Computational Optimization and Applications
https://doi.org/10.1007/s10589-023-00521-z

Stochastic inexact augmented Lagrangian method for nonconvex expectation constrained optimization

Zichong Li[1] · Pin-Yu Chen[2] · Sijia Liu[3] · Songtao Lu[2] · Yangyang Xu[1]

Received: 20 December 2022 / Accepted: 18 August 2023

Abstract
Many real-world problems not only have complicated nonconvex functional constraints but also use a large number of data points. This motivates the design of efficient stochastic methods on finite-sum or expectation constrained problems. In this paper, we design and analyze stochastic inexact augmented Lagrangian methods (Stoc-iALM) to solve problems involving a nonconvex composite (i.e. smooth + nonsmooth) objective and nonconvex smooth functional constraints. We adopt the standard iALM framework and design a subroutine by using the momentum-based variance-reduced proximal stochastic gradient method (PStorm) and a postprocessing step. Under certain regularity conditions (assumed also in existing works), to reach an ε-KKT point in expectation, we establish an oracle complexity result of $O(\varepsilon^{-5})$, which is better than the best-known $O(\varepsilon^{-6})$ result. Numerical experiments on the fairness constrained problem and the Neyman–Pearson classification problem with real data demonstrate that our proposed method outperforms an existing method with the previously best-known complexity result.

✉ Yangyang Xu
xuy21@rpi.edu

Zichong Li
liz19@rpi.edu

Pin-Yu Chen
pin-yu.chen@ibm.com

Sijia Liu
liusiji5@msu.edu

Songtao Lu
songtao@ibm.com

[1] Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, NY 12180, USA

[2] IBM Research, Thomas J. Watson Research Center, Yorktown Heights, NY 10598, USA

[3] Department of Computer Science and Engineering, Michigan State University, East Lansing, MI 48824, USA

Published online: 07 September 2023

Keywords First-order methods · Expectation constraint · Augmented Lagrangian · Nonconvex optimization · Momentum acceleration

1 Introduction

In the big-data era, many real-world applications are dealing with an extremely large amount of data. Many such applications involve nonconvex functional constraints. To compute solutions of these problems, using all data for each update (e.g., in a deterministic method) is prohibitively expensive. This motivates us to design stochastic methods to efficiently compute the solutions.

In this paper, we consider the nonconvex expectation constrained problem:

$$\begin{aligned} &f_0^* := \min_{\mathbf{x}\in\mathbb{R}^d} \big\{ f_0(\mathbf{x}) := g(\mathbf{x}) + h(\mathbf{x}),\ \text{s.t. } \mathbf{c}(\mathbf{x}) = \mathbf{0} \big\}, \\ &\text{with } \ g(\mathbf{x}) = \mathbb{E}_\xi[G_0(\mathbf{x};\xi)], \quad \mathbf{c}(\mathbf{x}) = \mathbb{E}_\xi[\mathbf{C}(\mathbf{x};\xi)] \in \mathbb{R}^m, \end{aligned} \tag{1}$$

where h is closed convex but possibly nonsmooth, and $\mathbb{E}_\xi$ denotes the expectation taken over the random variable ξ. Notice that it does not lose generality to use the same random variable ξ in the objective and constraints, because if they depend on two different random variables, we can represent ξ as the stack of the two random variables. We assume that $g(\cdot)$ and $\mathbf{c}(\cdot)$ are smooth (i.e., the gradient of g and the Jacobian matrix of $\mathbf{c}$ are Lipschitz continuous) but possibly nonconvex. When ξ follows the uniform distribution on $\{1, 2, \ldots, N\}$, the problem (1) reduces to a finite-sum structured problem:

$$\begin{aligned} &f_0^* := \min_{\mathbf{x}\in\mathbb{R}^d} \big\{ f_0(\mathbf{x}) := g(\mathbf{x}) + h(\mathbf{x}),\ \text{s.t. } \mathbf{c}(\mathbf{x}) = \mathbf{0} \big\}, \\ &\text{with } \ g(\mathbf{x}) = \frac{1}{N}\sum_{\xi=1}^{N} G_0(\mathbf{x};\xi), \quad \mathbf{c}(\mathbf{x}) = \frac{1}{N}\sum_{\xi=1}^{N} \mathbf{C}(\mathbf{x};\xi) \in \mathbb{R}^m, \end{aligned} \tag{2}$$

which arises from applications involving a large amount of pre-collected data.

Though only equality constraints are included, the formulation (1) is general enough. As shown in [15], an inequality constraint $t(\mathbf{x}) \le 0$ can be equivalently formulated as an equality constraint $t(\mathbf{x}) + s = 0$ by enforcing the nonnegativity of s, and the Karush–Kuhn–Tucker (KKT) conditions of the reformulation are equivalent to those of the original one. Also, a simple convex constraint set $\mathcal{X}$ can be included in (1) by setting (part of) h to the indicator function $\mathbf{1}_{\mathcal{X}}(\mathbf{x}) = 0$ if $\mathbf{x} \in \mathcal{X}$ and $+\infty$ otherwise. Many applications can be formulated to (1), such as Neyman-Pearson classification [26, 29] and the fairness constrained problem [21].

Due to the presence of nonconvexity and stochasticity in both objective and constraints, solving (1) is very challenging. Only a few works (e.g., [2, 21]) have proposed and analyzed methods to solve such a problem. However, no existing methods have fully exploited the structure of (1). We will present a stochastic method for (1) under the general expectation setting, and establish its oracle complexity result, where the oracle can return the function value and gradient of G_0 and $\mathbf{C}$ at any point $\mathbf{x}$ and a sample of

ξ. We follow the iALM framework and adopt the momentum-based variance-reduced proximal stochastic gradient method (PStorm) [39] to design a subroutine.

1.1 Contributions

Our contributions are two-fold. First, we propose novel stochastic gradient-type methods, based on the framework of the inexact augmented Lagrangian method (Stoc-iALM), for solving nonconvex composite optimization problems with nonlinear nonconvex (but smooth) expectation constraints, in the form of (1). By exploiting the so-called mean-squared smoothness structure, we apply PStorm [39] together with a proposed postprocessing step to design a subroutine within the framework of Stoc-iALM. The subroutine design is crucial to yield our complexity result that is better than existing best-known results and for good numerical performance, as its complexity has low-order dependence not only on a target error tolerance but also on other quantities such as the smoothness constant, variance bound, and initial objective gap.

Second, we conduct complexity analysis on the proposed Stoc-iALM with the designed subroutine. Under a regularity condition (that was also assumed in many existing works [15, 16, 18, 31]), we obtain an $O(\varepsilon^{-5})$ oracle complexity result for the expectation-constrained problem (1). Our $O(\varepsilon^{-5})$ result yields a substantial improvement over the best-known $\tilde{O}(\varepsilon^{-6})$ and $O(\varepsilon^{-6})$ complexity results[1] of the proximal-point methods in [21] and [2], which iteratively perturb both the objective and constraints and solve a perturbed convex constrained subproblem. For two tested problems, we verify the assumed regularity condition numerically and prove it when a ball constraint is imposed and the involved data are preprocessed appropriately.

1.2 Related works

In this subsection, we discuss related works on the inexact augmented Lagrangian method (iALM) and other first-order methods (FOMs) on functional constrained optimization.

The iALM has been popularly used for solving constrained problems. It alternatingly updates the primal variable by approximately minimizing the augmented Lagrangian function and the Lagrangian multiplier (also called dual variable) by dual gradient ascent [9, 30]. For deterministic convex linear and/or nonlinear constrained problems, the iALM-based FOM in [12, 17] and the proximal-iALM-based one in [14] obtain an ε-KKT point with $O\left(\varepsilon^{-1}\log\frac{1}{\varepsilon}\right)$ gradient evaluations, and the AL-based FOMs in [14, 25, 27, 37, 38] obtain an ε-optimal solution with $O(\varepsilon^{-1})$ gradient evaluations. For strongly-convex problems, the results are reduced to $O\left(\varepsilon^{-0.5}\log\frac{1}{\varepsilon}\right)$ and $O(\varepsilon^{-0.5})$ respectively, e.g., in [14, 17, 24, 25, 38]. For deterministic nonconvex problems with nonlinear convex constraints, when Slater's condition holds, $\tilde{O}(\varepsilon^{-\frac{5}{2}})$ complexity results are obtained by the AL or penalty based FOMs in [17, 18] and the proximal ALM-based FOM in [23]. If the constraints are polyhedral and the objective is smooth, the complexity can be reduced to $O(\varepsilon^{-2})$ with a hidden constant dependent

[1] In this paper, we use $\tilde{O}$ to suppress all logarithmic terms of ε from the big-O notation.

on the so-called Hoffman's bound of the polyhedral set [42]. Different from Slater's condition, a regularity condition is assumed in [14], which obtains an $\tilde{O}(\varepsilon^{-\frac{5}{2}})$ result by an iALM-based FOM. The regularity condition is used to guarantee near feasibility from near stationarity of the AL function. Assuming a similar regularity condition, [14] and [18] both achieve $\tilde{O}(\varepsilon^{-3})$ results for deterministic problems with nonconvex constraints, by an iALM based FOM and a proximal-point penalty based FOM respectively. A single-loop FOM is given in [19] for solving problems with nonconvex constraints.

There are many papers studying FOMs on convex stochastic constrained problems (e.g., [13, 36, 40]). Also, a few papers (e.g., [11, 32, 35]) have studied FOMs for nonconvex optimization with stochastic objective but deterministic constraints, either based on an exact-penalty framework or ALM. However, few papers have studied FOMs for the nonconvex expectation constrained problems. On solving inequality expectation constrained nonconvex optimization, both [21] and [2] design stochastic first-order methods in the framework of the proximal-point (PP) method. They achieve $\tilde{O}(\varepsilon^{-6})$ and $O(\varepsilon^{-6})$ complexity results respectively, which are higher than our $O(\varepsilon^{-5})$ result. Both PP-based methods in [2, 21] iteratively perturb the nonconvex objective and constraint functions to be strongly convex and inexactly solve the constrained convex subproblems. To achieve their results, the PP-based method in [21] uses the online stochastic subgradient subroutine in [41], while the one in [2] designs a constraint extrapolation (ConEx) subroutine. Note that nonconvex structures that we assume are different from those in [2] and [21]. While we assume a nonconvex composite objective and smooth constraints, the method in [2] applies to nonconvex problems where both the objective and constraint functions can be nonconvex composite, and [21] only assumes weak convexity[2] on the objective and constraint functions. However, even with the nonconvex structures that we assume, the methods in [21] and [2] can still only achieve the $\tilde{O}(\varepsilon^{-6})$ and $O(\varepsilon^{-6})$ complexity results, as they do not exploit the smoothness structure in their subroutines.

Stochastic FOMs have also been proposed for minimax problems (e.g., [10, 20, 33]). The work [33] gives a hybrid variance-reduced stochastic gradient method for nonconvex-linear minimax problems with a compact domain of dual variables and establishes an $O(\varepsilon^{-5})$ complexity result to find an ε-stationary point. Although a nonlinear-constrained problem can be formulated as a nonconvex-linear minimax problem by the ordinary Lagrangian function, KKT conditions of the former are stronger than stationarity conditions of the latter that assumes a compact dual domain. This is due to the fact that the stationary point of a nonconvex-concave minimax problem with a compact dual domain may not be primal feasible. Both of [10, 20] assume strong concavity on the dual side. Let κ be the condition number of the dual part. The method in [20] needs $O(\kappa^3\varepsilon^{-3})$ sample complexity to produce an ε-stationary solution, while the complexity result in [10] is $\tilde{O}(\kappa^{\frac{9}{2}}\varepsilon^{-3})$. In order to obtain an ε-KKT point of the problem (1) that we consider, under the regularity condition in Assumption 3 below, we can apply the methods in [10, 20] to a penalized problem $\min_{\mathbf{x}} \left\{ f_0(\mathbf{x}) + \frac{\rho}{2}\|\mathbf{c}(\mathbf{x})\|^2 \right\}$ with $\rho = \Theta(\varepsilon^{-1})$, which is equivalent to the nonconvex strongly-concave minimax problem $\min_{\mathbf{x}} \max_{\mathbf{y}} \left\{ f_0(\mathbf{x}) + \mathbf{y}^\top \mathbf{c}(\mathbf{x}) - \frac{1}{2\rho}\|\mathbf{y}\|^2 \right\}$. The

[2] A function f is ρ-weakly convex for some $\rho > 0$, if $f(\cdot) + \frac{\rho}{2}\|\cdot\|^2$ is convex.

resulting complexity results will be $O(\varepsilon^{-6})$ by the method in [20] and $O(\varepsilon^{-\frac{15}{2}})$ by the method in [10], as the condition number of the equivalent minimax problem is $\Theta(\varepsilon^{-1})$.

1.3 Notations

We use $\|\cdot\|$ for the Euclidean norm of a vector and the spectral norm of a matrix. The notation $[n]$ denotes the set $\{1, \ldots, n\}$. For any $a \in \mathbb{R}$, $[a]_{1+} := \max\{a, 1\}$. The natural logarithmic function is $\ln(\cdot)$, and $e = 2.71828...$ represents its base. We denote $J_{\mathbf{c}}(\mathbf{x})$ as the Jacobian matrix of $\mathbf{c}$ at $\mathbf{x}$ and $J_{\mathbf{C}}(\mathbf{x}; \xi)$ the Jacobian matrix of $\mathbf{C}(\cdot; \xi)$ at $\mathbf{x}$. The distance between a vector $\mathbf{x}$ and a set $\mathcal{X}$ is denoted as $\mathrm{dist}(\mathbf{x}, \mathcal{X}) = \min_{\mathbf{y}\in\mathcal{X}} \|\mathbf{x}-\mathbf{y}\|$. The proximal operator of a convex function r is defined as $\mathbf{prox}_r(\mathbf{x}) := \arg\min_{\mathbf{u}}\{r(\mathbf{u}) + \frac{1}{2}\|\mathbf{u}-\mathbf{x}\|^2\}$. $\mathbb{E}_{\xi_1,\xi_2}$ takes expectation about ξ_1 and ξ_2, and we always assume that ξ_1 and ξ_2 are independent and follow the same distribution as ξ in (1). We use ∂f to denote the subdifferetial of a function f. The augmented Lagrangian (AL) function of (1) is

$$\mathcal{L}_\beta(\mathbf{x}, \mathbf{y}) = f_0(\mathbf{x}) + \mathbf{y}^\top \mathbf{c}(\mathbf{x}) + \tfrac{\beta}{2}\|\mathbf{c}(\mathbf{x})\|^2, \tag{3}$$

where $\beta > 0$ is the penalty parameter, and $\mathbf{y} \in \mathbb{R}^m$ is the multiplier or the dual variable.

Definition 1 (ε-KKT point in expectation) Given $\varepsilon \geq 0$, a point $\mathbf{x} \in \mathbb{R}^d$ is called an ε-KKT point in expectation to (1) if there is a vector $\mathbf{y} \in \mathbb{R}^m$ such that

$$\mathbb{E}\big[\|\mathbf{c}(\mathbf{x})\|^2\big] \leq \varepsilon^2, \quad \mathbb{E}\left[\mathrm{dist}\left(\mathbf{0}, \partial f_0(\mathbf{x}) + J_{\mathbf{c}}^\top(\mathbf{x})\,\mathbf{y}\right)^2\right] \leq \varepsilon^2.$$

2 Stochastic iALM and its outer iteration complexity

To efficiently find a near KKT-point of (1), we design a stochastic gradient-type method based on the framework of the stochastic inexact augmented Lagrangian method (Stoc-iALM), which is given in Algorithm 1. Because of nonconvexity, we can only produce a near-stationary point of each subproblem, as required in (4). Though the condition in (4) is not checkable (due to taking expectation), it can be guaranteed from the convergence rate result of the subroutine that we will give in Sect. 3. The update to the multiplier is inspired by [15, 31] and adapts to the estimated primal infeasibility. With an appropriate choice of γ_k, we can ensure $\frac{\|\mathbf{y}^k\|}{\beta_k} \to 0$, which is crucial in our analysis. The iALM framework that we adopt here is similar to those in [15, 31]. The main difference is the choice of subroutine. Our method only uses stochastic gradient/value information of the objective and constraint functions in the updates of both primal and dual variables, while the methods in [15, 31] need exact gradients/function values.

Without specifying a subroutine to obtain $\mathbf{x}^{k+1}$, we first establish the outer iteration complexity result of Algorithm 1, by following the analysis in [15, 18]. Throughout this paper, we make the following assumptions about (1).

Algorithm 1: Stochastic inexact augmented Lagrangian method (Stoc-iALM) for solving (1)

Initialization: given $\varepsilon > 0$, set $\mathbf{y}^0 = \mathbf{0}$ and choose $\mathbf{x}^0 \in \mathrm{dom}(f_0)$, $\beta_0 > 0$, $\sigma > 1$, and an integer sequence $\{M_k\}$

for $k = 0, 1, \ldots,$ **do**

Let $\beta_k = \beta_0 \sigma^k$.

Obtain $\mathbf{x}^{k+1}$ (by a subroutine) satisfying

$$\mathbb{E}\left[\mathrm{dist}(\mathbf{0}, \partial_x \mathcal{L}_{\beta_k}(\mathbf{x}^{k+1}, \mathbf{y}^k))^2 \,|\, \mathbf{y}^k\right] \le \varepsilon^2. \tag{4}$$

Obtain i.i.d samples $\{\xi_i^k\}_{i=1}^{M_k}$ and set $\tilde{\mathbf{c}}(\mathbf{x}^{k+1}) = \frac{1}{M_k}\sum_{i=1}^{M_k} \mathbf{C}(\mathbf{x}^{k+1}, \xi_i^k)$.

Update $\mathbf{y}$ by

$$\mathbf{y}^{k+1} = \mathbf{y}^k + \min\left\{\beta_k, \frac{\gamma_k}{\|\tilde{\mathbf{c}}(\mathbf{x}^{k+1})\|}\right\}\tilde{\mathbf{c}}(\mathbf{x}^{k+1}). \tag{5}$$

Assumption 1 (stochastic first-order oracle) For the problem (1), a stochastic first-order oracle can be accessed. At any $\mathbf{x} \in \mathrm{dom}(h)$, the oracle can obtain a sample ξ and return $(\nabla G_0(\mathbf{x}, \xi), \mathbf{C}(\mathbf{x}, \xi), J_{\mathbf{C}}(\mathbf{x}, \xi))$.

Remark 1 The overall complexity result of our algorithm will be measured by the total number of stochastic first-order oracles that are called. Though the oracle can return a tuple $(\nabla G_0(\mathbf{x}, \xi), \mathbf{C}(\mathbf{x}, \xi), J_{\mathbf{C}}(\mathbf{x}, \xi))$, our algorithm may only use part of it during one update. However, even if part of an oracle is used, one oracle will be counted in measuring the complexity result.

Assumption 2 (structured bounded domain) The domain of h, denoted as $\mathcal{X} := \mathrm{dom}(h)$, is compact. Moreover, for some $M > 0$, it holds that $\partial h(\mathbf{x}) \subseteq \mathcal{N}_{\mathcal{X}}(\mathbf{x}) + \mathcal{B}_M, \forall \mathbf{x} \in \mathcal{X}$, where $\mathcal{N}_{\mathcal{X}}(\mathbf{x})$ denotes the normal cone of $\mathcal{X}$ at $\mathbf{x}$, and $\mathcal{B}_M$ denotes a closed ball of radius M centering at the origin.

Remark 2 Assumption 2 holds for rather general choices of $h(\cdot)$. For example, it holds for any $h(\cdot) := r(\cdot) + \mathbf{1}_{\mathcal{X}}(\cdot)$ as long as ∂r is bounded everywhere (e.g., the ℓ_p-norm for $p \ge 1$), where $\mathbf{1}_{\mathcal{X}}$ denotes the indicator function on $\mathcal{X}$. Under Assumption 2, there must exist finite constants B_0 and B_c such that

$$B_0 \ge \max_{\mathbf{x}\in\mathrm{dom}(h)} \max\left\{|f_0(\mathbf{x})|, \|\nabla g(\mathbf{x})\|\right\}, \tag{6a}$$

$$B_c \ge \max_{\mathbf{x}\in\mathrm{dom}(h)} \|J_{\mathbf{c}}(\mathbf{x})\|. \tag{6b}$$

Due to nonconvexity of the constraints in (1), one may not even find a near-feasible point in polynomial time. Therefore, following [15, 16, 18, 31], we assume a regularity condition on the constraints in (1), which ensures that a near-stationary point of the AL function is near feasible to (1), if the penalty parameter is big. Note that knowledge of v below is not required in Algorithm 1.

Assumption 3 (regularity condition) There is a constant $v > 0$ such that for any $\mathbf{x} \in \mathcal{X} = \mathrm{dom}(h)$,

$$v\|\mathbf{c}(\mathbf{x})\| \leq \mathrm{dist}\left(-J_{\mathbf{c}}(\mathbf{x})^{\top}\mathbf{c}(\mathbf{x}), \mathcal{N}_{\mathcal{X}}(\mathbf{x})\right). \tag{7}$$

Remark 3 Here, we give a couple of remarks about the regularity condition. First, this regularity condition has been proven for many applications. For example, [15] shows that it holds for all affine-equality constrained problems with possibly additional polyhedral or ball constraint sets. Other examples are given in [18, 31]. In the appendix, we show that the regularity condition can also hold for certain instances of the two problems that we test in Sect. 4. Second, to find a near KKT point of a nonconvex expectation constrained problem, the two existing works [22] and [2] also need a certain regularity condition. Different from Assumption 3, a uniform Slater's condition is assumed in [22], and a strong MFCQ condition is assumed in [2]. Those conditions are neither strictly stronger nor strictly weaker than Assumption 3, as shown in [18].

The next lemma will be used to upper bound $\frac{\|\mathbf{y}^k\|}{\beta_k}$.

Lemma 1 *For any constants $\alpha > 1$ and $\sigma > 1$, if $\alpha \geq \frac{8}{\ln\sigma}$ and $\ln\sigma \geq \frac{8}{e^4}$, then it holds $\alpha \geq \log_\sigma \alpha^2$. In addition, for any $x \geq \log_\sigma \alpha^2$, it holds $\frac{\sigma^x}{x} \geq \alpha$.*

Proof Define $\phi(\alpha) = \alpha - \frac{2\ln\alpha}{\ln\sigma}$. Then $\phi'(\alpha) = 1 - \frac{2}{\alpha\ln\sigma} > 0, \forall\, \alpha > \frac{2}{\ln\sigma}$. Hence, $\phi(\cdot)$ is increasing on $(\frac{2}{\ln\sigma}, \infty)$. In addition, the condition $\ln\sigma \geq \frac{8}{e^4}$ implies $\phi(\frac{8}{\ln\sigma}) \geq 0$. Thus, for $\alpha \geq \frac{8}{\ln\sigma}$, it holds $\phi(\alpha) \geq 0$ that is equivalent to $\alpha \geq \log_\sigma \alpha^2$.

Now define $\psi(x) = \sigma^x - \alpha x$. Then $\psi'(x) = \sigma^x \cdot \ln\sigma - \alpha$, and thus for any $x \geq \log_\sigma \alpha^2$, we have $\psi'(x) \geq \psi'(\log_\sigma \alpha^2) = \alpha^2 \ln\sigma - \alpha \geq 7\alpha > 0$, where we have used $\alpha \geq \frac{8}{\ln\sigma}$. Hence, $\psi(\cdot)$ is increasing on $[\log_\sigma \alpha^2, \infty)$, and for any $x \geq \log_\sigma \alpha^2$, it holds $\psi(x) \geq \psi(\log_\sigma \alpha^2) = \alpha^2 - \alpha \log_\sigma \alpha^2 \geq 0$. This completes the proof. □

In Lemma 1, the requirement of $\ln\sigma \geq \frac{8}{e^4}$ (that holds if $\sigma \geq 1.158$) is just for ease of the analysis of our algorithm. The complexity results that we establish later will hold for any $\sigma > 1$ but can have an additional constant factor if the condition is not satisfied. The theorem below gives the outer iteration number of Algorithm 1 to produce an ε-KKT point in expectation of (1).

Theorem 1 (Outer iteration complexity of Stoc-iALM) *In (5), set $\gamma_k = \gamma_0, \forall\, k \geq 0$ for some $\gamma_0 > 0$ such that $\frac{\sqrt{8}B_c\gamma_0}{\beta_0 v\varepsilon} \geq \frac{8}{\ln\sigma}$. Then under Assumptions 2 and 3, Algorithm 1 needs at most K outer iterations to find an ε-KKT point in expectation of (1), where*

$$K = \max\left\{\left\lceil \log_\sigma \frac{\sqrt{8}\sqrt{\varepsilon^2 + B_0^2 + M^2}}{\beta_0 v\varepsilon}\right\rceil, \left\lceil 2\log_\sigma \frac{\sqrt{8}B_c\gamma_0}{\beta_0 v\varepsilon}\right\rceil\right\} + 1. \tag{8}$$

Proof First, by $\mathbf{y}^0 = \mathbf{0}$, the $\mathbf{y}$-update in (5), and the choice of γ_k, we have from the triangle inequality that for any $k \geq 0$,

$$\|\mathbf{y}^k\| \leq \sum_{t=0}^{k-1} \gamma_t = k\gamma_0, \tag{9}$$

where by the convention we define $\sum_{t=0}^{k-1}\gamma_t = 0$ if $k = 0$.

Second, from (7), we have

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{c}(\mathbf{x}^k)\|^2\right] &\le \frac{1}{v^2}\mathbb{E}\left[\operatorname{dist}\left(-J_{\mathbf{c}}(\mathbf{x}^k)^\top\mathbf{c}(\mathbf{x}^k), \mathcal{N}_{\mathcal{X}}(\mathbf{x}^k)\right)^2\right]\\ &= \frac{1}{v^2\beta_{k-1}^2}\mathbb{E}\left[\operatorname{dist}\left(-\beta_{k-1}J_{\mathbf{c}}(\mathbf{x}^k)^\top\mathbf{c}(\mathbf{x}^k), \beta_{k-1}\mathcal{N}_{\mathcal{X}}(\mathbf{x}^k)\right)^2\right]\\ &= \frac{1}{v^2\beta_{k-1}^2}\mathbb{E}\left[\operatorname{dist}\left(-\beta_{k-1}J_{\mathbf{c}}(\mathbf{x}^k)^\top\mathbf{c}(\mathbf{x}^k), \mathcal{N}_{\mathcal{X}}(\mathbf{x}^k)\right)^2\right], \end{aligned} \tag{10}$$

where the last equation follows from $\mathcal{N}_{\mathcal{X}}(\mathbf{x}) = \beta\cdot\mathcal{N}_{\mathcal{X}}(\mathbf{x}), \forall\,\beta > 0, \forall\,\mathbf{x}\in\mathcal{X}$. In addition, by $\partial h(\mathbf{x}) \subseteq \mathcal{N}_{\mathcal{X}}(\mathbf{x}) + \mathcal{B}_M$ from Assumption 2, it holds $\operatorname{dist}\big(\mathbf{z}, \mathcal{N}_{\mathcal{X}}(\mathbf{x}) + \mathcal{B}_M\big) \le \operatorname{dist}\big(\mathbf{z}, \partial h(\mathbf{x})\big)$ for any $\mathbf{z}\in\mathbb{R}^d$. Also, it holds from the triangle inequality that $\operatorname{dist}\big(\mathbf{z}, \mathcal{N}_{\mathcal{X}}(\mathbf{x})\big) \le \operatorname{dist}\big(\mathbf{z}, \mathcal{N}_{\mathcal{X}}(\mathbf{x}) + \mathcal{B}_M\big) + M$. Hence,

$$\operatorname{dist}\big(\mathbf{z}, \mathcal{N}_{\mathcal{X}}(\mathbf{x})\big) \le \operatorname{dist}\big(\mathbf{z}, \partial h(\mathbf{x})\big) + M, \forall\,\mathbf{x}\in\mathcal{X}, \forall\,\mathbf{z}\in\mathbb{R}^d. \tag{11}$$

Using (11) with $\mathbf{z} = -\beta_{k-1}J_{\mathbf{c}}(\mathbf{x}^k)^\top\mathbf{c}(\mathbf{x}^k)$ and noticing

$$\partial h(\mathbf{x}^k) = \partial_x\mathcal{L}_{\beta_{k-1}}(\mathbf{x}^k, \mathbf{y}^{k-1}) - \nabla g(\mathbf{x}^k) - J_{\mathbf{c}}(\mathbf{x}^k)^\top\mathbf{y}^{k-1} - \beta_{k-1}J_{\mathbf{c}}(\mathbf{x}^k)^\top\mathbf{c}(\mathbf{x}^k),$$

it holds

$$\begin{aligned}&\operatorname{dist}\left(-\beta_{k-1}J_{\mathbf{c}}(\mathbf{x}^k)^\top\mathbf{c}(\mathbf{x}^k), \mathcal{N}_{\mathcal{X}}(\mathbf{x}^k)\right)\\ &\le \operatorname{dist}\big(\mathbf{0}, \partial_x\mathcal{L}_{\beta_{k-1}}(\mathbf{x}^k, \mathbf{y}^{k-1}) - \nabla g(\mathbf{x}^k) - J_{\mathbf{c}}(\mathbf{x}^k)^\top\mathbf{y}^{k-1}\big) + M,\end{aligned}$$

which together with (10) gives

$$\mathbb{E}\left[\|\mathbf{c}(\mathbf{x}^k)\|^2\right] \le \frac{1}{v^2\beta_{k-1}^2}\mathbb{E}\left[\operatorname{dist}\big(0, \partial_x\mathcal{L}_{\beta_{k-1}}(\mathbf{x}^k, \mathbf{y}^{k-1}) - \nabla g(\mathbf{x}^k) - J_{\mathbf{c}}(\mathbf{x}^k)^\top\mathbf{y}^{k-1}\big) + M\right]^2.$$

Moreover, applying the triangle inequality to the right hand side of the inequality above, we have

$$\begin{aligned}&\mathbb{E}\left[\|\mathbf{c}(\mathbf{x}^k)\|^2\right]\\ &\le \frac{1}{v^2\beta_{k-1}^2}\mathbb{E}\Big[\Big(\operatorname{dist}\big(0, \partial_x\mathcal{L}_{\beta_{k-1}}(\mathbf{x}^k, \mathbf{y}^{k-1})\big)\\ &\quad + \|\nabla g(\mathbf{x}^k)\| + \|J_{\mathbf{c}}(\mathbf{x}^k)\|\|\mathbf{y}^{k-1}\| + M\Big)^2\Big], \forall\,k\ge 1.\end{aligned} \tag{12}$$

Now, using the Young's inequality and by (6a), (6b), (4) and (9), we obtain

$$\mathbb{E}[\|\mathbf{c}(\mathbf{x}^k)\|^2] \le \frac{4}{v^2\beta_{k-1}^2}(\varepsilon^2 + B_0^2 + B_c^2(k-1)^2\gamma_0^2 + M^2), \forall\,k\ge 1. \tag{13}$$

Since $\beta_k = \beta_0 \sigma^k$, we have from the choice of K in (8) that $\frac{4}{v^2 \beta_{K-1}^2}(\varepsilon^2 + B_0^2 + M^2) \le \frac{\varepsilon^2}{2}$. In addition, let $\alpha = \frac{\sqrt{8} B_c \gamma_0}{\beta_0 v \varepsilon}$. Then from the choice of K, it holds $K - 1 \ge \log_\sigma \alpha^2$, and thus from Lemma 1, we have $\frac{\sigma^{K-1}}{K-1} \ge \alpha$. Hence,

$$\frac{4B_c^2 (K-1)^2 \gamma_0^2}{v^2 \beta_{K-1}^2} = \frac{4B_c^2 (K-1)^2 \gamma_0^2}{v^2 \beta_0^2 \sigma^{2(K-1)}} \le \frac{4B_c^2 \gamma_0^2}{v^2 \beta_0^2 \alpha^2} = \frac{\varepsilon^2}{2}.$$

Thus it follows from (13) that $\mathbb{E}[\|\mathbf{c}(\mathbf{x}^K)\|^2] \le \varepsilon^2$.

Finally, it holds from (4) that

$$\mathbb{E}\left[\operatorname{dist}\left(\mathbf{0}, \partial f_0(\mathbf{x}^K) + J_{\mathbf{c}}(\mathbf{x}^K)^\top \left(\mathbf{y}^{K-1} + \beta_{K-1} \mathbf{c}(\mathbf{x}^K) \right) \right)^2 \right] \le \varepsilon^2.$$

Therefore, $\mathbf{x}^K$ is an ε-KKT point in expectation of (1) with the corresponding multiplier $\mathbf{y}^{K-1} + \beta_{K-1} \mathbf{c}(\mathbf{x}^K)$, according to Definition 1. □

Remark 4 A few remarks about Theorem 1 are as follows. First, the condition $\frac{\sqrt{8} B_c \gamma_0}{\beta_0 v \varepsilon} \ge \frac{8}{\ln \sigma}$ requires to know v. However, this is only for the ease of analysis. We do not actually need the exact value of v. Notice that we can assume $v \le 1$, because if (7) holds for some $v > 1$, it also holds with $v = 1$. In this case, it suffices to pick γ_0 such that $\frac{\sqrt{8} B_c \gamma_0}{\beta_0 \varepsilon} \ge \frac{8}{\ln \sigma}$, which does not involve v. Second, for convex cases where $\mathbf{c}(\cdot)$ consists of affine constraints, it is guaranteed that $\mathbf{c}(\mathbf{x}^{k+1}) = O(\frac{1}{\beta_k})$ if (4) holds in a deterministic way (i.e., without the expectation) and the strong duality holds for (1); see [38] for example. In this case, with $\gamma_k = \gamma_0, \forall k$ for an appropriate γ_0, the dual update will accept β_k as the stepsize for all k. Third, the result in Theorem 1 does not depend on the setting of $\tilde{\mathbf{c}}(\mathbf{x}^{k+1})$. For the special case in (2), we will set $\tilde{\mathbf{c}}(\mathbf{x}^{k+1}) = \mathbf{c}(\mathbf{x}^{k+1})$, and for the general problem (1), we will choose $M_k = \Theta(\frac{1}{\varepsilon^2})$ so that $\mathbb{E}[\|\tilde{\mathbf{c}}(\mathbf{x}^{k+1}) - \mathbf{c}(\mathbf{x}^{k+1})\|^2] = O(\frac{1}{M_k}) = O(\varepsilon^2)$ and thus $\tilde{\mathbf{c}}(\mathbf{x}^{k+1})$ will be close to $\mathbf{c}(\mathbf{x}^{k+1})$ in expectation. Notice that with a smaller M_k, say $M_k = O(1)$, our outer and overall complexity results still hold. However, the multiplier update with a too-small M_k will deviate too much from that by the *classic* ALM and can yield bad practical performance. Finally, we have not specified the subroutine. To have a low overall complexity in terms of the number of sample/component gradients, it is important to obtain each $\mathbf{x}^{k+1}$ efficiently. In the next section, we will exploit the problem structure and design an efficient subroutine to make (4) hold for each k.

3 Momentum-accelerated subroutine and overall oracle complexity

In this section, we give a subroutine to find each $\mathbf{x}^{k+1}$ in Algorithm 1 and thus have a complete algorithm. Besides Assumptions 1-3, we make the following assumptions.

Assumption 4 (mean-squared smoothness) For any $\mathbf{u}, \mathbf{v} \in \mathrm{dom}(h)$, $G_0(\cdot, \xi)$ and $\mathbf{C}(\cdot, \xi)$ satisfy the mean-squared smoothness conditions:

$$\begin{aligned}
&\mathbb{E}_\xi\left[\|\nabla G_0(\mathbf{u}, \xi) - \nabla G_0(\mathbf{v}, \xi)\|^2\right] \le L_0^2\|\mathbf{u} - \mathbf{v}\|^2,\\
&\mathbb{E}_\xi\left[\|J_{\mathbf{C}}(\mathbf{u}, \xi) - J_{\mathbf{C}}(\mathbf{v}, \xi)\|^2\right] \le L_J^2\|\mathbf{u} - \mathbf{v}\|^2,\\
&\mathbb{E}_{\xi_1,\xi_2}\left[\|J_{\mathbf{C}}(\mathbf{u}, \xi_1)^\top \mathbf{C}(\mathbf{u}, \xi_2) - J_{\mathbf{C}}(\mathbf{v}, \xi_1)^\top \mathbf{C}(\mathbf{v}, \xi_2)\|^2\right] \le L_J^2\|\mathbf{u} - \mathbf{v}\|^2,
\end{aligned}$$

where ξ_1 and ξ_2 are independent and follow the same distribution as ξ in (1).

Remark 5 Mean-squared smoothness is needed to have accelerated convergence for a stochastic gradient-type method on solving nonconvex stochastic problems [1, 5, 7, 34, 39]. It naturally holds for the special case in (2) if each component of the objective and constraint functions is smooth. This condition is crucial to obtain our $O(\varepsilon^{-5})$ complexity result.

Assumption 5 (unbiasedness and bounded variance) For any $\mathbf{x} \in \mathrm{dom}(h)$, the objective and constraint functions satisfy

$$\mathbb{E}_\xi[\nabla G_0(\mathbf{x}, \xi)] = \nabla g(\mathbf{x}), \quad \mathbb{E}_\xi[J_{\mathbf{C}}(\mathbf{x}, \xi)] = J_{\mathbf{c}}(\mathbf{x}). \tag{14}$$

Also, there exist $\sigma_g, \sigma_c > 0$ such that for any $\mathbf{x} \in \mathrm{dom}(h)$,

$$\begin{aligned}
&\mathbb{E}_\xi\left[\|\nabla G_0(\mathbf{x}, \xi) - \nabla g(\mathbf{x})\|^2\right] \le \sigma_g^2,\\
&\mathbb{E}_\xi\left[\|J_{\mathbf{C}}(\mathbf{x}, \xi) - J_{\mathbf{c}}(\mathbf{x})\|_2^2\right] \le \sigma_c^2,\\
&\mathbb{E}_{\xi_1,\xi_2}\left[\|J_{\mathbf{C}}(\mathbf{x}, \xi_1)^\top \mathbf{C}(\mathbf{x}, \xi_2) - J_{\mathbf{c}}(\mathbf{x})^\top \mathbf{c}(\mathbf{x})\|^2\right] \le \sigma_c^2,
\end{aligned}$$

where ξ_1 and ξ_2 are independent and follow the same distribution as ξ.

Remark 6 Under Assumption 4 and the unbiasedness condition (14), it can be easily shown that $g(\cdot)$ is L_0-smooth and $\mathbf{c}(\cdot)$ is L_J-smooth; see the arguments at the end of section 2.2 of [34]. Hence, the mean-squared smoothness condition that we assume is stronger than the smoothness condition in [2]. In addition, [21] does not assume smoothness. However, the methods in [2, 21] can still only achieve a result of $O(\varepsilon^{-6})$ even with the mean-squared smoothness condition, as they solve a sequence of strongly-convex subproblems and will not benefit from the mean-squared smoothness structure.

3.1 PStorm subroutine

The smooth part of the AL function $\mathcal{L}_{\beta_k}(\cdot, \mathbf{y}^k)$ has a smoothness parameter depending on β_k that eventually depends on a given tolerance ε. Hence, to achieve a low-order overall complexity result, we need a subroutine whose complexity result has a low-order dependence not only on the pre-given stationarity violation but also on the

smoothness parameter. With the mean-squared smoothness condition, the momentum-based variance-reduced proximal stochastic gradient method (PStorm) in [39] is the one that meets our requirements. Below we first give a modified PStorm with a post-processing step and then in the next subsection, discuss how to apply it to find $\mathbf{x}^{k+1}$ in Algorithm 1. The postprocessing step is necessary in order to produce a solution that meets the requirement in (4) of Algorithm 1. Without the sampling and postprocessing steps, our subroutine is a direct application of the PStorm in [39].

Consider the problem

$$F^* := \min_{\mathbf{x}\in\mathbb{R}^d} \big\{F(\mathbf{x}) := G(\mathbf{x}) + H(\mathbf{x})\big\}, \tag{15}$$

where H is a closed convex function, and G is smooth and possibly nonconvex. Let $\mathcal{A}(\mathbf{x}, \zeta)$ be a stochastic map that depends on a random variable ζ. Suppose the following conditions hold: for some finite constants L_G and σ_G,

$$\mathbb{E}_\zeta\big[\|\mathcal{A}(\mathbf{x}_1, \zeta) - \mathcal{A}(\mathbf{x}_2, \zeta)\|^2\big] \le L_G^2\|\mathbf{x}_1 - \mathbf{x}_2\|^2, \forall\, \mathbf{x}_1, \mathbf{x}_2 \in \mathrm{dom}(H), \tag{16a}$$

$$\mathbb{E}_\zeta\big[\mathcal{A}(\mathbf{x}, \zeta)\big] = \nabla G(\mathbf{x}),\ \mathbb{E}\big[\|\mathcal{A}(\mathbf{x}, \zeta) - \nabla G(\mathbf{x})\|^2\big] \le \sigma_G^2, \forall\, \mathbf{x} \in \mathrm{dom}(H). \tag{16b}$$

With $\mathcal{A}(\cdot, \zeta)$ that satisfies the conditions above, we give the modified PStorm in Algorithm 2 and the complexity result in Lemma 2.

Algorithm 2: PStorm$(G, H, \mathbf{x}^0, L_G, \sigma_G, T, m_0, \bar{\eta}, \delta, \mathcal{A}, \varepsilon)$ for solving (15)

Input: initial point $\mathbf{x}^0 \in \mathrm{dom}(H)$, max iteration number T, smoothness constant L_G, variance bound σ_G^2, step size $\bar{\eta}$, momentum parameter $\delta \in (0, 1)$, and an unbiased gradient estimator $\mathcal{A}$ satisfying (16)

Initialization: Let $\mathbf{d}^0 = \frac{1}{m_0}\sum_{\zeta\in B_0}\mathcal{A}(\mathbf{x}^0, \zeta)$ with B_0 containing m_0 i.i.d. samples.

for $t = 0, 1, \ldots, T-1$ **do**

$$\mathbf{x}^{t+1} = \mathbf{prox}_{\bar{\eta}H}(\mathbf{x}^t - \bar{\eta}\mathbf{d}^t). \tag{17}$$

Take a sample ζ^{t+1} of ζ and compute

$$\mathbf{v}^{t+1} = \mathcal{A}(\mathbf{x}^{t+1}, \zeta^{t+1}), \quad \mathbf{u}^{t+1} = \mathcal{A}(\mathbf{x}^t, \zeta^{t+1}).$$

Let $\mathbf{d}^{t+1} = \mathbf{v}^{t+1} + (1-\delta)(\mathbf{d}^t - \mathbf{u}^{t+1})$.

Choose $\mathbf{x}^\tau$ uniformly at random from $\{\mathbf{x}^0, \ldots, \mathbf{x}^{T-1}\}$.

Sampling: Set $m_1 = \left\lceil \frac{48\sigma_G^2}{\varepsilon^2} \right\rceil$, obtain a set B_1 of m_1 i.i.d samples of ζ, and compute $\mathbf{v} = \frac{1}{m_1}\sum_{\zeta\in B_1}\mathcal{A}(\mathbf{x}^\tau, \zeta)$.

Postprocessing: output $\hat{\mathbf{x}} = \mathbf{prox}_{\bar{\eta}H}(\mathbf{x}^\tau - \bar{\eta}\mathbf{v})$.

Lemma 2 *Assume the conditions in (16). Given an error tolerance $\varepsilon > 0$, choose parameters of Algorithm 2 as follows:*

$$\begin{aligned} \bar{\eta} &= \frac{\eta}{L_G \sqrt[3]{T}},\ \delta = \frac{4\eta^2 + 10\eta^2(2 - \eta T^{-\frac{1}{3}})}{T^{\frac{2}{3}} + 4\eta^2},\ m_0 = \lceil c_0 \sqrt[3]{T} \rceil \\ T &= \left\lceil \frac{48^{\frac{3}{2}} 40^{\frac{3}{2}} \left(\frac{L_G [F(\mathbf{x}^0) - F^*]_{1+}}{\eta} + \frac{\sigma_G^2}{20 c_0 \eta^2} + \frac{24^2 \sigma_G^2 \eta^2}{10} \right)^{\frac{3}{2}}}{\varepsilon^3} \right\rceil, \end{aligned} \tag{18}$$

for some positive constants η and c_0, where $[a]_{1+} := \max\{a, 1\}$ for any $a \in \mathbb{R}$. If $\eta \le \frac{\sqrt[3]{T}}{10}$, then the output $\hat{\mathbf{x}}$ satisfies $\mathbb{E}\big[\mathrm{dist}(\mathbf{0}, \partial F(\hat{\mathbf{x}}))^2\big] \le \varepsilon^2$.

Proof First, directly from Corollary 2.2 of [39], we have

$$\mathbb{E}\left[\left\| \frac{1}{\bar{\eta}} \Big(\mathbf{x}^\tau - \mathbf{prox}_{\bar{\eta} H} \big(\mathbf{x}^\tau - \bar{\eta} \nabla G(\mathbf{x}^\tau) \big) \Big) \right\|^2 \right] \le \frac{\varepsilon^2}{48}. \tag{19}$$

Hence, by the postprocessing step of Algorithm 2, it holds that

$$\begin{aligned} \mathbb{E}\left[\left\| \frac{\hat{\mathbf{x}} - \mathbf{x}^\tau}{\bar{\eta}} \right\|^2 \right] &= \mathbb{E}\left[\left\| \frac{1}{\bar{\eta}} \Big(\mathbf{x}^\tau - \mathbf{prox}_{\bar{\eta} H} (\mathbf{x}^\tau - \bar{\eta} \mathbf{v}) \Big) \right\|^2 \right] \\ &\le \mathbb{E}\left[\left(\left\| \frac{1}{\bar{\eta}} \Big(\mathbf{x}^\tau - \mathbf{prox}_{\bar{\eta} H} \big(\mathbf{x}^\tau - \bar{\eta} \nabla G(\mathbf{x}^\tau) \big) \Big) \right\| + \| \mathbf{v} - \nabla G(\mathbf{x}^\tau) \| \right)^2 \right] \\ &\le 2\mathbb{E}\left[\left\| \frac{1}{\bar{\eta}} \Big(\mathbf{x}^\tau - \mathbf{prox}_{\bar{\eta} H} \big(\mathbf{x}^\tau - \bar{\eta} \nabla G(\mathbf{x}^\tau) \big) \Big) \right\|^2 \right] + 2\mathbb{E}\big[\| \mathbf{v} - \nabla G(\mathbf{x}^\tau) \|^2 \big], \end{aligned}$$

where the first inequality follows from the nonexpansiveness of the proximal gradient mapping, and the second inequality holds due to the Young's inequality. Now by (19) and noticing $\mathbb{E}[\| \mathbf{v} - \nabla G(\mathbf{x}^\tau) \|^2 \le \frac{\sigma_G^2}{m_1} \le \frac{\varepsilon^2}{48}$, we have from the inequality above that

$$\mathbb{E}\left[\left\| \frac{\hat{\mathbf{x}} - \mathbf{x}^\tau}{\bar{\eta}} \right\|^2 \right] \le \frac{\varepsilon^2}{12}. \tag{20}$$

In addition, we have $\frac{\mathbf{x}^\tau - \hat{\mathbf{x}}}{\bar{\eta}} + \nabla G(\hat{\mathbf{x}}) - \mathbf{v} \in \partial F(\hat{\mathbf{x}})$, and thus

$$\begin{aligned} \mathbb{E}\big[\mathrm{dist}(\mathbf{0}, \partial F(\hat{\mathbf{x}}))^2 \big] &\le \mathbb{E}\left[\left\| \frac{\mathbf{x}^\tau - \hat{\mathbf{x}}}{\bar{\eta}} + \nabla G(\hat{\mathbf{x}}) - \nabla G(\mathbf{x}^\tau) + \nabla G(\mathbf{x}^\tau) - \mathbf{v} \right\|^2 \right] \\ &\le 3\mathbb{E}\left[\left\| \frac{\mathbf{x}^\tau - \hat{\mathbf{x}}}{\bar{\eta}} \right\|^2 \right] + 3\mathbb{E}\big[\| \nabla G(\hat{\mathbf{x}}) - \nabla G(\mathbf{x}^\tau) \|^2 \big] \\ &\quad + 3\mathbb{E}\big[\| \nabla G(\mathbf{x}^\tau) - \mathbf{v} \|^2 \big] \end{aligned}$$

$$\leq (3 + 3L_G^2\bar{\eta}^2)\mathbb{E}\left[\left\|\frac{\hat{\mathbf{x}} - \mathbf{x}^\tau}{\bar{\eta}}\right\|^2\right] + \frac{3\varepsilon^2}{48}$$
$$\leq \frac{6\varepsilon^2}{12} + \frac{3\varepsilon^2}{6} = \varepsilon^2,$$

where we have used Young's inequality in the second inequality, the third inequality follows from the L_G-smoothness of G and $\mathbb{E}[\|\mathbf{v} - \nabla G(\mathbf{x}^\tau)\|^2 \leq \frac{\varepsilon^2}{48}$, and the fourth inequality holds because of (20) and $\bar{\eta} \leq \frac{1}{L_G}$. This completes the proof. □

Below we choose an appropriate η and c_0 in Lemma 2 to obtain a complexity result that has a low-order dependence on σ_G, L_G and $F(\mathbf{x}^0) - F^*$.

Lemma 3 *Assume the conditions in (16). Let $\varepsilon > 0$ be given and satisfy $\varepsilon \leq \frac{\sigma_G}{10}\sqrt{1920}\sqrt{3}\left(\frac{24^2}{10}\right)^{\frac{1}{2}}$. Choose parameters of Algorithm 2 as those in (18) with*

$$\eta = \left(\frac{10}{24^2}\right)^{\frac{1}{3}} \frac{\left(L_G[F(\mathbf{x}^0) - F^*]_{1+}\right)^{\frac{1}{3}}}{\sigma_G^{\frac{2}{3}}}, \quad c_0 = \left(\frac{24^2}{10}\right)^{\frac{1}{3}} \frac{\sigma_G^{\frac{8}{3}}}{20\left(L_G[F(\mathbf{x}^0) - F^*]_{1+}\right)^{\frac{4}{3}}}. \tag{21}$$

Then $\mathbb{E}\left[\text{dist}(\mathbf{0}, \partial F(\hat{\mathbf{x}}))^2\right] \leq \varepsilon^2$. The total number of calls to $\mathcal{A}$ is

$$\text{Total}_{\mathcal{A}} = \Theta\left(\frac{\sigma_G L_G[F(\mathbf{x}^0) - F^*]_{1+}}{\varepsilon^3} + \frac{\sigma_G^3}{\varepsilon L_G[F(\mathbf{x}^0) - F^*]_{1+}} + \frac{\sigma_G^{\frac{8}{3}}}{\left(L_G[F(\mathbf{x}^0) - F^*]_{1+}\right)^{\frac{4}{3}}}\right),$$

where $[a]_{1+} := \max\{a, 1\}$ for any $a \in \mathbb{R}$.

Proof First, plugging the chosen η and c_0 into (18), we have

$$T = \left\lceil \frac{1920^{\frac{3}{2}} 3^{\frac{3}{2}} \left(\frac{24^2}{10}\right)^{\frac{1}{2}} \sigma_G L_G[F(\mathbf{x}^0) - F^*]_{1+}}{\varepsilon^3} \right\rceil, \tag{22}$$

and it is straightforward to verify $\eta \leq \frac{\sqrt[3]{T}}{10}$ by the condition $\varepsilon \leq \frac{\sigma_G}{10}\sqrt{1920}\sqrt{3}\left(\frac{24^2}{10}\right)^{\frac{1}{2}}$. Hence, from Lemma 2, it follows that $\mathbb{E}\left[\text{dist}(\mathbf{0}, \partial F(\hat{\mathbf{x}}))^2\right] \leq \varepsilon^2$.

Second, notice that Algorithm 2 calls $\mathcal{A}$ twice for each iteration. Hence, accounting the calls to $\mathcal{A}$ in the initial and postprocessing steps, we obtain the total number of calls to $\mathcal{A}$ is

$$2T + m_0 + m_1 = 2T + \left\lceil c_0\sqrt[3]{T}\right\rceil + \left\lceil \frac{48\sigma_G^2}{\varepsilon^2}\right\rceil \leq 2T + c_0\sqrt[3]{T} + \frac{48\sigma_G^2}{\varepsilon^2} + 2$$

$$
\begin{aligned}
&\leq 2\frac{1920^{\frac{3}{2}}3^{\frac{3}{2}}\left(\frac{24^2}{10}\right)^{\frac{1}{2}}\sigma_G L_G[F(\mathbf{x}^0)-F^*]_{1+}}{\varepsilon^3}\\
&\quad + c_0\left(\frac{1920^{\frac{3}{2}}3^{\frac{3}{2}}\left(\frac{24^2}{10}\right)^{\frac{1}{2}}\sigma_G L_G[F(\mathbf{x}^0)-F^*]_{1+}}{\varepsilon^3}\right)^{\frac{1}{3}} + \frac{48\sigma_G^2}{\varepsilon^2} + c_0 + 4\\
&\leq 2\cdot 1920^{\frac{3}{2}}3^{\frac{3}{2}}\left(\frac{24^2}{10}\right)^{\frac{1}{2}}\left(\frac{\sigma_G L_G[F(\mathbf{x}^0)-F^*]_{1+}}{\varepsilon^3}\right.\\
&\quad \left. + \frac{\sigma_G^3}{\varepsilon L_G[F(\mathbf{x}^0)-F^*]_{1+}} + \frac{\sigma_G^2}{\varepsilon^2} + \frac{\sigma_G^{\frac{8}{3}}}{\left(L_G[F(\mathbf{x}^0)-F^*]_{1+}\right)^{\frac{4}{3}}}\right) + 4.
\end{aligned}
$$

Now notice $\frac{\sigma_G L_G[F(\mathbf{x}^0)-F^*]_{1+}}{\varepsilon^3} + \frac{\sigma_G^3}{\varepsilon L_G[F(\mathbf{x}^0)-F^*]_{1+}} \geq \frac{2\sigma_G^2}{\varepsilon^2}$ and absorb universal constants into Θ. We obtain the desired result and complete the proof. □

Remark 7 In the choice of η and c_0 in Lemma 3, we have implicitly assumed $\sigma_G > 0$. Hence, the claimed result does not apply to a deterministic scenario. Also, the setting of η and c_0 in (21) needs the value of F^* that is unknown. However, we can replace $F(\mathbf{x}^0) - F^*$ by its upper bound, which can be easily obtained, as we will see for the subproblems of Algorithm 1.

3.2 Overall complexity

To apply Algorithm 2 to find each $\mathbf{x}^{k+1}$, the key is to build a stochastic map that satisfies conditions similar to those in (16). The following lemma gives the key.

Lemma 4 *Under Assumptions 1, 4 and 5, given $\mathbf{y} \in \mathbb{R}^m$, let $\Phi(\cdot) := \mathcal{L}_\beta(\cdot, \mathbf{y}) - h(\cdot)$ and*

$$
\Delta(\mathbf{x}, \zeta) := \nabla G_0(\mathbf{x}, \xi_1) + J_{\mathbf{C}}(\mathbf{x}, \xi_1)^\top \mathbf{y} + \beta J_{\mathbf{C}}(\mathbf{x}, \xi_1)^\top \mathbf{C}(\mathbf{x}, \xi_2), \tag{23}
$$

where $\zeta = (\xi_1, \xi_2)$, and ξ_1 and ξ_2 are two independent random variables that follow the same distribution as ξ in (1). Then it holds

$$
\mathbb{E}_\zeta\left[\|\Delta(\mathbf{x}_1, \zeta) - \Delta(\mathbf{x}_2, \zeta)\|^2\right] \leq L_\Phi^2 \|\mathbf{x}_1 - \mathbf{x}_2\|^2, \forall\, \mathbf{x}_1, \mathbf{x}_2 \in \mathrm{dom}(h), \tag{24a}
$$

$$
\mathbb{E}_\zeta\left[\Delta(\mathbf{x}, \zeta)\right] = \nabla\Phi(\mathbf{x}),\ \mathbb{E}\left[\|\Delta(\mathbf{x}, \zeta) - \nabla\Phi(\mathbf{x})\|^2\right] \leq \sigma_\Phi^2, \forall\, \mathbf{x} \in \mathrm{dom}(h), \tag{24b}
$$

where

$$
L_\Phi = \sqrt{3L_0^2 + 3L_J^2\|\mathbf{y}\|^2 + 3\beta^2 L_J^2},\quad \sigma_\Phi = \sqrt{3\sigma_g^2 + 3\sigma_c^2\|\mathbf{y}\|^2 + 3\beta^2\sigma_c^2}.
$$

Proof Because ξ_1 and ξ_2 are independent, it holds

$$
\mathbb{E}_\zeta\left[\Delta(\mathbf{x}, \zeta)\right] = \mathbb{E}_{\xi_1}\left[\nabla G_0(\mathbf{x}, \xi_1)\right] + \mathbb{E}_{\xi_1}\left[J_{\mathbf{C}}(\mathbf{x}, \xi_1)^\top \mathbf{y}\right]
$$

$$+\mathbb{E}_{\xi_1}\big[\beta J_{\mathbf{C}}(\mathbf{x},\xi_1)\big]^\top \mathbb{E}_{\xi_2}\big[\mathbf{C}(\mathbf{x},\xi_2)\big].$$

Since ξ_1 and ξ_2 both follow the distribution of ξ, we have from the definition of $\mathbf{c}(\cdot)$ in (1) and (14) that

$$\mathbb{E}_\zeta\big[\Delta(\mathbf{x},\zeta)\big] = \nabla g(\mathbf{x}) + J_{\mathbf{c}}(\mathbf{x})^\top \mathbf{y} + \beta J_{\mathbf{c}}(\mathbf{x})^\top \mathbf{c}(\mathbf{x}) = \nabla \Phi(\mathbf{x}).$$

In addition, by the Young's inequality, it follows that

$$\begin{aligned}
&\|\Delta(\mathbf{x}_1,\zeta) - \Delta(\mathbf{x}_2,\zeta)\|^2 \\
&\quad \le 3\|\nabla G_0(\mathbf{x}_1,\xi_1) - \nabla G_0(\mathbf{x}_2,\xi_1)\|^2 + 3\|J_{\mathbf{C}}(\mathbf{x}_1,\xi_1)^\top \mathbf{y} - J_{\mathbf{C}}(\mathbf{x}_2,\xi_1)^\top \mathbf{y}\|^2 \\
&\qquad + 3\beta^2 \|J_{\mathbf{C}}(\mathbf{x}_1,\xi_1)^\top \mathbf{C}(\mathbf{x}_1,\xi_2) - J_{\mathbf{C}}(\mathbf{x}_2,\xi_1)^\top \mathbf{C}(\mathbf{x}_2,\xi_2)\|^2,
\end{aligned}$$

which together with Assumption 4 gives

$$\mathbb{E}_\zeta\big[\|\Delta(\mathbf{x}_1,\zeta) - \Delta(\mathbf{x}_2,\zeta)\|^2\big] \le 3(L_0^2 + L_J^2\|\mathbf{y}\|^2 + \beta^2 L_J^2)\|\mathbf{x}_1 - \mathbf{x}_2\|^2.$$

Hence, (24a) holds. Similarly, by the Young's inequality and Assumption 5, we can show (24b) and complete the proof. □

With Lemma 4, we are able to apply Algorithm 2 to find each $\mathbf{x}^{k+1}$. The theorem below gives the oracle complexity for the k-th outer iteration of Algorithm 1.

Theorem 2 (Oracle complexity per outer iteration) *Let* $(\mathbf{x}^k, \mathbf{y}^k)$ *be the k-th iterate generated in Algorithm 1 with a given tolerance* $\varepsilon > 0$. *Define*

$$F_k(\mathbf{x}) := \mathcal{L}_{\beta_k}(\mathbf{x}, \mathbf{y}^k), \quad F_k^* := \min_{\mathbf{x}} F_k(\mathbf{x}), \quad \Phi_k(\mathbf{x}) := F_k(\mathbf{x}) - h(\mathbf{x}).$$

Under Assumptions 1, 4 and 5, if $\varepsilon \le \frac{3\sqrt{\sigma_g^2+\beta_0^2\sigma_c^2}}{10}\sqrt{1920}\big(\frac{24^2}{10}\big)^{\frac{1}{2}}$, *then we can find* $\mathbf{x}^{k+1}$ *that satisfies* (4) *by Algorithm 2 with the following call*

$$\mathbf{x}^{k+1} \leftarrow \mathrm{PStorm}(\Phi_k, h, \mathbf{x}^k, L_{\Phi_k}, \sigma_{\Phi_k}, T_k, m_{0,k}, \bar{\eta}_k, \delta_k, \Delta_k, \varepsilon), \tag{25}$$

where

$$L_{\Phi_k} = \sqrt{3L_0^2 + 3L_J^2\|\mathbf{y}^k\|^2 + 3\beta_k^2 L_J^2}, \quad \sigma_{\Phi_k} = \sqrt{3\sigma_g^2 + 3\sigma_c^2\|\mathbf{y}^k\|^2 + 3\beta_k^2\sigma_c^2}, \tag{26}$$

$$\Delta_k(\mathbf{x},\zeta) := \nabla G_0(\mathbf{x},\xi_1) + J_{\mathbf{C}}(\mathbf{x},\xi_1)^\top \mathbf{y}^k + \beta_k J_{\mathbf{C}}(\mathbf{x},\xi_1)^\top \mathbf{C}(\mathbf{x},\xi_2). \tag{27}$$

$$\bar{\eta}_k = \frac{\eta_k}{L_{\Phi_k}\sqrt[3]{T_k}}, \ \delta_k = \frac{4\eta_k^2 + 10\eta_k^2(2 - \eta_k T_k^{-\frac{1}{3}})}{T_k^{\frac{2}{3}} + 4\eta_k^2}, \ m_{0,k} = \lceil c_{0,k}\sqrt[3]{T_k}\rceil \tag{28}$$

$$T_k = \left\lceil \frac{48^{\frac{3}{2}}40^{\frac{3}{2}}\left(\frac{L_{\Phi_k}[F_k(\mathbf{x}^k) - F_k^*]_{1+}}{\eta} + \frac{\sigma_{\Phi_k}^2}{20c_{0,k}\eta_k^2} + \frac{24^2\sigma_{\Phi_k}^2\eta_k^2}{10}\right)^{\frac{3}{2}}}{\varepsilon^3} \right\rceil \tag{29}$$

with

$$\eta_k = \Big(\frac{10}{24^2}\Big)^{\frac{1}{3}} \frac{\big(L_{\Phi_k}[F_k(\mathbf{x}^k) - F_k^*]_{1+}\big)^{\frac{1}{3}}}{\sigma_{\Phi_k}^{\frac{2}{3}}}, \quad c_{0,k} = \Big(\frac{24^2}{10}\Big)^{\frac{1}{3}} \frac{\sigma_{\Phi_k}^{\frac{8}{3}}}{20\big(L_{\Phi_k}[F_k(\mathbf{x}^k) - F_k^*]_{1+}\big)^{\frac{4}{3}}}. \quad (30)$$

In addition, $\mathcal{T}_k$ calls to the stochastic first-order oracle that is defined in Assumption 1 will be enough to produce $\mathbf{x}^{k+1}$, where

$$\mathcal{T}_k = O\left(\frac{\sigma_{\Phi_k} L_{\Phi_k}[F_k(\mathbf{x}^k) - F_k^*]_{1+}}{\varepsilon^3} + \frac{\sigma_{\Phi_k}^3}{\varepsilon L_{\Phi_k}} + \frac{\sigma_{\Phi_k}^{\frac{8}{3}}}{L_{\Phi_k}^{\frac{4}{3}}}\right). \quad (31)$$

Proof By Lemma 4, it holds

$$\mathbb{E}_\zeta\big[\|\Delta_k(\mathbf{x}_1,\zeta) - \Delta_k(\mathbf{x}_2,\zeta)\|^2\big] \le L_{\Phi_k}^2 \|\mathbf{x}_1 - \mathbf{x}_2\|^2, \forall\, \mathbf{x}_1, \mathbf{x}_2 \in \mathrm{dom}(h), \quad (32a)$$

$$\mathbb{E}_\zeta\big[\Delta_k(\mathbf{x},\zeta)\big] = \nabla\Phi_k(\mathbf{x}),\ \mathbb{E}\big[\|\Delta_k(\mathbf{x},\zeta) - \nabla\Phi_k(\mathbf{x})\|^2\big] \le \sigma_{\Phi_k}^2, \forall\, \mathbf{x} \in \mathrm{dom}(h), \quad (32b)$$

where L_{Φ_k} and σ_{Φ_k} are given in (26). Hence, $\varepsilon \le \frac{\sigma_{\Phi_k}}{10}\sqrt{1920}\sqrt{3}\big(\frac{24^2}{10}\big)^{\frac{1}{2}}$ by $\beta_0 \le \beta_k, \forall\, k \ge 0$ and the assumed condition on ε. Thus, from Lemma 3, the point $\mathbf{x}^{k+1}$ returned by Algorithm 2 with the call in (25) is an ε-stationary point $\mathbf{x}^{k+1}$ of $F_k(\cdot)$ in expectation, i.e., $\mathbb{E}\big[\mathrm{dist}(\mathbf{0}, \partial_x \mathcal{L}_{\beta_k}(\mathbf{x}^{k+1}, \mathbf{y}^k))^2 \,|\, \mathbf{y}^k\big] \le \varepsilon^2$ and, the total number of calls to Δ_k is

$$\begin{aligned}\mathrm{Total}_{\Delta_k} =& \Theta\left(\frac{\sigma_{\Phi_k} L_{\Phi_k}[F_k(\mathbf{x}^k) - F_k^*]_{1+}}{\varepsilon^3} + \frac{\sigma_{\Phi_k}^3}{\varepsilon L_{\Phi_k}[F_k(\mathbf{x}^k) - F_k^*]_{1+}}\right.\\ &\left. + \frac{\sigma_{\Phi_k}^{\frac{8}{3}}}{\big(L_{\Phi_k}[F_k(\mathbf{x}^k) - F_k^*]_{1+}\big)^{\frac{4}{3}}}\right)\\ =& O\left(\frac{\sigma_{\Phi_k} L_{\Phi_k}[F_k(\mathbf{x}^k) - F_k^*]_{1+}}{\varepsilon^3} + \frac{\sigma_{\Phi_k}^3}{\varepsilon L_{\Phi_k}} + \frac{\sigma_{\Phi_k}^{\frac{8}{3}}}{L_{\Phi_k}^{\frac{4}{3}}}\right), \end{aligned} \quad (33)$$

where we have used $[F_k(\mathbf{x}^k) - F_k^*]_{1+} \ge 1$ in the second equation. Since each call to Δ_k will need two stochastic first-order oracles as we assumed in Assumption 1, we have $\mathcal{T}_k = 2 \cdot \mathrm{Total}_{\Delta_k}$ and thus complete the proof. □

Remark 8 In the parameter settings of (26)–(30), we used the unknown value F_k^*. As we point out in Remark 7, we can replace $F_k(\mathbf{x}^k) - F_k^*$ by its upper bound such as the one we establish in (36) below.

The lemma below is used to bound $\mathbb{E}[F_k(\mathbf{x}^k) - F_k^*]$.

Lemma 5 *Let $F_k(\cdot)$ and F_k^* be defined in Theorem 2. If (9) and (13) hold, then*

$$F_0(\mathbf{x}^0) - F_0^* \le 2B_0 + \frac{\beta_0}{2}\|\mathbf{c}(\mathbf{x}^0)\|^2, \tag{34}$$

and for any $k \ge 1$,

$$\begin{aligned}
&\mathbb{E}[F_k(\mathbf{x}^k) - F_k^*] \\
&\le B_F := 2B_0 + \frac{\varepsilon^2 + B_0^2 + M^2}{v\beta_0}\Big(\frac{2\sigma}{v} + 1\Big) \\
&\quad + \frac{\gamma_0^2}{2v\beta_0}\Big(\frac{2\sigma}{v} + 1\Big)\big(v + 2B_c^2\big)\frac{4}{(\ln\sigma)^2\sigma^{\frac{2}{\ln\sigma}}}.
\end{aligned} \tag{35}$$

Proof Suppose $\hat{\mathbf{x}}^k = \arg\min_{\mathbf{x}} F_k(\mathbf{x})$, i.e., $F_k^* = F_k(\hat{\mathbf{x}}^k)$. Then it holds

$$\begin{aligned}
F_k(\mathbf{x}^k) - F_k^* &= f_0(\mathbf{x}^k) + \langle \mathbf{y}^k, \mathbf{c}(\mathbf{x}^k)\rangle + \frac{\beta_k}{2}\|\mathbf{c}(\mathbf{x}^k)\|^2 - f_0(\hat{\mathbf{x}}^k) \\
&\quad - \langle \mathbf{y}^k, \mathbf{c}(\hat{\mathbf{x}}^k)\rangle - \frac{\beta_k}{2}\|\mathbf{c}(\hat{\mathbf{x}}^k)\|^2 \\
&\le f_0(\mathbf{x}^k) + \langle \mathbf{y}^k, \mathbf{c}(\mathbf{x}^k)\rangle + \frac{\beta_k}{2}\|\mathbf{c}(\mathbf{x}^k)\|^2 - f_0(\hat{\mathbf{x}}^k) + \frac{1}{2\beta_k}\|\mathbf{y}^k\|^2 \\
&\le 2B_0 + k\gamma_0\|\mathbf{c}(\mathbf{x}^k)\| + \frac{\beta_k}{2}\|\mathbf{c}(\mathbf{x}^k)\|^2 + \frac{k^2\gamma_0^2}{2\beta_k}, \forall\, k \ge 0,
\end{aligned} \tag{36}$$

where the first inequality is by the Young's inequality, and the second inequality follows from (6a). Hence for $k = 0$, taking expectation on both sides of (36) gives (34), and for $k \ge 1$,

$$\begin{aligned}
\mathbb{E}[F_k(\mathbf{x}^k) - F_k^*] \le& 2B_0 + k\gamma_0\mathbb{E}\big[\|\mathbf{c}(\mathbf{x}^k)\|\big] + \frac{\beta_k}{2}\mathbb{E}\big[\|\mathbf{c}(\mathbf{x}^k)\|^2\big] + \frac{k^2\gamma_0^2}{2\beta_k} \\
\le& 2B_0 + \frac{2k\gamma_0\sqrt{\varepsilon^2 + B_0^2 + B_c^2(k-1)^2\gamma_0^2 + M^2}}{v\beta_{k-1}} \\
&+ \frac{2\beta_k}{v^2\beta_{k-1}^2}\left(\varepsilon^2 + B_0^2 + B_c^2(k-1)^2\gamma_0^2 + M^2\right) + \frac{k^2\gamma_0^2}{2\beta_k} \\
\le& 2B_0 + \frac{k^2\gamma_0^2}{v\beta_{k-1}} + \left(\frac{2\beta_k}{v^2\beta_{k-1}^2} + \frac{1}{v\beta_{k-1}}\right) \\
&(\varepsilon^2 + B_0^2 + B_c^2(k-1)^2\gamma_0^2 + M^2) + \frac{k^2\gamma_0^2}{2\beta_k},
\end{aligned} \tag{37}$$

where in the second inequality we have used (9), (13) and the Jensen's inequality, and the third inequality follows from the Young's inequality.

Notice that $\frac{x^2}{\sigma^x}$ attains its maximum at $x = \frac{2}{\ln\sigma}$ for $x > 0$. Thus

$$\frac{x^2}{\sigma^x} \le \frac{4}{(\ln\sigma)^2\sigma^{\frac{2}{\ln\sigma}}}, \forall\, x > 0. \tag{38}$$

Now by $\beta_k = \beta_0\sigma^k$, we have for any $k \ge 1$,

$$\begin{aligned}
&\frac{2\beta_k}{v^2\beta_{k-1}^2} + \frac{1}{v\beta_{k-1}} = \frac{2\sigma}{v^2\beta_0\sigma^{k-1}} + \frac{1}{v\beta_0\sigma^{k-1}} \le \frac{2\sigma}{v^2\beta_0} + \frac{1}{v\beta_0},\\
&\frac{k^2\gamma_0^2}{v\beta_{k-1}} + \frac{k^2\gamma_0^2}{2\beta_k} = \Big(\frac{\sigma\gamma_0^2}{v\beta_0} + \frac{\gamma_0^2}{2\beta_0}\Big)\frac{k^2}{\sigma^k} \le \Big(\frac{\sigma\gamma_0^2}{v\beta_0} + \frac{\gamma_0^2}{2\beta_0}\Big)\frac{4}{(\ln\sigma)^2\sigma^{\frac{2}{\ln\sigma}}},\\
&\left(\frac{2\beta_k}{v^2\beta_{k-1}^2} + \frac{1}{v\beta_{k-1}}\right)B_c^2(k-1)^2\gamma_0^2 \le B_c^2\gamma_0^2\Big(\frac{2\sigma}{v^2\beta_0} + \frac{1}{v\beta_0}\Big)\frac{4}{(\ln\sigma)^2\sigma^{\frac{2}{\ln\sigma}}}.
\end{aligned}$$

Plugging the three inequalities above into (37) and using the definition of B_F, we obtain (35). □

By Theorem 2 and Lemma 5, we are ready to show the overall oracle complexity of Algorithm 1 to produce an ε-KKT point of (1) in expectation.

Theorem 3 (Overall complexity of Algorithm 1) *Under Assumptions 1 through 5, let* $\varepsilon \in (0, 1)$ *be a given tolerance. Suppose* $\varepsilon \le \frac{3\sqrt{\sigma_g^2+\beta_0^2\sigma_c^2}}{10}\sqrt{1920}\big(\frac{24^2}{10}\big)^{\frac{1}{2}}$. *In Algorithm 1, set* $M_k = \Theta(\varepsilon^{-2})$ *and* $\gamma_k = \gamma_0, \forall\, k \ge 0$ *for some* $\gamma_0 > 0$ *such that* $\frac{\sqrt{8}B_c\gamma_0}{\beta_0 v\varepsilon} \ge \frac{8}{\ln\sigma}$. *Then it can produce an* ε*-KKT point of* (1) *in expectation, by using Algorithm 2 as the subroutine and calling it via* (25). *In addition, the total number* $\mathrm{Oracle}_{\mathrm{total}}$ *of calls to the oracle defined in Assumption 1 satisfies*

$$\mathbb{E}\big[\mathrm{Oracle}_{\mathrm{total}}\big] = O(\varepsilon^{-5}).$$

Proof From Theorem 1, we know that $\mathbf{x}^K$ is an ε-KKT point of (1) in expectation, where K is given in (8). Hence, the oracle complexity $\mathrm{Oracle}_{\mathrm{total}}$ of Algorithm 1 is upper bounded by $\sum_{k=0}^{K-1}(\mathcal{T}_k + M_k)$ with $\mathcal{T}_k$ defined in (31). To upper bound $\mathbb{E}\big[\mathrm{Oracle}_{\mathrm{total}}\big]$, it suffices to upper bound $\mathbb{E}\big[\mathcal{T}_k\big]$ for each $k < K$.

By $\beta_k = \beta_0\sigma^k$ and (9), we have from (26) that

$$\sqrt{3}\beta_0 L_J\sigma^k \le L_{\Phi_k} \le \sqrt{3L_0^2 + 3L_J^2k^2\gamma_0^2 + 3\beta_0^2L_J^2\sigma^{2k}} \le \sqrt{3}\big(L_0 + L_Jk\gamma_0 + \beta_0L_J\sigma^k\big). \tag{39}$$

and

$$\sigma_{\Phi_k} \le \sqrt{3\sigma_g^2 + 3\sigma_c^2k^2\gamma_0^2 + 3\beta_0^2\sigma_c^2\sigma^{2k}} \le \sqrt{3}\big(\sigma_g + \sigma_ck\gamma_0 + \beta_0\sigma_c\sigma^k\big). \tag{40}$$

Hence,

$$\begin{aligned}\frac{\sigma_{\Phi_k}^3}{\varepsilon L_{\Phi_k}}+\frac{\sigma_{\Phi_k}^{\frac{8}{3}}}{L_{\Phi_k}^{\frac{4}{3}}}&=\frac{\sigma_{\Phi_k}}{\varepsilon}\frac{\sigma_{\Phi_k}^2}{L_{\Phi_k}}+\Big(\frac{\sigma_{\Phi_k}^2}{L_{\Phi_k}}\Big)^{\frac{4}{3}}\\&\le\frac{\sqrt{3}\big(\sigma_g+\sigma_c k\gamma_0+\beta_0\sigma_c\sigma^k\big)}{\varepsilon}\cdot\frac{3\sigma_g^2+3\sigma_c^2k^2\gamma_0^2+3\beta_0^2\sigma_c^2\sigma^{2k}}{\sqrt{3}\beta_0L_J\sigma^k}\\&\quad+\Big(\frac{3\sigma_g^2+3\sigma_c^2k^2\gamma_0^2+3\beta_0^2\sigma_c^2\sigma^{2k}}{\sqrt{3}\beta_0L_J\sigma^k}\Big)^{\frac{4}{3}}.\end{aligned}\tag{41}$$

By (38), it holds $\frac{3\sigma_c^2k^2\gamma_0^2}{\sqrt{3}\beta_0L_J\sigma^k}\le\frac{3\sigma_c^2\gamma_0^2}{\sqrt{3}\beta_0L_J}\frac{4}{(\ln\sigma)^2\sigma^{\frac{2}{\ln\sigma}}}$ and $k\le\sigma^k\frac{4}{(\ln\sigma)^2\sigma^{\frac{2}{\ln\sigma}}}$. Hence,

$$\sqrt{3}\big(\sigma_g+\sigma_c k\gamma_0+\beta_0\sigma_c\sigma^k\big)=\Theta(\sigma^k),\quad\frac{3\sigma_g^2+3\sigma_c^2k^2\gamma_0^2+3\beta_0^2\sigma_c^2\sigma^{2k}}{\sqrt{3}\beta_0L_J\sigma^k}=\Theta(\sigma^k).\tag{42}$$

Therefore, (41) implies

$$\sum_{k=0}^{K-1}\left(\frac{\sigma_{\Phi_k}^3}{\varepsilon L_{\Phi_k}}+\frac{\sigma_{\Phi_k}^{\frac{8}{3}}}{L_{\Phi_k}^{\frac{4}{3}}}\right)=O\left(\sum_{k=0}^{K-1}\Big(\frac{\sigma^{2k}}{\varepsilon}+\sigma^{\frac{4k}{3}}\Big)\right)=O\left(\frac{\sigma^{2K}}{\varepsilon}\right)\tag{43}$$

In addition, $\mathbb{E}\big[\sigma_{\Phi_k}L_{\Phi_k}[F_k(\mathbf{x}^k)-F_k^*]_{1+}\big]\le\mathbb{E}\big[\sigma_{\Phi_k}L_{\Phi_k}(F_k(\mathbf{x}^k)-F_k^*)\big]+\sigma_{\Phi_k}L_{\Phi_k}$. Hence, from Lemma 5, (39), and (40), we obtain

$$\begin{aligned}&\mathbb{E}\Big[\sigma_{\Phi_k}L_{\Phi_k}[F_k(\mathbf{x}^k)-F_k^*]_{1+}\Big]\\&\quad\le3\big(L_0+L_Jk\gamma_0+\beta_0L_J\sigma^k\big)\big(\sigma_g+\sigma_ck\gamma_0+\beta_0\sigma_c\sigma^k\big)\\&\qquad\left(1+\max\left\{B_F,2B_0+\frac{\beta_0}{2}\|\mathbf{c}(\mathbf{x}^0)\|^2\right\}\right),\end{aligned}$$

which together with (42) gives

$$\sum_{k=0}^{K-1}\mathbb{E}\Big[\sigma_{\Phi_k}L_{\Phi_k}[F_k(\mathbf{x}^k)-F_k^*]_{1+}\Big]=O\left(\sum_{k=0}^{K-1}\sigma^{2k}\right)=O(\sigma^{2K}).\tag{44}$$

Therefore, from (31), (43), and (44), it follows that $\mathbb{E}\big[\sum_{k=0}^{K-1}\mathcal{T}_k\big]=O\left(\frac{\sigma^{2K}}{\varepsilon^3}\right)$. Plugging K given in (8), we have $\mathbb{E}\big[\sum_{k=0}^{K-1}\mathcal{T}_k\big]=O(\varepsilon^{-5})$. Since $M_k=O(\varepsilon^{-2})$ for all k, we obtain $\mathbb{E}\big[\text{Oracle}_{\text{total}}\big]=\mathbb{E}\big[\sum_{k=0}^{K-1}(\mathcal{T}_k+M_k)\big]=O(\varepsilon^{-5})$ and complete the proof. □

Before concluding this section, we make some discussions on certain other frameworks and their oracle complexity for solving (1). Both of [15, 18] focus on

deterministic FOMs, but their frameworks can be easily adapted to stochastic versions. Different from the proposed Stoc-iALM in Algorithm 1, the methods in [15, 18] solve a sequence of strongly-convex subproblems based on the proximal-point (PP) framework. Hence, they can be adapted to stochastic FOMs if a stochastic gradient-type method is applied to those strongly-convex subproblems. However, the resulting stochastic methods will have higher-order complexity than our proposed method on solving (1). Let us focus on the adaptation of the method in [15] while a similar discussion applies to that in [18]. Instead of using a stochastic method to directly find a stationary point of each ALM subproblem, [15] applies the PP method, and each PP subproblem is in the form of

$$\min_{\mathbf{x}} U(\mathbf{x}) := \mathcal{L}_\beta(\mathbf{x}, \mathbf{y}) + \frac{\rho}{2}\|\mathbf{x} - \hat{\mathbf{x}}\|^2, \tag{45}$$

where $\hat{\mathbf{x}}$ is the proximal center that is set to the approximate solution of the previous PP subproblem, and $\rho > 0$ is chosen to make the objective in (45) to be strongly convex. In order to find an ε-stationary point of $\mathcal{L}_\beta(\cdot, \mathbf{y})$, the PP method needs to approximately solve $\Theta(\frac{\rho}{\varepsilon^2})$ PP subproblems, for each of which an $\frac{\varepsilon}{4}$-stationary solution is required. Suppose U is L_U-smooth in (45) and has the unique minimizer $\mathbf{x}_U$. Then it holds $\|\nabla U(\mathbf{x})\| \leq L_U\|\mathbf{x} - \mathbf{x}_U\|$. Now let us apply the result in [28, Lemma 1] for the stochastic gradient method (SGM) on solving a smooth strongly-convex stochastic problem. The SGM needs $\Theta(\frac{L_U^2 G_U^2}{\rho^2\varepsilon^2})$ iterations to produce a point $\mathbf{x}$ such that $L_U\mathbb{E}\big[\|\mathbf{x} - \mathbf{x}_U\|\big] \leq \frac{\varepsilon}{4}$, where G_U is a bound of ∇U. Therefore, for each ALM subproblem, the PP method together with the (optimal) SGM as a subroutine will incur $\Theta(\frac{L_U^2 G_U^2}{\rho\varepsilon^4})$ oracle calls. Since the penalty parameter β will geometrically increase to $\Theta(\frac{1}{\varepsilon})$, the constants L_U, ρ and G_U of the corresponding ALM subproblem will eventually all be in the order of $\frac{1}{\varepsilon}$. This way, we obtain the oracle complexity of $\tilde{O}(\varepsilon^{-7})$ if the method in [15] is adapted to a stochastic version.

4 Numerical results

In this section, we demonstrate the numerical performance of the proposed Stoc-iALM in Algorithm 1 (with PStorm in Algorithm 2 as the subroutine) on solving a fairness constrained problem and a Neyman-Pearson Classification problem. We compare it to the IPC method in [21] that achieves the state-of-the-art complexity for solving (1). All the tests were performed in MATLAB 2019b on a Macbook Pro with 4 cores and 16GB memory.

4.1 Nonconvex fairness constrained problem

Let $\mathbf{x}$ denote the parameters of a linear model and $f(\mathbf{x}; \mathbf{a}, b) = \phi_\alpha(l(\mathbf{x}; \mathbf{a}, b))$ be the truncated logistic loss function, where $l(\mathbf{x}; \mathbf{a}, b) = \log(1 + \exp(-b\mathbf{a}^\top\mathbf{x}))$, $\phi_\alpha(s) = \alpha\log(1 + \frac{s}{\alpha})$, and $\alpha = 2$ is set in our tests. Suppose there is a labeled dataset $D =$

$\{(\mathbf{a}_i, b_i)\}_{i=1}^{|D|}$, a possibly unlabeled dataset $S = \{\mathbf{a}_j\}_{j=1}^{|S|}$, and a subset $S_{\min} \subseteq S$ of the minority population in S. Then the problem of training $\mathbf{x}$ using the loss $f(\mathbf{x}; \mathbf{a}, b)$ with a fairness constraint [21] can be formulated as

$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{R}^d} \ & f_0(\mathbf{x}) := \frac{1}{|D|} \sum_{(\mathbf{a},b)\in D} f(\mathbf{x}; \mathbf{a}, b), \\ \text{s.t.} \ & f_1(\mathbf{x}) := c \sum_{\mathbf{a}\in S} \sigma(\mathbf{a}^\top \mathbf{x}) - \sum_{\mathbf{a}\in S_{\min}} \sigma(\mathbf{a}^\top \mathbf{x}) \le 0, \end{aligned} \tag{46}$$

where $c \in (0, 1)$ is a fairness parameter and $\sigma(s) = \frac{\exp(s)}{1+\exp(s)}$. The fairness constraint above aims at forcing the classifier to have a positive prediction on the minority group often enough. Following [21], we use three data sets: *bank-marketing* from UCI repository [6] (shortened as *bank* below) with $d = 81$ and $(|D|, |S|, |S_{\min}|) = (22605, 22605, 233)$, *a9a* from LIBSVM library [4] with $d = 123$ and $(|D|, |S|, |S_{\min}|) = (32561, 16281, 1561)$, and *loan* from LendingClub (which contains the information of 128375 loans issued in the fourth quarter of 2018; see [21] for more description) with $d = 250$ and $(|D|, |S|, |S_{\min}|) = (64485, 63890, 31966)$. We set the fairness parameter to $c = 0.4$ for the *bank* dataset, $c = 0.1$ for the *a9a* dataset, and $c = 0.6$ for the *loan* dataset.

To solve (46) by the proposed Stoc-iALM, we reformulate its inequality constraint to an equality constraint $f_1(\mathbf{x}) + \mathbf{v} = 0$ where $\mathbf{v} \ge 0$ is enforced. Notice that the reformulation has equivalent stationarity conditions to the original model (46) as shown in [15]. The IPC method in [21] is applied directly to (46), and following [21], we adopt its deterministic version. We only compare to the IPC, as it is demonstrated in [21] to outperform other methods on solving (46) such as the Penalty with trust region method in [3] and the subgradient method in [41].

The tolerance is set to $\varepsilon = 0.01$ in all tests. Our proposed Stoc-iALM is terminated, if both primal residual (computed as $[f_1(\mathbf{x})]_+$) and dual residual (computed as $\|\nabla f_0(\mathbf{x}) + y\nabla f_1(\mathbf{x})\|$) are below ε, where $\mathbf{x}$ and y are the primal and dual iterate generated by the algorithm. The IPC method does not generate a dual iterate, so for a fair comparison, we compute an optimal dual variable $\mathbf{z} \ge 0$ that minimizes the squared sum of the violation to the dual feasibility and the complementary slackness conditions of (46):

$$\min_{\mathbf{z}\ge 0} \|\nabla f_0(\mathbf{x}) + J_{f_1}(\mathbf{x})^\top \mathbf{z}\|^2 + |\mathbf{z}^\top f_1(\mathbf{x})|^2. \tag{47}$$

Since f_1 is a scalar function in (46), it is not difficult to have the optimal $\mathbf{z} = \left[-\frac{\nabla f_0(\mathbf{x})^\top \nabla f_1(\mathbf{x})}{f_1(\mathbf{x})^2 + \|\nabla f_1(\mathbf{x})\|^2}\right]_+$. Given this $\mathbf{z}$, the IPC is terminated if both primal residual $[f_1(\mathbf{x})]_+$ and dual residual $\|\nabla f_0(\mathbf{x}) + \mathbf{z}\nabla f_1(\mathbf{x})\|$ are below the given tolerance ε. For Stoc-iALM, at the k-th outer iteration, we set $\beta_k = 2.5^k$ and the smoothness parameter to $10 + \beta_k$, and we set $\tilde{\mathbf{c}}(\mathbf{x}^{k+1}) = \mathbf{c}(\mathbf{x}^{k+1})$ in the $\mathbf{y}$-update (5). In the PStorm subroutine, we set the mini-batch size to 30 for all three data sets. The parameter settings of the IPC exactly follow from the code of [21] that was kindly provided by the authors. The primal and dual residuals are recorded after every 50 inner iterations for

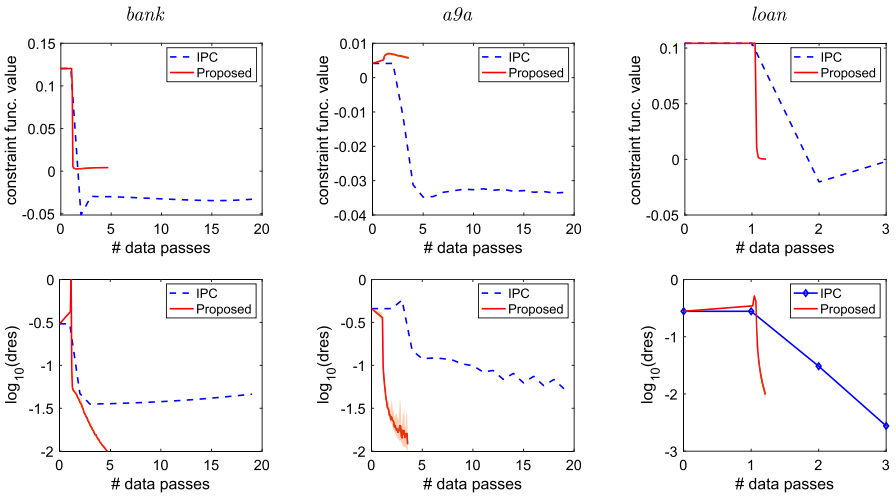

Fig. 1 Mean curves and standard deviation (by shadow area[†]) of 10 different trials by the proposed Stoc-iALM and the curves by the IPC in [21] on solving the fairness constrained problem (46) with *bank*, *a9a*, and *loan* data sets (from left to right). [†] Some shadow areas are invisible because the deviations are too small

Stoc-iALM, and after every data pass for IPC. Both methods start from a zero vector. As the proposed method is randomized, we run it for 10 independent trials by using different random seeds, while IPC is deterministic and thus we only perform one trial.

Table 1 lists the violation of primal feasibility and the violation to the dual feasibility at the produced ε-KKT point, and the number of data passes (shortened by `pres`, `dres` and `#data` respectively) for each method to produce such a point. Figure 1 plots the curves of the constraint function value and `dres` at generated iterates by the proposed Stoc-iALM and the IPC, where the solid red curve shows the average results and the shadow area represents the standard deviation for the proposed method. To clearly show the difference of the results by the proposed method and IPC, we only plot the curves by IPC up to 20 number of data passes for the *bank* and *a9a* data sets. From the results in Table 1 and Fig. 1, we see that both methods can reduce `pres` and `dres` below the given tolerance. However, the IPC needs significantly more data passes, especially for the *bank* and *a9a* data sets. In addition, the proposed method can perform well for all 10 trials.

4.2 Nonconvex Neyman–Pearson classification

In this subsection, we test our proposed Stoc-iALM (Algorithm 1) with the PStorm (Algorithm 2) subroutine on solving the nonconvex Neyman-Pearson classification problem [29, 40]. The problem aims at minimizing the false-negative error subject to a constraint on the level of false-positive error. It can be formulated as

Table 1 The violation of primal feasibility and the violation to the dual feasibility at the produced ε-KKT point with $\varepsilon = 10^{-2}$, and the number of data passes (shortened by pres, dres and #data[†] respectively) of 10 trials with different random seeds by the proposed Stoc-iALM and the IPC in [21] on solving the fairness constrained problem (46) with *a9a*, *bank*, and *loan* data sets (from left to right)

Method [*a9a*]	Pres	Dres	#Data	Method [*bank*]	Pres	Dres	#Data	Method [*loan*]	Pres	Dres	#Data
Stoc-iALM (1)	5.7e−3	9.3e−3	3.83	Stoc-iALM (1)	4.1e−3	9.7e−3	4.78	Stoc-iALM (1)	2.2e−4	9.1e−3	1.21
Stoc-iALM (2)	5.8e−3	9.5e−3	3.90	Stoc-iALM (2)	4.1e−3	9.9e−3	4.72	Stoc-iALM (2)	2.6e−4	9.7e−3	1.23
Stoc-iALM (3)	5.7e−3	9.4e−3	3.97	Stoc-iALM (3)	4.2e−3	1.00e−2	4.72	Stoc-iALM (3)	2.3e−4	9.3e−3	1.21
Stoc-iALM (4)	6.0e−3	1.00e−2	3.63	Stoc-iALM (4)	4.2e−3	9.9e−3	4.72	Stoc-iALM (4)	2.4e−4	9.6e−3	1.21
Stoc-iALM (5)	5.8e−3	9.2e−3	3.83	Stoc-iALM (5)	4.1e−3	9.5e−3	4.85	Stoc-iALM (5)	1.9e−4	8.3e−3	1.21
Stoc-iALM (6)	5.7e−3	9.9e−3	3.56	Stoc-iALM (6)	4.1e−3	1.00e−2	4.72	Stoc-iALM (6)	2.5e−4	9.2e−3	1.23
Stoc-iALM (7)	5.6e−3	9.5e−3	3.69	Stoc-iALM (7)	4.1e−3	9.8e−3	4.72	Stoc-iALM (7)	2.1e−4	9.0e−3	1.21
Stoc-iALM (8)	5.7e−3	9.3e−3	3.90	Stoc-iALM (8)	4.1e−3	9.9e−3	4.72	Stoc-iALM (8)	2.6e−4	9.8e−3	1.21
Stoc-iALM (9)	5.5e−3	8.6e−3	4.11	Stoc-iALM (9)	4.0e−3	9.9e−3	4.72	Stoc-iALM (9)	2.0e−4	8.8e−3	1.23
Stoc-iALM (10)	5.7e−3	9.9e−3	4.46	Stoc-iALM (10)	4.1e−3	9.8e−3	4.72	Stoc-iALM (10)	2.5e−4	1.00e−2	1.21
IPC	0	9.1e−3	111	IPC	0	9.7e−3	107	IPC	0	2.8e−3	3

[†]#Data by Stoc-iALM are fractional because the subroutine PStorm uses minibatch of data points to compute sample gradients and we record pres and dres after every 50 inner iterations

$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{R}^d} \ & f_0(\mathbf{x}) := \frac{1}{n^+}\sum_{i=1}^{n^+}\phi(\mathbf{x}^\top \mathbf{a}_i^+), \\ \text{s.t.} \ & f_1(\mathbf{x}) := \frac{1}{n^-}\sum_{i=1}^{n^-}\phi(-\mathbf{x}^\top \mathbf{a}_i^-) - \hat{c} \le 0, \end{aligned} \tag{48}$$

where $\{\mathbf{a}_i^+\}_{i=1}^{n^+}$ and $\{\mathbf{a}_i^-\}_{i=1}^{n^-}$ denotes the positive-class samples and negative-class samples of the training data set. The parameter $\hat{c}$ controls the level of the false-positive error. In (48), we set $\phi(\cdot)$ to the sigmoid function: $\phi(u) = 1/(1+\exp(u))$. We use three data sets: *spambase* [6] with $d = 57$ and $(n^+, n^-) = (1813, 2788)$, *madelon* [8] with $d = 500$ and $(n^+, n^-) = (1300, 1300)$, and *gisette* [8] with $d = 2000$ and $(n^+, n^-) = (3500, 3500)$. To make sure the feasibility of the problem, we set the false-positive error parameter to $\hat{c} = 0.2$ for *spambase* and *gisette*, and $\hat{c} = 0.4$ for *madelon*. Following [40], before feeding each data set into the solvers, we preprocess them by first normalizing it feature-wisely to have mean 0 and variance 1, and then scaling each sample to have unit 2-norm.

Similar to Sect. 4.1, we reformulate the inequality constraint in (48) to an equality constraint $f_1(\mathbf{x}) + \mathbf{v} = 0$ for our method Stoc-iALM, where $\mathbf{v} \ge 0$ is enforced. The compared IPC method in [21] is applied directly to (48). The tolerance is again set to $\varepsilon = 10^{-2}$ in all tests. Both methods are terminated if the violation of primal and dual feasibility is below ε, where the dual variable of the IPC is computed by (47) as in Sect. 4.1. For Stoc-iALM, at the k-th outer iteration, we set $\beta_k = 2^k$ and the smoothness constant to $\frac{\beta_k+1}{2}$, and again we set $\tilde{\mathbf{c}}(\mathbf{x}^{k+1}) = \mathbf{c}(\mathbf{x}^{k+1})$ in the $\mathbf{y}$-update (5). In the PStorm subroutine, we set the mini-batch size to 10 for *spambase*, and 30 for *madelon* and *gisette*. The parameter settings of the IPC exactly follow from the code of [21] provided by its authors. Again, we record the primal and dual residuals after every 50 inner iterations in Stoc-iALM, and after every data pass in IPC. Both methods start from a zero vector for each data set, and we perform 10 independent trials by using different random seeds for the proposed method.

Table 2 gives `pres` and `dres` at the produced ε-KKT point and `#data` by each method to produce such a point. Figure 2 plots the curves of the constraint function value and `dres`. Again, we see that our proposed method Stoc-iALM needs significantly fewer data passes to produce a KKT point with the same-level error tolerance.

5 Conclusion

We have presented a stochastic inexact augmented Lagrangian method (Stoc-iALM) for solving nonconvex expectation constrained optimization. To handle nonconvex stochastic iALM subproblems, we apply a momentum-based variance-reduced proximal stochastic gradient method (PStorm) subroutine with a proposed post-processing step. To reach an ε-KKT solution in expectation, we establish an oracle complexity of $O(\varepsilon^{-5})$, which improves over the state-of-the-art complexity of $O(\varepsilon^{-6})$. Numerically,

Table 2 The violation of primal feasibility and the violation to the dual feasibility at the produced ε-KKT point with $\varepsilon = 10^{-2}$, and the number of data passes (shortened by pres, dres and #data[†] respectively) of 10 trials with different random seeds by the proposed Stoc-iALM and the IPC in [21] on solving the Neyman-Pearson classification problem (48) with *spambase*, *madelon*, and *gisette* data sets (from left to right)

Method [*spambase*]	Pres	Dres	#Data	Method [*madelon*]	Pres	Dres	#Data	Method [*gisette*]	Pres	Dres	#Data
Stoc-iALM (1)	0	7.8e−3	18.75	Stoc-iALM (1)	0	1.00e−2	336.08	Stoc-iALM (1)	0	1.00e−2	234.04
Stoc-iALM (2)	0	6.1e−3	39.23	Stoc-iALM (2)	0	1.00e−2	324.77	Stoc-iALM (2)	0	1.00e−2	235.24
Stoc-iALM (3)	0	8.5e−3	13.74	Stoc-iALM (3)	0	1.00e−2	330.08	Stoc-iALM (3)	0	1.00e−2	234.04
Stoc-iALM (4)	0	7.6e−3	16.93	Stoc-iALM (4)	0	1.00e−2	314.15	Stoc-iALM (4)	0	1.00e−2	235.24
Stoc-iALM (5)	0	9.0e−3	11.01	Stoc-iALM (5)	0	1.00e−2	323.15	Stoc-iALM (5)	0	1.00e−2	234.04
Stoc-iALM (6)	0	6.5e−3	21.48	Stoc-iALM (6)	0	1.00e−2	322.92	Stoc-iALM (6)	0	1.00e−2	235.24
Stoc-iALM (7)	0	6.8e−3	28.76	Stoc-iALM (7)	0	1.00e−2	317.85	Stoc-iALM (7)	0	1.00e−2	235.24
Stoc-iALM (8)	0	9.9e−3	22.85	Stoc-iALM (8)	0	1.00e−2	332.38	Stoc-iALM (8)	0	1.00e−2	235.24
Stoc-iALM (9)	0	9.3e−3	11.47	Stoc-iALM (9)	0	1.00e−2	326.38	Stoc-iALM (9)	0	1.00e−2	234.04
Stoc-iALM (10)	0	7.1e−3	16.47	Stoc-iALM (10)	0	1.00e−2	341.15	Stoc-iALM (10)	0	1.00e−2	234.04
IPC	0	9.7e−3	37	IPC	0	1.00e−2	1804	IPC	0	1.00e−2	650

[†]#Data by Stoc-iALM are fractional because the subroutine PStorm uses minibatch of data points to compute sample gradients and we record pres and dres after every 50 inner iterations

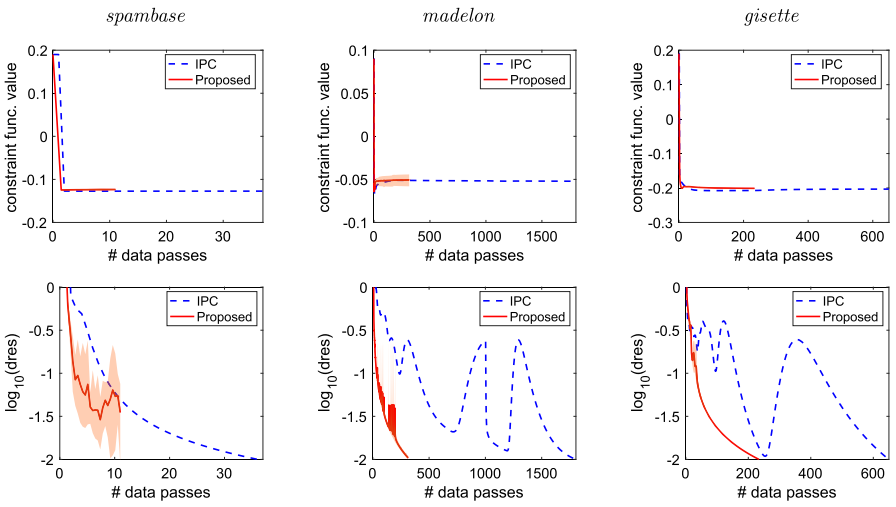

Fig. 2 Mean curves and standard deviation (by shadow area†) of 10 different trials by the proposed Stoc-iALM and the curves by the IPC in [21] on solving the Neyman-Pearson classification problem (48) with *spambase*, *madelon*, and *gisette* data sets (from left to right). † Some shadow areas are invisible because the deviations are too small

we have demonstrated that the proposed Stoc-iALM can significantly outperform one state-of-the-art method.

Acknowledgements The authors would like to thank two anonymous reviewers for their constructive comments and suggestions. This work is partly supported by NSF Grants DMS-2053493 and DMS-2208394 and the ONR award N00014-22-1-2573, and also by the Rensselaer-IBM AI Research Collaboration, part of the IBM AI Horizons Network.

Data availability statements The data used for numerical tests are from UCI repository at https://archive.ics.uci.edu/ml/index.php and LIBSVM at https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/.

Declarations

Conflict of interest There is no potential conflict of interest.

A regularity condition for ball-constrained fairness and Neyman–Pearson problems

In this section, we show that the regularity condition in (7) holds for the nonconvex fairness problem in (46) and the Neyman-Pearson classification problem in (48), when a ball constraint is imposed and the input data satisfies certain conditions. We consider the two problems together in the following form:

$$\min_{\mathbf{x},s} f_0(\mathbf{x}),\ \text{s.t. } \hat{f}_1(\mathbf{x},s) := f_1(\mathbf{x}) + s = 0, s \geq 0, \mathbf{x} \in \mathcal{X} := \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| \leq \lambda\}, \tag{49}$$

where $\lambda > 0$, and f_0 and f_1 are the functions defined in (46) and (48) respectively for the fairness problem and the Neyman-Pearson classification problem. The slack variable s is used to reformulate (46) and (48) into equality-constrained problems. We did not include the constraint $\mathbf{x} \in \mathcal{X}$ in our experiments but the generated iterate sequence remained bounded.

Let $\mathcal{N}_{\mathcal{X}}(\mathbf{x})$ be the normal cone of $\mathcal{X}$ at $\mathbf{x} \in \mathcal{X}$ and $\mathcal{N}_+(s)$ the normal cone of $\mathbb{R}_+$ at $s \geq 0$. Then the regularity condition in (7) for the problem (49) becomes: there exists $\nu > 0$ such that

$$\nu^2(f_1(\mathbf{x})+s)^2 \leq \operatorname{dist}\left(-(f_1(\mathbf{x})+s)\begin{bmatrix}\nabla f_1(\mathbf{x})\\ 1\end{bmatrix},\ \mathcal{N}_{\mathcal{X}}(\mathbf{x})\otimes\mathcal{N}_+(s)\right)^2, \forall\, \mathbf{x}\in\mathcal{X}, s\geq 0. \quad (50)$$

Notice that $\mathcal{N}_{\mathcal{X}}(\mathbf{x}) = \{\mathbf{0}\}$ if $\|\mathbf{x}\| < \lambda$ and $\mathcal{N}_{\mathcal{X}}(\mathbf{x}) = \{\alpha\mathbf{x} : \alpha \geq 0\}$ if $\|\mathbf{x}\| = \lambda$. Also, $\mathcal{N}_+(s) = \{0\}$ if $s > 0$ and $\mathcal{N}_+(s) = \mathbb{R}_-$ if $s = 0$. Hence, for any $(\mathbf{x}, s) \in \mathcal{X} \otimes \mathbb{R}_+$,

$$\begin{aligned}
&\operatorname{dist}\left(-(f_1(\mathbf{x})+s)\begin{bmatrix}\nabla f_1(\mathbf{x})\\ 1\end{bmatrix},\ \mathcal{N}_{\mathcal{X}}(\mathbf{x})\otimes\mathcal{N}_+(s)\right)^2\\
&= \begin{cases}
(f_1(\mathbf{x})+s)^2 + \operatorname{dist}\big(-(f_1(\mathbf{x})+s)\nabla f_1(\mathbf{x}),\ \mathcal{N}_{\mathcal{X}}(\mathbf{x})\big)^2, & \text{if } s>0\\
\min\big(f_1(\mathbf{x}),0\big)^2 + |f_1(\mathbf{x})|^2\|\nabla f_1(\mathbf{x})\|^2, & \text{if } s=0,\ \|\mathbf{x}\|<\lambda\\
\min\big(f_1(\mathbf{x}),0\big)^2 + \min_{\alpha\geq 0}\|f_1(\mathbf{x})\nabla f_1(\mathbf{x})+\alpha\mathbf{x}\|^2, & \text{if } s=0,\ \|\mathbf{x}\|=\lambda.
\end{cases}
\end{aligned} \quad (51)$$

From (51), we can easily have that when $s > 0$ or when $s = 0$ and $f_1(\mathbf{x}) \leq 0$, it holds

$$(f_1(\mathbf{x})+s)^2 \leq \operatorname{dist}\left(-(f_1(\mathbf{x})+s)\begin{bmatrix}\nabla f_1(\mathbf{x})\\ 1\end{bmatrix},\ \mathcal{N}_{\mathcal{X}}(\mathbf{x})\otimes\mathcal{N}_+(s)\right)^2.$$

Thus we only need to show the regularity condition at $(\mathbf{x}, 0)$ with $\mathbf{x} \in \mathcal{X}$ such that $f_1(\mathbf{x}) > 0$. We make the following assumption about the data involved in (46) and (48).

Assumption 6 The feature vectors in (46) and (48) satisfy:

(i) In (46), $\|\mathbf{a}\| = q, \forall\, \mathbf{a} \in S$ for some $q > 0$ and $\langle \mathbf{a}_1, \mathbf{a}_2\rangle \geq 0$ for any $\mathbf{a}_1, \mathbf{a}_2 \in S$. In addition,

$$\frac{e^{\lambda q}}{(1+e^{\lambda q})^2}\sqrt{\sum_{\mathbf{a}_1,\mathbf{a}_2\in S\setminus S_{\min}}\mathbf{a}_1^\top\mathbf{a}_2} > \frac{1-c}{4c}\sqrt{\sum_{\mathbf{a}_1,\mathbf{a}_2\in S_{\min}}\mathbf{a}_1^\top\mathbf{a}_2}. \quad (52)$$

(ii) In (48), $\|\mathbf{a}_i^-\| = q, \forall\, i$ for some $q > 0$, and $\langle \mathbf{a}_i^-, \mathbf{a}_j^-\rangle \geq 0$ for any i, j.

The above assumption will hold if each data point is first normalized and then appended by 1 at the end, which is equivalent to having an intercept term in the model, and in addition, for (46) the minority group $S_{\min}$ is only a small fraction of S.

Claim A1 *Under Assumption 6, let f_1 be given in (46) or (48). Then $\nu_1 := \min_{\|\mathbf{x}\|\le\lambda} \|\nabla f_1(\mathbf{x})\| > 0$.*

Proof We first prove the claim for (48), for which case,

$$\nabla f_1(\mathbf{x}) = -\frac{1}{n^-}\sum_{i=1}^{n^-}\phi'(-\mathbf{x}^\top \mathbf{a}_i^-)\mathbf{a}_i^-. \tag{53}$$

Suppose $\nu_1 = \|\nabla f_1(\tilde{\mathbf{x}})\|$, i.e., the minimum is reached at $\tilde{\mathbf{x}}$. Notice $\phi'(u) = -\frac{e^u}{(1+e^u)^2} < 0$. Thus

$$\nu_1^2 = \frac{1}{(n^-)^2}\sum_{i,j=1}^{n^-}\phi'(-\tilde{\mathbf{x}}^\top \mathbf{a}_i^-)\phi'(-\tilde{\mathbf{x}}^\top \mathbf{a}_j^-)\langle \mathbf{a}_i^-, \mathbf{a}_j^-\rangle \ge \frac{1}{(n^-)^2}\sum_{i=1}^{n^-}\left[\phi'(-\tilde{\mathbf{x}}^\top \mathbf{a}_i^-)\right]^2\|\mathbf{a}_i^-\|^2,$$

where the inequality follows from $\langle \mathbf{a}_i^-, \mathbf{a}_j^-\rangle \ge 0$ for all i, j. Hence, $\nu_1 > 0$ must hold by Assumption 6(ii).

Next we prove the claim for (46). When f_1 is the function in (46), it holds

$$\nabla f_1(\mathbf{x}) = c\sum_{\mathbf{a}\in S\setminus S_{\min}}\sigma'(\mathbf{a}^\top\mathbf{x})\mathbf{a} - (1-c)\sum_{\mathbf{a}\in S_{\min}}\sigma'(\mathbf{a}^\top\mathbf{x})\mathbf{a}.$$

Again, suppose $\nu_1 = \|\nabla f_1(\tilde{\mathbf{x}})\|$. Notice $\sigma'(u) = \frac{e^u}{(1+e^u)^2}$ is decreasing on $[0, +\infty)$ and increasing on $(-\infty, 0]$. Also, by Assumption 6(i) and $\|\tilde{\mathbf{x}}\| \le \lambda$, we have $|\mathbf{a}^\top\tilde{\mathbf{x}}| \le q\lambda$. Hence, $\frac{e^{q\lambda}}{(1+e^{q\lambda})^2} \le \sigma'(\mathbf{a}^\top\tilde{\mathbf{x}}) \le \frac{1}{4}$. Thus

$$\begin{aligned}\left\|c\sum_{\mathbf{a}\in S\setminus S_{\min}}\sigma'(\mathbf{a}^\top\tilde{\mathbf{x}})\mathbf{a}\right\|^2 &= c^2\sum_{\mathbf{a}_1,\mathbf{a}_2\in S\setminus S_{\min}}\sigma'(\mathbf{a}_1^\top\tilde{\mathbf{x}})\sigma'(\mathbf{a}_2^\top\tilde{\mathbf{x}})\mathbf{a}_1^\top\mathbf{a}_2\\ &\ge \frac{c^2e^{2q\lambda}}{(1+e^{q\lambda})^4}\sum_{\mathbf{a}_1,\mathbf{a}_2\in S\setminus S_{\min}}\mathbf{a}_1^\top\mathbf{a}_2,\end{aligned}$$

and

$$\left\|(1-c)\sum_{\mathbf{a}\in S_{\min}}\sigma'(\mathbf{a}^\top\mathbf{x})\mathbf{a}\right\|^2 \le \frac{(1-c)^2}{16}\sum_{\mathbf{a}_1,\mathbf{a}_2\in S_{\min}}\mathbf{a}_1^\top\mathbf{a}_2.$$

By the triangle inequality, it holds that

$$\|\nabla f_1(\mathbf{x})\| \ge c\left\|\sum_{\mathbf{a}\in S\setminus S_{\min}}\sigma'(\mathbf{a}^\top\mathbf{x})\mathbf{a}\right\| - (1-c)\left\|\sum_{\mathbf{a}\in S_{\min}}\sigma'(\mathbf{a}^\top\mathbf{x})\mathbf{a}\right\|.$$

Therefore, from (52), we obtain $\nu_1 > 0$ and complete the proof. □

Claim A2 *Suppose Assumption 6*(ii) *holds and in addition, the origin is a feasible point of* (48)*, i.e.,* $f_1(\mathbf{0}) \leq 0$*. Then it holds*

$$\nu_2 := \min_{\alpha \geq 0, \mathbf{x}} \left\{ \|f_1(\mathbf{x})\nabla f_1(\mathbf{x}) + \alpha \mathbf{x}\| : \|\mathbf{x}\| = \lambda,\ f_1(\mathbf{x}) \geq 0 \right\} > 0. \tag{54}$$

Proof Suppose that the minimum in (54) is reached at $\tilde{\mathbf{x}}$, i.e., $\|\tilde{\mathbf{x}}\| = \lambda$, $f_1(\tilde{\mathbf{x}}) \geq 0$, and

$$\nu_2 = \min_{\alpha \geq 0} \|f_1(\tilde{\mathbf{x}})\nabla f_1(\tilde{\mathbf{x}}) + \alpha \tilde{\mathbf{x}}\|.$$

If $\nu_2 = 0$, then we must have $\tilde{\mathbf{x}} = -\lambda \frac{\nabla f_1(\tilde{\mathbf{x}})}{\|\nabla f_1(\tilde{\mathbf{x}})\|}$ and the optimal $\alpha = \frac{\|\nabla f_1(\tilde{\mathbf{x}})\|}{\lambda}$. By (53), we have

$$-\tilde{\mathbf{x}}^\top \mathbf{a}_j^- = -\frac{\lambda}{\|\nabla f_1(\tilde{\mathbf{x}})\|} \frac{1}{n^-} \sum_{i=1}^{n^-} \phi'(-\mathbf{x}^\top \mathbf{a}_i^-) \langle \mathbf{a}_i^-, \mathbf{a}_j^- \rangle > 0,$$

where the inequality follows from $\phi'(u) < 0, \forall\, u$ and Assumption 6(ii). Now notice that $\phi(u)$ is an decreasing function. We have $f_1(\tilde{\mathbf{x}}) < f_1(\mathbf{0}) \leq 0$, which contradicts to $f_1(\tilde{\mathbf{x}}) \geq 0$. Therefore, we must have $\nu_2 > 0$ and thus complete the proof. □

By Claims A1 and A2, we immediately obtain the theorem below.

Theorem 4 *Suppose Assumption 6 holds and in addition, the origin is feasible in* (48)*. Then there must exist a constant* $\nu > 0$ *such that if* f_1 *is given in* (46)*,*

$$\nu |\hat{f}_1(\mathbf{x}, s)| \leq \operatorname{dist}\left(-\hat{f}_1(\mathbf{x}, s)\nabla \hat{f}_1(\mathbf{x}, s),\ \mathcal{N}_{\mathcal{X}}(\mathbf{x}) \otimes \mathcal{N}_+(s) \right), \forall\, \mathbf{x} \in \operatorname{int}(\mathcal{X}), s \geq 0, \tag{55}$$

and if f_1 *is given in* (48)*,*

$$\nu |\hat{f}_1(\mathbf{x}, s)| \leq \operatorname{dist}\left(-\hat{f}_1(\mathbf{x}, s)\nabla \hat{f}_1(\mathbf{x}, s),\ \mathcal{N}_{\mathcal{X}}(\mathbf{x}) \otimes \mathcal{N}_+(s) \right), \forall\, \mathbf{x} \in \mathcal{X}, s \geq 0, \tag{56}$$

where $\hat{f}_1$ *and* $\mathcal{X}$ *are defined in* (49)*.*

Proof From (50) and (51), we obtain (55) with $\nu = \min\{1, \nu_1\}$, where ν_1 is defined in Claim A1, and we obtain (56) with $\nu = \min\{1, \nu_1, \nu_2\}$, where ν_1 and ν_2 are defined in Claims A1 and A2. □

Remark 9 From Theorem 4, we see that the regularity condition will hold for the ball constrained version of (48) under a certain data preprocessing and an origin feasibility condition. It can almost hold for a ball constrained version of (46), except for the points on the sphere of the constraint ball. For the tested instances of (46) and (48) without a ball constraint, we checked the regularity condition at the iterates (because we actually only need the condition at the generated iterates). We found that for (46), the introduced slack variable s would always be positive during the iterations, and for

(48), the iterates became feasible after just a few iterations; see Fig. 2. Hence, for the instances tested in our experiments, the regularity condition in (7) holds with $\nu = 1$ for $\mathbf{x}$ being a generated iterate.

References

1. Arjevani, Y., Carmon, Y., Duchi, J.C., Foster, D.J., Srebro, N., Woodworth, B.: Lower bounds for non-convex stochastic optimization. Math. Program. **199**, 165–214 (2023)
2. Boob, D., Deng, Q., Lan, G.: Stochastic first-order methods for convex and nonconvex functional constrained optimization. Math. Program. **197**, 215–279 (2023)
3. Cartis, C., Gould, N.I., Toint, P.L.: On the evaluation complexity of composite function minimization with applications to nonconvex nonlinear programming. SIAM J. Optim. **21**(4), 1721–1739 (2011)
4. Chang, C.-C., Lin, C.-J.: Libsvm: a library for support vector machines. ACM Transa. Intell. Syst. Technol. (TIST) **2**(3), 1–27 (2011)
5. Cutkosky, A., Orabona, F.: Momentum-based variance reduction in non-convex sgd. In: Advances in Neural Information Processing Systems, vol. 32 (2019)
6. Dua, D., Graff, C.: UCI machine learning repository (2017)
7. Fang, C., Li, C.J., Lin, Z., Zhang, T.: Spider: near-optimal non-convex optimization via stochastic path-integrated differential estimator. In: Advances in Neural Information Processing Systems, vol. 31 (2018)
8. Guyon, I., Gunn, S., Ben-Hur, A., Dror, G.: Result analysis of the nips 2003 feature selection challenge. In: Advances in neural information processing systems, vol. 17 (2004)
9. Hestenes, M.R.: Multiplier and gradient methods. J. Optim. Theory Appl. **4**(5), 303–320 (1969)
10. Huang, F., Gao, S., Pei, J., Huang, H.: Accelerated zeroth-order and first-order momentum methods from mini to minimax optimization. J. Mach. Learn. Res. **23**, 1–36 (2022)
11. Jin, L., Wang, X.: A stochastic primal-dual method for a class of nonconvex constrained optimization. Comput. Optim. Appl. **83**(1), 143–180 (2022)
12. Lan, G., Monteiro, R.D.: Iteration-complexity of first-order augmented Lagrangian methods for convex programming. Math. Program. **155**(1–2), 511–547 (2016)
13. Lan, G., Zhou, Z.: Algorithms for stochastic optimization with function or expectation constraints. Comput. Optim. Appl. **76**(2), 461–498 (2020)
14. Li, F., Qu, Z.: An inexact proximal augmented Lagrangian framework with arbitrary linearly convergent inner solver for composite convex optimization. Math. Program. Comput. **13**(3), 583–644 (2021)
15. Li, Z., Chen, P.-Y., Liu, S., Lu, S., Xu, Y.: Rate-improved inexact augmented Lagrangian method for constrained nonconvex optimization. In: International Conference on Artificial Intelligence and Statistics, pp. 2170–2178. PMLR (2021)
16. Li, Z., Chen, P.-Y., Liu, S., Lu, S., Xu, Y.: Zeroth-order optimization for composite problems with functional constraints. Proc. AAAI Conf. Artif. Intell. **36**, 7453–7461 (2022)
17. Li, Z., Xu, Y.: Augmented Lagrangian-based first-order methods for convex-constrained programs with weakly convex objective. Informs J. Optim. **3**(4), 373–397 (2021)
18. Lin, Q., Ma, R., Xu, Y.: Complexity of an inexact proximal-point penalty method for constrained smooth non-convex optimization. Comput. Optim. Appl. **82**(1), 175–224 (2022)
19. Lu, S.: A single-loop gradient descent and perturbed ascent algorithm for nonconvex functional constrained optimization. In: International Conference on Machine Learning, pp. 14315–14357. PMLR (2022)
20. Luo, L., Ye, H., Huang, Z., Zhang, T.: Stochastic recursive gradient descent ascent for stochastic nonconvex-strongly-concave minimax problems. Adv. Neural. Inf. Process. Syst. **33**, 20566–20577 (2020)
21. Ma, R., Lin, Q., Yang, T.: Proximally constrained methods for weakly convex optimization with weakly convex constraints. arXiv:1908.01871 (2019)
22. Ma, R., Lin, Q., Yang, T.: Quadratically regularized subgradient methods for weakly convex optimization with weakly convex constraints. In: International Conference on Machine Learning, pp. 6554–6564. PMLR (2020)

23. Melo, J.G., Monteiro, R.D., Wang, H.: Iteration-complexity of an inexact proximal accelerated augmented Lagrangian method for solving linearly constrained smooth nonconvex composite optimization problems. Optimization Online (2020)
24. Necoara, I., Nedelcu, V.: Rate analysis of inexact dual first-order methods application to dual decomposition. IEEE Trans. Autom. Control **59**(5), 1232–1243 (2014)
25. Nedelcu, V., Necoara, I., Tran-Dinh, Q.: Computational complexity of inexact gradient augmented Lagrangian methods: application to constrained mpc. SIAM J. Control. Optim. **52**(5), 3109–3134 (2014)
26. Neyman, J., Pearson, E.S.: Containing papers of a mathematical or physical character. Ix. on the problem of the most efficient tests of statistical hypotheses. Philos. Trans. Roy. Soc. Lond. Ser. A **231**(694–706), 289–337 (1933)
27. Ouyang, Y., Chen, Y., Lan, G., Pasiliao, E., Jr.: An accelerated linearized alternating direction method of multipliers. SIAM J. Imag. Sci. **8**(1), 644–681 (2015)
28. Rakhlin, A., Shamir, O., Sridharan, K.: Making gradient descent optimal for strongly convex stochastic optimization. In: Proceedings of the 29th International Conference on International Conference on Machine Learning, pp. 1571–1578 (2012)
29. Rigollet, P., Tong, X.: Neyman–Pearson classification, convexity and stochastic constraints. J. Mach. Learn. Res. **12**(Oct), 2831–2855 (2011)
30. Rockafellar, R.T.: A dual approach to solving nonlinear programming problems by unconstrained optimization. Math. Program. **5**(1), 354–373 (1973)
31. Sahin, M.F., Alacaoglu, A., Latorre, F., Cevher, V. et al:. An inexact augmented Lagrangian framework for nonconvex optimization with nonlinear constraints. In: Advances in Neural Information Processing Systems, pp. 13943–13955 (2019)
32. Shi, Q., Wang, X., Wang, H.: A momentum-based linearized augmented Lagrangian method for nonconvex constrained stochastic optimization (2022)
33. Tran Dinh, Q., Liu, D., Nguyen, L.: Hybrid variance-reduced sgd algorithms for minimax problems with nonconvex-linear function. Adv. Neural. Inf. Process. Syst. **33**, 11096–11107 (2020)
34. Tran-Dinh, Q., Pham, N.H., Phan, D.T., Nguyen, L.M.: A hybrid stochastic optimization framework for composite nonconvex optimization. Math. Program. **191**(2), 1005–1071 (2022)
35. Wang, X., Ma, S., Yuan, Y.-X.: Penalty methods with stochastic approximation for stochastic nonlinear programming. Math. Comput. **86**(306), 1793–1820 (2017)
36. Xu, Yangyang: Primal-dual stochastic gradient method for convex programs with many functional constraints. SIAM J. Optim. **30**(2), 1664–1692 (2020). https://doi.org/10.1137/18M1229869
37. Xu, Y.: First-order methods for constrained convex programming based on linearized augmented Lagrangian function. Informs J. Optim. **3**(1), 89–117 (2021)
38. Xu, Y.: Iteration complexity of inexact augmented Lagrangian methods for constrained convex programming. Math. Program. **185**(1), 199–244 (2021)
39. Xu, Y., Xu, Y.: Momentum-based variance-reduced proximal stochastic gradient method for composite nonconvex stochastic optimization. J. Optim. Theory Appl. **196**(1), 266–297 (2023)
40. Yan, Y., Xu, Y.: Adaptive primal-dual stochastic gradient method for expectation-constrained convex stochastic programs. Math. Program. Comput. **14**, 319–363 (2022)
41. Yu, H., Neely, M., Wei, X.: Online convex optimization with stochastic constraints. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
42. Zhang, J., Luo, Z.-Q.: A global dual error bound and its application to the analysis of linearly constrained nonconvex optimization. SIAM J. Optim. **32**(3), 2319–2346 (2022)