RSG: BEATING SUBGRADIENT METHOD WITHOUT SMOOTHNESS AND STRONG CONVEXITY

TIANBAO YANG* AND QIHANG LIN†

Abstract. In this paper, we study the efficiency of a **R**estarted **S**ub**G**radient (RSG) method that periodically restarts the standard subgradient method (SG). We show that, when applied to a broad class of convex optimization problems, RSG method can find an ϵ-optimal solution with a low complexity than SG method. In particular, we first show that RSG can reduce the dependence of SG's iteration complexity on the distance between the initial solution and the optimal set to that between the ϵ-level set and the optimal set. In addition, we show the advantages of RSG over SG in solving three different families of convex optimization problems. (a) For the problems whose epigraph is a polyhedron, RSG is shown to converge linearly. (b) For the problems with local quadratic growth property, RSG has an $O(\frac{1}{\epsilon}\log(\frac{1}{\epsilon}))$ iteration complexity. (c) For the problems that admit a local Kurdyka-Łojasiewicz property with a power constant of $\beta \in [0, 1)$, RSG has an $O(\frac{1}{\epsilon^{2\beta}}\log(\frac{1}{\epsilon}))$ iteration complexity. On the contrary, with only the standard analysis, the iteration complexity of SG is known to be $O(\frac{1}{\epsilon^2})$ for these three classes of problems. The novelty of our analysis lies at exploiting the lower bound of the first-order optimality residual at the ϵ-level set. It is this novelty that allows us to explore the local properties of functions (e.g., local quadratic growth property, local Kurdyka-Łojasiewicz property, more generally local error bounds) to develop the improved convergence of RSG. We demonstrate the effectiveness of the proposed algorithms on several machine learning tasks including regression and classification.

Key words. subgradient method, improved convergence, local error bound, machine learning

AMS subject classifications. 90C25, 68Q25

1. Introduction.

We consider the following generic optimization problem

$$f_* := \min_{\mathbf{w}\in\Omega} f(\mathbf{w}) \tag{1}$$

where $f : \mathbb{R}^d \to (-\infty, +\infty]$ is an extended-valued, lower semicontinuous and convex function, and $\Omega \subseteq \mathbb{R}^d$ is a closed convex set such that $\Omega \subseteq \mathrm{dom}(f)$. Here, we do not assume the smoothness of f on $\mathrm{dom}(f)$. During the past several decades, many fast convergent (especially linearly convergent) optimization algorithms have been developed for (1) when f is smooth and/or strongly convex. On the contrary, there are relatively fewer techniques for solving generic non-smooth and non-strongly convex optimization problems which still have many applications in machine learning, statistics, computer vision, and etc. To solve (1) with f being potentially non-smooth and non-strongly convex, one of the simplest algorithms to use is the subgradient (SG) method. When f is Lipschitz-continuous, it is known that SG method requires $O(1/\epsilon^2)$ iterations for obtaining an ϵ-optimal solution [47, 36]. It has been shown that this iteration complexity is unimprovable for general non-smooth and non-strongly convex problems in a black-box first-order oracle model of computation [35]. However, better iteration complexity can be achieved by other first-order algorithms for certain class of f where additional structural information is available [37, 20, 16, 44, 45, 46].

In this paper, we present a generic restarted subgradient (RSG) method for solving (1) which runs in multiple stages with each stage warm-started by the solution from the previous stage. Within each stage, the standard projected subgradient descent is performed for a fixed number of iterations with a constant step size. This step size is reduced geometrically from stage to stage. With these schemes, we show that

*Department of Computer Science, The University of Iowa (tianbao-yang@uiowa.edu).
†Department of Management Sciences, The University of Iowa (qihang-lin@uiowa.edu).

RSG can achieve a lower iteration complexity than the classical SG method when f belongs to some classes of functions. In particular, we summarize the main results and properties of RSG below:

- For the general problem (1), under mild assumptions (see Assumption 1), RSG has an iteration complexity of $O(\frac{1}{\epsilon^2}\log(\frac{1}{\epsilon}))$ which has an addition $\log(\frac{1}{\epsilon})$ term but has significantly smaller constant in $O(\cdot)$ compared to SG. In particular, SG's complexity depends on the distance from the initial solution to the optimal set while RSG's complexity only depends on the distance from the ϵ-level set (defined in (2)) to the optimal set, which is much smaller than the former distance.
- When f is locally quadratically growing (see Definition 11), which is a weaker condition than strong convexity, RSG can achieve an $O(\frac{1}{\epsilon}\log(\frac{1}{\epsilon}))$ iteration complexity.
- When f admits a local Kurdyka-Łojasiewicz property (see Definition 14) with a power desingularizing function of degree $1-\beta$ where $\beta \in [0, 1)$, RSG can achieve an $O(\frac{1}{\epsilon^{2\beta}}\log(\frac{1}{\epsilon}))$ complexity.
- When the epigraph of f over Ω is a polyhedron, RSG can achieve linear convergence, i.e., an $O(\log(\frac{1}{\epsilon}))$ iteration complexity.

These results, except for the first one, are derived from a generic complexity of RSG for the problem satisfying a *local error condition* (16), which has a close connection to the existing error bound conditions and growth conditions [42, 41, 31, 33, 6] in the literature. In spite of its simplicity, the analysis of RSG provides additional insight on improving first-order methods' iteration complexity via restarting. It is known that restarting can improve the theoretical complexity of (stochastic) SG method for non-smooth problem when strongly convexity is assumed [19, 10, 23] but we show that restarting can be still helpful for SG methods under other (weaker) assumptions. Although we focus on SG methods, the similar idea can be incorporated into various existing algorithms, leading to different variants of restarted first-order methods. In particular, built on the groundwork (in particular Lemma 5 and Lemma 17) laid in this paper [1], several pieces of studies have improved the convergence of non-smooth optimization with a special structure [55], stochastic subgradient methods [54] and projection-reduced first-order methods [57].

We organize the reminder of the paper as follows. Section 2 reviews some related work. Section 3 presents some preliminaries and notations. Section 4 presents the algorithm of RSG and the general theory of convergence in the Euclidean norm. Section 5 considers several classes of non-smooth and non-strongly convex problems and shows the improved iteration complexities of RSG. Section 6 generalizes the algorithm and theory into the p-norm space using dual averaging method in each stage. Section 7 presents a parameter-free variant of RSG. Section 8 presents some experimental results. Finally, we conclude in Section 9.

2. Related Work. Smoothness and strong convexity are two key properties of a convex optimization problem that determine the iteration complexity of finding an ϵ-optimal solution by first-order methods. In general, a lower iteration complexity is expected when the problem is either smooth or strongly convex. We refer the reader to [36, 35] for the optimal iteration complexity of first-order methods when applied to the problems with different properties of smoothness and convexity. Recently there emerges a surge of interests in further accelerating first-order methods for non-strongly

[1]The first version of this paper appeared on arXiv on December 9, 2015 and a complete version of this paper appeared on arXiv on April 5, 2016.

convex or non-smooth problems that satisfy some particular conditions [5, 53, 49, 24, 63, 22, 20, 16]. The key condition for us to develop the improved complexity is a local error bound condition (16) which is closely related to the error bound conditions in [41, 42, 31, 33, 6, 60].

Various error bounds have been exploited in many studies to analyze the convergence of optimization algorithms. For example, Luo and Tseng [29, 30, 31] established the asymptotic linear convergence of a class of feasible descent algorithms for smooth optimization, including coordinate descent method and projected gradient method, based on a local error bound condition. Their results on coordinate descent method were further extended for a more general class of objective functions and constraints in [51, 52]. Wang and Lin [53] showed that a global error bound holds for a family of non-strongly convex and smooth objective functions for which feasible descent methods can achieve a global linear convergence rate. Recently, these error bounds have been generalized and leveraged to show faster convergence for structured convex optimization that consists of a smooth function and a simple non-smooth function [24, 62, 63]. We would like to emphasize that the aforementioned error bounds are different from the local error bound explored in this paper. They bound the distance of a point to the optimal set by the norm of the projected gradient or proximal gradient at the point, thus requiring the smoothness of the objective function. In contrast, we bound the distance of a point to the optimal set by its objective residual with respect to the optimal value, covering a much broader family of functions. More recently, there appear many studies that consider smooth optimization or composite smooth optimization problems whose objective functions satisfy different error bound conditions, growth conditions or other non-degeneracy conditions and established the linear convergence rates of several first-order methods including proximal-gradient method, accelerated gradient method, prox-linear method and so on [22, 33, 61, 60, 25, 13, 12, 24, 63]. The relative strength and relationships between some of those conditions are studied by Necoara et al. [33] and Zhang [60]. For example, the authors in [33] showed that under the smoothness assumption the second order growth condition is equivalent to the error bound condition in [53]. It was brought to our attention that the local error bound condition in the present paper is closely related to metric subregularity of subdifferentials [1, 28, 14, 32]. Nevertheless, to our best knowledge, this is the first work that leverages the considered local error bound to improve the convergence of subgradient method for non-smooth and non-strongly convex optimization.

The aforementioned works assume the objective function is smooth or is a summation of a smooth function and a simple non-smooth function, which is not only critical for defining some error bounds they assumed but also necessary for obtaining the linear rates for the algorithms they studied. On the contrary, we focus on minimizing non-smooth objective functions with less structure using the SG method. Exploiting the error bound (16), we develop restarting scheme in the RSG method and obtain different convergence rates from the works mentioned above. Moreover, even without condition (16) or any other error bounds, we still obtain a new complexity that can be better than the standard complexity of SG method in some scenarios (see Corollary 6 and the remark thereafter.)

Gilpin et al [20] established a polyhedral error bound condition (presented in Lemma 9 below). Using this polyhedral error bound condition, they study a two-person zero-sum game and proposed an restarted first-order method based on Nesterov's smoothing technique [37] that can find the Nash equilibrium in a linear convergence rate. This error bound condition was shown in [20] to hold for the objective function whose epigraph is polyhedral and the domain is a bounded polytope [20]. Here, we slightly generalize their result to allow the domain to be an unbounded polyhedron which is the case for many important applications. In addition, we consider a general condition that contains this polyhedral error bound condition as a special case and we try to solve the general problem (1) rather than the bilinear saddle-point problem in [20].

In his recent work [44, 45, 46], Renegar presented a framework of applying first-order methods to general conic optimization problems by transforming the original problem into an equivalent convex optimization problem with only linear equality constraints and a Lipschitz-continuous objective function. This framework greatly extends the applicability of first-order methods to the problems with general linear inequality constraints and leads to new algorithms and new iteration complexity. One of his results related to this work is Corollary 3.4 of [45], which implies, if the objective function has a polyhedral epigraph and the optimal objective value is known beforehand, a subgradient method can have a linear convergence rate. Compared to his work, our method does not need to the know the optimal value but instead requires an upper bound of the initial optimality gap and knowing a growth constant in the error bound condition of the polyhedral objective function. In addition, our results include improved iteration complexities for a broader family of objective functions than the ones with a polyhedral epigraph.

More recently, Freund and Lu [16] proposed a new SG method by assuming that a strict lower bound of f_*, denoted by f_{slb}, is known and f satisfies a growth condition, $\|\mathbf{w}-\mathbf{w}^*\|_2 \leq \mathcal{G}\cdot(f(\mathbf{w})-f_{slb})$, where $\mathbf{w}^*$ is the optimal solution closest to $\mathbf{w}$ and $\mathcal{G}$ is a growth rate constant depending on f_{slb}. Using a novel step size that incorporates f_{slb}, for non-smooth optimization, their SG method achieves an iteration complexity of $O(\mathcal{G}^2(\frac{\log H}{\epsilon'}+\frac{1}{\epsilon'^2}))$ for finding a solution $\hat{\mathbf{w}}$ such that $f(\hat{\mathbf{w}})-f_* \leq \epsilon'(f_*-f_{slb})$, where $H=\frac{f(\mathbf{w}_0)-f_{slb}}{f_*-f_{slb}}$ and $\mathbf{w}_0$ is the initial solution. We note that there are several key differences in the theoretical properties and implementations between our work and [16]: (i) Their growth condition has a similar form to the inequality (7) we prove for a general function but there are still noticeable differences in the both sides and the growth constants. (ii) The convergence results in [16] are established based on finding an solution $\hat{\mathbf{w}}$ with a relative error of ϵ' while we consider absolute error. (iii) By rewriting the convergence results in [16] in terms of absolute accuracy ϵ with $\epsilon=\epsilon'(f_*-f_{slb})$, the complexity in [16] will strongly depend on f_*-f_{slb} and can be worse than ours if f_*-f_{slb} is large. (iv) Their SG descent method keeps track of the best solution up to the current iteration while our method maintains the average of all solutions within each epoch. We will compared our RSG method with the method in [16] with more details in Section 4.

Restarting and multi-stage strategies have been utilized to achieve the (uniformly) optimal theoretical complexity of (stochastic) SG methods when f is strongly convex [19, 10, 23] or uniformly convex [39]. Here, we show that restarting can be still helpful even without uniform or strong convexity. Furthermore, in all the algorithms proposed in [19, 10, 23, 39], the number of iterations per stage increases between stages while our algorithm uses the same number of iterations in all stages. This provides a different possibility of designing restarted algorithms for a better complexity.

3. Preliminaries. In this section, we define some notations used in this paper and present the main assumptions needed to establish our results. We use $\partial f(\mathbf{w})$ to denote the set of subgradients (the subdifferential) of f at $\mathbf{w}$. Let $\mathcal{G}(\cdot)$ denote a first-order oracle that returns a subgradient of $f(\cdot)$, namely, $\mathcal{G}(\mathbf{w}) \in \partial f(\mathbf{w})$ for any

$\mathbf{w} \in \Omega$. Since the objective function is not necessarily strongly convex, the optimal solution is not necessarily unique. We denote by Ω_* the optimal solution set and by f_* the unique optimal objective value. We denote by $\|\cdot\|_2$ the Euclidean norm in $\mathbb{R}^d$.

Throughout the paper, we make the following assumptions unless stated otherwise.

ASSUMPTION 1. *For the convex minimization problem (1), we assume*

a. *For any $\mathbf{w}_0 \in \Omega$, we know a constant $\epsilon_0 \geq 0$ such that $f(\mathbf{w}_0) - f_* \leq \epsilon_0$.*

b. *There exists a constant G such that $\|\mathcal{G}(\mathbf{w})\|_2 \leq G$ for any $\mathbf{w} \in \Omega$.*

c. *Ω_* is a non-empty convex compact set.*

We make several remarks about the above assumptions: (i) Assumption 1.a is equivalent to assuming we know a lower bound of f_* which is one of the assumptions made in [16]. In machine learning applications, f_* is usually bounded below by zero, i.e., $f_* \geq 0$, so that $\epsilon_0 = f(\mathbf{w}_0)$ for any $\mathbf{w}_0 \in \mathbb{R}^d$ will satisfy the condition; (ii) Assumption 1.b is a standard assumption also made in many previous subgradient-based methods. (iii) Assumption 1.c simply assumes the optimal set is closed and bounded.

Let $\mathbf{w}^*$ denote the closest optimal solution in Ω_* to $\mathbf{w}$ measured in terms of norm $\|\cdot\|_2$, i.e.,

$$\mathbf{w}^* := \arg\min_{\mathbf{u}\in\Omega_*} \|\mathbf{u} - \mathbf{w}\|_2^2.$$

Note that $\mathbf{w}^*$ is uniquely defined for any $\mathbf{w}$ due to the convexity of Ω_* and that $\|\cdot\|_2^2$ is strongly convex. We denote by $\mathcal{L}_\epsilon$ the ϵ-level set of $f(\mathbf{w})$ and by $\mathcal{S}_\epsilon$ the ϵ-sublevel set of $f(\mathbf{w})$, respectively, i.e.,

$$\mathcal{L}_\epsilon := \{\mathbf{w} \in \Omega : f(\mathbf{w}) = f_* + \epsilon\} \quad \text{and} \quad \mathcal{S}_\epsilon := \{\mathbf{w} \in \Omega : f(\mathbf{w}) \leq f_* + \epsilon\}. \tag{2}$$

Since Ω_* is assumed to be bounded and f is convex, it follows from [47] (Corollary 8.7.1) that the sublevel set $\mathcal{S}_\epsilon$ must be bounded for any $\epsilon \geq 0$ and so is the level set $\mathcal{L}_\epsilon$. Let B_ϵ be the maximum distance between the points in the ϵ-level set $\mathcal{L}_\epsilon$ and the optimal set Ω_*, i.e.,

$$B_\epsilon := \max_{\mathbf{w}\in\mathcal{L}_\epsilon} \min_{\mathbf{u}\in\Omega_*} \|\mathbf{w} - \mathbf{u}\|_2 = \max_{\mathbf{w}\in\mathcal{L}_\epsilon} \|\mathbf{w} - \mathbf{w}^*\|_2. \tag{3}$$

We follow the convention that $B_\epsilon = 0$ if $\mathcal{L}_\epsilon = \emptyset$. Because $\mathcal{L}_\epsilon$ and Ω_* are bounded, B_ϵ is finite. Let $\mathbf{w}_\epsilon^\dagger$ denote the closest point in the ϵ-sublevel set to $\mathbf{w}$, i.e.,

$$\mathbf{w}_\epsilon^\dagger := \arg\min_{\mathbf{u}\in\mathcal{S}_\epsilon} \|\mathbf{u} - \mathbf{w}\|_2^2 \tag{4}$$

Denote by $\Omega\backslash\mathcal{S} = \{\mathbf{w} \in \Omega : \mathbf{w} \notin \mathcal{S}\}$. It is easy to show that $\mathbf{w}_\epsilon^\dagger \in \mathcal{L}_\epsilon$ when $\mathbf{w} \in \Omega\backslash\mathcal{S}_\epsilon$ (using the optimality condition of (4)).

Given $\mathbf{w} \in \Omega$, we denote the normal cone of Ω at $\mathbf{w}$ by $\mathcal{N}_\Omega(\mathbf{w})$. Formally, $\mathcal{N}_\Omega(\mathbf{w}) = \{\mathbf{v} \in \mathbb{R}^d : \mathbf{v}^\top(\mathbf{u} - \mathbf{w}) \leq 0, \forall \mathbf{u} \in \Omega\}$. Define $\|\partial f(\mathbf{w}) + \mathcal{N}_\Omega(\mathbf{w})\|_2$ as

$$\|\partial f(\mathbf{w}) + \mathcal{N}_\Omega(\mathbf{w})\|_2 := \min_{\mathbf{g}\in\partial f(\mathbf{w}), \mathbf{v}\in\mathcal{N}_\Omega(\mathbf{w})} \|\mathbf{g} + \mathbf{v}\|_2. \tag{5}$$

Note that $\mathbf{w} \in \Omega_*$ if and only if $\|\partial f(\mathbf{w}) + \mathcal{N}_\Omega(\mathbf{w})\|_2 = 0$. Therefore, we call $\|\partial f(\mathbf{w}) + \mathcal{N}_\Omega(\mathbf{w})\|$ the *first-order optimality residual* of (1) at $\mathbf{w} \in \Omega$. Given any $\epsilon > 0$ such that $\mathcal{L}_\epsilon \neq \emptyset$, we define a constant ρ_ϵ as

$$\rho_\epsilon := \min_{\mathbf{w}\in\mathcal{L}_\epsilon} \|\partial f(\mathbf{w}) + \mathcal{N}_\Omega(\mathbf{w})\|_2. \tag{6}$$

Given the notations above, we provide the following lemma which is the key to our analysis.

LEMMA 2. *For any $\epsilon > 0$ such that $\mathcal{L}_\epsilon \neq \emptyset$ and any $\mathbf{w} \in \Omega$, we have*

$$\|\mathbf{w} - \mathbf{w}_\epsilon^\dagger\|_2 \leq \frac{1}{\rho_\epsilon}(f(\mathbf{w}) - f(\mathbf{w}_\epsilon^\dagger)). \tag{7}$$

Proof. Since the conclusion holds trivially if $\mathbf{w} \in \mathcal{S}_\epsilon$ (so that $\mathbf{w}_\epsilon^\dagger = \mathbf{w}$), we assume $\mathbf{w} \in \Omega \backslash \mathcal{S}_\epsilon$. According to the first-order optimality conditions of (4), there exist a scalar $\zeta \geq 0$ (the Lagrangian multiplier of the constraint $f(\mathbf{u}) \leq f_* + \epsilon$ in (4)), a subgradient $\mathbf{g} \in \partial f(\mathbf{w}_\epsilon^\dagger)$ and a vector $\mathbf{v} \in \mathcal{N}_\Omega(\mathbf{w}_\epsilon^\dagger)$ such that

$$\mathbf{w}_\epsilon^\dagger - \mathbf{w} + \zeta \mathbf{g} + \mathbf{v} = 0. \tag{8}$$

The definition of normal cone leads to $(\mathbf{w}_\epsilon^\dagger - \mathbf{w})^\top \mathbf{v} \geq 0$. This inequality and the convexity of $f(\cdot)$ imply

$$\zeta\left(f(\mathbf{w}) - f(\mathbf{w}_\epsilon^\dagger)\right) \geq \zeta(\mathbf{w} - \mathbf{w}_\epsilon^\dagger)^\top \mathbf{g} \geq (\mathbf{w} - \mathbf{w}_\epsilon^\dagger)^\top(\zeta \mathbf{g} + \mathbf{v}) = \|\mathbf{w} - \mathbf{w}_\epsilon^\dagger\|_2^2$$

where the equality is due to (8). Since $\mathbf{w} \in \Omega \backslash \mathcal{S}_\epsilon$, we must have $\|\mathbf{w} - \mathbf{w}_\epsilon^\dagger\|_2 > 0$ so that $\zeta > 0$. Therefore, $\mathbf{w}_\epsilon^\dagger \in \mathcal{L}_\epsilon$ by complementary slackness. Dividing the inequality above by ζ gives

$$f(\mathbf{w}) - f(\mathbf{w}_\epsilon^\dagger) \geq \frac{\|\mathbf{w} - \mathbf{w}_\epsilon^\dagger\|_2^2}{\zeta} = \|\mathbf{w} - \mathbf{w}_\epsilon^\dagger\|_2 \|\mathbf{g} + \mathbf{v}/\zeta\|_2 \geq \rho_\epsilon \|\mathbf{w} - \mathbf{w}_\epsilon^\dagger\|_2, \tag{9}$$

where the equality is due to (8) and the last inequality is due to the definition of ρ_ϵ in (6). The lemma is then proved. ∎

The inequality (7) takes a similar form to the growth condition,

$$\|\mathbf{w} - \mathbf{w}^*\|_2 \leq \mathcal{G} \cdot (f(\mathbf{w}) - f_{slb}), \tag{10}$$

where f_{slb} is a strict lower bound of f^*, by Freund and Lu [16] but with some striking differences: the left-hand side is the distance of $\mathbf{w}$ to the optimal set in (10) while it is the distance of $\mathbf{w}$ to the ϵ-sublevel set in (7); the right-hand side is the objective gap with respect to f_{slb} in (10) and it is the objective gap with respect to f^* in (7); the growth constant $\mathcal{G}$ in (10) varies with f_{slb} and ρ_ϵ in (7) may depend on ϵ in general.

The inequality in (7) is the key to achieve improved convergence by RSG, which hinges on the condition that the first-order optimality residual on the ϵ-level set is lower bounded. It is important to note that (i) the above result depends on f rather than the optimization algorithm applied; and (ii) the above result can be generalized to use other norm such as the p-norm $\|\mathbf{w}\|_p$ ($p \in (1, 2]$) to measure the distance between $\mathbf{w}$ and $\mathbf{w}_\epsilon^\dagger$ (Section 6) and use the corresponding dual norm to define the lower bound of the residual in (5) and (6). This generalization allows us to design mirror decent [34] variant of RSG. We defer the details into Section 6. To our best knowledge, this is the first work that leverages the lower bound of the optimal residual to improve the convergence for non-smooth convex optimization.

In next several sections, we will exhibit the value of ρ_ϵ for different classes of problems and discuss its impact on the convergence.

4. Restarted SubGradient (RSG) Method and Its Complexity for General Problem. In this section, we present the proposed restarted subgradient (RSG) method and prove its general convergence result using Lemma 2. In next sections, we will present improved convergence of RSG for problems of different classes.

Algorithm 1 SG: $\widehat{\mathbf{w}}_T = \mathrm{SG}(\mathbf{w}_1, \eta, T)$

Input: a step size η, the number of iterations T, and the initial solution $\mathbf{w}_1 \in \Omega$
for $t = 1, \ldots, T$ **do**
  Query the subgradient oracle to obtain $\mathcal{G}(\mathbf{w}_t)$
  Update $\mathbf{w}_{t+1} = \Pi_\Omega[\mathbf{w}_t - \eta\mathcal{G}(\mathbf{w}_t)]$
end for
Output: $\widehat{\mathbf{w}}_T = \sum_{t=1}^{T} \frac{\mathbf{w}_t}{T}$

Algorithm 2 RSG: $\mathbf{w}_K = \mathrm{RSG}(\mathbf{w}_0, K, t, \alpha)$

Input: the number of stages K and the number of iterations t per-stage, $\mathbf{w}_0 \in \Omega$, and $\alpha > 1$.
Set $\eta_1 = \epsilon_0/(\alpha G^2)$, where ϵ_0 is from Assumption 1.a
for $k = 1, \ldots, K$ **do**
  Call subroutine SG to obtain $\mathbf{w}_k = \mathrm{SG}(\mathbf{w}_{k-1}, \eta_k, t)$
  Set $\eta_{k+1} = \eta_k/\alpha$
end for
Output: $\mathbf{w}_K$

The steps of RSG are presented in Algorithm 2 where SG is a subroutine of projected subgradient descent given in Algorithm 1 and $\Pi_\Omega[\mathbf{w}]$ is defined as

$$\Pi_\Omega[\mathbf{w}] = \arg\min_{\mathbf{u}\in\Omega} \|\mathbf{u} - \mathbf{w}\|_2^2.$$

The values of K and t in RSG will be revealed later for proving the convergence of RSG to an 2ϵ-optimal solution. The RSG algorithm runs in stages and calls SG once in each stage. The subroutine SG performs projected subgradient descent with a fixed step size η for a **fixed number of iterations** t, using the solution returned by SG from the previous stage of RSG as the starting point. The RSG algorithm geometrically decreases the step size η_k between stages. The output solution of RSG is the solution returned by SG in the K-th stage, i.e., $\mathbf{w}_K$. The number of iterations t is the only varying parameter in RSG that depends on the classes of problems. The parameter α could be any value larger than 1 (e.g., 2) and it only has a small influence on the iteration complexity.

We emphasize that (i) RSG is a generic algorithm that is applicable to a broad family of non-smooth and/or non-strongly convex problems without changing updating schemes except for one tuning parameter, the number of iterations per stage, whose best value varies with problems; (ii) RSG has different variants with different subroutine in stages. In fact, we can use other optimization algorithms than subgradient descent as the subroutine in Algorithm 2, as long as a similar convergence result to Lemma 3 is guaranteed. Examples include SG as shown in Algorithm 1, dual averaging [38], and the regularized dual averaging [10] [2]. In the following discussions, we will focus on using SG as the subroutine unless stated otherwise.

It was also brought to our attention that the current common practice for training deep neural network (yet with no formal justification) is very similar to the proposed RSG. In particular, one usually employs stochastic subgradient with momentum (with similar convergence guarantee to SG for non-smooth optimization [56]) with a fixed

[2]Without assuming smoothness and strong convexity.

step size by a certain number of iterations and then decreases the step size by a certain number of times [27].

Next, we establish the convergence of RSG. It relies on the convergence result of the SG subroutine which is given in the lemma below.

LEMMA 3. *[64, 36] If Algorithm 1 runs for T iterations, we have, for any $\mathbf{w} \in \Omega$,*

$$f(\widehat{\mathbf{w}}_T) - f(\mathbf{w}) \leq \frac{G^2\eta}{2} + \frac{\|\mathbf{w}_1 - \mathbf{w}\|_2^2}{2\eta T}.$$

We omit the proof because it follows a standard analysis and can be found in cited papers. With the above lemma, we can prove the following convergence of RSG. We would like to emphasize here the difference between our analysis of convergence for RSG and previous analysis of SG. In the previous analysis, $\mathbf{w}$ is set to an optimal solution $\mathbf{w}_*$ in Lemma 3. Assuming $\|\mathbf{w}_1 - \mathbf{w}_*\|_2 \leq D$, one can optimize η to obtain an $O(GD/\sqrt{T})$ convergence rate for SG. In our analysis, we leverage Lemma 2 to bound $\|\mathbf{w}_{k-1} - \mathbf{w}\|_2$ for the k-th stage and prove the convergence of RSG by induction.

THEOREM 4. *Suppose Assumption 1 holds. If $t \geq \frac{\alpha^2 G^2}{\rho_\epsilon^2}$ and $K = \lceil \log_\alpha(\frac{\epsilon_0}{\epsilon}) \rceil$ in Algorithm 2, with at most K stages, Algorithm 2 returns a solution $\mathbf{w}_K$ such that $f(\mathbf{w}_K) - f_* \leq 2\epsilon$. In other word, the total number of iterations for Algorithm 2 to find an 2ϵ-optimal solution is at most $T = O\left(t\lceil \log_\alpha(\frac{\epsilon_0}{\epsilon})\rceil\right)$ where $t \geq \frac{\alpha^2 G^2}{\rho_\epsilon^2}$. In particular, if $\frac{\alpha^2 G^2}{\rho_\epsilon^2} \leq t = O\left(\frac{\alpha^2 G^2}{\rho_\epsilon^2}\right)$, the total number of iterations for Algorithm 2 to find an 2ϵ-optimal solution is at most $T = O\left(\frac{\alpha^2 G^2}{\rho_\epsilon^2}\lceil \log_\alpha(\frac{\epsilon_0}{\epsilon})\rceil\right)$.*

Proof. Let $\mathbf{w}_{k,\epsilon}^\dagger$ denote the closest point to $\mathbf{w}_k$ in the ϵ-sublevel set. Let $\epsilon_k := \frac{\epsilon_0}{\alpha^k}$ so that $\eta_k = \epsilon_k/G^2$ because $\eta_1 = \epsilon_0/(\alpha G^2)$ and $\eta_{k+1} = \eta_k/\alpha$. We will show by induction that

$$f(\mathbf{w}_k) - f_* \leq \epsilon_k + \epsilon \tag{11}$$

for $k = 0, 1, \ldots, K$ which leads to our conclusion if we let $k = K$.

Note that (11) holds obviously for $k = 0$. Suppose it holds for $k - 1$, namely, $f(\mathbf{w}_{k-1}) - f_* \leq \epsilon_{k-1} + \epsilon$. We want to prove (11) for k. We apply Lemma 3 to the k-th stage of Algorithm 2 and get

$$f(\mathbf{w}_k) - f(\mathbf{w}_{k-1,\epsilon}^\dagger) \leq \frac{G^2\eta_k}{2} + \frac{\|\mathbf{w}_{k-1} - \mathbf{w}_{k-1,\epsilon}^\dagger\|_2^2}{2\eta_k t}. \tag{12}$$

We now consider two cases for $\mathbf{w}_{k-1}$. First, assume $f(\mathbf{w}_{k-1}) - f_* \leq \epsilon$, i.e., $\mathbf{w}_{k-1} \in \mathcal{S}_\epsilon$. Then $\mathbf{w}_{k-1,\epsilon}^\dagger = \mathbf{w}_{k-1}$ and $f(\mathbf{w}_k) - f(\mathbf{w}_{k-1,\epsilon}^\dagger) \leq \frac{G^2\eta_k}{2} = \frac{\epsilon_k}{2}$. As a result,

$$f(\mathbf{w}_k) - f_* \leq f(\mathbf{w}_{k-1,\epsilon}^\dagger) - f_* + \frac{\epsilon_k}{2} \leq \epsilon + \epsilon_k$$

Next, we consider the case that $f(\mathbf{w}_{k-1}) - f_* > \epsilon$, i.e., $\mathbf{w}_{k-1} \notin \mathcal{S}_\epsilon$. Then we have $f(\mathbf{w}_{k-1,\epsilon}^\dagger) = f_* + \epsilon$. By Lemma 2, we have

$$\begin{aligned}\|\mathbf{w}_{k-1} - \mathbf{w}_{k-1,\epsilon}^\dagger\|_2 &\leq \frac{1}{\rho_\epsilon}(f(\mathbf{w}_{k-1}) - f(\mathbf{w}_{k-1,\epsilon}^\dagger)) = \frac{f(\mathbf{w}_{k-1}) - f_* + (f_* - f(\mathbf{w}_{k-1,\epsilon}^\dagger))}{\rho_\epsilon} \\ &\leq \frac{\epsilon_{k-1} + \epsilon - \epsilon}{\rho_\epsilon}.\end{aligned} \tag{13}$$

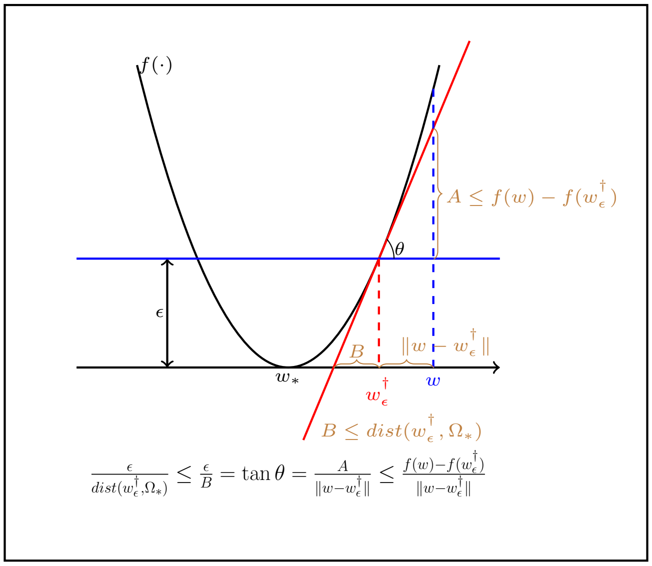


FIG. 1. *A geometric illustration of the inequality (14), where* $dist(w_\epsilon^\dagger, \Omega_*) = |w_\epsilon^\dagger - w_\epsilon^*|$.

Combining (12) and (13) and using the facts that $\eta_k = \frac{\epsilon_k}{G^2}$ and $t \ge \frac{\alpha^2 G^2}{\rho_\epsilon^2}$, we have

$$f(\mathbf{w}_k) - f(\mathbf{w}_{k-1,\epsilon}^\dagger) \le \frac{\epsilon_k}{2} + \frac{\epsilon_{k-1}^2}{2\epsilon_k \alpha^2} = \epsilon_k$$

which, together with the fact that $f(\mathbf{w}_{k-1,\epsilon}^\dagger) = f_* + \epsilon$, implies (11) for k. Therefore, by induction, we have (11) holds for $k = 1, 2, \ldots, K$ so that

$$f(\mathbf{w}_K) - f_* \le \epsilon_K + \epsilon = \frac{\epsilon_0}{\alpha^K} + \epsilon \le 2\epsilon,$$

where the last inequality is due to the definition of K. ◻

In Theorem 4, the iteration complexity of RSG for the general problem (1) is given in terms of ρ_ϵ. Next, we show that $\rho_\epsilon \ge \frac{\epsilon}{B_\epsilon}$ which allows us to choose the number of iterations t in each stage according to B_ϵ instead of ρ_ϵ so that corresponding complexity can be given in terms of B_ϵ.

LEMMA 5. *For any* $\epsilon > 0$ *such that* $\mathcal{L}_\epsilon \ne \emptyset$*, we have* $\rho_\epsilon \ge \frac{\epsilon}{B_\epsilon}$*, where* B_ϵ *is defined in (3), and for any* $\mathbf{w} \in \Omega$

$$(14) \qquad \|\mathbf{w} - \mathbf{w}_\epsilon^\dagger\|_2 \le \frac{\|\mathbf{w}_\epsilon^\dagger - \mathbf{w}_\epsilon^*\|_2}{\epsilon}(f(\mathbf{w}) - f(\mathbf{w}_\epsilon^\dagger)) \le \frac{B_\epsilon}{\epsilon}(f(\mathbf{w}) - f(\mathbf{w}_\epsilon^\dagger)),$$

where $\mathbf{w}_\epsilon^*$ *is the closest point in* Ω_* *to* $\mathbf{w}_\epsilon^\dagger$.

Proof. Given any $\mathbf{u} \in \mathcal{L}_\epsilon$, let $\mathbf{g}_\mathbf{u}$ be any subgradient in $\partial f(\mathbf{u})$ and $\mathbf{v}_\mathbf{u}$ be any vector in $\mathcal{N}_\Omega(\mathbf{u})$. By the convexity of $f(\cdot)$ and the definition of normal cone, we have

$$f(\mathbf{u}^*) - f(\mathbf{u}) \ge (\mathbf{u}^* - \mathbf{u})^\top \mathbf{g}_\mathbf{u} \ge (\mathbf{u}^* - \mathbf{u})^\top (\mathbf{g}_\mathbf{u} + \mathbf{v}_\mathbf{u}),$$

where $\mathbf{u}^*$ is the closest point in Ω_* to $\mathbf{u}$. This inequality further implies

$$(15) \quad \|\mathbf{u}^* - \mathbf{u}\|_2 \|\mathbf{g}_\mathbf{u} + \mathbf{v}_\mathbf{u}\|_2 \ge f(\mathbf{u}) - f(\mathbf{u}^*) = \epsilon, \quad \forall \mathbf{g}_\mathbf{u} \in \partial f(\mathbf{u}) \text{ and } \mathbf{v}_\mathbf{u} \in \mathcal{N}_\Omega(\mathbf{u})$$

where the equality is because $\mathbf{u} \in \mathcal{L}_\epsilon$. By (15) and the definition of B_ϵ, we obtain

$$B_\epsilon \|\mathbf{g_u} + \mathbf{v_u}\|_2 \geq \epsilon \Longrightarrow \|\mathbf{g_u} + \mathbf{v_u}\|_2 \geq \epsilon / B_\epsilon.$$

Since $\mathbf{g_u} + \mathbf{v_u}$ can be any element in $\partial f(\mathbf{u}) + \mathcal{N}_\Omega(\mathbf{u})$, we have $\rho_\epsilon \geq \frac{\epsilon}{B_\epsilon}$ by the definition (6).

To prove (14), we assume $\mathbf{w} \in \Omega \backslash \mathcal{S}_\epsilon$ and thus $\mathbf{w}_\epsilon^\dagger \in \mathcal{L}_\epsilon$; otherwise it is trivial. In the proof of Lemma 2, we have shown that (see (9)) there exists $\mathbf{g} \in \partial f(\mathbf{w}_\epsilon^\dagger)$ and $\mathbf{v} \in \mathcal{N}_\Omega(\mathbf{w}_\epsilon^\dagger)$ such that $f(\mathbf{w}) - f(\mathbf{w}_\epsilon^\dagger) \geq \|\mathbf{w} - \mathbf{w}_\epsilon^\dagger\|_2 \|\mathbf{g} + \mathbf{v}/\zeta\|_2$, which, according to (15) with $\mathbf{u} = \mathbf{w}_\epsilon^\dagger$, $\mathbf{g_u} = \mathbf{g}$ and $\mathbf{v_u} = \mathbf{v}/\zeta$, leads to (14). ∎

A geometric explanation of the inequality (14) in one dimension is shown in Figure 1.

With Lemma 5, the iteration complexity of RSG can be stated in terms of B_ϵ in the following corollary of Theorem 4.

COROLLARY 6. *Suppose Assumption 1 holds. The iteration complexity of RSG for obtaining an 2ϵ-optimal solution is $O(\frac{\alpha^2 G^2 B_\epsilon^2}{\epsilon^2}\lceil \log_\alpha(\frac{\epsilon_0}{\epsilon})\rceil)$ provided $\frac{\alpha^2 G^2 B_\epsilon^2}{\epsilon^2} \leq t = O\left(\frac{\alpha^2 G^2 B_\epsilon^2}{\epsilon^2}\right)$ and $K = \lceil \log_\alpha(\frac{\epsilon_0}{\epsilon})\rceil$.*

Remark: Compared to the standard SG whose iteration complexity is known as $O(\frac{G^2\|\mathbf{w}_0 - \mathbf{w}_0^*\|_2^2}{\epsilon^2})$ for achieving an 2ϵ-optimal solution, RSG's iteration complexity in Corollary 6 depends on B_ϵ^2 instead of $\|\mathbf{w}_0 - \mathbf{w}_0^*\|_2^2$ and only has a logarithmic dependence on ϵ_0, the upper bound of $f(\mathbf{w}_0) - f_*$. When the initial solution is far from the optimal set so that $B_\epsilon^2 \ll \|\mathbf{w}_0 - \mathbf{w}_0^*\|_2^2$, the proposed RSG can be much faster. In some special cases, e.g., when f satisfies the local error bound condition (16) so that $B_\epsilon = \Theta(\epsilon^\theta)$ with $\theta \in (0, 1]$, RSG only needs $O\left(\frac{1}{\epsilon^{2(1-\theta)}}\log\left(\frac{1}{\epsilon}\right)\right)$ iterations (see Section 5.1), which as a better dependency on ϵ than the complexity of standard SG method.

Remark: Compared to the SG method in [16] whose iteration complexity is $O(G^2\mathcal{G}^2(\frac{\log H}{\epsilon'} + \frac{1}{\epsilon'^2}))$ for finding a solution $\hat{\mathbf{w}}$ such that $f(\hat{\mathbf{w}}) - f_* \leq \epsilon'(f_* - f_{slb})$, where f_{slb} and $\mathcal{G}$ are defined in (10) and $H = \frac{f(\mathbf{w}_0) - f_{slb}}{f_* - f_{slb}}$, our RSG can be better if $f_* - f_{slb}$ is large. To see this, we represent the complexity in [16] in terms of the absolute error ϵ with $\epsilon = \epsilon'(f_* - f_{slb})$ and obtain $O(G^2\mathcal{G}^2(\frac{(f_* - f_{slb})\log H}{\epsilon} + \frac{(f_* - f_{slb})^2}{\epsilon^2}))$. If the gap $f_* - f_{slb}$ is large, e.g., $O(f(\mathbf{w}_0) - f_{slb})$, the second term is dominating, which is at least $O(\frac{G^2\|\mathbf{w}_0 - \mathbf{w}_0^*\|_2^2}{\epsilon^2})$ due to the definition of $\mathcal{G}$ in (10). This complexity has the same order of magnitude as the standard SG method so that RSG can be better for the reason in the last remark. When the gap $f_* - f_{slb}$ is small, e.g. $O(\epsilon)$, the first term is dominating, which is $O(G^2\mathcal{G}^2\log(\frac{\epsilon_0}{\epsilon}))$. In this case, RSG has a lower complexity when $\mathcal{G}^2 \geq \frac{B_\epsilon^2}{\epsilon^2}$. Note that, in general, $\mathcal{G}$ is non-increasing in $f_* - f_{slb}$ but the exact dependency is not clear.

However, compare to the standard SG and the method in [16], RSG does require knowing additional information about f. In particular, the issue with RSG is that its improved complexity is obtained by choosing $t = O\left(\frac{\alpha^2 G^2 B_\epsilon^2}{\epsilon^2}\right)$ which requires knowing the order of magnitude of B_ϵ, if not its exact value. To address the issue of unknown B_ϵ for general problems, in the next section, we consider the family of problems that admit a local error bound and show that the requirement of knowing B_ϵ is relaxed to knowing some particular parameters related to the local error bound.

5. Complexity for Some Classes of Non-smooth Non-strongly Convex Optimization. In this section, we consider a particular family of problems that admit local error bounds and show the improved iteration complexities of RSG compared to standard SG method.

5.1. Complexity for the Problems with Local Error Bounds. We first define local error bound of the objective function.

DEFINITION 7. *We say $f(\cdot)$ admits a* ***local error bound*** *on the ϵ-sublevel set $\mathcal{S}_\epsilon$ if*

$$\|\mathbf{w}-\mathbf{w}^*\|_2 \leq c(f(\mathbf{w})-f_*)^\theta, \quad \forall \mathbf{w}\in\mathcal{S}_\epsilon, \tag{16}$$

where $\mathbf{w}^*$ is the closet point in Ω_* to $\mathbf{w}$, $\theta\in(0,1]$ and $c>0$ are constants.

Because $\mathcal{S}_{\epsilon_2}\subset\mathcal{S}_{\epsilon_1}$ for $\epsilon_2\leq\epsilon_1$, if (16) holds for some ϵ, it will always hold when ϵ decreases to zero with the same θ and c, which is the case we are interested in.

If the problem admits a local error bound like (16), RSG can achieve a better iteration complexity than $\widetilde{O}(1/\epsilon^2)$. In particular, the property (16) implies

$$B_\epsilon \leq c\epsilon^\theta. \tag{17}$$

Replacing B_ϵ in Corollary 6 by this upper bound and choosing $t=\frac{\alpha^2G^2c^2}{\epsilon^{2(1-\theta)}}$ in RSG if c and θ are known, we obtain the following complexity of RSG.

COROLLARY 8. *Suppose Assumption 1 holds and $f(\cdot)$ admits a local error bound on $\mathcal{S}_\epsilon$. The iteration complexity of RSG for obtaining an 2ϵ-optimal solution is $O\left(\frac{\alpha^2G^2c^2}{\epsilon^{2(1-\theta)}}\log_\alpha\left(\frac{\epsilon_0}{\epsilon}\right)\right)$ provided $t=\frac{\alpha^2G^2c^2}{\epsilon^{2(1-\theta)}}$ and $K=\lceil\log_\alpha(\frac{\epsilon_0}{\epsilon})\rceil$.*

Juditsky and Nesterov [39] considered subgradient methods for (1) with f being *uniformly convex*, namely,

$$f(\alpha\mathbf{w}+(1-\alpha)\mathbf{v})\leq\alpha f(\mathbf{w})+(1-\alpha)f(\mathbf{v})-\frac{1}{2}\mu\alpha(1-\alpha)[\alpha^{\rho-1}+(1-\alpha)^{\rho-1}]\|\mathbf{w}-\mathbf{v}\|_2^\rho$$

for any $\mathbf{w}$ and $\mathbf{v}$ in Ω and any $\alpha\in[0,1]$[3], where $\rho\in[2,+\infty]$ and $\mu\geq0$. In this case, the method by [39] has an iteration complexity of $O\left(\frac{G^2}{\mu^{2/\rho}\epsilon^{2(\rho-1)/\rho}}\right)$. The uniform convexity of f further implies $f(\mathbf{w})-f_*\geq\frac{1}{2}\mu\|\mathbf{w}-\mathbf{w}^*\|_2^\rho$ for any $\mathbf{w}\in\Omega$ so that $f(\cdot)$ admits a local error bound on the ϵ-sublevel set $\mathcal{S}_\epsilon$ with $c=\left(\frac{2}{\mu}\right)^{\frac{1}{\rho}}$ and $\theta=\frac{1}{\rho}$. Therefore, our RSG has a complexity of $O\left(\frac{G^2}{\mu^{2/\rho}\epsilon^{2(\rho-1)/\rho}}\log(\frac{\epsilon_0}{\epsilon})\right)$ according to Corollary 8. Compared to [39], our complexity is higher by a logarithmic factor. However, we only require the local error bound property of f that is weaker than uniform convexity.

Next, we will consider different convex optimization problems that admit a local error bound on $\mathcal{S}_\epsilon$ with different c and θ and show the faster convergence of RSG when applied to these problems.

5.2. Linear Convergence for Polyhedral Convex Optimization. In this subsection, we consider a special family of non-smooth and non-strongly convex problems where the epigraph of $f(\cdot)$ over Ω is a polyhedron. In this case, we call (1) a **polyhedral convex minimization** problem. We show that, in polyhedral convex minimization problem, $f(\cdot)$ has a linear growth property and admits a local error bound with $\theta=1$ so that $B_\epsilon\leq c\epsilon$ for a constant c.

[3]The Euclidean norm in the definition here can be replaced by a general norm as in [39].

Lemma 9 (Polyhedral Error Bound Condition). *Suppose Ω is a polyhedron and the epigraph of $f(\cdot)$ is also polyhedron. There exists a constant $\kappa > 0$ such that*

$$\|\mathbf{w} - \mathbf{w}^*\|_2 \leq \frac{f(\mathbf{w}) - f_*}{\kappa}, \quad \forall \mathbf{w} \in \Omega.$$

Thus, $f(\cdot)$ admits a local error bound on $\mathcal{S}_\epsilon$ with $\theta = 1$ and $c = \frac{1}{\kappa}$ [4] *(so $B_\epsilon \leq \frac{\epsilon}{\kappa}$) for any $\epsilon > 0$.*

Remark: We remark that the above result can be extended to any valid norm to measure the distance between $\mathbf{w}$ and $\mathbf{w}_*$. The proof is included in the appendix. Lemma 9 above generalizes Lemma 4 by Gilpin et al [20], which requires Ω to be a bounded polyhedron, to a similar result where Ω can be an unbounded polyhedron. This generalization is simple but useful because it helps the development of efficient algorithms based on this error bound for unconstrained problems without artificially including a box constraint.

Different from [20] that used their Lemma 4 and Nesterov's smoothing technique [37] to develop a linearly convergent algorithm for solving the Nash equilibrium of a two-person zero-sum games, we show below that Lemma 9 provides the basis for RSG to achieve a linear convergence for the polyhedral convex minimization problems. In fact, the following linear convergence of RSG can be obtained if we plugin the values of $\theta = 1$ and $c = \frac{1}{\kappa}$ into Corollary 8.

Corollary 10. *Suppose Assumption 1 holds and* (1) *is a polyhedral convex minimization problem. The iteration complexity of RSG for obtaining an ϵ-optimal solution is $O(\frac{\alpha^2 G^2}{\kappa^2}\lceil\log_\alpha(\frac{\epsilon_0}{\epsilon})\rceil)$ provided $t = \frac{\alpha^2 G^2}{\kappa^2}$ and $K = \lceil\log_\alpha(\frac{\epsilon_0}{\epsilon})\rceil$.*

We want to point out that Corollary 10 can be proved directly by replacing $\mathbf{w}^\dagger_{k-1,\epsilon}$ by $\mathbf{w}^*_{k-1}$ and replacing ρ_ϵ by κ in the proof of Theorem 4. Here, we derive it as a corollary of a more general result. We also want to mention that, as shown by Renegar [45], the linear convergence rate in Corollary 10 can be also obtained by the SG method for the historically best solution, provided either κ or f^* is known.

Examples. Many non-smooth and non-strongly convex machine learning problems satisfy the assumptions of Corollary 10, for example, ℓ_1 **or** ℓ_∞ **constrained or regularized piecewise linear loss minimization**. In many machine learning tasks (e.g., classification and regression), there exists a set of data $\{(\mathbf{x}_i, y_i)\}_{i=1,2,\ldots,n}$ and one often needs to solve the following empirical risk minimization problem

$$\min_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w}) \triangleq \frac{1}{n}\sum_{i=1}^n \ell(\mathbf{w}^\top\mathbf{x}_i, y_i) + R(\mathbf{w}),$$

where $R(\mathbf{w})$ is a regularization term and $\ell(z, y)$ denotes a loss function. We consider a special case where (a) $R(\mathbf{w})$ is a ℓ_1 regularizer, ℓ_∞ regularizer or an indicator function of a ℓ_1/ℓ_∞ ball centered at zero; and (b) $\ell(z, y)$ is any piecewise linear loss function, including hinge loss $\ell(z, y) = \max(0, 1-yz)$, absolute loss $\ell(z, y) = |z-y|$, ϵ-insensitive loss $\ell(z, y) = \max(|z-y| - \epsilon, 0)$, and etc [58]. It is easy to show that the epigraph of $f(\mathbf{w})$ is a polyhedron if $f(\mathbf{w})$ is defined as a sum of any of these regularization terms and any of these loss functions. In fact, a piecewise linear loss functions can be generally written as

$$\ell(\mathbf{w}^\top\mathbf{x}, y) = \max_{1\leq j\leq m} a_j\mathbf{w}^\top\mathbf{x} + b_j, \tag{18}$$

[4] In fact, this property of $f(\cdot)$ is a global error bond on Ω.

where (a_j, b_j) for $j = 1, 2, \ldots, m$ are finitely many pairs of scalars. The formulation (18) indicates that $\ell(\mathbf{w}^\top \mathbf{x}, y)$ is a piecewise affine function so that its epigraph is a polyhedron. In addition, the ℓ_1 or ℓ_∞ norm is also a polyhedral function because we can represent them as

$$\|\mathbf{w}\|_1 = \sum_{i=1}^{d} \max(w_i, -w_i), \quad \|\mathbf{w}\|_\infty = \max_{1\leq i\leq d} |w_i| = \max_{1\leq i\leq d} \max(w_i, -w_i).$$

Since the sum of finitely many polyhedral functions is also a polyhedral function, the epigraph of $f(\mathbf{w})$ is a polyhedron.

Another important family of problems whose objective function has a polyhedral epigraph is **submodular function minimization**. Let $V = \{1, \ldots, d\}$ be a set and 2^V denote its power set. A submodular function $F(A) : 2^V \to \mathbb{R}$ is a set function such that $F(A) + F(B) \geq F(A \cup B) + F(A \cap B)$ for all subsets $A, B \subseteq V$ and $F(\emptyset) = 0$. A submodular function minimization can be cast into a non-smooth convex optimization using the Lovász extension [4]. In particular, let the base polyhedron $B(F)$ be defined as

$$B(F) = \{\mathbf{s} \in \mathbb{R}^d, \mathbf{s}(V) = F(V), \forall A \subseteq V, \mathbf{s}(A) \leq F(A)\},$$

where $\mathbf{s}(A) = \sum_{i\in A} s_i$. Then the Lovász extension of $F(A)$ is $f(\mathbf{w}) = \max_{\mathbf{s}\in B(F)} \mathbf{w}^\top \mathbf{s}$, and $\min_{A\subseteq V} F(A) = \min_{\mathbf{w}\in[0,1]^d} f(\mathbf{w})$. As a result, a submodular function minimization is essentially a non-smooth and non-strongly convex optimization with a polyhedral epigraph.

5.3. Improved Convergence for Locally Semi-Strongly Convex Problems. First, we give a definition of local semi-strong convexity.

DEFINITION 11. *A function $f(\mathbf{w})$ is semi-strongly convex on the ϵ-sublevel set $\mathcal{S}_\epsilon$ if there exists $\lambda > 0$ such that*

$$\frac{\lambda}{2}\|\mathbf{w} - \mathbf{w}^*\|_2^2 \leq f(\mathbf{w}) - f(\mathbf{w}^*), \quad \forall \mathbf{w} \in \mathcal{S}_\epsilon \tag{19}$$

where $\mathbf{w}^*$ is the closest point to $\mathbf{w}$ in the optimal set.

We refer to the property (19) as *local* semi-strong convexity when $\mathcal{S}_\epsilon \neq \Omega$. The two papers [22, 33] have explored the semi-strong convexity on the whole domain Ω to prove linear convergence of smooth optimization problems. In [33], the inequality (19) is also called **second-order growth property**. They have also shown that a class of problems satisfy (19) (see examples given below). The inequality (19) indicates that $f(\cdot)$ admits a local error bound on $\mathcal{S}_\epsilon$ with $\theta = \frac{1}{2}$ and $c = \sqrt{\frac{2}{\lambda}}$, which leads to the following the corollary about the iteration complexity of RSG for locally semi-strongly convex problems.

COROLLARY 12. *Suppose Assumption 1 holds and $f(\mathbf{w})$ is semi-strongly convex on $\mathcal{S}_\epsilon$. Then $B_\epsilon \leq \sqrt{\frac{2\epsilon}{\lambda}}$ [5] and the iteration complexity of RSG for obtaining an 2ϵ-optimal solution is $O(\frac{2\alpha^2G^2}{\lambda\epsilon}\lceil\log_\alpha(\frac{\epsilon_0}{\epsilon})\rceil)$ provided $t = \frac{2\alpha^2G^2}{\lambda\epsilon}$ and $K = \lceil\log_\alpha(\frac{\epsilon_0}{\epsilon})\rceil$.*

Remark: To the best of our knowledge, the previous subgradient methods can achieve the $O(1/\epsilon)$ iteration complexity only by assuming strong convexity [18, 19, 10, 23]. Here, we obtain an $\widetilde{O}(1/\epsilon)$ iteration complexity ($\widetilde{O}(\cdot)$ suppresses constants and logarithmic terms) only with local semi-strong convexity. It is obvious that strong convexity implies local semi-strong convexity [23] but not vice versa.

[5] Recall (17).

Examples. Consider a family of functions in the form of $f(\mathbf{w}) = h(X\mathbf{w}) + r(\mathbf{w})$, where $X \in \mathbb{R}^{n\times d}$, $h(\cdot)$ is smooth and *strongly convex* on any *compact set* and $r(\cdot)$ has a polyhedral epigraph. According to [22, 33], such a function $f(\mathbf{w})$ satisfies (19) for any $\epsilon \le \epsilon_0$ with a constant value for λ. Although smoothness is assumed for $h(\cdot)$ in [22, 33], we find that it is not necessary for proving (19). We state this result as the lemma below.

LEMMA 13. *Suppose Assumption 1 holds,* $\Omega = \{\mathbf{w} \in \mathbb{R}^d | C\mathbf{w} \le \mathbf{b}\}$ *with* $C \in \mathbb{R}^{k\times d}$ *and* $\mathbf{b} \in \mathbb{R}^k$*, and* $f(\mathbf{w}) = h(X\mathbf{w}) + r(\mathbf{w})$ *where* $h : \mathbb{R}^n \to \mathbb{R}$ *satisfies dom*$(h) = \mathbb{R}^k$ *and is a strongly convex function on any compact set in* $\mathbb{R}^n$*, and* $r(\mathbf{w})$ *has a polyhedral epigraph. Then,* $f(w)$ *satisfies (19) for any* $\epsilon \le \epsilon_0$*.*

Proof. The proof of this lemma is almost identical to the proof of Lemma 1 in [22] which assumes $h(\cdot)$ is smooth. Here, we show that a similar result holds without the smoothness of $h(\cdot)$.

Since $r(\cdot)$ has a polyhedral epigraph, we can assume $r(\mathbf{w}) = \mathbf{q}^T\mathbf{w}$ for some $q \in \mathbb{R}^d$ without loss of generality. In fact, when $r(\cdot)$ has a polyhedral epigraph and $\Omega = \mathbb{R}^d$, we can introduce a new variable $w \in \mathbb{R}$ so that (1) can be equivalently represented as

$$\min_{(\mathbf{w},w)\in\mathbb{R}^{d+1}} \quad h(X\mathbf{w}) + w$$
$$\text{s.t. } \mathbf{w} \in \Omega, r(\mathbf{w}) \le w,$$

where feasible set is a polyhedron in $\mathbb{R}^{d+1}$ and the corresponding $r(\cdot)$ becomes linear.

Since $h(\cdot)$ is a strongly convex function on any compact set, following a standard argument (see [22] for example), we can show that there exist $\mathbf{r}^* \in \mathbb{R}^n$ and $s^* \in \mathbb{R}$ such that

$$\Omega_* = \{\mathbf{w} \in \mathbb{R}^d | X\mathbf{w} = \mathbf{r}^*, \mathbf{q}^T\mathbf{w} = s^*, C\mathbf{w} \le \mathbf{b}\},$$

which is a polyhedron. By Hoffman's bound, there exists a constant $\zeta > 0$ such that, for any $\mathbf{w} \in \Omega$, we have

$$\|\mathbf{w} - \mathbf{w}^*\|^2 \le \zeta^2 \left(\|X\mathbf{w} - \mathbf{r}^*\|^2 + (\mathbf{q}^T\mathbf{w} - s^*)^2\right). \tag{20}$$

where $\mathbf{w}^*$ is the closest point to $\mathbf{w}$ in Ω_*.

By the compactness of $\mathcal{S}_\epsilon$ and the strong convexity of $h(\cdot)$ on $\mathcal{S}_\epsilon$, there exists a constant $\mu > 0$ such that

$$\begin{aligned} f(\mathbf{w}) - f(\mathbf{w}^*) &\ge \xi_{\mathbf{w}_*}^T(X\mathbf{w} - X\mathbf{w}^*) + \mathbf{q}^T(\mathbf{w} - \mathbf{w}^*) + \frac{\mu}{2}\|X\mathbf{w} - X\mathbf{w}^*\|^2 \\ &\ge \frac{\mu}{2}\|X\mathbf{w} - \mathbf{r}^*\|^2 \end{aligned} \tag{21}$$

for some $\xi_{\mathbf{w}^*} \in \partial h(\mathbf{r}^*)$ for any $\mathbf{w} \in \mathcal{S}_\epsilon$, where the second inequality is due to the optimality condition of $\mathbf{w}^*$. Note that $\xi_{\mathbf{w}^*}$ in (21) may change with $\mathbf{w}^*$ (and thus with $\mathbf{w}$). With the same $\xi_{\mathbf{w}^*}$ and $\mathbf{w}$ as above, we can also show that

$$\begin{aligned} (\mathbf{q}^T\mathbf{w} - s^*)^2 &= ((X^T\xi_{\mathbf{w}^*} + \mathbf{q} - X^T\xi_{\mathbf{w}^*})^T(\mathbf{w} - \mathbf{w}^*))^2 \\ &\le 2((X^T\xi_{\mathbf{w}^*} + \mathbf{q})^T(\mathbf{w} - \mathbf{w}^*))^2 + 2(\xi_{\mathbf{w}^*}^T(X\mathbf{w} - X\mathbf{w}^*))^2 \\ &\le 2(f(\mathbf{w}) - f^*)^2 + 2\|\partial h(\mathbf{r}^*)\|^2\|X\mathbf{w} - \mathbf{r}^*\|^2. \end{aligned} \tag{22}$$

Here, $\|\partial h(\mathbf{r}^*)\|^2 < +\infty$ because $\mathrm{dom}(h) = \mathbb{R}^n$. Applying (22) and (21) to (20), we obtain for any $\mathbf{w} \in \mathcal{S}_\epsilon$ that

$$\|\mathbf{w} - \mathbf{w}^*\|^2 \le 2\zeta^2 \left(\frac{1 + 2\|\partial h(\mathbf{r}^*)\|^2}{\mu}(f(\mathbf{w}) - f^*) + (f(\mathbf{w}) - f^*)^2\right),$$

which further implies

$$\frac{1}{2}\|\mathbf{w}-\mathbf{w}^*\|^2 \leq \zeta^2\left(\frac{1+2\|\partial h(\mathbf{r}^*)\|^2}{\mu}+\epsilon\right)(f(\mathbf{w})-f^*).$$

using the fact that $f(\mathbf{w})-f^* \leq \epsilon$ for $\mathbf{w}\in\mathcal{S}_\epsilon$. ☐

The function of this type covers some commonly used loss functions and regularization terms in machine learning and statistics. See the following examples.

Robust regression

$$\min_{\mathbf{w}\in\mathbb{R}^d}\frac{1}{n}\sum_{i=1}^{n}|\mathbf{x}_i^\top\mathbf{w}-y_i|^p, \tag{23}$$

where $p\in(1,2)$, $\mathbf{x}_i\in\mathbb{R}^d$ denotes the feature vector and y_i is the target output. The objective function is in the form of $h(X\mathbf{w})$ where X is a $n\times d$ matrix with $\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n$ being its rows and $h(\mathbf{u}) := \sum_{i=1}^n|u_i-y_i|^p$. According to [21], $h(\mathbf{u})$ is a strongly convex function on any compact set so that the objective function above is semi-strongly convex on $\mathcal{S}_\epsilon$ for any $\epsilon\leq\epsilon_0$.

5.4. Improved Convergence for Convex Problems with KL property. Lastly, we consider a family of non-smooth functions with a local Kurdyka-Łojasiewicz (KL) property. The definition of KL property is given below.

Definition 14. *The function $f(\mathbf{w})$ has the Kurdyka - Łojasiewicz (KL) property at $\bar{\mathbf{w}}$ if there exist $\eta\in(0,\infty]$, a neighborhood $U_{\bar{\mathbf{w}}}$ of $\bar{\mathbf{w}}$ and a continuous concave function $\varphi:[0,\eta)\to\mathbb{R}_+$ such that (i) $\varphi(0)=0$; (ii) φ is continuous on $(0,\eta)$; (iii)for all $s\in(0,\eta)$, $\varphi'(s)>0$; (iv) and for all $\mathbf{w}\in U_{\bar{\mathbf{w}}}\cap\{\mathbf{w}: f(\bar{\mathbf{w}})<f(\mathbf{w})<f(\bar{\mathbf{w}})+\eta\}$, the Kurdyka - Łojasiewicz (KL) inequality holds*

$$\varphi'(f(\mathbf{w})-f(\bar{\mathbf{w}}))\|\partial f(\mathbf{w})\|_2\geq 1, \tag{24}$$

where $\|\partial f(\mathbf{w})\|_2 := \min_{\mathbf{g}\in\partial f(\mathbf{w})}\|\mathbf{g}\|_2$.

The function φ is called the **desingularizing function** of f at $\bar{\mathbf{w}}$, which sharpens the function $f(\mathbf{w})$ by reparameterization. An important desingularizing function is in the form of $\varphi(s)=cs^{1-\beta}$ for some $c>0$ and $\beta\in[0,1)$, by which, (24) gives the KL inequality

$$\|\partial f(\mathbf{w})\|_2\geq\frac{1}{c(1-\beta)}(f(\mathbf{w})-f(\bar{\mathbf{w}}))^\beta.$$

Note that all semi-algebraic functions satisfy the KL property at any point [9]. Indeed, all the concrete examples given before satisfy the Kurdyka - Łojasiewicz property. For more discussions about the KL property, we refer readers to [9, 7, 48, 3, 6]. The following corollary states the iteration complexity of RSG for unconstrained problems that have the KL property at each $\bar{\mathbf{w}}\in\Omega_*$.

Corollary 15. *Suppose Assumption 1 holds, $f(\mathbf{w})$ satisfies a (uniform) Kurdyka - Łojasiewicz property at any $\bar{\mathbf{w}}\in\Omega_*$ with the same desingularizing function φ and constant η, and*

$$\mathcal{S}_\epsilon\subset\cup_{\bar{\mathbf{w}}\in\Omega_*}\left[U_{\bar{\mathbf{w}}}\cap\{\mathbf{w}: f(\bar{\mathbf{w}})<f(\mathbf{w})<f(\bar{\mathbf{w}})+\eta\}\right]. \tag{25}$$

The iteration complexity of RSG for obtaining an 2ϵ-optimal solution is given by $O\left(\alpha^2G^2(\varphi(\epsilon)/\epsilon)^2\lceil\log_\alpha(\frac{\epsilon_0}{\epsilon})\rceil\right)$ provided $\alpha^2G^2(\varphi(\epsilon)/\epsilon)^2\leq t=O\left(\alpha^2G^2(\varphi(\epsilon)/\epsilon)^2\right)$. In addition, if $\varphi(s)=cs^{1-\beta}$ for some $c>0$ and $\beta\in[0,1)$, the iteration complexity of RSG is $O(\frac{\alpha^2G^2c^2(1-\beta)^2}{\epsilon^{2\beta}}\lceil\log_\alpha(\frac{\epsilon_0}{\epsilon})\rceil)$ provided $t=\frac{\alpha^2G^2c^2}{\epsilon^{2\beta}}$ and $K=\lceil\log_\alpha(\frac{\epsilon_0}{\epsilon})\rceil$.

Proof. We can prove the above corollary following a result in [8] as presented in Proposition 24 in the appendix. According to Proposition 24, if $f(\cdot)$ satisfies the KL property at $\bar{\mathbf{w}}$, then for all $\mathbf{w} \in U_{\bar{\mathbf{w}}} \cap \{\mathbf{w} : f(\bar{\mathbf{w}}) < f(\mathbf{w}) < f(\bar{\mathbf{w}}) + \eta\}$ it holds that $\|\mathbf{w} - \mathbf{w}^*\|_2 \leq \varphi(f(\mathbf{w}) - f(\bar{\mathbf{w}}))$. It then, under the uniform condition in (25), implies that, for any $\mathbf{w} \in \mathcal{S}_\epsilon$

$$\|\mathbf{w} - \mathbf{w}^*\|_2 \leq \varphi(f(\mathbf{w}) - f_*) \leq \varphi(\epsilon),$$

where we use the monotonic property of φ. Then the first conclusion follows similarly as Corollary 6 by noting $B_\epsilon \leq \varphi(\epsilon)$. The second conclusion immediately follows by setting $\varphi(s) = cs^{1-\beta}$ in the first conclusion. ∎

While the conclusion in Corollary 15 hinges on a condition in (25), in practice many convex functions satisfy the KL property with $U = \mathbb{R}^d$ and $\eta = \infty$ [2]. It is worth mentioning that to our best knowledge, the present work is the first to leverage the KL property for developing improved subgradient methods, though it has been explored in non-convex and convex optimization for deterministic descent methods for smooth optimization [8, 9, 2, 25].

6. A variant of RSG with p-norm and Dual Averaging Method. To demonstrate the flexibility of our restarting strategy, in this section, we present a variant of RSG that utilizes the p-norm with $p \in (1, 2]$ and a Nesterov's dual averaging (DA) algorithm as its subroutine.

Let $\|\mathbf{w}\|_p = (\sum_{i=1}^d |w_i|^p)^{1/p}$ denote the p-norm. It is known that $\frac{1}{2}\|\mathbf{w}\|_p^2$ is $(p-1)$-strongly convex. Let q and p be conjugate constants, namely, $1/p + 1/q = 1$. Using the p-norm, we redefine some of the previous notations:

$$\mathbf{w}_\epsilon^\dagger := \arg\min_{\mathbf{u}\in\mathcal{S}_\epsilon} \|\mathbf{u} - \mathbf{w}\|_p^2 = \arg\min_{\mathbf{u}\in\Omega} \|\mathbf{u} - \mathbf{w}\|_p^2, \quad \text{s.t.} \quad f(\mathbf{u}) \leq f_* + \epsilon,$$

$$\mathbf{w}^* := \arg\min_{\mathbf{u}\in\Omega_*} \|\mathbf{u} - \mathbf{w}\|_p^2,$$

$$B_\epsilon := \max_{\mathbf{w}\in\mathcal{L}_\epsilon} \min_{\mathbf{u}\in\Omega_*} \|\mathbf{w} - \mathbf{u}\|_p = \max_{\mathbf{w}\in\mathcal{L}_\epsilon} \|\mathbf{w} - \mathbf{w}^*\|_p,$$

$$\|\partial f(\mathbf{w}) + \mathcal{N}_\Omega(\mathbf{w})\|_q := \min_{\mathbf{g}\in\partial f(\mathbf{w}), \mathbf{v}\in\mathcal{N}_\Omega(\mathbf{w})} \|\mathbf{g} + \mathbf{v}\|_q$$

$$\rho_\epsilon := \min_{\mathbf{w}\in\mathcal{L}_\epsilon} \|\partial f(\mathbf{w}) + \mathcal{N}_\Omega(\mathbf{w})\|_q.$$

We first generalize Lemma 2 with the p-norm.

LEMMA 16. *For any $\epsilon > 0$ such that $\mathcal{L}_\epsilon \neq \emptyset$ and any $\mathbf{w} \in \Omega$, we have*

$$\|\mathbf{w} - \mathbf{w}_\epsilon^\dagger\|_p \leq \frac{1}{\rho_\epsilon}(f(\mathbf{w}) - f(\mathbf{w}_\epsilon^\dagger)).$$

Proof. Let $[\mathbf{w}]_i$ represent the ith coordinate of $\mathbf{w}$. Note that

$$\frac{\partial\|\mathbf{w}\|_p}{\partial w_i} = \frac{|w_i|^{p-1}\text{sign}(w_i)}{\|\mathbf{w}\|_p^{p-1}}, \quad \left[\nabla\|\mathbf{w}\|_p^2\right]_i = 2\|\mathbf{w}\|_p^{2-p}|w_i|^{p-1}\text{sign}(w_i).$$

Since the conclusion holds trivially if $\mathbf{w} \in \mathcal{S}_\epsilon$ (so that $\mathbf{w}_\epsilon^\dagger = \mathbf{w}$), we assume $\mathbf{w} \in \Omega\backslash\mathcal{S}_\epsilon$. According to the definition of $\mathbf{w}_\epsilon^\dagger$ and the associated first-order optimality conditions, there exist a scalar $\zeta \geq 0$, a subgradient $\mathbf{g} \in \partial f(\mathbf{w}_\epsilon^\dagger)$ and a vector $\mathbf{v} \in \mathcal{N}_\Omega(\mathbf{w}_\epsilon^\dagger)$ such that

$$\|\mathbf{w}_\epsilon^\dagger - \mathbf{w}\|_p^{2-p}|[\mathbf{w}_\epsilon^\dagger - \mathbf{w}]_i|^{p-1}\text{sign}([\mathbf{w}_\epsilon^\dagger - \mathbf{w}]_i) + \zeta[\mathbf{g}]_i + [\mathbf{v}]_i = 0, \quad i = 1, 2, \ldots, d. \tag{26}$$

By the convexity of $f(\cdot)$ and the definition of normal cone, we have

$$\begin{aligned}\zeta(f(\mathbf{w}) - f(\mathbf{w}_\epsilon^\dagger)) &\geq (\mathbf{w} - \mathbf{w}_\epsilon^\dagger)^\top(\zeta\mathbf{g} + \mathbf{v}) \\ &= \|\mathbf{w}_\epsilon^\dagger - \mathbf{w}\|_p^{2-p} \sum_{i=1}^d |[\mathbf{w}_\epsilon^\dagger - \mathbf{w}]_i|^{p-1}\mathrm{sign}([\mathbf{w}_\epsilon^\dagger - \mathbf{w}]_i)[\mathbf{w}_\epsilon^\dagger - \mathbf{w}]_i \\ &= \|\mathbf{w}_\epsilon^\dagger - \mathbf{w}\|_p^{2-p}\|\mathbf{w}_\epsilon^\dagger - \mathbf{w}\|_p^p = \|\mathbf{w}_\epsilon^\dagger - \mathbf{w}\|_p^2.\end{aligned} \tag{27}$$

It is clear that $\zeta > 0$, since, otherwise, $\mathbf{w} = \mathbf{w}_\epsilon^\dagger$ which contradicts with $\mathbf{w} \notin \mathcal{S}_\epsilon$. It is from (26) that

$$\begin{aligned}&\zeta^q \sum_i |[\mathbf{g} + \mathbf{v}/\zeta]_i|^q = \|\mathbf{w}_\epsilon^\dagger - \mathbf{w}\|_p^{q(2-p)} \sum_{i=1}^d |[\mathbf{w}_\epsilon^\dagger - \mathbf{w}]_i|^{q(p-1)} \\ &= \|\mathbf{w}_\epsilon^\dagger - \mathbf{w}\|_p^{q(2-p)} \sum_{i=1}^d |[\mathbf{w}_\epsilon^\dagger - \mathbf{w}]_i|^p = \|\mathbf{w}_\epsilon^\dagger - \mathbf{w}\|_p^{q(2-p)+p} = \|\mathbf{w}_\epsilon^\dagger - \mathbf{w}\|_p^{p/(p-1)}.\end{aligned} \tag{28}$$

Organizing terms in the equality above gives

$$\frac{1}{\zeta} = \frac{\|\mathbf{g} + \mathbf{v}/\zeta\|_q}{\|\mathbf{w}_\epsilon^\dagger - \mathbf{w}\|_p^{p/(q(p-1))}} \geq \frac{\rho_\epsilon}{\|\mathbf{w}_\epsilon^\dagger - \mathbf{w}\|_p}, \tag{29}$$

which further implies

$$f(\mathbf{w}) - f(\mathbf{w}_\epsilon^\dagger) \geq \rho_\epsilon\|\mathbf{w}_\epsilon^\dagger - \mathbf{w}\|_p.$$

∎

Similar to Lemma 5, we can also lower bound ρ_ϵ by ϵ/B_ϵ. In particular, we have the following lemma.

LEMMA 17. *For any $\epsilon > 0$ such that $\mathcal{L}_\epsilon \neq \emptyset$ and any $\mathbf{w} \in \Omega$, we have*

$$\|\mathbf{w} - \mathbf{w}_\epsilon^\dagger\|_p \leq \frac{\|\mathbf{w}_\epsilon^\dagger - \mathbf{w}_\epsilon^*\|_p}{\epsilon}(f(\mathbf{w}) - f(\mathbf{w}_\epsilon^\dagger)) \leq \frac{B_\epsilon}{\epsilon}(f(\mathbf{w}) - f(\mathbf{w}_\epsilon^\dagger)).$$

where $\mathbf{w}_\epsilon^*$ is the closest point in Ω_* to $\mathbf{w}_\epsilon^\dagger$ measured in the p-norm.

Proof. Since the conclusion holds trivially if $\mathbf{w} \in \mathcal{S}_\epsilon$ (so that $\mathbf{w}_\epsilon^\dagger = \mathbf{w}$), we assume $\mathbf{w} \in \Omega\backslash\mathcal{S}_\epsilon$ so that $\mathbf{w}_\epsilon^\dagger \in \mathcal{L}_\epsilon$. From (28) in the proof of Lemma 16, we have $\|\mathbf{g}+\mathbf{v}/\zeta\|_q = \|\mathbf{w}_\epsilon^\dagger - \mathbf{w}\|_p/\zeta$, where $\zeta > 0$, $\mathbf{g} \in \partial f(\mathbf{w}_\epsilon^\dagger)$ and $\mathbf{v} \in \mathcal{N}_\Omega(\mathbf{w}_\epsilon^\dagger)$. Applying this relationship to (27) in the proof of Lemma 16, we have

$$f(\mathbf{w}) - f(\mathbf{w}_\epsilon^\dagger) \geq \|\mathbf{g} + \mathbf{v}/\zeta\|_q\|\mathbf{w} - \mathbf{w}_\epsilon^\dagger\|_p.$$

By the convexity of $f(\mathbf{w})$ and the definition of normal cone, we have

$$f(\mathbf{w}_\epsilon^*) - f(\mathbf{w}_\epsilon^\dagger) \geq (\mathbf{w}_\epsilon^* - \mathbf{w}_\epsilon^\dagger)^\top\mathbf{g} \geq (\mathbf{w}_\epsilon^* - \mathbf{w}_\epsilon^\dagger)^\top\left(\mathbf{g} + \mathbf{v}/\zeta\right),$$

Then,

$$\|\mathbf{g} + \mathbf{v}/\zeta\|_q\|\mathbf{w}_\epsilon^* - \mathbf{w}_\epsilon^\dagger\|_p \geq f(\mathbf{w}_\epsilon^\dagger) - f(\mathbf{w}_\epsilon^*) = \epsilon$$

As a result,

$$f(\mathbf{w}) - f(\mathbf{w}_\epsilon^\dagger) \geq \frac{\epsilon}{\|\mathbf{w}_\epsilon^* - \mathbf{w}_\epsilon^\dagger\|_p}\|\mathbf{w} - \mathbf{w}_\epsilon^\dagger\|_p$$

which completes the proof. ∎

Algorithm 3 DA_p: $\widehat{\mathbf{w}}_T = \text{DA}_p(\mathbf{w}_1, \eta, T)$

1: **Input**: a step size η, the number of iterations T, and the initial solution $\mathbf{w}_1$,
2: $\widehat{\mathbf{g}}_1 = 0$, $\Lambda_0 = 0$
3: **for** $t = 1, \ldots, T$ **do**
4: Query the gradient oracle to obtain $\mathcal{G}(\mathbf{w}_t)$
5: Set $\lambda_t = 1$ or $\lambda_t = 1/\|\mathcal{G}(\mathbf{w}_t)\|_q$, and $\Lambda_t = \Lambda_{t-1} + \lambda_t$
6: $\widehat{\mathbf{g}}_{t+1} = \widehat{\mathbf{g}}_t + \lambda_t \mathcal{G}(\mathbf{w}_t)$
7: update $\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}} \eta \widehat{\mathbf{g}}_{t+1}^\top \mathbf{w} + \frac{1}{2}\|\mathbf{w} - \mathbf{w}_1\|_p^2$
8: **end for**
9: **Output**: $\widehat{\mathbf{w}}_T = \frac{\sum_{t=1}^T \lambda_t \mathbf{w}_t}{\Lambda_T}$

Algorithm 4 RSG-DA_p

Input: the number of iterations t per-epoch, $\mathbf{w}_0 \in \mathbb{R}^d$, $\alpha > 1$.
Set $\eta_1 = \epsilon_0(p-1)/(\alpha G)$ or $\eta_1 = \epsilon_0(p-1)/(\alpha G^2)$ corresponding to the two choices of λ_t in DA_p, where ϵ_0 is from Assumption 20.a
for $k = 1, \ldots, K$ **do**
Call subroutine to obtain $\mathbf{w}_k = \text{DA}_p(\mathbf{w}_{k-1}, \eta_k, t)$
Set $\eta_{k+1} = \eta_k/\alpha$
end for
Output: $\mathbf{w}_K$

In Algorithm 4, we present a modified version of RSG that uses Algorithm 3 as a subroutine. Here, Algorithm 3 is a variant of Nesterov's dual averaging (DA) algorithm. We denote Algorithm 4 by RSG-DA_p to distinguish it from the RSG method in Algorithm 2.

Before presenting the convergence property of RSG-DA_p, we first need to address two issues: (i) how to compute $\mathbf{w}_{t+1}$ in Line 7 of DA_p and (ii) what the convergence rate of DA_p is. The proposition below shows that $\mathbf{w}_{t+1}$ can be computed in a closed form when $\Omega = \mathbb{R}^d$.

PROPOSITION 18. *The solution to* $\mathbf{w}^+ = \arg\min_{\mathbf{u}} \mathbf{g}^\top \mathbf{u} + \frac{1}{2}\|\mathbf{u} - \mathbf{w}\|_p^2$ *is given by*

$$[\mathbf{w}^+]_i = [\mathbf{w}]_i - \|\mathbf{g}\|_q^{(p-q)/p} sign([\mathbf{g}]_i)|[\mathbf{g}]_i|^{q-1}, \quad i = 1, 2, \ldots, d.$$

Proof. By the optimality condition, $\mathbf{w}^+$ satisfies

$$\|\mathbf{w}^+ - \mathbf{w}\|_p^{2-p}|[\mathbf{w}^+ - \mathbf{w}]_i|^{p-1}\text{sign}([\mathbf{w}^+ - \mathbf{w}]_i) + [\mathbf{g}]_i = 0, \quad i = 1, 2, \ldots, d. \tag{30}$$

It is a simple exercise to verify that $\|\mathbf{g}\|_q = \|\mathbf{w}^+ - \mathbf{w}\|_p$. From (30), we have

$$(\mathbf{w} - \mathbf{w}^+)^\top \mathbf{g} = \|\mathbf{w}^+ - \mathbf{w}\|_p^2 = \|\mathbf{w}^+ - \mathbf{w}\|_p \|\mathbf{g}\|_q.$$

This means the Holder's inequality, $(\mathbf{w} - \mathbf{w}^+)^\top \mathbf{g} \leq \|\mathbf{w}^+ - \mathbf{w}\|_p \|\mathbf{g}\|_q$, holds as an equality and this can happens only if $|[\mathbf{w}^+ - \mathbf{w}]_i| = c|[\mathbf{g}]_i|^{q-1}$ for a constant c. From (30), we can also see that $[\mathbf{w}^+ - \mathbf{w}]_i$ has the same sign as $-[\mathbf{g}]_i$ so that

$$[\mathbf{w}^+]_i - [\mathbf{w}]_i = -c\,\text{sign}([\mathbf{g}]_i)|[\mathbf{g}]_i|^{q-1}.$$

It remains to derive the value of c. To do so, we observe that

$$\sum_{i=1}^{d} |[\mathbf{w}^+ - \mathbf{w}]_i|^p = c^p \sum_{i=1}^{d} |[\mathbf{g}]_i|^{p(q-1)} = c^p \sum_{i=1}^{d} |[\mathbf{g}]_i|^q,$$

which indicates $\|\mathbf{w}^+ - \mathbf{w}\|_p^p = c^p \|\mathbf{g}\|_q^q$. Since $\|\mathbf{w}^+ - \mathbf{w}\|_p = \|\mathbf{g}\|_q$, we have

$$c = \left(\frac{\|\mathbf{g}\|_q^p}{\|\mathbf{g}\|_q^q}\right)^{1/p} = \|\mathbf{g}\|_q^{(p-q)/p}.$$

□

The following Lemma by Nesterov [38] characterizes the convergence property of DA_p.

LEMMA 19. *[38] Suppose $\|\mathcal{G}(\mathbf{w}_t)\|_q \leq G$. Let Algorithm 3 run for T iterations. Then, for any $\mathbf{w} \in \Omega$, we have*

$$f(\widehat{\mathbf{w}}_T) - f(\mathbf{w}) \leq \frac{\|\mathbf{w} - \mathbf{w}_1\|_p^2}{2\eta\Lambda_T} + \frac{\eta \sum_{t=1}^T \lambda_t^2 \|\mathcal{G}(\mathbf{w}_t)\|_q^2}{2(p-1)\Lambda_T}.$$

If $\lambda_t = 1$, we have

$$f(\widehat{\mathbf{w}}_T) - f(\mathbf{w}) \leq \frac{\|\mathbf{w} - \mathbf{w}_1\|_p^2}{2\eta T} + \frac{\eta G^2}{2(p-1)}.$$

If $\lambda_t = 1/\|\mathcal{G}(\mathbf{w}_t)\|_q$, we have

$$f(\widehat{\mathbf{w}}_T) - f(\mathbf{w}) \leq \frac{G\|\mathbf{w} - \mathbf{w}_1\|_p^2}{2\eta T} + \frac{\eta G}{2(p-1)}.$$

We omit the proof because it can be found in [38]. To state the convergence of RSG-DA_p. We need the following assumptions which is the same as Assumption 1 except that the Euclidean norm is replaced by the q-norm in item **b**:

ASSUMPTION 20. *For a convex minimization problem (1), we assume*
- **a.** *There exist $\mathbf{w}_0 \in \Omega$ and $\epsilon_0 \geq 0$ such that $f(\mathbf{w}_0) - \min_{\mathbf{w}\in\Omega} f(\mathbf{w}) \leq \epsilon_0$.*
- **b.** *There exists a constant G such that $\|\mathcal{G}(\mathbf{w})\|_q \leq G$ for any $\mathbf{w} \in \Omega$.*
- **c.** *Ω_* is a non-empty convex compact set.*

The following theorem characterizes the iteration complexity of RSG-DA_p whose proof is similar to that of Theorem 4 and is thus omitted.

THEOREM 21. *Suppose Assumption 20 hods. If $t \geq \frac{\alpha^2 G^2 B_\epsilon^2}{\epsilon^2}$ and $K = \lceil \log_\alpha(\frac{\epsilon_0}{\epsilon}) \rceil$ in Algorithm 4, with at most K stages, Algorithm 4 returns a solution $\mathbf{w}_K$ such that $f(\mathbf{w}_K) - f_* \leq 2\epsilon$. In other word, the total number of iterations for Algorithm 4 to find an 2ϵ-optimal solution is at most $T = O\left(t\lceil \log_\alpha(\frac{\epsilon_0}{\epsilon})\rceil\right)$ where $t \geq \frac{\alpha^2 G^2 B_\epsilon^2}{\epsilon^2}$. In particular, if $\frac{\alpha^2 G^2 B_\epsilon^2}{\epsilon^2} \leq t = O\left(\frac{\alpha^2 G^2 B_\epsilon^2}{\epsilon^2}\right)$, the total number of iterations for Algorithm 4 to find an 2ϵ-optimal solution is at most $T = O\left(\frac{\alpha^2 G^2 B_\epsilon^2}{\epsilon^2}\lceil \log_\alpha(\frac{\epsilon_0}{\epsilon})\rceil\right)$.*

We want to emphasis that, although the conclusion of Theorem 21 is similar to Corollary 6, B_ϵ here is defined using p-norm and G here is defined using the q-norm. Depending on the problem, Algorithm 4 may have a lower complexity than Algorithm 2. For an example, we can consider empirical loss minimization in machine learning for finding a sparse model with high-dimensional data. Suppose the loss function $\ell(z, y)$ is 1-Lipschitz continuous, if $f(\mathbf{w})$ is the average of $\ell(\mathbf{w}^\top \mathbf{x}_i, y_i)$ over training examples $(\mathbf{x}_i, y_i), i = 1, \ldots, n$, where $\mathbf{x}_i \in \mathbb{R}^d$ $(d \gg 1)$, then $G = \max_i \|\mathbf{x}_i\|_q$. Let $p = \frac{2\ln d}{2\ln d - 1}$ and $q = 2\ln d$ [6]. Then, using the p-norm to define B_ϵ and q-norm to define G, we

[6] Other choices are possible [15].

have $B_\epsilon G = \max_{\mathbf{w}\in\mathcal{L}_\epsilon} \|\mathbf{w}-\mathbf{w}^*\|_p \max_i \|\mathbf{x}_i\|_q \approx \max_{\mathbf{w}\in\mathcal{L}_\epsilon} \|\mathbf{w}-\mathbf{w}^*\|_1 \max_{\mathbf{x}_i} \|\mathbf{x}\|_\infty$. In contrast, if we use the Euclidean norm definitions, we have $B_\epsilon G = \max_{\mathbf{w}\in\mathcal{L}_\epsilon} \|\mathbf{w} - \mathbf{w}^*\|_2 \max_i \|\mathbf{x}\|_2$. If we assume $\|\mathbf{x}_i\|_\infty \leq 1$ and $\mathbf{w} \in \mathcal{S}_\epsilon$ is approximately s-sparse with $s \ll d$ such that $\frac{\|\mathbf{w}-\mathbf{w}^*\|_1}{\|\mathbf{w}-\mathbf{w}^*\|_2} \approx \sqrt{s}$ [59, 43]. In light of $\frac{\|\mathbf{x}_i\|_\infty}{\|\mathbf{x}_i\|_2} \approx \frac{1}{\sqrt{d}}$, the magnitude of $B_\epsilon G$ using the special p-norm and q-norm definitions is much smaller than $B_\epsilon G$ using the Euclidean norm definitions, making RSG-DA$_p$ converge faster than RSG.

7. Variants of RSG without knowing the constant c and the exponent θ in the local error bound. In Section 5, we have discussed the local error bound and presented several classes of problems to reveal the magnitude of B_ϵ, i.e., $B_\epsilon = c\epsilon^\theta$. For some problems, the value of θ is exhibited. However, the value of the constant c could be still difficult to estimate, which renders it challenging to set the appropriate value $t = \frac{\alpha^2 c^2 G^2}{\epsilon^{2(1-\theta)}}$ for inner iterations of RSG. In practice, one might use a sufficiently large c to set up the value of t. However, such an approach is vulnerable to both over-estimation and under-estimation of t. Over-estimating the value of t leads to a waste of iterations while under-estimation leads to an less accurate solution that might not reach to the target accuracy level. In addition, for other problems the value of θ is still an open problem. One interesting family of objective functions in machine learning is the sum of piecewise linear loss over training data and overlapped or non-overlapped group lasso. In this section, we present variants of RSG that can be implemented without knowing the value of c in the local error bound condition and even the value of exponent θ, and prove their improved convergence.

7.1. RSG without knowing c. The key idea is to use an increasing sequence of t and another level of restarting for RSG. The detailed steps are presented in Algorithm 5, to which we refer as R^{2}SG. With mild conditions on t_1 in R^{2}SG, the complexity of R^{2}SG for finding an ϵ solution is given by the theorem below.

THEOREM 22. *Suppose $\epsilon \leq \epsilon_0/4$ and $K = \lceil \log_\alpha(\epsilon_0/\epsilon)\rceil$. Let t_1 in Algorithm 5 be large enough so that there exists $\hat\epsilon_1$ such that $\epsilon \leq \hat\epsilon_1 \leq \epsilon_0/2$ and $t_1 = \frac{\alpha^2 c^2 G^2}{\hat\epsilon_1^{2(1-\theta)}}$. In addition, suppose $f(\cdot)$ admits an error bound condition with $\theta \in (0,1)$ on $\mathcal{S}_{\hat\epsilon_1}$. Then, with at most $S = \lceil \log_2(\hat\epsilon_1/\epsilon)\rceil + 1$ calls of RSG in Algorithm 5, we find a solution $\mathbf{w}^S$ such that $f(\mathbf{w}^S) - f_* \leq 2\epsilon$. The total number of iterations of R^{2}SG for obtaining 2ϵ-optimal solution is upper bounded by $T_S = O\left(\frac{c^2G^2}{\epsilon^{2(1-\theta)}}\lceil \log_\alpha(\frac{\epsilon_0}{\epsilon})\rceil\right)$.*

Proof. Since $K = \lceil \log_\alpha(\epsilon_0/\epsilon)\rceil \geq \lceil \log_\alpha(\epsilon_0/\hat\epsilon_1)\rceil$ and $t_1 = \frac{\alpha^2 c^2 G^2}{\hat\epsilon_1^{2(1-\theta)}}$, we can apply Corollary 8 with $\epsilon = \hat\epsilon_1$ to the first call of RSG in Algorithm 5 so that the output $\mathbf{w}^1$ satisfies

$$f(\mathbf{w}^1) - f_* \leq 2\hat\epsilon_1. \tag{31}$$

Then, we consider the second call of RSG with the initial solution $\mathbf{w}^1$ satisfying (31). By the setup $K = \lceil \log_\alpha(\epsilon_0/\epsilon)\rceil \geq \lceil \log_\alpha(2\hat\epsilon_1/(\hat\epsilon_1/2))\rceil$ and $t_2 = t_1 2^{2(1-\theta)} = \frac{c^2G^2}{(\hat\epsilon_1/2)^{2(1-\theta)}}$, we can apply Corollary 8 with $\epsilon = \hat\epsilon_1/2$ and $\epsilon_0 = 2\hat\epsilon_1$ so that the output $\mathbf{w}^2$ of the second call satisfies $f(\mathbf{w}^2) - f_* \leq \hat\epsilon_1$. By repeating this argument for all the subsequent calls of RSG, with at most $S = \lceil \log_2(\hat\epsilon_1/\epsilon)\rceil + 1$ calls, Algorithm 5 ensures that

$$f(\mathbf{w}^S) - f_* \leq 2\hat\epsilon_1/2^{S-1} \leq 2\epsilon$$

Algorithm 5 RSG with restarting: $\mathrm{R}^2\mathrm{SG}$

Input: the number of iterations t_1 in each stage of the first call of RSG and the number of stages K in each call of RSG
Initialization: $\mathbf{w}^0 \in \Omega$;
for $s = 1, 2 \ldots,$ **do**
  Let $\mathbf{w}^s = \mathrm{RSG}(\mathbf{w}^{s-1}, K, t_s, \alpha)$
  Let $t_{s+1} = t_s 2^{2(1-\theta)}$
  If a stopping criterion is satisfied, **then** terminate the algorithm
end for

The total number of iterations during the S calls of RSG is bounded by

□

$$T_S = K\sum_{s=1}^{S} t_s = K\sum_{s=1}^{S} t_1 2^{2(s-1)(1-\theta)} = Kt_1 2^{2(S-1)(1-\theta)} \sum_{s=1}^{S}\left(\frac{1}{2^{2(1-\theta)}}\right)^{S-s}$$
$$\leq \frac{Kt_1 2^{2(S-1)(1-\theta)}}{1 - 1/2^{2(1-\theta)}} \leq O\left(Kt_1\left(\frac{\hat{\epsilon}_1}{\epsilon}\right)^{2(1-\theta)}\right) = O\left(\frac{c^2G^2}{\epsilon^{2(1-\theta)}}\lceil \log_\alpha(\frac{\epsilon_0}{\epsilon})\rceil\right).$$

Remark: We make several remarks about Algorithm 5 and Theorem 22: (i) Theorem 22 applies only when $\theta \in (0, 1)$. If $\theta = 1$, in order to have an increasing sequence of t_s, we can set θ in Algorithm 5 to a little smaller value than 1 in practical implementation. (ii) the ϵ_0 in the implementation of RSG (Algorthm 2) can be re-calibrated for $s \geq 2$ to improve the performance (e.g., one can use the relationship $f(\mathbf{w}_{s-1}) - f_* = f(\mathbf{w}_{s-2}) - f_* + f(\mathbf{w}_{s-1}) - f(\mathbf{w}_{s-2})$ to do re-calibration); (iii) as a tradeoff, the exiting criterion of $\mathrm{R}^2\mathrm{SG}$ is not as automatic as RSG. In fact, the total number of calls S of RSG for obtaining an 2ϵ-optimal solution depends on an unknown parameter (namely $\hat{\epsilon}_1$). In practice, one could use other stopping criteria to terminate the algorithm. For example, in machine learning applications one can monitor the performance on the validation data set to terminate the algorithm. (vi) The quantities $\hat{\epsilon}_1$, S in the proof above are implicitly determined by t_1 and one do not need to compute $\hat{\epsilon}_1$ and S in order to apply Algorithm 5.

7.2. RSG for unknown θ and c. Without knowing $\theta \in (0, 1]$ and c to get a sharper local error bound, we can simply let $\theta = 0$ and $c = B_{\epsilon'}$ with $\epsilon' \geq \epsilon$, which still render the local error bound condition hold (c.f. Definition 7). Then we can employ the doubling trick to increase the values of t. In particular, we start with a sufficiently large value of t and run RSG with $K = \lceil \log_\alpha(\epsilon_0/\epsilon)\rceil$ stages, and then double the value of t and repeat the process.

THEOREM 23. *Suppose $\epsilon \leq \epsilon_0/4$ and $K = \lceil \log_\alpha(\epsilon_0/\epsilon)\rceil$. Let $\theta = 0$ and t_1 in Algorithm 5 be large enough so that there exists $\hat{\epsilon}_1$ such that $\epsilon \leq \hat{\epsilon}_1 \leq \epsilon_0/2$ and $t_1 = \frac{\alpha^2 B_{\hat{\epsilon}_1}^2 G^2}{\hat{\epsilon}_1^2}$. Then, with at most $S = \lceil \log_2(\hat{\epsilon}_1/\epsilon)\rceil + 1$ calls of RSG in Algorithm 5, we find a solution $\mathbf{w}^S$ such that $f(\mathbf{w}^S) - f_* \leq 2\epsilon$. The total number of iterations of R^2SG for obtaining 2ϵ-optimal solution is upper bounded by $T_S = O\left(\frac{B_{\hat{\epsilon}_1}^2 G^2}{\epsilon^2}\lceil \log_\alpha(\frac{\epsilon_0}{\epsilon})\rceil\right)$.*

Remark: Compared to the vanilla SG, the above iteration complexity is still an improved one with a smaller factor $B_{\hat{\epsilon}_1}^2$ compared to $\|\mathbf{w}_0 - \mathbf{w}_0^*\|_2^2$ in the iteration complexity of SG.

Proof. The proof is similar to that of Theorem 22 except that we let $c = B_{\hat\epsilon_1}$ and $\theta = 0$. Since $K = \lceil\log_\alpha(\epsilon_0/\epsilon)\rceil \geq \lceil\log_\alpha(\epsilon_0/\hat\epsilon_1)\rceil$ and $t_1 = \frac{\alpha^2 B_{\hat\epsilon_1}^2 G^2}{\hat\epsilon_1^2}$, we can apply Corollary 8 with $\epsilon = \hat\epsilon_1$ to the first call of RSG in Algorithm 5 so that the output $\mathbf{w}^1$ satisfies

$$f(\mathbf{w}^1) - f_* \leq 2\hat\epsilon_1. \tag{32}$$

Then, we consider the second call of RSG with the initial solution $\mathbf{w}^1$ satisfying (32). By the setup $K = \lceil\log_\alpha(\epsilon_0/\epsilon)\rceil \geq \lceil\log_\alpha(2\hat\epsilon_1/(\hat\epsilon_1/2))\rceil$ and $t_2 = t_1 2^2 = \frac{B_{\hat\epsilon_1}^2 G^2}{(\hat\epsilon_1/2)^2}$, we can apply Corollary 8 with $\epsilon = \hat\epsilon_1/2$ and $\epsilon_0 = 2\hat\epsilon_1$ so that the output $\mathbf{w}^2$ of the second call satisfies $f(\mathbf{w}^2) - f_* \leq \hat\epsilon_1$. By repeating this argument for all the subsequent calls of RSG, with at most $S = \lceil\log_2(\hat\epsilon_1/\epsilon)\rceil + 1$ calls, Algorithm 5 ensures that

$$f(\mathbf{w}^S) - f_* \leq 2\hat\epsilon_1/2^{S-1} \leq 2\epsilon$$

The total number of iterations during the S calls of RSG is bounded by

$$\begin{aligned} T_S &= K\sum_{s=1}^S t_s = K\sum_{s=1}^S t_1 2^{2(s-1)} = Kt_1 2^{2(S-1)} \sum_{s=1}^S \left(\frac{1}{2^2}\right)^{S-s} \\ &\leq \frac{Kt_1 2^{2(S-1)}}{1 - 1/2^2} \leq O\left(Kt_1\left(\frac{\hat\epsilon_1}{\epsilon}\right)^2\right) = O\left(\frac{B_{\hat\epsilon_1}^2 G^2}{\epsilon^2}\lceil\log_\alpha(\frac{\epsilon_0}{\epsilon})\rceil\right). \end{aligned}$$

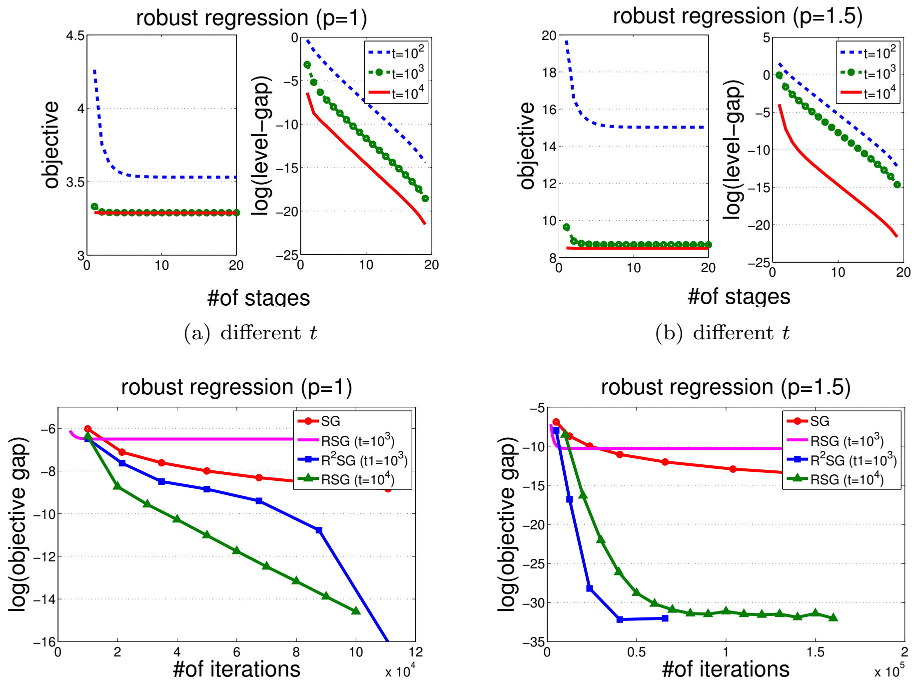
□

8. Experiments. In this section, we present some experiments to demonstrate the effectiveness of RSG. We focus on applications in machine learning, in particular regression and classification.

Robust Regression. The regression problem is to predict an output y based on a feature vector $\mathbf{x} \in \mathbb{R}^d$. Given a set of training examples $(\mathbf{x}_i, y_i), i = 1, \ldots, n$, a linear regression model can be found by solving the optimization problem in (23). We solve two instances of the problem with $p = 1$ and $p = 1.5$. We conduct experiments on two data sets from libsvm website [7], namely housing ($n = 506$ and $d = 13$) and space-ga ($n = 3107$ and $d = 6$). We first examine the convergence behavior of RSG with different values for the number of iterations per-stage $t = 10^2$, 10^3, and 10^4. The value of α is set to 2 in all experiments. The initial step size of RSG is set to be proportional to $\epsilon_0/2$ with the same scaling parameter for different variants. We plot the results on housing data in Figure 2 (a,b) and on space-ga data in Figure 3 (a,b). In each figure, we plot the objective value vs number of stages and the log difference between the objective value and the converged value (to which we refer as level gap). We can clearly see that with different t RSG converges to an ϵ-level set and the convergence rate is linear in terms of the number of stages, which is consistent with our theory.

Secondly, we compare with SG to verify the effectiveness of RSG. The baseline SG is implemented with a decreasing step size proportional to $1/\sqrt{\tau}$, where τ is the iteration index. The initial step size of SG is tuned in a wide range to give the fastest convergence. The initial step size of RSG is also tuned around the best initial step size of SG. The results are shown in Figure 2(c,d) and Figure 3(c,d), where we show RSG with two different values of t and also R^{2}SG with an increasing sequence

[7] https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/



(a) different t (b) different t

(c) different algorithms (d) different algorithms

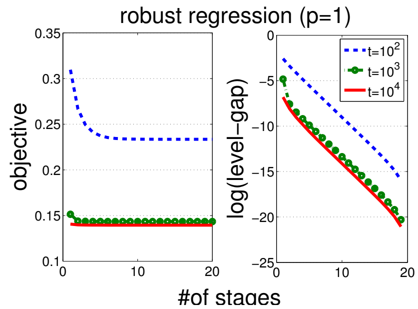
FIG. 2. *Comparison of RSG with different t and of different algorithms on the housing data. (t1 for R^2SG represents the initial value of t in the first call of RSG.)*

of t. In implementing R^{2}SG, we restart RSG for every 5 stages, and increase the number of iterations by a certain factor. In particular,we increase t by a factor of 1.15 and 1.5 respectively for $p = 1$ and $p = 1.5$. From the results, we can see that (i) RSG with a smaller value of $t = 10^3$ can quickly converge to an ϵ-level, which is less accurate than SG after running a sufficiently large number of iterations; (ii) RSG with a relatively large value $t = 10^4$ can converge to a much more accurate solution; (iv) R^{2}SG converges much faster than SG and can bridge the gap between RSG-$t = 10^3$ and RSG-$t = 10^4$.

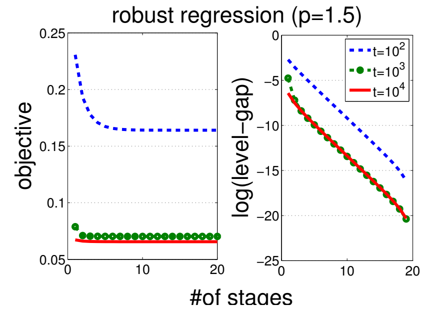
SVM Classification with a graph-guided fused lasso. The classification problem is to predict a binary class label $y \in \{1, -1\}$ based on a feature vector $\mathbf{x} \in \mathbb{R}^d$. Given a set of training examples $(\mathbf{x}_i, y_i), i = 1, \ldots, n$, the problem of training a linear classification model can be cast into

$$\min_{\mathbf{w}\in\mathbb{R}^d} F(\mathbf{w}) := \frac{1}{n}\sum_{i=1}^{n} \ell(\mathbf{w}^\top \mathbf{x}_i, y_i) + R(\mathbf{w}).$$

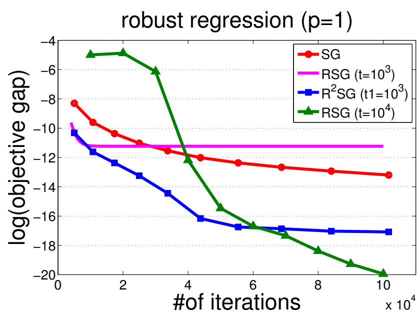
Here we consider the hinge loss $\ell(z, y) = \max(0, 1 - yz)$ as in support vector machine (SVM) and a graph-guided fused lasso (GFlasso) regularizer $R(\mathbf{w}) = \lambda\|F\mathbf{w}\|_1$ [26], where $F = [F_{ij}]_{m\times d} \in \mathbb{R}^{m\times d}$ encodes the edge information between variables. Suppose there is a graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ where nodes $\mathcal{V}$ are the attributes and each edge is assigned a weight s_{ij} that represents some kind of similarity between attribute i and attribute j. Let $\mathcal{E} = \{e_1, \ldots, e_m\}$ denote a set of m edges, where an edge $e_\tau = (i_\tau, j_\tau)$ consists of a tuple of two attributes. Then the τ-th row of F matrix can be formed by setting $F_{\tau,i_\tau} = s_{i_\tau,j_\tau}$ and $F_{\tau,j_\tau} = -s_{i_\tau,j_\tau}$ for $(i_\tau, j_\tau) \in \mathcal{E}$, and zeros for other entries.



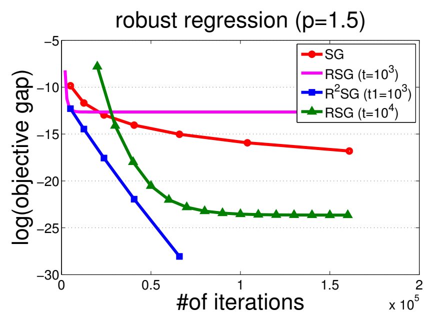
(a) different t



(b) different t

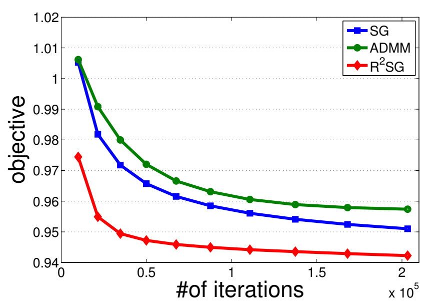


(c) different algorithms

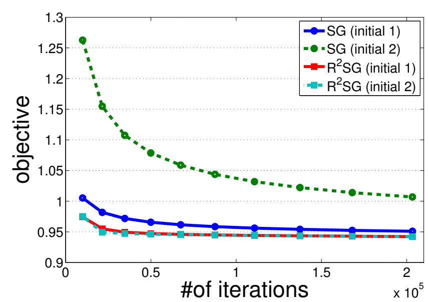


(d) different algorithms

FIG. 3. *Comparison of RSG with different t and of different algorithms on the space-ga data.*



(a) comparison between different algorithms



(b) sensitivity to initial solutions

FIG. 4. *Results for solving SVM classification with GFlassso regularizer. In (b), the objective values of the two initial solutions are* 1 *and* 73.75*.*

Then the GFlasso becomes $R(\mathbf{w}) = \lambda \sum_{(i,j)\in\mathcal{E}} s_{ij}|w_i - w_j|$. Previous studies have found that a carefully designed GFlasso regularization helps in reducing the risk of over-fitting. In this experiment, we follow [40] to generate a dependency graph by sparse inverse covariance selection [17]. To this end, we first generate a sparse inverse covariance matrix using the method in [17] and then assign an equal weight $s_{ij} = 1$ to all edges that have non-zero entries in the resulting inverse covariance matrix. We conduct the experiment on the dna data ($n = 2000$ and $d = 180$) from the libsvm website, which has three class labels. We solve the above problem to classify class 3 versus the rest. Besides SG, we also compare with another baseline, namely alternating direction method of multiplers (ADMM). A stochastic variant of ADMM has been employed to solve the above problem in [40] by writing the problem as

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n \ell(\mathbf{w}^\top\mathbf{x}_i, y_i) + \lambda\|\mathbf{u}\|_1, \quad s.t. \quad \mathbf{u} = F\mathbf{w}$$

For fairness, we compare with deterministic ADMM. Since the intermediate problems associated with $\mathbf{w}$ are also difficult to be solved, we therefore follow the approach in [50] [8] to linearize the hinge loss part at every iteration. The difference between the approach in [50] and [40] is that the former uses a special proximal term $\frac{1}{2\eta_\tau}(\mathbf{w}-\mathbf{w}_\tau)^\top G_\tau(\mathbf{w}-\mathbf{w}_\tau)$ to compute $\mathbf{w}_{\tau+1}$ at each iteration while the latter simply uses $\frac{1}{2\eta_\tau}\|\mathbf{w}-\mathbf{w}_\tau\|_2^2$, where η_τ is the step size and G_τ is a PSD matrix. We compared these two approaches and found that the variant in [50] works better for this problem and hence we only report its performance. The comparison between different algorithms starting from an initial solution with all zero entries for solving the above problem with $\lambda = 0.1$ is presented in Figure 4(a). For R^{2}SG, we start from $t_1 = 10^3$ and restart it every 10 stages with t increased by a factor of 1.15. The initial step sizes for all algorithms are tuned, and so is the penalty parameter in ADMM.

Finally, we compare the dependence of R^{2}SG's convergence on the initial solution with that of SG. We use two different initial solutions (the first initial solution $\mathbf{w}_0 = 0$ and the second initial solution $\mathbf{w}_0$ is generated once from a normal Gaussian distribution). The convergence curves of the two algorithms from the two different initial solutions are plotted in Figure 4(b). Note that the initial step sizes of SG and R^{2}SG are separately tuned for each initial solution. We can see that R^{2}SG is much less sensitive to a bad initial solution than SG consistent with our theory.

9. Conclusion. In this work, we have proposed a novel restarted subgradient method for non-smooth and/or non-strongly convex optimization for obtaining an ϵ-optimal solution. By leveraging the lower bound of the first-order optimality residual, we establish a generic complexity of RSG that improves over standard subgradient method. We have also considered several classes of non-smooth and non-strongly convex problems that admit a local error bound condition and derived the improved order of iteration complexities for RSG. Several extensions have been made to generalize the theory of RSG to a p-norm space and to design a parameter-free variant of RSG without requiring the knowledge of the multiple constant in the local error bound condition. Experimental results on several machine learning tasks have demonstrated the effectiveness of the proposed algorithms in comparison to the subgradient method.

Acknolwedgements. We thank James Renegar for pointing out the connection to his work and for his valuable comments on the difference between the two work. We also thank to Nghia T.A. Tran for pointing out the connection between the local error bound and metric subregularity of subdifferentials. Thanks to Mingrui Liu for spotting an error in the formulation of the F matrix for GFlasso in earlier versions. T. Yang is supported by NSF (1463988, 1545995).

Appendix A. Proof of Lemma 9. The proof below is motivated by the proof of Lemma 4 in [20]. However, since our Lemma 9 generalizes Lemma 4 in [20] by allowing the feasible set Ω to be unbounded, additional technical challenges are introduced in its proof, which lead to a parameter κ with a definition different from the parameter δ in [20]. In the proof, we let $\|\cdot\|$ denote any valid norm.

[8] The OPG variant.

Since the epigraph is polyhedron, by Minkowski-Weyl theorem [11], there exist finitely many feasible solutions $\{\mathbf{w}_i, i = 1,\ldots,M\}$ with the associated objective values $\{f_i = f_i(\mathbf{w}_i), i = 1,\ldots,M\}$ and finitely many directions $V = \left\{\begin{bmatrix}\mathbf{u}_1\\ s_1\end{bmatrix},\ldots,\begin{bmatrix}\mathbf{u}_E\\ s_E\end{bmatrix}\right\}$ in $\mathbb{R}^{p+1}$ such that

$$\begin{aligned}&\mathrm{epi}(f) = \mathrm{conv}\{(\mathbf{w}_i, f_i), i = 1,\ldots,M\} + \mathrm{cone}(V)\\ &= \left\{(\mathbf{w}, t) : \begin{pmatrix}\mathbf{w}\\ \\ t\end{pmatrix} = \begin{pmatrix}\sum_{i=1}^M \lambda_i \mathbf{w}_i + \sum_{j=1}^E \gamma_j \mathbf{u}_j\\ \\ \sum_{i=1}^M \lambda_i f_i + \sum_{j=1}^E \gamma_j s_j\end{pmatrix}, \lambda\in\Delta, \gamma\in\mathbb{R}_+\right\}.\end{aligned}$$

Thus, we can express $f(\mathbf{w})$ as

$$f(\mathbf{w}) = \min_{\lambda,\gamma}\left\{\sum_{i=1}^M \lambda_i f_i + \sum_{j=1}^E \gamma_j s_j : \mathbf{w} = \sum_{i=1}^M \lambda_i \mathbf{w}_i + \sum_{j=1}^E \gamma_j \mathbf{u}_j, \lambda\in\Delta, \gamma\in\mathbb{R}_+\right\}. \tag{33}$$

Since $\min_{\mathbf{w}\in\Omega} f(\mathbf{w}) = f_*$, we have $\min_{1\leq i\leq M} f_i = f_*$ and $s_j \geq 0, \forall j$. We temporarily assume that $f_1 \geq f_2 \geq \ldots \geq f_N > f_* = f_{N+1} = \ldots = f_M$ for $N \geq 1$. We denote by $\mathcal{S}\subset[E]$ the indices such that $s_j \neq 0$ and by $\mathcal{S}^c$ the complement. For any $\gamma\in\mathbb{R}^E_+$, we let $\gamma_{\mathcal{S}}$ denote a vector that contains elements γ_i such that $i\in\mathcal{S}$.

From (33), we can see that there exist $\lambda\in\Delta$ and $\gamma\in\mathbb{R}_+$, such that

$$\mathbf{w} = \sum_{i=1}^M \lambda_i\mathbf{w}_i + \sum_{j\in\mathcal{S}}\gamma_j\mathbf{u}_j + \sum_{j\in\mathcal{S}^c}\gamma_j\mathbf{u}_j, \quad \text{and} \quad f(\mathbf{w}) = \sum_{i=1}^N \lambda_i f_i + f_*\sum_{i=N+1}^M \lambda_i + \sum_{j\in\mathcal{S}}\gamma_j s_j.$$

Define $\mathbf{w}_s = \sum_{i=1}^M \lambda_i\mathbf{w}_i + \sum_{j\in\mathcal{S}}\gamma_j\mathbf{u}_j$. Then

$$\begin{aligned}\|\mathbf{w}-\mathbf{w}^+\| &\leq \min_{\lambda'\in\Delta,\gamma'\in\mathbb{R}_+}\left\|\mathbf{w} - \left(\sum_{i=N+1}^M \lambda'_i\mathbf{w}_i + \sum_{j\in\mathcal{S}^c}\gamma'_j\mathbf{u}_j\right)\right\|\\ &= \min_{\lambda'\in\Delta,\gamma'\in\mathbb{R}_+}\left\|\mathbf{w}_s + \sum_{j\in\mathcal{S}^c}\gamma_j\mathbf{u}_j - \left(\sum_{i=N+1}^M \lambda'_i\mathbf{w}_i + \sum_{j\in\mathcal{S}^c}\gamma'_j\mathbf{u}_j\right)\right\|\\ &\leq \min_{\lambda'\in\Delta}\|\mathbf{w}_s - \sum_{i=N+1}^M \lambda'_i\mathbf{w}_i\| + \min_{\gamma'\in\mathbb{R}_+}\|\sum_{j\in\mathcal{S}^c}\gamma_j\mathbf{u}_j - \sum_{j\in\mathcal{S}^c}\gamma'_j\mathbf{u}_j\|\\ &= \min_{\lambda'\in\Delta}\|\sum_{i=1}^M \lambda_i\mathbf{w}_i + \sum_{j\in\mathcal{S}}\gamma_j\mathbf{u}_j - \sum_{i=N+1}^M \lambda'_i\mathbf{w}_i\|\\ &\leq \min_{\lambda'\in\Delta}\|\sum_{i=1}^M \lambda_i\mathbf{w}_i - \sum_{i=N+1}^M \lambda'_i\mathbf{w}_i\| + \|\sum_{j\in\mathcal{S}}\gamma_j\mathbf{u}_j\|,\end{aligned} \tag{34}$$

where the first inequality is due to that $\sum_{i=N+1}^M \lambda'_i\mathbf{w}_i + \sum_{j\in\mathcal{S}^c}\gamma'_j\mathbf{u}_j\in\Omega_*$. Next, we will bound the two terms in the R.H.S of the above inequality.

To proceed, we construct δ and σ as follows:

$$\delta = \frac{f_N - f_*}{\max_{i,\mathbf{u}}\{\|\mathbf{w}_i - \mathbf{u}\| : i = 1,\ldots,N, \mathbf{u}\in\Omega_*\}} > 0, \quad \sigma = \min_{j\in\mathcal{S}:\|\mathbf{u}_j\|\neq 0}\frac{s_j}{\|\mathbf{u}_j\|} > 0.$$

Let $\mathbf{w}' = \sum_{i=1}^{M} \lambda_i \mathbf{w}_i$ and $\mu = \sum_{i=1}^{N} \lambda_i$.

Suppose $\mu > 0$. Let $\hat{\lambda}_i, i = 1, \ldots, N$ be defined as $\hat{\lambda}_i = \frac{\lambda_i}{\mu}, i = 1, \ldots, N$. Further define

$$\widehat{\mathbf{w}} = \sum_{i=1}^{N} \hat{\lambda}_i \mathbf{w}_i, \quad \text{and} \quad \widetilde{\mathbf{w}} = \begin{cases} \sum_{i=N+1}^{M} \mathbf{w}_i \frac{\lambda_i}{1-\mu} & \text{if } \mu < 1 \\ \mathbf{w}_M & \text{if } \mu = 1 \end{cases}.$$

As a result, $\mathbf{w}' = \mu\widehat{\mathbf{w}} + (1-\mu)\widetilde{\mathbf{w}}$.

For the first term in the R.H.S of (34), we have

$$\min_{\lambda' \in \Delta} \left\| \sum_{i=1}^{M} \lambda_i \mathbf{w}_i - \sum_{i=N+1}^{M} \lambda'_i \mathbf{w}_i \right\| = \min_{\lambda' \in \Delta} \left\| \mathbf{w}' - \sum_{i=N+1}^{M} \lambda'_i \mathbf{w}_i \right\| \leq \|\mathbf{w}' - \widetilde{\mathbf{w}}\|.$$

To continue,

$$\begin{aligned} \|\mathbf{w}' - \widetilde{\mathbf{w}}\| &= \mu \|\widehat{\mathbf{w}} - \widetilde{\mathbf{w}}\| \leq \mu \left\| \sum_{i=1}^{N} \hat{\lambda}_i (\mathbf{w}_i - \widetilde{\mathbf{w}}) \right\| \leq \mu \sum_{i=1}^{N} \hat{\lambda}_i \|\mathbf{w}_i - \widetilde{\mathbf{w}}\| \\ &\leq \mu \max_i \{\|\mathbf{w}_i - \widetilde{\mathbf{w}}\|, i = 1, \ldots, N, \widetilde{\mathbf{w}} \in \Omega_*\} \leq \frac{\mu(f_N - f_*)}{\delta}. \end{aligned}$$

where the last inequality is due to the definition of δ. On the other hand, we can represent

$$f(\mathbf{w}) = \sum_{i=1}^{M} \lambda_i f_i + \sum_{j \in \mathcal{S}} \gamma_j s_j = \mu \sum_{i=1}^{N} \hat{\lambda}_i f_i + f_*(1-\mu) + \sum_{j \in \mathcal{S}} \gamma_j s_j.$$

Since $s_j \geq 0$ for all j, we have

$$f(\mathbf{w}) - f_* \geq \mu \sum_{i=1}^{N} \hat{\lambda}_i (f_i - f_*) + \sum_{j \in \mathcal{S}} \gamma_j s_j \geq \max \left\{ \mu(f_N - f_*), \sum_{j \in \mathcal{S}} \gamma_j s_j \right\}.$$

Thus,

$$\min_{\lambda' \in \Delta} \left\| \sum_{i=1}^{M} \lambda_i \mathbf{w}_i - \sum_{i=N+1}^{M} \lambda'_i \mathbf{w}_i \right\| \leq \|\mathbf{w}' - \widetilde{\mathbf{w}}\| \leq \frac{f(\mathbf{w}) - f_*}{\delta}.$$

Suppose $\mu = 0$ (and thus $\lambda_i = 0$ for $i = 1, 2, \ldots, N$). We can still have

$$\min_{\lambda' \in \Delta} \left\| \sum_{i=1}^{M} \lambda_i \mathbf{w}_i - \sum_{i=N+1}^{M} \lambda'_i \mathbf{w}_i \right\| \leq \min_{\lambda' \in \Delta} \left\| \sum_{i=N+1}^{M} \lambda_i \mathbf{w}_i - \sum_{i=N+1}^{M} \lambda'_i \mathbf{w}_i \right\| = 0 \leq \frac{f(\mathbf{w}) - f_*}{\delta},$$

and can still represent

$$f(\mathbf{w}) = \sum_{i=1}^{M} \lambda_i f_i + \sum_{j \in \mathcal{S}} \gamma_j s_j = f_* + \sum_{j \in \mathcal{S}} \gamma_j s_j$$

so that $f(\mathbf{w}) - f_* \geq \sum_{j \in \mathcal{S}} \gamma_j s_j$. As a result, no matter $\mu = 0$ or $\mu > 0$, we can bound the first term in the R.H.S of (34) by $\frac{f(\mathbf{w}) - f_*}{\delta}$. For the second term in the R.H.S of (34), by the definition of σ , we have

$$\|\sum_{j \in \mathcal{S}} \gamma_j \mathbf{u}_j\| \leq \sum_{j \in \mathcal{S}} \gamma_j \|\mathbf{u}_j\| \leq \frac{1}{\sigma} \sum_{j \in \mathcal{S}} \gamma_j s_j \leq \frac{f(\mathbf{w}) - f_*}{\sigma},$$

Combining the results above with $\frac{1}{\kappa} = \left(\frac{1}{\delta} + \frac{1}{\sigma}\right)$, we have

$$\|\mathbf{w} - \mathbf{w}^+\| \leq \frac{1}{\kappa}(f(\mathbf{w}) - f_*).$$

Finally, we note that when $f_1 = \ldots = f_M = f_*$, the Lemma is trivially proved following the same analysis except that $\delta > 0$ can be any positive value.

Appendix B. A proposition needed to prove Corollary 15. The proof of Corollary 15 leverages the following result from [8].

PROPOSITION 24. *[8, Theorem 5] Let $f(x)$ be an extended-valued, proper, convex and lower semicontinuous function that satisfies the KL inequality (24) at $x_* \in \arg\min f(\cdot)$ for all $x \in U \cap \{x : f(x_*) < f(x) < f(x_*) + \eta\}$, where U is a neighborhood of x_*, then $dist(x, \arg\min f(\cdot)) \leq \varphi(f(x) - f(x_*))$ for all $x \in U \cap \{x : f(x_*) < f(x) < f(x_*) + \eta\}$.*

REFERENCES

[1] F. J. A. Artacho and M. H. Geoffroy, *Characterization of metric regularity of subdifferentials*, Journal of Convex Analysis, 15 (2008), pp. 365–380.

[2] H. Attouch, J. Bolte, P. Redont, and A. Soubeyran, *Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the kurdyka-lojasiewicz inequality*, Math. Oper. Res., 35 (2010), pp. 438–457.

[3] H. Attouch, J. Bolte, and B. F. Svaiter, *Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized gauss-seidel methods.*, Math. Program., 137 (2013), pp. 91–129.

[4] F. R. Bach, *Learning with submodular functions: A convex optimization perspective*, Foundations and Trends in Machine Learning, 6 (2013), pp. 145–373.

[5] F. R. Bach and E. Moulines, *Non-strongly-convex smooth stochastic approximation with convergence rate o(1/n)*, in Advances in Neural Information Processing Systems (NIPS), 2013, pp. 773–781.

[6] J. Bolte, A. Daniilidis, and A. Lewis, *The łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems*, SIAM J. on Optimization, 17 (2006), pp. 1205–1223.

[7] J. Bolte, A. Daniilidis, A. S. Lewis, and M. Shiota, *Clarke subgradients of stratifiable functions.*, SIAM Journal on Optimization, 18 (2007), pp. 556–572.

[8] J. Bolte, T. P. Nguyen, J. Peypouquet, and B. Suter, *From error bounds to the complexity of first-order descent methods for convex functions*, CoRR, abs/1510.08234 (2015).

[9] J. Bolte, S. Sabach, and M. Teboulle, *Proximal alternating linearized minimization for nonconvex and nonsmooth problems*, Mathematical Programming, 146 (2014), pp. 459–494.

[10] X. Chen, Q. Lin, and J. Pena, *Optimal regularized dual averaging methods for stochastic optimization*, in Advances in Neural Information Processing Systems (NIPS), 2012, pp. 395–403.

[11] G. Dahl, *An introduction to convexity, polyhedral theory and combinatorial optimization.*

[12] D. Drusvyatskiy and C. Kempton, *An accelerated algorithm for minimizing convex compositions*, arXiv:1605.00125, (2016).

[13] D. Drusvyatskiy and A. S. Lewis, *Error bounds, quadratic growth, and linear convergence of proximal methods*, arXiv:1602.06661, (2016).

[14] D. Drusvyatskiy, B. S. Mordukhovich, and T. T. A. Nhgia, *Second-order growth, tilt stability, and metric regularity of the subdifferential*, Journal of Convex Analysis, 21 (2014), pp. 1165 – 1192.

[15] J. C. Duchi, S. Shalev-Shwartz, Y. Singer, and A. Tewari, *Composite objective mirror descent*, in COLT 2010 - The 23rd Conference on Learning Theory, Haifa, Israel, June 27-29, 2010, 2010, pp. 14–26.

[16] R. M. Freund and H. Lu, *New computational guarantees for solving convex optimization problems with first order methods, via a function growth condition measure*, ArXiv e-prints, (2015).

[17] J. Friedman, T. Hastie, and R. Tibshirani, *Sparse inverse covariance estimation with the graphical lasso*, Biostatistics, 9 (2008).
[18] S. Ghadimi and G. Lan, *Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization I: A generic algorithmic framework*, SIAM Journal on Optimization, 22 (2012), pp. 1469–1492.
[19] S. Ghadimi and G. Lan, *Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, II: shrinking procedures and optimal algorithms*, SIAM Journal on Optimization, 23 (2013), pp. 2061–2089.
[20] A. Gilpin, J. Peña, and T. Sandholm, *First-order algorithm with log(1/epsilon) convergence for epsilon-equilibrium in two-person zero-sum games*, Math. Program., 133 (2012), pp. 279–298.
[21] R. Goebel and R. T. Rockafellar, *Local strong convexity and local lipschitz continuity of the gradient of convex functions*, Journal of Convex Analysis, (2007).
[22] P. Gong and J. Ye, *Linear convergence of variance-reduced projected stochastic gradient without strong convexity*, CoRR, abs/1406.1102 (2014).
[23] E. Hazan and S. Kale, *Beyond the regret minimization barrier: an optimal algorithm for stochastic strongly-convex optimization*, in Proceedings of the 24th Annual Conference on Learning Theory (COLT), 2011, pp. 421–436.
[24] K. Hou, Z. Zhou, A. M. So, and Z. Luo, *On the linear convergence of the proximal gradient method for trace norm regularization*, in Advances in Neural Information Processing Systems (NIPS), 2013, pp. 710–718.
[25] H. Karimi and M. Schmidt, *Linear convergence of proximal-gradient methods under the polyak-łojasiewicz condition*, The 8th NIPS Workshop on Optimization for Machine Learning, (2015).
[26] S. Kim, K. Sohn, and E. P. Xing, *A multivariate regression approach to association analysis of a quantitative trait network*, Bioinformatics, 25 (2009).
[27] A. Krizhevsky, I. Sutskever, and G. E. Hinton, *Imagenet classification with deep convolutional neural networks*, in Advances in Neural Information Processing Systems (NIPS), 2012, pp. 1106–1114.
[28] A. Y. Kruger, *Error bounds and hölder metric subregularity*, Set-Valued and Variational Analysis, 23 (2015), pp. 705–736.
[29] Z.-Q. Luo and P. Tseng, *On the convergence of coordinate descent method for convex differentiable minization*, Journal of Optimization Theory and Applications, 72 (1992), pp. 7–35.
[30] Z.-Q. Luo and P. Tseng, *On the linear convergence of descent methods for convex essenially smooth minization*, SIAM Journal on Control and Optimization, 30 (1992), pp. 408–425.
[31] Z.-Q. Luo and P. Tseng, *Error bounds and convergence analysis of feasible descent methods: a general approach*, Annals of Operations Research, 46 (1993), pp. 157–178.
[32] B. Mordukhovich and W. Ouyang, *Higher-order metric subregularity and its applications*, Journal of Global Optimization- An International Journal Dealing with Theoretical and Computational Aspects of Seeking Global Optima and Their Applications in Science, Management and Engineering, 63 (2015), pp. 777–795.
[33] I. Necoara, Y. Nesterov, and F. Glineur, *Linear convergence of first order methods for non-strongly convex optimization*, CoRR, abs/1504.06298 (2015).
[34] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, *Robust stochastic approximation approach to stochastic programming*, SIAM Journal on Optimization, 19 (2009), pp. 1574–1609.
[35] A. S. Nemirovsky A.S. and D. B. Yudin, *Problem complexity and method efficiency in optimization*, Wiley-Interscience series in discrete mathematics, Wiley, Chichester, New York, 1983. A Wiley-Interscience publication.
[36] Y. Nesterov, *Introductory lectures on convex optimization : a basic course*, Applied optimization, Kluwer Academic Publ., 2004.
[37] Y. Nesterov, *Smooth minimization of non-smooth functions*, Mathematical Programming, 103 (2005), pp. 127–152.
[38] Y. Nesterov, *Primal-dual subgradient methods for convex problems*, Mathematical Programming, 120 (2009), pp. 221–259.
[39] Y. Nesterov and A. Juditsky, *Primal-dual subgradient methods for minimizing uniformly convex functions*, arXiv:1401.1792, (2014).
[40] H. Ouyang, N. He, L. Tran, and A. G. Gray, *Stochastic alternating direction method of multipliers*, in Proceedings of the 30th International Conference on Machine Learning (ICML), 2013, pp. 80–88.
[41] J.-S. Pang, *A posteriori error bounds for the linearly-constrained variational inequality problem*, Mathematics of Operations Research, 12 (1987), pp. 474–484.

[42] J.-S. Pang, *Error bounds in mathematical programming*, Mathematical Programming, 79 (1997), pp. 299–332.
[43] Y. Plan and R. Vershynin, *One-bit compressed sensing by linear programming*, CoRR, abs/1109.4299 (2011).
[44] J. Renegar, *Efficient first-order methods for linear programming and semidefinite programming*, ArXiv e-prints, (2014).
[45] J. Renegar, *A framework for applying subgradient methods to conic optimization problems*, ArXiv e-prints, (2015).
[46] J. Renegar, *"efficient" subgradient methods for general convex optimization*, ArXiv e-prints, (2016).
[47] R. Rockafellar, *Convex Analysis*, Princeton mathematical series, Princeton University Press, 1970.
[48] R. Schneider and A. Uschmajew, *Convergence results for projected line-search methods on varieties of low-rank matrices via łojasiewicz inequality*, CoRR, abs/1402.5284 (2014).
[49] A. M. So, *Non-asymptotic convergence analysis of inexact gradient methods for machine learning without strong convexity*, CoRR, abs/1309.0113 (2013).
[50] T. Suzuki, *Dual averaging and proximal gradient descent for online alternating direction multiplier method.*, in Proceedings of the International Conference on Machine Learning (ICML), vol. 28, 2013, pp. 392–400.
[51] P. Tseng and Y. S., *A block-coordinate gradient descent method for linearly constrained nonsmooth separable optimization*, Journal of Optimization Theory Application, 140 (2009), pp. 513–535.
[52] P. Tseng and Y. S., *A coordinate gradient descent method for nonsmooth separable minimization*, Mathematical Programming, 117 (2009), pp. 387–423.
[53] P. Wang and C. Lin, *Iteration complexity of feasible descent methods for convex optimization*, Journal of Machine Learning Research, 15 (2014), pp. 1523–1548.
[54] Y. Xu, Q. Lin, and T. Yang, *Accelerate stochastic subgradient method by leveraging local error bound*, CoRR, abs/1607.01027 (2016).
[55] Y. Xu, Y. Yan, Q. Lin, and T. Yang, *Homotopy smoothing for non-smooth problems with lower complexity than 1/epsilon*, CoRR, abs/1607.03815 (2016).
[56] T. Yang, Q. Lin, and Z. Li, *Unified convergence analysis of stochastic momentum methods for convex and non-convex optimization*, coRR, (2016).
[57] T. Yang, Q. Lin, and L. Zhang, *A richer theory of convex constrained optimization with reduced projections and improved rates*, arXiv, (2016).
[58] T. Yang, M. Mahdavi, R. Jin, and S. Zhu, *An efficient primal-dual prox method for non-smooth optimization*, Machine Learning, (2014).
[59] T. Yang, L. Zhang, R. Jin, and S. Zhu, *Theory of dual-sparse regularized randomized reduction*, in Proceedings of the 32nd International Conference on Machine Learning (ICML), 2015, pp. 305–314.
[60] H. Zhang, *Characterization of linear convergence of gradient descent*, arXiv:1606.00269, (2016).
[61] H. Zhang and L. Cheng, *Restricted strong convexity and its applications to convergence analysis of gradient-type methods in convex optimization*, Optimization Letter, 9 (2015), pp. 961–979.
[62] Z. Zhou and A. M.-C. So, *A unified approach to error bounds for structured convex optimization problems*, arXiv:1512.03518, (2015).
[63] Z. Zhou, Q. Zhang, and A. M. So, *L1p-norm regularization: Error bounds and convergence rate analysis of first-order methods*, in Proceedings of the 32nd International Conference on Machine Learning, (ICML), 2015, pp. 1501–1510.
[64] M. Zinkevich, *Online convex programming and generalized infinitesimal gradient ascent*, in Proceedings of the International Conference on Machine Learning (ICML), 2003, pp. 928–936.