Stochastic Convex Optimization: Faster Local Growth Implies Faster Global Convergence

Yi Xu [1] **Qihang Lin** [2] **Tianbao Yang** [1]

Abstract

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 include 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. For example, when applied to the ℓ_1 regularized empirical polyhedral loss minimization (e.g., hinge loss, absolute loss), the proposed stochastic methods have a logarithmic iteration complexity.

[1]Department of Computer Science, The University of Iowa, Iowa City, IA 52242, USA [2]Department of Management Sciences, The University of Iowa, Iowa City, IA 52242, USA. Correspondence to: Tianbao Yang <tianbao-yang@uiowa.edu>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

1. Introduction

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

$$\min_{\mathbf{w}\in\mathcal{K}} F(\mathbf{w}) \triangleq \mathrm{E}_\xi[f(\mathbf{w};\xi)], \tag{1}$$

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) and F_* denote the optimal value.

Traditional stochastic subgradient (SSG) method updates the solution according to

$$\mathbf{w}_{t+1} = \Pi_{\mathcal{K}}[\mathbf{w}_t - \eta_t \partial f(\mathbf{w}_t;\xi_t)], \tag{2}$$

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 $t = 1, \ldots, T$ [1], and by setting the step size $\eta_t = \frac{B}{G\sqrt{T}}$ in (2), with a high probability $1 - \delta$ we have

$$F(\widehat{\mathbf{w}}_T) - F_* \leq O\left(GB(1 + \sqrt{\log(1/\delta)})/\sqrt{T}\right), \tag{3}$$

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 a $\mathbf{w}$ 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 (Roux et al., 2012; Zhang et al., 2013; Johnson & Zhang, 2013; Xiao & Zhang, 2014; Defazio et al.,

[1]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 $\mathcal{K} \cap \mathcal{B}(\mathbf{w}_1, B/2)$.

2014). 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 (Hazan & Kale, 2011; Juditsky & Nesterov, 2014) or a global growth condition (Ramdas & Singh, 2013) to control the distance of intermediate solutions to the *optimal* solution 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.

2. Related Work

The most similar work to the present one is (Ramdas & Singh, 2013), 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 (Ramdas & Singh, 2013) 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. Juditsky & Nesterov (2014) presented primal-dual subgradient and stochastic subgradient methods for solving problems under the uniform convexity assumption (see the definition under Observation 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 (Juditsky & Nesterov, 2014).

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 (Necoara et al., 2015; Liu et al., 2015; Zhang & Yin, 2013; Liu & Wright, 2015; Gong & Ye, 2014; Karimi et al., 2016; Zhang, 2016; Qu et al., 2016; Wang & Lin, 2014). Several recent works (Necoara et al., 2015; Karimi et al., 2016; Zhang, 2016) 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 (Qu et al., 2016).

Luo & Tseng (1992a;b; 1993) 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 (Hou et al., 2013; Zhou et al., 2015; Zhou & So, 2015) 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. Yang & Lin (2016); Xu et al. (2016) 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 (Nemirovsky A.S. & Yudin, 1983; Nesterov, 2004). 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 1(b)).

3. Preliminaries

Recall the notations $\mathcal{K}_*$ and F_* that denote the optimal set of (1) and the optimal value, respectively. For the optimization problem in (1), we make the following assumption throughout the paper.

Assumption 1. *For a stochastic optimization problem (1), we assume*

(a) there exist $\mathbf{w}_0 \in \mathcal{K}$ *and* $\epsilon_0 \geq 0$ *such that* $F(\mathbf{w}_0) - F_* \leq \epsilon_0$*;*

(b) $\mathcal{K}_*$ *is a non-empty compact set;*

(c) There exists a constant G *such that* $\|\partial f(\mathbf{w}; \xi)\|_2 \leq G$.

Remark: (a) 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). (b) simply assumes the optimal set is closed and bounded. This is a relaxed condition in contrast with most previous work that assume the domain $\mathcal{K}$ is bounded. Even if $\mathcal{K}$ is unbounded, as long as the function is a proper lower-semicontinuous convex and coercive function defined on a finite dimensional space, $\mathcal{K}_*$ is nonempty and compact (Bolte et al., 2015). Note that any norm-regularized loss function minimization problem on a finite dimensional space in machine learning satisfy this property. (c) is a standard assumption also made in many previous stochastic gradient-based methods. By Jensen's inequality, we also have $\|\partial F(\mathbf{w})\|_2 \leq G$.

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., $\mathcal{L}_\epsilon = \{\mathbf{w} \in \mathcal{K} : F(\mathbf{w}) = F_* + \epsilon\}$, $\mathcal{S}_\epsilon = \{\mathbf{w} \in \mathcal{K} : F(\mathbf{w}) \leq F_* + \epsilon\}$. Given $\mathcal{K}_*$ is bounded, it follows from (Rockafellar, 1970, Corollary 8.7.1) that the sublevel set $\mathcal{S}_\epsilon$ is bounded for any $\epsilon \geq 0$ and so as the level set $\mathcal{L}_\epsilon$. Let $\mathbf{w}_\epsilon^\dagger$ denote the closest point in the ϵ-sublevel set to $\mathbf{w}$, i.e.,

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

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, a function $F(\mathbf{w})$ has a local growth rate $\theta \in (0, 1]$ in the ϵ-sublevel set ($\epsilon \ll 1$) if there exists a constant $\lambda > 0$ such that:

$$\lambda\|\mathbf{w} - \mathbf{w}_*\|_2^{1/\theta} \leq F(\mathbf{w}) - F_*, \quad \forall \mathbf{w} \in \mathcal{S}_\epsilon, \tag{5}$$

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 (5) can be equivalently written as

$$\|\mathbf{w} - \mathbf{w}_*\|_2 \leq c(F(\mathbf{w}) - F_*)^\theta, \quad \forall \mathbf{w} \in \mathcal{S}_\epsilon, \tag{6}$$

where $c = 1/\lambda^\theta$, which is called as local error bound condition in (Yang & Lin, 2016). In this work, to avoid confusion with earlier work by Luo & Tseng (1992a;b; 1993) who also explored a related but different local error bound condition, we refer to the inequality in (5) or (6) as local growth condition (LGC). If the function $F(\mathbf{x})$ is assumed to satisfy (5) 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 (6) 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 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 $\mathbf{u} = \mathbf{w}$, $\mathbf{v} = \mathbf{w}_*$, and $\partial F(\mathbf{v}) = 0$, we have (5) with $\theta = 1/p \in (0, 1/2]$. Clearly LGC covers a broader family of functions than uniform convexity.

Observation 2. *The weak strong convexity (Necoara et al., 2015), essential strong convexity (Liu et al., 2015), restricted strong convexity (Zhang & Yin, 2013), optimal strong convexity (Liu & Wright, 2015), semi-strong convexity (Gong & Ye, 2014) and other error bound conditions considered in several recent work (Karimi et al., 2016; Zhang, 2016) 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 (Necoara et al., 2015), (Karimi et al., 2016) and (Zhang, 2016) for more discussions of these conditions.

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

Observation 3. *The LGC could induce a smaller constant c in (6) 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 (6) 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 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 (Li & Pong, 2016, 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 (Bolte et al., 2015, Theorem 5).

Observation 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 (6) holds, and vice versa.*

The above KL inequality has been established for continuous semi-algebraic and subanalytic functions (Attouch et al., 2013; Bolte et al., 2006; 2015), 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.

Lemma 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 (4).

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 & Lin (2016) have leveraged this relationship to improve the convergence of the standard subgradient method. In the sequel, we will build on this relationship to further develop novel stochastic optimization algorithms with faster convergence in high probability.

4. Main Results

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:

$$\mathbf{w}_{\tau+1} = \Pi_{\mathcal{K}\cap\mathcal{B}(\mathbf{w}_1, D)}[\mathbf{w}_\tau - \eta\nabla f(\mathbf{w}_\tau; \xi_\tau)]. \quad (7)$$

Lemma 2. *Given $\mathbf{w}_1 \in \mathcal{K}$, apply t-iterations of (7). 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 (Hazan & Kale, 2011). 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 of the stochastic subgradients cannot be reduced, we therefore propose to address this issue by reducing D in light of the inequality in Lemma 1.

The updates in (7) 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. In the sequel, all omitted proofs can be found in the supplement.

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 approx-

imately 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 (7). Thanks to Lemma 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 & Kale (2011) and the (multi-stage) AC-SA method with domain shrinkage by Ghadimi & Lan (2013) for stochastic strongly convex optimization. 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 1. The convergence of ASSG-c is presented in the theorem below.

Theorem 1. *Suppose Assumption 1 holds and $F(\mathbf{w})$ obeys the LGC (6). 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 $O(c^2 G^2 \lceil \log_2(\frac{\epsilon_0}{\epsilon}) \rceil \log(1/\delta)/\epsilon^{2(1-\theta)})$ 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 (Ramdas & Singh, 2013) 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 subsection 4.3, we alleviate this requirement to make the algorithm more practical.

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, we first present a common strategy that solves the non-strongly convex problem (1) by stochastic strongly convex optimization. The basic idea is from the classical deterministic *proximal point algorithm* (Rockafellar, 1976) which adds a strongly convex regularizer to the original problem and solve the resulting proximal problem. In particular, we construct a new problem

$$\min_{\mathbf{w}\in\mathcal{K}} \widehat{F}(\mathbf{w}) = F(\mathbf{w}) + \frac{1}{2\beta}\|\mathbf{w} - \mathbf{w}_1\|_2^2, \tag{8}$$

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 $\mathcal{K}$. We can employ the stochastic subgradient method suited for strongly convex problems to solve the above problem. The update is given by

$$\mathbf{w}_{t+1} = \Pi_{\mathcal{K}}[\mathbf{w}'_{t+1}] = \arg\min_{\mathbf{w}\in\mathcal{K}} \left\|\mathbf{w} - \mathbf{w}'_{t+1}\right\|_2^2, \tag{9}$$

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 $\eta_t = \frac{2\beta}{t}$ [2]. 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).

Lemma 3. *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.

Recall that the main iteration of the proximal point algorithm (Rockafellar, 1976) is

$$\mathbf{w}_k \approx \arg\min_{\mathbf{w}\in\mathcal{K}} F(\mathbf{w}) + \frac{1}{2\beta_k}\|\mathbf{w} - \mathbf{w}_{k-1}\|_2^2, \tag{10}$$

where $\mathbf{w}_k$ approximately solves the minimization problem above with β_k changing with k. With the same idea, our

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

Input: $\mathbf{w}_0 \in \mathcal{K}$, K, t, ϵ_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: $\mathbf{w}_K$

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

Algorithm 2 the ASSG-r algorithm for solving (1)

Input: $\mathbf{w}_0 \in \mathcal{K}$, K, t, ϵ_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 $\mathbf{w}'_{\tau+1} = \left(1 - \frac{2}{\tau}\right)\mathbf{w}_\tau^k + \frac{2}{\tau}\mathbf{w}_1^k - \frac{2\beta}{\tau}\partial f(\mathbf{w}_\tau^k; \xi_\tau^k)$
    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: $\mathbf{w}_K$

regularized variant of ASSG generates $\mathbf{w}_k$ from stage k by solving the minimization problem (10) approximately using (9). 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 (9) using the solution from the previous stage as the new regularization reference point. Despite of its similarity to the proximal point 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 2. *Suppose Assumption 1 holds and $F(\mathbf{w})$ obeys the LGC (6). 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 $O(c^2G^2\log(\epsilon_0/\epsilon)\log(1/\delta)/\epsilon^{2(1-\theta)})$ provided $\beta_1 = O(\frac{2c^2\epsilon_0}{\epsilon^{2(1-\theta)}})$.*

4.3. More 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 that depend on unknown c and potentially unknown θ. This subsection is devoted to more practical variants of ASSG. For ease of presentation, we focus on the constrained variant of ASSG.

When c is known, we present the details of a restarting variant of ASSG in Algorithm 3, to which we refer as RASSG. The key idea is to use an increasing sequence of t and another level of restarting for ASSG.

Theorem 3 (RASSG with unknown c)**.** *Let $\epsilon \leq \epsilon_0/4$, $\omega = 1$, and $K = \lceil \log_2(\frac{\epsilon_0}{\epsilon}) \rceil$ in Algorithm 3. 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 (6) 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 3 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 $T_S = O(\lceil \log_2(\frac{\epsilon_0}{\epsilon}) \rceil \log(1/\delta)/\epsilon^{2(1-\theta)})$.*

Algorithm 3 ASSG with Restarting: RASSG

Input: $\mathbf{w}^{(0)}$, K, $D_1^{(1)}$, t_1, ϵ_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: $\mathbf{w}^{(S)}$

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 3. 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.

When θ is unknown, we can set $\theta = 0$ and $c = B_\varepsilon$ with $\varepsilon \geq \epsilon$ in the LGC (6), 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 4 (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 3. 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 3 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 $T_S = O(\lceil \log_2(\frac{\epsilon_0}{\epsilon}) \rceil \log(1/\delta)\frac{G^2B_{\hat{\epsilon}_1}^2}{\epsilon^2})$.*

Remark: The Lemma 6 in the supplement 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 (3)), where B is the domain size or the distance of initial solution to the optimal set.

5. 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:

$$\min_{\mathbf{w}\in\mathbb{R}^d} F(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} \ell(\mathbf{w}^\top \mathbf{x}_i, y_i) + \lambda R(\mathbf{w}), \qquad (11)$$

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.

5.1. 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 (11) with a piecewise linear loss and ℓ_1, ℓ_∞ or $\ell_{1,\infty}$ regularizers.

Piecewise linear loss includes hinge loss, generalized hinge loss, absolute loss, and ϵ-insensitive loss. For particular forms of these loss functions, please refer to (Yang et al., 2014). 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 (Yang & Lin, 2016), 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 5. *Assume the loss function $\ell(z, y)$ is piecewise linear, then the problem in (11) with ℓ_1, ℓ_∞ or $\ell_{1,\infty}$ norm regularizer satisfy the LGC in (6) with $\theta = 1$. Hence ASSG can have an iteration complexity of $O(\log(1/\delta)\log(\epsilon_0/\epsilon))$ with a high probability $1 - \delta$.*

5.2. Piecewise Convex Quadratic Minimization

Next, 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 (Li, 2013). 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 (Li, 2013).

For examples of piecewise convex quadratic problems in machine learning, one can consider the problem (11) with a huber loss, squared hinge loss or square loss, and ℓ_1, ℓ_∞, $\ell_{1,\infty}$, or huber norm regularizer (Zadorozhnyi et al., 2016). The Huber function is defined as $\ell_\gamma(z) = \begin{cases} \frac{1}{2}z^2 & \text{if } |z| \leq \gamma, \\ \gamma(|z| - \frac{1}{2}\gamma) & \text{otherwise,} \end{cases}$, which is a piecewise convex quadratic function. The huber loss function $\ell(z, y) = \ell_\gamma(z - y)$ has been used for robust regression. A Huber regularizer is defined as $R(\mathbf{w}) = \sum_{i=1}^d \ell_\gamma(w_i)$.

It has been shown that (Li, 2013), if $F(\mathbf{w})$ is convex and piecewise convex quadratic, then it satisfies the LGC (6) with $\theta = 1/2$. The corollary below summarizes the iteration complexity of ASSG for solving these problems.

Corollary 6. *Assume the loss function $\ell(z, y)$ is a convex and piecewise convex quadratic, then the problem in (11) with ℓ_1, ℓ_∞, $\ell_{1,\infty}$ or huber norm regularizer satisfy the LGC in (6) with $\theta = 1/2$. Hence ASSG can have an iteration complexity of $\widetilde{O}(\frac{\log(1/\delta)}{\epsilon})$ with a high probability $1 - \delta$.*

5.3. Structured composite non-smooth problems

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

$$\min_{\mathbf{w}\in\mathbb{R}^d} F(\mathbf{w}) \triangleq h(X\mathbf{w}) + P(\mathbf{w}). \qquad (12)$$

Corollary 7. *Assume $h(\mathbf{u})$ is a strongly convex function on any compact set and $P(\mathbf{w})$ is polyhedral, then the problem in (12) satisfies the LGC in (6) 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 7 can be found in (Yang & Lin, 2016). 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 (Goebel & Rockafellar, 2007).

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 (Nyquist, 1983):

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

Liu & Yang (2016) have shown that the problem above satisfies the LGC in (6) with $\theta = \frac{1}{p}$.

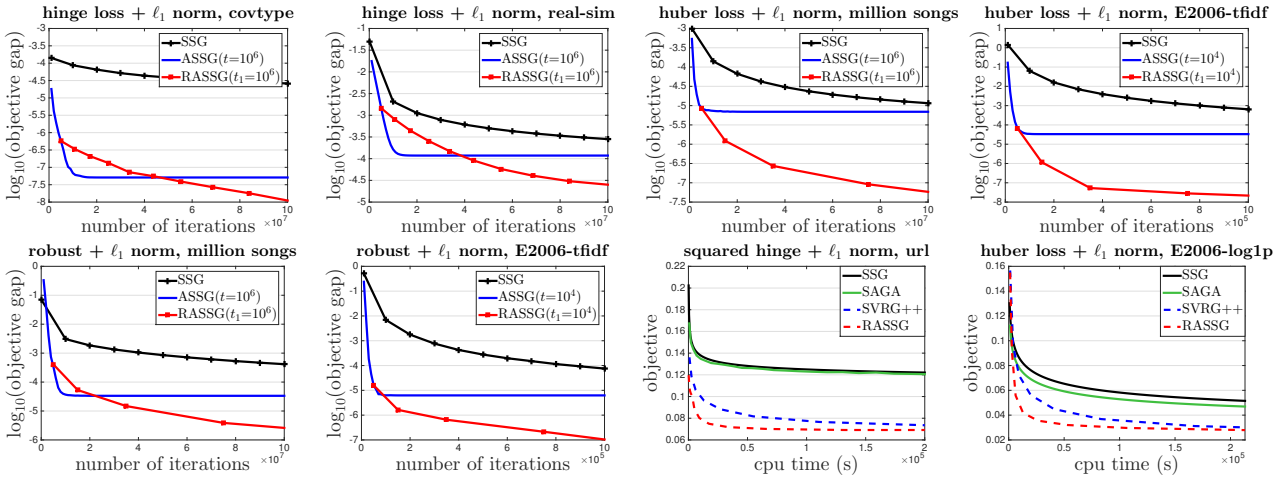

Figure 1. Comparison of different algorithms for solving different problems on different datasets.

6. Experiments

In this section, we perform some experiments to demonstrate effectiveness of proposed algorithms. We use very large-scale datasets from libsvm website in experiments, including covtype.binary ($n = 581012$), real-sim ($n = 72309$), url ($n = 2396130$) for classification, million songs ($n = 463715$), E2006-tfidf ($n = 16087$), E2006-log1p ($n = 16087$) for regression. The detailed statistics of these datasets are shown in the supplement.

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} in all tasks (We also perform the experiments with $\lambda = 10^{-2}$ and include the results in the supplement). 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 (first 6 figures), in which we plot the log difference between the objective value and the smallest obtained objective value (to which 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.

Finally, 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 SAGA (Defazio et al., 2014) and SVRG++ (Allen-Zhu & Yuan, 2016). We conduct experiments on two high-dimensional datasets url and E2006-log1p and fix the regularization parameter $\lambda = 10^{-4}$ (We also include the results for $\lambda = 10^{-2}$ in the supplement). 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 (Allen-Zhu & Yuan, 2016). We plot the objective versus the CPU time (second) in Figure 1 (last 2 figures). 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.

7. 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.

Acknowledgement

We thank the anonymous reviewers for their helpful comments. Y. Xu and T. Yang are partially supported by National Science Foundation (IIS-1463988, IIS-1545995). T. Yang would like to thank Lijun Zhang for pointing out (Kakade & Tewari, 2008) for his attention.

References

Allen-Zhu, Zeyuan and Yuan, Yang. Improved svrg for non-strongly-convex or sum-of-non-convex objectives. In *ICML*, pp. 1080–1089, 2016.

Attouch, Hedy, Bolte, Jérôme, and Svaiter, Benar Fux. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized gauss-seidel methods. *Math. Program.*, 137(1-2):91–129, 2013.

Bolte, Jérôme, Daniilidis, Aris, and Lewis, Adrian. The łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM J. on Optimization*, 17:1205–1223, 2006.

Bolte, Jérôme, Nguyen, Trong Phong, Peypouquet, Juan, and Suter, Bruce. From error bounds to the complexity of first-order descent methods for convex functions. *CoRR*, abs/1510.08234, 2015.

Defazio, Aaron, Bach, Francis R., and Lacoste-Julien, Simon. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *NIPS*, pp. 1646–1654, 2014.

Ghadimi, Saeed and Lan, Guanghui. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, ii: Shrinking procedures and optimal algorithms. *SIAM Journal on Optimization*, 23 (4):20612089, 2013.

Goebel, R. and Rockafellar, R. T. Local strong convexity and local lipschitz continuity of the gradient of convex functions. *Journal of Convex Analysis*, 2007.

Gong, Pinghua and Ye, Jieping. Linear convergence of variance-reduced projected stochastic gradient without strong convexity. *CoRR*, abs/1406.1102, 2014.

Hazan, Elad and Kale, Satyen. Beyond the regret minimization barrier: an optimal algorithm for stochastic strongly-convex optimization. In *COLT*, pp. 421–436, 2011.

Hou, Ke, Zhou, Zirui, So, Anthony Man-Cho, and Luo, Zhi-Quan. On the linear convergence of the proximal gradient method for trace norm regularization. In *NIPS*, pp. 710–718, 2013.

Johnson, Rie and Zhang, Tong. Accelerating stochastic gradient descent using predictive variance reduction. In *NIPS*, pp. 315–323, 2013.

Juditsky, Anatoli and Nesterov, Yuri. Deterministic and stochastic primal-dual subgradient algorithms for uniformly convex minimization. *Stoch. Syst.*, 4:44–80, 2014.

Kakade, Sham M. and Tewari, Ambuj. On the generalization ability of online strongly convex programming algorithms. In *NIPS*, pp. 801–808, 2008.

Karimi, Hamed, Nutini, Julie, and Schmidt, Mark W. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *ECML-PKDD*, pp. 795–811, 2016.

Li, Guoyin. Global error bounds for piecewise convex polynomials. *Math. Program.*, 137(1-2):37–64, 2013.

Li, Guoyin and Pong, Ting Kei. Calculus of the exponent of kurdyka- łojasiewicz inequality and its applications to linear convergence of first-order methods. *CoRR*, abs/1602.02915, 2016.

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

Liu, Ji, Wright, Stephen J., Ré, Christopher, Bittorf, Victor, and Sridhar, Srikrishna. An asynchronous parallel stochastic coordinate descent algorithm. *J. Mach. Learn. Res.*, 16:285–322, 2015. ISSN 1532-4435.

Liu, Mingrui and Yang, Tianbao. Adaptive accelerated gradient converging methods under holderian error bound condition. *CoRR*, abs/1611.07609, 2016.

Luo, Zhi-Quan and Tseng, Paul. On the convergence of coordinate descent method for convex differentiable minization. *Journal of Optimization Theory and Applications*, 72(1):7–35, 1992a.

Luo, Zhi-Quan and Tseng, Paul. On the linear convergence of descent methods for convex essenially smooth minization. *SIAM Journal on Control and Optimization*, 30(2): 408–425, 1992b.

Luo, Zhi-Quan and Tseng, Paul. Error bounds and convergence analysis of feasible descent methods: a general approach. *Annals of Operations Research*, 46:157–178, 1993.

Necoara, I., Nesterov, Yu., and Glineur, F. Linear convergence of first order methods for non-strongly convex optimization. *CoRR*, abs/1504.06298, 2015.

Nemirovsky A.S., Arkadii Semenovich. and Yudin, D. B. *Problem complexity and method efficiency in optimization*. Wiley-Interscience series in discrete mathematics. Wiley, Chichester, New York, 1983. ISBN 0-471-10345-4. A Wiley-Interscience publication.

Nesterov, Yurii. *Introductory lectures on convex optimization : a basic course*. Applied optimization. Kluwer Academic Publ., 2004. ISBN 1-4020-7553-7.

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

Qu, Chao, Xu, Huan, and Ong, Chong Jin. Fast rate analysis of some stochastic optimization algorithms. In *ICML*, pp. 662–670, 2016.

Ramdas, Aaditya and Singh, Aarti. Optimal rates for stochastic convex optimization under tsybakov noise condition. In *ICML*, pp. 365–373, 2013.

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

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

Roux, Nicolas Le, Schmidt, Mark W., and Bach, Francis. A stochastic gradient method with an exponential convergence rate for finite training sets. In *NIPS*, pp. 2672–2680, 2012.

Wang, Po-Wei and Lin, Chih-Jen. Iteration complexity of feasible descent methods for convex optimization. *Journal of Machine Learning Research*, 15(1):1523–1548, 2014.

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

Xu, Yi, Yan, Yan, Lin, Qihang, and Yang, Tianbao. Homotopy smoothing for non-smooth problems with lower complexity than $\mathrm{O}(1/\epsilon)$. In *NIPS*, pp. 1208–1216, 2016.

Yang, Tianbao and Lin, Qihang. Rsg: Beating sgd without smoothness and/or strong convexity. *CoRR*, abs/1512.03107, 2016.

Yang, Tianbao, Mahdavi, Mehrdad, Jin, Rong, and Zhu, Shenghuo. An efficient primal-dual prox method for non-smooth optimization. *Machine Learning*, 2014.

Zadorozhnyi, Oleksandr, Benecke, Gunthard, Mandt, Stephan, Scheffer, Tobias, and Kloft, Marius. Huber-norm regularization for linear prediction models. In *ECML-PKDD*, pp. 714–730, 2016.

Zhang, Hui. New analysis of linear convergence of gradient-type methods via unifying error bound conditions. *CoRR*, abs/1606.00269, 2016.

Zhang, Hui and Yin, Wotao. Gradient methods for convex minimization: better rates under weaker conditions. *CoRR*, abs/1303.4645, 2013.

Zhang, Lijun, Mahdavi, Mehrdad, and Jin, Rong. Linear convergence with condition number independent access of full gradients. In *NIPS*, pp. 980–988, 2013.

Zhou, Zirui and So, Anthony Man-Cho. A unified approach to error bounds for structured convex optimization problems. *CoRR*, abs/1512.03518, 2015.

Zhou, Zirui, Zhang, Qi, and So, Anthony Man-Cho. L1p-norm regularization: Error bounds and convergence rate analysis of first-order methods. In *ICML*, pp. 1501–1510, 2015.