Stochastic subGradient Methods with Linear Convergence for Polyhedral Convex Optimization

Tianbao Yang[†*] **Qihang Lin**[‡]
[†]Department of Computer Science
[‡]Department of Management Sciences
The University of Iowa, Iowa City, IA 52242
tianbao-yang@uiowa.edu, qihang-lin@uiowa.edu

Abstract

In this paper, we show that simple Stochastic subGradient Decent methods with multiple Restarting, named **RSGD**, can achieve a *linear convergence rate* for a class of non-smooth and non-strongly convex optimization problems where the epigraph of the objective function is a polyhedron, to which we refer as **polyhedral convex optimization**. Its applications in machine learning include ℓ_1 constrained or regularized piecewise linear loss minimization and submodular function minimization. To the best of our knowledge, this is the first result on the linear convergence rate of stochastic subgradient methods for non-smooth and non-strongly convex optimization problems.

1 Introduction

The subgradient descent algorithm and its stochastic version are classical first-order methods for optimizing non-smooth problems. When the objective function is non-strongly convex, their convergence rate is $O(1/\sqrt{T})$ with the T being the number of iterations. And it has been shown that this sublinear convergence rate is unimprovable for general non-smooth problems [6]. In this paper, we present linearly convergent stochastic subgradient methods as simple as standard stochastic subgradient descent methods for a class of non-smooth and non-strongly convex problems whose epigraph is a polyhedron. It can find applications in ℓ_1 or ℓ_∞ regularized/constrained piecewise linear loss (e.g., hinge loss, absolute loss) minimization [2] and submodular function minimization [1].

In the present paper, we show that for a family of non-smooth and non-strongly convex optimization problems, a simple restarting scheme can make stochastic subgradient descent (SGD) method converge linearly, given that the epigraph of the objective function is a polyhedron. This technique is based on the fact that, for such a problem, the distance of a solution to the optimal set can be bounded by a multiple of the difference between the objective value of this solution and the optimal objective value, as illustrated by Figure 1. We refer to this fact as **polyhedral error bound condition** and to the family of non-smooth and non-strongly convex optimization of interest as **polyhedral convex optimization**. Our work is motivated by [3] which established a version of the polyhedral error bound condition when the domain of the problem is a polytope and associated it with the Nesterov's smoothing method for solving the Nash equilibrium of a two-person zero-sum game. Their algorithm achieves a linear convergence rate. Compared to their method, the polyhedral error bound condition we consider here allows the domain to be a unbounded polyhedron, and thus, more general. We show that the stochastic subgradient methods can also benefit from this condition to obtain a linear convergence rate.

[*]A longer version of this extended abstract is available on arxiv at http://arxiv.org/abs/1510.01444.

2 Stochastic subGradient methods with Linear Convergence Rate

In this section, we describe the techniques that make stochastic subgradient descent methods converge linearly for a family of non-smooth and non-strongly convex problems. We consider the following optimization problem:

$$\min_{\mathbf{w}\in\Omega} f(\mathbf{w}) \tag{1}$$

where $f(\mathbf{w})$ is a non-smooth and non-strongly convex function and Ω is a closed convex set in $\mathbb{R}^d$. We denote by $\partial f(\mathbf{w})$ a subgradient of $f(\mathbf{w})$ and by $\partial f(\mathbf{w};\xi)$ a stochastic subgradient of $f(\mathbf{w})$ that depends on a random variable ξ such that $\mathrm{E}_\xi[\partial f(\mathbf{w};\xi)] = \partial f(\mathbf{w})$. Throughout the paper, we make the following assumptions unless stated otherwise.

Assumption 1. *For a convex minimization problem (1), we assume*

- **a.** *There exist known $\mathbf{w}_0 \in \Omega$ and $\epsilon_0 \geq 0$ such that $f(\mathbf{w}_0) - \min_{\mathbf{w}\in\Omega} f(\mathbf{w}) \leq \epsilon_0$.*
- **b.** *There exists a known constant G such that $\|\partial f(\mathbf{w})\|_2 \leq G$ or $\|\partial f(\mathbf{w};\xi)\|_2 \leq G$ almost surely.*
- **c.** *The epigraph of f over Ω is a polyhedron.*

We refer to the convex problem (1) that satisfies the **Assumption 1.c** as polyhedral convex optimization. We can show that many non-smooth and non-strongly convex machine learning problems satisfy the above assumptions, including

- ℓ_1 constrained/regularized hinge loss minimization, i.e.,

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} \max(0, 1 - y_i \mathbf{w}^\top \mathbf{x}_i) + R(\mathbf{w})$$

  where $R(\mathbf{w})$ is

$$R(\mathbf{w}) = \begin{cases} 0 & \text{if } \|\mathbf{w}\|_1 \leq B \\ \infty & \text{otherwise} \end{cases}, \text{ or } \quad R(\mathbf{w}) = \lambda\|\mathbf{w}\|_1$$

  for the constrained problems or for the regularized problems, respectively.

Another example is submodular function minimization [1] .

In the sequel, we denote by $\|\cdot\|$ a general vector norm and by $\|\cdot\|_2$ the Euclidean norm . Let Δ and $\mathbb{R}_+$ denote a simplex and a positive cone of an appropriate dimension, respectively. Since the objective function is not strongly convex, the optimal solutions may not be unique. Thus, we use Ω_* to denote the optimal solution set and use f_* to denote the unique optimal objective value. Let $\mathbf{w}^+$ denote the closest optimal solution in Ω_* to $\mathbf{w}$ measured in terms of norm $\|\cdot\|$, i.e.,

$$\mathbf{w}^+ = \min_{\mathbf{u}\in\Omega_*} \|\mathbf{w} - \mathbf{u}\|.$$

The following lemma is the key to our analysis that is a result of the Assumption **1.c**.

Lemma 1 (Polyhedral Error Bound Condition)**.** *Suppose Assumption* **1.c** *is satisfied, then there exists $\kappa > 0$ that depends on the definition of $\|\cdot\|$ such that*

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

Remark: Lemma 1 above generalizes the Lemma 4 in [3], which requires the feasible set to be a polytope (i.e., a bounded polyhedron), to a similar result where the feasible set can be a (unbounded) polyhedron. This generalization is essential because it allows the development of efficient algorithms for many unconstrained machine learning problems without artificially including a constraint. Different from [3] that used their Lemma 4 to develop a linearly convergent algorithm for solving the Nash equilibrium of a two-person zero-sum games based on Nesterov's smoothing technique [7], we show in this paper that Lemma 1 provides the basis for a stochastic gradient method with linear convergence for the polyhedral convex minimization problems. A graphical illustration of Lemma 1 for an one dimensional problem is shown in Figure 1.

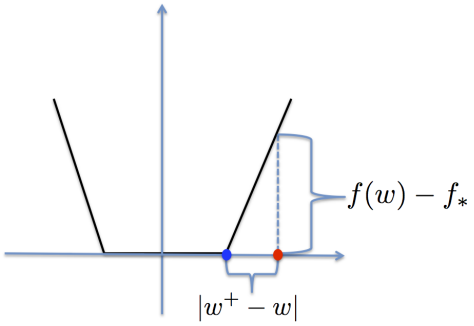

Figure 1: An illustration of Lemma 1 for a non-smooth and non-stronlgy convex problem. The value of $\kappa = \frac{f(w)-f_*}{|w-w^+|}$ is just the slope of the linear line, where w^+ (blue point) is the closest optimal solution to w (red point).

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

1: **Input**: a step size η, the number iterations T, and the initial solution $\mathbf{w}_1$,
2: **for** $t = 1, \ldots, T$ **do**
3: compute a subgradient or stochastic subgradient of $f(\mathbf{w})$ at $\mathbf{w}_t$ denoted by $\mathbf{g}_t$
4: update $\mathbf{w}_{t+1} = \Pi_\Omega[\mathbf{w}_t - \eta \mathbf{g}_t]$
5: **end for**
6: **Output**: $\widehat{\mathbf{w}}_T = \sum_{t=1}^T \frac{\mathbf{w}_t}{T}$

2.1 Restarted Stochastic subGradient Descent (RSGD) Method

In the sequel, we present all results **using Euclidean norm to define $\mathbf{w}^+$ and the parameter κ**. We first describe the vanilla stochastic subgradient descent method in Algorithm 1 that will serve as a subroutine in the proposed algorithm, where we make it an option to use either a subgradient or a stochastic subgradient and abuse the name SGD to denote both subgradient descent and stochastic subgradient descent methods. The step 4 in Algorithm 1 is a projection onto Ω defined as

$$\Pi_\Omega[\mathbf{w}] = \arg\min_{\mathbf{u}\in\Omega} \frac{1}{2}\|\mathbf{u} - \mathbf{w}\|_2^2$$

The following lemma [9] provides guarantee on the convergence of SGD methods. For the sake of completeness, we provide its proof in the Appendix.

Lemma 2. *Let SGD run T iterations. Then we have*

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

for using subgradients, and

$$\mathrm{E}\left[f(\widehat{\mathbf{w}}_T) - f_*\right] \leq \frac{G^2\eta}{2} + \frac{\mathrm{E}[\|\mathbf{w}_1 - \mathbf{w}_1^+\|_2^2]}{2\eta T}.$$

for using stochastic subgradients.

Now, we are ready to present the proposed RSGD method. The key steps are presented in Algorithm 2. The algorithm uses SGD as a subroutine in multiple epochs. In each epoch, it runs SGD for a fixed number of iterations t and restarts SGD using the averaged solution in the previous epoch as the starting point. The algorithm geometrically decreases the step size η_k between epochs. The returned solution $\mathbf{w}_K$ is the averaged solution of updates in the K-th epoch. We would like compare the proposed RSGD method with two other stochastic optimization algorithms that also run in epochs, namely Epoch-SGD for strongly convex optimization [4] and SVRG for smooth and strongly convex optimization [5]. Different from Epoch-SGD that needs to increase the number of iterations per-epoch geometrically, RSGD uses a constant number of iterations per epoch similarly

Algorithm 2 RSGD

Input: the number of epochs K and the number iterations t per-epoch,
Initialization: $\mathbf{w}_0 \in \Omega$ and ϵ_0 as in **Assumption 1.a**
for $k = 1, \ldots, K$ **do**
  Set $\eta_k = \epsilon_{k-1}/(2G^2)$
  Run SGD to obtain $\mathbf{w}_k = \text{SGD}(\mathbf{w}_{k-1}, \eta_k, t)$
  Set $\epsilon_k = \epsilon_{k-1}/2$
end for
Output: $\mathbf{w}_K$

as in SVRG. Different from SVRG that uses the constant step size due to the smoothness, RSGD requires a decreasing step size due to non-smoothness similarly as in Epoch-SGD.

The main theorem regarding the convergence of the proposed RSGD is presented below. The convergence result without expectation is for using subgradients to update the solution and that with expectation is for using stochastic subgradients, where the expectation is take over the randomness over the stochastic subgradients.

Theorem 1. *Suppose the* **Assumption 1** *is satisfied. Let Algorithm 2 run with a sufficiently large number of iterations per-epoch t such that $t \geq \frac{4G^2}{\kappa^2}$. Then, depending on using either subgradients or stochastic subgradients, we have*

$$f(\mathbf{w}_K) - f_* \leq \frac{\epsilon_0}{2^K}, \quad \textit{or} \quad \mathrm{E}[f(\mathbf{w}_K) - f_*] \leq \frac{\epsilon_0}{2^K},$$

In particular, if $K = \lceil \log_2(\frac{\epsilon_0}{\epsilon}) \rceil$, we have

$$f(\mathbf{w}_K) - f_* \leq \epsilon, \quad \textit{or} \quad \mathrm{E}[f(\mathbf{w}_K) - f_*] \leq \epsilon,$$

and the total number of iterations is $T = t\lceil \log_2(\frac{\epsilon_0}{\epsilon}) \rceil$.

Remark: Since the number of iterations per-epoch t is a constant independent of ϵ, the overall iteration complexity is $O(\log(1/\epsilon))$.

Proof. We prove the result for using the subgradients and the proof for stochastic subgradients is a straightforward extension. We prove the theorem by induction. The result holds obviously for $k = 0$. Assuming $f(\mathbf{w}_{k-1}) - f_* \leq \epsilon_{k-1}$, we only need to show that $f(\mathbf{w}_k) - f_* \leq \epsilon_k$. We first apply Lemma 2 to each epoch of Algorithm 2 and get

$$f(\mathbf{w}_k) - f_* \leq \frac{G^2\eta_k}{2} + \frac{\|\mathbf{w}_{k-1} - \mathbf{w}_{k-1}^+\|_2^2}{2\eta_k t}$$

By Lemma 1, we have

$$\|\mathbf{w}_{k-1} - \mathbf{w}_{k-1}^+\|_2 \leq \frac{1}{\kappa}(f(\mathbf{w}_{k-1}) - f_*) \leq \frac{\epsilon_{k-1}}{\kappa}$$

Choosing $\eta_k = \dfrac{\epsilon_{k-1}}{2G^2}$ and $t \geq \dfrac{4G^2}{\kappa^2}$, we have

$$f(\mathbf{w}_k) - f_* \leq \frac{\epsilon_{k-1}}{4} + \frac{\epsilon_{k-1}^2}{4\epsilon_{k-1}} \leq \frac{\epsilon_{k-1}}{2} = \epsilon_k.$$

As a result of induction, we have

$$f(\mathbf{w}_K) - f_* \leq \frac{\epsilon_0}{2^K}.$$

for any $K \geq 0$. □

3 Conclusions

In this paper, we have proposed a restarted stochastic subgradient descent method that restarts SGD updates after a fixed number of iterations with the averaged solution obtained from previous epoch as the starting point and with a geometrically decreasing step size. We prove that the proposed method achieves a linear convergence for a family of non-smooth and non-strongly convex problems including many examples from machine learning.

References

[1] F. R. Bach. Learning with submodular functions: A convex optimization perspective. *Foundations and Trends in Machine Learning*, 6(2-3):145–373, 2013.

[2] J. Duchi and Y. Singer. Efficient online and batch learning using forward backward splitting. *Journal of Machine Learning Research*, 10, 2009.

[3] A. Gilpin, J. Peña, and T. Sandholm. First-order algorithm with log(1/epsilon) convergence for epsilon-equilibrium in two-person zero-sum games. *Math. Program.*, 133(1-2):279–298, 2012.

[4] 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.

[5] R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *NIPS*, pages 315–323, 2013.

[6] A. S. Nemirovsky A.S. and D. B. IUdin. *Problem complexity and method efficiency in optimization*. Wiley-Interscience series in discrete mathematics. Wiley, Chichester, New York, 1983. A Wiley-Interscience publication.

[7] Y. Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1):127–152, 2005.

[8] J. Renegar. A framework for applying subgradient methods to conic optimization problems. *ArXiv e-prints*, 2015.

[9] M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 928–936, 2003.