Oracle Complexity of Single-Loop Switching Subgradient Methods for Non-Smooth Weakly Convex Functional Constrained Optimization

Yankun Huang
Department of Business Analytics
University of Iowa
Iowa City, IA 52242
yankun-huang@uiowa.edu

Qihang Lin
Department of Business Analytics
University of Iowa
Iowa City, IA 52242
qihang-lin@uiowa.edu

Abstract

We consider a non-convex constrained optimization problem, where the objective function is weakly convex and the constraint function is either convex or weakly convex. To solve this problem, we consider the classical switching subgradient method, which is an intuitive and easily implementable first-order method whose oracle complexity was only known for convex problems. This paper provides the first analysis on the oracle complexity of the switching subgradient method for finding a nearly stationary point of non-convex problems. Our results are derived separately for convex and weakly convex constraints. Compared to existing approaches, especially the double-loop methods, the switching gradient method can be applied to non-smooth problems and achieves the same complexity using only a single loop, which saves the effort on tuning the number of inner iterations.

1 Introduction

Continuous optimization with nonlinear constraints arises from many applications of machine learning and statistics. Examples include Neyman-Pearson classification [62] and learning with fairness constraints [77]. In this paper, we consider the following general nonlinear constrained optimization

$$f^* \equiv \min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}) \quad \text{s.t.} \quad g(\mathbf{x}) \leq 0, \tag{1}$$

where $\mathcal{X} \subset \mathbb{R}^d$ is a closed convex set that allows a computationally easy projection operator, f is weakly-convex, g is either convex or weakly convex, and functions f and g are not necessarily smooth. When $g(\mathbf{x}) \equiv \max_{i=1,\ldots,m} g_i(\mathbf{x})$, (1) is equivalent to an optimization problem with multiple nonlinear constraints $g_i(\mathbf{x}) \leq 0$ for $i = 1, \ldots, m$.

A weakly convex function can be non-convex, so computing an optimal solution of (1) is challenging in general even without constraints. For this reason, theoretical analysis for gradient-based algorithms for non-convex problems mostly focuses on an algorithm's (oracle) complexity for finding an ϵ-stationary solution for (1). When a problem is non-smooth, finding an ϵ-stationary solution is generally difficult even if the problem is convex [47]. Hence, in this paper, we consider finding a nearly ϵ-stationary solution for (1), whose definition will be stated later in Definition 3.2.

In the past decade, there have been many studies on non-convex constrained optimization. However, most of the existing algorithms and their theoretical complexity analysis are developed by assuming f and g_i's are all smooth or can be written as the sum of a smooth and a simple non-smooth functions. A non-exhaustive list of the works under such an assumption includes [34, 80, 81, 55, 44, 45, 43, 52, 63, 53, 50, 12, 22, 9, 21, 40, 51]. Their results cannot be applied to (1) due to non-smoothness in the problem.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Non-smoothness is common in optimization in machine learning, e.g., when a non-smooth loss function is applied, but there are much fewer studies on non-smooth non-convex constrained optimization. Under the weak-convexity assumption, an effective approach for solving a non-smooth non-convex problem with theoretical guarantees is the (inexact) proximal point method, where a quadratic proximal term is added to objective and constraint functions to construct a strongly convex constrained subproblem and then a sequence of solutions can be generated by solving this subproblem inexactly and updating the center of the proximal term. Oracle complexity for this method to find a nearly ϵ-stationary has been established by [13, 54, 39] under different constraint qualifications.

The inexact proximal point method is a *double-loop* algorithm where the inner loop is another optimization algorithm for solving the aforementioned strongly convex subproblems. The complexity results in [13, 54, 39] require the inner loop solves each subproblem to a targeted optimality gap. However, the optimality gap is hard to evaluate and thus cannot be used to terminate the inner loop. Although the number of inner iterations needed to achieve the targeted gap can be bounded theoretically, the bound usually involves some constants that are unknown or can only be estimated conservatively. Hence, using the theoretical iteration bound to stop the inner loop usually leads to significantly more inner iterations than needed, making the whole algorithm inefficient. In practices, users often need to tune the number of inner iterations to improve algorithm's efficiency, which is an inconvenience common to most double-loop methods.

On the contrary, a *single-loop* algorithm is usually easier to implement as it does not require tuning the number of inner iterations. Therefore, the **main contribution** of this paper is showing that a single-loop first-order algorithm can find a nearly ϵ-stationary point of (1) with oracle complexity $O(1/\epsilon^4)$, which matches state-of-the-art results obtained only by the double-loop methods [13, 54, 39]. The algorithm we study is the classical *switching subgradient* (SSG) method proposed by Polyak [59] in 1967, which is intuitive and easy to implement but has only been analyzed before in the convex case. We first show that the SSG method has complexity complexity $O(1/\epsilon^4)$ when g is convex. Then, we show that the same complexity result holds when g is weakly convex and appropriate constraint qualification holds. We also present some practical examples where all of our assumptions hold, including a fair classification problem subject to the demographic parity constraint [1]. Our **technical novelty** is inventing a *switching stepsize rule* to accompany the switching subgradient. In particular, we use a fixed stepsize when the solution is updated by the objective's subgradient while use an adaptive Polyak's stepsize [33, 60] when the solution is updated by the constraint's subgradient. This allows us to leverage the local error bound of the constraint function to keep the solution nearly feasible during the algorithm, which prevents the solution from being trapped at an infeasible stationry point. To the best of our knowledge, this paper is the first to establish the complexity of a single-loop first-order method for weakly convex non-smooth nonlinear constrained optimization. In the appendix, we also extend our algorithms and complexity analysis to the stochastic setting.

2   Related work

Non-convex constrained optimization has a long history [36, 16, 29, 17, 30, 6, 18] and the interest on this subject has been still growing in the machine learning community because of its new applications such as learning with fairness constraints (see e.g., [77]).

In recent literature, the prevalent classes of algorithms for non-convex constrained optimization include the augmented Lagrangian method (ALM), the penalty method [37, 79, 80, 81, 38, 42, 55, 44, 45, 69, 43, 52, 63, 53, 50, 40, 51], and the sequential quadratic programming method [11, 34, 12, 22, 10, 9, 8, 21]. Besides, an inexact projected gradient method is developed by [15] and a level conditional gradient method is developed by [20]. However, these works all focus on the case where g is smooth and f is either smooth or equals $f_1 + f_2$, where f_1 is smooth and non-convex while f_2 is non-smooth and convex and has a simple structure that allows a closed-form solution for the proximal mapping $\arg\min_{\mathbf{y}}\{f_2(\mathbf{y}) + \frac{\rho}{2}\|\mathbf{y} - \mathbf{x}\|^2\}$ for any $\mathbf{x}$. There are relatively fewer works on non-convex non-smooth constrained problems. An alternating direction method of multipliers (ADMM) and an ALM are studied by [75] and [78], respectively, for non-convex non-smooth problems with linear constraints while our study considers nonlinear non-smooth constraints. The methods by [20] and [12] can be extended to a structured non-smooth case where $f = f_1 + f_2$ with f_1 being smooth non-convex and $f_2 = \max_{\mathbf{y}}\{\mathbf{y}^\top A\mathbf{x} - \phi(\mathbf{y})\}$ with a convex ϕ, and g has a similar structure. The method by [14] can handle a specific non-smooth non-convex constraint, i.e., $g(\mathbf{x}) = \lambda\|\mathbf{x}\|_1 - h(\mathbf{x})$

2

where h is a convex and smooth. Compared to these works, our results apply to a more general non-smooth problem without those structure assumptions.

When f and g in (1) are weakly convex and non-smooth, the inexact proximal point method has been studied by [13, 54, 39] under different constraint qualifications and different notions of stationarity. Their complexity analysis utilizes the relationship between the gradient of the Moreau envelope of (1) and the near stationarity of a solution, which is originally used to analyze complexity of subgradient methods for weakly convex non-smooth unconstrained problems [24, 25, 27, 2, 28, 61, 83]. Our analysis utilizes a similar framework. The methods [13, 54, 39] are double-loop while our algorithm only uses a single loop and achieves the same complexity of $O(1/\epsilon^4)$ as them under similar assumptions.

The SSG algorithm is first proposed by Polyak [59]. It has been well-studied for convex problems [58, 7, 49, 67, 72, 73, 65, 68, 66, 3] and quasi-convex problems [67]. This paper provides the first complexity analysis for the SSG method under weak convexity assumption. Non-smooth non-convex optimization has also been studied without weak convexity assumption by [82, 46, 64, 47, 19, 70, 71]. These works analyze the complexity of first-order methods for computing an (ϵ, δ)-Goldstein approximate stationary point, which is a more general stationarity notation than what we consider here. However, these works only focus on unconstrained problems.

3   Algorithm and near stationarity

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

$$\partial h(\mathbf{x}) = \left\{ \boldsymbol{\zeta} \in \mathbb{R}^d \mid h(\mathbf{x}') \geq h(\mathbf{x}) + \langle \boldsymbol{\zeta}, \mathbf{x}' - \mathbf{x} \rangle + o(\|\mathbf{x}' - \mathbf{x}\|), \ \mathbf{x}' \to \mathbf{x} \right\},$$

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

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

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

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

for any $(\mathbf{x}, \mathbf{x}') \in \mathcal{X} \times \mathcal{X}$ and any $\boldsymbol{\zeta} \in \partial h(\mathbf{x}')$. We denote the normal cone of $\mathcal{X}$ at $\mathbf{x}$ by $\mathcal{N}_{\mathcal{X}}(\mathbf{x})$ and the relative interior of $\mathcal{X}$ by relint$(\mathcal{X})$. We say a point $\mathbf{x}$ is ϵ-feasible if $\mathbf{x} \in \mathcal{X}$ and $g(\mathbf{x}) \leq \epsilon$. Let $\delta_{\mathcal{X}}(\mathbf{x})$ be the zero-infinity characteristic function of set $\mathcal{X}$, proj$_{\mathcal{X}}(\cdot)$ be the projection mapping to $\mathcal{X}$, and dist$(\mathbf{x}, \mathcal{A}) := \min_{\mathbf{y} \in \mathcal{A}} \|\mathbf{x} - \mathbf{y}\|$ for set $\mathcal{A}$.

We make the following assumptions on (1) throughout the paper.

Assumption 3.1. Functions f and g are continuous with $\partial f(\mathbf{x}) \neq \emptyset$ and $\partial g(\mathbf{x}) \neq \emptyset$ on $\mathcal{X}$, and there exists M such that $\|\boldsymbol{\zeta}_f\| \leq M$ and $\|\boldsymbol{\zeta}_g\| \leq M$ for any $\mathbf{x} \in \mathcal{X}$, $\boldsymbol{\zeta}_f \in \partial f(\mathbf{x})$ and $\boldsymbol{\zeta}_g \in \partial g(\mathbf{x})$.

Since (1) is non-convex, finding an ϵ-optimal solution is intractable in general. For a non-convex problem, the target is typically to find a *stationary* point of (1), which is a point $\mathbf{x}^* \in \mathcal{X}$ that satisfies the following Karush-Kuhn-Tucker (KKT) conditions:

$$\boldsymbol{\zeta}_f^* + \lambda^* \boldsymbol{\zeta}_g^* \in -\mathcal{N}_{\mathcal{X}}(\mathbf{x}^*), \quad \lambda^* g(\mathbf{x}^*) = 0, \quad g(\mathbf{x}^*) \leq 0, \quad \lambda^* \geq 0,$$

where $\lambda^* \in \mathbb{R}$ is a Lagrangian multiplier, $\boldsymbol{\zeta}_f^* \in \partial f(\mathbf{x}^*)$ and $\boldsymbol{\zeta}_g^* \in \partial g(\mathbf{x}^*)$. Typically, an exact stationary point can only be approached by an algorithm at full convergence, which may require infinitely many iterations. Within a finite number of iterations, an algorithm can only generate an ϵ-*stationary* point [13], which is a point $\widehat{\mathbf{x}} \in \mathcal{X}$ satisfying

$$\text{dist}\left(\widehat{\boldsymbol{\zeta}}_f + \widehat{\lambda}\widehat{\boldsymbol{\zeta}}_g, -\mathcal{N}_{\mathcal{X}}(\widehat{\mathbf{x}})\right) \leq \epsilon, \quad |\widehat{\lambda}g(\widehat{\mathbf{x}})| \leq \epsilon^2, \quad g(\widehat{\mathbf{x}}) \leq \epsilon^2, \quad \widehat{\lambda} \geq 0, \tag{2}$$

where $\widehat{\lambda} \in \mathbb{R}$ is a Lagrangian multiplier, $\widehat{\boldsymbol{\zeta}}_f \in \partial f(\widehat{\mathbf{x}})$ and $\widehat{\boldsymbol{\zeta}}_g \in \partial g(\widehat{\mathbf{x}})$. However, because f and g are non-smooth, computing an ϵ-stationary point is still challenging even for an unconstrained problem. Nevertheless, under the weak convexity assumption, it is possible to compute a *nearly ϵ-stationary point*, which we will introduce next.

Given $\hat{\rho} \geq 0$, $\tilde{\rho} \geq 0$ and $\mathbf{x} \in \mathcal{X}$, we define a quadratically regularized problem of (1) as

3

Algorithm 1 Switching Subgradient (SSG) Method by Polyak [59]

1: **Input:** $\mathbf{x}^{(0)} \in \mathcal{X}$, total number of iterations T, tolerance of infeasibility $\epsilon_t \geq 0$, stepsize η_t, and a starting index S for generating outputs.
2: **Initialization:** $I = \emptyset$ and $J = \emptyset$.
3: **for** iteration $t = 0, 1, \cdots, T - 1$ **do**
4:    **if** $g(\mathbf{x}^{(t)}) \leq \epsilon_t$ **then**
5:       Set $\mathbf{x}^{(t+1)} = \mathrm{proj}_{\mathcal{X}}(\mathbf{x}^{(t)} - \eta_t \boldsymbol{\zeta}_f^{(t)})$ for any $\boldsymbol{\zeta}_f^{(t)} \in \partial f(\mathbf{x}^{(t)})$ and, if $t \geq S$, $I = I \cup \{t\}$.
6:    **else**
7:       Set $\mathbf{x}^{(t+1)} = \mathrm{proj}_{\mathcal{X}}(\mathbf{x}^{(t)} - \eta_t \boldsymbol{\zeta}_g^{(t)})$ for any $\boldsymbol{\zeta}_g^{(t)} \in \partial g(\mathbf{x}^{(t)})$ and, if $t \geq S$, $J = J \cup \{t\}$.
8:    **end if**
9: **end for**
10: **Output I:** $\mathbf{x}^{(\tau)}$ with τ randomly sampled from I using $\mathrm{Prob}(\tau = t) = \eta_t / \sum_{s \in I} \eta_s$.
11: **Output II:** $\mathbf{x}^{(\tau)}$ with τ randomly sampled from $I \cup J$ using $\mathrm{Prob}(\tau = t) = \eta_t / \sum_{s \in I \cup J} \eta_s$.

$$\varphi(\mathbf{x}) \equiv \min_{\mathbf{y} \in \mathcal{X}} \left\{ f(\mathbf{y}) + \frac{\hat{\rho}}{2} \|\mathbf{y} - \mathbf{x}\|^2, \ s.t. \ g(\mathbf{y}) + \frac{\tilde{\rho}}{2} \|\mathbf{y} - \mathbf{x}\|^2 \leq 0 \right\}, \tag{3}$$

$$\widehat{\mathbf{x}}(\mathbf{x}) \equiv \arg\min_{\mathbf{y} \in \mathcal{X}} \left\{ f(\mathbf{y}) + \frac{\hat{\rho}}{2} \|\mathbf{y} - \mathbf{x}\|^2, \ s.t. \ g(\mathbf{y}) + \frac{\tilde{\rho}}{2} \|\mathbf{y} - \mathbf{x}\|^2 \leq 0 \right\}. \tag{4}$$

Following the literature on weakly convex optimization [24, 27, 23, 13, 54, 39], we use the value of $\|\widehat{\mathbf{x}}(\mathbf{x}) - \mathbf{x}\|$ as a measure of the quality of a solution $\mathbf{x}$ because it can be interpreted as a stationarity measure. For the purpose of illustration, we assume for now that $\widehat{\mathbf{x}}(\mathbf{x})$ is uniquely defined and there exists a Lagrangian multiplier $\widehat{\lambda} \in \mathbb{R}$ such that the following KKT conditions of (4) holds.

$$\widehat{\boldsymbol{\zeta}}_f + \hat{\rho}(\widehat{\mathbf{x}}(\mathbf{x}) - \mathbf{x}) + \widehat{\lambda} \left(\widehat{\boldsymbol{\zeta}}_g + \tilde{\rho}(\widehat{\mathbf{x}}(\mathbf{x}) - \mathbf{x}) \right) \in -\mathcal{N}_{\mathcal{X}}(\widehat{\mathbf{x}}(\mathbf{x})),$$

$$\widehat{\lambda} \left(g(\widehat{\mathbf{x}}(\mathbf{x})) + \frac{\tilde{\rho}}{2} \|\widehat{\mathbf{x}}(\mathbf{x}) - \mathbf{x}\|^2 \right) = 0, \quad g(\widehat{\mathbf{x}}(\mathbf{x})) + \frac{\tilde{\rho}}{2} \|\widehat{\mathbf{x}}(\mathbf{x}) - \mathbf{x}\|^2 \leq 0, \quad \widehat{\lambda} \geq 0, \tag{5}$$

where $\widehat{\boldsymbol{\zeta}}_f \in \partial f(\widehat{\mathbf{x}}(\mathbf{x}))$ and $\widehat{\boldsymbol{\zeta}}_g \in \partial g(\widehat{\mathbf{x}}(\mathbf{x}))$. Therefore, as long as $\|\widehat{\mathbf{x}}(\mathbf{x}) - \mathbf{x}\| \leq \epsilon$, we have

$$\mathrm{dist}\left(\widehat{\boldsymbol{\zeta}}_f + \widehat{\lambda}\widehat{\boldsymbol{\zeta}}_g, -\mathcal{N}_{\mathcal{X}}(\widehat{\mathbf{x}}(\mathbf{x})) \right) \leq (\hat{\rho} + \widehat{\lambda}\tilde{\rho})\epsilon, \quad |\widehat{\lambda} g(\widehat{\mathbf{x}}(\mathbf{x}))| = \widehat{\lambda}\tilde{\rho}\epsilon^2/2, \quad g(\widehat{\mathbf{x}}(\mathbf{x})) \leq 0. \tag{6}$$

This means $\widehat{\mathbf{x}}(\mathbf{x})$ is an $O(\epsilon)$-stationary point of the original problem (1) in the sense of (2). Since $\mathbf{x}$ is within an ϵ-distance from $\widehat{\mathbf{x}}(\mathbf{x})$, we call such an $\mathbf{x}$ a *nearly ϵ-stationary point* of (1). We formalize this definition as follows.

Definition 3.2. Suppose $\widehat{\mathbf{x}}(\mathbf{x})$ is defined in (4) with $\epsilon \geq 0$. A (stochastic) point $\mathbf{x} \in \mathcal{X}$ is a (stochastic) nearly ϵ-stationary point of (1) if $\mathbb{E}[\|\widehat{\mathbf{x}}(\mathbf{x}) - \mathbf{x}\|] \leq \epsilon$.

Of course, we can claim $\widehat{\mathbf{x}}(\mathbf{x})$ is an $O(\epsilon)$-stationary point of (1) based on (6) only when $\widehat{\lambda}$ in (5) exists and does not go to infinity as ϵ approaches zero. Fortunately, we can show in Lemmas 4.2 and 5.2 that this is true under some constraint qualifications, which justifies Definition 3.2.

We present the SSG method in Algorithm 1 for finding a nearly ϵ-stationary point of (1). At iteration t, we check if the current solution $\mathbf{x}^{(t)}$ is nearly feasible in the sense that $g(\mathbf{x}^{(t)}) \leq \epsilon_t$ for a pre-determined tolerance of infeasibility ϵ_t. If yes, the algorithm performs a subgradient step along the subgradient of f. Otherwise, the algorithm switches the updating direction to the subgradient of g. The algorithm records the iteration indexes of the nearly feasible solutions in set I and other indexes in set J. The final output is randomly sampled from the iterates in I or $I \cup J$ with a distribution weighted by the stepsizes η_t's. An index S is set so the algorithm only starts to record I and J when $t \geq S$. We study the theoretical *oracle complexity* of Algorithm 1 for finding a nearly ϵ-stationary point, which is defined as the total number of times for which the algorithm queries the subgradient or function value of f or g. Our results are presented separately when g is convex and when g is weakly convex.

4 Convex constraints

In this section, we first consider a relatively easy case where f is weakly convex but g is convex. In particular, we make the following assumptions in addition to Assumption 3.1 in this section.

4

Assumption 4.1. The following statements hold:

    A. $f(\mathbf{x})$ is ρ-weakly convex on $\mathcal{X}$ and $g(\mathbf{x})$ is convex on $\mathcal{X}$.

    B. (Slater's condition) There exists $\mathbf{x}_{\text{feas}} \in \text{relint}(\mathcal{X})$ such that $g(\mathbf{x}_{\text{feas}}) < 0$.

    C. There exists $D \in \mathbb{R}$ such that $\|\mathbf{x} - \mathbf{x}'\| \leq D$ for any $\mathbf{x}$ and $\mathbf{x}'$ in $\mathcal{X}$.

In this section, we choose parameters in (3) such that

$$\hat{\rho} > \rho \text{ and } \tilde{\rho} = 0. \tag{7}$$

Under Assumption 4.1, (7) guarantees that (3) is strictly feasible, its objective function is $(\hat{\rho} - \rho)$-strongly convex and its constraint function is convex, so $\hat{\mathbf{x}}(\mathbf{x})$ in (4) is unique and $\hat{\lambda}$ in (5) exists. We first present an upper bound of $\hat{\lambda}$ that is independent of $\mathbf{x}$. The proof is in Section A.1.

Lemma 4.2. *Suppose Assumptions 3.1 and 4.1 hold. Given any $\mathbf{x} \in \mathcal{X}$, let $\hat{\mathbf{x}}(\mathbf{x})$ be defined as in (4) with $(\hat{\rho}, \tilde{\rho})$ satisfying (7) and $\hat{\lambda}$ be the associated Lagrangian multiplier satisfying (5). We have*

$$\hat{\lambda} \leq \Lambda := (MD + \hat{\rho}D^2)/(-g(\mathbf{x}_{\text{feas}})). \tag{8}$$

For simplicity of notation, we denote $\hat{\mathbf{x}}(\mathbf{x}^{(t)})$ defined in (4) by $\hat{\mathbf{x}}^{(t)}$. Let $\mathbb{E}_\tau[\cdot]$ be the expectation taken only over the random index τ when the algorithms stop. We present the convergence properties of Algorithm 1 when ϵ_t and η_t are static and diminishing. The proof is provided in Section A.4.

Theorem 4.3. *Suppose Assumptions 3.1 and 4.1 hold and Λ is as in (8). Let $\hat{\mathbf{x}}(\mathbf{x}^{(t)})$ be defined as in (4) with $(\hat{\rho}, \tilde{\rho})$ satisfying (7) and $\mathbf{x}^{(\tau)}$ is generated by Output I. Algorithm 1 guarantees $\mathbb{E}_\tau[\|\hat{\mathbf{x}}^{(\tau)} - \mathbf{x}^{(\tau)}\|] \leq \epsilon$ and $\mathbb{E}_\tau[g(\mathbf{x}^{(\tau)})] \leq \frac{\epsilon^2(\hat{\rho}-\rho)}{1+\Lambda}$ in either of the following cases.*

Case I: $S = 0$, $\epsilon_t = \frac{\epsilon^2(\hat{\rho}-\rho)}{1+\Lambda}$, $\eta_t = \frac{2\epsilon^2(\hat{\rho}-\rho)}{5(1+\Lambda)M^2}$ and $T \geq \frac{25M^2D^2(1+\Lambda)^2}{4\epsilon^4(\hat{\rho}-\rho)^2} = O(1/\epsilon^4)$.

Case II: $S = T/2$, $\epsilon_t = \frac{5MD}{\sqrt{t+1}}$, $\eta_t = \frac{D}{M\sqrt{t+1}}$ and $T \geq \frac{50M^2D^2(1+\Lambda)^2}{\epsilon^4(\hat{\rho}-\rho)^2} = O(1/\epsilon^4)$.

Algorithm 1 is single-loop with $O(1)$ oracle complexity per iteration, so its total complexity is just $T = O(1/\epsilon^4)$, which matches the start-of-the-art complexity by [13, 54, 39]. This result can be generalized to the case where there exist additional linear equality constraints $\mathbf{Ax} = \mathbf{b}$. See the extension of Lemma 4.2 in Section A.1 and Remark A.2.

Assumption 4.1C can be relaxed to only require that the feasible set, i.e., $\mathcal{S} = \{\mathbf{x} \in \mathcal{X} \mid g(\mathbf{x}) \leq 0\}$, is bounded instead of $\mathcal{X}$. The same complexity can be achieved by Algorithm 1 by using a switching step size rule similar to η_t in Proposition 5.5. This result is provided in Section A.8 but we recommend interested readers to read Section 5 first to get introduced to this special step size.

Remark 4.4. Property $\mathbb{E}_\tau[g(\mathbf{x}^{(\tau)})] \leq \frac{\epsilon^2(\hat{\rho}-\rho)}{1+\Lambda}$ in the theorems above is not required by Definition 3.2. By Assumption 3.1A, $\mathbb{E}_\tau[g(\mathbf{x}^{(\tau)})] \leq \mathbb{E}_\tau[g(\hat{\mathbf{x}}^{(\tau)})] + M\mathbb{E}_\tau[\|\hat{\mathbf{x}}^{(\tau)} - \mathbf{x}^{(\tau)}\|] \leq M\epsilon$, which means a nearly ϵ-stationary point must be $O(\epsilon)$-feasible by definition. Property $\mathbb{E}_\tau[g(\mathbf{x}^{(\tau)})] \leq \frac{\epsilon^2(\hat{\rho}-\rho)}{1+\Lambda}$ implies $O(\epsilon^2)$-feasibility for the output, which is even better.

When g is μ-strongly convex with $\mu > 0$, we can show that the complexity of Algorithm 1 is still $O(1/\epsilon^4)$ but one can simply set $\epsilon_t = 0$, which makes η_t the only tuning parameter. This makes this single-loop method even more attractive. Due to space limit, we include this result in Section A.5. We also extend our result to the stochastic case in Section A.6 and A.7.

5   Weakly convex constraints

Next we consider the case where both f and g are weakly convex but not necessarily convex. Let

$$g_+(\mathbf{x}) = \max\{g(\mathbf{x}), 0\}, \quad \mathcal{L} = \{\mathbf{x} \in \mathcal{X} \mid g(\mathbf{x}) = 0\} \text{ and } \mathcal{S} = \{\mathbf{x} \in \mathcal{X} \mid g(\mathbf{x}) \leq 0\}.$$

We make the following assumptions in addition to Assumption 3.1 in this section.

Assumption 5.1. The following statements hold:

5

A. $f(\mathbf{x})$ and $g(\mathbf{x})$ are ρ-weakly convex on $\mathcal{X}$.

B. There exist $\bar{\epsilon} > 0$, $\theta > 0$ and $\bar{\rho} > \rho$ such that, for any $\bar{\epsilon}^2$-feasible solution $\mathbf{x}$, there exists $\mathbf{y} \in \mathrm{relint}(\mathcal{X})$ such that $g(\mathbf{y}) + \frac{\bar{\rho}}{2}\|\mathbf{y} - \mathbf{x}\|^2 \leq -\theta$.

C. $\underline{f} := \inf_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}) > -\infty$.

Assumption 5.1B is called the uniform Slater's condition by [54]. We will present two real-world examples in Section B.1 that satisfy this assumption, including a fair classification problem under demographic parity constraint, which is one of the applications in our numerical experiments in Section 6.2. In this section, we choose parameters in (3) such that

$$\bar{\rho} \geq \hat{\rho} = \tilde{\rho} > \rho. \tag{9}$$

Under Assumption 5.1, (9) guarantees that (3) is uniformly strictly feasible for any ϵ^2-feasible $\mathbf{x}$, and the objective and constraint functions of (3) are both $(\hat{\rho} - \rho)$-strongly convex, so $\widehat{\mathbf{x}}(\mathbf{x})$ is uniquely defined and $\widehat{\lambda}$ in (5) exists. In addition, Assumption 5.1 has the following three implications that play important roles in our complexity analysis.

First, $\widehat{\lambda}$ in (5) can be bounded by a constant independent of $\mathbf{x}$ and ϵ as long as $\mathbf{x}$ is ϵ^2-feasible with $\epsilon \leq \bar{\epsilon}$. This result is similar to Lemma 1 by Ma et al. [54] except that they require $\mathcal{X}$ to be bounded but we do not. The proof is given in Section B.2.

Lemma 5.2. *Suppose Assumptions 3.1 and 5.1 hold. Given any ϵ^2-feasible $\mathbf{x}$ with any $\epsilon \leq \bar{\epsilon}$, let $\widehat{\mathbf{x}}(\mathbf{x})$ defined as in (4) satisfying (9) and $\widehat{\lambda}$ is the associated Lagrangian multiplier satisfying (5). We have*

$$\|\widehat{\mathbf{x}}(\mathbf{x}) - \mathbf{x}\| \leq M/\hat{\rho} \quad and \quad \widehat{\lambda} \leq \Lambda' := 2M/\sqrt{2\theta(\hat{\rho} - \rho)}. \tag{10}$$

Second, the subgradient of the constraint function $g(\mathbf{x}) + \delta_{\mathcal{X}}(\mathbf{x})$ on $\mathcal{L}$ is uniformly away from the origin. The proof is provided in Section B.2.

Lemma 5.3. *Suppose Assumptions 3.1 and 5.1 hold. It holds for any $\mathbf{x} \in \mathcal{L}$ that*

$$\min_{\boldsymbol{\zeta}_g \in \partial g(\mathbf{x}), \mathbf{u} \in \mathcal{N}_{\mathcal{X}}(\mathbf{x})} \|\boldsymbol{\zeta}_g + \mathbf{u}\| \geq \nu := \sqrt{2\theta(\hat{\rho} - \rho)}. \tag{11}$$

Lastly, note that $\mathcal{S}$ is the optimal set of $\min_{\mathbf{x} \in \mathcal{X}} g_+(\mathbf{x})$, which is a ρ-weakly convex non-smooth optimization problem with an optimal value of zero. Lemma 5.3 implies that $g_+(\mathbf{x})$ is sharp near the boundary of $\mathcal{S}$, meaning that $g_+(\mathbf{x})$ satisfies a linear error bound in an $O(1)$-neighborhood of $\mathcal{S}$. A similar result for a convex g_+ is given in Lemma 1 in [76]. In the lemma below, we extend their result for a weakly convex g_+. The proof is in Section B.2 and the second conclusion is directly from [26].

Lemma 5.4. *Suppose Assumptions 3.1 and 5.1 hold. It holds for any $\mathbf{x}$ satisfying $\mathrm{dist}(\mathbf{x}, \mathcal{S}) \leq \frac{\nu}{\rho}$ that*

$$(\nu/2)\mathrm{dist}(\mathbf{x}, \mathcal{S}) \leq g_+(\mathbf{x}). \tag{12}$$

Moreover, $\nu \leq 2M$ and $\min_{\mathbf{x} \in \mathcal{X}} g_+(\mathbf{x})$ has no stationary point satisfying $0 < \mathrm{dist}(\mathbf{x}, \mathcal{S}) < \frac{\nu}{\rho}$.

Since g is non-convex, Algorithm 1 may not even find a nearly feasible solution if $\mathbf{x}^{(t)}$ is trapped at a stationary point of g with $g(\mathbf{x}) > 0$, that is, a sub-optimal stationary point of $\min_{\mathbf{x} \in \mathcal{X}} g_+(\mathbf{x})$. Fortunately, the second conclusion of Lemma 5.4 indicates that this situation can be avoided by keeping $\mathrm{dist}(\mathbf{x}^{(t)}, \mathcal{S}) = O(\epsilon^2)$ during the algorithm. To do so, we start with $\mathbf{x}^{(0)} \in \mathcal{S}$ and use $\epsilon_t = O(\epsilon^2)$ in Algorithm 1. Moreover, we apply a switching stepsize rule that sets $\eta_t = O(\epsilon^2)$ when $t \in I$ and $\eta_t = g(\mathbf{x}^{(t)})/\|\boldsymbol{\zeta}_g^{(t)}\|^2$ when $t \in J$, the latter of which is known by the Polyak's stepsize [33, 60]. This way, when $g(\mathbf{x}^{(t)}) \leq \epsilon_t$, (12) ensures $\mathrm{dist}(\mathbf{x}^{(t)}, \mathcal{S}) = O(\epsilon^2)$. When $g(\mathbf{x}^{(t)}) > \epsilon_t$, (12) ensures $\mathrm{dist}(\mathbf{x}^{(t)}, \mathcal{S})$ Q-linearly converges to zero [26], which also guarantees $\mathrm{dist}(\mathbf{x}^{(t)}, \mathcal{S}) = O(\epsilon^2)$. As a result, we have $g(\mathbf{x}^{(t)}) \leq \epsilon^2$ for any t, so problem (3) with $\mathbf{x} = \mathbf{x}^{(t)}$ and $(\hat{\rho}, \tilde{\rho})$ satisfying (9) will be strictly feasible according to Assumption 5.1B. This finding is given in the proposition below with its proof in Section B.3.

Proposition 5.5. *Suppose Assumptions 3.1 and 5.1 hold and $\epsilon \leq \bar{\epsilon}$. Also, suppose $\mathbf{x}^{(t)}$ is generated by Algorithm 1 using $\mathbf{x}^{(0)} \in \mathcal{S}$, $\epsilon_t = \frac{\nu}{4}\min\{\epsilon^2/M, \nu/(4\rho)\}$ and*

$$\eta_t = \begin{cases} \frac{\nu}{4M^2}\min\{\epsilon^2/M, \nu/(4\rho)\} & \text{if } t \in I \\ g(\mathbf{x}^{(t)})/\|\boldsymbol{\zeta}_g^{(t)}\|^2 & \text{if } t \in J. \end{cases}$$

6

Then $\text{dist}(\mathbf{x}^{(t)}, \mathcal{S}) \leq \min\{\epsilon^2/M, \nu/(4\rho)\}$ *and* $g(\mathbf{x}^{(t)}) \leq \epsilon^2$ *for any* $t \geq 0$. *As a consequence,* $\mathbf{x}^{(t)}$ *is* ϵ^2-*feasible to* (3) *where* $\mathbf{x} = \mathbf{x}^{(t)}$ *and* $(\hat{\rho}, \tilde{\rho})$ *satisfies* (9).

Let $\widehat{\mathbf{x}}^{(t)}$ be $\widehat{\mathbf{x}}(\mathbf{x}^{(t)})$ defined in (4) with $(\hat{\rho}, \tilde{\rho})$ satisfying (9). The complexity result for the case of weakly convex constraints is as follows. The proof can be found in Section B.3.

Theorem 5.6. *Under the same assumptions as Proposition 5.5, Algorithm 1 guarantees* $\mathbb{E}_\tau[\|\widehat{\mathbf{x}}^{(\tau)} - \mathbf{x}^{(\tau)}\|] \leq C\epsilon$ *and* $\mathbb{E}_\tau[g(\mathbf{x}^{(\tau)})] \leq \epsilon^2$, *where* $C := 2\sqrt{\frac{1+\Lambda'}{\hat{\rho}-\rho}}$, *if we use Output II and set*

$$S = 0 \text{ and } T \geq \frac{8M^2\left(f(\mathbf{x}^{(0)}) - \underline{f} + 3M^2/(2\hat{\rho})\right)}{\hat{\rho}(1+\Lambda')\nu\epsilon^2\min\{\epsilon^2/M, \nu/(4\rho)\}} = O(1/\epsilon^4).$$

This theorem indicates that Algorithm 1 finds a nearly $(C\epsilon)$-stationary point with complexity $O(1/\epsilon^4)$. To obtain a nearly ϵ-stationary point with $\epsilon \leq \bar{\epsilon}$, one only needs to replace ϵ in η_t, ϵ_t and T in this theorem above by $\epsilon/\max\{C, 1\}$. This will only change the constant factor in the $O(1/\epsilon^4)$ complexity. This complexity matches the start-of-the-art complexity by [13, 54, 39].

6 Numerical experiments

We demonstrate the performance of the SSG method on two fairness-aware classification problems, which are instances of (1) with convex and weakly convex g's, respectively. We compare with two different double-loop inexact proximal point (IPP) methods [13, 54, 39]. The IPP method approximately solves a strongly convex constrained subproblem in each outer iteration, and we use the SSG method in this paper and the ConEx method in [13] as the solvers (inner loop) because they both have the best theoretical complexity for that subproblem. We use IPP-SSG and IPP-ConEx to denote these two implementations of the IPP method. All experiments are conducted using MATLAB 2022b on a computer with the CPU 3.20GHz x Intel Core i7-8700 and 16GB memory.

6.1 Classification problem with ROC-based fairness

Given a feature vector $\mathbf{a} \in \mathbb{R}^d$ and a class label $b \in \{1, -1\}$, the goal in a binary linear classification problem is to learn a model $\mathbf{x} \in \mathbb{R}^d$ to predict b based on the score $\mathbf{x}^\top\mathbf{a}$. Let $\mathcal{D} = \{(\mathbf{a}_i, b_i)\}_{i=1}^n$ be a training set and $\ell(\cdot)$ be a convex non-increasing loss function. Model $\mathbf{x}$ can be learned by solving

$$L^* = \min_{\mathbf{x}\in\mathcal{W}}\left\{L(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^n \ell(b_i\mathbf{x}^\top\mathbf{a}_i)\right\}, \tag{13}$$

where $\mathcal{W} = \{\mathbf{x} \in \mathbb{R}^d \mid \|\mathbf{x}\| \leq r\}$. Solving (13) may ensure good classification performance of $\mathbf{x}$ but not its fairness. Suppose there exist two additional datasets. One contains the feature vectors of a protected group, denoted by $\mathcal{D}_p = \{\mathbf{a}_i^p\}_{i=1}^{n_p}$, and the other one contains the feature vectors of an unprotected group, denoted by $\mathcal{D}_u = \{\mathbf{a}_i^u\}_{i=1}^{n_u}$. We want to enhance the fairness of $\mathbf{x}$ between these two groups using the ROC-based fairness metric proposed by [74]. Suppose we set a threshold θ and classify data $\mathbf{a}$ as positive if $\mathbf{x}^\top\mathbf{a} \geq \theta$ and as negative otherwise. The ROC-based fairness measure and its continuous approximation are defined as

$$\max_{\theta\in\Theta}\left|\frac{1}{n_p}\sum_{i=1}^{n_p}\mathbb{I}(\mathbf{x}^\top\mathbf{a}_i^p \geq \theta) - \frac{1}{n_u}\sum_{i=1}^{n_u}\mathbb{I}(\mathbf{x}^\top\mathbf{a}_i^u \geq \theta)\right|$$

$$\approx R(\mathbf{x}) := \max_{\theta\in\Theta}\left|\frac{1}{n_p}\sum_{i=1}^{n_p}\sigma(\mathbf{x}^\top\mathbf{a}_i^p - \theta) - \frac{1}{n_u}\sum_{i=1}^{n_u}\sigma(\mathbf{x}^\top\mathbf{a}_i^u - \theta)\right|, \tag{14}$$

where $\sigma(z) = \exp(z)/(1+\exp(z))$ is the sigmoid function and Θ is a finite set of thresholds. If the value of this measure is small, model $\mathbf{x}$ produces similar predicted positive rates for the protected and unprotected groups on various θ's, indicating the fairness of the model. To obtain a fair $\mathbf{x}$, we balance (13) and (14) by solving

$$\min_{\mathbf{x}\in\mathcal{W}} R(\mathbf{x}) \text{ s.t. } L(\mathbf{x}) \leq L^* + \kappa, \tag{15}$$

7

Table 1: Information of the datasets. Groups are males VS females in a9a, users with age within $[25, 60]$ VS outside $[25, 60]$ in bank, and caucasian VS non-caucasian in COMPAS.

DATASETS	n	d	LABEL	GROUPS
A9A	48,842	123	INCOME	GENDER
BANK	41,188	54	SUBSCRIPTION	AGE
COMPAS	6,172	16	RECIDIVISM	RACE
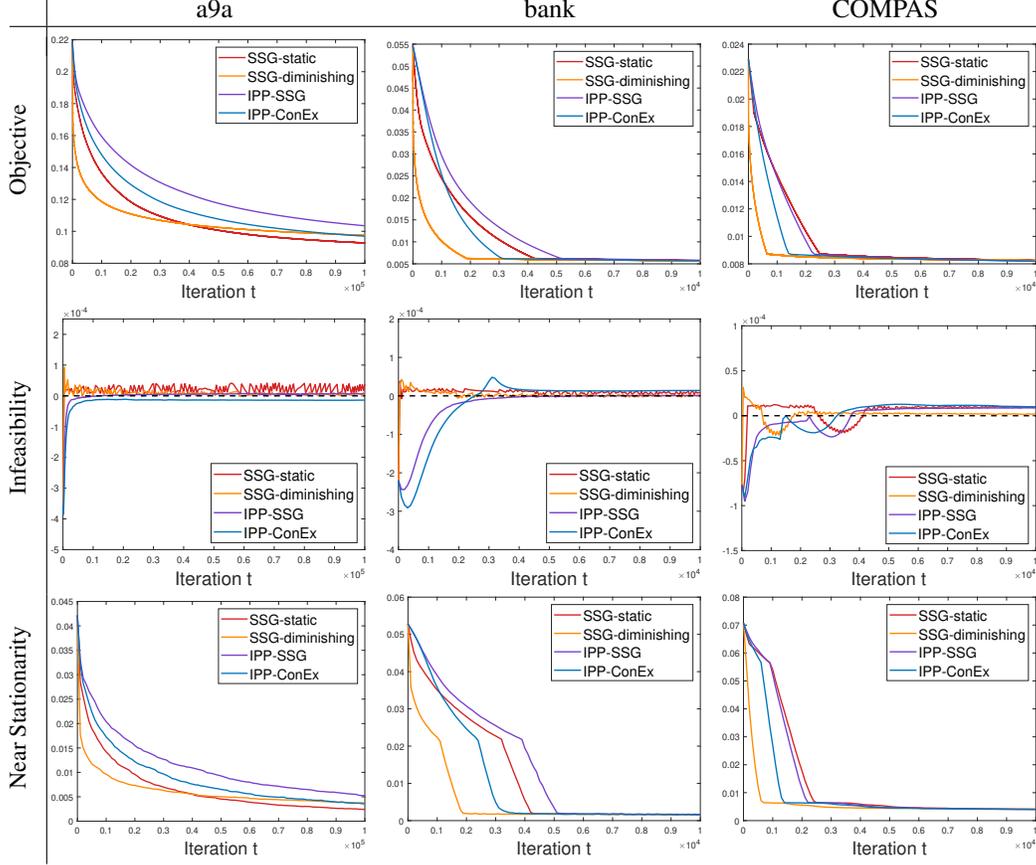
Figure 1: Performances vs number of iterations on classification problems with ROC-based fairness.

where κ is a slackness parameter indicating how much we are willing to increase the classification loss in order to reduce $R(\mathbf{x})$ to obtain a more fair model. Problem (15) is an instance of (1) satisfying Assumptions 3.1 and 4.1 with $\rho = \beta$ where β is defined in (77) in Section B.1.

We solve (15) on three datasets: *a9a* [41], *bank* [56] and *COMPAS* [4]. The information of these datasets is given in Table 1. We split each dataset into two subsets with a ratio of $2:1$. The larger set is used as $\mathcal{D}$ in the constraint and the smaller set is further split into $\mathcal{D}_p$ and $\mathcal{D}_u$ based on a binary group variable listed in Table 1.

In our experiments, we choose $\ell(z) = (1-z)_+$ and first solve (13) using the subgradient method with a large enough number of iterations to obtain L^* and a solution $\mathbf{x}_{\text{ERM}}$. Then we set $\kappa = 0.001 L^*$ and $r = 5\|\mathbf{x}_{\text{ERM}}\|$, and let Θ consist of 400 points equally spaced between $\min_i \mathbf{x}_{\text{ERM}}^\top \mathbf{a}_i - 0.5(\max_i \mathbf{x}_{\text{ERM}}^\top \mathbf{a}_i - \min_i \mathbf{x}_{\text{ERM}}^\top \mathbf{a}_i)$ and $\max_i \mathbf{x}_{\text{ERM}}^\top \mathbf{a}_i + 0.5(\max_i \mathbf{x}_{\text{ERM}}^\top \mathbf{a}_i - \min_i \mathbf{x}_{\text{ERM}}^\top \mathbf{a}_i)$.

All methods are initialized at $\mathbf{x}_{\text{ERM}}$. We implemented the SSG method with both static and diminishing stepsizes. For the static stepsize, we select ϵ_t from $\{10^{-6}, 2 \times 10^{-6}, 5 \times 10^{-6}, 10^{-5}\}$ and η_t from $\{2 \times 10^{-4}, 5 \times 10^{-4}, 10^{-3}, 2 \times 10^{-3}\}$. For the diminishing stepsize, we set $\epsilon_t = \frac{E_1}{\sqrt{t+1}}$ and $\eta_t = \frac{E_2}{\sqrt{t+1}}$ and select E_1 from $\{5 \times 10^{-5}, 10^{-4}, 2 \times 10^{-4}, 5 \times 10^{-4}\}$ and E_2 from $\{0.02, 0.05, 0.1, 0.2\}$. We

8

select the best set of parameters that produces the smallest objective value after 5000 iterations. For IPP, we select $\hat{\rho}$ from $\max\{\rho, 1\} \times \{1, 1.5, 2\}$ for all three datasets, and the proximal point subproblem is approximately solved by SSG and ConEx both with 100 iterations. For IPP-SSG, we apply a static stepsize with ϵ_t and η_t tuned in the same way as SSG. For IPP-ConEx, following the notation in [13], we set $\theta_t = \frac{t}{t+1}$, $\eta_t = c_1(t+1)$ and $\tau_t = \frac{c_2}{t+1}$, and select c_1 from $\{20, 50, 100, 200\}$ and c_2 from $\{0.002, 0.005, 0.01, 0.02\}$ by the same procedure adopted by SSG.

We report the performances versus number of iterations of all methods on each dataset in Figure 1. The performances versus the used CPU time will be given in Section C in the Appendix. For SSG, the x-axis represents the total number of iterations while, for IPP, it represents the total number of inner iterations across all outer iterations. The y-axis in each row represents the objective value, infeasibility and near stationarity achieved at each iteration, respectively. To measure near stationarity, we solve (3) with $\mathbf{x} = \mathbf{x}^{(t)}$ and parameters (7) using the SSG method with 2500 iterations and use the last iterate as an approximation of $\hat{\mathbf{x}}(\mathbf{x}^{(t)})$. We make sure that the change of $\|\hat{\mathbf{x}}(\mathbf{x}^{(t)}) - \mathbf{x}^{(t)}\|$ is less than 1% if the number of iterations is increased to 5000. Then we plot $\|\hat{\mathbf{x}}(\mathbf{x}^{(t)}) - \mathbf{x}^{(t)}\|$ as near stationarity in Figure 1. Since computing $\hat{\mathbf{x}}(\mathbf{x}^{(t)})$ with a high precision for each t is time-consuming, we only report near stationarity at 100 equally spaced iterations.

According to Figure 1, the SSG method with a diminishing stepsize has the best performance on all three datasets in the sense that it reduces the objective value and the (approximate) near stationarity measure faster than others while keeping the solutions nearly feasible. However, the SSG method with a static stepsize is not always better than the IPP methods. This is consistent with our theoretical finding that the SSG and IPP methods have the same oracle complexity.

6.2 Classification problem with demographic parity

Following the notation in the previous subsection, we consider a binary classification problem with a constraint enforcing demographic parity [1]. The measure of demographic parity and its continuous approximation are

$$\left| \frac{1}{n_p} \sum_{i=1}^{n_p} \mathbb{I}(\mathbf{x}^\top \mathbf{a}_i^p \geq 0) - \frac{1}{n_u} \sum_{i=1}^{n_u} \mathbb{I}(\mathbf{x}^\top \mathbf{a}_i^u \geq 0) \right| \approx R_0(\mathbf{x}) := \left| \frac{1}{n_p} \sum_{i=1}^{n_p} \sigma(\mathbf{x}^\top \mathbf{a}_i^p) - \frac{1}{n_u} \sum_{i=1}^{n_u} \sigma(\mathbf{x}^\top \mathbf{a}_i^u) \right|. \quad (16)$$

Fairness measure $R_0(\mathbf{x})$ is a special case of $R(\mathbf{x})$ with $\Theta = \{0\}$. If $R_0(\mathbf{x})$ is small, model $\mathbf{x}$ produces similar predicted positive rates for the protected and unprotected groups. To obtain a fair $\mathbf{x}$, we balance (13) and (16) by solving

$$\min L(\mathbf{x}) + \lambda \text{SCAD}(\mathbf{x}) \text{ s.t. } R_0(\mathbf{x}) \leq \kappa. \quad (17)$$

Different from (15), the fairness measure is used as the constraint in (17) while the objective function becomes the empirical hinge loss plus the smoothly clipped absolute deviation (SCAD) regularizer [35] for promoting a sparse solution. Here, λ is a regularization parameter and

$$\text{SCAD}(\mathbf{x}) := \sum_{i=1}^{d} s(x_i), \quad s(x_i) = \begin{cases} 2|x_i| & 0 \leq |x_i| \leq 1 \\ -x_i^2 + 4|x_i| + 1 & 1 < |x_i| \leq 2 \\ 3 & 2 < |x_i|. \end{cases} \quad (18)$$

We prove in Section B.1 that (17) is an instance of (1) satisfying Assumptions 3.1 and 5.1 with $\rho = \max\{2\lambda, \beta\}$ where β is defined in (77) in Section B.1.

We set $\lambda = 0.2$ for all datasets and set $\kappa = 0.005$, 0.02 and 0.02 for *a9a*, *bank* and *COMPAS*, respectively. For SSG, we select ϵ_t from $\{10^{-6}, 2 \times 10^{-6}, 5 \times 10^{-6}, 10^{-5}\}$ and select η_t from $\{10^{-4}, 2 \times 10^{-4}, 5 \times 10^{-4}, 7.5 \times 10^{-4}\}$ for $t \in I$ while set $\eta_t = g(\mathbf{x}^{(t)})/\|\zeta_g^{(t)}\|^2$ for $t \in J$. We select the best set of parameters that produces the smallest objective value after 50000 iterations. For IPP, we select $\hat{\rho}$ from $\max\{\rho, 1\} \times \{1, 1.5, 2\}$ and the proximal point subproblem is approximately solved by SSG and ConEx both with 600 iterations. For IPP-SSG, we apply a static stepsize with ϵ_t and η_t tuned in the same way as SSG. For IPP-ConEx, following the notation in [13], we set $\theta_t = \frac{t}{t+1}$, $\eta_t = c_1(t+1)$ and $\tau_t = \frac{c_2}{t+1}$ and select c_1 from $\{20, 50, 100, 200\}$ and c_2 from $\{0.002, 0.005, 0.01, 0.02\}$ by the same procedure adopted by SSG. Due to the limit of space, we present the performances vs number of iterations and CPU runtime of all methods in Figure 3 and Figure 4 respectively in Section C in the Appendix.

9

7 Conclusion

We study the oracle complexity of the switching subgradient (SSG) method for finding a nearly ϵ-stationary point of a non-smooth weakly convex constrained optimization problem. We show that the complexity of the SSG method matches the best result in literature that is achieved only by double-loop methods. On the contrary, the SSG method is single-loop and easier to implement with reduced tuning effort. This is the first complexity result for a single-loop first-order method for a weakly-convex non-smooth constrained problem.

Acknowledgements

This work was supported by NSF award 2147253.

References

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning*, pages 60–69. PMLR, 2018.

[2] Ahmet Alacaoglu, Yura Malitsky, and Volkan Cevher. Convergence of adaptive algorithms for constrained weakly convex optimization. *Advances in Neural Information Processing Systems*, 34:14214–14225, 2021.

[3] Mohammad S. Alkousa. On modification of an adaptive stochastic mirror descent algorithm for convex optimization problems with functional constraints. In *Computational Mathematics and Applications*, pages 47–63. Springer, 2020.

[4] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *ProPublica, May, 23*, 2016.

[5] Yossi Arjevani, Yair Carmon, John C. Duchi, Dylan J. Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199(1-2):165–214, 2023.

[6] Alfred Auslender. An extended sequential quadratically constrained quadratic programming algorithm for nonlinear, semidefinite, and second-order cone programming. *Journal of Optimization Theory and Applications*, 156(2):183–212, 2013.

[7] Anastasia Bayandina, Pavel Dvurechensky, Alexander Gasnikov, Fedor Stonyakin, and Alexander Titov. Mirror descent and convex optimization problems with non-smooth inequality constraints. In *Large-Scale and Distributed Optimization*, pages 181–213. Springer, 2018.

[8] Albert S. Berahas, Frank E. Curtis, Michael J. O'Neill, and Daniel P. Robinson. A stochastic sequential quadratic optimization algorithm for nonlinear equality constrained optimization with rank-deficient jacobians. *arXiv preprint arXiv:2106.13015*, 2021.

[9] Albert S. Berahas, Frank E. Curtis, Daniel Robinson, and Baoyu Zhou. Sequential quadratic optimization for nonlinear equality constrained stochastic optimization. *SIAM Journal on Optimization*, 31(2):1352–1379, 2021.

[10] Albert S. Berahas, Miaolan Xie, and Baoyu Zhou. A sequential quadratic programming method with high probability complexity bounds for nonlinear equality constrained stochastic optimization. *arXiv preprint arXiv:2301.00477*, 2023.

[11] Jérôme Bolte and Edouard Pauwels. Majorization-minimization procedures and convergence of sqp methods for semi-algebraic and tame programs. *Mathematics of Operations Research*, 41(2):442–465, 2016.

[12] Digvijay Boob, Qi Deng, and Guanghui Lan. Level constrained first order methods for function constrained optimization. *arXiv preprint arXiv:2205.08011*, 2022.

10

[13] Digvijay Boob, Qi Deng, and Guanghui Lan. Stochastic first-order methods for convex and nonconvex functional constrained optimization. *Mathematical Programming*, 197(1):215–279, 2023.

[14] Digvijay Boob, Qi Deng, Guanghui Lan, and Yilin Wang. A feasible level proximal point method for nonconvex sparse constrained optimization. *Advances in Neural Information Processing Systems*, 33:16773–16784, 2020.

[15] Morteza Boroun and Afrooz Jalilzadeh. Inexact-proximal accelerated gradient method for stochastic nonconvex constrained optimization problems. In *2021 Winter Simulation Conference (WSC)*, pages 1–12. IEEE, 2021.

[16] James V. Burke. A sequential quadratic programming method for potentially infeasible mathematical programs. *Journal of Mathematical Analysis and Applications*, 139(2):319–351, 1989.

[17] Coralia Cartis, Nicholas I.M. Gould, and Philippe L. Toint. On the evaluation complexity of composite function minimization with applications to nonconvex nonlinear programming. *SIAM Journal on Optimization*, 21(4):1721–1739, 2011.

[18] Coralia Cartis, Nicholas I.M. Gould, and Philippe L. Toint. Optimality of orders one to three and beyond: characterization and evaluation complexity in constrained nonconvex optimization. *Journal of Complexity*, 53:68–94, 2019.

[19] Lesi Chen, Jing Xu, and Luo Luo. Faster gradient-free algorithms for nonsmooth nonconvex stochastic optimization. *arXiv preprint arXiv:2301.06428*, 2023.

[20] Yi Cheng, Guanghui Lan, and H. Edwin Romeijn. Functional constrained optimization for risk aversion and sparsity control. *arXiv preprint arXiv:2210.05108*, 2022.

[21] Frank E. Curtis, Michael J. O'Neill, and Daniel P. Robinson. Worst-case complexity of an sqp method for nonlinear equality constrained stochastic optimization. *arXiv preprint arXiv:2112.14799*, 2021.

[22] Frank E. Curtis, Daniel P. Robinson, and Baoyu Zhou. Inexact sequential quadratic optimization for minimizing a stochastic objective function subject to deterministic nonlinear equality constraints. *arXiv preprint arXiv:2107.03512*, 2021.

[23] Damek Davis and Dmitriy Drusvyatskiy. Complexity of finding near-stationary points of convex functions stochastically. *arXiv preprint arXiv:1802.08556*, 2018.

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

[25] Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

[26] Damek Davis, Dmitriy Drusvyatskiy, Kellie J. MacPhee, and Courtney Paquette. Subgradient methods for sharp weakly convex functions. *Journal of Optimization Theory and Applications*, 179(3):962–982, 2018.

[27] Damek Davis and Benjamin Grimmer. Proximally guided stochastic subgradient method for nonsmooth, nonconvex problems. *SIAM Journal on Optimization*, 29(3):1908–1930, 2019.

[28] Qi Deng and Wenzhi Gao. Minibatch and momentum model-based methods for stochastic weakly convex optimization. *Advances in Neural Information Processing Systems*, 34:23115–23127, 2021.

[29] Gianni Di Pillo and Luigi Grippo. On the exactness of a class of nondifferentiable penalty functions. *Journal of Optimization Theory and Applications*, 57(3):399–410, 1988.

[30] Gianni Di Pillo and Luigi Grippo. Exact penalty functions in constrained optimization. *SIAM Journal on Control and Optimization*, 27(6):1333–1360, 1989.

11

[31] Yoel Drori and Ohad Shamir. The complexity of finding stationary points with stochastic gradient descent. In *International Conference on Machine Learning*, pages 2658–2667. PMLR, 2020.

[32] Dmitriy Drusvyatskiy and Courtney Paquette. Efficiency of minimizing compositions of convex functions and smooth maps. *Mathematical Programming*, 178(1–2):503–558, 2019.

[33] Ivan Ivanovich Eremin. The relaxation method of solving systems of inequalities with convex functions on the left-hand side. In *Doklady Akademii Nauk*, volume 160, pages 994–996. Russian Academy of Sciences, 1965.

[34] Francisco Facchinei, Vyacheslav Kungurtsev, Lorenzo Lampariello, and Gesualdo Scutari. Ghost penalties in nonconvex constrained optimization: Diminishing stepsizes and iteration complexity. *Mathematics of Operations Research*, 46(2):595–627, 2021.

[35] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

[36] Roger Fletcher. An ℓ_1 penalty method for nonlinear constraints. In *Numerical optimization 1984*, pages 26–40. SIAM Publications Philadelphia, 1985.

[37] Wenbo Gao, Donald Goldfarb, and Frank E. Curtis. Admm for multiaffine constrained optimization. *Optimization Methods and Software*, 35(2):257–303, 2020.

[38] Mingyi Hong, Zhi-Quan Luo, and Meisam Razaviyayn. Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems. *SIAM Journal on Optimization*, 26(1):337–364, 2016.

[39] Zhichao Jia and Benjamin Grimmer. First-order methods for nonsmooth nonconvex functional constrained optimization with or without slater points. *arXiv preprint arXiv:2212.00927*, 2022.

[40] Lingzi Jin and Xiao Wang. A stochastic primal-dual method for a class of nonconvex constrained optimization. *Computational Optimization and Applications*, 83(1):143–180, 2022.

[41] Ron Kohavi et al. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *KDD*, volume 96, pages 202–207, 1996.

[42] Weiwei Kong, Jefferson G. Melo, and Renato D.C. Monteiro. Complexity of a quadratic penalty accelerated inexact proximal point method for solving linearly constrained nonconvex composite programs. *SIAM Journal on Optimization*, 29(4):2566–2593, 2019.

[43] Weiwei Kong, Jefferson G. Melo, and Renato D.C. Monteiro. Iteration complexity of a proximal augmented lagrangian method for solving nonconvex composite optimization problems with nonlinear convex constraints. *Mathematics of Operations Research*, 48(2):1066–1094, 2023.

[44] Weiwei Kong, Jefferson G. Melo, and Renato D.C. Monteiro. Iteration complexity of an inner accelerated inexact proximal augmented lagrangian method based on the classical lagrangian function. *SIAM Journal on Optimization*, 33(1):181–210, 2023.

[45] Weiwei Kong and Renato D.C. Monteiro. An accelerated inexact dampened augmented lagrangian method for linearly-constrained nonconvex composite optimization problems. *Computational Optimization and Applications*, 85(2):509–545, 2023.

[46] Guy Kornowski and Ohad Shamir. Oracle complexity in nonsmooth nonconvex optimization. *Advances in Neural Information Processing Systems*, 34:324–334, 2021.

[47] Guy Kornowski and Ohad Shamir. On the complexity of finding small subgradients in nonsmooth optimization. *arXiv preprint arXiv:2209.10346*, 2022.

[48] Guanghui Lan, Arkadi Nemirovski, and Alexander Shapiro. Validation analysis of mirror descent stochastic approximation method. *Mathematical Programming*, 134(2):425–458, 2012.

[49] Guanghui Lan and Zhiqiang Zhou. Algorithms for stochastic optimization with function or expectation constraints. *Computational Optimization and Applications*, 76(2):461–498, 2020.

12

[50] Zichong Li, Pin-Yu Chen, Sijia Liu, Songtao Lu, and Yangyang Xu. Rate-improved inexact augmented lagrangian method for constrained nonconvex optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 2170–2178. PMLR, 2021.

[51] Zichong Li, Pin-Yu Chen, Sijia Liu, Songtao Lu, and Yangyang Xu. Zeroth-order optimization for composite problems with functional constraints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7453–7461, 2022.

[52] Zichong Li and Yangyang Xu. Augmented lagrangian–based first-order methods for convex-constrained programs with weakly convex objective. *INFORMS Journal on Optimization*, 3(4):373–397, 2021.

[53] Qihang Lin, Runchao Ma, and Yangyang Xu. Complexity of an inexact proximal-point penalty method for constrained smooth non-convex optimization. *Computational Optimization and Applications*, 82(1):175–224, 2022.

[54] Runchao Ma, Qihang Lin, and Tianbao Yang. Quadratically regularized subgradient methods for weakly convex optimization with weakly convex constraints. In *International Conference on Machine Learning*, pages 6554–6564. PMLR, 2020.

[55] Jefferson G. Melo, Renato D.C. Monteiro, and Hairong Wang. Iteration-complexity of an inexact proximal accelerated augmented lagrangian method for solving linearly constrained smooth nonconvex composite optimization problems. *arXiv preprint arXiv:2006.08048*, 2020.

[56] Sérgio Moro, Paulo Cortez, and Paulo Rita. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62:22–31, 2014.

[57] Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

[58] Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

[59] Boris T. Polyak. A general method of solving extremum problems. *Soviet Mathematics Doklady*, 8(3):593–597, 1967.

[60] Boris T. Polyak. Minimization of unsmooth functionals. *USSR Computational Mathematics and Mathematical Physics*, 9(3):14–29, 1969.

[61] Hassan Rafique, Mingrui Liu, Qihang Lin, and Tianbao Yang. Weakly-convex–concave min–max optimization: provable algorithms and applications in machine learning. *Optimization Methods and Software*, 37(3):1087–1121, 2022.

[62] Philippe Rigollet and Xin Tong. Neyman-pearson classification, convexity and stochastic constraints. *Journal of Machine Learning Research*, 12:2831–2855, 2011.

[63] Mehmet Fatih Sahin, Ahmet Alacaoglu, Fabian Latorre, Volkan Cevher, et al. An inexact augmented lagrangian framework for nonconvex optimization with nonlinear constraints. *Advances in Neural Information Processing Systems*, 32:13966–13978, 2019.

[64] Ohad Shamir. Can we find near-approximately-stationary points of nonsmooth nonconvex functions? *arXiv preprint arXiv:2002.11962*, 2020.

[65] Fedor S. Stonyakin, Mohammad S. Alkousa, Alexey N. Stepanov, and Aleksandr A. Titov. Adaptive mirror descent algorithms for convex and strongly convex optimization problems with functional constraints. *Journal of Applied and Industrial Mathematics*, 13(3):557–574, 2019.

[66] Fedor S. Stonyakin, Mohammad S. Alkousa, Alexander A. Titov, and Victoria V. Piskunova. On some methods for strongly convex optimization problems with one functional constraint. In *International Conference on Mathematical Optimization Theory and Operations Research*, pages 82–96. Springer, 2019.

[67] Fedor S. Stonyakin, Alexey N. Stepanov, Alexander A. Titov, and Alexander V. Gasnikov. Mirror descent for constrained optimization problems with large subgradient values. *arXiv preprint arXiv:1908.00218*, 2019.

13

[68] Fedor S. Stonyakin and Alexander A. Titov. One mirror descent algorithm for convex constrained optimization problems with non-standard growth properties. *arXiv preprint arXiv:1803.01329*, 2018.

[69] Arnesh Sujanani and Renato D.C. Monteiro. An adaptive superfast inexact proximal augmented lagrangian method for smooth nonconvex composite optimization problems. *Journal of Scientific Computing*, 97(2):34, 2023.

[70] Lai Tian and Anthony Man-Cho So. Computing goldstein (ϵ, δ)-stationary points of lipschitz functions in $\widetilde{O}(\epsilon^{-3}\delta^{-1})$ iterations via random conic perturbation. *arXiv preprint arXiv:2112.09002*, 2021.

[71] Lai Tian, Kaiwen Zhou, and Anthony Man-Cho So. On the finite-time complexity and practical computation of approximate stationarity concepts of lipschitz functions. In *International Conference on Machine Learning*, pages 21360–21379. PMLR, 2022.

[72] Alexander A. Titov, Fedor S. Stonyakin, Mohammad S. Alkousa, Seydamet S. Ablaev, and Alexander V. Gasnikov. Analogues of switching subgradient schemes for relatively lipschitz-continuous convex programming problems. In *International Conference on Mathematical Optimization Theory and Operations Research*, pages 133–149. Springer, 2020.

[73] Alexander A. Titov, Fedor S. Stonyakin, Alexander V. Gasnikov, and Mohammad S. Alkousa. Mirror descent and constrained online optimization problems. In *International Conference on Optimization and Applications*, pages 64–78. Springer, 2019.

[74] Robin Vogel, Aurélien Bellet, and Stéphan Clémençon. Learning fair scoring functions: Bipartite ranking under roc-based fairness constraints. In *International Conference on Artificial Intelligence and Statistics*, pages 784–792. PMLR, 2021.

[75] Yu Wang, Wotao Yin, and Jinshan Zeng. Global convergence of admm in nonconvex nonsmooth optimization. *Journal of Scientific Computing*, 78(1):29–63, 2019.

[76] Tianbao Yang and Qihang Lin. RSG: Beating subgradient method without smoothness and strong convexity. *Journal of Machine Learning Research*, 19(1):236–268, 2018.

[77] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. Fairness constraints: A flexible approach for fair classification. *Journal of Machine Learning Research*, 20(1):2737–2778, 2019.

[78] Jinshan Zeng, Wotao Yin, and Ding-Xuan Zhou. Moreau envelope augmented lagrangian method for nonconvex optimization with linear constraints. *Journal of Scientific Computing*, 91(2):61, 2022.

[79] Jiawei Zhang and Zhi-Quan Luo. A proximal alternating direction method of multiplier for linearly constrained nonconvex minimization. *SIAM Journal on Optimization*, 30(3):2272–2302, 2020.

[80] Jiawei Zhang and Zhi-Quan Luo. A global dual error bound and its application to the analysis of linearly constrained nonconvex optimization. *SIAM Journal on Optimization*, 32(3):2319–2346, 2022.

[81] Jiawei Zhang, Wenqiang Pu, and Zhi-Quan Luo. On the iteration complexity of smoothed proximal alm for nonconvex optimization problem with convex constraints. *arXiv preprint arXiv:2207.06304*, 2022.

[82] Jingzhao Zhang, Hongzhou Lin, Stefanie Jegelka, Suvrit Sra, and Ali Jadbabaie. Complexity of finding stationary points of nonconvex nonsmooth functions. In *International Conference on Machine Learning*, pages 11173–11182. PMLR, 2020.

[83] Siqi Zhang and Niao He. On the convergence rate of stochastic mirror descent for nonsmooth nonconvex optimization. *arXiv preprint arXiv:1806.04781*, 2018.

14

A   Complexity analysis for convex constraints

In this section, we present the theoretical complexity analysis for the SSG method when f is non-convex but g is convex. Our analysis will first focus on a more general setting where the oracles of the subgradients and function values of f and g are stochastic. Then the complexity for the case of deterministic oracles is derived as a special case.

A.1   Proof of Lemma 4.2 and extension with linear equality constraints

Proof of Lemma 4.2. For simplicity of notation, we denote $\widehat{\mathbf{x}}(\mathbf{x})$ in (4) by $\widehat{\mathbf{x}}$. By Assumption 4.1B, there exists a strictly feasible solution $\mathbf{x}_{\text{feas}} \in \text{relint}(\mathcal{X})$ with $g(\mathbf{x}_{\text{feas}}) < 0$. As a result, the Lagrangian multiplier $\widehat{\lambda} \geq 0$ corresponding to $\widehat{\mathbf{x}}$ is well-defined and satisfies (5), which means $\widehat{\lambda}g(\widehat{\mathbf{x}}) = 0$ and

$$\widehat{\boldsymbol{\zeta}}_f + \hat{\rho}(\widehat{\mathbf{x}} - \mathbf{x}) + \widehat{\lambda}\widehat{\boldsymbol{\zeta}}_g + \widehat{\mathbf{u}} = \mathbf{0}, \tag{19}$$

where $\widehat{\boldsymbol{\zeta}}_f \in \partial f(\widehat{\mathbf{x}})$, $\widehat{\boldsymbol{\zeta}}_g \in \partial g(\widehat{\mathbf{x}})$, $\widehat{\mathbf{u}} \in \mathcal{N}_{\mathcal{X}}(\widehat{\mathbf{x}})$ and $\mathcal{N}_{\mathcal{X}}(\widehat{\mathbf{x}})$ is the normal cone of $\mathcal{X}$ at $\widehat{\mathbf{x}}$.

If $\widehat{\lambda} = 0$, the conclusion holds trivially. Hence, we focus on the case that $\widehat{\lambda} > 0$. Note that, in this case, we must have $g(\widehat{\mathbf{x}}) = 0$. Taking the inner product between (19) and $\widehat{\mathbf{x}} - \mathbf{x}_{\text{feas}}$ gives

$$\frac{1}{\widehat{\lambda}}\langle\widehat{\boldsymbol{\zeta}}_f + \hat{\rho}(\widehat{\mathbf{x}} - \mathbf{x}), \widehat{\mathbf{x}} - \mathbf{x}_{\text{feas}}\rangle = \langle\widehat{\boldsymbol{\zeta}}_g + \widehat{\mathbf{u}}/\widehat{\lambda}, \mathbf{x}_{\text{feas}} - \widehat{\mathbf{x}}\rangle \leq g(\mathbf{x}_{\text{feas}}) - g(\widehat{\mathbf{x}}) = g(\mathbf{x}_{\text{feas}}), \tag{20}$$

where the inequality is because of convexity of g and the fact that $\widehat{\mathbf{u}}/\widehat{\lambda} \in \mathcal{N}_{\mathcal{X}}(\widehat{\mathbf{x}})$. By Assumption 3.1 and 4.1C, we have $\|\widehat{\boldsymbol{\zeta}}_f\| \leq M$, $\|\widehat{\mathbf{x}} - \mathbf{x}_{\text{feas}}\| \leq D$ and $\|\widehat{\mathbf{x}} - \mathbf{x}\| \leq D$, which imply from (20) that

$$\widehat{\lambda} \leq \frac{\langle\widehat{\boldsymbol{\zeta}}_f + \hat{\rho}(\widehat{\mathbf{x}} - \mathbf{x}), \mathbf{x}_{\text{feas}} - \widehat{\mathbf{x}}\rangle}{-g(\mathbf{x}_{\text{feas}})} \leq \frac{MD + \hat{\rho}D^2}{-g(\mathbf{x}_{\text{feas}})}.$$

$\square$

The extension of Lemma 4.2 when there exist linear constraints: Suppose there are linear equality constraints. We formulate problem (1) as

$$f^* \equiv \min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}) \quad \text{s.t.} \quad h(\mathbf{x}) \leq 0, \quad \mathbf{A}\mathbf{x} = \mathbf{b}, \tag{21}$$

where $\mathbf{A} \in \mathbb{R}^{l \times d}$ and $\mathbf{b} \in \mathbb{R}^{l}$. Moreover, we assume $\mathbf{A}$ has a full row rank. We then define

$$g(\mathbf{x}) = \max\{h(\mathbf{x}), \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_{\infty}\}. \tag{22}$$

With this g, Assumption 4.1 B cannot be satisfied. However, it is easy to show that $\widehat{\lambda}$ in (5) can be still bounded from above if the inequality constraint in (21) can be satisfied strictly.

To extend Lemma 4.2 for (21), we only need to change Assumption 4.1 B to the following

   B'. (Slater's condition) There exists $\mathbf{x}_{\text{feas}} \in \text{int}(\mathcal{X})$ such that $h(\mathbf{x}_{\text{feas}}) < 0$ and $\mathbf{A}\mathbf{x}_{\text{feas}} = \mathbf{b}$.

Then we can still derive an upper bound of $\widehat{\lambda}$ similar to the one in Lemma 4.2.

Lemma A.1. *Let g defined as in* (22). *Suppose Assumptions 3.1 and 4.1 hold except that Assumption 4.1 B is replaced by B' above. Given any $\mathbf{x} \in \mathcal{X}$, let $\widehat{\mathbf{x}}(\mathbf{x})$ be defined as in* (4) *with $(\hat{\rho}, \tilde{\rho})$ satisfying* (7) *and $\widehat{\lambda}$ be the associated Lagrangian multiplier satisfying* (5). *We have*

$$\widehat{\lambda} \leq \Lambda_{ec} := \frac{MD + \hat{\rho}D^2}{-h(\mathbf{x}_{\text{feas}})} + \sqrt{l}\|(\mathbf{A}\mathbf{A}^{\top})^{-1}\mathbf{A}\|\left(\frac{M^2D + \hat{\rho}MD^2}{-h(\mathbf{x}_{\text{feas}})} + \frac{MD + \hat{\rho}D^2}{\text{dist}(\mathbf{x}_{\text{feas}}, \partial\mathcal{X})} + M + \hat{\rho}D\right). \tag{23}$$

Proof. For simplicity of notation, we denote $\widehat{\mathbf{x}}(\mathbf{x})$ in (4) by $\widehat{\mathbf{x}}$. We first show that (5) holds for some Lagrangian multiplier $\widehat{\lambda}$ and $\widehat{\boldsymbol{\zeta}}_g \in \partial g(\widehat{\mathbf{x}})$. By Assumption B', there exist Lagrangian multipliers $\widehat{\lambda}_{in} \in \mathbb{R}_+$ and $\widehat{\boldsymbol{\lambda}}_{eq} \in \mathbb{R}^{l}$ such that

$$\widehat{\boldsymbol{\zeta}}_f + \hat{\rho}(\widehat{\mathbf{x}} - \mathbf{x}) + \widehat{\lambda}_{in}\widehat{\boldsymbol{\zeta}}_h + \mathbf{A}^{\top}\widehat{\boldsymbol{\lambda}}_{eq} + \widehat{\mathbf{u}} = \mathbf{0}, \tag{24}$$

15

where $\widehat{\boldsymbol{\zeta}}_h \in \partial h(\widehat{\mathbf{x}})$, and other notations are as in (19). It also holds that $g(\widehat{\mathbf{x}}) = 0$ and $\widehat{\lambda}_{in} h(\widehat{\mathbf{x}}) = 0$.
Since $h(\widehat{\mathbf{x}}) \leq 0$ and $\mathbf{A}\widehat{\mathbf{x}} = \mathbf{b}$, we have $\mathbf{0} \in \partial g(\widehat{\mathbf{x}})$ by the definition of g. Suppose $\widehat{\lambda}_{in} + \|\widehat{\boldsymbol{\lambda}}_{eq}\|_1 = 0$.
(24) implies (19) with $\widehat{\lambda} = 0$ and any $\widehat{\boldsymbol{\zeta}}_g \in \partial g(\widehat{\mathbf{x}})$. Suppose $\widehat{\lambda}_{in} + \|\widehat{\boldsymbol{\lambda}}_{eq}\|_1 > 0$. Let

$$\widehat{\boldsymbol{\zeta}}_g = \frac{\widehat{\lambda}_{in}\widehat{\boldsymbol{\zeta}}_h + \mathbf{A}^\top \widehat{\boldsymbol{\lambda}}_{eq}}{\widehat{\lambda}_{in} + \|\widehat{\boldsymbol{\lambda}}_{eq}\|_1} \quad \text{and} \quad \widehat{\lambda} = \widehat{\lambda}_{in} + \|\widehat{\boldsymbol{\lambda}}_{eq}\|_1.$$

It is easy to verify by definition that $\widehat{\boldsymbol{\zeta}}_g \in \partial g(\widehat{\mathbf{x}})$ and (19) also holds.

Next we derive the claimed upper bound of $\widehat{\lambda}$. When $\widehat{\lambda}_{in} + \|\widehat{\boldsymbol{\lambda}}_{eq}\|_1 = 0$ and thus $\widehat{\lambda} = 0$, the conclusion holds trivially so we focus on the case where $\widehat{\lambda} = \widehat{\lambda}_{in} + \|\widehat{\boldsymbol{\lambda}}_{eq}\|_1 > 0$ and $\widehat{\boldsymbol{\zeta}}_g$ is defined as above. In this case, we follow the same analysis in the proof of Lemma 3 in [53] to bound $\widehat{\lambda}_{in}$ and $\|\widehat{\boldsymbol{\lambda}}_{eq}\|_1$.

By the convexity of h and the fact that $\widehat{\lambda}_{in} \geq 0$, we can show that

$$\begin{aligned}
\widehat{\lambda}_{in} h(\mathbf{x}_{\text{feas}}) &\geq \widehat{\lambda}_{in} \left(h(\widehat{\mathbf{x}}) + \langle \widehat{\boldsymbol{\zeta}}_h, \mathbf{x}_{\text{feas}} - \widehat{\mathbf{x}} \rangle \right) \\
&= -\langle \widehat{\boldsymbol{\zeta}}_f + \widehat{\rho}(\widehat{\mathbf{x}} - \mathbf{x}) + \mathbf{A}^\top \widehat{\boldsymbol{\lambda}}_{eq} + \widehat{\mathbf{u}}, \mathbf{x}_{\text{feas}} - \widehat{\mathbf{x}} \rangle \\
&= -\langle \widehat{\boldsymbol{\zeta}}_f + \widehat{\rho}(\widehat{\mathbf{x}} - \mathbf{x}) + \widehat{\mathbf{u}}, \mathbf{x}_{\text{feas}} - \widehat{\mathbf{x}} \rangle, \quad (25)
\end{aligned}$$

where the first equality is due to (24) and the second equality is because $\mathbf{A}\widehat{\mathbf{x}} = \mathbf{A}\mathbf{x}_{\text{feas}} = \mathbf{b}$.

Next we bound the term $-\langle \widehat{\mathbf{u}}, \mathbf{x}_{\text{feas}} - \widehat{\mathbf{x}} \rangle$ in (25) from below. Suppose $\widehat{\mathbf{u}} \neq \mathbf{0}$. It must hold that $\widehat{\mathbf{x}} \in \partial \mathcal{X}$ since, otherwise, $\widehat{\mathbf{x}} \in \text{int}(\mathcal{X})$ so $\widehat{\mathbf{u}} = \mathbf{0}$. Let $\mathcal{H} = \{\mathbf{x} \in \mathbb{R}^d \,|\, \langle \widehat{\mathbf{u}}, \mathbf{x} - \widehat{\mathbf{x}} \rangle = 0\}$, which is a supporting hyperplane of $\mathcal{X}$ at $\widehat{\mathbf{x}}$. Since $\text{dist}(\mathbf{x}_{\text{feas}}, \mathcal{H}) = \langle \widehat{\mathbf{u}}, \widehat{\mathbf{x}} - \mathbf{x}_{\text{feas}} \rangle / \|\widehat{\mathbf{u}}\|$, we have

$$-\langle \widehat{\mathbf{u}}, \mathbf{x}_{\text{feas}} - \widehat{\mathbf{x}} \rangle = \text{dist}(\mathbf{x}_{\text{feas}}, \mathcal{H})\|\widehat{\mathbf{u}}\| \geq \text{dist}(\mathbf{x}_{\text{feas}}, \partial \mathcal{X})\|\widehat{\mathbf{u}}\|.$$

Applying the inequality to (25), we have

$$-\widehat{\lambda}_{in} h(\mathbf{x}_{\text{feas}}) + \text{dist}(\mathbf{x}_{\text{feas}}, \partial \mathcal{X})\|\widehat{\mathbf{u}}\| \leq \langle \widehat{\boldsymbol{\zeta}}_f + \widehat{\rho}(\widehat{\mathbf{x}} - \mathbf{x}), \mathbf{x}_{\text{feas}} - \widehat{\mathbf{x}} \rangle \leq MD + \widehat{\rho}D^2,$$

where the last inequality is by Assumption 3.1 and 4.1C, which imply $\|\widehat{\boldsymbol{\zeta}}_f\| \leq M$, $\|\widehat{\mathbf{x}} - \mathbf{x}_{\text{feas}}\| \leq D$ and $\|\widehat{\mathbf{x}} - \mathbf{x}\| \leq D$. The inequality above then implies

$$\widehat{\lambda}_{in} \leq \frac{MD + \widehat{\rho}D^2}{-h(\mathbf{x}_{\text{feas}})} \quad \text{and} \quad \|\widehat{\mathbf{u}}\| \leq \frac{MD + \widehat{\rho}D^2}{\text{dist}(\mathbf{x}_{\text{feas}}, \partial \mathcal{X})}. \quad (26)$$

Suppose $\widehat{\mathbf{u}} = \mathbf{0}$. We can also prove the first inequality in (26) by a similar argument and the second inequality in (26) holds trivially. Multiplying both sides of (24) by $\mathbf{A}$ and using the fact that $\mathbf{A}$ has a full row rank, we have

$$\widehat{\boldsymbol{\lambda}}_{eq} = -(\mathbf{A}\mathbf{A}^\top)^{-1}\mathbf{A}\left(\widehat{\boldsymbol{\zeta}}_f + \widehat{\rho}(\widehat{\mathbf{x}} - \mathbf{x}) + \widehat{\lambda}_{in}\widehat{\boldsymbol{\zeta}}_h + \widehat{\mathbf{u}} \right),$$

which, together with (26), further implies

$$\begin{aligned}
\|\widehat{\boldsymbol{\lambda}}_{eq}\|_1 &\leq \sqrt{l}\|\widehat{\boldsymbol{\lambda}}_{eq}\| \leq \sqrt{l}\|(\mathbf{A}\mathbf{A}^\top)^{-1}\mathbf{A}\|\left(M(1 + \widehat{\lambda}_{in}) + \widehat{\rho}D + \|\widehat{\mathbf{u}}\| \right) \\
&\leq \sqrt{l}\|(\mathbf{A}\mathbf{A}^\top)^{-1}\mathbf{A}\|\left(\frac{M^2D + \widehat{\rho}MD^2}{-h(\mathbf{x}_{\text{feas}})} + \frac{MD + \widehat{\rho}D^2}{\text{dist}(\mathbf{x}_{\text{feas}}, \partial \mathcal{X})} + M + \widehat{\rho}D \right).
\end{aligned}$$

Applying this upper bound of $\|\widehat{\boldsymbol{\lambda}}_{eq}\|$ and the upper bound of $\widehat{\lambda}_{in}$ in (26) to $\widehat{\lambda} = \widehat{\lambda}_{in} + \|\widehat{\boldsymbol{\lambda}}_{eq}\|_1$, we obtain the desired upper bound of $\widehat{\lambda}$. $\qquad \square$

Remark A.2. When there are linear equality constraints, we can just implement Algorithm 1 with the new definition of g in (22). Using the upper bound of $\widehat{\lambda}_t$ in (23) in place of the upper bound (8) in the proof of Theorem 4.3, we can show that Theorem 4.3 still holds except that Λ is repalced by Λ_{ec}.

16

Algorithm 2 Stochastic Switching Subgradient Method

1: **Input:** $\mathbf{x}^{(0)} \in \mathcal{X}$, total number of iterations T, tolerance of infeasibility $\epsilon_t \geq 0$, stepsize η_t, mini-batch size B, and a starting index S for generating outputs.
2: **Initialization:** $I = \emptyset$ and $J = \emptyset$.
3: **for** iteration $t = 0, 1, \cdots, T - 1$ **do**
4:     Generate a mini-batch $\{\omega_i^{(t)}\}_{i=1}^B$ of stochastic estimators of $g(\mathbf{x}^{(t)})$.
5:     Set $\bar{\omega}^{(t)} = \frac{1}{B} \sum_{i=1}^B \omega_i^{(t)}$.
6:     **if** $\bar{\omega}^{(t)} \leq \epsilon_t$ **then**
7:         Generate a stochastic subgradient $\boldsymbol{\xi}_f^{(t)}$ of f at $\mathbf{x}^{(t)}$.
8:         Set $\mathbf{x}^{(t+1)} = \text{proj}_{\mathcal{X}}(\mathbf{x}^{(t)} - \eta_t \boldsymbol{\xi}_f^{(t)})$ and, if $t \geq S$, $I = I \cup \{t\}$.
9:     **else**
10:        Generate a stochastic subgradient $\boldsymbol{\xi}_g^{(t)}$ of g at $\mathbf{x}^{(t)}$.
11:        Set $\mathbf{x}^{(t+1)} = \text{proj}_{\mathcal{X}}(\mathbf{x}^{(t)} - \eta_t \boldsymbol{\xi}_g^{(t)})$ and, if $t \geq S$, $I = I \cup \{t\}$.
12:     **end if**
13: **end for**
14: **Output:** $\mathbf{x}^{(\tau)}$ with τ randomly sampled from I using $\text{Prob}(\tau = t) = \eta_t / \sum_{s \in I} \eta_s$.

A.2 Stochastic switching subgradient method

In this section, we introduce a more general setting where the subgradients and function values of f and g can only be accessed through stochastic oracles, and then propose a stochastic switching subgradient method using these oracles. More specifically, in addition to Assumptions 3.1 and 4.1, we make the following assumption.

Assumption A.3. For any $\mathbf{x} \in \mathcal{X}$, stochastic subgradients $\boldsymbol{\xi}_f$ and $\boldsymbol{\xi}_g$ and a stochastic value ω can be generated independently such that $\mathbb{E}[\omega] = g(\mathbf{x})$, $\mathbb{E}[\boldsymbol{\xi}_f] \in \partial f(\mathbf{x})$ and $\mathbb{E}[\boldsymbol{\xi}_g] \in \partial g(\mathbf{x})$. Moreover, it holds that $\mathbb{E}\left[\exp\left(\|\boldsymbol{\xi}_f\|^2/M^2\right)\right] \leq \exp(1)$, $\mathbb{E}\left[\exp\left(\|\boldsymbol{\xi}_g\|^2/M^2\right)\right] \leq \exp(1)$, and $\mathbb{E}\left[\exp\left((\omega - g(\mathbf{x}))^2/\sigma^2\right)\right] \leq \exp(1)$ for a constant σ for any $\mathbf{x} \in \mathcal{X}$.

Assumption A.3 means the distributions of $\|\boldsymbol{\xi}_f\|^2$, $\|\boldsymbol{\xi}_g\|^2$ and $(\omega - g(\mathbf{x}))^2$ have light tails, which is a common assumption for proving a stochastic first-order method converges in a high probability. See (2.50) in [57] for an example. We follow the convention that $\sigma = 0$ when $w = g(\mathbf{x})$ deterministically.

Under Assumption A.3, we have to use stochastic subgradients $\boldsymbol{\xi}_f$ and $\boldsymbol{\xi}_g$ in the SSG method. Moreover, since $g(\mathbf{x}^{(t)})$ cannot be evaluated exactly, we sample a mini-batch $\{\omega_i^{(t)}\}_{i=1}^B$ of stochastic estimators of $g(\mathbf{x}^{(t)})$ with a batch size of B, and then construct an estimator of $g(\mathbf{x}^{(t)})$, denoted by $\bar{w}^{(t)}$, by averaging the samples. Then, we use the condition $\bar{w}^{(t)} \leq \epsilon_t$ in the SSG method to determine when the stochastic subgradient should be switched between $\boldsymbol{\xi}_f$ and $\boldsymbol{\xi}_g$. The SSG method with those modifications is called the stochastic SSG method and presented in Algorithm 2. Note that Algorithm 2 is reduced to Algorithm 1 when, for $t \geq 0$, it holds deterministically that

$$\boldsymbol{\xi}_f^{(t)} = \boldsymbol{\zeta}_f^{(t)}, \ \boldsymbol{\xi}_g^{(t)} = \boldsymbol{\zeta}_g^{(t)}, \ \bar{\omega}^{(t)} = g(\mathbf{x}^{(t)}), \ \sigma = 0 \ \text{and} \ B = 1. \tag{27}$$

The oracle complexity of Algorithm 1 is given in Theorem 4.3 while that of Algorithm 2 is given in Theorems A.13 and A.14.

A.3 Technical lemmas and propositions

In this section, we first introduce additional notations and then present a few technical lemmas and propositions which are needed to prove the convergence properties of Algorithms 1 and 2.

Let $\mathbb{I}(\cdot)$ be an indicator of an event, which equals one if the event occurs and zero otherwise. For each iterate $\mathbf{x}^{(t)}$ in Algorithms 1 and 2, let $\boldsymbol{\zeta}_f^{(t)} := \mathbb{E}[\boldsymbol{\xi}_f^{(t)}] \in \partial f(\mathbf{x}^{(t)})$ and $\boldsymbol{\zeta}_g^{(t)} := \mathbb{E}[\boldsymbol{\xi}_g^{(t)}] \in \partial g(\mathbf{x}^{(t)})$ and let $\widehat{\mathbf{x}}^{(t)} := \widehat{\mathbf{x}}(\mathbf{x}^{(t)})$ be defined as in (4) with $(\hat{\rho}, \tilde{\rho})$ satisfying (7) and let $\widehat{\lambda}_t \geq 0$ be the corresponding Lagrangian multiplier satisfying (5), which exists by Assumption 4.1B. Under Assumption A.3, let $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot \mid \bar{\omega}^{(0)}, \boldsymbol{\xi}_f^{(0)}, \boldsymbol{\xi}_g^{(0)}, \bar{\omega}^{(1)}, \boldsymbol{\xi}_f^{(1)}, \boldsymbol{\xi}_g^{(1)}, \ldots, \bar{\omega}^{(t-1)}, \boldsymbol{\xi}_f^{(t-1)}, \boldsymbol{\xi}_g^{(t-1)}]$, i.e., the conditional

17

expectation conditioning on all stochastic events before iteration t. Let $\mathbb{E}_\tau[\cdot]$ be the expectation taken only over the random index τ when the algorithm is terminated. The proposition below characterizes the relationship between $\|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\|$ and the parameters T, S, η_t and ϵ_t.

Proposition A.4. *Suppose Assumptions 3.1, 4.1 and A.3 hold and g is μ-strongly convex (μ can be zero). Algorithm 2 guarantees*

$$
\sum_{t=S}^{T-1} \left[\eta_t \hat{\rho} \widehat{\lambda}_t \mathbb{I}(\bar{\omega}^{(t)} \leq \epsilon_t) + \eta_t \hat{\rho} \mathbb{I}(\bar{\omega}^{(t)} > \epsilon_t) \right] \cdot \frac{\mu}{2} \|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\|^2
$$

$$
+ \sum_{t=S}^{T-1} \left[\eta_t \hat{\rho}(\hat{\rho} - \rho) \|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\|^2 - \eta_t \hat{\rho} \widehat{\lambda}_t g(\mathbf{x}^{(t)}) \right] \mathbb{I}(\bar{\omega}^{(t)} \leq \epsilon_t) + \sum_{t=S}^{T-1} \eta_t \hat{\rho} g(\mathbf{x}^{(t)}) \mathbb{I}(\bar{\omega}^{(t)} > \epsilon_t)
$$

$$
\leq \frac{\hat{\rho} D^2}{2} + \frac{\hat{\rho}}{2} \sum_{t=S}^{T-1} \left[\eta_t^2 \|\boldsymbol{\xi}_f^{(t)}\|^2 \mathbb{I}(\bar{\omega}^{(t)} \leq \epsilon_t) + \eta_t^2 \|\boldsymbol{\xi}_g^{(t)}\|^2 \mathbb{I}(\bar{\omega}^{(t)} > \epsilon_t) \right]
$$

$$
+ \sum_{t=S}^{T-1} \left[\eta_t \hat{\rho} \left\langle \boldsymbol{\xi}_f^{(t)} - \boldsymbol{\zeta}_f^{(t)}, \widehat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)} \right\rangle \mathbb{I}(\bar{\omega}^{(t)} \leq \epsilon_t) + \eta_t \hat{\rho} \left\langle \boldsymbol{\xi}_g^{(t)} - \boldsymbol{\zeta}_g^{(t)}, \widehat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)} \right\rangle \mathbb{I}(\bar{\omega}^{(t)} > \epsilon_t) \right].
$$

Proof. Let $\boldsymbol{\xi}^{(t)} = \boldsymbol{\xi}_f^{(t)}$ if $t \in I$ and $\boldsymbol{\xi}^{(t)} = \boldsymbol{\xi}_g^{(t)}$ if $t \in J$. Similarly, let $\boldsymbol{\zeta}^{(t)} = \boldsymbol{\zeta}_f^{(t)} \in \partial f(\mathbf{x}^{(t)})$ if $t \in I$ and $\boldsymbol{\zeta}^{(t)} = \boldsymbol{\zeta}_g^{(t)} \in \partial g(\mathbf{x}^{(t)})$ if $t \in J$. By the updating equation of $\mathbf{x}^{(t+1)}$, we have

$$
\|\mathbf{x}^{(t+1)} - \widehat{\mathbf{x}}^{(t)}\|^2 = \|\text{proj}_\mathcal{X}(\mathbf{x}^{(t)} - \eta_t \boldsymbol{\xi}^{(t)}) - \widehat{\mathbf{x}}^{(t)}\|^2 = \|\text{proj}_\mathcal{X}(\mathbf{x}^{(t)} - \eta_t \boldsymbol{\xi}^{(t)}) - \text{proj}_\mathcal{X}(\widehat{\mathbf{x}}^{(t)})\|^2
$$

$$
\leq \|\mathbf{x}^{(t)} - \eta_t \boldsymbol{\xi}^{(t)} - \widehat{\mathbf{x}}^{(t)}\|^2 = \|\mathbf{x}^{(t)} - \widehat{\mathbf{x}}^{(t)}\|^2 - 2\eta_t \left\langle \boldsymbol{\xi}^{(t)}, \mathbf{x}^{(t)} - \widehat{\mathbf{x}}^{(t)} \right\rangle + \eta_t^2 \|\boldsymbol{\xi}^{(t)}\|^2.
$$

Multiplying the inequality above by $\hat{\rho}/2$ and adding $f(\widehat{\mathbf{x}}^{(t)})$ to both sides, we obtain

$$
f(\widehat{\mathbf{x}}^{(t)}) + \frac{\hat{\rho}}{2} \|\mathbf{x}^{(t+1)} - \widehat{\mathbf{x}}^{(t)}\|^2
$$

$$
\leq f(\widehat{\mathbf{x}}^{(t)}) + \frac{\hat{\rho}}{2} \|\mathbf{x}^{(t)} - \widehat{\mathbf{x}}^{(t)}\|^2 - \eta_t \hat{\rho} \left\langle \boldsymbol{\xi}^{(t)}, \mathbf{x}^{(t)} - \widehat{\mathbf{x}}^{(t)} \right\rangle + \frac{\eta_t^2 \hat{\rho}}{2} \|\boldsymbol{\xi}^{(t)}\|^2
$$

$$
= \varphi(\mathbf{x}^{(t)}) - \eta_t \hat{\rho} \left\langle \boldsymbol{\zeta}^{(t)}, \mathbf{x}^{(t)} - \widehat{\mathbf{x}}^{(t)} \right\rangle + \frac{\eta_t^2 \hat{\rho}}{2} \|\boldsymbol{\xi}^{(t)}\|^2 - \eta_t \hat{\rho} \left\langle \boldsymbol{\xi}^{(t)} - \boldsymbol{\zeta}^{(t)}, \mathbf{x}^{(t)} - \widehat{\mathbf{x}}^{(t)} \right\rangle, \qquad (28)
$$

where the equality is by the definition of $\varphi(\mathbf{x})$ in (3). Since $\widehat{\mathbf{x}}^{(t)}$ is a feasible solution to problem (3) with $\mathbf{x} = \mathbf{x}^{(t+1)}$, we have

$$
\varphi(\mathbf{x}^{(t+1)}) \leq f(\widehat{\mathbf{x}}^{(t)}) + \frac{\hat{\rho}}{2} \|\mathbf{x}^{(t+1)} - \widehat{\mathbf{x}}^{(t)}\|^2,
$$

which, together with (28), implies

$$
\eta_t \hat{\rho} \left\langle \boldsymbol{\zeta}^{(t)}, \mathbf{x}^{(t)} - \widehat{\mathbf{x}}^{(t)} \right\rangle \leq \varphi(\mathbf{x}^{(t)}) - \varphi(\mathbf{x}^{(t+1)}) + \frac{\eta_t^2 \hat{\rho}}{2} \|\boldsymbol{\xi}^{(t)}\|^2 - \eta_t \hat{\rho} \left\langle \boldsymbol{\xi}^{(t)} - \boldsymbol{\zeta}^{(t)}, \mathbf{x}^{(t)} - \widehat{\mathbf{x}}^{(t)} \right\rangle. \quad (29)
$$

Next, we will bound $\left\langle \boldsymbol{\zeta}^{(t)}, \mathbf{x}^{(t)} - \widehat{\mathbf{x}}^{(t)} \right\rangle$ from below when $t \in I$ and $t \in J$, separately.

Suppose $t \in I$ so $\bar{\omega}^{(t)} \leq \epsilon_t$, $\boldsymbol{\zeta}^{(t)} = \boldsymbol{\zeta}_f^{(t)}$ and $\boldsymbol{\xi}^{(t)} = \boldsymbol{\xi}_f^{(t)}$. By the ρ-weak convexity of f, we have

$$
\left\langle \boldsymbol{\zeta}^{(t)}, \mathbf{x}^{(t)} - \widehat{\mathbf{x}}^{(t)} \right\rangle \geq f(\mathbf{x}^{(t)}) - f(\widehat{\mathbf{x}}^{(t)}) - \frac{\rho}{2} \|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\|^2
$$

$$
= f(\mathbf{x}^{(t)}) - f(\widehat{\mathbf{x}}^{(t)}) - \frac{\hat{\rho}}{2} \|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\|^2 + \frac{\hat{\rho} - \rho}{2} \|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\|^2. \qquad (30)
$$

Consider the convex optimization problem (3) with $\mathbf{x} = \mathbf{x}^{(t)}$. By Assumption 4.1B, there exists a Lagrangian multiplier $\widehat{\lambda}_t \geq 0$ such that $\widehat{\lambda}_t g(\widehat{\mathbf{x}}^{(t)}) = 0$ (complementary slackness) and

$$
\widehat{\mathbf{x}}^{(t)} = \arg\min_{\mathbf{x} \in \mathcal{X}} \left\{ f(\mathbf{x}) + \frac{\hat{\rho}}{2} \|\mathbf{x} - \mathbf{x}^{(t)}\|^2 + \widehat{\lambda}_t g(\mathbf{x}) \right\}.
$$

18

Since the objective function above is $(\hat\rho - \rho + \widehat\lambda_t\mu)$-strongly convex, we have

$$f(\mathbf{x}^{(t)}) + \widehat\lambda_t g(\mathbf{x}^{(t)}) = f(\mathbf{x}^{(t)}) + \frac{\hat\rho}{2}\|\mathbf{x}^{(t)} - \mathbf{x}^{(t)}\|^2 + \widehat\lambda_t g(\mathbf{x}^{(t)})$$

$$\geq f(\widehat{\mathbf{x}}^{(t)}) + \frac{\hat\rho}{2}\|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\|^2 + \widehat\lambda_t g(\widehat{\mathbf{x}}^{(t)}) + \frac{\hat\rho - \rho + \widehat\lambda_t\mu}{2}\|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\|^2,$$

which, by the fact that $\widehat\lambda_t g(\widehat{\mathbf{x}}^{(t)}) = 0$, implies

$$f(\mathbf{x}^{(t)}) - f(\widehat{\mathbf{x}}^{(t)}) - \frac{\hat\rho}{2}\|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\|^2 \geq -\widehat\lambda_t g(\mathbf{x}^{(t)}) + \frac{\hat\rho - \rho + \widehat\lambda_t\mu}{2}\|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\|^2.$$

Applying this inequality and inequality (30) to (29) leads to

$$\eta_t\hat\rho(\hat\rho - \rho)\|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\|^2 - \eta_t\hat\rho\widehat\lambda_t g(\mathbf{x}^{(t)}) + \frac{\eta_t\hat\rho\widehat\lambda_t\mu}{2}\|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\|^2$$

$$\leq \varphi(\mathbf{x}^{(t)}) - \varphi(\mathbf{x}^{(t+1)}) + \frac{\eta_t^2\hat\rho}{2}\|\boldsymbol{\xi}_f^{(t)}\|^2 - \eta_t\hat\rho\left\langle\boldsymbol{\xi}_f^{(t)} - \boldsymbol{\zeta}_f^{(t)}, \mathbf{x}^{(t)} - \widehat{\mathbf{x}}^{(t)}\right\rangle. \tag{31}$$

Suppose $t \in J$ so $\bar\omega^{(t)} > \epsilon_t$, $\boldsymbol{\zeta}^{(t)} = \boldsymbol{\zeta}_g^{(t)}$ and $\boldsymbol{\xi}^{(t)} = \boldsymbol{\xi}_g^{(t)}$. By the μ-strong convexity of g and the fact that $g(\widehat{\mathbf{x}}^{(t)}) \leq 0$, we have

$$\left\langle\boldsymbol{\zeta}^{(t)}, \mathbf{x}^{(t)} - \widehat{\mathbf{x}}^{(t)}\right\rangle - \frac{\mu}{2}\|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\|^2 \geq g(\mathbf{x}^{(t)}) - g(\widehat{\mathbf{x}}^{(t)}) \geq g(\mathbf{x}^{(t)}).$$

Applying this inequality to (29) leads to

$$\eta_t\hat\rho g(\mathbf{x}^{(t)}) + \frac{\eta_t\hat\rho\mu}{2}\|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\|^2$$

$$\leq \varphi(\mathbf{x}^{(t)}) - \varphi(\mathbf{x}^{(t+1)}) + \frac{\eta_t^2\hat\rho}{2}\|\boldsymbol{\xi}_g^{(t)}\|^2 - \eta_t\hat\rho\left\langle\boldsymbol{\xi}_g^{(t)} - \boldsymbol{\zeta}_g^{(t)}, \mathbf{x}^{(t)} - \widehat{\mathbf{x}}^{(t)}\right\rangle. \tag{32}$$

Recall that $\mathbb{I}(t \in I) = \mathbb{I}(\bar\omega^{(t)} \leq \epsilon_t)$ and $\mathbb{I}(t \in J) = \mathbb{I}(\bar\omega^{(t)} > \epsilon_t)$. Summing up (31) and (32) for $t = S, S+1, \ldots, T-1$, we have

$$\sum_{t=S}^{T-1}\left[\eta_t\hat\rho\widehat\lambda_t\mathbb{I}(\bar\omega^{(t)} \leq \epsilon_t) + \eta_t\hat\rho\mathbb{I}(\bar\omega^{(t)} > \epsilon_t)\right]\frac{\mu}{2}\|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\|^2$$

$$+ \sum_{t=S}^{T-1}\left[\eta_t\hat\rho(\hat\rho - \rho)\|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\|^2 - \eta_t\hat\rho\widehat\lambda_t g(\mathbf{x}^{(t)})\right]\mathbb{I}(\bar\omega^{(t)} \leq \epsilon_t) + \sum_{t=S}^{T-1}\eta_t\hat\rho g(\mathbf{x}^{(t)})\mathbb{I}(\bar\omega^{(t)} > \epsilon_t)$$

$$\leq \varphi(\mathbf{x}^{(S)}) - \varphi(\mathbf{x}^{(T)}) + \frac{\hat\rho}{2}\sum_{t=S}^{T-1}\left[\eta_t^2\|\boldsymbol{\xi}_f^{(t)}\|^2\mathbb{I}(\bar\omega^{(t)} \leq \epsilon_t) + \eta_t^2\|\boldsymbol{\xi}_g^{(t)}\|^2\mathbb{I}(\bar\omega^{(t)} > \epsilon_t)\right]$$

$$- \sum_{t=S}^{T-1}\left[\eta_t\hat\rho\left\langle\boldsymbol{\xi}_f^{(t)} - \boldsymbol{\zeta}_f^{(t)}, \mathbf{x}^{(t)} - \widehat{\mathbf{x}}^{(t)}\right\rangle\mathbb{I}(\bar\omega^{(t)} \leq \epsilon_t) + \eta_t\hat\rho\left\langle\boldsymbol{\xi}_g^{(t)} - \boldsymbol{\zeta}_g^{(t)}, \mathbf{x}^{(t)} - \widehat{\mathbf{x}}^{(t)}\right\rangle\mathbb{I}(\bar\omega^{(t)} > \epsilon_t)\right].$$

The conclusion is then implied by the facts that $\varphi(\mathbf{x}^{(T)}) = f(\widehat{\mathbf{x}}^{(T)}) + \frac{\hat\rho}{2}\|\widehat{\mathbf{x}}^{(T)} - \mathbf{x}^{(T)}\|^2 \geq f(\widehat{\mathbf{x}}^{(T)})$ and $\varphi(\mathbf{x}^{(S)}) = f(\widehat{\mathbf{x}}^{(S)}) + \frac{\hat\rho}{2}\|\widehat{\mathbf{x}}^{(S)} - \mathbf{x}^{(S)}\|^2 \leq f(\widehat{\mathbf{x}}^{(T)}) + \frac{\hat\rho}{2}\|\widehat{\mathbf{x}}^{(T)} - \mathbf{x}^{(S)}\|^2 \leq f(\widehat{\mathbf{x}}^{(T)}) + \frac{\hat\rho D^2}{2}$. $\square$

Lemma A.5. *Suppose Assumption A.3 holds. Given any $\mathbf{x} \in \mathcal{X}$, let $\{\omega_i\}_{i=1}^B$ be a mini-batch of stochastic estimators of g at $\mathbf{x}$ and $\bar\omega = \frac{1}{B}\sum_{i=1}^B \omega_i$. It holds that, for any $\delta \in (0, 1)$,*

$$\text{Prob}\left(\bar\omega > g(\mathbf{x}) + \sqrt{\frac{3}{B}}\sigma\sqrt{\ln(1/\delta)}\right) \leq \delta \ \text{ and } \ \text{Prob}\left(\bar\omega < g(\mathbf{x}) - \sqrt{\frac{3}{B}}\sigma\sqrt{\ln(1/\delta)}\right) \leq \delta.$$

Proof. The conclusion is guaranteed by Assumption A.3 and Lemma 2 (Case A) in [48] by choosing $\Omega = \sqrt{3\ln(1/\delta)}$ in their bound. $\square$

19

Lemma A.6. *Suppose Assumption A.3 holds. For any $\delta \in (0,1)$, Algorithm 2 guarantees with probability at least $1 - \delta/(T-S)$ that*

$$g(\mathbf{x}^{(t)})\mathbb{I}(\bar{\omega}^{(t)} \leq \epsilon_t) \leq \left(\epsilon_t + \sqrt{\frac{3}{B}}\sigma\sqrt{\ln((T-S)/\delta)} \right) \mathbb{I}(\bar{\omega}^{(t)} \leq \epsilon_t) \tag{33}$$

for $t = S, S+1, \ldots, T-1$ and, consequently, with probability at least $1 - \delta$ that

$$\sum_{t=S}^{T-1} \eta_t \hat{\rho}\widehat{\lambda}_t g(\mathbf{x}^{(t)})\mathbb{I}(\bar{\omega}^{(t)} \leq \epsilon_t) \leq \sum_{t=S}^{T-1} \eta_t \hat{\rho}\widehat{\lambda}_t \left(\epsilon_t + \sqrt{\frac{3}{B}}\sigma\sqrt{\ln((T-S)/\delta)} \right) \mathbb{I}(\bar{\omega}^{(t)} \leq \epsilon_t).$$

Similarly, Algorithm 2 guarantees with probability at least $1 - \delta$ that

$$\sum_{t=S}^{T-1} \eta_t \hat{\rho}g(\mathbf{x}^{(t)})\mathbb{I}(\bar{\omega}^{(t)} > \epsilon_t) \geq \sum_{t=S}^{T-1} \eta_t \hat{\rho} \left(\epsilon_t - \sqrt{\frac{3}{B}}\sigma\sqrt{\ln((T-S)/\delta)} \right) \mathbb{I}(\bar{\omega}^{(t)} > \epsilon_t).$$

Proof. For any t, $\mathbf{x}^{(t)}$ is determined by $\bar{\omega}^{(0)}, \boldsymbol{\xi}_f^{(0)}, \boldsymbol{\xi}_g^{(0)}, \bar{\omega}^{(1)}, \boldsymbol{\xi}_f^{(1)}, \boldsymbol{\xi}_g^{(1)}, \ldots, \bar{\omega}^{(t-1)}, \boldsymbol{\xi}_f^{(t-1)}$ and $\boldsymbol{\xi}_g^{(t-1)}$, while $\bar{\omega}^{(t)}$ is generated after $\mathbf{x}^{(t)}$. Hence, according to Lemma A.5, we have with a probability of at least $1 - \delta/(T-S)$ that $g(\mathbf{x}^{(t)})\mathbb{I}(\bar{\omega}^{(t)} \leq \epsilon_t) \leq \left(\bar{\omega}^{(t)} + \sqrt{\frac{3}{B}}\sigma\sqrt{\ln((T-S)/\delta)} \right) \mathbb{I}(\bar{\omega}^{(t)} \leq \epsilon_t)$, which implies (33). The first conclusion is then obtained by taking the union bound for the events above for $t = S, S+1, \ldots, T-1$. The second conclusion can be proved in a similar way. □

Lemma A.7. *Suppose Assumption A.3 holds. For any $\delta \in (0,1)$, Algorithm 2 guarantees with probability at least $1 - \delta$ that*

$$\sum_{t=S}^{T-1} \left[\eta_t^2 \|\boldsymbol{\xi}_f^{(t)}\|^2 \mathbb{I}(\bar{\omega}^{(t)} \leq \epsilon_t) + \eta_t^2 \|\boldsymbol{\xi}_g^{(t)}\|^2 \mathbb{I}(\bar{\omega}^{(t)} > \epsilon_t) \right]$$

$$\leq \sum_{t=S}^{T-1} \eta_t^2 M^2 + \max\left\{ \sqrt{12\ln(2/\delta)}, \frac{4}{3}\ln(2/\delta) \right\} \sqrt{\sum_{t=S}^{T-1} \eta_t^4 M^4}.$$

Proof. For any t, $\mathbf{x}^{(t)}$ is determined by $\bar{\omega}^{(0)}, \boldsymbol{\xi}_f^{(0)}, \boldsymbol{\xi}_g^{(0)}, \bar{\omega}^{(1)}, \boldsymbol{\xi}_f^{(1)}, \boldsymbol{\xi}_g^{(1)}, \ldots, \bar{\omega}^{(t-1)}, \boldsymbol{\xi}_f^{(t-1)}$ and $\boldsymbol{\xi}_g^{(t-1)}$. Also, $\bar{\omega}^{(t)}$, $\boldsymbol{\xi}_f^{(t)}$ and $\boldsymbol{\xi}_g^{(t)}$ are independent and generated after $\mathbf{x}^{(t)}$. Hence, by Assumption A.3, we have

$$\mathbb{E}_t \left[\exp\left(\frac{\eta_t^2 \|\boldsymbol{\xi}_f^{(t)}\|^2 \mathbb{I}(\bar{\omega}^{(t)} \leq \epsilon_t) + \eta_t^2 \|\boldsymbol{\xi}_g^{(t)}\|^2 \mathbb{I}(\bar{\omega}^{(t)} > \epsilon_t)}{\eta_t^2 M^2} \right) \right]$$

$$= \mathbb{E}_t \left[\mathbb{I}(\bar{\omega}^{(t)} \leq \epsilon_t) \exp\left(\frac{\|\boldsymbol{\xi}_f^{(t)}\|^2}{M^2} \right) \right] + \mathbb{E}_t \left[\mathbb{I}(\bar{\omega}^{(t)} > \epsilon_t) \exp\left(\frac{\|\boldsymbol{\xi}_g^{(t)}\|^2}{M^2} \right) \right]$$

$$= \mathbb{E}_t \left[\mathbb{I}(\bar{\omega}^{(t)} \leq \epsilon_t) \right] \mathbb{E}_t \left[\exp\left(\frac{\|\boldsymbol{\xi}_f^{(t)}\|^2}{M^2} \right) \right] + \mathbb{E}_t \left[\mathbb{I}(\bar{\omega}^{(t)} > \epsilon_t) \right] \mathbb{E}_t \left[\exp\left(\frac{\|\boldsymbol{\xi}_g^{(t)}\|^2}{M^2} \right) \right]$$

$$\leq \mathbb{E}_t \left[\mathbb{I}(\bar{\omega}^{(t)} \leq \epsilon_t) \right] \exp(1) + \mathbb{E}_t \left[\mathbb{I}(\bar{\omega}^{(t)} > \epsilon_t) \right] \exp(1) = \exp(1),$$

where the second equality is by the conditional independence between $\bar{\omega}^{(t)}$, $\boldsymbol{\xi}_f^{(t)}$ and $\boldsymbol{\xi}_g^{(t)}$. Then the conclusion is guaranteed by Lemma 2 (Case B) in [48] by choosing $\Omega = \max\{\sqrt{12\ln(2/\delta)}, (4/3) \cdot \ln(2/\delta)\}$ in their bound. □

20

Lemma A.8. *Suppose Assumptions 3.1 and A.3 hold. For any $\delta \in (0,1)$, Algorithm 2 guarantees with probability at least $1 - \delta$ that*

$$\sum_{t=S}^{T-1} \left[\eta_t \hat{\rho} \left\langle \boldsymbol{\xi}_f^{(t)} - \boldsymbol{\zeta}_f^{(t)}, \widehat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)} \right\rangle \mathbb{I}(\bar{\omega}^{(t)} \leq \epsilon_t) + \eta_t \hat{\rho} \left\langle \boldsymbol{\xi}_g^{(t)} - \boldsymbol{\zeta}_g^{(t)}, \widehat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)} \right\rangle \mathbb{I}(\bar{\omega}^{(t)} > \epsilon_t) \right]$$

$$\leq \sqrt{3 \ln(1/\delta)} \sqrt{\sum_{t=S}^{T-1} 4\eta_t^2 \hat{\rho}^2 M^2 D^2}.$$

Proof. By Assumption A.3 and Jensen's inequality, we have

$$\exp\left(\frac{\|\boldsymbol{\zeta}_f^{(t)}\|^2}{M^2}\right) \leq \mathbb{E}_t\left[\exp\left(\frac{\|\boldsymbol{\xi}_f^{(t)}\|^2}{M^2}\right)\right] \leq \exp(1)$$

$$\text{and} \quad \exp\left(\frac{\|\boldsymbol{\zeta}_g^{(t)}\|^2}{M^2}\right) \leq \mathbb{E}_t\left[\exp\left(\frac{\|\boldsymbol{\xi}_g^{(t)}\|^2}{M^2}\right)\right] \leq \exp(1),$$

which, by Jensen's inequality again, implies

$$\mathbb{E}_t\left[\exp\left(\frac{2\|\boldsymbol{\xi}_f^{(t)}\|^2 + 2\|\boldsymbol{\zeta}_f^{(t)}\|^2}{4M^2}\right)\right]$$

$$= \mathbb{E}_t\left[\exp\left(\frac{\|\boldsymbol{\xi}_f^{(t)}\|^2}{2M^2}\right)\exp\left(\frac{\|\boldsymbol{\zeta}_f^{(t)}\|^2}{2M^2}\right)\right] \leq \sqrt{\mathbb{E}_t\left[\exp\left(\frac{\|\boldsymbol{\xi}_f^{(t)}\|^2}{M^2}\right)\right]}\exp\left(\frac{1}{2}\right) \leq \exp(1) \quad (34)$$

$$\text{and} \quad \mathbb{E}_t\left[\exp\left(\frac{2\|\boldsymbol{\xi}_g^{(t)}\|^2 + 2\|\boldsymbol{\zeta}_g^{(t)}\|^2}{4M^2}\right)\right]$$

$$= \mathbb{E}_t\left[\exp\left(\frac{\|\boldsymbol{\xi}_g^{(t)}\|^2}{2M^2}\right)\exp\left(\frac{\|\boldsymbol{\zeta}_g^{(t)}\|^2}{2M^2}\right)\right] \leq \sqrt{\mathbb{E}_t\left[\exp\left(\frac{\|\boldsymbol{\xi}_g^{(t)}\|^2}{M^2}\right)\right]}\exp\left(\frac{1}{2}\right) \leq \exp(1). \quad (35)$$

For any t, $\mathbf{x}^{(t)}$ is determined by $\bar{\omega}^{(0)}, \boldsymbol{\xi}_f^{(0)}, \boldsymbol{\xi}_g^{(0)}, \bar{\omega}^{(1)}, \boldsymbol{\xi}_f^{(1)}, \boldsymbol{\xi}_g^{(1)}, \ldots, \bar{\omega}^{(t-1)}, \boldsymbol{\xi}_f^{(t-1)}$ and $\boldsymbol{\xi}_g^{(t-1)}$. Also, $\bar{\omega}^{(t)}, \boldsymbol{\xi}_f^{(t)}$ and $\boldsymbol{\xi}_g^{(t)}$ are independent and generated after $\mathbf{x}^{(t)}$. Hence, by Assumption A.3, we have

$$\mathbb{E}_t\left[\eta_t \hat{\rho}\left\langle \boldsymbol{\xi}_f^{(t)} - \boldsymbol{\zeta}_f^{(t)}, \widehat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)} \right\rangle \mathbb{I}(\bar{\omega}^{(t)} \leq \epsilon_t) + \eta_t \hat{\rho}\left\langle \boldsymbol{\xi}_g^{(t)} - \boldsymbol{\zeta}_g^{(t)}, \widehat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)} \right\rangle \mathbb{I}(\bar{\omega}^{(t)} > \epsilon_t)\right] = 0,$$

and

$$\mathbb{E}_t\left[\exp\left(\frac{\left[\eta_t \hat{\rho}\left\langle \boldsymbol{\xi}_f^{(t)} - \boldsymbol{\zeta}_f^{(t)}, \widehat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)} \right\rangle \mathbb{I}(\bar{\omega}^{(t)} \leq \epsilon_t) + \eta_t \hat{\rho}\left\langle \boldsymbol{\xi}_g^{(t)} - \boldsymbol{\zeta}_g^{(t)}, \widehat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)} \right\rangle \mathbb{I}(\bar{\omega}^{(t)} > \epsilon_t)\right]^2}{4\eta_t^2 \hat{\rho}^2 M^2 D^2}\right)\right]$$

$$= \mathbb{E}_t\left[\mathbb{I}(\bar{\omega}^{(t)} \leq \epsilon_t)\exp\left(\frac{\left[\left\langle \boldsymbol{\xi}_f^{(t)} - \boldsymbol{\zeta}_f^{(t)}, \widehat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)} \right\rangle\right]^2}{4M^2 D^2}\right)\right] + \mathbb{E}_t\left[\mathbb{I}(\bar{\omega}^{(t)} > \epsilon_t)\exp\left(\frac{\left[\left\langle \boldsymbol{\xi}_g^{(t)} - \boldsymbol{\zeta}_g^{(t)}, \widehat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)} \right\rangle\right]^2}{4M^2 D^2}\right)\right]$$

$$\leq \mathbb{E}_t\left[\mathbb{I}(\bar{\omega}^{(t)} \leq \epsilon_t)\right]\mathbb{E}_t\left[\exp\left(\frac{\|\boldsymbol{\xi}_f^{(t)} - \boldsymbol{\zeta}_f^{(t)}\|^2\|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\|^2}{4M^2 D^2}\right)\right]$$

$$+ \mathbb{E}_t\left[\mathbb{I}(\bar{\omega}^{(t)} > \epsilon_t)\right]\mathbb{E}_t\left[\exp\left(\frac{\|\boldsymbol{\xi}_g^{(t)} - \boldsymbol{\zeta}_g^{(t)}\|^2\|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\|^2}{4M^2 D^2}\right)\right]$$

$$\leq \mathbb{E}_t\left[\mathbb{I}(\bar{\omega}^{(t)} \leq \epsilon_t)\right]\mathbb{E}_t\left[\exp\left(\frac{2\|\boldsymbol{\xi}_f^{(t)}\|^2 + 2\|\boldsymbol{\zeta}_f^{(t)}\|^2}{4M^2}\right)\right] + \mathbb{E}_t\left[\mathbb{I}(\bar{\omega}^{(t)} > \epsilon_t)\right]\mathbb{E}_t\left[\exp\left(\frac{2\|\boldsymbol{\xi}_g^{(t)}\|^2 + 2\|\boldsymbol{\zeta}_g^{(t)}\|^2}{4M^2}\right)\right]$$

$$\leq \mathbb{E}_t\left[\mathbb{I}(\bar{\omega}^{(t)} \leq \epsilon_t)\right]\exp(1) + \mathbb{E}_t\left[\mathbb{I}(\bar{\omega}^{(t)} > \epsilon_t)\right]\exp(1) = \exp(1),$$

where the first inequality is by the Cauchy–Schwarz inequality and the conditional independence between $\bar{\omega}^{(t)}, \boldsymbol{\xi}_f^{(t)}$ and $\boldsymbol{\xi}_g^{(t)}$, the second inequality is by Assumption 3.1B and the last inequality is by (34) and (35). Then the conclusion is guaranteed by Lemma 2 (Case A) in [48] by choosing $\Omega = \sqrt{3\ln(1/\delta)}$ in their bound. $\qquad\square$

21

Suppose g is convex but not strongly convex, i.e., $\mu = 0$. Taking the union bound of the four events in Lemmas A.6, A.7 and A.8 with δ replaced by $\delta/4$ and applying the four inequalities (two from Lemma A.6, one from Lemma A.7 and one from Lemma A.8) holding in these events to Proposition A.4, we have the following bounds.

Proposition A.9. *Suppose Assumptions 3.1, 4.1 and A.3 hold and g is μ-strongly convex (μ can be zero). Algorithm 2 guarantees with probability at least $1 - \delta/(4(T - S))$ that*

$$g(\mathbf{x}^{(t)})\mathbb{I}(\bar{\omega}^{(t)} \le \epsilon_t) \le \left(\epsilon_t + \sqrt{\frac{3}{B}}\sigma\sqrt{\ln(4(T-S)/\delta)} \right)\mathbb{I}(\bar{\omega}^{(t)} \le \epsilon_t) \tag{36}$$

for $t = S, S+1, \ldots, T-1$ and, consequently, with probability at least $1 - \delta$ that

$$\sum_{t=S}^{T-1}\left[\eta_t\hat{\rho}\widehat{\lambda}_t\mathbb{I}(\bar{\omega}^{(t)} \le \epsilon_t) + \eta_t\hat{\rho}\mathbb{I}(\bar{\omega}^{(t)} > \epsilon_t)\right] \cdot \frac{\mu}{2}\|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\|^2$$

$$+ \sum_{t=S}^{T-1}\left[\eta_t\hat{\rho}(\hat{\rho} - \rho)\|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\|^2 - \eta_t\hat{\rho}\widehat{\lambda}_t\epsilon_t\right]\mathbb{I}(\bar{\omega}^{(t)} \le \epsilon_t) + \sum_{t=S}^{T-1}\eta_t\hat{\rho}\epsilon_t\mathbb{I}(\bar{\omega}^{(t)} > \epsilon_t)$$

$$\le \frac{\hat{\rho}D^2}{2} + \frac{\hat{\rho}}{2}\sum_{t=S}^{T-1}\eta_t^2 M^2 + \frac{\hat{\rho}}{2}\max\left\{\sqrt{12\ln(8/\delta)}, \frac{4}{3}\ln(8/\delta)\right\}\sqrt{\sum_{t=S}^{T-1}\eta_t^4 M^4} \tag{37}$$

$$+ \sqrt{3\ln(4/\delta)}\sqrt{\sum_{t=S}^{T-1}4\eta_t^2\hat{\rho}^2 M^2 D^2} + \sum_{t=S}^{T-1}\eta_t\hat{\rho}\sqrt{\frac{3}{B}}\sigma\sqrt{\ln(4(T-S)/\delta)}\left[\widehat{\lambda}_t\mathbb{I}(\bar{\omega}^{(t)} \le \epsilon_t) + \mathbb{I}(\bar{\omega}^{(t)} > \epsilon_t)\right].$$

A.4 Proof of Theorem 4.3

Although the technical results in the previous sections are derived for Algorithm 2 in the stochastic case, they all apply to Algorithm 1 which is a special case of Algorithm 2 when all oracles are deterministic, i.e., when (27) holds. Hence, we can uses these technical results to prove Theorem 4.3 as follows.

Proof of Theorem 4.3. Because Algorithm 1 is a deterministic special case of Algorithm 2 when (27) holds and also because we do not assume strong convexity in g ($\mu = 0$), we can simplify the inequality in Proposition A.4 as follows

$$\sum_{t=S}^{T-1}\left[\eta_t\hat{\rho}(\hat{\rho} - \rho)\|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\|^2 - \eta_t\hat{\rho}\widehat{\lambda}_t g(\mathbf{x}^{(t)})\right]\mathbb{I}(g(\mathbf{x}^{(t)}) \le \epsilon_t) + \sum_{t=S}^{T-1}\eta_t\hat{\rho}g(\mathbf{x}^{(t)})\mathbb{I}(g(\mathbf{x}^{(t)}) > \epsilon_t)$$

$$\le \frac{\hat{\rho}D^2}{2} + \frac{\hat{\rho}}{2}\sum_{t=S}^{T-1}\left[\eta_t^2\|\boldsymbol{\zeta}_f^{(t)}\|^2\mathbb{I}(g(\mathbf{x}^{(t)}) \le \epsilon_t) + \eta_t^2\|\boldsymbol{\zeta}_g^{(t)}\|^2\mathbb{I}(g(\mathbf{x}^{(t)}) > \epsilon_t)\right].$$

By the facts that $g(\mathbf{x}^{(t)})\mathbb{I}(g(\mathbf{x}^{(t)}) \le \epsilon_t) \le \epsilon_t\mathbb{I}(g(\mathbf{x}^{(t)}) \le \epsilon_t)$ and $g(\mathbf{x}^{(t)})\mathbb{I}(g(\mathbf{x}^{(t)}) > \epsilon_t) \ge \epsilon_t\mathbb{I}(g(\mathbf{x}^{(t)}) > \epsilon_t)$, the inequality above implies

$$\sum_{t=S}^{T-1}\left[\eta_t\hat{\rho}(\hat{\rho} - \rho)\|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\|^2 - \eta_t\hat{\rho}\widehat{\lambda}_t\epsilon_t\right]\mathbb{I}(g(\mathbf{x}^{(t)}) \le \epsilon_t) + \sum_{t=S}^{T-1}\eta_t\hat{\rho}\epsilon_t\mathbb{I}(g(\mathbf{x}^{(t)}) > \epsilon_t)$$

$$\le \frac{\hat{\rho}D^2}{2} + \frac{\hat{\rho}}{2}\sum_{t=S}^{T-1}\left[\eta_t^2\|\boldsymbol{\zeta}_f^{(t)}\|^2\mathbb{I}(g(\mathbf{x}^{(t)}) \le \epsilon_t) + \eta_t^2\|\boldsymbol{\zeta}_g^{(t)}\|^2\mathbb{I}(g(\mathbf{x}^{(t)}) > \epsilon_t)\right]$$

$$\le \frac{\hat{\rho}D^2}{2} + \frac{\hat{\rho}}{2}\sum_{t=S}^{T-1}\eta_t^2 M^2, \tag{38}$$

where the second inequality is because of Assumption 3.1.

We first prove that, if S, T, η_t and ϵ_t are chosen such that

$$\epsilon_t(1 + \widehat{\lambda}_t) \le \epsilon^2(\hat{\rho} - \rho) \tag{39}$$

and

$$\sum_{t=S}^{T-1}\eta_t\hat{\rho}\epsilon_t > \frac{\hat{\rho}D^2}{2} + \frac{\hat{\rho}}{2}\sum_{t=S}^{T-1}\eta_t^2 M^2, \tag{40}$$

22

we must have $g(\mathbf{x}^{(t)}) \le \epsilon_t$ for at least one t in $\{S, S+1, \ldots, T-1\}$ (i.e., $I \ne \emptyset$) and $\mathbb{E}_\tau[\|\widehat{\mathbf{x}}^{(\tau)} - \mathbf{x}^{(\tau)}\|^2] \le \epsilon^2$ (so $\mathbb{E}_\tau[\|\widehat{\mathbf{x}}^{(\tau)} - \mathbf{x}^{(\tau)}\|] \le \epsilon$).

Suppose (40) holds and $g(\mathbf{x}^{(t)}) > \epsilon_t$ for $t = S, S+1, \ldots, T-1$, i.e., $I = \emptyset$. (38) becomes exactly the opposite of (40). This contradiction means $g(\mathbf{x}^{(t)}) \le \epsilon_t$ for at least one t in $\{S, S+1, \ldots, T-1\}$. Suppose (39) and (40) hold but $\mathbb{E}_\tau[\|\widehat{\mathbf{x}}^{(\tau)} - \mathbf{x}^{(\tau)}\|^2] > \epsilon^2$. Since τ is generated by Output I, we have

$$\epsilon^2 < \mathbb{E}_\tau[\|\widehat{\mathbf{x}}^{(\tau)} - \mathbf{x}^{(\tau)}\|^2] = \frac{\sum_{t=S}^{T-1} \eta_t \mathbb{I}(g(\mathbf{x}^{(t)}) \le \epsilon_t)\|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\|^2}{\sum_{t=S}^{T-1} \eta_t \mathbb{I}(g(\mathbf{x}^{(t)}) \le \epsilon_t)}. \tag{41}$$

Note that the right-hand side of (41) is well-defined because we just proved $I \ne \emptyset$. (41) and (39) imply

$$\sum_{t=S}^{T-1} \left[\eta_t \hat{\rho}(\hat{\rho} - \rho)\|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\|^2 - \eta_t \hat{\rho}\widehat{\lambda}_t \epsilon_t \right] \mathbb{I}(g(\mathbf{x}^{(t)}) \le \epsilon_t) + \sum_{t=S}^{T-1} \eta_t \hat{\rho}\epsilon_t \mathbb{I}(g(\mathbf{x}^{(t)}) > \epsilon_t)$$

$$> \sum_{t=S}^{T-1} \left[\eta_t \hat{\rho}(\hat{\rho} - \rho)\epsilon^2 - \eta_t \hat{\rho}\widehat{\lambda}_t \epsilon_t \right] \mathbb{I}(g(\mathbf{x}^{(t)}) \le \epsilon_t) + \sum_{t=S}^{T-1} \eta_t \hat{\rho}\epsilon_t \mathbb{I}(g(\mathbf{x}^{(t)}) > \epsilon_t)$$

$$\ge \sum_{t=S}^{T-1} \eta_t \hat{\rho}\epsilon_t \mathbb{I}(g(\mathbf{x}^{(t)}) \le \epsilon_t) + \sum_{t=S}^{T-1} \eta_t \hat{\rho}\epsilon_t \mathbb{I}(g(\mathbf{x}^{(t)}) > \epsilon_t) = \sum_{t=S}^{T-1} \eta_t \hat{\rho}\epsilon_t, \tag{42}$$

where the second inequality is because of (39). Combining this inequality and (38) leads to the opposite of (40). This contradiction means $\mathbb{E}_\tau[\|\widehat{\mathbf{x}}^{(\tau)} - \mathbf{x}^{(\tau)}\|^2] \le \epsilon^2$.

Given the result above, we only need to show that the two choices of S, T, η_t and ϵ_t ensure (39) and (40).

In Case I, (39) holds because of Lemma 4.2 and the choice of ϵ_t. Let $\eta = \eta_t = \frac{2\epsilon^2(\hat{\rho}-\rho)}{5(1+\Lambda)M^2}$ for any t. Using Lemma 4.2 and plugging the values of S, T, η_t and ϵ_t in (40), we can show that (40) is equivalent to

$$\frac{T\eta\hat{\rho}\epsilon^2(\hat{\rho} - \rho)}{1 + \Lambda} > \frac{\hat{\rho}D^2}{2} + \frac{\hat{\rho}}{2}T\eta^2 M^2,$$

which, after dividing both sides by $T\eta\hat{\rho}$, can be equivalently written as

$$\frac{\epsilon^2(\hat{\rho} - \rho)}{1 + \Lambda} > \frac{D^2}{2T\eta} + \frac{\eta M^2}{2}.$$

By the values of η and T, each summand in the right-hand side of the inequality above is no more than $\frac{\epsilon^2(\hat{\rho}-\rho)}{5(1+\Lambda)}$ so the right-hand side of the inequality above no more than $\frac{2\epsilon^2(\hat{\rho}-\rho)}{5(1+\Lambda)}$ which is strictly less than the left-hand side. This means (40) holds with this choice of parameters and thus $\mathbb{E}_\tau[\|\widehat{\mathbf{x}}^{(\tau)} - \mathbf{x}^{(\tau)}\|] \le \epsilon$. By the convexity of g and the choices of η_t and ϵ_t, we have $\mathbb{E}_\tau[g(\mathbf{x}^{(\tau)})] \le \frac{\epsilon^2(\hat{\rho}-\rho)}{1+\Lambda}$.

In Case II, by the choices of ϵ_t and T, we have, for any $t \in \{S, S+1, \ldots, T-1\}$,

$$\epsilon_t = \frac{5MD}{\sqrt{t+1}} \le \frac{5MD}{\sqrt{S+1}} = \frac{5MD}{\sqrt{T/2+1}} \le \frac{\epsilon^2(\hat{\rho} - \rho)}{1 + \Lambda}. \tag{43}$$

This further implies (39) because of Lemma 4.2. Note that η_t and ϵ_t are decreasing in t. Hence, the left-hand side of (40) satisfies

$$\sum_{t=S}^{T-1} \eta_t \hat{\rho}\epsilon_t > \frac{T}{2}\eta_T \hat{\rho}\epsilon_T = \frac{5T}{2T+2}\hat{\rho}D^2 \ge \frac{5\hat{\rho}D^2}{4}. \tag{44}$$

The right-hand side of (40) satisfies

$$\frac{\hat{\rho}D^2}{2} + \frac{\hat{\rho}}{2}\sum_{t=S}^{T-1} \eta_t^2 M^2 = \frac{\hat{\rho}D^2}{2} + \frac{\hat{\rho}}{2}D^2 \sum_{t=S}^{T-1} \frac{1}{t+1} \le \hat{\rho}D^2, \tag{45}$$

where the equality is obtained by plugging in the definition of η_t and the inequality is because $\sum_{t=S}^{T-1} \frac{1}{t+1} \le \int_S^T \frac{1}{t}dt = \ln(T/S) = \ln(2) \le 1$. The right-hand side of (44) is strictly greater than the right-hand side (45), which means (40) holds and thus $\mathbb{E}_\tau[\|\widehat{\mathbf{x}}^{(\tau)} - \mathbf{x}^{(\tau)}\|] \le \epsilon$. By the convexity of g and the choices of η_t and ϵ_t, we have $\mathbb{E}_\tau[g(\mathbf{x}^{(\tau)})] \le \epsilon_S \le \frac{\epsilon^2(\hat{\rho}-\rho)}{1+\Lambda}$ according to (43). $\qquad \square$

23

A.5 Complexity analysis for strongly convex constraints

Suppose f and g in (1) are deterministic, i.e., (27) holds, and g is μ-strongly convex and $\mu > 0$. The complexity of Algorithm 1 is characterized by the following theorem.

Theorem A.10. *Suppose Assumptions 3.1 and 4.1 hold. Let $\widehat{\mathbf{x}}(\mathbf{x}^{(t)})$ be defined as in (4) with $(\hat{\rho}, \tilde{\rho})$ satisfying (7) and $\mathbf{x}^{(\tau)}$ is generated by Output II. Algorithm 1 guarantees $\mathbb{E}_\tau[\|\widehat{\mathbf{x}}^{(\tau)} - \mathbf{x}^{(\tau)}\|] \le \epsilon$ in either of the following cases.*

Case I: $S = 0$, $\epsilon_t = 0$, $\eta_t = \frac{\epsilon^2 \min\{\hat{\rho}-\rho, \mu/2\}}{M^2}$ and $T \ge \frac{M^2 D^2}{\epsilon^4 \min\{(\hat{\rho}-\rho)^2, \mu^2/4\}} = O(1/\epsilon^4).$

Case II: $S = T/2$, $\epsilon_t = 0$, $\eta_t = \frac{D}{M\sqrt{t+1}}$ and $T \ge \frac{4M^2 D^2}{\epsilon^4 \min\{(\hat{\rho}-\rho)^2, \mu^2/4\}} = O(1/\epsilon^4).$

Before presenting its proof, we would like to make a few remarks.

Remark A.11. According to Theorem A.10, strong convexity in the constraint function g does not reduce the $O(1/\epsilon^4)$ complexity of the SSG method for finding a nearly ϵ-stationary point. However, strong convexity brings benefit on other aspects. First, we can simply set $\epsilon_t = 0$ when g is strongly convex, which makes stepsize η_t the only tuning parameter. Second, the theoretical complexity no longer depends on Λ, the upper bound of the dual variables, so it can be strictly better than the one in Theorem 4.3 when Λ is very large.

Remark A.12. Theorem 4.3, A.13 and A.14 guarantee $\mathbb{E}_\tau[g(\mathbf{x}^{(\tau)})] \le O(\epsilon^2)$. This is no longer true in Theorem A.10 because τ can take value in the index set J on which $\mathbf{x}^{(t)}$ can be highly infeasible. However, with $\mathbb{E}_\tau[\|\widehat{\mathbf{x}}^{(\tau)} - \mathbf{x}^{(\tau)}\|] \le \epsilon$, we can at least derive $\mathbb{E}_\tau[g(\mathbf{x}^{(\tau)})] \le O(\epsilon)$ from Theorem A.10, which is good enough. See Remark 4.4 for the reason.

Proof of Theorem A.10. Since Algorithm 1 is fully deterministic, we can simplify the inequality in Proposition A.4 as follows

$$\sum_{t=S}^{T-1} \left[\eta_t \hat{\rho} \widehat{\lambda}_t \mathbb{I}(g(\mathbf{x}^{(t)}) \le 0) + \eta_t \hat{\rho} \mathbb{I}(g(\mathbf{x}^{(t)}) > 0) \right] \frac{\mu}{2} \|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\|^2$$

$$+ \sum_{t=S}^{T-1} \left[\eta_t \hat{\rho}(\hat{\rho} - \rho) \|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\|^2 - \eta_t \hat{\rho} \widehat{\lambda}_t g(\mathbf{x}^{(t)}) \right] \mathbb{I}(g(\mathbf{x}^{(t)}) \le 0) + \sum_{t=S}^{T-1} \eta_t \hat{\rho} g(\mathbf{x}^{(t)}) \mathbb{I}(g(\mathbf{x}^{(t)}) > 0)$$

$$\le \frac{\hat{\rho} D^2}{2} + \frac{\hat{\rho}}{2} \sum_{t=S}^{T-1} \left[\eta_t^2 \|\boldsymbol{\zeta}_f^{(t)}\|^2 \mathbb{I}(g(\mathbf{x}^{(t)}) \le 0) + \eta_t^2 \|\boldsymbol{\zeta}_g^{(t)}\|^2 \mathbb{I}(g(\mathbf{x}^{(t)}) > 0) \right].$$

After dropping some non-negative terms, the left-hand side of the inequality above can be bounded from below as

$$\sum_{t=S}^{T-1} \left[\eta_t \hat{\rho} \widehat{\lambda}_t \mathbb{I}(g(\mathbf{x}^{(t)}) \le 0) + \eta_t \hat{\rho} \mathbb{I}(g(\mathbf{x}^{(t)}) > 0) \right] \frac{\mu}{2} \|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\|^2$$

$$+ \sum_{t=S}^{T-1} \left[\eta_t \hat{\rho}(\hat{\rho} - \rho) \|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\|^2 - \eta_t \hat{\rho} \widehat{\lambda}_t g(\mathbf{x}^{(t)}) \right] \mathbb{I}(g(\mathbf{x}^{(t)}) \le 0) + \sum_{t=S}^{T-1} \eta_t \hat{\rho} g(\mathbf{x}^{(t)}) \mathbb{I}(g(\mathbf{x}^{(t)}) > 0)$$

$$\ge \sum_{t=S}^{T-1} \eta_t \hat{\rho} \mathbb{I}(g(\mathbf{x}^{(t)}) > 0) \frac{\mu}{2} \|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\|^2 + \sum_{t=S}^{T-1} \eta_t \hat{\rho}(\hat{\rho} - \rho) \|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\|^2 \mathbb{I}(g(\mathbf{x}^{(t)}) \le 0)$$

$$\ge \hat{\rho} \min\{\hat{\rho} - \rho, \mu/2\} \sum_{t=S}^{T-1} \eta_t \|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\|^2.$$

24

Combining the two inequalities above gives

$$\hat{\rho}\min\{\hat{\rho}-\rho,\mu/2\}\sum_{t=S}^{T-1}\eta_t\|\widehat{\mathbf{x}}^{(t)}-\mathbf{x}^{(t)}\|^2$$

$$\leq\frac{\hat{\rho}D^2}{2}+\frac{\hat{\rho}}{2}\sum_{t=S}^{T-1}\left[\eta_t^2\|\boldsymbol{\xi}_f^{(t)}\|^2\mathbb{I}(g(\mathbf{x}^{(t)})\leq 0)+\eta_t^2\|\boldsymbol{\zeta}_g^{(t)}\|^2\mathbb{I}(g(\mathbf{x}^{(t)})>0)\right]$$

$$\leq\frac{\hat{\rho}D^2}{2}+\frac{\hat{\rho}}{2}\sum_{t=S}^{T-1}\eta_t^2 M^2, \tag{46}$$

where the last inequality is because of Assumption 3.1. Since τ is generated by Output II, after organizing terms, we have

$$\mathbb{E}_\tau[\|\widehat{\mathbf{x}}^{(t)}-\mathbf{x}^{(t)}\|^2]\leq\frac{1}{\min\{\hat{\rho}-\rho,\mu/2\}}\left(\frac{D^2}{2}+\frac{1}{2}\sum_{t=S}^{T-1}\eta_t^2 M^2\right)\bigg/\left(\sum_{t=S}^{T-1}\eta_t\right). \tag{47}$$

In Case I, let $\eta=\eta_t=\epsilon^2\min\{\hat{\rho}-\rho,\mu/2\}/M^2$ for any t. By the choices of η_t, S and T, (47) implies

$$\mathbb{E}_\tau[\|\widehat{\mathbf{x}}^{(t)}-\mathbf{x}^{(t)}\|^2]\leq\frac{1}{\min\{\hat{\rho}-\rho,\mu/2\}}\left(\frac{D^2}{2T\eta}+\frac{\eta M^2}{2}\right)\leq\frac{\epsilon^2}{2}+\frac{\epsilon^2}{2}=\epsilon^2.$$

In Case II, by the choices of η_t, S and T, (47) implies

$$\mathbb{E}_\tau[\|\widehat{\mathbf{x}}^{(t)}-\mathbf{x}^{(t)}\|^2]\leq\frac{1}{\min\{\hat{\rho}-\rho,\mu/2\}}\left(\frac{DM}{\sqrt{T}}+\frac{DM}{\sqrt{T}}\right)\leq\frac{\epsilon^2}{2}+\frac{\epsilon^2}{2}=\epsilon^2,$$

where, in the first inequality, we use the facts that $\sum_{t=S}^{T-1}\eta_t\geq\frac{T}{2}\eta_{T-1}=\frac{\sqrt{T}D}{2M}$ and that $\sum_{t=S}^{T-1}\eta_t^2=\sum_{t=S}^{T-1}\frac{D^2}{M^2(t+1)}\leq\int_S^T\frac{D^2}{M^2 t}dt=\frac{D^2}{M^2}\ln(T/S)=\frac{D^2}{M^2}\ln(2)\leq\frac{D^2}{M^2}$, and the second inequality is by the choice of T. $\qquad\square$

A.6 Complexity analysis when subgradient oracle is stochastic

In this section, we will analyze the oracle complexity of Algorithm 2 under Assumption A.3 with an additional condition that

$$\bar{\omega}^{(t)}=g(\mathbf{x}^{(t)}),\ \sigma=0\ \text{and}\ B=1, \tag{48}$$

which is a relaxation of (27). In this case, the function value of g is can be accessed deterministically while the subgradients of f and g are still stochastic. This holds in the setting of zeroth-order or derivative-free optimization where one can evaluate f and g exactly as black boxes and then use their values to construct stochastic approximate subgradients. We will show that the complexity of Algorithm 2 when (48) holds is the same as Algorithm 1 (equivalent to Algorithm 2 when (27) holds). For simplicity of notation, we denote $\widehat{\mathbf{x}}(\mathbf{x}^{(t)})$ by $\widehat{\mathbf{x}}^{(t)}$.

Theorem A.13. *Suppose Assumptions 3.1, 4.1 and A.3 hold and Λ is as in (8). Also, suppose (48) holds. Let $\widehat{\mathbf{x}}(\mathbf{x}^{(t)})$ be defined as in (4) with $(\hat{\rho},\tilde{\rho})$ satisfying (7) and $\mathbf{x}^{(\tau)}$ is generated by Algorithm 2. Algorithm 2 guarantees $\mathbb{E}_\tau[\|\widehat{\mathbf{x}}^{(\tau)}-\mathbf{x}^{(\tau)}\|]\leq\epsilon$ and $\mathbb{E}_\tau[g(\mathbf{x}^{(\tau)})]\leq\frac{\epsilon^2(\hat{\rho}-\rho)}{1+\Lambda}$ with probability at least $1-\delta$ in either of the following cases.*

Case I: S, ϵ_t and η_t are chosen as Case I in Theorem 4.3 and

$$T\geq\left\{\frac{25M^2 D^2(1+\Lambda)^2}{4\epsilon^4(\hat{\rho}-\rho)^2},\ \max\left\{12\ln(8/\delta),\frac{16}{9}\ln^2(8/\delta)\right\},\ \frac{300\ln(4/\delta)M^2 D^2(1+\Lambda)^2}{\epsilon^4(\hat{\rho}-\rho)^2}\right\}=O(1/\epsilon^4).$$

Case II: S and η_t are chosen as Case II in Theorem 4.3, $\epsilon_t=\frac{EMD}{\sqrt{t+1}}$ where E is any positive constant such that

$$E\geq 4+\frac{2\pi}{\sqrt{6}}\max\left\{\sqrt{12\ln(8/\delta)},\frac{4}{3}\ln(8/\delta)\right\}+8\sqrt{3\ln(4/\delta)}$$

and $T\geq\frac{2E^2 M^2 D^2(1+\Lambda)^2}{\epsilon^4(\hat{\rho}-\rho)^2}=O(1/\epsilon^4)$.

25

This theorem shows that, the complexity remains $O(1/\epsilon^4)$ if the subgradient oracles are stochastic but the function value oracles remain deterministic. The only difference is that the result holds in a high probability. This complexity matches the lower-bound complexity for stochastic smooth non-convex unconstrained optimization [5, 31], so it is optimal.

Proof of Theorem A.13. By (48) and the fact that $\mu = 0$, we can drop the last term in the right-hand side of (37) and have

$$
\sum_{t=S}^{T-1} \left[\eta_t \hat{\rho}(\hat{\rho} - \rho) \|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\|^2 - \eta_t \hat{\rho} \widehat{\lambda}_t \epsilon_t \right] \mathbb{I}(g(\mathbf{x}^{(t)}) \le \epsilon_t) + \sum_{t=S}^{T-1} \eta_t \hat{\rho} \epsilon_t \mathbb{I}(g(\mathbf{x}^{(t)}) > \epsilon_t)
$$

$$
\le \frac{\hat{\rho} D^2}{2} + \frac{\hat{\rho}}{2} \sum_{t=S}^{T-1} \eta_t^2 M^2 + \frac{\hat{\rho}}{2} \max \left\{ \sqrt{12 \ln(8/\delta)}, \frac{4}{3} \ln(8/\delta) \right\} \sqrt{\sum_{t=S}^{T-1} \eta_t^4 M^4} \tag{49}
$$

$$
+ \sqrt{3 \ln(4/\delta)} \sqrt{\sum_{t=S}^{T-1} 4 \eta_t^2 \hat{\rho}^2 M^2 D^2}.
$$

with a probability of at least $1 - \delta$. In the rest of the proof, we always assume (49) holds. We first prove that, if S, T, η_t and ϵ_t are chosen such that (39) holds and

$$
\sum_{t=S}^{T-1} \eta_t \hat{\rho} \epsilon_t > \frac{\hat{\rho} D^2}{2} + \frac{\hat{\rho}}{2} \sum_{t=S}^{T-1} \eta_t^2 M^2 + \frac{\hat{\rho}}{2} \max \left\{ \sqrt{12 \ln(8/\delta)}, \frac{4}{3} \ln(8/\delta) \right\} \sqrt{\sum_{t=S}^{T-1} \eta_t^4 M^4} \tag{50}
$$

$$
+ \sqrt{3 \ln(4/\delta)} \sqrt{\sum_{t=S}^{T-1} 4 \eta_t^2 \hat{\rho}^2 M^2 D^2},
$$

we must have $g(\mathbf{x}^{(t)}) \le \epsilon_t$ for at least one t in $\{S, S+1, \ldots, T-1\}$ (i.e., $I \ne \emptyset$) and $\mathbb{E}_\tau[\|\widehat{\mathbf{x}}^{(\tau)} - \mathbf{x}^{(\tau)}\|^2] \le \epsilon^2$ (so $\mathbb{E}_\tau[\|\widehat{\mathbf{x}}^{(\tau)} - \mathbf{x}^{(\tau)}\|] \le \epsilon$).

Suppose (50) holds and $g(\mathbf{x}^{(t)}) > \epsilon_t$ for $t = S, S+1, \ldots, T-1$, i.e., $I = \emptyset$. (49) becomes exactly the opposite of (50). This contradiction means $g(\mathbf{x}^{(t)}) \le \epsilon_t$ for at least one t in $\{S, S+1, \ldots, T-1\}$. Suppose (39) and (50) hold but $\mathbb{E}_\tau[\|\widehat{\mathbf{x}}^{(\tau)} - \mathbf{x}^{(\tau)}\|^2] > \epsilon^2$. Since τ is generated by Algorithm 2, we have (41). Note that the right-hand side of (41) is well-defined because we just proved $I \ne \emptyset$. (41) and (39) imply (73). Combining (73) and (50) leads to the opposite of (49). This contradiction means $\mathbb{E}_\tau[\|\widehat{\mathbf{x}}^{(\tau)} - \mathbf{x}^{(\tau)}\|^2] \le \epsilon^2$.

Given the result above, we only need to show that the choices of S, T, η_t and ϵ_t ensure (39) and (50).

In Case I, (39) holds because of Lemma 4.2 and the choice of ϵ_t. Let $\eta = \eta_t = \frac{2\epsilon^2(\hat{\rho} - \rho)}{5(1+\Lambda)M^2}$ for any t. Using Lemma 4.2 and plugging the values of S, T, η_t and ϵ_t in (50), we can show that (50) is equivalent to

$$
\frac{T \eta \hat{\rho} \epsilon^2 (\hat{\rho} - \rho)}{1 + \Lambda} > \frac{\hat{\rho} D^2}{2} + \frac{\hat{\rho}}{2} T \eta^2 M^2 + \frac{\hat{\rho}}{2} \max \left\{ \sqrt{12 \ln(8/\delta)}, \frac{4}{3} \ln(8/\delta) \right\} \sqrt{T} \eta^2 M^2
$$

$$
+ 2 \sqrt{3 \ln(4/\delta)} \hat{\rho} \sqrt{T} \eta M D,
$$

which, after dividing both sides by $T \eta \hat{\rho}$, can be equivalently written as

$$
\frac{\epsilon^2 (\hat{\rho} - \rho)}{1 + \Lambda} > \frac{D^2}{2T\eta} + \frac{\eta M^2}{2} + \frac{1}{2} \max \left\{ \sqrt{12 \ln(8/\delta)}, \frac{4}{3} \ln(8/\delta) \right\} \frac{\eta M^2}{\sqrt{T}}
$$

$$
+ \frac{2 \sqrt{3 \ln(4/\delta)} M D}{\sqrt{T}}.
$$

By the values of η and T, each summand in the right-hand side of the inequality above is no more than $\frac{\epsilon^2(\hat{\rho} - \rho)}{5(1+\Lambda)}$ so the right-hand side of the inequality above no more than $\frac{4\epsilon^2(\hat{\rho} - \rho)}{5(1+\Lambda)}$ which is strictly less than the left-hand side. This means (50) holds with this choice of parameters and thus

26

$\mathbb{E}_\tau[\|\widehat{\mathbf{x}}^{(\tau)} - \mathbf{x}^{(\tau)}\|] \le \epsilon$. We can prove $\mathbb{E}_\tau[g(\mathbf{x}^{(\tau)})] \le \frac{\epsilon^2(\hat{\rho}-\rho)}{1+\lambda}$ in the same way as in the proof of Theorem 4.3.

In Case II, by the choices of ϵ_t and T, we have, for any $t \in \{S, S+1, \ldots, T-1\}$,

$$\epsilon_t = \frac{EMD}{\sqrt{t+1}} \le \frac{EMD}{\sqrt{S+1}} = \frac{EMD}{\sqrt{T/2+1}} \le \frac{\epsilon^2(\hat{\rho}-\rho)}{1+\Lambda}. \tag{51}$$

This further implies (39) because of Lemma 4.2. Note that η_t and ϵ_t are decreasing in t. Hence, the left-hand side of (50) satisfies

$$\sum_{t=S}^{T-1} \eta_t \hat{\rho}\epsilon_t > \frac{T}{2}\eta_T \hat{\rho}\epsilon_T = \frac{ET}{2T+2}\hat{\rho}D^2 \ge \frac{E\hat{\rho}D^2}{4}. \tag{52}$$

The right-hand side of (50) satisfies

$$\frac{\hat{\rho}D^2}{2} + \frac{\hat{\rho}}{2}\sum_{t=S}^{T-1}\eta_t^2 M^2 + \frac{\hat{\rho}}{2}\max\left\{\sqrt{12\ln(8/\delta)}, \frac{4}{3}\ln(8/\delta)\right\}\sqrt{\sum_{t=S}^{T-1}\eta_t^4 M^4}$$

$$+ \sqrt{3\ln(4/\delta)}\sqrt{\sum_{t=S}^{T-1}4\eta_t^2\hat{\rho}^2 M^2 D^2}$$

$$= \frac{\hat{\rho}D^2}{2} + \frac{\hat{\rho}}{2}D^2\sum_{t=S}^{T-1}\frac{1}{t+1} + \frac{\hat{\rho}}{2}\max\left\{\sqrt{12\ln(8/\delta)}, \frac{4}{3}\ln(8/\delta)\right\}D^2\sqrt{\sum_{t=S}^{T-1}\frac{1}{(t+1)^2}}$$

$$+ 2\sqrt{3\ln(4/\delta)}\hat{\rho}D^2\sqrt{\sum_{t=S}^{T-1}\frac{1}{t+1}}$$

$$\le \hat{\rho}D^2 + \frac{\hat{\rho}\pi}{2\sqrt{6}}\max\left\{\sqrt{12\ln(8/\delta)}, \frac{4}{3}\ln(8/\delta)\right\}D^2 + 2\sqrt{3\ln(4/\delta)}\hat{\rho}D^2, \tag{53}$$

where the equality is obtained by plugging in the definition of η_t and the inequality is because $\sum_{t=S}^{T-1}\frac{1}{t+1} \le \int_S^T \frac{1}{t}dt = \ln(T/S) = \ln(2) \le 1$ and $\sum_{t=S}^{T-1}\frac{1}{(t+1)^2} \le \pi^2/6$. By the condition satisfied by E, the right-hand side of (52) is greater than or equal to the right-hand side (53). This means (40) holds with this choice of parameters and thus $\mathbb{E}_\tau[\|\widehat{\mathbf{x}}^{(\tau)} - \mathbf{x}^{(\tau)}\|] \le \epsilon$. We can prove $\mathbb{E}_\tau[g(\mathbf{x}^{(\tau)})] \le \frac{\epsilon^2(\hat{\rho}-\rho)}{1+\lambda}$ in the same way as in the proof of Theorem 4.3. □

A.7 Complexity analysis when subgradient and function oracles are both stochastic

In this section, we will analyze the oracle complexity of Algorithm 2 under Assumption A.3 without requiring (27) or (48). In this case, not only the subgradient oracles are stochastic but also the function value g can only be accessed through a stochastic oracle. When the function value of g is stochastic, the complexity analysis becomes fundamentally more challenging than the case of a deterministic g. In fact, the challenge comes only from the stochastic function value of g instead of its stochastic subgradient. Note that, in the fully deterministic case, Algorithm 1 essentially updates $\mathbf{x}^{(t)}$ along a hybrid subgradient

$$\mathbb{I}(g(\mathbf{x}^{(t)}) \le \epsilon_t)\boldsymbol{\zeta}_f^{(t)} + \mathbb{I}(g(\mathbf{x}^{(t)}) > \epsilon_t)\boldsymbol{\zeta}_g^{(t)}. \tag{54}$$

If only the subgradients are stochastic but the function value $g(\mathbf{x}^{(t)})$ remains deterministic, the hybrid stochastic subgradient $\mathbb{I}(g(\mathbf{x}^{(t)}) \le \epsilon_t)\boldsymbol{\xi}_f^{(t)} + \mathbb{I}(g(\mathbf{x}^{(t)}) > \epsilon_t)\boldsymbol{\xi}_g^{(t)}$ provides an unbiased estimation of (54). In this case, we can still obtain complexity of $O(1/\epsilon^4)$ (see Theorem A.13) by a proof similar to the deterministic case. However, when $g(\mathbf{x}^{(t)})$ must be queried through some unbiased estimator $w^{(t)}$, the naively constructed direction

$$\mathbb{I}(w^{(t)} \le \epsilon_t)\boldsymbol{\xi}_f^{(t)} + \mathbb{I}(w^{(t)} > \epsilon_t)\boldsymbol{\xi}_g^{(t)}$$

27

is not an unbiased estimator of (54). To tackle this issue, we have to query a mini-batch of $w^{(t)}$ of size B, i.e., $\{\omega_i^{(t)}\}_{i=1}^B$ to construct $\bar{w}^{(t)}$ as an estimation of $g(\mathbf{x}^{(t)})$ with a high accuracy in a high probability. In this way, we can use

$$\mathbb{I}(\bar{w}^{(t)} \leq \epsilon_t)\boldsymbol{\xi}_f^{(t)} + \mathbb{I}(\bar{w}^{(t)} > \epsilon_t)\boldsymbol{\xi}_g^{(t)}$$

as a nearly unbiased estimator of (54), which leads to the switching condition in Algorithm 2. As a consequence, the oracle complexity of the function value of g increases from $O(1/\epsilon^4)$ to $\tilde{O}(1/\epsilon^8)$. [1] although the oracle complexity of stochastic subgradients remains $O(1/\epsilon^4)$.

Theorem A.14. *Suppose Assumptions 3.1, 4.1 and A.3 hold and Λ is as in (8). Let $\hat{\mathbf{x}}(\mathbf{x}^{(t)})$ be defined as in (4) with $(\hat{\rho}, \tilde{\rho})$ satisfying (7) and $\mathbf{x}^{(\tau)}$ is generated by Algorithm 2. Algorithm 2 guarantees $\mathbb{E}_\tau[\|\hat{\mathbf{x}}^{(\tau)} - \mathbf{x}^{(\tau)}\|] \leq \epsilon$ and $\mathbb{E}_\tau[g(\mathbf{x}^{(\tau)})] \leq \frac{2\epsilon^2(\hat{\rho}-\rho)}{1+\Lambda}$ with probability at least $1 - \delta$ in either of the following cases.*

Case I: If S, ϵ_t, η_t and T are chosen as Case I in Theorem A.13 and $B = \frac{300\sigma^2 \ln(4T/\delta)(1+\Lambda)^4}{\epsilon^4(\hat{\rho}-\rho)^2}$.

Case II: If S, ϵ_t, η_t and T are chosen as Case II in Theorem A.13 except that E is any positive constant such that

$$E \geq 8 + \frac{2\pi}{\sqrt{6}} \max\left\{\sqrt{12\ln(8/\delta)}, \frac{4}{3}\ln(8/\delta)\right\} + 8\sqrt{3\ln(4/\delta)}$$

and $B = \frac{3T\sigma^2 \ln(2T/\delta)(1+\Lambda)^2}{2M^2 D^2}$.

In each iteration of Algorithm 2, we query one stochastic subgradient of f or g but B stochastic function values of g. In both Case I and Case II, we have $T = O(1/\epsilon^4)$ and $B = \tilde{O}(1/\epsilon^4)$ so the subgradient oracle complexity is still $O(1/\epsilon^4)$ but the function value oracle complexity becomes $\tilde{O}(1/\epsilon^8)$, which is higher than the $O(1/\epsilon^6)$ complexity by the double-loop methods in [13, 54]. It is our future work to reduce the complexity when g is stochastic, for example, by a single-loop primal-dual method that uses a hybrid subgradient like $\boldsymbol{\xi}_f^{(t)} + \lambda_t\boldsymbol{\xi}_g^{(t)}$ with the dual variable λ_t updated by only one sample of $w^{(t)}$.

Proof of Theorem A.14. By Proposition A.9, with a probability of at least $1 - \delta$, we simultaneously have (37) with $\mu = 0$ and (36) for $t = S, S+1, \ldots, T-1$. In the rest of the proof, we assume (37) with $\mu = 0$ and (36) hold for $t = S, S+1, \ldots, T-1$.

We first prove that, if S, T, B, η_t and ϵ_t are chosen such that (39) holds and

$$\sum_{t=S}^{T-1} \eta_t\hat{\rho}\epsilon_t > \frac{\hat{\rho}D^2}{2} + \frac{\hat{\rho}}{2}\sum_{t=S}^{T-1}\eta_t^2 M^2 + \frac{\hat{\rho}}{2}\max\left\{\sqrt{12\ln(8/\delta)}, \frac{4}{3}\ln(8/\delta)\right\}\sqrt{\sum_{t=S}^{T-1}\eta_t^4 M^4} \quad (55)$$

$$+ \sqrt{3\ln(4/\delta)}\sqrt{\sum_{t=S}^{T-1}4\eta_t^2\hat{\rho}^2 M^2 D^2} + \sum_{t=S}^{T-1}\eta_t\hat{\rho}\sqrt{\frac{3}{B}}\sigma\sqrt{\ln(4(T-S)/\delta)}(1+\Lambda),$$

we must have $\bar{\omega}^{(t)} \leq \epsilon_t$ for at least one t in $\{S, S+1, \ldots, T-1\}$ (i.e., $I \neq \emptyset$) and $\mathbb{E}_\tau[\|\hat{\mathbf{x}}^{(\tau)} - \mathbf{x}^{(\tau)}\|^2] \leq \epsilon^2$ (so $\mathbb{E}_\tau[\|\hat{\mathbf{x}}^{(\tau)} - \mathbf{x}^{(\tau)}\|] \leq \epsilon$).

Suppose (55) holds and $\bar{\omega}^{(t)} > \epsilon_t$ for $t = S, S+1, \ldots, T-1$, i.e., $I = \emptyset$. (37) contradicts with (55) as $\Lambda \geq \hat{\lambda}_t$. This contradiction means $\bar{\omega}^{(t)} \leq \epsilon_t$ for at least one t in $\{S, S+1, \ldots, T-1\}$ so $I \neq \emptyset$. Suppose (39) and (55) hold but $\mathbb{E}_\tau[\|\hat{\mathbf{x}}^{(\tau)} - \mathbf{x}^{(\tau)}\|^2] > \epsilon^2$. Since τ is generated by Algorithm 2, we have

$$\epsilon^2 < \mathbb{E}_\tau[\|\hat{\mathbf{x}}^{(\tau)} - \mathbf{x}^{(\tau)}\|^2] = \frac{\sum_{t=S}^{T-1}\eta_t\mathbb{I}(\bar{\omega}^{(t)} \leq \epsilon_t)\|\hat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\|^2}{\sum_{t=S}^{T-1}\eta_t\mathbb{I}(\bar{\omega}^{(t)} \leq \epsilon_t)}. \quad (56)$$

---

[1] $\tilde{O}(\cdot)$ suppresses logarithmic factors in the order of complexity.

28

Note that the right-hand side of (56) is well-defined because we just proved $I \neq \emptyset$. (56) and (39) imply

$$
\begin{aligned}
&\sum_{t=S}^{T-1} \left[\eta_t \hat{\rho}(\hat{\rho} - \rho) \| \widehat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)} \|^2 - \eta_t \hat{\rho} \widehat{\lambda}_t \epsilon_t \right] \mathbb{I}(\bar{\omega}^{(t)} \leq \epsilon_t) + \sum_{t=S}^{T-1} \eta_t \hat{\rho} \epsilon_t \mathbb{I}(\bar{\omega}^{(t)} > \epsilon_t) \\
&> \sum_{t=S}^{T-1} \left[\eta_t \hat{\rho}(\hat{\rho} - \rho) \epsilon^2 - \eta_t \hat{\rho} \widehat{\lambda}_t \epsilon_t \right] \mathbb{I}(\bar{\omega}^{(t)} \leq \epsilon_t) + \sum_{t=S}^{T-1} \eta_t \hat{\rho} \epsilon_t \mathbb{I}(\bar{\omega}^{(t)} > \epsilon_t) \\
&\geq \sum_{t=S}^{T-1} \eta_t \hat{\rho} \epsilon_t \mathbb{I}(\bar{\omega}^{(t)} \leq \epsilon_t) + \sum_{t=S}^{T-1} \eta_t \hat{\rho} \epsilon_t \mathbb{I}(\bar{\omega}^{(t)} > \epsilon_t) \geq \sum_{t=S}^{T-1} \eta_t \hat{\rho} \epsilon_t, \quad (57)
\end{aligned}
$$

where the second inequality is because of (39). Combining (57) and (55) leads to the opposite of (37). This contradiction means $\mathbb{E}_\tau[\|\widehat{\mathbf{x}}^{(\tau)} - \mathbf{x}^{(\tau)}\|^2] \leq \epsilon^2$.

Given the result above, we only need to show that the choices of S, T, B, η_t and ϵ_t ensure (39) and (55).

In Case I, (39) holds because of Lemma 4.2 and the choice of ϵ_t. Let $\eta = \eta_t = \frac{2\epsilon^2(\hat{\rho} - \rho)}{5(1+\Lambda)M^2}$ for any t. Using Lemma 4.2 and plugging the values of S, T, B, η_t and ϵ_t in (55), we can show that (55) holds if

$$
\begin{aligned}
\frac{T\eta\hat{\rho}\epsilon^2(\hat{\rho} - \rho)}{1 + \Lambda} &> \frac{\hat{\rho}D^2}{2} + \frac{\hat{\rho}}{2}T\eta^2M^2 + \frac{\hat{\rho}}{2}\max\left\{ \sqrt{12\ln(8/\delta)}, \frac{4}{3}\ln(8/\delta) \right\} \sqrt{T}\eta^2M^2 \\
&\quad + 2\sqrt{3\ln(4/\delta)}\hat{\rho}\sqrt{T}\eta MD + T\eta\hat{\rho}\sqrt{\frac{3}{B}}\sigma\sqrt{\ln(4T/\delta)}(1+\Lambda),
\end{aligned}
$$

which, after dividing both sides by $T\eta\hat{\rho}$, can be equivalently written as

$$
\begin{aligned}
\frac{\epsilon^2(\hat{\rho} - \rho)}{1 + \Lambda} &> \frac{D^2}{2T\eta} + \frac{\eta M^2}{2} + \frac{1}{2}\max\left\{ \sqrt{12\ln(8/\delta)}, \frac{4}{3}\ln(8/\delta) \right\} \frac{\eta M^2}{\sqrt{T}} \\
&\quad + \frac{2\sqrt{3\ln(4/\delta)}MD}{\sqrt{T}} + \sqrt{\frac{3}{B}}\sigma\sqrt{\ln(4T/\delta)}(1+\Lambda).
\end{aligned}
$$

By the values of η, B and T, each of the first four summands on the right-hand side of the inequality above is no more than $\frac{\epsilon^2(\hat{\rho}-\rho)}{5(1+\Lambda)}$ while the last summand is no more than $\frac{\epsilon^2(\hat{\rho}-\rho)}{10(1+\Lambda)}$, so the right-hand side of the inequality above no more than $\frac{9\epsilon^2(\hat{\rho}-\rho)}{10(1+\Lambda)}$ which is strictly less than the left-hand side. This means (40) holds with this choice of parameters and thus $\mathbb{E}_\tau[\|\widehat{\mathbf{x}}^{(\tau)} - \mathbf{x}^{(\tau)}\|] \leq \epsilon$. Moreover, by (36), we have

$$
\begin{aligned}
\mathbb{E}_\tau[g(\mathbf{x}^{(\tau)})] &= \frac{\sum_{t=0}^{T-1} \eta_t g(\mathbf{x}^{(t)})\mathbb{I}(\bar{\omega}^{(t)} \leq \epsilon_t)}{\sum_{t=0}^{T-1} \eta_t \mathbb{I}(\bar{\omega}^{(t)} \leq \epsilon_t)} \leq \frac{\sum_{t=0}^{T-1} \eta_t \left(\epsilon_t + \sqrt{\frac{3}{B}}\sigma\sqrt{\ln(4T/\delta)} \right)\mathbb{I}(\bar{\omega}^{(t)} \leq \epsilon_t)}{\sum_{t=0}^{T-1} \eta_t \mathbb{I}(\bar{\omega}^{(t)} \leq \epsilon_t)} \\
&\leq \frac{\epsilon^2(\hat{\rho} - \rho)}{1 + \Lambda} + \sqrt{\frac{3}{B}}\sigma\sqrt{\ln(4T/\delta)} \leq \frac{2\epsilon^2(\hat{\rho} - \rho)}{1 + \Lambda},
\end{aligned}
$$

where the last inequality is because of the choice of B.

29

In Case II, by the choice of ϵ_t, we have (51) holds. This further implies (39) because of Lemma 4.2. Note that η_t and ϵ_t are decreasing in t. Hence, we also have (52). The right-hand side of (55) satisfies

$$\frac{\hat{\rho}D^2}{2} + \frac{\hat{\rho}}{2}\sum_{t=S}^{T-1}\eta_t^2 M^2 + \frac{\hat{\rho}}{2}\max\left\{\sqrt{12\ln(8/\delta)}, \frac{4}{3}\ln(8/\delta)\right\}\sqrt{\sum_{t=S}^{T-1}\eta_t^4 M^4}$$

$$+ \sqrt{3\ln(4/\delta)}\sqrt{\sum_{t=S}^{T-1}4\eta_t^2\hat{\rho}^2 M^2 D^2} + \sum_{t=S}^{T-1}\eta_t\hat{\rho}\sqrt{\frac{3}{B}}\sigma\sqrt{\ln(4(T-S)/\delta)}(1+\Lambda)$$

$$= \frac{\hat{\rho}D^2}{2} + \frac{\hat{\rho}}{2}D^2\sum_{t=S}^{T-1}\frac{1}{t+1} + \frac{\hat{\rho}}{2}\max\left\{\sqrt{12\ln(8/\delta)}, \frac{4}{3}\ln(8/\delta)\right\}D^2\sqrt{\sum_{t=S}^{T-1}\frac{1}{(t+1)^2}}$$

$$+ 2\sqrt{3\ln(4/\delta)}\hat{\rho}D^2\sqrt{\sum_{t=S}^{T-1}\frac{1}{t+1}} + \frac{\hat{\rho}D}{M}\sqrt{\frac{3}{B}}\sigma\sqrt{\ln(4(T-S)/\delta)}(1+\Lambda)\sum_{t=S}^{T-1}\frac{1}{\sqrt{t+1}}$$

$$\leq \hat{\rho}D^2 + \frac{\hat{\rho}\pi}{2\sqrt{6}}\max\left\{\sqrt{12\ln(8/\delta)}, \frac{4}{3}\ln(8/\delta)\right\}D^2 + 2\sqrt{3\ln(4/\delta)}\hat{\rho}D^2 + \hat{\rho}D^2, \tag{58}$$

where the equality is obtained by plugging in the definition of η_t and the inequality is because of the definition of B and the facts that $\sum_{t=S}^{T-1}\frac{1}{t+1} \leq \int_S^T \frac{1}{t}dt = \ln(T/S) = \ln(2) \leq 1$, $\sum_{t=S}^{T-1}\frac{1}{(t+1)^2} \leq \pi^2/6$ and $\sum_{t=S}^{T-1}\frac{1}{\sqrt{t+1}} \leq \int_S^T \frac{1}{\sqrt{t}}dt \leq \sqrt{T/2}$. By the condition of E, the right-hand side of (52) is strictly greater than the right-hand side (58). This means (55) holds with this choice of parameters and thus $\mathbb{E}_\tau[\|\hat{\mathbf{x}}^{(\tau)} - \mathbf{x}^{(\tau)}\|] \leq \epsilon$. Moreover, since (33) holds with δ replaced by $\delta/4$ for $t = S, S+1, \dots, T-1$, we have

$$\mathbb{E}_\tau[g(\mathbf{x}^{(\tau)})] = \frac{\sum_{t=S}^{T-1}\eta_t g(\mathbf{x}^{(t)})\mathbb{I}(\bar{\omega}^{(t)} \leq \epsilon_t)}{\sum_{t=S}^{T-1}\eta_t\mathbb{I}(\bar{\omega}^{(t)} \leq \epsilon_t)} \leq \frac{\sum_{t=S}^{T-1}\eta_t\left(\epsilon_t + \sqrt{\frac{3}{B}}\sigma\sqrt{\ln(2T/\delta)}\right)\mathbb{I}(\bar{\omega}^{(t)} \leq \epsilon_t)}{\sum_{t=S}^{T-1}\eta_t\mathbb{I}(\bar{\omega}^{(t)} \leq \epsilon_t)}$$

$$\leq \epsilon_{T/2} + \sqrt{\frac{3}{B}}\sigma\sqrt{\ln(2T/\delta)} \leq \frac{2\epsilon^2(\hat{\rho} - \rho)}{1 + \Lambda},$$

where the last inequality is because of (51) and the choices of B and T. □

Remark A.15. When g is μ-strongly convex with $\mu > 0$, under the same assumptions as Theorem A.13, we can establish $O(1/\epsilon^4)$ oracle complexity for Algorithm 2. Similarly, under the same assumptions as Theorem A.14, we can also establish $\tilde{O}(1/\epsilon^8)$ oracle complexity for Algorithm 2. These two complexity results of Algorithm 2 can be proved in a similar way as Theorem A.10. Since those results do not provide additional insights on the complexity and analysis for the SSG method, we do not include them in the paper. Similar to Theorem A.10, when $\mu > 0$, we only need to set $\epsilon_t = 0$ in Algorithm 2 and at last generate τ by Output II in Algorithm 1. This indicates that strong convexity in g also reduces the number of tuning parameters in Algorithm 2 just as it does for Algorithm 1.

A.8 Complexity analysis for convex constraints with bounded $\mathcal{S}$ (instead of $\mathcal{X}$)

In this section, we present the complexity analysis when f is weakly convex and g is convex after a small modification on Assumption 4.1. Recall that

$$g_+(\mathbf{x}) = \max\{g(\mathbf{x}), 0\}, \quad \mathcal{L} = \{\mathbf{x} \in \mathcal{X} \mid g(\mathbf{x}) = 0\} \text{ and } \mathcal{S} = \{\mathbf{x} \in \mathcal{X} \mid g(\mathbf{x}) \leq 0\}.$$

In this section, we choose parameters in (3) the same as in (7). We make the following assumptions in addition to Assumption 3.1 in this section.

Assumption A.16. The following statements hold:

   A. $f(\mathbf{x})$ is ρ-weakly convex on $\mathcal{X}$ and $g(\mathbf{x})$ is convex on $\mathcal{X}$.

   B. (Slater's condition) There exists $\mathbf{x}_{\text{feas}} \in \text{relint}(\mathcal{X})$ such that $g(\mathbf{x}_{\text{feas}}) < 0$.

   C. There exists $D > 0$ such that $\|\mathbf{x} - \mathbf{x}'\| \leq D$ for any $\mathbf{x}$ and $\mathbf{x}'$ in $\mathcal{S}$.

30

Under Assumption A.16, $\mathcal{X}$ is not necessarily bounded (e.g. $\mathcal{X} = \mathbb{R}^d$) while g must have a bounded 0-sublevel set over $\mathcal{X}$.

We next show that Assumptions A.16 B and C imply that subgradient of $g + \delta_{\mathcal{X}}$ can be bounded away from zero on $\mathcal{L}$. Here, $\delta_{\mathcal{X}}$ is the indicator function of $\mathcal{X}$. Given any $\mathbf{x} \in \mathcal{L}$, by Assumption A.16B and the convexity of $g + \delta_{\mathcal{X}}$, we have

$$g(\mathbf{x}_{\text{feas}}) \geq g(\mathbf{x}) + \langle \boldsymbol{\zeta}_g + \mathbf{u}, \mathbf{x}_{\text{feas}} - \mathbf{x} \rangle \geq -\|\boldsymbol{\zeta}_g + \mathbf{u}\| \cdot \|\mathbf{x}_{\text{feas}} - \mathbf{x}\|$$

for any $\boldsymbol{\zeta}_g \in \partial g(\mathbf{x})$ and $\mathbf{u} \in \mathcal{N}_{\mathcal{X}}(\mathbf{x})$. Since $\|\mathbf{x}_{\text{feas}} - \mathbf{x}\| \leq D$ by Assumption A.16C, it holds that

$$\min_{\boldsymbol{\zeta}_g \in \partial g(\mathbf{x}), \mathbf{u} \in \mathcal{N}_{\mathcal{X}}(\mathbf{x})} \|\boldsymbol{\zeta}_g + \mathbf{u}\| \geq \frac{-g(\mathbf{x}_{\text{feas}})}{\|\mathbf{x}_{\text{feas}} - \mathbf{x}\|} \geq \nu' := -g(\mathbf{x}_{\text{feas}})/D > 0. \tag{59}$$

This result is similar to (11).

Note that $\mathcal{S}$ is the optimal set of $\min_{\mathbf{x} \in \mathcal{X}} g_+(\mathbf{x})$, which is a convex non-smooth optimization problem with an optimal value of zero. By Lemma 1 in [76], (59) implies that $g_+(\mathbf{x})$ satisfies a global linear error bound on $\mathcal{X}$, namely, for any $\mathbf{x} \in \mathcal{X}$,

$$\nu' \text{dist}(\mathbf{x}, \mathcal{S}) \leq g_+(\mathbf{x}). \tag{60}$$

This result is stronger than (12) because (12) does not hold globally over $\mathcal{X}$. Moreover, this result implies $\nu' \leq M$.

Following almost the same proof as Lemma B.3, we obtain a result similar to (98).

Lemma A.17. *Suppose Assumptions 3.1 and A.16 hold. Also, suppose the sequence $\{\mathbf{x}^{(t)}\}_{t \geq 0}$ is generated by applying the projected subgradient method to $\min_{\mathbf{x} \in \mathcal{X}} g_+(\mathbf{x})$ using a Polyak's stepsize, namely,*

$$\mathbf{x}^{(t+1)} = \text{proj}_{\mathcal{X}}(\mathbf{x}^{(t)} - \eta_t \boldsymbol{\zeta}_g^{(t)}), \quad \eta_t = \begin{cases} v' g_+(\mathbf{x}^{(t)})/(2M\|\boldsymbol{\zeta}_g^{(t)}\|^2) & \text{if } \boldsymbol{\zeta}_g^{(t)} \neq \mathbf{0} \\ 0 & \text{if } \boldsymbol{\zeta}_g^{(t)} = \mathbf{0} \end{cases}, \quad \text{for } t = 0, 1, \ldots, \tag{61}$$

where $\boldsymbol{\zeta}_g^{(t)} \in \partial g_+(\mathbf{x}^{(t)})$. We have

$$\text{dist}^2(\mathbf{x}^{(t+1)}, \mathcal{S}) \leq \left(1 - \frac{3\nu'^3}{4M^3}\right) \text{dist}^2(\mathbf{x}^{(t)}, \mathcal{S}), \quad \forall t \geq 0. \tag{62}$$

Proof. Suppose $\boldsymbol{\zeta}_g^{(t)} = \mathbf{0}$ so $\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)}$. This happens only when $g_+(\mathbf{x}^{(t)}) = 0$. By (60), we must have $\text{dist}(\mathbf{x}^{(t)}, \mathcal{S}) = 0$ and thus $\text{dist}(\mathbf{x}^{(t+1)}, \mathcal{S}) = 0 \leq \left(1 - \frac{3\nu'^3}{4M^3}\right) \text{dist}(\mathbf{x}^{(t)}, \mathcal{S})$ and (62) holds.

Suppose $\boldsymbol{\zeta}_g^{(t)} \neq \mathbf{0}$. Let $\mathbf{x}^{\dagger(t)} = \text{proj}_{\mathcal{S}}(\mathbf{x}^{(t)})$ for $t \geq 0$. Following a proof similar to Lemma B.3, which is originally from [26], we have

$$\text{dist}^2(\mathbf{x}^{(t+1)}, \mathcal{S}) \leq \text{dist}^2(\mathbf{x}^{(t)}, \mathcal{S}) - 2\eta_t \left\langle \boldsymbol{\zeta}_g^{(t)}, \mathbf{x}^{(t)} - \mathbf{x}^{\dagger(t)} \right\rangle + \eta_t^2 \|\boldsymbol{\zeta}_g^{(t)}\|^2$$

$$\leq \text{dist}^2(\mathbf{x}^{(t)}, \mathcal{S}) + \frac{\nu' g_+(\mathbf{x}^{(t)})}{M\|\boldsymbol{\zeta}_g^{(t)}\|^2} \left(g_+(\mathbf{x}^{\dagger(t)}) - g_+(\mathbf{x}^{(t)})\right) + \frac{\nu'^2 g_+^2(\mathbf{x}^{(t)})}{4M^2\|\boldsymbol{\zeta}_g^{(t)}\|^2}$$

$$= \text{dist}^2(\mathbf{x}^{(t)}, \mathcal{S}) - \left(\frac{\nu'}{M} - \frac{\nu'^2}{4M^2}\right) \left(g_+(\mathbf{x}^{(t)})\right)^2 /\|\boldsymbol{\zeta}_g^{(t)}\|^2$$

$$\leq \text{dist}^2(\mathbf{x}^{(t)}, \mathcal{S}) - \frac{3\nu'^3}{4M^3} \|\mathbf{x}^{(t)} - \mathbf{x}^{\dagger(t)}\|^2,$$

where the equality is because $g_+(\mathbf{x}^{\dagger(t)}) = 0$, the second inequality is by the convexity of g_+ and the last is by (60) and the fact that $\nu' \leq M$. $\qquad\square$

Using almost the same proof as Proposition 5.5, we can show the following proposition.

Proposition A.18. *Suppose Assumptions 3.1 and A.16 hold and $\epsilon > 0$. Also, suppose that $\mathbf{x}^{(t)}$ is generated by Algorithm 1 using $\mathbf{x}^{(0)} \in \mathcal{S}$, $\epsilon_t = \frac{\nu'}{2} \min\{\epsilon^2/M, \nu'\}$ and*

$$\eta_t = \begin{cases} \frac{\nu'}{2M^2} \min\{\epsilon^2/M, \nu'\} & \text{if } t \in I \\ \nu' g(\mathbf{x}^{(t)})/(2M\|\boldsymbol{\zeta}_g^{(t)}\|^2) & \text{if } t \in J. \end{cases}$$

31

Then $\text{dist}(\mathbf{x}^{(t)}, \mathcal{S}) \leq \min\{\epsilon^2/M, \nu'\}$ *and* $g(\mathbf{x}^{(t)}) \leq \epsilon^2$ *for any* $t \geq 0$. *As a consequence,* $\mathbf{x}^{(t)}$ *is* ϵ^2*-feasible to* (3) *where* $\mathbf{x} = \mathbf{x}^{(t)}$ *and* $(\hat{\rho}, \tilde{\rho})$ *satisfies* (7).

Proof. The choices of ϵ_t and η_t imply

$$\epsilon_t/\nu' + \eta_t M^2/\nu' \leq \min\{\epsilon^2/M, \nu'\} \tag{63}$$

for $t \in I$. We prove $\text{dist}(\mathbf{x}^{(t)}, \mathcal{S}) \leq \min\{\epsilon^2/M, \nu'\}$ by induction on t. Since $\mathbf{x}^{(0)} \in \mathcal{S}$, this conclusion holds trivially for $t = 0$. Suppose it holds up to iteration t. We want to prove it also holds for iteration $t + 1$.

Suppose $t \in I$. Since $g(\mathbf{x}^{(t)}) \leq \epsilon_t$ for $t \in I$, we have

$$g(\mathbf{x}^{(t+1)}) \leq g(\mathbf{x}^{(t)}) + M\|\mathbf{x}^{(t+1)} - \mathbf{x}^{(t)}\| \leq g(\mathbf{x}^{(t)}) + M\|\eta_t\boldsymbol{\zeta}_f^{(t)}\| \leq \epsilon_t + \eta_t M^2.$$

By (60), we have $\text{dist}(\mathbf{x}^{(t+1)}, \mathcal{S}) \leq \epsilon_t/\nu' + \eta_t M^2/\nu' \leq \min\{\epsilon^2/M, \nu'\}$.

Suppose $t \in J$. Let t' be the largest index in I that is smaller than t. By the same proof as in the previous case, we have $\text{dist}(\mathbf{x}^{(t'+1)}, \mathcal{S}) \leq \min\{\epsilon^2/M, \nu'\}$. Since indexes $t' + 1$, $t' + 2$, ..., t are in J, Algorithm 1 essentially performs the projected subgradient method to $\min_{\mathbf{x} \in \mathcal{X}} g_+(\mathbf{x})$ using the Polyak's stepsize (61) during iterations $t' + 1$, $t' + 2$, ..., and t. Hence, by Lemma A.17, we have $\text{dist}(\mathbf{x}^{(t+1)}, \mathcal{S}) \leq \text{dist}(\mathbf{x}^{(t'+1)}, \mathcal{S}) \leq \min\{\epsilon^2/M, \nu'\}$.

By induction, we have prove that $\text{dist}(\mathbf{x}^{(t)}, \mathcal{S}) \leq \min\{\epsilon^2/M, \nu'\}$ for any $t \geq 0$. As a result, $g(\mathbf{x}^{(t)}) \leq M \cdot \text{dist}(\mathbf{x}^{(t)}, \mathcal{S}) \leq M \cdot \min\{\epsilon^2/M, \nu'\} \leq \epsilon^2$, meaning that $\mathbf{x}^{(t)}$ is ϵ^2-feasible for (3). $\square$

Similar to Lemma 4.2, we present an upper bound of $\widehat{\lambda}$ in (5) that is independent of $\mathbf{x}$.

Lemma A.19. *Suppose Assumptions 3.1 and A.16 hold. Given any* $\mathbf{x} \in \mathcal{X}$ *satisfying* $\text{dist}(\mathbf{x}, \mathcal{S}) \leq \nu'$, *let* $\widehat{\mathbf{x}}(\mathbf{x})$ *be defined as in* (4) *with* $(\hat{\rho}, \tilde{\rho})$ *satisfying* (7) *and* $\widehat{\lambda}$ *be the associated Lagrangian multiplier satisfying* (5). *We have*

$$\|\widehat{\mathbf{x}} - \mathbf{x}\| \leq D + \nu' \quad \text{and} \quad \widehat{\lambda} \leq \Lambda'' := \frac{(M + \hat{\rho}(D + \nu'))D}{-g(\mathbf{x}_{\text{feas}})}. \tag{64}$$

Proof. Since $\text{dist}(\mathbf{x}, \mathcal{S}) \leq \nu'$, we get by triangle inequality and Assumption A.16C that

$$\|\widehat{\mathbf{x}} - \mathbf{x}\| \leq \|\widehat{\mathbf{x}} - \mathbf{x}^\dagger\| + \|\mathbf{x}^\dagger - \mathbf{x}\| = \|\widehat{\mathbf{x}} - \mathbf{x}^\dagger\| + \text{dist}(\mathbf{x}, \mathcal{S}) \leq D + \nu'$$

where $\mathbf{x}^\dagger = \text{proj}_\mathcal{S}(\mathbf{x})$, which gives the first result in (64). Following the same steps as in the proof of Lemma 4.2, we obtain

$$\widehat{\lambda} \leq \left(\langle \widehat{\boldsymbol{\zeta}}_f + \hat{\rho}(\widehat{\mathbf{x}} - \mathbf{x}), \mathbf{x}_{\text{feas}} - \widehat{\mathbf{x}} \rangle \right) / (-g(\mathbf{x}_{\text{feas}})). \tag{65}$$

By Assumption 3.1 and A.16C, we have $\|\widehat{\boldsymbol{\zeta}}_f\| \leq M$ and $\|\widehat{\mathbf{x}} - \mathbf{x}_{\text{feas}}\| \leq D$, which imply from the first result in (64) and (65) that

$$\widehat{\lambda} \leq \frac{\langle \widehat{\boldsymbol{\zeta}}_f + \hat{\rho}(\widehat{\mathbf{x}} - \mathbf{x}), \mathbf{x}_{\text{feas}} - \widehat{\mathbf{x}} \rangle}{-g(\mathbf{x}_{\text{feas}})} \leq \frac{(M + \hat{\rho}(D + \nu'))D}{-g(\mathbf{x}_{\text{feas}})}.$$

$\square$

Let $\widehat{\mathbf{x}}^{(t)}$ be $\widehat{\mathbf{x}}(\mathbf{x}^{(t)})$ defined in (4) with $(\hat{\rho}, \tilde{\rho})$ satisfying (7). The following theorem provides the main result in this section.

Theorem A.20. *Under the same assumptions as Proposition A.18, Algorithm 1 guarantees* $\mathbb{E}_\tau[\|\widehat{\mathbf{x}}^{(\tau)} - \mathbf{x}^{(\tau)}\|] \leq C'\epsilon$ *and* $\mathbb{E}_\tau[g(\mathbf{x}^{(\tau)})] \leq \frac{\nu'}{2}\min\{\epsilon^2/M, \nu'\}$, *where* $C' := \sqrt{\frac{\nu'(1+\Lambda'')}{2M(\hat{\rho}-\rho)}}$, *if we use Output I and set*

$$S = 0 \text{ and } T > \frac{8M^3 D^2}{\nu'^3 \min\{\epsilon^4/M^2, \nu'^2\}} = O(1/\epsilon^4).$$

32

This theorem indicates that Algorithm 1 finds a nearly $(C'\epsilon)$-stationary point with complexity $O(1/\epsilon^4)$. To obtain a nearly ϵ-stationary point, one only needs to replace ϵ in η_t, ϵ_t and T in this theorem above by $\epsilon/\max\{C', 1\}$. This will only change the constant factor in the $O(1/\epsilon^4)$ complexity. Once again, this complexity matches the start-of-the-art complexity by [13, 54, 39].

Proof of Theorem A.20. Let $\boldsymbol{\zeta}^{(t)} = \boldsymbol{\zeta}_f^{(t)} \in \partial f(\mathbf{x}^{(t)})$ if $t \in I$ and $\boldsymbol{\zeta}^{(t)} = \boldsymbol{\zeta}_g^{(t)} \in \partial g(\mathbf{x}^{(t)})$ if $t \in J$. Let $\varphi(\mathbf{x})$ and $\widehat{\mathbf{x}}$ be defined in (3) and (4) with $(\hat{\rho}, \tilde{\rho})$ satisfying (7). By Assumption A.16B, problem (3) with $\mathbf{x} = \mathbf{x}^{(t)}$ is strongly convex and has a strictly feasible solution, so $\varphi(\mathbf{x}^{(t)})$ and $\widehat{\mathbf{x}}^{(t)}$ are well defined for any $t \geq 0$. For simplicity of notation, we denote $\widehat{\mathbf{x}}(\mathbf{x}^{(t)})$ by $\widehat{\mathbf{x}}^{(t)}$ and let $\widehat{\lambda}_t \geq 0$ be Lagrangian multiplier satisfying (5).

By following the proof of Proposition A.4 using deterministic subgradients and setting $\mu = 0$, we have

$$\sum_{t=S}^{T-1} \left[\eta_t \hat{\rho}(\hat{\rho} - \rho) \|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\|^2 - \eta_t \hat{\rho}\widehat{\lambda}_t g(\mathbf{x}^{(t)}) \right] \mathbb{I}(g(\mathbf{x}^{(t)}) \leq \epsilon_t) + \sum_{t=S}^{T-1} \eta_t \hat{\rho} g(\mathbf{x}^{(t)}) \mathbb{I}(g(\mathbf{x}^{(t)}) > \epsilon_t)$$

$$\leq \varphi(\mathbf{x}^{(S)}) - \varphi(\mathbf{x}^{(T)}) + \frac{\hat{\rho}}{2} \sum_{t=S}^{T-1} \left[\eta_t^2 \|\boldsymbol{\zeta}_f^{(t)}\|^2 \mathbb{I}(g(\mathbf{x}^{(t)}) \leq \epsilon_t) + \eta_t^2 \|\boldsymbol{\zeta}_g^{(t)}\|^2 \mathbb{I}(g(\mathbf{x}^{(t)}) > \epsilon_t) \right]. \tag{66}$$

Since $S = 0$ and $\mathbf{x}^{(0)} \in \mathcal{S}$, we then have $\varphi(\mathbf{x}^{(T)}) = f(\widehat{\mathbf{x}}^{(T)}) + \frac{\hat{\rho}}{2}\|\widehat{\mathbf{x}}^{(T)} - \mathbf{x}^{(T)}\|^2 \geq f(\widehat{\mathbf{x}}^{(T)})$ and $\varphi(\mathbf{x}^{(S)}) = f(\widehat{\mathbf{x}}^{(S)}) + \frac{\hat{\rho}}{2}\|\widehat{\mathbf{x}}^{(S)} - \mathbf{x}^{(S)}\|^2 \leq f(\widehat{\mathbf{x}}^{(T)}) + \frac{\hat{\rho}}{2}\|\widehat{\mathbf{x}}^{(T)} - \mathbf{x}^{(S)}\|^2 \leq f(\widehat{\mathbf{x}}^{(T)}) + \frac{\hat{\rho}D^2}{2}$. Note that D here is defined as in Assumption A.16 instead of Assumption 4.1. Applying these bounds to (66), we have

$$\sum_{t=S}^{T-1} \left[\eta_t \hat{\rho}(\hat{\rho} - \rho) \|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\|^2 - \eta_t \hat{\rho}\widehat{\lambda}_t g(\mathbf{x}^{(t)}) \right] \mathbb{I}(g(\mathbf{x}^{(t)}) \leq \epsilon_t) + \sum_{t=S}^{T-1} \eta_t \hat{\rho} g(\mathbf{x}^{(t)}) \mathbb{I}(g(\mathbf{x}^{(t)}) > \epsilon_t)$$

$$\leq \frac{\hat{\rho}D^2}{2} + \frac{\hat{\rho}}{2} \sum_{t=S}^{T-1} \left[\eta_t^2 \|\boldsymbol{\zeta}_f^{(t)}\|^2 \mathbb{I}(g(\mathbf{x}^{(t)}) \leq \epsilon_t) + \eta_t^2 \|\boldsymbol{\zeta}_g^{(t)}\|^2 \mathbb{I}(g(\mathbf{x}^{(t)}) > \epsilon_t) \right]. \tag{67}$$

For $t \in I$, we have from Assumption 3.1 and the definition of η_t that

$$\eta_t^2 \|\boldsymbol{\zeta}_f^{(t)}\|^2 \leq \eta_t^2 M^2 = \eta_t M^2 \frac{\nu'}{2M^2} \min\left\{\epsilon^2/M, \nu'\right\} = \eta_t \epsilon_t. \tag{68}$$

For $t \in J$, we have $\eta_t = \nu' g_+(\mathbf{x}^{(t)})/(2M\|\boldsymbol{\zeta}_g^{(t)}\|^2)$ so

$$\eta_t^2 \|\boldsymbol{\zeta}_g^{(t)}\|^2 = \eta_t \nu' g_+(\mathbf{x}^{(t)})/(2M) \leq \eta_t \nu' \mathrm{dist}(\mathbf{x}^{(t)}, \mathcal{S})/2 \leq \eta_t \epsilon_t, \tag{69}$$

where the first inequality is by M-Lipschitz continuity of g_+ and the last inequality is from Proposition A.18.

We first prove that, if S, T, η_t and ϵ_t are chosen such that

$$\epsilon_t(1 + \widehat{\lambda}_t) \leq C'^2 \epsilon^2 (\hat{\rho} - \rho) \tag{70}$$

and

$$\sum_{t=S}^{T-1} \eta_t \hat{\rho}\epsilon_t > \frac{\hat{\rho}D^2}{2} + \frac{\hat{\rho}}{2} \sum_{t=S}^{T-1} \left[\eta_t^2 \|\boldsymbol{\zeta}_f^{(t)}\|^2 \mathbb{I}(g(\mathbf{x}^{(t)}) \leq \epsilon_t) + \eta_t^2 \|\boldsymbol{\zeta}_g^{(t)}\|^2 \mathbb{I}(g(\mathbf{x}^{(t)}) > \epsilon_t) \right], \tag{71}$$

we must have $g(\mathbf{x}^{(t)}) \leq \epsilon_t$ for at least one t in $\{S, S+1, \ldots, T-1\}$ (i.e., $I \neq \emptyset$) and $\mathbb{E}_\tau[\|\widehat{\mathbf{x}}^{(\tau)} - \mathbf{x}^{(\tau)}\|^2] \leq C'^2 \epsilon^2$ (so $\mathbb{E}_\tau[\|\widehat{\mathbf{x}}^{(\tau)} - \mathbf{x}^{(\tau)}\|] \leq C'\epsilon$).

Suppose (71) holds and $g(\mathbf{x}^{(t)}) > \epsilon_t$ for $t = S, S+1, \ldots, T-1$, i.e., $I = \emptyset$. (67) becomes exactly the opposite of (71). This contradiction means $g(\mathbf{x}^{(t)}) \leq \epsilon_t$ for at least one t in $\{S, S+1, \ldots, T-1\}$. Suppose (70) and (71) hold but $\mathbb{E}_\tau[\|\widehat{\mathbf{x}}^{(\tau)} - \mathbf{x}^{(\tau)}\|^2] > C'^2 \epsilon^2$. Since τ is generated by Output I, we have

$$C'^2 \epsilon^2 < \mathbb{E}_\tau[\|\widehat{\mathbf{x}}^{(\tau)} - \mathbf{x}^{(\tau)}\|^2] = \frac{\sum_{t=S}^{T-1} \eta_t \mathbb{I}(g(\mathbf{x}^{(t)}) \leq \epsilon_t)\|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\|^2}{\sum_{t=S}^{T-1} \eta_t \mathbb{I}(g(\mathbf{x}^{(t)}) \leq \epsilon_t)}. \tag{72}$$

33

(72) and (70) imply

$$
\sum_{t=S}^{T-1} \left[\eta_t \hat{\rho}(\hat{\rho} - \rho) \|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\|^2 - \eta_t \hat{\rho} \widehat{\lambda}_t \epsilon_t \right] \mathbb{I}(g(\mathbf{x}^{(t)}) \le \epsilon_t) + \sum_{t=S}^{T-1} \eta_t \hat{\rho} \epsilon_t \mathbb{I}(g(\mathbf{x}^{(t)}) > \epsilon_t)
$$

$$
> \sum_{t=S}^{T-1} \left[\eta_t \hat{\rho}(\hat{\rho} - \rho) C'^2 \epsilon^2 - \eta_t \hat{\rho} \widehat{\lambda}_t \epsilon_t \right] \mathbb{I}(g(\mathbf{x}^{(t)}) \le \epsilon_t) + \sum_{t=S}^{T-1} \eta_t \hat{\rho} \epsilon_t \mathbb{I}(g(\mathbf{x}^{(t)}) > \epsilon_t)
$$

$$
\ge \sum_{t=S}^{T-1} \eta_t \hat{\rho} \epsilon_t \mathbb{I}(g(\mathbf{x}^{(t)}) \le \epsilon_t) + \sum_{t=S}^{T-1} \eta_t \hat{\rho} \epsilon_t \mathbb{I}(g(\mathbf{x}^{(t)}) > \epsilon_t) = \sum_{t=S}^{T-1} \eta_t \hat{\rho} \epsilon_t, \tag{73}
$$

where the second inequality is because of (70). Combining this inequality and (67) leads to the opposite of (71). This contradiction means $\mathbb{E}_\tau [\|\widehat{\mathbf{x}}^{(\tau)} - \mathbf{x}^{(\tau)}\|^2] \le C'^2 \epsilon^2$.

Given the result above, we only need to show that the the choices of S, T, η_t and ϵ_t ensure (70) and (71).

Plugging in the value of ϵ_t in Proposition A.18 and using the second result in Lemma A.19, we have

$$
\epsilon_t(1 + \widehat{\lambda}_t) \le \epsilon_t(1 + \Lambda'') \le \frac{\nu'(1 + \Lambda'')\epsilon^2}{2M}
$$

which implies (70) by the definition of C'. According to the definition of η_t for $t \in I$ and $t \in J$, we have

$$
\eta_t \ge \frac{\nu'}{2M} \eta_t = \frac{\nu'^2}{4M^3} \min\left\{ \epsilon^2/M, \nu' \right\}
$$

for $t \in I$ by noting that $\nu' \le M$ and

$$
\eta_t \ge \frac{\nu'}{2M} \cdot \frac{\epsilon_t}{M^2} = \frac{\nu'^2}{4M^3} \min\left\{ \epsilon^2/M, \nu' \right\}
$$

for $t \in J$. By the choice of T, we have

$$
\frac{\hat{\rho}}{2} \sum_{t=S}^{T-1} \eta_t \epsilon_t \ge \frac{\hat{\rho} T}{2} \frac{\nu'^3}{8M^3} \min\left\{ \epsilon^4/M^2, \nu'^2 \right\} > \frac{\hat{\rho} D^2}{2}
$$

which further implies (71) together with (68) and (69). Thus, we have $\mathbb{E}_\tau [\|\widehat{\mathbf{x}}^{(\tau)} - \mathbf{x}^{(\tau)}\|] \le C' \epsilon$. The inequality $\mathbb{E}_\tau [g(\mathbf{x}^{(\tau)})] \le \frac{\nu'}{2} \min\left\{ \epsilon^2/M, \nu' \right\}$ is a direct consequence of $g(\mathbf{x}^{(t)}) \le \epsilon_t$ for any $t \in I$ and the choice of ϵ_t. $\qquad \square$

B Convergence analysis for weakly convex constraints

In this section, we present two non-trivial practical examples that satisfy Assumption 5.1 and then analyze the complexity of Algorithm 1 when f and g are both weakly convex.

B.1 Examples that satisfy Assumption 5.1

In this section, we present two examples that satisfy Assumption 5.1, especially, Assumption 5.1B.

B.1.1 Demographic parity constraint

Problem (17) is an instance of (1) with

$$
\mathcal{X} = \mathbb{R}^d, \ f(\mathbf{x}) = L(\mathbf{x}) + \lambda \mathrm{SCAD}(\mathbf{x}), \ g(\mathbf{x}) = R_0(\mathbf{x}) - \kappa,
$$

where L, R_0 and SCAD are defined in (13), (16) and (18), respectively. It is easy to verify that problem (17) satisfies Assumption 3.1. In this section, we will prove that problem (17) also satisfies Assumption 5.1.

Since $L(\mathbf{x})$ is convex and $\mathrm{SCAD}(\mathbf{x})$ is 2-weakly convex, $f(\mathbf{x})$ is 2λ-weakly convex. Let

$$
h(\mathbf{x}) = \frac{1}{n_p} \sum_{i=1}^{n_p} \sigma(\mathbf{x}^\top \mathbf{a}_i^p) - \frac{1}{n_u} \sum_{i=1}^{n_u} \sigma(\mathbf{x}^\top \mathbf{a}_i^u) \tag{74}
$$

34

so $g(\mathbf{x}) = |h(\mathbf{x})| - \kappa$. Recalling that $\sigma(z) = \exp(z)/(1 + \exp(z))$, we have

$$\nabla h(\mathbf{x}) = \frac{1}{n_p} \sum_{i=1}^{n_p} \sigma(\mathbf{x}^\top \mathbf{a}_i^p)(1 - \sigma(\mathbf{x}^\top \mathbf{a}_i^p))\mathbf{a}_i^p - \frac{1}{n_u} \sum_{i=1}^{n_u} \sigma(\mathbf{x}^\top \mathbf{a}_i^u)(1 - \sigma(\mathbf{x}^\top \mathbf{a}_i^u))\mathbf{a}_i^u. \tag{75}$$

It is easy to prove that $h(\mathbf{x})$ is Lipschitz continuous with a constant of

$$\alpha = \frac{1}{4n_p} \sum_{i=1}^{n_p} \|\mathbf{a}_i^p\| + \frac{1}{4n_u} \sum_{i=1}^{n_u} \|\mathbf{a}_i^u\| \tag{76}$$

and that $\nabla h(\mathbf{x})$ is Lipschitz continuous with a constant of

$$\beta = \frac{1}{4n_p} \sum_{i=1}^{n_p} \|\mathbf{a}_i^p\|^2 + \frac{1}{4n_u} \sum_{i=1}^{n_u} \|\mathbf{a}_i^u\|^2, \tag{77}$$

so g is β-weakly convex according to Lemma 4.2 in [32]. Hence, (17) satisfies Assumption 5.1A with $\rho = \max\{2\lambda, \beta\}$.

Recall $\mathcal{D}_p = \{\mathbf{a}_i^p\}_{i=1}^{n_p}$ and $\mathcal{D}_u = \{\mathbf{a}_i^u\}_{i=1}^{n_u}$ are the feature vectors of protected and unprotected groups, respectively. We assume without generality that the first feature of $\mathbf{a}_i^p$, denoted by $[\mathbf{a}_i^p]_1$, equals one and the first feature of $\mathbf{a}_i^u$, denoted by $[\mathbf{a}_i^u]_1$, equals negative one. In fact, this feature can be the one used to split data into the protected group or the unprotected group (see the group variables in Table 1 for example). We can also simply add this feature to the data if it does not exist originally. With this feature, we can show that (17) satisfies Assumption 5.1B under mild conditions.

Theorem B.1. *Suppose* $[\mathbf{a}_i^p]_1 = 1$ *for* $i = 1, \dots, n_p$ *and* $[\mathbf{a}_i^u]_1 = -1$ *for* $i = 1, \dots, n_u$ *and*

$$\kappa \notin \mathcal{K} = \left\{ \left| \frac{i}{n_p} - \frac{j}{n_u} \right| \ \middle| \ i = 0, 1, \dots, n_p, \ j = 0, 1, \dots, n_u \right\}.$$

The following statements hold.

1. *There exist* $\overline{\kappa}$ *and* $\underline{\kappa}$ *in* $\mathcal{K}$ *such that* $(\underline{\kappa}, \overline{\kappa})$ *contains* κ *but no numbers in* $\mathcal{K}$.

2. *It holds that*

$$\underline{q} := \inf \left\{ q(\mathbf{x}) \ \middle| \ |h(\mathbf{x})| \in [(\kappa + \underline{\kappa})/2, (\kappa + \overline{\kappa})/2] \right\} > 0, \tag{78}$$

   where h is defined in (74) and

$$q(\mathbf{x}) = \max \left\{ \max_{i=1,\dots,n_p} \frac{\sigma(\mathbf{x}^\top \mathbf{a}_i^p)(1 - \sigma(\mathbf{x}^\top \mathbf{a}_i^p))}{n_p}, \max_{i=1,\dots,n_u} \frac{\sigma(\mathbf{x}^\top \mathbf{a}_i^u)(1 - \sigma(\mathbf{x}^\top \mathbf{a}_i^u))}{n_u} \right\}. \tag{79}$$

3. *Problem* (17) *satisfies Assumption 5.1B with*

$$\theta := \min \left\{ \frac{\kappa - \underline{\kappa}}{2}, \frac{\underline{q}^2}{4(\beta + \bar{\rho})}, \frac{\underline{q}^2 \kappa}{16\alpha^2} \right\}$$

   for any $\bar{\rho} > \rho$ and any $\bar{\epsilon} > 0$ that satisfy

$$\bar{\epsilon}^2 \leq \min \left\{ \frac{\overline{\kappa} - \kappa}{2}, \frac{\underline{q}^2}{4(\beta + \bar{\rho})}, \frac{\underline{q}^2 \kappa}{16\alpha^2} \right\}. \tag{80}$$

Proof. Since $\kappa \notin \mathcal{K}$ and $\mathcal{K}$ is discrete, Statement 1 holds trivially. In fact, one can sort the numbers in $\mathcal{K}$ and let $\overline{\kappa}$ and $\underline{\kappa}$ be the two consecutive numbers with $\kappa \in (\underline{\kappa}, \overline{\kappa})$.

Suppose $\underline{q} = 0$. There must exist a sequence $\{\mathbf{x}^{(t)}\}_{t \geq 0}$ such that $|h(\mathbf{x}^{(t)})| \in [(\kappa + \underline{\kappa})/2, (\kappa + \overline{\kappa})/2]$ for any $t \geq 0$ and $\lim_{t \to \infty} q(\mathbf{x}^{(t)}) = 0$. Since $\sigma(z) \in (0, 1)$ for any $z \in \mathbb{R}$, any sequence like $\sigma(z^{(t)})(1 - \sigma(z^{(t)}))$ will converge to zero only when all limiting points of $\sigma(z^{(t)})$ are in $\{0, 1\}$. By this observation and the definition of q in (79), after passing to a subsequence if necessary, we have

$$\lim_{t \to \infty} \sigma(\mathbf{x}^{(t)\top} \mathbf{a}_i^p) \in \{0, 1\}, i = 1, \dots, n_p \text{ and } \lim_{t \to \infty} \sigma(\mathbf{x}^{(t)\top} \mathbf{a}_i^u) \in \{0, 1\}, i = 1, \dots, n_u. \tag{81}$$

This means $\lim_{t \to \infty} |h(\mathbf{x}^{(t)})| \in \mathcal{K}$, contradicting with Statement 1 and the fact that $|h(\mathbf{x}^{(t)})| \in [(\kappa + \underline{\kappa})/2, (\kappa + \overline{\kappa})/2] \subset (\underline{\kappa}, \overline{\kappa})$. This contradiction indicates that $\underline{q} > 0$, which proves Statement 2.

35

Consider any $\bar{\rho} > \rho$ and any $\bar{\epsilon} > 0$ that satisfy (80). Because $[\mathbf{a}_i^p]_1 = 1$ for $i = 1, \ldots, n_p$ and $[\mathbf{a}_i^u]_1 = -1$ for $i = 1, \ldots, n_u$, it is easy to show that

$$\|\nabla h(\mathbf{x})\| \geq \left| \frac{1}{n_p} \sum_{i=1}^{n_p} \sigma(\mathbf{x}^\top \mathbf{a}_i^p)(1 - \sigma(\mathbf{x}^\top \mathbf{a}_i^p))[\mathbf{a}_i^p]_1 - \frac{1}{n_u} \sum_{i=1}^{n_u} \sigma(\mathbf{x}^\top \mathbf{a}_i^u)(1 - \sigma(\mathbf{x}^\top \mathbf{a}_i^u))[\mathbf{a}_i^u]_1 \right| \geq q(\mathbf{x}),$$

where ∇h is given in (75). As a result, it holds that

$$\inf \left\{ \|\nabla h(\mathbf{x})\| \mid g(\mathbf{x}) \in [(\underline{\kappa} - \kappa)/2, \bar{\epsilon}^2] \right\} \geq \inf \left\{ q(\mathbf{x}) \mid g(\mathbf{x}) \in [(\underline{\kappa} - \kappa)/2, \bar{\epsilon}^2] \right\}$$
$$= \inf \left\{ q(\mathbf{x}) \mid |h(\mathbf{x})| \in [(\kappa + \underline{\kappa})/2, \kappa + \bar{\epsilon}^2] \right\} \geq \inf \left\{ q(\mathbf{x}) \mid |h(\mathbf{x})| \in [(\kappa + \underline{\kappa})/2, (\kappa + \overline{\kappa})/2] \right\}$$
$$= \underline{q} > 0, \tag{82}$$

where the second inequality is because $\bar{\epsilon}^2 \leq (\overline{\kappa} - \kappa)/2$ so $\kappa + \bar{\epsilon}^2 \leq (\kappa + \overline{\kappa})/2$.

Consider a solution $\mathbf{x}$ with $g(\mathbf{x}) \leq \bar{\epsilon}^2$. Suppose $g(\mathbf{x}) \leq -(\kappa - \underline{\kappa})/2$. We can set $\mathbf{y} = \mathbf{x}$ and have

$$g(\mathbf{y}) + \frac{\bar{\rho}}{2}\|\mathbf{y} - \mathbf{x}\|^2 = g(\mathbf{x}) \leq -\frac{\kappa - \underline{\kappa}}{2} \leq -\theta. \tag{83}$$

Suppose $g(\mathbf{x}) \geq -(\kappa - \underline{\kappa})/2$ so $|h(\mathbf{x})| \geq (\underline{\kappa} + \kappa)/2 > 0$. We assume $h(\mathbf{x}) > 0$ and the proof when $h(\mathbf{x}) < 0$ is the same. Let $\mathbf{y}$ be generated from $\mathbf{x}$ by running a gradient descent step on $h(\mathbf{x})$ with a stepsize of $\eta > 0$, namely, $\mathbf{y} = \mathbf{x} - \eta \nabla h(\mathbf{x})$. We choose $\eta = \min\{1/(\beta + \bar{\rho}), \kappa/(4\alpha^2)\}$. Since $h(\mathbf{y}) + \frac{\bar{\rho}}{2}\|\mathbf{y} - \mathbf{x}\|^2$ is smooth and its gradient with respect to $\mathbf{y}$ is $(\beta + \bar{\rho})$-Lipschitz continuous, we have

$$h(\mathbf{y}) + \frac{\bar{\rho}}{2}\|\mathbf{y} - \mathbf{x}\|^2 \leq h(\mathbf{x}) + \frac{\bar{\rho}}{2}\|\mathbf{x} - \mathbf{x}\|^2 + \langle \nabla h(\mathbf{x}) + \bar{\rho}(\mathbf{x} - \mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\beta + \bar{\rho}}{2}\|\mathbf{y} - \mathbf{x}\|^2$$
$$\leq \kappa + \bar{\epsilon}^2 - \eta \left(1 - \frac{\beta + \bar{\rho}}{2}\eta \right) \|\nabla h(\mathbf{x})\|^2$$
$$\leq \kappa + \bar{\epsilon}^2 - \frac{1}{2} \min \left\{ \frac{1}{\beta + \bar{\rho}}, \frac{\kappa}{4\alpha^2} \right\} \|\nabla h(\mathbf{x})\|^2$$
$$\leq \kappa + \bar{\epsilon}^2 - \frac{\underline{q}^2}{2} \min \left\{ \frac{1}{\beta + \bar{\rho}}, \frac{\kappa}{4\alpha^2} \right\}, \tag{84}$$

where the second inequality is by the definition of $\mathbf{y}$ and the fact that $h(\mathbf{x}) = |h(\mathbf{x})| = g(\mathbf{x}) + \kappa \leq \kappa + \bar{\epsilon}^2$, the third inequality is because $\eta = \min\{1/(\beta + \bar{\rho}), \kappa/(4\alpha^2)\}$, and the last is by (82) and the fact that $g(\mathbf{x}) \in [(\underline{\kappa} - \kappa)/2, \bar{\epsilon}^2]$. Moreover, since $h(\mathbf{x}) = |h(\mathbf{x})| = g(\mathbf{x}) + \kappa \geq (\underline{\kappa} + \kappa)/2$, $\eta \leq \kappa/(4\alpha^2)$ and h is α-Lipschitz continuous, we have

$$h(\mathbf{y}) \geq h(\mathbf{x}) - \alpha\eta\|\nabla h(\mathbf{x})\| \geq \frac{\underline{\kappa} + \kappa}{2} - \eta\alpha^2 \geq \frac{\kappa}{4} > 0,$$

which indicates $g(\mathbf{y}) = |h(\mathbf{y})| - \kappa = h(\mathbf{y}) - \kappa$. Hence, (84) and (80) imply

$$g(\mathbf{y}) + \frac{\bar{\rho}}{2}\|\mathbf{y} - \mathbf{x}\|^2 \leq \bar{\epsilon}^2 - \frac{\underline{q}^2}{2} \min \left\{ \frac{1}{\beta + \bar{\rho}}, \frac{\kappa}{4\alpha^2} \right\} \leq -\frac{\underline{q}^2}{4} \min \left\{ \frac{1}{\beta + \bar{\rho}}, \frac{\kappa}{4\alpha^2} \right\} \leq -\theta. \tag{85}$$

Inequalities (83) and (85) mean (17) satisfies Assumption 5.1B, which proves Statement 3. □

B.1.2 Smoothly clipped absolute deviation constraint

Suppose $f(\mathbf{x})$ in (1) is any training loss that satisfies Assumptions 3.1, 5.1A and 5.1C. We consider a SCAD constraint for promoting sparsity of model $\mathbf{x}$. It is formulated as

$$g(\mathbf{x}) = \text{SCAD}(\mathbf{x}) - \kappa = \sum_{i=1}^d s(x_i) - \kappa \leq 0, \tag{86}$$

where $s(x_i)$ is defined in (18). Jia and Grimmer [39] show that, if κ is not divisible by three, constraint (86) satisfies the strong Mangasarian-Fromovitz constraint qualification (MFCQ), which is used in [39] to bound the Lagrangian multiplier in a way similar to (10). Motivated by the finding in [39], we show that (86) also satisfies the uniform Slater's condition in Assumption 5.1B.

Proposition B.2. *Suppose κ is not divisible by three. The following statements hold.*

1. *There exists a non-negative integer k such that $\kappa \in (3k, 3(k+1))$.*

36

2. *It holds that*

$$\inf\left\{\max_i \operatorname{dist}(x_i, \mathcal{Z}) \,\middle|\, \operatorname{SCAD}(\mathbf{x}) \in [(\kappa + 3k)/2, (\kappa + 3(k+1))/2]\right\} =: \underline{q} > 0, \quad (87)$$

where $\mathcal{Z} = (-\infty, -2] \cup \{0\} \cup [2, \infty)$.

3. *Constraint* (86) *satisfies Assumption 5.1B with*

$$\theta := \min\left\{\frac{\kappa - 3k}{2}, \frac{\underline{q}^2}{2 + \bar{\rho}}, \frac{\underline{q}^3}{2}\right\}$$

for any $\bar{\rho} > \rho$ *and any* $\bar{\epsilon} > 0$ *that satisfy*

$$\bar{\epsilon}^2 \le \min\left\{\frac{3(k+1) - \kappa}{2}, \frac{\underline{q}^2}{2 + \bar{\rho}}, \frac{\underline{q}^3}{2}\right\}. \quad (88)$$

Proof. Since κ is not divisible by three, Statement 1 holds trivially. In addition, by (88), we also have

$$3k < \frac{\kappa + 3k}{2} < \kappa < \kappa + \bar{\epsilon}^2 \le \frac{\kappa + 3(k+1)}{2} < 3(k+1). \quad (89)$$

As observed by [39], $0 \in \partial s(x_i)$ only when $x_i \in \mathcal{Z}$, or equivalently, $s(x_i) \in \{0, 3\}$ for any i. Suppose (87) is not true. There must exist a sequence $\{\mathbf{x}^{(t)}\}_{t \ge 0}$ such that $\operatorname{SCAD}(\mathbf{x}^{(t)}) \in [(\kappa + 3k)/2, (\kappa + 3(k+1))/2]$ for any $t \ge 0$ and $\lim_{t\to\infty} \operatorname{dist}(x_i^{(t)}, \mathcal{Z}) = 0$ for any i. Therefore, by passing to a subsequence if necessary, we have $\lim_{t\to\infty} s(x_i^{(t)}) \in \{0, 3\}$ for $i = 1, \ldots, d$, so $\lim_{t\to\infty} \operatorname{SCAD}(\mathbf{x}^{(t)})$ is divisible by three, contradicting with (89) and the fact that $\operatorname{SCAD}(\mathbf{x}^{(t)}) \in [(\kappa + 3k)/2, (\kappa + 3(k+1))/2]$. This contradiction proves Statement 2.

Consider a solution $\mathbf{x}$ with $g(\mathbf{x}) \le \bar{\epsilon}^2$. Suppose $g(\mathbf{x}) \le -(\kappa - 3k)/2$. We can set $\mathbf{y} = \mathbf{x}$ and have

$$g(\mathbf{y}) + \frac{\bar{\rho}}{2}\|\mathbf{y} - \mathbf{x}\|^2 = g(\mathbf{x}) \le -\frac{\kappa - 3k}{2} \le -\theta. \quad (90)$$

Suppose $g(\mathbf{x}) \ge -(\kappa - 3k)/2$ so $\operatorname{SCAD}(\mathbf{x}) \in [(\kappa + 3k)/2, (\kappa + 3(k+1))/2]$. By (87), there exists an index i such that $\operatorname{dist}(x_i, \mathcal{Z}) \ge \underline{q}$, which implies $x_i \in [\underline{q}, 2 - \underline{q}]$ or $x_i \in [-2 + \underline{q}, -\underline{q}]$. Without loss of generality, we assume $i = 1$ and $x_i \in [\underline{q}, 2 - \underline{q}]$ since the proof when $x_1 \in [-2 + \underline{q}, -\underline{q}]$ is the same. By the definition of s in (18), we have $s'(x_1) \in [2\underline{q}, 1]$. Let $y_i = x_i$ for $i \ne 1$ and $y_1 = x_1 - \eta s'(x_1)$ with $\eta = \min\{\underline{q}/2, 1/(2 + \bar{\rho})\}$. We then have $y_1 \ge x_1 - \underline{q}/2 \ge \underline{q}/2$. Since the function $s(y_1) + \frac{\bar{\rho}}{2}(y_1 - x_1)^2$ is smooth on $[\underline{q}/2, x_1]$ and its gradient with respect to y_1 is $(2 + \bar{\rho})$-Lipschitz continuous, we have

$$\begin{aligned}
g(\mathbf{y}) + \frac{\bar{\rho}}{2}\|\mathbf{y} - \mathbf{x}\|^2 &\le g(\mathbf{x}) + s'(x_1) \cdot (y_1 - x_1) + \frac{2 + \bar{\rho}}{2}(y_1 - x_1)^2 \\
&\le \bar{\epsilon}^2 - \eta\left(1 - \frac{2 + \bar{\rho}}{2}\eta\right)[s'(x_1)]^2 \\
&\le \bar{\epsilon}^2 - \frac{1}{2}\min\left\{\frac{\underline{q}}{2}, \frac{1}{2 + \bar{\rho}}\right\}[s'(x_1)]^2 \\
&\le \bar{\epsilon}^2 - 2\min\left\{\frac{\underline{q}}{2}, \frac{1}{2 + \bar{\rho}}\right\}\underline{q}^2 \le -\min\left\{\frac{\underline{q}}{2}, \frac{1}{2 + \bar{\rho}}\right\}\underline{q}^2 \le -\theta, \quad (91)
\end{aligned}$$

where the second inequality is by the definition of y_1 and the fact $g(\mathbf{x}) \le \bar{\epsilon}^2$, the third inequality is because $\eta = \min\{\underline{q}/2, 1/(2 + \bar{\rho})\}$, the fourth is because $s'(x_1) \in [2\underline{q}, 1]$, the fifth is because $\bar{\epsilon}^2 \le \min\{\underline{q}/2, 1/(2 + \bar{\rho})\}\underline{q}^2$, and the last by the definition of θ. Inequalities (90) and (91) mean (86) satisfies Assumption 5.1B, which proves Statement 3. $\square$

B.2 Proofs of Lemmas 5.2, 5.3 and 5.4

Proof of Lemma 5.2. For simplicity of notation, we denote $\hat{\mathbf{x}}(\mathbf{x})$ in (4) by $\hat{\mathbf{x}}$. Suppose $\mathbf{x}$ is ϵ^2-feasible. According to Assumption 5.1B, there exists $\mathbf{y} \in \operatorname{relint}(\mathcal{X})$ such that

$$g(\mathbf{y}) + \frac{\hat{\rho}}{2}\|\mathbf{y} - \mathbf{x}\|^2 \le g(\mathbf{y}) + \frac{\bar{\rho}}{2}\|\mathbf{y} - \mathbf{x}\|^2 \le -\theta. \quad (92)$$

37

This means (3) satisfies the Slater's condition, so there exists $\widehat{\lambda} \geq 0$ that satisfies the KKT condition together with $\widehat{\mathbf{x}}$. In particular, we have $\widehat{\lambda}(g(\widehat{\mathbf{x}}) + \frac{\hat{\rho}}{2}\|\widehat{\mathbf{x}} - \mathbf{x}\|^2) = 0$ and

$$\widehat{\boldsymbol{\zeta}}_f + \hat{\rho}(\widehat{\mathbf{x}} - \mathbf{x}) + \widehat{\lambda}(\widehat{\boldsymbol{\zeta}}_g + \hat{\rho}(\widehat{\mathbf{x}} - \mathbf{x})) + \widehat{\mathbf{u}} = \mathbf{0}, \tag{93}$$

where $\widehat{\boldsymbol{\zeta}}_f \in \partial f(\widehat{\mathbf{x}})$, $\widehat{\boldsymbol{\zeta}}_g \in \partial g(\widehat{\mathbf{x}})$, $\widehat{\mathbf{u}} \in \mathcal{N}_\mathcal{X}(\widehat{\mathbf{x}})$ and $\mathcal{N}_\mathcal{X}(\widehat{\mathbf{x}})$ is the normal cone of $\mathcal{X}$ at $\widehat{\mathbf{x}}$. Taking inner product between (93) and $\widehat{\mathbf{x}} - \mathbf{x}$ gives

$$0 \geq -\langle \widehat{\mathbf{u}}, \widehat{\mathbf{x}} - \mathbf{x} \rangle = \left\langle \widehat{\boldsymbol{\zeta}}_f + \widehat{\lambda}\widehat{\boldsymbol{\zeta}}_g, \widehat{\mathbf{x}} - \mathbf{x} \right\rangle + \hat{\rho}(1 + \widehat{\lambda})\|\widehat{\mathbf{x}} - \mathbf{x}\|^2$$

$$\geq -\frac{\|\widehat{\boldsymbol{\zeta}}_f + \widehat{\lambda}\widehat{\boldsymbol{\zeta}}_g\|^2}{2\hat{\rho}(1 + \widehat{\lambda})} - \frac{\hat{\rho}(1 + \widehat{\lambda})}{2}\|\widehat{\mathbf{x}} - \mathbf{x}\|^2 + \hat{\rho}(1 + \widehat{\lambda})\|\widehat{\mathbf{x}} - \mathbf{x}\|^2,$$

where the first inequality is because $\widehat{\mathbf{u}} \in \mathcal{N}_\mathcal{X}(\widehat{\mathbf{x}})$, the second inequality is by Young's inequality. Reorganizing the terms in this inequality and using the facts that $\|\widehat{\boldsymbol{\zeta}}_f\| \leq M$ and $\|\widehat{\boldsymbol{\zeta}}_g\| \leq M$, we obtain

$$\hat{\rho}(1 + \widehat{\lambda})\|\widehat{\mathbf{x}} - \mathbf{x}\|^2 \leq \frac{\|\widehat{\boldsymbol{\zeta}}_f + \widehat{\lambda}\widehat{\boldsymbol{\zeta}}_g\|^2}{\hat{\rho}(1 + \widehat{\lambda})} \leq \frac{(1 + \widehat{\lambda})M^2}{\hat{\rho}},$$

which further implies the first inequality in (10).

If $\widehat{\lambda} = 0$, the conclusion holds trivially. Hence, we focus on the case that $\widehat{\lambda} > 0$. Note that, in this case, we must have $g(\widehat{\mathbf{x}}) + \frac{\hat{\rho}}{2}\|\widehat{\mathbf{x}} - \mathbf{x}\|^2 = 0$.

Since $g(\mathbf{z}) + \frac{\hat{\rho}}{2}\|\mathbf{z} - \mathbf{x}\|^2 + \delta_\mathcal{X}(\mathbf{z})$ is $(\hat{\rho} - \rho)$-strongly convex in $\mathbf{z}$ and $\widehat{\mathbf{u}}/\widehat{\lambda} \in \mathcal{N}_\mathcal{X}(\widehat{\mathbf{x}}) = \partial\delta_\mathcal{X}(\widehat{\mathbf{x}})$, we have

$$g(\mathbf{y}) + \frac{\hat{\rho}}{2}\|\mathbf{y} - \mathbf{x}\|^2$$

$$\geq g(\widehat{\mathbf{x}}) + \frac{\hat{\rho}}{2}\|\widehat{\mathbf{x}} - \mathbf{x}\|^2 + \left\langle \widehat{\boldsymbol{\zeta}}_g + \hat{\rho}(\widehat{\mathbf{x}} - \mathbf{x}) + \widehat{\mathbf{u}}/\widehat{\lambda}, \mathbf{y} - \widehat{\mathbf{x}} \right\rangle + \frac{\hat{\rho} - \rho}{2}\|\mathbf{y} - \widehat{\mathbf{x}}\|^2$$

$$= \left\langle \widehat{\boldsymbol{\zeta}}_g + \hat{\rho}(\widehat{\mathbf{x}} - \mathbf{x}) + \widehat{\mathbf{u}}/\widehat{\lambda}, \mathbf{y} - \widehat{\mathbf{x}} \right\rangle + \frac{\hat{\rho} - \rho}{2}\|\mathbf{y} - \widehat{\mathbf{x}}\|^2.$$

Applying (92) to the inequality above and arranging terms give

$$-\theta \geq \left\langle \widehat{\boldsymbol{\zeta}}_g + \hat{\rho}(\widehat{\mathbf{x}} - \mathbf{x}) + \widehat{\mathbf{u}}/\widehat{\lambda}, \mathbf{y} - \widehat{\mathbf{x}} \right\rangle + \frac{\hat{\rho} - \rho}{2}\|\mathbf{y} - \widehat{\mathbf{x}}\|^2 \geq -\frac{\|\widehat{\boldsymbol{\zeta}}_g + \hat{\rho}(\widehat{\mathbf{x}} - \mathbf{x}) + \widehat{\mathbf{u}}/\widehat{\lambda}\|^2}{2(\hat{\rho} - \rho)},$$

which, together with (93) and the first inequality in (10), implies

$$\widehat{\lambda} = \frac{\|\widehat{\boldsymbol{\zeta}}_f + \hat{\rho}(\widehat{\mathbf{x}} - \mathbf{x})\|}{\|\widehat{\boldsymbol{\zeta}}_g + \hat{\rho}(\widehat{\mathbf{x}} - \mathbf{x}) + \widehat{\mathbf{u}}/\widehat{\lambda}\|} \leq \frac{2M}{\sqrt{2\theta(\hat{\rho} - \rho)}}.$$

$\square$

Proof of Lemma 5.3. Consider $\mathbf{x} \in \mathcal{L}$. Since $\mathbf{x}$ is ϵ^2-feasible, by Assumption 5.1B, there exists $\mathbf{y} \in \mathrm{relint}(\mathcal{X})$ such that $g(\mathbf{y}) + \frac{\bar{\rho}}{2}\|\mathbf{y} - \mathbf{x}\|^2 \leq -\theta$. Note that function $g(\mathbf{z}) + \frac{\hat{\rho}}{2}\|\mathbf{z} - \mathbf{x}\|^2 + \delta_\mathcal{X}(\mathbf{z})$ is $(\hat{\rho} - \rho)$-strongly convex with respect to $\mathbf{z}$ and its subdifferential with respect to $\mathbf{z}$ at location $\mathbf{z} = \mathbf{x}$ is $\partial g(\mathbf{x}) + \mathcal{N}_\mathcal{X}(\mathbf{x})$. We then have

$$g(\mathbf{x}) + \frac{\hat{\rho}}{2}\|\mathbf{x} - \mathbf{x}\|^2 + \langle \boldsymbol{\zeta}_g + \mathbf{u}, \mathbf{y} - \mathbf{x} \rangle + \frac{\hat{\rho} - \rho}{2}\|\mathbf{y} - \mathbf{x}\|^2 \leq g(\mathbf{y}) + \frac{\hat{\rho}}{2}\|\mathbf{y} - \mathbf{x}\|^2 \leq -\theta$$

for any $\boldsymbol{\zeta}_g \in \partial g(\mathbf{x})$ and any $\mathbf{u} \in \mathcal{N}_\mathcal{X}(\mathbf{x})$. Since $g(\mathbf{x}) = 0$ when $\mathbf{x} \in \mathcal{L}$, applying Young's inequality to the inequality above yields

$$-\frac{\|\boldsymbol{\zeta}_g + \mathbf{u}\|^2}{2(\hat{\rho} - \rho)} \leq \langle \boldsymbol{\zeta}_g + \mathbf{u}, \mathbf{y} - \mathbf{x} \rangle + \frac{\hat{\rho} - \rho}{2}\|\mathbf{y} - \mathbf{x}\|^2 \leq -\theta,$$

which implies the conclusion. $\square$

38

Proof of Lemma 5.4. For any $\mathbf{x} \in \mathcal{X}$ satisfying $\mathrm{dist}(\mathbf{x}, \mathcal{S}) \leq \frac{\nu}{\rho}$, we define

$$\mathbf{x}^\dagger = \mathrm{proj}_\mathcal{S}(\mathbf{x}) \in \arg\min_{\mathbf{y} \in \mathcal{X}, g(\mathbf{y}) \leq 0} \frac{1}{2}\|\mathbf{y} - \mathbf{x}\|^2. \tag{94}$$

Since g can be non-convex, $\mathbf{x}^\dagger$ is not necessarily unique, but this proof works for any $\mathbf{x}^\dagger$ satisfying (94). Since the first conclusion of this lemma holds trivially if $\mathbf{x} \in \mathcal{S}$, we assume $\mathbf{x} \in \mathcal{X}\backslash\mathcal{S}$, which implies $g(\mathbf{x}^\dagger) = 0$ and $\mathbf{x}^\dagger \in \mathcal{L}$. It is easy to show that $\mathbf{x}^\dagger$ is also the optimal solution of

$$\min_{\mathbf{y} \in \mathcal{X}} \left\{ \frac{1}{2}\|\mathbf{y} - \mathbf{x}\|^2, \ s.t. \ g(\mathbf{y}) + \frac{\bar{\rho}}{2}\|\mathbf{y} - \mathbf{x}^\dagger\|^2 \leq 0 \right\}, \tag{95}$$

which is a convex optimization problem. By Assumption 5.1B, there exists $\mathbf{y} \in \mathrm{relint}(\mathcal{X})$ such that $g(\mathbf{y}) + \frac{\bar{\rho}}{2}\|\mathbf{y} - \mathbf{x}^\dagger\|^2 \leq -\theta$. Hence, the Slater's condition holds for (95), so there exist a scalar $\lambda^\dagger \geq 0$, a subgradient $\boldsymbol{\zeta}_g \in \partial g(\mathbf{x}^\dagger)$ and a vector $\mathbf{u} \in \mathcal{N}_\mathcal{X}(\mathbf{x}^\dagger)$ such that

$$\mathbf{x}^\dagger - \mathbf{x} + \lambda^\dagger \boldsymbol{\zeta}_g + \mathbf{u} = \mathbf{0}. \tag{96}$$

Since $\langle \mathbf{u}, \mathbf{x} - \mathbf{x}^\dagger \rangle \leq 0$, we assert that $\lambda^\dagger > 0$ because, otherwise, (96) implies $0 = \langle \mathbf{x}^\dagger - \mathbf{x} + \mathbf{u}, \mathbf{x} - \mathbf{x}^\dagger \rangle \leq -\|\mathbf{x} - \mathbf{x}^\dagger\|^2 \leq 0$ and thus $\mathbf{x} = \mathbf{x}^\dagger$, contradicting with the fact that $\mathbf{x} \in \mathcal{X}\backslash\mathcal{S}$.

By the ρ-weak convexity of g and (96), we have

$$\lambda^\dagger \left(g(\mathbf{x}) - g(\mathbf{x}^\dagger) \right) \geq \lambda^\dagger \left(\langle \boldsymbol{\zeta}_g, \mathbf{x} - \mathbf{x}^\dagger \rangle - \frac{\rho}{2}\|\mathbf{x} - \mathbf{x}^\dagger\|^2 \right)$$

$$\geq \langle \lambda^\dagger \boldsymbol{\zeta}_g + \mathbf{u}, \mathbf{x} - \mathbf{x}^\dagger \rangle - \frac{\lambda^\dagger \rho}{2}\|\mathbf{x} - \mathbf{x}^\dagger\|^2 = \|\mathbf{x} - \mathbf{x}^\dagger\|^2 - \frac{\lambda^\dagger \rho}{2}\|\mathbf{x} - \mathbf{x}^\dagger\|^2.$$

Since $\lambda^\dagger > 0$, dividing both sides of the inequalities above leads to

$$g(\mathbf{x}) - g(\mathbf{x}^\dagger) \geq \frac{\|\mathbf{x} - \mathbf{x}^\dagger\|^2}{\lambda^\dagger} - \frac{\rho}{2}\|\mathbf{x} - \mathbf{x}^\dagger\|^2$$

$$= \|\mathbf{x} - \mathbf{x}^\dagger\| \cdot \|\boldsymbol{\zeta}_g + \mathbf{u}/\lambda^\dagger\| - \frac{\rho}{2}\|\mathbf{x} - \mathbf{x}^\dagger\|^2 \geq \nu\|\mathbf{x} - \mathbf{x}^\dagger\| - \frac{\rho}{2}\|\mathbf{x} - \mathbf{x}^\dagger\|^2 \geq \frac{\nu}{2}\|\mathbf{x} - \mathbf{x}^\dagger\|,$$

where the second inequality is by Lemma 5.3 and the last inequality is because $\|\mathbf{x} - \mathbf{x}^\dagger\| = \mathrm{dist}(\mathbf{x}, \mathcal{S}) \leq \frac{\nu}{\rho}$. The first conclusion is thus proved by the facts that $g(\mathbf{x}) = g_+(\mathbf{x})$ for $\mathbf{x} \in \mathcal{X}\backslash\mathcal{S}$ and that $g(\mathbf{x}^\dagger) = 0$. The second conclusion is directly from Lemma 3.1 and 3.2 in [26]. □

B.3 Proof of Proposition 5.5 and Theorem 5.6

To prove Proposition 5.5 and Theorem 5.6, we need the following lemma from Theorem 4.1 in [26] whose proof is provided only for completeness. It shows that, when the Polyak's stepsize is used, the subgradient method can utilize the local error bound condition in Lemma 5.4 to ensure $\mathrm{dist}(\mathbf{x}^{(t)}, \mathcal{S})$ Q-linearly converges to zero, which makes sure $\mathrm{dist}(\mathbf{x}^{(t)}, \mathcal{S})$ is small even when $t \in J$ and prevents $\mathbf{x}^{(t)}$ from being trapped in an infeasible stationary point given the second conclusion of Lemma 5.4.

Lemma B.3 ([26]). *Suppose Assumptions 3.1 and 5.1 hold. Also, suppose the sequence $\{\mathbf{x}^{(t)}\}_{t \geq 0}$ is generated by applying the projected subgradient method to $\min_{\mathbf{x} \in \mathcal{X}} g_+(\mathbf{x})$ using a Polyak's stepsize, namely,*

$$\mathbf{x}^{(t+1)} = \mathrm{proj}_\mathcal{X}(\mathbf{x}^{(t)} - \eta_t \boldsymbol{\zeta}_g^{(t)}), \quad \eta_t = \begin{cases} g_+(\mathbf{x}^{(t)})/\|\boldsymbol{\zeta}_g^{(t)}\|^2 & \text{if } \boldsymbol{\zeta}_g^{(t)} \neq \mathbf{0} \\ 0 & \text{if } \boldsymbol{\zeta}_g^{(t)} = \mathbf{0} \end{cases}, \quad \text{for } t = 0, 1, \ldots, \tag{97}$$

where $\boldsymbol{\zeta}_g^{(t)} \in \partial g_+(\mathbf{x}^{(t)})$. If $\mathrm{dist}(\mathbf{x}^{(0)}, \mathcal{S}) \leq \nu/(4\rho)$, we have

$$\mathrm{dist}^2(\mathbf{x}^{(t+1)}, \mathcal{S}) \leq \left(1 - \frac{\nu^2}{8M^2} \right) \mathrm{dist}^2(\mathbf{x}^{(t)}, \mathcal{S}) \tag{98}$$

and

$$\mathrm{dist}(\mathbf{x}^{(t)}, \mathcal{S}) \leq \mathrm{dist}(\mathbf{x}^{(0)}, \mathcal{S}), \quad \forall t \geq 0. \tag{99}$$

39

Proof. We prove (99) by induction. (99) holds trivially for $t = 0$. Suppose that (99) holds up to iteration t of (97). We want to prove that it also holds for iteration $t + 1$. By the induction hypothesis, we have $\text{dist}(\mathbf{x}^{(t)}, \mathcal{S}) \leq \nu/(4\rho)$.

Suppose $\boldsymbol{\zeta}_g^{(t)} = \mathbf{0}$ so $\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)}$. By Lemma 5.4, we must have $\text{dist}(\mathbf{x}^{(t)}, \mathcal{S}) = 0$ and thus $\text{dist}(\mathbf{x}^{(t+1)}, \mathcal{S}) = 0 \leq \text{dist}(\mathbf{x}^{(0)}, \mathcal{S})$ and (98) holds.

Suppose $\boldsymbol{\zeta}_g^{(t)} \neq \mathbf{0}$. Let $\mathbf{x}^{\dagger(t)} = \text{proj}_{\mathcal{S}}(\mathbf{x}^{(t)})$ for $t \geq 0$. By the nonexpansiveness of $\text{prox}_{\mathcal{X}}(\cdot)$, we have

$$
\begin{aligned}
\text{dist}^2(\mathbf{x}^{(t+1)}, \mathcal{S}) &\leq \|\mathbf{x}^{(t+1)} - \mathbf{x}^{\dagger(t)}\|^2 = \|\text{proj}_{\mathcal{X}}(\mathbf{x}^{(t)} - \eta_t \boldsymbol{\zeta}_g^{(t)}) - \text{proj}_{\mathcal{X}}(\mathbf{x}^{\dagger(t)})\|^2 \\
&\leq \|\mathbf{x}^{(t)} - \eta_t \boldsymbol{\zeta}_g^{(t)} - \mathbf{x}^{\dagger(t)}\|^2 \\
&= \text{dist}^2(\mathbf{x}^{(t)}, \mathcal{S}) - 2\eta_t \left\langle \boldsymbol{\zeta}_g^{(t)}, \mathbf{x}^{(t)} - \mathbf{x}^{\dagger(t)} \right\rangle + \eta_t^2 \|\boldsymbol{\zeta}_g^{(t)}\|^2 \\
&\leq \text{dist}^2(\mathbf{x}^{(t)}, \mathcal{S}) + \frac{2g_+(\mathbf{x}^{(t)})}{\|\boldsymbol{\zeta}_g^{(t)}\|^2} \left(g_+(\mathbf{x}^{\dagger(t)}) - g_+(\mathbf{x}^{(t)}) + \frac{\rho}{2}\|\mathbf{x}^{(t)} - \mathbf{x}^{\dagger(t)}\|^2 \right) + \frac{g_+^2(\mathbf{x}^{(t)})}{\|\boldsymbol{\zeta}_g^{(t)}\|^2} \\
&= \text{dist}^2(\mathbf{x}^{(t)}, \mathcal{S}) + \frac{g_+(\mathbf{x}^{(t)})}{\|\boldsymbol{\zeta}_g^{(t)}\|^2} \left(-g_+(\mathbf{x}^{(t)}) + \rho\|\mathbf{x}^{(t)} - \mathbf{x}^{\dagger(t)}\|^2 \right) \\
&\leq \text{dist}^2(\mathbf{x}^{(t)}, \mathcal{S}) + \frac{g_+(\mathbf{x}^{(t)})}{\|\boldsymbol{\zeta}_g^{(t)}\|^2} \left(-\frac{\nu}{2} + \rho\|\mathbf{x}^{(t)} - \mathbf{x}^{\dagger(t)}\| \right) \|\mathbf{x}^{(t)} - \mathbf{x}^{\dagger(t)}\| \\
&\leq \text{dist}^2(\mathbf{x}^{(t)}, \mathcal{S}) - \frac{g_+(\mathbf{x}^{(t)})}{\|\boldsymbol{\zeta}_g^{(t)}\|^2} \frac{\nu}{4} \|\mathbf{x}^{(t)} - \mathbf{x}^{\dagger(t)}\| \\
&\leq \text{dist}^2(\mathbf{x}^{(t)}, \mathcal{S}) - \frac{\nu^2}{8M^2} \|\mathbf{x}^{(t)} - \mathbf{x}^{\dagger(t)}\|^2,
\end{aligned}
$$

where the second equality is because $g_+(\mathbf{x}^{\dagger(t)}) = 0$, the third inequality is by the ρ-weak convexity of g_+, the fourth is by Lemma 5.4, the fifth is by the hypothesis that $\text{dist}(\mathbf{x}^{(t)}, \mathcal{S}) \leq \text{dist}(\mathbf{x}^{(0)}, \mathcal{S}) \leq \nu/(4\rho)$, and the last is by Lemma 5.4 and Assumption 3.1A. This inequality further implies

$$
\text{dist}^2(\mathbf{x}^{(t+1)}, \mathcal{S}) \leq \left(1 - \frac{\nu^2}{8M^2} \right) \text{dist}^2(\mathbf{x}^{(t)}, \mathcal{S}) \leq \text{dist}^2(\mathbf{x}^{(0)}, \mathcal{S}),
$$

which proves (98) and also proves (99) by induction.

$\square$

Proof of Proposition 5.5. The choices of ϵ_t and η_t imply

$$
2\epsilon_t/\nu + 2\eta_t M^2/\nu \leq \min\left\{ \epsilon^2/M, \nu/(4\rho) \right\} \tag{100}
$$

for $t \in I$. We first prove $\text{dist}(\mathbf{x}^{(t)}, \mathcal{S}) \leq \min\left\{ \epsilon^2/M, \nu/(4\rho) \right\}$ by induction on t. Since $\mathbf{x}^{(0)} \in \mathcal{S}$, this conclusion holds trivially for $t = 0$. Suppose it holds up to iteration t. We want to prove it also holds for iteration $t + 1$.

Suppose $t \in I$. Recall that $\nu \leq 2M$. We have $\eta_t M \leq 2\eta_t M^2/\nu \leq \nu/(4\rho)$ and $\|\mathbf{x}^{(t+1)} - \mathbf{x}^{(t)}\| \leq \|\eta_t \boldsymbol{\zeta}_f^{(t)}\| \leq \eta_t M$, which, by triangle inequality and the induction hypothesis, implies that

$$
\text{dist}(\mathbf{x}^{(t+1)}, \mathcal{S}) \leq \text{dist}(\mathbf{x}^{(t)}, \mathcal{S}) + \eta_t M \leq \frac{\nu}{4\rho} + \eta_t M \leq \frac{\nu}{2\rho}.
$$

Since $g(\mathbf{x}^{(t)}) \leq \epsilon_t$ for $t \in I$, we have $g(\mathbf{x}^{(t+1)}) \leq g(\mathbf{x}^{(t)}) + M\|\mathbf{x}^{(t+1)} - \mathbf{x}^{(t)}\| \leq \epsilon_t + \eta_t M^2$. By Lemma 5.4, we have $\text{dist}(\mathbf{x}^{(t+1)}, \mathcal{S}) \leq 2\epsilon_t/\nu + 2\eta_t M^2/\nu \leq \min\left\{ \epsilon^2/M, \nu/(4\rho) \right\}$.

Suppose $t \in J$. Let t' be the largest index in I that is smaller than t. By the same proof as in the previous case, we have $\text{dist}(\mathbf{x}^{(t'+1)}, \mathcal{S}) \leq \min\left\{ \epsilon^2/M, \nu/(4\rho) \right\}$. Since indexes $t' + 1, t' + 2, ...,$ t are in J, Algorithm 1 essentially performs the projected subgradient method to $\min_{\mathbf{x} \in \mathcal{X}} g_+(\mathbf{x})$ using a Polyak's stepsize during iterations $t' + 1, t' + 2, ...,$ and t. Hence, by Lemma B.3, we have $\text{dist}(\mathbf{x}^{(t+1)}, \mathcal{S}) \leq \text{dist}(\mathbf{x}^{(t'+1)}, \mathcal{S}) \leq \min\left\{ \epsilon^2/M, \nu/(4\rho) \right\}$.

By induction, we have prove that $\text{dist}(\mathbf{x}^{(t)}, \mathcal{S}) \leq \min\left\{ \epsilon^2/M, \nu/(4\rho) \right\}$ for any $t \geq 0$. As a result, $g(\mathbf{x}^{(t)}) + \frac{\tilde{\rho}}{2}\|\mathbf{x}^{(t)} - \mathbf{x}^{(t)}\|^2 = g(\mathbf{x}^{(t)}) \leq M \cdot \text{dist}(\mathbf{x}^{(t)}, \mathcal{S}) \leq M \cdot \min\left\{ \epsilon^2/M, \nu/(4\rho) \right\} \leq \epsilon^2$, meaning that $\mathbf{x}^{(t)}$ is ϵ^2-feasible for (3). $\square$

40

The following proposition provides the main inequality needed for proving Theorem 5.6.

Proposition B.4. *Under the same assumptions as Proposition 5.5, Algorithm 1 guarantees*

$$\sum_{t=S}^{T-1} \frac{\eta_t \hat{\rho}(\hat{\rho} - \rho)}{2} \|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\|^2 \leq f(\mathbf{x}^{(S)}) - \underline{f} + \frac{3M^2}{2\hat{\rho}} + \sum_{t\in I}(1 + \widehat{\lambda}_{t+1})\eta_t \hat{\rho}\widehat{\lambda}_t \epsilon_t$$

$$+ \frac{\hat{\rho}}{2}\sum_{t=S}^{T-1}(1 + \widehat{\lambda}_{t+1})\left[\eta_t^2\|\boldsymbol{\zeta}_f^{(t)}\|^2\mathbb{I}(g(\mathbf{x}^t) \leq \epsilon_t) + \eta_t^2\|\boldsymbol{\zeta}_g^{(t)}\|^2\mathbb{I}(g(\mathbf{x}^t) > \epsilon_t)\right].$$

Proof. Let $\boldsymbol{\zeta}^{(t)} = \boldsymbol{\zeta}_f^{(t)} \in \partial f(\mathbf{x}^{(t)})$ if $t \in I$ and $\boldsymbol{\zeta}^{(t)} = \boldsymbol{\zeta}_g^{(t)} \in \partial g(\mathbf{x}^{(t)})$ if $t \in J$. Let $\varphi(\mathbf{x})$ and $\widehat{\mathbf{x}}$ be defined in (3) and (4) with $(\hat{\rho}, \tilde{\rho})$ satisfying (9). For simplicity of notation, we denote $\widehat{\mathbf{x}}(\mathbf{x}^{(t)})$ by $\widehat{\mathbf{x}}^{(t)}$. Since $\mathbf{x}^{(t)}$ is ϵ^2-feasible by Proposition 5.5, we get from Assumption 5.1B that $\varphi(\mathbf{x}^{(t)})$ and $\widehat{\mathbf{x}}^{(t)}$ are well defined for any $t \geq 0$.

By Assumption 5.1B, problem (3) with $\mathbf{x} = \mathbf{x}^{(t)}$ is strongly convex and has a strictly feasible solution, so there exists a Lagrangian multiplier $\widehat{\lambda}_t \geq 0$ satisfying (5) and

$$\widehat{\mathbf{x}}^{(t)} = \arg\min_{\mathbf{x}\in\mathcal{X}}\left\{f(\mathbf{x}) + \frac{\hat{\rho}}{2}\|\mathbf{x} - \mathbf{x}^{(t)}\|^2 + \widehat{\lambda}_t\left(g(\mathbf{x}) + \frac{\hat{\rho}}{2}\|\mathbf{x} - \mathbf{x}^{(t)}\|^2\right)\right\}. \tag{101}$$

By the updating equation of $\mathbf{x}^{(t+1)}$, we have

$$\|\mathbf{x}^{(t+1)} - \widehat{\mathbf{x}}^{(t)}\|^2 = \|\text{proj}_{\mathcal{X}}(\mathbf{x}^{(t)} - \eta_t\boldsymbol{\zeta}^{(t)}) - \widehat{\mathbf{x}}^{(t)}\|^2 = \|\text{proj}_{\mathcal{X}}(\mathbf{x}^{(t)} - \eta_t\boldsymbol{\zeta}^{(t)}) - \text{proj}_{\mathcal{X}}(\widehat{\mathbf{x}}^{(t)})\|^2$$

$$\leq \|\mathbf{x}^{(t)} - \eta_t\boldsymbol{\zeta}^{(t)} - \widehat{\mathbf{x}}^{(t)}\|^2 = \|\mathbf{x}^{(t)} - \widehat{\mathbf{x}}^{(t)}\|^2 - 2\eta_t\left\langle\boldsymbol{\zeta}^{(t)}, \mathbf{x}^{(t)} - \widehat{\mathbf{x}}^{(t)}\right\rangle + \eta_t^2\|\boldsymbol{\zeta}^{(t)}\|^2.$$

Multiplying the inequality above by $\hat{\rho}(1 + \widehat{\lambda}_{t+1})/2$ and adding $f(\widehat{\mathbf{x}}^{(t)}) + \widehat{\lambda}_{t+1}g(\widehat{\mathbf{x}}^{(t)})$ to both sides, we obtain

$$f(\widehat{\mathbf{x}}^{(t)}) + \frac{\hat{\rho}}{2}\|\mathbf{x}^{(t+1)} - \widehat{\mathbf{x}}^{(t)}\|^2 + \widehat{\lambda}_{t+1}\left(g(\widehat{\mathbf{x}}^{(t)}) + \frac{\hat{\rho}}{2}\|\mathbf{x}^{(t+1)} - \widehat{\mathbf{x}}^{(t)}\|^2\right)$$

$$\leq f(\widehat{\mathbf{x}}^{(t)}) + \frac{\hat{\rho}}{2}\|\mathbf{x}^{(t)} - \widehat{\mathbf{x}}^{(t)}\|^2 + \widehat{\lambda}_{t+1}\left(g(\widehat{\mathbf{x}}^{(t)}) + \frac{\hat{\rho}}{2}\|\mathbf{x}^{(t)} - \widehat{\mathbf{x}}^{(t)}\|^2\right)$$

$$- (1 + \widehat{\lambda}_{t+1})\eta_t\hat{\rho}\left\langle\boldsymbol{\zeta}^{(t)}, \mathbf{x}^{(t)} - \widehat{\mathbf{x}}^{(t)}\right\rangle + \frac{(1 + \widehat{\lambda}_{t+1})\eta_t^2\hat{\rho}}{2}\|\boldsymbol{\zeta}^{(t)}\|^2$$

$$\leq \varphi(\mathbf{x}^{(t)}) - (1 + \widehat{\lambda}_{t+1})\eta_t\hat{\rho}\left\langle\boldsymbol{\zeta}^{(t)}, \mathbf{x}^{(t)} - \widehat{\mathbf{x}}^{(t)}\right\rangle + \frac{(1 + \widehat{\lambda}_{t+1})\eta_t^2\hat{\rho}}{2}\|\boldsymbol{\zeta}^{(t)}\|^2, \tag{102}$$

where the second inequality is by the definition of $\varphi(\mathbf{x})$ in (3) and the fact that $\widehat{\lambda}_{t+1} \geq 0$ and $g(\widehat{\mathbf{x}}^{(t)}) + \frac{\hat{\rho}}{2}\|\mathbf{x}^{(t)} - \widehat{\mathbf{x}}^{(t)}\|^2 \leq 0$.

By the definitions of $\varphi(\mathbf{x}^{(t+1)})$ and $\widehat{\lambda}_{t+1}$, we have the following equation similar to (101)

$$\varphi(\mathbf{x}^{(t+1)}) = f(\widehat{\mathbf{x}}^{(t+1)}) + \frac{\hat{\rho}}{2}\|\mathbf{x}^{(t+1)} - \widehat{\mathbf{x}}^{(t+1)}\|^2 + \widehat{\lambda}_{t+1}\left(g(\widehat{\mathbf{x}}^{(t+1)}) + \frac{\hat{\rho}}{2}\|\mathbf{x}^{(t+1)} - \widehat{\mathbf{x}}^{(t+1)}\|^2\right)$$

$$\leq f(\widehat{\mathbf{x}}^{(t)}) + \frac{\hat{\rho}}{2}\|\mathbf{x}^{(t+1)} - \widehat{\mathbf{x}}^{(t)}\|^2 + \widehat{\lambda}_{t+1}\left(g(\widehat{\mathbf{x}}^{(t)}) + \frac{\hat{\rho}}{2}\|\mathbf{x}^{(t+1)} - \widehat{\mathbf{x}}^{(t)}\|^2\right),$$

which, together with (102), implies

$$(1 + \widehat{\lambda}_{t+1})\eta_t\hat{\rho}\left\langle\boldsymbol{\zeta}^{(t)}, \mathbf{x}^{(t)} - \widehat{\mathbf{x}}^{(t)}\right\rangle \leq \varphi(\mathbf{x}^{(t)}) - \varphi(\mathbf{x}^{(t+1)}) + \frac{(1 + \widehat{\lambda}_{t+1})\eta_t^2\hat{\rho}}{2}\|\boldsymbol{\zeta}^{(t)}\|^2. \tag{103}$$

Next, we will bound $\left\langle\boldsymbol{\zeta}^{(t)}, \mathbf{x}^{(t)} - \widehat{\mathbf{x}}^{(t)}\right\rangle$ from below when $t \in I$ and $t \in J$, separately.

Suppose $t \in I$ so $g(\mathbf{x}^{(t)}) \leq \epsilon_t$ and $\boldsymbol{\zeta}^{(t)} = \boldsymbol{\zeta}_f^{(t)}$. By the ρ-weak convexity of f, we have

$$\left\langle\boldsymbol{\zeta}^{(t)}, \mathbf{x}^{(t)} - \widehat{\mathbf{x}}^{(t)}\right\rangle \geq f(\mathbf{x}^{(t)}) - f(\widehat{\mathbf{x}}^{(t)}) - \frac{\rho}{2}\|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\|^2$$

$$= f(\mathbf{x}^{(t)}) - f(\widehat{\mathbf{x}}^{(t)}) - \frac{\hat{\rho}}{2}\|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\|^2 + \frac{\hat{\rho} - \rho}{2}\|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\|^2. \tag{104}$$

41

Since the objective function in (101) is $(1 + \widehat{\lambda}_t)(\hat{\rho} - \rho)$-strongly convex, we have

$$f(\mathbf{x}^{(t)}) + \widehat{\lambda}_t g(\mathbf{x}^{(t)}) = f(\mathbf{x}^{(t)}) + \frac{\hat{\rho}}{2}\|\mathbf{x}^{(t)} - \mathbf{x}^{(t)}\|^2 + \widehat{\lambda}_t\left(g(\mathbf{x}^{(t)}) + \frac{\hat{\rho}}{2}\|\mathbf{x}^{(t)} - \mathbf{x}^{(t)}\|^2\right)$$

$$\geq f(\widehat{\mathbf{x}}^{(t)}) + \frac{\hat{\rho}}{2}\|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\|^2 + \widehat{\lambda}_t\left(g(\widehat{\mathbf{x}}^{(t)}) + \frac{\hat{\rho}}{2}\|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\|^2\right)$$

$$+ \frac{(1 + \widehat{\lambda}_t)(\hat{\rho} - \rho)}{2}\|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\|^2,$$

which, by the facts that $g(\mathbf{x}^{(t)}) \leq \epsilon_t$, $\widehat{\lambda}_t \geq 0$ and $\widehat{\lambda}_t\big(g(\widehat{\mathbf{x}}^{(t)}) + \frac{\hat{\rho}}{2}\|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\|^2\big) = 0$, implies

$$f(\mathbf{x}^{(t)}) - f(\widehat{\mathbf{x}}^{(t)}) - \frac{\hat{\rho}}{2}\|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\|^2 \geq -\widehat{\lambda}_t\epsilon_t + \frac{\hat{\rho} - \rho}{2}\|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\|^2.$$

Applying this inequality and inequality (104) to (103) leads to

$$(1 + \widehat{\lambda}_{t+1})\eta_t\hat{\rho}(\hat{\rho} - \rho)\|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\|^2$$

$$\leq \varphi(\mathbf{x}^{(t)}) - \varphi(\mathbf{x}^{(t+1)}) + \frac{(1 + \widehat{\lambda}_{t+1})\eta_t^2\hat{\rho}}{2}\|\boldsymbol{\zeta}_f^{(t)}\|^2 + (1 + \widehat{\lambda}_{t+1})\eta_t\hat{\rho}\widehat{\lambda}_t\epsilon_t. \tag{105}$$

Suppose $t \in J$ so $g(\mathbf{x}^{(t)}) > \epsilon_t$ and $\boldsymbol{\zeta}^{(t)} = \boldsymbol{\zeta}_g^{(t)}$. By the ρ-weak convexity of g and the fact that $g(\widehat{\mathbf{x}}^{(t)}) + \frac{\hat{\rho}}{2}\|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\|^2 \leq 0$, we have

$$\left\langle \boldsymbol{\zeta}^{(t)}, \mathbf{x}^{(t)} - \widehat{\mathbf{x}}^{(t)} \right\rangle \geq g(\mathbf{x}^{(t)}) - g(\widehat{\mathbf{x}}^{(t)}) - \frac{\rho}{2}\|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\|^2$$

$$= g(\mathbf{x}^{(t)}) - g(\widehat{\mathbf{x}}^{(t)}) - \frac{\hat{\rho}}{2}\|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\|^2 + \frac{\hat{\rho} - \rho}{2}\|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\|^2$$

$$> \epsilon_t + \frac{\hat{\rho} - \rho}{2}\|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\|^2.$$

Applying this inequality to (103) leads to

$$\frac{(1 + \widehat{\lambda}_{t+1})\eta_t\hat{\rho}(\hat{\rho} - \rho)}{2}\|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\|^2$$

$$\leq \varphi(\mathbf{x}^{(t)}) - \varphi(\mathbf{x}^{(t+1)}) + \frac{(1 + \widehat{\lambda}_{t+1})\eta_t^2\hat{\rho}}{2}\|\boldsymbol{\zeta}_g^{(t)}\|^2 - (1 + \widehat{\lambda}_{t+1})\eta_t\hat{\rho}\epsilon_t. \tag{106}$$

Summing up (105) and (106) for $t = S, S+1, \ldots, T-1$, we have

$$\sum_{t=S}^{T-1} \frac{\eta_t\hat{\rho}(\hat{\rho} - \rho)}{2}\|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\|^2 \leq \sum_{t=S}^{T-1} \frac{(1 + \widehat{\lambda}_{t+1})\eta_t\hat{\rho}(\hat{\rho} - \rho)}{2}\|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\|^2$$

$$\leq \varphi(\mathbf{x}^{(S)}) - \varphi(\mathbf{x}^{(T)}) + \frac{\hat{\rho}}{2}\sum_{t=S}^{T-1}(1 + \widehat{\lambda}_{t+1})\left[\eta_t^2\|\boldsymbol{\zeta}_f^{(t)}\|^2\mathbb{I}(g(\mathbf{x}^t) \leq \epsilon_t) + \eta_t^2\|\boldsymbol{\zeta}_g^{(t)}\|^2\mathbb{I}(g(\mathbf{x}^t) > \epsilon_t)\right]$$

$$+ \sum_{t\in I}\left[(1 + \widehat{\lambda}_{t+1})\eta_t\hat{\rho}\widehat{\lambda}_t\epsilon_t\right] + \sum_{t\in J}\left[-(1 + \widehat{\lambda}_{t+1})\eta_t\hat{\rho}\epsilon_t\right]$$

$$\leq \varphi(\mathbf{x}^{(S)}) - \varphi(\mathbf{x}^{(T)}) + \sum_{t\in I}(1 + \widehat{\lambda}_{t+1})\eta_t\hat{\rho}\widehat{\lambda}_t\epsilon_t$$

$$+ \frac{\hat{\rho}}{2}\sum_{t=S}^{T-1}(1 + \widehat{\lambda}_{t+1})\left[\eta_t^2\|\boldsymbol{\zeta}_f^{(t)}\|^2\mathbb{I}(g(\mathbf{x}^t) \leq \epsilon_t) + \eta_t^2\|\boldsymbol{\zeta}_g^{(t)}\|^2\mathbb{I}(g(\mathbf{x}^t) > \epsilon_t)\right]. \tag{107}$$

Finally, from Assumption 5.1C and the first inequality in (10), it can be easily shown that

$$\varphi(\mathbf{x}^{(T)}) = f(\widehat{\mathbf{x}}^{(T)}) + \frac{\hat{\rho}}{2}\|\widehat{\mathbf{x}}^{(T)} - \mathbf{x}^{(T)}\|^2 \geq \underline{f}$$

and

$$\varphi(\mathbf{x}^{(S)}) = f(\widehat{\mathbf{x}}^{(S)}) + \frac{\hat{\rho}}{2}\|\widehat{\mathbf{x}}^{(S)} - \mathbf{x}^{(S)}\|^2 \leq f(\mathbf{x}^{(S)}) + \frac{M^2}{\hat{\rho}} + \frac{\hat{\rho}}{2}\frac{M^2}{\hat{\rho}^2} \leq f(\mathbf{x}^{(S)}) + \frac{3M^2}{2\hat{\rho}}.$$

Then, the conclusion is derived from by applying these two inequalities to (107). $\qquad\square$

42

We are now ready to prove the main theorem for the weakly convex case.

Proof of Theorem 5.6. The inequality $\mathbb{E}_\tau[g(\mathbf{x}^{(\tau)})] \le \epsilon^2$ is a direct consequence of Proposition 5.5. Applying the upper bound Λ' of $\widehat{\lambda}_t$ from Lemma 5.2 to the inequality in Proposition B.4, we obtain

$$\sum_{t=S}^{T-1} \frac{\eta_t \hat{\rho}(\hat{\rho} - \rho)}{2}\|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\|^2 \le f(\mathbf{x}^{(S)}) - \underline{f} + \frac{3M^2}{2\hat{\rho}} + \sum_{t \in I}(1 + \Lambda')\eta_t \hat{\rho}\Lambda'\epsilon_t$$

$$+ \frac{\hat{\rho}}{2}\sum_{t=S}^{T-1}(1 + \Lambda')\left[\eta_t^2\|\boldsymbol{\zeta}_f^{(t)}\|^2\mathbb{I}(g(\mathbf{x}^t) \le \epsilon_t) + \eta_t^2\|\boldsymbol{\zeta}_g^{(t)}\|^2\mathbb{I}(g(\mathbf{x}^t) > \epsilon_t)\right]. \tag{108}$$

For $t \in I$, we have from Assumption 3.1 and the definition of η_t that

$$\eta_t^2\|\boldsymbol{\zeta}_f^{(t)}\|^2 \le \eta_t^2 M^2 \le \eta_t M \cdot \frac{2\eta_t M^2}{\nu} \le \eta_t \epsilon^2. \tag{109}$$

For $t \in J$, we have $\eta_t = g_+(\mathbf{x}^{(t)})/\|\boldsymbol{\zeta}_g^{(t)}\|^2$ so

$$\eta_t^2\|\boldsymbol{\zeta}_g^{(t)}\|^2 \le \eta_t g_+(\mathbf{x}^{(t)}) \le \eta_t \epsilon^2, \tag{110}$$

where the second inequality is from Proposition 5.5 that implies $g_+(\mathbf{x}^{(t)}) = g(\mathbf{x}^{(t)}) \le \epsilon^2$.

Applying (109) and (110) to (108) and organizing term lead to

$$\frac{\hat{\rho}(\hat{\rho} - \rho)}{2}\mathbb{E}_\tau[\|\widehat{\mathbf{x}}^{(\tau)} - \mathbf{x}^{(\tau)}\|^2] = \frac{\hat{\rho}(\hat{\rho} - \rho)}{2}\frac{\sum_{t=S}^{T-1}\eta_t\|\widehat{\mathbf{x}}^{(t)} - \mathbf{x}^{(t)}\|^2}{\sum_{t=S}^{T-1}\eta_t}$$

$$\le \frac{1}{\sum_{t=S}^{T-1}\eta_t}\left(f(\mathbf{x}^{(S)}) - \underline{f} + \frac{3M^2}{2\hat{\rho}}\right) + (1 + \Lambda')\hat{\rho}\frac{\sum_{t \in I}\eta_t\Lambda'\epsilon_t}{\sum_{t=S}^{T-1}\eta_t} + \frac{\hat{\rho}(1 + \Lambda')}{2}\epsilon^2$$

$$\le \frac{1}{\sum_{t=S}^{T-1}\eta_t}\left(f(\mathbf{x}^{(S)}) - \underline{f} + \frac{3M^2}{2\hat{\rho}}\right) + \frac{3\hat{\rho}(1 + \Lambda')}{2}\epsilon^2, \tag{111}$$

where the second inequality is by the definition of ϵ_t which implies $\epsilon_t \le \epsilon^2\nu/(2M) = \epsilon^2/\Lambda'$. According to the definition of η_t for $t \in I$ and $t \in J$, we have

$$\sum_{t=S}^{T-1}\eta_t \ge \frac{\nu|I|}{4M^2}\min\left\{\epsilon^2/M, \nu/(4\rho)\right\} + \sum_{t \in J}\frac{\epsilon_t}{\|\boldsymbol{\zeta}_g^{(t)}\|^2}$$

$$\ge \frac{\nu|I|}{4M^2}\min\left\{\epsilon^2/M, \nu/(4\rho)\right\} + \frac{\nu|J|}{4M^2}\min\left\{\epsilon^2/M, \nu/(4\rho)\right\}$$

$$= \frac{\nu(T - S)}{4M^2}\min\left\{\epsilon^2/M, \nu/(4\rho)\right\}. \tag{112}$$

Applying (112) as well as the definitions of S, η_t and ϵ_t to the right hand side of (111) gives

$$\mathbb{E}_\tau[\|\widehat{\mathbf{x}}^{(\tau)} - \mathbf{x}^{(\tau)}\|^2] \le \frac{8M^2\left(f(\mathbf{x}^{(0)}) - \underline{f} + 3M^2/(2\hat{\rho})\right)}{\hat{\rho}(\hat{\rho} - \rho)\nu\min\left\{\epsilon^2/M, \nu/(4\rho)\right\}T} + \frac{3(1 + \Lambda')}{\hat{\rho} - \rho}\epsilon^2,$$

which implies the conclusion of this theorem after plugging in the value of T. $\qquad\square$

C Additional numerical results

In this section, we first present the performances versus the used CPU time of the methods in comparisons on problem (15) in Figure 2. The meanings of y-axes in Figure 2 remain the same as in Figure 1, but the x-axis in each plot represents the CPU time in seconds the methods use. According to Figure 2, the relative speed among both SSG methods and IPP-SSG remain nearly the same as in Figure 1 while IPP-ConEx is about twice time-consuming compared to others. This can be explained by the algorithmic designs of these methods. In fact, the two SSG methods and IPP-SSG only need to compute either $\boldsymbol{\zeta}_f^{(t)}$ or $\boldsymbol{\zeta}_g^{(t)}$ in each (inner) iteration while IPP-SSG computes both $\boldsymbol{\zeta}_f^{(t)}$ and $\boldsymbol{\zeta}_g^{(t)}$ per inner iteration and thus induces additional computational cost.

43
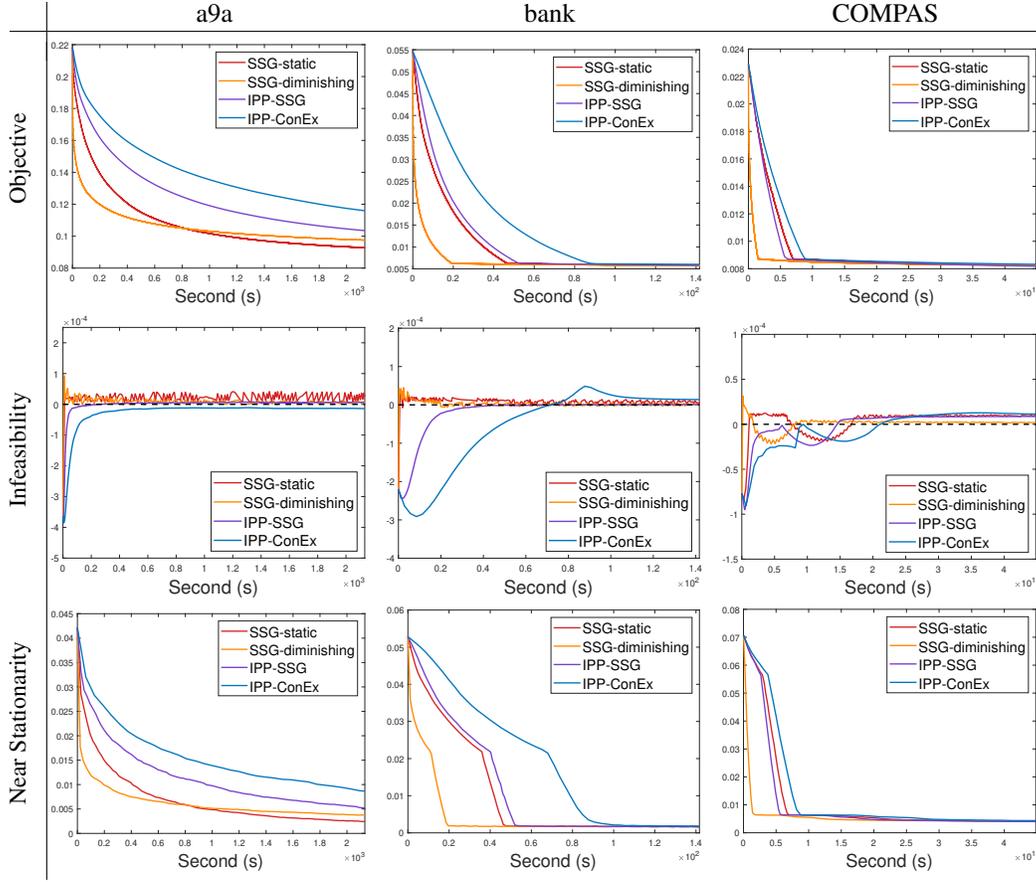
Figure 2: Performances vs CPU time on classification problems with ROC-based fairness.

Then, we present the numerical performances versus both the number of iterations and the CPU time of the methods in comparisons on problem (17). The results on each dataset are given in Figure 3 and Figure 4 where the meanings of axes and the procedure of calculating the (approximate) near stationarity are the same as in Figure 1 and Figure 2, respectively. Since computing $\widehat{\mathbf{x}}(\mathbf{x}^{(t)})$ with a high precision for each t is time-consuming, we only report near stationarity at 600 equally spaced iterations. Figure 3 shows that the SSG method is only a little faster than the other two methods in reducing the objective value and the (approximate) near stationarity while keeping the solutions nearly feasible. This is consistent with our theoretical finding that the three methods have the similar oracle complexity. Figure 4 shows that the relative speeds among the two SSG methods and IPP-SSG are nearly the same as in Figure 3 while IPP-ConEx is about twice time-consuming. This is again consistent with the algorithmic designs of these methods and can be explained by the same reasons we gave above for Figure 2.

44
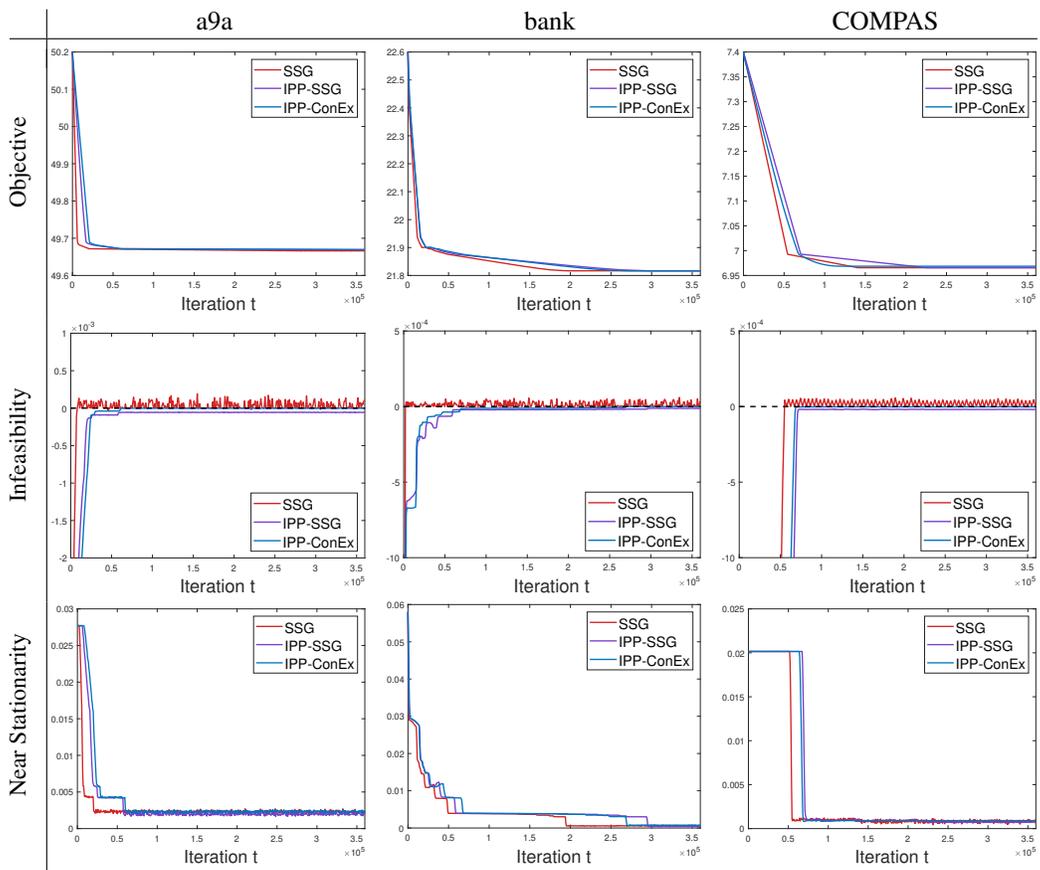
Figure 3: Performances vs number of iterations on classification problems with demographic equity constraint.

45
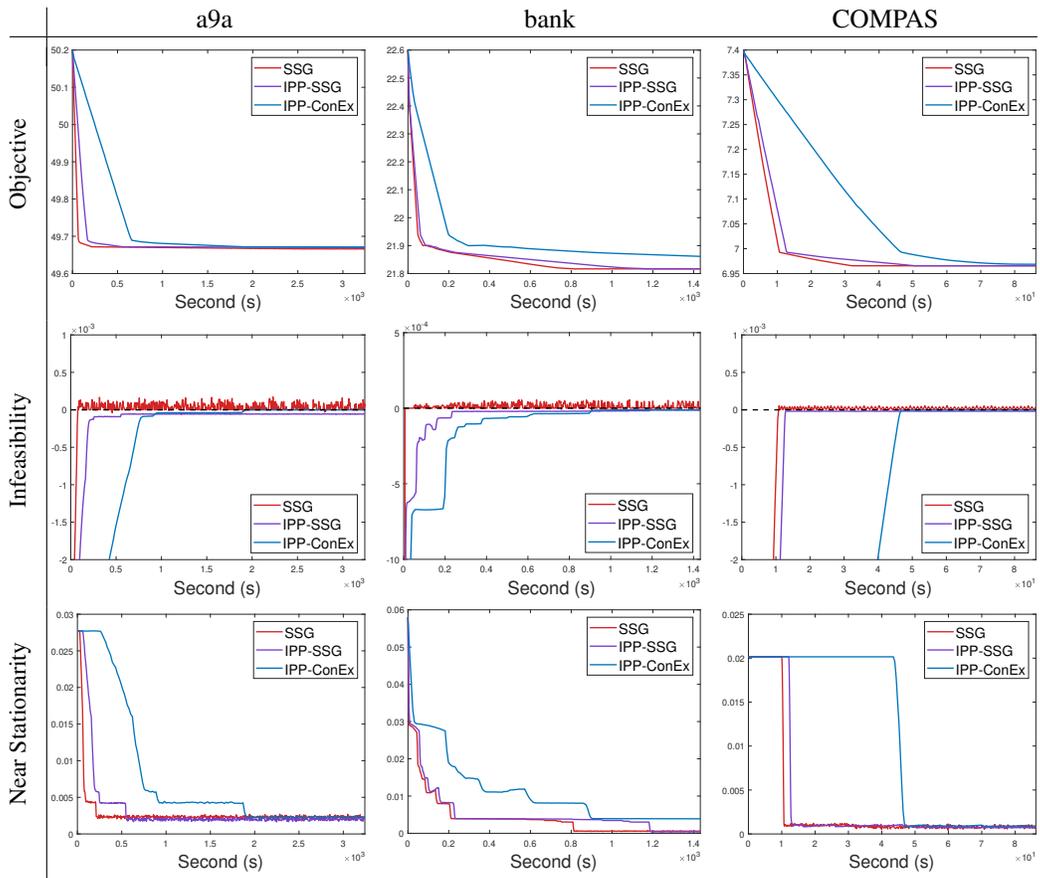
Figure 4: Performances vs CPU time on classification problems with demographic equity constraint.

46