Local Convergence Properties of SAGA/Prox-SVRG and Acceleration

Clarice Poon [*1] Jingwei Liang [*1] Carola-Bibiane Schönlieb [1]

Abstract

In this paper, we present a local convergence analysis for a class of stochastic optimisation methods: the proximal variance reduced stochastic gradient methods, and mainly focus on SAGA (Defazio et al., 2014) and Prox-SVRG (Xiao & Zhang, 2014). Under the assumption that the non-smooth component of the optimisation problem is partly smooth relative to a smooth manifold, we present a unified framework for the local convergence analysis 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. Furthermore, we discuss various possibilities for accelerating these algorithms, including adapting to better local parameters, and applying higher-order deterministic/stochastic optimisation methods which can achieve super-linear convergence. Several concrete examples arising from machine learning are considered to demonstrate the obtained result.

1. Introduction

1.1. Non-smooth Optimisation

Modern optimisation has become a core part of many fields, such as machine learning and signal/image processing. In a world of increasing data demands, there are two key driving forces behind modern optimisation.

- **Non-smooth regularisation.** We are often faced with models of high complexity, however, the solutions of interest often lie on a manifold of low dimension which is promoted by the non-smooth regularisers. There have been several recent studies explaining how proximal methods identify this low dimensional manifold and efficiently output solutions which take a certain structure; see for instance (Liang et al., 2017) for the case of deterministic proximal gradient methods.
- **Stochastic methods.** The past decades have seen an exponential growth in the data sizes that we have to handle, and stochastic methods have been popular due to their low computational cost; see for instance (Schmidt et al., 2017; Defazio et al., 2014; Xiao & Zhang, 2014) and references therein.

The purpose of this paper is to show that proximal variance reduced stochastic gradient methods allow to benefit from both efficient structure enforcement and low computational cost. In particular, we present a study of manifold identification and local acceleration properties of these methods when applied to the following minimisation problem:

$$\min_{x\in\mathbb{R}^n} \ \Phi(x) \overset{\text{def}}{=} R(x) + F(x), \qquad (\mathcal{P})$$

where $R(x)$ is a non-smooth structure imposing penalty term, and

$$F(x) \overset{\text{def}}{=} \frac{1}{m}\sum_{i=1}^{m} f_i(x),$$

where each f_i is continuously differentiable.

We are interested in the problems where the value of m is very large. A typical example of ($\mathcal{P}$) is the LASSO problem, which reads

$$\min_{x\in\mathbb{R}^n} \mu\|x\|_1 + \frac{1}{m}\sum_{i=1}^{m} \frac{1}{2}\|\mathcal{K}_i x - b_i\|^2,$$

where $\mu > 0$ is a positive trade-off parameter, $\mathcal{K}_i$ is the i^{th} row of a matrix $\mathcal{K} \in \mathbb{R}^{m\times n}$, and b_i is the i^{th} element of the vector $b \in \mathbb{R}^m$. Throughout this paper, we consider the following basic assumptions for problem ($\mathcal{P}$):

(A.1) $R : \mathbb{R}^n \to \mathbb{R}\cup\{+\infty\}$ is proper, convex and lower semi-continuous;

(A.2) $F : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable with ∇F being L_F-Lipschitz continuous. For each index $i = 1, \cdots, m$, f_i is continuously differentiable with L_i-Lipschitz continuous gradient;

(A.3) $\mathrm{Argmin}(\Phi) \neq \emptyset$, that is the set of minimisers is non-empty.

In addition to (**A.2**), define $L \overset{\text{def}}{=} \max_{i=\{1,\cdots,m\}} L_i$, which is the uniform Lipschitz continuity of functions f_i. Note that $L_F \le \frac{1}{m}\sum_i L_i \le L$ holds.

[*] Equal contribution [1] DAMTP, University of Cambridge, Cambridge, United Kingdom. Correspondence to: Jingwei Liang <jl993@cam.ac.uk>, Clarice Poon <C.M.H.S.Poon@maths.cam.ac.uk>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018. Copyright 2018 by the author(s).

1.2. Deterministic Forward–Backward Splitting

A classical approach to solve ($\mathcal{P}$) is the Forward–Backward splitting (FBS) method (Lions & Mercier, 1979), which is also known as *proximal gradient descent method*. The standard FBS iteration reads

$$x_{k+1} = \mathrm{prox}_{\gamma_k R}\big(x_k - \gamma_k \nabla F(x_k)\big),\ \ \gamma_k \in]0, \tfrac{2}{L_F}[, \quad (1)$$

where $\mathrm{prox}_{\gamma R}$ is the *proximity operator* of R defined by

$$\mathrm{prox}_{\gamma R}(\cdot) \overset{\text{def}}{=} \min_{x\in\mathbb{R}^n} \gamma R(x) + \tfrac{1}{2}\|x - \cdot\|^2. \quad (2)$$

In general, the advantages of FBS can be summarised as:

- **Robust convergence guarantees.** Both sequence and objective function are convergent for $0 < \underline{\gamma} \leq \gamma_k \leq \bar{\gamma} < 2/L_F$ where $\underline{\gamma}, \bar{\gamma} > 0$ (Combettes & Wajs, 2005);
- **Known convergence rates.** There holds $\|x_k - x_{k-1}\| = o(1/\sqrt{k})$ (Liang et al., 2016), and $\Phi(x_k) - \Phi(x^\star) = o(1/k)$ (Molinari et al., 2018) where $x^\star \in \mathrm{Argmin}(\Phi)$. These rates can be improved to linear[1] if for instance Φ is strongly convex (Nesterov, 2004);
- **Numerous acceleration techniques.** Such as the inertial schemes including inertial FBS (Liang et al., 2017), FISTA (Beck & Teboulle, 2009) and Nesterov's optimal methods (Nesterov, 2004);
- **Structure adaptivity.** Several recent work (Liang et al., 2014; 2017) studies the manifold identification properties of FBS. It is shown in (Liang et al., 2017) that within finite time, the FBS iterates x_k all lie on the same manifold as the optimal solution $x^\star$. Furthermore, upon identifying this manifold, the FBS iterates can be proved to converge linearly to the optimal solution $x^\star$.

However, despite these advantages, for problem ($\mathcal{P}$), when m is very large, computing $\nabla F(x_k)$ could be very expensive, which makes the deterministic FBS-type methods unsuitable for solving large-scale problems.

1.3. Proximal Stochastic Gradient Descent

The most straightforward extension of stochastic gradient descent to problem ($\mathcal{P}$) is the *proximal stochastic gradient descent* (Prox-SGD), which reads, for $k = 0, 1, 2, 3, \cdots$

$$\left[\begin{aligned} &\text{sample } i_k \text{ uniformly from } \{1, \cdots, m\} \\ &x_{k+1} = \mathrm{prox}_{\gamma_k R}\big(x_k - \gamma_k \nabla f_{i_k}(x_k)\big). \end{aligned}\right. \quad (3)$$

Different from FBS, Prox-SGD only evaluates the gradient of one sampled function f_{i_k} at each step. However, to ensure the convergence, the step-size γ_k of Prox-SGD has to converge to 0 at a proper speed (*e.g.* $\gamma_k = k^s$ for $s \in]1/2, 1]$), leading to only $O(1/\sqrt{k})$ convergence rate for $\Phi(x_k) - \Phi(x^\star)$. When Φ is strongly convex, the rate for the objective can only be improved to $O(1/k)$.

Perturbed Forward–Backward Splitting An alternative perspective of treating Prox-SGD is as a perturbation of the deterministic FBS scheme. More precisely, iteration (3) can be written as the inexact FBS iteration with stochastic approximation error for the gradient, for $k = 0, 1, 2, 3, \cdots$

$$\left[\begin{aligned} &\text{sample } \varepsilon_k \text{ from a finite distribution } \mathcal{D}_k, \\ &x_{k+1} = \mathrm{prox}_{\gamma_k R}\big(x_k - \gamma_k(\nabla F(x_k) + \varepsilon_k)\big). \end{aligned}\right. \quad (4)$$

For most existing stochastic gradient methods, $\mathbb{E}[\varepsilon_k] = 0$ and $\|\varepsilon_k\|^2$ is the variance of the stochastic gradient.

For Prox-SGD, the error ε_k takes the form

$$\varepsilon_k^{\mathrm{SGD}} \overset{\text{def}}{=} \nabla f_{i_k}(x_k) - \nabla F(x_k). \quad (5)$$

In general, the variance $\|\varepsilon_k^{\mathrm{SGD}}\|^2$ is only bounded, and does not vanish to 0 as $x_k \to x^\star$. One consequence of this non-vanishing variance is that, unlike FBS, in general manifold identification cannot occur. In Section 1 of the supplementary material, we give an example to illustrate this point.

1.4. Variance Reduced Stochastic Gradient Methods

To overcome the vanishing step-size and slow convergence speed of Prox-SGD caused by non-vanishing variance, several variance reduced schemes are proposed schemes (Defazio et al., 2014; Xiao & Zhang, 2014). The features of variance reduced techniques can be summarised as:

- Same as Prox-SGD, in expectation, the stochastic gradient remains an unbiased estimation of the full gradient;
- Different from Prox-SGD, the variance $\|\varepsilon_k\|^2$ converges to 0 when x_k approaches the solution $x^\star$.

SAGA Algorithm (Defazio et al., 2014) The key idea of SAGA for reducing the variance is utilising the gradient history for the evaluation of the current gradient.

Given an initial point x_0, define the individual gradient $g_{0,i} \overset{\text{def}}{=} \nabla f_i(x_0), i = 1, \cdots, m$. Then, for $k = 0, 1, 2, 3, \cdots$

$$\left[\begin{aligned} &\text{sample } i_k \text{ uniformly from } \{1, \cdots, m\}, \\ &w_k = x_k - \gamma_k\big(\nabla f_{i_k}(x_k) - g_{k,i_k} + \tfrac{1}{m}\textstyle\sum_{i=1}^m g_{k,i}\big), \\ &x_{k+1} = \mathrm{prox}_{\gamma_k R}(w_k), \\ &g_{k,i} = \begin{cases} \nabla f_i(x_k) & \text{if } i = i_k, \\ g_{k-1,i} & \text{o.w.} \end{cases} \end{aligned}\right. \quad (6)$$

In the context of (4), the stochastic approximation error ε_k of SAGA takes the form

$$\varepsilon_k^{\mathrm{SAGA}} \overset{\text{def}}{=} \nabla f_{i_k}(x_k) - g_{k,i_k} + \frac{1}{m}\sum\nolimits_{i=1}^m g_{k,i} - \nabla F(x_k). \quad (7)$$

Prox-SVRG Algorithm (Xiao & Zhang, 2014) Compared to SAGA, in stead of using the gradient history, Prox-

[1] Linear convergence is also known as geometric or exponential convergence.

SVRG computes the full gradient of a given point, and uses it for a certain number of iterations.

Let P be a positive integer, for $\ell = 0, 1, 2, \cdots$

$$\left\lfloor \begin{array}{l} \tilde{g}_\ell = \frac{1}{m}\sum_{i=1}^m \nabla f_i(\tilde{x}_\ell), x_{\ell,0} = \tilde{x}_\ell, \\ \text{For } p = 1, \cdots, P \\ \left\lfloor \begin{array}{l} \text{sample } i_p \text{ uniformly from } \{1, \cdots, m\}, \\ w_k = x_{\ell,p-1} - \gamma_k(\nabla f_{i_p}(x_{\ell,p-1}) - \nabla f_{i_p}(\tilde{x}_\ell) + \tilde{g}_\ell), \\ x_{\ell,p} = \mathrm{prox}_{\gamma_k R}(w_k). \end{array} \right. \\ \text{Option I} : \tilde{x}_{\ell+1} = x_{\ell,P}, \\ \text{Option II} : \tilde{x}_{\ell+1} = \frac{1}{P}\sum_{p=1}^P x_{\ell,p}. \end{array} \right. \tag{8}$$

In the context of (4), given $x_{\ell,p}$, denote $k = \ell P + p$, then we have $x_{\ell,p} = x_k$ and the stochastic approximation error ε_k of Prox-SVRG reads

$$\varepsilon_k^{\mathrm{SVRG}} \overset{\mathrm{def}}{=} \nabla f_{i_p}(x_k) - \nabla f_{i_p}(\tilde{x}_\ell) + \tilde{g}_\ell - \nabla F(x_k). \tag{9}$$

1.5. Contributions

In this paper, we present a very general framework for analysing the local convergence properties of variance reduced stochastic gradient. More precisely, this paper consists of the following contributions.

Convergence of Sequence for SAGA/Prox-SVRG Assuming only convexity, we prove the almost sure global convergence of the sequences generated by SAGA (Theorem 2.1) and Prox-SVRG with "Option I" (Theorem 2.2), which is new to the literature. Moreover, for Prox-SVRG with "Option I", an $O(1/k)$ ergodic convergence rate for the objective function is proved.

Finite Time Manifold Identification Let $x^\star$ be a global minimiser of problem ($\mathcal{P}$), and suppose that the sequence $\{x_k\}_{k\in\mathbb{N}}$ generated by the perturbed FBS iteration (4) converges to $x^\star$ almost surely. Then under the additional assumptions that the non-smooth function R is partly smooth at $x^\star$ relative to a C^2-smooth manifold $\mathcal{M}_{x^\star}$ (Definition 3.1) and a non-degeneracy condition (Eq. (ND)) holds at $x^\star$, in Theorem 3.2 we prove a general finite time manifold identification result for the perturbed FBS scheme (4). The manifold identification means that after a finite number of iterations, say K, there holds $x_k \in \mathcal{M}_{x^\star}$ for all $k \geq K$. Specialising the result to SAGA and Prox-SVRG algorithms, we prove the finite manifold identification of these two algorithms (Corollary 3.4).

Local Linear Convergence Building upon the manifold identification result, if moreover F is locally C^2-smooth along $\mathcal{M}_{x^\star}$ near $x^\star$ and a restricted injectivity condition (Eq. (RI)) is satisfied by the Hessian $\nabla^2 F(x^\star)$, we show that locally SAGA and Prox-SVRG converge linearly.

Local Accelerations Another important implication of manifold identification is that the global non-smooth optimisation problem becomes locally C^2-smooth along the manifold $\mathcal{M}_{x^\star}$, and moreover is locally strongly convex if the restricted injectivity condition (RI) is satisfied. This implies that locally we have many choices of acceleration to choose, for instance we can turn to higher-order optimisation methods, such as (quasi)-Newton methods or Riemannian manifold based optimisation methods which can lead to super linear convergence.

Lastly, for the numerical experiments considered in this paper, the corresponding MATLAB source code to reproduce the results is available online[2].

1.6. Notations

Throughout the paper, $\mathbb{N}$ denotes the set of non-negative integers and $k \in \mathbb{N}$ denotes the index. $\mathbb{R}^n$ is the Euclidean space of dimension n. For a nonempty convex set $\Omega \subset \mathbb{R}^n$, $\mathrm{ri}(\Omega)$ and $\mathrm{rbd}(\Omega)$ denote its relative interior and relative boundary respectively.

Let R be proper, convex and lower semi-continuous, the sub-differential is defined by $\partial R(x) \overset{\mathrm{def}}{=} \big\{g \in \mathbb{R}^n | R(y) \geq R(x) + \langle g,\, y - x\rangle,\, \forall y \in \mathbb{R}^n\big\}$. A function R is α-strongly convex for some $\alpha > 0$ if $R(x) - \frac{\alpha}{2}\|x\|^2$ is convex.

2. Global Convergence of SAGA/Prox-SVRG

In this section, we prove the convergence of the sequence generated by the SAGA and Prox-SVRG with "Option I" without strong convexity, while in their respective original work (Defazio et al., 2014; Xiao & Zhang, 2014), only the convergence of the objective function value is provided.

We present first the convergence result of the SAGA algorithm, recall that L is the uniform Lipschitz continuity of all element functions $f_i, i = 1, \cdots, m$.

Theorem 2.1 (Convergence of SAGA). *For problem ($\mathcal{P}$), suppose that conditions (**A.1**)-(**A.3**) hold. Let $\{x_k\}_{k\in\mathbb{N}}$ be the sequence generated by the SAGA algorithm (6) with $\gamma_k \equiv \frac{1}{3L}$, then there exists an $x^\star \in \mathrm{Argmin}(\Phi)$ such that almost surely $\Phi(x_k) \to \Phi(x^\star)$, $x_k \to x^\star$ and $\varepsilon_k^{\mathrm{SAGA}} \to 0$.*

Next we provide the convergence result of the Prox-SVRG algorithm with "Option I". Given $\ell \in \mathbb{N}$ and $p \in \{1, \cdots, P\}$, let $k = \ell P + p$, and denote $x_{\ell,p} = x_k$. For sequence $\{x_k\}_{k\in\mathbb{N}}$, define a new point by $\bar{x}_k \overset{\mathrm{def}}{=} \frac{1}{k}\sum_{\ell=1}^k x_\ell$.

Theorem 2.2 (Convergence of Prox-SVRG). *For problem ($\mathcal{P}$), suppose that conditions (**A.1**)-(**A.3**) hold. Let $\{x_k\}_{k\in\mathbb{N}}$ be the sequence generated by the Prox-SVRG algorithm (8) with "Option I". Then,*

[2] https://github.com/jliang993/Local-SAGA-ProxSVRG

(i) *If we fix* $\gamma_k \equiv \gamma$ *with* $\gamma \le \frac{1}{4L(P+2)}$, *then there exists a minimiser* $x^\star \in \mathrm{Argmin}(\Phi)$ *such that* $x_k \to x^\star$ *and* $\varepsilon_k^{\mathrm{SVRG}} \to 0$ *almost surely. Moreover, there holds for each* $k = \ell P$, $\mathbb{E}[\Phi(\bar{x}_k) - \Phi(x^\star)] = O(1/k)$*;*

(ii) *Suppose that* R, F *are moreover* α_R *and* α_F *strongly convex respectively, then if* $4L\gamma(P+1) < 1$, *there holds* $\mathbb{E}\big[\|\tilde{x}_\ell - x^\star\|^2\big] \le O(\rho_{\mathrm{SVRG}}^\ell)$, *where* $\rho_{\mathrm{SVRG}} = \max\{\frac{1-\gamma\alpha_F}{1+\gamma\alpha_R}, 4L\gamma(P+1)\}$.

Remark 2.3. To the best of our knowledge, the $O(1/k)$ ergodic convergence rate of $\{\mathbb{E}[\Phi(\bar{x}_k) - \Phi(x^\star)]\}_{k\in\mathbb{N}}$ is new to the literature, which also holds for SVRG (Johnson & Zhang, 2013).

3. Finite Time Manifold Identification

From this section, we move to the local convergence properties of the SAGA/Prox-SVRG algorithms.

3.1. Partial Smoothness

Let $\mathcal{M}_x$ be a C^2-smooth Riemannian manifold of $\mathbb{R}^n$ around a point x. Denote $\mathcal{T}_{\mathcal{M}_x}(x)$ the tangent space of $\mathcal{M}_x$ at a point x. Given a set S, $\mathrm{par}(S) \overset{\text{def}}{=} \mathbb{R}(S - S)$ is the subspace parallel to the affine hull of S, and $(\cdot)^\perp$ the orthogonal complement.

Definition 3.1 (Partial smoothness). Let $R : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be proper convex and lower semi-continuous. Then R is said to be partly smooth at x relative to a set $\mathcal{M}_x$ containing x if $\partial R(x) \neq \emptyset$, and moreover

Smoothness: $\mathcal{M}_x$ is a C^2-manifold around x, R restricted to $\mathcal{M}_x$ is C^2 around x.

Sharpness: the tangent space $\mathcal{T}_{\mathcal{M}_x}(x)$ coincides with $T_x \overset{\text{def}}{=} (\mathrm{par}(\partial R(x)))^\perp$.

Continuity: the set-valued mapping ∂R is continuous at x relative to $\mathcal{M}_x$.

The class of partly smooth functions at x relative to $\mathcal{M}_x$ is denoted as $\mathrm{PSF}_x(\mathcal{M}_x)$. Many popular non-smooth penalty functions are partly smooth, such as sparsity promoting ℓ_1-norm, group sparsity promoting $\ell_{1,2}$-norm, low rank promoting nuclear norm, etc; see Table 1 for more details. We refer to (Liang et al., 2017) and references therein for detailed properties of these partly smooth functions.

3.2. An Abstract Finite Time Manifold Identification

Recall the perturbed FBS iteration (4). We have the following abstract manifold identification.

Theorem 3.2 (Abstract manifold identification). *For problem* ($\mathcal{P}$)*, suppose that conditions* (**A.1**)-(**A.3**) *hold. For the perturbed FBS iteration* (4)*, suppose that:*

(**B.1**) *There exists* $\underline{\gamma} > 0$ *such that* $\liminf_{k\to+\infty} \gamma_k \ge \underline{\gamma}$*;*

(**B.2**) *The error* $\{\varepsilon_k\}_{k\in\mathbb{N}}$ *converges to* 0 *almost surely;*

(**B.3**) *There exists an* $x^\star \in \mathrm{Argmin}(\Phi)$ *such that* $\{x_k\}_{k\in\mathbb{N}}$ *converges to* $x^\star$ *almost surely.*

For the $x^\star$ *in* (**B.3**)*, suppose that* $R \in \mathrm{PSF}_{x^\star}(\mathcal{M}_{x^\star})$*, and the following non-degeneracy condition holds*

$$-\nabla F(x^\star) \in \mathrm{ri}\big(\partial R(x^\star)\big). \tag{ND}$$

Then, there exists a $K > 0$ *such that for all* $k \ge K$*, we have* $x_k \in \mathcal{M}_{x^\star}$ *almost surely.*

Remark 3.3. In the deterministic setting, the finite manifold identification of (4), *i.e.* ε_k is deterministic approximation error, is discussed in Section 3.3 of (Liang et al., 2017).

3.3. Manifold Identification of SAGA/Prox-SVRG

Specialising Theorem 3.2 to the case of SAGA/Prox-SVRG algorithms, we obtain the corollary below. For Prox-SVRG, recall that $x_{\ell,p}$ is denoted as x_k with $k = \ell P + p$.

Corollary 3.4. *For problem* ($\mathcal{P}$)*, suppose that conditions* (**A.1**)-(**A.3**) *hold. Suppose that*

- *SAGA is applied under the conditions of Theorem 2.1;*
- *Prox-SVRG is ran under the conditions of Theorem 2.2.*

Then there exists an $x^\star \in \mathrm{Argmin}(\Phi)$ *such that the sequence* $\{x_k\}_{k\in\mathbb{N}}$ *generated by either algorithm converges to* $x^\star$ *almost surely.*

If moreover, $R \in \mathrm{PSF}_{x^\star}(\mathcal{M}_{x^\star})$*, and the non-degeneracy condition* (ND) *holds. Then, there exists a* $K > 0$ *such that for all* $k \ge K$*,* $x_k \in \mathcal{M}_{x^\star}$ *almost surely.*

Remark 3.5. In (Lee & Wright, 2012; Duchi & Ruan, 2016), manifold identification properties of the *regularised dual averaging algorithm* (RDA) (Xiao, 2010) were reported. The main difference between the present work and those of (Lee & Wright, 2012; Duchi & Ruan, 2016) is that, RDA is proposed for solving ($\mathcal{P}$) with infinite sum problem, while in this work we mainly focus on the finite sum case. We remark also that although RDA has manifold identification properties, the method does not exhibit linear convergence under strong convexity, and hence, in contrast to variance reduction methods, there can be no local acceleration in the convergence rate upon manifold identification.

4. Local Linear Convergence

Now we turn to the local linear convergence properties of SAGA/Prox-SVRG algorithms. Throughout the section, $x^\star \in \mathrm{Argmin}(\Phi)$ denotes a global minimiser ($\mathcal{P}$), $\mathcal{M}_{x^\star}$ is a C^2-smooth manifold which contains $x^\star$, and $T_{x^\star}$ denotes the tangent space of $\mathcal{M}_{x^\star}$ at $x^\star$.

Table 1: Examples of partly smooth functions. For $x \in \mathbb{R}^n$ and some subset of indices $\mathcal{b} \subset \{1, \dots, n\}$, $x_{\mathcal{b}}$ is the restriction of x to the entries indexed in $\mathcal{b}$. D_{DIF} stands for the finite differences operator, $\mathrm{rank}(z)$ denotes the rank of a matrix.

FUNCTION	EXPRESSION	PARTIAL SMOOTH MANIFOLD
ℓ_1-NORM	$\|x\|_1 = \sum_{i=1}^{n} \lvert x_i \rvert$	$\mathcal{M}_x = T_x = \big\{z \in \mathbb{R}^n : \mathcal{I}_z \subseteq \mathcal{I}_x\big\}, \mathcal{I}_x = \big\{i : x_i \neq 0\big\}$
$\ell_{1,2}$-NORM	$\sum_{i=1}^{m} \|x_{\mathcal{b}_i}\|$	$\mathcal{M}_x = T_x = \big\{z \in \mathbb{R}^n : \mathcal{I}_z \subseteq \mathcal{I}_x\big\}, \mathcal{I}_x = \big\{i : x_{\mathcal{b}_i} \neq 0\big\}$
ℓ_∞-NORM	$\max_{i=\{1,\dots,n\}} \lvert x_i \rvert$	$\mathcal{M}_x = T_x = \big\{z \in \mathbb{R}^n : z_{\mathcal{I}_x} \in \mathbb{R}\mathrm{sign}(x_{\mathcal{I}_x})\big\}, \mathcal{I}_x = \big\{i : \lvert x_i \rvert = \|x\|_\infty\big\}$
TV SEMI-NORM	$\|x\|_{\mathrm{TV}} = \|D_{\mathrm{DIF}} x\|_1$	$\mathcal{M}_x = T_x = \big\{z \in \mathbb{R}^n : \mathcal{I}_{D_{\mathrm{DIF}} z} \subseteq \mathcal{I}_{D_{\mathrm{DIF}} x}\big\}, \mathcal{I}_{D_{\mathrm{DIF}} x} = \{i : (D_{\mathrm{DIF}} x)_i \neq 0\}$
NUCLEAR NORM	$\|x\|_* = \sum_{i=1}^{r} \sigma(x)$	$\mathcal{M}_x = \big\{z \in \mathbb{R}^{n_1 \times n_2} : \mathrm{rank}(z) = \mathrm{rank}(x) = r\big\}, \sigma(x)$ SINGULAR VALUES OF x

4.1. Local Linear Convergence

Similar to (Liang et al., 2017) for the deterministic FBS-type methods, the key assumption to establish local linear convergence for SAGA/Prox-SVRG is a so-called restricted injectivity condition, see below.

Restricted Injectivity Let F be locally C^2-smooth around the minimiser $x^\star$, and moreover the following restricted injectivity condition holds

$$\ker\big(\nabla^2 F(x^\star)\big) \cap T_{x^\star} = \{0\}. \tag{RI}$$

Owing to Proposition 12 of (Liang et al., 2017), it can be shown that under condition (RI), $x^\star$ actually is the unique minimiser of problem ($\mathcal{P}$), and Φ grows locally quadratic if moreover $R \in \mathrm{PSF}_{x^\star}(\mathcal{M}_{x^\star})$.

Lemma 4.1 (Local quadratic growth (Liang et al., 2017)). *For problem ($\mathcal{P}$), suppose that assumptions (**A.1**)-(**A.3**) hold. Let $x^\star \in \mathrm{Argmin}(\Phi)$ be a global minimiser such that conditions (ND) and (RI) are fulfilled and $R \in \mathrm{PSF}_{x^\star}(\mathcal{M}_{x^\star})$, then $x^\star$ is the unique minimiser of ($\mathcal{P}$) and there exist $\alpha > 0$ and $r > 0$ such that*

$$\Phi(x) - \Phi(x^\star) \geq \alpha\|x - x^\star\|^2 : \forall x \text{ s.t. } \|x - x^\star\| \leq r.$$

Remark 4.2. A similar result can be found in Theorem 5 of (Lee & Wright, 2012).

Lemma 4.1 implies that when a sequence convergent stochastic method is applied to solve ($\mathcal{P}$), eventually the generated sequence $\{x_k\}_{k\in\mathbb{N}}$ will enter a local neighbourhood of the solution $x^\star \in \mathrm{Argmin}(\Phi)$ where the function has the quadratic growth property. If moreover this method is linearly convergent under strong convexity, then locally it will also converge linearly under quadratic growth. As a consequence, we have the following propositions.

Proposition 4.3 (Local linear convergence of SAGA). *For problem ($\mathcal{P}$), suppose that conditions (**A.1**)-(**A.3**) hold, and the SAGA algorithm (6) is applied with $\gamma_k \equiv \frac{1}{3L}$. Then x_k converges to $x^\star \in \mathrm{Argmin}(\Phi)$ almost surely. If moreover, $R \in \mathrm{PSF}_{x^\star}(\mathcal{M}_{x^\star})$, and conditions (ND)-(RI) are satisfied, then there exists $K > 0$ such that for all $k \geq K$,*

$$\mathbb{E}\big[\|x_k - x^\star\|^2\big] = O(\rho_{\mathrm{SAGA}}^{k-K}),$$

where $\rho_{\mathrm{SAGA}} = 1 - \min\{\frac{1}{4m}, \frac{\alpha}{3L}\}$.

The claim is a direct consequence of Theorem 1 of (Defazio et al., 2014).

Proposition 4.4 (Local linear convergence of Prox-SVRG). *For problem ($\mathcal{P}$), suppose that conditions (**A.1**)-(**A.3**) hold, and the Prox-SVRG algorithm (8) is applied such that Theorem 2.2 holds. Then x_k converges to $x^\star \in \mathrm{Argmin}(\Phi)$ almost surely. If moreover, $R \in \mathrm{PSF}_{x^\star}(\mathcal{M}_{x^\star})$, and conditions (ND)-(RI) are satisfied, then there exists $K > 0$ such that for all $k \geq K$,*

$$\mathbb{E}\big[\|\tilde{x}_\ell - x^\star\|^2\big] = O(\rho_{\mathrm{SVRG}}^{\ell-K}),$$

where $\rho_{\mathrm{SVRG}} = \max\{\frac{1-\gamma\alpha_F}{1+\gamma\alpha_R}, 4L\gamma(P+1)\}$ and γ, P are chosen such that $\rho_{\mathrm{SVRG}} < 1$.

The claim is a direct consequence of Theorem 2.2(ii).

4.2. Better Linear Rate Estimation?

In this part, we discuss briefly the obtained linear rate estimations of SAGA/Prox-SVRG (*i.e.* $\rho_{\mathrm{SAGA}}, \rho_{\mathrm{SVRG}}$), in comparison to the one obtained in (Liang et al., 2017) for deterministic FBS.

For the deterministic FBS algorithm, when R is locally polyhedral around $x^\star$, Theorem 21 of (Liang et al., 2017) implies that the local convergence rate of FBS is

$$\rho_{\mathrm{FBS}} = 1 - \gamma\alpha.$$

While for $\rho_{\mathrm{SAGA}}, \rho_{\mathrm{SVRG}}$, their rate estimations both depend on the number of functions m, which means that tightness[3] of $\rho_{\mathrm{SAGA}}, \rho_{\mathrm{SVRG}}$ could be much worse than ρ_{FBS}.

This naturally leads to the question of whether the rate estimations for SAGA and Prox-SVRG can be improved. Numerically, the rate estimation seems to be problem dependent, that is for some problems ρ_{FBS} can be achieved; see Example 6.1 and Figure 1. While for some problems, the practical observation of SAGA/SVRG is much slower than ρ_{FBS}; see Section 4 of the supplementary material for details. Note that we are comparing the rate in terms of *per iteration*, not gradient evaluation complexity.

[3] Tightness means how close is the rate estimation to the practical observation of the algorithms.

5. Beyond Local Convergence Analysis

As already discussed, manifold identification implies that, the globally non-smooth problem $\min_{x\in\mathbb{R}^n}\Phi(x)$ locally becomes a C^2-smooth and possibly non-convex (*e.g.* nuclear norm) problem constrained on the identified manifold $\min_{x\in\mathcal{M}_{x^\star}}\Phi(x)$. Such a transition to local C^2-smoothness, provides various choices of acceleration.

5.1. Better Local Lipschitz Continuity

If the dimension of the manifold $\mathcal{M}_{x^\star}$ is much smaller than that of the whole space $\mathbb{R}^n$, then constrained to $\mathcal{M}_{x^\star}$, the Lipschitz property of the smooth part would become much better. For each $i\in\{1,\cdots,m\}$, denote by $L_{\mathcal{M}_{x^\star},i}$ the Lipschitz constant of ∇f_i along the manifold $\mathcal{M}_{x^\star}$, and let

$$L_{\mathcal{M}_{x^\star}} \overset{\text{def}}{=} \max_{i=1,\cdots,m} L_{\mathcal{M}_{x^\star},i}.$$

In general, locally around $x^\star$, we have $L_{\mathcal{M}_{x^\star}} \leq L$.

For SAGA/Prox-SVRG, and other stochastic methods which have manifold identification property, once the manifold is identified, one can adapt their step-sizes to the local Lipschitz constants. Since step-size is crucial to the convergence speed of these methods, the potential acceleration of such a local adaptive strategy can be significant. See Section 6.1 for numerical example, and the supplementary material on how to compute or approximate $L_{\mathcal{M}_{x^\star}}$.

5.2. Lower Computational Complexity

Another important aspect of the manifold identification property is that one can naturally reduce the computational cost, especially when $\mathcal{M}_{x^\star}$ is of very low dimension.

Take R as the ℓ_1-norm for example. Suppose that the solution $x^\star$ of Φ is κ-sparse, *i.e.* the number of non-zero entries of $x^\star$ is κ. We have two stages of gradient evaluation complexity for $\nabla f_i(x_k)$:

Before identification: $O(n)$,

After identification: $O(\kappa)$.

Now let R be the nuclear norm, and suppose F is quadratic. Let the solution$x^\star$ be of rank κ. We have two stages of gradient evaluation complexity for $\nabla f_i(x_k)$ (after identification, x_k is stored in terms of its SVD decomposition):

Before identification: $O(n^2)$,

After identification: $O(\kappa n)$.

For both cases, the reduction of computational cost depends on the ratio of n/κ.

5.3. Higher-order Acceleration

The last acceleration strategy is the Riemannian manifold based higher-order acceleration. Recently, various Riemannian manifold based optimisation methods have been proposed in the literature (Kressner et al., 2014; Ring & Wirth, 2012; Vandereycken, 2013; Boumal et al., 2014), particularly for low-rank matrix recovery. However, an obvious drawback of this class of methods is that the manifold should be known *a priori*, which limits the their applications.

The manifold identification of proximal methods implies that one can first use the proximal method to identify the correct manifold, and then turn to the manifold based optimisation methods. The higher-order methods that can be applied include Newton-type methods, when the restricted injectivity condition (RI) is satisfied, and Riemannian geometry based optimisation methods (Lemaréchal et al., 2000; Miller & Malick, 2005; Smith, 1994; Boumal et al., 2014; Vandereycken, 2013), for instance the non-linear conjugate gradient method (Smith, 1994). Stochastic Riemannian manifold based optimisation methods are also studied in the literature, for instance in (Zhang et al., 2016), the authors generalise SVRG to the manifold setting.

6. Numerical Experiments

We now consider several examples to verify the established results. Three examples for R are considered, sparsity promoting ℓ_1-norm, group sparsity promoting $\ell_{1,2}$-norm and low rank promoting nuclear norm.

As the main focus of this work is the theoretical properties of SAGA and Prox-SVRG algorithms, the scale of the problems considered are not very large. In the supplementary material, experiments on large scale real data are presented.

6.1. Local Linear Convergence

We consider the sparse logistic regression problem to demonstrate the manifold identification and local linear convergence of SAGA/Prox-SVRG algorithms. Moreover in this experiment, we provide only the rate estimation from the FBS scheme, which is $\rho_{\text{FBS}} = 1-\gamma\alpha$.

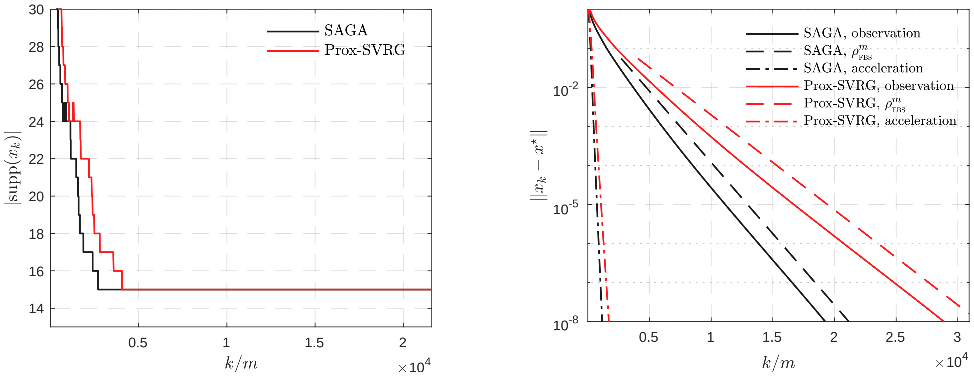
Example 6.1 (Sparse logistic regression). Let $m>0$ and $(z_i,y_i)\in\mathbb{R}^n\times\{\pm1\}$, $i=1,\cdots,m$ be the training set. The sparse logistic regression is to find a linear decision function which minimises the objective

$$\min_{(x,b)\in\mathbb{R}^n\times\mathbb{R}} \mu\|x\|_1 + \tfrac{1}{m}\textstyle\sum_{i=1}^m \log\big(1+e^{-y_i f(z_i;x,b)}\big),$$

where $f(z;x,b) = b + z^T x$.

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:

$$\text{SAGA}:\gamma=\tfrac{1}{2L};\quad \text{Prox-SVRG}:\gamma=\tfrac{1}{3L},\ \ P=m.$$



(a) $|\mathrm{supp}(x_k)|$

(b) $\|x_k - x^\star\|$

Figure 1: Finite manifold identification and local linear convergence of SAGA and Prox-SVRG for solving the sparse logistic regression problem in Example 6.1. (a) manifold identification; (b) local linear convergence. ρ_{FBS} is the rate estimation from FBS scheme, that is $\rho_{\mathrm{FBS}} = 1 - \gamma\alpha$, where γ is the step-size and α is from Lemma 4.1.

Remark 6.2. The reason of choosing different step-sizes for SAGA and Prox-SVRG is only to distinguish the red and black plots in Figure 1. As for the considered synthetic example, the performance of the two algorithms are almost the same under same step-size.

The observations of the experiments are shown in Figure 1. The observations of Prox-SVRG are for the inner loop sequence $x_{\ell,p}$, which is denoted as x_k by letting $k = \ell P + p$. The non-degeneracy condition (ND) and the restricted injectivity condition (RI) are checked *a posterior*, which are all satisfied for the tested example. The local quadratic growth parameter α and the local Lipschitz constant $L_{\mathcal{M}_{x^\star}}$ are

$$\alpha = 0.0156 \ \text{ and } \ L_{\mathcal{M}_{x^\star}} = 61.$$

Note that, locally the Lipschitz constant becomes about 19 times better.

Finite Manifold Identification In Figure 1(a), we plot the size of support of the sequence $\{x_k\}_{k\in\mathbb{N}}$ generated by the two algorithms. The lines are sub-sampled, one out of every m points.

The two algorithms are ran with the same initial point. It can be observed that SAGA shows slightly faster manifold identification than Prox-SVRG, this is due the fact that the step-size of SAGA (*i.e.* $\gamma = \frac{1}{2L}$) is larger than that of Prox-SVRG (*i.e.* $\gamma = \frac{1}{3L}$). As mentioned in Remark 6.2, the identification speed of the two algorithms will be rather similar if they are ran under the same choice of step-size.

Local Linear Convergence In Figure 1(b), we demonstrate the convergence rate of $\{\|x_k - x^\star\|\}_{k\in\mathbb{N}}$ of the two algorithms. The two *solid* lines are the practical observation of $\{\|x_k - x^\star\|\}_{k\in\mathbb{N}}$ generated by SAGA and Prox-SVRG, the two *dashed* lines are the theoretical estimations using ρ_{FBS}, and two *dot-dashed* lines are the practical observation of the acceleration of SAGA/Prox-SVRG based on the local Lipschitz continuity $L_{\mathcal{M}_{x^\star}}$. The lines are also sub-sampled, one out of every m points.

For the considered problem, given the values of α and γ above, we have that

$$\begin{aligned} \text{SAGA}: \rho_{\mathrm{FBS}} &= 0.999993, \ \rho^m_{\mathrm{FBS}} = 0.99916; \\ \text{Prox-SVRG}: \rho_{\mathrm{FBS}} &= 0.999995, \ \rho^m_{\mathrm{FBS}} = 0.99944. \end{aligned}$$

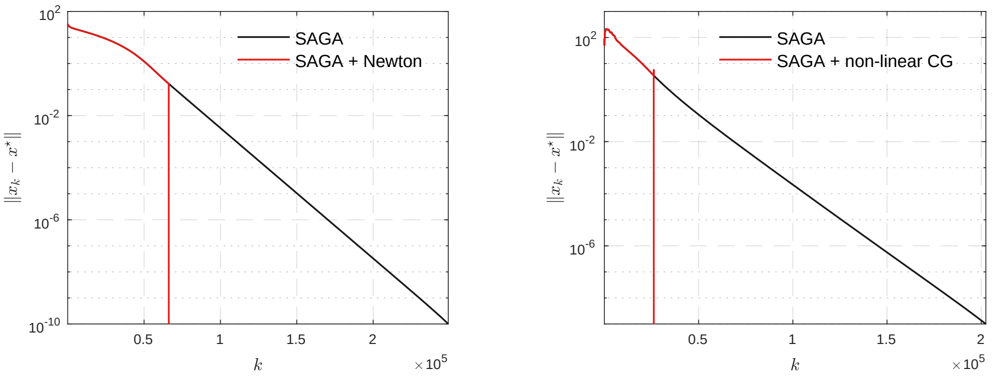
For the considered problem setting, the spectral radius quite matches the practical observations very well.

To conclude this part, we highlight the benefits of adapting to the local Lipschitz continuity of the problem. For both SAGA and Prox-SVRG, their adaptive schemes (*e.g. dot-dashed* lines) show 16 times faster performance compared to the non-adaptive ones (*e.g. solid* lines). Such an acceleration gain is on the same order of the difference between the global Lipschitz and local Lipschitz constants, which is 19 times. More importantly, the computational cost of evaluating the local Lipschitz constant is almost negligible, which makes the adaptive scheme more preferable in practice.

6.2. Local Higher-order Acceleration

We consider two problems of group sparse and low-rank regression to demonstrate local higher-order acceleration.

Example 6.3 (Group sparse and low-rank regression). Let $x_{\mathrm{ob}} \in \mathbb{R}^n$ be either a group sparse vector or a low-rank matrix (in a vectorised form), consider the following observation model $b = \mathcal{K}x_{\mathrm{ob}} + \omega$, where the entries of $\mathcal{K} \in \mathbb{R}^{m\times n}$ are sampled from an i.i.d. zero-mean and unit-variance Gaussian distribution, $\omega \in \mathbb{R}^m$ is an additive error with bounded ℓ_2-norm.



(a) Group sparsity

(b) Low rank

Figure 2: Local higher-order acceleration after manifold identification for Example 6.3. (a) Newton method is applied after the manifold is identified by SAGA; (b) non-linear conjugate gradient is applied after manifold identification. The black line is the observation of SAGA algorithm, and the red line is the observation of the "SAGA+higher-order" scheme. The black lines of the SAGA for both examples are not subsampled.

Let $\mu > 0$, and R be either the group sparsity promoting $\ell_{1,2}$-norm or the low rank promoting nuclear norm. Consider the problem to recover or approximate x_{ob},

$$\min_{x\in\mathbb{R}^n} \mu R(x) + \frac{1}{m}\sum\nolimits_{i=1}^{m} \frac{1}{2}\|\mathcal{K}_i x - b_i\|_2^2,$$

where $\mathcal{K}_i, b_i$ represent the i^{th} row and entry of $\mathcal{K}$ and b, respectively.

We have the following settings for the two examples of R:

$\ell_{1,2}$-norm: $(m, n) = (256, 512)$, x_{ob} has 8 non-zero blocks of block-size 4;

Nuclear norm: $(m, n) = (2048, 4096)$, $\mathrm{rank}(x_{\mathrm{ob}}) = 4$.

We consider only the SAGA algorithm for this test, as the main purpose is to highlight higher-order acceleration. For the $\ell_{1,2}$-norm, Newton method is applied after the manifold identification, while for nuclear norm, a non-linear conjugate gradient method (Boumal et al., 2014) is applied after manifold identification.

The numerical results are shown in Figure 2. For $\ell_{1,2}$-norm, the black line is the observation of the SAGA algorithm with $\gamma = \frac{1}{3L}$, the red line is the observation of the "SAGA+Newton" hybrid scheme. It should be noted that the lines are not subsampled.

For the hybrid scheme, SAGA is used for manifold identification, and Newton method is applied once the manifold is identified. As observed, the quadratically convergent Newton method converges in only a few steps. For nuclear norm, a non-linear conjugate gradient is applied when the manifold is identified. Similar to the observation of the $\ell_{1,2}$-norm, the super-linearly convergent non-linear conjugate gradient shows superior performance than SAGA.

7. Conclusion

In this paper, we proposed a unified framework of local convergence analysis for proximal variance reduced stochastic gradient methods, and especially focused on the SAGA and Prox-SVRG algorithms. Under partial smoothness, we established that these schemes identify the partial smooth manifold in finite time, and then converge locally linearly. Moreover, we proposed several practical acceleration approaches which can greatly improve the convergence speed of the algorithms.

Acknowledgements

The authors would like to thank F. Bach, J. Fadili and G. Peyré for helpful discussions.

References

Beck, A. and Teboulle, M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

Boumal, N., Mishra, B., Absil, P.-A., Sepulchre, R., et al. Manopt, a matlab toolbox for optimization on manifolds. *Journal of Machine Learning Research*, 15(1):1455–1459, 2014.

Combettes, P. L. and Wajs, V. R. Signal recovery by proximal Forward–Backward splitting. *Multiscale Modeling & Simulation*, 4(4):1168–1200, 2005.

Defazio, A., Bach, F., and Lacoste-Julien, S. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in*

Neural Information Processing Systems, pp. 1646–1654, 2014.

Duchi, J. and Ruan, F. Local asymptotics for some stochastic optimization problems: Optimality, constraint identification, and dual averaging. *arXiv preprint arXiv:1612.05612*, 2016.

Johnson, R. and Zhang, T. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in neural information processing systems*, pp. 315–323, 2013.

Kressner, D., Steinlechner, M., and Vandereycken, B. Low-rank tensor completion by riemannian optimization. *BIT Numerical Mathematics*, 54(2):447–468, 2014.

Lee, S. and Wright, S. J. Manifold identification in dual averaging for regularized stochastic online learning. *Journal of Machine Learning Research*, 13(Jun):1705–1744, 2012.

Lemaréchal, C., Oustry, F., and Sagastizábal, C. The U-Lagrangian of a convex function. *Trans. Amer. Math. Soc.*, 352(2):711–729, 2000.

Liang, J., Fadili, J., and Peyré, G. Local linear convergence of Forward–Backward under partial smoothness. In *Advances in Neural Information Processing Systems*, pp. 1970–1978, 2014.

Liang, J., Fadili, J., and Peyré, G. Convergence rates with inexact non-expansive operators. *Mathematical Programming*, 159(1):403–434, September 2016.

Liang, J., Fadili, J., and Peyré, G. Activity identification and local linear convergence of Forward–Backward-type methods. *SIAM Journal on Optimization*, 27(1):408–437, 2017.

Lions, P. L. and Mercier, B. Splitting algorithms for the sum of two nonlinear operators. *SIAM Journal on Numerical Analysis*, 16(6):964–979, 1979.

Miller, S. A. and Malick, J. Newton methods for nonsmooth convex minimization: connections among-Lagrangian, Riemannian Newton and SQP methods. *Mathematical programming*, 104(2-3):609–633, 2005.

Molinari, C., Liang, J., and Fadili, J. Convergence rates of Forward–Douglas–Rachford splitting method. *arXiv preprint arXiv:1801.01088*, 2018.

Nesterov, Y. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer, 2004.

Ring, W. and Wirth, B. Optimization methods on riemannian manifolds and their application to shape space. *SIAM Journal on Optimization*, 22(2):596–627, 2012.

Schmidt, M., Le Roux, N., and Bach, F. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1-2):83–112, 2017.

Smith, S. T. Optimization techniques on Riemannian manifolds. *Fields institute communications*, 3(3):113–135, 1994.

Vandereycken, B. Low-rank matrix completion by riemannian optimization. *SIAM Journal on Optimization*, 23(2): 1214–1236, 2013.

Xiao, L. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 11(Oct):2543–2596, 2010.

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

Zhang, H., Reddi, S. J., and Sra, S. Riemannian svrg: Fast stochastic optimization on riemannian manifolds. In *Advances in Neural Information Processing Systems*, pp. 4592–4600, 2016.