A Lyapunov Analysis of Loopless SARAH

Ayoub El Hanchi
McGill University
ayoub.elhanchi@mail.mcgill.ca

Abstract

Of the recently proposed variance-reduced optimization algorithms, SARAH, an outer-inner loop algorithm, is particularly popular due to its superior performance in the non-convex regime. Its convergence in the strongly-convex case, however, is still not as well understood as that of other variance-reduced algorithms. In this short note, we propose a new Lyapunov function to study the convergence of loopless SARAH, a slight modification of the original algorithm, and recover the optimal non-accelerated complexity $O((n+\kappa)\log\frac{1}{\varepsilon})$ to reach an ε-accurate solution without assuming strong-convexity of the individual functions.

1 Introduction

We consider the unconstrained optimization problem:

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

where we assume that $F:\mathbb{R}^d\to\mathbb{R}$ is μ-strongly-convex, and $f_i:\mathbb{R}^d\to\mathbb{R}$ is convex and L-smooth for all $i\in[n]$. Problems of this form are ubiquitous in machine learning where F is usually the empirical risk.

The classical algorithm to solve problems of the form (1) is stochastic gradient descent (SGD) [11]. However, its convergence rate of $O(1/k)$ can be significantly improved upon using variance-reduced algorithms which converge linearly to the optimum (see, e.g., [6, 2]). Inner-outer loop algorithms form a sub-family of variance-reduced methods, and are particularly popular due to their low memory overhead.

Stochastic variance-reduced gradient (SVRG) [6] is the earliest inner-outer loop algorithm. While it is able to achieve the optimal non-accelerated complexity $O((n+\kappa)\log\frac{1}{\varepsilon})$ to reach an ε-accurate solution, its analysis suffers from two problems. First, to achieve this rate, SVRG requires an inner loop size of order $O(\kappa)$ where $\kappa=\frac{L}{\mu}$ is the condition number which is usually unknown. Second, the original algorithm requires setting the reference point to the average of the iterates of the inner loop, which, empirically, is found to be worse than using the last iterate of the inner loop as the reference point [12].

These observations led to the development of a slight variation of SVRG named loopless SVRG (L-SVRG) [5, 7, 12] which, instead of having a fixed inner loop size, updates the full gradient with some small probability q at each iteration. This leads to a much simpler analysis based on a Lyapunov function, and solves both problems that SVRG suffers from.

Another inner-outer loop algorithm is the stochastic recursive gradient (SARAH) [10] which gathered a lot of attention recently due to its superior performance in the non-convex regime [4, 13]. In the strongly-convex case however, its original form and analysis also suffer from the same problems as the original SVRG. Recently, [8] considered the loopless version of this algorithm, which we will refer to as L-SARAH. Similar to L-SVRG, in L-SARAH the full gradient is recomputed with some small probability q at each iteration. The analysis of L-SARAH presented in [8] however is more cumbersome than that of L-SVRG [7], and only yields the suboptimal complexity of $O((n+\kappa^2)\log\frac{1}{\varepsilon})$ to reach an ε-accurate solution under our assumptions (see Remark 1, [8]).

In this short note, we fill this gap in the literature and propose a new Lyapunov function to study the convergence of L-SARAH. This yields a much cleaner analysis, and recovers the optimal non-accelerated rate without assuming strong-convexity of the individual functions f_i.

2 Algorithm

Loopless SARAH is given in Algorithm 1. It has two parameters: the step size $\alpha > 0$ and the probability $q \in (0, 1]$ of performing a full gradient update at a given iteration. Note that as oppose to the original SARAH algorithm [10] and the loopless version proposed in [8], we do not require v_0 to be initialized to $\nabla F(x_0)$. Furthermore, no ad-hoc "step-back" is needed as in the algorithm proposed in [8].

Algorithm 1: Loopless SARAH (L-SARAH)

Parameters: step size $\alpha > 0$, probability $q \in (0, 1]$
Initialization: $x_0 \in \mathbb{R}^d, v_0 \in \mathbb{R}^d$
for $k = 0, 1, 2, \ldots$ **do**
  $x_{k+1} = x_k - \alpha v_k$
  sample $b_k \sim$ Bernoulli(q)
  if $b_k = 1$ **then**
    $v_{k+1} = \nabla F(x_{k+1})$
  else
    sample i_k uniformly at random from [n]
    $v_{k+1} = v_k + \nabla f_{i_k}(x_{k+1}) - \nabla f_{i_k}(x_k)$
  end
end

3 Assumptions & inequalities

Before we proceed with the convergence analysis of Algorithm 1, we first state our assumptions precisely.

Assumption 1. *The function $F : \mathbb{R}^d \to \mathbb{R}$ is differentiable and μ-strongly convex, that is:*

$$F(y) \geq F(x) + \langle \nabla F(x), y - x \rangle + \frac{\mu}{2} \|y - x\|_2^2 \quad \forall x, y \in \mathbb{R}^d$$

Assumption 2. *For all $i \in [n]$, the functions $f_i : \mathbb{R}^d \to \mathbb{R}$ are differentiable and convex, that is:*

$$f_i(y) \geq f_i(x) + \langle \nabla f_i(x), y - x \rangle \quad \forall x, y \in \mathbb{R}^d$$

Assumption 3. *For all $i \in [n]$, the functions $f_i : \mathbb{R}^d \to \mathbb{R}$ are L-smooth, that is:*

$$f_i(y) \leq f_i(x) + \langle \nabla f_i(x), y - x \rangle + \frac{L}{2} \|y - x\|_2^2 \quad \forall x, y \in \mathbb{R}^d$$

Note that by averaging the n inequalities of Assumption 3, we have that $F : \mathbb{R}^d \to \mathbb{R}$ is also L-smooth. Furthermore, by combining Assumptions 2 and 3, we have the inequality (see [9], Theorem 2.1.5):

Lemma 1. *For all $i \in [n]$ we have:*

$$\langle \nabla f_i(y) - \nabla f_i(x), y - x \rangle \geq \frac{1}{L} \|\nabla f_i(y) - \nabla f_i(x)\|_2^2 \quad \forall x, y \in \mathbb{R}^d$$

Finally we will make use of a version of Young's inequality:

Lemma 2. *Let $a, b \in \mathbb{R}^d$, and let $\beta > 0$. Then:*

$$\|a + b\|_2^2 \leq (1 + \beta) \|a\|_2^2 + (1 + \beta^{-1}) \|b\|_2^2$$

4 Convergence Analysis

We propose the following Lyapunov function to study the convergence of L-SARAH:

$$T^k \coloneqq T(x_k, v_k) \coloneqq a \|\nabla F(x_k) - v_k\|_2^2 + b \|v_k\|_2^2 + \left(F(x_k) - F^* + \frac{\mu}{2} \|x_k - x^*\|_2^2 \right) \tag{2}$$

where x^* is the minimizer of F, $F^* \coloneqq F(x^*)$, and a and b are free parameters which will be set during the analysis. We note that the term in parentheses is the usual Lyapunov function used to study the convergence of gradient descent. Our main result is the following theorem:

Theorem 1. *Suppose x_k and v_k are evolved according to Algorithm 1, and suppose that Assumptions 1, 2 and 3 hold. Then for any $\alpha \leq \frac{1}{4L}$, $q \in (0, 1]$, and any $k \in \mathbb{N}$ we have:*

$$\mathbb{E}\left[T^k\right] \leq (1 - \rho)^k T^0$$

where $\rho \coloneqq \min\{\frac{\alpha\mu}{2}, \frac{q}{2}\}$, and the parameters a and b of the Lyapunov function are set to $a \coloneqq \frac{3\alpha}{q}$ and $b \coloneqq \frac{3\alpha}{7q}$. By taking $q = \frac{1}{n}$ and $\alpha = \frac{1}{4L}$, this implies an iteration complexity of $O((n + \kappa) \log \frac{1}{\varepsilon})$ to reach an ε-accurate solution using Algorithm 1.

4.1 Useful Lemmas

Before proceeding with the proof of the theorem, we state and prove a few useful lemmas whose proofs are taken from the original analysis of SARAH found in [10].

Lemma 3. *Conditional on $(i_t)_{t=0}^{k-1}$ and $(b_t)_{t=0}^{k-1}$, we have:*

$$\mathbb{E}\left[\|\nabla F(x_{k+1}) - v_{k+1}\|_2^2 \mid b_k = 0\right] = \|\nabla F(x_k) - v_k\|_2^2 + \left(\frac{1}{n} \sum_{i=1}^{n} \|\nabla f_i(x_{k+1}) - \nabla f_i(x_k)\|_2^2 \right) - \|\nabla F(x_{k+1}) - \nabla F(x_k))\|_2^2$$

where the expectation is taken with respect to i_k.

Proof of Lemma 3. This proof is taken from the proof of Lemma 2 in [10]. Conditional on $(i_t)_{t=0}^{k-1}$, $(b_t)_{t=0}^{k-1}$, and $b_k = 0$, we have:

$$\begin{aligned}
\|\nabla F(x_{k+1}) - v_{k+1}\|_2^2 &= \|[\nabla F(x_k) - v_k] + [\nabla F(x_{k+1}) - \nabla F(x_k)] - [v_{k+1} - v_k]\|_2^2 \\
&= \|\nabla F(x_k) - v_k\|_2^2 + \|\nabla F(x_{k+1}) - \nabla F(x_k)\|_2^2 + \|v_{k+1} - v_k\|_2^2 \\
&+ 2\langle \nabla F(x_k) - v_k, \nabla F(x_{k+1}) - \nabla F(x_k)\rangle \\
&- 2\langle \nabla F(x_k) - v_k, v_{k+1} - v_k\rangle \\
&- 2\langle \nabla F(x_{k+1}) - \nabla F(x_k), v_{k+1} - v_k\rangle
\end{aligned}$$

Taking expectation with respect to i_k we have:

$$\mathbb{E}\left[v_{k+1} - v_k\right] = \mathbb{E}\left[\nabla f_{i_k}(x_{k+1}) - \nabla f_{i_k}(x_k)\right] = \nabla F(x_{k+1}) - \nabla F(x_k)$$

and:

$$\mathbb{E}\left[\|v_{k+1} - v_k\|_2^2\right] = \mathbb{E}\left[\|\nabla f_{i_k}(x_{k+1}) - \nabla f_{i_k}(x_k)\|_2^2\right] = \frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(x_{k+1}) - \nabla f_i(x_k)\|_2^2$$

Replacing in the initial equality we get the result. □

Lemma 4. *Conditional on $(i_t)_{t=0}^{k-1}$ and $(b_t)_{t=0}^{k-1}$, we have:*

$$\mathbb{E}\left[\|v_{k+1}\|_2^2 \mid b_k = 0\right] \leq \|v_k\|_2^2 + \left(1 - \frac{2}{\alpha L}\right)\left(\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(x_{k+1}) - \nabla f_i(x_k)\|_2^2\right)$$

where the expectation is taken with respect to i_k.

Proof of Lemma 4. This proof is taken from the proof of Lemma 3 in [10]. Conditional on $(i_t)_{t=0}^{k-1}$, $(b_t)_{t=0}^{k-1}$, and $b_k = 0$, we have:

$$\begin{aligned}
\|v_{k+1}\|_2^2 &= \|v_k + [\nabla f_{i_k}(x_{k+1}) - \nabla f_{i_k}(x_k)]\|_2^2 \\
&= \|v_k\|_2^2 + 2\langle v_k, \nabla f_{i_k}(x_{k+1}) - \nabla f_{i_k}(x_k)\rangle + \|\nabla f_{i_k}(x_{k+1}) - \nabla f_{i_k}(x_k)\|_2^2 \\
&= \|v_k\|_2^2 - \frac{2}{\alpha}\langle x_{k+1} - x_k, \nabla f_{i_k}(x_{k+1}) - \nabla f_{i_k}(x_k)\rangle + \|\nabla f_{i_k}(x_{k+1}) - \nabla f_{i_k}(x_k)\|_2^2 \\
&\leq \|v_k\|_2^2 + \left(1 - \frac{2}{\alpha L}\right)\|\nabla f_{i_k}(x_{k+1}) - \nabla f_{i_k}(x_k)\|_2^2
\end{aligned}$$

where in the third line we use the identity $v_k = -\frac{x_{k+1} - x_k}{\alpha}$ and the last inequality uses Lemma 1. Taking expectation with respect to i_k gives the stated result. □

4.2 Proof of Theorem

We are now ready to prove Theorem 1. Recall the Lyapunov function under consideration:

$$T^k \coloneqq T(x_k, v_k) \coloneqq a\|\nabla F(x_k) - v_k\|_2^2 + b\|v_k\|_2^2 + \left(F(x_k) - F^* + \frac{\mu}{2}\|x_k - x^*\|_2^2\right)$$

Proof of Theorem 1. Our goal is to express T^{k+1} as T^k. All the expectations in this proof are conditional on $(i_t)_{t=0}^{k-1}$ and $(b_t)_{t=0}^{k-1}$.

Taking expectation with respect to i_k and b_k, the first term expands as:

$$\begin{aligned}
&\mathbb{E}\left[\|\nabla F(x_{k+1}) - v_{k+1}\|_2^2\right] = \mathbb{P}(b_k = 0)\mathbb{E}\left[\|\nabla F(x_{k+1}) - v_{k+1}\|_2^2 \mid b_k = 0\right] + \mathbb{P}(b_k = 1)\cdot 0 \\
&= (1-q)\mathbb{E}\left[\|\nabla F(x_{k+1}) - v_{k+1}\|_2^2 \mid b_k = 0\right] \\
&= (1-q)\|\nabla F(x_k) - v_k\|_2^2 \\
&\quad + (1-q)\left(\frac{1}{n}\sum_{i=1}^n \|\nabla f_i(x_{k+1}) - \nabla f_i(x_k)\|_2^2\right) - (1-q)\|\nabla F(x_{k+1}) - \nabla F(x_k))\|_2^2
\end{aligned}$$

where in the first line we used the fact that when $b_k = 1$ we have $v_{k+1} = \nabla F(x_{k+1})$, and the last equality follows from Lemma 3.

The second term expands as:

$$\begin{aligned}
&\mathbb{E}\left[\|v_{k+1}\|_2^2\right] = \mathbb{P}(b_k = 0)\mathbb{E}\left[\|v_{k+1}\|_2^2 \mid b_k = 0\right] + \mathbb{P}(b_k = 1)\|\nabla F(x_{k+1})\|_2^2 \\
&\le (1-q)\|v_k\|_2^2 + (1-q)\left(1 - \frac{2}{\alpha L}\right)\left(\frac{1}{n}\sum_{i=1}^n \|\nabla f_i(x_{k+1}) - \nabla f_i(x_k)\|_2^2\right) \\
&\quad + q(1+\gamma)\|\nabla F(x_{k+1}) - \nabla F(x_k)\|_2^2 + q(1+\gamma^{-1})\|\nabla F(x_k)\|_2^2
\end{aligned}$$

where the second line follows from Lemma 4 applied to the first term, and Lemma 2 applied to the decomposition $\|\nabla F(x_{k+1})\|_2^2 = \|\nabla F(x_{k+1}) - \nabla F(x_k) + \nabla F(x_{k+1})\|_2^2$.

The sub-optimality can be bounded by:

$$\begin{aligned}
&F(x_{k+1}) - F^* = F(x_{k+1}) - F(x_k) + F(x_k) - F^* \\
&\le \langle \nabla F(x_k), x_{k+1} - x_k\rangle + \frac{L}{2}\|x_{k+1} - x_k\|_2^2 + F(x_k) - F^* \\
&= F(x_k) - F^* - \alpha\langle v_k, \nabla F(x_k)\rangle + \frac{\alpha^2 L}{2}\|v_k\|_2^2 \\
&= F(x_k) - F^* - \frac{\alpha}{2}\|\nabla F(x_k)\|_2^2 + \frac{\alpha}{2}\|\nabla F(x_k) - v_k\|_2^2 + \left(\frac{\alpha^2 L}{2} - \frac{\alpha}{2}\right)\|v_k\|_2^2
\end{aligned}$$

where the second line follows from the L-smoothness of F, and the fourth line from the identity $2\langle a, b\rangle = \|a\|_2^2 + \|b\|_2^2 - \|a-b\|_2^2$

Finally, the last term is bounded by:

$$\begin{aligned}
&\|x_{k+1} - x^*\|_2^2 = \|x_k - \alpha v_k - x^*\|_2^2 \\
&= \|x_k - x^*\|_2^2 - 2\alpha\langle v_k, x_k - x^*\rangle + \alpha^2\|v_k\|_2^2 \\
&= \|x_k - x^*\|_2^2 - 2\alpha\langle \nabla F(x_k), x_k - x^*\rangle + 2\alpha\langle \nabla F(x_k) - v_k, x_k - x^*\rangle + \alpha^2\|v_k\|_2^2 \\
&\le (1+\alpha\beta)\|x_k - x^*\|_2^2 - 2\alpha\left[F(x_k) - F^* + \frac{\mu}{2}\|x_k - x^*\|_2^2\right] + \alpha\beta^{-1}\|\nabla F(x_k) - v_k\|_2^2 + \alpha^2\|v_k\|_2^2
\end{aligned}$$

where in the last line we use the strong-convexity of F to bound the first inner product, and we use the inequality $2\langle a, b\rangle = 2\langle \frac{a}{\sqrt{\beta}}, \sqrt{\beta} b\rangle \le \beta^{-1}\|a\|_2^2 + \beta\|b\|_2^2$ to bound the second inner product.

Combining all the bounds we have the global inequality:

$$\begin{aligned}\mathbb{E}\left[T^{k+1}\right] \leq & \left(1-q+\frac{\alpha}{2a}[1+\mu\beta^{-1}]\right) a\left\|\nabla F(x_k)-v_k\right\|_2^2 \\ & +\left(1-q+\frac{L+\mu}{2b}\alpha^2-\frac{\alpha}{2b}\right) b\left\|v_k\right\|_2^2 \\ & +\left(1+\alpha\beta-\alpha\mu\right)\left[F(x_k)-F^*+\frac{\mu}{2}\left\|x_k-x^*\right\|\right] \\ & +\left(qb(1+\gamma^{-1})-\frac{\alpha}{2}\right)\left\|\nabla F(x_k)\right\|_2^2 \\ & +\left(qb(1+\gamma)-(1-q)a\right)\left\|\nabla F(x_{k+1})-\nabla F(x_k)\right\|_2^2 \\ & +(1-q)\left(a+b\left[1-\frac{2}{\alpha L}\right]\right)\left[\frac{1}{n}\sum_{i=1}^n\left\|\nabla f_i(x_{k+1})-\nabla f_i(x_k)\right\|_2^2\right]\end{aligned}$$

By choosing the free parameters as:

$$\beta=\frac{\mu}{2},\quad \gamma=\frac{6}{q},\quad a=\frac{3\alpha}{q},\quad b=\frac{3\alpha}{7q}$$

and taking a step size:

$$\alpha\leq\frac{1}{4L}$$

The last three parentheses are guaranteed to be less than 0, the first two are less than $1-\frac{q}{2}$ and the third is less than $1-\frac{\alpha\mu}{2}$. Defining:

$$\rho=\min\left\{\frac{\alpha\mu}{2},\frac{q}{2}\right\}$$

We get the inequality:

$$\mathbb{E}\left[T^{k+1}\right]\leq(1-\rho)T^k$$

Applying this inequality recursively on the right-hand side and taking successive expectations, we get the desired result. □

5 Discussion

In this short note, we proposed and analyzed a loopless version of the SARAH algorithm for the optimization of finite-sum objectives of the form (1). To the best of our knowledge, this is the first convergence proof of loopless SARAH that achieves a complexity of $O((n+\kappa)\log\frac{1}{\varepsilon})$ to reach an ε-accurate solution and which does not assume strong-convexity of the individual functions f_i. Furthermore, the Lyapunov function we propose is new, and could potentially be used to gain insight into the behavior of SARAH. One peculiar property of this Lyapunov function is that it does not extend to arbitrary couplings: if two sequences $(x_k)_{k=0}^\infty$, $(y_k)_{k=0}^\infty$ evolve according to Algorithm 1, it is unclear how to modify the Lyapunov function to directly show a contraction between the two sequences. This is in contrast with the Lyapunov functions used to study the convergence of L-SVRG and SAGA, and makes it difficult to use L-SARAH for sampling as L-SVRG and SAGA are used [1, 3]. On the other hand, the fact that L-SARAH can only be used as an optimization algorithm and not as a sampling algorithm might help explain its superior performance in the non-convex case.

References

[1] Niladri S. Chatterji, Nicolas Flammarion, Yian Ma, Y. Ma, P. Bartlett, and Michael I. Jordan. On the theory of variance reduction for stochastic gradient monte carlo. *ArXiv*, abs/1802.05431, 2018.

[2] Aaron Defazio, Francis R. Bach, and S. Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *NIPS*, 2014.

[3] Kumar Avinava Dubey, S. Reddi, Sinead Williamson, B. Póczos, Alex Smola, and E. Xing. Variance reduction in stochastic gradient langevin dynamics. *Advances in neural information processing systems*, 29:1154–1162, 2016.

[4] C. Fang, C. Li, Zhouchen Lin, and T. Zhang. Spider: Near-optimal non-convex optimization via stochastic path integrated differential estimator. In *NeurIPS*, 2018.

[5] T. Hofmann, A. Lucchi, S. Lacoste-Julien, and Brian McWilliams. Variance reduced stochastic gradient descent with neighbors. In *NIPS*, 2015.

[6] R. Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *NIPS*, 2013.

[7] D. Kovalev, S. Horvath, and Peter Richtárik. Don't jump through hoops and remove those loops: Svrg and katyusha are better without the outer loop. In *ALT*, 2020.

[8] B. Li, Meng Ma, and Georgios B. Giannakis. On the convergence of sarah and beyond. In *AISTATS*, 2020.

[9] Yurii E. Nesterov. *Introductory Lectures on Convex Optimization - A Basic Course*, volume 87 of *Applied Optimization*. Springer, 2004.

[10] Lam M. Nguyen, J. Liu, K. Scheinberg, and Martin Takác. Sarah: A novel method for machine learning problems using stochastic recursive gradient. In *ICML*, 2017.

[11] H. Robbins. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 2007.

[12] Othmane Sebbouh, Nidham Gazagnadou, Samy Jelassi, Francis R. Bach, and Robert Mansel Gower. Towards closing the gap between the theory and practice of svrg. *ArXiv*, abs/1908.02725, 2019.

[13] Z. Wang, Kaiyi Ji, Yi Zhou, Y. Liang, and V. Tarokh. Spiderboost and momentum: Faster variance reduction algorithms. In *NeurIPS*, 2019.