Accelerate Stochastic Subgradient Method by Leveraging Local Error Bound

Yi Xu[†], **Qihang Lin**[♮], **Tianbao Yang**[†]
[†] Department of Computer Science, The University of Iowa, Iowa City, IA 52242
[♮] Department of Management Sciences, The University of Iowa, Iowa City, IA 52242
{yi-xu, qihang-lin, tianbao-yang}@uiowa.edu

Abstract

In this paper, we propose an **accelerated stochastic subgradient** method for stochastic non-strongly convex optimization problems by leveraging a generic local error bound condition. The novelty of the proposed method lies at smartly leveraging the recent historical solution to tackle the variance in the stochastic subgradient. The key idea of method is to iteratively solve the original problem approximately in a local region around a recent historical solution with size of the local region gradually decreasing as the solution approaches the optimal set. We establish the improved iteration complexity in a high probability for achieving an ϵ-optimal solution. Besides the improved order of iteration complexity with a high probability, the proposed algorithm also enjoys a logarithmic dependence on the distance of the initial solution to the optimal set. When applied to the ℓ_1 regularized polyhedral loss minimization (e.g., hinge loss, absolute loss), the proposed stochastic method has a logarithmic iteration complexity.

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 $\mathbf{w}$ 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 $\Pi_{\mathcal{K}}$ is a projection operator that projects a point into $\mathcal{K}$ (c.f. Eqn. (4)). Under the following assumptions i) $\|\partial f(\mathbf{w};\xi)\|_2 \le G$, ii) there exists $\mathbf{w}_* \in \mathcal{K}_*$ such that $\|\mathbf{w}_t - \mathbf{w}_*\|_2 \le B$ for $t = 1,\ldots,T$ [1], and by setting the step size $\eta_t = \frac{B}{G\sqrt{T}}$ in (2), we can show that with a high probability $1-\delta$

$$F(\hat{\mathbf{w}}_T) - F_* \le O\left(\left[GB(1+\sqrt{\log(1/\delta)})\right]/\sqrt{T}\right)$$

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

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

The slow convergence of SSG is due to the variance in the stochastic subgradient, 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 [16, 21, 8, 18, 3]. However, they all hinge on the smoothness assumption. In this paper, we tackle the issue of variance in stochastic **subgradient** without the smoothness assumption. The key idea is to iteratively solve the original problem approximately in a local region around a recent historical solution using the SSG method with an adaptive constant step size. By leveraging the local error bound, we gradually reduce the size of the local region and the step size as well in a stage-wise manner, which yields faster convergence. This strategy is fundamentally different from traditional SSG that reduces the step size after every iteration or simply adoptes a very small step size. This new strategy is the main message that we would like to convey. We refer to the proposed methods as **accelerated stochastic subgradient (ASSG)** methods.

In particular, we show that the proposed algorithm enjoys an iteration complexity of $\widetilde{O}\left(\frac{1}{\epsilon^{2(1-\theta)}}\right)$ [2] for obtaining an ϵ-optimal solution in a high probability $1-\delta$, where $\theta \in (0, 1]$ is a constant in the local error bound condition (Definition 2) that captures the local sharpness of the objective function $F(\mathbf{w})$ near the optimal set. Thus, for a family of problems with the constant $\theta = 1$ (e.g., ℓ_1 regularized empirical hinge loss minimization), the proposed algorithms have a logarithmic iteration complexity. To the best of our knowledge, this is the first work that improves the convergence of SSG method by exploring the local error bound, though it has been recently explored to improve the convergence of deterministic subgradient method in [19].

2 Preliminaries

We present some preliminaries in this section. For the optimization problem in (1), we make the following assumption throughout the paper.

Assumption 1. *For a stochastic optimization problem (1), we assume (i) there exist* $\mathbf{w}_0 \in \mathcal{K}$ *and* $\epsilon_0 \geq 0$ *such that* $F(\mathbf{w}_0) - F_* \leq \epsilon_0$*;(ii)* $\mathcal{K}_*$ *is a non-empty convex compact set; (iii) There exists a constant* G *such that* $\|\partial f(\mathbf{w}; \xi)\|_2 \leq G$.

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.,

$$\mathcal{L}_\epsilon = \{\mathbf{w} \in \mathcal{K} : F(\mathbf{w}) = F_* + \epsilon\}, \quad \mathcal{S}_\epsilon = \{\mathbf{w} \in \mathcal{K} : F(\mathbf{w}) \leq F_* + \epsilon\}.$$

Given $\mathcal{K}_*$ is bounded, it follows from [14, 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.,

$$\mathbf{w}_\epsilon^\dagger = \arg\min_{\mathbf{v}\in\mathcal{S}_\epsilon} \|\mathbf{v} - \mathbf{w}\|_2^2 \tag{3}$$

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 $\mathbf{w}$ with a radius r. Denote by $dist(w, \mathcal{K}_*)$ the distance between $\mathbf{w}$ and the set $\mathcal{K}_*$, i.e., $dist(\mathbf{w}, \mathcal{K}_*) = \min_{\mathbf{v}\in\mathcal{K}_*} \|\mathbf{w} - \mathbf{v}\|_2$. Let $\Pi_{\mathcal{K}}[\cdot]$ be a projection operator:

$$\Pi_{\mathcal{K}}[\mathbf{w}] = \arg\min_{\mathbf{v}\in\mathcal{K}} \|\mathbf{w} - \mathbf{v}\|_2^2 \tag{4}$$

The key to our development is to explore the local error bound condition, which is stated below.

Definition 2 (Local error bound (LEB))**.** *A function* $F(\mathbf{w})$ *is said to satisfy a local error bound condition on the* ϵ*-sublevel set if there exist* $\theta \in (0, 1]$ *and* $c > 0$ *such that for any* $\mathbf{w} \in \mathcal{S}_\epsilon$

$$dist(\mathbf{w}, \mathcal{K}_*) \leq c(F(\mathbf{w}) - F_*)^\theta. \tag{5}$$

Remark: We emphasize that the local error bound is a generic condition. A broad family of functions (including almost all commonly seen functions in machine learning) obey the local error bound condition. In literature, the inequality in (5) is also known as Hölderian error bound or Łojasiewicz error bound inequality. When functions are semi-algebraic and "regular" (for instance, continuous), the above inequality is known to hold on any compact set (c.f. [2] and references therein). For many functions, the constant c and the exponent θ are known [11, 10, 12, 2]. We give an example

[2] $\widetilde{O}()$ suppresses a logarithmic factor.

Algorithm 1 the ASSG algorithm for solving (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 $D_1 \geq \frac{c\epsilon_0}{\epsilon^{1-\theta}}$
for $k = 1, \ldots, K$ **do**
  Let $\mathbf{w}_1^k = \mathbf{w}_{k-1}$
  for $\tau = 1, \ldots, t$ **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 $\mathbf{w}_k = \frac{1}{t}\sum_{\tau=1}^t \mathbf{w}_\tau^k$
  Let $\eta_{k+1} = \eta_k/2$ and $D_{k+1} = D_k/2$.
end for
Output: $\mathbf{w}_K$

of its applications in machine learning. We can consider the polyhedral loss minimization with an ℓ_1 regularization: $\min_{\mathbf{w}\in\mathbb{R}^d} F(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^n \ell(\mathbf{w}^\top \mathbf{x}_i, y_i) + \tau\|\mathbf{w}\|_1$. The term $\ell(z, y)$ denotes a polyhedral loss, whose examples include hinge loss [17], generalized hinge loss [1], absolute loss [6], ϵ-insensitive loss [15], and piecewise linear loss [9]. For particular forms of these loss functions, please refer to [20]. The epigraph of $F(\mathbf{w})$ defined by sum of a polyhedral loss function and an ℓ_1 norm regularizer is still a polyhedron. According to the polyhedral error bound condition [5, 13, 19], there exists $c > 0$ such that $dist(\mathbf{w}, \mathcal{K}_*) \leq c(F(\mathbf{w}) - F_*)$ for any $\mathbf{w} \in \mathcal{K}$, meaning that $\theta = 1$. The local error bound will be explored with the following lemma in the proof of the main theorems.

Lemma 1 ([19]). *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).

3 Accelerated Stochastic Subgradient Method

In this section, we will present the proposed ASSG method and establish its improved iteration complexity with a high probability. The detailed steps are presented in Algorithm 1. The algorithm runs in stages and each stage runs t iterations of updates similar to the SSG update except that the intermediate solutions are projected into the intersection of the problem domain $\mathcal{K}$ and a ball $\mathcal{B}(\mathbf{w}_{k-1}, D_k)$. The radius D_k geometrically decreases as $\mathbf{w}_{k-1}$ approaches to the optimal set. The step size keeps the same during each stage and geometrically decreases between stages. It is notable that ASSG is similar to the Epoch-GD method by [7] and the (multi-stage) AC-SA method with domain shrinkage by [4] for stochastic strongly convex optimization, and is also similar to the restarted subgradient method (RSG) proposed by [19]. However, the difference between ASSG and Epoch-GD/AC-SA is that the number of iterations t for all stages are the same in ASSG while it geometrically increases between stages in Epoch-GD/AC-SA. 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 is presented in the theorem below.

Theorem 3. *Suppose Assumption 1 holds and $F(\mathbf{w})$ obeys the local error bound condition. 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 $t \geq \max\{1728\log(1/\tilde{\delta}), 9\}\frac{G^2 D_1^2}{\epsilon_0^2}$. 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 for achieving an 2ϵ-optimal solution with a high probability $1-\delta$ is $\widetilde{O}(\log(1/\delta)/\epsilon^{2(1-\theta)})$ provided $D_1 = O(\frac{c\epsilon_0}{\epsilon^{(1-\theta)}})$.*

Remark: 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 logarithmic dependence on ϵ_0. If we known $dist(\mathbf{w}_0, \mathcal{K}_*) \leq B$, then we can set $\epsilon_0 = GB$. Hence, the iteration complexity of ASSG has only a logarithmic dependence on the distance of the initial solution to the optimal set. It is notable that Epoch-GD [7] and AC-SA [4] have a complexity of $O(\log(1/\delta)/\epsilon)$ and $O((\log(1/\delta))^2/\epsilon)$

respectively. Compared to their results, our ASSG method can have a better complexity if $\theta > 1/2$ and we assume local error bound condition - a much more general condition than strong convexity.

To prove Theorem 3, we first present a lemma regarding each stage of ASSG.

Lemma 2. *Let D be the upper bound of $\|\mathbf{w}_1 - \mathbf{w}_{1,\epsilon}^\dagger\|_2$. Apply t-iterations of $\mathbf{w}_{\tau+1} = \Pi_{\mathcal{K}\cap\mathcal{B}(\mathbf{w}_1,D)}[\mathbf{w}_\tau - \eta\partial f(\mathbf{w}_\tau;\xi_\tau)]$. Given $\mathbf{w}_1 \in \mathcal{K}$, for any $\delta \in (0,1)$, with a probability at least $1-\delta$*

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

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

The proof of the above lemma follows similarly as that of Lemma 10 in [7]. Next, we prove the main theorem regarding the convergence of ASSG.

Proof of Theorem 3. 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 1, we have

$$\begin{aligned}
&\|\mathbf{w}_{k-1,\epsilon}^\dagger - \mathbf{w}_{k-1}\|_2 \leq \frac{dist(\mathbf{w}_{k-1,\epsilon}^\dagger, \mathcal{K}_*)}{\epsilon}(F(\mathbf{w}_{k-1}) - F(\mathbf{w}_{k-1,\epsilon}^\dagger)) \\
=&\frac{dist(\mathbf{w}_{k-1,\epsilon}^\dagger, \mathcal{K}_*)}{\epsilon}[F(\mathbf{w}_{k-1}) - F_* + (F_* - F(\mathbf{w}_{k-1,\epsilon}^\dagger))] \leq \frac{dist(\mathbf{w}_{k-1,\epsilon}^\dagger, \mathcal{K}_*)}{\epsilon}[\epsilon_{k-1} + \epsilon - \epsilon] \\
=&\frac{dist(\mathbf{w}_{k-1,\epsilon}^\dagger, \mathcal{K}_*)\epsilon_{k-1}}{\epsilon} \leq \frac{c(F(\mathbf{w}_{k-1,\epsilon}^\dagger) - F_*)^\theta \epsilon_{k-1}}{\epsilon} \leq \frac{c\epsilon^\theta \epsilon_{k-1}}{\epsilon} = \frac{c\epsilon_{k-1}}{\epsilon^{1-\theta}} \leq D_k
\end{aligned} \tag{6}$$

We apply Lemma 2 to the k-th stage of Algorithm 1 conditioned on randomness in previous stages. With a probability $1 - \tilde{\delta}$ we have

$$F(\mathbf{w}_k) - F(\mathbf{w}_{k-1,\epsilon}^\dagger) \leq \eta_k G^2/2 + \|\mathbf{w}_{k-1} - \mathbf{w}_{k-1,\epsilon}^\dagger\|_2^2/(2\eta_k t) + 4GD_k\sqrt{3\log(1/\tilde{\delta})}/\sqrt{t} \tag{7}$$

We now consider two cases for $\mathbf{w}_{k-1}$. First, we assume $F(\mathbf{w}_{k-1}) - F_* \leq \epsilon$, i.e. $\mathbf{w}_{k-1} \in \mathcal{S}_\epsilon$. Then we have $\mathbf{w}_{k-1,\epsilon}^\dagger = \mathbf{w}_{k-1}$ and $F(\mathbf{w}_k) - F(\mathbf{w}_{k-1,\epsilon}^\dagger) \leq \eta_k G^2/2 + 4GD_k\sqrt{3\log(1/\tilde{\delta})}/\sqrt{t}$. As a result,

$$F(\mathbf{w}_k) - F_* \leq F(\mathbf{w}_{k-1,\epsilon}^\dagger) - F_* + \frac{2\epsilon_k}{3} \leq \epsilon + \epsilon_k$$

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 (7) and (6) and using $\eta_k = \frac{2\epsilon_k}{3G^2}$ and $t \geq \max\{1728\log(1/\tilde{\delta}), 9\}\frac{G^2D_1^2}{\epsilon_0^2}$, we have

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

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. □

4 Conclusion

In this paper, we have proposed an accelerated stochastic subgradient method for solving general non-strongly convex stochastic optimization under the local error bound condition. The proposed method enjoys a lower iteration complexity than vanilla stochastic subgradient method and also a logarithmic dependence on the impact of the initial solution.

Acknowledgement

Y. Xu and T. Yang are partially supported by National Science Foundation (IIS-1463988, IIS-1545995).

References

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

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

[3] 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.

[4] 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):2061–2089, 2013.

[5] 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.

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

[7] 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.

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

[9] R. Koenker. *Quantile regression*. Number 38. Cambridge university press, 2005.

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

[11] Z.-Q. Luo and J. F. Sturm. Error bound for quadratic systems. *Applied Optimization*, 33:383–404, 2000.

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

[13] J. Renegar. Efficient first-order methods for linear programming and semidefinite programming. *ArXiv e-prints*, 2014.

[14] R. Rockafellar. *Convex Analysis*. Princeton mathematical series. Princeton University Press, 1970.

[15] 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.

[16] 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.

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

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

[19] T. Yang and Q. Lin. Rsg: Beating sgd without smoothness and/or strong convexity. *CoRR*, abs/1512.03107, 2016.

[20] T. Yang, M. Mahdavi, R. Jin, and S. Zhu. An efficient primal-dual prox method for non-smooth optimization. *Machine Learning*, 2014.

[21] L. Zhang, M. Mahdavi, and R. Jin. Linear convergence with condition number independent access of full gradients. In *Advances in Neural Information Processing Systems 26*, pages 980–988, 2013.