arXiv:1901.07648v1 [math.OC] 22 Jan 2019

Optimal Finite-Sum Smooth Non-Convex Optimization with SARAH

Lam M. Nguyen [1] **Marten van Dijk** [2] **Dzung T. Phan** [1] **Phuong Ha Nguyen** [2] **Tsui-Wei Weng** [3]
Jayant R. Kalagnanam [1]

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 non-convex objective function $F(w) = \frac{1}{n}\sum_{i=1}^{n} f_i(w)$ has been proven to be at least $\Omega(\sqrt{n}/\epsilon)$ where ϵ denotes the attained accuracy $\mathbb{E}[\|\nabla F(\tilde{w})\|^2] \leq \epsilon$ for the outputted approximation $\tilde{w}$ (Fang et al., 2018). This paper is the first to show that this lower bound is tight for the class of variance reduction methods which only assume the Lipschitz continuous gradient assumption. We prove this complexity result for a slightly modified version of the SARAH algorithm in (Nguyen et al., 2017a;b) – showing that SARAH is optimal and dominates all existing results. 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

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

where each f_i, $i \in [n] \overset{\text{def}}{=} \{1,\ldots,n\}$, has a Lipschitz continuous gradient with constant $L > 0$. Throughout the paper, we consider the case where F has a finite lower bound F^*.

[1]IBM Research, Thomas J. Watson Research Center, Yorktown Heights, USA. [2]Department of Electrical and Computer Engineering, University of Connecticut, USA. [3]Massachusetts Institute of Technology, Cambridge, USA. Correspondence to: Lam M. Nguyen <LamNguyen.MLTD@ibm.com>.

Problems of form (1) cover a wide range of convex and non-convex 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) (Robbins & Monro, 1951; Shalev-Shwartz et al., 2011; Bottou et al., 2016). In recent years, a large number of improved variants of stochastic gradient algorithms called variance reduction methods have emerged, in particular, SAG/SAGA (Schmidt et al., 2016; Defazio et al., 2014), SDCA (Shalev-Shwartz & Zhang, 2013), MISO (Mairal, 2013), SVRG/S2GD (Johnson & Zhang, 2013; Konečný & Richtárik, 2013), SARAH (Nguyen et al., 2017a), etc. These 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. (Reddi et al., 2016; Lei et al., 2017; Allen-Zhu, 2017a;b; Fang et al., 2018), etc.)

We study the SARAH algorithm (Nguyen et al., 2017a;b) 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

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

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

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

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 (Nguyen et al., 2017a) uses $\tilde{w}_s = w_t^{(s)}$ with t chosen uniformly at random from $\{0,1,\ldots,m\}$. The authors of (Nguyen et al., 2017a) 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 (Nguyen et al., 2017a) and call this SARAH.

Algorithm 1 SARAH (modified of (Nguyen et al., 2017a))

Parameters: the learning rate $\eta > 0$, the inner loop size m, and the outer loop size S
Initialize: $\tilde{w}_0$
Iterate:
for $s = 1, 2, \ldots, S$ **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 $t = 1, \ldots, m$ **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) where we *only* assume component functions having a finite Lipschitz continuous gradient L and no other assumptions:

Assumption 1 (L-smooth)**.** *Each $f_i : \mathbb{R}^d \to \mathbb{R}$, $i \in [n]$, is L-smooth, i.e., there exists a constant $L > 0$ such that, $\forall w, w' \in \mathbb{R}^d$,*

$$\|\nabla f_i(w) - \nabla f_i(w')\| \leq L\|w - w'\|. \tag{4}$$

We stress that our convergence analysis only relies on the above smooth assumption without bounded variance assumption (as required in (Lei et al., 2017; Zhou et al., 2018)) or Hessian-Lipschitz assumption (as required in (Fang et al., 2018)).

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 (Fang et al., 2018), is a random algorithm $\mathcal{A}$ that maps functions f to a sequence of iterates

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

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 (Fang et al., 2018) states the following lower bound: 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

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

Our main contribution is to meet this lower bound and show that in SARAH we can choose parameters S and m such that the total complexity is

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

or, equivalently,

$$\mathbb{E}[\|\nabla F(\tilde{w}_S)\|^2] \leq \epsilon = \mathcal{O}(\frac{L\sqrt{n}}{S \cdot (n + 2m)}).$$

This significantly improves over prior work which only achieves $\mathcal{T} = \mathcal{O}(n + L\sqrt{n}/\epsilon)$:

Related Work: The paper that introduces SARAH (Nguyen et al., 2017b) 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, (Fang et al., 2018) introduces SPIDER, as a variant of SARAH, which achieves to date the best known convergence result in the nonconvex case. SPIDER uses the SARAH update rule (2) as was originally proposed in (Nguyen et al., 2017a) and the mini-batch version of SARAH in (Nguyen et al., 2017b). 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 (Wang et al., 2018) 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

$$\mathcal{O}\left(n + L\sqrt{n}/\epsilon\right),$$

which is called "near-optimal" in (Fang et al., 2018) since, except for the $\mathcal{O}(n)$ term, it almost matches the lower bound.

Table 1[1] 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

[1] $a \wedge b$ is defined as $\min\{a, b\}$

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

Method	Complexity	Additional assumption
GD (Nesterov, 2004)	$\mathcal{O}\left(\frac{n}{\epsilon}\right)$	None
SVRG (Reddi et al., 2016)	$\mathcal{O}\left(n+\frac{n^{2/3}}{\epsilon}\right)$	None
SCSG (Lei et al., 2017)	$\mathcal{O}\left(\left(\frac{\sigma}{\epsilon}\wedge n\right)+\frac{1}{\epsilon}\left(\frac{\sigma}{\epsilon}\wedge n\right)^{2/3}\right)$ $\mathcal{O}\left(n+\frac{n^{2/3}}{\epsilon}\right)$	Bounded variance None ($\sigma\to\infty$)
SNVRG (Zhou et al., 2018)	$\mathcal{O}\left(\log^3\left(\frac{\sigma}{\epsilon}\wedge n\right)\left[\left(\frac{\sigma}{\epsilon}\wedge n\right)+\frac{1}{\epsilon}\left(\frac{\sigma}{\epsilon}\wedge n\right)^{1/2}\right]\right)$ $\mathcal{O}\left(\log^3(n)\left(n+\frac{\sqrt{n}}{\epsilon}\right)\right)$	Bounded variance None ($\sigma\to\infty$)
SPIDER (Fang et al., 2018)	$\mathcal{O}\left(n+\frac{\sqrt{n}}{\epsilon}\right)$	None
SpiderBoost (Wang et al., 2018)	$\mathcal{O}\left(n+\frac{\sqrt{n}}{\epsilon}\right)$	None
R-SPIDER (Zhang et al., 2018)	$\mathcal{O}\left(n+\frac{\sqrt{n}}{\epsilon}\right)$	None
SARAH (this paper)	$\mathcal{O}\left(\frac{\sqrt{n}}{\epsilon}\right)$	**None**

many algorithms have appeared during the past few years, we only compare algorithms having a convergence result which only supposes the smooth assumption. For example, (Fang et al., 2018) can also prove a total complexity of $\mathcal{O}\left(\sqrt{n}/\epsilon\right)$ by requiring an additional Hessian-Lipschitz assumption and adding dependence on the Hessian-Lipschitz constant to their analysis. For this reason, this result is not part of the table as it is weaker in that the analysis supposes an additional property of the component functions. (c) Among algorithms with convergence results that only suppose the smooth assumption, Table 1 only mentions recent state-of-the-art results. For example, we do not provide comparisons with SGD (Robbins & Monro, 1951) and SGD-like (e.g. (Duchi et al., 2011; Kingma & Ba, 2014)) since they achieve a much worse complexity of $\mathcal{O}(\frac{1}{\epsilon^2})$. (d) Although the bounded variance assumption $\mathbb{E}[\|\nabla f_i(w)-\nabla F(w)\|^2]\leq\sigma$ 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 $\sigma\to\infty$. (e) For completeness, incompatibility with assuming a bounded gradient $\mathbb{E}[\|\nabla f_i(w)\|^2]\leq\sigma$ has been discussed in (Nguyen et al., 2018a) for strongly convex objective functions.

According to the results in Table 1, we can observe that SARAH-type algorithms dominate SVRG-type algorithms. In fact this paper proves that SARAH (slightly modified as given in Algorithm 1) achieves the minimal possible total complexity among variance reduction techniques in the non-convex case for finding a first-order stationary point based on **only the smooth assumption**. This closes the gap of searching for "better" algorithms since the total complexity meets the lower bound $\Omega(\sqrt{n}/\epsilon)$.

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 (Nguyen et al., 2017b) which only analyses a single outer iteration). The convergence analysis **only** supposes the smooth assumption (Lipschitz continuous on the gradient) and proves that SARAH with multiple outer loops (which has not been analyzed before) attains the asymptotic *minimum possible total complexity* in the non-convex case (Theorem 1). We extend these results to the *mini-batch* case (Theorem 2).

Smooth Convex. In order to complete the picture, we study SARAH+ (Nguyen et al., 2017a) 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*

$$\mathbb{E}[F(w_{m+1}^{(s)})] \leq \mathbb{E}[F(w_0^{(s)})] - \frac{\eta}{2}\sum_{t=0}^{m}\mathbb{E}[\|\nabla F(w_t^{(s)})\|^2]. \tag{5}$$

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

$$\mathbb{E}[F(\tilde{w}_S)] \leq \mathbb{E}[F(\tilde{w}_0)] - \frac{\eta}{2}\sum_{s=1}^{S}\sum_{t=0}^{m}\mathbb{E}[\|\nabla F(w_t^{(s)})\|^2]. \tag{6}$$

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 (6) 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*

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

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*

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

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

$$v_t^{(s)} = \frac{1}{b}\sum_{i\in I_t}[\nabla f_i(w_t^{(s)}) - \nabla f_i(w_{t-1}^{(s)})] + v_{t-1}^{(s)}, \tag{7}$$

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 (7) 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, but this does not lead to additional insight; it results in the same minimal total complexity for ϵ-accurate solutions.

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

In this section, we propose a new variant of SARAH+ (Algorithm 2) (Nguyen et al., 2017a), 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, $\mathbb{E}[\|v_t^{(s)}\|^2] \to 0$ as $t \to \infty$ in the strongly convex case (Theorems 1a and 1b in (Nguyen et al., 2017a)). Therefore, it does not make any sense to update with tiny steps when $\|v_t^{(s)}\|^2$ is small. (We note that SVRG (Johnson & Zhang, 2013) does not have this property.) SARAH+ suggests to empirically choose parameter $\gamma = 1/8$ (Nguyen et al., 2017a) 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.

3.1. Details SARAH++ and Convergence Analysis

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

Algorithm 2 SARAH+ (Nguyen et al., 2017a)

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 $s = 1, 2, \ldots, S$ **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

Assumption 2 (μ-strongly convex)**.** *The function $F : \mathbb{R}^d \to \mathbb{R}$, 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 2, let us define the (unique) optimal solution of (1) as w_*. Then strong convexity of F implies that

$$2\mu[F(w)-F(w_*)] \leq \|\nabla F(w)\|^2, \ \forall w \in \mathbb{R}^d. \quad (8)$$

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 2 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 3. *Each function $f_i : \mathbb{R}^d \to \mathbb{R}$, $i \in [n]$, is convex, i.e.,*

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

SARAH++ is motivated by the following lemma.

Lemma 2. *Suppose that Assumptions 1 and 3 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*

$$\mathbb{E}[F(w_{t+1}^{(s)})-F(w_*)] \leq \mathbb{E}[F(w_t^{(s)})-F(w_*)] - \frac{\eta}{2}\mathbb{E}[\|\nabla F(w_t^{(s)})\|^2] + \frac{\eta}{2}\left(L\eta\mathbb{E}[\|v_0^{(s)}\|^2] - \mathbb{E}[\|v_t^{(s)}\|^2]\right), \quad (9)$$

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 (9) 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 $L\eta$.

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

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:

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

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.

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 3 in addition to smoothness we have the next theorem.

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

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

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 3 (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 2 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 1, 2 and 3 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*

$$\mathbb{E}[F(\hat{w}) - F(w_*)] \leq (1 - \mu\eta)^T[F(\tilde{w}_0) - F(w_*)]. \quad (10)$$

This leads to the following iteration complexity.

Corollary 4 (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 3 and 4 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 $T_i = 1$, $i = 1, \ldots, S$, 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 (Nguyen et al., 2017a) provides experiments showing good overall performance of SARAH over other algorithms such as SGD (Robbins & Monro, 1951), SAG (Le Roux et al., 2012), SVRG (Johnson & Zhang, 2013), 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 (Nguyen et al., 2017a). 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

$$f_i(w) = \log(1 + \exp(-y_i\langle x_i, w\rangle)) + \frac{\lambda}{2}\|w\|^2, \qquad (11)$$

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 (Le Roux et al., 2012; Nguyen et al., 2017a). 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 (Chang & Lin, 2011).

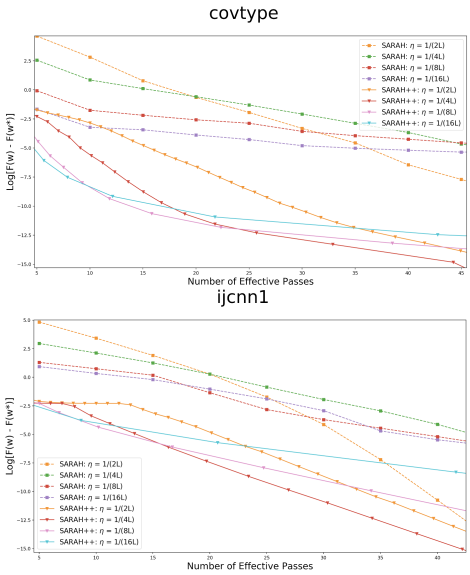

Figure 1: Comparisons of $\log[F(w) - F(w_*)]$ between SARAH++ and SARAH with different learning rates on *covtype* and *ijcnn1* datasets

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 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 (11) 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. We also provide experiments about the sensitivity of the maximum inner loop size in supplementary material.

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 $s = 1, 2, \ldots, S$ **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**
    $\eta_t^{(s)} = \frac{1}{L} \cdot \frac{\|v_t^{(s)}\|^2}{\|v_0^{(s)}\|^2}$ **(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
  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 $\eta \leq \frac{1}{L}$, (9) 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 (Nguyen et al., 2017a) 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.

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 the supplementary materials. We also present the numerical performance of SARAH Adaptive for different values of γ in the supplementary materials.

We note that additional experiments in this section on more data sets are performed in the supplementary material.

4. Conclusion and Future Research

Not known in prior literature, we have proven how to achieve optimal total complexity for smooth nonconvex problems in the finite-sum setting, which arises frequently in supervised learning applications. For convex problems, we proposed SARAH++ with theoretical convergence guarantee and showed improved performance over SARAH.

For future research, ideas in this paper may apply to general expectation minimization problems using an inexact version of the gradient (Nguyen et al., 2018b). It would also be noteworthy to investigate SARAH Adaptive in more detail since it has promising empirical results. Moreover, SARAH may open some new research directions because it could be reduced to Gradient Descent as shown in the paper.

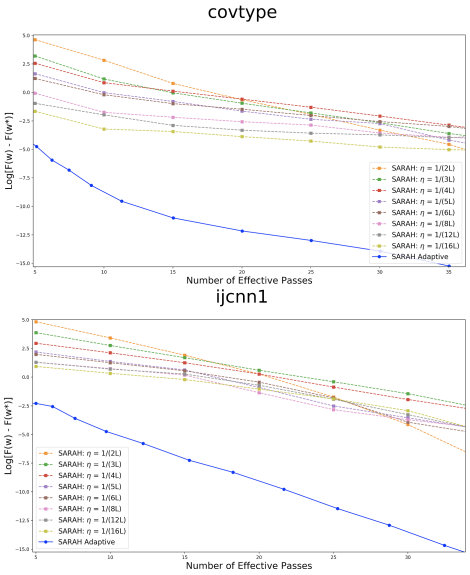

Figure 2: Comparisons of $\log[F(w) - F(w_*)]$ between SARAH Adaptive and SARAH with different learning rates on *covtype* and *ijcnn1* datasets

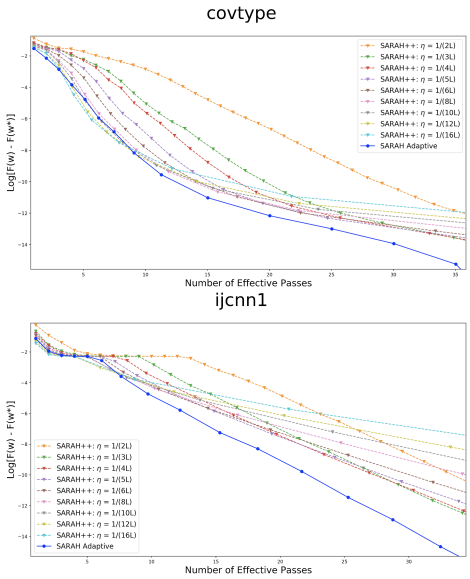

Figure 3: Comparisons of $\log[F(w) - F(w_*)]$ between SARAH Adaptive and SARAH++ with different learning rates on *covtype* and *ijcnn1* datasets

References

Allen-Zhu, Z. Natasha: Faster non-convex stochastic optimization via strongly non-convex parameter. *arXiv preprint arXiv:1702.00763*, 2017a.

Allen-Zhu, Z. Natasha 2: Faster non-convex optimization than sgd. *arXiv preprint arXiv:1708.08694*, 2017b.

Bottou, L., Curtis, F. E., and Nocedal, J. Optimization methods for large-scale machine learning. *arXiv:1606.04838*, 2016.

Chang, C.-C. and Lin, C.-J. 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.

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

Duchi, J., Hazan, E., and Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2011.

Fang, C., Li, C. J., Lin, Z., and Zhang, T. Spider: Near-optimal non-convex optimization via stochastic path integrated differential estimator. *arXiv preprint arXiv:1807.01695*, 2018.

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

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

Konečný, J. and Richtárik, P. Semi-stochastic gradient descent methods. *arXiv:1312.1666*, 2013.

Le Roux, N., Schmidt, M., and Bach, F. A stochastic gradient method with an exponential convergence rate for finite training sets. In *NIPS*, pp. 2663–2671, 2012.

Lei, L., Ju, C., Chen, J., and Jordan, M. I. Non-convex finite-sum optimization via SCSG methods. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 2348–2358. Curran Associates, Inc., 2017.

Mairal, J. Optimization with first-order surrogate functions. In *ICML*, pp. 783–791, 2013.

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

Nguyen, L., Nguyen, P. H., van Dijk, M., Richtarik, P., Scheinberg, K., and Takac, M. SGD and Hogwild! convergence without the bounded gradients assumption. In *ICML*, 2018a.

Nguyen, L. M., Liu, J., Scheinberg, K., and Takáč, M. SARAH: A novel method for machine learning problems using stochastic recursive gradient. In *ICML*, 2017a.

Nguyen, L. M., Liu, J., Scheinberg, K., and Takác, M. Stochastic recursive gradient algorithm for nonconvex optimization. *CoRR*, abs/1705.07261, 2017b.

Nguyen, L. M., Scheinberg, K., and Takac, M. Inexact SARAH algorithm for stochastic optimization. *arXiv preprint arXiv:1811.10105*, 2018b.

Reddi, S. J., Hefny, A., Sra, S., Póczos, B., and Smola, A. J. Stochastic variance reduction for nonconvex optimization. In *ICML*, pp. 314–323, 2016.

Robbins, H. and Monro, S. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3): 400–407, 1951.

Schmidt, M., Le Roux, N., and Bach, F. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, pp. 1–30, 2016.

Shalev-Shwartz, S. and Zhang, T. Stochastic dual coordinate ascent methods for regularized loss. *Journal of Machine Learning Research*, 14(1):567–599, 2013.

Shalev-Shwartz, S., Singer, Y., Srebro, N., and Cotter, A. Pegasos: Primal estimated sub-gradient solver for SVM. *Mathematical Programming*, 127(1):3–30, 2011.

Wang, Z., Ji, K., Zhou, Y., Liang, Y., and Tarokh, V. Spiderboost: A class of faster variance-reduced algorithms for nonconvex optimization. *arXiv preprint arXiv:1810.10690*, 2018.

Zhang, J., Zhang, H., and Sra, S. R-spider: A fast riemannian stochastic optimization algorithm with curvature independent rate. *arXiv preprint arXiv:1811.04194*, 2018.

Zhou, D., Xu, P., and Gu, Q. Stochastic nested variance reduction for nonconvex optimization. *arXiv preprint arXiv:1806.07811*, 2018.

Optimal Finite-Sum Smooth Non-Convex Optimization with SARAH Supplementary Material

Appendix

Useful Existing Results

Lemma 3 (Theorem 2.1.5 in (Nesterov, 2004))**.** *Suppose that f is L-smooth. Then, for any w, $w' \in \mathbb{R}^d$,*

$$f(w) \leq f(w') + \nabla f(w')^T(w - w') + \frac{L}{2}\|w - w'\|^2. \tag{12}$$

Lemma 4 (Lemma 2 in (Nguyen et al., 2017a) (or in (Nguyen et al., 2017b)))**.** *Suppose that Assumption 1 holds. Consider $v_t^{(s)}$ defined by (2) (or (7)) in SARAH (Algorithm 1) for any $s \geq 1$. Then for any $t \geq 1$,*

$$\mathbb{E}[\|\nabla F(w_t^{(s)}) - v_t^{(s)}\|^2] = \sum_{j=1}^{t} \mathbb{E}[\|v_j^{(s)} - v_{j-1}^{(s)}\|^2] - \sum_{j=1}^{t} \mathbb{E}[\|\nabla F(w_j^{(s)}) - \nabla F(w_{j-1}^{(s)})\|^2]. \tag{13}$$

Lemma 5 (Lemma 3 in (Nguyen et al., 2017a))**.** *Suppose that Assumptions 1 and 3 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 $t \geq 0$,*

$$\mathbb{E}[\|\nabla F(w_t^{(s)}) - v_t^{(s)}\|^2] \leq \frac{\eta L}{2 - \eta L}\left[\mathbb{E}[\|v_0^{(s)}\|^2] - \mathbb{E}[\|v_t^{(s)}\|^2]\right]. \tag{14}$$

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 (Nguyen et al., 2017b). By Assumption 1 and $w_{t+1}^{(s)} = w_t^{(s)} - \eta v_t^{(s)}$, for any $s \geq 1$, we have

$$\begin{aligned}\mathbb{E}[F(w_{t+1}^{(s)})] &\overset{(12)}{\leq} \mathbb{E}[F(w_t^{(s)})] - \eta\mathbb{E}[\nabla F(w_t^{(s)})^T v_t^{(s)}] + \frac{L\eta^2}{2}\mathbb{E}[\|v_t^{(s)}\|^2] \\ &= \mathbb{E}[F(w_t^{(s)})] - \frac{\eta}{2}\mathbb{E}[\|\nabla F(w_t^{(s)})\|^2] + \frac{\eta}{2}\mathbb{E}[\|\nabla F(w_t^{(s)}) - v_t^{(s)}\|^2] - \left(\frac{\eta}{2} - \frac{L\eta^2}{2}\right)\mathbb{E}[\|v_t^{(s)}\|^2],\end{aligned} \tag{15}$$

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

$$\mathbb{E}[F(w_{m+1}^{(s)})] \leq \mathbb{E}[F(w_0^{(s)})] - \frac{\eta}{2}\sum_{t=0}^{m} \mathbb{E}[\|\nabla F(w_t^{(s)})\|^2] + \frac{\eta}{2}\left(\sum_{t=0}^{m} \mathbb{E}[\|\nabla F(w_t^{(s)}) - v_t^{(s)}\|^2] - (1 - L\eta)\sum_{t=0}^{m} \mathbb{E}[\|v_t^{(s)}\|^2]\right). \tag{16}$$

Now, we would like to determine η such that the expression in (16)

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

We have

$$\|v_j^{(s)}-v_{j-1}^{(s)}\|^2\overset{(2)}{=}\|\nabla f_{i_j}(w_j^{(s)})-\nabla f_{i_j}(w_{j-1}^{(s)})\|^2\overset{(4)}{\leq}L^2\|w_j^{(s)}-w_{j-1}^{(s)}\|^2=L^2\eta^2\|v_{j-1}^{(s)}\|^2,\ j\geq 1. \tag{17}$$

Hence, by Lemma 4, we have

$$\mathbb{E}[\|\nabla F(w_t^{(s)})-v_t^{(s)}\|^2]\leq\sum_{j=1}^{t}\mathbb{E}[\|v_j^{(s)}-v_{j-1}^{(s)}\|^2]\overset{(17)}{\leq}L^2\eta^2\sum_{j=1}^{t}\mathbb{E}[\|v_{j-1}^{(s)}\|^2].$$

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

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

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

$$\begin{aligned}
&\sum_{t=0}^{m}\mathbb{E}[\|\nabla F(w_t^{(s)})-v_t^{(s)}\|^2]-(1-L\eta)\sum_{t=0}^{m}\mathbb{E}[\|v_t^{(s)}\|^2]\\
&\leq L^2\eta^2\Big[m\mathbb{E}\|v_0^{(s)}\|^2+(m-1)\mathbb{E}\|v_1^{(s)}\|^2+\cdots+\mathbb{E}\|v_{m-1}^{(s)}\|^2\Big]-(1-L\eta)\Big[\mathbb{E}\|v_0^{(s)}\|^2+\mathbb{E}\|v_1^{(s)}\|^2+\cdots+\mathbb{E}\|v_m^{(s)}\|^2\Big]\\
&\leq\Big[L^2\eta^2m-(1-L\eta)\Big]\sum_{t=1}^{m}\mathbb{E}[\|v_{t-1}^{(s)}\|^2]\leq 0,
\end{aligned}\tag{18}$$

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

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

$$\frac{2}{\eta[(m+1)S]}[F(\tilde{w}_0)-F^*]=\epsilon. \tag{19}$$

Let us choose η such that

$$\eta=\frac{2}{L(3\sqrt{m+1})}\overset{m\geq 0}{\leq}\frac{2}{L(2\sqrt{m+1}+1)}=\frac{2}{L(\sqrt{4m+4}+1)}\leq\frac{2}{L(\sqrt{1+4m}+1)}. \tag{20}$$

Hence, in order to achieve (19), we need

$$S = \frac{2}{\eta[(m+1)\epsilon]}[F(\tilde{w}_0) - F^*] \overset{(20)}{=} \frac{3L[F(\tilde{w}_0) - F^*]}{(\sqrt{m+1})}\frac{1}{\epsilon} = \mathcal{O}\left(\frac{1}{\sqrt{m+1}} \cdot \frac{1}{\epsilon}\right).$$

Therefore, the total complexity to achieve ϵ-accurate solution is $(n+2m)S = \mathcal{O}\left(\left(\frac{n+2m}{\sqrt{m+1}}\right)\frac{1}{\epsilon}\right)$. □

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 (7) 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 (16)

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

Let

$$\xi_t = \nabla f_t(w_j^{(s)}) - \nabla f_t(w_{j-1}^{(s)}). \tag{21}$$

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

$$\begin{aligned}
&\mathbb{E}[\|v_j^{(s)} - v_{j-1}^{(s)}\|^2|\mathcal{F}_j] - \|\nabla F(w_j^{(s)}) - \nabla F(w_{j-1}^{(s)})\|^2 \\
&\quad\overset{(7)}{=} \mathbb{E}\Big[\Big\|\frac{1}{b}\sum_{i\in I_j}[\nabla f_i(w_j^{(s)}) - \nabla f_i(w_{j-1}^{(s)})]\Big\|^2\Big|\mathcal{F}_j\Big] - \Big\|\frac{1}{n}\sum_{i=1}^{n}[\nabla f_i(w_j^{(s)}) - \nabla f_i(w_{j-1}^{(s)})]\Big\|^2 \\
&\quad\overset{(21)}{=} \mathbb{E}\Big[\Big\|\frac{1}{b}\sum_{i\in I_j}\xi_i\Big\|^2\Big|\mathcal{F}_j\Big] - \Big\|\frac{1}{n}\sum_{i=1}^{n}\xi_i\Big\|^2 \\
&\quad= \frac{1}{b^2}\mathbb{E}\Big[\sum_{i\in I_j}\sum_{k\in I_j}\xi_i^T\xi_k\Big|\mathcal{F}_j\Big] - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{k=1}^{n}\xi_i^T\xi_k \\
&\quad= \frac{1}{b^2}\mathbb{E}\Big[\sum_{i\neq k\in I_j}\xi_i^T\xi_k + \sum_{i\in I_j}\xi_i^T\xi_i\Big|\mathcal{F}_j\Big] - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{k=1}^{n}\xi_i^T\xi_k \\
&\quad= \frac{1}{b^2}\Big[\frac{b}{n}\frac{(b-1)}{(n-1)}\sum_{i\neq k}\xi_i^T\xi_k + \frac{b}{n}\sum_{i=1}^{n}\xi_i^T\xi_i\Big] - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{k=1}^{n}\xi_i^T\xi_k \\
&\quad= \frac{1}{b^2}\Big[\frac{b}{n}\frac{(b-1)}{(n-1)}\sum_{i=1}^{n}\sum_{k=1}^{n}\xi_i^T\xi_k + \Big(\frac{b}{n} - \frac{b}{n}\frac{(b-1)}{(n-1)}\Big)\sum_{i=1}^{n}\xi_i^T\xi_i\Big] - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{k=1}^{n}\xi_i^T\xi_k
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{bn}\Big[\Big(\frac{(b-1)}{(n-1)} - \frac{b}{n}\Big)\sum_{i=1}^{n}\sum_{k=1}^{n}\xi_i^T\xi_k + \frac{(n-b)}{(n-1)}\sum_{i=1}^{n}\xi_i^T\xi_i\Big] \\
&= \frac{1}{bn}\left(\frac{n-b}{n-1}\right)\Big[-\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{n}\xi_i^T\xi_k + \sum_{i=1}^{n}\xi_i^T\xi_i\Big] \\
&= \frac{1}{bn}\left(\frac{n-b}{n-1}\right)\Big[-n\Big\|\frac{1}{n}\sum_{i=1}^{n}\xi_i\Big\|^2 + \sum_{i=1}^{n}\|\xi_i\|^2\Big] \\
&\leq \frac{1}{b}\left(\frac{n-b}{n-1}\right)\frac{1}{n}\sum_{i=1}^{n}\|\xi_i\|^2 \\
&\overset{(21)}{=} \frac{1}{b}\left(\frac{n-b}{n-1}\right)\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(w_j^{(s)}) - \nabla f_i(w_{j-1}^{(s)})\|^2 \\
&\overset{(4)}{\leq} \frac{1}{b}\left(\frac{n-b}{n-1}\right)L^2\eta^2\frac{1}{n}\sum_{i=1}^{n}\|v_{j-1}^{(s)}\|^2 \\
&= \frac{1}{b}\left(\frac{n-b}{n-1}\right)L^2\eta^2\|v_{j-1}^{(s)}\|^2.
\end{aligned}
$$

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 4, for $t \geq 1$,

$$
\begin{aligned}
\mathbb{E}[\|\nabla F(w_t^{(s)}) - v_t^{(s)}\|^2] &= \sum_{j=1}^{t}\mathbb{E}[\|v_j^{(s)} - v_{j-1}^{(s)}\|^2] - \sum_{j=1}^{t}\mathbb{E}[\|\nabla F(w_j^{(s)}) - \nabla F(w_{j-1}^{(s)})\|^2] \\
&\leq \frac{1}{b}\left(\frac{n-b}{n-1}\right)L^2\eta^2\sum_{j=1}^{t}\mathbb{E}[\|v_{j-1}^{(s)}\|^2].
\end{aligned}
$$

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

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

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

$$
\begin{aligned}
&\sum_{t=0}^{m}\mathbb{E}[\|\nabla F(w_t^{(s)}) - v_t^{(s)}\|^2] - (1-L\eta)\sum_{t=0}^{m}\mathbb{E}[\|v_t^{(s)}\|^2] \\
&\leq \frac{1}{b}\left(\frac{n-b}{n-1}\right)L^2\eta^2\Big[m\mathbb{E}\|v_0^{(s)}\|^2 + (m-1)\mathbb{E}\|v_1^{(s)}\|^2 + \cdots + \mathbb{E}\|v_{m-1}^{(s)}\|^2\Big] \\
&- (1-L\eta)\Big[\mathbb{E}\|v_0^{(s)}\|^2 + \mathbb{E}\|v_1^{(s)}\|^2 + \cdots + \mathbb{E}\|v_m^{(s)}\|^2\Big] \\
&\leq \Big[\frac{1}{b}\left(\frac{n-b}{n-1}\right)L^2\eta^2 m - (1-L\eta)\Big]\sum_{t=1}^{m}\mathbb{E}[\|v_{t-1}^{(s)}\|^2] \leq 0,
\end{aligned} \tag{22}
$$

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

□

Convex SARAH++

Proof of Lemma 2

Lemma 2. *Suppose that Assumptions 1 and 3 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 (15) and adding $-F(w_*)$ for both sides, where $w_* = \arg\min_w F(w)$, we have

$$\begin{aligned}
\mathbb{E}[F(w_{t+1}^{(s)}) - F(w_*)] &\leq \mathbb{E}[F(w_t^{(s)}) - F(w_*)] - \frac{\eta}{2}\mathbb{E}[\|\nabla F(w_t^{(s)})\|^2] + \frac{\eta}{2}\mathbb{E}[\|\nabla F(w_t^{(s)}) - v_t^{(s)}\|^2] - \left(\frac{\eta}{2} - \frac{L\eta^2}{2}\right)\mathbb{E}[\|v_t^{(s)}\|^2] \\
&\overset{(14)}{\leq} \mathbb{E}[F(w_t^{(s)}) - F(w_*)] - \frac{\eta}{2}\mathbb{E}[\|\nabla F(w_t^{(s)})\|^2] \\
&\quad + \frac{\eta}{2}\frac{\eta L}{(2-\eta L)}\left(\mathbb{E}[\|v_0^{(s)}\|^2] - \mathbb{E}[\|v_t^{(s)}\|^2]\right) - \left(\frac{\eta}{2} - \frac{L\eta^2}{2}\right)\mathbb{E}[\|v_t^{(s)}\|^2] \\
&= \mathbb{E}[F(w_t^{(s)}) - F(w_*)] - \frac{\eta}{2}\mathbb{E}[\|\nabla F(w_t^{(s)})\|^2] \\
&\quad + \frac{\eta}{2}\left(\frac{\eta L}{(2-\eta L)}\left(\mathbb{E}[\|v_0^{(s)}\|^2] - \mathbb{E}[\|v_t^{(s)}\|^2]\right) - (1-L\eta)\,\mathbb{E}[\|v_t^{(s)}\|^2]\right) \\
&\overset{\eta\leq\frac{1}{L}}{\leq} \mathbb{E}[F(w_t^{(s)}) - F(w_*)] - \frac{\eta}{2}\mathbb{E}[\|\nabla F(w_t^{(s)})\|^2] \\
&\quad + \frac{\eta}{2}\left(\eta L\left(\mathbb{E}[\|v_0^{(s)}\|^2] - \mathbb{E}[\|v_t^{(s)}\|^2]\right) - (1-L\eta)\,\mathbb{E}[\|v_t^{(s)}\|^2]\right) \\
&= \mathbb{E}[F(w_t^{(s)}) - F(w_*)] - \frac{\eta}{2}\mathbb{E}[\|\nabla F(w_t^{(s)})\|^2] + \frac{\eta}{2}\left(L\eta\mathbb{E}[\|v_0^{(s)}\|^2] - \mathbb{E}[\|v_t^{(s)}\|^2]\right).
\end{aligned}$$

□

Proof of Theorem 3

Theorem 3 (Smooth general convex). *Suppose that Assumptions 1 and 3 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++*

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

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

$$T_s = \min_{t \geq 0} \left\{ t : \|v_t^{(s)}\|^2 < \gamma \|v_0^{(s)}\|^2 \right\} \ , \ s = 1, \ldots, S,$$

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

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

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

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

which is equivalent to

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

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

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

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). □

Proof of Corollary 3

Corollary 3 (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 $\eta = \mathcal{O}(\frac{1}{L})$. □

Proof of Theorem 4

Theorem 4 (Smooth strongly convex). *Suppose that Assumptions 1, 2 and 3 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,

$$\begin{aligned}
\mathbb{E}[F(w_{T_s}^{(s)}) - F(w_*)] &\leq \mathbb{E}[F(w_{T_s-1}^{(s)}) - F(w_*)] - \frac{\eta}{2}\mathbb{E}[\|\nabla F(w_{T_s-1}^{(s)})\|^2] \\
&\overset{(8)}{\leq} (1 - \mu\eta)\mathbb{E}[F(w_{T_s-1}^{(s)}) - F(w_*)] \\
&\leq (1 - \mu\eta)^{T_s}\mathbb{E}[F(w_0^{(s)}) - F(w_*)]
\end{aligned}$$

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

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

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 4

Corollary 4 (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

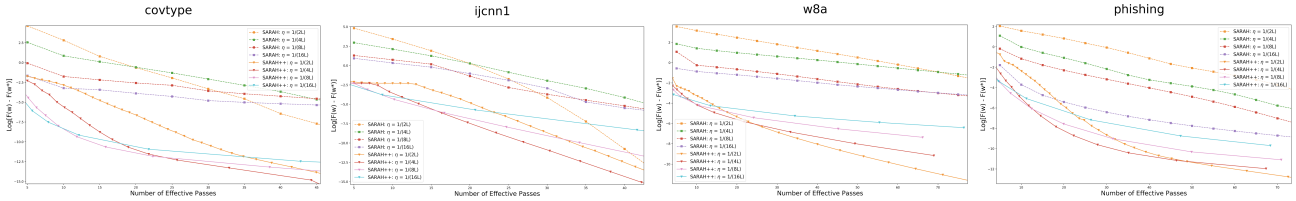

Figure 4: Comparisons of $\log[F(w) - F(w_*)]$ between SARAH++ and SARAH with different learning rates on *covtype*, *ijcnn1*, *w8a*, and *phishing* datasets

Additional experiments in Section 3.3

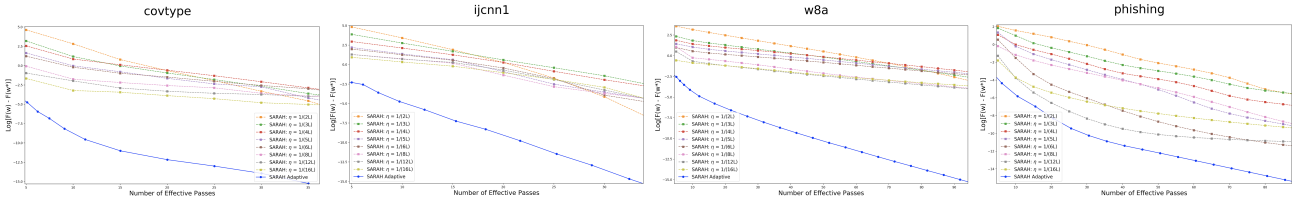

Figure 5: Comparisons of $\log[F(w) - F(w_*)]$ between SARAH Adaptive and SARAH with different learning rates on *covtype*, *ijcnn1*, *w8a*, and *phishing* datasets

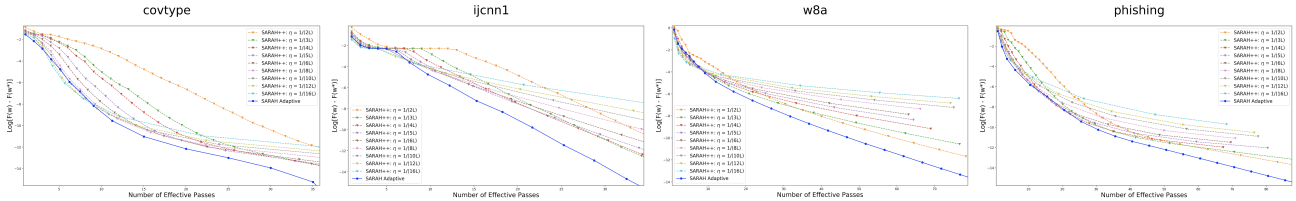

Figure 6: Comparisons of $\log[F(w) - F(w_*)]$ between SARAH Adaptive and SARAH++ with different learning rates on *covtype*, *ijcnn1*, *w8a*, and *phishing* data sets

Sensitivity of γ for SARAH Adaptive

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

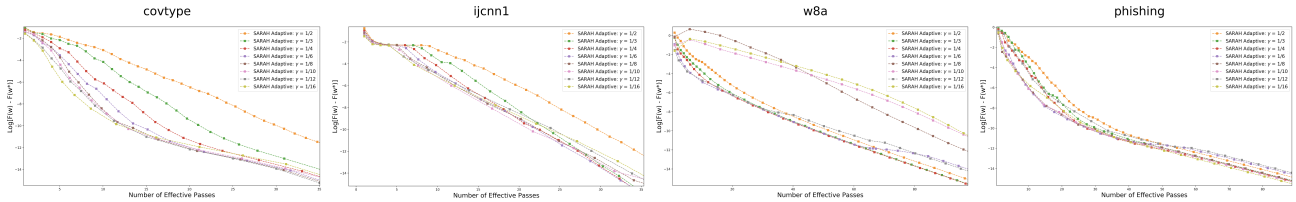

Figure 7: Comparisons of $\log[F(w) - F(w_*)]$ with different value of γ for SARAH Adaptive on *covtype*, *ijcnn1*, *w8a*, and *phishing* datasets

Maximum Inner Loop Size for SARAH++

We show the performance of SARAH++ for different values of the maximum inner loop size m. Figures 8 and 9 show that the effect of m is less important if m is large.

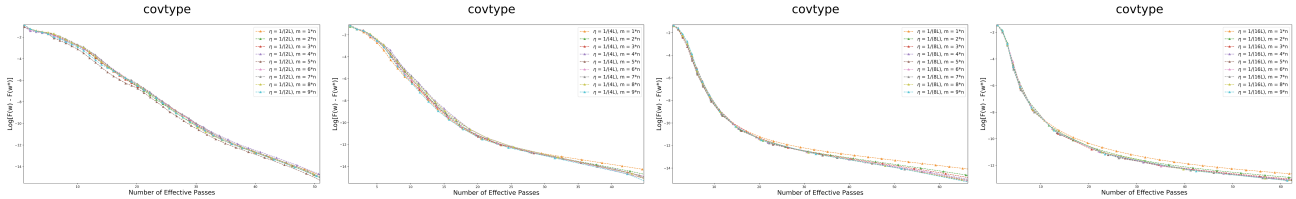

Figure 8: Comparisons of $\log[F(w) - F(w_*)]$ with different maximum inner loop size m and different learning rate of SARAH++ on *covtype* dataset

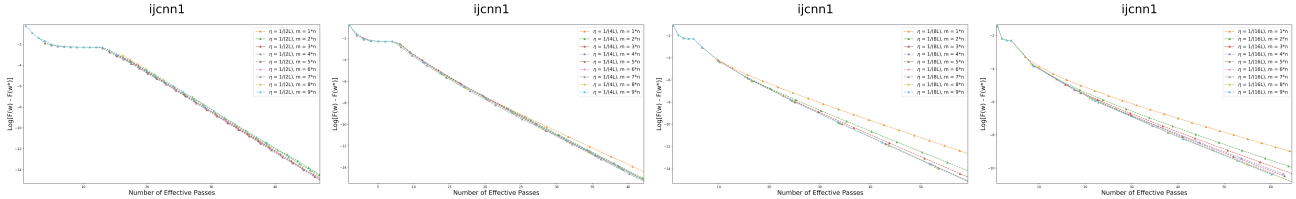

Figure 9: Comparisons of $\log[F(w) - F(w_*)]$ with different maximum inner loop size m and different learning rate of SARAH++ on *ijcnn1* dataset

Stopping of Outer S and Inner T_s for SARAH++

Let us choose $T = 25n$ and run SARAH++ and report the values of S and T_s. In Figures 10 and 11, the x-axis represents the number of outer iterations (s) and the y-axis represent the stopping time T_s corresponding to s.

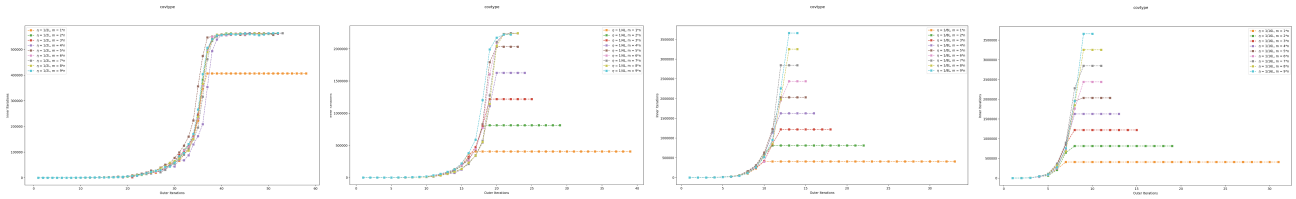

Figure 10: The stopping inner iterations T_s and the stopping outer iterations S for SARAH++ on *covtype* dataset

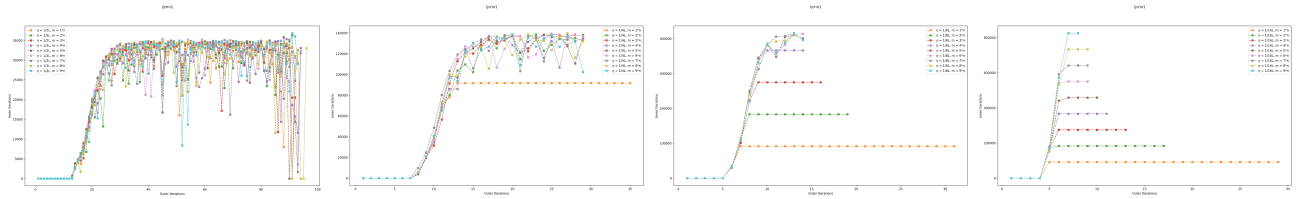

Figure 11: The stopping inner iterations T_s and the stopping outer iterations S for SARAH++ on *ijcnn1* dataset

Increasing controlled factor SARAH++

The performance of SARAH++ can be improved by properly choosing the controlled factor γ. In the following experiments, we change $\gamma = \gamma_s$ as a function of the outer loop iteration count s. Furthermore, we change the learning rate $\eta_s = \frac{\gamma_s}{L}$ at each outer iteration. We choose an initial $\gamma_1 = \frac{1}{16}$ at the 1-st outer iteration and then update the next ones with $\gamma_s = \min\{\gamma_{s-1}(1+\alpha), \beta\}$. In other words, we increase the value of γ_s at each outer iteration by multiplying with $(1+\alpha)$ until β is achieved. We denote this modification as *SARAH++ Adaptive* (Algorithm 5). We conduct experiments on the same problems and data sets as in the main paper. We tune among $\alpha = \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$ and $\beta = \{0.5, 0.6, 0.7, 0.8, 0.9\}$ in order to achieve the best performance.

Algorithm 5 SARAH++ Adaptive

Parameters: The total iteration $T > 0$, and the maximum inner loop size $m \leq T$, parameters $0 \leq \alpha < 1$ and $0 < \beta < 1$
Initialize: $\tilde{w}_0$, initial control factor γ_0
$G = 0$
Iterate:
$s = 0$
while $G < T$ **do**
  $s \leftarrow s + 1$
  $\gamma_s = \min\{\gamma_{s-1}(1+\alpha), \beta\}$
  Choose the learning rate $0 < \eta_s \leq \frac{\gamma_s}{L}$,
  $w_0^{(s)} = \tilde{w}_{s-1}$
  $v_0^{(s)} = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(w_0^{(s)})$
  $t = 0$
  while $\|v_t^{(s)}\|^2 \geq \gamma_s \|v_0^{(s)}\|^2$ **and** $t \leq m$ **do**
    $w_{t+1}^{(s)} = w_t^{(s)} - \eta_s v_t^{(s)}$
    $t \leftarrow t + 1$
    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
  $T_s = t$
  $\tilde{w}_s = w_{T_s}^{(s)}$
  $G \leftarrow G + T_s$
end while
$S = s$
Set $\hat{w} = \tilde{w}_S$

Figures 12, 13, and 14 show the performance of SARAH++ Adaptive for different values of α and β on different data sets.

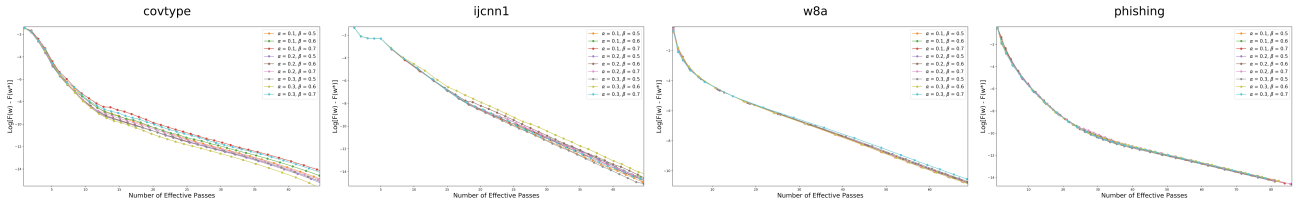

Figure 12: Comparisons of $\log[F(w) - F(w_*)]$ with different values of $\alpha = \{0.1, 0.2, 0.3\}$ and $\beta = \{0.5, 0.6, 0.7\}$ of SARAH++ on *covtype*, *ijcnn1*, *w8a*, and *phishing* datasets

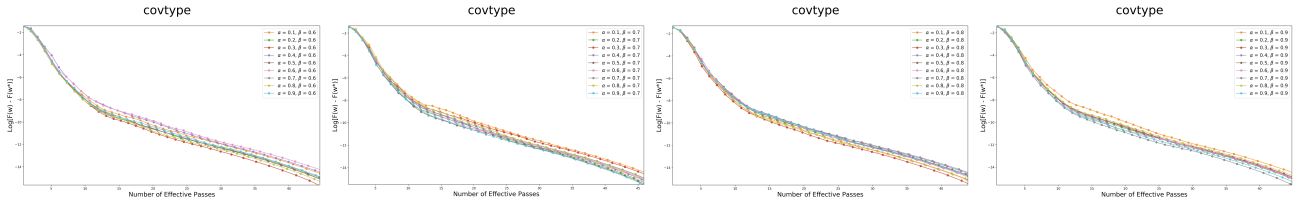

Figure 13: Comparisons of $\log[F(w) - F(w_*)]$ with different values of $\alpha = \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$ and $\beta = \{0.6, 0.7, 0.8, 0.9\}$ of SARAH++ on *covtype* dataset

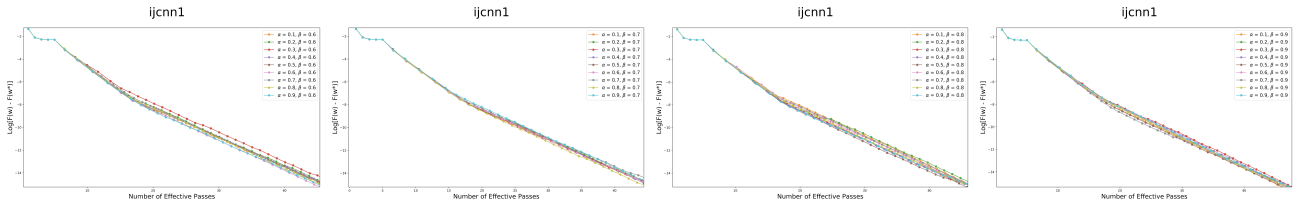

Figure 14: Comparisons of $\log[F(w) - F(w_*)]$ with different values of $\alpha = \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$ and $\beta = \{0.6, 0.7, 0.8, 0.9\}$ of SARAH++ on *ijcnn1* dataset

The limitation of this approach is how to properly choose values α and β. Indeed, the controlled factor γ is important and should be adaptive in order to improve the performance over SARAH++. The experimental results are promising and we leave this question as an open direction for future research.