Local Convergence Properties of SAGA/Prox-SVRG and Acceleration

Clarice Poon*† Jingwei Liang*‡ Carola-Bibiane Schönlieb§

Abstract. Over the past ten years, driven by large scale optimisation problems arising from machine learning, the development of stochastic optimisation methods have witnessed a tremendous growth. However, despite their popularity, the theoretical understandings of these methods are quite limited in contrast to the deterministic optimisation methods. In this paper, we present a local convergence analysis for a typical type of stochastic optimisation methods: proximal variance reduced stochastic gradient methods, and mainly focus on the SAGA [12] and Prox-SVRG [43] algorithms. 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 the SAGA/Prox-SVRG algorithms: (i) the sequences generated by the SAGA/Prox-SVRG are able to identify the smooth manifold in a finite number of iterations; (ii) then the sequence enters a local linear convergence regime. Beyond local convergence analysis, we also 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. Concrete examples arising from machine learning are considered to verify the obtained results.

Key words. Forward–Backward, stochastic optimisation, variance reduced technique, SAGA, Prox-SVRG, partial smoothness, finite activity identification, local linear convergence, acceleration

AMS subject classifications. 90C15, 90C25, 65K05, 49M37

1 Introduction

1.1 Non-smooth optimisation

Modern optimisation has become a core part of many fields in science and engineering, such as machine learning, inverse problem and signal/image processing, to name a few. 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 regulariser. There have been several recent studies explaining how proximal gradient methods identify this low dimensional manifold and efficiently output solutions which take a certain structure; see for instance [29] 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 [38, 12, 43] 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 structured minimisation problem:

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

*Equal contributions.
†DAMTP, University of Cambridge, Cambridge, UK. E-mail: C.M.H.S.Poon@maths.cam.ac.uk.
‡DAMTP, University of Cambridge, Cambridge, UK. E-mail: jl993@cam.ac.uk.
§DAMTP, University of Cambridge, Cambridge, UK. E-mail: cbs31@cam.ac.uk.

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

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

is the average of a finite sum, where each f_i is smooth differentiable. We are interested in the problems where the value of m is very large. A classic example of ($\mathcal{P}$) is ℓ_1-norm regularised least square estimation (*i.e.* 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 the 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 $b \in \mathbb{R}^m$. More examples of problem ($\mathcal{P}$) can be found in Section 5.

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

In addition to assumption (**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 $L_F \leq \frac{1}{m}\sum_i L_i \leq L$ holds.

1.2 Deterministic Forward–Backward splitting method

A classical approach to solve ($\mathcal{P}$) is the Forward–Backward splitting (FBS) method [30], which is also known as the *proximal gradient descent method*. Given a current point x_k, the standard non-relaxed Forward–Backward iteration updates the next point x_{k+1} based on the following rule,

$$x_{k+1} = \operatorname{prox}_{\gamma_k R}\big(x_k - \gamma_k \nabla F(x_k)\big),\ \ \gamma_k \in]0, 2/L_F[, \tag{1.1}$$

where $\operatorname{prox}_{\gamma R}$ is the *proximity operator* of R which is defined as

$$\operatorname{prox}_{\gamma R}(\cdot) \overset{\text{def}}{=} \min_{x\in\mathbb{R}^n} \gamma R(x) + \frac{1}{2}\|x - \cdot\|^2. \tag{1.2}$$

Throughout this paper, unless otherwise stated, "Forward–Backward splitting" or "FBS" refers to the deterministic Forward–Backward splitting scheme (1.1).

Since the original work [30], the properties of Forward–Backward splitting have been extensively studied in the literature. In general, the advantages of this method can be summarised as following:

- Robust convergence guarantees. The convergence of the method is guaranteed as long as $0 < \underline{\gamma} \leq \gamma_k \leq \bar{\gamma} < 2/L_F$ holds for some $\underline{\gamma}, \bar{\gamma} > 0$ [9], for both the sequence $\{x_k\}_{k\in\mathbb{N}}$ and the objective function value $\{\Phi(x_k)\}_{k\in\mathbb{N}}$;
- Known convergence rates. It is well established that the sequence of FBS scheme converges at the rate of $\|x_k - x_{k-1}\| = o(1/\sqrt{k})$ [28], while the objective function converges at the rate of $\Phi(x_k) - \Phi(x^\star) = o(1/k)$ [33] where $x^\star \in \operatorname{Argmin}(\Phi)$ is a global minimiser. These rates can be improved to linear[1] if for instance strong convexity is assumed [35];
- Numerous acceleration techniques. Extensive acceleration schemes have been proposed over the decades, for instance the inertial schemes which contains inertial FBS [34, 31, 29], FISTA [3] and Nesterov's optimal methods [35];

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

- Structure adaptivity. There has been several recent work [27, 29] exploring the manifold identification properties of FBS, in particular, under the non-degeneracy condition that

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

where $\mathrm{ri}(\partial R(x^\star))$ denotes the *relative interior* of the subdifferential $\partial R(x^\star)$. It is shown in [29] that after a finite number of iterations, the FBS iterates x_k all lie on the same manifold as the optimal solution $x^\star$. In the case of $R = \|\cdot\|_1$, this equates to saying that there exists some $K \in \mathbb{N}$ such that x_k has the same sparse pattern as $x^\star$ for all $k \geq K$. Furthermore, upon identifying this optimal manifold, the FBS iterates can be proved to converge linearly to the optimal solution $x^\star$.

However, despite the above advantages of FBS, for the considered problem ($\mathcal{P}$), when the value of m is very large, the computational cost of $\nabla F(x_k)$ could be very expensive, which makes the deterministic FBS-type methods unsuitable for solving the large-scale problems arising from machine learning.

1.3 Proximal variance reduced stochastic gradient methods

The most straightforward extension of stochastic gradient descent to the "smooth + non-smooth" setting is the *proximal stochastic gradient descent* (Prox-SGD), which reads

$$\begin{aligned} &\text{For } k = 0, 1, 2, 3, \cdots \\ &\quad \text{sample } i_k \text{ uniformly from } \{1, \cdots, m\} \\ &\quad x_{k+1} = \mathrm{prox}_{\gamma_k R}\big(x_k - \gamma_k \nabla f_{i_k}(x_k)\big). \end{aligned} \tag{1.3}$$

The advantage of Prox-SGD over FBS scheme is that at each iteration, Prox-SGD only evaluates the gradient of one sampled function f_{i_k}, while FBS needs to compute m gradients. However, to ensure the convergence of Prox-SGD, 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 $\Phi(x_k) - \Phi(x^\star)$. Moreover, when Φ is strongly convex, the rate for the objective can only be improved to $O(1/k)$ which is much slower than the linear rate of FBS.

Prox-SGD has no manifold identification properties Besides slow convergence speed, another disadvantage of Prox-SGD, when compared to FBS, is that the sequence $\{x_k\}_{k\in\mathbb{N}}$ generated by the method is unable to identify the structure of the problem, *i.e.* no finite time manifold identification property.

To give an intuitive explanation as to why the iterates of Prox-SGD are inherently unstructured, we first provide an alternative perspective of treating Prox-SGD, the perturbation of deterministic Forward–Backward splitting method. More precisely, this method can be written as the inexact Forward–Backward splitting method with stochastic approximation error on the gradient,

$$\begin{aligned} &\text{For } k = 0, 1, 2, 3, \cdots \\ &\quad \text{sample } \varepsilon_k \text{ from a finite distribution } \mathcal{D}_k, \\ &\quad x_{k+1} = \mathrm{prox}_{\gamma_k R}\big(x_k - \gamma_k(\nabla F(x_k) + \varepsilon_k)\big). \end{aligned} \tag{1.4}$$

For most stochastic gradient methods, we have $\mathbb{E}[\varepsilon_k] = 0$ and $\|\varepsilon_k\|^2$ is the variance of the stochastic gradient. The stochastic approximation error ε_k for Prox-SGD takes the form

$$\varepsilon_k^{\mathrm{SGD}} \overset{\text{def}}{=} \nabla f_{i_k}(x_k) - \nabla F(x_k). \tag{1.5}$$

Manifold identification for FBS can be guaranteed under the non-degeneracy condition (ND). In fact, from the definition of proximity operator (1.2), at each iteration, we have

$$g_k \overset{\text{def}}{=} -\frac{x_{k+1} - x_k}{\gamma_k} - \nabla F(x_k) - \varepsilon_k \in \partial R(x_{k+1})$$

and manifold identification can be guaranteed if $g_k \to -\nabla F(x^\star)$ as $k \to \infty$. The issue in the case of Prox-SGD is that although we have that in expectation $\mathbb{E}[\nabla f_{i_k}(x_k)] = \nabla F(x_k)$, the error $\varepsilon_k^{\text{SGD}}$ is only bounded and in general does not converge to 0 even if $\{x_k\}_{k\in\mathbb{N}}$ converges to a global minimiser $x^\star \in \operatorname{Argmin}(\Phi)$.

We present a simple example to illustrate that Prox-SGD does not have manifold identification properties in general. Consider the following LASSO problem in 3D,

$$\min_{x\in\mathbb{R}^3} \tfrac{1}{3}\|x\|_1 + \tfrac{1}{3}\sum_{i=1}^3 \tfrac{1}{2}\|\mathcal{K}_i x - b_i\|^2,$$

where

$$\mathcal{K} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \sqrt{2} & 0 \\ 0 & 0 & \sqrt{3} \end{bmatrix} \quad \text{and} \quad b = \begin{pmatrix} 2 \\ \sqrt{2}/3 \\ \sqrt{3}/4 \end{pmatrix}.$$

The optimal solution of this particular problem is $x^\star = (1,0,0)^T$ and writing $F(x) \overset{\text{def}}{=} \frac{1}{6}\|\mathcal{K}x - b\|^2$, we have that the non-degeneracy condition

$$-\nabla F(x^\star) = \frac{1}{3}\begin{pmatrix} 1 \\ \sqrt{2}/3 \\ \sqrt{3}/4 \end{pmatrix} \in \operatorname{ri}\left(\tfrac{1}{3}\partial\|x^\star\|_1\right), \quad \text{where} \quad (\partial\|x^\star\|)_i = \operatorname{sign}(x_i) = \begin{cases} +1 & : x_i > 0, \\ [-1,+1] & : x_i = 0, \\ -1 & : x_i < 0, \end{cases}$$

It is furthermore straightforward to verify that $\|\nabla f_i(x) - \nabla F(x)\| \geq \|\nabla F(x)\|$ for all $i = 1,2,3$. Moreover, if Prox-SGD is starting with $x_0 = (\mu,0,0)^T$ with $\mu \in \mathbb{R}$, then with probability $2/3$ the first iterate of the algorithm satisfies $x_1 \notin \mathcal{M}_{x^\star} \overset{\text{def}}{=} \{(x,0,0) : \ x \in \mathbb{R}\}$. In fact, x_1 will have 2 non-zero entries if $|\mu| > \gamma_1$ and $i_1 \in \{2,3\}$. Figure 1 shows the support sizes of the Prox-SGD iterates over 10^6 iterations.

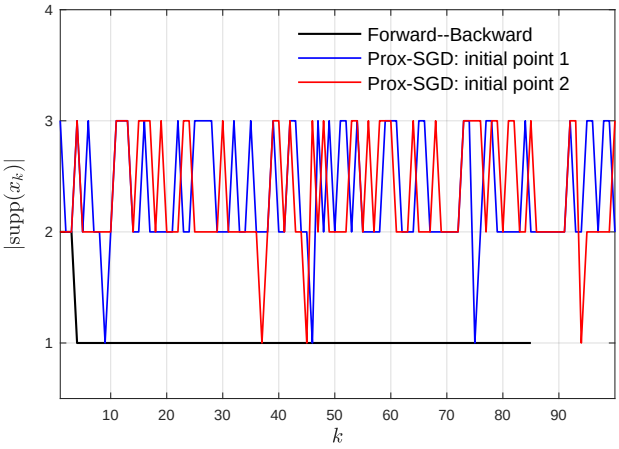


Figure 1: Support identification comparison between FBS and Prox-SGD. For Prox-SGD, "initial point 1" starts with an arbitrary point with all three elements non-zero; "initial point 2" starts with the point $10x^\star$. The maximum number of iteration for Prox-SGD is 10^6, the blue and red lines are sub-sampled, one out of every 10^4 points.

1.3.1 Variance reduced methods

To overcome the vanishing step-size and slow convergence speed of Prox-SGD, various (stochastic) incremental schemes are developed in literature; see for instance [4, 41, 21, 38, 12, 19, 43] and the references therein. Under stochastic setting, the variance reduced techniques are very popular approach, which have the following two main characteristics:

- 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 $x^\star$.

In the following, we introduce two well-known examples of variance reduced methods, the SAGA algorithm [12] and Prox-SVRG algorithm [43], which are the main targets of this paper.

SAGA algorithm [12] Similar to Prox-SGD algorithm, at each iteration k, the gradient of a sampled function $\nabla f_{i_k}(x_k)$ is computed by the SAGA algorithm where i_k is uniformly sampled from $\{1, \cdots, m\}$. In the meantime, let $\{\nabla f_{i_j}(x_{k-j})\}_{j=1,\cdots,m}$ be the gradients history over the past m steps, then the combination of these two aspects with additional debiasing yield the unbiased gradient approximation of the SAGA algorithm.

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

$$\begin{aligned}
&\text{For } k = 0, 1, 2, 3, \cdots \\
&\quad \text{sample } i_k \text{ uniformly from } \{1, \cdots, m\}, \\
&\quad x_{k+1} = \operatorname{prox}_{\gamma_k R}\big(x_k - \gamma_k(\nabla f_{i_k}(x_k) - g_{i_k,k} + \tfrac{1}{m}\textstyle\sum_{i=1}^m g_{i,k})\big), \\
&\quad \text{update the gradient history: } g_{k,i} = \begin{cases} \nabla f_i(x_k) & \text{if } i = i_k, \\ g_{k-1,i} & \text{o.w.} \end{cases}
\end{aligned} \tag{1.6}$$

SAGA successfully avoids the vanishing step-size of Prox-SGD, and has the same convergence rate as Forward–Backward splitting scheme. However, one distinctive drawback of SAGA is that, in general, its memory cost is proportional to the number of functions m.

In the context of (1.4), the stochastic approximation error ε_k of SAGA takes the form

$$\varepsilon_k^{\text{SAGA}} \overset{\text{def}}{=} \nabla f_{i_k}(x_k) - g_{k,i_k} + \frac{1}{m}\sum_{i=1}^m g_{k,i} - \nabla F(x_k). \tag{1.7}$$

Prox-SVRG algorithm The SVRG [19] (stochastic variance reduced gradient) method was originally proposed to solve ($\mathcal{P}$) with $R = 0$, later on in [43] it is extended to the case of R being non-trivial. Compared to SAGA, in stead of approximating the current gradient ∇F with the past m gradients ∇f_{i_k}, Prox-SVRG computes the full gradient of a given point along the iteration, and uses it for P iterations where P is on the order of m.

Let P be a positive integer. The iteration of the algorithm consists of two level of loops, for the sequence $\tilde{x}_\ell$ in the outer loop, full gradient $\nabla F(\tilde{x}_k)$ is computed. For the sequence $x_{\ell,p}$ in the inner loop, only the gradient of the sampled function is computed.

$$\begin{aligned}
&\text{For } \ell = 0, 1, 2, 3, \cdots \\
&\quad \tilde{g}_\ell = \tfrac{1}{m}\textstyle\sum_{i=1}^m \nabla f_i(\tilde{x}_\ell), x_{\ell,0} = \tilde{x}_\ell, \\
&\quad \text{For } p = 1, \cdots, P \\
&\quad\quad \text{sample } i_p \text{ uniformly from } \{1, \cdots, m\} \\
&\quad\quad x_{\ell,p} = \operatorname{prox}_{\gamma_k R}\big(x_{\ell,p-1} - \gamma_k(\nabla f_{i_p}(x_{\ell,p-1}) - \nabla f_{i_p}(\tilde{x}_\ell) + \tilde{g}_\ell)\big). \\
&\quad \text{Option I} : \tilde{x}_{\ell+1} = x_{\ell,P}, \\
&\quad \text{Option II} : \tilde{x}_{\ell+1} = \tfrac{1}{P}\textstyle\sum_{p=1}^P x_{\ell,p}.
\end{aligned} \tag{1.8}$$

Prox-SVRG can also afford non-vanishing step-size and has the same convergence rate as FBS scheme. It avoids the large memory cost of SAGA, however, the gradient evaluation complexity of Prox-SVRG is always higher than SAGA. For instance when $P = m$, the gradient evaluation of Prox-SVRG is three times that of SAGA.

In the context of (1.4), 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

$$\varepsilon_k^{\text{SVRG}} = \nabla f_{i_p}(x_k) - \nabla f_{i_p}(\tilde{x}_\ell) + \tilde{g}_\ell - \nabla F(x_k). \tag{1.9}$$

1.4 Contributions

In recent years, local linear convergence behaviors of the deterministic FBS-type methods have been studied under various scenarios. Particularly, in [29], based on the notion of partial smoothness (see Definition 3.1), the authors propose a unified framework for local linear convergence analysis of Forward–Backward splitting and its variants including inertial FBS and FISTA [3, 7].

In contrast to the deterministic setting, for the stochastic version of FBS scheme, very limited results of this nature have been reported in the literature. However, in practice local linear convergence of stochastic proximal gradient descent has been observed without global strong convexity. More importantly, the low dimensional property of partial smoothness naturally reduces the computational cost and provides rich possibilities of acceleration. As a consequence, the lack of uniform analysis framework and exploiting the local acceleration are the main motivations of this work.

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

Finite time manifold identification Let $x^\star \in \operatorname{Argmin}(\Phi)$ be a global minimiser of problem ($\mathcal{P}$), and suppose that the sequence $\{x_k\}_{k\in\mathbb{N}}$ generated by the perturbed Forward–Backward (1.4) 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 $\mathcal{M}_{x^\star}$ (see 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 Forward–Backward splitting scheme (1.4). 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 properties of them (see Corollary 3.4).

Local linear convergence for SAGA/Prox-SVRG 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 (see Eq. (RI)) is satisfied by the Hessian $\nabla^2 F(x^\star)$, we show that $x^\star$ is the unique minimiser of problem ($\mathcal{P}$) and moreover Φ has local quadratic grow property around $x^\star$. As a consequence, 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 Φ becomes C^2-smooth locally 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 (stochastic) Riemannian manifold based optimisation methods which can lead to super linear convergence speed.

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

1.4.1 Relation to previous work

Prior to our work, the identification properties of the *regularised dual averaging algorithm* (RDA) [42] were reported in [23, 14]. The RDA algorithm is also proposed for solving problem ($\mathcal{P}$), except that instead of being a finite sum, now the F takes the form

$$F(x) \overset{\text{def}}{=} \mathbb{E}_\xi\big[f(x;\xi)\big] = \int_\Omega f(x;\xi)\mathrm{d}\mathcal{D}(\xi),$$

where $\xi \in \mathbb{R}^m$ is a random vector whose probability distribution $\mathcal{D}$ is supported on the set $\Omega \subset \mathbb{R}^m$.

The RDA algorithm ([23, Algorithm 1]) for solving ($\mathcal{P}$) takes the following form, let $\xi_1 = 0$ and $g_0 = 0, \bar{g}_0 = 0$,

$$\begin{aligned}
&\text{For } k = 1, 2, 3, \cdots \\
&\quad \text{sample } \xi_k \text{ from the distribution } \mathcal{D}, \text{ and compute: } g_k = \nabla f(x_k;\xi_k); \\
&\quad \text{update the averaged gradient: } \bar{g}_k = \tfrac{k-1}{k}\bar{g}_{k-1} + \tfrac{1}{k} g_k; \\
&\quad \text{update new point: } x_{k+1} = \operatorname{prox}_{\frac{k}{\gamma_k}R}\Big(-\tfrac{k}{\gamma_k}\bar{g}_k\Big).
\end{aligned} \tag{1.10}$$

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

Though proposed for *infinite sum* problem, RDA can also applied to solve the *finite sum* problem, moreover the convergence properties establish in [23] remain hold. As a consequence, the identification property established there also holds true for the finite sum problem.

Compare the proposed work and those of [23, 14], there are several differences need to be pointed out:

- SAGA/Prox-SVRG and RDA are two very different types of methods. Although RDA can be applied to the finite sum problem, similarly to Prox-SGD, only $O(1/k)$ convergence rate can be achieved under strong convexity. While for the variance reduced SAGA/Prox-SVRG algorithms, linear convergence are available under strong convexity;
- For RDA algorithm, to the best of our knowledge, with only convexity assumption, so far there is no convergence result for the generated sequence $\{x_k\}_{k\in\mathbb{N}}$. While for SAGA/Prox-SVRG algorithms, in this paper we prove the convergence properties of their generated sequences under only convexity assumption, which are new to the literature.

1.5 Mathematical background

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 n dimension, and Id denotes the identity operator on $\mathbb{R}^n$. For a nonempty convex set $\Omega \subset \mathbb{R}^n$, $\mathrm{ri}(\Omega)$ and $\mathrm{rbd}(\Omega)$ denote its relative interior and boundary respectively, $\mathrm{aff}(\Omega)$ is its affine hull, and $\mathrm{par}(\Omega)$ is the subspace parallel to it. Denote P_Ω the orthogonal projector onto Ω.

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

Paper organisation The rest of the paper is organised as follows. In Section 2 we study the global convergence property of the sequence generated by SAGA and Prox-SVRG algorithms. The finite time manifold identification result is presented in Section 3. Local linear convergence and several local acceleration approaches are discussed in Section 4. We conclude the paper with various numerical experiments in Section 5. Several proofs of theorems are organised in Appendix A.

2 Global convergence of SAGA/Prox-SVRG

In literature, though the global almost sure convergence of the objective function value of SAGA/Prox-SVRG are well established [12, 43], the convergence properties of the generated sequences are not proved unless strong convexity is assumed. In this section, we prove the almost sure convergence of the sequence generated by SAGA and Prox-SVRG with "Option I" without strong convexity assumption. The proofs of the theorems are provided in Appendix A.1.

We present first the convergence 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 (1.6) with $\gamma_k \equiv \gamma = 1/(3L)$, then there exists an $x^\star \in \mathrm{Argmin}(\Phi)$ such that almost surely we have $\Phi(x_k) \to \Phi(x^\star)$, $x_k \to x^\star$ and $\varepsilon_k^{\mathrm{SAGA}} \to 0$.*

Next we provide the convergence result of the Prox-SVRG algorithm, and mainly focus on "Option I" for which convergence without strong convexity can be obtained. For "Option II", convergence of the sequence under strong convexity is discussed already in [43], hence we decide to skip the discussion here.

Given $\ell \in \mathbb{N}^+$ and $p \in \{1, \cdots, P\}$, denote $k = \ell P + p$, then $x_{\ell,p} = x_k$. For sequence $\{x_k\}_{k\in\mathbb{N}}$, define

$$\bar{x}_k \stackrel{\text{def}}{=} \frac{1}{k}\sum_{\ell=1}^{k} x_\ell.$$

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 (1.8) with "Option I". Then,*

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

$$\mathbb{E}\big[\Phi(\bar{x}_k) - \Phi(x^\star)\big] \leq \frac{1}{k\gamma^2}\big(\|\tilde{x}_0 - x^\star\|^2 + (2\gamma - \gamma^2)(\Phi(\tilde{x}_0) - \Phi(x^\star))\big). \tag{2.1}$$

(ii) *Suppose that R, F are moreover α_R and α_F strongly convex respectively, then if $4L\gamma(P+1) < 1$, there holds*

$$\mathbb{E}\big[\|\tilde{x}_\ell - x^\star\|^2\big] \leq \rho_{\mathrm{SVRG}}^{\ell}\Big(\|\tilde{x}_0 - x^\star\|^2 + \frac{2\gamma}{1+\gamma\alpha_R}\big(\Phi(\tilde{x}_0) - \Phi(x^\star)\big)\Big),$$

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 a new contribution to the literature.

3 Finite manifold identification of SAGA/Prox-SVRG

From this section, we turn to the local convergence properties of SAGA/Prox-SVRG algorithms. We first introduce the notion of partial smoothness, then present a general abstract finite manifold identification of the perturbed Forward–Backward splitting (1.4), and specialize the result to the case of the SAGA and Prox-SVRG algorithms.

3.1 Partial smoothness

The concept *partial smoothness* was first proposed in [25], which captures the essential features of the geometry of non-smoothness along the so-called active/identifiable manifold. Loosely speaking, a partly smooth function behaves smoothly along the manifold, and sharply normal to the manifold.

Let $\mathcal{M}_x$ be a C^2-smooth Riemannian manifold of $\mathbb{R}^n$ around a point x. Denotes $\mathcal{T}_{\mathcal{M}_x}(x')$ the tangent space of $\mathcal{M}_x$ at a point $x' \in \mathcal{M}_x$. Below we introduce the definition of partial smoothness for the class of proper convex and lower semi-continuous functions.

Definition 3.1 (Partly smooth function). Let function $R : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ 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 the subdifferential $\partial R(x) \neq \emptyset$, 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 $T_x \stackrel{\text{def}}{=} \operatorname{par}\big(\partial R(x)\big)^\perp$.
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 widely used non-smooth penalty functions in the literature 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 information. We refer to [29] and the references therein for more details of partly smooth functions.

3.2 An abstract finite manifold identification

Recall the perturbed Forward–Backward splitting iteration

$$x_{k+1} = \operatorname{prox}_{\gamma_k R}\big(x_k - \gamma_k(\nabla F(x_k) + \varepsilon_k)\big).$$

As discussed, the difference of stochastic optimisation methods in terms of perturbed Forward–Backward splitting is that each method has its own form of the perturbation error ε_k (*e.g.* $\varepsilon_k^{\mathrm{SGD}}$ in (1.5), $\varepsilon_k^{\mathrm{SAGA}}$ in (1.7) and $\varepsilon_k^{\mathrm{SVRG}}$ in (1.9)). We have the following abstract identification result for the perturbed Forward–Backward iteration.

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.

Function	Expression	Partial smooth manifold
ℓ_1-norm	$\|x\|_1 = \sum_{i=1}^n \|x_i\|$	$\mathcal{M}_x = T_x = \{z \in \mathbb{R}^n : \mathcal{I}_z \subseteq \mathcal{I}_x\}, \mathcal{I}_x = \{i : x_i \neq 0\}$
$\ell_{1,2}$-norm	$\sum_{i=1}^m \|x_{\mathcal{b}_i}\|$	$\mathcal{M}_x = T_x = \{z \in \mathbb{R}^n : \mathcal{I}_z \subseteq \mathcal{I}_x\}, \mathcal{I}_x = \{i : x_{\mathcal{b}_i} \neq 0\}$
ℓ_∞-norm	$\max_{i=\{1,\dots,n\}} \|x_i\|$	$\mathcal{M}_x = T_x = \{z \in \mathbb{R}^n : z_{\mathcal{I}_x} \in \mathbb{R}\text{sign}(x_{\mathcal{I}_x})\}, \mathcal{I}_x = \{i : \|x_i\| = \|x\|_\infty\}$
TV semi-norm	$\|x\|_{\text{TV}} = \|D_{\text{DIF}} x\|_1$	$\mathcal{M}_x = T_x = \{z \in \mathbb{R}^n : \mathcal{I}_{D_{\text{DIF}} z} \subseteq \mathcal{I}_{D_{\text{DIF}} x}\}, \mathcal{I}_{D_{\text{DIF}} x} = \{i : (D_{\text{DIF}} x)_i \neq 0\}$
Nuclear norm	$\|x\|_* = \sum_{i=1}^r \sigma(x)$	$\mathcal{M}_x = \{z \in \mathbb{R}^{n_1 \times n_2} : \text{rank}(z) = \text{rank}(x) = r\}$, $\sigma(x)$ singular values of x

Theorem 3.2 (Abstract finite manifold identification). *For problem ($\mathcal{P}$), suppose that conditions (A.1)-(A.3) hold. For the perturbed Forward–Backward splitting iteration (1.4), suppose that:*

- **(B.1)** *There exists $\underline{\gamma} > 0$ such that $\liminf_{k \to +\infty} \gamma_k \geq \underline{\gamma}$;*
- **(B.2)** *The perturbation error $\{\varepsilon_k\}_{k \in \mathbb{N}}$ converges to 0 almost surely;*
- **(B.3)** *There exists an $x^\star \in \text{Argmin}(\Phi)$ such that $\{x_k\}_{k \in \mathbb{N}}$ converges to $x^\star$ almost surely.*

For the $x^\star$ in (B.3), suppose that $R \in \text{PSF}_{x^\star}(\mathcal{M}_{x^\star})$, and the following non-degeneracy condition (ND) holds. Then, there exists a $K > 0$ such that for all $k \geq K$, we have $x_k \in \mathcal{M}_{x^\star}$ almost surely.

Remark 3.3.

(i) In the deterministic setting, the finite manifold identification property of (1.4), *i.e.* ε_k is not random error, is discussed in [29, Section 3.3].

(ii) From the convergence proof of Theorem 3.2, it can be observed that condition (**B.1**) can be relaxed if we have

$$\lim_{k \to +\infty} \frac{1}{\gamma_k} \|x_k - x_{k+1}\| = 0$$

holds almost surely, which means that $\mathbb{E}[\|x_k - x_{k+1}\|] = o(\gamma_k)$.

(iii) In Theorem 3.2, we only mention the existence of K after which the manifold identification happens and no estimation is provided. In [29] for the deterministic Forward–Backward splitting method, a lower bound of K is derived, though not very interesting from practical point of view. However, for the stochastic methods (*e.g.* SAGA and Prox-SVRG), even providing a lower bound for K is a challenging problem. More importantly, to provide a bound (either lower or upper) for K, $x^\star$ has to be involved; see [29, Proposition 3.6]. As a consequence, we decide to skip the discussion here.

Proof. First of all, the definition of proximity operator (1.2) and the update of x_{k+1} (1.4) entail that

$$\frac{x_k - x_{k+1}}{\gamma_k} - \nabla F(x_k) - \varepsilon_k \in \partial R(x_{k+1}), \tag{3.1}$$

from which we get

$$\begin{aligned} \text{dist}\big(-\nabla F(x^\star), \partial R(x_{k+1})\big) &\leq \|\tfrac{1}{\gamma_k}\big(x_k - x_{k+1}\big) - \nabla F(x_k) - \varepsilon_k + \nabla F(x^\star)\| \\ &\leq \tfrac{1}{\gamma_k}\|x_k - x_{k+1}\| + \|\nabla F(x_k) - \nabla F(x^\star)\| + \|\varepsilon_k\| \\ &\leq \tfrac{1}{\underline{\gamma}}\|x_{k+1} - x_k\| + L_F\|x_k - x^\star\| + \|\varepsilon_k\|, \end{aligned}$$

where lower boundedness of γ_k and the L_F-Lipschitz continuity of ∇F (see assumption (**A.2**)) is applied to get the last inequality. We have:

- The almost sure convergence of $\{x_k\}_{k \in \mathbb{N}}$ (condition (**B.3**)) ensures that $L_F\|x_k - x^\star\|$ converges to 0 almost surely. Owing to assumption (**A.1**), R is subdifferentially continuous at all the points of its domain, typically at $x^\star$ for $-\nabla F(x^\star)$, hence we have $R(x_k) \to R(x^\star)$ almost surely;

- Combine the almost sure convergence of $\{x_k\}_{k\in\mathbb{N}}$ and (**B.1**) the bounded from below property of $\{\gamma_k\}_{k\in\mathbb{N}}$, we have that $\frac{1}{\gamma}\|x_{k+1}-x_k\|$ converges to 0 almost surely;
- Condition (**B.2**) asserts that $\|\varepsilon_k\| \to 0$ almost surely.

Altogether, we have that

$$\operatorname{dist}\big(-\nabla F(x^\star), \partial R(x_{k+1})\big) \to 0 \ \text{ almost surely.}$$

To this end, all the conditions of [17, Theorem 5.3] are fulfilled almost surely on function $\langle \nabla F(x^\star), \cdot\rangle + R$, hence the identification result follows. □

3.3 Finite manifold identification of SAGA/Prox-SVRG

Now we specialise Theorem 3.2 to the case of SAGA/Prox-SVRG algorithms, which yields the proposition below. For Prox-SVRG, recall that in the convergence proof, we denote the inner iteration sequence $x_{\ell,p}$ as x_k with $k = \ell P + p$. It follows directly from Theorem 2.1 and Theorem 2.2 that the conditions of Theorem 3.2 are satisfied. Therefore, we have the following result.

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 applied under the conditions of Theorem 2.2.*

Then there exists an $x^\star \in \operatorname{Argmin}(\Phi)$ 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 $k \geq K$, $x_k \in \mathcal{M}_{x^\star}$ almost surely.

Remark 3.5. For the Prox-SVRG algorithm, since in Theorem 2.2 the convergence is obtained for "Option I", hence the sequence $\{\tilde{x}_\ell\}_{\ell\in\mathbb{N}}$ also has finite manifold identification property.

The situation however becomes complicated if "Option II" is applied. Suppose we have the convergence of the sequence generated by Prox-SVRG, the identification property of $\{x_{\ell,p}\}_{p=1,\cdots,P,\,\ell\in\mathbb{N}}$ is strateforward. However, for sequence $\{\tilde{x}_\ell\}_{\ell\in\mathbb{N}}$, unless $\mathcal{M}_{x^\star}$ is convex locally around $x^\star$, in general there is no identification guarantee for it. A typical example for which this is problematic is the nuclear norm, whose associated partial smooth manifold is a non-convex cone, hence there can be no identification result for the outer loop sequence $\{\tilde{x}_\ell\}_{\ell\in\mathbb{N}}$.

3.4 When non-degeneracy condition fails

In Theorem 3.2, besides the partial smoothness assumption of R, the non-degeneracy condition (ND) is crucial to the identification of the sequence $\{x_k\}_{k\in\mathbb{N}}$. Owing to the result of [26, 17, 18], it is a necessary condition for identification of the manifold $\mathcal{M}_{x^\star}$, and moreover ensures that the manifold $\mathcal{M}_{x^\star}$ is minimal and unique.

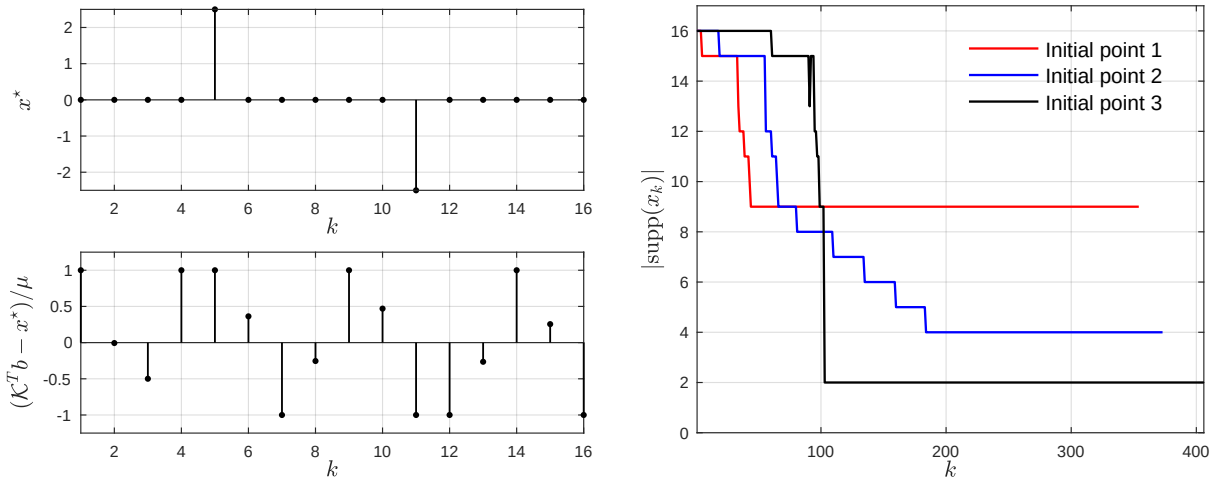
Recently, efforts are made to relax the non-degeneracy condition. In [15], under a so-called "mirror stratification condition", the authors manage to relax the non-degeneracy condition, however at the price that the manifold to be identified is no longer unique. More precisely, there will be another manifold $\overline{\mathcal{M}}_{x^\star}$, which includes $\mathcal{M}_{x^\star}$ and is determined by how (ND) is violated. The sequence $\{x_k\}_{k\in\mathbb{N}}$ will identify a manifold $\widetilde{\mathcal{M}}_{x^\star}$ such that

$$\mathcal{M}_{x^\star} \subseteq \widetilde{\mathcal{M}}_{x^\star} \subseteq \overline{\mathcal{M}}_{x^\star}.$$

Furthermore, the identification of $\{x_k\}_{k\in\mathbb{N}}$ could be unstable, that is $\{x_k\}_{k\in\mathbb{N}}$ may identify several different manifolds which are between $\mathcal{M}_{x^\star}$ and $\overline{\mathcal{M}}_{x^\star}$.

A degenerate LASSO problem We present a simple example of LASSO problem to demonstrate the unstable identification behaviour of $\{x_k\}_{k\in\mathbb{N}}$ when the non-degeneracy conditions fails. Consider the problem

$$\min_{x\in\mathbb{R}^n} \mu\|x\|_1 + \frac{1}{2}\|\mathcal{K}x - b\|^2, \tag{3.2}$$



(a) $x^\star$ and $(\mathcal{K}^T b - x^\star)/\mu$ (b) $|\mathrm{supp}(x_k)|$ under different starting point

Figure 2: Identification properties of deterministic Forward–Backward splitting method when the non-degeneracy condition (ND) fails.

where $\mu > 0$ is the penalty parameter, $\mathcal{K} \in \mathbb{R}^{n\times n}$ is a unitary matrix, and $b \in \mathbb{R}^n$ is a vector.

Since $\mathcal{K}$ is a unitary matrix, the solution of (3.2) is unique and can be given explicitly, which is

$$x^\star = \mathrm{sign}(\mathcal{K}^T b) \odot \max\left\{|\mathcal{K}^T b| - \mu, 0\right\}, \tag{3.3}$$

and $\odot$ denotes pointwise product. Moreover, we have the gradient at $x^\star$

$$-\nabla\left(\frac{1}{2}\|\mathcal{K}x^\star - b\|^2\right) = -\mathcal{K}^T(\mathcal{K}x^\star - b) = \mathcal{K}^T b - x^\star.$$

In the experiments, we set $\mu = 0.5$ and $n = 16$, and moreover the vector b is designed such that the non-degeneracy condition (ND) is violated. The two vectors $x^\star$ and $\mathcal{K}^T b - x^\star$ are shown in Figure 2(a), and it can be observed that $x^\star$ has only *two* non-zero elements, while $\mathcal{K}^T b - x^\star$ has *nine* saturated elements (the saturation means that the absolute value of corresponding element is equal to μ).

Though the solution $x^\star$ can be provided in closed form (3.3), we choose to solve (3.2) with deterministic Forward–Backward splitting with fixed step-size $\gamma = 0.05$, which is the following iteration

$$x_{k+1} = \mathrm{sign}(w_k) \odot \max\left\{|w_k| - \gamma\mu, 0\right\} \quad \text{where} \quad w_k = (1-\gamma)x_k - \mathcal{K}^T b. \tag{3.4}$$

Three different initial points for (3.4) are considered. For each starting point, the size of support of the sequence $\{x_k\}_{k\in\mathbb{N}}$, *i.e.* $\{|\mathrm{supp}(x_k)|\}_{k\in\mathbb{N}}$, is plotted in Figure 2(b). For all three cases, the iterations are ran until machine accuracy is reached. We obtain the following observations from the comparisons:

- "Initial point 1" and "Initial point 2" are unable to identify the support of the solution $x^\star$;
- "Initial point 1" identifies the largest manifold, *i.e.* $\overline{\mathcal{M}}_{x^\star}$. For "Initial point 2", the identification is not stable in the early iterations (*e.g.* $k \leq 190$) compared to the other cases, and eventually (*e.g.* $k \geq 190$) stablises onto a manifold $\widetilde{\mathcal{M}}_{x^\star}$ with $\mathcal{M}_{x^\star} \subset \widetilde{\mathcal{M}}_{x^\star} \subset \overline{\mathcal{M}}_{x^\star}$;
- "Initial point 3" manages to identify the smallest manifold, *i.e.* $\mathcal{M}_{x^\star}$.

We can conclude that the starting point is very crucial when the non-degeneracy condition (ND) fails.

4 Local linear convergence of SAGA/Prox-SVRG

Now we turn to the local linear convergence properties of SAGA/Prox-SVRG algorithms, the contents of this section consist of three main parts: local linear convergence of SAGA/Prox-SVRG, tightness of the rate estimation and more importantly acceleration techniques for these methods.

Throughout the section, $x^\star \in \operatorname{Argmin}(\Phi)$ 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 $x^\star$.

4.1 Local linear convergence

Similar to the result in [29] 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 defined 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 the local continuity of the Hessian of F, condition (RI) implies that there exist $\alpha > 0$ and $r > 0$ such that

$$\langle h,\, \nabla^2 F(x)h\rangle \geq \alpha\|h\|^2,\ \forall h \in T_{x^\star},\, \forall x \ \text{ s.t. } \|x - x^\star\| \leq r.$$

In [29, Proposition 12], it is shown that under the above condition, $x^\star$ 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 [29]). *For problem ($\mathcal{P}$), suppose that assumptions (A.1)-(A.3) hold. Let $x^\star \in \operatorname{Argmin}(\Phi)$ 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 also be found in [23, Theorem 5].

The local quadratic growth, implies that when a sequence convergent stochastic method is applied, and moreover the conditions of Lemma 4.1 are satisfied. Eventually, the method will enter a local neighborhood of the solution $x^\star$ where the function has the quadratic growth property. If moreover the 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 for SAGA and Prox-SVRG respectively.

Proposition 4.3 (Local linear convergence of SAGA). *For problem ($\mathcal{P}$), suppose that conditions (A.1)-(A.3) hold, and the SAGA algorithm (1.6) is applied with $\gamma_k \equiv \gamma = 1/(3L)$. Then x_k converges to $x^\star \in \operatorname{Argmin}(\Phi)$ 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 $k \geq K$,*

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

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

We refer to [12] for the proof of the proposition.

Remark 4.4. Follow the result of SAGA paper, if locally we change to $\gamma = 1/(2(\alpha m + L))$, then we have for ρ_{SAGA}

$$\rho_{\text{SAGA}} = 1 - \alpha\gamma = 1 - \frac{\alpha}{2(\alpha m + L)}.$$

It also should be noted that $\gamma = \frac{1}{3L}$ is the optimal step-size for SAGA as pointed out in [12].

Proposition 4.5 (Local linear convergence of Prox-SVRG). *For problem ($\mathcal{P}$), suppose that conditions (A.1)-(A.3) hold, and the Prox-SVRG algorithm (1.8) is applied such that Theorem 2.2 holds. Then x_k converges to $x^\star \in \operatorname{Argmin}(\Phi)$ 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 $k \geq K$,*

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

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

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

Remark 4.6.

(i) When P is large enough, then $\rho_{\mathrm{SVRG}} \approx \frac{1}{\alpha\gamma(1-4L\gamma)P} + \frac{4L\gamma}{1-4L\gamma}$, to make it strictly smaller than 1, we need $P \geq 32L/\alpha$ and moreover

$$\gamma \in \left[\frac{P\alpha - \sqrt{\Delta}}{16LP\alpha}, \frac{P\alpha + \sqrt{\Delta}}{16LP\alpha}\right] \quad \text{where} \quad \Delta = P\alpha(P\alpha - 32L).$$

(ii) In [16], the authors studied the linear convergence convergence of Prox-SVRG under a "semi-strongly convex" assumption. Our assumption for local linear convergence is very close to this one, however in stead of only allowing polyhedral functions (typically ℓ_1-norm), our analysis goes much further, for instance our result allows to analyse nuclear norm.

(iii) The above local linear convergence result is quite different from that of [29, Section 4] for the deterministic FBS-type methods, which can be summarised into the following steps:

Step 1. Locally along the identified $\mathcal{M}_{x^\star}$, the globally non-linear iteration (1.1) can be linearised, resulting in a linear matrix M_{FB};

Step 2. Spectral properties of M_{FB}, conditions such that the spectral radius $\rho(M_{\mathrm{FB}}) < 1$;

Step 3. Local linear convergence of FBS-type splitting schemes.

The advantage of this strategy is that it exploits explicitly the geometry of the manifold $\mathcal{M}_{x^\star}$ and encodes it into the matrix M_{FB}, which result in a very tight rate estimation.

The main difficulty of applying the above strategy to SAGA/Prox-SVRG is that, under the stochastic setting, the error ε_k in (1.4) cannot be controlled explicitly, which makes it impossible to use the spectral radius $\rho(M_{\mathrm{FB}})$ as rate estimation; see the section below for more details.

4.2 Better local rate estimation?

Consider FBS and SAGA algorithms, when Φ is α-strongly convex and the step-size is chosen as $\gamma = 1/(3L)$, then the convergence rate of $\|x_k - x^\star\|$ for these two algorithms are

$$\rho_{\mathrm{FBS}} = 1 - \frac{\alpha}{3L}, \quad \rho_{\mathrm{SAGA}} = \sqrt{1 - \min\{\tfrac{1}{4m}, \tfrac{\alpha}{3L}\}}.$$

Clearly, the rate estimation of FBS is better than that of SAGA. Note that here we are comparing the convergence rate *per iteration*, not based on gradient evaluation complexity. For the rest of this part, we will discuss the difficulties of improving the rate estimations for SAGA and Prox-SVRG.

4.2.1 Local linearised iteration

Follow the setting of [29], suppose that F locally around $x^\star$ is C^2-smooth, define the following matrices which are all *symmetric*:

$$H_F \stackrel{\text{def}}{=} \mathrm{P}_{T_{x^\star}} \nabla^2 F(x^\star) \mathrm{P}_{T_{x^\star}}, \quad G_F \stackrel{\text{def}}{=} \mathrm{Id} - \gamma H_F, \quad H_R \stackrel{\text{def}}{=} \nabla^2_{\mathcal{M}_{x^\star}} \Phi(x^\star) \mathrm{P}_{T_{x^\star}} - H_F, \tag{4.1}$$

where $\nabla^2_{\mathcal{M}_{x^\star}} \Phi$ is the Riemannian Hessian of Φ along the manifold $\mathcal{M}_{x^\star}$; see Lemma A.6.

Lemma 4.7 ([29, Lemma 13,14]). *For problem ($\mathcal{P}$), suppose that conditions (A.1)-(A.3) hold and $x^\star \in \mathrm{Argmin}(\Phi)$ such that $R \in \mathrm{PSF}_{x^\star}(\mathcal{M}_{x^\star})$, F is locally C^2 around $x^\star$ and conditions (ND) and (RI) hold.*

(i) *H_R is symmetric positive semi-definite, hence $\mathrm{Id} + \gamma H_R$ is invertible, and $W_R \stackrel{\text{def}}{=} (\mathrm{Id} + \gamma H_R)^{-1}$ is symmetric positive definite with eigenvalues in $]0, 1]$.*

(ii) *Define the matrix M_{FB} by*

$$M_{\mathrm{FB}} \stackrel{\text{def}}{=} W_R G_F. \tag{4.2}$$

For $\gamma \in]0, 1/L[$, M_{FB} has real eigenvalues lying in $]0, 1[$ with spectral radius $\rho(M_{\mathrm{FB}}) \leq 1 - \alpha\gamma$.

Proposition 4.8 (Local linearized iteration). *For problem ($\mathcal{P}$), suppose that conditions (A.1)-(A.3) hold. Assume the perturbed Forward–Backward iteration (1.4) is applied to create a sequence $\{x_k\}_{k\in\mathbb{N}}$ such that the conditions of Theorem 3.2 hold. Then there exists an $x^\star \in \operatorname{Argmin}(\Phi)$ such that $x_k \to x^\star$ almost surely and $x_k \in \mathcal{M}_{x^\star}$ for all k large enough.*

If moreover, F is locally C^2-smooth around $x^\star$ and $\gamma_k \to \gamma \in]0, 1/L[$, then with probability one, there exists $K \in \mathbb{N}$ such that for all $k \geq K$, we have

$$d_{k+1} = M_{\mathrm{FB}} d_k + \phi_k, \tag{4.3}$$

where $d_k \overset{\text{def}}{=} x_k - x^\star + o(\|x_k - x^\star\|)$ and $\phi_k = \gamma W_R \mathrm{P}_{T_{x^\star}} \varepsilon_k + o(\|\varepsilon_k\|)$.

See Appendix A.2 for the proof. Note that for ϕ_k, there still holds $\mathbb{E}[\phi_k] = 0$.

Remark 4.9. In [29], the linearisation of deterministic FBS scheme reads,

$$x_{k+1} - x^\star = M_{\mathrm{FB}}(x_k - x^\star) + o(\|x_k - x^\star\|),$$

which is much more straightforward than Theorem 4.8. The reason for such a difference is that the behaviour of deterministic FBS is monotonic, *e.g.* $\|x_{k+1} - x^\star\| \leq \|x_k - x^\star\|$, which allows us to encode all the small o-terms into $o(\|x_k - x^\star\|)$.

4.2.2 No better rate estimation

We discuss in short why the spectral radius of M_{FB} cannot serve as the local convergence rate of stochastic optimisation methods, which is different from the deterministic setting. Let $K \in \mathbb{N}$ be sufficiently large such that (4.3) holds. Then we get

$$d_{k+1} = M_{\mathrm{FB}}^{k+1-K} d_K + \textstyle\sum_{j=K}^{k} M_{\mathrm{FB}}^{k-j} \phi_j.$$

Take $\rho \in]\rho(M_{\mathrm{FB}}), 1[$, owing to Lemma 4.7, there exists a constant $C > 0$ such that

$$\begin{aligned}\mathbb{E}(\|d_{k+1}\|) &\leq C\rho^{k+1-K}\mathbb{E}(\|x_K - x^\star\|) + \textstyle\sum_{j=K}^{k} \rho^{k-j}\mathbb{E}(\|\phi_j\|) \\ &\leq C\rho^{k+1-K}\left(\mathbb{E}(\|x_K - x^\star\|) + \rho^{K-1}\textstyle\sum_{j=K}^{k} \dfrac{\mathbb{E}(\|\phi_j\|)}{\rho^j}\right).\end{aligned}$$

Now consider the SAGA algorithm, owing to Proposition 4.3, we have only that

$$\mathbb{E}(\|\phi_j\|) = O\big((\sqrt{\rho_{\mathrm{SAGA}}})^j\big),$$

which means $\lim_{k\to+\infty} \sum_{j=K}^{k} \frac{\mathbb{E}(\|\phi_j\|)}{\rho^j} < +\infty$ holds only for $\rho \in]\sqrt{\rho_{\mathrm{SAGA}}}, 1[$. As a consequence, we can only obtain the same rate estimation as the original SAGA.

Remark 4.10. The main message of the above discussion is: under a given step-size γ, the spectral radius $\rho(M_{\mathrm{FB}})$ is the optimal convergence rate can be achieved by SAGA/Prox-SVRG. However, depending on the problems to solve, the practical performance of these methods could be slower than $\rho(M_{\mathrm{FB}})$.

An overdetermined LASSO problem In Section 5.1, a sparse logistic regression problem is considered, where both SAGA and Prox-SVRG converge at the rate of $\rho(M_{\mathrm{FB}})$; see Figure 4. Below we design an example of LASSO problem, to discuss the situations where $\rho(M_{\mathrm{FB}})$ cannot be achieved.

Consider again the LASSO problem,

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

where now $\mathcal{K} \in \mathbb{R}^{m\times n}$ is a random Gaussian matrix with zero means and $b \in \mathbb{R}^m$. Moreover, we choose $m = 256, n = 32$, that is much more measurements than the size of the vector.

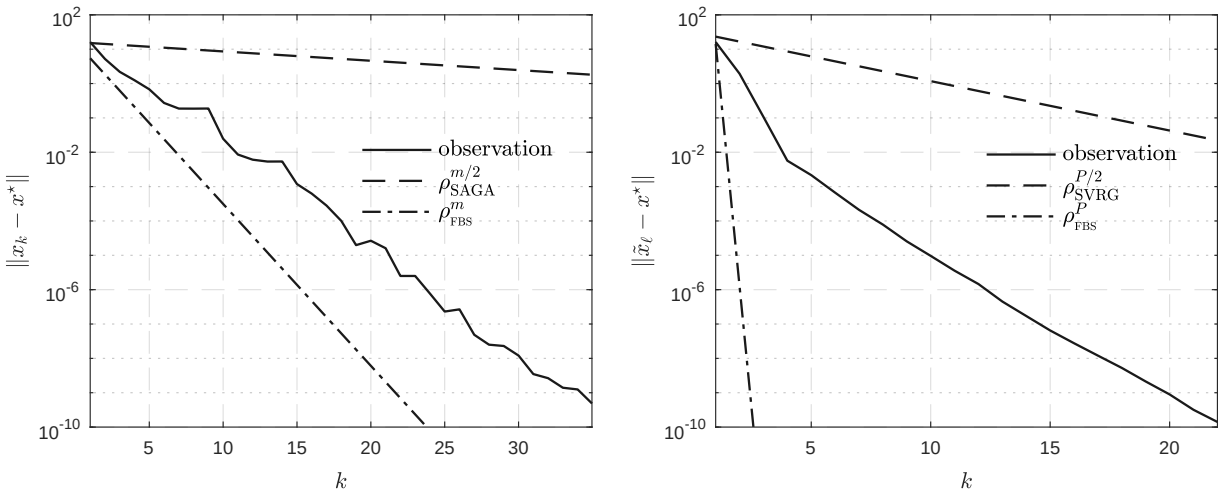
For the test example, we have $L = 0.2239$ and the local quadratic grow parameter $\alpha = 0.0032$. The parameter choices of SAGA and Prox-SVRG with "Option II" are:

$$\text{SAGA} : \gamma = \frac{1}{3L}; \qquad \text{Prox-SVRG} : \gamma = \frac{1}{10L}, \quad P = \frac{100L}{\alpha}.$$

We have $P \approx 27m$ which is quite large. As discussion in the original work [43], with the above parameters choices, $\rho_{\text{SVRG}} \approx \frac{5}{6}$.

The outcomes of the numerical experiments are shown in Figure 3, where the observation of $\{\|x_k - x^\star\|\}_{k\in\mathbb{N}}$ is provided for SAGA and $\{\|\tilde{x}_\ell - x^\star\|\}_{k\in\mathbb{N}}$ for Prox-SVRG. The *solid* lines stand for practical observations of the methods, the *dashed* lines are the theoretical estimation from Proposition 4.3 and 4.5, the *dot-dashed* lines are the estimation from $\rho(M_{\text{FB}})$. All the lines are sub-sampled, one out of every m points for SAGA and P points for Prox-SVRG. Note also that the observation is not in norm square.

For this example, both the convergence speeds of SAGA and Prox-SVRG are slower than the spectral radius $\rho(M_{\text{FB}})$. Empirically, the reason for SAGA is that the ratio of m/n much larger than 1, while for Prox-SVRG, the reason is that P/m is too large.



(a) SAGA, $\|x_k - x^\star\|$ (b) Prox-SVRG "Option II", $\|\tilde{x}_\ell - x^\star\|$

Figure 3: Convergence rate of SAGA and Prox-SVRG when solving an overdetermined LASSO problem. (a) convergence behaviour of $\|x_k - x^\star\|$ of SAGA; (b) convergence behaviour of $\|\tilde{x}_\ell - x^\star\|$ of Prox-SVRG. The *solid* lines stands for practical observations of the methods, the *dashed* lines are the theoretical estimation from Proposition 4.3 and 4.5, the *dot-dashed* lines are the estimation from $\rho(M_{\text{FB}})$. All the lines are sub-sampled, one out of every m points for SAGA and P points for Prox-SVRG. $\rho_{M_{\text{FB}}}$ denotes the spectral radius of M_{FB}.

4.3 Beyond local convergence analysis

As already pointed out, manifold identification (Theorem 3.2) implies that, the globally non-smooth problem locally becomes a C^2-smooth and possibly non-convex (*e.g.* nuclear norm) problem, constrained on the identified manifold, that is

$$\min_{x\in\mathbb{R}^n} \Phi \quad \text{non-smooth on } \mathbb{R}^n \qquad \xrightarrow{\text{Theorem 3.2}} \qquad \min_{x\in\mathcal{M}_{x^\star}} \Phi \quad C^2\text{-smooth on } \mathcal{M}_{x^\star}$$

Such a transition to local C^2-smoothness, provides various choices of acceleration. In the following, we discuss several practical acceleration strategies.

4.3.1 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

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

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

For SAGA/Prox-SVRG or other stochastic methods which have the manifold identification property, once the manifold is identified, they can adapt their step-sizes to the local Lipschitz of the problem once the manifold is identified, one can adapt their step-sizes to the local Lipschitz constants of the problem. Since step-size is crucial to the convergence speed of these algorithms, the potential acceleration of such as local adaptive strategy can be significant.

In the numerical experiments section, this strategy is applied to the sparse logistic regression problem. For the considered problem, we have $L/L_{\mathcal{M}_{x^\star}} \approx 16$, and the adaptive strategy achieves a 16 times acceleration. It is worth mentioning that, the computational cost for evaluating $\mathcal{M}_{x^\star}$ is negligible.

4.3.2 Lower computational complexity

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

Take $R = \|\cdot\|_1$ 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 $O(n)$ complexity;
After identification $O(\kappa)$ complexity;

The reduction of computational cost is decided by the ratio of n/κ. Depending on this ratio, either mini-batch based methods, or even deterministic methods with momentum acceleration can be applied (*e.g.* inertial Forward–Backward schemes [29], FISTA [3]).

4.3.3 Higher-order acceleration

The last acceleration strategy to discuss is the Riemannian manifold based higher-order acceleration. Recently, various the Riemannian manifold based optimisation methods are proposed in the literature [20, 37, 40, 5], particularly for low-rank matrix recovery. However, an obvious drawback of this class of methods is that the manifold should known *a priori*, which limits the applications of these methods.

The manifold identification property 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 method, when the restricted injectivity condition (RI) is satisfied, and Riemannian geometry based optimisation methods [24, 32, 39, 5, 40], for instance the non-linear conjugate gradient method [39]. Stochastic Riemannian manifold based optimisation methods are also studied in the literature, for instance in [44], the authors generalised the SVRG method to the manifold setting.

5 Numerical experiments

In this section, we consider several concrete examples to illustrate our results. Three examples of R are considered, sparsity promoting ℓ_1-norm, group sparsity promoting $\ell_{1,2}$-norm and low rank promoting nuclear norm. We refer to [29] and the references therein for the detailed properties of these functionals.

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.

5.1 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 spectral radius $\rho(M_{\mathrm{FB}})$.

Example 5.1 (Sparse logistic regression). Let $m > 0$ and $(z_i, y_i) \in \mathbb{R}^n \times \{\pm 1\}$, $i = 1, \cdots, m$ be the training set. The sparse logistic regression is to find a linear decision function which minimizes the objective

$$\min_{(x,b)\in\mathbb{R}^n\times\mathbb{R}} \mu\|x\|_1 + \frac{1}{m}\sum_{i=1}^m \log\left(1 + e^{-y_i f(z_i;x,b)}\right), \tag{5.1}$$

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

The setting of the experiment is: $n = 256$, $m = 128$, $\mu = 1/\sqrt{m}$ and $L = 1188$. Apparently, the dimension of the problem is larger than the number of training points. The parameters choices of SAGA and Prox-SVRG are:

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

Remark 5.2. The step-sizes of SAGA/Prox-SVRG exceeds the one allowed by Theorem 2.1 and 2.2, respectively. The reason of choosing different step-sizes for SAGA and Prox-SVRG is mostly for the visual quality of the graphs in Figure 4.

The observations of the experiments are shown in Figure 4. 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 4(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 started with the same initial point. It is observed that SAGA shows faster manifold identification than Prox-SVRG, this is mainly due the fact that the step-size of SAGA (*i.e.* $\gamma = \frac{1}{2L}$) is larger than that of Prox-SVRG (*i.e.* $\gamma = \frac{1}{3L}$). The identification speed of the two algorithms are very close if they are applied under the same choice of step-size.

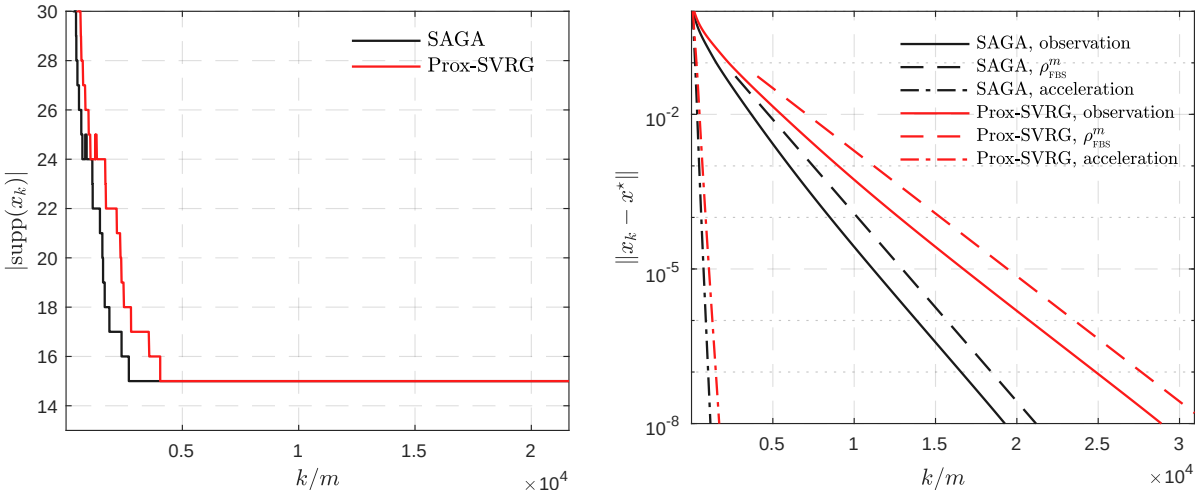
Local linear convergence In Figure 4(b), we demonstrate the convergence rate of $\{\|x_k - x^\star\|\}_{k\in\mathbb{N}}$ of the two algorithms. The two *solid* lines are the practical observation of $\{\|x_k - x^\star\|\}_{k\in\mathbb{N}}$ generated by SAGA and Prox-SVRG, the two *dashed* lines are the theoretical estimations using the spectral radius of M_{FB}, 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.

Since ℓ_1-norm is polyhedral, the spectral radius of M_{FB}, denoted by $\rho_{M_{\mathrm{FB}}}$, is determined by α and γ, that is $\rho_{M_{\mathrm{FB}}} = 1 - \gamma\alpha$. Given the values of α and γ of SAGA and Prox-SVRG, we have that

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

For the consider problem setting, the spectral radius quite matches the practical observations.

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) shows 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.



(a) $|\mathrm{supp}(x_k)|$ (b) $\|x_k - x^\star\|$

Figure 4: Finite manifold identification and local linear convergence of SAGA and Prox-SVRG for solving sparse logistic regression problem in Example 5.1. (a) finite manifold identification of SAGA/Prox-SVRG; (b) local linear convergence of SAGA/Prox-SVRG. $\rho_{M_{\mathrm{FB}}}$ denotes the spectral radius of M_{FB}.

5.2 Local higher-order acceleration

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

Example 5.3 (Group sparse and low-rank regression [13, 6]). 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 i.i.d. zero-mean and unit-variance Gaussian distribution, $\omega \in \mathbb{R}^m$ is an additive error with bounded ℓ_2-norm.

Let $\mu > 0$, and $R(x)$ 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},

$$\min_{x\in\mathbb{R}^n} \mu R(x) + \frac{1}{m}\sum_{i=1}^m \tfrac{1}{2}\|\mathcal{K}_i x - b_i\|_2^2, \tag{5.2}$$

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

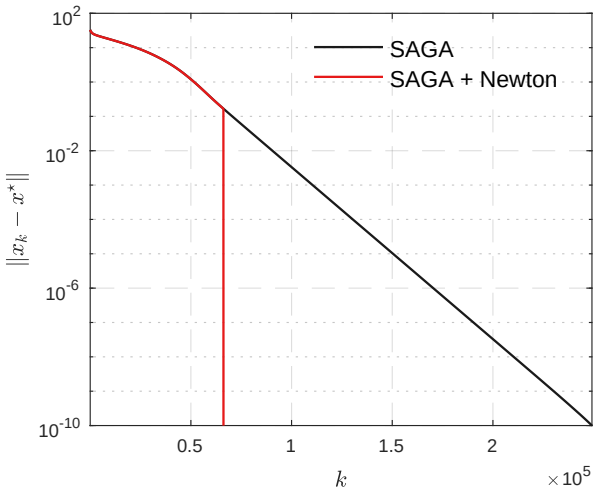
We have the following settings for the two examples of R:

Group sparsity: $n = 512$, $m = 256$, x_{ob} has 8 non-zero blocks of block-size 4;
Low rank: $n = 4096$, $m = 2048$, the rank of x_{ob} is 4.

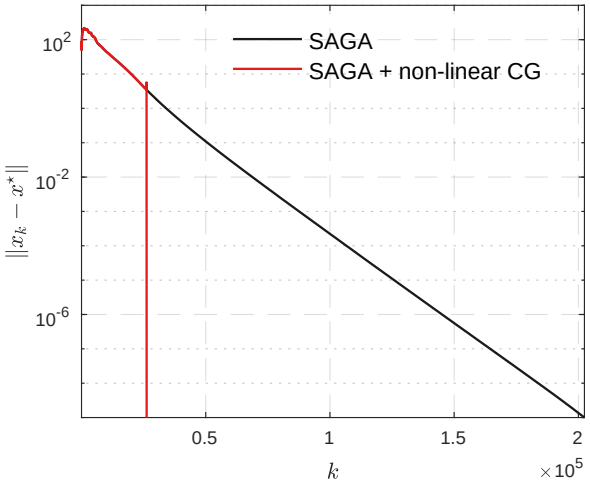
We consider only the SAGA algorithm for this test, as the main purpose is higher-order acceleration. For $\ell_{1,2}$-norm, Newton method is applied after the manifold identification, while for nuclear norm, a non-linear conjugate gradient [5] is applied after manifold identification.

The numerical results are shown in Figure 5. For $\ell_{1,2}$-norm, the black line is the observation of SAGA algorithm with $\gamma = \frac{1}{3L}$, the red line is the observation of "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 quadratic convergence Newton method converges in only few steps. For nuclear norm, a non-linear conjugate gradient is applied when the manifold is identified. Similar to the observation of $\ell_{1,2}$-norm, the super-linearly convergent non-linear conjugate gradient shows superior performance to SAGA.



(a) Group sparsity



(b) Low rank

Figure 5: Local higher-order acceleration after manifold identification in Example 5.3. (a) Newton method is applied after the manifold is identified by SAGA; (b) non-linear conjugate gradient is applied after manifold identification. Black line is the observation of SAGA algorithm, and the red line is the observation of SAGA+higher-order scheme. The black lines of SAGA for both examples are not subsampled.

6 Conclusion

In this paper, we proposed a unified framework of local convergence analysis for proximal stochastic variance reduced gradient methods, and typically focused on 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.

A Proofs of theorems

A.1 Proofs for Section 2

To prove Theorem 2.1 and 2.2, the lemma below is needed which is classical result from stochastic analysis [36].

Lemma A.1 (Supermartingale convergence). *Let Y_k, Z_k and W_k, $k = 0, 1, \ldots$, be three sequences of random variables and let $\mathcal{F}_k$, $k = 0, 1, \ldots$, be sets of random variables such that $\mathcal{F}_k \subset \mathcal{F}_{k+1}$ for all k. Suppose that:*

(i) *The random variables Y_k, Z_k and W_k are nonnegative, and are functions of the random variables in $\mathcal{F}_k$.*

(ii) *For each k, we have $\mathbb{E}(Y_{k+1}|\mathcal{F}_k) \leq Y_k - Z_k + W_k$.*

(iii) *With probability 1, $\sum_k W_k < \infty$.*

Then we have $\sum_k Z_k < \infty$ and the sequence Y_k converges to a nonnegative random variable Y with probability 1.

Proof of Theorem 2.1. The convergence of the objective function value for $\gamma_k \equiv \frac{1}{3L}$ is already studied in [12], here for the completeness of the proof, we shall keep the convergence proof of the objective function.

The proof of the theorem consists of several steps. First is the convergence of the objective function value. Let $\phi_{k,i}$ be the point such that $g_{k,i} = \nabla f_i(\phi_{k,i})$, then following the proof in the original SAGA paper [12], define the following Lyapunov

function $\mathcal{L}$,

$$\mathcal{L}_k \overset{\text{def}}{=} \mathcal{L}(x_k, \{\phi_{k,i}\}_{i=1}^m) \overset{\text{def}}{=} \frac{1}{m}\sum_{i=1}^m f_i(\phi_{k,i}) - F(x^\star) - \frac{1}{m}\sum_{i=1}^m \langle \nabla f_i(x^\star),\, \phi_{i,k} - x^\star\rangle + c\|x_k - x^\star\|^2$$

for some appropriate $c > 0$. Denote $\mathbb{E}_k[\cdot]$ the conditional expectation on step k. Then following the Appendix C of the supplementary material of [12], one can show that

$$\mathbb{E}_k[\mathcal{L}_{k+1}] \leq \mathcal{L}_k - \frac{1}{4m}\mathbb{E}_k[\Phi(x_{k+1}) - \Phi(x^\star)]. \tag{A.1}$$

Since $\mathbb{E}_k[\Phi(x_{k+1}) - \Phi(x^\star)]$ is a non-negative random variable of the k^{th} iteration, it then follows that $\{\mathcal{L}_k\}_{k\in\mathbb{N}}$ is a super-martingale owing to Lemma A.1. Therefore $\{\mathcal{L}_k\}_{k\in\mathbb{N}}$ converges to a non-negative random variable $\mathcal{L}^\star$ with probability 1. At the same time, with probability 1, $\|x_k - x^\star\|^2 \leq \frac{1}{c}\mathcal{L}_k$, hence $\{x_k\}_{k\in\mathbb{N}}$ is a bounded sequence and every cluster point of $\{x_k\}_{k\in\mathbb{N}}$ is a global minimiser of Φ. Moreover, from Lemma A.1 and (A.1), we have

$$\sum_{k=0}^\infty \big(\mathbb{E}_k[\Phi(x_{k+1}) - \Phi(x^\star)]\big) \leq \mathcal{L}_0 < +\infty$$

holds almost surely. Define a new random variable $y_j \overset{\text{def}}{=} \sum_{k\geq j}\mathbb{E}_k[\Phi(x_{k+1}) - \Phi(x^\star)]$, clearly we have $\{y_j\}_{k\in\mathbb{N}}$ is non-increasing and converges to 0 as $j \to +\infty$. As a consequence, by the monotone convergence theorem, we have

$$0 = \mathbb{E}\Big[\lim_{j\to+\infty} y_j\Big] = \lim_{j\to+\infty}\mathbb{E}[y_j] = \lim_{j\to+\infty}\sum\nolimits_{k\geq j}\mathbb{E}[\Phi(x_{k+1}) - \Phi(x^\star)] = \lim_{j\to+\infty}\mathbb{E}\Big[\sum\nolimits_{k\geq j}\big(\Phi(x_{k+1}) - \Phi(x^\star)\big)\Big],$$

which implies

$$\mathbb{E}\Big[\sum\nolimits_{k\geq j}\big(\Phi(x_{k+1}) - \Phi(x^\star)\big)\Big] < +\infty \Longrightarrow \sum\nolimits_k\big(\Phi(x_{k+1}) - \Phi(x^\star)\big) < +\infty \text{ almost surely}, \tag{A.2}$$

hence $\Phi(x_k) \to \Phi(x^\star)$ almost surely.

With the boundedness of $\{x_k\}_{k\in\mathbb{N}}$, the second step is to prove that $\{\|x_k - x^\star\|\}_{k\in\mathbb{N}}$ is convergent. Define a new sequence

$$w_k \overset{\text{def}}{=} \frac{1}{m}\sum_{i=1}^m f_i(\phi_{k,i}) - F(x^\star) - \frac{1}{m}\sum_{i=1}^m \langle \nabla f_i(x^\star),\, \phi_{i,k} - x^\star\rangle.$$

Observe that

$$\mathbb{E}_k[w_{k+1}] = \frac{1}{m}F(x_k) - F(x^\star) - \frac{1}{m}\langle \nabla F(x^\star),\, x_k - x^\star\rangle + \big(1 - \frac{1}{m}\big)w_k.$$

Since $x^\star \in \operatorname{Argmin}(\Phi)$ is a global minimiser, we have $-\nabla F(x^\star) \in \partial R(x^\star)$ and $\langle -\nabla F(x^\star),\, x_k - x^\star\rangle \leq R(x_k) - R(x^\star)$, therefore from above equality we further obtain

$$\mathbb{E}_k[w_{k+1}] \leq \frac{1}{m}\big(\Phi(x_k) - \Phi(x^\star)\big) + \big(1 - \frac{1}{m}\big)w_k.$$

Taking expectations over all previous steps for both sides and summing from $k = 0$ to j yields

$$\mathbb{E}[w_{j+1}] + \frac{1}{m}\sum_{k=1}^j \mathbb{E}[w_k] \leq \frac{1}{m}\sum_{k=0}^j \mathbb{E}\big[\Phi(x_k) - \Phi(x^\star)\big] + \big(1 - \frac{1}{m}\big)\mathbb{E}[w_0].$$

As a result, taking j to $+\infty$ implies that $\mathbb{E}[\sum_{k=1}^j w_k] < +\infty$, hence $\sum_{k=1}^j w_k < +\infty$ almost surely. Moreover, $w_k \to 0$ with probability 1. From the convergence result of $\{\mathcal{L}_k\}_{k\in\mathbb{N}}$ and $\{w_k\}_{k\in\mathbb{N}}$, we have that almost surely $\{\|x_k - x^\star\|\}_{k\in\mathbb{N}}$ is bounded and convergent.

Next we prove the almost sure convergence of the sequence $\{x_k\}_{k\in\mathbb{N}}$. Let $\{x_i^\star\}_i$ be a countable subset of the relative interior $\operatorname{ri}(\operatorname{Argmin}(\Phi))$ that is dense in $\operatorname{Argmin}(\Phi)$. From the almost sure convergence of $\|x_k - x^\star\|, x^\star \in \operatorname{Argmin}(\Phi)$, we have that for each i, the probability $\operatorname{Prob}(\{\|x_k - x_i^\star\|\}_{k\in\mathbb{N}}$ is not convergent$) = 0$. Therefore

$$\begin{aligned}\operatorname{Prob}\big(\forall i, \exists b_i \text{ s.t. } \lim_{k\to+\infty}\|x_k - x_i^\star\|\big) &= 1 - \operatorname{Prob}(\{\|x_k - x_i^\star\|\}_{k\in\mathbb{N}} \text{ is not convergent})\\ &\geq 1 - \sum\nolimits_i \operatorname{Prob}(\{\|x_k - x_i^\star\|\}_{k\in\mathbb{N}} \text{ is not convergent}) = 1,\end{aligned}$$

where the inequality follows from the union bound, *i.e.* for each i, $\{\|x_k - x_i^\star\|\}_{k\in\mathbb{N}}$ is a convergent sequence. For a contradiction, suppose that there are convergent subsequences $\{u_{k_j}\}_{k_j}$ and $\{v_{k_j}\}_{k_j}$ of $\{x_k\}_{k\in\mathbb{N}}$ which converge to their limiting points $u^\star$ and $v^\star$ respectively, with $\|u^\star - v^\star\| = r > 0$. Since $\Phi(x_k)$ converges to $\inf\Phi$, these two limiting points are necessarily

in $\operatorname{Argmin}(\Phi)$. Since $\{x_i^\star\}_i$ is dense in $\operatorname{Argmin}(\Phi)$, we may assume that for all $\epsilon > 0$, we have $x_{i_1}^\star$ and $x_{i_2}^\star$ are such that $\|x_{i_1}^\star - u^\star\| < \epsilon$ and $\|x_{i_2}^\star - v^\star\| < \epsilon$. Therefore, for all k_j sufficiently large,

$$\|u_{k_j} - x_{i_1}^\star\| \leq \|u_{k_j} - u^\star\| + \|u^\star + x_{i_1}^\star\| < \|u_{k_j} - u^\star\| + \epsilon.$$

On the other hand, for sufficiently large j, we have

$$\|v_{k_j} - x_{i_1}^\star\| \geq \|v^\star - u^\star\| - \|u^\star - x_{i_1}^\star\| - \|v_{k_j} - v^\star\| > r - \epsilon - \|v_{k_j} - v^\star\| > r - 2\epsilon.$$

This contradicts with the fact that $x_k - x_{i_1}^\star$ is convergent. Therefore, we must have $u^\star = v^\star$, hence there exists $\bar{x} \in \operatorname{Argmin}(\Phi)$ such that $x_k \to \bar{x}$.

Finally, to see that $\varepsilon_k^{\mathrm{SAGA}} \to 0$, from [12, Lemma 6],

$$\frac{1}{m}\sum_{i=1}^m \|\nabla f_i(\phi_{k,i}) - \nabla f_i(x^*)\|^2 \leq 2Lw_k \to 0,$$

therefore, combining this with the fact that ∇f_j is L-Lipschitz and $x_k \to x^*$, it follows that

$$\|\varepsilon_k^{\mathrm{SAGA}}\| \leq \|\nabla f_{i_k}(x_k) - \nabla f_{i_k}(\phi_{k,i})\| + \frac{1}{m}\sum_{j=1}^m \|\nabla f_j(\phi_{k,i}) - \nabla f_j(x_k)\| \to 0,$$

which concludes the proof. □

To prove Theorem 2.2, we require the following lemma, which is a direct consequence of Eq. (16) and Corollary 3 of [43].

Lemma A.2. *Assume that F is α_F-strongly convex and R is α_R-strongly convex. Let $\{x_{\ell,p}\}_{\ell,p}$ be the sequence generated by Prox-SVRG. Then, conditional on step $k = \ell P + p - 1$, we have*

$$\begin{aligned}&(1+\gamma\alpha_R)\mathbb{E}_k[\|x_{\ell,p} - x^\star\|^2]\\ &\leq (1-\gamma\alpha_F)\|x_{\ell,p-1} - x^\star\|^2 - 2\gamma\big(\Phi(x_{\ell,p}) - \Phi(x^\star)\big) + 8L\gamma^2\left(\Phi(x_{\ell,p-1}) - \Phi(x^\star) + \Phi(\tilde{x}_\ell) - \Phi(x^\star)\right).\end{aligned} \tag{A.3}$$

Proof of Theorem 2.2. We begin with the remark that following the arguments in the proof of Theorem 2.1, to show that $x_{\ell,p} \to x^\star$ almost surely for some $x^\star \in \operatorname{argmin}(\Phi)$, it is sufficient to prove that $\|x_{\ell,p} - x^\star\|$ is convergent. By Lemma A.2 with $\alpha_R = \alpha_F = 0$, we have that conditional on step $k = \ell P + p - 1$,

$$\mathbb{E}_k[\|x_{\ell,p} - x^\star\|^2] + 2\gamma\mathbb{E}_k[\Phi(x_{\ell,p}) - \Phi(x^\star)] \leq \|x_{\ell,p-1} - x^\star\|^2 + 8L\gamma^2\left(\Phi(x_{\ell,p-1}) - \Phi(x^\star) + \Phi(\tilde{x}_\ell) - \Phi(x^\star)\right). \tag{A.4}$$

Summing (A.4) over $p = 1, \ldots, P$ and taking expectation on the random variables $i_1, \ldots, i_P$, we obtain that

$$\begin{aligned}&\mathbb{E}[\|x_{\ell,P} - x^\star\|^2] + 2\gamma\mathbb{E}[\Phi(x_{\ell,P}) - \Phi(x^\star)] + 2\gamma(1-4L\gamma)\sum_{j=1}^{P-1}\mathbb{E}[\Phi(x_{\ell,j}) - \Phi(x^\star)]\\ &\leq \|\tilde{x}_\ell - x^\star\|^2 + 8L\gamma^2(P+1)\big(\Phi(\tilde{x}_\ell) - \Phi(x^\star)\big).\end{aligned} \tag{A.5}$$

Since $\gamma \leq \frac{1}{4L(P+2)}$, which yields $2\gamma(1-4L\gamma) \geq \gamma^2$, we obtain from (A.5)

$$\begin{aligned}&\mathbb{E}[\|x_{\ell,P} - x^\star\|^2] + (2\gamma - \gamma^2)\mathbb{E}[\Phi(x_{\ell,P}) - \Phi(x_*)] + \gamma^2\sum_{j=1}^{P}\mathbb{E}[\Phi(x_{\ell,j}) - \Phi(x^\star)]\\ &\leq \|\tilde{x}_\ell - x^\star\|^2 + 8L\gamma^2(P+1)(\Phi(\tilde{x}_\ell) - \Phi(x^\star)).\end{aligned}$$

Moreover, under "Option I", by defining the non-negative random variables

$$T_\ell \overset{\text{def}}{=} \|\tilde{x}_\ell - x^\star\|^2 + (2\gamma - \gamma^2)(\Phi(\tilde{x}_\ell) - \Phi(x^\star)) \text{ and } S_{\ell+1} \overset{\text{def}}{=} \sum_{j=1}^{P}\big(\Phi(x_{\ell,j}) - \Phi(x^\star)\big).$$

It follows from $8L\gamma^2(P+1) \leq 2\gamma - \gamma^2$ that

$$\mathbb{E}[T_{\ell+1}] \leq T_\ell - \gamma^2\mathbb{E}[S_{\ell+1}]. \tag{A.6}$$

So, by the super-martingale convergence theorem, $\{T_\ell\}_{\ell\in\mathbb{N}}$ converges to a nonnegative random variable and $\sum_\ell S_\ell < +\infty$ holds almost surely. In particular, we have $S_\ell \to 0$ as $\ell \to \infty$ and hence, $\Phi(\tilde{x}_\ell) \to \Phi(x^\star)$ as $\ell \to \infty$. Therefore, $\|\tilde{x}_\ell - x^\star\|^2$ converges almost surely. Following the proof of Theorem 2.1, we can then show that $\tilde{x}_\ell$ converges to an optimal point $x^\star$ almost surely.

Now we prove that the inner iteration sequence $\{x_{\ell,p}\}_{1\le p\le P,\,\ell\in\mathbb{N}}$ also converge to $x^\star$ as $\ell\to\infty$. Consider the inequality (A.4), and define the non-negative random variables

$$V_{\ell,p}\overset{\text{def}}{=}\|x_{\ell,p}-x^\star\|^2+2\gamma\big(\Phi(x_{\ell,p})-\Phi(x^\star)\big)\ \text{ and }\ W_{\ell,p}\overset{\text{def}}{=}8L\gamma^2\big(\Phi(\tilde{x}_\ell)-\Phi(x^\star)\big). \tag{A.7}$$

Equation (A.4) implies that

$$\mathbb{E}[V_{\ell,p}]\le V_{\ell,p-1}+W_{\ell,p-1},$$

and moreover $\sum_{\ell,p}W_{\ell,p}=\sum_\ell S_\ell<\infty$ holds almost surely. Therefore, the super martingale convergence theorem implies that $\{V_{\ell,p}\}_{p\in\{1,\cdots,P\},\ell\in\mathbb{N}}$ converges to a nonnegative random variable. Moreover, since $\Phi(x_{\ell,p})\to\Phi(x^\star)$, it follows that the sequence $\{\|x_{\ell,p}-x^\star\|\}_{p\in\{1,\cdots,P\},\ell\in\mathbb{N}}$ is convergent.

To prove the error rate (2.1), observe that by convexity of Φ and Jensen's inequality, we have

$$\mathbb{E}[S_{\ell+1}]\ge P\mathbb{E}\left[\Phi\left(\frac{1}{P}\sum_{j=1}^P x_{\ell,j}\right)-\Phi(x^\star)\right],$$

which further implies, owing to (A.6),

$$P\gamma^2\mathbb{E}\left[\Phi\left(\frac{1}{P}\sum_{j=1}^P x_{\ell,j}\right)-\Phi(x^\star)\right]\le\mathbb{E}[T_\ell]-\mathbb{E}[T_{\ell+1}].$$

Summing over $\ell=1,\ldots,Q$ and telescoping the right hand of the sum we arrive at

$$\begin{aligned}QP\gamma^2\mathbb{E}\Big[\Phi\big(\tfrac{1}{QP}\textstyle\sum_{\ell=1}^Q\sum_{j=1}^P x_{\ell,j}\big)-\Phi(x^\star)\Big]&\le QP\gamma^2\mathbb{E}\Big[\tfrac{1}{Q}\textstyle\sum_{\ell=1}^Q\Phi\left(\frac{1}{P}\sum_{j=1}^P x_{\ell,j}\right)-\Phi(x^\star)\Big]\\&\le\mathbb{E}[T_1]-\mathbb{E}[T_{Q+1}],\end{aligned}$$

where the first inequality follows from Jensen's inequality and convexity of Φ. Dividing both sides by $kP\gamma^2$ gives the required error bound. The convergence of $\varepsilon_k^{\text{SVRG}}$ is a straightforward consequence of the convergence of $x_{l,p}$.

Now we prove the second claim of the theorem. Taking expectation of both sides of (A.3) in Lemma A.2 and summing from $p=1,\cdots,P$ yields

$$\begin{aligned}&(1-\gamma\alpha_R)\mathbb{E}\big[\|x_{\ell,P}-x^\star\|^2\big]+2\gamma\mathbb{E}\big[\Phi(x_{\ell,P})-\Phi(x^\star)\big]\\&\le-(\alpha_F+\alpha_R)\textstyle\sum_{p=1}^P\mathbb{E}\big[\|x_{\ell,p}-x^\star\|^2\big]-\big(2\gamma-8\gamma^2L\big)\sum_{p=1}^{P-1}\mathbb{E}\big[\Phi(x_{\ell,p})-\Phi(x^\star)\big]\\&\quad+(1-\gamma\alpha_F)\mathbb{E}\big[\|x_{\ell,0}-x^\star\|^2\big]+8\gamma^2L\mathbb{E}\big[\Phi(x_{\ell,0})-\Phi(x^\star)\big]+8\gamma^2LP\mathbb{E}\big[\Phi(\tilde{x}_\ell)-\Phi(x^\star)\big].\end{aligned}$$

Since $\gamma L<\frac{1}{4(P+1)}<\frac{1}{4}$, we have $2\gamma-8\gamma^2L>0$, and we have from the above

$$(1-\gamma\alpha_R)\mathbb{E}\big[\|x_{\ell,P}-x^\star\|^2\big]+2\gamma\mathbb{E}\big[\Phi(x_{\ell,P})-\Phi(x^\star)\big]\le(1-\gamma\alpha_F)\mathbb{E}\big[\|x_{\ell,0}-x^\star\|^2\big]+8\gamma^2L(P+1)\mathbb{E}\big[\Phi(\tilde{x}_\ell)-\Phi(x^\star)\big].$$

Define

$$T_\ell\overset{\text{def}}{=}(1-\gamma\alpha_R)\mathbb{E}\big[\|x_{\ell,P}-x^\star\|^2\big]+2\gamma\mathbb{E}\big[\Phi(x_{\ell,P})-\Phi(x^\star)\big],$$

then there holds

$$\mathbb{E}[T_\ell]\le\max\left\{\frac{1-\gamma\alpha_F}{1+\gamma\alpha_R},4L\gamma(P+1)\right\}\mathbb{E}[T_{\ell-1}]$$

which implies the desired result. □

A.2 Proofs for Section 4

A.2.1 Riemannian Geometry

Let $\mathcal{M}$ be a C^2-smooth embedded submanifold of $\mathbb{R}^n$ around a point x. With some abuse of terminology, we shall state C^2-manifold instead of C^2-smooth embedded submanifold of $\mathbb{R}^n$. The natural embedding of a submanifold $\mathcal{M}$ into $\mathbb{R}^n$ permits to define a Riemannian structure and to introduce geodesics on $\mathcal{M}$, and we simply say $\mathcal{M}$ is a Riemannian manifold. We denote respectively $\mathcal{T}_{\mathcal{M}}(x)$ and $\mathcal{N}_{\mathcal{M}}(x)$ the tangent and normal space of $\mathcal{M}$ at point near x in $\mathcal{M}$.

Exponential map Geodesics generalize the concept of straight lines in $\mathbb{R}^n$, preserving the zero acceleration characteristic, to manifolds. Roughly speaking, a geodesic is locally the shortest path between two points on $\mathcal{M}$. We denote by $\mathfrak{g}(t;x,h)$ the value at $t\in\mathbb{R}$ of the geodesic starting at $\mathfrak{g}(0;x,h)=x\in\mathcal{M}$ with velocity $\dot{\mathfrak{g}}(t;x,h)=\frac{d\mathfrak{g}}{dt}(t;x,h)=h\in\mathcal{T}_{\mathcal{M}}(x)$ (which is uniquely defined). For every $h\in\mathcal{T}_{\mathcal{M}}(x)$, there exists an interval I around 0 and a unique geodesic $\mathfrak{g}(t;x,h):I\to\mathcal{M}$ such that $\mathfrak{g}(0;x,h)=x$ and $\dot{\mathfrak{g}}(0;x,h)=h$. The mapping

$$\mathrm{Exp}_x:\mathcal{T}_{\mathcal{M}}(x)\to\mathcal{M},\ \ h\mapsto\mathrm{Exp}_x(h)=\mathfrak{g}(1;x,h),$$

is called *Exponential map*. Given $x,x'\in\mathcal{M}$, the direction $h\in\mathcal{T}_{\mathcal{M}}(x)$ we are interested in is such that

$$\mathrm{Exp}_x(h)=x'=\mathfrak{g}(1;x,h).$$

Parallel translation Given two points $x,x'\in\mathcal{M}$, let $\mathcal{T}_{\mathcal{M}}(x),\mathcal{T}_{\mathcal{M}}(x')$ be their corresponding tangent spaces. Define

$$\tau:\mathcal{T}_{\mathcal{M}}(x)\to\mathcal{T}_{\mathcal{M}}(x'),$$

the parallel translation along the unique geodesic joining x to x', which is isomorphism and isometry w.r.t. the Riemannian metric.

Riemannian gradient and Hessian For a vector $v\in\mathcal{N}_{\mathcal{M}}(x)$, the Weingarten map of $\mathcal{M}$ at x is the operator $\mathfrak{W}_x(\cdot,v):\mathcal{T}_{\mathcal{M}}(x)\to\mathcal{T}_{\mathcal{M}}(x)$ defined by

$$\mathfrak{W}_x(\cdot,v)=-\mathrm{P}_{\mathcal{T}_{\mathcal{M}}(x)}\mathrm{d}V[h],$$

where V is any local extension of v to a normal vector field on $\mathcal{M}$. The definition is independent of the choice of the extension V, and $\mathfrak{W}_x(\cdot,v)$ is a symmetric linear operator which is closely tied to the second fundamental form of $\mathcal{M}$, see [8, Proposition II.2.1].

Let G be a real-valued function which is C^2 along the $\mathcal{M}$ around x. The covariant gradient of G at $x'\in\mathcal{M}$ is the vector $\nabla_{\mathcal{M}}G(x')\in\mathcal{T}_{\mathcal{M}}(x')$ defined by

$$\langle\nabla_{\mathcal{M}}G(x'),\,h\rangle=\frac{d}{dt}G\big(\mathrm{P}_{\mathcal{M}}(x'+th)\big)\big|_{t=0},\ \forall h\in\mathcal{T}_{\mathcal{M}}(x'),$$

where $\mathrm{P}_{\mathcal{M}}$ is the projection operator onto $\mathcal{M}$. The covariant Hessian of G at x' is the symmetric linear mapping $\nabla^2_{\mathcal{M}}G(x')$ from $\mathcal{T}_{\mathcal{M}}(x')$ to itself which is defined as

$$\langle\nabla^2_{\mathcal{M}}G(x')h,\,h\rangle=\frac{d^2}{dt^2}G\big(\mathrm{P}_{\mathcal{M}}(x'+th)\big)\big|_{t=0},\ \forall h\in\mathcal{T}_{\mathcal{M}}(x'). \tag{A.8}$$

This definition agrees with the usual definition using geodesics or connections [32]. Now assume that $\mathcal{M}$ is a Riemannian embedded submanifold of $\mathbb{R}^n$, and that a function G has a C^2-smooth restriction on $\mathcal{M}$. This can be characterized by the existence of a C^2-smooth extension (representative) of G, *i.e.* a C^2-smooth function $\widetilde{G}$ on $\mathbb{R}^n$ such that $\widetilde{G}$ agrees with G on $\mathcal{M}$. Thus, the Riemannian gradient $\nabla_{\mathcal{M}}G(x')$ is also given by

$$\nabla_{\mathcal{M}}G(x')=\mathrm{P}_{\mathcal{T}_{\mathcal{M}}(x')}\nabla\widetilde{G}(x'), \tag{A.9}$$

and $\forall h\in\mathcal{T}_{\mathcal{M}}(x')$, the Riemannian Hessian reads

$$\begin{aligned}\nabla^2_{\mathcal{M}}G(x')h&=\mathrm{P}_{\mathcal{T}_{\mathcal{M}}(x')}\mathrm{d}(\nabla_{\mathcal{M}}G)(x')[h]=\mathrm{P}_{\mathcal{T}_{\mathcal{M}}(x')}\mathrm{d}\big(x'\mapsto\mathrm{P}_{\mathcal{T}_{\mathcal{M}}(x')}\nabla_{\mathcal{M}}\widetilde{G}\big)[h]\\&=\mathrm{P}_{\mathcal{T}_{\mathcal{M}}(x')}\nabla^2\widetilde{G}(x')h+\mathfrak{W}_{x'}\big(h,\mathrm{P}_{\mathcal{N}_{\mathcal{M}}(x')}\nabla\widetilde{G}(x')\big),\end{aligned} \tag{A.10}$$

where the last equality comes from [2, Theorem 1]. When $\mathcal{M}$ is an affine or linear subspace of $\mathbb{R}^n$, then obviously $\mathcal{M}=x+\mathcal{T}_{\mathcal{M}}(x)$, and $\mathfrak{W}_{x'}(h,\mathrm{P}_{\mathcal{N}_{\mathcal{M}}(x')}\nabla\widetilde{G}(x'))=0$, hence (A.10) reduces to

$$\nabla^2_{\mathcal{M}}G(x')=\mathrm{P}_{\mathcal{T}_{\mathcal{M}}(x')}\nabla^2\widetilde{G}(x')\mathrm{P}_{\mathcal{T}_{\mathcal{M}}(x')}.$$

See [22, 8] for more materials on differential and Riemannian manifolds.

The following lemmas summarize two key properties that we will need throughout.

Lemma A.3 ([29, Lemma B.1]). *Let $x\in\mathcal{M}$, and x_k a sequence converging to x in $\mathcal{M}$. Denote $\tau_k:\mathcal{T}_{\mathcal{M}}(x)\to\mathcal{T}_{\mathcal{M}}(x_k)$ be the parallel translation along the unique geodesic joining x to x_k. Then, for any bounded vector $u\in\mathbb{R}^n$, we have*

$$(\tau_k^{-1}\mathrm{P}_{\mathcal{T}_{\mathcal{M}}(x_k)}-\mathrm{P}_{\mathcal{T}_{\mathcal{M}}(x)})u=o(\|u\|).$$

Lemma A.4 ([29, Lemma B.2]). *Let x, x' be two close points in $\mathcal{M}$, denote $\tau : \mathcal{T}_{\mathcal{M}}(x) \to \mathcal{T}_{\mathcal{M}}(x')$ the parallel translation along the unique geodesic joining x to x'. The Riemannian Taylor expansion of $\Phi \in C^2(\mathcal{M})$ around x reads,*

$$\tau^{-1}\nabla_{\mathcal{M}}\Phi(x') = \nabla_{\mathcal{M}}\Phi(x) + \nabla^2_{\mathcal{M}}\Phi(x)\mathrm{P}_{\mathcal{T}_{\mathcal{M}}(x)}(x'-x) + o(\|x'-x\|).$$

Lemma A.5 (Local normal sharpness [25, Proposition 2.10]). *If $R \in \mathrm{PSF}_x(\mathcal{M})$, then all $x' \in \mathcal{M}$ near x satisfy $\mathcal{T}_{\mathcal{M}}(x') = T_{x'}$. In particular, when $\mathcal{M}$ is affine or linear, then $T_{x'} = T_x$.*

Next we provide expressions of the Riemannian gradient and Hessian for the case of partly smooth functions relative to a C^2-smooth submanifold. This is summarized in the following proposition which follows by combining Eq. (A.9) and (A.10), Definition 3.1, Lemma A.5 and [10, Proposition 17] (or [32, Lemma 2.4]).

Lemma A.6 (Riemannian gradient and Hessian). *If $R \in \mathrm{PSF}_x(\mathcal{M})$, then for any $x' \in \mathcal{M}$ near x*

$$\nabla_{\mathcal{M}}R(x') = \mathrm{P}_{T_{x'}}(\partial R(x')).$$

For all $h \in T_{x'}$,

$$\nabla^2_{\mathcal{M}}R(x')h = \mathrm{P}_{T_{x'}}\nabla^2\widetilde{R}(x')h + \mathfrak{W}_{x'}\big(h, \mathrm{P}_{T_{x'}^{\perp}}\nabla\widetilde{R}(x')\big),$$

where $\widetilde{R}$ is a smooth representation of R on $\mathcal{M}$, and $\mathfrak{W}_x(\cdot,\cdot) : T_x \times T_x^{\perp} \to T_x$ is the Weingarten map of $\mathcal{M}$ at x.

A.2.2 Proofs

Proof of Proposition 4.8. By virtue the definition of proximity operator and the update of x_{k+1} in (1.4), we have

$$x_k - x_{k+1} - \gamma_k\big(\nabla F(x_k) - \nabla F(x_{k+1})\big) - \gamma_k\varepsilon_k \in \gamma_k\partial\Phi(x_{k+1}).$$

Given a global minimiser $x^\star \in \mathrm{Argmin}(\Phi)$, the classic optimality condition entails that

$$0 \in \gamma_k\partial\Phi(x^\star).$$

Projecting the above two inclusions on to $T_{x_{k+1}}$ and $T_{x^\star}$, respectively and using Lemma A.6, lead to

$$\begin{aligned}\gamma_k\tau_{k+1}^{-1}\nabla_{\mathcal{M}_{x^\star}}\Phi(x_{k+1}) &= \tau_{k+1}^{-1}\mathrm{P}_{T_{x_{k+1}}}\big(x_k - x_{k+1} - \gamma_k(\nabla F(x_k) - \nabla F(x_{k+1})) - \gamma_k\varepsilon_k\big)\\ \gamma_k\nabla_{\mathcal{M}_{x^\star}}\Phi(x^\star) &= 0.\end{aligned}$$

Adding both identities, and subtracting $\tau_{k+1}^{-1}\mathrm{P}_{T_{x_{k+1}}}x^\star$ on both sides, we get

$$\begin{aligned}&\tau_{k+1}^{-1}\mathrm{P}_{T_{x_{k+1}}}(x_{k+1} - x^\star) + \gamma_k\big(\tau_{k+1}^{-1}\nabla_{\mathcal{M}_{x^\star}}\Phi(x_{k+1}) - \nabla_{\mathcal{M}_{x^\star}}\Phi(x^\star)\big)\\ &= \tau_{k+1}^{-1}\mathrm{P}_{T_{x_{k+1}}}(x_k - x^\star) - \gamma_k\tau_{k+1}^{-1}\mathrm{P}_{T_{x_{k+1}}}\big(\nabla F(x_k) - \nabla F(x_{k+1})\big) - \gamma_k\tau_{k+1}^{-1}\mathrm{P}_{T_{x_{k+1}}}\varepsilon_k.\end{aligned} \tag{A.11}$$

For each term of (A.11), we have the following result

(i) By virtue of Lemma A.3, we get

$$\begin{aligned}\tau_{k+1}^{-1}\mathrm{P}_{T_{x_{k+1}}}(x_{k+1} - x^\star) &= \mathrm{P}_{T_{x^\star}}(x_{k+1} - x^\star) + (\tau_{k+1}^{-1}\mathrm{P}_{T_{x_{k+1}}} - \mathrm{P}_{T_{x^\star}})(x_{k+1} - x^\star)\\ &= \mathrm{P}_{T_{x^\star}}(x_{k+1} - x^\star) + o(\|x_{k+1} - x^\star\|).\end{aligned}$$

With the help of [27, Lemma 5.1], that $x_{k+1} - x^\star = \mathrm{P}_{T_{x^\star}}(x_{k+1} - x^\star) + o(\|x_{k+1} - x^\star\|)$, we further derive

$$\tau_{k+1}^{-1}\mathrm{P}_{T_{x_{k+1}}}(x_{k+1} - x^\star) = (x_{k+1} - x^\star) + o(\|x_{k+1} - x^\star\|). \tag{A.12}$$

Similarly for x_k, we have $\tau_{k+1}^{-1}\mathrm{P}_{T_{x_{k+1}}}(x_k - x^\star) = (x_k - x^\star) + o(\|x_k - x^\star\|)$.

(ii) Owing to Lemma A.4, we have for $\tau_{k+1}^{-1}\nabla_{\mathcal{M}_{x^\star}}\Phi(x_{k+1}) - \nabla_{\mathcal{M}_{x^\star}}\Phi(x^\star)$,

$$\tau_{k+1}^{-1}\nabla_{\mathcal{M}_{x^\star}}\Phi(x_{k+1}) - \nabla_{\mathcal{M}_{x^\star}}\Phi(x^\star) = \nabla^2_{\mathcal{M}_{x^\star}}\Phi(x^\star)\mathrm{P}_{T_{x^\star}}(x_{k+1} - x^\star) + o(\|x_{k+1} - x^\star\|). \tag{A.13}$$

(iii) Using Lemma A.4 again together with the local C^2-smoothness of F, we have

$$\begin{aligned}
&\tau_{k+1}^{-1}\mathrm{P}_{T_{x_{k+1}}}\big(\nabla F(x_k)-\nabla F(x_{k+1})\big)\\
&=\mathrm{P}_{T_{x^\star}}\big(\nabla F(x_k)-\nabla F(x_{k+1})\big)+o(\|\nabla F(x_k)-\nabla F(x_{k+1})\|)\\
&=\mathrm{P}_{T_{x^\star}}\big((\nabla F(x_k)-\nabla F(x^\star))-(\nabla F(x_{k+1})-\nabla F(x^\star))\big)+o(\|\nabla F(x_k)-\nabla F(x^\star)\|+\|\nabla F(x_{k+1})-\nabla F(x^\star)\|)\\
&=\mathrm{P}_{T_{x^\star}}\nabla^2F(x^\star)\mathrm{P}_{T_{x^\star}}(x_k-x^\star)-\mathrm{P}_{T_{x^\star}}\nabla^2F(x^\star)\mathrm{P}_{T_{x^\star}}(x_{k+1}-x^\star)+o(\|x_k-x^\star\|)+o(\|x_{k+1}-x^\star\|).
\end{aligned} \tag{A.14}$$

(iv) Owing to Lemma A.3, we have $\tau_{k+1}^{-1}\mathrm{P}_{T_{x_{k+1}}}\varepsilon_k=\mathrm{P}_{T_{x^\star}}\varepsilon_k+o(\|\varepsilon_k\|)$.
Combining the above relations with (A.11) we obtain

$$\begin{aligned}
&\big(\mathrm{Id}+\gamma_k\nabla^2_{\mathcal{M}_{x^\star}}\Phi(x^\star)\mathrm{P}_{T_{x^\star}}-\gamma_k\mathrm{P}_{T_{x^\star}}\nabla^2F(x^\star)\mathrm{P}_{T_{x^\star}}\big)(x_{k+1}-x^\star)+o(\|x_{k+1}-x^\star\|)\\
&=(\mathrm{Id}+\gamma_kH_R)(x_{k+1}-x^\star)+o(\|x_{k+1}-x^\star\|)\\
&=(\mathrm{Id}+\gamma H_R)(x_{k+1}-x^\star)+(\gamma_k-\gamma)H_R(x_{k+1}-x^\star)+o(\|x_{k+1}-x^\star\|)\\
&=(x_k-x^\star)-\gamma_kH_F(x_k-x^\star)+o(\|x_k-x^\star\|)-(\gamma_k\mathrm{P}_{T_{x^\star}}\varepsilon_k+o(\|\varepsilon_k\|))\\
&=(x_k-x^\star)-\gamma H_F(x_k-x^\star)-(\gamma_k-\gamma)H_F(x_k-x^\star)+o(\|x_k-x^\star\|)-(\gamma\mathrm{P}_{T_{x^\star}}\varepsilon_k+o(\|\varepsilon_k\|))-(\gamma_k-\gamma)\mathrm{P}_{T_{x^\star}}\varepsilon_k.
\end{aligned} \tag{A.15}$$

Since we have $\gamma_k\to\gamma$ and H_R is bounded, we have

$$\lim_{k\to+\infty}\frac{\|(\gamma_k-\gamma)H_R(x_{k+1}-x^\star)\|}{\|x_{k+1}-x^\star\|}\leq\lim_{k\to+\infty}\frac{|\gamma_k-\gamma|\|H_R\|\|x_{k+1}-x^\star\|}{\|x_{k+1}-x^\star\|}=0,$$

hence $(\gamma_k-\gamma)H_R(x_{k+1}-x^\star)=o(\|x_{k+1}-x^\star\|)$. Using the same arguments lead to $(\gamma_k-\gamma)H_F(x_k-x^\star)=o(\|x_k-x^\star\|)$ and $(\gamma_k-\gamma)\mathrm{P}_{T_{x^\star}}\varepsilon_k=o(\|\varepsilon_k\|)$. Therefore, from (A.15), we obtain

$$(\mathrm{Id}+\gamma H_R)(x_{k+1}-x^\star)+o(\|x_{k+1}-x^\star\|)=G_F(x_k-x^\star)+o(\|x_k-x^\star\|)-(\gamma\mathrm{P}_{T_{x^\star}}\varepsilon_k+o(\|\varepsilon_k\|)). \tag{A.16}$$

Inverting $\mathrm{Id}+\gamma H_R$ (which is possible owing to Lemma 4.7), we obtain

$$\begin{aligned}
x_{k+1}-x^\star+W_Ro(\|x_{k+1}-x^\star\|)&=W_RG_F(x_k-x^\star)+W_Ro(\|x_k-x^\star\|)-W_R(\gamma\mathrm{P}_{T_{x^\star}}\varepsilon_k+o(\|\varepsilon_k\|))\\
&=M_{\mathrm{FB}}(x_k-x^\star)+W_Ro(\|x_k-x^\star\|)-W_R(\gamma\mathrm{P}_{T_{x^\star}}\varepsilon_k+o(\|\varepsilon_k\|)).
\end{aligned} \tag{A.17}$$

Since W_R, G_F are non-expansive, we have

$$x_{k+1}-x^\star+W_Ro(\|x_{k+1}-x^\star\|)=x_{k+1}-x^\star+o(\|x_{k+1}-x^\star\|)=d_{k+1},$$

and similarly, we have $M_{\mathrm{FB}}(x_k-x^\star)+W_Ro(\|x_k-x^\star\|)=M_{\mathrm{FB}}(x_k-x^\star)+o(\|x_k-x^\star\|)=M_{\mathrm{FB}}d_k$ and $\gamma W_R\mathrm{P}_{T_{x^\star}}\varepsilon_k+W_Ro(\|\varepsilon_k\|)=\gamma W_R\mathrm{P}_{T_{x^\star}}\varepsilon_k+o(\|\varepsilon_k\|)=\phi_k$. Substituting back into (A.17) we conclude the proof. □

References

[1] P-A. Absil, R. Mahony, and R. Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.

[2] P-A. Absil, R. Mahony, and J. Trumpf. An extrinsic look at the Riemannian Hessian. In *Geometric Science of Information*, pages 361–368. Springer, 2013.

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

[4] D. Blatt, A. O. Hero, and H. Gauchman. A convergent incremental gradient method with a constant step size. *SIAM Journal on Optimization*, 18(1):29–51, 2007.

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

[6] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

[7] A. Chambolle and C. Dossal. On the convergence of the iterates of the "fast iterative shrinkage/thresholding algorithm". *Journal of Optimization Theory and Applications*, 166(3):968–982, 2015.

[8] I. Chavel. *Riemannian geometry: a modern introduction*, volume 98. Cambridge University Press, 2006.

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

[10] A. Daniilidis, W. Hare, and J. Malick. Geometrical interpretation of the predictor-corrector type algorithms in structured optimization problems. *Optimization: A Journal of Mathematical Programming & Operations Research*, 55(5-6):482–503, 2009.

[11] A. Defazio. *New Optimisation Methods for Machine Learning*. PhD thesis, Australian National University, 2014. http://www.aarondefazio.com/pubs.html.

[12] A. Defazio, F. Bach, and S. Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems*, pages 1646–1654, 2014.

[13] D. L. Donoho, M. Elad, and V. N. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Trans. Inform. Theory*, 52(1):6–18, 2006.

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

[15] J. Fadili, J. Malick, and G. Peyré. Sensitivity analysis for mirror-stratifiable convex functions. *arXiv preprint arXiv:1707.03194*, 2017.

[16] P. Gong and J. Ye. Linear convergence of variance-reduced stochastic gradient without strong convexity. *arXiv preprint arXiv:1406.1102*, 2014.

[17] W. L. Hare and A. S. Lewis. Identifying active constraints via partial smoothness and prox-regularity. *Journal of Convex Analysis*, 11(2):251–266, 2004.

[18] W. L. Hare and A. S. Lewis. Identifying active manifolds. *Algorithmic Operations Research*, 2(2):75–82, 2007.

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

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

[21] N. Le Roux, M. Schmidt, and F. R. Bach. A stochastic gradient method with an exponential convergence _rate for finite training sets. In *Advances in Neural Information Processing Systems*, pages 2663–2671, 2012.

[22] J. M. Lee. *Smooth manifolds*. Springer, 2003.

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

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

[25] A. S. Lewis. Active sets, nonsmoothness, and sensitivity. *SIAM Journal on Optimization*, 13(3):702–725, 2003.

[26] A. S. Lewis and S. Zhang. Partial smoothness, tilt stability, and generalized Hessians. *SIAM Journal on Optimization*, 23(1):74–94, 2013.

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

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

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

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

[31] D. A. Lorenz and T. Pock. An inertial forward-backward algorithm for monotone inclusions. *Journal of Mathematical Imaging and Vision*, 51(2):311–325, 2015.

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

[33] C. Molinari, J. Liang, and J. Fadili. Convergence rates of forward–douglas–rachford splitting method. *arXiv preprint arXiv:1801.01088*, 2018.

[34] A. Moudafi and M. Oliny. Convergence of a splitting inertial proximal method for monotone operators. *Journal of Computational and Applied Mathematics*, 155(2):447–454, 2003.

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

[36] J. Neveu. *Discrete-parameter martingales*, volume 10. Elsevier, 1975.

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

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

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

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

[41] N. D. Vanli, M. Gurbuzbalaban, and A. Ozdaglar. Global convergence rate of proximal incremental aggregated gradient methods. *arXiv preprint arXiv:1608.01713*, 2016.

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

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

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