Quadratically Regularized Subgradient Methods for Weakly Convex Optimization with Weakly Convex Constraints

Runchao Ma [1] **Qihang Lin** [1] **Tianbao Yang** [2]

Abstract

Optimization models with non-convex constraints arise in many tasks in machine learning, e.g., learning with fairness constraints or Neyman-Pearson classification with non-convex loss. Although many efficient methods have been developed with theoretical convergence guarantees for non-convex unconstrained problems, it remains a challenge to design provably efficient algorithms for problems with non-convex functional constraints. This paper proposes a class of subgradient methods for constrained optimization where the objective function and the constraint functions are weakly convex and nonsmooth. Our methods solve a sequence of strongly convex subproblems, where a quadratic regularization term is added to both the objective function and each constraint function. Each subproblem can be solved by various algorithms for strongly convex optimization. Under a uniform Slater's condition, we establish the computation complexities of our methods for finding a nearly stationary point.

1. Introduction

Continuous optimization models with nonlinear constraints have been widely used in many disciplines including machine learning, statistics, and data mining with many real-world applications. A general optimization problem with inequality constraints is formulated as

$$\begin{aligned} f^* \equiv \min_{\mathbf{x}\in\mathcal{X}} \{f(\mathbf{x}) \equiv f_0(\mathbf{x})\} \\ \text{s.t.} \quad g(\mathbf{x}) \equiv \max_{i=1,\dots,m} f_i(\mathbf{x}) \le 0 \end{aligned} \tag{1}$$

Here, we assume that $\mathcal{X} \subset \mathbb{R}^d$ is a compact convex set that allows for a simple projection and f_i for $i = 0, \dots, m$ are weakly-convex (potentially non-smooth) functions. A solution $\bar{\mathbf{x}} \in \mathcal{X}$ is *ε-optimal* if $f(\bar{\mathbf{x}}) - f^* \le \varepsilon$ and *ε-feasible* if $\bar{\mathbf{x}} \in \mathcal{X}$ and $g(\bar{\mathbf{x}}) \le \varepsilon$. Many optimization models in machine learning contain nonlinear constraints. Examples include Neyman-Pearson classification (Rigollet & Tong, 2011) and learning with dataset constraints (Goh et al., 2016) (e.g. fairness constraints and churn rate constraints).

Optimization problems with a convex objective function and convex constraints have been well studied in literature with many efficient algorithms and their theoretical complexity developed (Bertsekas, 2014; 1999; Nocedal & Wright, 2006). However, the parallel development for optimization with non-convex objective functions and non-convex constraints, especially for theoretically provable algorithms, remains limited, restricting the practices of statistical modeling and decision making in many disciplines. It is well-known that finding a global minimizer for a general non-convex function without any constraints has been intractable (Sahni, 1974). The difficulty will increase when constraints appear and will increase even further when those constraints are non-convex.

Therefore, when designing an algorithm for (1) with non-convex objective and constraint functions, the first question to be addressed is what kind of solutions can the algorithm guarantees and what complexity the algorithm has in order to find such solutions. In the recent studies on unconstrained or simply constrained[1] non-convex minimization (Davis & Drusvyatskiy, 2018c;a;b; Davis & Grimmer, 2017; Drusvyatskiy, 2017; Drusvyatskiy & Paquette, 2018; Ghadimi & Lan, 2013; 2016; Lan & Yang, 2018; Paquette et al., 2018; Reddi et al., 2016a;b), algorithms have been proposed to find a *nearly stationary point*, which is a feasible solution close to another feasible solution where the subdifferential of the objective function almost contains zero. However, these methods and analysis cannot be applied to (1) as they require the exact projection to the feasible set which is hard to perform for (1) due to the functional constraints. To address

[1]Department of Business Analytics, University of Iowa, Iowa City, IA, USA [2]Department of Computer Science, University of Iowa, Iowa City, IA, USA. Correspondence to: Qihang Lin <qihang-lin@uiowa.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

[1]Here, being simply constrained means the feasible set is a simple set, e.g., a box or a ball, that allows for a closed-form for the projection mapping.

this issue, in this work, we propose a class of first-order methods for (1) where the objective function and the constraint functions are all weakly convex. Our methods solve a sequence of strongly convex subproblems where, different from the traditional proximal-point method, a quadratic regularization term is also added to each constraint function instead of just the objective function. Each subproblem can be solved by an algorithm for strongly convex optimization. Under a uniform Slater's condition, we establish the complexities of our methods for finding a nearly stationary point. We will discuss some applications of (1) in machine learning next.

1.1. Optimization Problems in Machine Learning with Nonlinear Constraints

Multi-class Neyman-Pearson Classification: In multi-class classification, there exist K classes of data, denoted by ξ_k for $k = 1, 2, \ldots, K$, each of which has its own distribution. To classify each data into one of the K classes, one can rely on K linear models $\mathbf{x}_k$, $k = 1, 2, \ldots, K$ and predict the class of a data point ξ as $\arg\max_{k=1,2,\ldots,K} \mathbf{x}_k^\top \xi$. To achieve a high classification accuracy, we would like the value $\mathbf{x}_k^\top \xi_k - \mathbf{x}_l^\top \xi_k$ with $k \neq l$ to be positively large (Weston & Watkins, 1998; Crammer & Singer, 2002), which can be achieved by minimizing the expected loss $\mathbb{E}\phi(\mathbf{x}_k^\top \xi_k - \mathbf{x}_l^\top \xi_k)$, where ϕ is a non-increasing potentially non-convex loss function and $\mathbb{E}$ is the expectation taken over ξ_k. When training these K linear models, one can prioritize minimizing the loss on class 1 while control the losses on other classes by solving

$$\min_{\|\mathbf{x}_k\|_2 \leq \lambda, k=1,\ldots,K} \sum_{l \neq 1} \mathbb{E}[\phi(\mathbf{x}_1^\top \xi_1 - \mathbf{x}_l^\top \xi_1)]$$
$$s.t. \sum_{l \neq k} \mathbb{E}[\phi(\mathbf{x}_k^\top \xi_k - \mathbf{x}_l^\top \xi_k)] \leq r_k \quad k = 2, 3, \ldots, K,$$

where r_k controls the loss for class k and λ is a regularization parameter. When ξ follows the empirical distribution over a finite dataset, the expectations above are essentially sample averages so that this problem becomes a deterministic optimization problem.

Learning Data-Driven Constraints: Problem (1) also covers many machine learning models with data-driven constraints (Goh et al., 2016). The examples include the constraints that impose conditions on the coverage rates, churn rates, or fairness of a predictive model. More details can be found in (Goh et al., 2016). Here, we focus on learning a classifier with parity-based fairness constraints (Goh et al., 2016; Zafar et al., 2015; 2017). Suppose $(\mathbf{a}, b)$ is a point from a distribution $\mathcal{D}$ where $b \in \{1, -1\}$ is the label. Let $\mathcal{D}_M$ and $\mathcal{D}_F$ be two different distributions of points (not necessarily labeled), e.g., $\mathcal{D}_M$ and $\mathcal{D}_F$ may represent the male and female groups. The training of a classifier with fairness constraints can be formulated as

$$\min_{\|\mathbf{x}\|_2 \leq \lambda} \mathbb{E}_{(\mathbf{a},b)\sim\mathcal{D}}[\phi(-b\mathbf{a}^\top \mathbf{x})]$$
$$s.t. \ \mathbb{E}_{\mathbf{a}\sim\mathcal{D}_M}[\sigma(\mathbf{a}^\top \mathbf{x})] + \beta \mathbb{E}_{\mathbf{a}\sim\mathcal{D}_F}[\sigma(-\mathbf{a}^\top \mathbf{x})] \leq r$$
$$\mathbb{E}_{\mathbf{a}\sim\mathcal{D}_F}[\sigma(\mathbf{a}^\top \mathbf{x})] + \beta \mathbb{E}_{\mathbf{a}\sim\mathcal{D}_M}[\sigma(-\mathbf{a}^\top \mathbf{x})] \leq r$$

where ϕ is a non-increasing potentially non-convex loss function, $\sigma = \frac{\exp(z)}{1+\exp(z)}$, λ is a regularization parameter, β is a positive balance parameter and r is a constraint parameter. The objective function is the training loss of $\mathbf{x}$. The terms $\sigma(\mathbf{a}^\top \mathbf{x})$ and $\sigma(-\mathbf{a}^\top \mathbf{x})$ represent the predicted probabilities of $\mathbf{a}$ being in the positive and the negative class, respectively. The left hand side of the first constraint will be large if the model $\mathbf{x}$ is very "unfair" in the sense that it makes $\mathbf{a}^\top \mathbf{x}$ very negative for most of $\mathbf{a}$ from $\mathcal{D}_M$ but very positive for most of $\mathbf{a}$ from $\mathcal{D}_F$. The second constraint can be interpreted similarly. Choosing appropriate r forces the left hand sides of both constraints low so that the obtained model will be fair to both $\mathcal{D}_M$ and $\mathcal{D}_F$.

1.2. Contributions

We summarize our contributions as follows.

- We propose a class of algorithms (Algorithm 1) for (1) when all f_i are weakly convex. This method approximately solves a strongly convex subproblem (9) in each main iteration with precision $O(\epsilon^2)$ using a suitable first-order method. We show that our method finds a nearly ϵ-stationary point (Definition 1) for (1) in $O(\frac{1}{\epsilon^2})$ main iterations.
- When each f_i is a deterministic function, we develop a new variant of the switching subgradient method (Polyak, 1967) to solve (9). We show that the complexity of Algorithm 1 for finding a nearly ϵ-stationary point is $O(\frac{1}{\epsilon^4})$.
- When each f_i is given as an expectation of a stochastic function, we directly use the stochastic subgradient method by (Yu et al., 2017) to solve (9). We show that the complexity of Algorithm 1 for finding a nearly ϵ-stationary point is $\tilde{O}(\frac{1}{\epsilon^6})$.[2]

2. Related Work

There has been growing interest in first-order algorithms for non-convex minimization problems with no constraints or simple constraints in both stochastic and deterministic setting. Initially, the research in this direction mainly focus on the problem with a smooth objective function (Ghadimi & Lan, 2013; Yang et al., 2016; Ghadimi & Lan, 2016; Reddi et al., 2016b;a; Lan & Yang, 2018; Allen-Zhu, 2017;

[2] In this paper, $\tilde{O}(\cdot)$ suppresses all logarithmic factors of ϵ.

Allen-Zhu & Hazan, 2016; Lacoste-Julien, 2016). Recently, more studies have been developed on the algorithms and theories for non-convex minimization problems with non-smooth objective functions after assuming the objective function is weakly convex (Davis & Drusvyatskiy, 2018a;c; Drusvyatskiy & Paquette, 2018; Davis & Grimmer, 2017; Chen et al., 2018; Zhang & He, 2018). These works tackle the non-smoothness of objective function by introducing the Moreau envelope of objective function and analyze the complexity of finding a nearly stationary point. However, these methods are not directly applicable to (1) because of the functional constraints.

The studies on convex optimization with functional constraints have a long history (Bertsekas, 2014; 1999; Nocedal & Wright, 2006; Ruszczyński, 2006, and references therein). The recent development in the first-order methods for convex optimization with convex constraints include (Mahdavi et al., 2012; Zhang et al., 2013; Chen et al., 2016; Yang et al., 2017; Wei et al., 2018; Xu, 2018; 2017b;a; Yu et al., 2017; Lin et al., 2018c;b; Bayandina et al., 2018; Fercoq et al., 2019) for deterministic constraints and (Lan & Zhou, 2016; Yu & Neely, 2017) for stochastic constraints. (Wei & Neely, 2018) propose a primal-dual Frank-Wolfe method for (1) with non-convex f_0 but linear f_i for $i = 1, 2, \ldots, m$. Different from these works, this paper study the problems where the objective function and the constaints are all non-convex. (Sahin et al., 2019) propose an inexact augmented Lagrangian method for (1) with non-convex f_0 and nonlinear equality constraints. A complexity $\tilde{O}(\frac{1}{\epsilon^3})$ is claimed in Corollary 4.2 in (Sahin et al., 2019), but there is an error in its proof. The authors claimed the complexity of solving their subproblem is $\tilde{O}(\frac{\lambda_{\beta_k}^2 \rho^2}{\epsilon_{k+1}})$ but it should be $\tilde{O}(\frac{\lambda_{\beta_k}^2 \rho^2}{\epsilon_{k+1}^2})$. (See (Sahin et al., 2019) for the definitions of λ_{β_k}, ρ, and ϵ_{k+1}). After correcting this error, following the same proof they used gives a total complexity of $\tilde{O}(\frac{1}{\epsilon^4})$. (Nguyen, 2018) study the problem (1) with non-convex f_i for $i = 0, \ldots, m$, but only from the perspective of optimality conditions. Optimization algorithms and convergence analysis are not considered in their work.

We realize a paper by Boob et al. (Boob et al., 2019) was posted online simultaneously as our paper. The main algorithms (Algorithm 1 and 2 in (Boob et al., 2019)) they proposed are similar to our Algorithm 1 in the sense that a similar subproblem (9) is solved in each main iteration. The main difference between our paper and (Boob et al., 2019) is the assumptions made to ensure the boundness of the dual variables of subproblem (9), which is critical to the convergence analysis. The authors of (Boob et al., 2019) establish convergence result under various constraint qualification conditions including, Mangasarian-Fromovitz constraint qualification (MFCQ), strong MFCQ, and strong feasibility while we only consider a uniform Slater's condition (Assumption 1B). Strong feasibility condition is stronger than our uniform Slater's condition but, on the other hand, is easier to verify. The relative strength between (strong) MFCQ and the uniform Slater's condition is unknown. In addition, we focus on the cases where the objective and constraint functions are either all deterministic or all stochastic while (Boob et al., 2019) considers an additional case where only the objective is stochastic. In the stochastic case, we require the stochastic gradients to be bounded (Assumption 2) while (Boob et al., 2019) assume the boundness of the second moment of the stochastic gradients. The complexities of our methods and theirs for finding an ϵ-nearly stationary point are the same in the dependency on ϵ in the dominating terms. Their complexity is more general in the sense that it involves non-dominating terms that depend on the smoothness parameters of the smooth components of the functions which we do not consider.

3. Preliminaries

Let $\|\cdot\|$ be the ℓ_2-norm. For $h : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$, the subdifferential of h at $\mathbf{x}$ is

$$\partial h(\mathbf{x}) = \big\{\boldsymbol{\zeta} \in \mathbb{R}^d \big| h(\mathbf{x}') \geq h(\mathbf{x}) + \boldsymbol{\zeta}^\top(\mathbf{x}' - \mathbf{x}) + o(\|\mathbf{x}' - \mathbf{x}\|),\ \mathbf{x}' \to \mathbf{x}\big\},$$

where $\boldsymbol{\zeta} \in \partial h(\mathbf{x})$ is a subgradient of h at $\mathbf{x}$. We say h is *μ-strongly convex ($\mu \geq 0$)* on $\mathcal{X}$ if

$$h(\mathbf{x}) \geq h(\mathbf{x}') + \boldsymbol{\zeta}^\top(\mathbf{x} - \mathbf{x}') + \frac{\mu}{2}\|\mathbf{x} - \mathbf{x}'\|^2$$

for any $(\mathbf{x}, \mathbf{x}') \in \mathcal{X} \times \mathcal{X}$ and any $\boldsymbol{\zeta} \in \partial h(\mathbf{x}')$. We say h is *ρ-weakly convex* ($\rho \geq 0$) on $\mathcal{X}$ if

$$h(\mathbf{x}) \geq h(\mathbf{x}') + \boldsymbol{\zeta}^\top(\mathbf{x} - \mathbf{x}') - \frac{\rho}{2}\|\mathbf{x} - \mathbf{x}'\|_2^2$$

for any $(\mathbf{x}, \mathbf{x}') \in \mathcal{X} \times \mathcal{X}$ and any $\boldsymbol{\zeta} \in \partial h(\mathbf{x}')$. We denote the normal cone of $\mathcal{X}$ at $\mathbf{x}$ by $\mathcal{N}_{\mathcal{X}}(\mathbf{x})$ and the distance from $\mathbf{x}$ to a set S by $\text{Dist}(\mathbf{x}, S) = \min_{\mathbf{y} \in S} \|\mathbf{x} - \mathbf{y}\|$.

The following assumptions about (1) are made throughout the paper:

Assumption 1. *The following statements hold:*

A. *$f_i(\mathbf{x})$ is closed and ρ-weakly convex with $\partial f_i(\mathbf{x}) \neq \emptyset$ on any $\mathbf{x} \in \mathcal{X}$ for $i = 0, 1, \ldots, m$.*

B. *$\min_{\mathbf{y} \in \mathcal{X}} \{g(\mathbf{y}) + \frac{\rho + \rho_\epsilon}{2}\|\mathbf{y} - \mathbf{x}\|^2\} < -\sigma_\epsilon$ for any ϵ^2-feasible solution $\mathbf{x}$ ($\mathbf{x} \in \mathcal{X}$ and $g(\mathbf{x}) \leq \epsilon^2$) for some positive constants σ_ϵ and ρ_ϵ. We call this condition* uniform Slater's condition.[3]

[3] The original Slater's condition states that $g(\bar{\mathbf{y}}) < 0$ for some $\bar{\mathbf{y}} \in X$. Here, our assumption is stronger because it includes the term $\rho_\epsilon\|\bar{\mathbf{y}} - \mathbf{x}\|^2$ and requires that inequality holds for any ϵ^2-feasible solution $\mathbf{x}$.

C. *The domain $\mathcal{X}$ is compact such that* $\max_{\mathbf{x},\mathbf{x}'\in\mathcal{X}} \|\mathbf{x}-\mathbf{x}'\| \le D$ *for some constant* D.

D. $f_{lb} \equiv \min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}) > -\infty$.

E. *We have access to an initial ϵ^2-feasible solution $\mathbf{x}_{feas}$ with $\mathbf{x}_{feas} \in \mathcal{X}$ and $g(\mathbf{x}_{feas}) \le \epsilon^2$.*

F. $\|\boldsymbol{\zeta}\| \le M$ *for a constant* M *for any* $\boldsymbol{\zeta} \in \partial f_i(\mathbf{x})$, $\mathbf{x} \in \mathcal{X}$, *and* $i = 0, \ldots, m$.

A function is ρ-weakly convex if it is differentiable and the gradient is ρ-Lipchitz continuous. Hence, the two applications given in Section 1.1 satisfy Assumption 1A when the loss functions ϕ and σ are smooth. It is easy to show that g defined in (1) is also ρ-weakly convex under Assumption 1A. A discussion about Assumption 1E is given in Remark 1.

Under Assumption 1, (1) is a non-convex constrained optimization problem so that even finding an ϵ-feasible solution is difficult in general, let alone a globally optimal solution. For a non-convex problem, one alternative goal is to find a *stationary* point of (1), i.e., a point $\mathbf{x}_* \in \mathcal{X}$ that satisfies the following Karush-Kuhn-Tucker conditions (KKT) conditions (Rockafellar, 1970, Theorem 28.3)

$$-\boldsymbol{\zeta}_0^* - \sum_{i=1}^m \lambda_i^* \boldsymbol{\zeta}_i^* \in \mathcal{N}_{\mathcal{X}}(\mathbf{x}_*), \quad \lambda_i^* f_i(\mathbf{x}_*) = 0, \quad f_i(\mathbf{x}_*) \le 0, \quad \lambda_i^* \ge 0, \tag{2}$$

where λ_i^* is the Lagrangian multiplier corresponding to the constraint $f_i(\mathbf{x}) \le 0$ for $i = 1, \ldots, m$ and $\boldsymbol{\zeta}_i^* \in \partial f_i(\mathbf{x}^*)$ for $i = 0, 1, \ldots, m$. Since an exact stationary point is hard to find with a finite number of iterations by many algorithms, it is more common to aim at finding an ϵ-*stationary* point, i.e., a point $\widehat{\mathbf{x}} \in \mathcal{X}$ satisfying

$$\mathrm{Dist}\Big(-\widehat{\boldsymbol{\zeta}}_0 - \sum_{i=1}^m \widehat{\lambda}_i \widehat{\boldsymbol{\zeta}}_i, \mathcal{N}_{\mathcal{X}}(\widehat{\mathbf{x}})\Big) \le \epsilon, \quad |\widehat{\lambda}_i f_i(\widehat{\mathbf{x}})| \le \epsilon, \quad f_i(\widehat{\mathbf{x}}) \le \epsilon, \quad \widehat{\lambda}_i \ge 0, \tag{3}$$

where $\widehat{\lambda}_i$ is a Lagrangian multiplier corresponding to the constraint $f_i(\widehat{\mathbf{x}}) \le 0$ for $i = 1, \ldots, m$ and $\widehat{\boldsymbol{\zeta}}_i \in \partial f_i(\widehat{\mathbf{x}})$ for $i = 0, 1, \ldots, m$. However, there are two difficulties that prevent algorithms from finding an ϵ-stationary: **(i) Non-smoothness:** When f_0 is non-smooth, computing an ϵ-stationary point with finitely many iterations is challenging even if f_0 is convex and there is no constraint, e.g., $\min_{x\in\mathbb{R}} |x|$, where 0 is an exact stationary point while an algorithm may still return an $x \approx 0$ but $\neq 0$ which is not ϵ-stationary for any $\epsilon < 1$. **(ii) Non-convex constraints:** When non-convex constraints appear, it is difficult to numerically find a point $\widehat{\mathbf{x}}$ that satisfies the third inequality in (3). With a highly infeasible $\widehat{\mathbf{x}}$, the other two inequalities in (3) become less meaningful.

Therefore, to study (3) in a more tractable setting, we follow (Davis & Drusvyatskiy, 2018a; Davis & Grimmer, 2017; Davis & Drusvyatskiy, 2018b; Zhang & He, 2018) to make the weak convexity assumption in Assumption 1A and consider a function $\varphi_{\hat\rho}$ and a solution $\widehat{\mathbf{x}}$ defined as

$$\varphi_{\hat\rho}(\mathbf{x}) \equiv \min_{\mathbf{y}\in\mathcal{X}} \big\{ f(\mathbf{y}) + \frac{\hat\rho}{2}\|\mathbf{y}-\mathbf{x}\|^2, \quad s.t. \quad g(\mathbf{y}) + \frac{\hat\rho}{2}\|\mathbf{y}-\mathbf{x}\|^2 \le 0 \big\}, \tag{4}$$

$$\widehat{\mathbf{x}} \equiv \arg\min_{\mathbf{y}\in\mathcal{X}} \big\{ f(\mathbf{y}) + \frac{\hat\rho}{2}\|\mathbf{y}-\mathbf{x}\|^2, \quad s.t. \quad g(\mathbf{y}) + \frac{\hat\rho}{2}\|\mathbf{y}-\mathbf{x}\|^2 \le 0 \big\}, \tag{5}$$

where $\hat\rho \ge 0$ is a *regularization parameter*, g and f are defined as in (1). It is important to point out that $\varphi_{\hat\rho}$ is different from the Moreau envelope of the function[4] $f(\mathbf{x}) + \mathbf{1}_{\mathcal{X},g\le 0}(\mathbf{x})$ which is defined as

$$\tilde\varphi_{\hat\rho}(\mathbf{x}) \equiv \min_{\mathbf{y}\in\mathcal{X}} \Big\{ f(\mathbf{y}) + \frac{\hat\rho}{2}\|\mathbf{y}-\mathbf{x}\|^2, \quad \text{s.t.} \quad g(\mathbf{y}) \le 0 \Big\}. \tag{6}$$

The function $\tilde\varphi_{\hat\rho}$ was considered in (Davis & Drusvyatskiy, 2018a; Davis & Grimmer, 2017; Davis & Drusvyatskiy, 2018b; Zhang & He, 2018; Rafique et al., 2018) and their algorithm and analysis are based on the fact that (6) is a convex minimization problem when there is no g, f is ρ-weakly convex, and $\hat\rho \ge \rho$. However, for our problem (1) where g exists and is ρ-weakly convex, (6) is hard to evaluate even only approximately. Therefore, we include the term $\frac{\hat\rho}{2}\|\mathbf{y}-\mathbf{x}\|^2$ in the constraint of (4) and (5) so that the minimization problem has a $(\hat\rho-\rho)$-strongly convex objective function and $(\hat\rho-\rho)$-strongly convex constraints when $\hat\rho \ge \rho$. As a result of strong convexity, the solution $\widehat{\mathbf{x}}$ defined in (5) is unique and can be closely approximated by solving (4) or (5).

As an extension to the findings in (Davis & Drusvyatskiy, 2018a; Davis & Grimmer, 2017; Davis & Drusvyatskiy, 2018b; Zhang & He, 2018; Rafique et al., 2018), the quantity $\|\mathbf{x}-\widehat{\mathbf{x}}\|$ with $\widehat{\mathbf{x}}$ defined in (5) can be used as a measure of the quality of a solution $\mathbf{x}$. More specifically, let $\widehat{\lambda}$ be the Lagrangian multiplier that satisfies the following KKT conditions together with $\widehat{\mathbf{x}}$ in (5):

$$-\widehat{\boldsymbol{\zeta}}_0 - \hat\rho(\widehat{\mathbf{x}}-\mathbf{x}) - \widehat{\lambda}\Big(\widehat{\boldsymbol{\zeta}} + \hat\rho(\widehat{\mathbf{x}}-\mathbf{x})\Big) \in \mathcal{N}_{\mathcal{X}}(\widehat{\mathbf{x}}), \quad \widehat{\lambda}\Big(g(\widehat{\mathbf{x}}) + \frac{\hat\rho}{2}\|\widehat{\mathbf{x}}-\mathbf{x}\|^2\Big) = 0, \quad g(\widehat{\mathbf{x}}) + \frac{\hat\rho}{2}\|\widehat{\mathbf{x}}-\mathbf{x}\|^2 \le 0, \quad \widehat{\lambda} \ge 0 \tag{7}$$

[4] Here, $\mathbf{1}_{\mathcal{X},g\le 0}(\mathbf{x})$ denotes the indicator function of the feasible set $\{\mathbf{x} \in \mathcal{X} | g(\mathbf{x}) \le 0\}$

where $\widehat{\boldsymbol{\zeta}}_0 \in \partial f_0(\widehat{\mathbf{x}})$ and $\widehat{\boldsymbol{\zeta}} \in \partial g(\widehat{\mathbf{x}})$. These conditions imply

$$\begin{aligned} \text{Dist}(-\widehat{\boldsymbol{\zeta}}_0 - \widehat{\lambda}\widehat{\boldsymbol{\zeta}}, \mathcal{N}_{\mathcal{X}}(\widehat{\mathbf{x}})) &\le (1+\widehat{\lambda})\hat{\rho}\|\widehat{\mathbf{x}} - \mathbf{x}\|, \\ |\widehat{\lambda} g(\widehat{\mathbf{x}})| &= \frac{\widehat{\lambda}\hat{\rho}}{2}\|\widehat{\mathbf{x}} - \mathbf{x}\|^2, \\ g(\widehat{\mathbf{x}}) &\le 0, \\ \widehat{\lambda} &\ge 0. \end{aligned}$$

Therefore, in the scenario where $\|\widehat{\mathbf{x}} - \mathbf{x}\| \le \epsilon$, $\hat{\rho} = O(1)$ and $\widehat{\lambda} = O(1)$, we have $\text{Dist}(-\widehat{\boldsymbol{\zeta}}_0 - \widehat{\lambda}\widehat{\boldsymbol{\zeta}}, \mathcal{N}_{\mathcal{X}}(\widehat{\mathbf{x}})) = O(\epsilon)$, $|\widehat{\lambda} g(\widehat{\mathbf{x}})| = O(\epsilon^2)$, and $g(\widehat{\mathbf{x}}) \le 0$, which means $\widehat{\mathbf{x}}$ is feasible and satisfies the optimality conditions of the original problem (1) with $O(\epsilon)$ precision and $\mathbf{x}$ is only ϵ-away from $\widehat{\mathbf{x}}$. With this property, we can say $\mathbf{x}$ is near to an ϵ-stationary point (i.e., $\widehat{\mathbf{x}}$) of (1). In Lemma 1 below, we will show that $\widehat{\lambda} = O(1)$ when $\hat{\rho} \in (\rho, \rho + \rho_\epsilon]$ and $\mathbf{x}$ is ϵ^2-feasible under Assumption 1B. We formally define the solution we want to compute as follows.

Definition 1. *A point $\mathbf{x} \in \mathcal{X}$ is called a **nearly ϵ-stationary point** of (1) if $\|\widehat{\mathbf{x}} - \mathbf{x}\| \le \epsilon$ where $\widehat{\mathbf{x}}$ is defined in (5) with respect to $\mathbf{x}$ and $\hat{\rho}$.*

Next, we propose a numerical method for finding a nearly ϵ-stationary point of (1) with theoretical complexity analysis. The proofs for all theoretical results are given in the supplementary file.

4. Inexact Quadratically Regularized Constrained Method

The method we proposed is motivated by the recent studies on the inexact proximal methods by (Davis & Grimmer, 2017; Rafique et al., 2018; Lin et al., 2018a) which originates from the proximal point method (Rockafellar, 1976). The authors of (Davis & Grimmer, 2017) consider $\min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x})$ with a ρ-weakly convex and non-smooth $f(\mathbf{x})$. In their approach, given the iterate $\mathbf{x}_t \in \mathcal{X}$, they generate the next iterate $\mathbf{x}_{t+1}$ by approximately solving the following convex subproblem

$$\mathbf{x}_{t+1} \approx \arg\min_{\mathbf{y}\in\mathcal{X}} f(\mathbf{y}) + \frac{\hat{\rho}}{2}\|\mathbf{y} - \mathbf{x}_t\|_2^2 \tag{8}$$

using the standard stochastic subgradient (SSG) method. Then, $\mathbf{x}_{t+1}$ will be used to construct the next subproblem in a similar way. Similar approaches have been developed for solving non-convex non-concave min-max problems by (Rafique et al., 2018; Lin et al., 2018a).

Similar to their approaches, we will generate $\mathbf{x}_{t+1}$ from $\mathbf{x}_t$ by approximately solving

$$\begin{aligned} \mathbf{x}_{t+1} \approx \widehat{\mathbf{x}}_t \equiv \arg\min_{\mathbf{y}\in\mathcal{X}} \big\{ & f(\mathbf{y}) + \frac{\hat{\rho}}{2}\|\mathbf{y} - \mathbf{x}_t\|^2, \\ & s.t. \quad g(\mathbf{y}) + \frac{\hat{\rho}}{2}\|\mathbf{y} - \mathbf{x}_t\|^2 \le 0 \big\}. \end{aligned} \tag{9}$$

Algorithm 1 Inexact Quadratically Regularized Constrained (IQRC) Method

1: **Input:** An ϵ^2-feasible solution $\mathbf{x}_0 = \mathbf{x}_{\text{feas}}$ (Assumption 1E), $\hat{\rho} > \rho$, $\delta \in (0,1)$, $\hat{\epsilon} = \min\left\{1, \sqrt{\frac{\hat{\rho}-\rho}{4}}\left(\frac{M+\hat{\rho}D}{\sqrt{2\sigma_\epsilon(\hat{\rho}-\rho)}} + 1\right)^{-\frac{1}{2}}\right\}\epsilon$, the number of iterations T, and an oracle $\mathcal{A}$ for (9).
2: **for** $t = 0, \ldots, T-1$ **do**
3: $\quad \mathbf{x}_{t+1} = \mathcal{A}(\mathbf{x}_t, \hat{\rho}, \hat{\epsilon}, \frac{\delta}{T})$
4: **end for**
5: **Output:** $\mathbf{x}_R$ where R is a random index uniformly sampled from $\{0, \ldots, T\}$.

However, the SSG method cannot be directly applied to (9) due to the constraints $g(\mathbf{y}) + \frac{\hat{\rho}}{2}\|\mathbf{y} - \mathbf{x}_t\|^2 \le 0$. Thanks to the recent development in the first-order methods for nonlinear constrained convex optimization, there are existing techniques that can potentially be used as a subroutine to solve (9) in our main algorithm. To facilitate the description of our main algorithm and its anlaysis, we formally define the subroutine with the property we need as follows.

Definition 2. *An algorithm $\mathcal{A}$ is called an **oracle for** (9) if, for any $t \ge 0$, $\hat{\rho} > 0$, $\hat{\epsilon} > 0$, $\delta \in (0,1)$, and $\mathbf{x}_t \in \mathcal{X}$, it finds (potentially stochastic)[5] $\mathbf{x}_{t+1} \in \mathcal{X}$ such that, with a probability of at least $1 - \delta$,*

$$\begin{aligned} f(\mathbf{x}_{t+1}) + \frac{\hat{\rho}}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 - f(\widehat{\mathbf{x}}_t) - \frac{\hat{\rho}}{2}\|\widehat{\mathbf{x}}_t - \mathbf{x}_t\|^2 &\le \hat{\epsilon}^2, \\ g(\mathbf{x}_{t+1}) + \frac{\hat{\rho}}{2}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 &\le \hat{\epsilon}^2 \end{aligned}$$

where $\widehat{\mathbf{x}}_t$ is defined in (9). We denote the output of $\mathcal{A}$ by $\mathbf{x}_{t+1} = \mathcal{A}(\mathbf{x}, \hat{\rho}, \hat{\epsilon}, \delta)$.

Before we discuss which algorithms to use as the orcale, we first present the main algorithm, the inexact quadratically regularized constrained (IQRC) method, in Algorithm 1 and analyze the number of iterations it needs for finding a nearly ϵ-stationary point.

The following lemma shows that the optimal Lagrangian multiplier of (9) is uniformly bounded for all t under Assumption 1. This is critical for establishing the convergence of Algorithm 1.

Lemma 1. *Suppose $\hat{\rho} \in (\rho, \rho + \rho_\epsilon]$. Let $\mathbf{x}_t$ be generated by Algorithm 1, $\widehat{\mathbf{x}}_t$ be defined in (9), and λ_t be the Lagrangian multiplier in the KKT conditions (7) of (9) satisfied by $\widehat{\mathbf{x}}_t$. We have $\lambda_t \le \frac{M+\hat{\rho}D}{\sqrt{2\sigma_\epsilon(\hat{\rho}-\rho)}}$ for $t = 0, 1, 2, \ldots, T-1$ with a probability of at least $1 - \delta$.*

Theorem 1. *Under Assumption 1, Algorithm 1 guarantees $\mathbb{E}_R\|\mathbf{x}_R - \widehat{\mathbf{x}}_R\|^2 \le \epsilon^2$ with a probability of at least $1 - \delta$ if $T \ge \frac{4(f(\mathbf{x}_0) - f_{lb})}{\epsilon^2(\hat{\rho}-\rho)}$, where the expectation is taken over R.*

[5] Here, we allow $\mathcal{A}$ to be a stochastic algorithm.

Remark 1. *Algorithm 1 requires the access to an ϵ^2-feasible solution $\mathbf{x}_0 = \mathbf{x}_{feas}$ (Assumption 1E). When solving (1) without an initial feasible solution, a typical guarantee of an algorithm (e.g. (Cartis et al., 2011; 2014)) is that it either finds an ϵ-feasible and ϵ-stationary point of (1) or finds a point which is an ϵ-stationary point of g but infeasible to (1). In the later case, the solution is typically trapped in a local minimum of g where $g(\mathbf{x})$ is not small, which can happen due to non-convexity of g. Therefore, when $\mathbf{x}_{feas}$ is not available, our method will have such type of guarantee as long as a subgadient method (e.g. (Davis & Drusvyatskiy, 2018a)) is first applied to $\min_{\mathbf{x}\in\mathcal{X}} g(\mathbf{x})$ with $\mathcal{O}(\frac{1}{\epsilon^4})$ iterations, which will return a nearly ϵ-stationary point of g, denoted by $\mathbf{x}_{temp}$. Then if $g(\mathbf{x}_{temp}) \leq \epsilon^2$, we start Algorithm 1 with $\mathbf{x}_0 = \mathbf{x}_{temp}$. If not, we are in the second case mentioned above, namely, we have found a nearly ϵ-stationary point of g which is infeasible to (1), and $\mathbf{x}_{temp}$ is returned as the final output. Adding this step to our method does not change the order of magnitude of its complexity.*

According to Theorem 1, in order to find an nearly ϵ-stationary point in expectation, we have to call the oracle $\mathcal{A}$ $O(1/\epsilon^2)$ times. Therefore, the totally complexity of Algorithm 1 highly depends on the complexity of $\mathcal{A}$ for a given ϵ. In the next sections, we will discuss the methods that can be used as $\mathcal{A}$ when f_i have different properties.

4.1. Oracle for Deterministic Problem

In this section, we assume that we can calculate any $\boldsymbol{\zeta} \in \partial f_i(\mathbf{x})$ for any $\mathbf{x} \in \mathcal{X}$. We define

$$\begin{aligned} F(\mathbf{x}) &:= f(\mathbf{x}) + \frac{\hat{\rho}}{2}\|\mathbf{x} - \mathbf{x}_t\|^2, \\ G(\mathbf{x}) &:= g(\mathbf{x}) + \frac{\hat{\rho}}{2}\|\mathbf{x} - \mathbf{x}_t\|^2 \end{aligned} \tag{10}$$

so that problem (9) becomes $\min_{\mathbf{x}\in\mathcal{X}} F(\mathbf{x})$ s.t. $G(\mathbf{x}) \leq 0$. We define $F'(\mathbf{x})$ and $G'(\mathbf{x})$ as any subgradient of F and G, respectively. Under Assumption 1C and Assumption 1F, we have $\|F'(\mathbf{x})\| \leq M + \hat{\rho}D$ and $\|G'(\mathbf{x})\| \leq M + \hat{\rho}D$ for any $\mathbf{x} \in \mathcal{X}$.

Because problem (9) is non-smooth, we consider the Polyak's switching subgradient method (Polyak, 1967), which is also analyzed in (Nesterov, 2013) and recently extended by (Bayandina et al., 2018; Lan & Zhou, 2016). The method we propose here is a new variant of that method for a strongly convex problem. The details are given in Algorithm 2 where $\text{Proj}_{\mathcal{X}}(\mathbf{x})$ represents the projection of $\mathbf{x}$ to $\mathcal{X}$. Different from (Bayandina et al., 2018), our Algorithm 2 only uses a single loop instead of double loops. It is also different from (Lan & Zhou, 2016) in the sense that our method keeps every intermediate solution ϵ-feasible for (4) while the method in (Lan & Zhou, 2016) only ensures ϵ-feasibility after a fixed number of iterations. Moreover,

Algorithm 2 Switching subgradient method for the subproblem (9)

Input: $\mathbf{z}_0 = \mathbf{x}_t \in \mathcal{X}$, $\hat{\rho} > \rho$ and $\hat{\epsilon} > 0$.
Set $I = \emptyset$ and F and G as in (10).
Set $K = \left\lceil \frac{4(M^2+\hat{\rho}D^2)}{(\hat{\rho}-\rho)\hat{\epsilon}^2} \right\rceil$
for $k = 0, \ldots, K-1$ **do**
  $\gamma_k = \frac{2}{(\hat{\rho}-\rho)(k+2)}$
  if $G(\mathbf{z}_k) \leq \hat{\epsilon}^2$ **then**
    $I \leftarrow I \cup \{k\}$.
    $\mathbf{z}_{k+1} = \text{Proj}_{\mathcal{X}}(\mathbf{z}_k - \gamma_k F'(\mathbf{z}_k))$
  else
    $\mathbf{z}_{k+1} = \text{Proj}_{\mathcal{X}}(\mathbf{z}_k - \gamma_k G'(\mathbf{z}_k))$
  end if
end for
Output: $\mathbf{x}_{t+1} = \frac{\sum_{k\in I}(k+1)\mathbf{z}_k}{\sum_{k\in I}(k+1)}$.

(Lan & Zhou, 2016) requires knowing the total number of iterations before hand in order to design the step size γ_k while our method does not.

The convergence of Algorithm 2 is given below whose proof follows the idea of Section 3.2 in (Lacoste-Julien et al., 2012). However, the original analysis in (Lacoste-Julien et al., 2012) is for the subgradient method applied to unconstrained problems while our analysis is for the switching subgradient method applied to constrained problems.

Theorem 2. *Under Assumption 1, Algorithm 2 guarantees $F(\mathbf{x}_{t+1}) - F(\widehat{\mathbf{x}}_t) \leq \hat{\epsilon}^2$ and $G(\mathbf{x}_{t+1}) \leq \hat{\epsilon}^2$ deterministically and can be used as an oracle $\mathcal{A}$ for (9). The complexity of Algorithm 1 using Algorithm 2 as an oracle is therefore $O(\frac{1}{\epsilon^4})$.*

Remark 2. *Although the main focus of this paper is the case when f_i is non-smooth in (1) for $i = 0, \ldots, m$, our results can be easily extended to the case where each f_i is differentiable with an L-Lipschitz continuous gradient. In this case, the subproblem (9) is written as computing*

$$\begin{aligned} \mathbf{x}_{t+1} \approx \widehat{\mathbf{x}}_t \equiv \arg\min_{\mathbf{y}\in\mathcal{X}} \big\{ & f_0(\mathbf{y}) + \frac{\hat{\rho}}{2}\|\mathbf{y} - \mathbf{x}_t\|^2, \\ & s.t. \quad f_i(\mathbf{y}) + \frac{\hat{\rho}}{2}\|\mathbf{y} - \mathbf{x}_t\|^2 \leq 0,\ i = 1, \ldots, m \big\}. \end{aligned}$$

Since the objective function and constraint functions here are all strongly convex and smooth, there exist some algorithms that can be used as an oracle for (9) satisfying Definition (2). The examples include the level-set method (Lin et al., 2018c) and the augmented Lagrangian method (Xu, 2017b) whose complexity for computing $\mathbf{x}_{t+1} = \mathcal{A}(\mathbf{x}, \hat{\rho}, \hat{\epsilon}, \delta)$ is $O(\frac{1}{\hat{\epsilon}})$. Since $\hat{\epsilon} = O(\epsilon)$, the complexity of Algorithm 1 using (Lin et al., 2018c) or (Xu, 2017b) as the oracle is $O(\frac{1}{\epsilon^2}) \times O(\frac{1}{\hat{\epsilon}}) = O(\frac{1}{\epsilon^3})$.

4.2. Oracle for Stochastic Problem

In this section, we consider the scenario where only a stochastic unbiased estimation for the subgradient of f_i is available. In addition to Assumption 1, we make the following assumption.

Assumption 2. *For any $\mathbf{x} \in \mathcal{X}$ and any $i = 0, 1 \ldots, m$, we can compute a stochastic estimation $\theta_i(\mathbf{x})$ and a stochastic gradient $\boldsymbol{\zeta}_i(\mathbf{x})$ of f_i such that $\mathbb{E}\theta_i(\mathbf{x}) = f_i(\mathbf{x})$ and $\mathbb{E}\boldsymbol{\zeta}_i(\mathbf{x}) \in \partial f_i(\mathbf{x})$. Moreover, there exist constants M_0 and M_1 such that $\|(\theta_1(\mathbf{x}), \theta_2(\mathbf{x}), \ldots, \theta_m(\mathbf{x}))\| \le M_0$ and $\|\boldsymbol{\zeta}_i(\mathbf{x})\| \le M_1$ for any $\mathbf{x}$ almost surely.*

A typical situation where this assumption holds is the stochastic optimization where $f_i \equiv \mathbb{E}F_i(\mathbf{x}, \xi)$ and ξ is a random variable. In that case, we can sample ξ and compute $\theta_i(\mathbf{x}) = F_i(\mathbf{x}, \xi)$ and compute $\boldsymbol{\zeta}_i(\mathbf{x})$ as a subgradient of $F_i(\mathbf{x}, \xi)$ with respect to $\mathbf{x}$.

Under this setting, when solving the subproblem (9), it is not possible to construct an unbiased stochastic subgradient for g in (1) or G in (10) due to the maximization operator in their definitions. Hence, we treat each f_i as an individual function and define

$$F_i(\mathbf{x}) := f_i(\mathbf{x}) + \frac{\hat{\rho}}{2}\|\mathbf{x} - \mathbf{x}_t\|^2 \text{ for } i = 0, 1, \ldots, m$$

so that problem (9) becomes $\min_{\mathbf{x}\in\mathcal{X}} F(\mathbf{x})$ s.t. $F_i(\mathbf{x}) \le 0$ for $i = 0, 1, \ldots, m$. Note that F_i is still $(\hat{\rho}-\rho)$-strongly convex. Its stochastic estimation is $\theta_i(\mathbf{x}) + \frac{\hat{\rho}}{2}\|\mathbf{x} - \mathbf{x}_t\|^2$ which satisfies $\|(\theta_i(\mathbf{x}) + \frac{\hat{\rho}}{2}\|\mathbf{x} - \mathbf{x}_t\|^2)_{i=1}^m\| \le \|(\theta_i(\mathbf{x}))_{i=1}^m\| + \frac{\hat{\rho}\sqrt{m}}{2}\|\mathbf{x} - \mathbf{x}_t\|^2 \le M_0 + \frac{\hat{\rho}\sqrt{m}D^2}{2} \equiv \tilde{M}_0$. Its stochastic gradient is $\boldsymbol{\zeta}_i(\mathbf{x}) + \hat{\rho}(\mathbf{x} - \mathbf{x}_t)$ which satisfies $\|\boldsymbol{\zeta}_i(\mathbf{x}) + \hat{\rho}(\mathbf{x} - \mathbf{x}_t)\| \le M_1 + \hat{\rho}D \equiv \tilde{M}_1$.

The switching subgradient method (Algorithm 2) and its variants (Bayandina et al., 2018; Lan & Zhou, 2016) cannot handle stochastic constraints functions unless a large high-cost mini-batch is used per iteration (Lan & Zhou, 2016). Therefore, we consider using the online stochatsic subgradient method by (Yu et al., 2017) which allows for both stochastic objective function and stochastic constraints. We present their method in Algorithm 3 and analyze the complexity of Algorithm 1 when using their method as the oracle.

Theorem 3. *Under Assumption 1 and 2, Algorithm 3 guarantees $F(\mathbf{x}_{t+1}) - F(\widehat{\mathbf{x}}_t) \le \mathcal{B}_1(D, \tilde{M}_0, \tilde{M}_1, m, \sigma_\epsilon, K, \delta)$ and $F_i(\mathbf{x}_{t+1}) \le \mathcal{B}_2(D, \tilde{M}_0, \tilde{M}_1, m, \sigma_\epsilon, K, \delta)$ for functions $\mathcal{B}_1(D, \tilde{M}_0, \tilde{M}_1, m, \sigma_\epsilon, K, \delta) = O(\frac{\log(K/\delta)}{\sqrt{K}})$ and $\mathcal{B}_1(D, \tilde{M}_0, \tilde{M}_1, m, \sigma_\epsilon, K, \delta) = O(\frac{\log(K/\delta)}{\sqrt{K}})$ with a probability of at least $1 - \delta$. As a consequence, when K is large enough (i.e. $K = \tilde{O}(\frac{1}{\hat{\epsilon}^4}\log(\frac{1}{\delta}))$) so that $\mathcal{B}_1 \le \hat{\epsilon}^2$ and $\mathcal{B}_2 \le \hat{\epsilon}^2$, Algorithm 3 can be used as an oracle $\mathcal{A}$ for (9). The complexity of Algorithm 1 using Algorithm 3 as an oracle is therefore $\tilde{O}(\frac{1}{\epsilon^6})$.*

Algorithm 3 Online stochastic subgradient method by (Yu et al., 2017) for the subproblem (9)

Input: $\mathbf{z}_0 = \mathbf{x}_t \in \mathcal{X}$, $\hat{\rho} > \rho$, $\hat{\epsilon} > 0$, and the number of iterations K.
Set $V = \sqrt{K}$ and $\alpha = K$.
Set $Q_0^i = 0$ for $i = 1, \ldots, m$.
for $k = 0, \ldots, K-1$ **do**
  $\tilde{\theta}_i^k = \theta_i(\mathbf{z}_k) + \frac{\hat{\rho}}{2}\|\mathbf{z}_k - \mathbf{x}_t\|^2$ and $\tilde{\boldsymbol{\zeta}}_i^k = \boldsymbol{\zeta}_i(\mathbf{z}_k) + \hat{\rho}(\mathbf{z}_k - \mathbf{x}_t)$ for $i = 0, 1, 2, \ldots, m$.
  $\mathbf{z}_{k+1} = \arg\min_{\mathbf{z}\in\mathcal{X}} \left\{ \begin{array}{l} (V\boldsymbol{\zeta}_0^k + \sum_{i=1}^m Q_k^i \boldsymbol{\zeta}_i^k)^\top (\mathbf{z} - \mathbf{z}_k) \\ +\alpha\|\mathbf{z} - \mathbf{z}_k\|^2 \end{array} \right\}$
  $Q_{k+1}^i = \max\{Q_k^i + \tilde{\theta}_i^k + (\tilde{\boldsymbol{\zeta}}_i^k)^\top(\mathbf{z}_{k+1} - \mathbf{z}_k), 0\}$ $i = 1, 2, \ldots, m$.
end for
Output: $\mathbf{x}_{t+1} = \frac{1}{K}\sum_{k=0}^{K-1} \mathbf{z}_k$.

Since functions $\mathcal{B}_1$ and $\mathcal{B}_2$ are complicated, we put them in (12) and (17) in the supplementary file.

5. Numerical Experiments

In this section, we evaluate the numerical performance of the proposed methods on a multi-class Neyman-Pearson classification (mNPC) problem with nonconvex loss. Let ξ_k for $k = 1, 2, \ldots, K$ denote K classes of data, each of which belongs to a subset of training data D_k. We train K linear models $\mathbf{x}_k$, $k = 1, 2, \ldots K$, and then predict the class of data ξ by $\arg\max_{k=1,2,\ldots,K} \mathbf{x}_k^\top \xi$. To achieve a high classification accuracy, the value $\mathbf{x}_k^\top \xi_k - \mathbf{x}_l^\top \xi_k$ needs to be positively large for any $k \ne l$ and $\xi_k \in D_k$ (Weston & Watkins, 1998; Crammer & Singer, 2002), which can be achieved by minimizing the following loss

$$\frac{1}{|D_k|}\sum_{l\ne k}\sum_{\xi_k\in D_k} \phi(\mathbf{x}_k^\top \xi_k - \mathbf{x}_l^\top \xi_k),$$

where ϕ is a non-increasing potentially non-convex loss function. mNPC prioritizes minimizing the loss on one class, which is class 1 in our formulation, and then controls the loss on all other classes by solving

$$\begin{aligned} &\min_{\|\mathbf{x}_k\|_2\le\lambda, k=1,\ldots,K} \frac{1}{|\mathcal{D}_1|}\sum_{l\ne k}\sum_{\xi\in\mathcal{D}_1} \phi(\mathbf{x}_k^\top \xi - \mathbf{x}_l^\top \xi), \\ &\text{s.t. } \frac{1}{|\mathcal{D}_k|}\sum_{l\ne k}\sum_{\xi\in\mathcal{D}_k} \phi(\mathbf{x}_k^\top \xi - \mathbf{x}_l^\top \xi) \le r_k, \quad k = 2, 3, \ldots, K, \end{aligned} \tag{11}$$

where r_k controls the loss of class k and λ is the regularization parameter. In the experiment, the function ϕ in (11) is chosen as the sigmoid function $1/(1 + \exp(z))$.

We compare our IQRC method to the exact penalty method proposed in (Cartis et al., 2011). Both methods are implemented in Matlab on a 64-bit MacOS Catalina machine with a 2.90 Ghz Intel Core i7-6920HQ CPU and 16GB of memory. We conduct experiments on three LIBSVM

multi-class classification datasets *pendigits*, *segment*, and *usps*. *pendigits* dataset has 7494 instances and 10 classes, while each instance is represented by a feature vector of dimension 16. *segment* dataset has 2310 instances and 7 classes, and each instance is represented by a feature vector of dimension 7. *usps* dataset has 7291 instances and 10 classes while each instance has 256 number of features. We choose $r_k = 4.5, k = 2, \ldots, K$ for *pendigits* and *usps*, and $r_k = 3, k = 2, \ldots, K$ for *segment*. λ is selected to be 0.1 for all datasets. For both algorithms in comparison, the initial solution $\mathbf{x} = \mathbf{0}$ is chosen. It is easy to verify that $\mathbf{x} = \mathbf{0}$ is a feasible solution given the λ and r_k we choose. For both algorithms, we tune hyper-parameters from a discrete set of choices.

To simplify notations, let $f_0(\mathbf{x})$ denote the objective function of problem (11). Similarly, let $f_i(\mathbf{x}) \leq 0$ for $i = 1, \ldots, 2K-1$ denote all constraints of problem (11) including both functional constraints and compact set constraints. Then exact penalty method proposed by (Cartis et al., 2011) first derives a direction $\mathbf{s}_k$ by solving

$$\mathbf{s}_k \in \arg\min_{\|\mathbf{s}\| \leq \Delta_k} \Big\{ f_0(\mathbf{x}_k) + \nabla f_0(\mathbf{x}_k)^\top \mathbf{s} + p \sum_{i=1}^{2K-1} \Big| \max\{0, f_i(\mathbf{x}_k) + \nabla f_i(\mathbf{x}_k)^\top \mathbf{s}\} \Big| \Big\}, \tag{12}$$

where $p > 0$ is the penalty parameter and $\Delta_k > 0$ is the radius. When $\mathbf{s}_k$ is derived, a trust region method is used to update the current solution $\mathbf{x}_k$ by performing $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{s_k}$ if this update can significantly reduce the value of the function

$$f_0(\mathbf{x}) + p \sum_{i=1}^{2K-1} |\max\{0, f_i(\mathbf{x})\}|.$$

In our implementation, we reformulate problem (12) as a linear programming problem and then use the Matlab built-in solver to obtain $\mathbf{s}_k$. The exact penalty method in (Cartis et al., 2011) requires several hyper-parameters including a steering parameter ξ, an increase factor τ to update the penalty parameter, an initial penalty parameter p_{-1}, and the tolerance ϵ. After tunning ξ and τ, we set $\xi = 0.1, \tau = 100, p_{-1} = 1/\xi$, and $\epsilon = 0.001$ for *pendigits* and *segment*. For *usps*, we choose $\xi = 0.02, \tau = 100, p_{-1} = 1/\xi$, and $\epsilon = 0.001$. The trust-region algorithm in (Cartis et al., 2011) specifies only the interval in which the trust-region radius should fall in, and there are many possible choices. In the experiment, we specifically follow the rule that $\Delta_{k+1} = \Delta_k$ if $r_k \geq \eta_2$, $\Delta_{k+1} = \gamma_2 \Delta_k$ if $r_k \in [\eta_1, \eta_2]$, and $\Delta_{k+1} = \gamma_1 \Delta_k$ if $r_k < \eta_1$. The trust region subproblem also requires several control parameters. For all datasets, we use $\Delta_0 = 1, \eta_1 = 0.3, \eta_2 = 0.7, \gamma_1 = 0.5, \gamma_2 = 1$. (See (Cartis et al., 2011) for the definitions of $\Delta_0, \eta_1, \eta_2, \gamma_1$, and γ_2).

For our proposed method, the subproblem in the IQRC method is solved using the switching subgradient method (Algorithm 2). After tunning regularization parameter $\hat{\rho}$ and inner iteration number K, we set $\hat{\rho} = 1$, $K = 20000$, and tolerance $\epsilon = 0.001$ for all datasets.

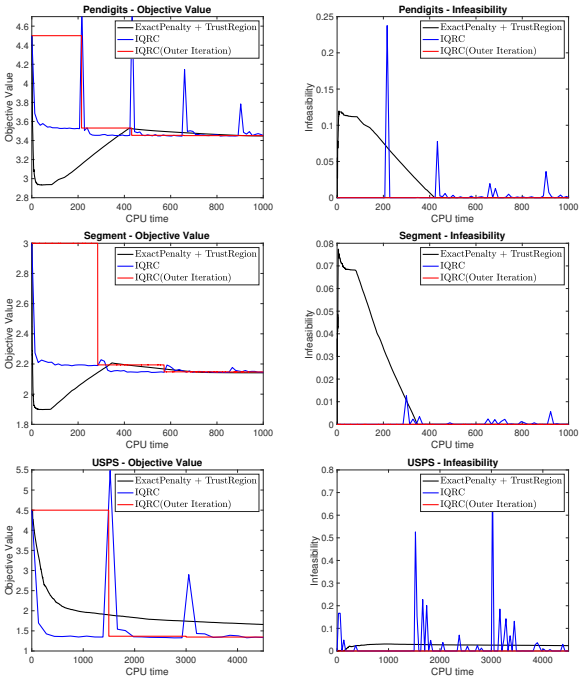

Figure 1. Comparison between the IQRC method and the exact penalty method by (Cartis et al., 2011) for solving multi-class Neyman-Pearson classification problem (11).

The numerical results are presented in Figure 1. The x-axis represents the CPU time that each algorithm took. The y-axis on left column of Figure 1 represents the objective value of (11) and the y-axis on right column of Figure 1 represents infeasibility, i.e., $\max\{\max_{i=1,\ldots,m} f_i(\mathbf{x}), \max_{k=1,\ldots,K} \|\mathbf{x}_k\| - \lambda, 0\}$, of the iterates. The red line shows the performance of the outer iteration solution $\mathbf{x}_t$ in Algorithm 1, which is the solution users need in practice. The blue line evaluates the performance of inner iteration $\mathbf{z}_t$ in Algorithm 2. We show it here only for reader's curiosity. The black line represents the performance of the exact penalty method proposed by (Cartis et al., 2011). We conclude from Figure 1 that, for these three instances, the IQRC method outperformed the exact penalty method in terms of the capability of reducing the objective value and infeasiblity.

6. Conclusion

Continuous optimization models with nonlinear constraints have been widely used in many areas including machine learning, statistics and operations research. When non-convex functional constraints appear, even finding a feasible solution is challenging. In this paper, we proposed a class of quadratically regularized subgradient method which can

find a nearly stationary point for functional constrained non-convex problems. The complexity to find a nearly stationary point for both deterministic case and stochastic case are derived: (i) when each function f_i is a deterministic function, we proposed a new variant of switching subgradient to solve strongly convex subproblem and the total complexity of Algorithm 1 for finding a nearly ϵ-stationary point is $\mathcal{O}(\frac{1}{\epsilon^4})$, (ii) when each function f_i is given as an expectation of a stochastic function, we analyzed stochastic subgradient method by (Yu et al., 2017) for solving subproblem and the total complexity of Algorithm 1 for finding a nearly ϵ-stationary point is $\tilde{\mathcal{O}}(\frac{1}{\epsilon^6})$.

ACKNOWLEDGMENTS

T. Yang is partially supported by National Science Foundation CAREER Award 1844403.

References

Allen-Zhu, Z. Natasha: Faster non-convex stochastic optimization via strongly non-convex parameter. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pp. 89–97, 2017.

Allen-Zhu, Z. and Hazan, E. Variance reduction for faster non-convex optimization. In *Proceedings of the 33nd International Conference on Machine Learning (ICML)*, pp. 699–707, 2016.

Bayandina, A., Dvurechensky, P., Gasnikov, A., Stonyakin, F., and Titov, A. Mirror descent and convex optimization problems with non-smooth inequality constraints. In *Large-Scale and Distributed Optimization*, pp. 181–213. Springer, 2018.

Bertsekas, D. P. *Nonlinear programming*. Athena scientific Belmont, 1999.

Bertsekas, D. P. *Constrained optimization and Lagrange multiplier methods*. Academic press, 2014.

Boob, D., Deng, Q., and Lan, G. Optimization with convex and nonconvex functional constraints. *arXiv preprint arXiv:1908.02734*, 2019.

Cartis, C., Gould, N. I., and Toint, P. L. On the evaluation complexity of composite function minimization with applications to nonconvex nonlinear programming. *SIAM Journal on Optimization*, 21(4):1721–1739, 2011.

Cartis, C., Gould, N. I., and Toint, P. L. On the complexity of finding first-order critical points in constrained nonlinear optimization. *Mathematical Programming*, 144(1-2):93–106, 2014.

Chen, J., Yang, T., Lin, Q., Zhang, L., and Chang, Y. Optimal stochastic strongly convex optimization with a logarithmic number of projections. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, pp. 122–131. AUAI Press, 2016.

Chen, Z., Yang, T., Yi, J., Zhou, B., and Chen, E. Universal stagewise learning for non-convex problems with convergence on averaged solutions. *CoRR*, /abs/1808.06296, 2018.

Crammer, K. and Singer, Y. On the learnability and design of output codes for multiclass problems. *Machine learning*, 47(2-3):201–233, 2002.

Davis, D. and Drusvyatskiy, D. Stochastic subgradient method converges at the rate $o(k^{-1/4})$ on weakly convex functions. *arXiv preprint arXiv:1802.02988*, 2018a.

Davis, D. and Drusvyatskiy, D. Complexity of finding near-stationary points of convex functions stochastically. *arXiv preprint arXiv:1802.08556*, 2018b.

Davis, D. and Drusvyatskiy, D. Stochastic model-based minimization of weakly convex functions. *arXiv preprint arXiv:1803.06523*, 2018c.

Davis, D. and Grimmer, B. Proximally guided stochastic subgradient method for nonsmooth, nonconvex problems. *arXiv preprint arXiv:1707.03505*, 2017.

Drusvyatskiy, D. The proximal point method revisited. *arXiv preprint arXiv:1712.06038*, 2017.

Drusvyatskiy, D. and Paquette, C. Efficiency of minimizing compositions of convex functions and smooth maps. *Mathematical Programming*, Jul 2018.

Fercoq, O., Alacaoglu, A., Necoara, I., and Cevher, V. Almost surely constrained convex optimization. *arXiv preprint arXiv:1902.00126*, 2019.

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

Ghadimi, S. and Lan, G. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Math. Program.*, 156(1-2):59–99, 2016.

Goh, G., Cotter, A., Gupta, M., and Friedlander, M. P. Satisfying real-world goals with dataset constraints. In *Advances in Neural Information Processing Systems*, pp. 2415–2423, 2016.

Lacoste-Julien, S. Convergence rate of frank-wolfe for non-convex objectives. *arXiv preprint arXiv:1607.00345*, 2016.

Lacoste-Julien, S., Schmidt, M., and Bach, F. A simpler approach to obtaining an $\mathcal{O}(1/t)$ convergence rate for the projected stochastic subgradient method. *arXiv preprint arXiv:1212.2002*, 2012.

Lan, G. and Yang, Y. Accelerated stochastic algorithms for nonconvex finite-sum and multi-block optimization. *CoRR*, abs/1805.05411, 2018.

Lan, G. and Zhou, Z. Algorithms for stochastic optimization with expectation constraints. *arXiv preprint arXiv:1604.03887*, 2016.

Lin, Q., Liu, M., Rafique, H., and Yang, T. Solving weakly-convex-weakly-concave saddle-point problems as weakly-monotone variational inequality. *arXiv preprint arXiv:1810.10207*, 2018a.

Lin, Q., Ma, R., and Yang, T. Level-set methods for finite-sum constrained convex optimization. In *International Conference on Machine Learning*, pp. 3118–3127, 2018b.

Lin, Q., Nadarajah, S., and Soheili, N. A level-set method for convex optimization with a feasible solution path. *SIAM Journal on Optimization*, 28(4):3290–3311, 2018c.

Mahdavi, M., Yang, T., Jin, R., and Zhu, S. Stochastic gradient descent with only one projection. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 503–511, 2012.

Nesterov, Y. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

Nguyen, L. T. Dc approach to weakly convex optimization and nonconvex quadratic optimization problems. *Optimization*, 67(2):195–216, 2018.

Nocedal, J. and Wright, S. *Numerical optimization*. Springer Science & Business Media, 2006.

Paquette, C., Lin, H., Drusvyatskiy, D., Mairal, J., and Harchaoui, Z. Catalyst for gradient-based nonconvex optimization. pp. 1–10, 2018.

Polyak, B. A general method of solving extremum problems. *Soviet Mathematics Doklady*, 8(3):593–597, 1967.

Rafique, H., Liu, M., Lin, Q., and Yang, T. Non-convex min-max optimization: Provable algorithms and applications in machine learning. *arXiv preprint arXiv:1810.02060*, 2018.

Reddi, S. J., Hefny, A., Sra, S., Póczós, B., and Smola, A. Stochastic variance reduction for nonconvex optimization. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning (ICML)*, pp. 314–323. JMLR.org, 2016a.

Reddi, S. J., Sra, S., Póczos, B., and Smola, A. J. Fast incremental method for smooth nonconvex optimization. In *55th IEEE Conference on Decision and Control (CDC)*, pp. 1971–1977, 2016b.

Rigollet, P. and Tong, X. Neyman-pearson classification, convexity and stochastic constraints. *Journal of Machine Learning Research*, 12(Oct):2831–2855, 2011.

Rockafellar, R. *Convex Analysis*. Princeton mathematical series. Princeton University Press, 1970.

Rockafellar, R. T. Monotone operators and the proximal point algorithm. *SIAM J. on Control and Optimization*, 14(5), 1976. URL http://epubs.siam.org/sicon/resource/1/sjcodc/v14/i5/p877_s1.

Ruszczyński, A. P. *Nonlinear optimization*, volume 13. Princeton university press, 2006.

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. 13965–13977, 2019.

Sahni, S. Computationally related problems. *SIAM Journal on Computing*, 3(4):262–279, 1974.

Wei, X. and Neely, M. J. Primal-dual frank-wolfe for constrained stochastic programs with convex and non-convex objectives. *arXiv preprint arXiv:1806.00709*, 2018.

Wei, X., Yu, H., Ling, Q., and Neely, M. Solving non-smooth constrained programs with lower complexity than $o(1/\varepsilon)$: A primal-dual homotopy smoothing approach. In *Advances in Neural Information Processing Systems*, pp. 3995–4005, 2018.

Weston, J. and Watkins, C. Multi-class support vector machines. Technical report, Citeseer, 1998.

Xu, Y. First-order methods for constrained convex programming based on linearized augmented lagrangian function. *arXiv preprint arXiv:1711.08020*, 2017a.

Xu, Y. Global convergence rates of augmented lagrangian methods for constrained convex programming. *arXiv preprint arXiv:1711.05812*, 2017b.

Xu, Y. Primal-dual stochastic gradient method for convex programs with many functional constraints. *arXiv preprint arXiv:1802.02724*, 2018.

Yang, T., Lin, Q., and Li, Z. Unified convergence analysis of stochastic momentum methods for convex and non-convex optimization. volume abs/1604.03257, 2016.

Yang, T., Lin, Q., and Zhang, L. A richer theory of convex constrained optimization with reduced projections and improved rates. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 3901–3910. JMLR. org, 2017.

Yu, H. and Neely, M. J. A simple parallel algorithm with an o(1/t) convergence rate for general convex programs. *SIAM Journal on Optimization*, 27(2):759–783, 2017.

Yu, H., Neely, M., and Wei, X. Online convex optimization with stochastic constraints. In *Advances in Neural Information Processing Systems*, pp. 1428–1438, 2017.

Zafar, M. B., Valera, I., Rodriguez, M. G., and Gummadi, K. P. Fairness constraints: Mechanisms for fair classification. *arXiv preprint arXiv:1507.05259*, 2015.

Zafar, M. B., Valera, I., Rodriguez, M., Gummadi, K., and Weller, A. From parity to preference-based notions of fairness in classification. In *Advances in Neural Information Processing Systems*, pp. 229–239, 2017.

Zhang, L., Yang, T., Jin, R., and He, X. O (logt) projections for stochastic optimization of smooth and strongly convex functions. *arXiv preprint arXiv:1304.0740*, 2013.

Zhang, S. and He, N. On the convergence rate of stochastic mirror descent for nonsmooth nonconvex optimization. *arXiv preprint arXiv:1806.04781*, 2018.