

Finite-Sum Smooth Optimization with SARAH

Lam M. Nguyen

Marten van Dijk

Dzung T. Phan

Phuong Ha Nguyen

Tsui-Wei Weng

Jayant R. Kalagnanam

April 23, 2019

Abstract

The total complexity (measured as the total number of gradient computations) of a stochastic first-order optimization algorithm that finds a first-order stationary point of a finite-sum smooth nonconvex objective function $F(w) = \frac{1}{n} \sum_{i=1}^n f_i(w)$ has been proven to be at least $\Omega(\sqrt{n}/\epsilon)$ for $n \leq \mathcal{O}(\epsilon^{-2})$ where ϵ denotes the attained accuracy $\mathbb{E}[\|\nabla F(\tilde{w})\|^2] \leq \epsilon$ for the outputted approximation \tilde{w} [6]. In this paper, we provide a convergence analysis for a slightly modified version of the SARAH algorithm [14, 15] and achieve total complexity that matches the lower-bound worst case complexity in [6] up to a constant factor when $n \leq \mathcal{O}(\epsilon^{-2})$ for nonconvex problems. For convex optimization, we propose SARAH++ with sublinear convergence for general convex and linear convergence for strongly convex problems; and we provide a practical version for which numerical experiments on various datasets show an improved performance.

1 Introduction

We are interested in solving the *finite-sum smooth* minimization problem

$$\min_{w \in \mathbb{R}^d} \left\{ F(w) = \frac{1}{n} \sum_{i=1}^n f_i(w) \right\}, \quad (1)$$

where each f_i , $i \in [n] \stackrel{\text{def}}{=} \{1, \dots, n\}$, has a Lipschitz continuous gradient. Throughout the paper, we consider the case where F has a finite lower bound F^* .

Problems of form (1) cover a wide range of convex and nonconvex problems in machine learning applications including but not limited to logistic regression, neural networks, multi-kernel learning, etc. In many of these applications, the number of component functions n is very large, which makes the classical Gradient Descent (GD) method less efficient since it requires to compute a full gradient many times. Instead, a traditional alternative is to employ stochastic gradient descent (SGD) [17, 19, 3]. In recent years, a large number of improved variants of stochastic gradient algorithms called variance reduction methods have emerged, in particular, SAG/SAGA [18, 5], SDCA [20], MISO [11], SVRG/S2GD [7, 8], SARAH [14], etc. These

Correspondence to: **Lam M. Nguyen**, IBM Research, Thomas J. Watson Research Center, Yorktown Heights, NY, USA.
Email: LamNguyen.MLTD@ibm.com

methods were first analyzed for strongly convex problems of form (1). Due to recent interest in deep neural networks, *nonconvex* problems of form (1) have been studied and analyzed by considering a number of different approaches including many variants of variance reduction techniques (see e.g. [16, 10, 2, 1, 6], etc.)

We study the SARAH algorithm [14, 15] depicted in Algorithm 1, slightly modified. We use upper index s to indicate the s -th outer loop and lower index t to indicate the t -th iteration in the inner loop. The key update rule is

$$v_t^{(s)} = \nabla f_{i_t}(w_t^{(s)}) - \nabla f_{i_t}(w_{t-1}^{(s)}) + v_{t-1}^{(s)}. \quad (2)$$

The computed $v_t^{(s)}$ is used to update

$$w_{t+1}^{(s)} = w_t^{(s)} - \eta v_t^{(s)}. \quad (3)$$

After m iteration in the inner loop, the outer loop remembers the last computed $w_{m+1}^{(s)}$ and starts its loop anew – first with a full gradient computation before again entering the inner loop with updates (2). Instead of remembering $\tilde{w}_s = w_{m+1}^{(s)}$ for the next outer loop, the original SARAH algorithm in [14] uses $\tilde{w}_s = w_t^{(s)}$ with t chosen uniformly at random from $\{0, 1, \dots, m\}$. The authors of [14] chose to do this in order to being able to analyze the convergence rate for a single outer loop – since in practice it makes sense to keep the last computed $w_{m+1}^{(s)}$ if multiple outer loop iterations are used, we give full credit of Algorithm 1 to [14] and call this SARAH. We would like to attain an ϵ -accurate solution satisfying $\mathbb{E}[\|\nabla F(\tilde{w})\|^2] \leq \epsilon$ for the outputted approximation \tilde{w} .

Algorithm 1 SARAH (modified of [14])

Parameters: the learning rate $\eta > 0$, the inner loop size m , and the outer loop size S

Initialize: \tilde{w}_0

Iterate:

for $s = 1, 2, \dots, S$ **do**

$$w_0^{(s)} = \tilde{w}_{s-1}$$

$$v_0^{(s)} = \frac{1}{n} \sum_{i=1}^n \nabla f_i(w_0^{(s)})$$

$$w_1^{(s)} = w_0^{(s)} - \eta v_0^{(s)}$$

Iterate:

for $t = 1, \dots, m$ **do**

 Sample i_t uniformly at random from $[n]$

$$v_t^{(s)} = \nabla f_{i_t}(w_t^{(s)}) - \nabla f_{i_t}(w_{t-1}^{(s)}) + v_{t-1}^{(s)}$$

$$w_{t+1}^{(s)} = w_t^{(s)} - \eta v_t^{(s)}$$

end for

 Set $\tilde{w}_s = w_{m+1}^{(s)}$ (**modified point**)

end for

We will analyze SARAH for smooth nonconvex optimization, i.e., we study (1) with the following assumption

Assumption 1 (average- L -smooth). *The objective function F is L -average-smooth, i.e., there exists a constant $L > 0$ such that, $\forall w, w' \in \mathbb{R}^d$,*

$$\mathbb{E}_i[\|\nabla f_i(w) - \nabla f_i(w')\|^2] = \frac{1}{n} \sum_{i=1}^n \|\nabla f_i(w) - \nabla f_i(w')\|^2 \leq L^2 \|w - w'\|^2. \quad (4)$$

We notice that, the above assumption is weaker than the assumption on L -smoothness of each f_i , $i = 1, \dots, n$. Throughout this paper for non-convex results, we only consider Assumption 1 and no other assumptions. We stress that our convergence analysis only relies on the above average smooth assumption without bounded variance assumption (as required in [10, 22]).

We note that Assumption 1 implies that F is L -smooth, that is, there exists a constant $L > 0$ such that, $\forall w, w' \in \mathbb{R}^d$, $\|\nabla F(w) - \nabla F(w')\| \leq L\|w - w'\|$. By Theorem 2.1.5 in [12], we could obtain

$$F(w) \leq F(w') + \nabla F(w')^T(w - w') + \frac{L}{2}\|w - w'\|^2. \quad (5)$$

We measure the convergence rate in terms of total complexity \mathcal{T} , i.e., the total number of gradient computations. For SARAH we have

$$\mathcal{T} = S \cdot (n + 2m).$$

We notice that SARAH, using the notation and definition of [6], is a random algorithm \mathcal{A} that maps functions f to a sequence of iterates

$$[\mathbf{x}^k; i_k] = \mathcal{A}^k(\xi, \nabla f_{i_0}(\mathbf{x}^0), \nabla f_{i_1}(\mathbf{x}^1), \dots, \nabla f_{i_{k-1}}(\mathbf{x}^{k-1})),$$

where \mathcal{A}^{k-1} is a measure mapping, i_k is the individual function chosen by \mathcal{A} at iteration k , and ξ is a uniform random vector with entries in $[0, 1]$. Rephrasing Theorem 3 in [6] states the following lower bound: For $n \leq \mathcal{O}(\epsilon^{-2})$, there exists a function f such that in order to find a point $\tilde{\mathbf{x}}$ for which accuracy $\|\nabla F(\tilde{\mathbf{x}})\|^2 \leq \epsilon$, \mathcal{A} must have a total complexity \mathcal{T} of at least $\Omega(L\sqrt{n}/\epsilon)$ stochastic gradient computations. Applying this bound to SARAH tells us that if the final output \tilde{w}_S has

$$\mathbb{E}[\|\nabla F(\tilde{w}_S)\|^2] \leq \epsilon \text{ then } \mathcal{T} = S \cdot (n + 2m) = \Omega(L\sqrt{n}/\epsilon), \quad n \leq \mathcal{O}(\epsilon^{-2}).$$

In this paper, we show that in SARAH we can choose parameters S and m such that, for $n \leq \mathcal{O}(\epsilon^{-2})$, the total complexity is

$$\mathcal{T} = S \cdot (n + 2m) = \mathcal{O}(L\sqrt{n}/\epsilon).$$

Related Work: The paper that introduces SARAH [15] is only able to analyze convergence of a single outer loop giving a total complexity of $\mathcal{O}(n + \frac{L^2}{\epsilon^2})$.

Besides the lower bound, [6] introduces SPIDER, as a variant of SARAH, which achieves the best known convergence result in the nonconvex case. SPIDER uses the SARAH update rule (2) as was originally proposed in [14] and the mini-batch version of SARAH in [15]. SPIDER and SARAH are different in terms of iteration (3), which are $w_{t+1} = w_t - \eta(v_t/\|v_t\|)$ and $w_{t+1} = w_t - \eta v_t$, respectively. Also, SPIDER does not divide into outer loop and inner loop as SARAH does although SPIDER does also perform a full gradient update after a certain fixed number of iterations. A recent technical report [21] provides an improved version of SPIDER called SpiderBoost which allows a larger learning rate. Both SPIDER and SpiderBoost are able to show for smooth nonconvex optimization a total complexity of $\mathcal{O}(n + L\sqrt{n}/\epsilon)$.

Table 1¹ shows the comparison of results on the total complexity for smooth nonconvex optimization. (a) Each of the complexities in Table 1 also depends on the Lipschitz constant L , however, since we consider smooth optimization and it is custom to assume/design $L = \mathcal{O}(1)$, we ignore the dependency on L in the complexity results. (b) Although many algorithms have appeared during the past few years, we only compare

¹ $a \wedge b$ is defined as $\min\{a, b\}$ and $a \vee b$ is defined as $\max\{a, b\}$

Table 1: Comparison of results on the total complexity for smooth nonconvex optimization

Method	Complexity	Additional assumption
GD [12]	$\mathcal{O}\left(\frac{n}{\epsilon}\right)$	None
SVRG [16]	$\mathcal{O}\left(n + \frac{n^{2/3}}{\epsilon}\right)$	None
SCSG [10]	$\mathcal{O}\left((\frac{\sigma}{\epsilon} \wedge n) + \frac{1}{\epsilon} (\frac{\sigma}{\epsilon} \wedge n)^{2/3}\right)$	Bounded variance
	$\mathcal{O}\left(n + \frac{n^{2/3}}{\epsilon}\right)$	None ($\sigma \rightarrow \infty$)
SNVRG [22]	$\mathcal{O}\left(\log^3(\frac{\sigma}{\epsilon} \wedge n) \left[(\frac{\sigma}{\epsilon} \wedge n) + \frac{1}{\epsilon} (\frac{\sigma}{\epsilon} \wedge n)^{1/2} \right]\right)$	Bounded variance
	$\mathcal{O}\left(\log^3(n) \left(n + \frac{\sqrt{n}}{\epsilon} \right)\right)$	None ($\sigma \rightarrow \infty$)
SPIDER [6]	$\mathcal{O}\left(n + \frac{\sqrt{n}}{\epsilon}\right)$	None
SpiderBoost [21]	$\mathcal{O}\left(n + \frac{\sqrt{n}}{\epsilon}\right)$	None
SARAH (this paper)	$\mathcal{O}\left(\frac{\sqrt{n}}{\epsilon} \vee n\right)$	None

algorithms having a convergence result which only supposes the smooth assumption. (c) Among algorithms with convergence results that only suppose the smooth assumption, Table 1 only mentions recent state-of-the-art results. (d) Although the bounded variance assumption $\mathbb{E}[\|\nabla f_i(w) - \nabla F(w)\|^2] \leq \sigma^2$ is acceptable in many existing literature, this additional assumption limits the applicability of these convergence results since it adds dependence on σ which can be arbitrarily large. For fair comparison with convergence analysis without the bounded variance assumption, σ must be set to go to infinity – and this is what is mentioned in Table 1. As an example, from Table 1 we observe that SCSG has an advantage over SVRG only if $\sigma = \mathcal{O}(1)$ but, theoretically, it has the same total complexity as SVRG if $\sigma \rightarrow \infty$. (e) For completeness, incompatibility with assuming a bounded gradient $\mathbb{E}[\|\nabla f_i(w)\|^2] \leq \sigma$ has been discussed in [13] for strongly convex objective functions.

According to the results in Table 1, we can observe that SARAH enjoys the same fast convergence rate as those of SPIDER and SpiderBoost in the nonconvex case for finding a first-order stationary point based on only the average smooth assumption. Its complexity matches the lower-bound worst case complexity in [6] up to a constant factor when $n \leq \mathcal{O}(\epsilon^{-2})$.

Contributions: We summarize our key contributions as follows.

- *Smooth Non-Convex.* We provide a convergence analysis for the full SARAH algorithm with multiple outer iterations for nonconvex problems (unlike in [15] which only analyses a single outer iteration). Its complexity matches the lower-bound worst case complexity in [6] up to a constant factor when $n \leq \mathcal{O}(\epsilon^{-2})$. The convergence analysis **only** supposes the average smooth assumption (which is weaker than Lipschitz continuous assumption on each component gradient) in the non-convex case (Theorem 1). We extend this result to the *mini-batch* case (Theorem 2).
- *Smooth Convex.* In order to complete the picture, we study SARAH+ [14] which was designed as a variant of SARAH for convex optimization. We propose a novel variant of SARAH+ called SARAH++. Here, we study the *iteration complexity* measured by the total number of iterations (which counts one full gradient computation as adding one iteration to the complexity) – and leave an analysis of the total complexity as an open problem. For SARAH++ we show a sublinear convergence rate in the general convex case (Theorem 3) and a linear convergence rate in the strongly convex case (Theorem 4). SARAH itself may already lead to good convergence and there may no need to introduce SARAH++; in numerical experiments we show the advantage of SARAH++ over SARAH. We further propose a practical version called *SARAH Adaptive* which improves the performance of SARAH and SARAH++ for convex problems – numerical experiments on various data sets show good overall

performance.

- For the convergence analysis of SARAH for the non-convex case and SARAH++ for the convex case we show that the analysis generalizes the total complexity of Gradient Descent (GD) (Remarks 1 and 2), i.e., the analysis reproduces known total complexity results of GD. Up to the best of our knowledge, this is the first variance reduction method having this property.

2 Non-Convex Case: Convergence Analysis of SARAH

SARAH is very different from other algorithms since it has a *biased* estimator of the gradient. Therefore, in order to analyze SARAH's convergence rate, it is non-trivial to use existing proof techniques from unbiased estimator algorithms such as SGD, SAGA, and SVRG.

2.1 A single batch case

We start analyzing SARAH (Algorithm 1) for the case where we choose a single sample i_t uniformly at random from $[n]$ in the inner loop.

Lemma 1. *Suppose that Assumption 1 holds. Consider a single outer loop iteration in SARAH (Algorithm 1) with $\eta \leq \frac{2}{L(\sqrt{1+4m}+1)}$. Then, for any $s \geq 1$, we have*

$$\mathbb{E}[F(w_{m+1}^{(s)})] \leq \mathbb{E}[F(w_0^{(s)})] - \frac{\eta}{2} \sum_{t=0}^m \mathbb{E}[\|\nabla F(w_t^{(s)})\|^2]. \quad (6)$$

The above result is for a single outer loop iteration of SARAH, which includes a full gradient step together with the inner loop. Since the outer loop iteration concludes with $\tilde{w}_s = w_{m+1}^{(s)}$, and $\tilde{w}_{s-1} = w_0^{(s)}$, we have

$$\mathbb{E}[F(\tilde{w}_s)] \leq \mathbb{E}[F(\tilde{w}_{s-1})] - \frac{\eta}{2} \sum_{t=0}^m \mathbb{E}[\|\nabla F(w_t^{(s)})\|^2].$$

Summing over $1 \leq s \leq S$ gives

$$\mathbb{E}[F(\tilde{w}_S)] \leq \mathbb{E}[F(\tilde{w}_0)] - \frac{\eta}{2} \sum_{s=1}^S \sum_{t=0}^m \mathbb{E}[\|\nabla F(w_t^{(s)})\|^2]. \quad (7)$$

This proves our main result:

Theorem 1 (Smooth nonconvex). *Suppose that Assumption 1 holds. Consider SARAH (Algorithm 1) with $\eta \leq \frac{2}{L(\sqrt{1+4m}+1)}$. Then, for any given \tilde{w}_0 , we have*

$$\frac{1}{(m+1)S} \sum_{s=1}^S \sum_{t=0}^m \mathbb{E}[\|\nabla F(w_t^{(s)})\|^2] \leq \frac{2}{\eta[(m+1)S]} [F(\tilde{w}_0) - F^*],$$

where F^* is any lower bound of F , and $w_t^{(s)}$ is the result of the t -th iteration in the s -th outer loop.

The proof easily follows from (7) since F^* is a lower bound of F (that is, $\mathbb{E}[F(\tilde{w}_S)] \geq F^*$). We note that the term

$$\frac{1}{(m+1)S} \sum_{s=1}^S \sum_{t=0}^m \mathbb{E}[\|\nabla F(w_t^{(s)})\|^2]$$

is simply the average of the expectation of the squared norms of the gradients of all the iteration results generated by SARAH. For nonconvex problems, our goal is to achieve

$$\frac{1}{(m+1)S} \sum_{s=1}^S \sum_{t=0}^m \mathbb{E}[\|\nabla F(w_t^{(s)})\|^2] \leq \epsilon.$$

We note that, for simplicity, if \bar{w}_s is chosen uniformly at random from all the iterations generated by SARAH, we are able to have accuracy $\mathbb{E}[\|\nabla F(\bar{w}_s)\|^2] \leq \epsilon$.

Corollary 1. *Suppose that Assumption 1 holds. Consider SARAH (Algorithm 1) with $\eta = \mathcal{O}(\frac{1}{L\sqrt{m+1}})$ where m is the inner loop size. Then, in order to achieve an ϵ -accurate solution, the total complexity is*

$$\mathcal{O}\left(\left[\left(\frac{n+2m}{\sqrt{m+1}}\right)\frac{1}{\epsilon}\right] \vee [n+2m]\right).$$

The total complexity can be minimized over the inner loop size m . By choosing $m = n$, we achieve the minimal total complexity:

Corollary 2. *Suppose that Assumption 1 holds. Consider SARAH (Algorithm 1) with $\eta = \mathcal{O}(\frac{1}{L\sqrt{m+1}})$ where m is the inner loop size and chosen equal to $m = n$. Then, in order to achieve an ϵ -accurate solution, the total complexity is*

$$\mathcal{O}\left(\frac{\sqrt{n}}{\epsilon} \vee n\right).$$

Remark 1. *The total complexity in Corollary 1 covers all choices for the inner loop size m . For example, in the case of $m = 0$, SARAH recovers the Gradient Descent (GD) algorithm which has total complexity $\mathcal{O}(\frac{n}{\epsilon})$. Theorem 1 for $m = 0$ also recovers the requirement on the learning rate for GD, which is $\eta \leq \frac{1}{L}$.*

The above results explain the relationship between SARAH and GD and explains the advantages of the inner loop and outer loop of SARAH. SARAH becomes more beneficial in ML applications where n is large.

2.2 Mini-batch case

The above results can be extended to the *mini-batch* case where instead of choosing a single sample i_t , we choose b samples uniformly at random from $[n]$ for updating v_t in the inner loop. We then replace v_t in Algorithm 1 by

$$v_t^{(s)} = \frac{1}{b} \sum_{i \in I_t} [\nabla f_i(w_t^{(s)}) - \nabla f_i(w_{t-1}^{(s)})] + v_{t-1}^{(s)}, \quad (8)$$

where we choose a mini-batch $I_t \subseteq [n]$ of size b uniformly at random at each iteration of the inner loop. The result of Theorem 1 generalizes as follows.

Theorem 2 (Smooth nonconvex with mini-batch). *Suppose that Assumption 1 holds. Consider SARAH (Algorithm 1) by replacing v_t in the inner loop size by (8) with*

$$\eta \leq \frac{2}{L \left(\sqrt{1 + \frac{4m}{b} \left(\frac{n-b}{n-1} \right)} + 1 \right)}.$$

Then, for any given \tilde{w}_0 , we have

$$\frac{1}{(m+1)S} \sum_{s=1}^S \sum_{t=0}^m \mathbb{E}[\|\nabla F(w_t^{(s)})\|^2] \leq \frac{2}{\eta[(m+1)S]} [F(\tilde{w}_0) - F^*],$$

where F^ is any lower bound of F , and $w_t^{(s)}$ is the t -th iteration in the s -th outer loop.*

We can again derive similar corollaries as was done for Theorem 1.

Corollary 3. *For the conditions in Theorem 2, in order to achieve an ϵ -accurate solution, the total complexity is*

$$\mathcal{O} \left(\left[\left(\frac{n+2bm}{m+1} \right) \left(\sqrt{1 + \frac{4m}{b} \left(\frac{n-b}{n-1} \right)} \right) \frac{1}{\epsilon} \right] \vee [n+2bm] \right).$$

Corollary 4. *For the conditions in Theorem 2 and Corollary 3 with $b = n^\alpha$ and $m = n^\beta$ where $\alpha + \beta = 1$ with $\beta \geq 1/2$ and $0 \leq \alpha \leq 1/2$, in order to achieve an ϵ -accurate solution, the total complexity is*

$$\mathcal{O} \left(\frac{\sqrt{n}}{\epsilon} \vee n \right).$$

3 Convex Case: SARAH++: A New Variant of SARAH+

In this section, we propose a new variant of SARAH+ (Algorithm 2) [14], called SARAH++ (Algorithm 3), for convex problems of form (1).

Different from SARAH, SARAH+ provides a stopping criteria for the inner loop; as soon as

$$\|v_{t-1}^{(s)}\|^2 \leq \gamma \|v_0^{(s)}\|^2,$$

the inner loop finishes. This idea originates from the property of SARAH that, for each outer loop iteration s , $\mathbb{E}[\|v_t^{(s)}\|^2] \rightarrow 0$ as $t \rightarrow \infty$ in the strongly convex case (Theorems 1a and 1b in [14]). Therefore, it does not make any sense to update with tiny steps when $\|v_t^{(s)}\|^2$ is small. (We note that SVRG [7] does not have this property.) SARAH+ suggests to empirically choose parameter $\gamma = 1/8$ [14] without theoretical guarantee.

Here, we modify SARAH+ (Algorithm 2) into SARAH++ (Algorithm 3) by choosing the stopping criteria for the inner loop as

$$\|v_{t-1}^{(s)}\|^2 < \gamma \|v_0^{(s)}\|^2 \text{ where } \gamma \geq L\eta$$

and by introducing a stopping criteria for the outer loop.

Algorithm 2 SARAH+ [14]

Parameters: the learning rate $\eta > 0$, $0 < \gamma \leq 1$, the maximum inner loop size m , and the outer loop size S

Initialize: \tilde{w}_0

Iterate:

```

for  $s = 1, 2, \dots, S$  do
     $w_0^{(s)} = \tilde{w}_{s-1}$ 
     $v_0^{(s)} = \frac{1}{n} \sum_{i=1}^n \nabla f_i(w_0^{(s)})$ 
     $w_1^{(s)} = w_0^{(s)} - \eta v_0^{(s)}$ 
     $t = 1$ 
    while  $\|v_{t-1}^{(s)}\|^2 > \gamma \|v_0^{(s)}\|^2$  and  $t \leq m$  do
        Sample  $i_t$  uniformly at random from  $[n]$ 
         $v_t^{(s)} = \nabla f_{i_t}(w_t^{(s)}) - \nabla f_{i_t}(w_{t-1}^{(s)}) + v_{t-1}^{(s)}$ 
         $w_{t+1}^{(s)} = w_t^{(s)} - \eta v_t^{(s)}$ 
         $t \leftarrow t + 1$ 
    end while
    Set  $\tilde{w}_s = w_t^{(s)}$ 
end for

```

3.1 Details SARAH++ and Convergence Analysis

Before analyzing and explaining SARAH++ in detail, we introduce the following assumptions used in this section.

Assumption 2 (L -smooth). *Each $f_i : \mathbb{R}^d \rightarrow \mathbb{R}$, $i \in [n]$, is L -smooth, i.e., there exists a constant $L > 0$ such that, $\forall w, w' \in \mathbb{R}^d$,*

$$\|\nabla f_i(w) - \nabla f_i(w')\| \leq L\|w - w'\|. \quad (9)$$

Assumption 3 (μ -strongly convex). *The function $F : \mathbb{R}^d \rightarrow \mathbb{R}$, is μ -strongly convex, i.e., there exists a constant $\mu > 0$ such that $\forall w, w' \in \mathbb{R}^d$,*

$$F(w) \geq F(w') + \nabla F(w')^T(w - w') + \frac{\mu}{2}\|w - w'\|^2.$$

Under Assumption 3, let us define the (unique) optimal solution of (1) as w_* . Then strong convexity of F implies that

$$2\mu[F(w) - F(w_*)] \leq \|\nabla F(w)\|^2, \quad \forall w \in \mathbb{R}^d. \quad (10)$$

We note here, for future use, that for strongly convex functions of the form (1), arising in machine learning applications, the condition number is defined as $\kappa \stackrel{\text{def}}{=} L/\mu$. Assumption 3 covers a wide range of problems, e.g. l_2 -regularized empirical risk minimization problems with convex losses.

We separately assume the special case of strong convexity of all f_i 's with $\mu = 0$, called the general convexity assumption, which we will use for convergence analysis.

Assumption 4. *Each function $f_i : \mathbb{R}^d \rightarrow \mathbb{R}$, $i \in [n]$, is convex, i.e.,*

$$f_i(w) \geq f_i(w') + \nabla f_i(w')^T(w - w').$$

Algorithm 3 SARAH++

Parameters: The controlled factor $0 < \gamma \leq 1$, the learning rate $0 < \eta \leq \frac{\gamma}{L}$, the total iteration $T > 0$, and the maximum inner loop size $m \leq T$.

Initialize: \tilde{w}_0

$G = 0$

Iterate:

$s = 0$

while $G < T$ **do**

- $s \leftarrow s + 1$
- $w_0^{(s)} = \tilde{w}_{s-1}$
- $v_0^{(s)} = \frac{1}{n} \sum_{i=1}^n \nabla f_i(w_0^{(s)})$
- $t = 0$
- while** $\|v_t^{(s)}\|^2 \geq \gamma \|v_0^{(s)}\|^2$ **and** $t \leq m$ **do**

 - $w_{t+1}^{(s)} = w_t^{(s)} - \eta v_t^{(s)}$
 - $t \leftarrow t + 1$
 - if** $m \neq 0$ **then**

 - Sample i_t uniformly at random from $[n]$
 - $v_t^{(s)} = \nabla f_{i_t}(w_t^{(s)}) - \nabla f_{i_t}(w_{t-1}^{(s)}) + v_{t-1}^{(s)}$

 - end if**

- end while**
- $T_s = t$
- $\tilde{w}_s = w_{T_s}^{(s)}$
- $G \leftarrow G + T_s$

end while

$S = s$

Set $\hat{w} = \tilde{w}_S$

SARAH++ is motivated by the following lemma.

Lemma 2. Suppose that Assumptions 2 and 4 hold. Consider a single outer loop iteration in SARAH (Algorithm 1) with $\eta \leq \frac{1}{L}$. Then, for $t \geq 0$ and any $s \geq 1$, we have

$$\mathbb{E}[F(w_{t+1}^{(s)}) - F(w_*)] \leq \mathbb{E}[F(w_t^{(s)}) - F(w_*)] - \frac{\eta}{2} \mathbb{E}[\|\nabla F(w_t^{(s)})\|^2] + \frac{\eta}{2} \left(L\eta \mathbb{E}[\|v_0^{(s)}\|^2] - \mathbb{E}[\|v_t^{(s)}\|^2] \right), \quad (11)$$

where w_* is any optimal solution of F .

Clearly, if

$$L\eta \mathbb{E}[\|v_0^{(s)}\|^2] - \mathbb{E}[\|v_t^{(s)}\|^2] \leq \gamma \mathbb{E}[\|v_0^{(s)}\|^2] - \mathbb{E}[\|v_t^{(s)}\|^2] \leq 0,$$

where $\eta \leq \frac{\gamma}{L}$, inequality (11) implies

$$\mathbb{E}[F(w_{t+1}^{(s)}) - F(w_*)] \leq \mathbb{E}[F(w_t^{(s)}) - F(w_*)] - \frac{\eta}{2} \mathbb{E}[\|\nabla F(w_t^{(s)})\|^2].$$

For this reason, we choose the stopping criteria for the inner loop in SARAH++ as $\|v_t^{(s)}\|^2 < \gamma \|v_0^{(s)}\|^2$ with $\gamma \geq L\eta$. Unlike SARAH+, for analyzing the convergence rate γ can be as small as $L\eta$.

The above discussion leads to SARAH++ (Algorithm 3). In order to analyze its convergence for convex problems, we define random variable T_s as the stopping time of the inner loop in the s -th outer iteration:

$$T_s = \min \left\{ \min_{t \geq 0} \left\{ t : \|v_t^{(s)}\|^2 < \gamma \|v_0^{(s)}\|^2 \right\}, m + 1 \right\}, \quad s = 1, 2, \dots$$

Note that T_s is at least 1 since at $t = 0$, the condition $\|v_0^{(s)}\|^2 \geq \gamma \|v_0^{(s)}\|^2$ always holds (and $m \geq 0$).

Let random variable S be the stopping time of the outer iterations as a function of an algorithm parameter $T > 0$:

$$S = \min_{\hat{S}} \left\{ \hat{S} : \sum_{s=1}^{\hat{S}} T_s \geq T \right\}.$$

Notice that SARAH++ maintains a running sum $G = \sum_{j=1}^s T_i$ against which parameter T is compared in the stopping criteria of the outer loop.

For the general convex case which supposes Assumption 4 in addition to smoothness we have the next theorem.

Theorem 3 (Smooth general convex). *Suppose that Assumptions 2 and 4 hold. Consider SARAH++ (Algorithm 3) with $\eta \leq \frac{\gamma}{L}$, $0 < \gamma \leq 1$. Then,*

$$\mathbb{E} \left[\frac{1}{T_1 + \dots + T_S} \sum_{s=1}^S \sum_{t=0}^{T_s-1} \mathbb{E}[\|\nabla F(w_t^{(s)})\|^2 | T_1, \dots, T_S] \right],$$

the expectation of the average of the squared norm of the gradients of all iterations generated by SARAH++, is bounded by

$$\frac{2}{T\eta} [F(\tilde{w}_0) - F(w_*)].$$

The theorem leads to the next corollary about iteration complexity, i.e., we bound T which is the total number of iterations performed by the inner loop across all outer loop iterations. This is different from the total complexity since T does not separately count the n gradient evaluations when the full gradient is computed in the outer loop.

Corollary 5 (Smooth general convex). *For the conditions in Theorem 3 with $\eta = \mathcal{O}(\frac{1}{L})$, we achieve an ϵ -accurate solution after $\mathcal{O}(\frac{1}{\epsilon})$ inner loop iterations.*

By supposing Assumption 3 in addition to the smoothness and general convexity assumptions, we can prove a linear convergence rate. For strongly convex objective functions we have the following result.

Theorem 4 (Smooth strongly convex). *Suppose that Assumptions 2, 3 and 4 hold. Consider SARAH++ (Algorithm 3) with $\eta \leq \frac{\gamma}{L}$, $0 < \gamma \leq 1$. Then, for the final output \hat{w} of SARAH++, we have*

$$\mathbb{E}[F(\hat{w}) - F(w_*)] \leq (1 - \mu\eta)^T [F(\tilde{w}_0) - F(w_*)]. \quad (12)$$

This leads to the following iteration complexity.

Corollary 6 (Smooth strongly convex). *For the conditions in Theorem 4 with $\eta = \mathcal{O}(\frac{1}{L})$, we achieve $\mathbb{E}[F(\hat{w}) - F(w_*)] \leq \epsilon$ after $\mathcal{O}(\kappa \log(\frac{1}{\epsilon}))$ total iterations, where $\kappa = L/\mu$ is the condition number.*

Remark 2. *The proofs of the above results hold for any $m \leq T$. If we choose $m = 0$, then SARAH++ reduces to the Gradient Descent algorithm since the inner “while” loop stops right after updating $w_1^{(s)} = w_0^{(s)} - \eta v_0^{(s)}$. In this case, Corollaries 5 and 6 recover the rate of convergence and complexity of GD.*

In this section, we showed that SARAH++ has a guarantee of theoretical convergence (see Theorems 3 and 4) while SARAH+ does not have such a guarantee.

An interesting open question we would like to discuss here is the total complexity of SARAH++. Although we have shown the convergence results of SARAH++ in terms of the iteration complexity, the total complexity which is computed as the total number of evaluations of the component gradient functions still remains an open question. It is clear that the total complexity must depend on the learning rate η (or γ) – the factor that decides when to stop the inner iterations.

We note that T can be “closely” understood as the total number of updates $w_{t+1}^{(s)}$ of the algorithm. The total complexity is equal to $\sum_{i=1}^S (n+2(T_i-1))$. For the special case $T_i = 1, i = 1, \dots, S$, the algorithm recovers the GD algorithm with $T = \sum_{i=1}^S T_s = S$. Since each full gradient takes n gradient evaluations, the total complexity for this case is equal to $nS = \mathcal{O}(\frac{n}{\epsilon})$ (in the general convex case) and $nS = \mathcal{O}(n\kappa \log(\frac{1}{\epsilon}))$ (in the strongly convex case).

However, it is non-trivial to derive the total complexity of SARAH++ since it should depend on the learning rate η . We leave this question as an open direction for future research.

3.2 Numerical Experiments

Paper [14] provides experiments showing good overall performance of SARAH over other algorithms such as SGD [17], SAG [9], SVRG [7], etc. For this reason, we provide experiments comparing SARAH++ directly with SARAH. We notice that SARAH (with multiple outer loops) like SARAH++ has theoretical guarantees with sublinear convergence for general convex and linear convergence for strongly convex problems as proved in [14]. Because of these theoretical guarantees (which SARAH+ does not have), SARAH itself may already perform well for convex problems and the question is whether SARAH++ offers an advantage.

We consider ℓ_2 -regularized logistic regression problems with

$$f_i(w) = \log(1 + \exp(-y_i \langle x_i, w \rangle)) + \frac{\lambda}{2} \|w\|^2, \quad (13)$$

where $\{x_i, y_i\}_{i=1}^n$ is the training data and the regularization parameter λ is set to $1/n$, a widely-used value in literature [9, 14]. The condition number is equal to $\kappa = L/\mu = n$. We conducted experiments to demonstrate the advantage in performance of SARAH++ over SARAH for convex problems on popular data sets including *covtype* ($n = 406,708$ training data; estimated $L \simeq 1.90$) and *ijcnn1* ($n = 91,701$ training data; estimated $L \simeq 1.77$) from LIBSVM [4].

Figure 1 shows comparisons between SARAH++ and SARAH for different values of learning rate η . We depicted the value of $\log[F(w) - F(w_*)]$ (i.e. $F(w) - F(w_*)$ in log scale) for the y -axis and “number of effective passes” (or number of epochs, where an epoch is the equivalent of n component gradient

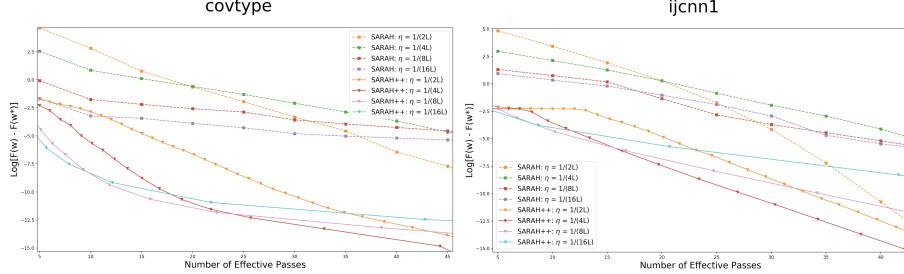


Figure 1: Comparisons of $\log[F(w) - F(w_*)]$ between SARAH++ and SARAH with different learning rates on *covtype* and *ijcn1* datasets

evaluations or one full gradient computation) for the x -axis. For SARAH, we choose the outer loop size $S = 10$ and tune the inner loop size $m = \{0.5n, n, 2n, 3n, 4n\}$ to achieve the best performance. The optimal solution w_* of the strongly convex problem in (13) is found by using Gradient Descent with stopping criterion $\|\nabla F(w)\|^2 \leq 10^{-15}$. We observe that, SARAH++ achieves improved overall performance compared to regular SARAH as shown in Figure 1. From the experiments we see that the stopping criteria $\|v_t^{(s)}\|^2 < \gamma \|v_0^{(s)}\|^2$ ($\gamma = L\eta$) of SARAH++ is indeed important. The stopping criteria helps the inner loop to prevent updating tiny redundant steps.

3.3 SARAH Adaptive: A New Practical Variant

We now propose a practical adaptive method which aims to improve performance. Although we do not have any theoretical result for this adaptive method, numerical experiments are very promising and they heuristically show the improved performance on different data sets.

Algorithm 4 SARAH Adaptive

Parameters: The maximum inner loop size m , and the outer loop size S , the factor $0 < \gamma \leq 1$.

Initialize: \tilde{w}_0

Iterate:

for $s = 1, 2, \dots, S$ **do**

$$w_0^{(s)} = \tilde{w}_{s-1}$$

$$v_0^{(s)} = \frac{1}{n} \sum_{i=1}^n \nabla f_i(w_0^{(s)})$$

$$t = 0$$

while $\|v_t^{(s)}\|^2 \geq \gamma \|v_0^{(s)}\|^2$ **and** $t \leq m$ **do**

$$\eta_t^{(s)} = \frac{1}{L} \cdot \frac{\|v_t^{(s)}\|^2}{\|v_0^{(s)}\|^2} \text{ (adaptive)}$$

$$w_{t+1}^{(s)} = w_t^{(s)} - \eta_t^{(s)} v_t^{(s)}$$

$$t \leftarrow t + 1$$

if $m \neq 0$ **then**

 Sample i_t uniformly at random from $[n]$

$$v_t^{(s)} = \nabla f_{i_t}(w_t^{(s)}) - \nabla f_{i_t}(w_{t-1}^{(s)}) + v_{t-1}^{(s)}$$

end if

end while

$$\text{Set } \tilde{w}_s = w_t^{(s)}$$

end for

The motivation of this algorithm comes from the intuition of Lemma 2 (for convex optimization). For a

single outer loop with $\eta \leq \frac{1}{L}$, (11) holds for SARAH (Algorithm 1). Hence, for any s , we intentionally choose $\eta = \eta_t^{(s)} = \frac{\|v_t^{(s)}\|^2}{L\|v_0^{(s)}\|^2}$ such that $L\eta\mathbb{E}[\|v_0^{(s)}\|^2] - \mathbb{E}[\|v_t^{(s)}\|^2] = 0$. Since $\|v_t^{(s)}\|^2 \leq \|v_0^{(s)}\|^2$, $t \geq 0$, in [14] for convex problems, we have $\eta_t^{(s)} \leq \frac{1}{L}$, $t \geq 0$. We also stop the inner loop by the stopping criteria $\|v_t^{(s)}\|^2 < \gamma\|v_0^{(s)}\|^2$ for some $0 < \gamma \leq 1$. SARAH Adaptive is given in detail in Algorithm 4 without convergence analysis.

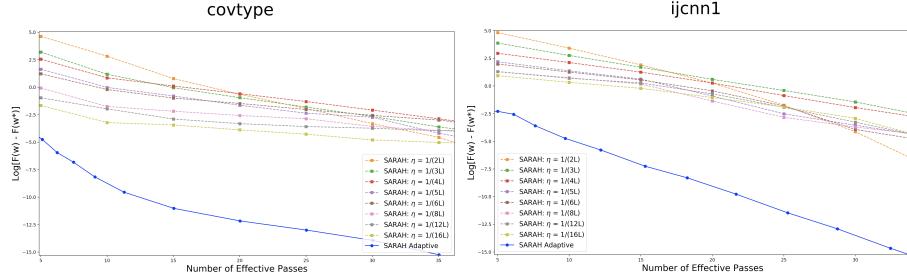


Figure 2: Comparisons of $\log[F(w) - F(w_*)]$ between SARAH Adaptive and SARAH with different learning rates on *covtype* and *ijcn1* datasets

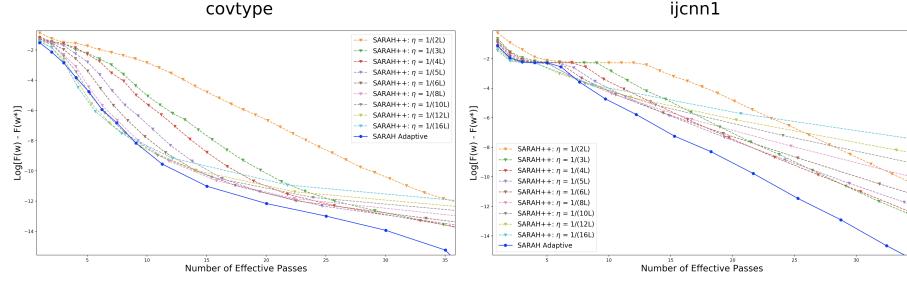


Figure 3: Comparisons of $\log[F(w) - F(w_*)]$ between SARAH Adaptive and SARAH++ with different learning rates on *covtype* and *ijcn1* datasets

We have conducted numerical experiments on the same datasets and problems as introduced in the previous subsection. Figures 2 and 3 show the comparison between SARAH Adaptive and SARAH and SARAH++ for different values of η . We observe that SARAH Adaptive has an improved performance over SARAH and SARAH++ (without tuning learning rate). We also present the numerical performance of SARAH Adaptive for different values of γ in Appendix. We also present the numerical performance of SARAH Adaptive for different values of γ in Appendix.

We note that additional experiments in this section on more data sets are performed in Appendix.

References

- [1] Zeyuan Allen-Zhu. Natasha 2: Faster non-convex optimization than sgd. *arXiv preprint arXiv:1708.08694*, 2017.
- [2] Zeyuan Allen-Zhu. Natasha: Faster non-convex stochastic optimization via strongly non-convex parameter. *arXiv preprint arXiv:1702.00763*, 2017.
- [3] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *arXiv:1606.04838*, 2016.

- [4] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- [5] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *NIPS*, pages 1646–1654, 2014.
- [6] Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. Spider: Near-optimal non-convex optimization via stochastic path integrated differential estimator. *arXiv preprint arXiv:1807.01695*, 2018.
- [7] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *NIPS*, pages 315–323, 2013.
- [8] Jakub Konečný and Peter Richtárik. Semi-stochastic gradient descent methods. *arXiv:1312.1666*, 2013.
- [9] Nicolas Le Roux, Mark Schmidt, and Francis Bach. A stochastic gradient method with an exponential convergence rate for finite training sets. In *NIPS*, pages 2663–2671, 2012.
- [10] Lihua Lei, Cheng Ju, Jianbo Chen, and Michael I Jordan. Non-convex finite-sum optimization via SCSG methods. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 2348–2358. Curran Associates, Inc., 2017.
- [11] Julien Mairal. Optimization with first-order surrogate functions. In *ICML*, pages 783–791, 2013.
- [12] Yurii Nesterov. *Introductory lectures on convex optimization : a basic course*. Applied optimization. Kluwer Academic Publ., Boston, Dordrecht, London, 2004.
- [13] Lam Nguyen, Phuong Ha Nguyen, Marten van Dijk, Peter Richtarik, Katya Scheinberg, and Martin Takac. SGD and Hogwild! convergence without the bounded gradients assumption. In *ICML*, 2018.
- [14] Lam M. Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. SARAH: A novel method for machine learning problems using stochastic recursive gradient. In *ICML*, 2017.
- [15] Lam M. Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. Stochastic recursive gradient algorithm for nonconvex optimization. *CoRR*, abs/1705.07261, 2017.
- [16] Sashank J. Reddi, Ahmed Hefny, Suvrit Sra, Barnabás Póczos, and Alexander J. Smola. Stochastic variance reduction for nonconvex optimization. In *ICML*, pages 314–323, 2016.
- [17] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.
- [18] Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, pages 1–30, 2016.
- [19] Shai Shalev-Shwartz, Yoram Singer, Nathan Srebro, and Andrew Cotter. Pegasos: Primal estimated sub-gradient solver for SVM. *Mathematical Programming*, 127(1):3–30, 2011.
- [20] Shai Shalev-Shwartz and Tong Zhang. Stochastic dual coordinate ascent methods for regularized loss. *Journal of Machine Learning Research*, 14(1):567–599, 2013.

- [21] Zhe Wang, Kaiyi Ji, Yi Zhou, Yingbin Liang, and Vahid Tarokh. Spiderboost: A class of faster variance-reduced algorithms for nonconvex optimization. *arXiv preprint arXiv:1810.10690*, 2018.
- [22] Dongruo Zhou, Pan Xu, and Quanquan Gu. Stochastic nested variance reduction for nonconvex optimization. *arXiv preprint arXiv:1806.07811*, 2018.

Appendix

Useful Existing Results

Lemma 3 (Lemma 2 in [14] (or in [15])). Consider $v_t^{(s)}$ defined by (2) (or (8)) in SARAH (Algorithm 1) for any $s \geq 1$. Then for any $t \geq 1$,

$$\mathbb{E}[\|\nabla F(w_t^{(s)}) - v_t^{(s)}\|^2] = \sum_{j=1}^t \mathbb{E}[\|v_j^{(s)} - v_{j-1}^{(s)}\|^2] - \sum_{j=1}^t \mathbb{E}[\|\nabla F(w_j^{(s)}) - \nabla F(w_{j-1}^{(s)})\|^2]. \quad (14)$$

Lemma 4 (Lemma 3 in [14]). Suppose that Assumptions 2 and 4 hold. Consider $v_t^{(s)}$ defined as (2) in SARAH (Algorithm 1) with $\eta < 2/L$ for any $s \geq 1$. Then we have that for any $t \geq 0$,

$$\mathbb{E}[\|\nabla F(w_t^{(s)}) - v_t^{(s)}\|^2] \leq \frac{\eta L}{2 - \eta L} \left[\mathbb{E}[\|v_0^{(s)}\|^2] - \mathbb{E}[\|v_t^{(s)}\|^2] \right]. \quad (15)$$

Nonconvex SARAH

Proof of Lemma 1

Lemma 1. Suppose that Assumption 1 holds. Consider SARAH (Algorithm 1) within a single outer loop with $\eta \leq \frac{2}{L(\sqrt{1+4m+1})}$. Then, for any $s \geq 1$, we have

$$\mathbb{E}[F(w_{m+1}^{(s)})] \leq \mathbb{E}[F(w_0^{(s)})] - \frac{\eta}{2} \sum_{t=0}^m \mathbb{E}[\|\nabla F(w_t^{(s)})\|^2].$$

Proof. We use some parts of the proof in [15]. By Assumption 1 and $w_{t+1}^{(s)} = w_t^{(s)} - \eta v_t^{(s)}$, for any $s \geq 1$, we have

$$\begin{aligned} \mathbb{E}[F(w_{t+1}^{(s)})] &\stackrel{(5)}{\leq} \mathbb{E}[F(w_t^{(s)})] - \eta \mathbb{E}[\nabla F(w_t^{(s)})^T v_t^{(s)}] + \frac{L\eta^2}{2} \mathbb{E}[\|v_t^{(s)}\|^2] \\ &= \mathbb{E}[F(w_t^{(s)})] - \frac{\eta}{2} \mathbb{E}[\|\nabla F(w_t^{(s)})\|^2] + \frac{\eta}{2} \mathbb{E}[\|\nabla F(w_t^{(s)}) - v_t^{(s)}\|^2] - \left(\frac{\eta}{2} - \frac{L\eta^2}{2} \right) \mathbb{E}[\|v_t^{(s)}\|^2], \end{aligned} \quad (16)$$

where the last equality follows from the fact $a^T b = \frac{1}{2} [\|a\|^2 + \|b\|^2 - \|a - b\|^2]$, for any $a, b \in \mathbb{R}^d$. By summing over $t = 0, \dots, m$, we have

$$\begin{aligned} \mathbb{E}[F(w_{m+1}^{(s)})] &\leq \mathbb{E}[F(w_0^{(s)})] - \frac{\eta}{2} \sum_{t=0}^m \mathbb{E}[\|\nabla F(w_t^{(s)})\|^2] \\ &\quad + \frac{\eta}{2} \left(\sum_{t=0}^m \mathbb{E}[\|\nabla F(w_t^{(s)}) - v_t^{(s)}\|^2] - (1 - L\eta) \sum_{t=0}^m \mathbb{E}[\|v_t^{(s)}\|^2] \right). \end{aligned} \quad (17)$$

Now, we would like to determine η such that the expression in (17)

$$\sum_{t=0}^m \mathbb{E}[\|\nabla F(w_t^{(s)}) - v_t^{(s)}\|^2] - (1 - L\eta) \sum_{t=0}^m \mathbb{E}[\|v_t^{(s)}\|^2] \leq 0.$$

Let $\mathcal{F}_j = \sigma(w_0, w_1, \dots, w_j)$ be the σ -algebra generated by w_0, w_1, \dots, w_j . Note that \mathcal{F}_j also contains all information of v_0, \dots, v_{j-1} . We have

$$\begin{aligned} \mathbb{E}[\|v_j^{(s)} - v_{j-1}^{(s)}\|^2 | \mathcal{F}_j] &\stackrel{(2)}{=} \mathbb{E}[\|\nabla f_{i_j}(w_j^{(s)}) - \nabla f_{i_j}(w_{j-1}^{(s)})\|^2 | \mathcal{F}_j] \\ &\stackrel{(4)}{\leq} L^2 \|w_j^{(s)} - w_{j-1}^{(s)}\|^2 = L^2 \eta^2 \|v_{j-1}^{(s)}\|^2, \quad j \geq 1. \end{aligned}$$

Taking the expectations to both sides yields

$$\mathbb{E}[\|v_j^{(s)} - v_{j-1}^{(s)}\|^2] \leq L^2 \eta^2 \mathbb{E}[\|v_{j-1}^{(s)}\|^2]. \quad (18)$$

Hence, by Lemma 3, we have

$$\mathbb{E}[\|\nabla F(w_t^{(s)}) - v_t^{(s)}\|^2] \leq \sum_{j=1}^t \mathbb{E}[\|v_j^{(s)} - v_{j-1}^{(s)}\|^2] \stackrel{(18)}{\leq} L^2 \eta^2 \sum_{j=1}^t \mathbb{E}[\|v_{j-1}^{(s)}\|^2].$$

Note that $\|\nabla F(w_0^{(s)}) - v_0^{(s)}\|^2 = 0$. Hence, by summing over $t = 0, \dots, m$ ($m \geq 1$), we have

$$\sum_{t=0}^m \mathbb{E}[\|\nabla F(w_t^{(s)}) - v_t^{(s)}\|^2] \leq L^2 \eta^2 \left[m \mathbb{E}\|v_0^{(s)}\|^2 + (m-1) \mathbb{E}\|v_1^{(s)}\|^2 + \dots + \mathbb{E}\|v_{m-1}^{(s)}\|^2 \right].$$

By choosing $\eta \leq \frac{2}{L(\sqrt{1+4m+1})}$, we have

$$\begin{aligned} &\sum_{t=0}^m \mathbb{E}[\|\nabla F(w_t^{(s)}) - v_t^{(s)}\|^2] - (1 - L\eta) \sum_{t=0}^m \mathbb{E}[\|v_t^{(s)}\|^2] \\ &\leq L^2 \eta^2 \left[m \mathbb{E}\|v_0^{(s)}\|^2 + (m-1) \mathbb{E}\|v_1^{(s)}\|^2 + \dots + \mathbb{E}\|v_{m-1}^{(s)}\|^2 \right] - (1 - L\eta) \left[\mathbb{E}\|v_0^{(s)}\|^2 + \mathbb{E}\|v_1^{(s)}\|^2 + \dots + \mathbb{E}\|v_m^{(s)}\|^2 \right] \\ &\leq \left[L^2 \eta^2 m - (1 - L\eta) \right] \sum_{t=1}^m \mathbb{E}[\|v_{t-1}^{(s)}\|^2] \leq 0, \end{aligned} \quad (19)$$

since $\eta = \frac{2}{L(\sqrt{1+4m+1})}$ is a root of equation $L^2 \eta^2 m - (1 - L\eta) = 0$. Therefore, with $\eta \leq \frac{2}{L(\sqrt{1+4m+1})}$, we have

$$\mathbb{E}[F(w_{m+1}^{(s)})] \leq \mathbb{E}[F(w_0^{(s)})] - \frac{\eta}{2} \sum_{t=0}^m \mathbb{E}[\|\nabla F(w_t^{(s)})\|^2].$$

□

Proof of Corollary 1

Corollary 1. Suppose that Assumption 1 holds. Consider SARAH (Algorithm 1) with $\eta = \mathcal{O}(\frac{1}{L\sqrt{m+1}})$ where m is the inner loop size. Then, in order to achieve ϵ -accurate solution, the total complexity is $\mathcal{O}\left(\left[\left(\frac{n+2m}{\sqrt{m+1}}\right)\frac{1}{\epsilon}\right] \vee [n+2m]\right)$.

Proof. In order to achieve

$$\frac{1}{(m+1)S} \sum_{s=1}^S \sum_{t=0}^m \mathbb{E}[\|\nabla F(w_t^{(s)})\|^2] \leq \epsilon,$$

we need

$$\frac{2}{\eta[(m+1)S]} [F(\tilde{w}_0) - F^*] = \epsilon. \quad (20)$$

Let us choose η such that

$$\eta = \frac{2}{L(3\sqrt{m+1})} \stackrel{m \geq 0}{\leq} \frac{2}{L(2\sqrt{m+1} + 1)} = \frac{2}{L(\sqrt{4m+4} + 1)} \leq \frac{2}{L(\sqrt{1+4m} + 1)}. \quad (21)$$

Hence, in order to achieve (20), we need

$$S = \frac{2}{\eta[(m+1)\epsilon]} [F(\tilde{w}_0) - F^*] \stackrel{(21)}{=} \frac{3L[F(\tilde{w}_0) - F^*]}{(\sqrt{m+1})} \frac{1}{\epsilon} = \mathcal{O}\left(\left[\frac{1}{\sqrt{m+1}} \cdot \frac{1}{\epsilon}\right] \vee 1\right)$$

since $S \geq 1$. Therefore, the total complexity to achieve ϵ -accurate solution is

$$(n+2m)S = \mathcal{O}\left(\left[\left(\frac{n+2m}{\sqrt{m+1}}\right) \frac{1}{\epsilon}\right] \vee [n+2m]\right).$$

□

Proof of Theorem 2

Theorem 2 (Smooth nonconvex with mini-batch). *Suppose that Assumption 1 holds. Consider SARAH (Algorithm 1) by replacing v_t in the inner loop size by (8) with*

$$\eta \leq \frac{2}{L \left(\sqrt{1 + \frac{4m}{b} \left(\frac{n-b}{n-1} \right)} + 1 \right)}.$$

Then, for any given \tilde{w}_0 , we have

$$\frac{1}{(m+1)S} \sum_{s=1}^S \sum_{t=0}^m \mathbb{E}[\|\nabla F(w_t^{(s)})\|^2] \leq \frac{2}{\eta[(m+1)S]} [F(\tilde{w}_0) - F^*],$$

where F^* is any lower bound of F , and $w_t^{(s)}$ is the t -th iteration in the s -th outer loop.

Proof. Following the proof of Lemma 1, we would like to determine η such that the expression in (17)

$$\sum_{t=0}^m \mathbb{E}[\|\nabla F(w_t^{(s)}) - v_t^{(s)}\|^2] - (1 - L\eta) \sum_{t=0}^m \mathbb{E}[\|v_t^{(s)}\|^2] \leq 0.$$

Let

$$\xi_t = \nabla f_t(w_j^{(s)}) - \nabla f_t(w_{j-1}^{(s)}). \quad (22)$$

Let $\mathcal{F}_j = \sigma(w_0^{(s)}, I_1, I_2, \dots, I_{j-1})$ be the σ -algebra generated by $w_0^{(s)}, I_1, I_2, \dots, I_{j-1}$; $\mathcal{F}_0 = \mathcal{F}_1 = \sigma(w_0^{(s)})$. Note that \mathcal{F}_j also contains all the information of $w_0^{(s)}, \dots, w_j^{(s)}$ as well as $v_0^{(s)}, \dots, v_{j-1}^{(s)}$. We have

$$\begin{aligned}
& \mathbb{E}[\|v_j^{(s)} - v_{j-1}^{(s)}\|^2 | \mathcal{F}_j] - \|\nabla F(w_j^{(s)}) - \nabla F(w_{j-1}^{(s)})\|^2 \\
& \stackrel{(8)}{=} \mathbb{E}\left[\left\|\frac{1}{b} \sum_{i \in I_j} [\nabla f_i(w_j^{(s)}) - \nabla f_i(w_{j-1}^{(s)})]\right\|^2 | \mathcal{F}_j\right] - \left\|\frac{1}{n} \sum_{i=1}^n [\nabla f_i(w_j^{(s)}) - \nabla f_i(w_{j-1}^{(s)})]\right\|^2 \\
& \stackrel{(22)}{=} \mathbb{E}\left[\left\|\frac{1}{b} \sum_{i \in I_j} \xi_i\right\|^2 | \mathcal{F}_j\right] - \left\|\frac{1}{n} \sum_{i=1}^n \xi_i\right\|^2 \\
& = \frac{1}{b^2} \mathbb{E}\left[\sum_{i \in I_j} \sum_{k \in I_j} \xi_i^T \xi_k | \mathcal{F}_j\right] - \frac{1}{n^2} \sum_{i=1}^n \sum_{k=1}^n \xi_i^T \xi_k \\
& = \frac{1}{b^2} \mathbb{E}\left[\sum_{i \neq k \in I_j} \xi_i^T \xi_k + \sum_{i \in I_j} \xi_i^T \xi_i | \mathcal{F}_j\right] - \frac{1}{n^2} \sum_{i=1}^n \sum_{k=1}^n \xi_i^T \xi_k \\
& = \frac{1}{b^2} \left[\frac{b}{n(n-1)} \sum_{i \neq k} \xi_i^T \xi_k + \frac{b}{n} \sum_{i=1}^n \xi_i^T \xi_i \right] - \frac{1}{n^2} \sum_{i=1}^n \sum_{k=1}^n \xi_i^T \xi_k \\
& = \frac{1}{b^2} \left[\frac{b}{n(n-1)} \sum_{i=1}^n \sum_{k=1}^n \xi_i^T \xi_k + \left(\frac{b}{n} - \frac{b}{n(n-1)} \right) \sum_{i=1}^n \xi_i^T \xi_i \right] - \frac{1}{n^2} \sum_{i=1}^n \sum_{k=1}^n \xi_i^T \xi_k \\
& = \frac{1}{bn} \left[\left(\frac{(b-1)}{(n-1)} - \frac{b}{n} \right) \sum_{i=1}^n \sum_{k=1}^n \xi_i^T \xi_k + \frac{(n-b)}{(n-1)} \sum_{i=1}^n \xi_i^T \xi_i \right] \\
& = \frac{1}{bn} \left(\frac{n-b}{n-1} \right) \left[-\frac{1}{n} \sum_{i=1}^n \sum_{k=1}^n \xi_i^T \xi_k + \sum_{i=1}^n \xi_i^T \xi_i \right] \\
& = \frac{1}{bn} \left(\frac{n-b}{n-1} \right) \left[-n \left\| \frac{1}{n} \sum_{i=1}^n \xi_i \right\|^2 + \sum_{i=1}^n \|\xi_i\|^2 \right] \\
& \leq \frac{1}{b} \left(\frac{n-b}{n-1} \right) \frac{1}{n} \sum_{i=1}^n \|\xi_i\|^2 \\
& \stackrel{(22)}{=} \frac{1}{b} \left(\frac{n-b}{n-1} \right) \frac{1}{n} \sum_{i=1}^n \|\nabla f_i(w_j^{(s)}) - \nabla f_i(w_{j-1}^{(s)})\|^2 \\
& \stackrel{(9)}{\leq} \frac{1}{b} \left(\frac{n-b}{n-1} \right) L^2 \eta^2 \|v_{j-1}^{(s)}\|^2
\end{aligned}$$

Hence, by taking expectation, we have

$$\mathbb{E}[\|v_j^{(s)} - v_{j-1}^{(s)}\|^2] - \mathbb{E}[\|\nabla F(w_j^{(s)}) - \nabla F(w_{j-1}^{(s)})\|^2] \leq \frac{1}{b} \left(\frac{n-b}{n-1} \right) L^2 \eta^2 \mathbb{E}[\|v_{j-1}^{(s)}\|^2].$$

By Lemma 3, for $t \geq 1$,

$$\mathbb{E}[\|\nabla F(w_t^{(s)}) - v_t^{(s)}\|^2] = \sum_{j=1}^t \mathbb{E}[\|v_j^{(s)} - v_{j-1}^{(s)}\|^2] - \sum_{j=1}^t \mathbb{E}[\|\nabla F(w_j^{(s)}) - \nabla F(w_{j-1}^{(s)})\|^2]$$

$$\leq \frac{1}{b} \left(\frac{n-b}{n-1} \right) L^2 \eta^2 \sum_{j=1}^t \mathbb{E}[\|v_{j-1}^{(s)}\|^2].$$

Note that $\|\nabla F(w_0^{(s)}) - v_0^{(s)}\|^2 = 0$. Hence, by summing over $t = 0, \dots, m$ ($m \geq 1$), we have

$$\sum_{t=0}^m \mathbb{E}[\|\nabla F(w_t^{(s)}) - v_t^{(s)}\|^2] \leq \frac{1}{b} \left(\frac{n-b}{n-1} \right) L^2 \eta^2 \left[m \mathbb{E}\|v_0^{(s)}\|^2 + (m-1) \mathbb{E}\|v_1^{(s)}\|^2 + \dots + \mathbb{E}\|v_{m-1}^{(s)}\|^2 \right].$$

By choosing $\eta \leq \frac{2}{L(\sqrt{1+\frac{4m}{b}(\frac{n-b}{n-1})}+1)}$, we have

$$\begin{aligned} & \sum_{t=0}^m \mathbb{E}[\|\nabla F(w_t^{(s)}) - v_t^{(s)}\|^2] - (1-L\eta) \sum_{t=0}^m \mathbb{E}[\|v_t^{(s)}\|^2] \\ & \leq \frac{1}{b} \left(\frac{n-b}{n-1} \right) L^2 \eta^2 \left[m \mathbb{E}\|v_0^{(s)}\|^2 + (m-1) \mathbb{E}\|v_1^{(s)}\|^2 + \dots + \mathbb{E}\|v_{m-1}^{(s)}\|^2 \right] \\ & \quad - (1-L\eta) \left[\mathbb{E}\|v_0^{(s)}\|^2 + \mathbb{E}\|v_1^{(s)}\|^2 + \dots + \mathbb{E}\|v_m^{(s)}\|^2 \right] \\ & \leq \left[\frac{1}{b} \left(\frac{n-b}{n-1} \right) L^2 \eta^2 m - (1-L\eta) \right] \sum_{t=1}^m \mathbb{E}[\|v_{t-1}^{(s)}\|^2] \leq 0, \end{aligned} \tag{23}$$

since $\eta = \frac{2}{L(\sqrt{1+\frac{4m}{b}(\frac{n-b}{n-1})}+1)}$ is a root of equation $\frac{1}{b} \left(\frac{n-b}{n-1} \right) L^2 \eta^2 m - (1-L\eta) = 0$.

Therefore, with $\eta \leq \frac{2}{L(\sqrt{1+\frac{4m}{b}(\frac{n-b}{n-1})}+1)}$, we have

$$\mathbb{E}[F(w_{m+1}^{(s)})] \leq \mathbb{E}[F(w_0^{(s)})] - \frac{\eta}{2} \sum_{t=0}^m \mathbb{E}[\|\nabla F(w_t^{(s)})\|^2].$$

Following the same derivation of Theorem 1, we could achieve the desired result as follows for any given \tilde{w}_0 .

$$\frac{1}{(m+1)S} \sum_{s=1}^S \sum_{t=0}^m \mathbb{E}[\|\nabla F(w_t^{(s)})\|^2] \leq \frac{2}{\eta[(m+1)S]} [F(\tilde{w}_0) - F^*],$$

where F^* is any lower bound of F , and $w_t^{(s)}$ is the t -th iteration in the s -th outer loop. \square

Proof of Corollary 3

Corollary 3. *For the conditions in Theorem 2, in order to achieve an ϵ -accurate solution, the total complexity is*

$$\mathcal{O} \left(\left[\left(\frac{n+2bm}{m+1} \right) \left(\sqrt{1 + \frac{4m}{b} \left(\frac{n-b}{n-1} \right)} \right) \frac{1}{\epsilon} \right] \vee [n+2bm] \right).$$

Proof. By Theorem 2, let

$$\eta = \frac{2}{L \left(\sqrt{1 + \frac{4m}{b} \left(\frac{n-b}{n-1} \right)} + 1 \right)}.$$

Hence, we have

$$\begin{aligned} \frac{1}{(m+1)S} \sum_{s=1}^S \sum_{t=0}^m \mathbb{E}[\|\nabla F(w_t^{(s)})\|^2] &\leq \frac{2}{\eta[(m+1)S]} [F(\tilde{w}_0) - F^*] \\ &= \frac{\left(\sqrt{1 + \frac{4m}{b} \left(\frac{n-b}{n-1} \right)} + 1 \right)}{(m+1)S} L [F(\tilde{w}_0) - F^*] \\ &\leq \frac{\left(\sqrt{1 + \frac{4m}{b} \left(\frac{n-b}{n-1} \right)} \right)}{(m+1)S} 2L [F(\tilde{w}_0) - F^*] = \epsilon. \end{aligned}$$

In order to achieve the ϵ -accurate solution, we need

$$S = \frac{\left(\sqrt{1 + \frac{4m}{b} \left(\frac{n-b}{n-1} \right)} \right)}{(m+1)\epsilon} 2L [F(\tilde{w}_0) - F^*] = \mathcal{O} \left(\frac{\left(\sqrt{1 + \frac{4m}{b} \left(\frac{n-b}{n-1} \right)} \right)}{(m+1)\epsilon} \vee 1 \right),$$

since $S \geq 1$. Therefore, the total complexity is

$$(n + b \cdot 2m)S = \mathcal{O} \left(\left[\left(\frac{n+2bm}{m+1} \right) \left(\sqrt{1 + \frac{4m}{b} \left(\frac{n-b}{n-1} \right)} \right) \frac{1}{\epsilon} \right] \vee [n+2bm] \right).$$

□

Proof of Corollary 4

Corollary 4. *For the conditions in Theorem 2 and Corollary 3 with $b = n^\alpha$ and $m = n^\beta$ where $\alpha + \beta = 1$ with $\beta \geq 1/2$ and $0 \leq \alpha \leq 1/2$, in order to achieve an ϵ -accurate solution, the total complexity is*

$$\mathcal{O} \left(\frac{\sqrt{n}}{\epsilon} \vee n \right).$$

Proof. Let $b = n^\alpha$, $\alpha < 1$, and $m = n^\beta$, we have

$$\begin{aligned} \left(\frac{n+2bm}{m+1} \right) \sqrt{1 + \frac{4m}{b} \left(\frac{n-b}{n-1} \right)} &= \left(\frac{n+2n^{\alpha+\beta}}{n^\beta+1} \right) \sqrt{1 + 4n^{\beta-\alpha} \left(\frac{n-n^\alpha}{n-1} \right)} \\ &\leq \frac{n+2n^{\alpha+\beta}}{n^\beta} 2\sqrt{1 + n^{\beta-\alpha}}. \end{aligned}$$

If $\beta \geq \alpha$, we have

$$\frac{n + 2n^{\alpha+\beta}}{n^\beta} 2\sqrt{1+n^{\beta-\alpha}} \leq 2\sqrt{2} \left(\frac{n + 2n^{\alpha+\beta}}{n^\beta} \right) n^{(\beta-\alpha)/2} = 2\sqrt{2}(n^{1-\alpha/2-\beta/2} + 2n^{\alpha/2+\beta/2}).$$

In order to minimize the order of n , we need to choose $1 - \alpha/2 - \beta/2 = \alpha/2 + \beta/2$, which is equivalent to $\alpha + \beta = 1$ with $\beta \geq \alpha$. The best option is to choose $\alpha + \beta = 1$ with $\beta \geq 1/2$ and $0 \leq \alpha \leq 1/2$ in order to achieve $\mathcal{O}(n^{1/2})$.

If $\beta \leq \alpha$, we have

$$\frac{n + 2n^{\alpha+\beta}}{n^\beta} 2\sqrt{1+n^{\beta-\alpha}} \leq 2\sqrt{2}(n^{1-\beta} + 2n^\alpha).$$

In order to minimize the order of n , we need to choose $1 - \beta = \alpha$, which is equivalent to $\alpha + \beta = 1$ with $\beta \leq \alpha$. The best option is to choose $\beta = 1/2$ and $\alpha = 1/2$ in order to achieve $\mathcal{O}(n^{1/2})$.

Therefore, with $b = n^\alpha$ and $m = n^\beta$ where $\alpha + \beta = 1$ with $\beta \geq 1/2$ and $0 \leq \alpha \leq 1/2$, we have

$$\left(\frac{n + 2bm}{m + 1} \right) \sqrt{1 + \frac{4m}{b} \left(\frac{n - b}{n - 1} \right)} = \mathcal{O}(n^{1/2}).$$

By Corollary 3 with $bm = n^{\alpha+\beta} = n$, it implies the total complexity

$$(n + b \cdot 2m)S = \mathcal{O}\left(\frac{\sqrt{n}}{\epsilon} \vee n\right).$$

□

Convex SARAH++

Proof of Lemma 2

Lemma 2. Suppose that Assumptions 2 and 4 holds. Consider SARAH (Algorithm 1) within a single outer loop with $\eta \leq \frac{1}{L}$. Then, for $t \geq 0$ and any $s \geq 1$, we have

$$\mathbb{E}[F(w_{t+1}^{(s)}) - F(w_*)] \leq \mathbb{E}[F(w_t^{(s)}) - F(w_*)] - \frac{\eta}{2} \mathbb{E}[\|\nabla F(w_t^{(s)})\|^2] + \frac{\eta}{2} \left(L\eta \mathbb{E}[\|v_0^{(s)}\|^2] - \mathbb{E}[\|v_t^{(s)}\|^2] \right),$$

where w_* is any optimal solution of F .

Proof. By using (16) and adding $-F(w_*)$ for both sides, where $w_* = \arg \min_w F(w)$, we have

$$\begin{aligned} \mathbb{E}[F(w_{t+1}^{(s)}) - F(w_*)] &\leq \mathbb{E}[F(w_t^{(s)}) - F(w_*)] - \frac{\eta}{2} \mathbb{E}[\|\nabla F(w_t^{(s)})\|^2] + \frac{\eta}{2} \mathbb{E}[\|\nabla F(w_t^{(s)}) - v_t^{(s)}\|^2] \\ &\quad - \left(\frac{\eta}{2} - \frac{L\eta^2}{2} \right) \mathbb{E}[\|v_t^{(s)}\|^2] \\ &\stackrel{(15)}{\leq} \mathbb{E}[F(w_t^{(s)}) - F(w_*)] - \frac{\eta}{2} \mathbb{E}[\|\nabla F(w_t^{(s)})\|^2] \end{aligned}$$

$$\begin{aligned}
& + \frac{\eta}{2} \frac{\eta L}{(2 - \eta L)} \left(\mathbb{E}[\|v_0^{(s)}\|^2] - \mathbb{E}[\|v_t^{(s)}\|^2] \right) - \left(\frac{\eta}{2} - \frac{L\eta^2}{2} \right) \mathbb{E}[\|v_t^{(s)}\|^2] \\
& = \mathbb{E}[F(w_t^{(s)}) - F(w_*)] - \frac{\eta}{2} \mathbb{E}[\|\nabla F(w_t^{(s)})\|^2] \\
& \quad + \frac{\eta}{2} \left(\frac{\eta L}{(2 - \eta L)} \left(\mathbb{E}[\|v_0^{(s)}\|^2] - \mathbb{E}[\|v_t^{(s)}\|^2] \right) - (1 - L\eta) \mathbb{E}[\|v_t^{(s)}\|^2] \right) \\
& \stackrel{\eta \leq \frac{1}{L}}{\leq} \mathbb{E}[F(w_t^{(s)}) - F(w_*)] - \frac{\eta}{2} \mathbb{E}[\|\nabla F(w_t^{(s)})\|^2] \\
& \quad + \frac{\eta}{2} \left(\eta L \left(\mathbb{E}[\|v_0^{(s)}\|^2] - \mathbb{E}[\|v_t^{(s)}\|^2] \right) - (1 - L\eta) \mathbb{E}[\|v_t^{(s)}\|^2] \right) \\
& = \mathbb{E}[F(w_t^{(s)}) - F(w_*)] - \frac{\eta}{2} \mathbb{E}[\|\nabla F(w_t^{(s)})\|^2] + \frac{\eta}{2} \left(L\eta \mathbb{E}[\|v_0^{(s)}\|^2] - \mathbb{E}[\|v_t^{(s)}\|^2] \right).
\end{aligned}$$

□

Proof of Theorem 3

Theorem 3 (Smooth general convex). *Suppose that Assumptions 2 and 4 holds. Consider SARAH++ (Algorithm 3) with $\eta \leq \frac{\gamma}{L}$, $0 < \gamma \leq 1$. Then, the expectation of the average of squared norm of gradient of all iterations generated by SARAH++*

$$\mathbb{E} \left[\frac{1}{T_1 + \dots + T_S} \sum_{s=1}^S \sum_{t=0}^{T_s-1} \mathbb{E}[\|\nabla F(w_t^{(s)})\|^2 | T_1, \dots, T_S] \right] \leq \frac{2}{T\eta} [F(\tilde{w}_0) - F(w_*)].$$

Proof. We recall the following definitions. T_s is the stopping time (a random variable) of the s -th outer iteration such that

$$T_s = \min \left\{ \min_{t \geq 0} \left\{ t : \|v_t^{(s)}\|^2 < \gamma \|v_0^{(s)}\|^2 \right\}, m+1 \right\}, \quad s = 1, 2, \dots$$

and S is the stopping time of the outer iterations (a random variable) and such that for some $T > 0$

$$S = \min_{\hat{S}} \left\{ \hat{S} : \sum_{s=1}^{\hat{S}} T_s \geq T \right\}.$$

Note that $T_s \geq 1$ is the first time such that $\|v_{T_s}^{(s)}\|^2 < \gamma \|v_0^{(s)}\|^2$. Hence, for a given T_s , we have $\|v_t^{(s)}\|^2 \geq \gamma \|v_0^{(s)}\|^2$, for $0 \leq t \leq T_s - 1$, and

$$\begin{aligned}
\mathbb{E}[F(w_{T_s}^{(s)}) - F(w_*)] & \leq \mathbb{E}[F(w_{T_s-1}^{(s)}) - F(w_*)] - \frac{\eta}{2} \mathbb{E}[\|\nabla F(w_{T_s-1}^{(s)})\|^2] + \frac{\eta}{2} \left(L\eta \mathbb{E}[\|v_0^{(s)}\|^2] - \mathbb{E}[\|v_{T_s-1}^{(s)}\|^2] \right) \\
& \stackrel{\eta \leq \frac{\gamma}{L}}{\leq} \mathbb{E}[F(w_{T_s-1}^{(s)}) - F(w_*)] - \frac{\eta}{2} \mathbb{E}[\|\nabla F(w_{T_s-1}^{(s)})\|^2] + \frac{\eta}{2} \left(\gamma \mathbb{E}[\|v_0^{(s)}\|^2] - \mathbb{E}[\|v_{T_s-1}^{(s)}\|^2] \right) \\
& \leq \mathbb{E}[F(w_{T_s-1}^{(s)}) - F(w_*)] - \frac{\eta}{2} \mathbb{E}[\|\nabla F(w_{T_s-1}^{(s)})\|^2] \\
& \leq \mathbb{E}[F(w_0^{(s)}) - F(w_*)] - \frac{\eta}{2} \sum_{t=0}^{T_s-1} \mathbb{E}[\|\nabla F(w_t^{(s)})\|^2].
\end{aligned}$$

Since $\tilde{w}_s = w_{T_s}^{(s)}$ and $\tilde{w}_{s-1} = w_0^{(s)}$, for given T_1, \dots, T_S , we have

$$\begin{aligned}\mathbb{E}[F(\tilde{w}_S) - F(w_*)] &\leq \mathbb{E}[F(\tilde{w}_{S-1}) - F(w_*)] - \frac{\eta}{2} \sum_{t=0}^{T_S-1} \mathbb{E}[\|\nabla F(w_t^{(s)})\|^2] \\ &\leq \mathbb{E}[F(\tilde{w}_0) - F(w_*)] - \frac{\eta}{2} \sum_{s=1}^S \sum_{t=0}^{T_s-1} \mathbb{E}[\|\nabla F(w_t^{(s)})\|^2].\end{aligned}$$

Since $F(\tilde{w}_S) \geq F(w_*)$, bringing the second term of the RHS to the LHS. For any given \tilde{w}_0 , we have

$$\frac{\eta}{2} \sum_{s=1}^S \sum_{t=0}^{T_s-1} \mathbb{E}[\|\nabla F(w_t^{(s)})\|^2 | T_1, \dots, T_S] \leq [F(\tilde{w}_0) - F(w_*)],$$

which is equivalent to

$$\begin{aligned}\frac{1}{T_1 + \dots + T_S} \sum_{s=1}^S \sum_{t=0}^{T_s-1} \mathbb{E}[\|\nabla F(w_t^{(s)})\|^2 | T_1, \dots, T_S] &\leq \frac{1}{T_1 + \dots + T_S} \frac{2}{\eta} [F(\tilde{w}_0) - F(w_*)] \\ &\leq \frac{2}{\eta T} [F(\tilde{w}_0) - F(w_*)],\end{aligned}$$

where the last inequality follows since $\sum_{s=1}^S T_s \geq T$. Hence, by taking the expectation to both sides, we could have

$$\mathbb{E} \left[\frac{1}{T_1 + \dots + T_S} \sum_{s=1}^S \sum_{t=0}^{T_s-1} \mathbb{E}[\|\nabla F(w_t^{(s)})\|^2 | T_1, \dots, T_S] \right] \leq \frac{2}{\eta T} [F(\tilde{w}_0) - F(w_*)].$$

Therefore, we achieve the desired result since the LHS is the expectation of the average of squared norm of gradient of all iterations generated by SARAH++ (Algorithm 3). \square

Proof of Corollary 5

Corollary 5 (Smooth general convex). *Consider the conditions in Theorem 3 with $\eta = \mathcal{O}(\frac{1}{L})$. Then we could achieve the ϵ -accurate solution after $\mathcal{O}(\frac{1}{\epsilon})$ total iterations.*

Proof. The proof is trivial since we want

$$\frac{2}{\eta T} [F(\tilde{w}_0) - F(w_*)] = \epsilon,$$

which requires $T = \frac{2[F(\tilde{w}_0) - F(w_*)]}{\eta} \cdot \frac{1}{\epsilon} = \mathcal{O}(\frac{1}{\epsilon})$ iterations, where we could choose $\eta = \mathcal{O}(\frac{1}{L})$. \square

Proof of Theorem 4

Theorem 4 (Smooth strongly convex). *Suppose that Assumptions 2, 3 and 4 holds. Consider SARAH++ (Algorithm 3) with $\eta \leq \frac{\gamma}{L}$, $0 < \gamma \leq 1$. Then, for the final output \hat{w} of SARAH++, we have*

$$\mathbb{E}[F(\hat{w}) - F(w_*)] \leq (1 - \mu\eta)^T [F(\tilde{w}_0) - F(w_*)].$$

Proof. Following the beginning part of the proof of Theorem 3, we have, for a given T_s ,

$$\begin{aligned}\mathbb{E}[F(w_{T_s}^{(s)}) - F(w_*)] &\leq \mathbb{E}[F(w_{T_s-1}^{(s)}) - F(w_*)] - \frac{\eta}{2} \mathbb{E}[\|\nabla F(w_{T_s-1}^{(s)})\|^2] \\ &\stackrel{(10)}{\leq} (1 - \mu\eta) \mathbb{E}[F(w_{T_s-1}^{(s)}) - F(w_*)] \\ &\leq (1 - \mu\eta)^{T_s} \mathbb{E}[F(w_0^{(s)}) - F(w_*)]\end{aligned}$$

Since $\tilde{w}_s = w_{T_s}^{(s)}$ and $\tilde{w}_{s-1} = w_0^{(s)}$, for given T_1, \dots, T_S , we have

$$\begin{aligned}\mathbb{E}[F(\hat{w}) - F(w_*)|T_1, \dots, T_S] &= \mathbb{E}[F(\tilde{w}_S) - F(w_*)|T_1, \dots, T_S] \\ &\leq (1 - \mu\eta)^{T_1 + \dots + T_S} [F(\tilde{w}_0) - F(w_*)] \\ &\leq (1 - \mu\eta)^T [F(\tilde{w}_0) - F(w_*)],\end{aligned}$$

where the last inequality follows since $\sum_{s=1}^S T_s \geq T$. Hence, by taking the expectation to both sides, we could have

$$\mathbb{E}[F(\hat{w}) - F(w_*)] \leq (1 - \mu\eta)^T [F(\tilde{w}_0) - F(w_*)].$$

□

Proof of Corollary 6

Corollary 6 (Smooth strongly convex). *Consider the conditions in Theorem 4 with $\eta = \mathcal{O}(\frac{1}{L})$. Then we could achieve $\mathbb{E}[F(\hat{w}) - F(w_*)] \leq \epsilon$ after $\mathcal{O}(\kappa \log(\frac{1}{\epsilon}))$ total iterations, where $\kappa = L/\mu$ is the condition number.*

Proof. We want

$$(1 - \mu\eta)^T [F(\tilde{w}_0) - F(w_*)] = \epsilon.$$

Hence,

$$T = -\frac{1}{\log(1 - \mu\eta)} \log\left(\frac{[F(\tilde{w}_0) - F(w_*)]}{\epsilon}\right).$$

Note that: $-\frac{1}{x} - 1 \leq -\frac{1}{\log(1+x)} \leq -\frac{1}{x}$, $-1 < x < 0$. We can have

$$\left(\frac{1}{\mu\eta} - 1\right) \log\left(\frac{[F(\tilde{w}_0) - F(w_*)]}{\epsilon}\right) \leq T \leq \frac{1}{\mu\eta} \log\left(\frac{[F(\tilde{w}_0) - F(w_*)]}{\epsilon}\right).$$

By choosing $\eta = \mathcal{O}(\frac{1}{L})$, we have $T = \mathcal{O}(\kappa \log(\frac{1}{\epsilon}))$.

□

Additional Experiments

We provide more experiments in this section on popular data sets with diverse size n including *covtype* ($n = 406, 708$ training data; estimated $L \simeq 1.90$), *ijcnn1* ($n = 91, 701$ training data; estimated $L \simeq 1.77$), *w8a* ($n = 49, 749$ training data, estimated $L \simeq 7.05$) and *phishing* ($n = 7, 738$ training data, estimated $L \simeq 7.49$) from LIBSVM.

Additional experiments in Section 3.2

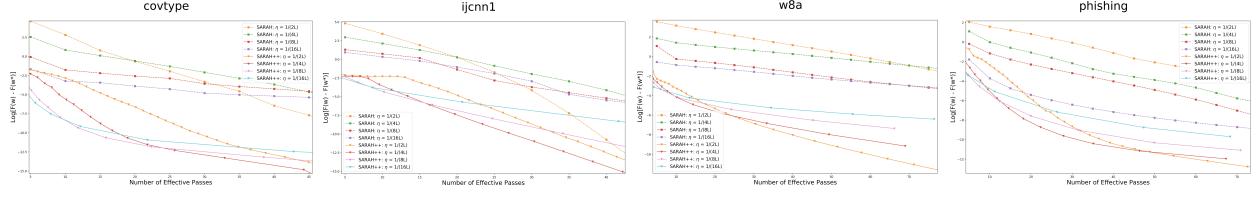


Figure 4: Comparisons of $\log[F(w) - F(w_*)]$ between SARAH++ and SARAH with different learning rates on *covtype*, *ijcnn1*, *w8a*, and *phishing* datasets

Additional experiments in Section 3.3

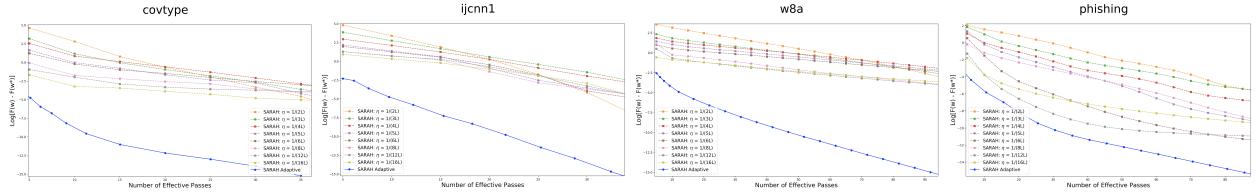


Figure 5: Comparisons of $\log[F(w) - F(w_*)]$ between SARAH Adaptive and SARAH with different learning rates on *covtype*, *ijcnn1*, *w8a*, and *phishing* datasets

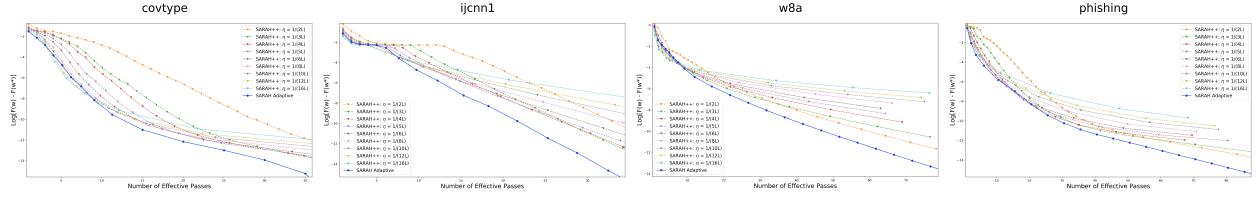


Figure 6: Comparisons of $\log[F(w) - F(w_*)]$ between SARAH Adaptive and SARAH++ with different learning rates on *covtype*, *ijcnn1*, *w8a*, and *phishing* data sets

Sensitivity of γ for SARAH Adaptive

In Figure 7 we present the numerical performance of SARAH Adaptive for different values of $\gamma = \{\frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{6}, \frac{1}{8}, \frac{1}{10}, \frac{1}{12}, \frac{1}{16}\}$ on *covtype*, *ijcnn1*, *w8a*, and *phishing* data sets.

