Local Linear Convergence of Variance Reduced Stochastic Gradient Methods for Low Complexity Regularization

Clarice Poon  
Department of Mathematical Sciences, University of Bath  
Claverton Down, Bath, BA2 7AY, United Kingdom  
Email: C.M.H.S.Poon@maths.cam.ac.uk

Jingwei Liang, Carola-Bibiane Schönlieb  
DAMTP, University of Cambridge  
Wilberforce Road, Cambridge, CB3 0WA, United Kingdom  
Email: {jq993, cbs31}@cam.ac.uk

Abstract—We present a local convergence analysis for proximal variance reduced stochastic gradient methods: SAGA [4] and Prox-SVRG [6]. Under the assumption that the non-smooth function is partly smooth, we present an analysis for the local convergence behaviours of SAGA/Prox-SVRG: (i) the sequences generated by the methods are able to identify the smooth manifold in a finite number of iterations; (ii) then the sequence enters a local linear convergence regime. The result is illustrated by several concrete examples and supported by numerical experiments.

I. INTRODUCTION

Consider the following convex structured optimization problem
\[
\min_{x \in R^n} \{ \Phi(x) \triangleq R(x) + \frac{1}{m} \sum_{i=1}^{m} f_i(x) \}, \quad (P)
\]
where \( R \in \Gamma_c(R^n) \), the set of proper convex and lower semi-continuous on \( R^n \), and for each \( i = 1, \ldots, m \), \( f_i \in C^1(R^n) \) is convex with \( \nabla f_i \) being \( L \)-Lipschitz. We assume that \( \text{Argmin} \Phi \neq \emptyset \).

In this paper, we consider two variance reduced stochastic gradient methods which are widely used in machine learning: SAGA [4] and Prox-SVRG [6] to solve (P). Given an initial point \( x_0 \), define the individual gradient \( g_{0,i} \equiv \nabla f_i(x_0), i = 1, \ldots, m \). Then, the iteration of SAGA reads
\[
\begin{aligned}
\text{sample } i_k \text{ uniformly from } \{1, \ldots, m\}, \\
w_k = x_k - \gamma_k (\nabla f_i(x_k) - g_k,i) + \frac{1}{m} \sum_{i=1}^{m} g_k,i, \\
x_{k+1} = \text{prox}_{\gamma_k R}(w_k), \\
g_{k+1} = \begin{cases} 
\nabla f_i(x_{k+1}) & \text{if } i = i_k, \\
g_k - i_k, & \text{o.w.}
\end{cases}
\end{aligned}
\] (I.1)

For \( \gamma > 0 \), \( \text{prox}_{\gamma R}(x) \equiv \text{argmin}_{z \in R^n} \gamma R(z) + \frac{1}{2z} \|z - x\|^2 \) is called the proximal mapping of \( \gamma R \). Now let \( P \) be a positive integer, the iteration of Prox-SVRG is
\[
\begin{aligned}
\tilde{g}_k & = \frac{1}{m} \sum_{i=1}^{m} \nabla f_i(\tilde{x}_k), \\
x_{k+1} = x_k - \gamma_k (\tilde{g}_k - g_k) - \gamma_k (\nabla f_i)p(x_k) - \nabla f_{i_k}(\tilde{x}_k) - g_k, \\
\tilde{x}_{k+1} = \text{prox}_{\gamma_k R}(w_k).
\end{aligned}
\] (I.2)

II. PARTIALLY SMOOTHNESS AND FINITE IDENTIFICATION

The class of partly smooth functions [2], is specialized here to functions in \( \Gamma_0(R^n) \). Denote \( \text{par}(C) \) the linear subspace parallel to the non-empty convex set \( C \subseteq R^n \), and \( \text{ri}(C) \) its relative interior.

Definition II.1. Let \( R \in \Gamma_0(R^n) \) and \( x \in R^n \) such that \( \partial R(x) \neq \emptyset \). \( R \) is partly smooth at \( x \) relative to a set \( M \) containing \( x \) if
(Smoothness) \( M \) is a \( C^2 \)-manifold, \( R|_M \in C^2 \) around \( x \).
(Sharpness) The tangent space \( T_M(x) = T_x \text{ par}(\partial R(x)) \).
(Continuity) The \( \partial R \) is continuous at \( x \) relative to \( M \).

Examples of such functions are given in Section IV, see also [3].

Theorem II.2 (Finite activity identification). The sequence \( x_k \) generated by SAGA/Prox-SVRG converges almost surely to a minimizer \( x^* \) of (P). Suppose that \( R \) is partly smooth at \( x^* \) relative to \( M_{x^*} \), and \( -\nabla F(x^*) \in \text{ri}(\partial R(x^*)) \), then there exists a \( K > 0 \) such that for all \( k \geq K \), \( x_k \in M_{x^*} \), almost surely.

Condition \(-\nabla F(x^*) \in \text{ri}(\partial R(x^*))\) can be viewed as a geometric generalization of the strict complementarity of non-linear programming, and is almost necessary for the finite identification [2].

III. LOCAL LINEAR CONVERGENCE

We now turn to the local linear convergence of the SAGA/Prox-SVRG with partly smooth functions. Denote \( T_{x^*} \), the orthonormal projection operator onto \( T_{x^*} \), and \( \text{Id} \) the identity operator.

Theorem III.1. We assume the conditions of Theorem II.2 hold. If moreover \( F \in C^2 \) near \( x^* \) and there exists \( \alpha > 0 \) such that \( \text{Pr}_{T_{x^*}} \nabla^2 F(x^*)_{T_{x^*}} \geq \alpha I \). Then for all \( k \) large enough, we have
1) SAGA: if \( \gamma_k \equiv \frac{\alpha}{\gamma} \), then
\[
\text{E}[\|x_k - x^*\|^2] = O(\gamma^k),
\]
where \( \gamma = 1 - \min \left\{ \frac{1}{\alpha}, \frac{\gamma}{\alpha} \right\} \).
2) Prox-SVRG: if \( \gamma, P \) are chosen such that \( \eta = \max \{1 - \gamma \alpha, \gamma \alpha L(P + 1) \} < 1 \), then
\[
\text{E}[\|\tilde{x}_k - x^*\|^2] = O(\gamma^k).
\]

IV. NUMERICAL EXPERIMENTS

Example IV.1. \( \ell_1 \)-norm is partly smooth relative to \( M = \{ u \in R^n : \text{supp}(u) \subseteq \text{supp}(x) \} \). Example IV.2. \( \ell_1,2 \)-norm is partly smooth relative to \( M = \{ u \in R^n : \text{supp}(u) \subseteq \text{supp}(x) \} \), where \( \text{supp}(u) = \bigcup_{b \in B} \{ x : x_b \neq 0 \} \), and \( \bigcup_{b \in B} B = \{ 1, \ldots, n \} \).

Example IV.3. TV semi-norm \( \|x\|_{TV} = \|\nabla x\|_1 \) is partly smooth relative to \( M = \{ u \in R^n : \text{supp}(\nabla u) \subseteq I \} \). (Supp(\nabla x).)

Example IV.4. Nuclear norm is partly smooth relative to the manifold of fixed rank matrices, \( M = \{ z \in R^{n \times m} : \text{rank}(z) = r \} \).

Let \( m > 0 \) and \( \{ z_i, y_i \} \in R^{n \times m} \times \{ \pm 1 \} \), \( i = 1, \ldots, m \) be the training set. The sparse logistic regression is to find a linear decision function which minimises the objective
\[
\min_{x,b} \frac{1}{m} \sum_{i=1}^{m} \log (1 + e^{-y_i f(x_i,x;b)}),
\]
where \( f(z;x,b) = b^T z^T x \). The setting of the experiment is: \( n = 256 \), \( m = 128 \), \( \mu = 1/\sqrt{m} \) and \( L = 1188 \). Notice that, the dimension of the problem is larger than the number of training points. The parameters choices of SAGA and Prox-SVRG are:

SAGA: \( \gamma = \frac{1}{L^2} \), Prox-SVRG: \( \gamma = \frac{1}{L^2}, P = m \).

The convergence profiles are depicted in Figure 1.
Fig. 1: Finite manifold identification and local linear convergence of SAGA and Prox-SVRG for solving the sparse logistic regression problem. (a) manifold identification; (b) local linear convergence. Restricted to the manifold $M_{x^*}$, functions $f_i, i = 1, ..., m$ will have better Lipschitz constant, which allows us to accelerate SAGA/Prox-SVRG; See the dashed line of the right figure. The reason of choosing different step-sizes for SAGA and Prox-SVRG is only to distinguish the red and black plots above. For the considered synthetic example, the performance of the two algorithms are almost the same under same step-size.

REFERENCES