Journal of Computational Mathematics
Vol.xx, No.x, 2024, 1–30.

http://www.global-sci.org/jcm
doi:10.4208/jcm.2208-m2022-0035

A STOCHASTIC AUGMENTED LAGRANGIAN METHOD FOR STOCHASTIC CONVEX PROGRAMMING*

Jiani Wang
Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China
Email: wjiani@lsec.cc.ac.cn
Liwei Zhang[1)]
School of Mathematical Sciences, Dalian University of Technology, Dalian 116024, China
Email: lwzhang@dlut.edu.cn

Abstract

In this paper, we analyze the convergence properties of a stochastic augmented Lagrangian method for solving stochastic convex programming problems with inequality constraints. Approximation models for stochastic convex programming problems are constructed from stochastic observations of real objective and constraint functions. Based on relations between solutions of the primal problem and solutions of the dual problem, it is proved that the convergence of the algorithm from the perspective of the dual problem. Without assumptions on how these random models are generated, when estimates are merely sufficiently accurate to the real objective and constraint functions with high enough, but fixed, probability, the method converges globally to the optimal solution almost surely. In addition, sufficiently accurate random models are given under different noise assumptions. We also report numerical results that show the good performance of the algorithm for different convex programming problems with several random models.

Mathematics subject classification: 49N15, 90C15, 90C25.
Key words: Stochastic convex optimization, Stochastic approximation, Augmented Lagrangian method, Duality theory.

1. Introduction

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

$$\begin{aligned} \min_{x\in X_0} \quad & f(x)=\mathbb{E}[F(x,\xi)] \\ \text{s.t.} \quad & g_i(x)=\mathbb{E}[G_i(x,\xi)]\le 0, \quad i=1,\ldots,p. \end{aligned} \tag{1.1}$$

Here $X_0 \subset \mathbb{R}^n$ is a nonempty bounded closed convex set, $\xi:\Omega\to\Xi$ is a random vector defined on a given probability space $(\Omega,\mathcal{F},\mathbb{P})$ and $F: O\times\Xi\to\mathbb{R}, G_i: O\times\Xi\to\mathbb{R}, i{=}1,\ldots,p$, where $O{\supset}X_0$ is an open convex set and Ξ is a measurable space. Without loss of generality, we assume that expectations $\mathbb{E}[F(x,\xi)]$ and $\mathbb{E}[G_i(x,\xi)]$ are well defined and finite valued for every $x\in O$ and the expected value function $f(\cdot)$ and $g_i(\cdot)$ are continuous and convex on O. Any algorithm for solving problem (1.1) has to be faced with the difficulty that the full evaluations of expectations $\mathbb{E}[F(x,\xi)]$ and $\mathbb{E}[G_i(x,\xi)]$ are either impossible or expensive in practice. There are two types of methods to resolve this problem: the sample average approximation (SAA) method and the

* Received February 9, 2022 / Revised version received June 6, 2022 / Accepted August 31, 2022 / Published online January 19, 2024 /

[1)] Corresponding author

stochastic approximation (SA). The SAA method usually solves the random approximation model through sample averaging estimators of random variables. Let $\xi_1, \cdots, \xi_N$ be an i.i.d. sample of realizations of random vector ξ of size N and the average sample approximation model is defined as

$$\begin{aligned} &\min_{x\in X_0} \ \frac{1}{N}\sum_{m=1}^{N} F(x,\xi_m) \\ &\text{s.t.} \ \ \frac{1}{N}\sum_{m=1}^{N} G_i(x,\xi_m) \le 0, \quad i=1,\dots,p. \end{aligned} \tag{1.2}$$

Usually, the convergence of SAA depends on the special choice of parameters and the expensive iteration cost, like [18, 26, 28]. However, so far no study has applied sample averaging methods to the case of biased noise for stochastic convex optimization problem (1.1).

On the other hand, the stochastic approximation, in most studies (for example, [17, 31]), is to generate stochastic oracles $\mathbb{F}^{t_k} : \mathbb{R}^n \times \Xi \to \mathbb{R}$ and $\mathbb{G}_i^{s_k} : \mathbb{R}^n \times \Xi \to \mathbb{R}$ of the stochastic function values of f and g_i. More specifically, the random approximation model of (1.1) is defined as

$$\begin{aligned} &\min_{x\in X_0} \ \mathbb{F}^{t_k}(x) \\ &\text{s.t.} \ \ \mathbb{G}_i^{s_k}(x) \le 0, \quad i=1,\dots,p, \end{aligned} \tag{1.3}$$

where $\mathbb{F}^{t_k}$ and $\mathbb{G}_i^{s_k}$ are function which are constructed by one or mini-batches of random samples t_k, s_k of stochastic function [5, 29]. For each $k, \mathbb{F}^{t_k}$ and $\mathbb{G}_i^{s_k}$ are continuous on $x \in O$. Obviously, the iterate $x_{k+1} = x_{k+1}(\xi_{[k]})$ can be seen as a function of the history $\xi_{[k]} := (\xi_1, \cdots, \xi_k)$ of the generated random process. The above two random models are considered as the noisy computable version of the real optimization problem (1.1) and the convergence of both methods relies on zero-mean noise with bounded variance (or even with decreasing variance), so estimators in these random models need to be carefully chosen [1, 30]. To the best of our knowledge, no study so far has mentioned the convergence of the stochastic convex programming with inequality constraints under the above two methods, for the case of biased noise. Regardless of the random approximation model (1.2) or (1.3), we propose a stochastic augmented Lagrange method and prove that when the random models are merely sufficiently close to the real optimization problems with high enough, but fixed, probability, the sequence generated by the stochastic algorithm converges to the optimal solution almost surely. In this paper, we consider a general random approximation model of (1.1) as follow:

$$\begin{aligned} &\min_{x\in X_0} \ f^k(x) \\ &\text{s.t.} \ \ G^k(x) \le 0. \end{aligned} \tag{1.4}$$

For each k, f^k and $G^k =: (g_1^k, \cdots, g_p^k)$ are stochastic approximations of f and g_i and continuous on $x \in O$. The augmented Lagrangian function of problem (1.4) is defined by

$$\mathcal{L}_r^k(x,\lambda) = f^k(x) + \frac{1}{2r}\Big[\big\|\Pi_{\mathbb{R}_+^p}\big(\lambda + rG^k(x)\big)\big\|^2 - \|\lambda\|^2\Big], \quad \forall\,(x,\lambda) \in \mathbb{R}^n \times \mathbb{R}^p, \tag{1.5}$$

where $\Pi_{\mathbb{R}_+^p}(y)$ represents the projection of y onto $\mathbb{R}_+^p$ for any $y \in \mathbb{R}^p$. In the following we denote $[y]_+ := \Pi_{\mathbb{R}_+^p}(y)$. The stochastic augmented Lagrangian method for solving (1.1) with the help of the random model (1.4) can be described as Algorithm 1.1.

The augmented Lagrangian method for solving the optimization problem with constraints can be traced back to the pioneering paper by Rockafellar [23]. Since the augmented Lagrangian

Algorithm 1.1: Stochastic Augmented Lagrangian Method (SALM).

```
  Require: Given parameter $r > 0$, the initial multiplier $\lambda^0 \in \mathbb{R}^p$ and the initial point
           $x^0 \in \mathbb{R}^n$. Let $k = 0$.
1 for $k = 0, 1, \ldots$ do
2     If $x^k$ satisfies the termination criterion, then stop and return $x^k$.
3     Select the estimation models $f^k$ and $G^k$ and compute
          $x^{k+1} = \operatorname{argmin}\left\{\mathcal{L}_r^k(x, \lambda^k),\ x \in X_0\right\},$   (1.6)
          $\lambda^{k+1} = [\lambda^k + rG^k(x^{k+1})]_+.$   (1.7)
4     Let $k = k + 1$ and go to Step 1.
5 end
```

method is interpreted as the proximal point method applied to the dual optimality, a superlinear convergence can be reached under certain conditions [4], which is widely used on various optimization problems [2,9,11,22,27]. However, there is little research on stochastic nonlinear programming. In the current research, [14] analyzes the convergence of the optimal value depending on the sufficiently large sample and the boundedness of the gradient of the constraint functions; [15] only obtains the convergence for the general Lipschitz continuous objective functions with a high probability; [30] establishes convergence based on the assumption of unbiased estimates. Inspired by [3, 10], which analyze the convergence of the unconstrained stochastic optimization problems under the trust-region method, we show the global convergence of Algorithm 1.1 by analyzing the characteristics of the dual problem. Compared to assumptions in classical stochastic algorithms in prior work, our conditions are weaker, which just assume that:

- For each k, the estimation models f^k and G^k of objective and constraint functions are sufficiently accurate with sufficiently high probability.
- For each k, the estimates of the gradient of the objective function ϕ_r of the dual problem are sufficiently accurate at the current iterate with sufficiently high probability.

In particular, we do not assume that:

- The probabilities of generating sufficiently accurate optimization models and estimates are increasing (they only need to be higher than a certain constant).
- The distribution of the random models and estimates has certain specific characteristics. (That is, the inaccuracy of the random models can be arbitrary, and estimates can have biased or unbiased noise.)

On this basis, the paper makes the following contributions to the convergence analysis of this model:

- Under conditions that models and estimates are sufficiently accurate with sufficiently high, but fixed probability and the optimal values of approximate dual problems are bounded, the multiplier sequence $\{\lambda^k\}$ generated by the stochastic augmented Lagrangian method converges to the optimal solution of the dual problem with probability 1.

- Moreover, if generalized Slater condition holds for (1.1), we prove that the sequence $\{x^k\}$ converges to the optimal solution of the primal problem (1.1) with probability 1.
- In addition, different settings of noise (biased or unbiased) are discussed. In order to ensure the convergence, the selections of the parameters of the algorithm and the probability of the accuracy of the models in different situations are given in this paper.

The paper is organized as follows. In the next section, using the duality theorem, the gradient of the objective function of the dual problem is calculated and the significance of the algorithm is discussed in depth in terms of duality. The global convergence results of our algorithm are given in Section 3. In Section 4, various noise scenarios are discussed and to achieve convergence, different accuracies of models and estimates are also constructed. Further, in Section 5, we show the performance of the convergence of SALM in the stochastic convex programming problems and the effect of the selection of parameters and sample sizes on the convergence in some specific cases. Finally, the conclusion is given in Section 6.

Notations. We use the following notation throughout the paper. $\mathbb{R}^n_+$ represents n-dimensional positive octant space and $\|\cdot\|$ represents the ℓ^2 vector norm. For a set C and a point $x_0, \mathrm{dist}(C-x_0)$ denotes the distance between C and x_0, i.e. $\mathrm{dist}(C-x_0)=\sup_{x_1\in C}\|x_1-x_0\|$. ri C represents the relative interior of set C. The conjugate function of the function $f:\mathbb{R}^n\to\mathbb{R}$ is denoted as $f^*:\mathbb{R}^n\to\mathbb{R}$. The optimal solution set of a minimization problem f and a maximization problem g are expressed as argmin f and argmax g, respectively. Let $(\Omega,\mathcal{F},\mathbb{P})$ be a given probability space. We write $\xi\in\mathcal{F}$ for "ξ is $\mathcal{F}$-measurable". We use $\mathcal{B}(\mathbb{R}^n)$ to denote the Borel σ-algebra of $\mathbb{R}^n$ and $\sigma(\xi^1,\cdots,\xi^k)$ denoted as the σ-algebra generated by the family of random variables $\xi^1,\cdots,\xi^k$. For a random variable ξ and a sub-σ-algebra $\mathcal{S}\subset\mathcal{F}$, the conditional expectation of ξ given $\mathcal{S}$ is denoted by $\mathbb{E}[\xi|\mathcal{S}]$. The abbreviations "a.s." stand for "almost surely".

2. The Dual Problem of the Convex Problem

In this section, we discuss some properties of the dual problem of (1.4) and give an alternative interpretation of the augmented Lagrangian method from the perspective of the dual problem. Considering the Lagrangian function of (1.4) is

$$\mathcal{L}^k(x,\lambda)=f^k(x)+\lambda^TG^k(x),\quad \forall\,(x,\lambda)\in\mathbb{R}^n\times\mathbb{R}^p.$$

So the dual problem of (1.4) is

$$(\mathrm{D}^k_0)\qquad \max_{\lambda\in\mathbb{R}^p_+}\ \phi^k_0(\lambda):=\inf_{x\in X_0}\ \mathcal{L}^k(x,\lambda). \tag{2.1}$$

For any $r\geq 0$, we consider

$$(\mathrm{D}^k_r)\qquad \max_{\lambda\in\mathbb{R}^p}\ \phi^k_r(\lambda):=\inf_{x\in X_0}\ \mathcal{L}^k_r(x,\lambda). \tag{2.2}$$

The following proposition discusses the relationship between the optimal solution sets of (D^k_r) and (D^k_0) and the expression of the gradient of the concave function ϕ^k_r.

Proposition 2.1. *For any k-th iteration and each $r>0, \phi^k_r(\cdot)$ is concave on $\lambda\in\mathbb{R}^p$ and satisfies*

$$\phi^k_r(\lambda)=\max_z\left\{\phi^k_0(z)-\frac{1}{2r}\|z-\lambda\|^2\right\}. \tag{2.3}$$

Therefore, the dual problems (D_r^k) *and* (D_0^k) *have the same optimal solution. Further, if* $\phi_0^k \not\equiv -\infty$*, then* $\phi_r^k(\cdot)$ *is finite continuous differentiable on* $\mathbb{R}^p$*. In particular, if for a given* $\bar{\lambda}$*, the function* $\mathcal{L}_r^k(\cdot,\bar{\lambda})$ *minimizes at* $\bar{x}$ *(not necessarily unique), then for* $i = 1, \ldots, p$*,*

$$\frac{\partial \phi_r^k(\bar{\lambda})}{\partial \lambda_i} = \frac{\partial \mathcal{L}_r^k(\bar{x},\bar{\lambda})}{\partial \lambda_i} = \max\left\{-\frac{\bar{\lambda}_i}{r},\ g_i^k(\bar{x})\right\}. \tag{2.4}$$

Proof. For any $r \geq 0$ and $u^k = (u_1^k, \cdots, u_p^k) \in \mathbb{R}^p$, let

$$p_r^k(u^k) := \inf_{x\in X_0}\ l_r^k(x,u^k),$$

where $l_r^k(x,u^k)$ is defined as

$$l_r^k(x,u^k) = \begin{cases} f^k(x) + \dfrac{r}{2}\displaystyle\sum_{i=1}^{p}\left(u_i^k\right)^2, & \text{if} \quad g_i^k(x) \leq u_i^k, \quad i = 1,\ldots,p,\\ +\infty, & \text{otherwise.}\end{cases}$$

Since $l_r^k(x,u^k)$ is convex on $(x,u^k) \in X_0 \times \mathbb{R}^p$, $p_r^k(u^k)$ is convex for any $u^k \in \mathbb{R}^p$. Noticed that for any given $k \in \mathbb{N}$,

$$p_r^k = p_0^k + rq^k,$$

where

$$q^k(u^k) = \frac{1}{2}\|u^k\|^2.$$

By [21, Theorem 3.1] and the definition of the conjugate function, we obtain that

$$\phi_r^k(\lambda) = \inf_{u^k}\left\{p_r^k(u^k) + \langle\lambda,u^k\rangle\right\} = -\left(p_r^k\right)^*(-\lambda). \tag{2.5}$$

Using the formula that calculates the conjugate function of the sum of convex functions (see [20, Theorem 16.4]), it has

$$-\left(p_r^k\right)^*(-\lambda) = -\left(p_0^k + rq^k\right)^*(-\lambda) = \max_{z\in\mathbb{R}^p}\left\{-\left(p_0^k\right)^*(-z) - (rq^k)^*(z-\lambda)\right\}. \tag{2.6}$$

In addition, by simply calculating, we have $(rq^k)^*(y) = r(q^k)^*(y/r)$ and $(q^k)^* = q^k$. Hence, combining (2.5) and (2.6), it yields

$$\phi_r^k(\lambda) = \max_{z\in\mathbb{R}^p}\left\{-(p_0^k)^*(-z) - rq^k\left(\frac{z-\lambda}{r}\right)\right\} = \max_{z\in\mathbb{R}^p}\left\{\phi_0^k(z) - \frac{1}{2r}\|z-\lambda\|^2\right\},$$

which implies (2.3). According to the definition of Moreau envelope function in [24, Definition 1.22], $\phi_r^k(\lambda)$ is Moreau envelope of the function $\phi_0^k(\lambda)$, i.e.

$$\phi_r^k(\lambda) = -e_r\left[-\phi_0^k\right](\lambda).$$

Then by [24, Theorem 2.26], $-\phi_r^k(\lambda)$ is convex and continuous differentiable on $\mathbb{R}^p$, which implies that $\phi_r^k(\lambda)$ is concave and

$$\nabla\phi_r^k(\lambda) = -\nabla e_r\left[-\phi_0^k\right](\lambda) = -\frac{1}{r}\left[\lambda - P_r\left[-\phi_0^k\right](\lambda)\right] = \frac{1}{r}\left[P_r\left[-\phi_0^k\right](\lambda) - \lambda\right], \tag{2.7}$$

where $P_r[-\phi_0^k](\lambda)$ is the proximal mapping of the function $-\phi_0^k$ and the parameter r. In particular, for the given $\bar{\lambda}$, let $\bar{x}$ be the minimum of $\mathcal{L}_r^k(\cdot,\bar{\lambda})$, then it has

$$\phi_r^k(\bar{\lambda}) = \mathcal{L}_r^k(\bar{x},\bar{\lambda}). \tag{2.8}$$

Hence, by the definition of the proximal mapping in [24, Definition 1.22], we get

$$P_r\big[-\phi_0^k\big](\bar{\lambda}) = \operatorname{argmin}_{z\geq 0}\left\{-f^k(\bar{x}) - z^T G^k(\bar{x}) + \frac{1}{2r}\|z-\bar{\lambda}\|^2\right\}. \tag{2.9}$$

Notice that for general convex programming problems $\min\psi(x)$ s.t. $x\geq 0$, x^* is the optimal solution if and only if $0\leq x^*\perp\nabla\psi(x^*)\geq 0$. So the optimal solution z^* in (2.9) is

$$P_r\big[-\phi_0^k\big](\bar{\lambda}) = \max\big\{0, \bar{\lambda} + rG^k(\bar{x})\big\}. \tag{2.10}$$

Therefore, taking the gradient of (2.8), combined with (2.7) and(2.10), the conclusions are proved. □

The above proposition shows that the optimal solution $\bar{\lambda}$ of (D_r^k) satisfies for any $i=1,\ldots,p$,

$$\frac{\partial\phi_r^k(\bar{\lambda})}{\partial\lambda_i} = \max\left\{-\frac{\bar{\lambda}_i}{r},\ g_i^k(\bar{x})\right\} = 0, \tag{2.11}$$

where $\bar{x}$ be the minimum of $\mathcal{L}_r^k(\cdot,\bar{\lambda})$. We look at Algorithm 1.1 again, and from the perspective of the dual problem, (1.7) is equivalent to

$$\begin{aligned}\lambda^{k+1} &= \lambda^k + \big(\max\big\{0,\ \lambda^k + rG^k(x^{k+1})\big\} - \lambda^k\big)\\ &= \lambda^k + r\max\left\{-\frac{\lambda^k}{r}, G^k(x^{k+1})\right\}\\ &= \lambda^k - r\nabla\big(-\phi_r^k\big)(\lambda^k),\end{aligned}$$

where

$$x^{k+1} = \operatorname{argmin}\{\mathcal{L}_r^k(x,\lambda^k),\ x\in X_0\}.$$

According to Proposition 2.1 and the definition of Moreau envelope function in [24, Definition 1.22], $\phi_r^k(\lambda)$ is Moreau envelope of the function $\phi_0^k(\lambda)$, and (D_r^k) and (D_0^k) have the same optimal solution set. Therefore, (1.7) can also be seen as the stochastic negative gradient method with constant step size used to solve the dual problem of (1.1). The main purpose of our algorithm is to achieve the optimal solution x^* of the stochastic problem (1.1). Therefore, the following proposition discusses the equivalence condition between the minimum of $\mathcal{L}_r^k(\cdot,\bar{\lambda})$ and the optimal solution of (1.4). Undoubtedly, this applies to the case of the real problem (1.1).

Proposition 2.2. *For any k-th iteration, suppose that the dual gap between* (1.4) *and its dual problem* (D_0^k) *is zero. Let $\bar{\lambda}$ be any dual optimal solution of* (D_0^k)*. Let $r>0$ be a given positive number. Then $\bar{x}$ is an optimal solution to* (1.4) *if and only if $\bar{x}$ is the minimum of the function $\mathcal{L}_r^k(\cdot,\bar{\lambda})$ on X_0.*

Proof. Since $\bar{\lambda}$ is a Kuhn-Tucker vector, by [20, Theorem 28.3], $\bar{x}$ is an optimal solution to (1.4) if and only if $(\bar{x},\bar{\lambda})$ is a saddle point of $\mathcal{L}_r^k(\cdot,\cdot)$, which indicates that $\bar{x}$ is the minimum of the function $\mathcal{L}_r^k(\cdot,\bar{\lambda})$ (see [16, Theorem SP2]). On the other hand, from Proposition 2.1, if $\bar{x}$ is the minimum of the function $\mathcal{L}_r^k(\cdot,\bar{\lambda})$, it has $\nabla_\lambda\mathcal{L}_r^k(\bar{x},\bar{\lambda}) = \nabla\phi_r^k(\bar{\lambda}) = 0$. Hence, $\bar{\lambda}$ is the maximum of the function $\mathcal{L}_r^k(\bar{x},\cdot)$, which demonstrates that $(\bar{x},\bar{\lambda})$ is a saddle point of $\mathcal{L}_r^k$. This yields that $\bar{x}$ is the optimal solution to (1.4). □

3. Convergence Analysis

For the stochastic convex optimization problem (1.1), consider the following problem:

$$(\mathrm{D}_r) \qquad \max_{\lambda\in\mathbb{R}^p} \phi_r(\lambda) := \inf_{x\in X_0} \mathcal{L}_r(x,\lambda), \tag{3.1}$$

where the augmented Lagrangian function of (1.1) is defined as

$$\mathcal{L}_r(x,\lambda) = f(x) + \frac{1}{2r}\Big[\big\|\Pi_{\mathbb{R}^p_+}\big(\lambda + rG(x)\big)\big\|^2 - \|\lambda\|^2\Big],$$

where $G(x) := (g_1(x),\cdots,g_p(x))$ as the constraint function. There is no doubt that (D_0) is the dual problem of (1.1). The proof of convergence will be divided into the following two parts. Firstly, from Proposition 2.1, it is easy to deduce that, with similar proof, if the algorithm converges to the optimal solution λ^* of the problem (3.1) almost surely, then λ^* is also the optimal solution to the dual problem of (1.1) almost surely. Secondly, since the problem considered here is convex, if generalized Slater condition[1] holds for (1.1), there is no dual gap between (1.1) and the dual problem (D_0). Hence, from Proposition 2.2, we only need to prove that any cluster point $\bar{x}$ generated by the iterative sequence $\{x^k\}$ satisfies that $\bar{x}$ is the minimum of $\mathcal{L}_r(\cdot,\lambda^*)$. In order to prove that the cluster points of the iterative sequence generated in Algorithm 1.1 are the optimal solutions for the real problem (1.1), we require the dual estimate ϕ_r^k is sufficiently accurate. It can be established by sufficient accuracy of the objective estimate f^k and the constraint estimate G^k. Modifying the definition of accurate estimates in [13] and [10], we give the definitions of accurate estimates under the stochastic convex optimization problem as follows.

Definition 3.1. *For each $k\in\mathbb{N}$, the objective estimation model f^k is said to be κ_f-accurate estimation model of f with a given boundary M_k^f, if for any $x\in O$,*

$$|f(x) - f^k(x)| \le \kappa_f\big(M_k^f\big)^2, \tag{3.2}$$

where O is an open neighborhood containing X_0.

Definition 3.2. *For each $k\in\mathbb{N}$, the constraint estimation model G^k is said to be κ_g-accurate estimation model of G with a given boundary M_k^g, if for any $x\in O$,*

$$\max\big\{\|G(x)-G^k(x)\|, \big|\|G(x)\|^2 - \|G^k(x)\|^2\big|\ \big\} \le \kappa_g \min\big\{M_k^g, (M_k^g)^2\big\}, \tag{3.3}$$

where O is an open neighborhood containing X_0.

Here it needs to emphasize that the boundary M_k^f or M_k^g may be random and may change with k. If f^k is κ_f-accurate estimation model and G^k is κ_g-accurate estimation model, then for a given λ, $\phi_r^k(\lambda)$ is an accurate estimate of $\phi_r(\lambda)$. In fact,

$$\begin{aligned}
\big|\phi_r(\lambda) - \phi_r^k(\lambda)\big| &= \Big|\inf_{x\in X_0}\mathcal{L}_r(x,\lambda) - \inf_{x\in X_0}\mathcal{L}_r^k(x,\lambda)\Big| \\
&\le \sup_{x\in X_0}\big|\mathcal{L}_r(x,\lambda) - \mathcal{L}_r^k(x,\lambda)\big| \\
&\le \sup_{x\in X_0}\left|\big(f(x)-f^k(x)\big) + \lambda^T\left(\max\left\{-\frac{\lambda}{r}, G(x)\right\} - \max\left\{-\frac{\lambda}{r}, G^k(x)\right\}\right)\right.
\end{aligned}$$

[1] Generalized Slater condition holds for (1.1) if there exists a $x_0 \in \operatorname{ri} X_0$ such that $g_i(x_0) < 0, i = 1,\ldots,p$.

$$
\begin{aligned}
&+\frac{r}{2}\left(\left\|\max\left\{-\frac{\lambda}{r},G(x)\right\}\right\|^2-\left\|\max\left\{-\frac{\lambda}{r},G^k(x)\right\}\right\|^2\right)\Bigg|\\
&\le \sup_{x\in X_0}|f(x)-f^k(x)|+\|\lambda\|\sup_{x\in X_0}\left\|\max\left\{-\frac{\lambda}{r},G(x)\right\}-\max\left\{-\frac{\lambda}{r},G^k(x)\right\}\right\|\\
&\quad+\frac{r}{2}\sup_{x\in X_0}\left|\left\|\max\left\{-\frac{\lambda}{r},G(x)\right\}\right\|^2-\left\|\max\left\{-\frac{\lambda}{r},G^k(x)\right\}\right\|^2\right|\\
&\le \sup_{x\in X_0}\left|f(x)-f^k(x)\right|+\|\lambda\|\sup_{x\in X_0}\|G(x)-G^k(x)\|\\
&\quad+\frac{r}{2}\sup_{x\in X_0}\left|Q^\lambda\big(G(x)\big)-Q^\lambda\big(G^k(x)\big)\right|, \qquad (3.4)
\end{aligned}
$$

where

$$Q^\lambda(y):=\left\|\max\left\{-\frac{\lambda}{r},y\right\}\right\|^2.$$

In the last inequality above, we use the nonexpansivity of the projection, i.e.

$$
\begin{aligned}
&\sup_{x\in X_0}\left\|\max\left\{-\frac{\lambda}{r},G(x)\right\}-\max\left\{-\frac{\lambda}{r},G^k(x)\right\}\right\|\\
&=\sup_{x\in X_0}\left\|\Pi_{\mathbb{R}^p_+}\left(G(x)+\frac{\lambda}{r}\right)-\Pi_{\mathbb{R}^p_+}\left(G^k(x)+\frac{\lambda}{r}\right)\right\|\\
&\le \sup_{x\in X_0}\|G(x)-G^k(x)\|.
\end{aligned}
$$

Now let us estimate the upper bound of $|Q^\lambda(G(x))-Q^\lambda(G^k(x))|$. The discussion will be divided into three situations.

Case 1. $-\lambda/r\le\min\{G(x),G^k(x)\}$:

$$\left|Q^\lambda\big(G(x)\big)-Q^\lambda\big(G^k(x)\big)\right|=\left|\|G(x)\|^2-\|G^k(x)\|^2\right|\le\kappa_g\big(M_k^g\big)^2.$$

Case 2. $-\lambda/r\ge\max\{G(x),G^k(x)\}$:

$$\left|Q^\lambda\big(G(x)\big)-Q^\lambda\big(G^k(x)\big)\right|=0\le\kappa_g\big(M_k^g\big)^2.$$

Case 3. $\min\{G(x),G^k(x)\}\le-\lambda/r\le\max\{G(x),G^k(x)\}$: If $\min\{G(x),G^k(x)\}\ge 0$ or $\max\{G(x),G^k(x)\}\le 0$, then it has

$$
\begin{aligned}
\left|Q^\lambda\big(G(x)\big)-Q^\lambda\big(G^k(x)\big)\right|&=\left|\|\max\left\{G(x),G^k(x)\right\}\|^2-\left\|\frac{\lambda}{r}\right\|^2\right|\\
&\le\left|\|G(x)\|^2-\|G^k(x)\|^2\right|\le\kappa_g\big(M_k^g\big)^2.
\end{aligned}
$$

If $\min\{G(x),G^k(x)\}\le 0\le\max\{G(x),G^k(x)\}$, then $\|\max\{G(x),\ G^k(x)\}\|\le\kappa_g M_k^g$, since $\|G(x)-G^k(x)\|\le\kappa_g M_k^g$. So it implies that

$$
\begin{aligned}
\left|Q^\lambda\big(G(x)\big)-Q^\lambda\big(G^k(x)\big)\right|&=\left|\|\max\{G(x),G^k(x)\}\|^2-\left\|\frac{\lambda}{r}\right\|^2\right|\\
&\le\|\max\{G(x),G^k(x)\}\|^2\le\kappa_g^2\big(M_k^g\big)^2.
\end{aligned}
$$

Let $\kappa_g \in (0,1)$, then $\kappa_g^2 \le \kappa_g$. In summary, the upper bound of $|Q^\lambda(G(x)) - Q^\lambda(G^k(x))|$ is estimated by

$$\left|Q^\lambda\big(G(x)\big) - Q^\lambda\big(G^k(x)\big)\right| \le \kappa_g \big(M_k^g\big)^2.$$

Combined with (3.4), $\phi_r^k(\lambda)$ is accurate estimate of $\phi_r(\lambda)$ with the boundary

$$\left|\phi_r(\lambda) - \phi_r^k(\lambda)\right| \le \kappa_f \big(M_k^f\big)^2 + \kappa_g \left(\|\lambda\| + \frac{r}{2}\right) \big(M_k^g\big)^2. \tag{3.5}$$

To ensure convergence, the algorithm further need that the distance between $\nabla\phi_r$ and $\nabla\phi_r^k$ is not too far. So we give the following definition.

Definition 3.3. *For each $k \in \mathbb{N}$, the estimate $\phi_r^k(\lambda^k)$ is said to be μ_ϕ-accurate gradient estimate of $\phi_r(\lambda^k)$ with a given boundary M_k^ϕ, if*

$$\left\|\nabla\phi_r(\lambda^k) - \nabla\phi_r^k(\lambda^k)\right\| \le \mu_\phi M_k^\phi. \tag{3.6}$$

Definitions 3.1 and 3.2 propose the criteria for the estimation model (i.e. conditions must hold on all $x \in O$), while Definition 3.3 simply requires that the function $\nabla\phi_r^k$ be the accurate estimate at λ^k. However, it is not easy to calculate the difference between $\nabla\phi_r$ and $\nabla\phi_r^k$ at λ^k. So the following proposition provides a sufficient condition that $\nabla\phi_r^k$ is μ_ϕ-accurate estimate at λ^k.

Proposition 3.1. *Suppose that the "true" constraint function G is Lipschitz continue with modulus L_g and G^k is κ_g-accurate estimation model of G with the boundary M_k^g. If the distance between the optimal solution set of $\inf_{x\in X_0} \mathcal{L}_r(x, \lambda^k)$ and x^{k+1} has the bound*

$$\operatorname{dist}\big(x^{k+1}, \operatorname{agrmin}_{x\in X_0} \mathcal{L}_r(x,\lambda^k)\big) \le \mu_L M_k^g, \tag{3.7}$$

then ϕ_r^k is μ_ϕ-accurate gradient estimate of $\phi_r(\lambda^k)$ with the boundary $M_k^\phi = M_k^g$, where $\mu_\phi = \mu_L L_g + \kappa_g$.

Proof. Proposition 2.1 provides that $\nabla(\phi_r^k)(\lambda^k) = \max\{-\lambda^k/r,\ G^k(x^{k+1})\}$ and $\nabla(\phi_r)(\lambda^k) = \max\{-\lambda^k/r, G(\bar{x})\}$, where $\bar{x} \in \operatorname{agrmin}\mathcal{L}_r(x,\lambda^k)$. By the Lipschitz continue of G and (3.3), it is derived that

$$\begin{aligned}
\|\nabla\phi_r(\lambda^k) - \nabla\phi_r^k(\lambda^k)\| &= \left\|\max\left\{-\frac{\lambda^k}{r}, G(\bar{x})\right\} - \max\left\{-\frac{\lambda^k}{r}, G^k(x^{k+1})\right\}\right\| \\
&\le \left\|G(\bar{x}) - G^k(x^{k+1})\right\| \\
&\le \left\|G(\bar{x}) - G(x^{k+1})\right\| + \left\|G(x^{k+1}) - G^k(x^{k+1})\right\| \\
&\le L_g\|\bar{x} - x^{k+1}\| + \kappa_g M_k^g \\
&\le L_g \operatorname{dist}\big(x^{k+1}, \operatorname{agrmin}_{x\in X_0}\mathcal{L}_r(x,\lambda^k)\big) + \kappa_g M_k^g \\
&\le (L_g\mu_L + \kappa_g) M_k^g,
\end{aligned}$$

which prove the conclusion. □

Remark 3.1. The condition (3.7) can be satisfied under certain conditions. One of the sufficient conditions is proposed by [6, Theorem 5.53]. Suppose that the true problem (1.1) has unique optimal solution, Mangasarian-Fromovitz constraint qualification holds at the optimal solution, the set of Lagrange multipliers is nonempty, the strong second order sufficient conditions are satisfied and the feasible set of the approximation problem (1.4) is nonempty and uniformly bounded for any κ_f-accurate estimation model f^k and κ_g-accurate estimation model G^k, then there exists $\mu_L > 0$ such that the condition (3.7) is satisfied. Although there are very strong conditions, in the proof of the convergence, we only need to assume that (3.7) holds with high enough, but fixed, probability. Therefore, the above conditions only need to be established with high probability.

Our models $\{f^k, G^k\}$ are generated by some random samples of stochastic function f and G on each iteration. Hence, the models themselves are random and contribute to the randomness of the iterates X_k and the multipliers Λ_k. Let $x_k = X_k(\omega)$ and $\lambda_k = \Lambda_k(\omega)$ denote their respective realizations. Moreover denote

$$\Delta^k := \max\left\{-\frac{\Lambda^k}{r}, G^k(X^{k+1})\right\}, \quad \Phi^k_{r,\lambda} := \phi^k_r(\Lambda^k)$$

and their realizations $\delta^k = \Delta^k(\omega), \phi^k_{r,\lambda} = \Phi^k_{r,\lambda}(\omega)$. We now combine Definitions 3.1-3.3 and extend the definitions of probabilistically accurate estimation models and probabilistically accurate gradient estimates that is used in the remainder of the paper.

Definition 3.4. *A sequence of random function estimation models $\{f^k, G^k\}$ is said to be α-probabilistically κ-accurate-model with the boundary $\|\Delta^k\|$ where $\kappa = (\kappa_f, \kappa_g)$ and*

$$\Delta^k = \max\left\{-\frac{\Lambda^k}{r}, G^k(X^{k+1})\right\}$$

if the events

$$\begin{aligned} I_k = \big\{ & f^k \text{ is } \kappa_f\text{-accurate estimation model with the boundary } \|\Delta^k\| \text{ and} \\ & G^k \text{ is } \kappa_g\text{-accurate estimation model with the boundary } \|\Delta^k\|\big\} \end{aligned} \tag{3.8}$$

satisfy the condition

$$P\left(I_k \,|\, \mathcal{F}^{k-1}\right) \geq \alpha,$$

where κ_f and κ_g are fixed constants and $\mathcal{F}^{k-1}$ is the σ-algebra generated by $f^0, \cdots, f^{k-1}$ and $G^0, \cdots, G^{k-1}$.

Definition 3.5. *A sequence of random function estimates $\{\Phi^k_{r,\lambda}\}$ is said to be β-probabilistically μ_ϕ-accurate-gradient with the boundary $\|\Delta^k\|$ where $\Delta^k = \max\{-\Lambda^k/r, G^k(X^{k+1})\}$ if the events*

$$J_k = \big\{\Phi^k_{r,\lambda} \text{ is } \mu_\phi\text{-accurate gradient estimate with the boundary } \|\Delta^k\|\big\} \tag{3.9}$$

satisfy the condition

$$P\left(J_k \,|\, \mathcal{F}^{k-1}\right) \geq \beta,$$

where μ_ϕ is a fixed constant and $\mathcal{F}^{k-1}$ is the σ-algebra generated by $f^0, \cdots, f^{k-1}$ and $G^0, \cdots, G^{k-1}$.

For the parameters in the above definition, we make some comments. Firstly, to achieve the global convergence of Algorithm 1.1, in the analysis, this is only required that $\kappa_f, \kappa_g, \mu_\phi$ are fixed constants which have an upper bound and α, β are sufficiently large, but fixed, constants.

Secondly, the boundary $\|\Delta^k\|$ seems to be an unknown value. In fact, according to Proposition 2.1, $\Delta^k = \nabla_\lambda(\Phi^k_{r,\lambda})$. Here $\Lambda^k \in \mathcal{F}^{k-1}$ is a known, determined vector. Therefore, Δ^k is just a concrete representation of the gradient of ϕ^k_r at Λ^k. On the other hand,

$$\|\Delta^k\| = \frac{1}{r}\| \left[\Lambda^k + rG^k(X^{k+1})\right]_+ - \Lambda^k\| = \frac{1}{r}\|\Lambda^{k+1} - \Lambda^k\|,$$

which can be thought of as the distance from the next iteration point to Λ^k, then the key thought in the proof of convergence is $\|\Delta^k\|$ converges to 0 almost surely. Finally, during the iteration, it may happen that $\|\Delta^k\| = 0$. At this point, κ-accurate estimate for random function estimate $\{f^k, G^k\}$ is a too strict condition. Since I_k only needs to occur with a certain probability α, the convergence can be achieved even if some $\{f^k, G^k\}$ do not satisfy the condition. This conclusion also applies to estimates $\{\Phi^k_{r,\lambda}\}$. The following lemma provide conditions to guarantee the decrease of the dual problem $\phi_r(\lambda)$ in Algorithm 1.1.

Lemma 3.1. *Suppose that $\phi^k_r \not\equiv -\infty$ and estimates $\{f^k, G^k\}$ are κ-accurate with $\kappa = (\kappa_f, \kappa_g)$ with the boundary $\|\delta^k\|$ where $\delta^k = \max\{-\lambda^k/r,\ G^k(x^{k+1})\}$. Let $\kappa_g \in (0, 1/2)$. If at the k-th iteration,*

$$r \geq \frac{2\big(3\kappa_f + \kappa_g(\|\lambda^k\| + \|\lambda^{k+1}\|)\big)}{1 - 2\kappa_g}, \tag{3.10}$$

then the improvement of ϕ_r is bounded below as follows:

$$\phi_r(\lambda^k) - \phi_r(\lambda^{k+1}) \leq -\kappa_f\|\delta^k\|^2. \tag{3.11}$$

Proof. Let $z(\lambda^k)$ is the optimal solution of the maximization problem at the right side of (2.3) with $\lambda = \lambda^k$, which derives that

$$\phi^k_r(\lambda^k) = \phi^k_0\big(z(\lambda^k)\big) - \frac{1}{2r}\|z(\lambda^k) - \lambda^k\|^2. \tag{3.12}$$

From the definition of the proximal mapping, we have

$$\begin{aligned} P_r\big[-\phi^k_0\big](\lambda^k) &= \operatorname{argmin}_z \left\{-\phi^k_0(z) + \frac{1}{2r}\|z - \lambda^k\|^2\right\} \\ &= \operatorname{argmax}_z \left\{\phi^k_0(z) - \frac{1}{2r}\|z - \lambda^k\|^2\right\} = z(\lambda^k), \end{aligned}$$

so combined with (2.7), we get

$$\nabla\phi^k_r(\lambda^k) = \frac{1}{r}\big[P_r\big[-\phi^k_0\big](\lambda^k) - \lambda^k\big] = \frac{1}{r}\big[z(\lambda^k) - \lambda^k\big],$$

which yields from (3.12) that

$$\begin{aligned} &\phi^k_0\big(z(\lambda^k)\big) - \frac{1}{2r}\big\|z(\lambda^k) - \lambda'\big\|^2 \\ =\ &\phi^k_0\big(z(\lambda^k)\big) - \frac{1}{2r}\big\|z(\lambda^k) - \lambda^k\big\|^2 + \frac{1}{2r}\big\|z(\lambda^k) - \lambda^k\big\|^2 - \frac{1}{2r}\big\|z(\lambda^k) - \lambda'\big\|^2 \\ =\ &\phi^k_r(\lambda^k) - \frac{1}{r}\big\langle\lambda' - \lambda^k, z(\lambda^k) - \lambda^k\big\rangle - \frac{1}{2r}\|\lambda' - \lambda^k\|^2 \\ =\ &\phi^k_r(\lambda^k) + \big\langle\lambda' - \lambda^k, \nabla\phi^k_r(\lambda^k)\big\rangle - \frac{1}{2r}\|\lambda' - \lambda^k\|^2. \end{aligned}$$

For any $\lambda' \in \mathbb{R}^p$, from the expression of ϕ_r^k in (2.3), it has

$$\phi_r^k(\lambda') \geq \phi_0^k\big(z(\lambda^k)\big) - \frac{1}{2r}\big\|z(\lambda^k) - \lambda'\big\|^2, \tag{3.13}$$

which derives that

$$\phi_r^k(\lambda') \geq \phi_r^k(\lambda^k) + \big\langle \lambda' - \lambda^k, \nabla\phi_r^k(\lambda^k)\big\rangle - \frac{1}{2r}\|\lambda' - \lambda^k\|^2. \tag{3.14}$$

In the above derivation, taking $\lambda' = \lambda^{k+1}$, with $\delta^k = 1/r(\lambda^{k+1} - \lambda^k)$ and $\nabla\phi_r^k(\lambda^k) = \delta^k$, it implies that

$$\phi_r^k(\lambda^k) - \phi_r^k(\lambda^{k+1}) \leq -\frac{r}{2}\|\delta^k\|^2. \tag{3.15}$$

Since estimates $\{f^k, G^k\}$ are κ-accurate with the boundary $\|\delta^k\|$, by the analysis of (3.5), it infers that for any λ^k,

$$\big|\phi_r(\lambda^k) - \phi_r^k(\lambda^k)\big| \leq \kappa_f\|\delta^k\|^2 + \kappa_g\left(\|\lambda^k\| + \frac{r}{2}\right)\|\delta^k\|^2. \tag{3.16}$$

Combined with (3.15), (3.16) and the restriction of r, the improvement in ϕ_r can be bounded as

$$\begin{aligned}\phi_r(\lambda^k) - \phi_r(\lambda^{k+1}) &\leq \phi_r^k(\lambda^k) - \phi_r^k(\lambda^{k+1}) + \big(2\kappa_f + \kappa_g(\|\lambda^k\| + \|\lambda^{k+1}\| + r)\big)\|\delta^k\|^2\\ &\leq -\frac{r}{2}\|\delta^k\|^2 + \big(2\kappa_f + \kappa_g(\|\lambda^k\| + \|\lambda^{k+1}\| + r)\big)\|\delta^k\|^2\\ &\leq -\kappa_f\|\delta^k\|^2.\end{aligned}$$

The proof is complete. □

In the above analysis, no assumptions are made on κ_f. Furthermore, it is observed that the greater κ_f is, the weaker the condition for estimates f^k is, and the faster the decrease of ϕ_r can be obtained, but the stronger the condition for the parameter r is. In the convergence theorem, it can be seen that r determines the choice of the probability α and β. Therefore, in the convergence analysis, the greater κ_f is obtained, the higher the requirement of the probability α and β will be.

To prove convergence of Algorithm 1.1, we need that $\|\Delta^k\|$ converges to 0 with probability 1, which has a close relationship with the boundedness of $\{\|\Lambda^k\|\}$. Hence, we would like to discuss under what conditions can guarantee the boundedness of $\{\|\Lambda^k\|\}$.

Lemma 3.2. *Suppose that the sequence of the optimal value $\{\|\max\phi_r^k\|\}$ of the problem (2.2) is bounded. Moreover, there exist a positive sequence $\varepsilon_k \downarrow 0$ and $0 < \bar{k} \in \mathbb{N}$ such that for any $k \geq \bar{k}$, the selected estimation models f^k and G^k satisfy*

$$\begin{aligned}&\sup_{x\in X_0} f^{k-1}(x) - f^k(x) \leq \varepsilon_k,\\ &\sup_{x\in X_0}\left\|\Pi_{\mathbb{R}^p_+}\left(\frac{\Lambda^k}{r} + G^{k-1}(x)\right)\right\|^2 - \left\|\Pi_{\mathbb{R}^p_+}\left(\frac{\Lambda^k}{r} + G^k(x)\right)\right\|^2 \leq \varepsilon_k.\end{aligned} \tag{3.17}$$

Then $\{\|\Lambda^k\|\}$ is bounded.

Proof. We prove the conclusion now by a contradiction. Since G^k is continue and X_0 is nonempty bounded set, there exists a positive constant M_X so that $sup_{x\in X_0}\ \|G^k(x)\| \leq M_X$. Suppose that there exist a subsequence $\{k_i\}_i$ and a subsequence realization $\{\bar{\xi}^{k_i}\}$ such that

$\{\|\lambda^{k_i}(\bar{\xi}^{k_i})\|\}_i$ is unbounded, but for any $j \notin \{k_i\}_i$ and all corresponding realization $\bar{\xi}$, there is a positive constant M_λ with $\|\lambda^j(\bar{\xi})\| \le M_\lambda$. Therefore, there exists an $i_0 \in \mathbb{N}$ such that for any $i > i_0$, $\lambda^{k_i}(\bar{\xi}^{k_i}) > 2(M_\lambda + rM_X)$, but $\lambda^{k_i-1}(\bar{\xi}^{k_i-1}) \le M_\lambda$. On the other hand,

$$\begin{aligned}\|\lambda^{k_i}(\bar{\xi}^{k_i})\| &\le \|\lambda^{k_i-1}(\bar{\xi}^{k_i-1}) + rG^{k_i-1}(x^{k_i}(\bar{\xi}^{k_i}))\| \\ &\le \|\lambda^{k_i-1}(\bar{\xi}^{k_i-1})\| + r\|G^{k_i-1}(x^{k_i}(\bar{\xi}^{k_i}))\| \le M_\lambda + rM_X,\end{aligned}$$

which contradicts the assumption of the unboundedness of the subsequence $\{\|\lambda^{k_i}\|\}_i$.

So, in the following, we assume that there exists a sequence realization $\{\bar{\xi}^k\}$ the sequence such that $\{\|\lambda^k(\bar{\xi}^k)\|\}$ is unbounded with the probability α_0. Algorithm 1.1 can unconditionally guarantee that (3.15) is established. Let

$$\delta^k(\bar{\xi}^k, \bar{\xi}^{k+1}) = \max\left\{-\frac{\lambda^k(\bar{\xi}^k)}{r}, G^k(x^{k+1}(\bar{\xi}^{k+1}))\right\}.$$

It yields that

$$\begin{aligned}\phi_r^k(\lambda^{k+1}(\bar{\xi}^{k+1})) &\ge \phi_r^k(\lambda^k(\bar{\xi}^k)) + \frac{r}{2}\|\delta^k(\bar{\xi}^k, \bar{\xi}^{k+1})\|^2 \\ &\ge \left(\phi_r^k(\lambda^k(\bar{\xi}^k)) - \phi_r^{k-1}(\lambda^k(\bar{\xi}^k))\right) + \phi_r^{k-1}(\lambda^{k-1}(\bar{\xi}^{k-1})) \\ &\quad + \frac{r}{2}\left(\|\delta^{k-1}(\bar{\xi}^{k-1}, \bar{\xi}^k)\|^2 + \|\delta^k(\bar{\xi}^k, \bar{\xi}^{k+1})\|^2\right).\end{aligned}$$

Thus, for any $k \ge \bar{k}$, it has

$$\begin{aligned}\phi_r^k(\lambda^{k+1}(\bar{\xi}^{k+1})) \ge &\sum_{i=\bar{k}}^{k}\left(\phi_r^i(\lambda^i(\bar{\xi}^i)) - \phi_r^{i-1}(\lambda^i(\bar{\xi}^i))\right) + \phi_r^{\bar{k}-1}(\lambda^{\bar{k}-1}(\bar{\xi}^{\bar{k}-1})) \\ &+ \frac{r}{2}\sum_{i=\bar{k}-1}^{k}\|\delta^k(\bar{\xi}^i, \bar{\xi}^{i+1})\|^2. \qquad (3.18)\end{aligned}$$

Since (3.17) holds, for the sequence realization $\{\bar{\xi}^k\}$, we have

$$\begin{aligned}&\phi_r^{k-1}(\lambda^k(\bar{\xi}^k)) - \phi_r^k(\lambda^k(\bar{\xi}^k)) \\ &= \inf_{x\in X_0}\mathcal{L}_r^{k-1}(x, \lambda^k(\bar{\xi}^k)) - \inf_{x\in X_0}\mathcal{L}_r^k(x, \lambda^k(\bar{\xi}^k)) \\ &\le \sup_{x\in \operatorname{argmin}\mathcal{L}_r^k(x,\lambda^k(\bar{\xi}^k))}\left\{\mathcal{L}_r^{k-1}(x, \lambda^k(\bar{\xi}^k)) - \mathcal{L}_r^k(x, \lambda^k(\bar{\xi}^k))\right\} \\ &\le \sup_{x\in X_0}\left\{(f^{k-1}(x) - f^k(x)) + \frac{1}{2r}\left(\|\Pi_{\mathbb{R}_+^p}(\lambda^k(\bar{\xi}^k) + rG^{k-1}(x))\|^2 - \|\Pi_{\mathbb{R}_+^p}(\lambda^k(\bar{\xi}^k) + rG^k(x))\|^2\right)\right\} \\ &\le \sup_{x\in X_0}\{f^{k-1}(x) - f^k(x)\} + \frac{r}{2}\sup_{x\in X_0}\left\{\left\|\Pi_{\mathbb{R}_+^p}\left(\frac{\lambda^k(\bar{\xi}^k)}{r} + G^{k-1}(x)\right)\right\|^2 - \left\|\Pi_{\mathbb{R}_+^p}\left(\frac{\lambda^k(\bar{\xi}^k)}{r} + G^k(x)\right)\right\|^2\right\} \\ &\le \frac{2+r}{2}\varepsilon_k.\end{aligned}$$

From the conclusion of Theorem 3.1, the sequence $\{\|\lambda^k(\bar{\xi}^k)\|\}$ is unbounded if and only if there exists a positive constant τ so that for any $k \in \mathbb{N}$, $\|\delta^k(\bar{\xi}^k, \bar{\xi}^{k+1})\| \ge \tau > 0$. Since $\varepsilon_k \downarrow 0$, there exists $\tilde{k} \ge \bar{k}$ so that $\varepsilon_k \le r\tau^2/(2(2+r))$. That is,

$$\phi_r^{k-1}(\lambda^k(\bar{\xi}^k)) - \phi_r^k(\lambda^k(\bar{\xi}^k)) \le \frac{r\tau^2}{4} \le \frac{r}{4}\|\delta^k(\bar{\xi}^k, \bar{\xi}^{k+1})\|^2.$$

Therefore, combined with (3.18), for any $k \geq \tilde{k}$, it implies that

$$\begin{aligned}
\phi_r^k(\lambda^{k+1}(\bar{\xi}^{k+1})) \geq& \sum_{i=\tilde{k}}^{k} \Big(\phi_r^i(\lambda^i(\bar{\xi}^i)) - \phi_r^{i-1}(\lambda^i(\bar{\xi}^i))\Big) + \phi_r^{\tilde{k}-1}(\lambda^{\tilde{k}-1}(\bar{\xi}^{\tilde{k}-1})) \\
&+ \frac{r}{2} \sum_{i=\tilde{k}-1}^{k} \|\delta^i(\bar{\xi}^i, \bar{\xi}^{i+1})\|^2 \\
\geq& -\sum_{i=\tilde{k}}^{k} \frac{r}{4}\|\delta^i(\bar{\xi}^i, \bar{\xi}^{i+1})\|^2 + \phi_r^{\tilde{k}-1}(\lambda^{\tilde{k}-1}(\bar{\xi}^{\tilde{k}-1})) \\
&+ \frac{r}{2} \sum_{i=\tilde{k}-1}^{k} \|\delta^i(\bar{\xi}^i, \bar{\xi}^{i+1})\|^2 \\
\geq& \phi_r^{\tilde{k}-1}(\lambda^{\tilde{k}-1}(\bar{\xi}^{\tilde{k}-1})) + \frac{r}{4}\sum_{i=\tilde{k}}^{k} \|\delta^i(\bar{\xi}^i, \bar{\xi}^{i+1})\|^2 \\
\geq& \phi_r^{\tilde{k}-1}(\lambda^{\tilde{k}-1}(\bar{\xi}^{\tilde{k}-1})) + \frac{r}{4}\sum_{i=\tilde{k}}^{k} \tau^2 \quad \to \quad +\infty.
\end{aligned}$$

As $k \to \infty$, the right side of the above formula tends to be positive infinity, so $\phi_r^k(\lambda^{k+1}(\bar{\xi}^{k+1})) \to +\infty$, which contradicts the assumption of the boundedness of the optimal value $\{\| \max \phi_r^k \|\}$. As a consequence, $\{\|\lambda^k\|\}$ is bounded. □

It should be emphasized that the conditions given above are only one of the sufficient conditions for the boundedness of $\|\Lambda^k\|$. Since f^{k-1}, G^{k-1} and Λ^k are both determined in the k-th iteration, (3.17) is an operable condition.

The following theorem states that as long as the probability α can be chosen specially, under the mild assumptions of Lemma 3.1, $\|\Delta^k\|$ converges to 0 is equivalent to the sequence $\{\|\Lambda^k\|\}$ is bounded.

Theorem 3.1. *Suppose that ϕ_r is bounded from above on $\mathbb{R}^p$ and $\phi_r^k \not\equiv -\infty$. Moreover, estimates $\{f^k, G^k\}$ are α-probabilistically κ-accurate-model and $\{\Phi_{r,\lambda}^k\}$ is β-probabilistically μ_ϕ-accurate-gradient both with the boundary $\|\Delta^k\|$ where $\Delta^k = \max\{-\Lambda^k/r,\ G^k(X^{k+1})\}$. Let $\kappa_g \in (0, 1/2)$ and α, β satisfies*

$$\alpha\beta > \frac{r}{\nu + r}, \tag{3.19}$$

where

$$\nu = \frac{2\kappa_f}{2 + (1 + \mu_\phi)^2}.$$

Then the sequence $\{\|\Lambda^k\|\}$ is bounded almost surely if and only if there exists a constant $\bar{r}$ such that for any $r > \bar{r}$,

$$\sum_{k=0}^{\infty} \|\Delta^k\|^2 < \infty \tag{3.20}$$

holds almost surely.

Proof. We first prove the sufficiency and argue the conclusion by a contradiction. We assume that if (3.20) holds almost surely, there exists a sequence realization $\{\bar{\xi}^k\}$ such that the sequence $\{\|\lambda^k(\bar{\xi}^k)\|\}$ is unbounded with the probability α_0. Then, from (1.7) and

$$\delta^k = \max\left\{-\frac{\lambda^k}{r}, G^k(x^{k+1})\right\} = r(\lambda^{k+1} - \lambda^k),$$

for any $r \geq \bar{r}$, it holds that

$$\|\lambda^{k+1}\| \leq \|\lambda^k\| + r\|\delta^k\| \leq \|\lambda^0\| + r\sum_{i=0}^{k}\|\delta^i\|^2.$$

Since the sequence $\{\|\lambda^k(\bar{\xi}^k)\|\}$ is unbounded with the probability α_0, the sequence $\sum_{i=0}^{k}\|\delta^i(\bar{\xi}^i)\|^2$ is unbounded with the probability α_0, which contradicts the assumption (3.20). Hence, the sequence $\{\|\Lambda^k\|\}$ is bounded almost surely.

The following proves the necessity. As usual, let $x_k, \lambda_k, \delta^k, \phi^k_{r,\lambda}$ denote realizations of random quantities $X_k, \Lambda_k, \Delta^k, \Phi^k_{r,\lambda}$, respectively. Now we consider all realizations of Algorithm 1.1 and we estimate $\phi_r(\lambda^k) - \phi_r(\lambda^{k+1})$ in two cases.

(a) The events I_k and J_k both occur. If $\{\|\Lambda^k\|\}$ is bounded almost surely, i.e. the probability that $\{\|\Lambda^k\|\}$ is unbounded equals to zero, then there exists a boundary M_λ such that $\mathbb{P}(\|\Lambda^k\| \leq M_\lambda) = 1$. Choose

$$\bar{r} = \frac{2(3\kappa_f + 2\kappa_g M_\lambda)}{1 - 2\kappa_g},$$

by Lemma 3.1, for any $r \geq \bar{r}$, (3.11) holds with probability α. The event J_k implies that

$$\left\|\nabla\phi_r(\lambda^k) - \delta^k\right\| = \left\|\nabla\phi_r(\lambda^k) - \nabla\phi^k_r(\lambda^k)\right\| \leq \mu_\phi\|\delta^k\|.$$

Hence, $\|\nabla\phi_r(\lambda^k)\| \leq (1 + \mu_\phi)\|\delta^k\|$ and (3.11) is rewritten as

$$\phi_r(\lambda^k) - \phi_r(\lambda^{k+1}) \leq -\kappa_f\|\delta^k\|^2 \leq -\nu\|\delta^k\|^2 - \frac{\kappa_f - \nu}{(1+\mu_\phi)^2}\left\|\nabla\phi_r(\lambda^k)\right\|^2 < 0. \tag{3.21}$$

(b) At least one of the event I_k or J_k does not hold. Observe that the proof of formula (3.14) is similarly applied to ϕ_r. With $\lambda' = \lambda^{k+1}$, it has

$$\phi_r(\lambda^k) - \phi_r(\lambda^{k+1}) \leq \left\langle\lambda^k - \lambda^{k+1}, \nabla\phi_r(\lambda^k)\right\rangle + \frac{1}{2r}\left\|\lambda^{k+1} - \lambda^k\right\|^2.$$

Hence, the change in function ϕ_r is bounded by

$$\phi_r(\lambda^k) - \phi_r(\lambda^{k+1}) \leq -r\left\langle\delta^k, \nabla\phi_r(\lambda^k)\right\rangle + \frac{r}{2}\|\delta^k\|^2. \tag{3.22}$$

Combining the above two cases, since the events I_k and J_k both occur at least with the probability $\alpha\beta$, it implies that

$$\begin{aligned}
&\mathbb{E}\left[\phi_r(\Lambda^k) - \phi_r(\Lambda^{k+1}) \,|\, \mathcal{F}^{k-1}\right] \\
&\leq -\alpha\beta\nu\|\Delta^k\|^2 - \alpha\beta\frac{\kappa_f - \nu}{(1+\mu_\phi)^2}\left\|\nabla\phi_r(\Lambda^k)\right\|^2 \\
&\quad - (1-\alpha\beta)r\left\langle\Delta^k, \nabla\phi_r(\Lambda^k)\right\rangle + \frac{(1-\alpha\beta)r}{2}\|\Delta^k\|^2.
\end{aligned}$$

From the selection of α, β in (3.19), it has $\alpha\beta\nu/2 > (1-\alpha\beta)r/2$. With

$$\frac{\nu}{2} = \frac{\kappa_f - \nu}{(1+\mu_\phi)^2},$$

so, it yields that

$$\begin{aligned}
&\mathbb{E}\left[\phi_r(\Lambda^k) - \phi_r(\Lambda^{k+1}) \,|\, \mathcal{F}^{k-1}\right] \\
&\leq -\frac{\alpha\beta\nu}{2}\|\Delta^k\|^2 - \frac{\alpha\beta\nu}{2}\left\|\nabla\phi_r(\Lambda^k)\right\|^2 - (1-\alpha\beta)r\left\langle\Delta^k, \nabla\phi_r(\Lambda^k)\right\rangle \\
&= -\left(\frac{\alpha\beta\nu}{2} - \frac{(1-\alpha\beta)r}{2}\right)(\|\Delta^k\|^2 + \|\nabla\phi_r(\Lambda^k)\|^2) \\
&\quad - \frac{(1-\alpha\beta)r}{2}\left\|\Delta^k + \nabla\phi_r(\Lambda^k)\right\|^2 < 0
\end{aligned} \tag{3.23}$$

holds for any $k \in \mathbb{N}$. Since ϕ_r has an upper bound, summing (3.23) over $k \in (1, \infty)$ and taking expectation on both sides, it can conclude that (3.20) holds with probability 1. □

Remark 3.2. From the above analysis, it also can infer that

$$\sum_{k=0}^{\infty}\left\|\nabla\phi_r(\Lambda^k)\right\|^2 < \infty, \quad \sum_{k=0}^{\infty}\left\|\Delta^k - \nabla\phi_r(\Lambda^k)\right\|^2 < \infty \tag{3.24}$$

almost sure holds with probability 1. In fact,

$$\begin{aligned}
\sum_{k=0}^{\infty}\left\|\Delta^k - \nabla\phi_r(\Lambda^k)\right\|^2 &\leq \sum_{k=0}^{\infty}\left(2\left\|\Delta^k + \nabla\phi_r(\Lambda^k)\right\|^2 + 8\left\|\nabla\phi_r(\Lambda^k)\right\|^2\right) \\
&\leq 2\sum_{k=0}^{\infty}\left\|\Delta^k + \nabla\phi_r(\Lambda^k)\right\|^2 + 8\sum_{k=0}^{\infty}\left\|\nabla\phi_r(\Lambda^k)\right\|^2 \leq \infty
\end{aligned}$$

holds with probability 1.

Remark 3.3. Since there exists a constant $\bar{r}$ such that for any $r > \bar{r}$, the probability α and β are totally constants and selected as $\alpha = \beta = \sqrt{r/(\nu + r)}$. In addition, observe that the parameter μ_ϕ has no restrictions in the above proof. Therefore, in theory, the convergence only need the existence of $\mu_\phi > 0$. However, in actual situations, the larger μ_ϕ is, the greater the probability $\alpha\beta$ of the occurrence of events I_k and J_k need to be.

In the last part of this section, the global convergence of the augmented Lagrange method is established.

Theorem 3.2. *Let the assumptions of Theorem* 3.1 *hold. Moreover, suppose that generalized Slater condition holds for* (1.1) *and the multiplier set* $\{\|\Lambda^k\|\}$ *is bounded almost surely. Then any cluster point* $\bar{X}$ *of the sequence* $\{X^k\}$ *generated by Algorithm* 1.1 *is the optimal solution of* (1.1) *almost surely.*

Proof. Replacing ϕ_r^k in Propositions 2.1 and 2.2 with ϕ_r, the conclusions about ϕ_r are still established. Hence, for any $r > 0$, the dual problems (D_r) and (D_0) have the same optimal solution. Furthermore, let Λ^* be the cluster point of the sequence $\{\Lambda^k\}$. If Λ^* is the dual optimal solution of (D_r), Λ^* is also the dual optimal solution of (D_0). Since generalized Slater condition holds for (1.1), the dual gap between (1.1) and its dual problem (D_0) is zero. Based on Proposition 2.2, if Λ^* is the dual optimal solution of (D_0), then $\bar{X}$ is an optimal solution to (1.1) if and only if $\bar{X}$ is the minimum of the function $\mathcal{L}_r(\cdot, \Lambda^*)$ on X_0. Based on the above analysis, it only need to prove $\bar{X}$ is the minimum of the function $\mathcal{L}_r(\cdot, \Lambda^*)$, where Λ^* is the dual optimal solution of (D_r).

With the assumptions of Theorem 3.1, (3.20) and (3.24) hold almost surely. (3.20) implies that

$$\lim_{k\to\infty} \|\Lambda^{k+1} - \Lambda^k\| = \lim_{k\to\infty} r\|\Delta^k\| = 0 \quad \text{a.s.}$$

Since $\{\|\Lambda^k\|\}$ is bounded almost surely, $\{\Lambda^k\}$ converges to a certain limit point Λ^* almost surely. $\sum_{k=0}^{\infty} \|\nabla\phi_r(\Lambda^{k+1})\|^2 < \infty$ with probability 1 in (3.24) implies that

$$\lim_{k\to\infty} \big\|\nabla\phi_r(\Lambda^{k+1})\big\| = \|\nabla\phi_r(\Lambda^*)\| = 0$$

holds almost surely. That deduces that Λ^* is the dual optimal solution of (D_r) almost surely. On the other hand,

$$\lim_{k\to\infty} \big\|\nabla\phi_r(\Lambda^{k+1})\big\| = \lim_{k\to\infty} \left\|\max\left\{-\frac{\Lambda^k}{r}, G(X^{k+1})\right\}\right\| = \left\|\max\left\{-\frac{\Lambda^*}{r}, G(\bar{X})\right\}\right\| = 0 \tag{3.25}$$

holds with probability 1. (3.25) implies that $\nabla\phi_r(\Lambda^*) = \max\{-\Lambda^*/r, G(\bar{X})\}$ holds with probability 1. By Proposition 2.1, $\bar{X}$ must be the minimum of the function $\mathcal{L}_r(\cdot, \Lambda^*)$ almost surely. The conclusion is proved. □

4. Stochastic Noise in Different Settings

In this section, we discuss specific conditions where α-probabilistically κ-accurate-model and β-probabilistically μ_ϕ-accurate-gradient are satisfied under various settings of stochastic noise. The first one is unbiased stochastic noise, that is, functions are estimated by zero-mean noise with bounded variance. We construct a random approximation model and achieve the convergence by selecting the appropriate sample size. Secondly, we consider convex optimizations with biased stochastic noise, where the random approximation models have noise with some positive probability. We obtain α-probabilistically κ-accurate-model and β-probabilistically μ_ϕ-accurate-gradient by controlling the probability of the error of approximation models.

In the random approximation model (1.4), for each iteration, we select the noisy versions of the objective function f and the constraint function g_i. Let ξ be a random vector defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Each time the sample is selected to generate an estimated model, let $f^k(x) := f(x, \xi)$ and $G^k(x) := G(x, \xi)$.

4.1. Unbiased stochastic noise

In this part, we discuss conditions for α-probabilistically κ-accurate-model and β-probabilistically μ_ϕ-accurate-gradient under unbiased stochastic noise, which is most common in stochastic optimization. The accurate estimate of ϕ_r^k is closely related to the convergence of Algorithm 1.1. Recall that the accurate estimate of ϕ_r^k is obtained through f^k being κ_f-accurate estimation model and G^k being κ_g-accurate estimation model. In the following proposition, other conditions are given to ensure the accurate estimate of ϕ_r^k.

Proposition 4.1. *Suppose that for each $k \in \mathbb{N}$, the estimation model $\{f^k, G^k\}$ is $\kappa_{f,g}$-accurate estimation conjunctive model of $\{f, G\}$ with a given boundary $M_k^{f,g}$, i.e. for any $x \in X_0$,*

$$\max\Big\{\big|f(x) - f^k(x)\big|, \big|q_\lambda\big(G(x)\big) - q_\lambda\big(G^k(x)\big)\big|\Big\} \le \kappa_{f,g}\big(M_k^{f,g}\big)^2, \tag{4.1}$$

where the function

$$q_\lambda(y) := \left\| \Pi_{\mathbb{R}^p_+}\left(\frac{\lambda}{r} + y\right)\right\|^2.$$

Then for a given $\lambda, \phi_r^k(\lambda)$ *is* $(2+r)/2\kappa_{f,g}$*-accurate estimate of* $\phi_r(\lambda)$ *with the boundary* $M_k^{f,g}$*.*

Proof. For a given λ, similar to the analysis in (3.5), it has

$$\begin{aligned}\left|\phi_r(\lambda) - \phi_r^k(\lambda)\right| &\leq \sup_{x\in X_0}\left|\left(f(x) - f^k(x)\right) + \frac{1}{2r}\left(\left\|\Pi_{\mathbb{R}^p_+}\left(\lambda + rG(x)\right)\right\|^2 - \left\|\Pi_{\mathbb{R}^p_+}\left(\lambda + rG^k(x)\right)\right\|^2\right)\right| \\ &\leq \sup_{x\in X_0}\left|f(x) - f^k(x)\right| + \frac{r}{2}\sup_{x\in X_0}\left|q_\lambda\big(G(x)\big) - q_\lambda\big(G^k(x)\big)\right| \\ &\leq \frac{2+r}{2}\kappa_{f,g}\big(M_k^{f,g}\big)^2, \end{aligned} \tag{4.2}$$

which prove the conclusion. □

It is not hard to see that the parameter $\kappa_{f,g}$ depends on $\|\lambda\|$ and r. If $\{\|\lambda^k\|\}$ generated by Algorithm 1.1 is bounded and r is selected as (3.10), then for any k-th iteration in Algorithm 1.1, the parameter $\kappa_{f,g}$ is chosen by

$$\kappa_{f,g} = \min\{\kappa_f, \kappa_g\}. \tag{4.3}$$

Hence, (3.5) can be obtained, so the convergence analysis in this case is the same as Theorem 3.1 with $M_k^{f,g} = \|\delta^k\|$. Hence, our purpose is to construct a model so that the estimation model $\{f^k, G^k\}$ can satisfy $\kappa_{f,g}$-accurate-conjunctive-model with some sufficiently high probability α. One of the ideas is that the standard stochastic approximation is used to obtain effective models. In particular, the algorithm choose the i.i.d. realizations ξ_j of the random vector ξ. Therefore, let

$$f^k(x) = \frac{1}{m}\sum_{j=1}^m F(x,\xi_j), \quad q_\lambda^k(G)(x) = \frac{1}{m}\sum_{j=1}^m q_\lambda\big(\mathcal{G}(x,\xi_j)\big), \tag{4.4}$$

where $\mathcal{G}(x,\xi_j) = (G_1(x,\xi_j), \cdots, G_p(x,\xi_j))$, and

$$\phi_r^k(\lambda) = \frac{1}{m}\sum_{j=1}^m \hat{\phi}_r(\lambda,\xi_j) := \frac{1}{m}\sum_{j=1}^m \inf_{x\in X_0}\left\{F(x,\xi_j) + \frac{r}{2}\left[q_\lambda\big(\mathcal{G}(x,\xi_j)\big) - \left\|\frac{\lambda}{r}\right\|^2\right]\right\}, \tag{4.5}$$

which implies that

$$\nabla\phi_r^k(\lambda) = \frac{1}{m}\sum_{j=1}^m \nabla\hat{\phi}_r(\lambda,\xi_j).$$

We now give several mild assumptions for the random approximation model.

Assumption 4.1. *(A1) Functions f and G are estimated by zero-mean noise with bounded variance, i.e. there exist positive constants V_f and V_g such that for any $j = 1, \ldots, m, \lambda \in \mathbb{R}^p, x \in X_0$,*

$$\mathbb{E}_\xi\left[F(x,\xi_j)\right] = f(x), \qquad \mathbb{E}_\xi\left[q_\lambda\big(\mathcal{G}(x,\xi_j)\big)\right] = q_\lambda\big(G(x)\big), \tag{4.6}$$

$$\mathbb{E}_\xi\left[|F(x,\xi_j) - f(x)|^2\right] \leq V_f, \quad \mathbb{E}_\xi\left[\left|q_\lambda\big(\mathcal{G}(x,\xi_j)\big) - q_\lambda\big(G(x)\big)\right|^2\right] \leq V_g, \tag{4.7}$$

where the function

$$q_\lambda(y) := \left\| \Pi_{\mathbb{R}^p_+}\left(\frac{\lambda}{r} + y\right)\right\|^2.$$

(A2) The gradient of ϕ_r is estimated by stochastic first-order oracles with zero-mean noise and bounded variance, i.e. there exists a positive constant V_ϕ such that for all $k \in \mathbb{N}$ and any $j = 1, \ldots, m, \lambda \in \mathbb{R}^p,\ x \in X_0$,

$$\mathbb{E}_\xi\left[\nabla_\lambda \hat{\phi}_r\left(\lambda^k, \xi_j\right)\right] = \nabla_\lambda \phi_r(\lambda^k), \quad \mathbb{E}_\xi\left[\left\|\nabla_\lambda \hat{\phi}_r\left(\lambda^k, \xi_j\right) - \nabla_\lambda \phi_r(\lambda^k)\right\|^2\right] \le V_\phi.$$

Proposition 4.2. *Suppose that the random approximation models* (4.4) *and* (4.5) *satisfy Assumption* 4.1*, and the sample size m selected by Algorithm* 1.1 *satisfies*

$$m \ge \max\left\{\frac{\max\{V_f, V_g\}}{(1-\alpha)\kappa_{f,g}^2 \hat{\delta}_k^4}, \frac{V_\phi}{(1-\beta)\mu_\phi^2 \hat{\delta}_k^2}\right\}, \tag{4.8}$$

where $0 < \hat{\delta}_k \le \|\delta^k\|$ and $\delta^k = \max\{-\lambda^k/r,\ G^k(x^{k+1})\}$. Then estimates $\{f^k, G^k\}$ are α-probabilistically $\kappa_{f,g}$-accurate-conjunctive-model and $\{\phi_{r,\lambda}^k\}$ is β-probabilistically μ_ϕ-accurate-gradient both with the boundary $\|\delta^k\|$.

Proof. By Chebyshev inequality, for any $\theta > 0$,

$$\mathbb{P}\big(|f^k(x) - f(x)| > \theta\big) = \mathbb{P}\Big(\big|f^k(x) - \mathbb{E}_\xi[F(x,\xi)]\big| > \theta\Big) \le \frac{V_f}{m\theta^2},$$
$$\mathbb{P}\Big(\big|q_\lambda^k(G)(x) - q_\lambda\big(G(x)\big)\big| > \theta\Big) = \mathbb{P}\Big(\big|q_\lambda^k(G)(x) - \mathbb{E}_\xi\left[q_\lambda\big(\mathcal{G}(x,\xi)\big)\right]\big| > \theta\Big) \le \frac{V_g}{m\theta^2}.$$

Choose $\theta = \kappa_{f,g}\hat{\delta}_k^2$ for some special $0 < \hat{\delta}_k \le \|\delta^k\|$ and m satisfies

$$\max\left\{\frac{V_f}{m\theta^2}, \frac{V_g}{m\theta^2}\right\} \le 1 - \alpha.$$

So the random approximation model (4.4) is $\kappa_{f,g}$-accurate-conjunctive-model with probability α provided with

$$m \ge \frac{\max\{V_f, V_g\}}{(1-\alpha)\kappa_{f,g}^2 \hat{\delta}_k^4}.$$

Reuse the extension of the Chebyshev inequality, then $\{\phi_r^k\}$ is μ_ϕ-accurate-gradient with probability β if

$$m \ge \frac{V_\phi}{(1-\beta)\mu_\phi^2 \hat{\delta}_k^2}.$$

As a result, the random approximation model (4.4) by Algorithm 1.1 guarantees convergence provided with

$$m \ge \max\left\{\frac{\max\{V_f, V_g\}}{(1-\alpha)\kappa_{f,g}^2 \hat{\delta}_k^4}, \frac{V_\phi}{(1-\beta)\mu_\phi^2 \hat{\delta}_k^2}\right\}.$$

The proof is complete. □

Proposition 4.2 states that if we select the appropriate sample size in Algorithm 1.1, the random approximation models (4.4) and (4.5) are guaranteed to estimate the problem (1.1) with sufficient accuracy, which can establish the convergence under the boundedness of the multiplier set. We summarize as the following theorem.

Theorem 4.1. *Suppose that generalized Slater condition holds for* (1.1), *and ϕ_r is bounded from above on $\mathbb{R}^p$. Moreover, the random approximation models $\{f^k, G^k\}$ in* (4.4) *and $\{\phi_r^k\}$ in* (4.5) *satisfy Assumption* 4.1, *and $\phi_r^k \not\equiv -\infty$. Let $\kappa_{f,g} \in (0, 1/2)$,*

$$r \geq \bar{r} := \frac{6\kappa_{f,g}}{1 - 2\kappa_{f,g}},$$

and α, β satisfies (3.19), *where*

$$\nu = \frac{2\kappa_{f,g}}{2 + (1 + \mu_\phi)^2},$$

and the sample size m selected by Algorithm 1.1 *satisfies* (4.8), *where $0 < \hat{\delta}_k \leq \|\delta^k\|$ and $\delta^k = \max\{-\lambda^k/r,\ G^k(x^{k+1})\}$. If the multiplier set $\{\|\lambda^k\|\}$ is bounded almost surely, any cluster point $\bar{x}$ of the sequence $\{x^k\}$ generated by Algorithm* 1.1 *is the optimal solution of* (1.1) *almost surely.*

Proof. By (3.15), (4.2) and the restriction of r, the improvement in ϕ_r can be bounded as

$$\begin{aligned}\phi_r(\lambda^k) - \phi_r(\lambda^{k+1}) &\leq \phi_r^k(\lambda^k) - \phi_r^k(\lambda^{k+1}) + \left|\phi_r(\lambda^k) - \phi_r^k(\lambda^k)\right| + \left|\phi_r(\lambda^{k+1}) - \phi_r^k(\lambda^{k+1})\right| \\ &\leq -\frac{r}{2}\|\delta^k\|^2 + (2 + r)\kappa_{f,g}\|\delta^k\|^2 \\ &\leq -\kappa_{f,g}\|\delta^k\|^2.\end{aligned}$$

It yields from Proposition 4.2 that $\{f^k, G^k\}$ are α-probabilistically $\kappa_{f,g}$-accurate-conjunctive-model and $\{\phi_{r,\lambda}^k\}$ is β-probabilistically μ_ϕ-accurate-gradient both with the boundary $\|\delta^k\|$. Similar to the proof of Theorem 3.1, we can prove that $\sum_{k=0}^{\infty} \|\delta^k\|^2 < \infty$ holds almost surely. Then the conclusion can be proved by Theorem 3.2. □

4.2. Biased stochastic noise

In many applications of economics and machine learning, there are many stochastic problems with complex noise structures. One example is in portfolio problems which focus on minimizing the variance subject to budget constraints as follow:

$$\begin{aligned}\min_{w \in \mathbb{R}_+^d, \sum_{i=1}^d w_i = 1} \quad & \langle w, \mathbb{E}[aa^T]w\rangle \\ \text{s.t.} \quad & \mathbb{E}[\langle a, w\rangle] \geq \gamma,\end{aligned}$$

where parameters a in the convex optimization models are random because ξ and the probability distributions are even unknown. Therefore, the full evaluation of the objective and constraint functions are impossible to obtain in practice. Other examples like Neyman-Pearson classification optimization models [19] and some online convex optimizations [14]. In these stochastic optimization models, due to unknown random components, the random approximation models $\{f^k, G^k\}$ may have large noise with some positive probability under some numerical methods. More likely, the probability of these deviations caused by the noises depends on x (see [10]). So it is reasonable to discuss a random approximation model as follows:

$$\begin{aligned}\min_{x \in X_0} \quad & f^k(x) = f(x, \xi) = \begin{cases} f(x) & \text{with probability} \quad 1 - \gamma_f(x), \\ \xi_f(x) \leq V_f & \text{with probability} \quad \gamma_f(x), \end{cases} \\ \text{s.t.} \quad & G^k(x) = G(x, \xi) = \begin{cases} 0 & \text{with probability} \quad 1 - \gamma_g(x), \\ \xi_g(x) \leq V_g & \text{with probability} \quad \gamma_g(x), \end{cases}\end{aligned} \tag{4.9}$$

where the probability $\gamma_f(x)$ and $\gamma_g(x)$ depend on x so that the function f and G can not compute accurately, and $\xi_f(x)$ and $\xi_g(x)$ are some random functions of x which are controlled by the known upper bound V_f and V_g. (4.9) is an idealized model due to the exact computation of f and G with probability $(1-\gamma_f(x))(1-\gamma_g(x))$. In practice, a sufficiently small error is allowed between the random approximation model and f, G under probability $(1-\gamma_f(x))(1-\gamma_g(x))$.

Obviously, the model is not unbiased, i.e.

$$\mathbb{E}_\xi[f(x,\xi)] = \big(1-\gamma_f(x)\big)f(x) + \gamma_f(x)\mathbb{E}[\xi_f(x)] \neq f(x), \quad \mathbb{E}_\xi[G(x,\xi)] \neq G(x).$$

Hence, traditional stochastic approximation techniques (like [7,8,12,25]), can not used in this situation. However, this model satisfies our convergence conditions by assuming that probability $\min\{\gamma_f(x),\ \gamma_g(x)\} \leq \bar{\gamma}$ is small enough, for any $x \in X_0$. When $(1-\bar{\gamma})^2 \geq \max\{\alpha,\ \beta\}$, $\{f^k, G^k\}$ is the exact computation of f and G with probability $(1-\bar{\gamma})^2$. Hence, with probability $(1-\bar{\gamma})^2$, $\{f^k, G^k\}$ is κ-accurate-model and estimates $\{\phi_r^k\}$ is μ_ϕ-accurate-gradient with $\kappa = \mu_\phi = 0$. As a consequence, if the sequence $\{\|\lambda^k\|\}$ is bounded and choose r to be a sufficiently large constant and α, β satisfy (3.19), under generalized Slater condition, the sequence $\{x^k\}$ generated by Algorithm 1.1 converges to the optimal solution of (1.1) with probability 1. We conclude this settings with the following theorem.

Theorem 4.2. *Suppose that generalized Slater condition holds for* (1.1)*, and* ϕ_r *is bounded from above on* $\mathbb{R}^p$*. Moreover, the random approximation models* $\{f^k, G^k\}$ *in* (4.9) *satisfy*

$$(1-\bar{\gamma})^2 \geq \max\{\alpha,\ \beta\},$$

where $\min\{\gamma_f(x), \gamma_g(x)\} \leq \bar{\gamma}$ *for any* $x \in X_0$ *and* $\phi_r^k \not\equiv -\infty$*. Let*

$$r \geq \bar{r} := 2\kappa_f,$$

and α, β *satisfies* (3.19)*, where* $\nu = 2\kappa_f/3$*. If the multiplier set* $\{\|\lambda^k\|\}$ *is bounded almost surely, any cluster point* $\bar{x}$ *of the sequence* $\{x^k\}$ *generated by Algorithm* 1.1 *is the optimal solution of* (1.1) *almost surely.*

Proof. Since functions$f$ and G can be exactly estimated by the random approximation models $\{f^k, G^k\}$ (4.9) with probability $(1-\bar{\gamma})^2$, it implies from (3.15) that

$$\phi_r(\lambda^k) - \phi_r(\lambda^{k+1}) \leq -\frac{r}{2}\|\delta^k\|^2 \leq -\kappa_f\|\delta^k\|^2$$

holds with probability $(1-\bar{\gamma})^2$. Then the conclusion can be proved by Theorems 3.2 and 3.1. The proof is complete. □

5. Numerical Experiments

We concentrate on numerical experiments to verify the performance of the stochastic augmented Lagrange method (SALM) for stochastic convex optimization problems with inequality constraints under various noisy situations discussed in the previous section. All numerical experiments throughout this section are performed using MATLAB R2019a on a laptop with Intel(R) Core(TM) i5-6200U 2.30 GHz and 8 GB memory.

5.1. Random models

In the following, we consider the minimization of a sum of problems of the form

$$\begin{aligned} \min \quad & f(x) = \sum_{i=1}^{m_f} f_i(x) \\ \text{s.t.} \quad & G(x) = \sum_{j=1}^{m_g} G_j(x) \leq 0, \end{aligned} \tag{5.1}$$

where $G(x) = (g_1(x), \cdots, g_p(x)) \in \mathbb{R}^p$. For each $i \in \{1, \ldots, m_f\}$ and $j \in \{1, \ldots, m_g\}$, f_i and G_j are both smooth and convex mappings. Three different forms of noise will be discussed in this section, namely multiplicative noise, additive noise and probability noise. The multiplicative noise is composed of two groups of random variables ξ_i and ξ_j, where $i \in \{1, \ldots, m_f\}, j \in \{1, \ldots, m_g\}$, that follow the uniform distribution on $[-\sigma_f, \sigma_f]$ and $[-\sigma_g, \sigma_g]$, respectively. The parameters $\sigma_f, \sigma_g \in (0, 1)$ are the main factors causing the instability of the functions f and G. The random model generated by the multiplicative noise is expressed as follows:

$$\begin{aligned} \min \quad & f(x, \xi) = \sum_{i=1}^{m_f} (1 - \xi_i) f_i(x) \\ \text{s.t.} \quad & G(x, \xi) = \sum_{j=1}^{m_g} (1 - \xi_j) G_j(x) \leq 0. \end{aligned} \tag{5.2}$$

Obviously, for each x, it has $\mathbb{E}_\xi[f(x, \xi)] = f(x)$ and $\mathbb{E}_\xi[G(x, \xi)] = G(x)$ which are different from the assumptions of Proposition 4.1. Although the convergence under these assumptions is difficult to prove by our theory, from the point of view of the numerical performance, as long as the appropriate sample size is chosen, the gap between the stochastic augmented Lagrange function generated in the random model and the "true" problem is sufficiently small. As a consequence, the stochastic augmented Lagrange method ensures that the sequence $\{x^k\}$ converges to the optimal solution almost surely.

For the second type of noise – additive noise, we are going to look at a more complex random form. We suppose that the objective function in the problem (5.1) can be represented as the sum of the squares of some functions, i.e.

$$f(x) = \sum_{i=1}^{m_f} f_i(x) = \sum_{i=1}^{m_f} \left(\bar{f}_i(x) \right)^2.$$

For example, (5.1) is a quadratic programming problem, and so on. In additive noise, for each $i \in \{1, \ldots, m_f\}$, $j \in \{1, \ldots, m_g\}$, we also generate random variables ξ_i and ξ_j from the uniform distribution on $[-\sigma_f, \sigma_f]$ and $[-\sigma_g, \sigma_g]$, respectively, where the parameters $\sigma_f, \sigma_g \in (0, 1)$. The additive random model is written in the following form:

$$\begin{aligned} \min \quad & f(x, \xi) = \sum_{i=1}^{m_f} (\bar{f}_i(x) + \xi_i)^2 \\ \text{s.t.} \quad & G(x, \xi) = \sum_{j=1}^{m_g} G_j(x) + \xi_j \leq 0. \end{aligned} \tag{5.3}$$

Different from the multiplicative noise, for each x, it has

$$\mathbb{E}_\xi[f(x,\xi)] = f(x) + \sum_{i=1}^{m_f} \mathbb{E}(\xi_i)^2.$$

Because of

$$\text{argmin}_x \mathbb{E}_\xi[f(x,\xi)] = \text{argmin}_x f(x),$$

hence, the constant bias has no impact on the optimization process and the algorithm still converges to the optimal solution.

The last type of noise, unlike the above two kinds of noise, is biased noise. In actual data statistics, there may be some abnormal or missing data. These data are sometimes eliminated during the calculation process. In another way, we can mark the missing data as a constant that is quite different from the actual statistical data. At this time, there is a large deviation in the computation of function values with a small probability in the optimization problem, that is, for each component in the sum of the objective function in (5.1), for some parameter ϵ, if $|f_i(x)| < \epsilon$, the value of $f_i(x)$ is computed as

$$f_i(x) = \begin{cases} f_i(x) & \text{with probability} \quad 1-P, \\ V & \text{with probability} \quad P, \end{cases} \tag{5.4}$$

where the parameter $P > 0$ is the probability of function computation failures and V is a very large constant. If $|f_i(x)| > \epsilon$, then the computation of $f_i(x)$ is deterministic and accurate. In addition, the constraint function $G(x)$ is considered to be accurately computation. Obviously, this noise is biased and the bias depends on x. It can be seen that the convergence of this random model depends on the number of iterations, the accuracy of the convergence chosen by the termination criterion, ϵ, P and the parameter r, but does not depend on V. Therefore, we might as well choose $V = -10^5$ in the following experiment.

5.2. Test problems

For the problem (5.1), we test two kinds of functions. One is the quadratic programming problem, i.e. for each $i \in \{1,\dots,m_f\}$, $f_i(x)$ is a convex quadratic function and for each $j \in \{1,\dots,m_g\}$, $G_j(x)$ is a linear mapping. It can be expressed as

$$f_i(x) = x^T Q_i x + c_i^T x + d_i, \quad G_j(x) = A_j x - b_j, \tag{5.5}$$

where $Q_i \in \mathbb{R}^{n\times n}, c_i \in \mathbb{R}^n, d_i \in \mathbb{R}, A_j \in \mathbb{R}^{p\times n}, b_j \in \mathbb{R}^p$ are randomly selected and the problem (5.1) is guaranteed to be convex and has the optimal solution x^*. Choose each Q_i to be a symmetric positive semi-definite matrix, so that Q_i can be decomposed into $Q_i = L_i^T L_i$, where L_i is an upper triangular matrix. The objective functions in (5.5) can be rewrite as

$$f_i(x) = (L_i x + \bar{c}_i)^T (L_i x + \bar{c}_i) + \bar{d}_i,$$

where

$$2\bar{c}_i^T L_i x = c_i^T x, \quad \bar{d}_i = d_i - \bar{c}_i^T \bar{c}_i.$$

Hence, the objective function of the quadratic programming problem can be represented as the sum of the squares of some functions. The other problem is selected as the polynomial programming problem, where $f_i(x)$ is a convex quartic function and $G_j(x)$ is still a linear mapping, i.e.

$$f_i(x) = \left(x^T Q_i x + c_i^T x + d_i\right)^2, \quad G_j(x) = A_j x - b_j. \tag{5.6}$$

SALM is used to solve all three random models in both the quadratic and polynomial programming problem with the dimension $n = 10, p = 5$ and the number of functions m_f=100, m_g=20, while the true problem is solved by the augmented Lagrange method (ALM).

5.3. Algorithms and numerical results

In this part, we illustrate the performance of Algorithm 1.1 for different convex programming problems with several random models and compare it with other stochastic convex programming algorithms. We list the algorithms present in the following numerical experiments.

SALM. Algorithm 1.1.

SALM-SAA. Combining the standard sample averaging approximation techniques with the stochastic augmented Lagrange method, the Algorithm is shown in Algorithm 5.1 for solving the random model generated by multiplicative noise and additive noise.

SPDO. Stochastic primal-dual optimization with multiple objectives in [15], the optimal primal and dual solutions are obtained by using the gradient descent method for the convex-concave optimization problem, where the objective function is the Lagrange function of random approximation models.

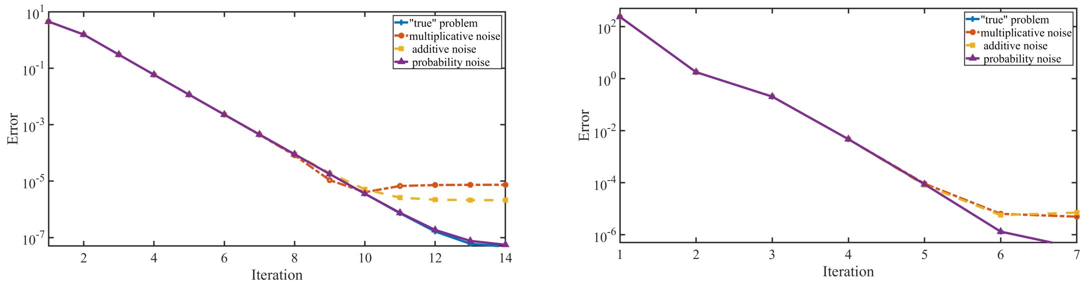
SPDA. Stochastic primal-dual algorithm in [15], the optimal primal and dual solutions are obtained by exactly solving the convex-concave optimization problem, where the objective function is the Lagrange function of random approximation models.

Algorithm 5.1: SALM-SAA.

```
Require: The parameter $r > 0$ and $\varepsilon > 0$, the sample size $N$, the initial multiplier
         $\lambda^0 \in \mathbb{R}^p$ and the initial point $x^0 \in \mathbb{R}^n$. Let $k = 0$.
1 for $k = 0, 1, \ldots$ do
2     If $x^k$ and $\lambda^k$ satisfies the termination criterion
          $\left\|\nabla_x L(x^k, \lambda^k) = \nabla_x f(x^k) + (\lambda^k)^T \nabla_x G(x^k)\right\| \leq \varepsilon,$   (5.7)
      then stop and return $x^k$.
3     Randomly choose sample sets $N_f = \{\xi^1, \cdots, \xi^N\}$, $N_g = \{\xi^1, \cdots, \xi^N\}$ with sizes $N$,
      then compute $f^k$ and $G^k$ as follows:
          $f^k(x) = \frac{1}{N}\sum_{i=1}^{N} f(x, \xi^i), \quad G^k(x) = \frac{1}{N}\sum_{i=1}^{N} G(x, \xi^i).$   (5.8)
4     Compute
          $x^{k+1} = \operatorname{argmin}\left\{\mathcal{L}_r^k(x, \lambda^k),\ x \in X_0\right\},$   (5.9)
          $\lambda^{k+1} = \left[\lambda^k + rG^k(x^{k+1})\right]_+.$   (5.10)
5     Let $k = k + 1$ and go to Step 1.
6 end
```

The convergence trend of SALM under the four models is shown in Fig. 5.1. For the same given initial multiplier λ^0 and initial point x^0, we use the nonlinear programming solver "fminunc" in MATLAB to solve the inner problem (1.6) in SALM and (5.7) in SALM-SAA and the parameters are chosen as $r = 10, \varepsilon = 10^{-7}$. The sample size in multiplicative noise and additive noise is selected as $N = 10^6$, the parameter $\sigma_f = \sigma_g = 0.01$ and parameters in probabilistic noise are chosen as $\epsilon = 0.3, P = 10^{-4}$. Since the probability that each component in the sum of the function fails to calculate is very small, in each iteration of SALM for solving the random model (5.4), with a high probability, the approximations satisfy $f^k = f, G^k = G$. This implies that in each iteration, SALM solves the deterministic problem (5.1) with a high probability. Therefore, the convergence curve of the true problem coincides with the random model (5.4), with a high probability, in both the quadratic and polynomial programming problem. However, for multiplicative noise and additive noise, even if a considerable sample size is selected, there is still a certain error between the approximation function f^k, G^k and f, G, so the accuracy of convergence is not as high as the true problem (5.1).

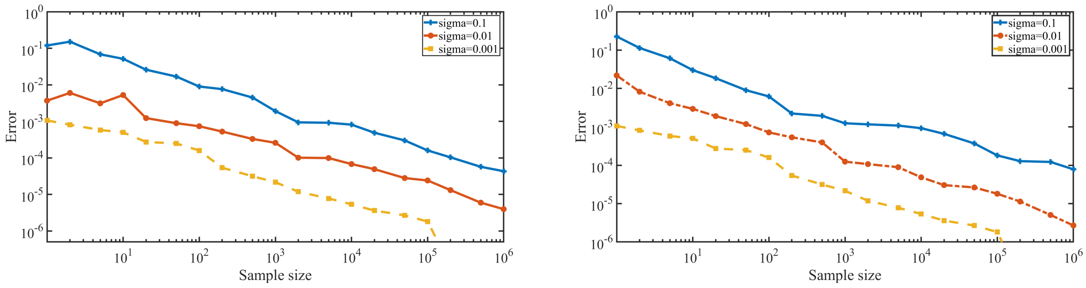
The random variable ξ and the sample size N have a huge impact on the convergence in SALM-SAA 5.1, which is illustrated in Figs. 5.2 and 5.3. We test SALM-SAA (Algorithm 5.1) for the stochastic quadratic programming model with multiplicative noise and additive noise. Choose the parameter $r = 10$ in the augmented Lagrange function and let $\sigma_f = \sigma_g = \sigma$. For the given uniform distribution on $[-\sigma, \sigma]$, as the sample size N increases, Algorithm 5.1 converges more accurately. Fig. 5.3 attempts to interpret the relationship of the variance of the random variable ξ and the sample size N. When the variance becomes larger, i.e. the



(a) quadratic programming

(b) polynomial programming

Fig. 5.1. The trend of the error by SALM under the true problem and three random models (5.1)-(5.4) for two convex problems.



(a) multiplicative noise

(b) additive noise

Fig. 5.2. The effect of the variance of the random variable ξ and the sample size on the accuracy of convergence under two random models for the quadratic programming problem.

parameters σ become larger, in order to achieve the same convergence accuracy 10^{-6} chosen in the termination criterion, SALM-SAA requires a larger sample size N.

In order to verify that SALM converges to the optimal solution with probability 1 when the random models are sufficiently accurate with high enough but fixed probability, Fig. 5.4 shows the relationship between the probability of function computation failures P and the probability of successful convergence of SALM for the random model (5.4). It is said that SALM converges successfully for an optimization problem, if SALM terminates within K iterations and satisfies the error of the gradient of Lagrange function $L(x, \lambda)$ within ε. In our experiment, choose the objective and constraint functions as (5.5) and the parameters $r = 10$ and $\epsilon = 0.3$. Let $K = 30$ and $\varepsilon = 10^{-7}$. In the left figure of Fig. 5.4, for a given P, SALM is used to solve the random model (5.4) repeatedly 50 times, and we record the number of the successful convergence S_k. The probability of the successful convergence of SALM is expressed as $P_S = S_k/50$. From the curve on the left in Fig. 5.4, it can be seen that when the probability of error P is reduced to 0.09, SALM guarantees to converge to the optimal solution with probability 1. Although SALM can be guaranteed to converge when the probability $P \leq 0.09$, the convergence rate depends on the value of P. From the curve on the right in Fig. 5.4, it can be seen that $P = 0.08$ and $P = 0.0001$ both can guarantee the successful convergence. However, $P = 0.08$ means that the probability of making mistakes is greater, so the convergence curve has large fluctuations and it converges much slower than $P = 0.0001$. On the other hand, when the probability of making a mistake $P = 0.5$, the algorithm cannot converge.

In Fig. 5.5, we discuss the influence of the parameter r in the augmented Lagrange function on SALM. From a theoretical point of view, it is not difficult to find that when r is larger,

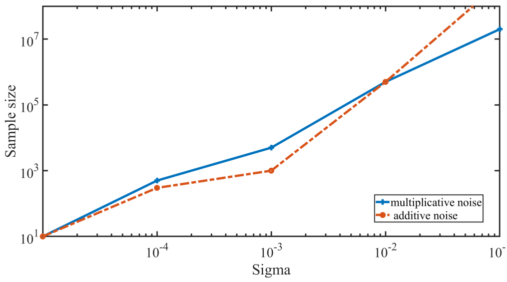


Fig. 5.3. The relationship between the variance of the random variable ξ and the sample size selected by Algorithm 1.1 for the quadratic programming problem with the same convergence accuracy 10^{-6}.

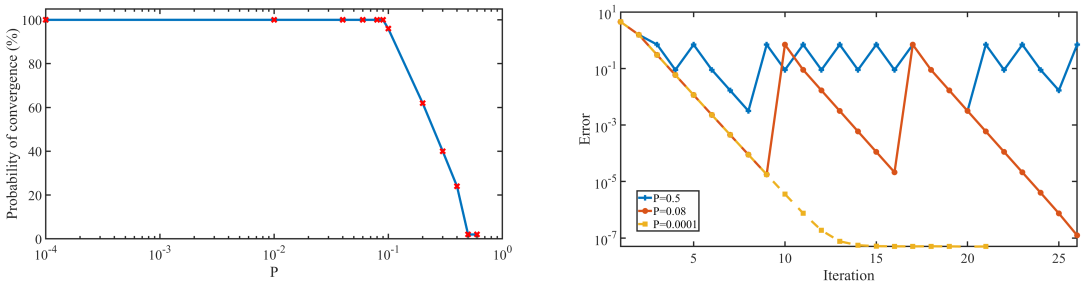
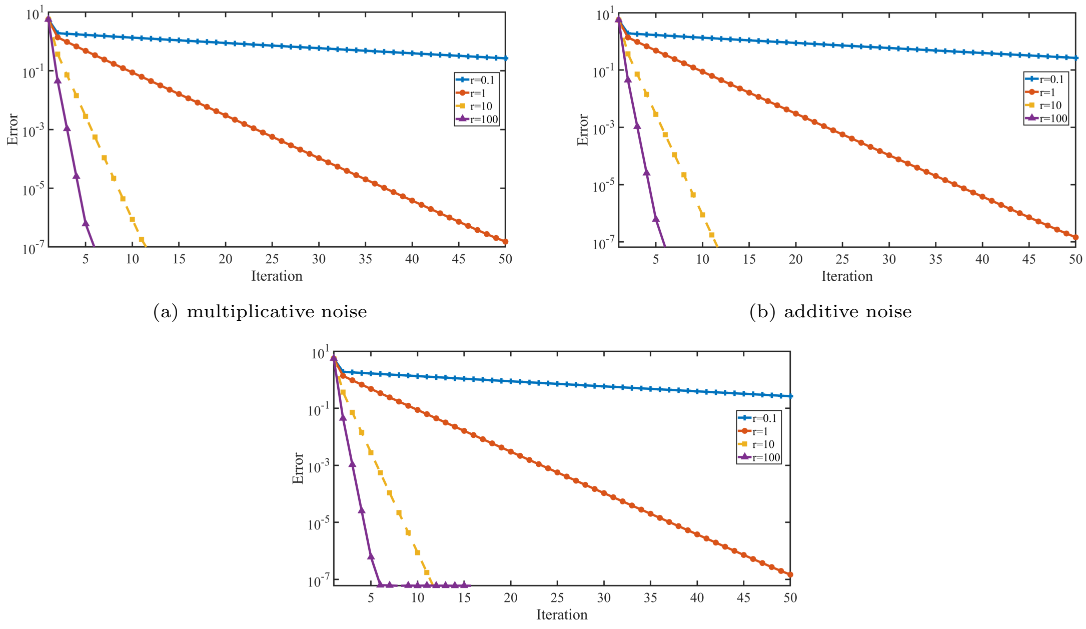


Fig. 5.4. The trend of the convergence probability where with probability P, the objective function in the quadratic programming problem is computed incorrectly (left) and the trend of the error at the three specific values of the probability P (right).

the convergence rate of SALM is faster. In the numerical simulation, we can easily see this phenomenon. Looking at Fig. 5.5, we test the effect of the value of r on the convergence rate under three random models. At this time, for the quadratic programming problem, to better illustrate the effect of r, we substitute $\|x^k - x^*\|$ for $\|\nabla_x L(x^k, \lambda^k)\|$ for the error and get the results in Fig. 5.5, where x^* is the optimal solution of the problem (5.1). Here, except for r, the parameters are chosen as the same as those used in the test in Fig. 5.1. Another fact we want to explain is that although the increase in r can make the algorithm converge faster, it will lead to a decrease in the accuracy of the convergence. When $\|\nabla_x L(x^k, \lambda^k)\|$ is used as the error, we can see from Fig. 5.6 that when $r = 100$, the error flattens out after it drops rapidly to 10^{-6}. However, when $r = 10$, the error can be reduced to 10^{-8}. The reason is that when the optimal solution x^* satisfies, for some $i \in \{1, \dots, p\}, g_i(x^*) = 0$, the accuracy of the machine can usually only reach $g_i(x^k) \approx 10^{-10} \neq 0$. This leads to an error in calculating the



(a) multiplicative noise

(b) additive noise

(c) probabilistic noise

Fig. 5.5. The trend of the error at four different values of r under three random models for the quadratic programming problem with respect to iteration number.

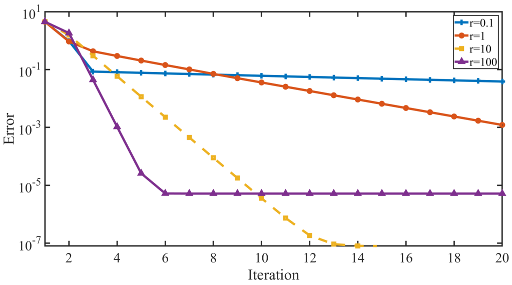
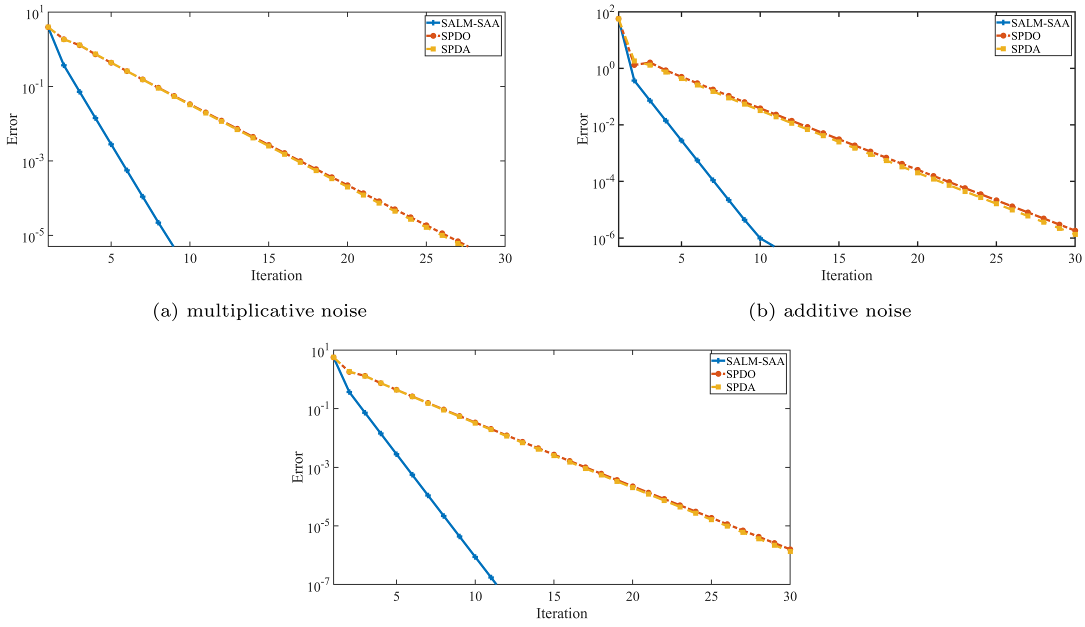


Fig. 5.6. The trend of the error at four different values of r for the stochastic quadratic programming problem with multiplicative noise with respect to iteration number.

multiplier λ in SALM-SAA (Algorithm 5.1), i.e. at this time, $rg_i(x^k) \neq 0$, when the parameter r is too large. Therefore, the appropriate r should be selected in the algorithm, neither too small leading to slow convergence rate, nor too large leading to low accuracy.

Finally, we show the comparison performances of SALM, SPDO and SPDA for the quadratic programming problem in Fig. 5.7. For the same given initial multiplier λ^0 and initial point x^0, we use the nonlinear programming solver `fminunc` in MATLAB to solve the inner problem in SALM and SPDA. The parameters in SALM are chosen as $r = 10$, $\varepsilon = 10^{-7}$ and the step size of SPDO is tuned for best performance. For the random models with multiplicative noise and additive noise, the sample size in SALM-SAA is selected as $N = 10^3$ and the parameter $\sigma_f = \sigma_g = 0.01$. For optimization with probabilistic noise, parameters in SALM are chosen as $\epsilon = 0.3$, $P = 10^{-4}$. From the numerical results, SALM converges more rapidly than SPDO and SPDA in general for all random models.



(a) multiplicative noise

(b) additive noise

(c) probabilistic noise

Fig. 5.7. Comparison of three algorithms for solving the quadratic programming problem with respect to iteration.

6. Conclusion

In this paper, a stochastic augmented Lagrange method is constructed basing on a class of random approximation models for stochastic convex optimization problems with inequality constraints. The convergence of the stochastic augmented Lagrange method depends on how the estimates of the approximation model being sufficiently close to the true problem with high but fixed probability. Without assuming the expectation and variance of the models and estimates, if the coefficient r in the augmented Lagrange function is selected as an appropriate constant, and the models and estimates are sufficiently accurate with high enough but fixed probability, $\{x^k\}$ can converge to the optimal solution of the true problem almost surely. In addition, some special approximation models are discussed under biased or unbiased noise assumptions. From numerical experiments, with high enough but fixed probability (which is guaranteed by the sample sizes being large enough or the probability of making a mistake being small enough), the distance between the stochastic augmented Lagrange function generated in the random model and the true problem is sufficiently close, and the appropriate parameter r is selected, the stochastic augmented Lagrange method can accurately and quickly converge to the optimal solution.

Acknowledgments. This research was supported by the National Key R&D Program of China (Project No. 2022YFA1004000) and by the Natural Science Foundation of China (Project No. 12371298).

References

[1] Z. Akhtar, A.S. Bedi, and K. Rajawat, Conservative stochastic optimization with expectation constraints, *IEEE Trans. Signal Process.*, **69** (2021), 3190–3205.

[2] P. Apkarian, D. Noll, and H.D. Tuan, Fixed-order H_∞ control design via a partially augmented Lagrangian method, *Int. J. Robust Nonlinear Control.*, **13**:12 (2003), 1137–1148.

[3] E. Bergou, S. Gratton, and L.N. Vicente, Levenberg-Marquardt methods based on probabilistic gradient models and inexact subproblem solution with application to data assimilation, *SIAM-ASA J. Uncertain.*, **4**:1 (2016), 924–951.

[4] D.P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, 1982.

[5] R. Bollapragada, R.H. Byrd, and J. Nocedal, Exact and inexact subsampled Newton methods for optimization, *IMA J. Numer. Anal.*, **39**:2 (2019), 545–548.

[6] J.F. Bonnans and A. Shapiro, *Perturbation Analysis of Optimization Problems*, Springer, 2000.

[7] R.H. Byrd, G.M. Chin, W. Neveitt, and J. Nocedal, On the use of stochastic Hessian information in optimization methods for machine learning, *SIAM J. Optim.*, **21**:3 (2011), 977–995.

[8] R.H. Byrd, G.M. Chin, J. Nocedal, and Y. Wu, Sample size selection in optimization methods for machine learning, *Math. Program.*, **134**:1 (2012), 127–155.

[9] N. Chatzipanagiotis, D. Dentcheva, and M.M. Zavlanos, An augmented Lagrangian method for distributed optimization, *Math. Program.*, **152**:1-2 (2015), 405–434.

[10] R. Chen, M. Menickelly, and K. Scheinberg, Stochastic optimization using a trust-region method and random models, *Math. Program.*, **169**:2 (2018), 447–487.

[11] 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, *Math. Program.*, **178**:1 (2019), 381–415.

[12] R. Johnson and T. Zhang, Accelerating stochastic gradient descent using predictive variance reduction, in: *Proceedings of the 26th International Conference on Neural Information Processing Systems*, **1**:3 (2013), 315–323.

[13] J. Larson and S.C. Billups, Stochastic derivative-free optimization using a trust region framework, *Comput. Optim. Appl.*, **64**:3 (2016), 1–27.

[14] M. Mahdavi, R. Jin, and T. Yang, Trading regret for efficiency: Online convex optimization with long term constraints, *J. Mach. Learn. Res.*, **13**:3 (2011), 2503–2528.

[15] M. Mahdavi, T. Yang, and R. Jin, Stochastic convex optimization with multiple objectives, in: *Advances of Neural Information Processing Systems*, (2013), 1115–1123.

[16] R.E. Miller, *Optimization: Foundations and Applications*, John Wiley and Sons, 2011.

[17] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, Robust stochastic approximation approach to stochastic programming, *SIAM J. Optim.*, **19**:4 (2009), 1574–1609.

[18] R. Pasupathy and S. Ghosh, Simulation optimization: A concise overview and implementation guide, in: *INFORMS TutORials in Operations Research*, INFORMS, (2013), 122–150.

[19] P. Rigollet and X. Tong, Neyman-pearson classification, convexity and stochastic constraints, *J. Mach. Learn. Res.*, **12** (2011), 2831–2855.

[20] R.T. Rockafellar, *Convex Analysis*, Princeton University Press, 1970.

[21] R.T. Rockafellar, A dual approach to solving nonlinear programming problems by unconstrained optimization, *Math. Program.*, **5**:1 (1973), 354–373.

[22] R.T. Rockafellar, Monotone operators and the proximal point algorithm, *SIAM J. Control Optim.*, **14** (1976), 877–898.

[23] R.T. Rockafellar, Augmented Lagrangians and applications of the proximal point algorithm in convex programming, *Math. Oper. Res.*, **1** (1976), 97–116.

[24] R.T. Rockafellar and R.J.B. Wets, *Variational Analysis*, Springer, Vol. 317, 2009.

[25] M. Schmidt, N. Le Roux, and F. Bach, Minimizing finite sums with the stochastic average gradient, *Math. Program.*, **162**:1-2 (2017), 1–30.

[26] A. Shapiro, Statistical inference of semidefinite programming, *Math. Program.*, **174**:1 (2009), 77–97.

[27] D. Sun, J. Sun, and L. Zhang, The rate of convergence of the augmented Lagrangian method for nonlinear semidefinite programming, *Math. Program.*, **114**:2 (2008), 349–391.

[28] W. Wang and S. Ahmed, Sample average approximation of expected value constrained stochastic programs,*Oper. Res. Lett.*, **36**:5 (2008), 515–519.

[29] Y. Xu and W. Yin, Block stochastic gradient iteration for convex and nonconvex optimization, *SIAM J. Optim.*, **25**:3 (2015), 1686–1716.

[30] L. Zhang, Y. Zhang, X. Xiao, and J. Wu, Stochastic approximation proximal method of multipliers for convex stochastic programming, *Math. Oper. Res.*, **48**:1 (2022), 177–193.

[31] Z. Zhou, P. Mertikopoulos, N. Bambos, S. Boyd, and P. Glynn, On the convergence of mirror descent beyond stochastic convex programming, *SIAM J. Optim.*, **30**:1 (2020), 687–716.