inexactly. In addition, it is also worth noting that when M is large, the computation of the function value and gradient of the augmented Lagrangian function (2.1) is expensive or even intractable. To cope with this difficulty, we may rewrite the following summation in the form of an expectation, i.e.,

$$\begin{aligned}\frac{c^k}{2}\|\big(h(x)+\frac{y^k}{c^k}\big)_+\|^2 - \frac{1}{2c^k}\|y^k\|^2 &= \sum_{j=1}^{M}\frac{c^k}{2}\big(h_j(x)+\frac{y_j^k}{c^k}\big)_+^2 - \frac{1}{2c^k}{y_j^k}^2 \\ &= \frac{1}{M}\sum_{j=1}^{M}\tilde{h}(x,\zeta_j,y^k,c^k) = \mathbb{E}(\tilde{h}(x,\zeta,y^k,c^k)),\end{aligned}$$

where for each j, $\tilde{h}(x,\zeta_j,y^k,c^k) = M\big(\frac{c^k}{2}\big(h_j(x)+\frac{y_j^k}{c^k}\big)_+^2 - \frac{1}{2c^k}{y_j^k}^2\big)$ denotes the j-th item in this summation and ζ is a random variable, which follows a discrete uniform empirical density distribution Z of $\{\zeta_1,...,\zeta_M\}$. Since the independence of ξ and ζ, we may further rewrite the augmented Lagrangian function (2.1) in the expectation form, $L(x,y^k,c^k) = \mathbb{E}[\tilde{L}(x,y^k,c^k,\xi,\zeta)]$, where $\tilde{L}(x,y^k,c^k,\xi,\zeta) = f_0(x) + F(x,\xi) + \tilde{h}(x,\zeta,y^k,c^k)$. Thus, the k-th iteration of SALM (2.2) takes the following inexact form:

$$x^{k+1} \approx \arg\min_{x\in X}\mathbb{E}(\tilde{L}(x,y^k,c^k,\xi,\zeta)), \tag{3.1a}$$

$$y^{k+1} = \max\big\{0, y^k + c^k h(x^{k+1})\big\}, \tag{3.1b}$$

where the subproblem (3.1a) is solved by the Robbins-Monro type method [27]: for each k and a given integer $S^{k+1} > 1$,

$$\begin{aligned} &w_1^k = x^k, \\ &w_{s+1}^k = \mathrm{prox}_X\{w_s^k - \gamma_s^k\nabla_w\tilde{L}(w_s^k,y^k,c^k,\xi_s^k,\zeta_s^k)\}, \quad s = 1,\dots,S^{k+1}-1, \\ &x^{k+1} = w_{S^{k+1}}^k, \end{aligned} \tag{3.2}$$

where $\gamma_s^k > 0$ a given constant. Overall, the Robbins-Monro augmented Lagrangian method (RMALM) for solving the stochastic convex optimization (1.1) is summarized as Algorithm 3.1.

Traditional inexact ALM stops the subproblem by suitable stopping criteria. However, due to expectations, verifying the stopping criteria for stochastic convex optimization is usually computationally intractable. In our method, the subproblem iterates a given number of steps S^k. In the next subsection, we shall study the convergence of the RMALM with the fixed $S^k \equiv S$ and the increasing S^k, respectively.

3.1. Convergence analysis of RMALM. Since there is no guarantee that x^{k+1} is an unbiased estimate of the exact solution $\hat{x}^{k+1}$ of the augmented Lagrangian subproblem (2.2a), in stead of studying the almost sure convergence of the iteration sequence $\{y^k\}$ as in the Theorem 2.3, we shall study the convergence properties of the RMALM in the sense of expectation.

To conclude, we need the following assumption on the dual essential objective function g.

Algorithm 3.1 Robbins-Monro Augmented Lagrangian Method

Initial $x^0 \in \mathbb{R}^n$, $y^0 \in \mathbb{R}^M_+$. Define $\gamma_s^k := \tau_s \eta^k/(n+\beta)$, where $\beta > 0$ and $\{\tau_s\}$ and $\{\eta^k\}$ are bounded sequences. $\{c^k\}$ is a given positive sequence.
for $k = 0, 1, 2, \ldots, K-1$ **do**
  $w_1^k = x^k$.
  while $1 \le n \le S^{k+1} - 1$ **do**
    Random sample ξ_s^k and ζ_s^k.
    $w_{s+1}^k = \mathrm{prox}_X\{w_s^k - \gamma_s^k \nabla_w \tilde{L}(w_s^k, y^k, c^k, \xi_s^k, \zeta_s^k)\}$.
  end while
  $x^{k+1} = w_{S^{k+1}}^k$.
  $y^{k+1} = \max\left\{0, y^k + c^k h(x^{k+1})\right\}$.
end for
return x^K

ASSUMPTION 3.1. *g is a $\alpha-$strongly concave function.*

The above assumption is natural, and we will give the following examples where Assumption 3.1 may satisfy.

(1) Linear constraint: when the constraint $h(x) = Ax - b \le 0$, from [16, Proposition 2.7] we have g is strongly concave on $\mathbb{R}^M$ with strong concavity $\frac{\lambda_{\min}(AA^T)}{L_f}$, where L_f is Lipschitz constant of ∇f.

(2) Quadratically constrained quadratic program: when the problem is formulated as follows

$$\left\{ \begin{array}{ll} \min\limits_{x\in\mathbb{R}^n} & f(x) := \frac{1}{2}x^T Q_0 x + a_0^T x + b_0, \\ \text{s.t.} & h_1(x) := \frac{1}{2}x^T Q_1 x + a_1^T x + b_1 \le 0. \end{array} \right. \tag{3.3}$$

Assume that Q_0, Q_1 are positive definite, $a_0 \neq Q_0 Q_1^{-1} a_1$, and there exists x_0 such that $h_1(x_0) < 0$. Let l be any lower bound of the optimal value about (3.3), and

$$\bar{\mu} = (l - f(x_0))/h_1(x_0) \ge 0.$$

Then from [16, Proposition 2.8], the dual essential objective function g is strongly concave on the interval $[0, \bar{\mu}]$ with constant of strong concavity

$$\alpha_D = \left(Q_1^{-1/2}\left(a_0 - Q_0 Q_1^{-1} a_1\right)\right)^T \left(Q_1^{-1/2} Q_0 Q_1^{-1/2} + \bar{\mu} I_n\right)^{-3} Q_1^{-1/2}\left(a_0 - Q_0 Q_1^{-1} a_1\right) > 0.$$

(3) General nonlinear constraint: for general nonlinear constraint, we usually need the linear independence of $\{\nabla h_j(x) : j = 1, 2, \ldots M\}$ at optimal points by [16, Theorem 2.10], thus M cannot be larger than n theoretically. In practical, tests in Section 4 also show favorable experimental results on $n \ll M$.

Recall that for each k, $\hat{x}^{k+1}$ and $\hat{y}^{k+1}$ are the exact solutions of (2.2a) and (2.6). The following lemma is on the iteration error estimations between $\{\hat{x}^{k+1}, \hat{y}^{k+1}\}$ and the inexact solutions $\{x^{k+1}, y^{k+1}\}$.

LEMMA 3.1. *For each k, let $\hat{x}^{k+1}$ and $\hat{y}^{k+1}$ be the exact solutions of* (2.2a) *and* (2.6)*, respectively. Suppose Assumption 3.1 holds. Then, we have for each k,*

$$\|\hat{y}^{k+1} - y^*\|^2 \le \theta^k \|y^k - y^*\|^2, \tag{3.4}$$

where $\theta^k = \left(1 + \alpha c^k\right)^{-2} < 1$ and y^* is the unique dual optimal solution of (1.1).

Let T_l be the mapping defined by (2.5)*. If we further assume that the inverse T_l^{-1} is Lipschitz continuous at the origin with modulus $a_l > 0$, then for each k,*

$$\|\hat{x}^{k+1} - x^*\|^2 \le \theta^{k'} \|y^k - y^*\|^2, \tag{3.5}$$

where $\theta^{k\prime} = \left[\frac{(2+\alpha c^k)a_l}{c^k+\alpha(c^k)^2}\right]^2$ *and* x^* *is the unique optimal solution to* (1.1).

Proof. It follows from Assumption 3.1 and [33, Exercise 12.59] that $T_g = -\partial g$ is strongly monotone operator with modulus α. It then follows from [31, (1.15)] that there exist the unique solution y^* satisfying $0 \in T_g(y^*)$ and for each k,

$$\|\hat{y}^{k+1} - y^*\| = \|P^k\big(y^k\big) - P^k\big(y^*\big)\| \leq \big(1+\alpha c^k\big)^{-1}\|y^k - y^*\|,$$

which implies (3.4) holds.

Since that T_l^{-1} is Lipschitz continuous at origin with modulus $a_l > 0$, we know from Definition 2.1 that the unique primal and dual solution x^* and y^* of (1.1) satisfy $T_l^{-1}(0) = (x^*, y^*)$ and for each k,

$$\|(\hat{x}^{k+1}, \hat{y}^{k+1}) - (x^*, y^*)\| \leq a_l \operatorname{dist}((0,0), T_l(\hat{x}^{k+1}, \hat{y}^{k+1})). \tag{3.6}$$

For each k, donate

$$\Phi_k(x) = \begin{cases} L(x, y^k, c^k), & \text{if } x \in X, \\ \infty, & \text{otherwise.} \end{cases}$$

Then, by [30, (4.21)] we obtain that for each k,

$$\operatorname{dist}((0,0), T_l(\hat{x}^{k+1}, \hat{y}^{k+1})) \leq \big(\operatorname{dist}^2(0, \partial\Phi_k(\hat{x}^{k+1}) + (c^k)^{-2}\|\hat{y}^{k+1} - y^k\|^2\big)^{1/2}. \tag{3.7}$$

Since $\hat{x}^{k+1}$ is the optimal solution of (2.2a), we have $0 \in \partial\Phi_k(\hat{x}^{k+1})$. Then, we have for each k,

$$\operatorname{dist}\big((0,0), T_l(\hat{x}^{k+1}, \hat{y}^{k+1})\big) \leq {c^k}^{-1}\|\hat{y}^{k+1} - y^k\|.$$

This, together with (3.6), yields that for each k,

$$\|(\hat{x}^{k+1}, \hat{y}^{k+1}) - (x^*, y^*)\| \leq \frac{a_l}{c^k}||\hat{y}^{k+1} - y^k||.$$

It then follows from (3.4) that for each k,

$$\begin{aligned}\|(\hat{x}^{k+1}, \hat{y}^{k+1}) - (x^*, y^*)\| &\leq \frac{a_l}{c^k}(\|\hat{y}^{k+1} - y^*|| + ||y^k - y^*\|) \leq \frac{a_l}{c^k}(1+\sqrt{\theta^k})\|y^k - y^*\| \\ &= \frac{(2+\alpha c^k)a_l}{c^k+\alpha(c^k)^2}\|y^k - y^*\|.\end{aligned}$$

This completes the proof. ∎

Under the following assumption, we know that the augmented Lagrangian function L defined by (2.1) is also strongly convex for all $x \in X$ with modulus μ.

ASSUMPTION 3.2. *The convex objective function* f *of* (1.1) *is* $\mu-$*strongly convex for all* $x \in X$.

The following assumption on the stochastic gradients of augmented Lagrangian function L is standard in the convergence analysis of the Robbins-Monro type method.

ASSUMPTION 3.3. *Suppose that there exists a constant* $\sigma > 0$ *such that for each* k *and* $1 \leq s \leq S^{k+1} - 1$, $\nabla\tilde{L}(w, y^k, c^k, \xi_s^k, \zeta_s^k)$ *in* (3.2) *satisfies*

$$\begin{aligned}&\mathbb{E}(\nabla\tilde{L}(w, y^k, c^k, \xi_s^k, \zeta_s^k)|\mathcal{H}_{s-1}^k) = \nabla L(w, y^k, c^k), \\ &\mathbb{E}(\|\nabla\tilde{L}(w, y^k, c^k, \xi_s^k, \zeta_s^k)\|^2|\mathcal{H}_{s-1}^k) \leq \sigma^2,\end{aligned} \quad \forall\, w \in X, \tag{3.8}$$

where $\mathcal{H}_{s-1}^k$ *denotes the* σ*-algebra generated by* $\{\xi_1^0, \zeta_1^0, \ldots, \xi_{S^1-1}^0, \zeta_{S^1-1}^0, \ldots, \xi_{s-1}^k, \zeta_{s-1}^k\}$.

The following lemma is on the non-asymptotic estimation on the distance between an inexact solution x^{k+1} generated by the Robbins-Monro type method (3.2) and the exact solution $\hat{x}^{k+1}$ of (3.1a).

LEMMA 3.2. *Suppose Assumption* 3.2 *and Assumption* 3.3 *hold. For each* k *and* $1 \le s \le S^{k+1}-1$, *let* $\gamma_s^k = \tau_s \eta^k/(n+\beta)$ *with* $0 < \underline{\tau} < \tau_s < \bar{\tau}$, $1/(\mu\underline{\tau}) < \eta^k < \eta$ *and* $\beta > 2\mu\eta\bar{\tau} - 1$. *Then, for each* k, *the* k*-iteration* x^{k+1} *generated by* (3.2) *satisfies the following inequality*

$$\mathbb{E}\big(\|x^{k+1} - \hat{x}^{k+1}\|^2\big) \le \frac{v}{S^{k+1}+\beta} \tag{3.9}$$

where

$$v := \max\left\{\frac{\eta^2\bar{\tau}^2\sigma^2}{2\mu\eta\underline{\tau}-1}, (\beta+1)d^2\right\} \quad \textit{with} \quad d := \max_{\forall x,x' \in X} \|x - x'\|. \tag{3.10}$$

Proof. For each k and $1 \le s \le S^{k+1}-1$, denote $B_s^k := \left\|w_s^k - \hat{w}^k\right\|^2$ and $b_s^k := \mathbb{E}\left(B_s^k\right) = \mathbb{E}\left(\left\|w_s^k - \hat{w}^k\right\|^2\right)$, where $\hat{w}^k$ represents the optimal of (3.1a). By (3.2) and the non-expansion property of the proximal mapping and noting that $\mathrm{prox}_X(\hat{w}^k) = \hat{w}^k$, we have for each k and s,

$$\begin{aligned} B_{s+1}^k &= \|\,\mathrm{prox}_X\{w_s^k - \gamma_s^k \nabla_w \tilde{L}(w_s^k, y^k, c^k, \xi_s^k, \zeta_s^k)\} - \hat{w}^k\|^2 \\ &\le \|w_s^k - \gamma_s^k \nabla_w \tilde{L}(w_s^k, y^k, c^k, \xi_s^k, \zeta_s^k) - \hat{w}^k\|^2 \\ &= B_s^k + {\gamma_s^k}^2 \|\nabla_w \tilde{L}(w_s^k, y^k, c^k, \xi_s^k, \zeta_s^k)\|^2 - 2\gamma_s^k (w_s^k - \hat{w}^k)^T \nabla_w \tilde{L}(w_s^k, y^k, c^k, \xi_s^k, \zeta_s^k). \end{aligned} \tag{3.11}$$

Under Assumption 3.3, since w_s^k is $\mathcal{H}_{s-1}^k$-measurable, we obtain that for each k and s,

$$\begin{aligned} &\mathbb{E}\big((w_s^k - \hat{w}^k)^T \nabla_w \tilde{L}(w_s^k, y^k, c^k, \xi_s^k, \zeta_s^k)\big) \\ &= \mathbb{E}\big(\mathbb{E}((w_s^k - \hat{w}^k)^T \nabla_w \tilde{L}(w_s^k, y^k, c^k, \xi_s^k, \zeta_s^k) \mid \mathcal{H}_{s-1}^k)\big) \\ &= \mathbb{E}\big((w_s^k - \hat{w}^k)^T \mathbb{E}(\nabla_w \tilde{L}(w_s^k, y^k, c^k, \xi_s^k, \zeta_s^k) \mid \mathcal{H}_{s-1}^k)\big) \\ &= \mathbb{E}\big((w_s^k - \hat{w}^k)^T \nabla_w L(w_s^k, y^k, c^k)\big). \end{aligned} \tag{3.12}$$

It then follows from the μ-strong convexity of $L(w, y^k, c^k)$ that the minimizer $\hat{w}^k$ is unique, and for each k,

$$(w - \hat{w}^k)^T \big(\nabla_w L(w, y^k, c^k) - \nabla_w L(\hat{w}^k, y^k, c^k)\big) \ge \mu \|w - \hat{w}^k\|^2 \quad \forall\, w \in X. \tag{3.13}$$

Since $\hat{w}^k$ is the optimal of (3.1a), we have $(w - \hat{w}^k)^T \nabla_w L(\hat{w}^k, y^k, c^k) \ge 0$ for any $w \in X$. Thus, it follows from (3.13) that $(w - \hat{w}^k)^T \nabla_w L(w, y^k, c^k) \ge \mu \|w - \hat{w}^k\|^2$. Therefore, we have for each k and s,

$$\mathbb{E}\big((w_s^k - \hat{w}^k)^T \nabla_w L(w_s^k, y^k, c^k)\big) \ge \mu \mathbb{E}\big(\|w_s^k - \hat{w}^k\|^2\big) = \mu b_s^k. \tag{3.14}$$

Due to Assumption 3.3, we have that for each k and s,

$$\mathbb{E}\big(\|\nabla_w \tilde{L}(w_s^k, y^k, c^k, \xi_s^k, \zeta_s^k)\|^2\big) = \mathbb{E}\big(\mathbb{E}\big(\|\nabla_w \tilde{L}(w_s^k, y^k, c^k, \xi_s^k, \zeta_s^k)\|^2 \mid \mathcal{H}_{s-1}^k\big)\big) \le \sigma^2. \tag{3.15}$$

By taking the expectation of both sides of (3.11), we know from (3.12), (3.14) and (3.15) that for each k and s,

$$b_{s+1}^k \le (1 - 2\mu\gamma_s^k) b_s^k + (\gamma_s^k)^2 \sigma^2. \tag{3.16}$$

Next, we shall show the following inequality by induction

$$b_s^k \le \frac{V(\eta^k)}{s+\beta}, \quad s = 1, 2, \ldots \tag{3.17}$$

where $V(\eta^k) = \max\big\{\frac{(\eta^k)^2\bar{\tau}^2\sigma^2}{2\mu\eta^k\underline{\tau}-1}, (\beta+1)d^2\big\}$. In fact, for $s=1$, it is clear from the definition of $V(\eta^k)$ that (3.17) holds. Suppose (3.17) holds for $s \geq 1$. Denote $\hat{s} := s+\beta$. By noting that $1 - \dfrac{2\mu\eta^k\tau_s}{\hat{s}} \geq 0$ (since $\beta > 2\mu\eta\bar{\tau} - 1$) and $\hat{s}^2 \geq (\hat{s}+1)(\hat{s}-1)$, we obtain from (3.16) that

$$\begin{aligned} b_{s+1}^k &\leq (1 - \frac{2\mu\eta^k\tau_s}{\hat{s}})b_s^k + \frac{(\eta^k)^2\tau_s^2\sigma^2}{\hat{s}^2} \leq (1 - \frac{2\mu\eta^k\tau_s}{\hat{s}})\frac{V(\eta^k)}{\hat{s}} + \frac{(\eta^k)^2\tau_s^2\sigma^2}{\hat{s}^2} \\ &\leq (\frac{\hat{s} - 2\mu\eta^k\tau_s}{\hat{s}^2})V(\eta^k) + \frac{(\eta^k)^2\tau_s^2\sigma^2}{\hat{s}^2} \leq (\frac{\hat{s}-1}{\hat{s}^2})V(\eta^k) - \frac{2\mu\eta^k\tau_s - 1}{\hat{s}^2}V(\eta^k) + \frac{(\eta^k)^2\tau_s^2\sigma^2}{\hat{s}^2} \\ &\leq \frac{V(\eta^k)}{\hat{s}+1}. \end{aligned}$$

Thus, we know the inequality (3.17) holds.

Let $\psi(\eta^k) = \frac{(\eta^k)^2\bar{\tau}^2\sigma^2}{2\mu\eta^k\underline{\tau}-1}$. By noting that $\psi'(\eta^k) > 0$ for all $\eta^k > 1/\mu\underline{\tau}$, we have $\psi(\eta^k) < \psi(\eta)$ since $\eta^k < \eta$. Thus, we obtain that

$$b_s^k \leq \frac{v}{s+\beta}.$$

By noting that $x^{k+1} = w_{S^{k+1}}^k$ and $\hat{x}^{k+1} = \hat{w}^k$, we then obtain (3.9). This completes the proof. ∎

The following theorem is on the non-asymptotic convergence property of the sequence $\{(x^k, y^k)\}$ with a fixed subproblem iteration number, i.e., for each k, $S^k \equiv S$, where $1 < S$ is a given integer.

THEOREM 3.3. *Let $\{(x^k, y^k)\}$ be the sequence generated by Algorithm 3.1. Suppose that Assumption 2.2, Assumption 3.1, Assumption 3.2 and Assumption 3.3 hold. Let T_l be defined by (2.5). Suppose that T_l^{-1} is Lipschitz continuous at origin with modulus $a_l > 0$. For each k and $1 \leq s \leq S^{k+1} - 1$, let $\gamma_s^k = \tau_s\eta^k/(n+\beta)$ with $\underline{\tau} < \tau_s < \bar{\tau}$, $1/(\mu\underline{\tau}) < \eta^k < \eta$ and $\beta > 2\mu\eta\bar{\tau} - 1$. Let $c > 0$ be such that $\theta := (1+\alpha c)^{-2} < \frac{1}{2}$ and $c^k \equiv c$ for each k. Denote $\rho := 2\theta < 1$ and $\theta' := \big[\frac{(2+\alpha c)a_l}{c+\alpha c^2}\big]^2$. Then, we have for each k,*

$$\mathbb{E}\big(\|y^k - y^*\|^2\big) \leq \frac{2c^2L_h^2v}{S^k} + 2\theta\mathbb{E}\big(\|y^{k-1} - y^*\|^2\big) \tag{3.18}$$

and

$$\mathbb{E}\big(\|x^k - x^*\|^2\big) \leq \frac{2v}{S^k} + 2\theta'\mathbb{E}\big(\|y^{k-1} - y^*\|^2\big), \tag{3.19}$$

where v is the constant given by (3.10).

If we further assume $S^k \equiv S$ for all k with a given integer $S > 1$, then the following inequalities hold for each k,

$$\mathbb{E}\big(\|y^k - y^*\|^2\big) \leq \frac{2c^2L_h^2v}{(1-\rho)S} + \rho^k\|y^0 - y^*\|^2 \tag{3.20}$$

and

$$\mathbb{E}\big(\|x^k - x^*\|^2\big) \leq \frac{2v}{S} + \frac{4c^2L_h^2\theta'v}{(1-\rho)S} + 2\theta'\rho^{k-1}\|y^0 - y^*\|^2. \tag{3.21}$$

Proof. It follows from Lemma 3.1, Lemma 3.2 and Assumption 2.2 that for each k,

$$\begin{aligned} \mathbb{E}\big(\|y^k - y^*\|^2\big) &\leq 2\mathbb{E}\big(\|y^k - \hat{y}^k\|^2\big) + 2\mathbb{E}\big(\|\hat{y}^k - y^*\|^2\big) \\ &\leq 2\mathbb{E}\big(\|y^k - \hat{y}^k\|^2\big) + 2\theta\mathbb{E}\big(\|y^{k-1} - y^*\|^2\big) \\ &\leq 2\mathbb{E}\big(\|y^{k-1} + ch(x^k) - y^{k-1} - ch(\hat{x}^k)\|^2\big) + 2\theta\mathbb{E}\big(\|y^{k-1} - y^*\|^2\big) \\ &\leq 2c^2L_h^2\mathbb{E}\big(\|x^k - \hat{x}^k\|^2\big) + 2\theta\mathbb{E}\big(\|y^{k-1} - y^*\|^2\big) \\ &\leq 2c^2L_h^2\frac{v}{S^k} + 2\theta\mathbb{E}\big(\|y^{k-1} - y^*\|^2\big), \end{aligned}$$

which implies (3.18) holds. Using Lemma 3.1 and Lemma 3.2 again, we obtain that for each k,

$$\begin{aligned}\mathbb{E}\big(\|x^k - x^*\|^2\big) &\le 2\mathbb{E}\big(\|x^k - \hat{x}^k\|^2\big) + 2\mathbb{E}\big(\|\hat{x}^k - x^*\|^2\big)\\ &\le 2\mathbb{E}\big(\|x^k - \hat{x}^k\|^2\big) + 2\theta'\mathbb{E}\big(\|y^{k-1} - y^*\|^2\big)\\ &\le \frac{2v}{S^k} + 2\theta'\mathbb{E}\big(\|y^{k-1} - y^*\|^2\big),\end{aligned}$$

which implies (3.19) holds. In addition, if $S^k \equiv S$ for all k, since $\rho = 2\theta = 2(1+\alpha c)^{-2} < 1$, we obtain from the recursion of (3.18) that for each k,

$$\mathbb{E}\big(\|y^k - y^*\|^2\big) \le \frac{2c^2L_h^2}{1-\rho}\frac{v}{S} + \rho^k\|y^0 - y^*\|^2.$$

Thus, we know that (3.20) holds. It then follows from the recursion of (3.19) and (3.20) that for each k,

$$\mathbb{E}\big(\|x^k - x^*\|^2\big) \le \frac{2v}{S} + \frac{4c^2L_h^2\theta'}{1-\rho}\frac{v}{S} + 2\theta'\rho^{k-1}\|y^0 - y^*\|^2.$$

This completes the proof. ∎

From Theorem 3.3, we know that the RMALM (Algorithm 3.1) converges with the fixed subproblem iteration number $S^k \equiv S$. Roughly speaking, the equation (3.20) and (3.21) imply that $\mathbb{E}(\|y^k - y^*\|^2)$ and $\mathbb{E}(\|x^k - x^*\|^2)$ converge at the rate of $\mathcal{O}(1/S)$ in terms of S. Therefore, the ε-solution (i.e., $\mathbb{E}(\|x^k - x^*\|^2) < \varepsilon$, $\mathbb{E}(\|y^k - y^*\|^2) < \varepsilon$) can be obtained by choosing $k = \mathcal{O}(\log\frac{1}{\varepsilon})$ and $S = \mathcal{O}(\frac{1}{\varepsilon})$. Overall, the total iteration number to obtain a ε-solution is $\mathcal{O}(\frac{1}{\varepsilon}\log\frac{1}{\varepsilon})$. On the other hand, the convergence results obtained in Theorem 3.3 imply that the subproblem iteration number S must converge to infinity, which is unpractical. However, we do not need S^k to be very large at the beginning of the ALM algorithm. In intuition, as the algorithm proceeds, the subproblem of the ALM will require higher accuracy, which means more iterations. Therefore, we can set S^k to get progressively larger, and obtain a practical complexity by choosing a more suitable subproblem iteration number S^k. Below we present the convergence results of RMALM with the increasing S^k.

THEOREM 3.4. *Suppose the conditions in Theorem* 3.3 *hold. Let* $S^k = \lceil S^0\rho^{-k(1+q)}\rceil$, *where* $S^0 > 1$, $q > 0$ *are given constants and* $\lceil a\rceil$ *denotes the smallest integer larger than* a *for any number* a. *Let* $\{(x^k, y^k)\}$ *be the sequence generated by Algorithm* 3.1. *Then we have the linear convergence rate of* $\{y^k\}$

$$\mathbb{E}\big(\|y^k - y^*\|^2\big) = D\rho^k, \tag{3.22}$$

where $D := \dfrac{2c^2L_h^2v\rho^q}{S^0(1-\rho^q)} + \|y_0 - y^*\|^2$.

The ε*-solution for* x^k *and* y^k, *that satisfies* $\mathbb{E}(\|x^k - x^*\|^2) < \varepsilon$ *and* $\mathbb{E}(\|y^k - y^*\|^2) < \varepsilon$ *respectively, need* $\mathcal{O}((\frac{1}{\varepsilon})^{1+q})$ *iterations both.*

Proof. It follows from the recursion of (3.18) that for each k,

$$\mathbb{E}\big(\|y^k - y^*\|^2\big) \le 2c^2L_h^2v\big(\frac{1}{S^k} + \frac{\rho}{S^{k-1}} + \cdots + \frac{\rho^{k-1}}{S^1}\big) + \rho^k\|y_0 - y^*\|^2.$$

Let $S^k = \lceil S^0\rho^{-k(1+q)}\rceil$, where $S^0 > 1$ and $q > 0$ are given constants. We have

$$\begin{aligned}\mathbb{E}\big(\|y^k - y^*\|^2\big) &\le \big(2c^2L_h^2\frac{v}{S^0}(\rho^{kq} + \cdots + \rho^q) + \|y_0 - y^*\|^2\big)\rho^k\\ &\le \big(2c^2L_h^2\frac{v}{S^0}\frac{\rho^q}{1-\rho^q} + \|y_0 - y^*\|^2\big)\rho^k,\end{aligned}$$

which implies that the y^k has a linear convergence rate (3.22).

Under the requirement $\mathbb{E}(\|y^k - y^*\|^2) \le \varepsilon$, we have $k \ge (\ln \rho)^{-1} \ln \frac{\varepsilon}{D}$. Sum up the total iterations for $K = (\ln \rho)^{-1} \ln \frac{\varepsilon}{D}$, we obtain that

$$\begin{aligned}\sum_{k=1}^{K} S^k =& \sum_{k=1}^{K} \lceil S^0 \rho^{-k(1+q)} \rceil \le S^0 \sum_{k=1}^{K} [\rho^{-(1+q)}]^k + K = S^0 \frac{\rho^{-(1+q)}(1-\rho^{-(1+q)\cdot K})}{1-\rho^{-(1+q)}} + K \\ =& \frac{S^0 \cdot \rho^{-(1+q)}}{\rho^{-(1+q)}-1} \rho^{-(1+q)\cdot K} - \frac{S^0 \rho^{-(1+q)}}{\rho^{-(1+q)}-1} + K = \mathcal{O}(\rho^{-(1+q)K}) + \mathcal{O}(K) \\ =& \mathcal{O}(\rho^{-(1+q)\cdot (\ln \rho)^{-1} \ln \frac{\varepsilon}{D}}) + \mathcal{O}(\ln \frac{1}{\epsilon}) = \mathcal{O}\big((\frac{1}{\varepsilon})^{1+q}\big),\end{aligned} \tag{3.23}$$

which implies the $\mathcal{O}((\frac{1}{\varepsilon})^{1+q})$ iteration complexity of $\mathbb{E}(\|y^{k-1} - y^*\|^2)$.

It follows from (3.19) that

$$\mathbb{E}(\|x^k - x^*\|^2) \le \frac{2v}{S^k} + 2\theta' \mathbb{E}(\|y^{k-1} - y^*\|^2) \le \frac{2v}{S^0} \rho^{k(1+q)} + 2\theta' D \rho^{k-1} \le 2(\frac{v}{S^0} + \frac{\theta' D}{\rho}) \rho^k.$$

Denote $D_0 := 2(\frac{v}{S^0} + \frac{\theta' D}{\rho})$. Under the requirement $\mathbb{E}(\|x^k - x^*\|^2) \le \varepsilon$, we have $k \ge (\ln \rho)^{-1} \ln \frac{\varepsilon}{D_0}$. Similar to (3.23), sum up the total iterations for $K = (\ln \rho)^{-1} \ln \frac{\varepsilon}{D_0}$, we also obtain that

$$\sum_{k=1}^{K} S^k = \sum_{k=1}^{K} \lceil S^0 \rho^{-k(1+q)} \rceil = \mathcal{O}(\rho^{-(1+q)\cdot (\ln \rho)^{-1} \ln \frac{\varepsilon}{D_0}}) + \mathcal{O}(\ln \frac{1}{\epsilon}) = \mathcal{O}\big((\frac{1}{\varepsilon})^{1+q}\big),$$

which implies the $\mathcal{O}((\frac{1}{\varepsilon})^{1+q})$ iteration complexity of $\mathbb{E}(\|x^{k-1} - x^*\|^2)$. ∎

Two results are stated in Theorem 3.4. Firstly, by setting $S^k = \lceil S^0 \rho^{-k(1+q)} \rceil$, where $S^0 > 1$ and $q > 0$ are given constants, we can guarantee the linear convergence rate of the RMALM without the stopping criteria. Without the verifying the stopping criteria, the whole algorithm is more simpler and practical. This idea can also be used for other subproblem-solving algorithms. Moreover, we obtain the total complexity of $\mathbb{E}(\|x^k - x^*\|^2)$ is arbitrarily close to $\mathcal{O}(1/\varepsilon)$.

4. Numerical experiments. This section tests the proposed method (RMALM) on the stochastic convex QCQP, a two-stage stochastic program, and a stochastic portfolio optimization problem. We compare our method to four existing methods, the CSA method in [19], the MSA in [23], the PDSG-adp method in [40] and the APriD method in [41]. In [19], the output of CSA is the weighted average of x^t over $t \in \mathcal{B}^k = \{t = 1, 2, \dots k \mid \widehat{G}_t \le \eta_t\}$. Note that $\mathcal{B}^k$ may be empty for a small k. Therefore, we also compute the weighted average of x^t overall $t = 1, 2, \dots k$ as in [41] and name the results of CSA as CSA1, CSA2 respectively. The parameters in the stochastic convex QCQP are the same as in the experiments of [41]. Specifically, we take $s = 1$ and $J_g = 100$ in CSA; $\beta_1 = 0.9, \beta_2 = 0.99$ in APriD. In our algorithm, we set $S^k = \lceil 5 \times 1.7^{k(1+0.0001)} \rceil$. Other parameters are different in each experiment, and we take the optimal parameters according to the experiment for comparison.

In all experiments, our comparisons contain the objective value $f(x^k)$, the averaged constraint violation measured by $\frac{1}{M} \sum_{j=1}^{M} \big[h_j(x^k)\big]_+$, the maximum constraint violation measured by $\max_{j \in \{1,\dots,M\}} \big[h_j(x^k)\big]_+$, the iteration error measured by $\|x^k - x_{opt}\|^2$ and the averaged error $\|\bar{x}^k - x_{opt}\|^2$ which has the better performance for other algorithms, where x_{opt} is the optimal solution in every experiment and $\bar{x}^k$ is a kind of average of the history iteration points according to the algorithms. All the tests are performed in MATLAB R2021b installed on Linux with Intel Xeon(R) Gold 6230R CPU @ 2.10GHz.

4.1. QCQP with expectation objective. In this subsection, we test the algorithms on the stochastic convex QCQP in the following form:

$$\begin{aligned} &\min_{x\in X}\ f(x)=\mathbb{E}\big(\frac{1}{2}\left\|\xi_H x-\xi_c\right\|^2\big),\\ &\text{s.t. } h_j(x)=\frac{1}{2}x^\top Q_j x+a_j^\top x\le b_j,\ j=1,\dots,M. \end{aligned} \tag{4.1}$$

Here $X=[-10,10]^n$, $\xi_H\in\mathbb{R}^{p\times n}$ and $\xi_c\in\mathbb{R}^p$ are randomly generated, and their components are generated by standard Gaussian distribution and then normalized. For each $j\in\{1,\dots,M\}$, $Q_j\in\mathbb{R}^{n\times n}$ is a randomly generated symmetric positive semidefinite matrix with unit 2-norm; a_j is randomly generated independently by the standard Gaussian distribution and then normalized; b_j is generated from the uniform distribution on $[0.1,1.1]$.

In the experiment, we test on QCQP instances of size $(n,p)=(10,5)$ and $(200,150)$ and $M=5$ and 10000, respectively. In both instances, we set batch size is 50 and run 5×10^4 iterations. For the instances with small data size, we solve the approximation problem for the generated 10^5 samples by using CVX[14, 15] to obtain the optimal solution x_{opt}. When running the code, we obtain the unbiased estimate of the gradient and function values by sampling the random variable ξ_H,ξ_c over the above distribution. For the instances with large data sizes, we employ an estimated optimal solution x_{opt} which has the smallest objective value in the feasible set among all iterations of RMALM, APriD, CSA, MSA, and PDSG-adp.

In Figure 1, we report the objective value, the averaged constraint violation, the maximum constraint violation, the last iteration error, and the averaged error by iteration and time. In the first four columns of the Figure 1, the results of the other four algorithms are about $\bar{x}^k$, which have convergence guarantees. However, the results of our algorithm are only about the current iteration point x^k. The distance from the current iteration point x^k to the optimal point for all algorithms is shown in the last column , and we can see that the other four algorithms start to oscillate at a certain distance from the optimal point and fail to converge to the optimal point, while our algorithm converges steadily to the optimal point. The running time for the algorithms in the different settings is shown as Table 1. All results show that the RMALM outperforms the other methods in different stochastic convex QCQP instances.

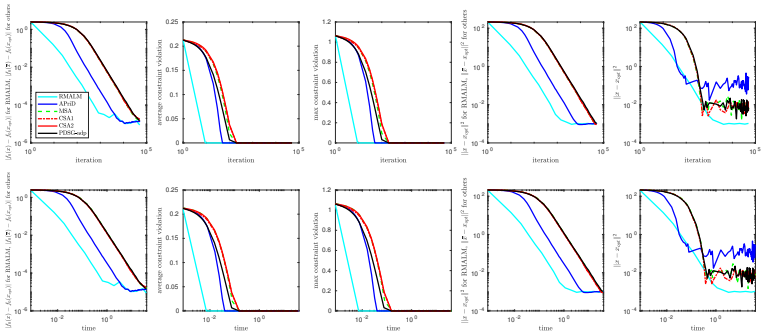
TABLE 1
Running time (in seconds) for QCQP (4.1).

(n,p)	$(10,5)$		$(200,150)$	
number of constraints	$M=5$	$M=10000$	$M=5$	$M=10000$
RMALM	**31.7**	**40.2**	**9449.4**	**10798.9**
ApriD	41.7	51.6	10829.0	11724.5
MSA	41.6	47.1	10661.5	11563.9
CSA	40.8	53.2	10661.1	12453.6
PDSG_adp	42.8	44.2	10530.5	11010.2

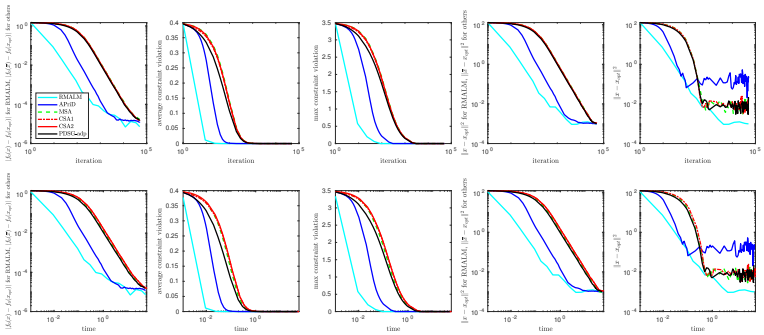
4.2. QCQP with finite-sum objective. In this subsection, we test the algorithms on the QCQP with a finite-sum objective and some constraints:

$$\begin{aligned} &\min_{x\in X}\ f(x)=\frac{1}{2N}\sum_{i=1}^{N}\left\|H_i x-c_i\right\|^2,\\ &\text{s.t. } h_j(x)=\frac{1}{2}x^\top Q_j x+a_j^\top x\le b_j,\ j=1,\dots,M. \end{aligned} \tag{4.2}$$

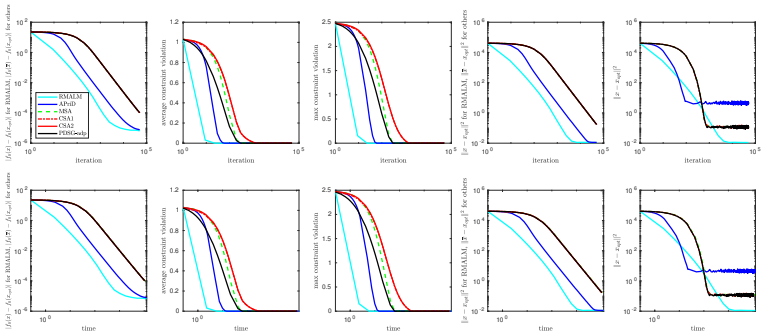
Here $X=[-10,10]^n$. H_i,c_i $(i=1,\dots,N)$ are independently generated from the same distribution as ξ_H,ξ_c in Section 4.1 and Q_j,a_j,b_j $(j=1,\dots,M))$ are generated in the same way as in Section 4.1. In this experiment, we test on QCQP instances of size $(n,p)=(10,5)$



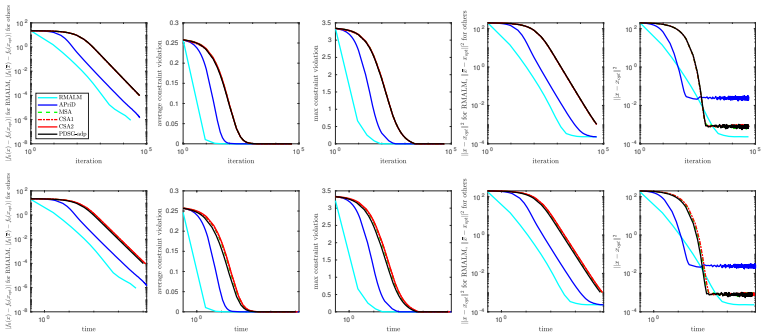
(a) $n = 10$, $p = 5$ and $M = 5$



(b) $n = 10$, $p = 5$ and $M = 10000$



(c) $n = 200$, $p = 150$ and $M = 5$



(d) $n = 200$, $p = 150$ and $M = 10000$

FIG. 1. *In each subplot, the objective error (Left1), averaged constraint violation (Left2), maximum constraint violation (Middle), x^k error for RMALM and $\bar{x}^k$ error for others (Right2), x^k error (Right1) by five methods on solving QCQP instances of* (4.1). *Rows 1 is with respect to iteration; rows 2 is with respect to time (in seconds).*

and $(200, 150)$ and $M = 5$ and 10000. In both instances, we set $N = 10^4$, $batchsize = 50$ and run 5×10^4 iterations.

Optimal solution x_{opt} is obtained in the same way as in Section 4.1. We record the errors and constraints violations in Figure 2 and the running time for the algorithms in Table 2. From the results, we can also notice the better performance of the RMALM for different data sizes and number of constraints.

TABLE 2
Running time (in seconds) for QCQP (4.2).

(n, p)	$(10, 5)$		$(200, 150)$	
number of constraints	$M = 5$	$M = 10000$	$M = 5$	$M = 10000$
RMALM	**3.0**	**9.1**	**235.5**	**793.2**
ApriD	7.2	17.5	482.2	1798.2
MSA	5.3	14.8	259.2	1558.0
CSA	5.9	20.7	461.6	2579.9
PDSG_adp	5.0	13.6	251.2	1492.0

4.3. Two-stage stochastic program. We perform the RMALM on a specific example of the two-stage stochastic program introduced in Section 1. Given by [16], the program is:

$$\begin{aligned} &\min_{x_1 \in \mathbb{R}^n} \; c^T x_1 + \mathbb{E}\left(\mathfrak{Q}\left(x_1, \xi\right)\right) \\ &\text{s.t.} \quad \left\|x_1 - x_0\right\|_2 \leq 1, \end{aligned} \tag{4.3}$$

where cost-to-go function $\mathfrak{Q}\left(x_1, \xi\right)$ has nonlinear objective and constraint coupling functions and is given by

$$\begin{aligned} \mathfrak{Q}\left(x_1, \xi\right) := \min_{x_2 \in \mathbb{R}^n} \; & \frac{1}{2} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}^T \left(\xi\xi^T + \lambda I_{2n}\right) \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \xi^T \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\ \text{s.t.} \quad & \frac{1}{2}\left\|x_2 - y_0\right\|_2^2 + \frac{1}{2}\left\|x_1 - x_0\right\|_2^2 - \frac{R^2}{2} \leq 0. \end{aligned} \tag{4.4}$$

For both problems, $\xi \in \mathbb{R}^{2n}$ is generated from the Gaussian distribution and $\lambda > 0$. The components of ξ are independent with means and standard deviations randomly generated in intervals $[5, 25]$ and $[5, 15]$. We consider two instances of these problem with $n = 5, 30$ and a large sample of size $N = 20,000$ of ξ. We fix $\lambda = 2$ while the components of c are generated randomly in interval $[1, 3]$.

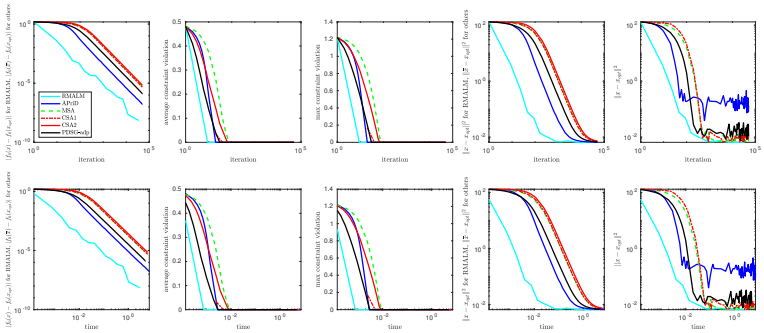
For specific N, we transform the problem to a single quadratic program

$$\begin{aligned} \min_{x_1, y_1, \ldots, y_N} \; & c^T x_1 + \frac{1}{N}\sum_{i=1}^{N} \frac{1}{2} \begin{pmatrix} x_1 \\ y_i \end{pmatrix}^T \left(\xi\xi^T + \lambda I_{2n}\right) \begin{pmatrix} x_1 \\ y_i \end{pmatrix} + \xi^T \begin{pmatrix} x_1 \\ y_i \end{pmatrix} \\ \text{s.t.} \quad & \left\|x_1 - x_0\right\|_2 \leq 1, \; \frac{1}{2}\left\|y_i - y_0\right\|_2^2 + \frac{1}{2}\left\|x_1 - x_0\right\|_2^2 - \frac{R^2}{2} \leq 0, \quad i = 1, 2, \ldots, N, \end{aligned} \tag{4.5}$$

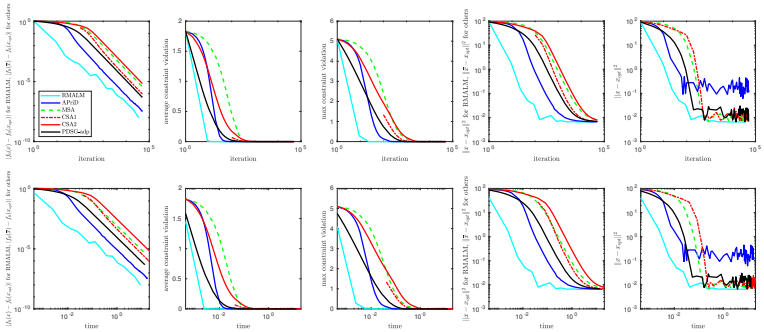
where y_i represents the second stage decision corresponding to ξ_i. For problem (4.5) we take $R = 5$ and $x_0(i) = y_0(i) = 10, i = 1, \ldots, n$. In both instances, we set batch size $= 100$ and run 5×10^4 iterations.

We record the running time and optimal value in Table 3. All constraint violations in five algorithms reach 0. It shows that the RMALM can get optimal value in a faster time.

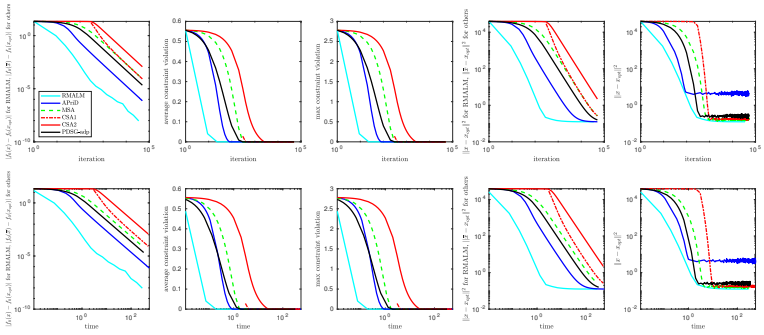
4.4. Stochastic portfolio optimization. We perform the RMALM on the portfolio optimization problem involving Conditional Value at Risk (CVaR) shown as (1.10) and



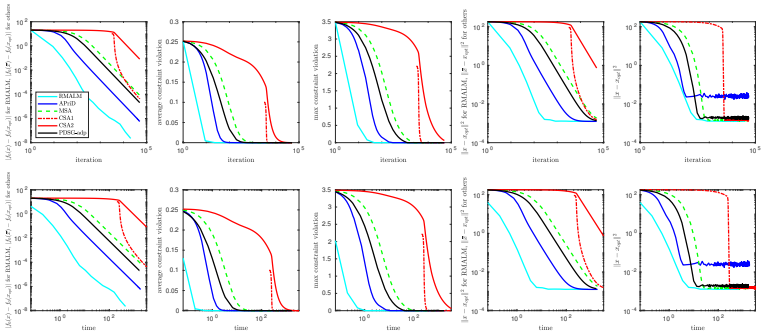
(a) $n = 10$, $p = 5$ and $M = 5$



(b) $n = 10$, $p = 5$ and $M = 10000$



(c) $n = 200$, $p = 150$ and $M = 5$



(d) $n = 200$, $p = 150$ and $M = 10000$

FIG. 2. *In each subplot, the objective error (Left1), averaged constraint violation (Left2), maximum constraint violation (Middle), x^k error for RMALM and $\bar{x}^k$ error for others (Right2), x^k error (Right1) by five methods on solving QCQP instances of* (4.2). *Rows 1 is with respect to iteration; rows 2 is with respect to time (in seconds).*

TABLE 3

CPU time in seconds, approximate optimal value of instances about problems (4.3)-(4.4) *(for* $n = 5$ *or* 30 *and* $N = 20000$*)*

method	time(s)	optimal value
RMALM	**16.8**	**170.78**
ApriD	108.6	171.06
MSA	49.6	179.86
CSA1	53.7	171.12
CSA2	53.7	171.65
PDSG_adp	52.8	180.69

(a) $n = 5$

method	time(s)	optimal value
RMALM	**52.0**	**1941.41**
ApriD	505.1	1944.32
MSA	137.1	2061.07
CSA1	132.5	1969.41
CSA2	132.5	1974.48
PDSG_adp	150.1	2077.19

(b) $n = 30$

(4.6) on the finite dataset

$$\begin{aligned} &\min_{a,x\in X,y} \; a + \frac{1}{(1-p)N}\sum_{i=1}^{N} y_i \\ &\text{s.t.} \quad y_i \geq -\xi_i^T x - a, \quad y_i \geq 0, \quad i = 1,\ldots,N, \end{aligned} \tag{4.6}$$

where x represents the portfolio, ξ_i denotes the rate of return corresponding to the investment at the i-th sample, $p \in (0,1)$ is a safety (reliability) level chosen by users, a is a threshold of loss, N represents the number of samples. Together with the feasible set (1.8), we can rewrite (4.6) as:

$$\begin{aligned} &\min_{a,x,y} \; a + \frac{1}{(1-p)N}\sum_{i=1}^{N} y_i \\ &\text{s.t. } y_i \geq -x^T\xi_i - a, \quad y_i \geq 0, \quad i = 1,\ldots,N, \\ &\qquad -m^T x \leq -R, \quad \sum_{j=1}^{n} x_j = 1, \quad 0 \leq x_j \leq 1, \end{aligned} \tag{4.7}$$

where $m := \mathbb{E}(\xi)$ is the average return, R encodes a minimum desired return. Without loss of generality, we set the desired return as the average return of overall assets in the training set, i.e., $R := \text{mean}(m)$.

We test on five different real portfolio datasets: Dow Jones industrial average (DJIA, with 30 stocks for 507 days), Standard & Poor's 500 (SP500, with 25 stocks for 1276 days), Toronto stock exchange (TSE, with 88 stocks for 1258 days), New York stock exchange (NYSE, with 36 stocks for 5651 days) which are also used in [4, 43]; and one dataset Fama and French (FF100, 100 portfolios formed on size and book-to-market, 25,251 days from July 1926 to May 2022) which is commonly used in financial literature, e.g., [6, 22]. We complete the missing data in FF100 using the K-nearest neighbor method with Euclidean distance.

In both instances, we set batch size = 100 and run 5×10^4 iterations. Then, for the p-values 0.95, we calculated the p-CVaR of the optimal portfolio x^* from the formulas in (4.7), obtaining the results in Table 4, which records the running time, approximate optimal value, and averaged constraint violation for different datasets. The table shows that the RMALM algorithm can optimize the objective function value in a faster time for the same magnitude of constraint violation in these five real datasets.

5. Conclusions. We present a hybrid method of stochastic approximation technique and augmented Lagrangian method for constrained stochastic convex optimization. The complexity is shown to be comparable with the existing related stochastic methods. Numerical experiments also demonstrate superiority in comparison with the first-order stochastic methods. Thus, both theoretical and numerical results suggest that the proposed algorithm is efficient for solving stochastic convex optimization with hard projection constraints.

TABLE 4

CPU time in seconds, approximate optimal value and averaged constraint violation of problem (4.7) *(for DJIA, SP500, TSE, NYSE and FF100)*

method	time(s)	optimal value	averaged constraint violation
RMALM	**1.16**	**-0.9747**	**3.3e-6**
ApriD	2.53	-0.9114	4.2e-6
MSA	2.13	-0.9057	6.2e-6
CSA1	3.41	-0.8457	8.3e-5
CSA2	3.41	-0.6794	2.5e-4
PDSG_adp	2.10	-0.9730	7.4e-6

(a) DJIA, $(N, n) = (507, 30)$

method	time(s)	optimal value	averaged constraint violation
RMALM	**2.24**	**-0.9499**	**1.1e-6**
ApriD	5.20	-0.9283	1.4e-5
MSA	3.69	-0.8853	1.2e-5
CSA1	5.99	-0.8588	9.2e-6
CSA2	5.99	-0.6048	5.0e-5
PDSG_adp	3.67	-0.9453	**1.1e-6**

(b) SP500, $(N, n) = (1276, 25)$

method	time(s)	optimal value	averaged constraint violation
RMALM	**2.85**	**-0.9650**	**7.1e-6**
ApriD	7.08	-0.8777	8.9e-6
MSA	5.59	-0.8633	9.1e-6
CSA1	10.76	-0.8763	1.3e-5
CSA2	10.76	-0.6300	1.9e-4
PDSG_adp	5.83	-0.9590	7.3e-6

(c) TSE, $(N, n) = (1258, 88)$

method	time(s)	optimal value	averaged constraint violation
RMALM	**3.98**	**-1.0024**	**7.0e-6**
ApriD	14.36	-0.9229	8.2e-6
MSA	6.61	-0.5617	7.1e-6
CSA1	9.45	-0.8684	8.4e-6
CSA2	9.45	-0.5896	2.1e-5
PDSG_adp	6.93	-0.9992	7.3e-6

(d) NYSE, $(N, n) = (5651, 36)$

method	time(s)	optimal value	averaged constraint violation
RMALM	**10.60**	**5.1800**	**4.1e-6**
ApriD	42.43	5.9690	4.4e-6
MSA	19.99	15.4154	**4.1e-6**
CSA1	25.85	24.3529	4.8e-6
CSA2	25.85	20.5283	7.9e-6
PDSG_adp	18.87	5.3721	5.4e-6

(e) FF100, $(N, n) = (25251, 100)$

Our algorithm can also be extended to solve online constrained problems with determinate constraints. However, there are still several important issues to be studied. Firstly, our algorithm currently considers stochastic convex optimization with determinate constraints. Secondly, the convergence analysis of RMALM is guaranteed by the strong concavity of the dual essential objective function, and we will further consider weakening this assumption. Another interesting topic is how to use the techniques in this paper to deal with nonconvex constrained stochastic optimization, such as training neural networks with constraints.

REFERENCES

[1] K. Basu and P. Nandy, *Optimal convergence for stochastic optimization with multiple expectation constraints*, arXiv preprint arXiv:1906.03401, (2019).

[2] J. R. Birge, INFORMS journal on computing, 9 (1997), pp. 111–133.

[3] D. Boob, Q. Deng, and G. Lan, *Stochastic first-order methods for convex and nonconvex functional constrained optimization*, Mathematical Programming, (2022), https://doi.org/10.1007/s10107-021-01742-y.

[4] A. Borodin, R. El-Yaniv, and V. Gogan, *Can we learn to beat the best stock*, Advances in Neural Information Processing Systems, 16 (2003).

[5] L. Bottou, F. E. Curtis, and J. Nocedal, *Optimization methods for large-scale machine learning*, Siam Review, 60 (2018), pp. 223–311.

[6] J. Brodie, I. Daubechies, C. De Mol, D. Giannone, and I. Loris, *Sparse and stable markowitz portfolios*, Proceedings of the National Academy of Sciences, 106 (2009), pp. 12267–12272.

[7] C. Chen, F. Tung, N. Vedula, and G. Mori, *Constraint-aware deep neural network compression*, in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 400–415.

[8] Y. S. Chow and H. Teicher, *Probability theory: independence, interchangeability, martingales*, Springer Science & Business Media, 2003.

[9] Y. Cui, C. Ding, and X. Zhao, *Quadratic growth conditions for convex matrix optimization problems associated with spectral functions*, SIAM Journal on Optimization, 27 (2017), pp. 2332–2355.

[10] Y. Cui, D. Sun, and K.-C. Toh, *On the r-superlinear convergence of the kkt residuals generated by the augmented lagrangian method for convex composite conic programming*, Mathematical Programming, 178 (2019), pp. 381–415.

[11] J. M. Danskin, *The theory of max-min, with applications*, SIAM Journal on Applied Mathematics, 14 (1966), pp. 641–664.

[12] F. Facchinei and J.-S. Pang, *Finite-dimensional variational inequalities and complementarity problems*, Springer, 2003.

[13] J. E. Falk, *Lagrange multipliers and nonlinear programming*, Journal of Mathematical Analysis and Applications, 19 (1967), pp. 141–159.

[14] M. Grant and S. Boyd, *Graph implementations for nonsmooth convex programs*, in Recent Advances in Learning and Control, V. Blondel, S. Boyd, and H. Kimura, eds., Lecture Notes in Control and Information Sciences, Springer-Verlag Limited, 2008, pp. 95–110. http://stanford.edu/~boyd/graph_dcp.html.

[15] M. Grant and S. Boyd, *CVX: Matlab software for disciplined convex programming, version 2.1*. http://cvxr.com/cvx, Mar. 2014.

[16] V. Guigues, *Inexact stochastic mirror descent for two-stage nonlinear stochastic programs*, Mathematical Programming, 187 (2021), pp. 533–577.

[17] M. R. Hestenes, *Multiplier and gradient methods*, Journal of optimization theory and applications, 4 (1969), pp. 303–320.

[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 (2002), pp. 479–502.

[19] G. Lan and Z. Zhou, *Algorithms for stochastic optimization with function or expectation constraints*, Computational Optimization and Applications, 76 (2020), pp. 461–498.

[20] X. Li, D. Sun, and K.-C. Toh, *A highly efficient semismooth newton augmented lagrangian method for solving lasso problems*, SIAM Journal on Optimization, 28 (2018), pp. 433–458.

[21] A. Milzarek, F. Schaipp, and M. Ulbrich, *A semismooth newton stochastic proximal point algorithm with variance reduction*, arXiv preprint arXiv:2204.00406, (2022).

[22] K. Nakagawa and K. Ito, *Taming tail risk: Regularized multiple β worst-case cvar portfolio*, Symmetry, 13 (2021), p. 922.

[23] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, *Robust stochastic approximation approach to stochastic programming*, SIAM Journal on optimization, 19 (2009), pp. 1574–1609.

[24] G. C. Pflug, *Optimization of stochastic models: the interface between simulation and optimization*, vol. 373, Springer Science & Business Media, 2012.

[25] G. C. Pflug and A. Pichler, *Multistage stochastic optimization*, vol. 1104, Springer, 2014.

[26] M. J. Powell, *A method for nonlinear constraints in minimization problems*, Optimization, (1969), pp. 283–298.

[27] H. Robbins and S. Monro, *A stochastic approximation method*, The annals of mathematical

statistics, (1951), pp. 400–407.
[28] H. Robbins and D. Siegmund, *A convergence theorem for non negative almost supermartingales and some applications*, in Optimizing methods in statistics, Elsevier, 1971, pp. 233–257.
[29] R. T. Rockafellar, *Convex analysis*, vol. 18, Princeton university press, 1970.
[30] R. T. Rockafellar, *Augmented lagrangians and applications of the proximal point algorithm in convex programming*, Mathematics of operations research, 1 (1976), pp. 97–116.
[31] R. T. Rockafellar, *Monotone operators and the proximal point algorithm*, SIAM journal on control and optimization, 14 (1976), pp. 877–898.
[32] R. T. Rockafellar, S. Uryasev, et al., *Optimization of conditional value-at-risk*, Journal of risk, 2 (2000), pp. 21–42.
[33] R. T. Rockafellar and R. J.-B. Wets, *Variational analysis*, vol. 317, Springer Science & Business Media, 2009.
[34] E. K. Ryu and S. Boyd, *Stochastic proximal iteration: a non-asymptotic improvement upon stochastic gradient descent*, Author website, early draft, (2014).
[35] A. Shapiro, *Monte carlo sampling methods*, Handbooks in operations research and management science, 10 (2003), pp. 353–425.
[36] A. Shapiro, D. Dentcheva, and A. Ruszczynski, *Lectures on stochastic programming: modeling and theory*, SIAM, 2021.
[37] P. Toulis, T. Horel, and E. M. Airoldi, *The proximal robbins–monro method*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 83 (2021), pp. 188–212.
[38] L. Xiao and T. Zhang, *A proximal stochastic gradient method with progressive variance reduction*, SIAM Journal on Optimization, 24 (2014), pp. 2057–2075.
[39] X. Xiao, *Penalized stochastic gradient methods for stochastic convex optimization with expectation constraints*, Optimization-online, (2019).
[40] Y. Xu, *Primal-dual stochastic gradient method for convex programs with many functional constraints*, SIAM Journal on Optimization, 30 (2020), pp. 1664–1692.
[41] Y. Yan and Y. Xu, *Adaptive primal-dual stochastic gradient method for expectation-constrained convex stochastic programs*, Mathematical Programming Computation, (2022), pp. 1–45.
[42] H. Yu, M. Neely, and X. Wei, *Online convex optimization with stochastic constraints*, Advances in Neural Information Processing Systems, 30 (2017).
[43] A. Yurtsever, B. C. Vu, and V. Cevher, *Stochastic three-composite convex minimization*, Advances in Neural Information Processing Systems, 29 (2016).
[44] L. Zhang, Y. Zhang, J. Wu, and X. Xiao, *Solving stochastic optimization with expectation constraints efficiently by a stochastic augmented lagrangian-type algorithm*, (2021).
[45] X.-Y. Zhao, T. Cai, and D. Xu, *A newton-cg augmented lagrangian method for convex quadratically constrained quadratic semidefinite programs*, in Advances in Global Optimization, Springer, 2015, pp. 337–345.
[46] Y. Zhou, C. Bao, C. Ding, and J. Zhu, *A semi-smooth newton based augmented lagrangian method for nonsmooth optimization on matrix manifolds*, arXiv preprint arXiv:2103.02855, (2021).