Analysis and Applications
© World Scientific Publishing Company

Accelerate Stochastic Subgradient Method by Leveraging Local Growth Condition

Yi Xu

Department of Computer Science
The University of Iowa, Iowa City, IA 52242, USA
yi-xu@uiowa.edu

Qihang Lin

Department of Management Sciences
The University of Iowa, Iowa City, IA 52242, USA
qihang-lin@uiowa.edu

Tianbao Yang

Department of Computer Science
The University of Iowa, Iowa City, IA 52242, USA
tianbao-yang@uiowa.edu

Received (Day Month Year)
Revised (Day Month Year)

In this paper, a new theory is developed for first-order stochastic convex optimization, showing that the global convergence rate is sufficiently quantified by a local growth rate of the objective function in a neighborhood of the optimal solutions. In particular, if the objective function $F(\mathbf{w})$ in the ϵ-sublevel set grows as fast as $\|\mathbf{w}-\mathbf{w}_*\|_2^{1/\theta}$, where $\mathbf{w}_*$ represents the closest optimal solution to $\mathbf{w}$ and $\theta \in (0,1]$ quantifies the local growth rate, the iteration complexity of first-order stochastic optimization for achieving an ϵ-optimal solution can be $\widetilde{O}(1/\epsilon^{2(1-\theta)})$, which is *optimal at most* up to a logarithmic factor. To achieve the faster global convergence, we develop two different **accelerated stochastic subgradient** methods by iteratively solving the original problem approximately in a local region around a historical solution with the size of the local region gradually decreasing as the solution approaches the optimal set. Besides the theoretical improvements, this work also includes new contributions towards making the proposed algorithms practical: (i) we present practical variants of accelerated stochastic subgradient methods that can run without the knowledge of multiplicative growth constant and even the growth rate θ; (ii) we consider a broad family of problems in machine learning to demonstrate that the proposed algorithms enjoy faster convergence than traditional stochastic subgradient method. We also characterize the complexity of the proposed algorithms for ensuring the gradient is small without the smoothness assumption.

Keywords: Convex Optimization; Stochastic Subgradient; Local Growth Condition.

Mathematics Subject Classification 2000: 46N10, 60H30, 49J52

1. Introduction

In this paper, we are interested in solving the following stochastic optimization problem:

$$\min_{\mathbf{w}\in\mathcal{K}} F(\mathbf{w}) \triangleq \mathrm{E}_{\xi}[f(\mathbf{w};\xi)], \tag{1.1}$$

where ξ is a random variable, $f(\mathbf{w};\xi)$ is a convex function of $\mathbf{w}$, $\mathrm{E}_{\xi}[\cdot]$ is the expectation over ξ and $\mathcal{K}$ is a convex domain. We denote by $\partial f(\mathbf{w};\xi)$ a subgradient of $f(\mathbf{w};\xi)$. Let $\mathcal{K}_*$ denote the optimal set of (1.1) and F_* denote the optimal value.

In recent years, it becomes very important to develop efficient and effective optimization algorithms for solving large-scale machine learning problems [12,17,6]. Traditional stochastic subgradient (SSG) method updates the solution according to

$$\mathbf{w}_{t+1} = \Pi_{\mathcal{K}}[\mathbf{w}_t - \eta_t \partial f(\mathbf{w}_t;\xi_t)], \tag{1.2}$$

for $t = 1, \ldots, T$, where ξ_t is a sampled value of ξ at t-th iteration, η_t is a step size and $\Pi_{\mathcal{K}}[\mathbf{w}] = \arg\min_{\mathbf{v}\in\mathcal{K}} \|\mathbf{w}-\mathbf{v}\|_2^2$ is a projection operator that projects a point into $\mathcal{K}$. Previous studies have shown that under the following assumptions i) $\|\partial f(\mathbf{w};\xi)\|_2 \leq G$, ii) there exists $\mathbf{w}_* \in \mathcal{K}_*$ such that $\|\mathbf{w}_t - \mathbf{w}_*\|_2 \leq B$ for $t = 1, \ldots, T$ [a], and by setting the step size $\eta_t = \frac{B}{G\sqrt{T}}$ in (1.2), with a high probability $1-\delta$ we have

$$F(\widehat{\mathbf{w}}_T) - F_* \leq O\left(GB(1+\sqrt{\log(1/\delta)})/\sqrt{T}\right), \tag{1.3}$$

where $\widehat{\mathbf{w}}_T = \sum_{t=1}^{T} \mathbf{w}_t/T$. The above convergence implies that in order to obtain an ϵ-optimal solution by SSG, i.e., finding an $\mathbf{w}$ such that $F(\mathbf{w}) - F_* \leq \epsilon$ with a high probability $1-\delta$, one needs at least $T = O(G^2B^2(1+\sqrt{\log(1/\delta)})^2/\epsilon^2)$ in the worst-case.

It is commonly known that the slow convergence of SSG is due to the variance in the stochastic subgradient and the non-smoothness nature of the problem as well, which therefore requires a decreasing step size or a very small step size. Recently, there emerges a stream of studies on various variance reduction techniques to accelerate stochastic **gradient** method [43,56,21,47,10]. However, they all hinge on the smoothness assumption. The proposed algorithms in this work tackle the issue of variance of **stochastic subgradient** without the smoothness assumption from another pespective.

The main motivation for addressing this problem is from a key observation: a high probability analysis of the SSG method shows that the variance term of the stochastic subgradient is accompanied by an upper bound of distance of intermediate solutions to the *target* solution. This observation has also been leveraged in previous analysis to design faster convergence for stochastic convex optimization that use a strong or uniform convexity condition [19,22] or a global growth condition [40] to control the distance of intermediate solutions to the *optimal* solution

[a] This holds if we assume the domain $\mathcal{K}$ is bounded such that $\max_{\mathbf{w},\mathbf{v}\in\mathcal{K}} \|\mathbf{w}-\mathbf{v}\|_2 \leq B$ or if assume $dist(\mathbf{w}_1,\mathcal{K}_*) \leq B/2$ and project every solution $\mathbf{w}_t$ into $\mathcal{K}\cap\mathcal{B}(\mathbf{w}_1, B/2)$.

by their functional residuals. However, we find these global assumptions are completely unnecessary, which may not only restrict their applications to a broad family of problems but also worsen the convergence rate due to the larger multiplicative growth constant that could be domain-size dependent. In contrast, we develop a new theory only relying on the local growth condition to control the distance of intermediate solutions to the ϵ-*optimal* solution by their functional residuals but achieving a fast global convergence.

Besides the fundamental difference, the present work also possesses several unique algorithmic contributions compared with previous similar work on stochastic optimization: (i) we have two different ways to control the distance of intermediate solutions to the ϵ-*optimal* solution, one by explicitly imposing a bounded ball constraint and another one by implicitly regularizing the intermediate solutions, where the later one could be more efficient if the projection into the intersection of a bounded ball and the problem domain is complicated; (ii) we develop more practical variants that can be run without knowing the multiplicative growth constant though under a slightly stringent condition; (iii) for problems whose local growth rate is unknown we still develop an improved convergence result of the proposed algorithms comparing with the SSG method. In addition, the present work will demonstrate the improved results and practicability of the proposed algorithms for many problems in machine learning, which is lacking in similar previous work.

We summarize the main results below. The proposed algorithms and their analysis are developed under the following generic local growth condition (LGC):

$$\|\mathbf{w} - \mathbf{w}_*\|_2 \leq c(F(\mathbf{w}) - F_*)^{\theta}, \quad \forall \mathbf{w} \in \mathcal{S}_{\epsilon}, \tag{1.4}$$

where $\theta \in (0, 1]$, $c > 0$ and $\mathcal{S}_\epsilon$ denotes the ϵ-sublevel set with ϵ being a small value.

- In Section 4, we present two variants of accelerated stochastic subgradient (ASSG) methods and analyze their iteration complexities for finding an ϵ-optimal solution with high probability. The two variants use different ways to mitigate the effect of variance of stochastic subgradient with one using shrinking ball constraints and the second variant using increasing regularization. With complete knowledge of c and θ, we show that both variants can find an ϵ-optimal solution with a complexity of $\widetilde{O}(1/\epsilon^{2(1-\theta)})$ for $\theta \in (0, 1]$, where $\widetilde{O}$ suppresses a logarithmic factor in terms of $1/\epsilon$.
- In Section 5, we present a practical variant of ASSG with partial or no knowledge about the LGC. In particular, when c is unknown and $\theta \in (0, 1)$ is known the practical variant of ASSG enjoys an improved complexity of $\widetilde{O}(1/\epsilon^{2(1-\theta)})$. When both c and θ are unknown, we show that the practical variant still enjoys a better complexity than that of traditional SSG. In particular, the dependence on the distance from the initial solution to the optimal set of SSG's complexity is reduced to a much smaller distance multiplied by a logarithmic factor dependent on the quality of the initial solution.

- In Section 6, we consider an extension to proximal algorithms that handle non-smooth but simple regularizers by a proximal mapping. In Section 7, we consider the complexity of the proposed ASSG algorithms for ensuing the gradient of the objective function is small.
- In Section 8, we consider the applications in machine learning and present many examples with the local growth rate θ explicitly exhibited. In Section 9, we present numerical experiments for demonstrating the effectiveness of the proposed algorithms.

2. Related Work

The most similar work to the present one is [40], which studied stochastic convex optimization under a global growth condition, which they called Tsybakov noise condition. One major difference from their result is that we achieve the same order of iteration complexity up to a logarithmic factor under only a local growth condition. As observed later on, the multiplicative growth constant in local growth condition is domain-size independent that is smaller than that in global growth condition, which could be domain-size dependent. Besides, the stochastic optimization algorithm in [40] assume the *optimization domain $\mathcal{K}$ is bounded*, which is removed in this work. In addition, they do not address the issue when the multiplicative constant is unknown and lack study of applicability for machine learning problems. [22] presented primal-dual subgradient and stochastic subgradient methods for solving problems under the uniform convexity assumption (see the definition under Observation 3.1). As exhibited shortly, the uniform convexity condition covers only a smaller family of problems than the considered local growth condition. However, when the problem is uniform convex, the iteration complexity obtained in this work resembles that in [22].

Recently, there emerge a wave of studies that attempt to improve the convergence of existing algorithms under no strong convexity assumption by considering certain weaker conditions than strong convexity [34,29,55,28,16,24,53,38,45]. Several recent works [34,24,53] have unified many of these conditions, implying that they are a kind of global growth condition with $\theta = 1/2$. Unlike the present work, most of these developments require certain smoothness assumption except [38].

Luo and Tseng [31,32,33] pioneered the idea of using local error bound condition to show faster convergence of gradient descent, proximal gradient descent, and many other methods for a family of structured composite problems (e.g., the LASSO problem). Many follow-up works [20,58,57] have considered different regularizers (e.g., $\ell_{1,2}$ regularizer, nuclear norm regularizer). However, these works only obtained asymptotically faster (i.e., linear) convergence and they hinge on the smoothness on some parts of the problem. [50,49] have considered the same local growth condition (aka local error bound condition in their work) for developing faster deterministic algorithms for non-smooth optimization. However, they did not address the problem of stochastic convex optimization, which restricts their applicability to large-scale problems in machine learning.

Finally, we note that the improved iteration complexity in this paper does not contradict to the lower bound in [35,36]. The bad examples constructed to derive the lower bound for general non-smooth optimization do not satisfy the assumptions made in this work (in particular Assumption 3.1(b)). Recently, [59] characterize the local minimax complexity of stochastic convex optimization by introducing modulus of continuity that measures the size of the "flat set" where the magnitude of the subderivative is a small value. They established a local minimax complexity result when the modulus of continuity has polynomial growth and proposed an adaptive stochastic optimization algorithm for only one-dimensional problems that achieves the local minimax complexity upto a logarithmic factor. It remains unclear which is more generic between LGC and the polynomial growing modulus of continuity.

3. Preliminaries

Recall the notations $\mathcal{K}_*$ and F_* that denote the optimal set of (1.1) and the optimal value, respectively. For the optimization problem in (1.1), we make the following assumption throughout the paper.

Assumption 3.1. *For a stochastic optimization problem (1.1), we assume*

(1) there exist $\mathbf{w}_0 \in \mathcal{K}$ and $\epsilon_0 \geq 0$ such that $F(\mathbf{w}_0) - F_* \leq \epsilon_0$;
(2) There exists a constant G such that $\|\partial f(\mathbf{w}; \xi)\|_2 \leq G$.

Remark: (1) essentially assumes the availability of a lower bound of the optimal objective value, which usually holds for machine learning problems (due to non-negativeness of the objective function). (2) is a standard assumption also made in many previous stochastic gradient-based methods [19,39,40]. By Jensen's inequality, we also have $\|\partial F(\mathbf{w})\|_2 \leq G$. It is notable that unlike previous analysis of SSG, we do not assume the domain $\mathcal{K}$ is bounded. Instead, we will assume the problem satisfies a generic local growth condition as presented shortly.

For any $\mathbf{w} \in \mathcal{K}$, let $\mathbf{w}^*$ denote the closest optimal solution in $\mathcal{K}_*$ to $\mathbf{w}$, i.e., $\mathbf{w}^* = \arg\min_{\mathbf{v}\in\mathcal{K}_*} \|\mathbf{v} - \mathbf{w}\|_2^2$, which is unique. 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., $\mathcal{L}_\epsilon = \{\mathbf{w} \in \mathcal{K} : F(\mathbf{w}) = F_* + \epsilon\}$, $\mathcal{S}_\epsilon = \{\mathbf{w} \in \mathcal{K} : F(\mathbf{w}) \leq F_* + \epsilon\}$. Let $\mathbf{w}_\epsilon^\dagger$ denote the closest point in the ϵ-sublevel set to $\mathbf{w}$, i.e.,

$$\mathbf{w}_\epsilon^\dagger = \arg\min_{\mathbf{v}\in\mathcal{S}_\epsilon} \|\mathbf{v} - \mathbf{w}\|_2^2. \tag{3.1}$$

It is easy to show that $\mathbf{w}_\epsilon^\dagger \in \mathcal{L}_\epsilon$ when $\mathbf{w} \notin \mathcal{S}_\epsilon$ (using the KKT condition). Let $\mathcal{B}(\mathbf{w}, r) = \{\mathbf{u} \in \mathbb{R}^d : \|\mathbf{u} - \mathbf{w}\|_2 \leq r\}$ denote an Euclidean ball centered at $\mathbf{w}$ with a radius r. Denote by $dist(\mathbf{w}, \mathcal{K}_*) = \min_{\mathbf{v}\in\mathcal{K}_*} \|\mathbf{w} - \mathbf{v}\|_2$ the distance between $\mathbf{w}$ and the set $\mathcal{K}_*$, by $\partial^0 F(\mathbf{w})$ the projection of 0 onto the nonempty closed convex set $\partial F(\mathbf{w})$, i.e., $\|\partial^0 F(\mathbf{w})\|_2 = \min_{\mathbf{v}\in\partial F(\mathbf{w})} \|\mathbf{v}\|_2$.

3.1. *Functional Local Growth Rate*

We quantify the functional local growth rate by measuring how fast the functional value increase when moving a point away from the optimal solution in the ϵ-sublevel set. In particular, we state the local growth condition in the following assumption.

Assumption 3.2. *The objective function $F(\cdot)$ satisfies a local growth condition on $\mathcal{S}_\epsilon$ if there exists a constant $c > 0$ and $\theta \in (0, 1]$ such that:*

$$\|\mathbf{w} - \mathbf{w}_*\|_2 \leq c(F(\mathbf{w}) - F_*)^\theta, \quad \forall \mathbf{w} \in \mathcal{S}_\epsilon, \tag{3.2}$$

where $\mathbf{w}_*$ is the closest solution in the optimal set $\mathcal{K}_*$ to $\mathbf{w}$.

Note that the local growth rate θ is at most 1. This is due to that $F(\mathbf{w})$ is G-Lipschitz continuous and $\lim_{\mathbf{w}\to\mathbf{w}_*} \|\mathbf{w} - \mathbf{w}_*\|_2^{1-\alpha} = 0$ if $\alpha < 1$. The inequality in (3.2) is also called as local error bound condition in [50]. In this work, to avoid confusion with earlier work by [31,32,33] who also explored a related but different local error bound condition, we refer to the inequality in (3.2) or (1.4) as local growth condition (LGC). It is worth noting that LGC is a general condition, comparing with several other error bound conditions. For example, the polyhedral error bound condition [50] implies LGC with $\theta = 1$; while the function has a Lipschitz-continuous gradient, then the Polyak-Łojasiewicz condition is equivalent to the LGC with $\theta = 1/2$. In Section 8, we will present several applications in risk minimization problems that satisfying LGC. For more details about the relationship between LGC and other conditions, we refer the reader to [24,8,54,50]. If the function $F(\mathbf{x})$ is assumed to satisfy (3.2) for all $\mathbf{w} \in \mathcal{K}$, it is referred to as global growth condition (GGC). Note that since we do not assume a bounded $\mathcal{K}$, the GGC might be ill posed. In the following discussions, when compared with GGC we simply assume the domain is bounded.

Below, we present several observations mostly from existing work to clarify the relationship between the LGC (1.4) and previous conditions, and also justify our choice of LGC that covers a much broader family of functions than previous conditions and induces a smaller multiplicative growth constant c than that induced by GGC.

Observation 3.1. Strong convexity or uniform convexity condition implies LGC with $\theta = 1/2$, but not vice versa.

$F(\mathbf{w})$ is said to satisfy a uniform convexity condition on $\mathcal{K}$ with convexity parameters $p \geq 2$ and μ if:

$$F(\mathbf{u}) \geq F(\mathbf{v}) + \partial F(\mathbf{v})^\top(\mathbf{u} - \mathbf{v}) + \frac{\mu\|\mathbf{u} - \mathbf{v}\|_2^p}{2}, \forall \mathbf{u}, \mathbf{v} \in \mathcal{K}.$$

If we let $\mathbf{u} = \mathbf{w}$, $\mathbf{v} = \mathbf{w}_*$, then $\partial F(\mathbf{w}_*)^\top(\mathbf{w} - \mathbf{w}_*) \geq 0$ for any $\mathbf{w} \in \mathcal{K}$, and we have (3.2) with $\theta = 1/p \in (0, 1/2]$. Clearly LGC covers a broader family of functions than uniform convexity.

Observation 3.2. The weak strong convexity [34], essential strong convexity [29], restricted strong convexity [55], optimal strong convexity [28], semi-strong convexity [16] and other error bound conditions considered in several recent work [24,53] imply a GGC on the entire optimization domain $\mathcal{K}$ with $\theta = 1/2$ for a convex function.

Some of these conditions are also equivalent to the GGC with $\theta = 1/2$. We refer the reader to [34], [24] and [53] for more discussions of these conditions.

The third observation shows that LGC could imply faster convergence than that induced by GGC.

Observation 3.3. The LGC could induce a smaller constant c in (1.4) that is domain-size independent than that induced by the GGC on the entire optimization domain $\mathcal{K}$.

To illustrate this, we consider a function $f(x) = x^2$ if $|x| \leq 1$ and $f(x) = |x|$ if $1 < |x| \leq s$, where s specifies the size of the domain. In the ϵ-sublevel set ($\epsilon < 1$), the LGC (1.4) holds with $\theta = 1/2$ and $c = 1$. In order to make the inequality $|x| \leq cf(x)^{1/2}$ hold for all $x \in [-s, s]$, we can see that $c = \max_{|x|\leq s} \frac{|x|}{f(x)^{1/2}} = \max_{|x|\leq s} \sqrt{|x|} = \sqrt{s}$. As a result, GGC induces a larger c that depends on the domain size.

The next observation shows that Luo-Tseng's local error bound condition is closely related to the LGC with $\theta = 1/2$. To this end, we first give the definition of Luo-Tseng's local error bound condition. Let $F(\mathbf{w}) = h(\mathbf{w}) + P(\mathbf{w})$, where $h(\mathbf{w})$ is a proper closed function with an open domain containing $\mathcal{K}$ and is continuously differentiable with a locally Lipschitz continuous gradient on any compact set within $dom(h)$ and $P(\mathbf{w})$ is a proper closed convex function. Such a function $F(\mathbf{w})$ is said to satisfy Luo-Tseng's local error bound if for any $\zeta > 0$, there exists $c, \varepsilon > 0$ so that

$$\|\mathbf{w} - \mathbf{w}_*\|_2 \leq c\|\text{prox}_P(\mathbf{w} - \nabla h(\mathbf{w})) - \mathbf{w}\|_2,$$

whenever $\|\text{prox}_P(\mathbf{w} - \nabla h(\mathbf{w})) - \mathbf{w}\|_2 \leq \varepsilon$ and $F(\mathbf{w}) - F_* \leq \zeta$, where $\text{prox}_P(\mathbf{w}) = \arg\min_{\mathbf{u}\in\mathcal{K}} \frac{1}{2}\|\mathbf{u} - \mathbf{w}\|_2^2 + P(\mathbf{w})$.

Observation 3.4. If $F(\mathbf{w}) = h(\mathbf{w}) + P(\mathbf{w})$ is defined above and satisfies the Luo-Tseng's local error bound condition, it then implies that there exists a sufficiently small $\epsilon' > 0$ and $C > 0$ such that $\|\mathbf{w} - \mathbf{w}_*\|_2 \leq C(F(\mathbf{w}) - F_*)^{1/2}$ for any $\mathbf{w} \in \mathcal{B}(\mathbf{w}_*, \epsilon')$.

This observation was established in [27, Theorem 4.1]. Note that the LGC condition with $\epsilon = G\epsilon'$ and $\theta = 1/2$ also implies that $\|\mathbf{w} - \mathbf{w}_*\|_2 \leq C(F(\mathbf{w}) - F_*)^{1/2}$ for any $\mathbf{w} \in \mathcal{B}(\mathbf{w}_*, \epsilon')$. Nonetheless, Luo-Tseng's local error bound imposes some smoothness assumption on $h(\mathbf{w})$.

The last observation is that the LGC is equivalent to a Kurdyka - Łojasiewicz inequality (KL), which was proved in [8, Theorem 5].

Observation 3.5. If $F(\mathbf{w})$ satisfies a KL inequality, i.e., $\varphi'(F(\mathbf{w}) - F_*)\|\partial^0 F(\mathbf{w})\|_2 \geq 1$ for $\mathbf{w} \in \{\mathbf{x} \in \mathcal{K}, F(\mathbf{x}) - F_* < \epsilon\}$ with $\varphi(s) = cs^\theta$, then LGC (1.4) holds, and vice versa.

The above KL inequality has been established for continuous semi-algebraic and subanalytic functions [3,7,8], which cover a broad family of functions therefore justifying the generality of the LGC.

Finally, we present a key lemma that can leverage the LGC to control the distance of intermediate solutions to an ϵ-optimal solution, which is due to [50].

Lemma 3.1. *For any* $\mathbf{w} \in \mathcal{K}$ *and* $\epsilon > 0$*, we have*

$$\|\mathbf{w} - \mathbf{w}_\epsilon^\dagger\|_2 \leq \frac{dist(\mathbf{w}_\epsilon^\dagger, \mathcal{K}_*)}{\epsilon}(F(\mathbf{w}) - F(\mathbf{w}_\epsilon^\dagger)),$$

where $\mathbf{w}_\epsilon^\dagger \in \mathcal{S}_\epsilon$ *is the closest point in the* ϵ*-sublevel set to* $\mathbf{w}$ *as defined in (3.1).*

Remark: In view of LGC, we can see that $\|\mathbf{w} - \mathbf{w}_\epsilon^\dagger\|_2 \leq \frac{c}{\epsilon^{1-\theta}}(F(\mathbf{w}) - F(\mathbf{w}_\epsilon^\dagger))$ for any $\mathbf{w} \in \mathcal{K}$. Yang and Lin [50] have leveraged this relationship to improve the convergence of the standard subgradient method. In this work, we will build on this relationship to further develop novel stochastic optimization algorithms with faster convergence in high probability.

4. Accelerated Stochastic Subgradient Methods under LGC

In this section, we will present the proposed accelerated stochastic subgradient (ASSG) methods and establish their improved iteration complexity with a high probability. The key to our development is to control the distance of intermediate solutions to the ϵ-*optimal* solution by their functional residuals that are decreasing as the solutions approach the optimal set. It is this decreasing factor that help mitigate the non-vanishing variance issue in the stochastic subgradient. To formally illustrate this, we consider the following stochastic subgradient update:

$$\mathbf{w}_{\tau+1} = \Pi_{\mathcal{K}\cap\mathcal{B}(\mathbf{w}_1, D)}[\mathbf{w}_\tau - \eta\nabla f(\mathbf{w}_\tau; \xi_\tau)]. \tag{4.1}$$

Then we present a lemma regarding the update of (4.1).

Lemma 4.1. *Given* $\mathbf{w}_1 \in \mathcal{K}$*, apply* t *iterations of (4.1). For any fixed* $\mathbf{w} \in \mathcal{K} \cap \mathcal{B}(\mathbf{w}_1, D)$ *and* $\delta \in (0, 1)$*, with a probability at least* $1 - \delta$*, the following inequality holds*

$$F(\widehat{\mathbf{w}}_t) - F(\mathbf{w}) \leq \frac{\eta G^2}{2} + \frac{\|\mathbf{w}_1 - \mathbf{w}\|_2^2}{2\eta t} + \frac{4GD\sqrt{3\log(\frac{1}{\delta})}}{\sqrt{t}},$$

where $\widehat{\mathbf{w}}_t = \sum_{\tau=1}^t \mathbf{w}_t / t$.

Remark: The proof of the above lemma follows similarly as that of Lemma 10 in [19]. We note that the last term is due to the variance of the stochastic subgradients. In fact, due to the non-smoothness nature of the problem the variance

Algorithm 1 ASSG-c($\mathbf{w}_0, K, t, D_1, \epsilon_0$)

```
Input: $\mathbf{w}_0 \in \mathcal{K}$, $K$, $t$, $\epsilon_0$ and $D_1 \geq \frac{c\epsilon_0}{\epsilon^{1-\theta}}$
Set $\eta_1 = \epsilon_0/(3G^2)$
for $k = 1, \ldots, K$ do
    Let $\mathbf{w}_1^k = \mathbf{w}_{k-1}$
    for $\tau = 1, \ldots, t-1$ do
        $\mathbf{w}_{\tau+1}^k = \Pi_{\mathcal{K}\cap\mathcal{B}(\mathbf{w}_{k-1}, D_k)}[\mathbf{w}_\tau^k - \eta_k \partial f(\mathbf{w}_\tau^k; \xi_\tau^k)]$
    end for
    Let $\mathbf{w}_k = \frac{1}{t}\sum_{\tau=1}^t \mathbf{w}_\tau^k$
    Let $\eta_{k+1} = \eta_k/2$ and $D_{k+1} = D_k/2$.
end for
Output: $\mathbf{w}_K$
```

of the stochastic subgradients cannot be reduced, we therefore propose to address this issue by reducing D in light of the inequality in Lemma 3.1.

The updates in (4.1) can be also understood as approximately solving the original problem in the neighborhood of $\mathbf{w}_1$. In light of this, we will also develop a regularized variant of the proposed method.

4.1. *Accelerated Stochastic Subgradient Method: the Constrained variant (ASSG-c)*

In this subsection, we present the constrained variant of ASSG that iteratively solves the original problem approximately in an explicitly constructed local neighborhood of the recent historical solution. The detailed steps are presented in Algorithm 1. We refer to this variant as ASSG-c. The algorithm runs in stages and each stage runs t iterations of updates similar to (4.1). Thanks to Lemma 3.1, we gradually decrease the radius D_k in a stage-wise manner. The step size keeps the same during each stage and geometrically decreases between stages. We notice that ASSG-c is similar to the Epoch-GD method by Hazan and Kale [19] and the (multi-stage) AC-SA method with domain shrinkage by Chadimi and Lan [14] for stochastic strongly convex optimization, and is also similar to the restarted subgradient method (RSG) proposed by Yang and Lin [50]. However, the difference between ASSG and Epoch-GD/AC-SA lies at the initial radius D_1 and the number of iterations per-stage, which is due to difference between the strong convexity assumption and Lemma 3.1. Compared to RSG, the solutions updated along gradient direction in ASSG are projected back into a local neighborhood around $\mathbf{w}_{k-1}$, which is the key to establish the faster convergence of ASSG. The convergence of ASSG-c is presented in the theorem below.

Theorem 4.1. *Suppose Assumptions 3.1 and 3.2 hold for a target $\epsilon \ll 1$. Given $\delta \in (0,1)$, let $\tilde{\delta} = \delta/K$, $K = \lceil \log_2(\frac{\epsilon_0}{\epsilon}) \rceil$, $D_1 \geq \frac{c\epsilon_0}{\epsilon^{1-\theta}}$ and t be the smallest integer such that $t \geq \max\{9, 1728\log(1/\tilde{\delta})\}\frac{G^2 D_1^2}{\epsilon_0^2}$. Then ASSG-c guarantees that,* *with a probability* $1-\delta$, $F(\mathbf{w}_K) - F_* \leq 2\epsilon$. *As a result, the iteration complexity of ASSG-c for achieving an* 2ϵ*-optimal solution with a high probability* $1-\delta$ *is* $O(c^2G^2\lceil\log_2(\frac{\epsilon_0}{\epsilon})\rceil\log(1/\delta)/\epsilon^{2(1-\theta)})$ *provided* $D_1 = O(\frac{c\epsilon_0}{\epsilon^{(1-\theta)}})$.

Remark: It is notable that the faster local growth rate θ implies the faster global convergence, i.e., lower iteration complexity. In light of the lower bound presented in [40] under a GGC, our iteration complexity under the LGC is optimal up to at most a logarithmic factor. It is worth mentioning that unlike traditional high-probability analysis of SSG that usually requires the domain to be bounded, the convergence analysis of ASSG does not rely on such a condition. Furthermore, the iteration complexity of ASSG has a better dependence on the quality of the initial solution or the size of domain if it is bounded. In particular, if we let $\epsilon_0 = GB$ assuming $dist(\mathbf{w}_0, \mathcal{K}_*) \leq B$, though this is not necessary in practice, then the iteration complexity of ASSG has only a logarithmic dependence on the distance of the initial solution to the optimal set, while that of SSG has a quadratic dependence on this distance. The above theorem requires a target precision ϵ in order to set D_1. In Section 5, we alleviate this requirement to make the algorithm more practical. Next, we prove Theorem 4.1 regarding the convergence of ASSG-c.

Proof. Let $\mathbf{w}_{k,\epsilon}^\dagger$ denote the closest point to $\mathbf{w}_k$ in $\mathcal{S}_\epsilon$. Define $\epsilon_k = \frac{\epsilon_0}{2^k}$. Note that $D_k = \frac{D_1}{2^{k-1}} \geq \frac{c\epsilon_{k-1}}{\epsilon^{1-\theta}}$ and $\eta_k = \frac{\epsilon_{k-1}}{3G^2}$. We will show by induction that $F(\mathbf{w}_k) - F_* \leq \epsilon_k + \epsilon$ for $k = 0, 1, \ldots$ with a high probability, which leads to our conclusion when $k = K$. The inequality holds obviously for $k = 0$. Conditioned on $F(\mathbf{w}_{k-1}) - F_* \leq \epsilon_{k-1} + \epsilon$, we will show that $F(\mathbf{w}_k) - F_* \leq \epsilon_k + \epsilon$ with a high probability. By Lemma 3.1, we have

$$\|\mathbf{w}_{k-1,\epsilon}^\dagger - \mathbf{w}_{k-1}\|_2 \leq \frac{c}{\epsilon^{1-\theta}}(F(\mathbf{w}_{k-1}) - F(\mathbf{w}_{k-1,\epsilon}^\dagger)) \leq \frac{c\epsilon_{k-1}}{\epsilon^{1-\theta}} \leq D_k. \tag{4.2}$$

We apply Lemma 4.1 to the k-th stage of Algorithm 1 conditioned on randomness in previous stages. With a probability $1-\tilde{\delta}$ we have

$$F(\mathbf{w}_k) - F(\mathbf{w}_{k-1,\epsilon}^\dagger) \leq \frac{\eta_k G^2}{2} + \frac{\|\mathbf{w}_{k-1} - \mathbf{w}_{k-1,\epsilon}^\dagger\|_2^2}{2\eta_k t} + \frac{4GD_k\sqrt{3\log(1/\tilde{\delta})}}{\sqrt{t}}. \tag{4.3}$$

Combining (4.2) and (4.3), we get

$$F(\mathbf{w}_k) - F(\mathbf{w}_{k-1,\epsilon}^\dagger) \leq \frac{\eta_k G^2}{2} + \frac{D_k^2}{2\eta_k t} + \frac{4GD_k\sqrt{3\log(1/\tilde{\delta})}}{\sqrt{t}}.$$

Since $\eta_k = \frac{2\epsilon_k}{3G^2}$ and $t \geq \max\{9, 1728\log(1/\tilde{\delta})\}\frac{G^2D_1^2}{\epsilon_0^2}$, we have each term in the R.H.S of above inequality bounded by $\epsilon_k/3$. As a result,

$$F(\mathbf{w}_k) - F(\mathbf{w}_{k-1,\epsilon}^\dagger) \leq \epsilon_k,$$

which together with the fact that $F(\mathbf{w}_{k-1,\epsilon}^\dagger) - F_* \leq \epsilon$ by definition of $\mathbf{w}_{k-1,\epsilon}^\dagger$ implies

$$F(\mathbf{w}_k) - F_* \leq \epsilon + \epsilon_k.$$

Therefore by induction, with a probability at least $(1-\tilde{\delta})^K$ we have

$$F(\mathbf{w}_K) - F_* \leq \epsilon_K + \epsilon \leq 2\epsilon.$$

Since $\tilde{\delta} = \delta/K$, then $(1-\tilde{\delta})^K \geq 1-\delta$ and we complete the proof. □

Theorem 4.1 shows the high probability convergence bound for ASSG-c. We also prove the following expectational convergence bound, which is an immediate consequence of Theorem 4.1. Its proof is provided in Appendix A.

Corollary 4.1. *Suppose Assumptions 3.1 and 3.2 hold for a target $\epsilon \ll 1$. Given $\delta \in (0,1)$, let $\delta \leq \frac{\epsilon}{2GD_1+\epsilon_0}$, $K = \lceil \log_2(\frac{\epsilon_0}{\epsilon}) \rceil$, $D_1 \geq \frac{c\epsilon_0}{\epsilon^{1-\theta}}$ and t be the smallest integer such that $t \geq \max\{9, 1728\log(K/\delta)\}\frac{G^2D_1^2}{\epsilon_0^2}$. Then ASSG-c achieves that $\mathbb{E}\left[F(\mathbf{w}_K) - F_*\right] \leq 2\epsilon$ using at most $O\left(\lceil \log_2(\frac{\epsilon_0}{\epsilon}) \rceil \log\left(\frac{2GD_1+\epsilon_0}{\epsilon}\right) c^2G^2/\epsilon^{2(1-\theta)}\right)$ iterations provided $D_1 = O(\frac{c\epsilon_0}{\epsilon^{(1-\theta)}})$.*

4.2. *Accelerated Stochastic Subgradient Method: the Regularized variant (ASSG-r)*

One potential issue of ASSG-c is that the projection into the intersection of the problem domain and an Euclidean ball might increase the computational cost per-iteration depending on the problem domain $\mathcal{K}$. To address this issue, we present a regularized variant of ASSG. Before delving into the details of ASSG-r (Algorithm 2), we first present a common strategy that solves the non-strongly convex problem (1.1) by stochastic strongly convex optimization. The basic idea is from the classical deterministic *proximal point algorithm* [41] which adds a strongly convex regularizer to the original problem and solve the resulting proximal problem. In particular, we construct a new problem

$$\min_{\mathbf{w}\in\mathcal{K}} \widehat{F}(\mathbf{w}) = F(\mathbf{w}) + \frac{1}{2\beta}\|\mathbf{w}-\mathbf{w}_1\|_2^2,$$

where $\mathbf{w}_1 \in \mathcal{K}$ is called the regularization reference point. Let $\widehat{\mathbf{w}}_*$ denote the optimal solution to the above problem given $\mathbf{w}_1$. It is easy to know $\widehat{F}(\mathbf{w})$ is a $\frac{1}{\beta}$-strongly convex function on $\mathcal{K}$. There are many stochastic methods can be used to solve the above strongly convex optimization problem with an $\widetilde{O}(\beta/T)$ convergence, including stochastic subgradient, proximal stochastic subgradient [11], Epoch-GD [19], stochastic dual averaging [46], etc. We employ the stochastic subgradient method suited for strongly convex problems to solve the above problem. The update is given by

$$\mathbf{w}_{t+1} = \Pi_{\mathcal{K}}[\mathbf{w}'_{t+1}] = \arg\min_{\mathbf{w}\in\mathcal{K}} \left\|\mathbf{w}-\mathbf{w}'_{t+1}\right\|_2^2, \tag{4.4}$$

where $\mathbf{w}'_{t+1} = \mathbf{w}_t - \eta_t(\partial f(\mathbf{w}_t;\xi_t) + \frac{1}{\beta}(\mathbf{w}_t - \mathbf{w}_1))$, and $\eta_t = \frac{2\beta}{t}$ [b]. We present a lemma below to bound $\|\widehat{\mathbf{w}}_* - \mathbf{w}_t\|_2$ and $\|\mathbf{w}_t - \mathbf{w}_1\|_2$ by the above update, which will be used in the proof of convergence of ASSG-r for solving (1.1).

[b] The factor 2 in the step size is used for proving the high probability convergence.

Algorithm 2 the ASSG-r algorithm for solving (1.1)

```
Input: $\mathbf{w}_0 \in \mathcal{K}$, $K$, $t$, $\epsilon_0$ and $\beta_1 \geq \frac{2c^2\epsilon_0}{\epsilon^{2(1-\theta)}}$
for $k = 1, \ldots, K$ do
    Let $\mathbf{w}_1^k = \mathbf{w}_{k-1}$
    for $\tau = 1, \ldots, t-1$ do
        Let $\mathbf{w}'_{\tau+1} = \left(1 - \frac{2}{\tau}\right)\mathbf{w}_\tau^k + \frac{2}{\tau}\mathbf{w}_1^k - \frac{2\beta_k}{\tau}\partial f(\mathbf{w}_\tau^k; \xi_\tau^k)$
        Let $\mathbf{w}_{\tau+1}^k = \Pi_{\mathcal{K}}(\mathbf{w}'_{\tau+1})$
    end for
    Let $\mathbf{w}_k = \frac{1}{t}\sum_{\tau=1}^{t}\mathbf{w}_\tau^k$, and $\beta_{k+1} = \beta_k/2$
end for
Output: $\mathbf{w}_K$
```

Lemma 4.2. *For any $t \geq 1$, we have $\|\widehat{\mathbf{w}}_* - \mathbf{w}_t\|_2 \leq 3\beta G$ and $\|\mathbf{w}_t - \mathbf{w}_1\|_2 \leq 2\beta G$.*

Remark: The lemma implies that the regularization term implicitly imposes a constraint on the intermediate solutions to center around the regularization reference point, which achieves a similar effect as the ball constraint in Algorithm 1. We include its proof in Appendix B.

Next, we present a high probability convergence bound, whose proof can be found in Appendix C.

Lemma 4.3. *Given $\mathbf{w}_1 \in \mathcal{K}$, apply T-iterations of (4.4). For any fixed $\mathbf{w} \in \mathcal{K}$, $\delta \in (0,1)$, and $T \geq 3$, with a probability at least $1-\delta$, following inequality holds*

$$F(\widehat{\mathbf{w}}_T) - F(\mathbf{w}) \leq \frac{\|\mathbf{w} - \mathbf{w}_1\|_2^2}{2\beta} + \frac{34\beta G^2\left(1 + \log T + \log(4\log T/\delta)\right)}{T},$$

where $\widehat{\mathbf{w}}_t = \sum_{\tau=1}^{t}\mathbf{w}_t/t$.

Remark: From the above result, we can see that one can set β to be a large value to ensure convergence. In particular, by assuming that $dist(\mathbf{w}_1, \mathcal{K}_*) \leq B$, we can set $\beta = \frac{B^2}{\epsilon}$ and $T \geq \frac{68G^2B^2(1+\log(4\log T/\delta)+\log T)}{\epsilon^2}$ so as to obtain $F(\widehat{\mathbf{w}}_T) - F_* \leq \epsilon$ with a high probability $1-\delta$, which yields the same order of iteration complexity to SSG for directly solving (1.1).

Recall that the main iteration of the proximal point algorithm [41] is

$$\mathbf{w}_k \approx \arg\min_{\mathbf{w}\in\mathcal{K}} F(\mathbf{w}) + \frac{1}{2\beta_k}\|\mathbf{w} - \mathbf{w}_{k-1}\|_2^2, \tag{4.5}$$

where $\mathbf{w}_k$ approximately solves the minimization problem above with β_k changing with k. With the same idea, our regularized variant of ASSG generates $\mathbf{w}_k$ from stage k by solving the minimization problem (4.5) approximately using (4.4). The detailed steps are presented in Algorithm 2, which starts from a relatively large value of the parameter $\beta = \beta_1$ and gradually decreases β by a constant factor after running a number of t iterations (4.4) using the solution from the previous stage as the new regularization reference point. Despite of its similarity to the proximal point

Algorithm 3 ASSG-s($\mathbf{w}_0, K, t, \epsilon_0$)

```
Input: $\mathbf{w}_0 \in \mathcal{K}$, $K$, $t$, $\epsilon_0$
Set $\eta_1 = \epsilon_0/(3G^2)$
for $k = 1, \ldots, K$ do
    Let $\mathbf{w}_1^k = \mathbf{w}_{k-1}$
    for $\tau = 1, \ldots, t-1$ do
        $\mathbf{w}_{\tau+1}^k = \Pi_{\mathcal{K}}[\mathbf{w}_\tau^k - \eta_k \partial f(\mathbf{w}_\tau^k; \xi_\tau^k)]$
    end for
    Let $\mathbf{w}_k = \frac{1}{t}\sum_{\tau=1}^t \mathbf{w}_\tau^k$ and $\eta_{k+1} = \eta_k/2$.
end for
Output: $\mathbf{w}_K$
```

algorithm, ASSG-r incorporates the LGC into the choices of β_k and the number of iterations per-stage and obtains new iteration complexity described below.

Theorem 4.2. *Suppose Assumptions 3.1 and 3.2 hold for a target $\epsilon \ll 1$. Given $\delta \in (0, 1/e)$, let $\tilde{\delta} = \delta/K$, $K = \lceil \log_2(\frac{\epsilon_0}{\epsilon}) \rceil$, $\beta_1 \geq \frac{2c^2\epsilon_0}{\epsilon^{2(1-\theta)}}$ and t be the smallest integer such that $t \geq \max\{3, \frac{136\beta_1 G^2(1+\log(4\log t/\tilde{\delta})+\log t)}{\epsilon_0}\}$. Then ASSG-r guarantees that, with a probability $1-\delta$, $F(\mathbf{w}_K) - F_* \leq 2\epsilon$. As a result, the iteration complexity of ASSG-r for achieving an 2ϵ-optimal solution with a high probability $1-\delta$ is $O(c^2G^2\lceil \log_2(\frac{\epsilon_0}{\epsilon}) \rceil \log(1/\delta)/\epsilon^{2(1-\theta)})$ provided $\beta_1 = O(\frac{2c^2\epsilon_0}{\epsilon^{2(1-\theta)}})$.*

With Lemma 4.3, the proof of Theorem 4.2 is similar to the proof of Theorem 4.1. For completeness, we include it in Appendix D.

4.3. *A Simple Variant of ASSG under GGC*

As a byproduct of similar analysis, we can show that a simpler variant of ASSG without using shrinking domain constraint or increasing regularization can have an improved complexity in expectation under GGC for $\theta \in (0, 1/2]$. When the problems satisfy GGC with $\theta \in (1/2, 1]$ and $F(\mathbf{w}) - F_*$ is bounded over $\mathcal{K}$, one can always show that the problem satisfies a GGC with $\theta = 1/2$ [48]. The details of updates are presented in Algorithm 3, which is referred to ASSG-s. The algorithm is almost the same to Algorithm 1 except that the projection is simply done onto the original domain $\mathcal{K}$ without intersecting with a bounded ball at each epoch. At each epoch, the update is exactly the same to the stochastic subgradient update

$$\mathbf{w}_{\tau+1} = \Pi_{\mathcal{K}}[\mathbf{w}_\tau - \eta \partial f(\mathbf{w}_\tau; \xi_\tau)]. \tag{4.6}$$

To establish the convergence result, we first need the following lemma, whose proof is included in Appendix E.

Lemma 4.4. *Given $\mathbf{w}_1 \in \mathcal{K}$, apply t iterations of (4.6). For any fixed $\mathbf{w} \in \mathcal{K}$, the* *following inequality holds*

$$\mathrm{E}[F(\widehat{\mathbf{w}}_t) - F(\mathbf{w})] \leq \frac{\eta G^2}{2} + \frac{\|\mathbf{w}_1 - \mathbf{w}\|_2^2}{2\eta t},$$

where $\widehat{\mathbf{w}}_t = \sum_{\tau=1}^{t} \mathbf{w}_t / t$.

We then give the convergence result of ASSG-s in the following theorem.

Theorem 4.3. *Suppose Assumption 3.1 holds and* $F(\mathbf{w})$ *obeys GGC (1.4) with* $\theta \in (0, 1/2]$. *Given* $\epsilon > 0$, $K = \lceil \log_2(\frac{\epsilon_0}{\epsilon}) \rceil$ *and* t *be the smallest integer such that* $t \geq \frac{18c^2G^2}{\epsilon^{2(1-\theta)}}$. *Then ASSG-s guarantees that* $\mathrm{E}[F(\mathbf{w}_K) - F_*] \leq \epsilon$. *As a result, the iteration complexity of ASSG-s for achieving an* ϵ*-optimal solution is* $O(c^2G^2\lceil \log_2(\frac{\epsilon_0}{\epsilon}) \rceil / \epsilon^{2(1-\theta)})$ *in expectation.*

Proof. Let define $\epsilon_k = \frac{\epsilon_0}{2^k}$. Note that $\eta_k = \frac{\epsilon_{k-1}}{3G^2}$. We will show by induction that $\mathrm{E}[F(\mathbf{w}_k) - F_*] \leq \epsilon_k$ for $k = 0, 1, \ldots,$ which leads to our conclusion when $k = K$. The inequality holds obviously for $k = 0$. Conditioned on $\mathrm{E}[F(\mathbf{w}_{k-1}) - F_*] \leq \epsilon_{k-1}$, we will show that $\mathrm{E}[F(\mathbf{w}_k) - F_*] \leq \epsilon_k$. By GGC, we have for any $\mathbf{w}_{k-1} \in \mathcal{K}$,

$$\|\mathbf{w}_{k-1} - \mathbf{w}_{k-1}^*\|_2 \leq c(F(\mathbf{w}_{k-1}) - F_*)^{\theta}.$$

Then by the condition $\mathrm{E}[F(\mathbf{w}_{k-1}) - F_*] \leq \epsilon_{k-1}$, we have

$$\mathrm{E}[\|\mathbf{w}_{k-1} - \mathbf{w}_{k-1}^*\|_2] \leq c\epsilon_{k-1}^{\theta}. \tag{4.7}$$

We apply Lemma 4.4 to the k-th stage of Algorithm 3 conditioned on randomness in previous stages. For any $\mathbf{w}_* \in \mathcal{K}_*$ we have

$$\mathrm{E}[F(\mathbf{w}_k) - F(\mathbf{w}_*)] \leq \frac{\eta_k G^2}{2} + \frac{\mathrm{E}[\|\mathbf{w}_{k-1} - \mathbf{w}_*\|_2^2]}{2\eta_k t}. \tag{4.8}$$

By using GGC (1.4) with $\theta \in (0, 1/2]$ we have

$$\begin{aligned}
\mathrm{E}[F(\mathbf{w}_k) - F(\mathbf{w}_*)] \leq & \frac{\eta_k G^2}{2} + \frac{c^2\mathrm{E}[F(\mathbf{w}_{k-1}) - F(\mathbf{w}_*)]^{2\theta}}{2\eta_k t} \\
\leq & \frac{\eta_k G^2}{2} + \frac{c^2\{\mathrm{E}[F(\mathbf{w}_{k-1}) - F(\mathbf{w}_*)]\}^{2\theta}}{2\eta_k t} \\
\leq & \frac{\eta_k G^2}{2} + \frac{c^2\epsilon^{2\theta}}{2\eta_k t} \leq \frac{\epsilon_k}{3} + \frac{\epsilon_k}{3} \leq \epsilon_k,
\end{aligned}$$

where the second inequality uses the concavity of $\mathrm{E}[X^{\alpha}] \leq \{\mathrm{E}[X]\}^{\alpha}$ whith $0 < \alpha \leq 1$; the fourth inequality using the fact that $\eta_k = \frac{\epsilon_k}{3G^2}$ and $t \geq \frac{18c^2G^2}{\epsilon^{2(1-\theta)}}$ with $\epsilon \leq \epsilon_k$. Therefore by induction, we have

$$\mathrm{E}[F(\mathbf{w}_K) - F_*] \leq \epsilon_K \leq \epsilon. \qquad \square$$

Algorithm 4 ASSG with Restarting: RASSG

```
**Input**: $\mathbf{w}^{(0)}$, $K$, $D_1^{(1)}$, $t_1$, $\epsilon_0$ and $\omega \in (0, 1]$
Set $\epsilon_0^{(1)} = \epsilon_0$, $\eta_1 = \epsilon_0/(3G^2)$
**for** $s = 1, 2, \ldots, S$ **do**
    Let $\mathbf{w}^{(s)}$ =ASSG-c($\mathbf{w}^{(s-1)}, K, t_s, D_1^{(s)}, \epsilon_0^{(s)}$)
    Let $t_{s+1} = t_s 2^{2(1-\theta)}$, $D_1^{(s+1)} = D_1^{(s)} 2^{1-\theta}$, and $\epsilon_0^{(s+1)} = \omega\epsilon_0^{(s)}$
**end for**
**Output**: $\mathbf{w}^{(S)}$
```

5. Practical Variants of ASSG

Readers may have noticed that the presented algorithms require appropriately setting up the initial values of D_1 or β_1 or t that depend on potentially unknown c and unknown θ. As we show later, the value of θ is exhibited for many problems. However, the parameter c is usually difficult to estimate, which leads to a challenge to set the value of t. Overestimate of t leads to waste of iterations while underestimate of t leads to a less accurate solution so that it may not reach the target level of accuracy. This section is devoted to more practical variants of ASSG that can be implemented without knowing parameter c or θ. For ease of presentation, we focus on the constrained variant of ASSG. Similar extensions can be made for the regularized variant ASSG-r and the simple variant ASSG-s, which are omitted here. In the following subsections, we divide the problem into two cases: (1) unknown c; (2) unknown θ.

5.1. *ASSG with unknown c*

When c is unknown, we present the details of a restarting variant of ASSG in Algorithm 4, to which we refer as RASSG. When discussing the restarting variants of ASSG-c, ASSG-r and ASSG-s, we refer to them as RSSG-c, RSSG-r, and RSSG-s, respectively, for clarity. The key idea is to use an increasing sequence of t and another level of restarting for ASSG. The convergence analysis for RASSG without knowing c is presented in the following theorem.

Theorem 5.1 (RASSG with unknown c). *Let $\epsilon \leq \epsilon_0/4$, $\omega = 1$, and $K = \lceil \log_2(\frac{\epsilon_0}{\epsilon}) \rceil$ in Algorithm 4. Suppose $D_1^{(1)}$ is sufficiently large so that there exists $\hat{\epsilon}_1 \in [\epsilon, \epsilon_0/2]$, with which $F(\cdot)$ satisfies a LGC (3.2) on $\mathcal{S}_{\hat{\epsilon}_1}$ with $\theta \in (0, 1)$ and the constant c, and $D_1^{(1)} = \frac{c\epsilon_0}{\hat{\epsilon}_1^{1-\theta}}$. Let $\hat{\delta} = \frac{\delta}{K(K+1)}$, and $t_1 = \max\{9, 1728\log(1/\hat{\delta})\}\left(GD_1^{(1)}/\epsilon_0\right)^2$. Then with at most $S = \lceil \log_2(\hat{\epsilon}_1/\epsilon) \rceil + 1$ calls of ASSG-c, Algorithm 4 finds a solution $\mathbf{w}^{(S)}$ such that $F(\mathbf{w}^{(S)}) - F_* \leq 2\epsilon$ with probability $1 - \delta$. The total number of iterations of RASSG for obtaining 2ϵ-optimal solution is upper bounded by $T_S = O(\lceil \log_2(\frac{\epsilon_0}{\epsilon}) \rceil \log(1/\delta) c^2/\epsilon^{2(1-\theta)})$.*

Remark: The above theorem requires a slightly stringent LGC condition on $\mathcal{S}_{\hat{\epsilon}_1}$ that is induced by the initial value of D_1. If the problem satisfies the LGC with $\theta = 1$, we can give a slightly smaller value for θ in order to run Algorithm 4. If the target precision ϵ is not specified, we can give it a sufficiently small value ϵ' (e.g., the machine precision) that only affects K marginally. The corresponding iteration complexity for achieving an ϵ-optimal solution is given by $O(\lceil \log_2(\frac{\epsilon_0}{\epsilon'})\rceil \log(1/\delta)/\epsilon^{2(1-\theta)})$. The parameter $\omega \in (0, 1]$ is introduced to increase the practical performance of RASSG, which accounts for decrease of the objective gap of the initial solutions for each call of ASSG-c.

Proof. Since $K = \lceil \log_2(\frac{\epsilon_0}{\epsilon})\rceil \geq \lceil \log_2(\frac{\epsilon_0}{\hat{\epsilon}_1})\rceil$, $D_1^{(1)} = \frac{c\epsilon_0}{\hat{\epsilon}_1^{1-\theta}}$, and $t_1 = \max\{9, 1728 \log(1/\hat{\delta})\} \left(\frac{GD_1^{(1)}}{\epsilon_0}\right)^2$, following the proof of Theorem 4.1, we can show that with a probability $1 - \frac{\delta}{K+1}$,

$$F(\mathbf{w}^{(1)}) - F_* \leq 2\hat{\epsilon}_1. \tag{5.1}$$

By running ASSG-c starting from $\mathbf{w}^{(1)}$ which satisfies (5.1) with $K = \lceil \log_2(\frac{\epsilon_0}{\epsilon})\rceil \geq \lceil \log_2(\frac{2\hat{\epsilon}_1}{\hat{\epsilon}_1/2})\rceil$, $D_1^{(2)} = \frac{c\epsilon_0}{(\hat{\epsilon}_1/2)^{1-\theta}} \geq \frac{c2\hat{\epsilon}_1}{(\hat{\epsilon}_1/2)^{1-\theta}}$, and $t_2 = \max\{9, 1728 \log(1/\hat{\delta})\} \left(GD_1^{(2)}/\epsilon_0\right)^2$, Theorem 4.1 ensures that

$$F(\mathbf{w}^{(2)}) - F_* \leq \hat{\epsilon}_1$$

with a probability at least $(1 - \delta/(K+1))^2$. By continuing the process, with $S = \lceil \log_2(\hat{\epsilon}_1/\epsilon)\rceil + 1$ we can prove that with a probability at least $(1 - \delta/(K+1))^S \geq 1 - \delta\frac{S}{K+1} \geq 1 - \delta$,

$$F(\mathbf{w}^{(S)}) - F_* \leq 2\hat{\epsilon}_1/2^{S-1} \leq 2\epsilon.$$

The total number of iterations for the S calls of ASSG-c is bounded by

$$T_S = K\sum_{s=1}^{S} T_s = K\sum_{s=1}^{S} t_1 2^{2(s-1)(1-\theta)} = Kt_1 2^{2(S-1)(1-\theta)} \sum_{s=1}^{S} \left(1/2^{2(1-\theta)}\right)^{S-s}$$

$$\leq \frac{Kt_1 2^{2(S-1)(1-\theta)}}{1 - 1/2^{2(1-\theta)}} \leq O\left(Kt_1 \left(\frac{\hat{\epsilon}_1}{\epsilon}\right)^{2(1-\theta)}\right) \leq \widetilde{O}(\log(1/\delta)/\epsilon^{2(1-\theta)}). \qquad \square$$

As a corollary of the above theorem, we present a result of RASSG for problems satisfying GGC with $\theta = 1/2$ but without knowing the value of c (or satisfying strong convexity but without knowing the strong convexity parameter), which is of interest to a broad audience who are familiar with stochastic strongly convex optimization. It has been shown many machine learning problems satisfy GGC with $\theta = 1/2$ (see examples presented in Section 8). Almost all existing algorithms and analysis for stochastic strongly convex optimization or problems satisfying GGC with $\theta = 1/2$ require knowing the value of strong convexity parameter in order to run the algorithms [19,39]. The result is presented below.

Corollary 5.1. *Suppose $F(\cdot)$ satisfies a GGC on $\mathcal{K}$ with $\theta = 1/2$ and some unknown constant $c > 0$. Let $\epsilon \leq \epsilon_0/4$, $\omega = 1$, and $K = \lceil \log_2(\frac{\epsilon_0}{\epsilon}) \rceil$ in Algorithm 4. Suppose $D_1^{(1)}$ is sufficiently large so that there exists $\hat{\epsilon}_1 \in [\epsilon, \epsilon_0/2]$ such that $D_1^{(1)} = \frac{c\epsilon_0}{\sqrt{\hat{\epsilon}_1}}$. Let $\hat{\delta} = \frac{\delta}{K(K+1)}$, and $t_1 = \max\{9, 1728\log(1/\hat{\delta})\}\left(GD_1^{(1)}/\epsilon_0\right)^2$. Then with at most $S = \lceil \log_2(\hat{\epsilon}_1/\epsilon) \rceil + 1$ calls of ASSG-c, Algorithm 4 finds a solution $\mathbf{w}^{(S)}$ such that $F(\mathbf{w}^{(S)}) - F_* \leq 2\epsilon$ with probability $1 - \delta$. The total number of iterations of RASSG for obtaining 2ϵ-optimal solution is upper bounded by $T_S = \widetilde{O}(\log(1/\delta)c^2/\epsilon)$.*

Remark: It is notable that when the objective function is λ-strongly convex, then $c^2 = 1/\lambda$ and the above complexity $\widetilde{O}(\log(1/\delta)/\lambda\epsilon)$ is optimal up to a logarithmic factor. The advantage of RASSG over previous stochastic algorithms for strongly convex optimization is that RASSG does not need to know the value of strong convexity parameter.

5.2. *ASSG with unknown θ*

When θ is unknown, we can set $\theta = 0$. Then the problem will satisfy the LGC (3.2) with $\theta = 0$ and $c = B_\varepsilon$ with any $\varepsilon \geq \epsilon$, where $B_\varepsilon = \max_{\mathbf{w}\in\mathcal{L}_\varepsilon} \min_{\mathbf{v}\in\mathcal{K}_*} \|\mathbf{w} - \mathbf{v}\|_2$ is the maximum distance between the points in the ε-level set $\mathcal{L}_\varepsilon$ and the optimal set $\mathcal{K}_*$. The following theorem states the convergence result.

Theorem 5.2 (RASSG with unknown θ). *Let $\theta = 0$, $\epsilon \leq \epsilon_0/4$, $\omega = 1$, and $K = \lceil \log_2(\frac{\epsilon_0}{\epsilon}) \rceil$ in Algorithm 4. Assume $D_1^{(1)}$ is sufficiently large so that there exists $\hat{\epsilon}_1 \in [\epsilon, \epsilon_0/2]$ rendering that $D_1^{(1)} = \frac{B_{\hat{\epsilon}_1}\epsilon_0}{\hat{\epsilon}_1}$. Let $\hat{\delta} = \frac{\delta}{K(K+1)}$, and $t_1 = \max\{9, 1728\log(1/\hat{\delta})\}\left(GD_1^{(1)}/\epsilon_0\right)^2$. Then with at most $S = \lceil \log_2(\hat{\epsilon}_1/\epsilon) \rceil + 1$ calls of ASSG-c, Algorithm 4 finds a solution $\mathbf{w}^{(S)}$ such that $F(\mathbf{w}^{(S)}) - F_* \leq 2\epsilon$. The total number of iterations of RASSG for obtaining 2ϵ-optimal solution is upper bounded by $T_S = O(\lceil \log_2(\frac{\epsilon_0}{\epsilon}) \rceil \log(1/\delta)\frac{G^2 B_{\hat{\epsilon}_1}^2}{\epsilon^2})$.*

Remark: The Lemma Appendix F.1 shows that $\frac{B_\epsilon}{\epsilon}$ is a monotonically decreasing function in terms of ϵ, which guarantees the existence of $\hat{\epsilon}_1$ given a sufficiently large $D_1^{(1)}$. The iteration complexity of RASSG could be still better with a smaller factor $B_{\hat{\epsilon}_1}$ than the B in the iteration complexity of SSG (see (1.3)), where B is the domain size or the distance of initial solution to the optimal set.

Proof. The proof is similar to the proof of Theorem 5.1, and we reprove it for completeness. It is easy to show that $t_1 \geq \frac{136\beta_1^{(1)}G^2(1+\log(4\log t_1/\hat{\delta})+\log t_1)}{\epsilon_0}$. Following the proof of Theorem 4.2, we then can show that with a probability $1 - \frac{\delta}{S}$,

$$F(\mathbf{w}^{(1)}) - F_* \leq 2\hat{\epsilon}_1 \tag{5.2}$$

with $K = \lceil \log_2(\frac{\epsilon_0}{\epsilon}) \rceil \geq \lceil \log_2(\frac{\epsilon_0}{\hat{\epsilon}_1}) \rceil$ and $\beta_1^{(1)} = \frac{2c^2\epsilon_0}{\hat{\epsilon}_1^{2(1-\theta)}}$. By running ASSG-r starting from $\mathbf{w}^{(1)}$ which satisfies (5.2) with $K = \lceil \log_2(\frac{\epsilon_0}{\epsilon}) \rceil \geq \lceil \log_2(\frac{2\hat{\epsilon}_1}{\hat{\epsilon}_1/2}) \rceil$, $t_2 = t_1 2^{2(1-\theta)} \geq \frac{136\beta_1^{(2)} G^2 (1+\log(4\log t_2/\hat{\delta})+\log t_2)}{\epsilon_0}$ and $\beta_1^{(2)} = \frac{2c^2\epsilon_0}{(\hat{\epsilon}_1/2)^{2(1-\theta)}} \geq \frac{2c^2\hat{\epsilon}_1/2}{(\hat{\epsilon}_1/2)^{2(1-\theta)}}$, Theorem 4.2 ensures that

$$F(\mathbf{w}^{(2)}) - F_* \leq \hat{\epsilon}_1$$

with a probality at least $(1-\delta/S)^2$. By continuing the process, with $S = \lceil \log_2(\hat{\epsilon}_1/\epsilon) \rceil + 1$, we can prove that with a probality at least $(1-\delta/S)^S \geq 1-\delta$

$$F(\mathbf{w}^{(S)}) - F_* \leq 2\hat{\epsilon}_1/2^{S-1} \leq 2\epsilon$$

The total number of iterations for the S calls of ASSG-c is bounded by

$$T_S = K\sum_{s=1}^{S} T_s = K\sum_{s=1}^{S} t_1 2^{2(s-1)(1-\theta)} = Kt_1 2^{2(S-1)(1-\theta)} \sum_{s=1}^{S} \left(1/2^{2(1-\theta)}\right)^{S-s}$$

$$\leq \frac{Kt_1 2^{2(S-1)(1-\theta)}}{1-1/2^{2(1-\theta)}} \leq O\left(Kt_1 \left(\frac{\hat{\epsilon}_1}{\epsilon}\right)^{2(1-\theta)}\right) \leq \widetilde{O}(\log(1/\delta)/\epsilon^{2(1-\theta)}) \qquad \square$$

Finally, we make several remarks about the Algorithm 4: (1) if $\theta = 1$, in order to obtain an increasing sequence of t_s, θ can be set to a little smaller value than 1 (for example, 0.95); (2) if $D_1^{(1)}$ in RASSG-c and $\beta_1^{(1)}$ in RASSG-r are determined, the starting number of iterations t_1 can be automatically set since $t_1 \propto D_1^{(1)}$ in RASSG-c and $t_1 \propto \beta_1^{(1)}$ in RASSG-r; (3) after the first call of ASSG, one can re-calibrate the ϵ_0 in the implementation to improve the performance or equivalently tune ω in practice; (4) the tradeoff is that the stopping criterion for RASSG is not as automatic as ASSG.

6. Proximal ASSG for Non-smooth Composite Optimization

To obtain solutions with certain structures, many machine learning problems add a regularizer to the objective function (e.g., adding ℓ_1 regularizer for sparsity). When the regularizers are non-smooth but have closed form of proximal mapping, some proximal algorithms can be employed to solve the regularized problems. As an extension of ASSG, in this section, we will present a proximal variant of ASSG for solving the following non-smooth composite optimization problem:

$$\min_{\mathbf{w}\in\mathbb{R}^d} F(\mathbf{w}) \triangleq \underbrace{\mathrm{E}_\xi[f(\mathbf{w};\xi)]}_{f(\mathbf{w})} + R(\mathbf{w}), \tag{6.1}$$

where both $f(\mathbf{w})$ and $R(\mathbf{w})$ are non-smooth convex functions. The above problem commonly appears in machine learning, which is also known as regularized risk minimization. We assume that the function $R(\mathbf{w})$ is simple enough such that the proximal mapping given below is easy to compute

$$\mathrm{Prox}_{\Omega}^{\eta,R}[\mathbf{w}] = \arg\min_{\mathbf{u}\in\Omega} \frac{1}{2}\|\mathbf{u}-\mathbf{w}\|_2^2 + \eta R(\mathbf{u}),$$

where $\Omega \subseteq \mathbb{R}^d$ is a bounded ball. An example of $R(\mathbf{w})$ is the ℓ_1-norm $R(\mathbf{w}) = \lambda\|\mathbf{w}\|_1$. We also make the following assumption throughout this section.

Assumption 6.1. *For a stochastic optimization problem (6.1), we assume*

(1) there exist $\mathbf{w}_0 \in \mathbb{R}^d$ and $\epsilon_0 \geq 0$ such that $F(\mathbf{w}_0) - F_* \leq \epsilon_0$;
(2) There exist two constants G and ρ such that $\|\partial f(\mathbf{w};\xi)\|_2 \leq G$ and $\|\partial R(\mathbf{w};\xi)\|_2 \leq \rho$.

Assumption 6.1 is quite similar as Assumption 3.1 except for an additional assumption of $\|\partial R(\mathbf{w};\xi)\|_2 \leq \rho$.

We present the detail steps of proximal ASSG (ProxASSG) in Algorithm 5, which is similar to Algorithm 1 except that Step 5 is replaced by a proximal mapping:

$$\mathbf{w}_{\tau+1}^k = \mathrm{Prox}_{\Omega_k}^{\eta_k,R}\left[\mathbf{w}_\tau^k - \eta_k \partial f(\mathbf{w}_\tau^k;\xi_\tau^k)\right],$$

where Ω_k is a ball centered at $\mathbf{w}_{k-1}$ with a radius D_k. The convergence result is stated in the following theorem:

Theorem 6.1. *Suppose Assumptions 6.1 and 3.2 hold for a target $\epsilon \ll 1$. Given $\delta \in (0,1)$, let $\tilde{\delta} = \delta/K$, $K = \lceil \log_2(\frac{\epsilon_0}{\epsilon}) \rceil$, $D_1 \geq \frac{c\epsilon_0}{\epsilon^{1-\theta}}$ and t be the smallest integer such that $t \geq \max\left\{\max(16, 3072\log(1/\tilde{\delta}))\frac{G^2D_1^2}{\epsilon_0^2}, \frac{8\rho D_1}{\epsilon_0}\right\}$. Then ProxASSG guarantees that, with a probability $1-\delta$,*

$$F(\mathbf{w}_K) - F_* \leq 2\epsilon.$$

As a result, the iteration complexity of ProxASSG for achieving an 2ϵ-optimal solution with a high probability $1-\delta$ is $\widetilde{O}(\log(1/\delta)/\epsilon^{2(1-\theta)})$ provided $D_1 = O(\frac{c\epsilon_0}{\epsilon^{(1-\theta)}})$.

To prove Theorem 6.1, we need the following lemma for each stage of ProxASSG.

Lemma 6.1. *Let D be the upper bound of $\|\mathbf{w}_1 - \mathbf{w}_{1,\epsilon}^\dagger\|_2$. Apply t-iterations of following steps:*

$$\mathbf{w}_{\tau+1} = \arg\min_{\mathbf{w}\in\mathcal{B}(\mathbf{w}_1,D)} \frac{1}{2}\|\mathbf{w}-\mathbf{w}_\tau\|_2^2 + \eta\partial f(\mathbf{w}_\tau;\xi_\tau)^\top \mathbf{w} + \eta R(\mathbf{w}).$$

Given $\mathbf{w}_1 \in \mathbb{R}^d$, for any $\delta \in (0,1)$, with a probability at least $1-\delta$,

$$F(\widehat{\mathbf{w}}_t) - F(\mathbf{w}_{1,\epsilon}^\dagger) \leq \frac{\eta G^2}{2} + \frac{\|\mathbf{w}_1 - \mathbf{w}_{1,\epsilon}^\dagger\|_2^2}{2\eta t} + \frac{4GD\sqrt{3\log(1/\delta)}}{\sqrt{t}} + \frac{\rho D}{t},$$

where $\widehat{\mathbf{w}}_t = \sum_{\tau=1}^t \mathbf{w}_t/t$.

Algorithm 5 the ProxASSG algorithm for solving (6.1)

```
Input: the number of stages $K$, the number of iterations $t$ per stage, and the initial solution $\mathbf{w}_0$, $\eta_1 = \epsilon_0/(4G^2)$ and $D_1 \geq \frac{c\epsilon_0}{\epsilon^{1-\theta}}$
for $k = 1, \ldots, K$ do
    Let $\mathbf{w}_1^k = \mathbf{w}_{k-1}$, $\Omega_k = \mathcal{B}(\mathbf{w}_{k-1}, D_k)$
    for $\tau = 1, \ldots, t$ do
        Update $\mathbf{w}_{\tau+1}^k = \text{Prox}_{\Omega_k}^{\eta_k, R}\left[\mathbf{w}_\tau^k - \eta_k \partial f(\mathbf{w}_\tau^k; \xi_\tau^k)\right]$
    end for
    Let $\mathbf{w}_k = \frac{1}{t}\sum_{\tau=1}^t \mathbf{w}_\tau^k$
    Let $\eta_{k+1} = \eta_k/2$ and $D_{k+1} = D_k/2$.
end for
Output: $\mathbf{w}_K$
```

The proof of Lemma 6.1 is deferred to Appendix G. With the above lemma, the proof of Theorem 6.1 is similar to that of Theorem 4.1. We include the details in Appendix H.

Before ending this section, we note that the presented ProxASSG algorithm in Algorithm 5 is based on the constrained version of ASSG. One can also develop a proximal variant based on the regularized version of ASSG. We include the details in Appendix I. However, the convergence guarantee of proximal ASSG based on the regularized version is slightly worse than that based on the constrained version by a constant factor depending on G and ρ.

7. Complexity of ASSG for Ensuing the Gradient is Small

Recently, there has been an increasing interest in the complexity of stochastic algorithms for finding a solution for a convex optimization problem with a small gradient [1,13]. However, these studies assume the smoothness of the objective function. The non-smoothness of the objective function make it more challenging to design stochastic algorithms and characterize their complexity of making the gradient small.

The first challenge is how to quantify the convergence in terms of gradient for a non-smooth problem. A traditional measure is using the distance from 0 to the subgradient (a set) of the objective function at a solution $\mathbf{x} \in \mathcal{K}$, i.e., $\text{dist}(0, \partial(f(\mathbf{x}) + 1_{\mathcal{K}}(\mathbf{x}))$, where $1_{\mathcal{K}}$ is the indicator function of the domain $\mathcal{K}$. However, for a non-smooth function finding an ϵ-level stationary point (i.e., $\text{dist}(0, \partial(f(\mathbf{x}) + 1_{\mathcal{K}}(\mathbf{x})) \leq \epsilon$) is difficult. For example, considering the simple function $f(x) = |x|$, as long as $x \neq 0$ the traditional measure $\text{dist}(0, \partial(f(\mathbf{x}) + 1_{\mathcal{K}}(\mathbf{x})) = 1$ is never 0. To address this challenge, previous studies on non-smooth optimization have used a new convergence measure based on the Moreau envelop of the objective function. A Moreau envelope of $F(\mathbf{w})$ associated with a positive constant $\lambda > 0$ is defined as:

$$F_\lambda(\mathbf{w}) = \min_{\mathbf{v}\in\mathcal{K}} \left\{ F(\mathbf{v}) + \frac{\lambda}{2}\|\mathbf{v}-\mathbf{w}\|^2 \right\}, \tag{7.1}$$

and the associated proximal mapping is defined as

$$\widetilde{\mathbf{w}} = \mathrm{Prox}_{F/\lambda}(\mathbf{w}) := \arg\min_{\mathbf{v}\in\mathcal{K}} \left\{ F(\mathbf{v}) + \frac{\lambda}{2}\|\mathbf{v}-\mathbf{w}\|^2 \right\}. \tag{7.2}$$

It is easy to show that $F_\lambda(\cdot)$ is a smooth function whose gradient is λ-Lipchitz continuous [5] and $\widetilde{\mathbf{w}}$ satisfies [9]:

$$\begin{aligned} F(\widetilde{\mathbf{w}}) &\leq F(\mathbf{w}), \\ \nabla F_\lambda(\mathbf{w}) &= \lambda(\mathbf{w}-\widetilde{\mathbf{w}}) \\ \mathrm{dist}(0; \partial F(\widetilde{\mathbf{w}})) &\leq \|\nabla F_\lambda(\mathbf{w})\|. \end{aligned}$$

It means that if $\|\nabla F_\lambda(\mathbf{w})\| \leq \epsilon$ then $\mathbf{w}$ is close to some point $\widetilde{\mathbf{w}}$ that is an ϵ-stationary solution for the problem (1.1). This gives a new convergence measure in terms of gradient for a non-smooth function. We call a solution $\mathbf{w}$ an ϵ-nearly stationary point if the following inequality holds for some constant $\lambda > 0$:

$$\|\nabla F_\lambda(\mathbf{w})\| \leq \epsilon. \tag{7.3}$$

It is also notable that when F is L-smooth [c] and the constraint domain is the whole space $\mathcal{K} = \mathbb{R}^d$, then an ϵ-nearly stationary point $\mathbf{w}$ also implies that it is $O(\epsilon)$-stationary in the traditional sense, i.e., $\|\nabla F(\mathbf{w})\| \leq O(\epsilon)$. This can be easily seen from $\|\nabla F(\mathbf{w})\| \leq \|\nabla F(\widetilde{\mathbf{w}})\| + \|\nabla F(\mathbf{w}) - \nabla F(\widetilde{\mathbf{w}})\| \leq \|\nabla F_\lambda(\mathbf{w})\| + L\|\mathbf{w}-\widetilde{\mathbf{w}}\| = (1+\frac{L}{\lambda})\|\nabla F_\lambda(\mathbf{w})\| \leq (1+L/\lambda)\epsilon$.

Next, we give a simple lemma that will be useful for our analysis later.

Lemma 7.1. *For any* $\mathbf{w}\in\mathcal{K}$, *it holds*

$$\|\nabla F_\lambda(\mathbf{w})\|^2 \leq 2\lambda(F(\mathbf{w}) - F(\mathbf{w}_*)), \tag{7.4}$$

where $\mathbf{w}_* \in \mathcal{K}_*$.

Proof. We first show that $\arg\min_{\mathbf{w}\in\mathcal{K}} F_\lambda(\mathbf{w}) = \arg\min_{\mathbf{w}\in\mathcal{K}} F(\mathbf{w})$. Let us consider any $\widetilde{\mathbf{w}}_* \in \widetilde{\mathcal{K}}_*$. Then for any $\mathbf{v},\mathbf{w}\in\mathcal{K}$, $F_\lambda(\widetilde{\mathbf{w}}_*) \leq F_\lambda(\mathbf{w}) \leq F(\mathbf{v}) + \frac{\lambda}{2}\|\mathbf{v}-\mathbf{w}\|^2$. Let $\mathbf{v}=\mathbf{w}=\mathbf{w}_*$, we have

$$F_\lambda(\widetilde{\mathbf{w}}_*) \leq F_\lambda(\mathbf{w}_*) \leq F(\mathbf{w}_*). \tag{7.5}$$

On the other hand, if we let $\widehat{\mathbf{v}} := \arg\min_{\mathbf{v}\in\mathcal{K}}\{F(\mathbf{v}) + \frac{\lambda}{2}\|\mathbf{v}-\widetilde{\mathbf{w}}_*\|^2\}$, then

$$F(\mathbf{w}_*) \leq F(\widehat{\mathbf{v}}) \leq F(\widehat{\mathbf{v}}) + \frac{\lambda}{2}\|\widehat{\mathbf{v}}-\widetilde{\mathbf{w}}_*\|^2 = F_\lambda(\widetilde{\mathbf{w}}_*). \tag{7.6}$$

[c] whose gradient is L-Lipchitz continuous.

Therefore, by (7.5) and (7.6) we have $F(\mathbf{w}_*) = F_\lambda(\widetilde{\mathbf{w}}_*)$. Next, let $\widetilde{\mathbf{w}} := \arg\min_{\mathbf{v}\in\mathcal{K}} F(\mathbf{v}) + \frac{\lambda}{2}\|\mathbf{v}-\mathbf{w}\|^2$. By the smoothness of $F_\lambda(\mathbf{w})$, we have

$$\begin{aligned} F_\lambda(\widetilde{\mathbf{w}}) - F_\lambda(\mathbf{w}) \leq & \nabla F_\lambda(\mathbf{w})^\top(\widetilde{\mathbf{w}}-\mathbf{w}) + \frac{\lambda}{2}\|\widetilde{\mathbf{w}}-\mathbf{w}\|^2 \\ = & -\frac{1}{\lambda}\|\nabla F_\lambda(\mathbf{w})\|^2 + \frac{1}{2\lambda}\|\nabla F_\lambda(\mathbf{w})\|^2 \end{aligned}$$

Rewriting above inequality and combining with $F_\lambda(\widetilde{\mathbf{w}}_*) \leq F_\lambda(\widetilde{\mathbf{w}})$ we get

$$\frac{1}{2\lambda}\|\nabla F_\lambda(\mathbf{w})\|^2 \leq F_\lambda(\mathbf{w}) - F_\lambda(\widetilde{\mathbf{w}}_*).$$

By the definition of $F_\lambda(\mathbf{w})$, for any $\mathbf{w}\in\mathcal{K}$, we have $F_\lambda(\mathbf{w}) \leq F(\mathbf{w})$. Therefore, we have

$$\frac{1}{2\lambda}\|\nabla F_\lambda(\mathbf{w})\|^2 \leq F(\mathbf{w}) - F(\mathbf{w}_*).$$ □

Next, we will characterize the complexity of ASSG for finding an ϵ-nearly stationary point for the problem (1.1) under the LGC by leveraging the result in Lemma 7.1.

Theorem 7.1. *Under the same setting in Theorem 4.1 or Theorem 4.2, then with a high probability $1-\delta$, ASSG-c or ASSG-r guarantees that $\|\nabla F_{1/4}(\mathbf{w}_K)\| \leq \epsilon$ with the iteration complexity of $\widetilde{O}(1/\epsilon^{4(1-\theta)})$.*

Remark. Allen-Zhu [1] considered stochastic gradient descent (SGD) with recursive regularization to solve smooth and convex problems and provided a complexity $\widetilde{O}(1/\epsilon^2)$ for achieving an ϵ-stationary point. In contrast, we focus on non-smooth problems in this paper. When $\theta > \frac{1}{2}$, our methods achieve better complexities.

Proof. Let $\lambda = \frac{1}{4}$ and $\mathbf{w} = \mathbf{w}_K$ in (7.4) of Lemma 7.1, we get

$$\|\nabla F_{1/4}(\mathbf{w}_K)\|^2 \leq \frac{1}{2}(F(\mathbf{w}_K) - F(\mathbf{w}_*)). \tag{7.7}$$

Let $\epsilon = \epsilon^2$ in Theorem 4.1 or Theorem 4.2, we know that with high probability $1-\delta$,

$$F(\mathbf{w}_K) - F(\mathbf{w}_*) \leq 2\epsilon^2. \tag{7.8}$$

By (7.7) and (7.8) we have

$$\|\nabla F_{1/4}(\mathbf{w}_K)\| \leq \epsilon. \tag{7.9}$$ □

8. Applications in Risk Minimization

In this section, we present some applications of the proposed ASSG to risk minimization in machine learning. Let $(\mathbf{x}_i, y_i), i = 1, \ldots, n$ denote a set of pairs of feature vectors and labels that follow a distribution $\mathcal{P}$, where $\mathbf{x}_i \in \mathcal{X} \subset \mathbb{R}^d$ and $y_i \in \mathcal{Y}$. Many machine learning problems end up solving the regularized empirical risk minimization problem:

$$\min_{\mathbf{w}\in\mathbb{R}^d} F(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} \ell(\mathbf{w}^\top \mathbf{x}_i, y_i) + \lambda R(\mathbf{w}), \tag{8.1}$$

where $R(\mathbf{w})$ is a regularizer, λ is the regularization parameter and $\ell(z, y)$ is a loss function. Below we will present several examples in machine learning that enjoy faster convergence by the proposed ASSG than by SSG.

8.1. *Piecewise Linear Minimization*

First, we consider some examples of non-smooth and non-strongly convex problems such that ASSG can achieve linear convergence. In particular, we consider the problem (8.1) with a piecewise linear loss and ℓ_1, ℓ_∞ or $\ell_{1,\infty}$ regularizers.

Piecewise linear loss includes hinge loss [44], generalized hinge loss [4], absolute loss [18], and ϵ-insensitive loss [42]. For particular forms of these loss functions, please refer to [51]. The epigraph of $F(\mathbf{w})$ defined by sum of a piecewise linear loss function and an ℓ_1, ℓ_∞ or $\ell_{1,\infty}$ norm regularizer is a polyhedron. According to the polyhedral error bound condition [50], for any $\epsilon > 0$ there exists a constant $0 < c < \infty$ such that

$$dist(\mathbf{w}, \mathcal{K}_*) \leq c(F(\mathbf{w}) - F_*)$$

for any $\mathbf{w} \in \mathcal{S}_\epsilon$, meaning that the proposed ASSG has an $O(\log(\epsilon_0/\epsilon))$ iteration complexity for solving such family of problems. Formally, we state the result in the following corollary.

Corollary 8.1. *Assume the loss function $\ell(z, y)$ is piecewise linear, then the problem in (8.1) with ℓ_1, ℓ_∞ or $\ell_{1,\infty}$ norm regularizer satisfy the LGC in (1.4) with $\theta = 1$. Hence ASSG can have an iteration complexity of $O(\log(1/\delta)\log(\epsilon_0/\epsilon))$ with a high probability $1 - \delta$.*

8.2. *Piecewise Convex Quadratic Minimization*

In this subsection, we consider some examples of piecewise quadratic minimization problems in machine learning and show that ASSG enjoys an iteration complexity of $\widetilde{O}\left(\frac{1}{\epsilon}\right)$. We first give an definition of piecewise convex quadratic functions, which is from [26]. A function $g(\mathbf{w})$ is a real polynomial if there exists $k \in \mathbb{N}^+$ such that $g(\mathbf{w}) = \sum_{0\leq|\alpha^j|\leq k} \lambda_j \prod_{i=1}^d w_i^{\alpha_i^j}$, where $\lambda_j \in \mathbb{R}$ and $\alpha_i^j \in \mathbb{N}^+ \cup \{0\}$, $\alpha^j = (\alpha_1^j, \ldots, \alpha_d^j)$, and $|\alpha^j| = \sum_{i=1}^d \alpha_i^j$. The constant k is called the degree of g. A continuous function $F(\mathbf{w})$ is said to be a piecewise convex polynomial if there exist finitely many polyhedra $P_1, \ldots, P_m$ with $\cup_{j=1}^m P_j = \mathbb{R}^d$ such that the restriction of F on each P_j is a convex polynomial. Let F_j be the restriction of F on P_j. The degree of a piecewise convex polynomial function F is the maximum of the degree of each F_j. If the degree is 2, the function is referred to as a piecewise convex quadratic function. Note that a piecewise convex quadratic function is not necessarily a smooth function nor a convex function [26].

For examples of piecewise convex quadratic problems in machine learning, one can consider the problem (8.1) with a huber loss, squared hinge loss or square loss, and ℓ_1, ℓ_∞, $\ell_{1,\infty}$, or huber norm regularizer [52]. The huber function is defined as

$$\ell_\delta(z) = \begin{cases} \frac{1}{2}z^2 \text{ if } |z| \leq \delta, \\ \delta(|z| - \frac{1}{2}\delta)\text{otherwise}, \end{cases}$$

which is a piecewise convex quadratic function. The huber loss function $\ell(z, y) = \ell_\delta(z - y)$ has been used for robust regression. A huber regularizer is defined as $R(\mathbf{w}) = \sum_{i=1}^d \ell_\delta(w_i)$.

It has been shown that [26], if $F(\mathbf{w})$ is convex and piecewise convex quadratic, then it satisfies the LGC (1.4) with $\theta = 1/2$. The corollary below summarizes the iteration complexity of ASSG for solving these problems.

Corollary 8.2. *Assume the loss function $\ell(z, y)$ is a convex and piecewise convex quadratic, then the problem in (8.1) with ℓ_1, ℓ_∞, $\ell_{1,\infty}$ or huber norm regularizer satisfy the LGC in (1.4) with $\theta = 1/2$. Hence ASSG can have an iteration complexity of $\widetilde{O}(\frac{\log(1/\delta)}{\epsilon})$ with a high probability $1 - \delta$.*

Remark: The Lipschitz continuity assumption for some loss functions (e.g., squared hinge loss and square loss) can be easily satisfied by adding a boundness constraint on the solution. We note that a recent work [30] also studied the piecewise convex quadratic minimization problems under the error bound condition. They explore the smoothness of the loss functions and develop deterministic accelerated gradient methods with a linear convergence. In contrast, the proposed ASSG is a stochastic algorithm and does not rely on the smoothness assumption. One might also notice that several recent works [16,24] have showed the linear convergence of SVRG by exploring the smoothness of the loss function and a similar condition as in (1.4) with $\theta = 1/2$. However, their required condition is a global growth condition that is required to hold for any $\mathbf{w} \in \mathbb{R}^d$.

Indeed, a convex and piecewise convex quadratic function enjoy a global growth condition [26]:

$$dist(\mathbf{w}, \mathcal{K}_*) \leq c[F(\mathbf{w}) - F_* + (F(\mathbf{w}) - F_*)^{1/2}], \quad \forall \mathbf{w} \in \mathbb{R}^d.$$

It remains an open problem that how to leverage such a global growth condition to develop a linear convergence for SVRG and other similar algorithms for solving finite-sum smooth problems, which is beyond the scope of this work. Nevertheless, using the above global growth condition we can reduce the iteration complexity by a $\log(\epsilon_0/\epsilon)$ factor for ASSG. We include the details in Appendix J.

8.3. *Structured composite non-smooth problems*

Next, we present a corollary of our main result regarding the following structured problem:

$$\min_{\mathbf{w}\in\mathbb{R}^d} F(\mathbf{w}) \triangleq h(X\mathbf{w}) + R(\mathbf{w}). \tag{8.2}$$

where $X \in \mathbb{R}^{n\times d}$, $h(\mathbf{u})$ is a strongly convex function (not necessarily a smooth function) on any compact set and $R(\mathbf{w})$ is ℓ_1, ℓ_∞ or $\ell_{1,\infty}$ norm regularizer. The corollary below formally states the LGC of the above problem and the iteration complexity of ASSG.

Corollary 8.3. *Assume $h(\mathbf{u})$ is a strongly convex function on any compact set and $P(\mathbf{w})$ is polyhedral, then the problem in (8.2) satisfies the LGC in (1.4) with $\theta = 1/2$. Hence ASSG can have an iteration complexity of $\widetilde{O}(\frac{\log(1/\delta)}{\epsilon})$ with a high probability $1-\delta$.*

The proof of the first part of Corollary 8.3 can be found in [50]. One example of $h(\mathbf{u})$ is p-norm error ($p \in (0,1)$), where $h(\mathbf{u}) = \frac{1}{n}\sum_{i=1}^n |u_i - y_i|^p$. The local strong convexity of the p-norm error ($p \in (1,2)$) is shown in [15].

Finally, we give an example that satisfies the LGC with intermediate values $\theta \in (0, 1/2)$. We can consider an ℓ_1 constrained ℓ_p norm regression [37]:

$$\min_{\|\mathbf{w}\|_1\leq s} F(\mathbf{w}) \triangleq \frac{1}{n}\sum_{i=1}^n (\mathbf{x}_i^\top \mathbf{w} - y_i)^p, \quad p \in 2\mathbb{N}^+.$$

[30] have shown that the problem above satisfies the LGC in (1.4) with $\theta = \frac{1}{p}$.

Table 1. Statistics of real datasets

Dataset	#Training (n)	#Features (d)	Problem Type
covtype.binary	581,012	54	Classification
real-sim	72,309	20,958	Classification
url	2,396,130	3,231,961	Classification
avazu	40,428,967	1,000,000	Classification
gisette	6,000	5,000	Classification
kdd 2010 raw	19,264,097	1,163,024	Classification
news20.binary	19,996	1,355,191	Classification
rcv1.binary	20,242	47,236	Classification
webspam	350,000	16,609,143	Classification
million songs	463,715	90	Regression
E2006-tfidf	16,087	150,360	Regression
E2006-log1p	16,087	4,272,227	Regression

9. Experiments

In this section, we perform some experiments to demonstrate effectiveness of proposed algorithms. For the first two experimens, we use very large-scale datasets

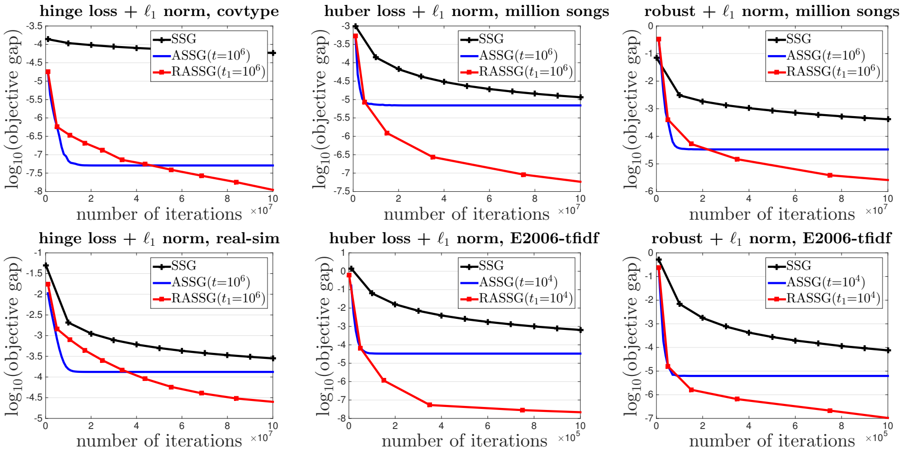


Fig. 1. Comparison of different algorithms for solving different problems on different datasets ($\lambda = 10^{-4}$).

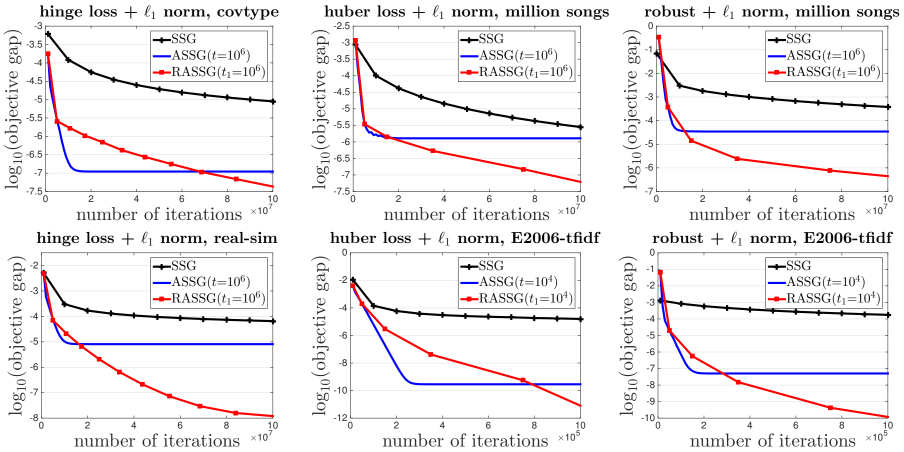
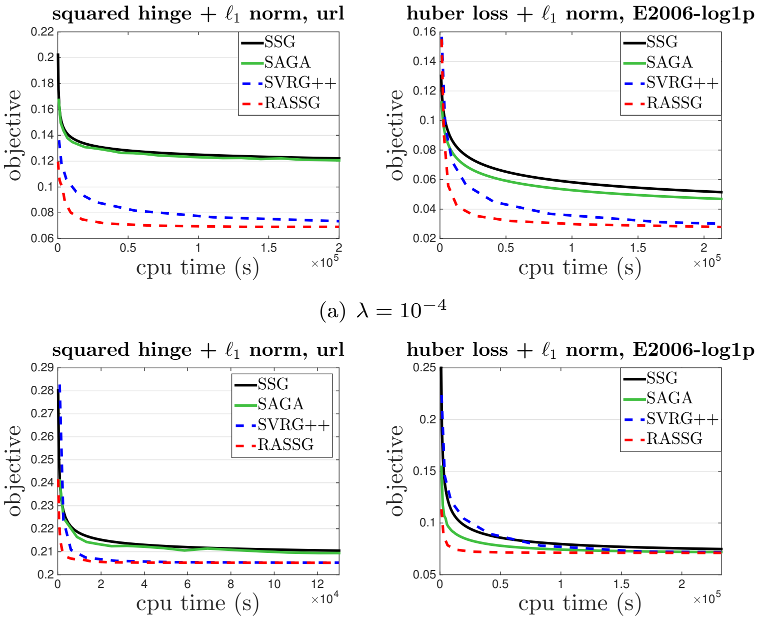


Fig. 2. Comparison of different algorithms for solving different problems on different datasets ($\lambda = 10^{-2}$).

from libsvm website in experiments, including covtype.binary, real-sim, url for classification, million songs, E2006-tfidf, E2006-log1p for regression. While for the last experimenst, we only consider classification problem and use nine datasets from libsvm website including covtype.binary, real-sim, avazu, gisette, kdd 2010 raw, news20.binary, rcv1.binary, url and webspam. The detailed statistics of these datasets are shown in Table 8.3.



(a) $\lambda = 10^{-4}$

(b) $\lambda = 10^{-2}$

Fig. 3. Comparison of different algorithms for solving different problems on different datasets.

Effectiveness of ASSG-c and RASSG-c for non-smooth problems. We first compare ASSG with SSG on three tasks: ℓ_1 norm regularized hinge loss minimization for linear classification, ℓ_1 norm regularized Huber loss minimization for linear regression, and ℓ_1 norm regularized p-norm robust regression with a loss function $\ell(\mathbf{w}^\top\mathbf{x}_i, y_i) = |\mathbf{w}^\top\mathbf{x}_i - y_i|^p$. The regularization parameter λ is set to be $10^{-4}/10^{-2}$ in all tasks. We set $\gamma = 1$ in Huber loss and $p = 1.5$ in robust regression. In all experiments, we use the constrained variant of ASSG, i.e., ASSG-c. For fairness, we use the same initial solution with all zero entries for all algorithms. We use a decreasing step size proportional to $1/\sqrt{\tau}$ (τ is the iteration index) in SSG. The initial step size of SSG is tuned in a wide range to obtain the fastest convergence. The step size of ASSG in the first stage is also tuned around the best initial step size of SSG. The value of D_1 in both ASSG and RASSG is set to 100 for all problems. In implementing the RASSG, we restart every 5 stages with t increased by a factor of 1.15, 2 and 2 respectively for hinge loss, Huber loss and robust regression. We tune the parameter ω among $\{0.3, 0.6, 0.9, 1\}$. We report the results of ASSG with a fixed number of iterations per-stage t and RASSG with an increasing sequence of t. The results are plotted in Figure 1 and Figure 2 in which we plot the log difference between the objective value and the smallest obtained objective value (to which

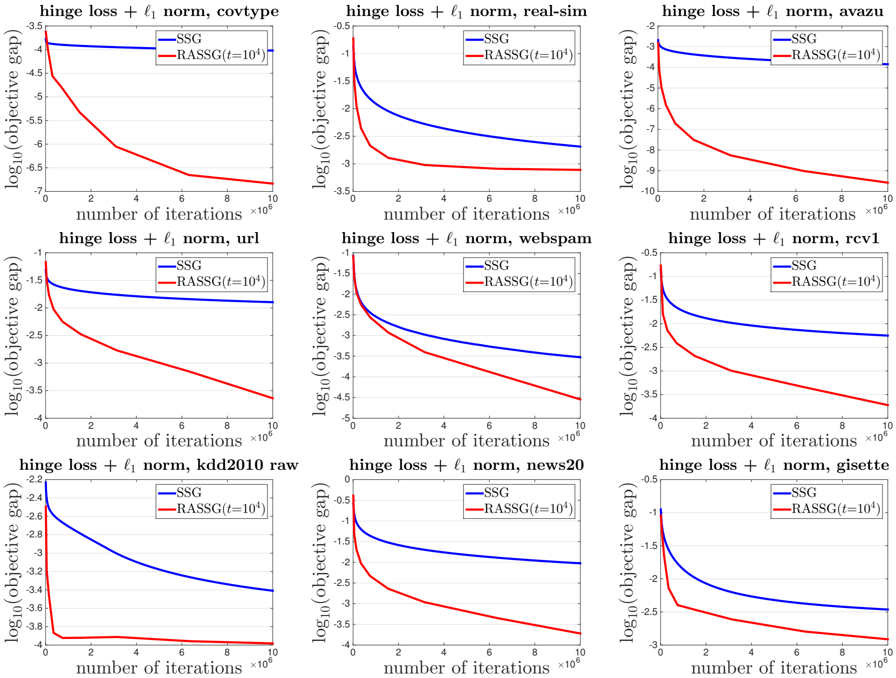


Fig. 4. Comparison of SSG and RASSG-s on different datasets ($\lambda = 10^{-4}$).

we refer as objective gap) versus number of iterations. The figures show that (i) ASSG can quickly converge to a certain level set determined implicitly by t; (ii) RASSG converges much faster than SSG to more accurate solutions; (iii) RASSG can gradually decrease the objective value.

Effectiveness of ASSG-c and RASSG-c for smooth problems. Second, we compare RASSG with state-of-art stochastic optimization algorithms for solving a finite-sum problem with a smooth piecewise quadratic loss (e.g., squared hinge loss, huber loss) and an ℓ_1 norm regularization. In particular, we compare with two variance-reduction algorithms that leverage the smoothness of the function, namely SAGA [10] and SVRG++ [2]. We conduct experiments on two high-dimensional datasets url and E2006-log1p and fix the regularization parameter $\lambda = 10^{-4}$ or $\lambda = 10^{-2}$. We use $\delta = 1$ in Huber loss. For RASSG, we start from $D_1 = 100$ and $t_1 = 10^3$, then restart it every 5 stages with t increased by a factor of 2. We tune the initial step sizes for all algorithms in a wide range and set the values of parameters in SVRG++ followed by [2]. We plot the objective versus the CPU time (second) in Figure 3. The results show that RASSG converges faster than other three algorithms for the two tasks. This is not surprising considering that RASSG, SAGA and SVRG++ suffer from an iteration complexity of $\widetilde{O}(1/\epsilon)$, $O(n/\epsilon)$, and $O(n\log(1/\epsilon) + 1/\epsilon)$, respectively.

Effectiveness of RASSG-s. Finally, we compare RASSG-s with SSG on ℓ_1 norm regularized hinge loss minimization for linear classification. The regularization parameter λ is set to be 10^{-4}, and the initial iteration number of RASSG-s is set to be $10,000$. We fixed the total number of iterations as $1,000,000$ both for SSG and RASSG-s. Although the parameter $\theta = 1$ in the considered task, we can always reduce it to $\theta = \frac{1}{2}$ [48]. Thus we set GGC parameter $\theta = \frac{1}{2}$ in this experiment. The other parameters of SSG and RASSG-s are set as same as the first experiment. The results are presented in Figure 4, showing that RASSG-s converges much faster than SSG to more accurate solutions.

10. Conclusion

In this paper, we have proposed accelerated stochastic subgradient methods for solving general non-strongly convex stochastic optimization under the functional local growth condition. The proposed methods enjoy a lower iteration complexity than vanilla stochastic subgradient method and also a logarithmic dependence on the impact of the initial solution. We have also made an extension by developing a more practical variant. Applications in machine learning have demonstrated the faster convergence of the proposed methods.

Appendix A. Proof of Corollary 4.1

Proof. First, we show that for any $1 \leq k \leq K$,

$$\|\mathbf{w}_k - \mathbf{w}_0\|_2 \leq 2D_1 \tag{A.1}$$

When $k = 1$, it is easy to show that $\|\mathbf{w}_1 - \mathbf{w}_0\|_2 \leq D_1$, which satisfies inequality (A.1). When $k \geq 2$, we have

$$\begin{aligned}
&\|\mathbf{w}_k - \mathbf{w}_0\|_2\\
\leq&\|\mathbf{w}_k - \mathbf{w}_{k-1}\|_2 + \|\mathbf{w}_{k-1} - \mathbf{w}_{k-2}\|_2 + \cdots + \|\mathbf{w}_2 - \mathbf{w}_1\|_2 + \|\mathbf{w}_1 - \mathbf{w}_0\|_2\\
\leq&D_1/2^{k-1} + D_1/2^{k-2} + \cdots + D_1/2 + D_1 \leq 2D_1
\end{aligned}$$

where the second inequality is based on the updates of Algorithm 1. With probability 1, we have

$$\begin{aligned}
F(\mathbf{w}_k) - F(\mathbf{w}_0) =&F(\mathbf{w}_k) - F_* + F_* - F(\mathbf{w}_0)\\
\leq&\|\partial F(\mathbf{w}_k)\|_2\|\mathbf{w}_k - \mathbf{w}_0\|_2 + F_* - F(\mathbf{w}_0) \leq 2GD_1 + \epsilon_0
\end{aligned} \tag{A.2}$$

where the last inequality using the fact that $\|\partial F(\mathbf{w}_k)\|_2 \leq G$, inequality (A.1) and Assumption 3.1 (a). Based on Theorem 4.1, ASSG-c guarantees that

$$Prob(F(\mathbf{w}_K) - F_*) \leq \epsilon) \geq 1 - \delta \tag{A.3}$$

Then

$$\begin{aligned}\mathbb{E}\left[F(\mathbf{w}_K) - F_*\right] =&\mathbb{E}\left[F(\mathbf{w}_K) - F_* | F(\mathbf{w}_K) - F_* \leq \epsilon\right] Prob(F(\mathbf{w}_K) - F_*) \leq \epsilon) \\ &+ \mathbb{E}\left[F(\mathbf{w}_K) - F_* | F(\mathbf{w}_K) - F_* \geq \epsilon\right] Prob(F(\mathbf{w}_K) - F_*) \geq \epsilon) \\ \leq& \epsilon + (2GD_1 + \epsilon_0)\delta \leq 2\epsilon\end{aligned}$$

where the first inequality uses inequalities (A.2) and (A.3), and the second inequalty is due to $\delta \leq \frac{\epsilon}{2GD_1+\epsilon_0}$. Therefore, ASSG-c achieves that $\mathbb{E}\left[F(\mathbf{w}_K) - F_*\right] \leq 2\epsilon$ using at most $O\left(\lceil \log_2(\frac{\epsilon_0}{\epsilon})\rceil \log\left(\frac{2GD_1+\epsilon_0}{\epsilon}\right) c^2G^2/\epsilon^{2(1-\theta)}\right)$ iterations provided $D_1 = O(\frac{c\epsilon_0}{\epsilon^{(1-\theta)}})$. □

Appendix B. Proof of Lemma 4.2

Proof. By the optimality of $\widehat{\mathbf{w}}_*$, we have for any $\mathbf{w} \in \mathcal{K}$

$$\left(\partial F(\widehat{\mathbf{w}}_*) + \frac{1}{\beta}(\widehat{\mathbf{w}}_* - \mathbf{w}_1)\right)^\top (\mathbf{w} - \widehat{\mathbf{w}}_*) \geq 0.$$

Let $\mathbf{w} = \mathbf{w}_1$, we have

$$\partial F(\widehat{\mathbf{w}}_*)^\top (\mathbf{w}_1 - \widehat{\mathbf{w}}_*) \geq \frac{\|\widehat{\mathbf{w}}_* - \mathbf{w}_1\|_2^2}{\beta}.$$

Because $\|\partial F(\widehat{\mathbf{w}}_*)\|_2 \leq G$ due to $\|\partial f(\mathbf{w}; \xi)\|_2 \leq G$, then

$$\|\widehat{\mathbf{w}}_* - \mathbf{w}_1\|_2 \leq \beta G.$$

Next, we bound $\|\mathbf{w}_t - \mathbf{w}_1\|_2$. According to the update of $\mathbf{w}_{t+1}$ we have

$$\|\mathbf{w}_{t+1} - \mathbf{w}_1\|_2 \leq \|\mathbf{w}'_{t+1} - \mathbf{w}_1\|_2 = \| - \eta_t \partial f(\mathbf{w}_t; \xi_t) + (1 - \eta_t/\beta)(\mathbf{w}_t - \mathbf{w}_1)\|_2.$$

We prove $\|\mathbf{w}_t - \mathbf{w}_1\|_2 \leq 2\beta G$ by induction. First, we consider $t = 1$, where $\eta_t = 2\beta$, then

$$\|\mathbf{w}_2 - \mathbf{w}_1\|_2 \leq \|2\beta \partial f(\mathbf{w}_t; \xi_t)\|_2 \leq 2\beta G.$$

Then we consider any $t \geq 2$, where $\eta_t/\beta \leq 1$. Then

$$\begin{aligned}\|\mathbf{w}_{t+1} - \mathbf{w}_1\|_2 \leq& \left\| -\frac{\eta_t}{\beta}\beta \partial f(\mathbf{w}_t; \xi_t) + \left(1 - \frac{\eta_t}{\beta}\right)(\mathbf{w}_t - \mathbf{w}_1)\right\|_2 \\ \leq& \frac{\eta_t}{\beta}\beta G + \left(1 - \frac{\eta_t}{\beta}\right) 2\beta G \leq 2\beta G.\end{aligned}$$

Therefore

$$\|\widehat{\mathbf{w}}_* - \mathbf{w}_t\|_2 \leq 3\beta G.$$ □

Appendix C. Proof of Lemma 4.3

In this proof, we need the following lemma.

Lemma Appendix C.1. *(Lemma 3 [23]) Suppose $X_1, \ldots, X_T$ is a martingale difference sequence with $|X_t| \leq b$. Let*

$$Var_t X_t = Var(X_t | X_1, \ldots, X_{t-1}).$$

where Var denotes the variance. Let $V = \sum_{t=1}^T Var_t X_t$ be the sum of conditional variance of X_t's. Further, let $\sigma = \sqrt{V}$. Then we have for any $\delta < 1/e$ and $T \geq 3$,

$$\Pr\left(\sum_{t=1}^T X_t > \max\{2\sigma, 3b\sqrt{\log(1/\delta)}\}\sqrt{\log(1/\delta)}\right) \leq 4\delta \log T.$$

Then, let us start the proof of Lemma 4.3.

Proof. Let $\mathbf{g}_t = \partial f(\mathbf{w}_t; \xi_t) + (\mathbf{w}_t - \mathbf{w}_1)/\beta$ and $\partial \widehat{F}(\mathbf{w}_t) = \partial F(\mathbf{w}_t) + (\mathbf{w}_t - \mathbf{w}_1)/\beta$. Note that $\|\mathbf{g}_t\|_2 \leq 3G$. According to the standard analysis for the stochastic gradient method we have

$$\mathbf{g}_t^\top (\mathbf{w}_t - \widehat{\mathbf{w}}_*) \leq \frac{1}{2\eta_t}\|\mathbf{w}_t - \widehat{\mathbf{w}}_*\|_2^2 - \frac{1}{2\eta_t}\|\mathbf{w}_{t+1} - \widehat{\mathbf{w}}_*\|_2^2 + \frac{\eta_t}{2}\|\mathbf{g}_t\|_2^2.$$

Then

$$\begin{aligned}\partial \widehat{F}(\mathbf{w}_t)^\top (\mathbf{w}_t - \widehat{\mathbf{w}}_*) \leq & \frac{1}{2\eta_t}\|\mathbf{w}_t - \widehat{\mathbf{w}}_*\|_2^2 - \frac{1}{2\eta_t}\|\mathbf{w}_{t+1} - \widehat{\mathbf{w}}_*\|_2^2 + \frac{\eta_t}{2}\|\mathbf{g}_t\|_2^2 \\ & + (\partial \widehat{F}(\mathbf{w}_t) - \mathbf{g}_t)^\top (\mathbf{w}_t - \widehat{\mathbf{w}}_*).\end{aligned}$$

By strong convexity of $\widehat{F}$ we have

$$\widehat{F}(\widehat{\mathbf{w}}_*) - \widehat{F}(\mathbf{w}_t) \geq \partial \widehat{F}(\mathbf{w}_t)^\top (\widehat{\mathbf{w}}_* - \mathbf{w}_t) + \frac{1}{2\beta}\|\widehat{\mathbf{w}}_* - \mathbf{w}_t\|_2^2.$$

Then

$$\begin{aligned}\widehat{F}(\mathbf{w}_t) - \widehat{F}(\widehat{\mathbf{w}}_*) \leq & \frac{1}{2\eta_t}\|\mathbf{w}_t - \widehat{\mathbf{w}}_*\|_2^2 - \frac{1}{2\eta_t}\|\mathbf{w}_{t+1} - \widehat{\mathbf{w}}_*\|_2^2 + \frac{\eta_t}{2}\|\mathbf{g}_t\|_2^2 \\ & + (\partial \widehat{F}(\mathbf{w}_t) - \mathbf{g}_t)^\top (\mathbf{w}_t - \widehat{\mathbf{w}}_*) - \frac{1}{2\beta}\|\widehat{\mathbf{w}}_* - \mathbf{w}_t\|_2^2 \\ \leq & \frac{1}{2\eta_t}\|\mathbf{w}_t - \widehat{\mathbf{w}}_*\|_2^2 - \frac{1}{2\eta_t}\|\mathbf{w}_{t+1} - \widehat{\mathbf{w}}_*\|_2^2 + \frac{\eta_t}{2}\|\mathbf{g}_t\|_2^2 \\ & + \underbrace{(\partial F(\mathbf{w}_t) - \partial f(\mathbf{w}_t; \xi_t))^\top (\mathbf{w}_t - \widehat{\mathbf{w}}_*)}_{\zeta_t} - \frac{1}{2\beta}\|\widehat{\mathbf{w}}_* - \mathbf{w}_t\|_2^2.\end{aligned}$$

By summing the above inequalities across $t = 1, \dots, T$, we have

$$\begin{aligned}\sum_{t=1}^{T}(\widehat{F}(\mathbf{w}_t) - \widehat{F}(\widehat{\mathbf{w}}_*)) \leq & \sum_{t=1}^{T-1}\frac{1}{2}\left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} - \frac{1}{2\beta}\right)\|\widehat{\mathbf{w}}_* - \mathbf{w}_{t+1}\|_2^2 + \sum_{t=1}^{T}\zeta_t \\ & - \frac{1}{4\beta}\sum_{t=1}^{T}\|\widehat{\mathbf{w}}_* - \mathbf{w}_t\|_2^2 - \frac{1}{4\beta}\|\widehat{\mathbf{w}}_* - \mathbf{w}_1\|_2^2 + \frac{1}{2\eta_1}\|\widehat{\mathbf{w}}_* - \mathbf{w}_1\|_2^2 + \frac{9G^2}{2}\sum_{t=1}^{T}\eta_t \\ \leq & \sum_{t=1}^{T}\zeta_t - \frac{1}{4\beta}\sum_{t=1}^{T}\|\widehat{\mathbf{w}}_* - \mathbf{w}_t\|_2^2 + 9\beta G^2(1 + \log T).\end{aligned}$$

where the last inequality uses $\eta_t = \frac{2\beta}{t}$.

Next, we bound R.H.S of the above inequality by using Lemma Appendix C.1. To proceed the proof of Lemma 4.3. We let $X_t = \zeta_t$ and $D_T = \sum_{t=1}^{T}\|\mathbf{w}_t - \widehat{\mathbf{w}}_*\|_2^2$. Then $X_1, \dots, X_T$ is a martingale difference sequence. Let $D = 3\beta G$. Note that $|\zeta_t| \leq 2GD$. By Lemma Appendix C.1, for any $\delta < 1/e$ and $T \geq 3$, with a probability $1 - \delta$ we have

$$\sum_{t=1}^{T}\zeta_t \leq \max\left\{2\sqrt{\log(4\log T/\delta)}\sqrt{\sum_{t=1}^{T}\mathrm{Var}_t\zeta_t}, 6GD\log(4\log T/\delta)\right\}.$$

Note that

$$\sum_{t=1}^{T}\mathrm{Var}_t\zeta_t \leq \sum_{t=1}^{T}\mathrm{E}_t[\zeta_t^2] \leq 4G^2\sum_{t=1}^{T}\|\mathbf{w}_t - \widehat{\mathbf{w}}_*\|_2^2 = 4G^2 D_T.$$

As a result, with a probability $1 - \delta$,

$$\begin{aligned}\sum_{t=1}^{T}\zeta_t \leq & 4G\sqrt{\log(4\log T/\delta)}\sqrt{D_T} + 6GD\log(4\log T/\delta) \\ \leq & 16\beta G^2\log(4\log T/\delta) + \frac{1}{4\beta}D_T + 6GD\log(4\log T/\delta).\end{aligned}$$

As a result, with a probability $1 - \delta$,

$$\begin{aligned}\sum_{t=1}^{T}\zeta_t - \frac{1}{4\beta}\sum_{t=1}^{T}\|\widehat{\mathbf{w}}_* - \mathbf{w}_t\|_2^2 \leq & 16\beta G^2\log(4\log T/\delta) + 6GD\log(4\log T/\delta) \\ = & 34\beta G^2\log(4\log T/\delta).\end{aligned}$$

Thus, with a probability $1 - \delta$

$$\begin{aligned}\widehat{F}(\widehat{\mathbf{w}}_T) - \widehat{F}(\widehat{\mathbf{w}}_*) \leq & \frac{34\beta G^2\log(4\log T/\delta)}{T} + \frac{9\beta G^2(1 + \log T)}{T} \\ \leq & \frac{34\beta G^2(1 + \log T + \log(4\log T/\delta))}{T}.\end{aligned}$$

Using the facts that $F(\widehat{\mathbf{w}}_T) \leq \widehat{F}(\widehat{\mathbf{w}}_T)$ and $\widehat{F}(\widehat{\mathbf{w}}_*) \leq \widehat{F}(\mathbf{w}) = F(\mathbf{w}) + \frac{\|\mathbf{w}-\mathbf{w}_1\|_2^2}{2\beta}$, we have

$$F(\widehat{\mathbf{w}}_T) - F(\mathbf{w}) - \frac{\|\mathbf{w}-\mathbf{w}_1\|_2^2}{2\beta} \leq \frac{34\beta G^2(1+\log T + \log(4\log T/\delta))}{T}. \qquad \square$$

Appendix D. Proof of Theorem 4.2

Proof. Let $\mathbf{w}_{k,\epsilon}^{\dagger}$ denote the closest point to $\mathbf{w}_k$ in the ϵ sublevel set. Define $\epsilon_k \triangleq \frac{\epsilon_0}{2^k}$. First, we note that $\beta_k \geq \frac{2c^2\epsilon_{k-1}}{\epsilon^{2(1-\theta)}}$. We will show by induction that $F(\mathbf{w}_k) - F_* \leq \epsilon_k + \epsilon$ for $k = 0, 1, \ldots$ with a high probability, which leads to our conclusion when $k = K$. The inequality holds obviously for $k = 0$. Conditioned on $F(\mathbf{w}_{k-1}) - F_* \leq \epsilon_{k-1} + \epsilon$, we will show that $F(\mathbf{w}_k) - F_* \leq \epsilon_k + \epsilon$ with a high probability. We apply Lemma 4.3 to the k-th stage of Algorithm 2 conditioned on the randomness in previous stages. With a probability at least $1 - \tilde{\delta}$ we have

$$F(\mathbf{w}_k) - F(\mathbf{w}_{k-1,\epsilon}^{\dagger}) \leq \frac{\|\mathbf{w}_{k-1,\epsilon}^{\dagger} - \mathbf{w}_{k-1}\|_2^2}{2\beta_k} + \frac{34\beta_k G^2\left(1 + \log t + \log\left(\frac{4\log t}{\tilde{\delta}}\right)\right)}{t}. \tag{D.1}$$

Similar to the proof of Theorem 4.1, by Lemma 3.1, we have

$$\|\mathbf{w}_{k-1,\epsilon}^{\dagger} - \mathbf{w}_{k-1}\|_2 \leq \frac{c\epsilon_{k-1}}{\epsilon^{1-\theta}}. \tag{D.2}$$

Combining (D.1) and (D.2), we have

$$F(\mathbf{w}_k) - F(\mathbf{w}_{k-1,\epsilon}^{\dagger}) \leq \frac{1}{2\beta_k}\left(\frac{c\epsilon_{k-1}}{\epsilon^{1-\theta}}\right)^2 + \frac{34\beta_k G^2(1+\log t + \log(4\log t/\tilde{\delta}))}{t}.$$

Using the fact that $\beta_k \geq \frac{2c^2\epsilon_{k-1}}{\epsilon^{2(1-\theta)}}$ and $t \geq \frac{68\beta_k G^2(1+\log t+\log(4\log t/\tilde{\delta}))}{\epsilon_k} = \frac{136\beta_1 G^2(1+\log t+\log(4\log t/\tilde{\delta}))}{\epsilon_0}$, we get

$$F(\mathbf{w}_k) - F(\mathbf{w}_{k-1,\epsilon}^{\dagger}) \leq \frac{\epsilon_{k-1}}{4} + \frac{\epsilon_k}{2} = \epsilon_k,$$

which together with the fact that $F(\mathbf{w}_{k-1,\epsilon}^{\dagger}) - F_* \leq \epsilon$ by definition of $\mathbf{w}_{k-1,\epsilon}^{\dagger}$ implies

$$F(\mathbf{w}_k) - F_* \leq \epsilon + \epsilon_k.$$

Therefore by induction, we have with a probability at least $(1-\tilde{\delta})^K$,

$$F(\mathbf{w}_K) - F_* \leq \epsilon_K + \epsilon = \frac{\epsilon_0}{2^K} + \epsilon \leq 2\epsilon,$$

where the last inequality is due to the value of $K = \lceil \log_2(\frac{\epsilon_0}{\epsilon}) \rceil$. Since $\tilde{\delta} = \delta/K$, then $(1-\tilde{\delta})^K \geq 1 - \delta$. $\square$

Appendix E. Proof of Lemma 4.4

Proof. Let rewrite the update of $\mathbf{w}_{\tau+1}$ in k-th epoch as

$$\mathbf{w}' = \mathbf{w}_\tau - \eta\partial f(\mathbf{w}_\tau;\xi_\tau), \mathbf{w}_{\tau+1} = \Pi_{\mathcal{K}}[\mathbf{w}'].$$

Then for any fixed $\mathbf{w} \in \mathcal{K}$ we have

$$\begin{aligned}\frac{1}{2\eta}\|\mathbf{w}_{\tau+1} - \mathbf{w}\|^2 \leq & \frac{1}{2\eta}\|\mathbf{w}' - \mathbf{w}\|^2 = \frac{1}{2\eta}\|\mathbf{w}_\tau - \eta\partial f(\mathbf{w}_\tau;\xi_\tau) - \mathbf{w}\|^2 \\ = & \frac{1}{2\eta}\|\mathbf{w}_\tau - \mathbf{w}\|^2 - \langle\partial f(\mathbf{w}_\tau;\xi_\tau), \mathbf{w}_\tau - \mathbf{w}\rangle + \frac{\eta}{2}\|\partial f(\mathbf{w}_\tau;\xi_\tau)\|^2,\end{aligned}$$

which implies

$$\begin{aligned}\langle\partial F(\mathbf{w}_\tau), \mathbf{w}_\tau - \mathbf{w}\rangle \leq & \frac{1}{2\eta}\|\mathbf{w}_\tau - \mathbf{w}\|^2 - \frac{1}{2\eta}\|\mathbf{w}_{\tau+1} - \mathbf{w}\|^2 + \frac{\eta}{2}\|\partial f(\mathbf{w}_\tau;\xi_\tau)\|^2 \\ & - \langle\partial f(\mathbf{w}_\tau;\xi_\tau) - \partial F(\mathbf{w}_\tau), \mathbf{w}_\tau - \mathbf{w}\rangle.\end{aligned}$$

By the convexity of $F(\mathbf{w})$, i.e., $F(\mathbf{w}) - F(\mathbf{w}_\tau) \geq \langle\partial F(\mathbf{w}_\tau), \mathbf{w} - \mathbf{w}_\tau\rangle$, and Assumption 3.1 (c), then

$$\begin{aligned}F(\mathbf{w}_\tau) - F(\mathbf{w}) \leq & \frac{1}{2\eta}\|\mathbf{w}_\tau - \mathbf{w}\|^2 - \frac{1}{2\eta}\|\mathbf{w}_{\tau+1} - \mathbf{w}\|^2 + \frac{\eta G^2}{2} \\ & - \langle\partial f(\mathbf{w}_\tau;\xi_\tau) - \partial F(\mathbf{w}_\tau), \mathbf{w}_\tau - \mathbf{w}\rangle.\end{aligned}$$

Taking expectation over $1, \ldots, \tau$, we have

$$\mathrm{E}[F(\mathbf{w}_\tau) - F(\mathbf{w})] \leq \frac{1}{2\eta}\mathrm{E}[\|\mathbf{w}_\tau - \mathbf{w}\|^2] - \frac{1}{2\eta}\mathrm{E}[\|\mathbf{w}_{\tau+1} - \mathbf{w}\|^2] + \frac{\eta G^2}{2},$$

where uses the fact that $\mathrm{E}[\langle\partial f(\mathbf{w}_\tau;\xi_\tau) - \partial F(\mathbf{w}_\tau), \mathbf{w}_\tau - \mathbf{w}\rangle] = 0$. By summing the above inequalities across $\tau = 1, \ldots, t$, we have

$$\sum_{\tau=1}^{t}(\mathrm{E}[F(\mathbf{w}_\tau) - F(\mathbf{w})]) \leq \frac{1}{2\eta}\|\mathbf{w}_1 - \mathbf{w}\|^2 + \frac{\eta G^2 t}{2}.$$

It implies

$$\mathrm{E}\left[\frac{1}{t}\sum_{\tau=1}^{t}(F(\mathbf{w}_\tau) - F(\mathbf{w}))\right] \leq \frac{1}{2\eta t}\|\mathbf{w}_1 - \mathbf{w}\|^2 + \frac{\eta G^2}{2}.$$

We complete the proof by using the convexity of $F(\mathbf{w})$. □

Appendix F. Monotonicity of B_ϵ/ϵ

Lemma Appendix F.1. $\frac{B_\epsilon}{\epsilon}$ *is monotonically decreasing in* ϵ.

Proof. Consider $\epsilon' > \epsilon > 0$. Let $\mathbf{x}_{\epsilon'}$ be any point on $\mathcal{L}_{\epsilon'}$ such that $dist(\mathbf{x}_{\epsilon'}, \Omega_*) = B_{\epsilon'}$ and $\mathbf{x}^*_{\epsilon'}$ be the closest point to $\mathbf{x}_{\epsilon'}$ in Ω_* so that $\|\mathbf{x}^*_{\epsilon'} - \mathbf{x}_{\epsilon'}\| = B_{\epsilon'}$. We define a new point between $\mathbf{x}_{\epsilon'}$ and $\mathbf{x}^*_{\epsilon'}$ as

$$\bar{\mathbf{x}} = \frac{B_\epsilon}{B_{\epsilon'}}\mathbf{x}_{\epsilon'} + \frac{B_{\epsilon'} - B_\epsilon}{B_{\epsilon'}}\mathbf{x}^*_{\epsilon'}.$$

Since $0 < B_\epsilon < B_{\epsilon'}$, $\bar{\mathbf{x}}$ is strictly between $\mathbf{x}_{\epsilon'}$ and $\mathbf{x}^*_{\epsilon'}$ and $dist(\bar{\mathbf{x}}, \Omega_*) = \|\mathbf{x}^*_{\epsilon'} - \bar{\mathbf{x}}\| = \frac{B_\epsilon}{B_{\epsilon'}}\|\mathbf{x}^*_{\epsilon'} - \mathbf{x}_{\epsilon'}\| = B_\epsilon$. By the convexity of F, we have

$$\frac{F(\bar{\mathbf{x}}) - F_*}{dist(\bar{\mathbf{x}}, \Omega_*)} \leq \frac{F(\mathbf{x}_{\epsilon'}) - F_*}{dist(\mathbf{x}_{\epsilon'}, \Omega_*)} = \frac{\epsilon'}{B_{\epsilon'}}.$$

Note that we must have $F(\bar{\mathbf{x}}) - F_* \geq \epsilon$ since, otherwise, we can move $\bar{\mathbf{x}}$ towards $\mathbf{x}_{\epsilon'}$ until $F(\bar{\mathbf{x}}) - F_* = \epsilon$ but $dist(\bar{\mathbf{x}}, \Omega_*) > B_\epsilon$, contradicting with the definition of B_ϵ. Then, the proof is completed by applying $F(\bar{\mathbf{x}}) - F_* \geq \epsilon$ and $dist(\bar{\mathbf{x}}, \Omega_*) = B_\epsilon$ to the previous inequality. □

Appendix G. Proof of Lemma 6.1

In this proof, we need the following lemma.

Lemma Appendix G.1 (Lemma 2 of [25]). *Let $X_1, \ldots, X_t$ be a martingale difference sequence, i.e.* $\mathrm{E}[X_\tau | X_1, \ldots, X_{\tau-1}] = 0$ *for all τ. Suppose that for some values σ_τ, for $\tau = 1, \ldots, t$, we have* $\mathrm{E}\left[\exp\left(\frac{X_\tau^2}{\sigma_\tau^2}\right)|X_1, \ldots, X_{\tau-1}\right] \leq \exp(1)$. *Then with probability at least $1 - \delta$, we have*

$$\sum_{\tau=1}^{t} X_\tau \leq \sqrt{3\log(1/\delta)\sum_{\tau=1}^{t}\sigma_\tau^2}.$$

Then, let us start the proof of Lemma 6.1.

Proof. Based on the fact that $\frac{1}{2}\|\mathbf{w} - \mathbf{w}_\tau\|_2^2 + \eta\partial f(\mathbf{w}_\tau; \xi_\tau)^\top \mathbf{w} + \eta R(\mathbf{w})$ is $\frac{1}{2}$-stongly convex in terms of $\mathbf{w}$, then for any $\mathbf{w} \in \mathcal{B}(\mathbf{w}_1, D)$, we have

$$\begin{aligned}&\frac{1}{2}\|\mathbf{w}_{\tau+1} - \mathbf{w}_\tau\|_2^2 + \eta\partial f(\mathbf{w}_\tau; \xi_\tau)^\top \mathbf{w}_{\tau+1} + \eta R(\mathbf{w}_{\tau+1})\\ \leq&\frac{1}{2}\|\mathbf{w} - \mathbf{w}_\tau\|_2^2 + \eta\partial f(\mathbf{w}_\tau; \xi_\tau)^\top \mathbf{w} + \eta R(\mathbf{w}) - \frac{1}{2}\|\mathbf{w} - \mathbf{w}_{\tau+1}\|_2^2.\end{aligned}$$

Rewrite the inequality and then it becomes

$$\begin{aligned}&\partial f(\mathbf{w}_\tau; \xi_\tau)^\top(\mathbf{w}_{\tau+1} - \mathbf{w}) + R(\mathbf{w}_{\tau+1}) - R(\mathbf{w})\\ \leq&\frac{1}{2\eta}\|\mathbf{w} - \mathbf{w}_\tau\|_2^2 - \frac{1}{2\eta}\|\mathbf{w} - \mathbf{w}_{\tau+1}\|_2^2 - \frac{1}{2\eta}\|\mathbf{w}_{\tau+1} - \mathbf{w}_\tau\|_2^2.\end{aligned} \tag{G.1}$$

Then we can lower bound the first term, that is

$$\begin{aligned}
&\partial f(\mathbf{w}_\tau;\xi_\tau)^\top(\mathbf{w}_{\tau+1}-\mathbf{w})\\
=&\partial f(\mathbf{w}_\tau;\xi_\tau)^\top(\mathbf{w}_{\tau+1}-\mathbf{w}_\tau)+\partial f(\mathbf{w}_\tau;\xi_\tau)^\top(\mathbf{w}_\tau-\mathbf{w})\\
=&\partial f(\mathbf{w}_\tau;\xi_\tau)^\top(\mathbf{w}_{\tau+1}-\mathbf{w}_\tau)+[\partial f(\mathbf{w}_\tau;\xi_\tau)-\partial f(\mathbf{w}_\tau)]^\top(\mathbf{w}_\tau-\mathbf{w})\\
&+\partial f(\mathbf{w}_\tau)^\top(\mathbf{w}_\tau-\mathbf{w})\\
\geq&\partial f(\mathbf{w}_\tau;\xi_\tau)^\top(\mathbf{w}_{\tau+1}-\mathbf{w}_\tau)+[\partial f(\mathbf{w}_\tau;\xi_\tau)-\partial f(\mathbf{w}_\tau)]^\top(\mathbf{w}_\tau-\mathbf{w})+f(\mathbf{w}_\tau)-f(\mathbf{w}).
\end{aligned}\tag{G.2}$$

The last inequality uses the convexity of $f(\mathbf{w})$. Plugging inequality (G.2) into (G.1), we get

$$\begin{aligned}
&f(\mathbf{w}_\tau)-f(\mathbf{w})+R(\mathbf{w}_{\tau+1})-R(\mathbf{w})\\
\leq&\frac{1}{2\eta}\|\mathbf{w}-\mathbf{w}_\tau\|_2^2-\frac{1}{2\eta}\|\mathbf{w}-\mathbf{w}_{\tau+1}\|_2^2-\frac{1}{2\eta}\|\mathbf{w}_{\tau+1}-\mathbf{w}_\tau\|_2^2\\
&-\partial f(\mathbf{w}_\tau;\xi_\tau)^\top(\mathbf{w}_{\tau+1}-\mathbf{w}_\tau)+\underbrace{[\partial f(\mathbf{w}_\tau)-\partial f(\mathbf{w}_\tau;\xi_\tau)]^\top(\mathbf{w}_\tau-\mathbf{w})}_{\zeta_\tau(\mathbf{w})}.
\end{aligned}\tag{G.3}$$

On the other hand, by the Cauchy-Shwartz Inequality,

$$\begin{aligned}
-\partial f(\mathbf{w}_\tau;\xi_\tau)^\top(\mathbf{w}_{\tau+1}-\mathbf{w}_\tau)\leq&\frac{1}{2\eta}\|\mathbf{w}_{\tau+1}-\mathbf{w}_\tau\|_2^2+\frac{1}{2}\eta\|\partial f(\mathbf{w}_\tau;\xi_\tau)\|_2^2\\
\leq&\frac{1}{2\eta}\|\mathbf{w}_{\tau+1}-\mathbf{w}_\tau\|_2^2+\frac{1}{2}\eta G^2.
\end{aligned}\tag{G.4}$$

Combining inequalities (G.3) and (G.4) it will have

$$\begin{aligned}
&f(\mathbf{w}_\tau)+R(\mathbf{w}_{\tau+1})-f(\mathbf{w})-R(\mathbf{w})\\
\leq&\frac{1}{2\eta}\|\mathbf{w}-\mathbf{w}_\tau\|_2^2-\frac{1}{2\eta}\|\mathbf{w}-\mathbf{w}_{\tau+1}\|_2^2+\frac{1}{2}\eta G^2+\zeta_\tau(\mathbf{w}).
\end{aligned}$$

Taking summation over τ from 1 to t and dividing by t on both sides of the inequality, then

$$\begin{aligned}
&\frac{1}{t}\sum_{\tau=1}^{t}F(\mathbf{w}_\tau)-F(\mathbf{w})\\
\leq&\frac{1}{t}(R(\mathbf{w}_1)-R(\mathbf{w}_{t+1}))+\frac{1}{2\eta t}\|\mathbf{w}-\mathbf{w}_1\|_2^2+\frac{\eta G^2}{2}+\frac{1}{t}\sum_{\tau=1}^{t}\zeta_\tau(\mathbf{w}).
\end{aligned}$$

Since $\|\partial R(\mathbf{w})\|_2\leq\rho$ and the convexity of $F(\mathbf{w})$, let $\mathbf{w}=\mathbf{w}_{1,\epsilon}^\dagger$, then we get

$$\begin{aligned}
F(\widehat{\mathbf{w}}_t)-F(\mathbf{w}_{1,\epsilon}^\dagger)\leq&\frac{\rho\|\mathbf{w}_1-\mathbf{w}_{t+1}\|_2}{t}+\frac{\|\mathbf{w}_1-\mathbf{w}_{1,\epsilon}^\dagger\|_2^2}{2\eta t}+\frac{\eta G^2}{2}+\frac{1}{t}\sum_{\tau=1}^{t}\zeta_\tau(\mathbf{w}_{1,\epsilon}^\dagger)\\
\leq&\frac{\rho D}{t}+\frac{\|\mathbf{w}_1-\mathbf{w}_{1,\epsilon}^\dagger\|_2^2}{2\eta t}+\frac{\eta G^2}{2}+\frac{1}{t}\sum_{\tau=1}^{t}\zeta_\tau(\mathbf{w}_{1,\epsilon}^\dagger).
\end{aligned}\tag{G.5}$$

Next, we will use the Lemma Appendix G.1 of martingale inequality to upper bound $\sum_{\tau=1}^{t}\zeta_\tau(\mathbf{w}_{1,\epsilon}^{\dagger})$ with a high probability. By using the Jensen's inequality, we have $\|\partial f(\mathbf{w}_\tau)\|_2 = \|\mathrm{E}[\partial f(\mathbf{w}_\tau;\xi_\tau)]\|_2 \le \mathrm{E}[\|\partial f(\mathbf{w}_\tau;\xi_\tau)\|_2] \le G$. Let's denote $X_\tau = \zeta_\tau(\mathbf{w}_{1,\epsilon}^{\dagger}) = [\partial f(\mathbf{w}_\tau) - \partial f(\mathbf{w}_\tau;\xi_\tau)]^\top(\mathbf{w}_\tau - \mathbf{w}_{1,\epsilon}^{\dagger})$, then $\mathrm{E}\,[X_\tau]=0$ and

$$\begin{aligned}|X_\tau| \le& \|\partial f(\mathbf{w}_\tau) - \partial f(\mathbf{w}_\tau;\xi_\tau)\|_2\|\mathbf{w}_\tau - \mathbf{w}_{1,\epsilon}^{\dagger}\|_2\\ \le& \left(\|\partial f(\mathbf{w}_\tau)\|_2 + \|\partial f(\mathbf{w}_\tau;\xi_\tau\|_2\right)\left(\|\mathbf{w}_\tau - \mathbf{w}_1\|_2 + \|\mathbf{w}_1 - \mathbf{w}_{1,\epsilon}^{\dagger}\|_2\right) \le 4GD,\end{aligned}$$

where we use the fact that $\mathbf{w}_\tau \in \mathcal{B}(\mathbf{w}_1, D)$ and $\|\mathbf{w}_1 - \mathbf{w}_{1,\epsilon}\|_2 \le D$. This implies that

$$\mathrm{E}\left[\exp\left(\frac{X_\tau^2}{16G^2D^2}\right)\right] \le \exp(1).$$

Then with probability at least $1-\delta$, we have

$$\sum_{\tau=1}^{t} X_\tau \le \sqrt{3\log(1/\delta)\sum_{\tau=1}^{t}16G^2D^2} = 4GD\sqrt{3\log(1/\delta)t}. \tag{G.6}$$

We complete the proof by combining (G.5) and (G.6). □

Appendix H. Proof of Theorem 6.1

Proof. This proof is similar to that of Theorem 4.1. Let $\mathbf{w}_{k,\epsilon}^{\dagger}$ denote the closest point to $\mathbf{w}_k$ in $\mathcal{S}_\epsilon$. Define $\epsilon_k = \frac{\epsilon_0}{2^k}$. Note that $D_k = \frac{D_1}{2^{k-1}} \ge \frac{c\epsilon_{k-1}}{\epsilon^{1-\theta}}$ and $\eta_k = \frac{\epsilon_{k-1}}{4G^2}$. We will show by induction that $F(\mathbf{w}_k) - F_* \le \epsilon_k + \epsilon$ for $k = 0, 1, \dots$ with a high probability, which leads to our conclusion when $k = K$. The inequality holds obviously for $k = 0$. Conditioned on $F(\mathbf{w}_{k-1}) - F_* \le \epsilon_{k-1} + \epsilon$, we will show that $F(\mathbf{w}_k) - F_* \le \epsilon_k + \epsilon$ with a high probability. By Lemma 3.1, we have

$$\|\mathbf{w}_{k-1,\epsilon}^{\dagger} - \mathbf{w}_{k-1}\|_2 \le \frac{c\epsilon_{k-1}}{\epsilon^{1-\theta}} \le D_k. \tag{H.1}$$

We apply Lemma 6.1 to the k-th stage of Algorithm 1 conditioned on randomness in previous stages. With a probability $1-\tilde{\delta}$ we have

$$F(\mathbf{w}_k) - F(\mathbf{w}_{k-1,\epsilon}^{\dagger}) \le \frac{\rho D_k}{t} + \frac{\eta_k G^2}{2} + \frac{\|\mathbf{w}_{k-1} - \mathbf{w}_{k-1,\epsilon}^{\dagger}\|_2^2}{2\eta_k t} + \frac{4GD_k\sqrt{3\log(1/\tilde{\delta})}}{\sqrt{t}}. \tag{H.2}$$

We now consider two cases for $\mathbf{w}_{k-1}$. First, we assume $F(\mathbf{w}_{k-1}) - F_* \le \epsilon$, i.e. $\mathbf{w}_{k-1} \in \mathcal{S}_\epsilon$. Then we have $\mathbf{w}_{k-1,\epsilon}^{\dagger} = \mathbf{w}_{k-1}$ and

$$F(\mathbf{w}_k) - F(\mathbf{w}_{k-1,\epsilon}^{\dagger}) \le \frac{\rho D_k}{t} + \frac{\eta_k G^2}{2} + \frac{4GD_k\sqrt{3\log(1/\tilde{\delta})}}{\sqrt{t}} \le \frac{\epsilon_k}{4} + \frac{\epsilon_k}{4} + \frac{\epsilon_{k-1}}{8} = \frac{3\epsilon_k}{4}.$$

The second inequality using the fact that $\eta_k = \frac{\epsilon_k}{2G^2}$, $t \geq 3072\log(1/\tilde{\delta})\frac{G^2D_1^2}{\epsilon_0^2}$ and $t \geq \frac{8\rho D_1}{\epsilon_0}$. As a result,

$$F(\mathbf{w}_k) - F_* \leq F(\mathbf{w}_{k-1,\epsilon}^{\dagger}) - F_* + \frac{3\epsilon_k}{4} \leq \epsilon + \epsilon_k.$$

Next, we consider $F(\mathbf{w}_{k-1}) - F_* > \epsilon$, i.e. $\mathbf{w}_{k-1} \notin \mathcal{S}_\epsilon$. Then we have $F(\mathbf{w}_{k-1,\epsilon}^{\dagger}) - F_* = \epsilon$. Combining (H.1) and (H.2), we get

$$F(\mathbf{w}_k) - F(\mathbf{w}_{k-1,\epsilon}^{\dagger}) \leq \frac{\rho D_k}{t} + \frac{\eta_k G^2}{2} + \frac{D_k^2}{2\eta_k t} + \frac{4GD_k\sqrt{3\log(1/\tilde{\delta})}}{\sqrt{t}}.$$

Since $\eta_k = \frac{\epsilon_k}{2G^2}$ and $t \geq \max\left\{\max\{16, 3072\log(1/\tilde{\delta})\}\frac{G^2D_1^2}{\epsilon_0^2}, \frac{8\rho D_1}{\epsilon_0}\right\}$, we have

$$\begin{aligned}
&\frac{\rho D_k}{t} \leq \frac{\rho D_k \epsilon_0}{8\rho D_1} = \frac{\epsilon_k}{4},\\
&\frac{\eta_k G^2}{2} = \frac{\epsilon_k}{4},\\
&\frac{D_k^2}{2\eta_k t} \leq \frac{(D_1/2^{k-1})^2}{2\epsilon_k/(2G^2)}\frac{\epsilon_0^2}{16G^2D_1^2} = \frac{\epsilon_k}{4},\\
&\frac{4GD_k\sqrt{3\log(1/\tilde{\delta})}}{\sqrt{t}} \leq \frac{4G(D_1/2^{k-1})\sqrt{3\log(1/\tilde{\delta})}\epsilon_0}{GD_1\sqrt{3072\log(1/\tilde{\delta})}} = \frac{\epsilon_k}{4}.
\end{aligned}$$

As a result,

$$F(\mathbf{w}_k) - F(\mathbf{w}_{k-1,\epsilon}^{\dagger}) \leq \epsilon_k \Rightarrow F(\mathbf{w}_k) - F_* \leq \epsilon_k + \epsilon.$$

with a probability $1 - \tilde{\delta}$. Therefore by induction, with a probability at least $(1-\tilde{\delta})^K$ we have,

$$F(\mathbf{w}_K) - F_* \leq \epsilon_K + \epsilon \leq 2\epsilon.$$

Since $\tilde{\delta} = \delta/K$, then $(1-\tilde{\delta})^K \geq 1 - \delta$ and we complete the proof. □

Appendix I. Proximal ASSG based on the regularized variant

In this section, we will present a proximal ASSG based on the regularized variant, which is referred to ProxASSG-r. Similar to ASSG-r, we construct a new problem by adding a strongly convex term $\frac{1}{2\beta}\|\mathbf{w} - \mathbf{w}_1\|_2^2$ to the original problem (6.1):

$$\min_{\mathbf{w}\in\mathbb{R}^d} \widehat{F}(\mathbf{w}) = F(\mathbf{w}) + \frac{1}{2\beta}\|\mathbf{w} - \mathbf{w}_1\|_2^2, \tag{I.1}$$

where $F(\mathbf{w})$ is defined in (6.1). We denote $\widehat{\mathbf{w}}_*$ the optimal solution to problem (I.1) given the regularization reference point $\mathbf{w}_1$. We first extend SSGS to its proximal version as presented in Algorithm 6. To give the convergence analysis of ProxASSG-r for solving (6.1), we first present a lemma below to bound $\|\widehat{\mathbf{w}}_* - \mathbf{w}_t\|_2$ and $\|\mathbf{w}_t - \mathbf{w}_1\|_2$.

Algorithm 6 Proxmal SSG for solving (6.1) with a Strongly convex regularizer: ProxSSGS($\mathbf{w}_1, \beta, T$)

```
for $t = 1, \ldots, T$ do
    Compute $\mathbf{w}_{t+1} = \mathrm{Prox}_{\mathbb{R}^d}^{2\beta/t, R}\left[(1 - \frac{2}{t})\mathbf{w}_t + \frac{2}{t}\mathbf{w}_1 - \frac{2\beta}{t}\partial f(\mathbf{w}_t; \xi_t)\right]$
end for
Output: $\widehat{\mathbf{w}}_T = \sum_{t=1}^{T} \mathbf{w}_t / T$
```

Lemma Appendix I.1. *For any $t \geq 1$, we have $\|\widehat{\mathbf{w}}_* - \mathbf{w}_t\|_2 \leq 3\beta(G + \rho)$ and $\|\mathbf{w}_t - \mathbf{w}_1\|_2 \leq 2\beta(G + \rho)$.*

Proof. By the optimality of $\widehat{\mathbf{w}}_*$, we have for any $\mathbf{w} \in \mathbb{R}^d$

$$(\partial F(\widehat{\mathbf{w}}_*) + (\widehat{\mathbf{w}}_* - \mathbf{w}_1)/\beta)^\top (\mathbf{w} - \widehat{\mathbf{w}}_*) \geq 0.$$

Let $\mathbf{w} = \mathbf{w}_1$, we have

$$\partial F(\widehat{\mathbf{w}}_*)^\top (\mathbf{w}_1 - \widehat{\mathbf{w}}_*) \geq \frac{\|\widehat{\mathbf{w}}_* - \mathbf{w}_1\|_2^2}{\beta}.$$

We have $\|\partial F(\widehat{\mathbf{w}}_*)\|_2 \leq G + \rho$ due to $\|\partial g(\mathbf{w}; \xi)\|_2 \leq G$ and $\|\partial R(\mathbf{w})\|_2 \leq \rho$, then

$$\|\widehat{\mathbf{w}}_* - \mathbf{w}_1\|_2 \leq \beta(G + \rho).$$

Next, we bound $\|\mathbf{w}_t - \mathbf{w}_1\|_2$. According to the update of $\mathbf{w}_{t+1}$, there exists a subgradient $\partial R(\mathbf{w}_{t+1})$ such that

$$\mathbf{w}_{t+1} - \left(\mathbf{w}_t - \eta_t[\partial f(\mathbf{w}_t; \xi_t) + \frac{1}{\beta}(\mathbf{w}_t - \mathbf{w}_1)]\right) + \eta_t \partial R(\mathbf{w}_{t+1}) = 0,$$

where $\eta_t = \frac{2\beta}{t}$. Thus,

$$\|\mathbf{w}_{t+1} - \mathbf{w}_1\|_2 = \| - \eta_t(\partial f(\mathbf{w}_t; \xi_t) + \partial R(\mathbf{w}_{t+1})) + (1 - \eta_t/\beta)(\mathbf{w}_t - \mathbf{w}_1)\|_2.$$

We prove $\|\mathbf{w}_t - \mathbf{w}_1\|_2 \leq 2\beta(G + \rho)$ by induction. First, we consider $t = 1$, where $\eta_t = 2\beta$, then

$$\|\mathbf{w}_2 - \mathbf{w}_1\|_2 = \|2\beta(\partial f(\mathbf{w}_t; \xi_t) + \partial R(\mathbf{w}_{t+1}))\|_2 \leq 2\beta(G + \rho).$$

Then we consider any $t \geq 2$, where $\frac{\eta_t}{\beta} \leq 1$. Then

$$\begin{aligned}\|\mathbf{w}_{t+1} - \mathbf{w}_1\|_2 =& \left\| - \frac{\eta_t}{\beta}\beta(\partial f(\mathbf{w}_t; \xi_t) + \partial R(\mathbf{w}_{t+1})) + (1 - \frac{\eta_t}{\beta})(\mathbf{w}_t - \mathbf{w}_1)\right\|_2 \\ \leq& \frac{\eta_t}{\beta}\beta(G + \rho) + (1 - \frac{\eta_t}{\beta})2\beta(G + \rho) \leq 2\beta(G + \rho).\end{aligned}$$

Therefore

$$\|\widehat{\mathbf{w}}_* - \mathbf{w}_t\|_2 \leq 3\beta(G + \rho). \qquad \square$$

Next, we present a high probability convergence bound of ProxSSGS for optimizing $\widehat{F}(\mathbf{w})$.

Theorem Appendix I.1. *Suppose Assumption 3.2.c holds. Let $\widehat{\mathbf{w}}_T$ be the returned solution of Algorithm 6. Given $\mathbf{w}_1 \in \mathbb{R}^d$, $\delta < 1/e$ and $T \geq 3$, with a high probability $1-\delta$ we have*

$$F(\widehat{\mathbf{w}}_T) - F(\mathbf{w}) \leq \frac{1}{2\beta}\|\mathbf{w}-\mathbf{w}_1\|_2^2 + \frac{34\beta(G+\rho)^2(1+\log T + \log(4\log T/\delta))}{T}.$$

Proof. Based on the update of $\mathbf{w}_{t+1}$ and the fact that $\frac{1}{2}\|\mathbf{w}-\mathbf{w}_t\|_2^2+\eta_t[\partial f(\mathbf{w}_t;\xi_t)+(\mathbf{w}_t-\mathbf{w}_1)/\beta]^\top\mathbf{w}+\eta_t R(\mathbf{w})$ is $\frac{1}{2}$-stongly convex in terms of $\mathbf{w}$, then for any $\mathbf{w}\in\mathbb{R}^d$, we have

$$\begin{aligned}&\frac{1}{2}\|\mathbf{w}_{t+1}-\mathbf{w}_t\|_2^2 + \eta_t[\partial f(\mathbf{w}_t;\xi_t)+(\mathbf{w}_t-\mathbf{w}_1)/\beta]^\top\mathbf{w}_{t+1} + \eta_t R(\mathbf{w}_{t+1})\\ \leq&\frac{1}{2}\|\mathbf{w}-\mathbf{w}_t\|_2^2 + \eta_t[\partial f(\mathbf{w}_t;\xi_t)+(\mathbf{w}_t-\mathbf{w}_1)/\beta]^\top\mathbf{w} + \eta_t R(\mathbf{w}) - \frac{1}{2}\|\mathbf{w}-\mathbf{w}_{t+1}\|_2^2.\end{aligned}$$

Rearranging the inequality gives

$$\begin{aligned}&\eta_t[\partial f(\mathbf{w}_t)+(\mathbf{w}_t-\mathbf{w}_1)/\beta]^\top(\mathbf{w}_{t+1}-\mathbf{w}) + \eta_t(R(\mathbf{w}_{t+1})-R(\mathbf{w}))\\ \leq&\frac{1}{2}\|\mathbf{w}-\mathbf{w}_t\|_2^2 - \frac{1}{2}\|\mathbf{w}-\mathbf{w}_{t+1}\|_2^2 - \frac{1}{2}\|\mathbf{w}_{t+1}-\mathbf{w}_t\|_2^2\\ &+\eta_t[\partial f(\mathbf{w}_t)-\partial f(\mathbf{w}_t;\xi_t)]^\top(\mathbf{w}_{t+1}-\mathbf{w}).\end{aligned}\tag{I.2}$$

By the strong convexity of $f(\mathbf{w})+\frac{1}{2\beta}\|\mathbf{w}-\mathbf{w}_1\|_2^2$, we have

$$\begin{aligned}&\eta_t[\partial f(\mathbf{w}_t)+(\mathbf{w}_t-\mathbf{w}_1)/\beta]^\top(\mathbf{w}_{t+1}-\mathbf{w})\\ =&\eta_t[\partial f(\mathbf{w}_t)+(\mathbf{w}_t-\mathbf{w}_1)/\beta]^\top(\mathbf{w}_t-\mathbf{w}) + \eta_t[\partial f(\mathbf{w}_t)+(\mathbf{w}_t-\mathbf{w}_1)/\beta]^\top(\mathbf{w}_{t+1}-\mathbf{w}_t)\\ \geq&\eta_t\left[f(\mathbf{w}_t)+\frac{1}{2\beta}\|\mathbf{w}_t-\mathbf{w}_1\|_2^2\right] - \eta_t\left[f(\mathbf{w})+\frac{1}{2\beta}\|\mathbf{w}-\mathbf{w}_1\|_2^2)\right] + \frac{\eta_t}{2\beta}\|\mathbf{w}_t-\mathbf{w}\|_2^2\\ &+\eta_t[\partial f(\mathbf{w}_t)+(\mathbf{w}_t-\mathbf{w}_1)/\beta]^\top(\mathbf{w}_{t+1}-\mathbf{w}_t).\end{aligned}\tag{I.3}$$

Plugging inequality (I.3) into (I.2), we get

$$\begin{aligned}&\eta_t\left[f(\mathbf{w}_t)+\frac{1}{2\beta}\|\mathbf{w}_t-\mathbf{w}_1\|_2^2\right] - \eta_t\left[f(\mathbf{w})+\frac{1}{2\beta}\|\mathbf{w}-\mathbf{w}_1\|_2^2\right] + \eta_t(R(\mathbf{w}_{t+1})-R(\mathbf{w}))\\ \leq&\frac{1}{2}\|\mathbf{w}-\mathbf{w}_t\|_2^2 - \frac{1}{2}\|\mathbf{w}-\mathbf{w}_{t+1}\|_2^2 + \eta_t[\partial f(\mathbf{w}_t)-\partial f(\mathbf{w}_t;\xi_t)]^\top(\mathbf{w}_{t+1}-\mathbf{w})\\ &-\frac{1}{2}\|\mathbf{w}_{t+1}-\mathbf{w}_t\|_2^2 - \frac{\eta_t}{2\beta}\|\mathbf{w}_t-\mathbf{w}\|_2^2 - \eta_t[\partial f(\mathbf{w}_t)+(\mathbf{w}_t-\mathbf{w}_1)/\beta]^\top(\mathbf{w}_{t+1}-\mathbf{w}_t).\end{aligned}\tag{I.4}$$

On the other hand, by the Cauchy-Shwartz inequality we have

$$\begin{aligned}&-\eta_t[\partial f(\mathbf{w}_t)+(\mathbf{w}_t-\mathbf{w}_1)/\beta]^\top(\mathbf{w}_{t+1}-\mathbf{w}_t)\\ \leq&\frac{1}{4}\|\mathbf{w}_{t+1}-\mathbf{w}_t\|_2^2 + \eta_t^2\|\partial f(\mathbf{w}_t)+(\mathbf{w}_t-\mathbf{w}_1)/\beta\|_2^2\\ \leq&\frac{1}{4}\|\mathbf{w}_{t+1}-\mathbf{w}_t\|_2^2 + 2\eta_t^2[G^2+4(G+\rho)^2]\end{aligned}\tag{I.5}$$

and

$$\begin{aligned}
&\eta_t[\partial f(\mathbf{w}_t) - \partial f(\mathbf{w}_t;\xi_t)]^\top(\mathbf{w}_{t+1} - \mathbf{w})\\
=&\eta_t[\partial f(\mathbf{w}_t) - \partial f(\mathbf{w}_t;\xi_t)]^\top(\mathbf{w}_t - \mathbf{w}) + \eta_t[\partial f(\mathbf{w}_t) - \partial f(\mathbf{w}_t;\xi_t)]^\top(\mathbf{w}_{t+1} - \mathbf{w}_t)\\
\leq&\eta_t[\partial f(\mathbf{w}_t) - \partial f(\mathbf{w}_t;\xi_t)]^\top(\mathbf{w}_t - \mathbf{w}) + \frac{1}{4}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2 + \eta_t^2\|\partial f(\mathbf{w}_t) - \partial f(\mathbf{w}_t;\xi_t)\|_2^2\\
\leq&\eta_t[\partial f(\mathbf{w}_t) - \partial f(\mathbf{w}_t;\xi_t)]^\top(\mathbf{w}_t - \mathbf{w}) + \frac{1}{4}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2 + 4\eta_t^2 G^2. \qquad \text{(I.6)}
\end{aligned}$$

Plugging inequalities (I.5) and (I.6) into inequality (I.4), we get

$$\begin{aligned}
&\left[f(\mathbf{w}_t) + R(\mathbf{w}_{t+1}) + \frac{1}{2\beta}\|\mathbf{w}_t - \mathbf{w}_1\|_2^2\right] - \left[f(\mathbf{w}) + R(\mathbf{w}) + \frac{1}{2\beta}\|\mathbf{w} - \mathbf{w}_1\|_2^2\right]\\
\leq&\frac{1}{2\eta_t}\|\mathbf{w} - \mathbf{w}_t\|_2^2 - \frac{1}{2\eta_t}\|\mathbf{w} - \mathbf{w}_{t+1}\|_2^2 + \underbrace{[\partial f(\mathbf{w}_t) - \partial f(\mathbf{w}_t;\xi_t)]^\top(\mathbf{w}_t - \mathbf{w})}_{\zeta_t(\mathbf{w})}\\
&-\frac{1}{2\beta}\|\mathbf{w} - \mathbf{w}_t\|_2^2 + 2\eta_t[3G^2 + 4(G+\rho)^2].
\end{aligned}$$

By summing the above inequalities across $t = 1, \ldots, T$ and setting $\mathbf{w} = \widehat{\mathbf{w}}_*$, we have

$$\begin{aligned}
&\sum_{t=1}^T(\widehat{F}(\mathbf{w}_t) - \widehat{F}(\widehat{\mathbf{w}}_*))\\
\leq&\sum_{t=1}^{T-1}\frac{1}{2}\left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} - \frac{1}{2\beta}\right)\|\widehat{\mathbf{w}}_* - \mathbf{w}_{t+1}\|_2^2 + \sum_{t=1}^T\zeta_t(\widehat{\mathbf{w}}_*) - \frac{1}{4\beta}\sum_{t=1}^T\|\widehat{\mathbf{w}}_* - \mathbf{w}_t\|_2^2\\
&-\frac{1}{4\beta}\|\widehat{\mathbf{w}}_* - \mathbf{w}_1\|_2^2 + \frac{1}{2\eta_1}\|\widehat{\mathbf{w}}_* - \mathbf{w}_1\|_2^2 + (6G^2 + 8(G+\rho)^2)\sum_{t=1}^T\eta_t + R(\mathbf{w}_1) - R(\mathbf{w}_{T+1})\\
\leq&\sum_{t=1}^T\zeta_t(\widehat{\mathbf{w}}_*) - \frac{1}{4\beta}\sum_{t=1}^T\|\widehat{\mathbf{w}}_* - \mathbf{w}_t\|_2^2 + \beta[(12G^2 + 16(G+\rho)^2)(1+\log T) + 2\rho(G+\rho)],
\end{aligned}$$

where the last inequality uses $\eta_t = \frac{2\beta}{t}$ and $R(\mathbf{w}_1) - R(\mathbf{w}_{T+1}) \leq \rho\|\mathbf{w}_1 - \mathbf{w}_{T+1}\|_2 \leq 2\beta\rho(G+\rho)$. Next, we bound R.H.S of the above inequality. By using Lemma Appendix C.1, we employ the same technique in the proof of Theorem 4.2 to proceed our proof. The only difference is that we set $D = 3\beta(G+\rho)$. We omit the detailed steps but present the key results: with a probability $1-\delta$, we have

$$\begin{aligned}
\sum_{t=1}^T\zeta_t - \frac{1}{4\beta}\sum_{t=1}^T\|\widehat{\mathbf{w}}_* - \mathbf{w}_t\|_2^2 \leq&16\beta G^2\log(4\log T/\delta) + 6GD\log(4\log T/\delta)\\
=&\beta(34G^2 + 18G\rho)\log(4\log T/\delta).
\end{aligned}$$

Algorithm 7 the ProxASSG-r algorithm for solving (6.1)

```
Input: the number of stages $K$ and the number of iterations $t$ per-stage, the initial solution $\mathbf{w}_0 \in \mathcal{K}$, and $\beta_1 \geq \frac{2c^2\epsilon_0}{\epsilon^{2(1-\theta)}}$
for $k = 1, \ldots, K$ do
    Let $\mathbf{w}_k = \text{ProxSSGS}(\mathbf{w}_{k-1}, \beta_k, t)$
    Update $\beta_{k+1} = \beta_k/2$
end for
Output: $\mathbf{w}_K$
```

Thus, with a probability $1-\delta$,

$$\begin{aligned}\widehat{F}(\widehat{\mathbf{w}}_T) - \widehat{F}(\widehat{\mathbf{w}}_*) \leq & \beta\frac{(34G^2 + 18G\rho)\log(4\log T/\delta)}{T}\\ & + \beta\frac{(12G^2 + 16(G+\rho)^2)(1+\log T) + 2\rho(G+\rho)}{T}\\ \leq & \frac{34\beta(G+\rho)^2(1+\log T + \log(4\log T/\delta))}{T}.\end{aligned}$$

We complete the proof by using the facts that $F(\widehat{\mathbf{w}}_T) \leq \widehat{F}(\widehat{\mathbf{w}}_T)$ and $\widehat{F}(\widehat{\mathbf{w}}_*) \leq F(\mathbf{w}) + \frac{\lambda}{2}\|\mathbf{w} - \mathbf{w}_1\|_2^2$. □

Finally, we present ProxASSG-r in Algorithm 7 and its convergence guaratnee is presented in theorem below.

Theorem Appendix I.2. *Suppose Assumptions 3.1 and 3.2 hold for a target $\epsilon \ll 1$. Given $\delta \in (0, 1/e)$, let $\tilde{\delta} = \delta/K$ and $K = \lceil \log_2(\frac{\epsilon_0}{\epsilon}) \rceil$ and t be the smallest integer such that $t \geq \max\{\frac{136\beta_1(G+\rho)^2(1+\log(4\log t/\tilde{\delta}) + \log t)}{\epsilon_0}, 3\}$. Then ProxASSG-r guarantees that, with a probability $1-\delta$,*

$$F(\mathbf{w}_K) - F_* \leq 2\epsilon.$$

As a result, the iteration complexity of ASSG-r for achieving an 2ϵ-optimal solution with a high probability $1-\delta$ is $\widetilde{O}(\log(1/\delta)/\epsilon^{2(1-\theta)})$ provided $\beta_1 = \Omega(\frac{2c^2\epsilon_0}{\epsilon^{2(1-\theta)}})$.

Proof. The proof is the same to the proof of Theorem 4.2 by replacing G by $G+\rho$. □

Appendix J. ASSG for Piecewise Convex Quadratic Minimization

In this section, we develop an ASSG for piecewise convex quadratic minimization under the global error bound condition:

$$dist(\mathbf{w}, \mathcal{K}_*) \leq c[F(\mathbf{w}) - F_* + (F(\mathbf{w}) - F_*)^{1/2}], \quad \forall \mathbf{w} \in \mathbb{R}^d. \tag{J.1}$$

We assume that an upper bound of $c \leq \hat{c}$ is given. Here, we only show the results for the constrained variant of ASSG, which is presented in Algorithm 8. The regularized variant is a simple exercise.

Algorithm 8 the ASSG-c algorithm under the global error bound condition (J.1)

```
Input: the number of stages $K$, the number of iterations $t$ per stage, and the
initial solution $\mathbf{w}_0$, $\eta_1 = \epsilon_0/(3G^2)$ and $\hat{c} \geq c$
for $k = 1, \ldots, K$ do
    Let $\mathbf{w}_1^k = \mathbf{w}_{k-1}$ and $D_k \geq \hat{c}(\epsilon_{k-1} + \sqrt{\epsilon_{k-1}})$
    for $\tau = 1, \ldots, t_k$ do
        Update $\mathbf{w}_{\tau+1}^k = \Pi_{\mathcal{K}\cap\mathcal{B}(\mathbf{w}_{k-1}, D_k)}[\mathbf{w}_\tau^k - \eta_k \partial f(\mathbf{w}_\tau^k; \xi_\tau^k)]$
    end for
    Let $\mathbf{w}_k = \frac{1}{t_k}\sum_{\tau=1}^{t_k} \mathbf{w}_\tau^k$
    Let $\eta_{k+1} = \eta_k/2$
end for
Output: $\mathbf{w}_K$
```

Theorem Appendix J.1. *Suppose Assumption 3.1 holds and $F(\mathbf{w})$ is convex and piecewise convex quadratic function. Given $\delta \in (0,1)$, let $\tilde{\delta} = \delta/K$, $K = \lceil \log_2(\frac{\epsilon_0}{\epsilon}) \rceil$, and t_k be the smallest integer such that $t_k \geq 6912G^2\hat{c}^2\log(1/\tilde{\delta})\max\{1, \frac{1}{\epsilon_k}\}$. Then Algorithm 8 guarantees that, with a probability $1-\delta$,*

$$F(\mathbf{w}_K) - F_* \leq \epsilon.$$

As a result, the iteration complexity of Algorithm 8 for achieving an ϵ-optimal solution with a high probability $1-\delta$ is $O(\log(1/\delta)/\epsilon)$.

Proof. Define $\epsilon_k = \frac{\epsilon_0}{2^k}$. Note that $\eta_k = \frac{\epsilon_{k-1}}{3G^2}$. We will show by induction that $F(\mathbf{w}_k) - F_* \leq \epsilon_k$ for $k = 0, 1, \ldots$ with a high probability, which leads to our conclusion when $k = K$. The inequality holds obviously for $k = 0$. Conditioned on $F(\mathbf{w}_{k-1}) - F_* \leq \epsilon_{k-1}$, we will show that $F(\mathbf{w}_k) - F_* \leq \epsilon_k$ with a high probability. First, we have

$$\begin{aligned}\|\mathbf{w}_{k-1}^* - \mathbf{w}_{k-1}\|_2 \leq & c[F(\mathbf{w}_{k-1}) - F_* + (F(\mathbf{w}_{k-1}) - F_*)^{1/2}] \\ \leq & c(\epsilon_{k-1} + \sqrt{\epsilon_{k-1}}) \leq D_k,\end{aligned}$$

where $\mathbf{w}_{k-1}^* \in \mathcal{K}_*$ is the closest point to $\mathbf{w}_{k-1}$ in the optimal set, the second inequality follows the global error bound (J.1) and the last inequality uses the value of D_k. We apply Lemma 4.1 replacing $\mathbf{w}_{1,\epsilon}^\dagger$ with $\mathbf{w}_1^*$ to the k-th stage of Algorithm 1 conditioned on randomness in previous stages. With a probability $1-\tilde{\delta}$ we have

$$\begin{aligned}F(\mathbf{w}_k) - F_* \leq & \frac{\eta_k G^2}{2} + \frac{\|\mathbf{w}_{k-1} - \mathbf{w}_*\|_2^2}{2\eta_k t_k} + \frac{4GD_k\sqrt{3\log(1/\tilde{\delta})}}{\sqrt{t_k}} \\ \leq & \frac{\eta_k G^2}{2} + \frac{D_k^2}{\eta_k t_k} + \frac{4GD_k\sqrt{3\log(1/\tilde{\delta})}}{\sqrt{t_k}}.\end{aligned}$$

Since $\eta_k = \frac{2\epsilon_k}{3G^2}$ and $t_k \geq 6912G^2\hat{c}^2\log(1/\tilde{\delta})\max\{1, \frac{1}{\epsilon_k}\}$, we can derive that $F(\mathbf{w}_k) - F_* \leq \epsilon_k$ with a probability $1-\tilde{\delta}$. Therefore by induction, with a probability at least

$(1-\tilde{\delta})^K$ we have $F(\mathbf{w}_K) - F_* \leq \epsilon_K \leq \epsilon$. Since $\tilde{\delta} = \delta/K$, then $(1-\tilde{\delta})^K \geq 1-\delta$ and we complete the proof. In fact, the total number of iterations of ASSG-c is bounded by $T = \sum_{k=1}^K t_k \leq O\left(\log(1/\tilde{\delta})\sum_{k=1}^K \frac{1}{\epsilon_k}\right) = O\left(\frac{\log(1/\tilde{\delta})}{\epsilon}\right)$. □

References

[1] Z. Allen-Zhu. How to make the gradients small stochastically: Even faster convex and nonconvex SGD. In *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*, pages 1165–1175, 2018.

[2] Z. Allen-Zhu and Y. Yuan. Improved svrg for non-strongly-convex or sum-of-non-convex objectives. In *International Conference on Machine Learning (ICML)*, pages 1080–1089, 2016.

[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. *Mathematical Programming*, 137(1-2):91–129, 2013.

[4] P. L. Bartlett and M. H. Wegkamp. Classification with a reject option using a hinge loss. *Journal of Machine Learning Research*, 9:1823–1840, 2008.

[5] H. H. Bauschke and P. L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer Publishing Company, Incorporated, 1st edition, 2011.

[6] G. Blanchard and N. Krämer. Convergence rates of kernel conjugate gradient for random design regression. *Analysis and Applications*, 14(06):763–794, 2016.

[7] J. Bolte, A. Daniilidis, and A. Lewis. The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization*, 17:1205–1223, 2006.

[8] J. Bolte, T. P. Nguyen, J. Peypouquet, and B. W. Suter. From error bounds to the complexity of first-order descent methods for convex functions. *Mathematical Programming*, 165(2):471–507, 2017.

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

[10] A. Defazio, F. R. Bach, and S. Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1646–1654, 2014.

[11] J. C. Duchi, S. Shalev-Shwartz, Y. Singer, and A. Tewari. Composite objective mirror descent. In *Proceedings of the 23rd Annual Conference on Learning Theory (COLT)*, pages 14–26, 2010.

[12] Q. Fang, M. Xu, and Y. Ying. Faster convergence of a randomized coordinate descent method for linearly constrained optimization problems. *Analysis and Applications*, 16(05):741–755, 2018.

[13] D. J. Foster, A. Sekhari, O. Shamir, N. Srebro, K. Sridharan, and B. E. Woodworth. The complexity of making the gradient small in stochastic convex optimization. *arXiv preprint arXiv:1902.04686*, 2019.

[14] 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(4):20612089, 2013.

[15] R. Goebel and R. T. Rockafellar. Local strong convexity and local lipschitz continuity of the gradient of convex functions. *Journal of Convex Analysis*, 15(2):263, 2008.

[16] P. Gong and J. Ye. Linear convergence of variance-reduced projected stochastic gradient without strong convexity. *arXiv preprint arXiv:1406.1102*, 2014.

[17] Z.-C. Guo, D.-H. Xiang, X. Guo, and D.-X. Zhou. Thresholded spectral algorithms for sparse approximations. *Analysis and Applications*, 15(03):433–455, 2017.

[18] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning.* Springer Series in Statistics. Springer, 2009.

[19] 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)*, pages 421–436, 2011.

[20] 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)*, pages 710–718, 2013.

[21] R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems (NIPS)*, pages 315–323, 2013.

[22] A. Juditsky and Y. Nesterov. Deterministic and stochastic primal-dual subgradient algorithms for uniformly convex minimization. *Stochastic Systems*, 4:44–80, 2014.

[23] S. M. Kakade and A. Tewari. On the generalization ability of online strongly convex programming algorithms. In *Advances in Neural Information Processing Systems (NIPS)*, pages 801–808, 2008.

[24] H. Karimi, J. Nutini, and M. W. Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD)*, pages 795–811, 2016.

[25] G. Lan, A. Nemirovski, and A. Shapiro. Validation analysis of mirror descent stochastic approximation method. *Mathematical programming*, 134(2):425–458, 2012.

[26] G. Li. Global error bounds for piecewise convex polynomials. *Mathematical programming*, 137(1-2):37–64, 2013.

[27] G. Li and T. K. Pong. Calculus of the exponent of Kurdyka-Łojasiewicz inequality and its applications to linear convergence of first-order methods. *Foundations of Computational Mathematics*, pages 1–34, 2017.

[28] J. Liu and S. J. Wright. Asynchronous stochastic coordinate descent: Parallelism and convergence properties. *SIAM Journal on Optimization*, 25:351–376, 2015.

[29] J. Liu, S. J. Wright, C. Ré, V. Bittorf, and S. Sridhar. An asynchronous parallel stochastic coordinate descent algorithm. *Journal Machine Learning Research*, 16:285–322, 2015.

[30] M. Liu and T. Yang. Adaptive accelerated gradient converging method under holderian error bound condition. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3107–3117, 2017.

[31] Z.-Q. Luo and P. Tseng. On the convergence of coordinate descent method for convex differentiable minization. *Journal of Optimization Theory and Applications*, 72(1):7–35, 1992.

[32] 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(2):408–425, 1992.

[33] Z.-Q. Luo and P. Tseng. Error bounds and convergence analysis of feasible descent methods: a general approach. *Annals of Operations Research*, 46:157–178, 1993.

[34] I. Necoara, Y. Nesterov, and F. Glineur. Linear convergence of first order methods for non-strongly convex optimization. *Mathematical Programming*, pages 1–39, 2016.

[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.

[36] Y. Nesterov. *Introductory lectures on convex optimization : a basic course*. Applied optimization. Kluwer Academic Publ., 2004.

[37] H. Nyquist. The optimal lp norm estimator in linear regression models. *Communications in Statistics - Theory and Methods*, 12(21):2511–2524, 1983.

[38] C. Qu, H. Xu, and C. J. Ong. Fast rate analysis of some stochastic optimization algorithms. In *International Conference on Machine Learning (ICML)*, pages 662–670, 2016.

[39] A. Rakhlin, O. Shamir, and K. Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *International Conference on Machine Learning (ICML)*, pages 1571–1578, 2012.

[40] A. Ramdas and A. Singh. Optimal rates for stochastic convex optimization under Tsybakov noise condition. In *International Conference on Machine Learning (ICML)*, pages 365–373, 2013.

[41] R. T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14:877–898, 1976.

[42] L. Rosasco, E. De Vito, A. Caponnetto, M. Piana, and A. Verri. Are loss functions all the same? *Neural Computation*, 16(5):1063–1076, 2004.

[43] N. L. Roux, M. W. Schmidt, and F. Bach. A stochastic gradient method with an exponential convergence rate for finite training sets. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2672–2680, 2012.

[44] V. N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, 1998.

[45] P. Wang and C. Lin. Iteration complexity of feasible descent methods for convex optimization. *Journal of Machine Learning Research*, 15(1):1523–1548, 2014.

[46] L. Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 11(Oct):2543–2596, 2010.

[47] L. Xiao and T. Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, 2014.

[48] Y. Xu, Q. Lin, and T. Yang. Adaptive svrg methods under error bound conditions with unknown growth parameter. In *Advances In Neural Information Processing Systems 30 (NIPS)*, pages 3279–3289, 2017.

[49] Y. Xu, Y. Yan, Q. Lin, and T. Yang. Homotopy smoothing for non-smooth problems with lower complexity than $O(1/\epsilon)$. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1208–1216, 2016.

[50] T. Yang and Q. Lin. Rsg: Beating subgradient method without smoothness and strong convexity. *The Journal of Machine Learning Research*, 19(1):236–268, 2018.

[51] T. Yang, M. Mahdavi, R. Jin, and S. Zhu. An efficient primal dual prox method for non-smooth optimization. *Machine Learning*, 98(3):369–406, 2015.

[52] O. Zadorozhnyi, G. Benecke, S. Mandt, T. Scheffer, and M. Kloft. Huber-norm regularization for linear prediction models. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD)*, pages 714–730, 2016.

[53] H. Zhang. New analysis of linear convergence of gradient-type methods via unifying error bound conditions. *Mathematical Programming*, pages 1–46, 2016.

[54] H. Zhang. The restricted strong convexity revisited: analysis of equivalence to error bound and quadratic growth. *Optimization Letters*, 11(4):817–833, 2017.

[55] H. Zhang and W. Yin. Gradient methods for convex minimization: better rates under weaker conditions. *arXiv preprint arXiv:1303.4645*, 2013.

[56] L. Zhang, M. Mahdavi, and R. Jin. Linear convergence with condition number independent access of full gradients. In *Advances in Neural Information Processing Systems (NIPS)*, pages 980–988, 2013.

[57] Z. Zhou and A. M.-C. So. A unified approach to error bounds for structured convex optimization problems. *Mathematical Programming*, 165(2):689–728, 2017.

[58] Z. Zhou, Q. Zhang, and A. M. So. L1p-norm regularization: Error bounds and convergence rate analysis of first-order methods. In *International Conference on Machine Learning (ICML)*, pages 1501–1510, 2015.

[59] Y. Zhu, S. Chatterjee, J. C. Duchi, and J. D. Lafferty. Local minimax complexity of stochastic convex optimization. In *Advances In Neural Information Processing Systems (NIPS)*, pages 3423–3431, 2016.