Inexact proximal stochastic second-order methods for nonconvex composite optimization *

Xiao Wang † Hongchao Zhang ‡

Abstract

In this paper, we propose a framework of Inexact Proximal Stochastic Second-order (IPSS) methods for solving nonconvex optimization problems, whose objective function consists of an average of finitely many, possibly weakly, smooth functions and a convex but possibly nonsmooth function. At each iteration, IPSS inexactly solves a proximal subproblem constructed by using some positive definite matrix which could capture the second-order information of original problem. Proper tolerances are given for the subproblem solution in order to maintain global convergence and the desired overall complexity of the algorithm. Under mild conditions, we analyze the computational complexity related to the evaluations on the component gradient of the smooth function. We also investigate the number of evaluations of subgradient when using an iterative subgradient method to solve the subproblem. In addition, based on IPSS, we propose a linearly convergent algorithm under the proximal Polyak-Łojasiewicz condition. Finally, we extend the analysis to problems with weakly smooth function and obtain the computational complexity accordingly.

Keywords: Stochastic gradient, second-order approximation, Polyak-Łojasiewicz (PL) inequality, inexact subproblem solution, (weakly) smooth function, variance reduction, complexity, nonconvex

Mathematics Subject Classification 2010: 49M37, 65K05, 90C30

1 Introduction

In this paper, we consider the following nonconvex optimization problem:

$$\min_{x\in\mathbb{R}^d} \quad F(x) = f(x) + h(x). \tag{1.1}$$

Here, $h : \mathbb{R}^d \to \mathbb{R}$ is a proper, convex function but possibly nonsmooth, function and f is an average of many, but a finite number of, smooth component functions, i.e.,

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x), \tag{1.2}$$

where $f_i : \mathbb{R}^d \to \mathbb{R}$, $i = 1, \ldots, n$, is first-order continuously differentiable, but possibly nonconvex. We will also consider the case that the component functions are only weakly smooth. We assume that the optimal objective function value F^* of (1.1) is finite and the number n of component functions is

*May 15, 2019. This research was partially supported by the National Natural Science Foundation of China (11871453, 11731013) and by the Young Elite Scientists Sponsorship Program by China Association for Science and Technology, and by the USA National Science Foundation (1522654, 1819161).

†wangxiao@ucas.ac.cn, School of Mathematical Sciences, University of Chinese Academy of Sciences, No.19A Yuquan Road, Beijing 100049, China.

‡hozhang@math.lsu.edu, http://www.math.lsu.edu/∼hozhang, Department of Mathematics, Louisiana State University, Baton Rouge 70803, United States.

very large or even huge such that it is very expensive or even impossible to evaluate the gradient of f at each iteration. To deal with this challenge, stochastic methods, utilizing stochastic gradient to approximate the exact gradient of f, have attracted more and more attention. There are mainly two type of stochastic methods for solving (1.1): proximal stochastic first-order methods and proximal stochastic second-order methods.

Proximal stochastic first-order methods, also known as proximal stochastic gradient methods, normally solve the following type of subproblem based on stochastic gradient of f at each iteration:

$$\min_{x\in\mathbb{R}^d} \quad \langle g_k, x - x_k\rangle + \frac{1}{2}\|x - x_k\|_2^2 + h(x)$$

where $g_k \in \mathbb{R}^d$ is a stochastic gradient of f at x_k. Xiao and Zhang [29] proposed a proximal SVRG (Prox-SVRG) method for solving (1.1)-(1.2). Their method has $\mathcal{O}(n\log(1/\epsilon))$ component gradient complexity when f is strongly convex and $(n + 1/\varepsilon)\mathcal{O}(\log(1/\epsilon))$ complexity when f is only convex. Prox-SVRG extends the original SVRG method proposed for solving smooth problem by Johnson and Zhang [11], where an unbiased stochastic gradient is constructed by combining the the exact gradient of f at certain particularly chosen points. Later, [23] extends proximal SVRG method to solve nonconvex problems, with complexity $\mathcal{O}(n + n^{2/3}/\epsilon)$. Other related works include Prox-SAGA [23] and prox-SARAH [21], RSPG [25]. However, all the above algorithms need solving the subproblems exactly at each iteration. Wang *et al.* [27] proposed an inexact proximal stochastic gradient (IPSG) method for convex composite optimization. At each iteration, the subproblem is solved up to some pre-given tolerance. Although the subproblems are solved inexactly, IPSG can still keep global convergence with desirable computational complexity. Inspired by this work, in this paper we propose an algorithm, which not only allows to solve the proximal subproblem inexactly but also has the ability to explore the second-order information of the objective function of (1.1). Due to the difference of proximal subproblems, we will generalize the original definition of inexact subproblem solutions proposed in [27]. And more importantly, our method proposed in this paper is designed for nonconvex composite optimization.

Proximal Second-order Method (PSM) is another type of popular methods for solving (1.1). At each iteration of PSM, the following type of subproblem is solved to update its iterates:

$$\min_{x\in\mathbb{R}^d} \quad q_k(x) := \langle g_k, x - x_k\rangle + \frac{1}{2}\|x - x_k\|_{B_k}^2 + h(x), \tag{1.3}$$

where $B_k \in \mathbb{R}^{d\times d}$ is a positive definite matrix often carrying some second-order information of f at x_k and $g_k \in \mathbb{R}^d$ is still a stochastic gradient of f at x_k. Literatures on deterministic proximal second-order methods for solving (1.1) include [2,8,17,18]. For particular h with simple structure, e.g. $h(x) = \lambda\|x\|_1$, several proximal second-order methods were proposed to explore the problem structure, such as [4,5,7,9,12]. However, all these previous work focuses on (1.1) when its objective is convex or strongly convex. Recently, exploring the summation structure of f as (1.2), Wang *et al.* [28] proposed a proximal stochastic quasi-Newton method for (1.1)-(1.2), where the subproblem (1.3) was solved using $B_k = (1/\eta_k)B$, but with a fixed B, at each iteration. With proper choice of stepsize parameter η_k, Wang *et al.* [28] proved the theoretical convergence as well as the computational complexity of their proposed algorithm. However, each subproblem in [28] is also required be solved exactly, which in many scenarios will be very time consuming or even theoretically impossible, for example, when h is a general nonsmooth function and/or B is a complicated matrix such that there is no closed-form subproblem solution. In this paper, we will extend this algorithm to an inexact proximal stochastic second-order algorithm which allows to solve the subproblems inexactly and has more flexibilities for choosing the matrix B_k.

In the literature, deterministic algorithms have been also proposed to solve problem (1.1) under the assumption that $\nabla f : \mathbb{R}^d \to \mathbb{R}^d$ is Hölder continuous, i.e. there exist $\gamma > 0$ and $\nu \in (0, 1]$ such that $\|\nabla f(x) - \nabla f(y)\| \le \frac{1}{\gamma}\|x - y\|_*^\nu$ for any x, y, where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$, such as [1, 6, 10, 11, 16, 19, 20, 26]. When $\nu < 1$, f is usually called weakly smooth. Recently, Ghadimi [24] proposes a conditional gradient method for solving (1.1) under stochastic setting, where the information of f is obtained through a stochastic oracle. However, the summation structure of f

as (1.2) is not explored in [24]. Our IPSS algorithm proposed in this paper is based on stochastic variance reduced gradient (SVRG) [11] by taking advantage of the summation structure of f. We will also study the theoretical properties of IPSS for solving (1.1)-(1.2) when f is only weakly smooth.

Contributions

This paper proposes a framework of inexact proximal stochastic second-order (IPSS) algorithm for solving general nonconvex composite optimization problems. Moreover, IPSS allows inexact solution for each proximal subproblem and also allows to explore the second-order information of the objective function. Wang *et al.* [27] first studied inexact proximal stochastic gradient method for solving convex and strongly convex composite optimization by inexactly solving subproblems. But the algorithm proposed in this paper differs from [27] in generalizing the definition of inexact subproblem solution, since the subproblem of IPSS is in a more general setting for possibly using the second-order information. More importantly, the new algorithm IPSS is designed for solving nonconvex composite optimization (1.1) with function f being smooth or weakly smooth settings, while the algorithm in [27] can only be applied for solving convex composite optimization.

We explore the theoretical properties of IPSS for solving (1.1) under different smooth assumptions of the objective function. Although Wang *et al.* [28] studied proximal stochastic quasi-Newton methods for nonconvex composite optimization, this paper differs and generalizes the analysis in [28] in the following aspects. Firstly, IPSS allows to solve the subproblems inexactly which is often critical when the subproblem becomes more difficult in the second-order methods. Secondly, the smoothness of function f is more generally defined with respect to a symmetric positive definite matrix M (See assumptions A1 and A3) and more general theoretical requirements on second-order approximation matrix B are derived connected with this matrix M. Thirdly, for analyzing IPSS in this paper, we propose a new operator $\mathcal{G}$ (defined in (2.11)) to measure the first-order optimality of the optimization problem. This operator $\mathcal{G}$ also relies on the matrix M, which inherits the smooth information of f. In addition, an iterative subgradient method is proposed to solve subproblems to the required accuracy at each iteration and the complexity with respect to evaluations of subgradients of function h is analyzed. Furthermore, based on IPSS, we propose a linearly convergent algorithm when the objective function satisfies the so-called proximal Polyak-Łojasiewicz (PPL) inequality. Finally, we explore theoretical properties of IPSS for solving optimization problems (1.1) when f is only weakly smooth. Our analysis shows that under mild conditions the theoretical properties of IPSS for (1.1) with smooth f can be properly extended to the case when f is weakly smooth.

Notations and Organizations

The gradient of f at a point x is denoted by $\nabla f(x)$ and $\partial f(x)$ represents the subdifferential of f at x when f is a proper convex function. Given $x, y \in \mathbb{R}^d$, $\langle x, y\rangle = x^{\mathsf{T}} y$ is the standard Euclidean inner product in $\mathbb{R}^n$. S^d_{++} denotes the set of all $d \times d$ symmetric positive definite matrices. For two matrices $A, B \in \mathbb{R}^{d\times d}$, $A \succeq B$ means $A - B$ is positive semidefinite. Given $B \in S^d_{++}$ and $x \in \mathbb{R}^d$, $\|x\|^2_B$ is defined to be $x^{\mathsf{T}} B x$. Given a real number a, $\lfloor a \rfloor$ means the largest integer less than or equal to a. For a random variable or vector X, its expectation is denoted as $\mathbb{E}[X]$, while its expectation conditioned on another random variable or vector Y is denoted as $\mathbb{E}[X|Y]$.

The remainder of this paper is organized as follows. In Section 2, we propose the framework of an inexact proximal stochastic second-order (IPSS) algorithm and give some preliminary results and backgrounds. The Section 3 is divided to three subsections. In Subsection 3.1, we discuss the convergence properties of IPSS in details for solving (1.1) with smooth f. In Subsection 3.2, we propose a particular subgradient method to solve the subproblems of IPSS inexactly and discuss the overall algorithm complexity. A linearly convergent algorithm, called PPL-IPSS, is proposed in Subsection 3.3 to solve (1.1) under the proximal PL-inequality. In Section 4, we analyze the convergence properties of IPSS for solving (1.1) when f in the objective function is only weakly smooth. We finally summarizes the paper in the last Section 5.

2 Framework of IPSS algorithm

In this section, we propose a framework of inexact proximal stochastic second-order (IPSS) algorithm to solve (1.1)-(1.2). The key subproblem to solve in each iteration is the following type of proximal subproblem

$$\min_{x\in\mathbb{R}^d} q(x) := h(x) + \frac{1}{2}\|x-w\|_B^2. \tag{2.1}$$

Here, $B \in S_{++}^d$ could be used to capture certain second-order curvature information of f. Due to the introduction of B matrix and the existence of possibly nonsmooth function h, it might be too expensive or impossible to find the exact solution of (2.1). Under this circumstance, inexact solutions of (2.1) will be necessary in practice to solve the problem (1.1)-(1.2). In this paper, we will propose proper criteria for inexact solutions of subproblem (2.1) while still maintain global convergence of the algorithm with the desired computational complexity.

Before defining the inexact solution of (2.1), we first recall the concept of the ε-subdifferential of a convex function [3].

DEFINITION **2.1.** *Given a convex function $\phi : \mathbb{R}^d \to \mathbb{R}$, its ε-subdifferential at x, denoted as $\partial_\varepsilon\phi(x)$, is defined as*

$$\partial_\varepsilon\phi(x) = \{z : \phi(y) - \phi(x) \geq \langle z, y-x\rangle - \varepsilon \text{ for all } y \in \mathbb{R}^d\}.$$

We now give our definition of inexact solution of (2.1). Notice that this definition is a generalization of that in [27], since the definition of inexact solution in [27] simply corresponds to the case that B is an identify matrix.

DEFINITION **2.2.** *Given $\bar{\varepsilon} > 0$ and $\hat{\varepsilon} > 0$, we call z to be an $(\bar{\varepsilon}, \hat{\varepsilon})$-solution of the problem (2.1), if there exists $u \in \mathbb{R}^d$ such that*

$$\|u\|_B \leq \sqrt{2\bar{\varepsilon}} \quad \text{and} \quad B(w-z-u) \in \partial_{\hat{\varepsilon}} h(z).$$

According to Definition 2.1, we can have the following lemma.

LEMMA **2.1.** *Let $l(x) := \frac{1}{2}\|x-w\|_B^2$ where $B \in S_{++}^d$. Given $\varepsilon > 0$,*

$$\partial_\varepsilon l(x) = \left\{z : z = B(x-w+u) \text{ with } \tfrac{1}{2}\|u\|_B^2 \leq \varepsilon\right\}.$$

Proof. Following from Definition 2.1, we have that $z \in \partial_\varepsilon l(x)$ if and only if

$$\varepsilon \geq \frac{1}{2}\|x-w\|_B^2 - \frac{1}{2}\|w-y\|_B^2 + \langle z, y-x\rangle, \quad \forall y \in \mathbb{R}^d. \tag{2.2}$$

Notice that

$$\begin{aligned}\frac{1}{2}\|w-y\|_B^2 - \langle z, y-x\rangle &= \frac{1}{2}\|y\|_B^2 - \langle y, Bw+z\rangle + \frac{1}{2}\|w\|_B^2 + \langle z, x\rangle \\ &= \frac{1}{2}\|y-(w+B^{-1}z)\|_B^2 - \frac{1}{2}\|w+B^{-1}z\|_B^2 + \frac{1}{2}\|w\|_B^2 + \langle z, x\rangle.\end{aligned}$$

Then, (2.2) is equivalent to

$$\varepsilon \geq \frac{1}{2}\|x-(w+B^{-1}z)\|_B^2 - \frac{1}{2}\|y-(w+B^{-1}z)\|_B^2, \quad \forall y \in \mathbb{R}^d.$$

Hence, we obtain

$$\partial_\varepsilon l(x) = \left\{z : \frac{1}{2}\|x-(w+B^{-1}z)\|_B^2 \leq \varepsilon\right\} = \left\{z : z = B(x-w+u) \text{ with } \frac{1}{2}\|u\|_B^2 \leq \varepsilon\right\}.$$

□

Algorithm 2.1 **IPSS**$(\tilde{x}_0, B_1^1, \bar{\varepsilon}, \hat{\varepsilon}, S, m)$

Input: Maximum outer iteration number S and inner iteration number m, batch sizes $\{b_k\}$, inexactness tolerances $\bar{\varepsilon} = \{\bar{\varepsilon}_t^k\}$, $\hat{\varepsilon} = \{\hat{\varepsilon}_t^k\}$, initial iterate $\tilde{x}_0 \in \mathbb{R}^d$ and $B_1^1 \in S_{++}^d$; Randomly generate a vector $R = (R_k, R_t)$ according to certain probabilistic distribution P.

1: **for** $k = 1, \ldots, S$ **do**
2: Set $x_0^k = \tilde{x}_{k-1}$.
3: Calculate $\tilde{v} = \nabla f(\tilde{x}_{k-1})$.
4: **for** $t = 1, 2, \ldots, m$ **do**
5: If $(k, t) = (R_k, R_t)$, stop the algorithm and return $x_R = x_{t-1}^k$.
6: Randomly choose a sample set $\mathcal{K} \subset [1, 2, \ldots, n]$ with size b_k, such that the probability of each index being picked is b_k/n.
7: Calculate

$$v_{t-1}^k = \nabla f_{\mathcal{K}}(x_{t-1}^k) - \nabla f_{\mathcal{K}}(\tilde{x}_{k-1}) + \tilde{v}$$

with $\nabla f_{\mathcal{K}}(\cdot) = \frac{1}{b_k} \sum_{i \in \mathcal{K}} \nabla f_i(\cdot)$.
8: Compute $(\bar{\varepsilon}_t^k, \hat{\varepsilon}_t^k)$-solution x_t^k of subproblem

$$\min_{x \in \mathbb{R}^d} q_t^k(x) = \langle v_{t-1}^k, x \rangle + \frac{1}{2}\|x - x_{t-1}^k\|_{B_t^k}^2 + h(x). \tag{2.3}$$

9: Generate $B_{t+1}^k \in S_{++}^d$.
10: **end for**
11: Set $\tilde{x}_k = x_m^k$.
12: **end for**

Output: Return x_R.

By Definition 2.2, given $\bar{\varepsilon} > 0$ and $\hat{\varepsilon} > 0$, if z is an $(\bar{\varepsilon}, \hat{\varepsilon})$-solution of (2.1), it follows from Lemma 2.1 that

$$0 \in \partial_{\bar{\varepsilon}} l(z) + \partial_{\hat{\varepsilon}} h(z) \subseteq \partial_{\bar{\varepsilon}+\hat{\varepsilon}} q(z)$$

which implies that for any $x \in \mathbb{R}^d$ we have

$$q(x) - q(z) \geq \langle 0, x - z \rangle - (\bar{\varepsilon} + \hat{\varepsilon}),$$

or equivalently, for any $x \in \mathbb{R}^d$

$$q(z) \leq q(x) + (\bar{\varepsilon} + \hat{\varepsilon}). \tag{2.4}$$

Now, supposing $R = (R_k, R_t)$ is a random vector supported on $\{(k, t) : k = 1, \ldots, S \text{ and } t = 1, \ldots, m\}$, our IPSS algorithm is presented in Algorithm IPSS. One difference of IPSS from the algorithm proposed by Wang *et al.* [28] is that the subproblem (2.3) is designed without introducing another stepsize parameter η_k. The second difference is that the subproblem (2.3) is allowed to be solved inexactly, which not only saves the computation for finding subproblem solution but also gives more flexibility of choosing the matrix B_t^k. To ensure global convergence, we only require B_t^k satisfies certain upper and lower bounds (See Assumptions A2 and A4). By specifying proper probabilistic distribution P for returning the output iterate, we also analyze the theoretical performance of IPSS when the objective function is only weakly smooth. The detailed analysis is given in Section 4. We want to mention that similar sampling strategy for choosing sample sets $\mathcal{K}$ in Step 5 has been also considered in the literature, such as Wang *et al.* [27] and Zhang *et al.* [31]. According to the computation of the stochastic gradient v_{t-1}^k in Step 6, one can obtain that

$$\mathbb{E}[v_{t-1}^k | x_{t-1}^k] = \nabla f(x_{t-1}^k), \quad \mathbb{E}[\|v_{t-1}^k - \nabla f(x_{t-1}^k)\|^2 | x_{t-1}^k] \leq \frac{1}{b_k} \mathbb{E}[\|\nabla f_i(x_{t-1}^k) - \nabla f_i(\tilde{x}_{k-1}) | x_{t-1}^k\|^2].$$

(Interested readers are referred to Lemma 2.1 in [27] for more details.) Moreover, since x_t^k is an $(\bar{\varepsilon}_t^k, \hat{\varepsilon}_t^k)$-solution of subproblem (2.3), by Definition 2.2 there exists an u_t^k such that

$$\|u_t^k\|_{B_t^k} \leq \sqrt{2\bar{\varepsilon}_t^k} \quad \text{and} \quad B_t^k(x_{t-1}^k - x_t^k - u_t^k) - v_{t-1}^k \in \partial_{\hat{\varepsilon}_t^k} h(x_t^k).$$

Let

$$\varepsilon_t^k = \bar{\varepsilon}_t^k + \hat{\varepsilon}_t^k. \tag{2.5}$$

Then, (2.4) indicates

$$q_t^k(x_t^k) \le \min q_t^k(x) + \varepsilon_t^k. \tag{2.6}$$

For the following analysis, let us define $\bar{x}_t^k$ be the exact solution of proximal full gradient step, namely,

$$\bar{x}_t^k = \arg\min_{y\in\mathbb{R}^n} \langle \nabla f(x_{t-1}^k), x\rangle + \frac{1}{2}\|x - x_{t-1}^k\|_{B_t^k}^2 + h(x). \tag{2.7}$$

Then, the following lemma shows the relationship between x_t^k and $\bar{x}_t^k$. The proof is a generalization of that for Lemma 2.2 in [27].

LEMMA **2.2.** *Let $\bar{x}_t^k$ be the exact solution of* (2.7)*, i.e., the exact solution of* (2.3) *with v_{t-1}^k replaced by $\nabla f(x_{t-1}^k)$. Then*

$$\|x_t^k - \bar{x}_t^k\|_{B_t^k} \le \sqrt{2\varepsilon_t^k} + \|v_{t-1}^k - \nabla f(x_{t-1}^k)\|_{(B_t^k)^{-1}}. \tag{2.8}$$

Proof. Let $\hat{x}_t^k$ be the exact solution of (2.3). Then due to the strong convexity of q_t^k, we have

$$q_t^k(x_t^k) - q_t^k(\hat{x}_t^k) \ge \frac{1}{2}\|x_t^k - \hat{x}_t^k\|_{B_t^k}^2$$

which together with (2.6) yields that

$$\|x_t^k - \hat{x}_t^k\|_{B_t^k} \le \sqrt{2\varepsilon_t^k}. \tag{2.9}$$

Note that $\bar{x}_t^k$ and $\hat{x}_t^k$ satisfy

$$B_t^k(x_{t-1}^k - \bar{x}_t^k) - \nabla f(x_{t-1}^k) \in \partial h(\bar{x}_t^k) \quad \text{and} \quad B_t^k(x_{t-1}^k - \hat{x}_t^k) - v_{t-1}^k \in \partial h(\hat{x}_t^k),$$

which imply

$$h(\hat{x}_t^k) - h(\bar{x}_t^k) \ge \left\langle B_t^k(x_{t-1}^k - \bar{x}_t^k) - \nabla f(x_{t-1}^k), \hat{x}_t^k - \bar{x}_t^k \right\rangle$$

and

$$h(\bar{x}_t^k) - h(\hat{x}_t^k) \ge \left\langle B_t^k(x_{t-1}^k - \hat{x}_t^k) - v_{t-1}^k, \bar{x}_t^k - \hat{x}_t^k \right\rangle.$$

Summing up the above two inequalities yields

$$0 \ge \langle \nabla f(x_{t-1}^k) - v_{t-1}^k, \bar{x}_t^k - \hat{x}_t^k\rangle + \langle B_t^k(\bar{x}_t^k - \hat{x}_t^k), \bar{x}_t^k - \hat{x}_t^k\rangle.$$

Then if follows that

$$\|\bar{x}_t^k - \hat{x}_t^k\|_{B_t^k}^2 \le \langle v_{t-1}^k - \nabla f(x_{t-1}^k), \bar{x}_t^k - \hat{x}_t^k\rangle \le \|\bar{x}_t^k - \hat{x}_t^k\|_{B_t^k}\|v_{t-1}^k - \nabla f(x_{t-1}^k)\|_{(B_t^k)^{-1}},$$

which indicates

$$\|\bar{x}_t^k - \hat{x}_t^k\|_{B_t^k} \le \|v_{t-1}^k - \nabla f(x_{t-1}^k)\|_{(B_t^k)^{-1}}. \tag{2.10}$$

Therefore, we obtain (2.8) by adding (2.9) and (2.10). □

To characterize the computational complexity of IPSS in later analysis, we define the operator $\mathcal{G}_h(\cdot, g, B)$ by

$$\mathcal{G}_h(x, g, B) = B(x - \bar{x}), \tag{2.11}$$

where

$$\bar{x} = \arg\min_{y\in\mathbb{R}^d} \langle g, y - x\rangle + \frac{1}{2}\|y - x\|_B^2 + h(y).$$

For the operator $\mathcal{G}_h$, we first give an important lemma.

LEMMA **2.3.** *For any given $B \in S^d_{++}$, x^* is a stationary point of $f+h$ if and only if*

$$\mathcal{G}_h(x^*, \nabla f(x^*), B) = 0.$$

Moreover, when h vanishes,

$$\mathcal{G}_0(x, g, B) = g, \quad \forall g \in \mathbb{R}^d.$$

Proof. On the one hand, if x^* is a stationary point of $f+h$, there exists $p^* \in \partial h(x^*)$ such that $\nabla f(x^*) + p^* = 0$. Define $\bar{x}^*$ by

$$\bar{x}^* := \arg\min_{y \in \mathbb{R}^d} \langle \nabla f(x^*), y - x^* \rangle + \frac{1}{2}\|y - x^*\|_B^2 + h(y). \tag{2.12}$$

Then, we have

$$h(x^*) \geq \langle \nabla f(x^*), \bar{x}^* - x^* \rangle + \frac{1}{2}\|\bar{x}^* - x^*\|_B^2 + h(\bar{x}^*).$$

By the convexity of h, we have

$$h(\bar{x}^*) \geq h(x^*) + \langle p^*, \bar{x}^* - x^* \rangle.$$

Summing up above two inequalities yields $\|\bar{x}^* - x^*\|_B = 0$, which by the positive definiteness of B implies $\bar{x}^* = x^*$. Hence, $\mathcal{G}_h(x^*, \nabla f(x^*), B) = B(x^* - \bar{x}^*) = 0$.

On the other hand, if $\mathcal{G}_h(x^*, \nabla f(x^*), B) = B(x^* - \bar{x}^*) = 0$, we have $x^* = \bar{x}^*$, where $\bar{x}^*$ is defined by (2.12). Then, we have from the first-order optimality condition that

$$0 \in \nabla f(x^*) + B(\bar{x}^* - x^*) + \partial h(\bar{x}^*) = \nabla f(x^*) + \partial h(x^*),$$

which indicates that x^* is a stationary point of $f+h$.

Moreover, it is straightforward to verify that when h vanishes $\mathcal{G}_0(x, g, B) = g$ for any $g \in \mathbb{R}^d$. □

For notation simplicity, in the following we simply use $\mathcal{G}(x^k_{t-1})$ to denote $\mathcal{G}_h(x^k_{t-1}, \nabla f(x^k_{t-1}), B^k_t)$. By Lemma 2.3 and the Lipschitz continuity assumption of ∇f, it is appropriate to use $\|\mathcal{G}(\cdot)\|^2_{M^{-1}}$ as an operator to measure the first-order optimality condition at the iterate, where $M \in S^d_{++}$ is the matrix associated with the smoothness assumption of f (see assumptions **A1** and **A3**.) Given $\epsilon > 0$ and x_R returned through a random process, we call x_R an ϵ-solution of (1.1)-(1.2), if

$$\mathbb{E}[\|\mathcal{G}(x_R)\|^2_{M^{-1}}] < \epsilon,$$

where the expectation is taken with respect to all the random variables generated in this random process. To discuss the computational complexity, we use $\mathcal{SFO}$ and $\mathcal{PO}$ complexity to mean the maximum number of component gradients of f and proximal subproblem solutions need to be computed by the algorithm, respectively, to obtain an ϵ-solution.

In the following two sections, we will analyze theoretical properties of IPSS in two cases separately. In the first case we consider (1.1)-(1.2) when each f_i is smooth, while in the second case we consider the case that f_i is only weakly smooth.

3 IPSS for solving (1.1)-(1.2) with smooth f

In this section, we investigate the convergence properties of IPSS under the assumption that that f_i, $i = 1, \ldots, n$, is smooth. To continue, we first give two assumptions that will be used throughout this section.

A1 The function f_i, $i = 1, \ldots, n$, is $1/\gamma$-smooth with respect to a matrix $M \in S^d_{++}$, that is, $f_i \in \mathcal{C}^1(\mathbb{R}^d)$[1] and ∇f_i satisfies

$$\|\nabla f_i(x) - \nabla f_i(y)\|_{M^{-1}} \leq \frac{1}{\gamma}\|x - y\|_M, \quad \forall x, y \in \mathbb{R}^d.$$

[1] $f \in \mathcal{C}^1(\mathbb{R}^d)$ means that $f : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable.

A2 There exist two positive constants κ and $\bar{\kappa}$ such that

$$\bar{\kappa} I \succeq M^{-\frac{1}{2}} B_t^k M^{-\frac{1}{2}} \succeq \kappa I$$

for all $k = 1, \ldots, S$ and $t = 1, \ldots, m$, where M is the matrix in **A1**.

Notice that under the assumption **A1**, we have

$$f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + \frac{1}{2\gamma} \|x - y\|_M^2 \quad \forall x, y \in \mathbb{R}^d. \tag{3.1}$$

Under assumption **A2**, it is easy to obtain that

$$\kappa^{-1} I \succeq M^{\frac{1}{2}} (B_t^k)^{-1} M^{\frac{1}{2}} \succeq \bar{\kappa}^{-1} I \quad \text{and} \quad \kappa^{-1} I \succeq (B_t^k)^{-\frac{1}{2}} M (B_t^k)^{-\frac{1}{2}} \succeq \bar{\kappa}^{-1} I.$$

Furthermore, it implies that for any $x \in \mathbb{R}^d$, we have

$$\|x\|_M^2 \leq \kappa^{-1} \|x\|_{B_t^k}^2 \quad \text{and} \quad \|x\|_{(B_t^k)^{-1}}^2 \leq \kappa^{-1} \|x\|_{M^{-1}}^2.$$

3.1 Theoretical properties

In this subsection, we discuss the convergence properties of IPSS. We denote in the following analysis that

$$R_t^k = \mathbb{E}[F(x_t^k) + c_t^k \|x_t^k - \tilde{x}_{k-1}\|_M^2]$$

for $k = 1, \ldots, S$ and $t = 1, \ldots, m$, where the expectation is taken with respect to all the random variables generated by IPSS. The following Lemma 3.2 provides recursive relationship between R_t^k and R_{t-1}^k, which plays a key role in our convergence analysis. To prepare it, we first introduce the following operator:

$$\mathcal{D}_h(x, g, B, \alpha) = -2\alpha \min_{y \in \mathbb{R}^d} \left\{ \langle g, y - x \rangle + \frac{\alpha}{2} \|y - x\|_B^2 + h(y) - h(x) \right\}, \quad \forall \alpha > 0. \tag{3.2}$$

This operator was first given in [13,14]. The lemma below shows its two important properties [28].

LEMMA **3.1.** *(Lemma 3 in [28]) The following properties hold.*

(a) For any fixed $B \in S_{++}^d$, we have

$$\mathcal{D}_h(x, g, B, \alpha) \geq \alpha^2 \|x - \bar{x}\|_B^2, \quad \forall x \in \mathbb{R}^d, \alpha > 0,$$

where $\bar{x} = \arg\min_y \langle g, y - x \rangle + \frac{\alpha}{2} \|y - x\|_B^2 + h(y)$.

(b) For any fixed x, g and $B \in S_{++}^d$, $\mathcal{D}_h(\cdot, \alpha)$ is non-decreasing with respect to $\alpha > 0$, i.e.,

$$\mathcal{D}_h(x, g, B, \alpha_2) \geq \mathcal{D}_h(x, g, B, \alpha_1), \quad \forall \alpha_2 \geq \alpha_1 > 0.$$

LEMMA **3.2.** *Suppose that assumptions **A1** and **A2** hold. By setting $c_m^k = 0$, we have*

$$\begin{aligned} R_t^k \leq R_{t-1}^k &+ \left(\frac{1}{2\gamma} + c_t^k \left(1 + \frac{1}{\beta} \right) \right) \mathbb{E}[\|x_t^k - x_{t-1}^k\|_M^2] - \frac{1}{2} \mathbb{E}[\|x_t^k - x_{t-1}^k\|_{B_t^k}^2] \\ &- \left(\frac{1}{2} - \frac{1}{\gamma \kappa} \right) \mathcal{D}_h(x_{t-1}^k, \nabla f(x_{t-1}^k), B_t^k, 1) + \varepsilon_t^k + 2\sqrt{\varepsilon_t^k \bar{\varepsilon}_t^k} \end{aligned} \tag{3.3}$$

for any $k = 1, \ldots, S$ and $t = 1, \ldots, m$, where $c_{t-1}^k = \frac{1}{\kappa \gamma^2 b_k} + c_t^k (1 + \beta)$ and $\beta > 0$ is any constant.

Proof. By Definition 2.2, we have that for any $z \in \mathbb{R}^d$,

$$\begin{aligned}
&h(z) + \frac{1}{2}\|z - x_{t-1}^k\|_{B_t^k}^2 \\
\geq\, &h(x_t^k) + \left\langle B_t^k(x_{t-1}^k - x_t^k - u_t^k) - v_{t-1}^k, z - x_t^k \right\rangle - \hat{\varepsilon}_t^k + \frac{1}{2}\|z - x_{t-1}^k\|_{B_t^k}^2 \\
=\, &h(x_t^k) - \langle B_t^k u_t^k, z - x_t^k \rangle - \langle v_{t-1}^k, z - x_t^k \rangle - \hat{\varepsilon}_t^k + \frac{1}{2}\|z - x_t^k\|_{B_t^k}^2 + \frac{1}{2}\|x_t^k - x_{t-1}^k\|_{B_t^k}^2,
\end{aligned}$$

which implies that

$$\begin{aligned}
h(x_t^k) \leq h(z) + \langle B_t^k u_t^k, z - x_t^k \rangle + \langle v_{t-1}^k, z - x_t^k \rangle + \frac{1}{2}\|z - x_{t-1}^k\|_{B_t^k}^2 \\
- \frac{1}{2}\|z - x_t^k\|_{B_t^k}^2 - \frac{1}{2}\|x_t^k - x_{t-1}^k\|_{B_t^k}^2 + \hat{\varepsilon}_t^k.
\end{aligned} \tag{3.4}$$

By (3.1), we have

$$f(x_t^k) \leq f(x_{t-1}^k) + \langle \nabla f(x_{t-1}^k), x_t^k - x_{t-1}^k \rangle + \frac{1}{2\gamma}\|x_t^k - x_{t-1}^k\|_M^2$$

and for any $z \in \mathbb{R}^d$,

$$f(x_{t-1}^k) \leq f(z) + \langle \nabla f(x_{t-1}^k), x_{t-1}^k - z \rangle + \frac{1}{2\gamma}\|x_{t-1}^k - z\|_M^2.$$

Summing up above two inequalities yields that

$$f(x_t^k) \leq f(z) + \langle \nabla f(x_{t-1}^k), x_t^k - z \rangle + \frac{1}{2\gamma}\|x_t^k - x_{t-1}^k\|_M^2 + \frac{1}{2\gamma}\|x_{t-1}^k - z\|_M^2. \tag{3.5}$$

Summing up (3.4) and (3.5) gives

$$\begin{aligned}
F(x_t^k) \leq F(z) + \langle x_t^k - z, \nabla f(x_{t-1}^k) - v_{t-1}^k \rangle + \frac{1}{2\gamma}\|x_t^k - x_{t-1}^k\|_M^2 - \frac{1}{2}\|x_t^k - x_{t-1}^k\|_{B_t^k}^2 \\
+ \frac{1}{2\gamma}\|x_{t-1}^k - z\|_M^2 + \frac{1}{2}\|z - x_{t-1}^k\|_{B_t^k}^2 - \frac{1}{2}\|z - x_t^k\|_{B_t^k}^2 + \langle B_t^k u_t^k, z - x_t^k \rangle + \hat{\varepsilon}_t^k.
\end{aligned}$$

Setting $z = \bar{x}_t^k$ with $\bar{x}_t^k$ being defined in (2.7), we have

$$\begin{aligned}
F(x_t^k) \leq F(\bar{x}_t^k) + \langle x_t^k - \bar{x}_t^k, \nabla f(x_{t-1}^k) - v_{t-1}^k \rangle + \frac{1}{2\gamma}\|x_t^k - x_{t-1}^k\|_M^2 - \frac{1}{2}\|x_t^k - x_{t-1}^k\|_{B_t^k}^2 \\
+ \frac{1}{2\gamma}\|x_{t-1}^k - \bar{x}_t^k\|_M^2 + \frac{1}{2}\|x_{t-1}^k - \bar{x}_t^k\|_{B_t^k}^2 - \frac{1}{2}\|\bar{x}_t^k - x_t^k\|_{B_t^k}^2 + \langle B_t^k u_t^k, \bar{x}_t^k - x_t^k \rangle + \hat{\varepsilon}_t^k.
\end{aligned} \tag{3.6}$$

Now, according to the definition of $\bar{x}_t^k$ and smooth property (3.1), we have

$$\begin{aligned}
F(\bar{x}_t^k) &= f(\bar{x}_t^k) + h(\bar{x}_t^k) \\
&\leq F(x_{t-1}^k) + \langle \nabla f(x_{t-1}^k), \bar{x}_t^k - x_{t-1}^k \rangle + \frac{1}{2\gamma}\|\bar{x}_t^k - x_{t-1}^k\|_M^2 + h(\bar{x}_t^k) - h(x_{t-1}^k) \\
&\leq F(x_{t-1}^k) + \langle \nabla f(x_{t-1}^k), \bar{x}_t^k - x_{t-1}^k \rangle + \frac{1}{2}\|\bar{x}_t^k - x_{t-1}^k\|_{B_t^k}^2 + h(\bar{x}_t^k) - h(x_{t-1}^k) \\
&\quad + \frac{1}{2\gamma}\|\bar{x}_t^k - x_{t-1}^k\|_M^2 - \frac{1}{2}\|\bar{x}_t^k - x_{t-1}^k\|_{B_t^k}^2 \\
&= F(x_{t-1}^k) - \frac{1}{2}\mathcal{D}_h(x_{t-1}^k, \nabla f(x_{t-1}^k), B_t^k, 1) + \frac{1}{2\gamma}\|\bar{x}_t^k - x_{t-1}^k\|_M^2 - \frac{1}{2}\|\bar{x}_t^k - x_{t-1}^k\|_{B_t^k}^2,
\end{aligned}$$

which together with (3.6) gives that

$$F(x_t^k) \le F(x_{t-1}^k) + T_1 + T_2 + \frac{1}{2\gamma}\|x_t^k - x_{t-1}^k\|_M^2 - \frac{1}{2}\|x_t^k - x_{t-1}^k\|_{B_t^k}^2 + \frac{1}{\gamma}\|x_{t-1}^k - \bar{x}_t^k\|_M^2 - \frac{1}{2}\mathcal{D}_h(x_{t-1}^k, \nabla f(x_{t-1}^k), B_t^k, 1) - \frac{1}{2}\|\bar{x}_t^k - x_t^k\|_{B_t^k}^2 + \hat{\varepsilon}_t^k,$$

where

$$T_1 = \langle x_t^k - \bar{x}_t^k, \nabla f(x_{t-1}^k) - v_{t-1}^k\rangle \quad \text{and} \quad T_2 = \langle B_t^k u_t^k, \bar{x}_t^k - x_t^k\rangle. \tag{3.7}$$

Notice that

$$\begin{aligned} T_1 &= \langle x_t^k - \bar{x}_t^k, \nabla f(x_{t-1}^k) - v_{t-1}^k\rangle \\ &\le \|x_t^k - \bar{x}_t^k\|_{B_t^k}\|\nabla f(x_{t-1}^k) - v_{t-1}^k\|_{(B_t^k)^{-1}} \\ &\le \frac{1}{2}\|x_t^k - \bar{x}_t^k\|_{B_t^k}^2 + \frac{1}{2}\|\nabla f(x_{t-1}^k) - v_{t-1}^k\|_{(B_t^k)^{-1}}^2. \end{aligned}$$

And by Lemma 2.2 and $\|u_t^k\|_{B_t^k} \le \sqrt{2\bar{\varepsilon}_t^k}$, we have

$$\begin{aligned} T_2 &= \langle B_t^k u_t^k, \bar{x}_t^k - x_t^k\rangle \\ &\le \|u_t^k\|_{B_t^k}\|\bar{x}_t^k - x_t^k\|_{B_t^k} \\ &\le \|u_t^k\|_{B_t^k}\|v_{t-1}^k - \nabla f(x_{t-1}^k)\|_{(B_t^k)^{-1}} + \sqrt{2\varepsilon_t^k}\|u_t^k\|_{B_t^k} \\ &\le \frac{\|u_t^k\|_{B_t^k}^2}{2} + \frac{1}{2}\|v_{t-1}^k - \nabla f(x_{t-1}^k)\|_{(B_t^k)^{-1}}^2 + \sqrt{2\varepsilon_t^k}\|u_t^k\|_{B_t^k} \\ &\le \frac{1}{2}\|v_{t-1}^k - \nabla f(x_{t-1}^k)\|_{(B_t^k)^{-1}}^2 + \bar{\varepsilon}_t^k + 2\sqrt{\varepsilon_t^k\bar{\varepsilon}_t^k}. \end{aligned}$$

Then, we can derive

$$T_1 + T_2 \le \frac{1}{2}\|x_t^k - \bar{x}_t^k\|_{B_t^k}^2 + \|\nabla f(x_{t-1}^k) - v_{t-1}^k\|_{(B_t^k)^{-1}}^2 + \bar{\varepsilon}_t^k + 2\sqrt{\varepsilon_t^k\bar{\varepsilon}_t^k}.$$

In addition, it follows from **A1** and Lemma 3.1 that

$$\frac{1}{\gamma}\|x_{t-1}^k - \bar{x}_t^k\|_M^2 \le \frac{1}{\gamma\kappa}\|x_{t-1}^k - \bar{x}_t^k\|_{B_t^k}^2 \le \frac{1}{\gamma\kappa}\mathcal{D}_h(x_{t-1}^k, \nabla f(x_{t-1}^k), B_t^k, 1). \tag{3.8}$$

It thus follows from (3.7) that

$$\begin{aligned} F(x_t^k) &\le F(x_{t-1}^k) + \|\nabla f(x_{t-1}^k) - v_{t-1}^k\|_{(B_t^k)^{-1}}^2 + \frac{1}{2\gamma}\|x_t^k - x_{t-1}^k\|_M^2 - \frac{1}{2}\|x_t^k - x_{t-1}^k\|_{B_t^k}^2 \\ &\quad + \frac{1}{\gamma}\|x_{t-1}^k - \bar{x}_t^k\|_M^2 - \frac{1}{2}\mathcal{D}_h(x_{t-1}^k, \nabla f(x_{t-1}^k), B_t^k, 1) + \varepsilon_t^k + 2\sqrt{\varepsilon_t^k\bar{\varepsilon}_t^k} \\ &\le F(x_{t-1}^k) + \|\nabla f(x_{t-1}^k) - v_{t-1}^k\|_{(B_t^k)^{-1}}^2 + \frac{1}{2\gamma}\|x_t^k - x_{t-1}^k\|_M^2 - \frac{1}{2}\|x_t^k - x_{t-1}^k\|_{B_t^k}^2 \\ &\quad - \left(\frac{1}{2} - \frac{1}{\gamma\kappa}\right)\mathcal{D}_h(x_{t-1}^k, \nabla f(x_{t-1}^k), B_t^k, 1) + \varepsilon_t^k + 2\sqrt{\varepsilon_t^k\bar{\varepsilon}_t^k}, \end{aligned}$$

where the second inequality is due to (3.8). Consequently, it follows from

$$\|x_t^k - \tilde{x}_{k-1}\|_M^2 \le (1+\beta)\|x_{t-1}^k - \tilde{x}_{k-1}\|_M^2 + \left(1 + \frac{1}{\beta}\right)\|x_t^k - x_{t-1}^k\|_M^2,$$

where $\beta > 0$ is any constant, and

$$\begin{aligned} \mathbb{E}[\|v_{t-1}^k - \nabla f(x_{t-1}^k)\|_{(B_t^k)^{-1}}^2] &= \frac{1}{b_k}\mathbb{E}[\|\nabla f_i(x_{t-1}^k) - \nabla f_i(\tilde{x}_{k-1})\|_{(B_t^k)^{-1}}^2] \\ &\le \frac{1}{\kappa b_k}\mathbb{E}[\|\nabla f_i(x_{t-1}^k) - \nabla f_i(\tilde{x}_{k-1})\|_{M^{-1}}^2] \\ &\le \frac{1}{\kappa\gamma^2 b_k}\mathbb{E}[\|x_{t-1}^k - \tilde{x}_{k-1}\|_M^2] \end{aligned}$$

that

$$\begin{aligned}R_t^k &= \mathbb{E}[F(x_t^k) + c_t^k\|x_t^k - \tilde{x}_{k-1}\|_M^2] \\ &\le \mathbb{E}[F(x_{t-1}^k)] + \left(\frac{1}{\kappa\gamma^2 b_k} + c_t^k(1+\beta)\right)\mathbb{E}[\|x_{t-1}^k - \tilde{x}_{k-1}\|_M^2] \\ &+ \left(\frac{1}{2\gamma} + c_t^k\left(1+\frac{1}{\beta}\right)\right)\mathbb{E}[\|x_t^k - x_{t-1}^k\|_M^2]] - \frac{1}{2}\mathbb{E}[\|x_t^k - x_{t-1}^k\|_{B_t^k}^2 \\ &- \left(\frac{1}{2} - \frac{1}{\gamma\kappa}\right)\mathbb{E}[\mathcal{D}_h(x_{t-1}^k, \nabla f(x_{t-1}^k), B_t^k, 1)] + \varepsilon_t^k + 2\sqrt{\varepsilon_t^k\bar{\varepsilon}_t^k},\end{aligned}$$

which completes the proof by the definition of R_t^k and the relation of c_{t-1}^k with c_t^k. □

We are now ready to state the main theorem of this subsection as follows.

THEOREM **3.1.** *Under the conditions in Lemma 3.2, if P has a uniform distribution, that is*

$$\text{Prob}(R_k = k \text{ and } R_t = t) = \frac{1}{T},$$

where $T = mS$, $b_k = b$ for any k, and $\gamma\kappa \ge \max\{4, 1 + 2m/\sqrt{b}\}$, then

$$\mathbb{E}[\|\mathcal{G}(x_R)\|_{M^{-1}}^2] \le \frac{4\bar{\kappa}^2}{\kappa}\frac{F(\tilde{x}_0) - F^* + \sum_{(k=1,t=1)}^{(S,m)}(\varepsilon_t^k + 2\sqrt{\varepsilon_t^k\bar{\varepsilon}_t^k})}{T}. \tag{3.9}$$

Proof. Recall that $c_{t-1}^k = \frac{1}{\kappa\gamma^2 b} + c_t^k(1+\beta)$ and $c_m^k = 0$. Then it is easy to obtain

$$c_t^k = \frac{1}{\kappa\gamma^2 b}\frac{(1+\beta)^{m-t} - 1}{\beta} \le \frac{m(e-1)}{\kappa\gamma^2 b}$$

with $\beta = \frac{1}{m}$. Then, by $\gamma\kappa \ge 4$ and

$$\left(\frac{1}{2\gamma} + c_t^k\left(1+\frac{1}{\beta}\right)\right)\mathbb{E}[\|x_t^k - x_{t-1}^k\|_M^2]] - \frac{1}{2}\mathbb{E}[\|x_t^k - x_{t-1}^k\|_{B_t^k}^2 \le 0,$$

we have from (3.3) that

$$\frac{1}{4}\mathbb{E}[\mathcal{D}_h(x_{t-1}^k, \nabla f(x_{t-1}^k), B_t^k, 1)] \le R_{t-1}^k - R_t^k + \varepsilon_t^k + 2\sqrt{\varepsilon_t^k\bar{\varepsilon}_t^k}. \tag{3.10}$$

By Lemma 3.1, we have

$$\begin{aligned}\|\mathcal{G}_h(x_{t-1}^k, \nabla f(x_{t-1}^k), B_t^k)\|_{M^{-1}}^2 &= (x_{t-1}^k - \bar{x}_t^k)^T B_t^k M^{-1} B_t^k (x_{t-1}^k - \bar{x}_t^k) \\ &\le \|M^{-\frac{1}{2}} B M^{-\frac{1}{2}}\|^2 \|x_{t-1}^k - \bar{x}_t^k\|_M^2 \\ &\le \frac{\bar{\kappa}^2}{\kappa}\|x_{t-1}^k - \bar{x}_t^k\|_{B_t^k}^2 \\ &\le \frac{\bar{\kappa}^2}{\kappa}\mathcal{D}_h(x_{t-1}^k, \nabla f(x_{t-1}^k), B_t^k, 1).\end{aligned} \tag{3.11}$$

Hence, we have from (3.10) that

$$\mathbb{E}[\|\mathcal{G}_h(x_{t-1}^k, \nabla f(x_{t-1}^k), B_t^k)\|_{M^{-1}}^2] \le \frac{4\bar{\kappa}^2}{\kappa}\left(R_{t-1}^k - R_t^k + \varepsilon_t^k + 2\sqrt{\varepsilon_t^k\bar{\varepsilon}_t^k}\right).$$

So, by summation of the above inequality for $k = 1, \ldots, S$ and $t = 1, \ldots, m$, we have

$$\frac{1}{T}\sum_{(k=1,t=1)}^{(S,m)}\mathbb{E}[\|\mathcal{G}_h(x_{t-1}^k, \nabla f(x_{t-1}^k), B_t^k)\|_{M^{-1}}^2] \le \frac{4\bar{\kappa}^2}{\kappa}\frac{F(\tilde{x}_0) - F^* + \sum_{(k=1,t=1)}^{(S,m)}(\varepsilon_t^k + 2\sqrt{\varepsilon_t^k\bar{\varepsilon}_t^k})}{T}.$$

Then, (3.9) follows from the uniform distribution of P. □

The corollary bellow shows the complexity property of IPSS to achieve an ϵ-solution x_R, i.e. $\mathbb{E}[\|\mathcal{G}(x_R)\|^2_{M^{-1}}] < \epsilon$.

Corollary 3.2. *Under the conditions of Theorem 3.1, if we choose $b = \lfloor n^{\frac{2}{3}} \rfloor$, $m = \lfloor n^{\frac{1}{3}} \rfloor$ and $\sum_{(k=1,t=1)}^{(S,m)} \varepsilon_t^k < +\infty$, then to achieve an ϵ-solution, the $\mathcal{SFO}$ and $\mathcal{PO}$ complexity of IPSS is of order*

$$\mathcal{O}\left(n + \frac{\bar{\kappa}^2 n^{2/3}}{\kappa\epsilon}\right) \quad \text{and} \quad \mathcal{O}\left(\frac{\bar{\kappa}^2}{\kappa\epsilon}\right), \tag{3.12}$$

respectively.

Proof. Since $\kappa > 4/\gamma$, by Theorem 3.1 and ε_t^k being summable, we have

$$\mathbb{E}[\|\mathcal{G}(x_R)\|^2_{M^{-1}}] \leq \frac{4\bar{\kappa}^2}{\kappa} \frac{F(\tilde{x}_0) - F^* + 3\tau}{T}, \tag{3.13}$$

where $\tau = \sum_{(k=1,t=1)}^{(S,m)} \varepsilon_t^k < +\infty$. Hence, to achieve an ϵ-solution, we need T to be large enough such that

$$T > \frac{4\bar{\kappa}^2}{\kappa} \frac{F(\tilde{x}_0) - F^* + 3\tau}{\epsilon}.$$

Thus the total maximum number of component gradients and proximal subproblem solutions computed, i.e., the $\mathcal{SFO}$ and $\mathcal{PO}$ complexity, are $n + (b + n/m)T$ and T, respectively, which are of order (3.12). □

Remark 3.1. *Besides summable tolerances ε_t^k, we could also obtain the complexity property related with some nonsummable ε_t^k. For example, it follows from Theorem 3.1 that*

$$\mathbb{E}[\|\mathcal{G}(x_R)\|^2_{M^{-1}}] = \begin{cases} \mathcal{O}\left(\frac{\bar{\kappa}^2}{\kappa T^\delta}\right), & \text{if } \varepsilon_t^k = \frac{1}{((k-1)m+t)^\delta} \text{ with } 0 < \delta < 1, \\ \mathcal{O}\left(\frac{\bar{\kappa}^2}{\kappa} \frac{\log T}{T}\right), & \text{if } \varepsilon_t^k = \frac{1}{(k-1)m+t}. \end{cases}$$

Then, similar to the analysis in Corollary 3.2, to achieve ϵ-solution, we have that the $\mathcal{SFO}$ complexity is of order

$$\begin{cases} \mathcal{O}\left(n + \frac{\bar{\kappa}^2}{\kappa} n^{\frac{2}{3}} / \epsilon^{\frac{1}{\delta}}\right), & \text{if } \varepsilon_t^k = \frac{1}{((k-1)m+t)^\delta} \text{ with } 0 < \delta < 1, \\ \mathcal{O}\left(n + \frac{\bar{\kappa}^2 n^{\frac{2}{3}}}{\kappa\epsilon} \log(1/\epsilon)\right), & \text{if } \varepsilon_t^k = \frac{1}{(k-1)m+t}; \end{cases}$$

and the $\mathcal{PO}$ complexity is of order

$$\begin{cases} \mathcal{O}\left(\frac{\bar{\kappa}^2}{\kappa} / \epsilon^{\frac{1}{\delta}}\right), & \text{if } \varepsilon_t^k = \frac{1}{((k-1)m+t)^\delta} \text{ with } 0 < \delta < 1, \\ \mathcal{O}\left(\frac{\bar{\kappa}^2}{\kappa\epsilon} \log(1/\epsilon)\right), & \text{if } \varepsilon_t^k = \frac{1}{(k-1)m+t}. \end{cases}$$

3.2 Subgradient method for solving subproblem (2.3)

There is often a tradeoff between using the matrix B and the difficulty of solving subproblem (2.3). We expect the matrix B to capture some second-order curvature information of function f. But, introduction of a complicated matrix B other than using an identity matrix in (2.3) may increase the difficulty of solving subproblem (2.3). Depending on the specific structures of matrix B and function h, there may exist highly efficient algorithms to solve the subproblem (2.3) to the required accuracy. One practical example on the CUR-like factorization optimization can be found in [27], where the function h is complicated so that subproblem (2.3) does not have a closed form solution. In this subsection, we would like to introduce an iterative subgradient method for solving the subproblem (2.3) in its general form ignoring the problem structures.

From (3.9) we can see that the computational complexity of Algorithm 2.1 depends very much on the subproblem tolerance ε_t^k. As a matter of fact, for q defined in (2.1) and $\varepsilon > 0$, following from Theorem 2.8.7 in [30], we have

$$\partial_\varepsilon q(z) = \bigcup \left\{ \partial_{\bar{\varepsilon}} \left\{ \frac{1}{2}\|x - w\|_B^2 \right\} \Big|_{x=z} + \partial_{\hat{\varepsilon}} h(z) : \ \bar{\varepsilon}, \hat{\varepsilon} \geq 0, \ \bar{\varepsilon} + \hat{\varepsilon} = \varepsilon \right\}.$$

Therefore, in particular if $0 \in \partial_\varepsilon q(z)$, there exist $\bar{\varepsilon}, \hat{\varepsilon} > 0$ such that $\varepsilon = \bar{\varepsilon} + \hat{\varepsilon}$ and

$$0 \in \partial_{\bar{\varepsilon}} \left\{ \frac{1}{2}\|x - w\|_B^2 \right\} \Big|_{x=z} + \partial_{\hat{\varepsilon}} h(z).$$

So, we do not specify the choice of $\bar{\varepsilon}_t^k$ and $\hat{\varepsilon}_t^k$. It suffices to obtain an inexact solution x_t^k of (2.3) satisfying

$$q_t^k(x_t^k) \leq \min_x q_t^k(x) + \varepsilon_t^k, \tag{3.14}$$

where ε_t^k is given in (2.5). We now state a subgradient method as follows to solve (2.3).

Algorithm 3.1 Subgradient method for solving (2.3)

Input: T_t^k, y_0, α_i

1: **for** $i = 1, \ldots, T_t^k$ **do**

2: $y_i = y_{i-1} + \alpha_i(v_{t-1}^k + B_t^k(y_{i-1} - x_{t-1}^k) + p_{i-1})$, where $p_{i-1} \in \partial h(y_{i-1})$,

3: **end for**

Output: $\frac{2}{T_t^k(T_t^k+1)} \sum_{i=0}^{T_t^k-1} y_i$.

Notice that under assumption **A2**, it is easy to obtain that $\|B_t^k\|_2 \geq \kappa/\|M^{-1}\|_2$. So, q_t^k is strongly convex with modulus no smaller than $\kappa/\|M^{-1}\|_2$. Hence, we can have the following theorem, from which an explicit bound on the number of iterations by Algorithm 3.1 to achieve (3.14) can be derived. For detailed proof of this algorithm, one may refer to Section 3.2 in [15].

THEOREM **3.3.** *Suppose that assumption* **A2** *holds and Algorithm 3.1 is applied to solve subproblem* (2.3). *If there exists a* $\chi > 0$ *such that* $\|p_{i-1}\| \leq \chi$ *for any* $i = 1, \ldots, T_t^k$, *by choosing* $\alpha_i = \frac{2\|M^{-1}\|_2}{\kappa(i+1)}$, *we have*

$$q_t^k\left(\frac{2}{T_t^k(T_t^k+1)} \sum_{i=0}^{T_t^k-1} y_i\right) - \min_x q_t^k(x) \leq \frac{2\chi^2\|M^{-1}\|_2}{\kappa(T_t^k+1)}.$$

Hence, to achieve (3.14), by Theorem 3.3 we can set x_t^k as

$$x_t^k = \frac{2}{T_t^k(T_t^k+1)} \sum_{i=0}^{T_t^k-1} y_i,$$

and choose T_t^k large enough such that

$$\frac{2\chi^2\|M^{-1}\|_2}{\kappa(T_t^k+1)} \leq \varepsilon_t^k,$$

which equivalently requires

$$T_t^k \geq \frac{2\chi^2\|M^{-1}\|_2}{\kappa\varepsilon_t^k}.$$

Therefore, under same conditions as Theorem 3.1, with particular settings of ε_t^k, the total number of subgradient evaluations of h, that is $\sum_{(k=1,t=1)}^{(S,m)} T_t^k$, is of order

$$\mathcal{O}(T^{1+\delta}) = \begin{cases} \mathcal{O}\left(\left(\frac{\bar{\kappa}^2}{\kappa}/\epsilon\right)^{1+\delta}\right), & \text{if } \varepsilon_t^k = \frac{1}{((k-1)m+t)^\delta} \text{ with } \delta > 1, \\ \mathcal{O}\left(\left(\frac{\bar{\kappa}^2}{\kappa}/\epsilon^{\frac{1}{\delta}}\right)^{1+\delta}\right), & \text{if } \varepsilon_t^k = \frac{1}{((k-1)m+t)^\delta} \text{ with } 0 < \delta < 1, \\ \mathcal{O}\left(\left(\frac{\bar{\kappa}^2}{\kappa\epsilon}\log(1/\epsilon)\right)^{1+\delta}\right), & \text{if } \varepsilon_t^k = \frac{1}{(k-1)m+t}. \end{cases}$$

3.3 Complexity under PPL inequality

In this subsection, we propose a globally linearly convergent algorithm to solve a class of nonconvex composite optimization problems satisfying the proximal Polyak-Łojasiewicz inequality. Polyak-Łojasiewicz inequality was first proposed by Polyak [22] to show the linear convergence rate of gradient methods for solving unconstrained smooth optimization problems. For a continuously differentiable function f with f^* as its minimum value, it is called to satisfy the Polyak-Łojasiewicz inequality if there exists $\mu > 0$ such that

$$\frac{1}{2}\|\nabla f(x)\|^2 \geq \mu(f(x) - f(x^*)).$$

For nonsmooth problems, in order to analyze proximal-type algorithms, a generalization of Polyak-Łojasiewicz inequality, called *proximal Polyak-Łojasiewicz inequality* with definition given below, was studied in [12, 13], where relevant examples satisfying this inequality are also discussed. Interested readers are referred to these two papers for details.

Definition **3.1.** *Consider* (1.1) *with f satisfying assumption* **A1**. *We say F satisfies proximal Polyak-Łojasiewicz (PPL) inequality if there exists $\mu > 0$ such that*

$$\frac{1}{2}\mathcal{D}_h(x, \nabla f(x), M, 1/\gamma) \geq \mu(F(x) - F^*)$$

where $\mathcal{D}_h$ is defined in (3.2) *and F^* is the optimal objective function value of* (1.1).

We now present our PPL-IPSS algorithm and analyze its convergent properties.

Algorithm 3.2 PPL-IPSS algorithm

Input: Initial point $x^0 \in \mathbb{R}^d$, initial matrix $B^0 \in \mathcal{S}_{++}^d$, inexact tolerances $\bar{\varepsilon}, \hat{\varepsilon}$, parameters S and m

1: **for** $s = 1, \ldots, N$ **do**
2: $x^s = \text{IPSS}(x^{s-1}, B^{s-1}, \bar{\varepsilon}^{s-1}, \hat{\varepsilon}^{s-1}, S, m)$;
3: **end for**

Output: x^N.

In this subsection, to emphasize the relationship of the inexactness tolerance ε_t^k in Step 2 of Algorithm 3.2 with iteration index s, we refer ε_t^k as $\varepsilon_{t,s}^k$ and $\bar{\varepsilon}_t^k$ as $\varepsilon_{t,s}^k$.

Theorem **3.4.** *Under the conditions of Theorem 3.1, if the PPL-inequality holds for the objective function F, then*

$$\mathbb{E}[F(x^s) - F^*] \leq \frac{2\bar{\kappa}}{\mu T}\mathbb{E}[F(x^{s-1}) - F^*] + \frac{2\bar{\kappa}}{\mu T}A_s, \quad s = 1, \ldots, N, \tag{3.15}$$

where $T = Sm$ and $A_s = \sum_{(k=1,t=1)}^{(S,m)} \varepsilon_{t,s}^k + 2\sqrt{\varepsilon_{t,s}^k \bar{\varepsilon}_{t,s}^k}$.

Proof. In the s-th iteration of PPL-IPSS, it follows from Lemma 3.1(b) that

$$\begin{aligned}\mathcal{D}_h(x_{t-1}^k, \nabla f(x_{t-1}^k), B_t^k, 1) &\geq -2\min_y \Big\{\langle \nabla f(x_{t-1}^k), y\rangle + \frac{\bar{\kappa}}{2}\|y - x\|_M^2 + h(y) - h(x)\Big\} \\ &= \frac{1}{\bar{\kappa}}\mathcal{D}_h(x_{t-1}^k, \nabla f(x_{t-1}^k), M, \bar{\kappa}) \\ &\geq \frac{1}{\bar{\kappa}}\mathcal{D}_h(x_{t-1}^k, \nabla f(x_{t-1}^k), M, \frac{1}{\gamma}) \\ &\geq \frac{2\mu}{\bar{\kappa}}(F(x_{t-1}^k) - F^*).\end{aligned}$$

Hence, it implies from (3.10) that

$$\frac{\mu}{2\bar{\kappa}}\mathbb{E}[F(x_{t-1}^k) - F^*] \leq R_{t-1}^k - R_t^k + \varepsilon_{t,s}^k + 2\sqrt{\varepsilon_{t,s}^k \bar{\varepsilon}_{t,s}^k}.$$

Summing up the above inequality for $t = 1, \ldots, m$ and $k = 1, \ldots, S$ yields

$$\frac{\mu}{2\bar{\kappa}} \sum_{(k=1,t=1)}^{(S,m)} \mathbb{E}[F(x_{t-1}^k) - F^*] \leq \sum_{(k=1,t=1)}^{(S,m)} (R_{t-1}^k - R_t^k) + A_s.$$

Notice that

$$R_0^k = \mathbb{E}[F(x_0^k)] = \mathbb{E}[F(\tilde{x}_{k-1})], \quad \text{and} \quad R_m^k = \mathbb{E}[F(x_m^k)] = \mathbb{E}[F(\tilde{x}_k)],$$

where the second equality is due to $c_m^k = 0$. Therefore, by Algorithm 3.2, it yields that

$$\frac{\mu}{2\bar{\kappa}} \sum_{(k=1,t=1)}^{(S,m)} (F(x_{t-1}^k) - F^*) \leq \mathbb{E}[F(\tilde{x}_0)] - F^* + A_s.$$

Thus, from $F(\tilde{x}_0) = F(x^{s-1})$ we have

$$\mathbb{E}[F(x^s) - F^*] \leq \frac{2\bar{\kappa}}{\mu T}\mathbb{E}[F(x^{s-1}) - F^*] + \frac{2\bar{\kappa}}{\mu T}A_s.$$

Hence, (3.15) holds. □

We now analyze the computational complexity of PPL-IPSS. Given an $\epsilon > 0$, in the following we study the total number of gradient evaluations of f as well as the number of subgradient evaluations of h to obtain x^N such that $\mathbb{E}[F(x^N) - F^*] < \epsilon$. Since the computational complexity relies on the settings of inexactness tolerance $\varepsilon_{t,s}^k$, we next classify the analysis into several cases with different specifications of $\varepsilon_{t,s}^k$.

Case 1. In this case, we set

$$\varepsilon_{t,s}^k = \alpha^{(1+\theta)s} \cdot \frac{1}{((k-1)m + t)^\delta}, \quad \theta > 0, \delta > 0, 1 > \alpha > 0.$$

Depending on the value of δ, we further specify the analysis into three subcases.

Subcase 1. $\delta > 1$. Then

$$A_s \leq 3 \sum_{(k=1,t=1)}^{(S,m)} \varepsilon_{t,s}^k \leq \frac{3}{\delta - 1}\alpha^{(1+\theta)s},$$

which yields

$$\mathbb{E}[F(x^s) - F^*] \leq \frac{2\bar{\kappa}}{\mu T}\mathbb{E}[F(x^{s-1}) - F^*] + \frac{6\bar{\kappa}}{\mu(\delta - 1)T}\alpha^{(1+\theta)s}.$$

Subcase 2. $\delta = 1$. Then

$$A_s \leq 3 \sum_{(k=1,t=1)}^{(S,m)} \varepsilon_{t,s}^k \leq 3\alpha^{(1+\theta)s} \log T,$$

which yields

$$\mathbb{E}[F(x^s) - F^*] \leq \frac{2\bar{\kappa}}{\mu T}\mathbb{E}[F(x^{s-1}) - F^*] + \frac{6\bar{\kappa}\log T}{\mu T}\alpha^{(1+\theta)s}.$$

Subcase 3. $\delta < 1$. Then

$$A_s \leq 3 \sum_{(k=1,t=1)}^{(S,m)} \varepsilon_{t,s}^k \leq \frac{3T^{1-\delta}}{1-\delta}\alpha^{(1+\theta)s},$$

which yields

$$\mathbb{E}[F(x^s) - F^*] \leq \frac{2\bar{\kappa}}{\mu T}\mathbb{E}[F(x^{s-1}) - F^*] + \frac{6\bar{\kappa}T^{1-\delta}}{\mu(1-\delta)T}\alpha^{(1+\theta)s}.$$

In all the above three subcases, we can always choose T sufficiently large such that

$$\mathbb{E}[F(x^s) - F^*] \leq \alpha\mathbb{E}[F(x^{s-1}) - F^*] + \alpha\alpha^{(1+\theta)s} \leq B_\alpha \cdot \alpha^s,$$

where B_α is a positive constant depending on α. Hence $\mathbb{E}[F(x^s) - F^*]$ converges to zero linearly. Therefore, to achieve $\mathbb{E}[F(x^N) - F^*] < \epsilon$, the outer iteration number N should satisfy that $\alpha^N < \epsilon$, namely,

$$N = \mathcal{O}(\log(1/\epsilon)).$$

Hence, the $\mathcal{SFO}$ complexity is in the order of $\mathcal{O}(N(n + n^{2/3}T))$, same as $\mathcal{O}((n + n^{2/3}T)\log(1/\epsilon))$, where $T = mS$.

We now consider the total number of subgradient evaluations of h when Algorithm 3.1 is applied to solve the subproblem (2.3). Notice that in the s-th iteration of applying IPSS algorithm in Algorithm 3.2, we have the total number of subgradient evaluations of h is

$$T_{t,s}^k = \frac{1}{\varepsilon_{t,s}^k} = \frac{((k-1)m + t)^\delta}{\alpha^{(1+\theta)s}}.$$

Consequently, the total number of subgradient evaluations of h is

$$\sum_{s=1}^{N} \sum_{(k=1,t=1)}^{(S,m)} T_{t,s}^k = \sum_{s=1}^{N} \alpha^{-(1+\theta)s} \sum_{(k=1,t=1)}^{(S,m)} ((k-1)m + t)^\delta = \mathcal{O}(T^{1+\delta}\alpha^{-(1+\theta)N}) = \mathcal{O}(T^{1+\delta}/\epsilon^{1+\theta}).$$

Noticing that $T = \mathcal{O}(1)$. Therefore, we obtain the number of component gradient evaluations of f and the number of subgradient evaluations of h are in the order of

$$\mathcal{O}(n\log(1/\epsilon)) \quad \text{and} \quad \mathcal{O}(1/\epsilon^{1+\theta}),$$

respectively.

Case 2. In this case, we set

$$\varepsilon_{t,s}^k = \alpha^s \cdot \frac{1}{((k-1)m + t)^\delta}, \quad \delta > 0, 1 > \alpha > 0,$$

then similar to *Case 1*, we can choose T sufficiently large such that

$$\mathbb{E}[F(x^s) - F^*] \leq \alpha\mathbb{E}[F(x^{s-1}) - F^*] + \alpha A_s \leq B_\alpha \cdot s\alpha^s.$$

Hence, to achieve $\mathbb{E}[F(x^N) - F^*] < \epsilon$, it should have $N\alpha^N = \mathcal{O}(\epsilon)$. By defining $\rho = \alpha^{-N}$, we have $N = \log\rho$ and $\rho^{-1}\log\rho = \mathcal{O}(\epsilon)$. Then it is sufficient to require $\rho = \mathcal{O}\left(\epsilon^{-1}\log(1/\epsilon)\right)$, which yields

$$N = \mathcal{O}(\log(1/\epsilon) + \log\log(1/\epsilon)).$$

Hence, the $\mathcal{SFO}$ complexity and the total number of subgradient evaluations of h are in the order of

$$\mathcal{O}\left(n\log(1/\epsilon) + n\log\log(1/\epsilon)\right) \quad \text{and} \quad \mathcal{O}\left(\epsilon^{-1}\log(1/\epsilon)\right),$$

respectively.

4 IPSS for (1.1)-(1.2) with weakly smooth f

In this section, we assume that f_i, $i = 1, \ldots, n$, is only weakly smooth and study the theoretical properties of IPSS in this case. To be clear, we first make the following assumptions.

A3 The function f_i, $i = 1, \ldots, n$, is $1/\gamma$-weakly smooth with respect to a matrix $M \in S^d_{++}$, that is $f_i \in \mathcal{C}^1(\mathbb{R}^d)$ and ∇f_i satisfies

$$\|\nabla f_i(x) - \nabla f_i(y)\|_{M^{-1}} \leq \frac{1}{\gamma}\|x - y\|_M^{\nu}, \quad \forall x, y \in \mathbb{R}^d,$$

where $\nu \in (0, 1)$.

A4 There exist two positive constants κ_t^k and $\bar{\kappa}_t^k$ such that

$$\bar{\kappa}_t^k I \succeq M^{-\frac{1}{2}} B_t^k M^{-\frac{1}{2}} \succeq \kappa_t^k I$$

for all $k = 1, \ldots, S$ and $t = 1, \ldots, m$, where M is the matrix in **A3**.

It follows from assumption **A3** that for any x and y,

$$f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{1}{\gamma(1+\nu)}\|y - x\|_M^{1+\nu}.$$

Notice that different from the assumption **A2**, **A4** assumes that the lower and upper bounds vary along with k and t.

We now give the main theorem in this section.

THEOREM **4.1.** *Suppose assumptions* **A3** *and* **A4** *hold, $b_k = b$ for any k, there exists a constant C such that $\bar{\kappa}_t^k \leq C\kappa_t^k$ with*

$$\kappa_t^k = \max\left\{\frac{4}{\gamma}(s_t^k)^{1-\nu}, \frac{1}{\gamma}(s_t^k)^{1-\nu} + \frac{2\nu}{\gamma^2 b}m(m+1)(e-1)\right\}, \tag{4.1}$$

where $s_t^k = (k+(1-\nu)S-1)m+t$ and e is the Euler's number, and P has the probabilistic distribution

$$\text{Prob}(R_k = k \text{ and } R_t = t) = \frac{(\kappa_t^k)^{-1}}{\sum_{(k=1,t=1)}^{(S,m)}(\kappa_t^k)^{-1}}.$$

Then, we have

$$\mathbb{E}[\|\mathcal{G}(x_R)\|_{M^{-1}}^2] \leq \left(\tilde{C} + 12C^2 \sum_{(k=1,t=1)}^{(S,m)} \varepsilon_t^k\right)\left(\frac{16}{\gamma}T^{-\nu} + \frac{4\nu m^2}{\gamma^2 b}T^{-1}\right), \tag{4.2}$$

where $\tilde{C} = 4C^2(F(\tilde{x}_0) - F^*) + \frac{4C^2(1-\nu)}{\gamma^2 b}\log\frac{2-\nu}{1-\nu} + \frac{6C^2(1-\nu)}{\gamma\nu}$.

Proof. Similar to (3.5), we have

$$f(x_t^k) \leq f(z) + \langle \nabla f(x_{t-1}^k), x_t^k - z\rangle + \frac{1}{\gamma(1+\nu)}\|x_t^k - x_{t-1}^k\|_M^{1+\nu} + \frac{1}{\gamma(1+\nu)}\|x_{t-1}^k - z\|_M^{1+\nu}. \tag{4.3}$$

Then, summing up (4.3) and (3.4) with $z = \bar{x}_t^k$ provides that

$$\begin{aligned} F(x_t^k) \leq\ & F(\bar{x}_t^k) + \langle x_t^k - \bar{x}_t^k, \nabla f(x_{t-1}^k) - v_{t-1}^k\rangle + \frac{1}{\gamma(1+\nu)}\|x_t^k - x_{t-1}^k\|_M^{1+\nu} - \frac{1}{2}\|x_t^k - x_{t-1}^k\|_{B_t^k}^2 \\ & + \frac{1}{\gamma(1+\nu)}\|x_{t-1}^k - \bar{x}_t^k\|_M^{1+\nu} + \frac{1}{2}\|x_{t-1}^k - \bar{x}_t^k\|_{B_t^k}^2 - \frac{1}{2}\|\bar{x}_t^k - x_t^k\|_{B_t^k}^2 + \langle B_t^k u_t^k, \bar{x}_t^k - x_t^k\rangle + \hat{\varepsilon}_t^k. \end{aligned} \tag{4.4}$$

Following from the definition of $\bar{x}_t^k$, we have that

$$F(\bar{x}_t^k) \leq F(x_{t-1}^k) + \frac{1}{\gamma(1+\nu)}\|\bar{x}_t^k - x_{t-1}^k\|_M^{1+\nu} - \frac{1}{2}\|\bar{x}_t^k - x_{t-1}^k\|_{B_t^k}^2 - \frac{1}{2}\|\bar{x}_t^k - x_{t-1}^k\|_{B_t^k}^2. \tag{4.5}$$

So, summing up (4.5) with (4.4) gives that

$$\begin{aligned} F(x_t^k) \leq F(x_{t-1}^k) + T_1 + T_2 + \frac{1}{\gamma(1+\nu)}\|x_t^k - x_{t-1}^k\|_M^{1+\nu} - \frac{1}{2}\|x_t^k - x_{t-1}^k\|_{B_t^k}^2 \\ + \frac{2}{\gamma(1+\nu)}\|x_{t-1}^k - \bar{x}_t^k\|_M^{1+\nu} - \frac{1}{2}\|x_{t-1}^k - \bar{x}_t^k\|_{B_t^k}^2 - \frac{1}{2}\|\bar{x}_t^k - x_t^k\|_{B_t^k}^2 + \hat{\varepsilon}_t^k, \end{aligned}$$

where T_1 and T_2 are defined in (3.7). Then, it follows from the bounds on T_1 and T_2 right after (3.7) that

$$\begin{aligned} F(x_t^k) \leq F(x_{t-1}^k) + \|v_{t-1}^k - \nabla f(x_{t-1}^k)\|_{(B_t^k)^{-1}}^2 + \frac{1}{\gamma(1+\nu)}\|x_t^k - x_{t-1}^k\|_M^{1+\nu} - \frac{1}{2}\|x_t^k - x_{t-1}^k\|_{B_t^k}^2 \\ + \frac{2}{\gamma(1+\nu)}\|x_{t-1}^k - \bar{x}_t^k\|_M^{1+\nu} - \frac{1}{2}\|x_{t-1}^k - \bar{x}_t^k\|_{B_t^k}^2 + \varepsilon_t^k + 2\sqrt{\varepsilon_t^k \bar{\varepsilon}_t^k}. \end{aligned} \tag{4.6}$$

Taking expectation on both sides of (4.6) yields that

$$\begin{aligned} \mathbb{E}[F(x_t^k)] \leq \mathbb{E}[F(x_{t-1}^k)] + \mathbb{E}[\|v_{t-1}^k - \nabla f(x_{t-1}^k)\|_{(B_t^k)^{-1}}^2] + \frac{1}{\gamma(1+\nu)}\mathbb{E}[\|x_t^k - x_{t-1}^k\|_M^{1+\nu}] - \frac{1}{2}\mathbb{E}[\|x_t^k - x_{t-1}^k\|_{B_t^k}^2] \\ + \frac{2}{\gamma(1+\nu)}\mathbb{E}[\|x_{t-1}^k - \bar{x}_t^k\|_M^{1+\nu}] - \frac{1}{2}\mathbb{E}[\|x_{t-1}^k - \bar{x}_t^k\|_{B_t^k}^2] + \varepsilon_t^k + 2\sqrt{\varepsilon_t^k \bar{\varepsilon}_t^k}. \end{aligned}$$

Since

$$\mathbb{E}[\|v_{t-1}^k - \nabla f(x_{t-1}^k)\|]_{(B_t^k)^{-1}}^2 \leq \frac{1}{\gamma^2 \kappa_t^k b_k}\|x_{t-1}^k - \tilde{x}_{k-1}\|_M^{2\nu} = \frac{1}{\gamma^2 \kappa_t^k b}\|x_{t-1}^k - \tilde{x}_{k-1}\|_M^{2\nu},$$

we have

$$\begin{aligned} \mathbb{E}[F(x_t^k)] \leq \mathbb{E}[F(x_{t-1}^k)] + \frac{1}{\gamma^2 \kappa_t^k b}\mathbb{E}[\|x_{t-1}^k - \tilde{x}_{k-1}\|_M^{2\nu}] + \frac{1}{\gamma(1+\nu)}\mathbb{E}[\|x_t^k - x_{t-1}^k\|_M^{1+\nu}] - \frac{1}{2}\mathbb{E}[\|x_t^k - x_{t-1}^k\|_{B_t^k}^2] \\ + \frac{2}{\gamma(1+\nu)}\mathbb{E}[\|x_{t-1}^k - \bar{x}_t^k\|_M^{1+\nu}] - \frac{1}{2}\mathbb{E}[\|x_{t-1}^k - \bar{x}_t^k\|_{B_t^k}^2] + \varepsilon_t^k + 2\sqrt{\varepsilon_t^k \bar{\varepsilon}_t^k}. \end{aligned}$$

By the inequality $ab \leq a^p/p + b^q/q$ with $p = \frac{2}{1+\nu}$, $q = \frac{2}{1-\nu}$, $a = \|x_t^k - x_{t-1}^k\|_M^{1+\nu}(s_t^k)^{\frac{(1+\nu)(1-\nu)}{2}}$ and $b = (s_t^k)^{-\frac{(1+\nu)(1-\nu)}{2}}$, we have

$$\|x_t^k - x_{t-1}^k\|_M^{1+\nu} \leq \|x_t^k - x_{t-1}^k\|_M^2 \cdot (s_t^k)^{1-\nu} \cdot \frac{1+\nu}{2} + (s_t^k)^{-(1+\nu)} \cdot \frac{1-\nu}{2}$$

and

$$\|x_{t-1}^k - \bar{x}_t^k\|_M^{1+\nu} \leq \|x_{t-1}^k - \bar{x}_t^k\|_M^2 \cdot (s_t^k)^{1-\nu} \cdot \frac{1+\nu}{2} + (s_t^k)^{-(1+\nu)} \cdot \frac{1-\nu}{2}.$$

Similarly, from $ab \leq a^p/p + b^q/q$ with $p = \frac{1}{v}$, $q = \frac{1}{1-v}$, $a = \|x_{t-1}^k - \tilde{x}_{k-1}\|_M^{2\nu}(\kappa_t^k)^\nu$ and $b = (\kappa_t^k)^{-\frac{\nu}{1-\nu}}$, it yields that

$$\|x_{t-1}^k - \tilde{x}_{k-1}\|_M^{2\nu} \leq \|x_{t-1}^k - \tilde{x}_{k-1}\|_M^2 \cdot \kappa_t^k \nu + (\kappa_t^k)^{-\frac{\nu}{1-\nu}} \cdot (1-\nu).$$

Then, we have

$$\begin{aligned} \mathbb{E}[F(x_t^k)] \leq \mathbb{E}[F(x_{t-1}^k)] + \frac{1}{\gamma^2 b}\nu\mathbb{E}[x_t^k - \tilde{x}_{k-1}\|_M^2] + \frac{1-\nu}{\gamma^2 b}(\kappa_t^k)^{-\frac{1}{1-\nu}} \\ + \frac{1}{2\gamma}\|x_t^k - x_{t-1}^k\|_M^2 (s_t^k)^{1-\nu} + \frac{1-\nu}{2\gamma(1+\nu)}(s_t^k)^{-(1+\nu)} - \frac{1}{2}\mathbb{E}[\|x_t^k - x_{t-1}^k\|_{B_t^k}^2] \\ + \frac{1}{\gamma}\|x_{t-1}^k - \bar{x}_t^k\|_M^2 (s_t^k)^{1-\nu} + \frac{1-\nu}{\gamma(1+\nu)}(s_t^k)^{-(1+\nu)} - \frac{1}{2}\mathbb{E}[\|x_{t-1}^k - \bar{x}_t^k\|_{B_t^k}^2] \\ + \varepsilon_t^k + 2\sqrt{\varepsilon_t^k \bar{\varepsilon}_t^k}. \end{aligned}$$

Therefore,

$$\begin{aligned}
R_t^k &= \mathbb{E}[F(x_t^k) + c_t^k \|x_t^k - \tilde{x}_{k-1}\|_M^2] \\
&\leq \mathbb{E}[F(x_t^k)] + c_t^k \left(1 + \frac{1}{\beta}\right) \mathbb{E}[\|x_t^k - x_{t-1}^k\|_M^2] + c_t^k (1+\beta) \mathbb{E}[\|x_{t-1}^k - \tilde{x}_{k-1}\|_M^2] \\
&\leq \mathbb{E}[F(x_{t-1}^k)] + \left(\frac{\nu}{\gamma^2 b} + c_t^k (1+\beta)\right) \mathbb{E}[x_t^k - \tilde{x}_{k-1}\|_M^2] \\
&\quad + \left(\frac{1}{2\gamma}(s_t^k)^{1-\nu} + c_t^k \left(1 + \frac{1}{\beta}\right)\right) \|x_t^k - x_{t-1}^k\|_M^2 - \frac{1}{2}\mathbb{E}[\|x_t^k - x_{t-1}^k\|_{B_t^k}^2] \\
&\quad + \frac{1}{\gamma}(s_t^k)^{1-\nu} \|x_{t-1}^k - \bar{x}_t^k\|_M^2 - \frac{1}{2}\mathbb{E}[\|x_{t-1}^k - \bar{x}_t^k\|_{B_t^k}^2] \\
&\quad + \frac{1-\nu}{\gamma^2 b}(\kappa_t^k)^{-\frac{1}{1-\nu}} + \frac{3(1-\nu)}{2\gamma(1+\nu)}(s_t^k)^{-(1+\nu)} + \varepsilon_t^k + 2\sqrt{\varepsilon_t^k \bar{\varepsilon}_t^k}.
\end{aligned}$$

Let

$$c_{t-1}^k = \frac{\nu}{\gamma^2 b} + c_t^k (1+\beta).$$

By setting $c_m^k = 0$, we obtain that

$$c_t^k \leq c_{t-1}^k \leq c_0^k = \frac{\nu}{\gamma^2 b} \frac{(1+\beta)^m - 1}{\beta}.$$

Furthermore, by setting $\beta = 1/m$, we have

$$c_t^k \leq \frac{\nu}{\gamma^2 b} m(e-1).$$

It thus yields

$$\left(\frac{1}{2\gamma}(s_t^k)^{1-\nu} + c_t^k \left(1 + \frac{1}{\beta}\right)\right) \|x_t^k - x_{t-1}^k\|_M^2 < \frac{1}{2}\mathbb{E}[\|x_t^k - x_{t-1}^k\|_{B_t^k}^2]$$

and

$$\frac{1}{\gamma}(s_t^k)^{1-\nu} \|x_{t-1}^k - \bar{x}_t^k\|_M^2 - \frac{1}{2}\mathbb{E}[\|x_{t-1}^k - \bar{x}_t^k\|_{B_t^k}^2] < -\frac{1}{4}\mathbb{E}[\|x_{t-1} - \bar{x}_t^k\|_{B_t^k}^2].$$

So, we have

$$\frac{1}{4}\mathbb{E}[\|x_{t-1}^k - \bar{x}_t^k\|_{B_t^k}^2] \leq R_{t-1}^k - R_t^k + \frac{1-\nu}{\gamma^2 b}(\kappa_t^k)^{-\frac{1}{1-\nu}} + \frac{3(1-\nu)}{2\gamma(1+\nu)}(s_t^k)^{-(1+\nu)} + 3\varepsilon_t^k.$$

Following from (3.11) and the condition that $\bar{\kappa}_t^k \leq C\kappa_t^k$, we have

$$(\kappa_t^k)^{-1}\mathbb{E}[\|\mathcal{G}_h(x_{t-1}^k, \nabla f(x_{t-1}^k), B_t^k)\|_{M^{-1}}^2] \leq C^2 \mathbb{E}[\|x_{t-1}^k - \bar{x}_t^k\|_{B_t^k}^2].$$

Now, summing up the above inequality for $t = 1, \ldots, m$ and $k = 1, \ldots, S$ yields that

$$\begin{aligned}
&\sum_{(k=1,t=1)}^{(S,m)} (\kappa_t^k)^{-1}\mathbb{E}[\|\mathcal{G}_h(x_{t-1}^k, \nabla f(x_{t-1}^k), B_t^k)\|_{M^{-1}}^2] \\
&\leq 4C^2(F(\tilde{x}_0) - F^*) + 4C^2 \frac{1-\nu}{\gamma^2 b} \sum_{(k=1,t=1)}^{(S,m)} (\kappa_t^k)^{-\frac{1}{1-\nu}} + \frac{6C^2(1-\nu)}{\gamma(1+\nu)} \sum_{(k=1,t=1)}^{(S,m)} (s_t^k)^{-(1+\nu)} + 12C^2 \sum_{(k=1,t=1)}^{(S,m)} \varepsilon_t^k \\
&\leq 4C^2(F(\tilde{x}_0) - F^*) + \frac{4C^2(1-\nu)}{\gamma^2 b} \sum_{(k=1,t=1)}^{(S,m)} (s_t^k)^{-1} + \frac{6C^2(1-\nu)}{\gamma\nu} + 12C^2 \sum_{(k=1,t=1)}^{(S,m)} \varepsilon_t^k \\
&\leq \tilde{C} + 12C^2 \sum_{(k=1,t=1)}^{(S,m)} \varepsilon_t^k, \qquad (4.7)
\end{aligned}$$

where the last inequality follows from

$$\sum_{(k=1,t=1)}^{(S,m)} (s_t^k)^{-1} \leq \sum_{(k=(1-\nu)S,t=1)}^{((2-\nu)S,m)} (km+t)^{-1} \leq \log \frac{2-\nu}{1-\nu}.$$

Notice that by the IPSS algorithm and the probability distribution of P, we have

$$\text{Prob}(x_R = x_{t-1}^k) = \frac{(\kappa_t^k)^{-1}}{\sum_{(k=1,t=1)}^{(S,m)} (\kappa_t^k)^{-1}}$$

for all $t = 1, \ldots, m$; and $k = 1, \ldots, N$. Hence, we have

$$\sum_{(k=1,t=1)}^{(S,m)} (\kappa_t^k)^{-1} \mathbb{E}[\|\mathcal{G}_R\|^2] = \sum_{(k=1,t=1)}^{(S,m)} (\kappa_t^k)^{-1} \mathbb{E}[\|\mathcal{G}_h(x_{t-1}^k, \nabla f(x_{t-1}^k), B_t^k)\|_{M^{-1}}^2]. \tag{4.8}$$

Notice that

$$\sum_{(k=1,t=1)}^{(S,m)} (\kappa_t^k)^{-1} \mathbb{E}[\|\mathcal{G}_R\|_{M^{-1}}^2] \geq \frac{T^2}{\sum_{(k=1,t=1)}^{(S,m)} \kappa_t^k} \mathbb{E}[\|\mathcal{G}(x_R)\|_{M^{-1}}^2]. \tag{4.9}$$

It is easy to obtain from (4.1) that

$$\kappa_t^k \leq \frac{4}{\gamma}(s_t^k)^{1-\nu} + \frac{4\nu m^2}{\gamma^2 b}.$$

Hence,

$$\sum_{(k=1,t=1)}^{(S,m)} \kappa_t^k \leq \frac{4(2-\nu)^{2-\nu}}{\gamma} T^{2-\nu} + \frac{4\nu m^2}{\gamma^2 b} T \leq \frac{16}{\gamma} T^{2-\nu} + \frac{4\nu m^2}{\gamma^2 b} T. \tag{4.10}$$

By combining the inequalities (4.7)-(4.10), we have (4.2). □

Corollary 4.2. *Under the same assumptions as Theorem 4.1, if ε_t^k is summable, $m = \lfloor (n/\nu)^{\frac{1}{2+\nu}} \rfloor$ and $b = \lfloor n/m \rfloor$, then to achieve an ϵ-solution of (1.1)-(1.2), the $\mathcal{SFO}$ and $\mathcal{PO}$ complexity of IPSS are*

$$\mathcal{O}\left(n + n^{\frac{1+\nu}{2+\nu}} / \epsilon^{\frac{1}{\nu}}\right) \quad \text{and} \quad \mathcal{O}\left(1/\epsilon^{\frac{1}{\nu}}\right), \tag{4.11}$$

respectively.

Proof. It is easy to obtain from the setting of m that $bT^{-\nu} \geq \nu m^2 T^{-1}$, which together with (4.2) gives

$$\mathbb{E}[\|\mathcal{G}(x_R)\|_{M^{-1}}^2] \leq \left(\frac{16}{\gamma} + \frac{4}{\gamma^2}\right) \left(\tilde{C} + 12C^2 \sum_{(k=1,t=1)}^{(S,m)} \varepsilon_t^k\right) T^{-\nu}.$$

Therefore, to achieve an ϵ-solution, we have $T = \mathcal{O}(1/\epsilon^{\frac{1}{\nu}})$ if ε_t^k is summable, which implies that the $\mathcal{SFO}$ and $\mathcal{PO}$ complexity are those given in (4.11), respectively. □

Remark 4.1. *We would like to compare the complexity proved in Theorem 4.1 with that given in [24]. With the same settings as those in our paper, when f is nonconvex and h is convex, the $\mathcal{PO}$ computational complexity given in [24] is $\mathcal{O}(1/\epsilon^{\frac{1+3\nu}{\nu}})$, while in our paper it is only $\mathcal{O}(1/\epsilon^{\frac{1}{\nu}})$ when ε_t^k is summable.*

5 Summary

In this paper, we proposed a framework and studied the theoretical properties of an inexact proximal stochastic second-order (IPSS) algorithm for solving nonconvex composite optimization. The objective function of such an optimization problem consists of an average of many, possibly weakly, smooth nonconvex functions and a possibly nonsmooth convex function. This IPSS algorithm allows to incorporate second-order information through a positive-definite matrix in the proximal subproblem to accelerate the convergence. In addition, IPSS allows to solve the subproblem inexactly while still keeping desired computational complexity. We also give an iterative subgradient method to solve the subproblem to the required accuracy and discuss its overall computational complexity. When the objective function satisfies the PL inequality, based on the IPSS algorithm, we propose an algorithm, called PPL-IPSS, which has been shown to have linear convergence rate. Furthermore, we have investigated the convergence properties of IPSS when f in the objective function is only weakly smooth. In this paper, we have focused on the theoretical properties of a framework of IPSS algorithm. Extensive numerical experiments will be performed in the following work to verify the practical performance of IPSS by investigating proper strategies of setting the algorithm parameters.

6 Acknowledgements

Part of research by Xiao Wang was done during her working as a research fellow in University Research Facility in Big Data Analytics of the Hong Kong Polytechnic University. We would also like to thank two anonymous referees for providing us valuable comments and suggestions.

References

[1] M. Ahookhosh. Accelerated first-order methods for large-scale convex minimization, Mathematical Methods of Operations Research, 1-35, 2019.

[2] S. Becker and J. Fadili, A second order proximal splitting method, Advances in Neural Information Processing Systems, Lake Tahoe, 26182626, 2012.

[3] D. P. Bertsekas, A. Nedić and A. E. Ozdaglar, Convex analysis and optimization, Athena Scientific, Belmont, Mass. 02478-9998, U.S.A., 2003.

[4] R.H. Byrd, G.M. Chin, J. Nocedal and F. Oztoprak, A family of second-order methods for convex l1-regularized optimization, Math. Program., 159(12): 435467, 2016.

[5] R.H. Byrd, J. Nocedal, and F. Oztoprak, An inexact successive quadratic approximation method for convex l1 regularized optimization, arXiv:1309.3529, 2013.

[6] O. Devolder, F. Glineur, Y.E. Nesterov, First-order methods within exact oracle: the strongly convex case, CORE Discussion Paper 2013/16, 2013.

[7] J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani, Pathwise coordinate optimization, Ann. Appl. Stat. 1(2): 302332, 2007.

[8] H. Ghanbari and K. Scheinberg, Proximal second order methods for regularized convex optimization with linear and accelerated sublinear convergence rates, Comp. Opt. and Appl., DOI:10.1007/s10589-017-9964-z, 2018.

[9] C.J. Hsieh, M.A. Sustik, I.S. Dhillon, and P. Ravikumar, Sparse inverse covariance matrix estimation using quadratic approximation, In: NIPS, 23302338, 2011.

[10] M. Ito, New results on subgradient methods for strongly convex optimization problems with a unified analysis, Comp. Opt. and Appl., 65(1): 127-172, 2016.

[11] R. Johnson, T. Zhang: Accelerating stochastic gradient descent using predictive variance reduction. In: NIPS, 315323, 2013.

[12] Hamed Karimi and Mark Schmidt. Linear convergence of proximal-gradient methods under the polyak-Łojasiewicz condition. In NIPS Workshop, 2015.

[13] H. Karimi, J. Nutini, M. Schmidt: Linear convergence of gradient and proximal-gradient methods under the polyak-Łojasiewicz condition. In Joint European Conference on Ma- chine Learning and Knowledge Discovery in Databases (pp. 795-811). Springer, Cham, September, 2016.

[14] S. Karimiand and S. Vavasis, I mro: A proximal second order method for solving l_1-regularized least squares problems, SIAM J. Optim. 27(2): 583615, 2017.

[15] S. Lacoste-Julien, M. Schmidt and F. Bach, A simpler approach to obtaining an $\mathcal{O}(1/t)$ convergence rate for the projected stochastic subgradient method, arXiv: 1212.2002, 2012.

[16] G. Lan, Bundle-level type methods uniformly optimal for smooth and non-smooth convex optimization, Math. Program., 149(1)1-45, 2015.

[17] J. Lee, Y. Sun, and M. Saunders, Proximal Newton-type methods for minimizing composite functions, SIAM. J. Optim. 24(3): 14201443, 2014.

[18] H. Lin, J. Mairal, and Z. Harchaoui, A generic second order algorithm for faster gradient-based optimization, arXiv:1610.00960, 2016.

[19] A.S. Nemirovskii, Y.E. Nesterov: Optimal methods for smooth of order convex minimization. Zh. Vichisl. Mat. Fiz. 25, 356369 (1985). (In Russian)

[20] Y.E. Nesterov: Complexity bounds for primal-dual methods minimizing the model of objective function, Math. Program., 171: 311-330, 2018.

[21] N.H. Pham, L.M. Nguyen, D.T. Phan, Q. Tran-Dinh, ProxSARAH: An Efficient Algorithmic Framework for Stochastic Composite Nonconvex Optimization, arXiv:1902.05679v2 [math.OC] 29 Mar 2019.

[22] B.T. Polyak. Gradient methods for the minimization of functionals. USSR Comput, Math. & Math. Phys., 3(4): 864878, 1963.

[23] S.J. Reddi, S. Sra, B. Poczos and A. Smola, Fast stochastic methods for nonsmooth nonconvex optimization, preprint (2016). Available at arXiv:1605.06900.

[24] G. Saeed, Conditional gradient type methods for composite nonlinear and stochastic optimization, Math. Program., 173(1-2): 431-64, 2019.

[25] S. Ghadimi, G. Lan and H. Zhang, Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization, Math. Program. 155(1-2) (2016), pp. 267305.

[26] G. Saeed, G. Lan, and Hongchao Zhang, Generalized uniformly optimal methods for nonlinear programming, J. Sci. Comput. 79(3): 1854-81, 2019.

[27] X. Wang, S.X. Wang and H. Zhang, Inexact proximal stochastic gradient method for convex composite optimization, Comp. Opt. and Appl., 68(3): 579-618, 2017.

[28] X.Y. Wang, X. Wang and Y. Yuan, Stochastic proximal quasi-Newton methods for non-convex composite optimization, Optim. Methods Softw., 34: 922-948, 2019.

[29] L. Xiao, T. Zhang: A proximal stochastic gradient method with progressive variance reduction. SIAM J. Optim. 24, 2057-2075, 2014.

[30] C. Zălinescu: Convex Analysis in General Vector Spaces. World Scientific Publishing Co. Inc., Singapore, 2002.

[31] Y. Zhang, X. Lin, Stochastic primal dual coordinate method for regularized empirical risk minimization. In: ICML, 2015.