COMPLEXITY OF FINITE-SUM OPTIMIZATION WITH NONSMOOTH COMPOSITE FUNCTIONS AND NON-LIPSCHITIZ REGULARIZATION*

XIAO WANG† AND XIAOJUN CHEN‡

Abstract. In this paper we present complexity analysis of proximal inexact gradient methods for finite-sum optimization with a nonconvex nonsmooth composite function and non-Lipschitz regularization. By getting access to a convex approximation to the Lipschitz function and a Lipschitz continuous approximation to the non-Lipschitz regularizer, we construct a proximal subproblem at each iteration without using exact function values and gradients. With certain accuracy control on inexact gradients and subproblem solutions, we show that the oracle complexity in terms of total number of inexact gradient evaluations is in order $\mathcal{O}(\epsilon^{-2})$ to find an (ϵ, δ)-approximate first-order stationary point, ensuring that within a δ-ball centered at this point the maximum reduction of an approximation model does not exceed $\epsilon\delta$. This shows that we can have the same worse-case evaluation complexity order as [5, 12] even if we introduce the non-Lipschitz singularity and the nonconvex nonsmooth composite function in the objective function. Moreover, we establish that the oracle complexity regarding the total number of stochastic oracles is in order $\tilde{\mathcal{O}}(\epsilon^{-2})$ with high probability for stochastic proximal inexact gradient methods. We further extend the algorithm to adjust to solving stochastic problems with expectation form and derive the associated oracle complexity in order $\tilde{\mathcal{O}}(\epsilon^{-10/3})$ with high probability.

Key words. nonconvexity, nonsmoothness, non-Lipschitz regularization, inexact oracle, complexity

MSC codes. 90C30, 90C46, 65K05

1. Introduction.

In this paper, we consider the following nonconvex nonsmooth optimization problem:

$$\min_{x\in\mathcal{F}} \quad Q(x) := f(x) + h(c(x)) + \|Vx\|_p^p, \tag{1.1}$$

where $\mathcal{F} \subseteq \mathbb{R}^n$ is nonempty, bounded, closed and convex, $f : \mathbb{R}^n \to \mathbb{R}$ and $c : \mathbb{R}^n \to \mathbb{R}^r$ are continuously differentiable with Lipschitz continuous gradients over $\mathcal{F}$, $h : \mathbb{R}^r \to \mathbb{R}$ is Lipschitz continuous and convex but possibly nonsmooth, $V \in \mathbb{R}^{\bar{n}\times n}$ with $\bar{n} \le n$ and $p \in (0, 1)$. We assume that rows of V, denoted by v_i^T, $i = 1, \ldots, \bar{n}$, are orthonormal, without loss of generality. Problem (1.1) has numerous applications in data science, where f is a loss function, h is a penalty function and $\|\cdot\|_p^p$ is a sparse regularization. For instance, with the increasing interest of group sparsity regularization for neural networks (see e.g. [4, 20, 25, 31]), the loss function f may rely on a large data set and be defined in the form

$$f(x) = \frac{1}{N}\sum_{i=1}^{N} f_i(x), \tag{1.2}$$

where $f_i : \mathbb{R}^n \to \mathbb{R}, i = 1, \ldots, N$, are continuously differentiable and the sample size N can be very large such that it may be time-consuming and sometimes even prohibitive to access all component functions to compute the exact gradient of f at

*Submitted to the editors DATE.

Funding: This work was funded by by the Major Key Project of PCL (No. PCL2022A05), the Chinese NSF Grant 12271278 and Hong Kong Research Grants Council, grant PolyU15300219.

†Department of AI Computing, Peng Cheng Laboratory, Shenzhen, China (wangx07@pcl.ac.cn).

‡Department of Applied Mathematics, Hong Kong Polytechnic University, Kowloon, Hong Kong (maxjchen@polyu.edu.hk).

a query point. Moreover, constraints are often imposed to enforce specific conditions on variables. For example, constraints of the form $c(x) \leq 0$, where $c : \mathbb{R}^n \to \mathbb{R}^r$, are prevalent in a wide range of applications, including image restoration [1], film restoration [18] and SVM [21]. However, ensuring feasibility of iterates with respect to these constraints throughout the algorithmic process can be challenging. To tackle this issue, infeasible methods are commonly employed, which allow for violations of the constraints. Specifically, with the aid of a penalty function, for instance, ℓ_1 penalty function, one can remove the constraints by introducing a nonsmooth penalty term in the objective, e.g. $h(c(x)) = \rho\|(c(x))_+\|_1$ with $(c(x))_+ = \max(c(x), 0)$ and ρ being a penalty parameter. As studied in [10, 15], the resulting problem can be an exact penalty formulation of the original one to some extent, with nice properties regarding their minimizers. As is well studied in the literature, the nonconvex, nonsmooth and non-Lipschitz ℓ_p $(0 < p < 1)$ regularizer has shown a good performance for sparse variable selection. However, in general the non-Lipschitz regularized problems are strongly NP-hard [9]. Challenges often arise in algorithm design and analysis. The past decade has witnessed highly productive progress on the study of ℓ_p $(0 < p < 1)$ optimization and a surge of works has been proposed, to name a few but not limited to [2, 3, 13, 14, 16, 19, 21, 30].

Cartis *et al.* study the evaluation complexities of minimizing $f(x) + h(c(x))$ to reach the first-order critical measure within ϵ in [5] and to reach high-order approximate minimizers in [7]. In recent work [8], they consider minimizing $f(x)+h(c(x))$ over a convex set and apply high-order approximation model to reach high-order approximate minimizers. Gratton *et al.* in [17] propose an adaptive regularization algorithm using inexact function and gradient evaluations for minimizing $f(x)+h(c(x))$ and show that their algorithm needs at most $O(|\log(\epsilon)|\epsilon^{-2})$ evaluations of the functions and their derivatives for finding an ϵ-approximate first-order stationary point. In [11, 12], high-order algorithms for solving minimization problems with non-Lipschitzian group sparsity terms are studied, where the objective is the sum of a smooth function and a non-Lipschitz regularizer. Compared with problems studied in [5, 8, 11, 12, 17], the objective function in (1.1) has not only the nonsmooth function $h(c(x))$, but also the non-Lipschitz regularizer $\|Vx\|_p^p$, whose complexity has not been established in the literature to the best of our knowledge. We also notice that those algorithms studied in previous works [5, 8, 11, 12] rely on exact function values and gradients of f, which, however, are expensive to obtain in many scenarios with a large finite-sum structure. Inspired by above points, in this paper we will focus on complexity analysis for problem (1.1) to reach approximate first-order stationary point. We will investigate whether the absence of diffentiability of the Lipschitz term together with the existence of non-Lipschitz regularization will affect the worst-case complexity, compared with existing works.

Problems with f in finite-sum structure (1.2) face challenges when computing exact function information, due to the large number of component functions. To alleviate possible difficulties, stochastic oracles are normally called to approximate exact information. In the past decade, along with the development of data science, studies on stochastic approximation methods for nonlinear optimization grow rapidly in popularity, ranging from convex to nonconvex problems and from smooth to nonsmooth problems. Xu *et al.* [29] study a class of optimization problems with nonconvex, nonsmooth regularizer, namely minimizing $g(x) - h(x) + \Lambda(x)$, where g and h are both convex and Λ is a nonconvex and nonsmooth regularizer. Moreover, it requires in theoretical analysis that g be smooth and h be Hölder smooth. The proposed algorithm in [29] can be also applied to unconstrained $\ell_p(0 < p < 1)$ regularized optimization with g in the finite-sum form. The associated gradient complexity to find a nearly ϵ-critical point is in order $\mathcal{O}(\epsilon^{-4})$. Metel and Takeda [22] consider unconstrained optimization with a nonconvex but Lipschitz continuous regularizer. The proposed algorithm owns $\mathcal{O}(\epsilon^{-3})$ gradient-call complexity for finite-sum minimization when a variance reduction strategy is applied. However, the Lipschitz continuity assumption fails for ℓ_p $(0 < p < 1)$ regularizer. Cheng *et al.* [14] propose an interior stochastic gradient method for nonnegative constrained optimization with ℓ_p regularizer and investigate the oracle complexities to find an approximate stationary point. Xu *et al.* [28] propose stochastic proximal gradient methods for minimizing summation of a smooth function f and a nonsmooth nonconvex regularizer and show that the $\mathcal{O}(\epsilon^{-2})$ gradient complexity can be achieved to find an ϵ-stationary point. The proposed algorithm in [28] requires the proximal mapping of the nonconvex regularizer be easy to obtain. However, these existing results cannot be applied to problem (1.1) due to the nonsmoothness and nonconvexity of $f + h$ or the non-Lipschitz continuity.

Contribution. The main contribution of this paper lies in the complexity analysis of proximal inexact gradient methods for finite-sum optimization with a nonconvex nonsmooth composite function and non-Lipschitz regularization (1.1). By getting access to a convex approximation to the Lipschitz function in the objective, together with a Lipschitz continuous approximation to the non-Lipschitz regularizer, we build a proximal subproblem at each iteration without using exact function values and gradients of f. Under certain conditions on inexact gradients and inexact subproblem solutions, we prove that the oracle complexity in terms of the total number of inexact gradient evaluations to find an approximate (ϵ, δ)-approximate first-order stationary point is in order $\mathcal{O}(\epsilon^{-2})$. This verifies that adding the nonsmooth nonconvex composite function and non-Lipschitz regularizer and using inexact gradients do not affect the worst-case oracle complexity, compared with existing results [5, 8, 11, 12]. Furthermore, we use the finite-sum structure of f and propose a stochastic variant of the algorithm through calls to stochastic first-order oracles. We show that the corresponding oracle complexity in terms of total number of stochastic first-order oracles is in order $\tilde{\mathcal{O}}(\epsilon^{-2})$ with high probability, where we use $\tilde{\mathcal{O}}$ to hide the dependence on logarithmic factor in the complexity order. Furthermore, we extend the proposed algorithm to solve stochastic problems with f in expectation form and obtain the $\tilde{\mathcal{O}}(\epsilon^{-10/3})$-oracle complexity with high probability. We present more details on the significant differences from existing works.

(i) The related convergence and iteration complexity in [5, 8, 11, 12] are established within trust region schemes, which require accurate function values and derivatives of function f. However, those analysis cannot be applied to stochastic optimization problems, where only approximate or stochastic gradients are available. In this scenario, the behavior of the objective function can only be characterized based on inexact derivatives of f. Hence, it is imperative to modify the primary algorithmic framework to accommodate the reliance on approximate or stochastic gradients and the absence of a trust region scheme. This adaptation necessitates rigorous analysis under these altered conditions.

(ii) While our method draws inspiration from existing techniques, such as the convex approximation to the composite part and the Lipschitz continuous approximation to the non-Lipschitz regularizer, the coexistence of these two aspects brings significant challenges for the theoretical analysis, which makes it different from existing works. For instance, one particular challenge is about ensuring the existence of an inexact subproblem solution s_k that satisfies the required conditions, as presented in Lemma 2.5. Such detailed analysis, however, is not provided in [11, 12]. Another challenge

arises when considering the approximate criticality of the output of Algorithm 2.1. Due to the significant modifications made to adapt to the stochastic setting and the absence of a trust region scheme, the algorithm framework's analysis differs substantially from existing methods. In addition to addressing these challenges, we present a unified framework by incorporating various elements and leveraging the strengths of each element. This enables our algorithm to tackle a broader range of problems.
(iii) When adapting the deterministic proximal inexact gradient method to stochastic settings, including the finite-sum setting and expectation setting, it causes nontrivial challenges to the theoretical analysis. In our paper, we go beyond a simple replacement of the deterministic gradient with a stochastic gradient, recognizing the need for careful consideration of oracle complexity analysis in the stochastic counterpart, which contributes to the value of our work. Particularly, the extension of the analysis in [25] to non-Lipschitz regularized optimization and to the expectation case proves to be a nontrivial task. Our oracle complexity analysis heavily relies on the essential property of the proposed algorithm, specifically the boundedness of $\sum_{k\in\mathcal{K}} \|s_k\|^2$ as demonstrated in Theorem 3.8.

Organization. This paper is organized as follows. In Section 2 we present a detailed algorithmic framework for proximal inexact gradient methods for (1.1). In Section 3 we explore the oracle complexity of the proposed framework to find an (ϵ, δ)-approximate first-order stationary point. In Section 4 we propose a stochastic variant of the algorithm for problems with f in finite-sum structure (1.2) and establish the oracle complexity accordingly. In Section 5 we propose an extended stochastic variant for problems in expectation case and investigate the related oracle complexity. In Section 6 we illustrate our algorithm by a numerical example. Finally, concluding remarks are drawn in Section 7.

2. Algorithm description.

In this section, we will present an algorithmic framework for proximal inexact gradient methods for solving (1.1). As the objective function Q is nonconvex, nonsmooth and non-Lipschitz, it is generally intractable to approximately find a global or even a local minimizer. Thus our algorithm aims for an approximate first-order stationary point of (1.1). The core of our algorithm design is to construct a Lipschitz continuous approximation model of the objective function at each iteration. We then perform a search within a local neighborhood of the current iterate while aiming to minimize the approximation model as much as possible. The use of Lipschitz continuous approximation models helps us predict the behavior of the objective function while minimizing the impact of approximation errors in the optimization process. The proposed algorithm differs from existing works on complexity analysis, such as [5, 11, 12], where a trust region scheme is typically employed, requiring exact evaluations of the function value and its derivatives. In contrast, it only relies on getting access to inexact first-order derivatives of the objective function, which enables us to extend its applicability to stochastic variants. By utilizing these inexact derivatives, we can effectively navigate the search space and make progress towards the optimal solution without the need for precise function value and derivative evaluations. By adopting this approach, we strike a balance between computational efficiency and accuracy, making our algorithm more suitable for scenarios where exact evaluations may be costly or impractical.

We first define the following index sets at a point x for a given nonnegative constant ϵ:

$$\mathcal{A}(x,\epsilon) = \left\{ i \in [\bar{n}] : |v_i^T x| > \epsilon \right\}, \quad \mathcal{R}(x,\epsilon) = \bigcap_{i\in[\bar{n}]\setminus\mathcal{A}(x,\epsilon)} \ker\left(v_i^T\right),$$

where $[\bar{n}] := \{1,\ldots,\bar{n}\}$. Then for any $d \in \mathcal{R}(x,\epsilon)$, it holds that $v_i^T d = 0$, $i \in [\bar{n}]\backslash\mathcal{A}(x,\epsilon)$. Define the function

$$Q_\epsilon(x) := f(x) + h(c(x)) + \sum_{i\in\mathcal{A}(x,\epsilon)} |v_i^T x|^p.$$

Note that $|v_i^T x|^p$, $i \in \mathcal{A}(x,\epsilon)$, is differentiable at x, and Q_ϵ is a continuous lower approximation to Q. Also define

$$\psi_Q^{\epsilon,\delta}(x) := Q_\epsilon(x) - \min_{\substack{x+d\in\mathcal{F}\\ d\in\mathcal{R}(x,\epsilon),\, \|d\|\le\delta}} T_{Q_\epsilon}(x,d) \tag{2.1}$$

with

$$T_{Q_\epsilon}(x,d) := f(x) + \nabla f(x)^T d + h(c(x)+J(x)d) + \sum_{i\in\mathcal{A}(x,\epsilon)} (|v_i^T x|^p + \nabla(|v_i^T x|^p)^T d),$$

where $J(x) = (\nabla c_1(x),\ldots,\nabla c_r(x))^T$. Here, T_{Q_ϵ} is a convex approximation to Q_ϵ, obtained through linearization of smooth functions w.r.t. d, i.e., $f(x+d)$, $c(x+d)$ and $|v_i^T(x+d)|^p, i \in \mathcal{A}(x,\epsilon)$. The function $\psi_Q^{\epsilon,\delta}$ plays a crucial role in characterizing the optimality condition of a local minimizer of (1.1). It represents the maximum reduction of T_{Q_ϵ} within a neighborhood of current iterate. Intuitively, when current iterate x is a local minimizer of (1.1) and $\epsilon = 0$, around x there is no feasible point that can yield a greater reduction in the function value. By [8, Lemma 3.2] and [11, Theorem 2.1] we obtain the following lemma.

LEMMA 2.1. *Let x_* be a local minimizer of* (1.1). *Then there exists $\bar{\delta} \in (0,1]$ such that for any $\delta \in (0,\bar{\delta}]$, $\psi_Q^{0,\delta}(x_*) = 0$.*

Proof. As x_* is a local minimizer of (1.1), there exists $\delta_1 > 0$ such that x_* is a global minimizer of (1.1) on $\mathcal{B}(x_*,\delta_1)\cap\mathcal{F}$. Let

$$\delta_2 = \min\Big\{1, \min_{i\in\mathcal{A}(x_*,0)} |v_i^T x_*|\Big\}.$$

Obviously, $\delta_2 \in (0,1]$. Note that there exists $\bar{\delta} \in (0,\min(\delta_1,\delta_2))$ such that for any $x_* + d$ in the ball $\mathcal{B}(x_*,\bar{\delta})$,

$$|v_i^T(x_*+d)| \ge |v_i^T x_*| - |v_i^T d| \ge \delta_2 - \bar{\delta} > 0, \quad i \in \mathcal{A}(x_*,0).$$

Then $\sum_{i\in\mathcal{A}(x_*,0)} |v_i^T x|$ is continuously differentiable in $\mathcal{B}(x_*,\bar{\delta})$. Moreover, since h is Lipschitz continuous over $\mathcal{F}$ and x_* is the global minimizer of (1.1) on $\mathcal{B}(x_*,\bar{\delta})\cap\mathcal{F}$, it holds that for any $\delta \in (0,\bar{\delta}]$,

$$\begin{aligned}
Q(x_*) &= \min_{x_*+d\in\mathcal{F},\|d\|\le\delta} f(x_*+d) + h(c(x_*+d)) + \|V(x_*+d)\|_p^p \\
&\le \min_{\substack{x_*+d\in\mathcal{F}\\ d\in\mathcal{R}(x_*,0),\|d\|\le\delta}} f(x_*+d) + h(c(x_*+d)) + \|V(x_*+d)\|_p^p \\
&= \min_{\substack{x_*+d\in\mathcal{F}\\ d\in\mathcal{R}(x_*,0),\|d\|\le\delta}} f(x_*+d) + h(c(x_*+d)) + \sum_{i\in\mathcal{A}(x_*,0)} |v_i^T(x_*+d)|^p.
\end{aligned}$$

Note that the equality in above relations can be reachable at $d = 0$. Thus 0 is a global minimizer of the problem

$$\min_{\substack{x_*+d\in\mathcal{F}\\ d\in\mathcal{R}(x_*,0),\, \|d\|\le\delta}} f(x_*+d) + h(c(x_*+d)) + \sum_{i\in\mathcal{A}(x_*,0)} |v_i^T(x_*+d)|^p. \tag{2.2}$$

Then it yields from [8, Lemma 3.2] that $\psi_Q^{0,\delta}(x_*)=0$ which completes the proof. $\square$

We call $\bar{x}$ a **first-order stationary point of** (1.1), if $\psi_Q^{0,\delta}(\bar{x})=0$ for some $\delta\in(0,1]$.

Remark 2.2. We now show that if $\bar{x}$ is a first-order stationary point of (1.1) , i.e. $\psi_Q^{0,\delta}(\bar{x})=0$ for some $\delta\in(0,1]$, then $\bar{x}$ is a limiting stationary point for a practice example. The concept of a limiting stationary point for a proper lower semicontinuous function has been used in the study for non-Lipschitz continuous minimization [10]. We recall from [24, Definition 8.3] that for a proper lower semicontinuous function Φ, the limiting subdifferential is defined as

$$\partial\Phi(x):=\left\{v:\ \exists\, x^k\overset{\Phi}{\to}x,\ v^k\to v \text{ with } \liminf_{z\to x^k}\frac{\Phi(z)-\Phi(x^k)-\langle v^k,z-x^k\rangle}{\|z-x^k\|}\geq 0,\ \forall k\right\},$$

where $x^k\overset{\Phi}{\to}x$ means both $x^k\to x$ and $\Phi(x^k)\to\Phi(x)$. In [10], a first-order stationary condition using the limiting subdifferential for problem

$$\min\Theta(x):=\lambda[(\|Ax-b\|_2^2-\sigma^2)_+ + \|(Bx-h)_+\|_1]+\|x\|_p^p \tag{2.3}$$

is defined as

$$0\in\partial\lambda((\|Ax-b\|_2^2-\sigma^2)_+)+\partial\lambda\|(Bx-h)_+\|_1+\partial\|x\|_p^p, \tag{2.4}$$

where $A\in\mathbb{R}^{r\times n}, B\in\mathbb{R}^{l\times n}, b\in\mathbb{R}^r, h\in\mathbb{R}^r,\ p\in(0,1),\ \sigma\geq 0$, and $\lambda>0$. In [10], a point $\bar{x}$ is called a first-order stationary point of (2.3) if $\bar{x}$ satisfies (2.4). Let $\mathcal{A}(x)=\{i:|x_i|>0\}$. From Lemma 2.5 in [10], $\partial|t|^p=\mathbb{R}$ at $t=0$. Hence, the inclusion in (2.4) is trivial for $i\notin\mathcal{A}(x)$. Let

$$Q(x)=\lambda\left((\|Ax-b\|_2^2-\sigma^2)_+ + \|(Bx-h)_+\|_1\right)+\sum_{i\in\mathcal{A}(x)}|x_i|^p,$$

$$\begin{aligned}T_Q(x,d)=\lambda&\left((\|Ax-b\|_2^2-\sigma^2+2(Ax-b)^TAd)_+ + \|(Bx-h+Bd)_+\|_1\right)\\&+\sum_{i\in\mathcal{A}(x)}(|x_i|^p+p|x_i|^{p-1}\mathrm{sgn}(x_i)d_i)\end{aligned}$$

and

$$\psi_Q^{0,\delta}(x)=Q(x)-\min_{d\in\mathcal{R}(x),\|d\|\leq\delta}T_Q(x,d),\qquad \mathcal{R}(x)=\{d\in\mathbb{R}^n:\ e_i^Td=0, i\notin\mathcal{A}(x)\}.$$

Following the proof of Lemma 2.1, we can show that if $\bar{x}$ is a local minimizer of (2.3), then $\psi_Q^{0,\delta}(\bar{x})=0$ for some $\delta>0$. Now we show that if $\psi_Q^{0,\delta}(\bar{x})=0$ for some $\delta>0$, then $\bar{x}$ satisfies (2.4). Let $\bar{\delta}=\min_{i\in\mathcal{A}(\bar{x})}|\bar{x}_i|$. Then for any $\bar{x}+d\in\mathcal{B}(\bar{x},\delta)$ with $\delta\in(0,\bar{\delta})$, we have $|\bar{x}+d|_i\geq|\bar{x}|_i-|d|_i\geq\bar{\delta}-\delta>0, \forall\, i\in\mathcal{A}(\bar{x})$. Hence $\sum_{i\in\mathcal{A}(\bar{x})}|x_i|^p$ is differentiable in $\mathcal{B}(\bar{x},\delta)$. Moreover, we know that $(\|Ax-b\|_2^2-\sigma^2)_+$ and $\|(Bx-h)_+\|_1$ are directionally differentiable. Therefore, Q is directionally differentiable at $\bar{x}$ in the direction $d\in\mathcal{R}(\bar{x})$. Additionally, the directional derivative of Q at $\bar{x}$ in the direction $d\in\mathcal{R}(\bar{x})$ has the form

$$Q'(\bar{x};d)=\lambda[2v(\bar{x})(A\bar{x}-b)^TAd+u(\bar{x})^TBd]+\sum_{i\in\mathcal{A}(\bar{x})}p|\bar{x}_i|^{p-1}\mathrm{sgn}(\bar{x}_i)d_i,$$

where

$$v(\bar{x}) = \begin{cases} 1 & \text{if } \|A\bar{x}-b\|_2^2 > \sigma^2 \\ 0 & \text{if } \|A\bar{x}-b\|_2^2 < \sigma^2 \\ (\mathrm{sgn}((A\bar{x}-b)^T Ad))_+ & \text{if } \|A\bar{x}-b\|_2^2 = \sigma^2 \end{cases}$$

and

$$u_i(\bar{x}) = \begin{cases} 1 & \text{if } (B\bar{x}-h)_i > 0 \\ 0 & \text{if } (B\bar{x}-h)_i < 0 \\ (\mathrm{sgn}((Bd)_i))_+ & \text{if } (B\bar{x}-h)_i = 0, \quad i = 1, \dots, l. \end{cases}$$

Let $\hat{\delta} \in (0, \bar{\delta})$ such that

$$\hat{\delta} < \min\left\{ \frac{|\|A\bar{x}-b\|_2^2 - \sigma^2|}{\|(A\bar{x}-b)^T A\|_\infty}, \frac{|B\bar{x}-h|_i}{\|B\|_\infty} \right\}, \quad \text{for } \|A\bar{x}-b\|_2^2 - \sigma^2 \neq 0, (B\bar{x}-h)_i \neq 0.$$

Then it derives

$$T_Q(\bar{x}, d) = Q(\bar{x}) + Q'(\bar{x}; d), \qquad \forall d \in \mathcal{R}(\bar{x}),\ \|d\| \leq \hat{\delta}.$$

From $\psi_Q^{0,\hat{\delta}}(\bar{x}) = 0$, we have

$$0 = Q(\bar{x}) - \min_{d \in \mathcal{R}(\bar{x}), \|d\| \leq \hat{\delta}} T_Q(\bar{x}, d) = - \min_{d \in \mathcal{R}(\bar{x}), \|d\| \leq \hat{\delta}} Q'(\bar{x}; d),$$

which implies $Q'(\bar{x}; d) \geq 0$ for any $d \in \mathcal{R}(\bar{x})$. From $\Theta(\bar{x}+d) \geq Q(\bar{x}+d)$ for $d \in \mathbb{R}^n$ and $\Theta(\bar{x}) = Q(\bar{x})$, the subderivative function $\mathrm{d}\Theta(\bar{x})$ satisfies

$$\mathrm{d}\Theta(\bar{x})(d) = \liminf_{\substack{t \downarrow 0 \\ d' \to d}} \frac{\Theta(\bar{x}+td') - \Theta(\bar{x})}{t} \geq \liminf_{\substack{t \downarrow 0 \\ d' \to d}} \frac{Q(\bar{x}+td') - Q(\bar{x})}{t}.$$

Hence $\mathrm{d}\Theta(\bar{x})(d) \geq 0$ for $d \in \mathcal{R}(\bar{x})$ and $\mathrm{d}\Theta(\bar{x})(d) = +\infty$ for $d \notin \mathcal{R}(\bar{x})$. By [24, Exercise 8.4], we find that 0 is in the regular subdiffrential of Θ at $\bar{x}$, and thus by [24, Definition 8.3, Exercise 10.10], the inclusion in (2.4) holds at $\bar{x}$.

We now present the definition of an (ϵ, δ)-approximate first-order stationary point of (1.1).

Definition 2.3. *Given $\epsilon > 0$, we call $x \in \mathcal{F}$ an (ϵ, δ)-approximate first-order stationary point of* (1.1)*, if $\psi_Q^{\epsilon,\delta}(x) \leq \epsilon\delta$ for some $\delta \in (0, 1]$.*

The concept of (ϵ, δ)-approximate first-order stationary points has been used in [6, 7, 11], which generalizes the concept of ϵ-approximate first-order stationary points with $\delta = 1$ in some papers, e.g. [5, 12, 17]. Our definitions of first-order stationary point and (ϵ, δ)-approximate first-order stationary point are based on the concepts in [5, 6, 7, 8, 11, 12, 17] and related articles. In Lemma 2.1, we show that a local minimizer x^* of (1.1) is a $(0, \delta)$-approximate first-order stationary point of (1.1) for some $\delta > 0$, which implies that x^* is an (ϵ, δ)-approximate first-order stationary point of (1.1) for $\epsilon > 0$. Within a δ-ball centered at an (ϵ, δ)-approximate first-order stationary point, the maximum reduction of the approximation model does not exceed $\epsilon\delta$. In practice, the choice of (ϵ, δ) depends on the users' need for the quality of a computed solution. For each k, let x_k be an (ϵ_k, δ_k)-approximate first-order stationary point of (1.1) for some δ_k with $1 \geq \delta_k > 0$ and $\epsilon_k > 0$. If $\{\delta_k\}$ has a uniform positive lower bound as $\epsilon_k \to 0$, following the proof of [11, Theorem 2.2] we can obtain that any cluster point of $\{x_k\}$ is a first-order stationary point of (1.1).

In the following context, we consider $\epsilon > 0$. We now prepare for the design of the main algorithm. The main step of the algorithm is to construct a model function to predict the behavior of the objective function Q at current iterate x along a direction s. For the non-Lipschitz regularizer in the objective function, we focus on indices in $\mathcal{A}(x,\epsilon)$ and discard those close to non-Lipschitz continuity. We define the following Lipschitz continuous approximation of $|v_i^T(x+s)|^p$ in a similar approach in [11] and [12]:

$$m_i(x,s) := |v_i^T x|^p + p|v_i^T x|^{p-1}\left(|v_i^T(x+s)| - |v_i^T x|\right), \quad i \in \mathcal{A}(x,\epsilon). \tag{2.5}$$

Supposing that $v_i^T(x+s) \neq 0$, $i \in \mathcal{A}(x,\epsilon)$, as analyzed in [11], m_i is the first-order Taylor's expansion of $|v_i^T x + \zeta_i \frac{v_i^T x}{|v_i^T x|}|^p$ expressed as a function of the scalar $\zeta_i := |v_i^T(x+s)| - |v_i^T x| \geq -|v_i^T s|$. Regarding the smooth function f, the calculation of exact first-order derivatives of f can be expensive sometimes even impossible in many scenarios. We can only get access to approximate gradients of f. For ease of notations, given x_k and s_k we denote g_k as an approximation to ∇f at x_k, and

$$\mathcal{A}_k = \mathcal{A}(x_k,\epsilon),\ \mathcal{R}_k = \mathcal{R}(x_k,\epsilon),\ c_k = c(x_k),\ J_k = J(x_k) \text{ and } s_i^k = v_i^T s_k$$

for $i \in [\bar{n}]$. Due to existence of the convex but possibly nonsmooth function h, we design the following proximal type subproblem at kth iteration:

$$\min_{\substack{x_k+s\in\mathcal{F}\\ s\in\mathcal{R}_k}} \quad m(x_k,s) := g_k^T s + h(c_k + J_k s) + \sum_{i\in\mathcal{A}_k} m_i(x_k,s) + \frac{1}{2\eta}\|s\|^2, \tag{2.6}$$

where $\eta > 0$ is a proximal parameter. It is worth noting that subproblem (2.6) is a strongly convex minimization problem over a convex set, thus it admits a unique global minimizer. Note that resolution of (2.6) only involves matrx-vector products and does not affect the evaluations of (inexact) derivatives of f, thus has no impact on the iteration complexity and oracle complexity of the proposed algorithm. Moreover, when $\mathcal{F}$ and h exhibit polyhedral structures, for example, $\mathcal{F} = [b_l, b_u] \subseteq \mathbb{R}^n$ with $-b_l, b_u \in \mathbb{R}^n_+$, and $h(\cdot) = \|(\cdot)_+\|_1$, by introducing $\bar{z} = (c_k + J_k s)_+ \in \mathbb{R}^r$, (2.6) is equivalent to the following linearly constrained convex program:

$$\begin{aligned} \min_{s,\bar{z}} \quad & g_k^T s + e^T\bar{z} + \sum_{i\in\mathcal{A}_k} p|v_i^T x_k|^{p-1}|v_i^T(x_k+s)| + \frac{1}{2}\|s\|^2 \\ \text{s.t.} \quad & b_l \leq x_k + s \leq b_u, \quad 0 \leq \bar{z}, c_k + J_k s \leq \bar{z}, \quad v_i^T s = 0, i \notin \mathcal{A}_k, \end{aligned}$$

where $e = (1,1,\ldots,1)^T \in \mathbb{R}^r$. Numerous state-of-the-art approaches have been extensively studied for solving linearly constrained convex program in the literature.

In theoretical analysis, however, an inexact solution of (2.6) can be enough. Specifically, we solve (2.6) to look for s_k with $m(x_k,s_k) < m(x_k,0)$ such that the near optimality is achieved in that

$$\psi_m^{\epsilon,\delta}(x_k,s_k) \leq \min\left\{\theta\epsilon, p \min_{i\in\mathcal{A}(x_k+s_k,\epsilon)} |v_i^T(x_k+s_k)|\right\}\delta, \quad \text{for some } \delta \in (0,1], \tag{2.7}$$

where $\theta \in (0,1)$ and

$$\psi_m^{\epsilon,\delta}(x_k,s_k) := h(c_k + J_k s_k) - \min_{\substack{x_k+s_k+d\in\mathcal{F}\\ d\in\mathcal{R}(x_k+s_k,\epsilon),\ \|d\|\leq\delta}} T_m(x_k,s_k;d)$$

with $m_0(x_k, s) := g_k^T s + \frac{1}{2\eta}\|s\|^2$ and

$$T_m(x_k, s; d) := h(c_k + J_k(s+d)) + \nabla_s m_0(x_k, s)^T d + \sum_{i \in \mathcal{A}(x_k+s,\epsilon)} \nabla_s m_i(x_k, s)^T d.$$

It is noteworthy that $\psi_m^{\epsilon,\delta}$ describes the potential maximum reduction of T_m within a neighborhood of s_k with radius δ. This measure is defined in a similar way to that in Definition 2.3. When the reduction is below a certain level, s_k is regarded as an inexact minimizer of (2.4). Moreover, by the definition of $\mathcal{R}_k$, for any $i \in [\bar{n}]\backslash\mathcal{A}_k$, $v_i^T s_k = 0$, thus $v_i^T(x_k + s_k) = v_i^T x_k$. That is, once $|v_i^T x_k| \le \epsilon$ for some $i \in [\bar{n}]$, the value of $v_i^T(x_k + s_k)$ will be fixed and the remaining minimization will be carried out on $\mathcal{R}(x_k + s_k, \epsilon)$. Therefore, the following relations hold:

$$\mathcal{R}_k^+ := \mathcal{R}(x_k + s_k, \epsilon) \subseteq \mathcal{R}_k, \quad \mathcal{A}_k^+ := \mathcal{A}(x_k + s_k, \epsilon) \subseteq \mathcal{A}_k. \tag{2.8}$$

We are now ready to present the main algorithm framework for proximal inexact gradient methods for (1.1) as Algorithm 2.1.

Algorithm 2.1

Input: $x_0 \in \mathcal{F}, \epsilon \in (0,1], \eta > 0, \bar{\beta} \in (0, \bar{w})$ with $\bar{w} \in (0,1)$, $s_{-1} = 0$.
1: **for** $k = 0, 1, \ldots,$ **do**
2: Obtain g_k from **InexactOracle**.
3: Solve (2.6) to find an approximate minimizer s_k with $m(x_k, s_k) < m(x_k, 0)$ satisfying (2.7), then go to Step 5. If the solution of (2.6) is zero, then go to Step 4.
4: Set $s_k = 0$. If $s_{k-1} = 0$, terminate and return x_k; otherwise, go to Step 5.
5: Set $x_{k+1} = x_k + s_k$. If $\|s_k\| + \|s_{k-1}\| \le \bar{\beta}\epsilon$ and $\mathcal{A}_k\backslash\mathcal{A}_{k+1} = \emptyset$, terminate and return x_{k+1}.
6: $k := k + 1$.
7: **end for**

Remark 2.4. In Algorithm 2.1 two termination criteria are employed. One is $s_{k-1} = s_k = 0$ in Step 4. In this case, similar to [6, 11] we terminate the algorithm and return x_k. It will be shown in Lemma 3.1 that x_k is an (ϵ, δ)-approximate first-order stationary point of (1.1) for some $\delta \in (0, 1]$. On the other hand, if $\mathcal{A}_{k+1} = \mathcal{A}_k$ (and hence $\mathcal{R}_{k+1} = \mathcal{R}_k$) and $\|s_k\| + \|s_{k-1}\|$ is sufficiently small and $\mathcal{A}_k\backslash\mathcal{A}_{k+1} = \emptyset$ then there is no i s.t. $|v_i^T x_k| > \epsilon$ but $|v_i^T(x_k + s_k)| < \epsilon$, we return x_{k+1} and will prove that x_{k+1} is an approximate first-order stationary point when $s_k = 0$ in Lemma 3.1 and when $s_k \ne 0$ in Lemma 3.6, respectively. In addition, as $s_k \in \mathcal{R}_k$ for any $k \ge 1$, it follows from (2.8) that $\mathcal{A}_{k+1}\backslash\mathcal{A}_k = \emptyset$ for any $k \ge 1$. Moreover, in Algorithm 2.1 we obtain inexact gradient g_k through calling the subroutine **InexactOracle**, which may adopt different ways to generate an inexact gradient of f at the inquiry iterate x_k. So we simply omit the required inputs by **InexactOracle** here and specify them when necessary.

In the following, we denote the unique global minimizer of (2.6) by s_k^*. If $s_k^* \ne 0$, it obviously holds that $m(x_k, s_k^*) < m(x_k, 0)$. Moreover, we can guarantee that $s_k = s_k^*$ satisfies (2.7) for some $\delta \in (0, 1]$, as shown in the lemma below.

LEMMA 2.5. *Suppose that* $s_k^* \ne 0$. *Then there exists* $\underline{\mu}_k \in (0, 1]$ *such that* (2.7) *holds for* $s_k = s_k^*$ *and any* $\delta \in (0, \underline{\mu}_k]$.

Proof. Consider the auxiliary problem

$$\min_{\substack{x_k+s_k^*+d\in\mathcal{F}\\ d\in\mathcal{R}(x_k+s_k^*,\epsilon)}} \quad h(c_k+J_k(s_k^*+d))+m_0(x_k,s_k^*+d)+\sum_{i\in\mathcal{A}(x_k+s_k^*,\epsilon)} m_i\left(x_k,s_k^*+d\right). \tag{2.9}$$

Due to the strong convexity, (2.9) has a unique global minimizer, which is denoted by $\bar{s}_k$. As $\bar{s}_k\in\mathcal{R}(x_k+s_k^*,\epsilon)\subseteq\mathcal{R}_k$, we have $m_i(x_k,s_k^*)=m_i(x_k,s_k^*+\bar{s}_k)$ for any $i\in\mathcal{A}_k\backslash\mathcal{A}(x_k+s_k^*,\epsilon)$. Then it yields that

$$\begin{aligned}
&m(x_k,s_k^*+\bar{s}_k)\\
&=h(c_k+J_k(s_k^*+\bar{s}_k))+m_0(x_k,s_k^*+\bar{s}_k)+\sum_{i\in\mathcal{A}(x_k+s_k^*,\epsilon)} m_i(x_k,s_k^*+\bar{s}_k)\\
&\quad+\sum_{i\in\mathcal{A}_k\backslash\mathcal{A}(x_k+s_k^*,\epsilon)} m_i(x_k,s_k^*+\bar{s}_k)\\
&\le h(c_k+J_ks_k^*)+m_0(x_k,s_k^*)+\sum_{i\in\mathcal{A}(x_k+s_k^*,\epsilon)} m_i(x_k,s_k^*)\\
&\quad+\sum_{i\in\mathcal{A}_k\backslash\mathcal{A}(x_k+s_k^*,\epsilon)} m_i(x_k,s_k^*)=m(x_k,s_k^*),
\end{aligned}$$

where the inequality follows from the optimality of $\bar{s}_k$.

Due to the optimality and uniqueness of s_k^* as the global minimizer of (2.6), we obtain $\bar{s}_k=0$. Thus 0 is the global minimizer of (2.9). Then for any $d\in\mathcal{R}(x_k+s_k^*,\epsilon)$ satisfying $x_k+s_k^*+d\in\mathcal{F}$, it holds that

$$\begin{aligned}
&g_k^Ts_k^*+\frac{1}{2\eta}\|s_k^*\|^2+h(c_k+J_ks_k^*)+\sum_{i\in\mathcal{A}(x_k+s_k^*,\epsilon)} m_i(x_k,s_k^*)\\
&\le g_k^T(s_k^*+d)+\frac{1}{2\eta}\|s_k^*+d\|^2+h(c_k+J_k(s_k^*+d))+\sum_{i\in\mathcal{A}(x_k+s_k^*,\epsilon)} m_i(x_k,s_k^*+d)
\end{aligned}$$

which yields

$$\begin{aligned}
&h(c_k+J_ks_k^*)-h(c_k+J_k(s_k^*+d))-g_k^Td-\frac{1}{\eta}(s_k^*)^Td-\sum_{i\in\mathcal{A}(x_k+s_k^*,\epsilon)}\nabla m_i(x_k,s_k^*)^Td\\
&\le\sum_{i\in\mathcal{A}(x_k+s_k^*,\epsilon)}(m_i(x_k,s_k^*+d)-m_i(x_k,s_k^*)-\nabla_s m_i(x_k,s_k^*)^Td)+\frac{1}{2\eta}\|d\|^2.
\end{aligned} \tag{2.10}$$

Note that there exists $\hat{\mu}_k$ such that for any $d\in\mathcal{R}(x_k+s_k^*,\epsilon)$ with $\|d\|\le\hat{\mu}_k$ and $x_k+s_k^*+d\in\mathcal{F}$,

$$\operatorname{sgn}(v_i^T(x_k+s_k^*+d))=\operatorname{sgn}(v_i^T(x_k+s_k^*))\ \text{ and }\ |v_i^T(x_k+s_k^*+d)|>\epsilon,\quad\forall i\in\mathcal{A}(x_k+s_k^*,\epsilon), \tag{2.11}$$

which together with (2.10) indicate that

$$\begin{aligned}
&h(c_k+J_ks_k^*)-h(c_k+J_k(s_k^*+d))-g_k^Td-\frac{1}{\eta}{s_k^*}^Td-\sum_{i\in\mathcal{A}(x_k+s_k^*,\epsilon)}\nabla_s m_i(x_k,s_k^*)^Td\\
&\le\frac{1}{2\eta}\|d\|^2.
\end{aligned}$$

Hence, by the definition of $\psi_m^{\epsilon,\delta}(x_k, s_k^*)$, there exists $\underline{\mu}_k \in (0, \min\{1, \hat{\mu}_k\}]$ such that for any $\delta \in (0, \underline{\mu}_k]$,

$$\psi_m^{\epsilon,\delta}(x_k, s_k^*) \le \frac{1}{2\eta}\delta^2 \le \min\Big\{\theta\epsilon, p \min_{i\in\mathcal{A}(x_k+s_k^*,\epsilon)} |v_i^T(x_k + s_k^*)|\Big\}\delta. \tag{2.12}$$

∎

Define the set $\mathcal{S}_k := \{s : x_k + s \in \mathcal{F}\} \cap \{s : s - s_k^* \in \mathcal{R}(x_k + s_k^*, \epsilon)\}$. Obviously $s_k^* \in \mathcal{S}_k$. Without loss of generality, we assume in the following that $\mathcal{S}_k \backslash \{s_k^*\} \neq \emptyset$.

Lemma 2.6. *Suppose that $s_k^* \neq 0$. Then there exist $\tilde{\mu}_k, \bar{\mu}_k \in (0, 1]$ such that for any $\delta \in (0, \tilde{\mu}_k]$ and any $s \in S_k \cap \mathcal{B}(s_k^*, \bar{\mu}_k)$, we have*

$$m(x_k, s) < m(x_k, 0) \quad \textit{and} \quad \psi_m^{\epsilon,\delta}(x_k, s) \le \min\Big\{\theta\epsilon, p \min_{i\in\mathcal{A}(x_k+s,\epsilon)} \big|v_i^T(x_k + s)\big|\Big\}\delta. \tag{2.13}$$

Proof. Note that if $s_k^* \neq 0$, there exists $\bar{\mu}_k \in (0, 1]$ such that for any $s \in \mathcal{B}(s_k^*, \bar{\mu}_k)$, $m(x_k, s) < m(x_k, 0)$ and $\mathcal{A}(x_k+s_k^*, \epsilon) \subseteq \mathcal{A}(x_k+s, \epsilon)$. Hence, for any $s \in \mathcal{S}_k \cap \mathcal{B}(s_k^*, \bar{\mu}_k)$,

$$\mathcal{A}(x_k + s_k^*, \epsilon) = \mathcal{A}(x_k + s, \epsilon) \subseteq \mathcal{A}_k. \tag{2.14}$$

For any given $s \in \mathcal{S}_k \cap \mathcal{B}(s_k^*, \bar{\mu}_k)$, we define $\mathcal{F}_s := \{d : x_k + s + d \in \mathcal{F}\}$ which is obviously convex due to the convexity of $\mathcal{F}$. For any $d \in \mathcal{F}_s$, we denote its projection onto $\mathcal{F}_{s_k^*}$ as $\bar{d}$. If $d = \bar{d}$, then set $d_1 = d$. Otherwise, as $x_k + s_k^* + d \notin \mathcal{F}$, there exists $d_1 \in \mathcal{F}_{s_k^*}$ such that $x_k + s_k^* + d_1$ is the projection of $x_k + s_k^* + d$ onto $\mathcal{F}$. Then it follows from definition of the projection operator and $x_k + s + d \in \mathcal{F}$ that $\|d - \bar{d}\| \le \|d - d_1\|$ and

$$\|(x_k + s_k^* + d) - (x_k + s_k^* + d_1)\| \le \|(x_k + s_k^* + d) - (x_k + s + d)\| = \|s_k^* - s\|,$$

thus

$$\|d - \bar{d}\| \le \|s_k^* - s\|. \tag{2.15}$$

Then by definition of $T_m(x_k, s; d)$ and (2.14) we obtain that for any $d \in \mathcal{F}_s$,

$$\begin{aligned}
&h(c_k + J_k s) - T_m(x_k, s; d)\\
&= h(c_k + J_k s_k^*) + h(c_k + J_k s) - h(c_k + J_k s_k^*) - \Big[h(c_k + J_k(s_k^* + d))\\
&\quad + h(c_k + J_k(s + d)) - h(c_k + J_k(s_k^* + d)) + d^T\nabla_s m_0(x_k, s_k^*)\\
&\quad + d^T(\nabla_s m_0(x_k, s) - \nabla_s m_0(x_k, s_k^*)) + \sum_{i\in\mathcal{A}(x_k+s_k^*,\epsilon)} d^T\nabla_s m_i(x_k, s_k^*)\\
&\quad + \sum_{i\in\mathcal{A}(x_k+s_k^*,\epsilon)} d^T\nabla_s m_i(x_k, s) - \sum_{i\in\mathcal{A}(x_k+s_k^*,\epsilon)} d^T\nabla_s m_i(x_k, s_k^*)\Big]\\
&= h(c_k + J_k s_k^*) - \Big[h(c_k + J_k(s_k^* + d)) + d^T\nabla_s m_0(x_k, s_k^*)\\
&\quad + \sum_{i\in\mathcal{A}(x_k+s_k^*,\epsilon)} d^T\nabla_s m_i(x_k, s_k^*)\Big] + \Gamma_k\\
&= h(c_k + J_k s_k^*) - T_m(x_k, s_k^*; d) + \Gamma_k\\
&= h(c_k + J_k s_k^*) - T_m(x_k, s_k^*; \bar{d}) + \Gamma_k + T_m(x_k, s_k^*; \bar{d}) - T_m(x_k, s_k^*; d),
\end{aligned}$$

where

$$\Gamma_k = h(c_k + J_k s) - h(c_k + J_k s_k^*) - (h(c_k + J_k(s + d)) - h(c_k + J_k(s_k^* + d)))$$
$$- d^T(\nabla_s m_0(x_k, s) - \nabla_s m_0(x_k, s_k^*))$$
$$- \sum_{i\in\mathcal{A}(x_k+s_k^*,\epsilon)} d^T(\nabla_s m_i(x_k, s) - \nabla_s m_i(x_k, s_k^*)).$$

Note that, on the one hand, by the definition of m_0,

$$\|\nabla_s m_0(x_k, s) - \nabla_s m_0(x_k, s_k^*)\| \le \frac{1}{\eta}\|s - s_k^*\|,$$

while on the other hand, by (2.5) and (2.14),

$$\nabla_s m_i(x_k, s) = p|v_i^T x_k|^{p-1}\mathrm{sgn}(v_i^T x_k)v_i = \nabla_s m_i(x_k, s_k^*), \quad \forall i \in \mathcal{A}(x_k + s_k^*, \epsilon).$$

Recall that h is Lipschitz continuous over $\mathcal{F}$. It together with the boundedness of J_k derives

$$\Gamma_k = \mathcal{O}(\|s - s_k^*\|). \tag{2.16}$$

Besides, it indicates from definition of T_m that

$$T_m(x_k, s_k^*; \bar{d}) - T_m(x_k, s_k^*; d) = \mathcal{O}(\|d - \bar{d}\|) = \mathcal{O}(\|s - s_k^*\|).$$

Therefore, there exists $\tilde{\mu}_k \in (0, \min\{\underline{\mu}_k, \bar{\mu}_k\})$ such that $\tilde{\mu}_k + \tilde{\mu}_k^{(1+\varrho)} < \underline{\mu}_k$ with $\varrho > 0$, and for any $\delta \in (0, \tilde{\mu}_k]$ and $s \in \mathcal{S}_k \cap \mathcal{B}(s_k^*, \delta^{1+\varrho})$, the following relations can be derived:

$$h(c_k + J_k s) - \min_{\substack{x_k+s+d\in\mathcal{F} \\ d\in\mathcal{R}(x_k+s,\epsilon),\, \|d\|\le\delta}} T_m(x_k, s; d)$$
$$\le h(c_k + J_k s_k^*) - \min_{\substack{x_k+s_k^*+\bar{d}\in\mathcal{F} \\ \bar{d}\in\mathcal{R}(x_k+s_k^*,\epsilon),\, \|\bar{d}\|\le\delta+\delta^{1+\varrho}}} T_m(x_k, s_k^*; \bar{d}) + \mathcal{O}(\|s - s_k^*\|)$$
$$\le \psi_m^{\epsilon,\delta+\delta^{1+\varrho}}(x_k, s_k^*) + \mathcal{O}(\|s - s_k^*\|)$$
$$\le \frac{1}{2\eta}(\delta + \delta^{1+\varrho})^2 + \mathcal{O}(\delta^{1+\varrho}) \le \min\left\{\theta\epsilon, p \min_{i\in\mathcal{A}(x_k+s_k^*)} |v_i^T(x_k + s_k^*)|\right\}\delta, \tag{2.17}$$

where $\underline{\mu}_k$ is introduced in Lemma 2.5, the first inequality is due to $\|\bar{d}\| \le \|d\| + \|s - s_k^*\| \le \delta + \delta^{1+\varrho}$ and the third inequality follows from (2.12). The proof is completed.□

3. Oracle complexity. In this section, we will analyze the oracle complexity of Algorithm 2.1 in terms of the total number of inexact gradient evaluations until the algorithm terminates. In the following, we use $\mathcal{K}$ to denote the set of all iteration indices until the termination of Algorithm 2.1. Let $\{x_k\}$ be the iterate sequence generated during the algorithm. Since f and c are Lipschitz continuously differentiable and h is Lipschitz continuous over $\mathcal{F}$, there exist positive constants $M_F, \kappa, L_f, L_h, L_c^0, L_c^1$ such that for any $x, y \in \mathcal{F}$, $\|x\| \le M_F$ and $\|\nabla f(x)\| \le \kappa$ and

$$\|\nabla f(x) - \nabla f(y)\| \le L_f\|x - y\|, \quad |h(x) - h(y)| \le L_h\|x - y\|,$$
$$\|c(x) - c(y)\| \le L_c^0\|x - y\|, \quad \|\nabla c(x) - \nabla c(y)\| \le L_c^1\|x - y\|.$$

To derive desired theoretical properties of Algorithm 2.1, we lay out the following assumption on gradient approximations returned by **InexactOracle**.

ASSUMPTION 3.1. *For any $k \in \mathcal{K}$, the gradient approximation g_k satisfies*

$$\|g_k - \nabla f(x_k)\| \le \beta \max\left\{\bar{L}\min(\|s_{k-1}\|, D), \epsilon\right\}, \tag{3.1}$$

where $\beta \in (0, \frac{1-\theta}{17}), D > 0$ and $\bar{L} \in (0, \bar{\beta}^{-1})$.

The parameter θ in Assumption 3.1 was introduced initially in (2.7). And Assumption 3.1 ensures that g_k is uniformly upper bounded, namely,

$$\|g_k\| \le \chi := \kappa + \beta \max\left\{\bar{L}D, \epsilon\right\} \quad \text{for any } k \in \mathcal{K}. \tag{3.2}$$

We will show in Lemma 3.1 that x_k is an (ϵ, δ)-approximate first-order stationary point of (1.1) when Algorithm 2.1 terminates in Step 4 or in Step 5 with $s_k = 0$. When Algorithm 2.1 terminates in Step 5 with $s_k \neq 0$, we will show in Lemma 3.6 that the output x_{k+1} is an approximate first-order stationary point of (1.1).

LEMMA 3.1. *Suppose that Algorithm 2.1 terminates in Step 4 or in Step 5 with $s_k = 0$. Then x_k is an (ϵ, δ)-approximate first-order stationary point of* (1.1) *for some $\delta \in (0, 1]$.*

Proof. Whenever Algorithm 2.1 terminates in Step 4 or in Step 5 with $s_k = 0$, it holds that $s_k = 0$ and $\|s_{k-1}\| \le \bar{\beta}\epsilon$. As $s_k = 0$, by the algorithmic framework there is no step s_k with $m(x_k, s_k) < m(x_k, 0)$ satisfying (2.7). Then it follows from Lemma 2.5 that $m(x_k, d) \ge m(x_k, 0)$ for any $d \in \mathcal{R}_k$ satisfying $x_k + d \in \mathcal{F}$. And by the definition of T_{Q_ϵ} the following equalities hold:

$$\begin{aligned}
&m(x_k, d) - m(x_k, 0)\\
&= h(c_k + J_k d) - h(c_k) + g_k^T d + \frac{1}{2\eta}\|d\|^2 + \sum_{i\in\mathcal{A}_k}\left[m_i(x_k, d) - m_i(x_k, 0)\right]\\
&= -\Big[f(x_k) + h(c_k) + \sum_{i\in\mathcal{A}_k}|v_i^T x_k|^p - T_{Q_\epsilon}(x_k, d)\Big] + (g_k - \nabla f(x_k))^T d\\
&\quad + \frac{1}{2\eta}\|d\|^2 + \sum_{i\in\mathcal{A}_k}(m_i(x_k, d) - m_i(x_k, 0) - (\nabla|v_i^T x|^p|_{x=x_k})^T d)\\
&= -(Q_\epsilon(x_k) - T_{Q_\epsilon}(x_k, d)) + (g_k - \nabla f(x_k))^T d + \frac{1}{2\eta}\|d\|^2\\
&\quad + \sum_{i\in\mathcal{A}_k}(m_i(x_k, d) - m_i(x_k, 0) - (\nabla|v_i^T x|^p|_{x=x_k})^T d).
\end{aligned}$$

Note that there exists $\bar{\delta} \in (0, 1]$ such that for all d with $\|d\| \le \bar{\delta}$, $\mathrm{sgn}(v_i^T(x_k + d)) = \mathrm{sgn}(v_i^T x_k)$ for any $i \in \mathcal{A}_k$, then

$$\begin{aligned}
m_i(x_k, d) - m_i(x_k, 0) &= p|v_i^T x_k|^{p-1}(|v_i^T(x_k + d)| - |v_i^T x_k|)\\
&= \mathrm{sgn}(v_i^T x_k)\cdot p|v_i^T x_k|^{p-1} v_i^T d\\
&= (\nabla(|v_i^T x|^p|)_{x=x_k})^T d \quad \text{for any } i \in \mathcal{A}_k.
\end{aligned}$$

It follows from $\|s_{k-1}\| \le \bar{\beta}\epsilon$ and Assumption 3.1 that $\|g_k - \nabla f(x_k)\| \le \beta\max\{\bar{L}\bar{\beta}, 1\}\epsilon$. Hence, by $\beta < (1-\theta)/(32M_F + 1)$ and $\bar{L} < \bar{\beta}^{-1}$ we can choose $\delta < \bar{\delta}$ sufficiently small such that for any $d \in \mathcal{R}_k$ satisfying $x_k + d \in \mathcal{F}$ and $\|d\| \le \delta$,

$$Q_\epsilon(x_k) - T_{Q_\epsilon}(x_k, d) \le (g_k - \nabla f(x_k))^T d + \frac{\|d\|^2}{2\eta} \le \beta\max\{\bar{L}\bar{\beta}, 1\}\epsilon\|d\| + \frac{\|d\|^2}{2\eta} \le \epsilon\delta,$$

which yields the conclusion by Definition 2.3. $\square$

In the following, we assume that Algorithm 2.1 does not terminate at kth iteration with $s_k = 0$. It derives from definitions of Q_ϵ and $\mathcal{A}_k$ that for any $x \in \mathcal{F}$,

$$Q_\epsilon(x) = Q(x) - \sum_{i \in [\bar{n}] \backslash \mathcal{A}_k} |v_i^T x|^p \geq Q^* - |[\bar{n}] \backslash \mathcal{A}_k| \epsilon^p \geq Q^* - \bar{n}\epsilon^p =: Q_\epsilon^*,$$

where Q^* is the lower bound of Q on $\mathcal{F}$. The lemma below provides an upper bound on the accumulated square of step lengths.

Lemma 3.2. *Suppose that $\eta < (L_f + L_h L_c^1)^{-1}$. Then it holds that*

$$\left(\frac{1}{2\eta} - \frac{L_f + L_h L_c^1}{2}\right) \sum_{k \in \mathcal{K}} \|s_k\|^2 \leq \sum_{k \in \mathcal{K}} (\nabla f(x_k) - g_k)^T s_k + Q_\epsilon(x_0) - Q_\epsilon^*. \tag{3.3}$$

Proof. It follows from (2.8) and Lipschitz continuity of ∇f, h and ∇c that

$$\begin{aligned}
&Q_\epsilon(x_k + s_k) - Q_\epsilon(x_k) \\
&= f(x_k + s_k) + h(c(x_k + s_k)) - f(x_k) - h(c_k) + \sum_{i \in \mathcal{A}_k^+} |v_i^T (x_k + s_k)|^p - \sum_{i \in \mathcal{A}_k} |v_i^T x_k|^p \\
&\leq h(c(x_k + s_k)) - h(c_k) + (\nabla f(x_k))^T s_k + \frac{L_f}{2}\|s_k\|^2 + \sum_{i \in \mathcal{A}_k} (m_i(x_k, s_k) - m_i(x_k, 0)) \\
&= h(c_k + J_k s_k) - h(c_k) + g_k^T s_k + \frac{1}{2\eta}\|s_k\|^2 + \sum_{i \in \mathcal{A}_k} (m_i(x_k, s_k) - m_i(x_k, 0)) \\
&\quad + (\nabla f(x_k) - g_k)^T s_k + \left(\frac{L_f}{2} - \frac{1}{2\eta}\right)\|s_k\|^2 + h(c(x_k + s_k)) - h(c_k + J_k s_k) \\
&\leq m(x_k, s_k) - m(x_k, 0) + (\nabla f(x_k) - g_k)^T s_k + \left(\frac{L_f + L_h L_c^1}{2} - \frac{1}{2\eta}\right)\|s_k\|^2,
\end{aligned} \tag{3.4}$$

where the first inequality is due to $\mathcal{A}_k^+ \subseteq \mathcal{A}_k$ and [11, Lemma 3.2] which shows that $m_i(x_k, s_k) \geq |v_i^T (x_k + s_k)|^p$ for $i \in \mathcal{A}_k$. Then (3.4) indicates from $m(x_k, 0) \geq m(x_k, s_k)$ that

$$Q_\epsilon(x_k + s_k) - Q_\epsilon(x_k) \leq (\nabla f(x_k) - g_k)^T s_k + \left(\frac{L_f + L_h L_c^1}{2} - \frac{1}{2\eta}\right)\|s_k\|^2. \tag{3.5}$$

Hence, summing up (3.5) over $k \in \mathcal{K}$ and by $Q_\epsilon(x) \geq Q_\epsilon^*$ for all $x \in \mathcal{F}$ implies (3.3). □

For a given $\mu > 0$ which is independent of ϵ, we define

$$\mathcal{O}_{k,\mu} := \left\{ i \in \mathcal{A}_k^+ : \min\left\{ |v_i^T x_k|, |v_i^T (x_k + s_k)| \right\} \geq \mu \right\}, \tag{3.6}$$

$$\bar{Q}_{k,\mu}(x) := f(x) + \sum_{i \in \mathcal{O}_{k,\mu}} m_i(x, 0), \tag{3.7}$$

$$\bar{T}_{k,\mu}(x, s) := f(x) + \nabla f(x)^T s + \sum_{i \in \mathcal{O}_{k,\mu}} m_i(x, s). \tag{3.8}$$

The following lemma characterizes the relation between derivatives of $\bar{Q}_{k,\mu}$ and $\bar{T}_{k,\mu}$.

Lemma 3.3. *It holds that for any $k \geq 1$,*

$$\|\nabla \bar{Q}_{k,\mu}(x_k + s_k) - \nabla_s \bar{T}_{k,\mu}(x_k, s_k)\| \leq L(\mu)\|s_k\|, \tag{3.9}$$

where $L(\mu) := L_f + \frac{p(2-p)}{1-p}\mu^{p-2}$.

Proof. By definitions of $Q_{k,\mu}$ and $T_{k,\mu}$, it is easy to obtain

$$\begin{aligned}&\|\nabla \bar{Q}_{k,\mu}(x_k+s_k)-\nabla_s \bar{T}_{k,\mu}(x_k,s_k)\|\\ &\le \|\nabla f(x_k+s_k)-\nabla f(x_k)\|+\sum_{i\in\mathcal{O}_{k,\mu}}\|\nabla(|v_i^Tx|^p)|_{x=x_k+s_k}-\nabla_s m_i(x_k,s_k)\|.\end{aligned}\tag{3.10}$$

On the one hand, the Lipschitz continuity of ∇f ensures

$$\|\nabla f(x_k+s_k)-\nabla f(x_k)\|\le L_f\|s_k\|.\tag{3.11}$$

On the other hand, it follows from [11, Lemma 5.2] that

$$\begin{aligned}\sum_{i\in\mathcal{O}_{k,\mu}}\|\nabla(|v_i^Tx|^p)|_{x=x_k+s_k}-\nabla_s m_i(x_k,s_k)\| &\le \frac{p(2-p)}{1-p}\mu^{p-2}|v_i^Ts_k|\\ &\le \frac{p(2-p)}{1-p}\mu^{p-2}\|s_k\|.\end{aligned}\tag{3.12}$$

Hence, plugging (3.11) and (3.12) into (3.10) leads to the conclusion. ∎

To proceed, we assume the following assumption holds.

ASSUMPTION 3.2. *For problem* (1.1)*, it holds that*

$$0\in v_i^T\mathcal{F},\quad \mathrm{Proj}_{\ker(v_i^T)}\mathcal{F}\subseteq\mathcal{F},\quad i=1,...,\bar{n}.$$

Under Assumption 3.2 , it is easy to check that for any $x\in\mathcal{F}$, $(I-v_iv_i^T)x\in\mathcal{F}$ due to $\|v_i\|=1$, for any $i=1,\dots,\bar{n}$. A simple example of $\mathcal{F}$ satisfies Assumption 3.2 is that $\mathcal{F}=\{x|\underline{l}\le Vx\le\underline{u}\}$, where $-\underline{l},\underline{u}\in\mathbb{R}_+^{\bar{n}}$.

We now set ω satisfying

$$0<\omega<\min\left\{6^{\frac{1}{p-1}},\left(\frac{p}{2(L_hL_c^0+\chi+2M_F/\eta)}\right)^{\frac{1}{1-p}}\right\}.\tag{3.13}$$

Next lemma characterizes properties of points that are close to singularity.

LEMMA 3.4. *Suppose $\epsilon<\omega$, $|v_i^Tx_k|<\omega$ for some $i\in[\bar{n}]$ and $s_k\ne 0$. Then it holds that $|v_i^T(x_k+s_k)|\le\epsilon$ or $|v_i^T(x_k+s_k)|\ge\omega$.*

Proof. It is straightforward to obtain the conclusion if $i\in[\bar{n}]\backslash\mathcal{A}_k^+$. We now assume by contradiction that $|v_i^Tx_k|<\omega$ and

$$|v_i^T(x_k+s_k)|\in(\epsilon,\omega)\quad\text{ for some } i\in\mathcal{A}_k^+.\tag{3.14}$$

Besides, by (2.7) there exists $\delta_k\in(0,1]$ such that $\psi_m^{\epsilon,\delta_k}(x_k,s_k)\le p|v_i^T(x_k+s_k)|\delta_k$. As $v_i^T, i\in[\bar{n}]$ are orthogonal, by the definition of $\mathcal{R}_k^+$ and (2.8) we have $\mathcal{R}_{\{i\}}:=\mathrm{span}\{v_i\}\subseteq\mathcal{R}_k^+$. Consider the following minimization problem:

$$\min_{\substack{x_k+s_k+d\in\mathcal{F}\\ d\in\mathcal{R}_{\{i\}},\|d\|\le\delta_k}} q_k(d)\tag{3.15}$$

with

$$q_k(d):=h(c_k+J_k(s_k+d))-h(c_k+J_ks_k)+d^T\nabla_s m_0(x_k,s_k)+\sum_{i\in\mathcal{A}_k^+}d^T\nabla_s m_i(x_k,s_k).$$

It is worthy to note that $d = 0$ is a feasible point of (3.15). Then the optimal function value of (3.15) must be nonpositive, thus

$$\left| \min_{\substack{x_k+s_k+d\in\mathcal{F} \\ d\in\mathcal{R}_{\{i\}},\|d\|\le\delta_k}} q_k(d) \right| \le \left| \min_{\substack{x_k+s_k+d\in\mathcal{F} \\ d\in\mathcal{R}_k^+,\|d\|\le\delta_k}} q_k(d) \right| = \psi_m^{\epsilon,\delta_k}(x_k,s_k) \le p\delta_k|v_i^T(x_k+s_k)|. \tag{3.16}$$

Note that it follows from $\mathrm{Proj}_{\ker(v_i^T)}\mathcal{F} \subseteq \mathcal{F}$ and $x_k + s_k \in \mathcal{F}$ that

$$x_k + s_k - v_i^T(x_k+s_k)v_i = x_k + s_k - v_i v_i^T(x_k+s_k) = \mathrm{Proj}_{\ker(v_i^T)}(x_k+s_k) \in \mathcal{F}.$$

Then by the convexity of $\mathcal{F}$ and $\delta_k \in (0,1]$ we obtain $x_k + s_k + d \in \mathcal{F}$, where $d = -\delta_k(v_i^T(x_k+s_k))v_i$. Obviously, $d \in \mathrm{span}\{v_i\} = \mathcal{R}_{\{i\}}$. And it follows from (3.14) that $|v_i^T(x_k+s_k)| < \omega < 1$, thus $\|d\| = \delta_k|v_i^T(x_k+s_k)| < \delta_k$. Then d is a feasible point of problem (3.15). Moreover, it holds that

$$q_k(d) = -\delta_k(v_i^T(x_k+s_k))\bar{\mathcal{G}}_k, \tag{3.17}$$

where

$$\bar{\mathcal{G}}_k = -\frac{h(c_k+J_k(s_k+d)) - h(c_k+J_ks_k)}{\delta_k(v_i^T(x_k+s_k))} + v_i^T\Big(\nabla_s m_0(x_k,s_k) + \sum_{i\in\mathcal{A}_k^+}\nabla_s m_i(x_k,s_k)\Big).$$

We next derive a lower bound of $|\bar{\mathcal{G}}_k|$. By the definition of m_i and $s_i^k = v_i^T s_k$, we have

$$m_i(x_k,s_k) = |v_i^Tx_k|^p + p|v_i^Tx_k|^{p-1}\cdot\begin{cases} s_i^k, & \text{if } v_i^Tx_k>0,\ v_i^T(x_k+s_k)>0,\\ -2v_i^Tx_k - s_i^k, & \text{if } v_i^Tx_k>0,\ v_i^T(x_k+s_k)<0,\\ 2v_i^Tx_k + s_i^k, & \text{if } v_i^Tx_k<0,\ v_i^T(x_k+s_k)>0,\\ -s_i^k, & \text{if } v_i^Tx_k<0,\ v_i^T(x_k+s_k)<0,\end{cases}$$

which implies from $|v_i^Tx_k| < \omega$ that

$$\mathrm{sgn}(\nabla_{s_i}m_i(x_k,s_k)) = \mathrm{sgn}(v_i^T(x_k+s_k)) \quad\text{and}\quad |\nabla_{s_i}m_i(x_k,s_k)| = p|v_i^Tx_k|^{p-1} > p\omega^{p-1}. \tag{3.18}$$

As $\|x\| \le M_F$ for any $x \in \mathcal{F}$, $\|s_k\| \le 2M_F$. Then it indicates from (3.2) that

$$\|\nabla_s m_0(x_k,s_k)\| = \|g_k + \frac{1}{\eta}s_k\| \le \chi + \frac{2M_F}{\eta}.$$

It together with the Lipschitz continuity of h, $\|d\| = \delta_k|v_i^T(x_k+s_k)|$, (3.2), (3.18), $v_i^T\sum_{i\in\mathcal{A}_k^+}\nabla_s m_i(x_k,s_k) = \nabla_{s_i}m_i(x_k,s_k)$ and $\omega < (\frac{p}{2(L_hL_c^0+\chi+2M_F/\eta)})^{\frac{1}{1-p}}$ derives the following lower bound:

$$\begin{aligned}
|\bar{\mathcal{G}}_k| =& \Big|\frac{1}{-\delta_kv_i^T(x_k+s_k)}(h(c_k+J_k(s_k+d)) - h(c_k+J_ks_k)) + v_i^T\nabla_s m_0(x_k,s_k)\\
&+ \nabla_{s_i}m_i(x_k,s_k)\Big|\\
\ge& |\nabla_{s_i}m_i(x_k,s_k)| - \frac{1}{\delta_k|v_i^T(x_k+s_k)|}|h(c_k+J_k(s_k+d)) - h(c_k+J_ks_k)|\\
&- |v_i^T\nabla_s m_0(x_k,s_k)|\\
\ge& p|v_i^Tx_k|^{p-1} - L_hL_c^0 - \|\nabla_s m_0(x_k,s_k)\|\\
>& p\,\omega^{p-1} - \Big(L_hL_c^0 + \chi + \frac{2M_F}{\eta}\Big) \ge \frac{1}{2}p\,\omega^{p-1}.
\end{aligned} \tag{3.19}$$

Furthermore, (3.18) indicates $\mathrm{sgn}(\bar{\mathcal{G}}_k) = \mathrm{sgn}(\nabla_{s_i} m_i(x_k, s_k)) = \mathrm{sgn}(v_i^T(x_k + s_k))$, thus by (3.17), $q_k(d) = -\delta_k(v_i^T(x_k + s_k))\bar{\mathcal{G}}_k = -\delta_k|v_i^T(x_k + s_k)||\bar{\mathcal{G}}_k| < 0$.

We now denote by d^* the optimal solution of (3.15). Obviously, $d^* \neq 0$. As $d^* \in \mathcal{R}_{\{i\}}$, there exists $\alpha \in \mathbb{R}$ such that $d^* = \alpha v_i$, thus $\alpha \neq 0$ and $\|d^*\| = |\alpha|$. Then we obtain $q_k(d^*) = \alpha\mathcal{G}_k$, where

$$\mathcal{G}_k = \frac{1}{\alpha}(h(c_k + J_k(s_k + d^*)) - h(c_k + J_k s_k)) + v_i^T(\nabla_s m_0(x_k, s_k) + \sum_{i \in \mathcal{A}_k^+} \nabla_s m_i(x_k, s_k)).$$

Again by the negativeness of the optimal function value of (3.15) it holds that

$$\mathrm{sgn}(\alpha) = -\mathrm{sgn}(\mathcal{G}_k) \quad \text{and} \quad |q_k(d^*)| = |\alpha\mathcal{G}_k| = \|d^*\||\mathcal{G}_k|. \tag{3.20}$$

Meanwhile, by the optimality of d^* we obtain

$$\|d^*\||\mathcal{G}_k| \geq \delta_k|v_i^T(x_k + s_k)||\bar{\mathcal{G}}_k|. \tag{3.21}$$

We next derive a lower bound of $\|d^*\|$. From (3.19) it follows that

$$\left|-\frac{h(c_k + J_k(s_k + d)) - h(c_k + J_k s_k)}{\delta_k(v_i^T(x_k + s_k))} + v_i^T\nabla_s m_0(x_k, s_k)\right| \leq \frac{1}{2}|\nabla_{s_i} m_i(x_k, s_k)|.$$

Moreover, analogy to (3.19) we can obtain $|\mathcal{G}_k| \geq \frac{1}{2}p\omega^{p-1}$ and

$$\left|\frac{h(c_k + J_k(s_k + d^*)) - h(c_k + J_k s_k)}{\alpha} + v_i^T\nabla_s m_0(x_k, s_k)\right| \leq \frac{1}{2}|\nabla_{s_i} m_i(x_k, s_k)|.$$

Then by definitions of $\mathcal{G}_k$ and $\bar{\mathcal{G}}_k$, we have

$$\begin{aligned}\frac{|\bar{\mathcal{G}}_k|}{|\mathcal{G}_k|} &\geq \frac{|\nabla_{s_i} m_i(x_k, s_k)| - |-\frac{h(c_k+J_k(s_k+d))-h(c_k+J_ks_k)}{\delta_k(v_i^T(x_k+s_k))} + v_i^T\nabla_s m_0(x_k, s_k)|}{|\nabla_{s_i} m_i(x_k, s_k)| + |\frac{h(c_k+J_k(s_k+d^*))-h(c_k+J_ks_k)}{\alpha} + v_i^T\nabla_s m_0(x_k, s_k)|} \\ &\geq \frac{\frac{1}{2}|\nabla_{s_i} m_i(x_k, s_k)|}{\frac{3}{2}|\nabla_{s_i} m_i(x_k, s_k)|} = \frac{1}{3},\end{aligned}$$

which indicates from (3.21) that $\|d^*\| \geq \frac{1}{3}\delta_k|v_i^T(x_k + s_k)|$. Based on above inequality together with (3.16), (3.20) and $|\mathcal{G}_k| \geq \frac{1}{2}p\omega^{p-1}$ we obtain

$$\frac{1}{6}p\omega^{p-1}\delta_k|v_i^T(x_k + s_k)| \leq |q_k(d^*)| \leq p\delta_k|v_i^T(x_k + s_k)|,$$

which, however, contradicts $\omega < 6^{\frac{1}{p-1}}$. Thus, the conclusion is proved by contradiction. ∎

To analyze oracle complexity of Algorithm 2.1, we first introduce the following index sets:

$$\begin{aligned}&\mathcal{K}_u := \{k \in \mathcal{K} : x_k = x_{k+1}\}, && \mathcal{K}_\epsilon := \{k \in \mathcal{K}\backslash\mathcal{K}_u : \mathcal{A}_k\backslash\mathcal{A}_{k+1} \neq \emptyset\}, \\ &\mathcal{K}_\omega := \{k \in \mathcal{K}\backslash\mathcal{K}_u : \|s_k\| \geq \frac{1}{4}\omega\}, && \mathcal{K}_\heartsuit := \mathcal{K}\backslash(\mathcal{K}_u \cup \mathcal{K}_\epsilon \cup \mathcal{K}_\omega).\end{aligned}$$

Due to the monotonely non-increasing property of $\mathcal{A}_k$, it is easy to have

$$|\mathcal{K}_\epsilon| \leq \bar{n}. \tag{3.22}$$

Since Algorithm 2.1 terminates when both k and $k-1$ belong to $\mathcal{K}_u$, it must hold that

$$|\mathcal{K}_u| \leq |\mathcal{K}\backslash\mathcal{K}_u| + 2 \leq |\mathcal{K}_\heartsuit \cup \mathcal{K}_\omega| + \bar{n} + 2. \tag{3.23}$$

Define $\alpha = \frac{3}{4}\omega$, where ω satisfies (3.13). The following lemma shows properties of $\mathcal{A}_k$ and $\mathcal{A}_{k+1}$ with $k \in \mathcal{K}_\heartsuit$ which are also discussed in [11].

Lemma 3.5. *Suppose that $\epsilon < \alpha$. Then the following relations hold:*

$$\mathcal{A}_k = \mathcal{A}_{k+1} = \mathcal{O}_{k,\alpha}, \quad k \in \mathcal{K}_\heartsuit, \tag{3.24}$$

where $\mathcal{O}_{k,\alpha}$ is defined in (3.6).

Proof. By (2.8) and the definition of $\mathcal{K}_\heartsuit$, it is easy to have $\mathcal{A}_k = \mathcal{A}_{k+1}$ for any $k \in \mathcal{K}_\heartsuit$. For any $k \in \mathcal{K}_\heartsuit$, we partition $\mathcal{A}_k$ into the following sets:

$$\begin{aligned}
\mathcal{I}_{\heartsuit,k} &:= \left\{ i \in \mathcal{A}_k : \min\{|v_i^T x_k|, |v_i^T(x_k+s_k)|\} \geq \alpha \right\}, \\
\mathcal{I}_{\diamondsuit,k} &:= \big\{ i \in \mathcal{A}_k : \left(|v_i^T x_k| \geq \omega, |v_i^T(x_k+s_k)| \in (\epsilon,\alpha)\right) \\
&\qquad \text{or } \left(|v_i^T x_k| \in (\epsilon,\alpha), |v_i^T(x_k+s_k)| \geq \omega\right) \big\}, \\
\mathcal{I}_{\clubsuit,k} &:= \left\{ i \in \mathcal{A}_k : |v_i^T x_k| \in (\epsilon,\omega) \text{ and } |v_i^T(x_k+s_k)| \in (\epsilon,\omega) \right\}.
\end{aligned}$$

Note that for any $i \in \mathcal{I}_{\diamondsuit,k}$,

$$\|s_k\| \geq |v_i^T s_k| \geq \left| |v_i^T(x_k+s_k)| - |v_i^T x_k| \right| \geq \omega - \alpha = \frac{1}{4}\omega.$$

It then indicates $i \in \mathcal{K}_\omega$. Thus $\mathcal{I}_{\diamondsuit,k} = \emptyset$. Meanwhile, it follows from Lemma 3.4 that $\mathcal{I}_{\clubsuit,k} = \emptyset$. Thus, $\mathcal{A}_k = \mathcal{I}_{\heartsuit,k}$, namely, $\mathcal{A}_k = \left\{ i : \min\left\{|v_i^T x_k|, |v_i^T(x_k+s_k)|\right\} \geq \alpha \right\}$, $k \in \mathcal{K}_\heartsuit$. It then yields (3.24) by definition of $\mathcal{O}_{k,\alpha}$ in (3.6). ∎

Motivated by Lemma 2.1, we suppose that δ_k, $k \in \mathcal{K}$ is uniformly lower bounded by $\delta > 0$ which is independent of ϵ. Then by the boundedness of $\mathcal{F}$, there exists $M > 0$ such that $\|s_k\| \leq \|x_{k+1}\| + \|x_k\| \leq 2M_F \leq M\delta \leq M\delta_k$ for any $k \in \mathcal{K}$. The lemma below shows that when Algorithm 2.1 terminates at Step 5 with $s_k \neq 0$, the output is an approximate first-order stationary point of (1.1), provided that input $\bar{w}$ and $\bar{\beta}$ in Algorithm 2.1 satisfy

$$\bar{\beta} \leq \min\left\{ \frac{1}{3}\bar{w}, \frac{1-\beta-\theta}{\max(L(\alpha) + 1/\eta + L_h L_c^1(M+1), \bar{L})} \right\}. \tag{3.25}$$

We would like to mention that (3.25) can ensure $\bar{\beta}\bar{L} < 1$, which meets the requirement on $\bar{L}$ in Assumption 3.1.

Lemma 3.6. *Suppose that $\epsilon < \alpha$. If Algorithm 2.1 terminates at Step 5 with $s_k \neq 0$ and $\bar{\beta}$ satisfies* (3.25)*, then x_{k+1} is an (ϵ,δ)-approximate first-order stationary point of* (1.1).

Proof. When Algorithm 2.1 terminates at Step 5 with $s_k \neq 0$ and $k \notin \mathcal{K}_\epsilon$, $k \notin \mathcal{K}_u$. Besides, it follows from the algorithmic framework that

$$\|s_k\| + \|s_{k-1}\| \leq \bar{\beta}\epsilon \leq \frac{1}{3}\bar{w}\epsilon < \frac{1}{3}\alpha = \frac{1}{4}\omega,$$

which indicates $k \notin \mathcal{K}_\omega$, thus $k \in \mathcal{K}_\heartsuit$ and (3.24) holds. Recall that (2.7) holds with $\delta = \delta_k$, for some $\delta_k \in (0,1]$, i.e.

$$\psi_m^{\epsilon,\delta_k}(x_k, s_k) \leq \min\left\{ \theta\epsilon, p \min_{i \in \mathcal{A}_{k+1}} |v_i^T x_{k+1}| \right\}\delta_k, \quad \text{for some } \delta_k \in (0,1].$$

Note that by (2.1) and (3.24) as well as (3.7),

$$
\begin{aligned}
&\psi_Q^{\epsilon,\delta_k}(x_{k+1})\\
&= Q_\epsilon(x_{k+1}) - \min_{\substack{x_{k+1}+d\in\mathcal{F}\\ d\in\mathcal{R}_{k+1},\|d\|\le\delta_k}} T_{Q_\epsilon}(x_{k+1},d)\\
&= h(c_{k+1})\\
&\quad - \min_{\substack{x_{k+1}+d\in\mathcal{F}\\ d\in\mathcal{R}_{k+1},\|d\|\le\delta_k}} \Big\{h(c_{k+1}+J_{k+1}d) + d^T\nabla\Big(f(x)+\sum_{i\in\mathcal{A}_{k+1}}|v_i^Tx|^p\Big)|_{x=x_{k+1}}\Big\}\\
&= h(c_{k+1}) - \min_{\substack{x_{k+1}+d\in\mathcal{F}\\ d\in\mathcal{R}_{k+1},\|d\|\le\delta_k}} \big\{h(c_{k+1}+J_{k+1}d) + \nabla\bar{Q}_{k,\alpha}(x_{k+1})^Td\big\}.
\end{aligned}\tag{3.26}
$$

As the minimization problem in (3.26) is convex, it admits a global minimizer, which we still denote as d with a slight abuse of notation. Obviously, $\|d\|\le\delta_k$. We next show by contradiction that $\psi_Q^{\epsilon,\delta_k}(x_{k+1})\le\epsilon\delta_k$. We now assume that it were not true. Then it holds that $\psi_Q^{\epsilon,\delta_k}(x_{k+1}) = h(c_{k+1}) - h(c_{k+1}+J_{k+1}d) - \nabla\bar{Q}_{k,\alpha}\left(x_{k+1}\right)^T d > \epsilon\delta_k$. It can further derive

$$
\begin{aligned}
&\psi_Q^{\epsilon,\delta_k}\left(x_{k+1}\right)\\
&= -\left(\nabla\bar{Q}_{k,\alpha}\left(x_{k+1}\right)\right)^T d + \left(\nabla_s\bar{T}_{k,\alpha}\left(x_k,s_k\right)\right)^T d - \left(\nabla_s\bar{T}_{k,\alpha}\left(x_k,s_k\right)\right)^T d\\
&\quad - \frac{1}{2\eta}[\nabla\left(\|s\|^2\right)|_{s=s_k}]^T d + \frac{1}{2\eta}[\nabla\left(\|s\|^2\right)|_{s=s_k}]^T d + h(c_{k+1}) - h(c_{k+1}+J_{k+1}d)\\
&\le \|\nabla\bar{Q}_{k,\alpha}\left(x_{k+1}\right) - \nabla_s\bar{T}_{k,\alpha}\left(x_k,s_k\right)\|\|d\| - \Big[\nabla_s\big(\bar{T}_{k,\alpha}\left(x_k,s\right) + \frac{\|s\|^2}{2\eta}\big)\Big|_{s=s_k}\Big]^T d\\
&\quad + \frac{1}{\eta}s_k^T d + h(c_{k+1}) - h(c_{k+1}+J_{k+1}d)\\
&\le \Big(L\left(\alpha\right) + \frac{1}{\eta} + L_hL_c^1(M+1)\Big)\|s_k\|\delta_k + \|\nabla f\left(x_k\right) - g_k\|\delta_k + \theta\epsilon\delta_k,
\end{aligned}
$$

where the last inequality follows from $\|d\|\le\delta_k$, (3.9), (3.24), and

$$
\begin{aligned}
&h(c_{k+1}) - h(c_{k+1}+J_{k+1}d) - \Big[\nabla_s\Big(\bar{T}_{k,\alpha}\left(x_k,s\right) + \frac{1}{2\eta}\|s\|^2\Big)\Big|_{s=s_k}\Big]^T d\\
&= -\left(\nabla f\left(x_k\right) - g_k\right)^T d + h(c_{k+1}) - h(c_{k+1}+J_{k+1}d)\\
&\quad - \Big[\nabla_s m_0(x_k,s_k) + \sum_{i\in\mathcal{A}_k}\nabla_s m_i(x_k,s_k)\Big]^T d\\
&\le \|\nabla f\left(x_k\right) - g_k\|\|d\| + \max\Big\{0, h(c_{k+1}) - h(c_{k+1}+J_{k+1}d)\\
&\quad - \Big[\nabla_s m_0(x_k,s_k) + \sum_{i\in\mathcal{A}_{k+1}}\nabla_s m_i\left(x_k,s_k\right)\Big]^T d\Big\}\\
&\le \|\nabla f\left(x_k\right) - g_k\|\|d\| + \psi_m^{\epsilon,\delta_k}\left(x_k,s_k\right) + |h(c_{k+1}) - h(c_k+J_ks_k)|\\
&\quad + |h(c_k+J_k(s_k+d)) - h(c_{k+1}+J_{k+1}d)|\\
&\le \|\nabla f\left(x_k\right) - g_k\|\delta_k + \theta\epsilon\delta_k + L_hL_c^1(M+1)\|s_k\|\delta_k
\end{aligned}
$$

due to $\psi_m^{\epsilon,\delta_k}(x_k,s_k)\le\theta\epsilon\delta_k$,

$$
|h(c_{k+1}) - h(c_k+J_ks_k)| \le \frac{L_hL_c^1}{2}\|s_k\|^2 \le \frac{L_hL_c^1M}{2}\delta_k\|s_k\|
$$

and

$$
\begin{aligned}
&|h(c_k + J_k(s_k + d)) - h(c_{k+1} + J_{k+1}d)| \\
&\le L_h\|c_k + J_k(s_k + d) - c_{k+1} - J_{k+1}d\| \\
&\le L_h\|c_k + J_k s_k - c_{k+1}\| + L_h\|J_k - J_{k+1}\|\|d\| \\
&\le \frac{L_h L_c^1}{2}\|s_k\|^2 + L_h L_c^1\|s_k\|\|d\| \\
&\le L_h L_c^1\Big(\frac{M}{2} + 1\Big)\delta_k\|s_k\|.
\end{aligned}
$$

Then it follows from $\psi_Q^{\epsilon,\delta_k}(x_{k+1}) > \epsilon\delta_k$ and Assumption 3.1 with $\beta < 1-\theta$ that

$$\epsilon\delta_k < \Big(L(\alpha) + \frac{1}{\eta} + L_h L_c^1(M+1)\Big)\|s_k\|\delta_k + \bar{L}\|s_{k-1}\|\delta_k + (\beta+\theta)\epsilon\delta_k$$

which implies

$$(1-\beta-\theta)\,\epsilon < \max\Big\{L\,(\alpha) + \frac{1}{\eta} + L_h L_c^1(M+1), \bar{L}\Big\}(\|s_k\| + \|s_{k-1}\|).$$

However, this contradicts $\|s_k\| + \|s_{k-1}\| \le \bar{\beta}\epsilon$ by the setting of $\bar{\beta}$. Therefore, x_{k+1} is an (ϵ,δ)-approximate first-order stationary point of (1.1). ∎

Remark 3.7. Lemmas 3.1 and 3.6 show that Algorithm 2.1 can always return an approximate first-order stationary point of (1.1) when it terminates.

We now partition $\mathcal{K}_\heartsuit$ into $\mathcal{K}_\heartsuit^1 \cup \mathcal{K}_\heartsuit^2$, where

$$\mathcal{K}_\heartsuit^1 := \{k \in \mathcal{K}_\heartsuit : \|s_k\| + \|s_{k-1}\| \ge \bar{\beta}\epsilon\}, \quad \mathcal{K}_\heartsuit^2 := \{k \in \mathcal{K}_\heartsuit : \|s_k\| + \|s_{k-1}\| < \bar{\beta}\epsilon\}.$$

By the definition of $\mathcal{K}_\heartsuit$, Lemma 3.6 and termination conditions of Algorithm 2.1, we know that $|\mathcal{K}_\heartsuit^2| \le 1$, thus $|\mathcal{K}_\heartsuit| \le |\mathcal{K}_\heartsuit^1| + 1$. Then it together with (3.22) and (3.23) implies that the total number of iterations until Algorithm 2.1 terminates satisfies

$$
\begin{aligned}
|\mathcal{K}| \le |\mathcal{K}_u| + |\mathcal{K}_\heartsuit \cup \mathcal{K}_\omega| + |\mathcal{K}_\epsilon| &\le |\mathcal{K}_\heartsuit \cup \mathcal{K}_\omega| + \bar{n} + 2 + |\mathcal{K}_\heartsuit \cup \mathcal{K}_\omega| + \bar{n} \\
&\le 2|\mathcal{K}_\heartsuit^1 \cup \mathcal{K}_\omega| + 2\bar{n} + 4.
\end{aligned}
\tag{3.27}
$$

Based on above relations, to estimate the upper bound of $|\mathcal{K}|$, it suffices to derive an upper bound on $|\mathcal{K}_\heartsuit^1 \cup \mathcal{K}_\omega|$. Inspired by this, we establish the oracle complexity of Algorithm 2.1 in the theorem below. In the following we assume that the positive parameter η in (2.6) satisfies

$$\frac{1}{16\eta} - \frac{L_f + L_h L_c^1}{16} - \beta\bar{L} - \frac{\beta}{\bar{\beta}} \ge 1, \quad \frac{1}{4\eta} - \frac{L_f + L_h L_c^1}{4} - \beta\bar{L} - 3\beta \ge 1. \tag{3.28}$$

It is noteworthy that the setting of $\bar{\beta}$ in (3.25) together with (3.28) and Assumption 3.1 ensures the existence of desired input parameters $\bar{\omega}, \bar{\beta}, \eta, \bar{L}$ and β. We now proceed under such parameter settings.

THEOREM 3.8. *Suppose that $\epsilon < \alpha$. Then there exists a positive constant $C = \mathcal{O}(1)$ such that $\sum_{k\in\mathcal{K}} \|s_k\|^2 \le C$. Furthermore, the maximum iteration number until Algorithm 2.1 terminates is in order of $\mathcal{O}(\epsilon^{-2})$.*

Proof. It follows from Lemma 3.2, $s_{-1} = 0$ and Assumption 3.1 that

$$\begin{aligned}
&\Big(\frac{1}{2\eta} - \frac{L_f + L_h L_c^1}{2}\Big) \sum_{k\in\mathcal{K}} \|s_k\|^2 \\
&\le \sum_{k\in\mathcal{K}} \|\nabla f(x_k) - g_k\| \|s_k\| + Q_\epsilon(x_0) - Q_\epsilon^* \\
&\le \sum_{k\in\mathcal{K}} \beta \max\left\{\bar{L}\min\{\|s_{k-1}\|, D\}, \epsilon\right\} \|s_k\| + Q_\epsilon(x_0) - Q_\epsilon^* \\
&\le \sum_{k\in\mathcal{K}} \beta \left(\bar{L}\|s_{k-1}\|\|s_k\| + \epsilon\|s_k\|\right) + Q_\epsilon(x_0) - Q_\epsilon^* \\
&\le \sum_{k\in\mathcal{K}} \beta \Big(\frac{\bar{L}}{2}\left(\|s_k\|^2 + \|s_{k-1}\|^2\right) + \epsilon\|s_k\|\Big) + Q_\epsilon(x_0) - Q_\epsilon^* \\
&\le \sum_{k\in\mathcal{K}} \beta(\bar{L}\|s_k\|^2 + \epsilon\|s_k\|) + Q_\epsilon(x_0) - Q_\epsilon^*. \qquad (3.29)
\end{aligned}$$

As

$$\|s_k\| \ge \frac{1}{3}\alpha \ge \frac{1}{3}\epsilon, \quad k \in \mathcal{K}_\omega, \qquad (3.30)$$

it indicates

$$\bar{L}\|s_k\|^2 + \epsilon\|s_k\| \le (\bar{L}+3)\|s_k\|^2, \quad k \in \mathcal{K}_\omega. \qquad (3.31)$$

Moreover, by definition of $\mathcal{K}_\heartsuit^1$ we have

$$\|s_k\| + \|s_{k-1}\| \ge \bar{\beta}\epsilon, \quad k \in \mathcal{K}_\heartsuit^1, \qquad (3.32)$$

thus

$$\begin{aligned}
\bar{L}\|s_k\|^2 + \epsilon\|s_k\| &\le \bar{L}\|s_k\|^2 + \bar{\beta}^{-1}(\|s_k\| + \|s_{k-1}\|)\|s_k\| \\
&\le (\bar{L} + \bar{\beta}^{-1})(\|s_k\| + \|s_{k-1}\|)\|s_k\|, \quad k \in \mathcal{K}_\heartsuit^1. \qquad (3.33)
\end{aligned}$$

Since $s_k = 0$ for any $k \in \mathcal{K}_u$, plugging (3.31) and (3.33) into (3.29) yields

$$\begin{aligned}
&\Big(\frac{1}{2\eta} - \frac{L_f + L_h L_c^1}{2}\Big) \sum_{k\in\mathcal{K}} \|s_k\|^2 \\
&\le \sum_{k\in\mathcal{K}_\heartsuit^1} \beta(\bar{L} + \bar{\beta}^{-1})\left(\|s_k\| + \|s_{k-1}\|\right)\|s_k\| + \sum_{k\in\mathcal{K}_\omega} \beta\left(\bar{L}+3\right)\|s_k\|^2 \\
&\quad + \sum_{k\in\mathcal{K}_\epsilon\cup\mathcal{K}_\heartsuit^2} \beta\left(\bar{L}\|s_k\|^2 + \epsilon\|s_k\|\right) + Q_\epsilon(x_0) - Q_\epsilon^*. \qquad (3.34)
\end{aligned}$$

Recall that $\|s_k\| < \bar{\beta}\epsilon < 1$ for any $k \in \mathcal{K}_\heartsuit^2$. Besides, by the boundedness of $\mathcal{F}$ we have $\|s_k\| \le 2M_F$ for any $k \in \mathcal{K}_\epsilon$. Then it follows from (3.34) that

$$\begin{aligned}
&\Big(\frac{1}{2\eta} - \frac{L_f + L_h L_c^1}{2}\Big) \sum_{k\in\mathcal{K}} \|s_k\|^2 \\
&\le \sum_{k\in\mathcal{K}_\heartsuit^1} \beta(\bar{L} + \bar{\beta}^{-1})\left(\|s_k\| + \|s_{k-1}\|\right)^2 + \sum_{k\in\mathcal{K}_\omega} \beta\left(\bar{L}+3\right)\|s_k\|^2 \\
&\quad + \bar{n}\beta(4\bar{L}M_F^2 + 2\epsilon M_F) + \beta(\bar{L}+\epsilon) + Q_\epsilon(x_0) - Q_\epsilon^*, \qquad (3.35)
\end{aligned}$$

where the last term in above inequality uses the facts that $|\mathcal{K}_\epsilon| \le \bar{n}$ and $|\mathcal{K}_\heartsuit^2| \le 1$. Notice that

$$\begin{aligned}
\sum_{k\in\mathcal{K}} \|s_k\|^2 &= \frac{1}{2}\Big(\sum_{k\in\mathcal{K}} \|s_k\|^2 + \sum_{k\in\mathcal{K}} \|s_k\|^2\Big) \\
&\ge \frac{1}{4}\sum_{k\in\mathcal{K}} (\|s_k\|^2 + \|s_{k-1}\|^2) + \frac{1}{2}\sum_{k\in\mathcal{K}_\omega} \|s_k\|^2 \\
&\ge \frac{1}{8}\sum_{k\in\mathcal{K}} (\|s_k\| + \|s_{k-1}\|)^2 + \frac{1}{2}\sum_{k\in\mathcal{K}_\omega} \|s_k\|^2 \\
&\ge \frac{1}{8}\sum_{k\in\mathcal{K}_\heartsuit^1} (\|s_k\| + \|s_{k-1}\|)^2 + \frac{1}{2}\sum_{k\in\mathcal{K}_\omega} \|s_k\|^2,
\end{aligned}$$

which further derives

$$\begin{aligned}
&\Big(\frac{1}{2\eta} - \frac{L_f + L_h L_c^1}{2}\Big)\sum_{k\in\mathcal{K}} \|s_k\|^2 \\
&\ge \Big(\frac{1}{16\eta} - \frac{L_f + L_h L_c^1}{16}\Big)\sum_{k\in\mathcal{K}_\heartsuit^1} (\|s_k\| + \|s_{k-1}\|)^2 + \Big(\frac{1}{4\eta} - \frac{L_f + L_h L_c^1}{4}\Big)\sum_{k\in\mathcal{K}_\omega} \|s_k\|^2.
\end{aligned}$$

Then it together with (3.35) and the boundedness of $\mathcal{F}$ implies that

$$\begin{aligned}
&\Big(\frac{1}{16\eta} - \frac{L_f + L_h L_c^1}{16} - \beta\bar{L} - \frac{\beta}{\bar{\beta}}\Big)\sum_{k\in\mathcal{K}_\heartsuit^1} (\|s_k\| + \|s_{k-1}\|)^2 \\
&+ \Big(\frac{1}{4\eta} - \frac{L_f + L_h L_c^1}{4} - \beta\bar{L} - 3\beta\Big)\sum_{k\in\mathcal{K}_\omega} \|s_k\|^2 \le \bar{\Gamma}
\end{aligned}$$

with $\bar{\Gamma} = \bar{n}\beta(4\bar{L}M_F^2 + 2\epsilon M_F) + \beta(\bar{L}+\epsilon) + Q_\epsilon(x_0) - Q_\epsilon^*$. Furthermore, from the setting of η as in (3.28) we attain

$$\sum_{k\in\mathcal{K}_\heartsuit^1} (\|s_k\| + \|s_{k-1}\|)^2 + \sum_{k\in\mathcal{K}_\omega} \|s_k\|^2 \le \bar{\Gamma} \tag{3.36}$$

which leads to the conclusion from (3.35) with $C = \frac{\bar{\Gamma}(1+\beta\bar{L}+\beta\max\{\bar{\beta}^{-1},3\})}{1/(2\eta)-(L_f+L_hL_c^1)/2}$. Obviously, $C = \mathcal{O}(1)$.

Moreover, by (3.30) and (3.32) we obtain

$$\sum_{k\in\mathcal{K}_\heartsuit^1} (\|s_k\| + \|s_{k-1}\|)^2 + \sum_{k\in\mathcal{K}_\omega} \|s_k\|^2 \ge \bar{\beta}^2\epsilon^2|\mathcal{K}_\heartsuit^1| + \frac{1}{9}\epsilon^2|\mathcal{K}_\omega|.$$

Then it together with (3.36) implies $|\mathcal{K}_\heartsuit^1| + |\mathcal{K}_\omega| = \mathcal{O}(\epsilon^{-2})$. Hence by (3.27) the maximum iteration number until the termination of Algorithm 2.1 is in order $\mathcal{O}(\epsilon^{-2})$. □

Since only one inexact gradient is evaluated at each iteration, the oracle complexity of Algorithm 2.1 is in order $\mathcal{O}(\epsilon^{-2})$.

4. Stochastic variant.

For problem (1.1), when f owns a finite-sum structure (1.2), as the sample size N can be very large, it will be expensive to go through all component functions to compute exact gradients, thereby only approximate gradients are available. To cope with this type of problems, we propose a stochastic variant of Algorithm 2.1. The proposed algorithm follows the main framework of Algorithm 2.1, with **InexactOracle** specified in Algorithm 4.1. Here inexact gradients are computed by calling stochastic first-order oracles in a recursive way [23] and l is a positive integer.

Algorithm 4.1 InexactOracle(x_k, x_{k-1}, g_{k-1}, k, l)

Input: Index set $\mathcal{I}_k$ generated uniformly at random without replacement from $\{1, \ldots, N\}$.

1: **if** $\mod(k, l) = 0$ **then**
2: Compute $g_k = \frac{1}{|\mathcal{I}_k|}\sum_{i\in\mathcal{I}_k} \nabla f_i(x_k)$.
3: **else**
4: Compute $g_k = \frac{1}{|\mathcal{I}_k|}\sum_{i\in\mathcal{I}_k} (\nabla f_i(x_k) - \nabla f_i(x_{k-1})) + g_{k-1}$.
5: **end if**

Note that Assumption 3.1 plays a key role in supervising the analysis in previous section. In this section, adopting a proper sampling strategy we can guarantee Assumption 3.1 with high probability. We then establish the complexity of the proposed algorithm, in terms of number of stochastic first-order oracles, to find an approximate first-order stationary point. To proceed the analysis, we first introduce a lemma regarding the concentration inequality under sampling without replacement. As this lemma is a duplicate of [27, Theorem 4], we omit its proof here.

LEMMA 4.1. *Let $\mathcal{X} = \{X_i \in \mathbb{R}^n, i = 1, \ldots, N\}$. Suppose $\|X_i\| \le \sigma$ for all $i = 1, \ldots, N$ and some $\sigma > 0$. Denote $\lambda = \frac{1}{N}\sum_{i=1}^N X_i$. Let $A_1, \ldots, A_\nu$, $\nu < N$ be samples from $\mathcal{X}$ under the sampling without replacement. Then, for any $\epsilon > 0$, the following bound holds:*

$$\mathrm{Prob}\Big(\Big\|\frac{1}{\nu}\sum_{i=1}^{\nu} A_i - \lambda\Big\| \ge \epsilon\Big) \le 2\,(n+1)\exp\Big(-\frac{\nu\epsilon^2}{8\sigma^2\left(1+\frac{1}{\nu}\right)\left(1-\frac{\nu}{N}\right)}\Big).$$

Given $\zeta \in (0, 1)$, following Lemma 4.1, we can achieve

$$\mathrm{Prob}\Big(\Big\|\frac{1}{\nu}\sum_{i=1}^{\nu} A_i - \lambda\Big\| \le \epsilon\Big) \ge 1-\zeta\,,\ \text{if } \nu \ge \Big[\frac{1}{N} + \frac{\epsilon^2}{16\sigma^2\log\left(2\,(n+1)\,/\zeta\right)}\Big]^{-1}.$$

We assume that $\nabla f_i, i = 1, \ldots, N$ are Lipschitz continuously differentiable. With a slight abuse of notations, we still use L_f and κ to denote the Lipschitz constant and upper bound of ∇f_i, $i = 1, \ldots, N$ over $\mathcal{F}$. Then for any k with $\mod(k, l) = 0$, g_k generated by Algorithm 4.1 satisfies

$$\mathrm{Prob}\left(\|g_k - \nabla f(x_k)\| \le \beta\epsilon\right) \ge 1-\zeta,\ \text{if } |\mathcal{I}_k| \ge \Big[\frac{1}{N} + \frac{\beta^2\epsilon^2}{16\kappa^2\log\left(2\,(n+1)\,/\zeta\right)}\Big]^{-1}.$$

For those k with $\mod(k, l) \ne 0$, the lemma below provides a sampling strategy such that Assumption 3.1 can be satisfied with high probability.

LEMMA 4.2. *Under sampling without replacement, for any k with mod$(k,l) \neq 0$, g_k generated by Algorithm 4.1 satisfies Assumption 3.1 with probability at least $1-\zeta$, provided that*

$$(4.1)\quad |\mathcal{I}_j| \geq \begin{cases} \left[\frac{1}{N} + \frac{\beta^2\epsilon^2/l^2}{256L_f^2\|x_j - x_{j-1}\|^2\log(4(n+1)l/\zeta)}\right]^{-1}, & j = k, k-1, \dots, \lfloor k/l \rfloor l + 1, \\ \left[\frac{1}{N} + \frac{\beta^2\epsilon^2}{256\kappa^2\log(4(n+1)/\zeta)}\right]^{-1}, & j = \lfloor k/l \rfloor l. \end{cases}$$

Proof. For any k with mod$(k,l) \neq 0$, it follows from the algorithmic framework that

$$\begin{aligned} & g_k - \nabla f(x_k) \\ & = \frac{1}{|\mathcal{I}_k|}\sum_{i\in\mathcal{I}_k}[\nabla f_i(x_k) - \nabla f_i(x_{k-1})] + g_{k-1} - \nabla f(x_k) \\ & = \frac{1}{|\mathcal{I}_k|}\sum_{i\in\mathcal{I}_k}[\nabla f_i(x_k) - \nabla f_i(x_{k-1}) - (\nabla f(x_k) - \nabla f(x_{k-1}))] + g_{k-1} - \nabla f(x_{k-1}). \end{aligned}$$

We thus obtain $g_k - \nabla f(x_k) = \sum_{j=\lfloor k/l \rfloor l}^{k} Y_j$, where $Y_j := \frac{1}{|\mathcal{I}_j|}\sum_{i\in\mathcal{I}_j} Z_{j,i}$ with

$$Z_{j,i} := \begin{cases} \nabla f_i(x_j) - \nabla f_i(x_{j-1}) - (\nabla f(x_j) - \nabla f(x_{j-1})), & j = k, k-1, \dots, \lfloor k/l \rfloor l + 1, \\ \nabla f_i(x_j) - \nabla f(x_j), & j = \lfloor k/l \rfloor l \end{cases}$$

for $i = 1, \dots, N$. Define $\bar{\epsilon}_k := \beta \max\left\{\bar{L}\min(\|s_{k-1}\|, D), \epsilon\right\}$ with $\bar{L}$, D and β in Assumption 3.1. For notation simplicity we denote by B_j the event $\|Y_j\| \leq \frac{\bar{\epsilon}_k}{2(k-\lfloor k/l \rfloor l)}$ with $j = k, \dots, \lfloor k/l \rfloor l + 1$, and by B_j the event $\|Y_j\| \leq \frac{\bar{\epsilon}_k}{2}$ with $j = \lfloor k/l \rfloor l$. We use $\bar{B}_j$ to denote the complement of B_j. Then

$$\text{Prob}(\|g_k - \nabla f(x_k)\| \leq \bar{\epsilon}_k) \geq \text{Prob}\Big(\bigcap_{j=\lfloor k/l \rfloor l}^{k} B_j\Big) = 1 - \text{Prob}\Big(\bigcup_{j=\lfloor k/l \rfloor l}^{k} \bar{B}_j\Big)$$

which is no less than $1 - \sum_{j=\lfloor k/l \rfloor l}^{k} \text{Prob}(\bar{B}_j)$ by the union bound. Hence, to achieve that (3.1) holds with probability at least ζ, it suffices to require

$$(4.2)\quad \text{Prob}(\bar{B}_j) = \begin{cases} \text{Prob}\Big(\|Y_j\| > \frac{\bar{\epsilon}_k}{2(k-\lfloor k/l \rfloor l)}\Big) \leq \frac{\zeta}{2(k-\lfloor k/l \rfloor l)}, & j = k, \dots, \lfloor k/l \rfloor l + 1, \\ \text{Prob}\Big(\|Y_j\| > \frac{\bar{\epsilon}_k}{2}\Big) \leq \frac{\zeta}{2}, & j = \lfloor k/l \rfloor l. \end{cases}$$

Due to the smoothness of f_i, $\|Z_{j,i}\| \leq 2L_f\|x_j - x_{j-1}\|$, $j = k, k-1, \dots, \lfloor k/l \rfloor l + 1$ and $\|Z_{\lfloor k/l \rfloor l, i}\| \leq 2\kappa$, $i = 1, \dots, N$. As $\sum_{i=1}^{N} Z_{j,i} = 0$ for any $j = k, \dots, \lfloor k/l \rfloor l$, by Lemma 4.1 with $\lambda = 0$, $\nu = |\mathcal{I}_j|$ and $A_{i'} = Z_{j,i'}$, $i' = 1, \dots, \nu$, $i' \in [N]$, we obtain that (4.2) can be achieved provided that

$$|\mathcal{I}_j| \geq \begin{cases} \left[\frac{1}{N} + \frac{\beta^2\max\left\{\bar{L}^2\min(\|s_{k-1}\|^2, D^2), \epsilon^2\right\}/(2(k-\lfloor k/l \rfloor l))^2}{64L_f^2\|x_j - x_{j-1}\|^2\log(4(n+1)(k-\lfloor k/l \rfloor l)/\zeta)}\right]^{-1}, & j = k, \dots, \lfloor k/l \rfloor l + 1, \\ \left[\frac{1}{N} + \frac{\beta^2\max\left\{\bar{L}^2\min(\|s_{k-1}\|^2, D^2), \epsilon^2\right\}/4}{64\kappa^2\log(4(n+1)/\zeta)}\right]^{-1}, & j = \lfloor k/l \rfloor l, \end{cases}$$

which can be guaranteed by (4.1) due to $\bar{\epsilon}_k \geq \beta\epsilon$ and $k - \lfloor k/l \rfloor l \leq l$. ∎

We are now ready to present the oracle complexity in terms of total number of stochastic first-order oracles required to guarantee that Algorithm 2.1 can find an (ϵ, δ)-approximate first-order stationary point of (1.1)-(1.2).

THEOREM 4.3. *Suppose that conditions of Theorem 3.8 and Lemma 4.2 with* $l = \mathcal{O}(N^{1/3})$ *hold, and Algorithm 2.1 with Algorithm 4.1 called to compute inexact oracles terminates in finite iterations. Then for given* $\rho \in (0, 1)$*, with probability at least* $1-\rho$*, it returns an* (ϵ, δ)*-approximate first-order stationary point of* (1.1)-(1.2) *with the oracle complexity in order* $\mathcal{O}(N + N^{\frac{2}{3}}\epsilon^{-2}\log(\frac{4(n+1)N^{\frac{1}{3}}}{\epsilon^2\rho}))$*. Consequently, the oracle complexity of Algorithm 2.1 with Algorithm 4.1 is in order* $\tilde{\mathcal{O}}(\epsilon^{-2})$*.*

Proof. We still use $\mathcal{K}$ to denote the set of all iteration indices until termination. As can be seen from previous section, to make sure the algorithm returns an approximate stationary point with probability at least $1-\rho$, it suffices to guarantee with probability at least $1 - \rho$ that Assumption 3.1 holds for all iterations in $\mathcal{K}$. To realize this, Assumption 3.1 should be satisfied at each one of the iterations with probability at least $1 - \zeta$ for some $\zeta \in [0, 1]$ such that $1 - |\mathcal{K}|\zeta \geq 1 - \rho$. We may simply set $\zeta = \frac{\rho}{|\mathcal{K}|}$. Furthermore, to achieve Assumption 3.1 with probability at least $1-\zeta$ at jth iteration for any given $j \in \mathcal{K}$, by Lemma 4.2 the size of $\mathcal{I}_j$ can be equal to the right side of (4.1) after rounding up. With above settings, it holds with probability at least $1 - \rho$ that $|\mathcal{K}| = \mathcal{O}(\epsilon^{-2})$ and $\sum_{j\in\mathcal{K}} \|s_j\|^2 \leq C$, where $C = \mathcal{O}(1)$ by Theorem 3.8. Hence, to reach an (ϵ, δ)-approximate first-order stationary point with probability at least $1-\rho$, the total number of stochastic first-order oracles is bounded by

$$
\begin{aligned}
\sum_{i\in\mathcal{K}} |\mathcal{I}_i| &= \sum_{i:\mathrm{mod}(i,l)=0} |\mathcal{I}_i| + \sum_{i=\lfloor|\mathcal{K}|/l\rfloor l+1}^{|\mathcal{K}|} |\mathcal{I}_i| + \sum_{i=0}^{\lfloor|\mathcal{K}|/l\rfloor-1} \sum_{j=1}^{l-1} |\mathcal{I}_{il+j}| \\
&\leq \lceil\frac{|\mathcal{K}|}{l}\rceil N + \sum_{i=\lfloor|\mathcal{K}|/l\rfloor l+1}^{|\mathcal{K}|} \Big[\frac{1}{N} + \frac{\beta^2\epsilon^2/l^2}{256L_f^2\|x_i - x_{i-1}\|^2 \log\left(4\left(n+1\right)l/\zeta\right)}\Big]^{-1} \\
&\quad + \sum_{i=0}^{\lfloor|\mathcal{K}|/l\rfloor-1} \sum_{j=1}^{l-1} \Big[\frac{1}{N} + \frac{\beta^2\epsilon^2/l^2}{256L_f^2\|x_{il+j} - x_{il+j-1}\|^2 \log\left(4\left(n+1\right)l/\zeta\right)}\Big]^{-1} + |\mathcal{K}| \\
&\leq \lceil\frac{|\mathcal{K}|}{l}\rceil N + \sum_{i=\lfloor|\mathcal{K}|/l\rfloor l+1}^{|\mathcal{K}|} \frac{256L_f^2\|x_i - x_{i-1}\|^2 \log\left(4\left(n+1\right)l/\zeta\right)}{\beta^2\epsilon^2/l^2} \\
&\quad + \sum_{i=0}^{\lfloor|\mathcal{K}|/l\rfloor-1} \sum_{j=1}^{l-1} \frac{256L_f^2\|x_{il+j} - x_{il+j-1}\|^2 \log\left(4\left(n+1\right)l/\zeta\right)}{\beta^2\epsilon^2/l^2} + |\mathcal{K}| \\
&\leq \lceil\frac{|\mathcal{K}|}{l}\rceil N + 256Cl^2L_f^2\frac{1}{\beta^2\epsilon^2}\log\left(4\left(n+1\right)l/\zeta\right) + |\mathcal{K}|
\end{aligned}
$$

which derives the oracle complexity order by the setting of l. ∎

5. Extension to expectation case.

In this section, we focus on solving the problem with f in the expectation form, given by:

$$
\min_{x\in\mathcal{F}} \quad Q(x) := f(x) + h(c(x)) + \|Vx\|_p^p \quad \text{with} \quad f(x) := \mathbb{E}[\mathbf{F}(x, \xi)]. \tag{5.1}
$$

Here, $\xi \in \Xi$ represents a random variable following the probability function $\mathbb{P}$, and $\mathbf{F} : \mathbb{R}^n \times \Xi \to \mathbb{R}$ is continuously differentiable with respect to $x \in \mathcal{F}$ for almost every $\xi \in \Xi$. To address the challenges posed by problems in the expectation form, where the sample set can be infinite, we propose a modification to Algorithm 4.1 by randomly generating a subset of samples from Ξ, presented in Algorithm 5.1.

Algorithm 5.1 InexactOracle(x_k, x_{k-1}, g_{k-1}, k, l)

Input: Generate a sample subset ξ_k uniformly at random from Ξ.
1: **if** $\bmod(k,l) = 0$ **then**
2: Compute $g_k = \frac{1}{|\xi_k|}\sum_{\xi\in\xi_k} \nabla_x \mathbf{F}(x_k,\xi)$.
3: **else**
4: Compute $g_k = \frac{1}{|\xi_k|}\sum_{\xi\in\xi_k} (\nabla_x \mathbf{F}(x_k,\xi) - \nabla_x \mathbf{F}(x_{k-1},\xi)) + g_{k-1}$.
5: **end if**

The aim of this section is to investigate the oracle complexity of Algorithm 2.1 with Algorithm 5.1 called to compute stochastic first-order oracles. Before delving into the analysis, we introduce an assumption that stochastic oracles satisfy.

ASSUMPTION 5.1. *There exist $\Delta, L_f > 0$ such that for all $x \in \mathcal{F}$,*

$$\mathbb{E}[\nabla_x \mathbf{F}(x,\xi)] = \nabla f(x), \quad \|\nabla_x \mathbf{F}(x,\xi) - \nabla f(x)\| \le \Delta \textit{ almost surely},$$

and for any $x, y \in \mathcal{F}$, $\|\nabla_x \mathbf{F}(x,\xi) - \nabla_x \mathbf{F}(y,\xi)\| \le L_f\|x-y\|$ almost surely.

The following lemma presents the matrix Bernsterin inequality [26].

LEMMA 5.1. *Let $X_1,\dots,X_\nu$ be i.i.d. random vectors in $\mathbb{R}^n$, and satisfy $\mathbb{E}[X_i] = 0$ and $\|X_i\| \le \sigma$ almost surely for some $\sigma > 0$ and any $i = 1,\dots,\nu$. Define $M := \max(\|\sum_{i=1}^\nu \mathbb{E}[X_iX_i^T]\|, \|\sum_{i=1}^\nu \mathbb{E}[X_i^TX_i]\|)$. Then for any $t \ge 0$,*

$$\mathrm{Prob}\left(\left\|\sum_{i=1}^{\nu} X_i\right\| \ge t\right) \le (n+1)\cdot \exp\left(\frac{-t^2/2}{M+\sigma t/3}\right).$$

Note that $M \le \sum_{i=1}^\nu \mathbb{E}[\|X_i\|^2] \le \nu\sigma^2$. By Lemma 5.1, we obtain that for any $\epsilon > 0$,

$$\mathrm{Prob}\left(\frac{1}{\nu}\left\|\sum_{i=1}^{\nu} X_i\right\| \ge \epsilon\right) \le (n+1)\cdot \exp\left(\frac{-\nu\epsilon^2/2}{\sigma^2+\sigma\epsilon/3}\right).$$

Then for g_k, generated by Algorithm 5.1 with k s.t. $\bmod(k,l) = 0$, under Assumption 5.1 and by Lemma 5.1 we attain

$$\mathrm{Prob}\left(\|g_k - \nabla f(x_k)\| \le \beta\epsilon\right) \ge 1-\zeta, \text{ if } |\xi_k| \ge \left(\frac{2\Delta^2}{\beta^2\epsilon^2} + \frac{2\Delta}{3\beta\epsilon}\right)\log\left(\frac{n+1}{\zeta}\right).$$

For those k with $\bmod(k,l) \ne 0$, similar to Lemma 4.2, we can provide a sampling strategy such that Assumption 3.1 holds with high probability.

LEMMA 5.2. *Let g_k be generated by Algorithm 5.1. For any k with $\bmod(k,l) \ne 0$, Assumption 3.1 holds at kth iteration with probability at least $1-\zeta$, provided that*

$$|\xi_j| \ge \begin{cases} \left(\frac{32L_f^2\|x_j-x_{j-1}\|^2l^2}{\beta^2\epsilon^2} + \frac{8L_f\|x_j-x_{j-1}\|l}{3\beta\epsilon}\right)\log\left(\frac{2(n+1)l}{\zeta}\right), & j = k,\dots,\lfloor k/l\rfloor l+1, \\ \left(\frac{8\Delta^2}{\beta^2\epsilon^2} + \frac{4\Delta}{3\beta\epsilon}\right)\log\left(\frac{2(n+1)}{\zeta}\right), & j = \lfloor k/l\rfloor l. \end{cases} \tag{5.2}$$

Proof. By the computation of g_k in Algorithm 5.1 and $Y_j := \frac{1}{|\xi_j|}\sum_{\xi\in\xi_j} Z_j(\xi)$ with

$$Z_j(\xi) = \begin{cases} \nabla_x \mathbf{F}(x_j,\xi) - \nabla_x \mathbf{F}(x_{j-1},\xi) - \nabla f(x_j) + \nabla f(x_{j-1}), & j = k,\dots,\lfloor k/l\rfloor l+1, \\ \nabla_x \mathbf{F}(x_j,\xi) - \nabla f(x_j), & j = \lfloor k/l\rfloor l, \end{cases}$$

we obtain $g_k - \nabla f(x_k) = \sum_{j=\lfloor k/l\rfloor l}^{k} Y_j$. Under Assumption 5.1 and due to the smoothness of f, $\|Z_j(\xi)\| \le 2L_f\|x_j - x_{j-1}\|$, $j = k,\dots,\lfloor k/l\rfloor l+1$ and $\|Z_{\lfloor k/l\rfloor l}(\xi)\| \le \Delta$. Similar to the analysis of Lemma 5.2, the remainder is to ensure (4.2). It follows from Lemma 5.1 that to achieve (4.2) it suffices to require

$$|\xi_j| \ge \begin{cases} \left(\frac{32L_f^2\|x_j-x_{j-1}\|^2(k-\lfloor k/l\rfloor l)^2}{\bar\epsilon_k^2} + \frac{8L_f\|x_j-x_{j-1}\|(k-\lfloor k/l\rfloor l)}{3\bar\epsilon_k}\right)\log\left(\frac{2(n+1)(k-\lfloor k/l\rfloor l)}{\zeta}\right), \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad j = k,\dots,\lfloor k/l\rfloor l+1, \\ \left(\frac{8\Delta^2}{\bar\epsilon_k^2} + \frac{4\Delta}{3\bar\epsilon_k}\right)\log\left(\frac{2(n+1)}{\zeta}\right), \qquad\qquad j = \lfloor k/l\rfloor l, \end{cases}$$

which can be guaranteed by (5.2) and $\bar\epsilon_k \ge \beta\epsilon$. ∎

We slightly abuse the notation and continue to use $\mathcal{K}$ to represent all the iteration indices until Algorithm 2.1 terminates, with Algorithm 5.1 being called to compute inexact oracles. According to Lemma 5.2, in order to achieve an (ϵ,δ)-approximate first-order stationary point with a probability at least $1-\rho$, where $\rho\in(0,1)$, Assumption 3.1 must hold at each iteration with probability at least $1-\zeta$ for $\zeta\in(0,1)$ such that $1-|\mathcal{K}|\zeta \ge 1-\rho$. Therefore, we set $\zeta = \frac{\rho}{|\mathcal{K}|}$. Consequently, by applying Theorem 3.8 and setting $l = \mathcal{O}(|\mathcal{K}|^{1/3})$ we can conclude that the total number of stochastic first-order oracles is bounded by:

$$\begin{aligned} \sum_{i\in\mathcal{K}} |\xi_i| &= \sum_{i:\mathrm{mod(i,l)}=0} |\xi_i| + \sum_{i=\lfloor|\mathcal{K}|/l\rfloor l+1}^{|\mathcal{K}|} |\xi_i| + \sum_{i=0}^{\lfloor|\mathcal{K}|/l\rfloor-1}\sum_{j=1}^{l-1} |\xi_{il+j}| \\ &\le \lceil\frac{|\mathcal{K}|}{l}\rceil\left(\frac{8\Delta^2}{\beta^2\epsilon^2} + \frac{4\Delta}{3\beta\epsilon}\right)\log\left(\frac{2(n+1)}{\zeta}\right) \\ &\quad + \sum_{i=\lfloor|\mathcal{K}|/l\rfloor l+1}^{|\mathcal{K}|}\left(\frac{32L_f^2\|x_j-x_{j-1}\|^2 l^2}{\beta^2\epsilon^2} + \frac{8L_f\|x_j-x_{j-1}\| l}{3\beta\epsilon}\right)\log\left(\frac{2(n+1)l}{\zeta}\right) \\ &\quad + \sum_{i=0}^{\lfloor|\mathcal{K}|/l\rfloor-1}\sum_{j=1}^{l-1}\left(\frac{32L_f^2\|x_j-x_{j-1}\|^2 l^2}{\beta^2\epsilon^2} + \frac{8L_f\|x_j-x_{j-1}\| l}{3\beta\epsilon}\right)\log\left(\frac{2(n+1)l}{\zeta}\right) + |\mathcal{K}| \\ &= \mathcal{O}\left(\frac{|\mathcal{K}|}{l}\left(\frac{1}{\epsilon^2}+\frac{1}{\epsilon}\right) + \frac{l^2}{\epsilon^2} + |\mathcal{K}|^{1/2}\frac{l}{\epsilon}\right)\log\left(\frac{2(n+1)l}{\zeta}\right) + |\mathcal{K}| \\ &= \mathcal{O}\left(\epsilon^{-10/3}\log\left(\frac{1}{\rho\epsilon}\right)\right). \end{aligned}$$

We summarize above analysis into the following theorem.

THEOREM 5.3. *Suppose that conditions of Theorem 3.8 and Lemma 5.2 hold, with $l = \mathcal{O}(|\mathcal{K}|^{1/3})$, and Algorithm 2.1 with Algorithm 5.1 called to compute inexact oracles terminates in finite iterations. Then for given $\rho\in(0,1)$, with probability at least $1-\rho$, the algorithm returns an (ϵ,δ)-approximate first-order stationary point of* (5.1) *with the oracle complexity in order $\mathcal{O}(\epsilon^{-10/3}\log(1/(\rho\epsilon)))$, i.e., $\tilde{\mathcal{O}}(\epsilon^{-10/3})$.*

6. Numerical simulation. In this section, we consider the problem

$$\min_{x \in \mathcal{F}} \quad f(x) + \|Vx\|_p^p, \quad \text{s.t.} \quad Bx \le b, \tag{6.1}$$

where $\mathcal{F} = \{x \in \mathbb{R}^n : b_l \le x \le b_u\}$, $B \in \mathbb{R}^{r\times n}$, $b \in \mathbb{R}^r$, and $f(x) = \frac{1}{N}\sum_{i=1}^N ((A_i x - c_i)_+)^2$ with $A_i^T \in \mathbb{R}^n$ and $c_i \in \mathbb{R}$. By penalizing the constraints of (6.1) with τ being a penalty parameter, we obtain the penalty approximation problem in the form of (1.1)-(1.2):

$$\min_{x \in [b_l, b_u]} \quad \frac{1}{N}\sum_{i=1}^N ((A_i x - c_i)_+)^2 + \tau\|(Bx-b)_+\|_1 + \|Vx\|_p^p. \tag{6.2}$$

We apply Algorithm 2.1 to solve (6.2) by calling Algorithm 4.1 at kth iteration to compute inexact first-order oracle g_k, $k \ge 0$. Following (2.6), the subproblem at kth iteration is defined as

$$\begin{aligned} \min_{s \in \mathbb{R}^n} \quad & g_k^T s + \tau\|(Bx_k + Bs - b)_+\|_1 + \sum_{i \in \mathcal{A}_k} p|v_i^T x_k|^{p-1}|v_i^T(x_k+s)| + \frac{1}{2}\|s\|^2 \\ \text{s.t.} \quad & b_l \le x_k + s \le b_u, \quad v_i^T s = 0, \quad i \notin \mathcal{A}_k. \end{aligned}$$

By introducing $\bar{z} = (Bx_k + Bs - b)_+ \in \mathbb{R}^r$, and $\hat{z} = (\hat{z}_i, i \in \mathcal{A}_k)^T \in \mathbb{R}^{|\mathcal{A}_k|}$ with $\hat{z}_i = |v_i^T(x_k+s)|, i \in \mathcal{A}_k$, we obtain the following linearly constrained quadratic program:

$$\begin{aligned} \min_{s, \bar{z}, \hat{z}} \quad & g_k^T s + \tau e^T \bar{z} + \sum_{i \in \mathcal{A}_k} p|v_i^T x_k|^{p-1}\hat{z}_i + \frac{1}{2}\|s\|^2 \\ \text{s.t.} \quad & b_l \le x_k + s \le b_u, \quad v_i^T s = 0, \quad i \notin \mathcal{A}_k, \\ & 0 \le \bar{z}, \quad Bx_k + Bs - b \le \bar{z}, \quad -\hat{z}_i \le v_i^T(x_k+s) \le \hat{z}_i, \quad i \in \mathcal{A}_k. \end{aligned}$$

The numerical implementation was conducted in MATLAB R2022a on a PC with Intel I7-12700H 2.3GHZ CPU processor, 16GB RAM memory and a Windows operating system. We use Matlab default solver **quadprog** to solve each quadratic program. We generate the optimal solution x^* with $\|x^*\|_0 = K$ and set V, b_l, b_u, B, b, A, c as follows.

IndexK = randperm(n); x_0 = randn(n, 1); x^* = zeros(n, 1);
x^*(IndexK(1 : K)) = 2 ∗ (randn(K, 1) > 0.5) - 1; V= 0.1*eye(n);
b_l = -100 ∗ ones(n, 1); b_u = 100 ∗ ones(n, 1); B = rand(n, n); B = orth(B')';
b = B ∗x^*, A = randn(N, n); c = max(A ∗x^* + 0.01 ∗ randn(N, 1), 0);

In particular, we set parameters $n = 100, N = 10^5, K = 10, \epsilon = 10^{-4}, \bar{\beta} = 0.2, \tau = 200, \eta = 0.01, l = 10$ and the batch size as 1000. In Figure 1, we report the performances of the proposed algorithm. Specifically, Figures 1(a)-(d) showcase the behavior of different metrics, including the function value error $f(x_k) - f(x^*)$, the relative error between the iterate and x^* given by $\frac{\|x_k - x^*\|}{\|x_k\|}$, the number of nonzero entries in the iterate denoted as $\|x_k\|_0$, and the comparison between the nonzero entries of the output and x^*, respectively.

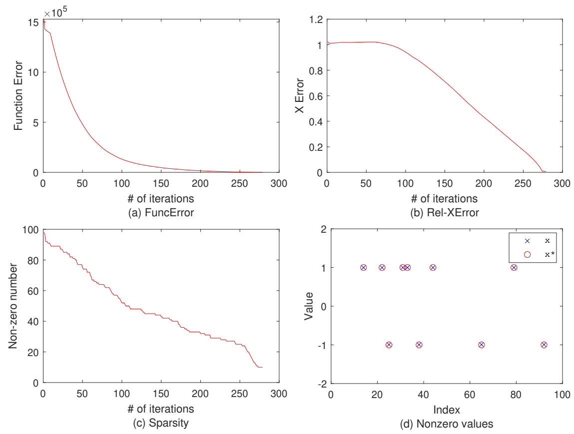

Fig. 1. *Numerical profiles on test problem* (6.2)

7. Conclusions. We present complexity analysis of proximal inexact gradient methods for finite-sum optimization with nonsmooth composite functions and a non-Lipschitz regularizer (1.1). Existence of the nonsmooth function h and non-Lipschitz term makes it inadequate to build an approximation model simply based on Taylor expansion as in [5, 8, 11, 12]. Moreover, those algorithms in [5, 8, 11, 12] rely on exact function values and gradients of f, which have difficulties in computation of problem (1.1) with the large scale finite-sum of f. In our Algorithm 2.1, we solve a strongly convex proximal subproblem (2.6) at each iteration without computing the function values and exact gradients of f, based on convex approximation to $f(x)+h(c(x))$ and a Lipschitz continuous approximation to $\|Vx\|_p^p$. By controlling inexactness of inexact gradients as well as subproblem solutions, we establish $\mathcal{O}(\epsilon^{-2})$ oracle complexity to find an (ϵ,δ)-approximate first-order stationary point of problem (1.1). This verifies that the worst-case oracle complexity still keeps the same with the absence of the differentiability of the Lipschitz term compared to [11, 12] and with the existence of non-Lipschitz regularizer in contrast to [5, 8]. Moreover, we propose a stochastic variant of Algorithm 2.1, by calling stochastic first-order oracles in a recursive way and applying a proper sampling strategy. We establish that the oracle complexity is in order $\tilde{\mathcal{O}}(\epsilon^{-2})$ to find an (ϵ,δ)-approximate first-order stationary point with high probability. We further extend the stochastic variant of algorithm to solve problems in the expectation form and derive the oracle complexity in order $\tilde{\mathcal{O}}(\epsilon^{-10/3})$ with high probability. Numerical performances of the proposed algorithm are also reported on a test problem.

REFERENCES

[1] W. Bian and X. Chen, *Linearly constrained non-Lipschitz optimization for image restoration*, SIAM J. Imaging Sci., 8 (2015), pp. 2294–2322.

[2] W. Bian and X. Chen, *Optimality and complexity for constrained optimization problems with nonconvex regularization*, Math. Oper. Res., 42 (2017), pp. 1063–1084.

[3] W. Bian, X. Chen, and Y. Ye, *Complexity analysis of interior point algorithms for non-Lipschitz and nonconvex minimization*, Math. Program., 149 (2015), pp. 301–327.

[4] K. Bui, F. Park, S. Zhang, Y. Qi, and J. Xin, *Structured sparsity of convolutional neural networks via nonconvex sparse group regularization*, Front. Appl. Math. Stat., (2021), https://doi.org/10.3389/fams.2020.529564.

[5] C. Cartis, N. Gould, and P. Toint, *On the evaluation complexity of composite function*

minimization with applications to nonconvex nonlinear programming, SIAM J. Optim., 21 (2011), pp. 1721–1739.

[6] C. Cartis, N. Gould, and P. Toint, *Sharp worst-case evaluation complexity bounds for arbitrary-order nonconvex optimization with inexpensive constraints*, SIAM J. Optim., 30 (2020), pp. 513–541.

[7] C. Cartis, N. Gould, and P. Toint, *Strong evaluation complexity of an inexact trust-region algorithm for arbitrary-order unconstrained nonconvex optimization*, https://doi.org/10.48550/arXiv.2011.00854, (2021).

[8] C. Cartis, N. Gould, and P. Toint, *Strong evaluation complexity bounds for arbitrary-order optimization of nonconvex nonsmooth composite functions*, Proceedings of the International Congress of Mathematicians (ICM 2022). EMS Press. https://doi.org/10.4171/ICM2022/95, (2022).

[9] X. Chen, D. Ge, Z. Wang, and Y. Ye, *Complexity of unconstrained ℓ_2-ℓ_p minimization*, Math. Program., 143 (2014), pp. 371–383.

[10] X. Chen, Z. Lu, and T. Pong, *Penalty methods for a class of non-Lipschitz optimization problems*, SIAM J. Optim., 26 (2016), pp. 1465–1492.

[11] X. Chen and P. Toint, *High-order evaluation complexity for convexly-constrained optimization with non-Lipschitzian group sparsity terms*, Math. Program., 187 (2020), pp. 47–78.

[12] X. Chen, P. Toint, and H. Wang, *Complexity of partially-separable convexly-constrained optimization with non-Lipschitzian singularities*, SIAM J. Optim., 29 (2019), pp. 874–903.

[13] X. Chen, F. Xu, and Y. Ye, *Lower bound theory of nonzero entries in solutions of ℓ_2-ℓ_p minimization*, SIAM J. Sci. Comput., 32 (2010), pp. 2832–2852.

[14] W. Cheng, X. Wang, and X. Chen, *An interior stochastic gradient method for a class of non-Lipschitz optimization problems*, J. Sci. Comput., 92 (2022).

[15] Y. Cui, Z. He, and J.-S. Pang, *Multicomposite nonconvex optimization for training deep neural networks*, SIAM J. Optim., 30 (2020), pp. 1693–1723.

[16] D. Ge, R. He, and S. He, *An improved algorithm for the L_2-L_p minimization problem*, Math. Program., 166 (2017), pp. 131–158.

[17] S. Gratton, E. Simon, and P. Toint, *An algorithm for the minimization of nonsmooth nonconvex functions using inexact evaluations and its worst-case complexity*, Math. Program., 187 (2021), pp. 1–24.

[18] A. Jain, *Fundamentals of Digital Image Processing*, Upper Saddle River, NJ: Prentice-Hall, 1989.

[19] D. Li, Z. Sun, and X. Zhang, *A constrained optimization reformulation and a feasible descent direction method for $L_{1/2}$ regularization*, Comput. Optim. Appl., 59 (2014), pp. 263–284.

[20] Y. Li, S. Gu, and R. T. C. Mayer, L. V. Gool, *Group sparsity: The hinge between filter pruning and decomposition for network compression*, CVPR, (2020).

[21] Y. Liu, S. Ma, Y. Dai, and S. Zhang, *A smoothing SQP framework for a class of composite L_q minimization over polyhedron*, Math. Program., 158 (2016), pp. 467–500.

[22] M. Metel and A. Takeda, *Simple stochastic gradient methods for non-smooth non-convex regularized optimization*, ICML, (2019), pp. 4537–4545.

[23] L. Nguyen, J. Liu, K. Scheinberg, and M. Takac, *SARAH: a novel method for machine learning problems using stochastic recursive gradient*, ICML, (2017), pp. 2613–2621.

[24] R. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer Verlag, Heidelberg, Berlin, New York, 1998.

[25] S. Scardapane, D. Comminiello, A. Hussain, and A. Uncini, *Group sparse regularization for deep neural networks*, Neurocomputing, 241 (2017), pp. 81–89.

[26] J. A. Tropp, *User-friendly tail bounds for sums of random matrices*, Found Comput Math, 12 (2012), pp. 389–434.

[27] Z. Wang, Y. Zhou, Y. Liang, and G. Lan, *Stochastic variance-reduced cubic regularization for nonconvex optimization*, Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics, PMLR, 89 (2019), pp. 2731–2740.

[28] Y. Xu, R. Jin, and T. Yang, *Non-asymptotic analysis of stochastic methods for non-smooth non-convex regularized problems*, in NeurIPS, H. Wallach, H. Larochelle, A. Beygelzimer, F. Alché-Buc, E. Fox, and R. Garnett, eds., vol. 32, Curran Associates, Inc., 2019.

[29] Y. Xu, Q. Qi, Q. Lin, R. Jin, and T. Yang, *Stochastic optimization for DC functions and non-smooth non-convex regularizers with non-asymptotic convergence*, ICML, (2019), pp. 6942–6951.

[30] Z. Xu, X. Chang, F. Xu, and H. Zhang, *$L_{1/2}$ regularization: A thresholding representation theory and a fast solver*, IEEE T. Neur. Net. Lear., 23 (2012), pp. 1013–1027.

[31] J. Yoon and S. Hwang, *Combined group and exclusive sparsity for deep neural networks*, ICML, (2017), pp. 3958–3966.