Comput Optim Appl
DOI 10.1007/s10589-017-9932-7

Inexact proximal stochastic gradient method for convex composite optimization

Xiao Wang[1] · **Shuxiong Wang**[2] · **Hongchao Zhang**[3]

Received: 9 September 2016
© Springer Science+Business Media, LLC 2017

Abstract We study an inexact proximal stochastic gradient (IPSG) method for convex composite optimization, whose objective function is a summation of an average of a large number of smooth convex functions and a convex, but possibly nonsmooth, function. Variance reduction techniques are incorporated in the method to reduce the stochastic gradient variance. The main feature of this IPSG algorithm is to allow solving the proximal subproblems inexactly while still keeping the global convergence with desirable complexity bounds. Different subproblem stopping criteria are proposed. Global convergence and the component gradient complexity bounds are derived for the both cases when the objective function is strongly convex or just generally convex. Preliminary numerical experiment shows the overall efficiency of the IPSG algorithm.

This research is partially supported by the National Natural Science Foundation of China 11301505 and the National Science Foundation of USA 1522654.

✉ Xiao Wang
wangxiao@ucas.ac.cn

Shuxiong Wang
wsx@lsec.cc.ac.cn

Hongchao Zhang
hozhang@math.lsu.edu
http://www.math.lsu.edu/∼hozhang

1 School of Mathematical Sciences, University of Chinese Academy of Sciences, No.19A Yuquan Road, Beijing 100049, China

2 Institute of Computational Mathematics and Scientific/Engineering Computing, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijng 100190, China

3 Department of Mathematics, Louisiana State University, 220 Lockett Hall, Baton Rouge, LA 70803-4918, USA

Published online: 11 August 2017

Keywords Convex composite optimization · Empirical risk minimization · Stochastic gradient · Inexact methods · Global convergence · Complexity bound

Mathematics Subject Classification 47N10 · 65K10

1 Introduction

In this paper, we consider the following composite optimization problem

$$\min_{w\in\mathbb{R}^d} \quad P(w) := f(w) + h(w), \tag{1.1}$$

where f is an average of many smooth component convex functions, that is

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

with $f_i\colon \mathbb{R}^d \to \mathbb{R}$, $i = 1, \ldots, n$, being convex (not necessarily strongly convex) and Lipschitz continuously differentiable, and $h : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is a convex, but possibly nonsmooth, function. In addition, we assume the solution set of (1.1) is not empty and achieves the optimal function value P^*. Problem (1.1)–(1.2) is also known as regularized empirical risk minimization [32] and can arise in many applications, such as many kinds of classification and regression models in machine learning. In these applications, a key feature of the problem (1.1)–(1.2) is that the total number of component functions, denoted by n, is very large or even huge such that it is often prohibitive to do the full function and/or gradient evaluations of the objective function at each iteration, which is often required by an iterative method.

Problem (1.1) can be solved by standard proximal gradient methods. Given a starting point $w_0 \in \mathbb{R}^d$, such methods update iterates through

$$w_{k+1} = \arg\min_{w\in\mathbb{R}^d} \left\{ \langle \nabla f(w_k), w - w_k\rangle + h(w) + \frac{1}{2\eta_k}\|w - w_k\|_2^2 \right\}, \tag{1.3}$$

where $\eta_k > 0$ is a proximal parameter. Since the full gradient of f is used in (1.3), the proximal gradient method is also referred as the proximal full gradient (Prox-FG) method. The Prox-FG methods have been shown superior to the classical subgradient methods for solving (1.1). To achieve an ε-solution $\bar{w}$ of the problem (1.1), i.e. $P(\bar{w}) - P^* < \varepsilon$, the standard iteration complexity bound for the Prox-FG method is $\mathcal{O}(\varepsilon^{-1})$, while for subgradient method it is $\mathcal{O}(\varepsilon^{-2})$. When combining with Nesterov's acceleration technique, the iteration complexity bound of the Prox-FG methods can achieve $\mathcal{O}(\varepsilon^{-0.5})$. Extensive research has been carried out on proximal gradient methods, such as [2,9,24]. However, in many cases it is hard or time-consuming to evaluate the full gradient ∇f. So, proximal stochastic gradient (Prox-SG) methods become popular for solving (1.1). These methods usually update the iterates by

$$w_{k+1} = \arg\min_{w\in\mathbb{R}^d} \left\{ \langle g_k, w - w_k\rangle + h(w) + \frac{1}{2\eta_k}\|w - w_k\|_2^2 \right\},$$

where g_k is an approximate gradient at w_k based on some random variable and its related expectation is equal to $\nabla f(x_k)$. For general convex case, based on Nesterov's optimal method, Lan [20] propose an unified optimal stochastic approximation algorithm which can achieve optimal complexity order $O(\epsilon^{-2})$ without particulary knowing the problem structure. In addition, the algorithm depends optimally on the Lipschitz-type constants of the problems. Then, Ghadimi and Lan [12,13] have studied the accelerated stochastic approximation methods for strongly convex case, which can achieve optimal or nearly optimal convergence rate with proper stepsizes. The Prox-SG methods can be also referred as randomized incremental proximal methods when g_k is defined as component gradient and the index is chosen in a certain random order. A good review of incremental proximal methods can be found in [4,5,34,35].

Although much computation could be saved by using the stochastic gradient at each iteration, its large variance may slow down the convergence speed. Hence, the variance reduction techniques aiming to reduce the variance of stochastic gradients have recently attracted much interest. The variance reduction techniques are originally proposed for improving classical stochastic gradient methods for solving (1.1)–(1.2) with $h \equiv 0$. A stochastic average gradient (SAG) method, which takes an average of historical stochastic gradients at each iteration, is proposed in [29]. The $\mathcal{O}(n\log(1/\varepsilon))$ component gradient complexity bound is achieved by this SAG method if f is strongly convex, Lipschitz continuously differentiable and $n \geq 8L/\mu$, where L is the Lipschitz constant and μ is the strongly convex modulus. When $f_i(w) = \phi_i(a_i^T w), i = 1,\ldots,n$, with ϕ_i being L_i-smooth and h being μ-strongly convex, a stochastic dual coordinate ascent (SDCA) method is proposed in [31] to solve (1.1)–(1.2). This SDCA method is shown to have the $\mathcal{O}((n + L_{max}/\mu)\log(1/\varepsilon))$ complexity, where $L_{max} = \max_i\{L_i\}$. Although both SAG and SDCA have lower computational cost at each iteration and achieve good convergence rate, they all have a relatively high $\mathcal{O}(nd)$ memory requirement. A stochastic variance reduced gradient (SVRG) method was introduced in [17] for solving smooth and strongly convex problem. This SVRG method uses an unbiased stochastic gradient constructed from the full gradient computed at a particular chosen point after every $m > 0$ iterations. With proper choice of m, the same convergence rate as [29,31] can be achieved, but only the $\mathcal{O}(d)$ memory is required. Another closely related algorithm is SAGA proposed in [8], which uses an unbiased stochastic gradient incorporated with average of past stochastic gradients. Hence, it could be considered as a method between SAG and SVRG. SAGA can be directly extended to solve the composite optimization when the objective function is not strongly convex. But, similar to SAG, SAGA also needs relatively high memory. Due to the small memory requirement and low component gradient complexity, SVRG-type methods have been paid more attention recently. In [36], a Prox-SVRG method is proposed to achieve $\mathcal{O}((n + L/\mu)\log(1/\varepsilon))$ component gradient complexity for solving strongly convex composite optimization. When the objective function is not strongly convex, by applying the method on a slightly perturbed strongly convex problem, the $(n + L/\varepsilon)\mathcal{O}(\log(1/\varepsilon))$ complexity bound can be obtained. On the other hand, in [40]

a SVRG based Prox-SG method is designed to solve the non-strongly convex composite optimization directly with an improved $\mathcal{O}(n\log(1/\varepsilon)+L/\varepsilon)$ complexity bound. For other Prox-SG type methods, one may refer to [14,19,25].

In theory, all the aforementioned stochastic gradient methods require to solve the proximal subproblems exactly. However, when the function h is complicated and/or does not have special structure, the proximal subproblem often does not have closed form solution. Many optimization problems including topographic dictionary learning [18,23], CUR-like factorization optimization [22,23] and multi-task learning problems [26,37], normally require the regularization term h reveal some group sparsity structure of variables. For example, it might take the form of summation of 2-norm. In this case, it would be expensive or sometimes impossible to solve the subproblems exactly. On the other hand, since the subproblems are built on stochastic (inexact) gradients, it is also unnecessary to spend great efforts for solving the subproblems exactly, especially in the early iterations of an algorithm. To the best of our knowledge, there is no work focusing on inexact Prox-SG methods yet. Hence, in this paper we would develop an inexact proximal stochastic gradient (IPSG) method which would guarantee global convergence and desirable complexity bounds. In particular, we show that although the subproblems are solved inexactly in the IPSG method, compared to the exact SVRG based Prox-SG methods, the same or almost the same component gradient complexity bounds can still be maintained. Our analysis is partly motivated from the deterministic inexact Prox-FG methods. One line of research on inexact Prox-FG methods focuses on using approximate proximal operators for which exact solution could be obtained relatively easily, such as the methods developed in [6,7,11,21]. The inexact Prox-FG methods developed in [15,28,30,33] are more closely related to our work. Instead of constructing an approximate proximal operator, these methods allow to solve subproblems inexactly up to certain accuracy derived from theoretical analysis. However, most of these Prox-FG methods use deterministic exact full gradients, which as mentioned previously is not suitable for solving our problem (1.1)–(1.2). Although the method in [30] allows inexact gradients, the analysis developed for Prox-FG method in [30] could not be directly applied to the inexact Prox-SG methods. One key reason is that the variance of stochastic gradient may not meet the accuracy required in [30] for the inexact gradients. For example, without combining with particular treatment, the variance of stochastic gradient may not even reduce. Hence, in the similar spirit to the SVRG methods, we also apply the variance reduction techniques to control the variance of stochastic gradients. As a result, in this paper, we would propose an inexact Prox-SG method with variance reduction techniques for solving (1.1)–(1.2) and develop its corresponding convergence theory.

The remainder of this paper is organized as follows. In Sect. 2, we present an inexact Prox-SG (IPSG) method for solving the convex composite optimization problem (1.1)–(1.2) and discuss some of its basic properties. In Sect. 3, we give the convergence analysis of this IPSG method when f is strongly convex and provide its convergence rate under different accuracy criteria of solving the subproblem. In Sect. 4, we extend our analysis to the case that f may not be strongly convex and give the corresponding convergence rate when the subproblems are solved inexactly. Some preliminary numerical experiment results are reported in Sect. 5. Finally, some conclusion remarks are given in Sect. 6.

Notation The gradient of f is denoted as ∇f. Given a vector $x \in \mathbb{R}^d$ and $p > 0$, $\|x\|_p$ denotes its p-norm, while $\|x\| = \|x\|_2$ represents its Euclidean norm. For $x, y \in \mathbb{R}^d$, we use $\langle x, y\rangle$ to denote the Euclidean inner product of x and y. Given a matrix $X \in \mathbb{R}^{m\times n}$, $\|X\|_F$ is its Frobenius norm. For a real number a, $\lfloor a \rfloor$ denotes the largest integer number less than or equal to a. Given a random variable X, $\mathbb{E}_\xi[X]$ denotes the expectation of X with respect to ξ.

2 Inexact Prox-SG method for (1.1)–(1.2)

In this section, we first give the definition of inexact solutions used in this paper for the proximal mapping problem:

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

Then, we will present our inexact Prox-SG algorithm for solving the convex composite optimization problem (1.1)–(1.2) and develop some of its important properties needed in the later analysis.

Let us first recall the definition of the ε-subdifferential of a convex function: Given a convex function $\phi : \mathbb{R}^d \to \mathbb{R}$, its ε-subdifferential at z, denoted as $\partial_\varepsilon \phi(z)$, is defined as

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

Now, we give our definition of inexact solutions used for the problem (2.1).

Definition 2.1 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

$$\|u\| \leq \sqrt{2\eta\bar{\varepsilon}} \quad \text{and} \quad \frac{y - z - u}{\eta} \in \partial_{\hat{\varepsilon}} h(z). \tag{2.2}$$

By the definition of ε-subdifferential, when $\|u\| \leq \sqrt{2\eta\bar{\varepsilon}}$, it is not difficult to verify that

$$\frac{z - y + u}{\eta} \in \partial_{\bar{\varepsilon}} \left\{ \frac{1}{2\eta}\|w - y\|^2 \right\}\Big|_{w=z}.$$

Hence, if z is an $(\bar{\varepsilon}, \hat{\varepsilon})$-solution of the problem (2.1), by (2.2) we have

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

So, by the definition of ε-subdifferential again we have

$$q(w) - q(z) \geq \langle 0, w - z\rangle - (\bar{\varepsilon} + \hat{\varepsilon}), \quad \text{for any } w \in \mathbb{R}^d,$$

which implies

$$q(z) \leq \min_{w\in\mathbb{R}^d} q(w) + (\bar{\varepsilon} + \hat{\varepsilon}). \tag{2.3}$$

Now, let us consider a particular case that h is an indicator function of a closed convex set $C \subseteq \mathbb{R}^d$. For this case, similar observations have been made in [27,33]. Our purpose here is to compare the inexact solutions given by Definition 2.1 and other characterization of inexact solutions in the literature. When $y \in C$, the solution of proximal mapping (2.1) is trivial. Hence, we assume $y \notin C$. Given a tolerance $\varepsilon > 0$, one often used condition for inexact solutions of (2.1) is to find z such that

$$0 \in \partial_\varepsilon q(z), \tag{2.4}$$

which is equivalent (see. e.g., [3, Section 4.3]) to require z satisfying

$$q(z) \leq \min_{w\in\mathbb{R}^d} q(w) + \varepsilon. \tag{2.5}$$

This condition (2.5) is used in [30]. Note that when $\bar{\varepsilon} + \hat{\varepsilon} \leq \varepsilon$, by (2.3), the inexact solution z given by Definition 2.1 will also satisfy (2.5). When z satisfies (2.5), we have $z \in C$ and

$$\begin{aligned} &\frac{1}{2\eta}\|w - y\|^2 + \varepsilon \geq \frac{1}{2\eta}\|z - y\|^2 \quad \text{for any } w \in C \\ &\iff \|z - y\|^2 \leq \|\mathrm{Proj}_C(y) - y\|^2 + 2\eta\varepsilon, \end{aligned}$$

where $\mathrm{Proj}_C(\cdot)$ is the projection on C. So, geometrically, z belongs to the intersection of C and the ball centered at y with radius $\sqrt{\|\mathrm{Proj}_C(y) - y\|^2 + 2\eta\varepsilon}$ (see Fig. 1a). Another type of condition often used for inexact solutions of (2.1) (see, e.g. in [27,33]) requires z to satisfy

$$\frac{y - z}{\eta} \in \partial_\varepsilon h(z), \tag{2.6}$$

which is equivalent to $z \in C$ and

$$\langle w - z, y - z\rangle \leq \eta\varepsilon, \quad \text{for any } w \in C. \tag{2.7}$$

Note that the condition (2.2) in Definition 2.1 will reduce to (2.6) by setting $\bar{\varepsilon} = 0$ and $\hat{\varepsilon} = \varepsilon$. Geometrically, z satisfying (2.6) belongs to the intersection of C and the half-space determined by (2.7), for which the separating hyperplane is normal to $y - z$ and with distance $\eta\varepsilon/\|y - z\|$ to z. (see Fig. 1b). The third approximation criterion used in literature (see, e.g. [27]) for inexact solutions of (2.1) is to require z satisfying

$$d(0, \partial q(z)) \leq \varepsilon, \tag{2.8}$$

which is equivalent to existing $e \in \mathbb{R}^d$ and $\|e\| \leq \varepsilon$ such that

$$e \in \partial q(z) \iff \frac{y - z}{\eta} + e \in \partial h(z)$$

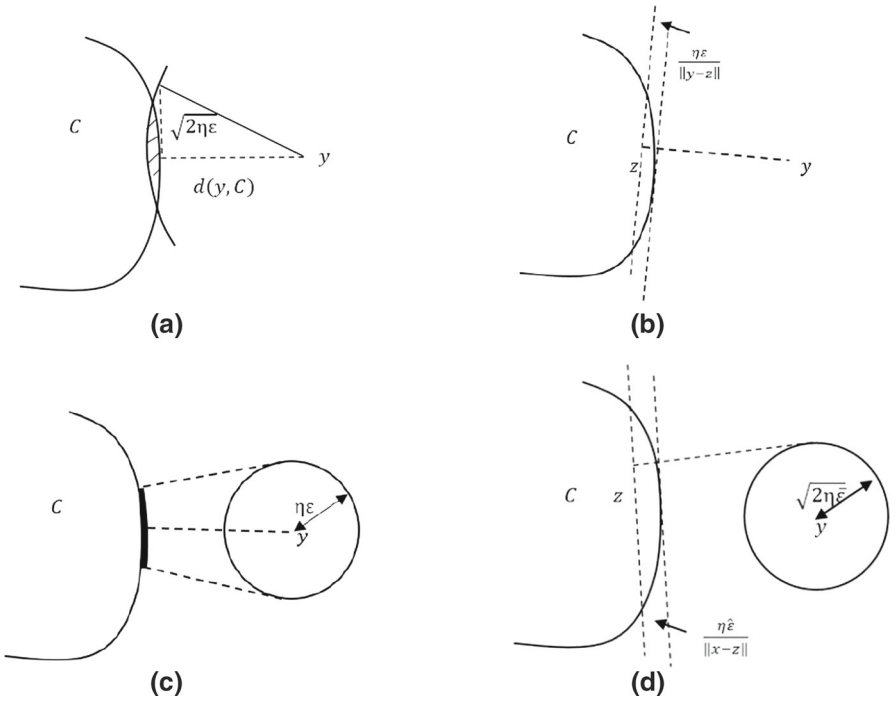

Fig. 1 Comparison of several inexact solution criteria. **a** Corresponds to (2.5), **b** corresponds to (2.6), **c** corresponds to (2.8), **d** corresponds to (2.2)

$$\begin{aligned} &\Longleftrightarrow \langle y + \eta e - z, w - z\rangle \le 0, \quad \text{for any } w \in C \\ &\Longleftrightarrow z = \mathrm{Proj}_C(y + \eta e). \end{aligned}$$

Therefore, geometrically, the approximate solution z satisfying (2.8) belongs to the projection of the ball centered at y with radius $\eta\varepsilon$ on C (see Fig. 1c). Note that the condition (2.2) in Definition 2.1 will reduce to (2.8) by setting $\hat{\varepsilon} = 0$ and $\bar{\varepsilon} = \eta\varepsilon^2/2$. Finally, the condition (2.2) in Definition 2.1 is equivalent to $z \in C$ and

$$\langle w - z, x - z\rangle \le \eta\hat{\varepsilon}, \quad \text{for any } w \in C, \tag{2.9}$$

where $x = y - u$. Geometrically, this means z belongs to the intersection of C and the half-space determined by (2.9), for which the separating hyperplane is normal to $x - z$ and with distance $\eta\hat{\varepsilon}/\|x - z\|$ to z and x is any point in the ball centered at y with radius $\sqrt{2\eta\bar{\varepsilon}}$ (see Fig. 1d).

Now, we give our inexact Prox-SG algorithm for solving (1.1)–(1.2).

Algorithm 2.1 Inexact Prox-SG method (IPSG) for (1.1)–(1.2)

Input: Maximum number of iterations N, integer parameters $\{m_k\}_{k=1}^N$, $\{s_k\}_{k=1}^N$, and tolerance sequence $\left\{\bar{\varepsilon}_t^k\right\}$ and $\left\{\hat{\varepsilon}_t^k\right\}$ for $k = 1, \ldots, N$ and $t = 1, \ldots, m_k$. Initialize $w_0^1 = \tilde{w}_0$.

1: **for** $k = 1, 2, \ldots, N$ **do**
2: Choose proximal parameter $\eta_k > 0$.
3: Calculate $\tilde{v} = \nabla f(\tilde{w}_{k-1})$.
4: **for** $t = 1, 2, \ldots, m_k$ **do**
5: Randomly choosing a sample set $\mathcal{K} \subseteq \{1, 2, \ldots, n\}$ with size s_k, such that the probability of each index being picked is s_k/n. Compute

$$v_{t-1}^k = \nabla f_{\mathcal{K}}\left(w_{t-1}^k\right) - \nabla f_{\mathcal{K}}(\tilde{w}_{k-1}) + \tilde{v},$$

where $\nabla f_{\mathcal{K}}(\cdot) = \frac{1}{s_k}\sum_{i\in\mathcal{K}} \nabla f_i(\cdot)$.

6: Find an $(\bar{\varepsilon}_t^k, \hat{\varepsilon}_t^k)$-solution w_t^k of the subproblem

$$\min_{w\in\mathbb{R}^d} q_t^k(w) := \left\langle v_{t-1}^k, w - w_{t-1}^k\right\rangle + h(w) + \frac{1}{2\eta_k}\left\|w - w_{t-1}^k\right\|^2. \tag{2.10}$$

7: **end for**
8: Set $\tilde{w}_k = \frac{1}{m_k}\sum_{t=1}^{m_k} w_t^k$.
9: Choose w_0^{k+1}.
10: **end for**

Output: $\tilde{w}_N$.

Notice that Algorithm 2.1 consists of a two-loop procedure. In the inner loop, a sequence of inexact proximal stochastic gradient steps are carried out. Because of expensive calculation of the full gradient, by randomly choosing a sample set $\mathcal{K}$, an inexpensive stochastic gradient of f is formed and used in calculating these proximal gradient steps. Meanwhile, the variance reduction technique proposed in SVRG [17] is applied to control the variance of the stochastic gradient generated in the inner iterations. In the outer loop, we choose the proximal parameter $\eta_k > 0$ and update $\tilde{w}_k$, where the full gradient $\nabla f(\tilde{w}_k)$ will be calculated. In Step 9, we do not specify how to choose the starting iterate w_0^{k+1} of the inner loop. As will be seen in the later sections, two different strategies on how to choose w_0^{k+1} will be given according to whether the objective function is strongly convex or simply convex. In Step 5, there are different ways of choosing the sample set $\mathcal{K}$ such that the probability of each index in $\{1, \ldots, n\}$ being picked is s_k/n. The following simple strategy is considered in [39] and used in our numerical experiments. First, partition all the indices $\{1, \ldots, n\}$ into s_k disjoint groups $\mathcal{K}_j$, $j = 1, \ldots, s_k$, with equal cardinality n/s_k(assuming s_k divides n). Then, $\mathcal{K}$ consists of picking one index from each group $\mathcal{K}_j$ with uniform probability s_k/n. The major difference between our Algorithm 2.1 and the existing Prox-SG methods is that the subproblem (2.10) is solved inexactly in Algorithm 2.1. Hence, the computational cost of solving each subproblem could be reduced greatly compared with exactly solving of (2.10) required by other methods. More precisely, we only find an $(\bar{\varepsilon}_t^k, \hat{\varepsilon}_t^k)$-solution w_t^k of the subproblem (2.10). By Definition 2.1, we want to find w_t^k such that there exists an u_t^k satisfying

$$\left\|u_t^k\right\| \leq \sqrt{2\eta_k \bar{\varepsilon}_t^k} \quad \text{and} \quad \frac{w_{t-1}^k - w_t^k - u_t^k}{\eta_k} - v_{t-1}^k \in \partial_{\bar{\varepsilon}_t^k} h\left(w_t^k\right). \tag{2.11}$$

For convenience, in the following, we let

$$\varepsilon_t^k = \bar{\varepsilon}_t^k + \hat{\varepsilon}_t^k. \tag{2.12}$$

Hence, by (2.3), we have

$$q_t^k\left(w_t^k\right) \leq \min_{w \in \mathbb{R}^d} q_t^k(w) + \varepsilon_t^k. \tag{2.13}$$

In the later sections, we would give the global convergence and the complexity analysis on the rate of $\mathbb{E}[P(\tilde{w}_N)] - P(w_*)$ converging to zero after N outer iterations of Algorithm 2.1, where w_* is an optimal solution of (1.1), i.e.,

$$w_* \in \arg\min_{w \in \mathbb{R}^d} P(w).$$

The following assumption is required throughout this paper.

AS.1 For any $i \in \{1, \ldots, n\}$, f_i is Lipschitz continuously differentiable with Lipschitz constant $L_i > 0$, i.e.,

$$f_i(x) - f_i(y) - \langle \nabla f_i(y), x - y \rangle \leq \frac{L_i}{2}\|x - y\|^2, \quad \text{for any } x, y \in \mathbb{R}^d.$$

In the following we denote

$$L = \max_{i \in \{1,\ldots,n\}} L_i. \tag{2.14}$$

Hence, under the Assumption **AS.1**, all the functions f_i, $i = 1, \ldots, n$, and f are Lipschitz continuously differentiable with constant L.

To facilitate the analysis, we first observe that for any $w \in \mathbb{R}^d$, $\nabla f_{\mathcal{K}}(w)$ is an unbiased estimate of $\nabla f(w)$. Indeed, since for any index i the probability of $i \in \mathcal{K}$ is s_k/n, by taking expectation of $\nabla f_{\mathcal{K}}(w)$ with respect to the random sample set $\mathcal{K}$, we have

$$\begin{aligned}\mathbb{E}_{\mathcal{K}}[\nabla f_{\mathcal{K}}(w)] &= \frac{1}{s_k}\mathbb{E}_{\mathcal{K}}\left[\sum_{i \in \mathcal{K}} \nabla f_i(w)\right] = \frac{1}{s_k}\sum_{i=1}^{n}\left[\frac{s_k}{n}\nabla f_i(w)\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(w) = \nabla f(w).\end{aligned} \tag{2.15}$$

So, by the definition of v_{t-1}^k in Algorithm 2.1, (2.15) implies that v_{t-1}^k is an unbiased estimate of $\nabla f\left(w_{t-1}^k\right)$, that is, $\mathbb{E}\left[v_{t-1}^k\right] = \nabla f\left(w_{t-1}^k\right)$. The variance of the stochastic gradient v_{t-1}^k can be bounded by the following lemma, whose proof is a proper modification of the proof of Lemma 3.4 in [36] due to the existence of batch size s_k.

Lemma 2.1 *Under assumption* **AS.1**, *we have*

$$\mathbb{E}_{\mathcal{K}}\left[\left\|v_{t-1}^k-\nabla f\left(w_{t-1}^k\right)\right\|^2\right]\leq\frac{4L}{s_k}\left[P\left(w_{t-1}^k\right)-P(w_*)+P(\tilde{w}_{k-1})-P(w_*)\right],$$

where L is defined in (2.14).

Proof Given any $i\in\{1,\ldots,n\}$, since f_i is Lipschitz continuously differentiable with constant L_i, for any $w\in\mathbb{R}^d$ we have

$$\|\nabla f_i(w)-\nabla f_i(w_*)\|^2\leq 2L_i[f_i(w)-f_i(w_*)-\langle\nabla f_i(w_*),w-w_*\rangle].$$

By the definition of L and summing up the above inequality over $i=1,\ldots,n$, we have

$$\frac{1}{n}\sum_{i=1}^n\|\nabla f_i(w)-\nabla f_i(w_*)\|^2\leq 2L[f(w)-f(w_*)-\langle\nabla f(w_*),w-w_*\rangle]. \tag{2.16}$$

From the optimality condition of (1.1), there exists a p_* such that for any $w\in\mathbb{R}^d$

$$\langle\nabla f(w_*)+p_*,w-w_*\rangle\geq 0,$$

which together with the convexity of h yields that

$$-\langle\nabla f(w_*),w-w_*\rangle\leq\langle p_*,w-w_*\rangle\leq h(w)-h(w_*).$$

Then, it follows from (2.16) that for any $w\in\mathbb{R}^d$ we have

$$\frac{1}{n}\sum_{i=1}^n\|\nabla f_i(w)-\nabla f_i(w_*)\|^2\leq 2L[P(w)-P(w_*)]. \tag{2.17}$$

Now, noticing that $\mathbb{E}_{\mathcal{K}}[\nabla f_{\mathcal{K}}(w_{t-1}^k)]=\nabla f(w_{t-1}^k),\mathbb{E}_{\mathcal{K}}[\nabla f_{\mathcal{K}}(\tilde{w}_{k-1})]=\nabla f(\tilde{w}_{k-1})$ and the fact that for any index i the probability of $i\in\mathcal{K}$ is s_k/n, we have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{K}}\left[\left\|v_{t-1}^k-\nabla f(w_{t-1}^k)\right\|^2\right]\\
&=\mathbb{E}_{\mathcal{K}}\left[\left\|\nabla f_{\mathcal{K}}\left(w_{t-1}^k\right)-\nabla f_{\mathcal{K}}(\tilde{w}_{k-1})+\nabla f(\tilde{w}_{k-1})-\nabla f\left(w_{t-1}^k\right)\right\|^2\right]\\
&=\frac{1}{s_k^2}\mathbb{E}_{\mathcal{K}}\left[\left\|\sum_{i\in\mathcal{K}}\left(\nabla f_i\left(w_{t-1}^k\right)-\nabla f_i(\tilde{w}_{k-1})+\nabla f(\tilde{w}_{k-1})-\nabla f\left(w_{t-1}^k\right)\right)\right\|^2\right]\\
&=\frac{1}{s_k^2}\mathbb{E}_{\mathcal{K}}\left[\sum_{i\in\mathcal{K}}\left\|\nabla f_i\left(w_{t-1}^k\right)-\nabla f_i(\tilde{w}_{k-1})+\nabla f(\tilde{w}_{k-1})-\nabla f\left(w_{t-1}^k\right)\right\|^2\right]
\end{aligned}$$

$$\leq \frac{1}{s_k^2}\mathbb{E}_{\mathcal{K}}\left[\sum_{i\in\mathcal{K}}\left\|\nabla f_i\left(w_{t-1}^k\right)-\nabla f_i(\tilde{w}_{k-1})\right\|^2\right]$$
$$= \frac{1}{s_k^2}\sum_{i=1}^{n}\frac{s_k}{n}\left\|\nabla f_i\left(w_{t-1}^k\right)-\nabla f_i(\tilde{w}_{k-1})\right\|^2.$$

Hence, it follows from (2.17) that

$$\mathbb{E}_{\mathcal{K}}\left[\left\|v_{t-1}^k-\nabla f\left(w_{t-1}^k\right)\right\|^2\right]$$
$$\leq \frac{2}{s_k n}\sum_{i=1}^{n}\left[\left\|\nabla f_i\left(w_{t-1}^k\right)-\nabla f_i(w_*)\right\|^2+\left\|\nabla f_i(\tilde{w}_{k-1})-\nabla f_i(w_*)\right\|^2\right]$$
$$\leq \frac{4L}{s_k}\left[P\left(w_{t-1}^k\right)-P(w_*)+P(\tilde{w}_{k-1})-P(w_*)\right].$$

□

Remark 2.1 The Lemma 2.1 provides an insight that the batch size s_k can help to further reduce the variance of the stochastic gradient. By the later analysis, although the overall component gradient complexity can not be reduced by increasing the batch size s_k, the stochastic gradient could be obtained by parallel computing when $s_k > 1$ and hence, the total number of subproblems to be solved can be reduced by a factor of s_k. Note that since we do not save the component gradients $\nabla f_i(\tilde{w}_{k-1})$, $i = 1, \ldots, n$, at Step 3 of Algorithm 2.1, in together $2s_k$ component gradients need to be computed at Step 5 of Algorithm 2.1.

The following lemma provides a bound on the difference between the inexact proximal stochastic gradient step given by (2.10) and the exact proximal full gradient step.

Lemma 2.2 *Under assumption* **AS.1**, *we have*

$$\left\|w_t^k-\bar{w}_t^k\right\| \leq \eta_k\left\|v_{t-1}^k-\nabla f\left(w_{t-1}^k\right)\right\|+\sqrt{2\eta_k\varepsilon_t^k},$$

where $\bar{w}_t^k$ *denotes the exact solution to (2.10) with* v_{t-1}^k *replaced by* $\nabla f\left(w_{t-1}^k\right)$.

Proof Since w_t^k is an $\left(\bar{\varepsilon}_t^k, \hat{\varepsilon}_t^k\right)$-solution of the subproblem (2.10), we have from (2.13) and the definition of ε_t^k in (2.12) that

$$q_t^k\left(w_t^k\right)-q_t^k\left(\hat{w}_t^k\right) \leq \varepsilon_t^k,$$

where $\hat{w}_t^k$ is the exact minimizer of q_t^k defined in (2.10). So, by the strong convexity of q_t^k we have

$$\frac{1}{2\eta_k}\left\|w_t^k-\hat{w}_t^k\right\|^2 \leq q_t^k\left(w_t^k\right)-q_t^k\left(\hat{w}_t^k\right) \leq \varepsilon_t^k,$$

which implies

$$\left\|w_t^k - \hat{w}_t^k\right\| \leq \sqrt{2\eta\varepsilon_t^k}. \tag{2.18}$$

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

$$\frac{w_{t-1}^k - \bar{w}_t^k}{\eta_k} - \nabla f\left(w_{t-1}^k\right) \in \partial h\left(\bar{w}_t^k\right) \quad \text{and} \quad \frac{w_{t-1}^k - \hat{w}_t^k}{\eta_k} - v_{t-1}^k \in \partial h\left(\hat{w}_t^k\right).$$

Hence, we have

$$\begin{aligned} h\left(\hat{w}_t^k\right) - h\left(\bar{w}_t^k\right) &\geq \left\langle \frac{w_{t-1}^k - \bar{w}_t^k}{\eta_k} - \nabla f\left(w_{t-1}^k\right), \hat{w}_t^k - \bar{w}_t^k \right\rangle \text{ and} \\ h\left(\bar{w}_t^k\right) - h\left(\hat{w}_t^k\right) &\geq \left\langle \frac{w_{t-1}^k - \hat{w}_t^k}{\eta_k} - v_{t-1}^k, \bar{w}_t^k - \hat{w}_t^k \right\rangle. \end{aligned}$$

Summing up the above two inequalities yields that

$$\begin{aligned} \left\|\bar{w}_t^k - \hat{w}_t^k\right\|^2 &\leq \eta_k \left\langle v_{t-1}^k - \nabla f\left(w_{t-1}^k\right), \bar{w}_t^k - \hat{w}_t^k \right\rangle \\ &\leq \eta_k \left\|v_{t-1}^k - \nabla f\left(w_{t-1}^k\right)\right\| \left\|\bar{w}_t^k - \hat{w}_t^k\right\|, \end{aligned}$$

which gives

$$\left\|\bar{w}_t^k - \hat{w}_t^k\right\| \leq \eta_k \left\|v_{t-1}^k - \nabla f\left(w_{t-1}^k\right)\right\|.$$

Therefore, by (2.18) we have

$$\left\|w_t^k - \bar{w}_t^k\right\| \leq \left\|\bar{w}_t^k - \hat{w}_t^k\right\| + \left\|w_t^k - \hat{w}_t^k\right\| \leq \eta_k \left\|v_{t-1}^k - \nabla f\left(w_{t-1}^k\right)\right\| + \sqrt{2\eta_k\varepsilon_t^k}.$$

□

Remark 2.2 From the above Lemma 2.2, we can derive the following inequality, which will be frequently used in the later analysis:

$$\begin{aligned} &\left\langle v_{t-1}^k - \nabla f\left(w_{t-1}^k\right), w_* - w_t^k \right\rangle \\ &\quad \leq \left\langle v_{t-1}^k - \nabla f\left(w_{t-1}^k\right), w_* - \bar{w}_t^k \right\rangle + \left\|v_{t-1}^k - \nabla f\left(w_{t-1}^k\right)\right\| \left\|\bar{w}_t^k - w_t^k\right\| \\ &\quad \leq \left\langle v_{t-1}^k - \nabla f\left(w_{t-1}^k\right), w_* - \bar{w}_t^k \right\rangle + \eta_k \left\|v_{t-1}^k - \nabla f\left(w_{t-1}^k\right)\right\|^2 \\ &\qquad + \sqrt{2\eta_k\varepsilon_t^k} \left\|v_{t-1} - \nabla f\left(w_{t-1}^k\right)\right\| \\ &\quad \leq \left\langle v_{t-1}^k - \nabla f\left(w_{t-1}^k\right), w_* - \bar{w}_t^k \right\rangle + 2\eta_k \left\|v_{t-1}^k - \nabla f\left(w_{t-1}^k\right)\right\|^2 + \frac{1}{2}\varepsilon_t^k. \end{aligned} \tag{2.19}$$

The following lemma gives a bound on the function value gap $P(w_t^k) - P(w_*)$.

Lemma 2.3 *Under assumption* **AS.1**, *if the proximal parameter* η_k *in Algorithm 2.1 satisfies* $\eta_k \leq 1/L$, *we have*

$$P\left(w_t^k\right) \leq P(w_*) + \frac{1}{2\eta_k}\left(\left\|w_* - w_{t-1}^k\right\|^2 - \left\|w_* - w_t^k\right\|^2\right) + \frac{1}{\eta_k}\left\langle u_t^k, w_* - w_t^k\right\rangle + \left\langle v_{t-1}^k - \nabla f(w_{t-1}^k), w_* - w_t^k\right\rangle + \hat{\varepsilon}_t^k, \quad (2.20)$$

where $\left\|u_t^k\right\| \leq \sqrt{2\eta_k \bar{\varepsilon}_t^k}$.

Proof Since w_t^k is an $\left(\bar{\varepsilon}_t^k, \hat{\varepsilon}_t^k\right)$-solution of the subproblem (2.10), we have from (2.11) that

$$\begin{aligned}
&h(w_*) + \frac{1}{2\eta_k}\left\|w_* - w_{t-1}^k\right\|^2 \\
&\geq h\left(w_t^k\right) + \left\langle \frac{w_{t-1}^k - w_t^k - u_t^k}{\eta_k} - v_{t-1}^k, w_* - w_t^k\right\rangle - \hat{\varepsilon}_t^k + \frac{1}{2\eta_k}\left\|w_* - w_{t-1}^k\right\|^2 \\
&= h\left(w_t^k\right) - \frac{1}{\eta_k}\left\langle u_t^k, w_* - w_t^k\right\rangle - \left\langle v_{t-1}^k, w_* - w_t^k\right\rangle - \hat{\varepsilon}_t^k + \frac{1}{2\eta_k}\left\|w_* - w_t^k\right\|^2 \\
&\quad + \frac{1}{2\eta_k}\left\|w_t^k - w_{t-1}^k\right\|^2,
\end{aligned}$$

where $\|u_t^k\| \leq \sqrt{2\eta_k \bar{\varepsilon}_t^k}$. By rearranging the terms, we have

$$\begin{aligned}
&\frac{1}{2\eta_k}\left\|w_* - w_t^k\right\|^2 \\
&\leq \frac{1}{2\eta_k}\left(\left\|w_* - w_{t-1}^k\right\|^2 - \left\|w_t^k - w_{t-1}^k\right\|^2\right) + h(w_*) - h\left(w_t^k\right) \\
&\quad + \left\langle v_{t-1}^k, w_* - w_t^k\right\rangle + \frac{1}{\eta_k}\left\langle u_t^k, w_* - w_t^k\right\rangle + \hat{\varepsilon}_t^k \\
&\leq \frac{1}{2\eta_k}\left(\left\|w_* - w_{t-1}^k\right\|^2 - \left\|w_t^k - w_{t-1}^k\right\|^2\right) + h(w_*) - h\left(w_t^k\right) \\
&\quad + \left\langle \nabla f\left(w_{t-1}^k\right), w_* - w_t^k\right\rangle + \left\langle v_{t-1}^k - \nabla f(w_{t-1}^k), w_* - w_t^k\right\rangle \\
&\quad + \frac{1}{\eta_k}\left\langle u_t^k, w_* - w_t^k\right\rangle + \hat{\varepsilon}_t^k. \qquad (2.21)
\end{aligned}$$

It follows from the convexity of f that

$$f(w_*) \geq f\left(w_{t-1}^k\right) + \left\langle \nabla f\left(w_{t-1}^k\right), w_* - w_{t-1}^k\right\rangle. \quad (2.22)$$

By the Lipschitz continuity of ∇f with Lipschitz constant L, we have

$$f\left(w_{t-1}^k\right) \geq f\left(w_t^k\right) - \left\langle \nabla f(w_{t-1}^k), w_t^k - w_{t-1}^k \right\rangle - \frac{L}{2}\left\| w_t^k - w_{t-1}^k \right\|^2. \tag{2.23}$$

Then, summing (2.22) and (2.23) yields

$$f(w_*) \geq f\left(w_t^k\right) + \left\langle \nabla f\left(w_{t-1}^k\right), w_* - w_t^k \right\rangle - \frac{L}{2}\left\| w_t^k - w_{t-1}^k \right\|^2.$$

Combing the above inequality and (2.21), we have

$$\begin{aligned}
&\frac{1}{2\eta_k}\left\| w_* - w_t^k \right\|^2 \\
&\quad \leq \frac{1}{2\eta_k}\left\| w_* - w_{t-1}^k \right\|^2 - \left(\frac{1}{2\eta_k} - \frac{L}{2}\right)\left\| w_t^k - w_{t-1}^k \right\|^2 + P(w_*) - P\left(w_t^k\right) \\
&\qquad + \left\langle v_{t-1}^k - \nabla f\left(w_{t-1}^k\right), w_* - w_t^k \right\rangle + \frac{1}{\eta_k}\left\langle u_t^k, w_* - w_t^k \right\rangle + \hat{\varepsilon}_t^k,
\end{aligned}$$

which gives (2.20) due to $\eta_k \leq 1/L$. □

3 Convergence properties for strongly convex case

In this section, we investigate the theoretical properties of Algorithm 2.1 with the additional assumption that f in the objective function (1.1) is μ-strongly convex. Since h in (1.1) is convex, the objective function P is μ-strongly convex, i.e., for any $w \in \mathbb{R}^d$ we have

$$P(w) - P(w_*) \geq \frac{\mu}{2}\|w - w_*\|^2. \tag{3.1}$$

For the strongly convex case in this section, the w_0^{k+1} in Step 8 of Algorithm 2.1 is chosen as

$$w_0^{k+1} = \tilde{w}_k. \tag{3.2}$$

The next theorem gives a recursive relation between $\mathbb{E}[P(\tilde{w}_k)] - P(w_*)$ and $\mathbb{E}[P(\tilde{w}_{k-1})] - P(w_*)$.

Theorem 3.1 *Under assumption* **AS.1**, *if f is μ-strongly convex and the proximal parameter η_k in Algorithm 2.1 satisfies*

$$\eta_k < \min\left\{\frac{s_k}{12L}, \frac{1}{L}\right\}$$

where $s_k \in \{1, \ldots, n\}$, the following property holds:

$$\begin{aligned}&\mathbb{E}[P(\tilde{w}_k)] - P(w_*)\\ &\leq \left(\frac{2s_k}{\mu\eta_k m_k(s_k - 12\eta_k L)} + \frac{12(m_k+1)\eta_k L}{m_k(s_k - 12\eta_k L)}\right)(\mathbb{E}[P(\tilde{w}_{k-1})] - P(w_*))\\ &\quad + \frac{s_k}{m_k(s_k - 12\eta_k L)}A_k,\end{aligned} \tag{3.3}$$

where

$$A_k = \left(\sum_{t=1}^{m_k}\sqrt{\bar{\varepsilon}_t^k}\right)^2 + \frac{3}{2}\sum_{t=1}^{m_k}\left(\sum_{i=t}^{m_k}\sqrt{\bar{\varepsilon}_i^k}\right)^2 + \sum_{t=1}^{m_k}\hat{\varepsilon}_t^k + \frac{3}{2}\sum_{t=1}^{m_k}\varepsilon_t^k \tag{3.4}$$

and the expectation is taken with respect to all the history random variables.

Proof Summing up (2.20) over $t = 1, \ldots, m_k$, we have

$$\begin{aligned}\sum_{t=1}^{m_k}\left(P\left(w_t^k\right) - P(w_*)\right) &\leq \frac{1}{2\eta_k}\left(\left\|w_0^k - w_*\right\|^2 - \left\|w_{m_k}^k - w_*\right\|^2\right)\\ &\quad + \frac{1}{\eta_k}\sum_{t=1}^{m_k}\left\langle u_t^k, w_* - w_t^k\right\rangle + \sum_{t=1}^{m_k}\hat{\varepsilon}_t^k + \sum_{t=1}^{m_k}\left\langle v_{t-1}^k - \nabla f\left(w_{t-1}^k\right), w_* - w_t^k\right\rangle,\end{aligned} \tag{3.5}$$

where $\left\|u_t^k\right\| \leq \sqrt{2\eta_k\bar{\varepsilon}_t^k}$.

We first bound the term $\frac{1}{\eta_k}\sum_{t=1}^{m_k}\left\langle u_t^k, w_* - w_t^k\right\rangle$. By (2.13), $\eta_k < \frac{1}{L}$ and the same proof of [30, Proposition 3], we can obtain

$$\begin{aligned}\left\|w_t^k - w_*\right\| &\leq (1-\mu\eta_k)^t\left(\left\|w_0^k - w_*\right\| + \sum_{i=1}^{t}(1-\mu\eta_k)^{-i}\right.\\ &\quad \left.\times\left(\eta_k\left\|v_{i-1}^k - \nabla f\left(w_{i-1}^k\right)\right\| + \sqrt{2\eta_k\varepsilon_t^k}\right)\right).\end{aligned}$$

Then, it follows from $\mu\eta_k < 1$ due to $\eta_k < \frac{1}{L} \leq \frac{1}{\mu}$ that

$$\left\|w_t^k - w_*\right\| \leq \left\|w_0^k - w_*\right\| + \sum_{i=1}^{t}\left(\eta_k\left\|v_{i-1}^k - \nabla f\left(w_{i-1}^k\right)\right\| + \sqrt{2\eta_k\varepsilon_t^k}\right).$$

Hence, we have

$$\begin{aligned}\frac{1}{\eta_k}\sum_{t=1}^{m_k}\left\langle u_t^k, w_* - w_t^k\right\rangle &\leq \frac{1}{\eta_k}\sum_{t=1}^{m_k}\left\|u_t^k\right\|\left\|w_* - w_t^k\right\|\\ &\leq \frac{1}{\eta_k}\left\|w_0^k - w_*\right\|\sum_{t=1}^{m_k}\sqrt{2\eta_k\bar{\varepsilon}_t^k} + \frac{1}{\eta_k}\sum_{t=1}^{m_k}\sqrt{2\eta_k\bar{\varepsilon}_t^k}\sum_{i=1}^{t}\end{aligned}$$

$$
\begin{aligned}
&\times \left(\eta_k \left\|v_{i-1}^k - \nabla f\left(w_{i-1}^k\right)\right\| + \sqrt{2\eta_k \varepsilon_i^k}\right) \\
&\le \left(\sum_{t=1}^{m_k} \sqrt{\bar{\varepsilon}_t^k}\right)^2 + \frac{1}{2\eta_k}\left\|w_0^k - w_*\right\|^2 + \frac{1}{\eta_k}\sum_{t=1}^{m_k} \\
&\times \left(\sum_{i=t}^{m_k} \sqrt{2\eta_k \bar{\varepsilon}_i^k}\right)\left(\eta_k \left\|v_{t-1}^k - \nabla f\left(w_{t-1}^k\right)\right\| + \sqrt{2\eta_k \varepsilon_t^k}\right) \\
&\le \left(\sum_{t=1}^{m_k} \sqrt{\bar{\varepsilon}_t^k}\right)^2 + \frac{1}{2\eta_k}\left\|w_0^k - w_*\right\|^2 + \sum_{t=1}^{m_k} \sqrt{2\eta_k} \\
&\times \left(\sum_{i=t}^{m_k} \sqrt{\bar{\varepsilon}_i^k}\right)\left\|v_{t-1}^k - \nabla f(w_{t-1}^k)\right\| + 2\sum_{t=1}^{m_k}\left(\sum_{i=t}^{m_k} \sqrt{\bar{\varepsilon}_i^k}\right)\sqrt{\varepsilon_t^k} \\
&\le \left(\sum_{t=1}^{m_k} \sqrt{\bar{\varepsilon}_t^k}\right)^2 + \frac{1}{2\eta_k}\left\|w_0^k - w_*\right\|^2 + \frac{1}{2}\sum_{t=1}^{m_k} \\
&\times \left(\sum_{i=t}^{m_k} \sqrt{\bar{\varepsilon}_i^k}\right)^2 + \eta_k \sum_{t=1}^{m_k}\left\|v_{t-1}^k - \nabla f(w_{t-1}^k)\right\|^2 \\
&+ \sum_{t=1}^{m_k}\left(\sum_{i=t}^{m_k} \sqrt{\bar{\varepsilon}_i^k}\right)^2 + \sum_{t=1}^{m_k} \varepsilon_t^k \\
&= \left(\sum_{t=1}^{m_k} \sqrt{\bar{\varepsilon}_t^k}\right)^2 + \frac{1}{2\eta_k}\left\|w_0^k - w_*\right\|^2 + \frac{3}{2}\sum_{t=1}^{m_k}\left(\sum_{i=t}^{m_k} \sqrt{\bar{\varepsilon}_i^k}\right)^2 + \sum_{t=1}^{m_k} \varepsilon_t^k \\
&+ \eta_k \sum_{t=1}^{m_k}\left\|v_{t-1}^k - \nabla f\left(w_{t-1}^k\right)\right\|^2. \qquad (3.6)
\end{aligned}
$$

It thus follows from (2.19), (3.5), (3.6) and the definition of A_k in (3.4) that

$$
\begin{aligned}
&\sum_{t=1}^{m_k}\left(P\left(w_t^k\right) - P(w_*)\right) \\
&\le \frac{1}{2\eta_k}\left(\left\|w_0^k - w_*\right\|^2 - \left\|w_{m_k}^k - w_*\right\|^2\right) + \frac{1}{2\eta_k}\left\|w_0^k - w_*\right\|^2 \\
&+ 3\eta_k \sum_{t=1}^{m_k}\left\|v_{t-1}^k - \nabla f\left(w_{t-1}^k\right)\right\|^2 \\
&+ \sum_{t=1}^{m_k}\left\langle v_{t-1}^k - \nabla f\left(w_{t-1}^k\right), w_* - \bar{w}_t^k\right\rangle + A_k \\
&\le \frac{1}{\eta_k}\left\|w_0^k - w_*\right\|^2 + 3\eta_k \sum_{t=1}^{m_k}\left\|v_{t-1}^k - \nabla f\left(w_{t-1}^k\right)\right\|^2
\end{aligned}
$$

$$
\begin{aligned}
&+\sum_{t=1}^{m_k}\left\langle v_{t-1}^k-\nabla f\left(w_{t-1}^k\right), w_*-\bar{w}_t^k\right\rangle+A_k \\
&\leq \frac{2}{\mu\eta_k}\left[P(\tilde{w}_{k-1})-P(w_*)\right]+3\eta_k\sum_{t=1}^{m_k}\left\|v_{t-1}^k-\nabla f\left(w_{t-1}^k\right)\right\|^2 \\
&+\sum_{t=1}^{m_k}\left\langle v_{t-1}^k-\nabla f\left(w_{t-1}^k\right), w_*-\bar{w}_t^k\right\rangle+A_k,
\end{aligned}
$$

where the last inequality follows from $w_0^k=\tilde{w}_{k-1}$ and the μ-strong convexity (3.1) of P. Taking expectation on both sides of the above inequality and noticing that

$$
\mathbb{E}\left[\left\langle v_{t-1}^k-\nabla f\left(w_{t-1}^k\right), w_*-\bar{w}_t^k\right\rangle\middle| w_{t-1}^k\right]=0, \tag{3.7}
$$

we obtain

$$
\begin{aligned}
\eta_k\sum_{t=1}^{m_k}\left(\mathbb{E}\left[P\left(w_t^k\right)\right]-P(w_*)\right) &\leq \frac{2}{\mu}\left(\mathbb{E}\left[P(\tilde{w}_{k-1})\right]-P(w_*)\right) \\
&+3\eta_k^2\sum_{t=1}^{m_k}\mathbb{E}\left[\left\|v_{t-1}^k-\nabla f\left(w_{t-1}^k\right)\right\|^2\right]+\eta_k A_k.
\end{aligned}
$$

Then it follows from Lemma 2.1 that

$$
\begin{aligned}
&\eta_k\sum_{t=1}^{m_k}\left(\mathbb{E}P\left(w_t^k\right)-P(w_*)\right) \\
&\leq \frac{2}{\mu}\left(\mathbb{E}\left[P\left(\tilde{w}_{k-1}\right)\right]-P(w_*)\right)+\frac{12\eta_k^2L}{s_k}\sum_{t=1}^{m_k}\left(\mathbb{E}\left[P\left(w_{t-1}^k\right)\right]-P(w_*)\right. \\
&\quad\left.+\mathbb{E}\left[P(\tilde{w}_{k-1})\right]-P(w_*)\right)+\eta_k A_k \\
&=\frac{2}{\mu}\left(\mathbb{E}\left[P(\tilde{w}_{k-1})\right]-P(w_*)\right)+\frac{12\eta_k^2L}{s_k}\left(\mathbb{E}\left[P\left(w_0^k\right)\right]-P(w_*)\right) \\
&\quad+\frac{12\eta_k^2L}{s_k}\sum_{t=2}^{m_k}\left(\mathbb{E}\left[P\left(w_{t-1}^k\right)\right]-P(w_*)\right) \\
&\quad+\frac{12m_k\eta_k^2L}{s_k}\left(\mathbb{E}\left[P(\tilde{w}_{k-1})\right]-P(w_*)\right)+\eta_k A_k.
\end{aligned}
$$

Noticing that $w_0^k=\tilde{w}_{k-1}$ and rearranging terms, we have

$$
\eta_k\left(1-\frac{12\eta_k L}{s_k}\right)\sum_{t=1}^{m_k}\left(\mathbb{E}\left[P\left(w_t^k\right)\right]-P(w_*)\right)\leq\left(\frac{2}{\mu}+\frac{12(m_k+1)\eta_k^2L}{s_k}\right)
$$

$$\times \left(\mathbb{E}\left[P(\tilde{w}_{k-1}) - P(w_*)\right]\right) + \eta_k A_k. \tag{3.8}$$

Since $\tilde{w}_k = \frac{1}{m_k}\sum_{t=1}^{m_k} w_t^k$, it follows from the convexity of P that

$$\sum_{t=1}^{m_k}\left(\mathbb{E}\left[P\left(w_t^k\right)\right] - P(w_*)\right) \geq m_k\left(\mathbb{E}\left[P(\tilde{w}_k)\right] - P(w_*)\right). \tag{3.9}$$

Then, (3.3) follows from $\eta_k < \frac{s_k}{12L}$ and (3.8). □

Motivated from Theorem 3.1, if we further set

$$\bar{\varepsilon}_t^k \leq \varepsilon_t^k\left(1-\sqrt{\mu\eta_k}\right)^{t-1}, \quad \text{for } k = 1, \ldots, N \text{ and } t = 1, \ldots, m_k,$$

then we have

$$\left(\sum_{t=1}^{m_k}\sqrt{\bar{\varepsilon}_t^k}\right)^2 \leq \left(\sum_{t=1}^{m_k}\varepsilon_t^k\right)\left(\sum_{t=1}^{m_k}(1-\sqrt{\mu\eta_k})^{t-1}\right) \leq \frac{1}{\sqrt{\mu\eta_k}}\sum_{t=1}^{m_k}\varepsilon_t^k,$$

and

$$\begin{aligned}\sum_{t=1}^{m_k}\left(\sum_{i=t}^{m_k}\sqrt{\bar{\varepsilon}_i^k}\right)^2 &\leq \sum_{t=1}^{m_k}\left[\left(\sum_{i=t}^{m_k}\varepsilon_i^k\right)\left(\sum_{i=t}^{m_k}(1-\sqrt{\mu\eta_k})^{i-1}\right)\right]\\ &\leq \frac{1}{\sqrt{\mu\eta_k}}\sum_{t=1}^{m_k}\left[\left(\sum_{i=t}^{m_k}\varepsilon_i^k\right)(1-\sqrt{\mu\eta_k})^{t-1}\right]\\ &= \frac{1}{\sqrt{\mu\eta_k}}\sum_{t=1}^{m_k}\left[\varepsilon_t^k\sum_{i=1}^{t}(1-\sqrt{\mu\eta_k})^{i-1}\right] \leq \frac{1}{\mu\eta_k}\sum_{t=1}^{m_k}\varepsilon_t^k.\end{aligned}$$

Then, due to $\mu\eta_k < 1$ and $\hat{\varepsilon}_t^k \leq \varepsilon_t^k$, A_k defined in (3.4) can be upper bounded by

$$A_k \leq \left(\frac{1}{\sqrt{\mu\eta_k}} + \frac{3}{2\mu\eta_k} + \frac{5}{2}\right)\sum_{t=1}^{m_k}\varepsilon_t^k \leq \frac{5}{\mu\eta_k}\sum_{t=1}^{m_k}\varepsilon_t^k.$$

Hence, by (3.3) we have

$$\begin{aligned}\mathbb{E}\left[P(\tilde{w}_k)\right] - P(w_*) \leq &\left(\frac{2s_k}{\mu\eta_k m_k(s_k - 12\eta_k L)} + \frac{12(m_k+1)\eta_k L}{m_k(s_k-12\eta_k L)}\right)\left(\mathbb{E}\left[P(\tilde{w}_{k-1})\right] - P(w_*)\right)\\ &+ \frac{5s_k}{\mu\eta_k m_k(s_k-12\eta_k L)}\sum_{t=1}^{m_k}\varepsilon_t^k,\end{aligned} \tag{3.10}$$

which would lead to the following theorem.

Theorem 3.2 *Under assumption* **AS.1**, *if f is μ-strongly convex and the parameters in Algorithm 2.1 are set as*

$$\eta_k < \min\left\{\frac{s_k}{40L}, \frac{1}{L}\right\}, \quad m_k > \frac{20\theta}{\mu\eta_k} \text{ for some } \theta > 1, \quad \text{and} \quad \bar{\varepsilon}_t^k = \varepsilon_t^k(1-\sqrt{\mu\eta_k})^{t-1}, \tag{3.11}$$

where $s_k \in \{1, \ldots, n\}$, then we have

$$\mathbb{E}[P(\tilde{w}_k)] - P(w_*) \le \gamma^k(\mathbb{E}[P(\tilde{w}_0)] - P(w_*)) + \frac{5}{2}\sum_{i=1}^{k}\sum_{t=1}^{m}\varepsilon_t^i\gamma^{k-i+1}, \tag{3.12}$$

where

$$\gamma = \frac{1+6\theta}{7\theta} < 1.$$

Proof Since $\eta_k < s_k/(40L)$, we have that

$$\frac{s_k}{s_k - 12\eta_k L} < \frac{10}{7} \quad \text{and} \quad \frac{\eta_k L}{s_k - 12\eta_k L} < \frac{1}{28}.$$

Hence, denoting

$$\gamma_k = \frac{2s_k}{\mu\eta_k m_k(s_k - 12L\eta_k)} + \frac{12(m_k+1)L\eta_k}{m_k(s_k - 12L\eta_k)}, \tag{3.13}$$

we have from $m_k > \frac{20\theta}{\mu\eta_k}$ and $\theta > 1$ that

$$\gamma_k < \frac{20}{7\mu\eta_k m_k} + \frac{12(m_k+1)}{28m_k} \le \frac{20}{7\mu\eta_k m_k} + \frac{24}{28} < \frac{1+6\theta}{7\theta} = \gamma < 1.$$

Notice that the coefficient of the last term in (3.10) is less than $\frac{5}{2}\gamma_k$. Hence, by (3.10) we have

$$\mathbb{E}[P(\tilde{w}_k)] - P(w_*) \le \gamma_k(\mathbb{E}[P(\tilde{w}_{k-1})] - P(w_*)) + \frac{5\gamma_k}{2}\sum_{t=1}^{m_k}\varepsilon_t^k. \tag{3.14}$$

Then, (3.12) follows from induction and $\gamma_k < \gamma < 1$. □

Remark 3.1 From (3.13), we can see that for fixed η_k and m_k, γ_k will decrease as s_k increases. Hence, by (3.14), increasing the sample size s_k when calculating the stochastic gradient in Step 5 of Algorithm 2.1 will generally improve the convergence speed of $\mathbb{E}[P(\tilde{w}_k)] - P(w_*)$ to zero.

By Theorem 3.2, to ensure $\mathbb{E}[P(\tilde{w}_k)] - P(w_*)$ converges to zero, it is sufficient to require $\sum_{i=1}^{k}\sum_{t=1}^{m_k}\varepsilon_t^i\gamma^{k-i+1}$ converges to zero as k increases to infinity. This gives

us certain freedom for the choices of ε_t^k. In the following, we analyze several different choices of ε_t^k and derive its corresponding complexity bounds.

Corollary 3.3 *Under assumption* **AS.1**, *suppose* f *is* μ*-strongly convex and the parameters in Algorithm 2.1 are set as in (3.11). Then,*

(1) *if the subproblem tolerance* ε_t^k *satisfies*

$$\varepsilon_t^k \leq \alpha^{k+t} \quad \text{with } \alpha \in (0, 1), \tag{3.15}$$

we have

$$\mathbb{E}[P(\tilde{w}_k)] - P(w_*) \leq \gamma^k (P(\tilde{w}_0) - P(w_*)) + \frac{5\alpha \max\{\gamma^2, \alpha\}}{2(1-\alpha)(1-\max\{\gamma, \sqrt{\alpha}\})} \max\{\sqrt{\alpha}, \gamma\}^k; \tag{3.16}$$

(2) *if the subproblem tolerance* ε_t^k *satisfies*

$$\varepsilon_t^k \leq \frac{1}{k^\beta} \cdot \alpha^t \quad \text{with } \alpha \in (0, 1) \quad \text{and} \quad \beta > 0, \tag{3.17}$$

we have

$$\mathbb{E}[P(\tilde{w}_k)] - P(w_*) \leq \gamma^k (P(\tilde{w}_0) - P(w_*)) + \frac{5\alpha\gamma}{(1-\alpha)(1-\gamma)} \xi(k)^{-\beta}, \tag{3.18}$$

where $\xi(k)$ *is a scalar satisfying*

$$\gamma^{k-\xi(k)} = \xi(k)^{-\beta}; \tag{3.19}$$

(3) *if the subproblem tolerance* ε_t^k *satisfies*

$$\varepsilon_t^k \leq \frac{1}{k^\beta} \cdot \frac{1}{t^{1+\phi}} \quad \text{with } \beta > 0 \quad \text{and} \quad \phi > 0, \tag{3.20}$$

we have

$$\mathbb{E}[P(\tilde{w}_k)] - P(w_*) \leq \gamma^k (P(\tilde{w}_0) - P(w_*)) + \frac{5\gamma(1+\phi^{-1})}{(1-\gamma)} \xi(k)^{-\beta}, \tag{3.21}$$

where $\xi(k)$ *is a scalar satisfying (3.19);*

(4) *if the subproblem tolerance* ε_t^k *satisfies*

$$\varepsilon_t^k \leq \frac{1}{(k+t)^\kappa} \quad \text{with } \kappa > 1, \tag{3.22}$$

we have

$$\mathbb{E}[P(\tilde{w}_k)] - P(w_*) \leq \gamma^k (P(\tilde{w}_0) - P(w_*)) + \frac{5\gamma}{(\kappa-1)(1-\gamma)} \xi(k)^{1-\kappa}, \quad (3.23)$$

where $\xi(k)$ is a scalar satisfying (3.19).

Proof We now analyze the estimate bound for $\mathbb{E}[P(\tilde{w}_k)] - P(w_*)$ case by case.

(1) If the tolerance ε_t^k satisfies (3.15), direct calculations show that $\sum_{t=1}^{m_k} \varepsilon_t^k \leq \frac{\alpha}{1-\alpha}\alpha^k$. Then

$$\begin{aligned}\sum_{i=1}^{k} \gamma^{k-i+1}\alpha^i &= \sum_{i=1}^{k} \gamma^{k-i+1} \left(\sqrt{\alpha}\right)^{2i} \leq \sum_{i=1}^{k} (\max\{\gamma, \sqrt{\alpha}\})^{k+i+1} \\ &\leq \frac{\max\{\gamma^2, \alpha\}}{1 - \max\{\gamma, \sqrt{\alpha}\}} \max\{\gamma, \sqrt{\alpha}\}^k.\end{aligned}$$

So, (3.16) follows from (3.12) directly.

(2) If the tolerance ε_t^k satisfies (3.17), then $\sum_{t=1}^{m_k} \varepsilon_t^k \leq \frac{\alpha}{1-\alpha} k^{-\beta}$. It follows from (3.12) that

$$\mathbb{E}[P(\tilde{w}_k)] - P(w_*) \leq \gamma^k (P(\tilde{w}_0) - P(w_*)) + \frac{5\alpha}{2(1-\alpha)} \sum_{i=1}^{k} \gamma^{k-i+1} i^{-\beta}.$$

We now compute $\sum_{i=1}^{k} \gamma^{k-i+1} i^{-\beta}$. Since $\xi(k)$ satisfies $\gamma^{k-\xi(k)} = \xi(k)^{-\beta}$, we have $\xi(k) \in (0, k)$, and

$$\begin{aligned}\sum_{i=1}^{k} \gamma^{k-i+1} i^{-\beta} &= \sum_{i=1}^{\lfloor \xi(k) \rfloor} \gamma^{k-i+1} i^{-\beta} + \sum_{i=\lfloor \xi(k) \rfloor + 1}^{k} \gamma^{k-i+1} i^{-\beta} \\ &\leq \sum_{i=1}^{\lfloor \xi(k) \rfloor} \gamma^{k-i+1} + (\lfloor \xi(k) \rfloor + 1)^{-\beta} \sum_{i=\lfloor \xi(k) \rfloor + 1}^{k} \gamma^{k-i+1} \\ &\leq \frac{\gamma^{k-\lfloor \xi(k) \rfloor + 1}}{1-\gamma} + \frac{\gamma}{1-\gamma} (\lfloor \xi(k) \rfloor + 1)^{-\beta} \\ &\leq \frac{2\gamma}{1-\gamma} \xi(k)^{-\beta}. \end{aligned} \quad (3.24)$$

Therefore, we obtain (3.18).

(3) If the tolerance ε_t^k satisfies (3.20), we have

$$\sum_{t=1}^{m_k} \varepsilon_t^k = k^{-\beta} \left(1 + \sum_{t=2}^{m_k} t^{-(1+\phi)}\right) \leq k^{-\beta} \left(1 + \int_1^{m_k} x^{-(1+\phi)} dx\right) \leq (1 + \phi^{-1}) k^{-\beta}. \quad (3.25)$$

Then, (3.21) follows from (3.12) and (3.24).

(4) If the tolerance ε_t^k satisfies (3.22), we have

$$\sum_{t=1}^{m_k} \varepsilon_t^k = \sum_{t=1}^{m_k} (k+t)^{-\kappa} \leq \int_0^{m_k} (k+x)^{-\kappa} dx = \frac{k^{1-\kappa}}{\kappa-1} - \frac{(k+m_k)^{1-\kappa}}{\kappa-1} \leq \frac{k^{1-\kappa}}{\kappa-1}. \tag{3.26}$$

Then, (3.23) follows from (3.12) and (3.24).

□

Remark 3.2 Notice that for any given $\lambda \in (0, 1)$, we have

$$\lim_{k\to\infty} \frac{\gamma^{(1-\lambda)k}}{(\lambda k)^{-\beta}} = 0.$$

So, by the definition of $\xi(k)$ in (3.19), there exists an integer $K > 0$ such that

$$\xi(k) \geq \lambda k \quad \text{for all } k \geq K,$$

which yields that

$$\xi(k)^{-\beta} \leq (\lambda k)^{-\beta}.$$

Therefore, all the upper bounds in (3.18), (3.21) and (3.23) converge to zero as k increases to infinity.

Remark 3.3 We are more interested in the complexity bounds of the total number of component gradient evaluations to reduce the function value gap $\mathbb{E}[P(\tilde{w}_k)] - P(w_*)$ below certain tolerance. This complexity bound is also called batch complexity, which often measures the dominating cost of solving original problem (1.1). From our theoretical analysis, it seems reasonable to choose the parameters m_k and s_k such that $m_k s_k = \mathcal{O}(L/\mu)$. With this choice of m_k and s_k, let us discuss the component gradient complexity bounds implied by Corollary 3.3 under different accuracies of solving the proximal subproblem (2.10).

- For the tolerance setting (3.15), to achieve $\mathbb{E}[P(\tilde{w}_N)] - P(w_*) < \varepsilon$ for some $\varepsilon > 0$, by (3.16) the maximum outer iteration number N should satisfy

  $$\max\{\sqrt{\alpha}, \gamma\}^N = \mathcal{O}(\varepsilon),$$

  which implies that $N = \mathcal{O}(\log(1/\varepsilon))$. Hence, by the fact that $m_k s_k$ is on the order L/μ, the total number of component gradient evaluations $Nn + 2\sum_{k=1}^N m_k s_k$ is in the order of

  $$\left(n + \frac{L}{\mu}\right) \mathcal{O}\left(\log\left(\frac{1}{\varepsilon}\right)\right). \tag{3.27}$$

  Note that when all the subproblems (2.10) are solved exactly, our Algorithm 2.1 will be reduced to the method proposed in [36]. The same complexity bound (3.27) was obtained in [36], but it required to solve all the subproblems exactly. Hence, based on our results, the algorithm could have the same complexity bounds even

without solving the subproblems exactly. When the subproblem does not have a closed-form solution or is too expensive to be solved exactly, allowance of solving the subproblems inexactly would be crucial for both efficiency and stableness of the algorithm.

- For the tolerance setting (3.17), to achieve $\mathbb{E}[P(\tilde{w}_N)] - P(w_*) < \varepsilon$, by (3.18) and Remark 3.2, when $\varepsilon > 0$ is sufficiently small, it suffices to require

$$\gamma^N = \mathcal{O}(\varepsilon) \quad \text{and} \quad N^{-\beta} = \mathcal{O}(\varepsilon).$$

This implies the outer iteration number N should be in the order of $\mathcal{O}(\varepsilon^{-1/\beta})$ when N is large. Thus, for a sufficiently small $\varepsilon > 0$, the component gradient complexity bound is

$$\left(n + \frac{L}{\mu}\right) \mathcal{O}\left(\frac{1}{\varepsilon^{1/\beta}}\right). \tag{3.28}$$

- For the tolerance setting (3.20), to achieve $\mathbb{E}[P(\tilde{w}_N)] - P(w_*) < \varepsilon$, by (3.21) and Remark 3.2, when $\varepsilon > 0$ is sufficiently small, it suffices to require

$$\gamma^N = \mathcal{O}(\varepsilon) \quad \text{and} \quad (1 + \phi^{-1})N^{-\beta} = \mathcal{O}(\varepsilon).$$

This implies the outer iteration number N should in the order of $\mathcal{O}((\rho\varepsilon)^{-1/\beta})$, where $\rho = \min\{1, \phi\}$. Thus, for a sufficient small $\varepsilon > 0$, the component gradient complexity bound is

$$\left(n + \frac{L}{\mu}\right) \mathcal{O}\left(\frac{1}{(\rho\varepsilon)^{1/\beta}}\right). \tag{3.29}$$

- For the tolerance setting (3.22), similar to above two cases, following from (3.23) and Remark 3.2, to obtain $\mathbb{E}[P(\tilde{w}_N)] - P(w_*) < \varepsilon$ for a sufficiently small $\varepsilon > 0$, it suffices to require

$$\gamma^N = \mathcal{O}(\varepsilon) \quad \text{and} \quad \frac{1}{\kappa - 1} N^{1-\kappa} = \mathcal{O}(\varepsilon).$$

This implies N should be in the order of $\mathcal{O}(((\kappa - 1)\varepsilon)^{-1/(\kappa-1)})$ and the component gradient complexity bound is

$$\left(n + \frac{L}{\mu}\right) \mathcal{O}\left(\frac{1}{((\kappa - 1)\varepsilon)^{1/(\kappa-1)}}\right). \tag{3.30}$$

Note that compared with the latter three cases, the tolerance (3.15) has the most accuracy. Intuitively, it should have the smallest complexity bound, which is verified by (3.27). Meanwhile, we can see that as less subproblem accuracy is required, the corresponding component gradient complexity bound increases. Hence, proper subproblem accuracy setting should depend on the balance of the cost of solving the subproblems and the cost of component gradient evaluations.

Remark 3.4 It also deserves to mention the inexact proximal gradient methods proposed in [30]. To obtain a linear convergence rate analogous to (3.16) when the

objective is strongly convex, the linear decrease rate on the subproblem accuracies similar to (3.15) is also required in [30]. But methods proposed in [30] are deterministic methods, which do not consider the particular summation structure of our objective function in (1.1). Furthermore, by combining with variance reduction techniques, we show the more relaxed subproblem accuracies, such as (3.17), (3.20) and (3.22), can be applied without loosing the convergence, which have not been discussed in [30].

4 Convergence properties for general convex case

In this section, we discuss the theoretical properties of Algorithm 2.1 for solving the problem (1.1)–(1.2) without assuming the strong convexity of f. Different from (3.2), for the general convex case in this section, the w_0^{k+1} in Step 8 of Algorithm 2.1 is chosen as

$$w_0^{k+1} = w_{m_k}^k. \tag{4.1}$$

And given an initial positive integer m_0, we require the succeeded inner iteration number m_k, $k = 1, 2, \ldots$, in Algorithm 2.1 is increasing and satisfies

$$m_k = (1 + \theta_k) m_{k-1}, \tag{4.2}$$

where $\theta_k > 0$ is a parameter.

First, the following lemma [30, Lemma 1] will be used in our analysis to bound $\|w_t^k - w_*\|$.

Lemma 4.1 *Assume that the nonnegative sequence $\{v_k\}$ satisfies the following recursion for all $k \geq 1$:*

$$v_k^2 \leq S_k + \sum_{i=1}^{k} \lambda_i v_i,$$

with $\{S_k\}$ an increasing sequence, $S_0 \geq u_0^2$ and $\lambda_i \geq 0$ for all i. Then for all $k \geq 1$,

$$v_k \leq \frac{1}{2} \sum_{i=1}^{k} \lambda_i + \left(S_k + \left(\frac{1}{2} \sum_{i=1}^{k} \lambda_i \right)^2 \right)^{1/2}.$$

Based on Lemma 4.1, we give the following bound on $\|w_t^k - w_*\|$.

Lemma 4.2 *Under assumption* **AS.1**, *if the proximal parameter η_k in Algorithm 2.1 satisfies $\eta_k \leq 1/L$, then*

$$\begin{aligned} \left\| w_t^k - w_* \right\| \leq & \sum_{i=1}^{t} \left(\eta_k \left\| v_{i-1}^k - \nabla f\left(w_{i-1}^k \right) \right\| + \sqrt{2\eta_k \bar{\varepsilon}_i^k} \right) \\ & + \left(\left\| w_0^k - w_* \right\|^2 + 2\eta_k \sum_{i=1}^{t} \hat{\varepsilon}_i^k + \left(\sum_{i=1}^{t} \left(\eta_k \left\| v_{i-1}^k - \nabla f\left(w_{i-1}^k \right) \right\| + \sqrt{2\eta_k \bar{\varepsilon}_i^k} \right) \right)^2 \right)^{1/2}. \end{aligned} \tag{4.3}$$

Proof Note that (2.20) together with $\left\|u_t^k\right\| \leq \sqrt{2\eta_k\bar{\varepsilon}_t^k}$ give

$$\begin{aligned}\left\|w_i^k - w_*\right\|^2 &\leq \left\|w_{i-1}^k - w_*\right\|^2 \\ &+ \left(2\eta_k \left\|v_{i-1}^k - \nabla f\left(w_{i-1}^k\right)\right\| + 2\sqrt{2\eta_k\bar{\varepsilon}_i^k}\right)\left\|w_i^k - w_*\right\| + 2\eta_k\hat{\varepsilon}_i^k.\end{aligned}$$

Summing up the above inequality over $i = 1, \ldots, t$, we have

$$\begin{aligned}\left\|w_t^k - w_*\right\|^2 \leq \left\|w_0^k - w_*\right\|^2 &+ \sum_{i=1}^{t}\left(2\eta_k \left\|v_{i-1}^k - \nabla f\left(w_{i-1}^k\right)\right\| + 2\sqrt{2\eta_k\bar{\varepsilon}_i^k}\right)\left\|w_i^k - w_*\right\| \\ &+ 2\eta_k\sum_{i=1}^{t}\hat{\varepsilon}_i^k.\end{aligned}$$

By letting $v_i = \left\|w_i^k - w_*\right\|$, $S_t = \left\|w_0^k - w_*\right\|^2 + 2\eta_k\sum_{i=1}^{t}\hat{\varepsilon}_i^k$ and $\lambda_i = 2\sqrt{2\eta_k\bar{\varepsilon}_i^k} + 2\eta_k\left\|v_{i-1}^k - \nabla f(w_{i-1}^k)\right\|$ in Lemma 4.1, we obtain (4.3). □

Now we give the following theorem which is analogous to Theorem 3.1 for the strongly convex case.

Theorem 4.1 *Under assumption* **AS.1**, *if the proximal parameter* η_k *in Algorithm 2.1 satisfies*

$$\eta_k < \min\left\{\frac{s_k}{16L(\theta_k+2)}, \frac{1}{L}\right\}, \tag{4.4}$$

the following property holds:

$$\begin{aligned}&\mathbb{E}[P(\tilde{w}_k)] - P(w_*) + \frac{s_k}{2\eta_k m_k(s_k - 16\eta_k L)}\mathbb{E}\left[\left\|w_0^{k+1} - w_*\right\|^2\right] \\ &\quad + \frac{16\eta_k L}{m_k(s_k - 16\eta_k L)}\left(\mathbb{E}\left[P\left(w_0^{k+1}\right)\right] - P(w_*)\right) \\ &\quad \leq \frac{1+\sqrt{\eta_k}}{1+\theta_k}R_k + \frac{2}{m_k}B_k,\end{aligned} \tag{4.5}$$

where

$$\begin{aligned}R_k = \mathbb{E}[P(\tilde{w}_{k-1})] - P(w_*) &+ \frac{s_k}{2\eta_k m_{k-1}(s_k - 16\eta_k L)}\mathbb{E}\left[\left\|w_0^k - w_*\right\|^2\right] \\ &+ \frac{16\eta_k L}{m_{k-1}(s_k - 16\eta_k L)}\left(\mathbb{E}\left[P\left(w_0^k\right)\right] - P(w_*)\right)\end{aligned} \tag{4.6}$$

and

$$B_k = \frac{1}{\sqrt{\eta_k}}\left(\sum_{t=1}^{m_k}\sqrt{\bar{\varepsilon}_t^k}\right)^2 + 4\sum_{t=1}^{m_k}\left(\sum_{i=t}^{m_k}\sqrt{\bar{\varepsilon}_i^k}\right)^2 + \frac{5}{2}\sum_{t=1}^{m_k}\varepsilon_t^k. \tag{4.7}$$

Here, the expectation is taken with respect to all the history random variables.

Proof Summing up (2.20) over $t = 1, \ldots, m_k$, due to $w_0^{k+1} = w_{m_k}^k$ we have

$$\begin{aligned}
&\sum_{t=1}^{m_k}\left(P\left(w_t^k\right) - P(w_*)\right) \\
&\quad\le \frac{1}{2\eta_k}\left(\left\|w_0^k - w_*\right\|^2 - \left\|w_0^{k+1} - w_*\right\|^2\right) + \frac{1}{\eta_k}\sum_{t=1}^{m_k}\left\langle u_t^k, w_* - w_t^k\right\rangle \\
&\qquad + \sum_{t=1}^{m_k}\left\langle v_{t-1}^k - \nabla f\left(w_{t-1}^k\right), w_* - w_t^k\right\rangle + \sum_{t=1}^{m_k}\hat{\varepsilon}_t^k,
\end{aligned} \tag{4.8}$$

where $\left\|u_t^k\right\| \le \sqrt{2\eta_k\bar{\varepsilon}_t^k}$.

We first give a bound on the term $\frac{1}{\eta_k}\sum_{t=1}^{m_k}\left\langle u_t^k, w_* - w_t^k\right\rangle$. The upper bound of $\|u_t^k\|$ in (4.8) and Lemma 4.2 give

$$\frac{1}{\eta_k}\sum_{t=1}^{m_k}\left\langle u_t^k, w_* - w_t^k\right\rangle \le \frac{1}{\eta_k}\sum_{t=1}^{m_k}\sqrt{2\eta\bar{\varepsilon}_t^k}\left\|w_* - w_t^k\right\| \le E_1 + E_2, \tag{4.9}$$

where

$$E_1 := \frac{1}{\eta_k}\sum_{t=1}^{m_k}\sqrt{2\eta_k\bar{\varepsilon}_t^k}\sum_{i=1}^{t}\left(\eta_k\left\|v_{i-1}^k - \nabla f\left(w_{i-1}^k\right)\right\| + \sqrt{2\eta_k\bar{\varepsilon}_i^k}\right),$$

$$E_2 := \frac{1}{\eta_k}\sum_{t=1}^{m_k}\sqrt{2\eta_k\bar{\varepsilon}_t^k}\left(\left\|w_0^k - w_*\right\|^2 + 2\eta_k\sum_{i=1}^{t}\hat{\varepsilon}_i^k + \left(\sum_{i=1}^{t}\left(\eta_k\left\|v_{i-1}^k - \nabla f\left(w_{i-1}^k\right)\right\| + \sqrt{2\eta_k\bar{\varepsilon}_i^k}\right)\right)^2\right)^{1/2}.$$

Now, let us first derive a bound on E_1:

$$\begin{aligned}
E_1 &= \sum_{t=1}^{m_k}\left(\sum_{i=t}^{m_k}\sqrt{2\eta_k\bar{\varepsilon}_i^k}\right)\left\|v_{t-1}^k - \nabla f\left(w_{t-1}^k\right)\right\| + 2\sum_{t=1}^{m_k}\sum_{i=t}^{m_k}\sqrt{\bar{\varepsilon}_i^k\bar{\varepsilon}_t^k} \\
&\le \frac{1}{2}\sum_{t=1}^{m_k}\left(\sum_{i=t}^{m_k}\sqrt{\bar{\varepsilon}_i^k}\right)^2 + \eta_k\sum_{t=1}^{m_k}\left\|v_{t-1}^k - \nabla f\left(w_{t-1}^k\right)\right\|^2 + \sum_{t=1}^{m_k}\bar{\varepsilon}_t^k + \sum_{t=1}^{m_k}\left(\sum_{i=t}^{m_k}\bar{\varepsilon}_i^k\right)^2 \\
&= \frac{3}{2}\sum_{t=1}^{m_k}\left(\sum_{i=t}^{m_k}\sqrt{\bar{\varepsilon}_i^k}\right)^2 + \eta_k\sum_{t=1}^{m_k}\left\|v_{t-1}^k - \nabla f\left(w_{t-1}^k\right)\right\|^2 + \sum_{t=1}^{m_k}\bar{\varepsilon}_t^k.
\end{aligned}$$

Noticing the definition of ε_t^k in (2.12), we have the following bound on E_2:

$$E_2 \le \frac{1}{\eta_k}\sum_{t=1}^{m_k}\sqrt{2\eta_k\bar{\varepsilon}_t^k}\left(\left\|w_0^k - w_*\right\| + \sqrt{2\eta_k}\sum_{i=1}^{t}\sqrt{\hat{\varepsilon}_i^k} + \sum_{i=1}^{t}\left(\eta_k\left\|v_{i-1}^k - \nabla f\left(w_{i-1}^k\right)\right\| + \sqrt{2\eta_k\bar{\varepsilon}_i^k}\right)\right)$$

$$
\begin{aligned}
&\leq \sum_{t=1}^{m_k}\sqrt{\frac{2\bar{\varepsilon}_t^k}{\eta_k}}\left\|w_0^k-w_*\right\|+2\sum_{t=1}^{m_k}\sum_{i=1}^{t}\sqrt{\bar{\varepsilon}_t^k\hat{\varepsilon}_i^k}+\sum_{t=1}^{m_k}\sum_{i=1}^{t}\sqrt{2\eta_k\bar{\varepsilon}_t^k}\left\|v_{i-1}^k-\nabla f\left(w_{i-1}^k\right)\right\|+2\sum_{t=1}^{m_k}\sum_{i=1}^{t}\sqrt{\bar{\varepsilon}_t^k\bar{\varepsilon}_i^k}\\
&\leq \frac{\left\|w_0^k-w_*\right\|^2}{2\sqrt{\eta_k}}+\frac{1}{\sqrt{\eta_k}}\left(\sum_{t=1}^{m_k}\sqrt{\bar{\varepsilon}_t^k}\right)^2+2\sum_{t=1}^{m_k}\sum_{i=t}^{m_k}\sqrt{\hat{\varepsilon}_t^k\bar{\varepsilon}_i^k}+\sum_{t=1}^{m_k}\left(\sum_{i=t}^{m_k}\sqrt{2\eta_k\bar{\varepsilon}_i^k}\right)\left\|v_{t-1}^k-\nabla f\left(w_{t-1}^k\right)\right\|\\
&\quad+2\sum_{t=1}^{m_k}\sum_{i=t}^{m_k}\sqrt{\bar{\varepsilon}_t^k\bar{\varepsilon}_i^k}\\
&\leq \frac{\left\|w_0^k-w_*\right\|^2}{2\sqrt{\eta_k}}+\frac{1}{\sqrt{\eta_k}}\left(\sum_{t=1}^{m_k}\sqrt{\bar{\varepsilon}_t^k}\right)^2+\sum_{t=1}^{m_k}\hat{\varepsilon}_t^k+\sum_{t=1}^{m_k}\left(\sum_{i=t}^{m_k}\sqrt{\bar{\varepsilon}_i^k}\right)^2+\eta_k\sum_{t=1}^{m_k}\left\|v_{t-1}^k-\nabla f\left(w_{t-1}^k\right)\right\|^2\\
&\quad+\frac{1}{2}\sum_{t=1}^{m_k}\left(\sum_{i=t}^{m_k}\sqrt{\bar{\varepsilon}_i^k}\right)^2+\sum_{t=1}^{m_k}\bar{\varepsilon}_t^k+\sum_{t=1}^{m_k}\left(\sum_{i=t}^{m_k}\sqrt{\bar{\varepsilon}_i^k}\right)^2\\
&\leq \frac{\left\|w_0^k-w_*\right\|^2}{2\sqrt{\eta_k}}+\frac{1}{\sqrt{\eta_k}}\left(\sum_{t=1}^{m_k}\sqrt{\bar{\varepsilon}_t^k}\right)^2+\sum_{t=1}^{m_k}\varepsilon_t^k+\frac{5}{2}\sum_{t=1}^{m_k}\left(\sum_{i=t}^{m_k}\sqrt{\bar{\varepsilon}_i^k}\right)^2+\eta_k\sum_{t=1}^{m_k}\left\|v_{t-1}^k-\nabla f\left(w_{t-1}^k\right)\right\|^2.
\end{aligned}
$$

By inserting the above two bounds on E_1 and E_2 into (4.9), we have

$$
\begin{aligned}
\frac{1}{\eta_k}\sum_{t=1}^{m_k}\left\langle u_t^k, w_*-w_t^k\right\rangle \leq{}& \frac{\left\|w_0^k-w_*\right\|^2}{2\sqrt{\eta_k}}+2\eta_k\sum_{t=1}^{m_k}\left\|v_{t-1}^k-\nabla f(w_{t-1}^k)\right\|^2\\
&+\frac{1}{\sqrt{\eta_k}}\left(\sum_{t=1}^{m}\sqrt{\bar{\varepsilon}_t^k}\right)^2+4\sum_{t=1}^{m_k}\left(\sum_{i=t}^{m_k}\sqrt{\bar{\varepsilon}_i^k}\right)^2\\
&+\sum_{t=1}^{m_k}\varepsilon_t^k+\sum_{t=1}^{m_k}\bar{\varepsilon}_t^k. \qquad (4.10)
\end{aligned}
$$

Combining (2.19), (4.8) and (4.10) gives

$$
\begin{aligned}
\sum_{t=1}^{m_k}\left(P\left(w_t^k\right)-P(w_*)\right)\leq{}& \frac{1+\sqrt{\eta_k}}{2\eta_k}\left\|w_0^k-w_*\right\|^2-\frac{1}{2\eta_k}\left\|w_0^{k+1}-w_*\right\|^2\\
&+4\eta_k\sum_{t=1}^{m_k}\left\|v_{t-1}^k-\nabla f\left(w_{t-1}^k\right)\right\|^2\\
&+\sum_{t=1}^{m_k}\left\langle v_{t-1}^k-\nabla f\left(w_{t-1}^k\right), w_*-\bar{w}_t^k\right\rangle+B_k, \qquad (4.11)
\end{aligned}
$$

where B_k is defined in (4.7). Taking expectation on both sides of (4.11) and noticing (3.7), it follows from Lemma 2.1 that

$$
\begin{aligned}
&\sum_{t=1}^{m_k}(\mathbb{E}[P(w_t^k)]-P(w_*))\\
&\quad\leq \frac{1+\sqrt{\eta_k}}{2\eta_k}\mathbb{E}[\|w_0^k-w_*\|^2]-\frac{1}{2\eta_k}\mathbb{E}[\|w_0^{k+1}-w_*\|^2]
\end{aligned}
$$

$$+\frac{16\eta_k L}{s_k}\sum_{t=1}^{m_k}(\mathbb{E}[P(w_{t-1}^k)]-P(w_*))$$
$$+\frac{16m_k\eta_k L}{s_k}(\mathbb{E}P(\tilde{w}_{k-1})-P(w_*))+B_k.$$

Then, by $\eta_k < \frac{s_k}{16L}$ and $w_0^{k+1}=w_{m_k}^k$, we have

$$\begin{aligned}&\left(1-\frac{16\eta_k L}{s_k}\right)\sum_{t=1}^{m_k}(\mathbb{E}[P(w_t^k)]-P(w_*))\\&\quad+\frac{1}{2\eta_k}\mathbb{E}[\|w_0^{k+1}-w_*\|^2]+\frac{16\eta_k L}{s_k}(\mathbb{E}[P(w_0^{k+1})]-P(w_*))\\&\le\frac{16m_k\eta_k L}{s_k}(\mathbb{E}[P(\tilde{w}_{k-1})]-P(w_*))\\&\quad+\frac{1+\sqrt{\eta_k}}{2\eta_k}\mathbb{E}[\|w_0^k-w_*\|^2]+\frac{16\eta_k L}{s_k}(\mathbb{E}[P(w_0^k)]-P(w_*))+B_k.\end{aligned}\tag{4.12}$$

Then, it follows from (3.9) and (4.12) that

$$\begin{aligned}&\mathbb{E}[P(\tilde{w}_k)]-P(w_*)+\frac{s_k}{2\eta_k m_k(s_k-16\eta_k L)}\mathbb{E}[\|w_0^{k+1}-w_*\|^2]\\&\quad+\frac{16\eta_k L}{m_k(s_k-16\eta_k L)}(\mathbb{E}[P(w_0^{k+1})]-P(w_*))\\&\le\frac{16\eta_k L}{s_k-16\eta_k L}(\mathbb{E}[P(\tilde{w}_{k-1})]-P(w_*))+\frac{s_k(1+\sqrt{\eta_k})}{2\eta_k m_k(s_k-16\eta_k L)}\mathbb{E}[\|w_0^k-w_*\|^2]\\&\quad+\frac{16\eta_k L}{m_k(s_k-16\eta_k L)}(\mathbb{E}[P(w_0^k)]-P(w_*))+\frac{s_k}{m_k(s_k-16\eta_k L)}B_k.\end{aligned}\tag{4.13}$$

Since $\eta_k < \frac{s_k}{16L(\theta_k+2)}$, we have $s_k-16\eta_k L > \frac{\theta_k+1}{\theta_k+2}s_k$. Therefore, we have

$$\frac{16\eta_k L}{s_k-16\eta_k L}<\frac{1}{1+\theta_k}\quad\text{and}\quad\frac{s_k}{s_k-16\eta_k L}=\frac{\theta_k+2}{\theta_k+1}<2,\tag{4.14}$$

which together with (4.13) and $m_k=(1+\theta_k)m_{k-1}$ imply (4.5). □

Motivated from Theorem 4.1, if we further set

$$\bar{\varepsilon}_t^k\le\varepsilon_t^k(1-\sqrt{\eta_k})^{t-1},\quad\text{for}\quad k=1,\ldots,N\quad\text{and}\quad t=1,\ldots,m_k,\tag{4.15}$$

then we have

$$\left(\sum_{t=1}^{m_k}\sqrt{\bar{\varepsilon}_t^k}\right)^2\le\left(\sum_{t=1}^{m_k}\varepsilon_t^k\right)\left(\sum_{t=1}^{m_k}\left(1-\sqrt{\eta_k}\right)^{t-1}\right)\le\frac{1}{\sqrt{\eta_k}}\sum_{t=1}^{m_k}\varepsilon_t^k,$$

and

$$\begin{aligned}\sum_{t=1}^{m_k}\left(\sum_{i=t}^{m_k}\sqrt{\bar{\varepsilon}_i^k}\right)^2 &\le \sum_{t=1}^{m_k}\left[\left(\sum_{i=t}^{m_k}\varepsilon_i^k\right)\left(\sum_{i=t}^{m_k}(1-\sqrt{\eta_k})^{i-1}\right)\right]\\ &\le \frac{1}{\sqrt{\eta_k}}\sum_{t=1}^{m_k}\left[\left(\sum_{i=t}^{m_k}\varepsilon_i^k\right)(1-\sqrt{\eta_k})^{t-1}\right]\\ &= \frac{1}{\sqrt{\eta_k}}\sum_{t=1}^{m_k}\left[\varepsilon_t^k\sum_{i=1}^{t}(1-\sqrt{\eta_k})^{i-1}\right] \le \frac{1}{\eta_k}\sum_{t=1}^{m_k}\varepsilon_t^k.\end{aligned}$$

Then, the B_k defined in (4.7) can be upper bounded by

$$B_k \le \left(\frac{5}{\eta_k}+\frac{5}{2}\right)\sum_{t=1}^{m_k}\varepsilon_t^k. \tag{4.16}$$

We now give the convergence property of Algorithm 2.1 for solving (1.1)–(1.2) without strongly convex assumption.

Theorem 4.2 *Under assumption* **AS.1**, *if the parameters* η_k *and* $\bar{\varepsilon}_t^k$ *in Algorithm 2.1 are set as*

$$\eta_k = \eta < \min\left\{\frac{s_k}{16L(\theta_k+2)}, \frac{1}{L}\right\} \quad \text{and} \quad \bar{\varepsilon}_t^k = \varepsilon_t^k(1-\sqrt{\eta})^{t-1}, \quad \text{for} \quad k = 1,\ldots,N \quad \text{and} \quad t = 1,\ldots,m_k, \tag{4.17}$$

the inner iteration number m_k *satisfies the condition (4.2) and the sample size* $\{s_k\}_{k=1}^N$ *is a nondecreasing sequence, then we have*

$$\mathbb{E}[P(\tilde{w}_k)] - P(w_*) \le \left(a_0 + \frac{10\eta^{-1}+5}{m_0}\sum_{i=1}^{k}(1+\sqrt{\eta})^{-i}\sum_{t=1}^{m_i}\varepsilon_t^i\right)\prod_{j=1}^{k}\frac{1+\sqrt{\eta}}{1+\theta_j}, \tag{4.18}$$

where

$$a_0 = \left(1+\frac{16\eta L}{m_0(s_1-16\eta L)}\right)[P(\tilde{w}_0)-P(w_*)] + \frac{s_1}{2\eta m_0(s_1-16\eta L)}\|w_0^1-w_*\|^2. \tag{4.19}$$

Proof First, for all $k \ge 1$, denote

$$\begin{aligned}a_k = \mathbb{E}[P(\tilde{w}_k)] - P(w_*) + \frac{s_k}{2\eta m_k(s_k-16\eta L)}\mathbb{E}[\|w_0^{k+1}-w_*\|^2]\\ +\frac{16\eta L}{m_k(s_k-16\eta L)}(\mathbb{E}[P(w_0^{k+1})]-P(w_*)).\end{aligned}$$

Then, it follows from Theorem 4.1, (4.17) and $\{s_k\}_{k=1}^N$ being a nondecreasing sequence that for all $k \geq 1$

$$a_k \leq \left(\frac{1+\sqrt{\eta}}{1+\theta_k}\right) a_{k-1} + \frac{2B_k}{m_k}$$

So, by induction we get

$$a_k \leq a_0 \prod_{j=1}^{k} \frac{1+\sqrt{\eta}}{1+\theta_j} + 2\sum_{i=1}^{k}\left(\prod_{j=i+1}^{k} \frac{1+\sqrt{\eta}}{1+\theta_j}\right)\frac{B_i}{m_i}.$$

Then, (4.18) follows from $m_i = m_0 \prod_{j=1}^{i}(1+\theta_j)$ and the upper bound (4.16) on B_i. □

Remark 4.1 Note that to ensure $\mathbb{E}[P(\tilde{w}_k)] - P(w_*)$ converges to zero as k goes to infinity, by Theorem 4.2, a sufficient condition is to require

$$\prod_{j=1}^{k} \frac{1+\sqrt{\eta}}{1+\theta_j} \to 0 \quad \text{and} \quad \left(\sum_{i=1}^{k}(1+\sqrt{\eta})^{-i}\sum_{t=1}^{m_i}\varepsilon_t^i\right)\prod_{j=1}^{k} \frac{1+\sqrt{\eta}}{1+\theta_j} \to 0,$$

as k goes to infinity.

Motivated from Remark 4.1, we have the following Corollary.

Corollary 4.3 *Under assumption* **AS.1**, *suppose the inner iteration number m_k satisfies the condition (4.2) with $\theta_k = \theta > 0$ and the parameters in Algorithm 2.1 are set as*

$$s_k = s, \quad \eta_k = \eta < \min\left\{\frac{s}{16L(\theta+2)}, \frac{1}{L}, \theta^2\right\} \quad \text{and} \quad \bar{\varepsilon}_t^k = \varepsilon_t^k(1-\sqrt{\eta})^{t-1}. \tag{4.20}$$

Then, we have

$$\mathbb{E}[P(\tilde{w}_k)] - P(w_*) \leq \left(a_0 + \frac{10\eta^{-1}+5}{m_0}\sum_{i=1}^{k}(1+\sqrt{\eta})^{-i}\sum_{t=1}^{m_i}\varepsilon_t^i\right)\gamma^k, \tag{4.21}$$

where a_0 is defined in (4.19) and

$$\gamma = \frac{1+\sqrt{\eta}}{1+\theta} < 1. \tag{4.22}$$

Consequently,

(1) *if the subproblem tolerance ε_t^k satisfies (3.15), then*

$$\mathbb{E}[P(\tilde{w}_k)] - P(w_*) \leq \left(a_0 + \frac{\alpha^2(10\eta^{-1}+5)}{m_0(1-\alpha)^2}\right)\gamma^k; \tag{4.23}$$

(2) *if the subproblem tolerance* ε_t^k *satisfies (3.17), then*

$$\mathbb{E}[P(\tilde{w}_k)] - P(w_*) \leq \left(a_0 + \frac{\alpha(10\eta^{-1}+5)}{m_0(1-\alpha)\sqrt{\eta}}\right)\gamma^k; \tag{4.24}$$

(3) *if the subproblem tolerance* ε_t^k *satisfies (3.20), then*

$$\mathbb{E}[P(\tilde{w}_k)] - P(w_*) \leq \left(a_0 + \frac{(1+\phi^{-1})(10\eta^{-1}+5)}{m_0\sqrt{\eta}}\right)\gamma^k; \tag{4.25}$$

(4) *if the subproblem tolerance* ε_t^k *satisfies (3.22), then*

$$\mathbb{E}[P(\tilde{w}_k)] - P(w_*) \leq \left(a_0 + \frac{10\eta^{-1}+5}{m_0(\kappa-1)\sqrt{\eta}}\right)\gamma^k. \tag{4.26}$$

Proof First, by Theorem 4.2, (4.21) follows from (4.18) directly.

(1) If the subproblem tolerance ε_t^k satisfies (3.15), then $\sum_{t=1}^{m_i}\varepsilon_t^i \leq \frac{\alpha}{1-\alpha}\alpha^i$, which implies that

$$\sum_{i=1}^{k}(1+\sqrt{\eta})^{-i}\sum_{t=1}^{m_i}\varepsilon_t^i = \frac{\alpha}{1-\alpha}\sum_{i=1}^{k}\left(\frac{\alpha}{1+\sqrt{\eta}}\right)^i \leq \frac{\alpha}{1-\alpha}\cdot\frac{\alpha}{1+\sqrt{\eta}-\alpha} \leq \frac{\alpha^2}{(1-\alpha)^2}.$$

Then, (4.23) follows from (4.21).

(2) If the subproblem tolerance ε_t^k satisfies (3.17), then $\sum_{t=1}^{m_i}\varepsilon_t^i \leq \frac{\alpha}{1-\alpha}i^{-\beta}$, which implies that

$$\sum_{i=1}^{k}(1+\sqrt{\eta})^{-i}\sum_{t=1}^{m_i}\varepsilon_t^i \leq \frac{\alpha}{1-\alpha}\sum_{i=1}^{k}(1+\sqrt{\eta})^{-i} \leq \frac{\alpha}{\sqrt{\eta}(1-\alpha)}.$$

Then, (4.24) follows from (4.21).

(3) If the subproblem tolerance ε_t^k satisfies (3.20), then (4.25) follows from (4.21) and the fact that

$$\sum_{t=1}^{m_i}\varepsilon_t^i \leq (1+\phi^{-1})i^{-\beta} \leq 1+\phi^{-1} \quad \text{and} \quad \sum_{i=1}^{k}(1+\sqrt{\eta})^{-i} \leq \frac{1}{\sqrt{\eta}},$$

where the first inequality is by (3.25).

(4) If the subproblem tolerance ε_t^k satisfies (3.22), then (4.26) follows from (4.21) and the fact that

$$\sum_{t=1}^{m_i}\varepsilon_t^i \leq \frac{1}{\kappa-1}i^{1-\kappa} \leq \frac{1}{\kappa-1},$$

where the first inequality is by (3.26). □

Remark 4.2 Following from Corollary 4.3, we have $\mathbb{E}[P(\tilde{w}_k)] - P(w_*)$ converges to zero as k goes to infinity. We now analyze the component gradient complexity of Algorithm 2.1 implied by Corollary 4.3. In the algorithm, suppose we set

$$m_0 = \mathcal{O}(L), \quad \eta = \mathcal{O}(L^{-1}) \text{ and } s = \mathcal{O}(1). \tag{4.27}$$

Then, it follows from (4.27) and (4.14) that

$$\frac{16\eta_k L}{s_k - 16\eta_k L} = \mathcal{O}(1), \quad \frac{s_k}{s_k - 16\eta_k L} = \mathcal{O}(1).$$

Hence, we have $a_0 = \mathcal{O}(1)$. In the following, let us denote

$$\tau = \frac{\log(1+\theta)}{\log(1+\theta) - \log(1+\sqrt{\eta})}. \tag{4.28}$$

Then, Corollary 4.3 would give the following complexity bounds.

- Suppose ε_t^k is set as (3.15). It follows from $a_0 = \mathcal{O}(1)$ and (4.27) that the coefficient of γ^k in (4.23) is $\mathcal{O}(1)$. Hence, to achieve $\mathbb{E}[P(\tilde{w}_N)] - P(w_*) < \varepsilon$ for a given $\varepsilon > 0$, the outer iteration number N should be in the order of $\mathcal{O}(\log_\gamma \varepsilon) = \mathcal{O}(\log(1/\varepsilon))$. So, by direct calculation, the total number of component gradient evaluations required by Algorithm 2.1 is

$$nN + 2\sum_{i=1}^{k} sm_i = nN + 2sm_0 \frac{(1+\theta)^{N+1} - (1+\theta)}{\theta} = \mathcal{O}\left(n \log\left(\frac{1}{\varepsilon}\right) + \frac{L}{\varepsilon^\tau}\right). \tag{4.29}$$

- Suppose ε_t^k is set as (3.17). It follows from $a_0 = \mathcal{O}(1)$ and (4.27) that the coefficient of γ^k in (4.24) is $\mathcal{O}(L^{1/2})$. Hence, to achieve $\mathbb{E}[P(\tilde{w}_N)] - P(w_*) < \varepsilon$ for a given $\varepsilon > 0$, the outer iteration number N should be in the order of $\mathcal{O}(\log_\gamma(\varepsilon/L^{1/2})) = \mathcal{O}(\log(L^{1/2}/\varepsilon))$. Therefore, the total number of component gradient evaluations required by Algorithm 2.1 is

$$\mathcal{O}\left(n \log\left(\frac{L^{1/2}}{\varepsilon}\right) + L\left(\frac{L^{1/2}}{\varepsilon}\right)^\tau\right).$$

- Suppose ε_t^k is set as (3.20). It follows from $a_0 = \mathcal{O}(1)$ and (4.27) that the coefficient of γ^k in (4.25) is $\mathcal{O}(L^{1/2}/\rho)$ where $\rho = \min\{1, \phi\}$. Hence, to achieve $\mathbb{E}[P(\tilde{w}_N)] - P(w_*) < \varepsilon$ for a given $\varepsilon > 0$, the outer iteration number N should be in the order of $\mathcal{O}(\log_\gamma(\rho\varepsilon/L^{1/2})) = \mathcal{O}(\log(L^{1/2}/(\rho\varepsilon)))$. Therefore, the total number of component gradient evaluations required by Algorithm 2.1 is

$$\mathcal{O}\left(n \log\left(\frac{L^{1/2}}{\rho\varepsilon}\right) + L\left(\frac{L^{1/2}}{\rho\varepsilon}\right)^\tau\right).$$

- Suppose ε_t^k is set as (3.22). It follows from $a_0 = \mathcal{O}(1)$ and (4.27) that the coefficient of γ^k in (4.26) is $\mathcal{O}(\sigma)$, where $\sigma = \max\{1, L^{1/2}/(\kappa - 1)\}$. Hence, to achieve $\mathbb{E}[P(\tilde{w}_N)] - P(w_*) < \varepsilon$ for a given $\varepsilon > 0$, the outer iteration number N should be in the order of $\mathcal{O}(\log_\gamma(\varepsilon/\sigma) = \mathcal{O}(\log(\sigma/\varepsilon))$. Therefore, the total number of component gradient evaluations required by Algorithm 2.1 is

$$\mathcal{O}\left(n \log\left(\frac{\sigma}{\varepsilon}\right) + L\left(\frac{\sigma}{\varepsilon}\right)^\tau\right).$$

Similar to the strongly convex case, we can see as the subproblem tolerances become less strict, the component gradient complexity bound would increase. Again, the proper subproblem accuracy setting should depend on the balance of the subproblem difficulty and the cost of component gradient evaluations.

Remark 4.3 When all the subproblems (2.10) are solved exactly, the complexity bound $\mathcal{O}(n \log(1/\varepsilon) + L/\varepsilon)$ is obtained in [40] for solving (1.1)–(1.2) without strongly convexity assumption. By comparison, the complexity bound (4.29) is worse since it has a power $\tau > 1$. This gap arises when the term $(1/\eta)\langle u_t^k, w_* - w_t^k\rangle$ in (4.8) is estimated. When the subproblems are solved exactly, this term vanishes. However, when the subproblems are solved inexactly, this term needs to be estimated as shown in (4.10), which leads to the the reduction ratio $(1 + \sqrt{\eta})/(1 + \theta)$ in (4.5), and the power $\tau > 1$ finally appear in the complexity result. But, on the other hand, when the Lipschitz constant L is very large, which is usually the case for ill-posed problems, we could have $\theta >> L^{-1}$ and τ would be close to 1. Then, the complexity bound (4.29) would be very close to the complexity bound obtained in [40], which assumes exactly solving of all the subproblems.

Observe that by the same analysis for establishing (4.16), we can replace (4.15) by

$$\bar{\varepsilon}_t^k = \varepsilon_t^k (1 - \sigma\sqrt{\eta_k})^{t-1}, \tag{4.30}$$

where $\sigma \in (0, 1)$ is a constant. In this case, the convergence complexity order of Algorithm 2.1 given in Theorem 4.2 and Corollary 4.3 would still be the same.

In addition, based on Theorem 4.2, we could also have a variable choice of θ_k as

$$\theta_k = \lambda\left(1 + \sqrt{\eta}\right)\frac{k+1}{k} - 1, \tag{4.31}$$

which gives

$$\frac{1 + \sqrt{\eta}}{1 + \theta_k} = \lambda\frac{k}{k+1},$$

where $\lambda > 0$ is a constant, Then, by (4.18) we have

$$\mathbb{E}[P(\tilde{w}_k)] - P(w_*) \le \left(a_0 + \frac{10\eta^{-1} + 5}{m_0}\sum_{i=1}^{k}(1 + \sqrt{\eta})^{-i}\sum_{t=1}^{m_i}\varepsilon_t^i\right)\frac{\lambda}{k}.$$

Same analysis as in the proof of Corollary 4.3 would show the factor in front of λ/k in the above estimate are bounded for all the four choices of tolerance settings (3.15), (3.17), (3.20) and (3.22). Hence, the convergence of $\mathbb{E}[P(\tilde{w}_k)] - P(w_*)$ to zero, as k goes to infinity, is still guaranteed.

5 Experiments

In this section, we consider solving the following CUR-like factorization optimization proposed in [23]:

$$\min_{X\in\mathbb{R}^{n_r\times n_c}} P(X) := \frac{1}{2}\|W - WXW\|_F^2 + \lambda_{row}\sum_{i=1}^{n_r}\|X^i\|_p + \lambda_c\sum_{j=1}^{n_c}\|X_j\|_p, \quad (5.1)$$

where $W \in \mathbb{R}^{n_c\times n_r}$ is a given matrix, $X^i = X(i,;)$ and $X_j = X(:,j)$ are the i-th row and j-th column of the matrix X, respectively. This optimization model (5.1) would generate a solution X with sparse rows and columns with different choices of $p > 0$. It can be seen that the problem (5.1) would have the formulation as (1.1) by setting

$$f(X) := \frac{1}{2}\|W - WXW\|_F^2 = \frac{1}{2}\sum_{i=1}^{n_c}\sum_{j=1}^{n_r} f_{i,j}(X) \quad \text{and}$$

$$h(X) := \lambda_{row}\sum_{i=1}^{n_r}\|X^i\|_p + \lambda_c\sum_{j=1}^{n_c}\|X_j\|_p,$$

where $f_{i,j}(X) = (W(i,j) - W^iXW_j)^2$, $W^i = W(i,:)$ and $W_j = W(:,j)$ are the i-th row and j-th column of the matrix W, respectively. In [23], the authors choose $p = \infty$ for which the proximal subproblem (2.10) has a closed-form solution. The more natural choice of $p = 2$ is discussed in [30]. However, in this case there is no closed-form formula to find the exact solution of (2.10). Hence, the proximal subproblem (2.10) is solved in [30] by the block coordinate descent (BCD) algorithm proposed in [16], which is also equivalent to Dykstra's algorithm for two monotone operators [1].

In our experiments, we also consider the choice of $p = 2$ and the proximal subproblem (2.10) is solved by the same BCD algorithm used in [30]. In the implementation of the BCD method, the proximal subproblem is minimized alternatively with respect to the rows and columns. Since the duality gap can be computed while applying the BCD algorithm, we can compute an approximate solution of the subproblem (2.10) until its function value duality gap is below any given tolerance $\varepsilon > 0$.

The same data sets in [23,30][1] are used in our experiments. We set $\lambda_{row} = 0.01$ and $\lambda_{col} = 0.01$ in (5.1) as those in [30], which would yield approximately 25–40% non-zero entries in the solution. The data sets is summarized in the following table:

[1] The datasets are available at http://www.gems-system.org.

Data sets	9_Tumors	Brain_Tumor1	Leukemia1	SRBCT
n_r	5726	5920	5328	2308
n_c	60	90	72	83

In the experiments, the size of the random sample index set in Step 5 of Algorithm 2.1 is set as $s_k = n_c$ for all k. The starting point w_0^{k+1} for the inner loop is selected by (4.1), i.e.,

$$w_0^{k+1} = w_{m_k}^k.$$

For the inner iteration number, we set $m_k = (1 + \theta_k)m_{k-1}$ with $m_0 = 1$, and as suggested by (4.22) and (4.31), we consider the following two choices of θ_k:

(1) $\theta_k = \lambda_1(1 + \sqrt{\eta_k}) - 1$ with $\lambda_1 = 1.001$;
(2) $\theta_k = \lambda_2 \frac{k+1}{k}(1 + \sqrt{\eta_k}) - 1$ with $\lambda_2 = 0.8$.

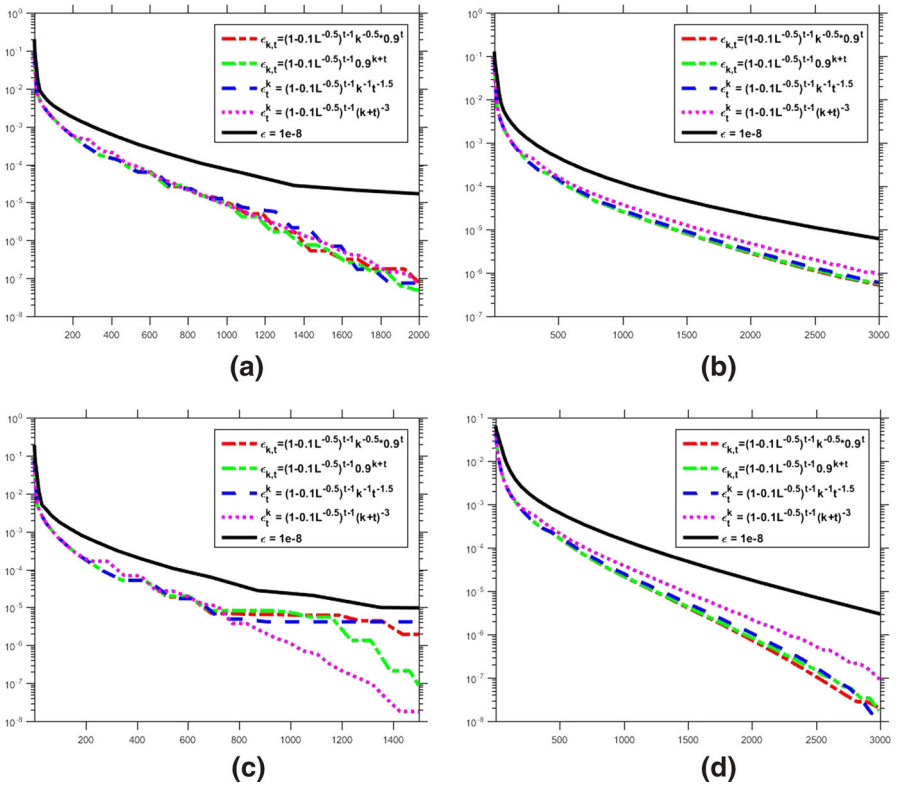

Fig. 2 Objective function value gap $P(\tilde{w}_k) - P(w_*)$ (*vertical axis*) against the number of BCD iterations (*horizontal axis*). In this figure, $\theta_k = \lambda_1(1 + \sqrt{\eta_k}) - 1$ and $\lambda_1 = 1.001$. **a** 9_Tumors, **b** Brain_Tumor1, **c** Leukemia1, **d** SRBCT

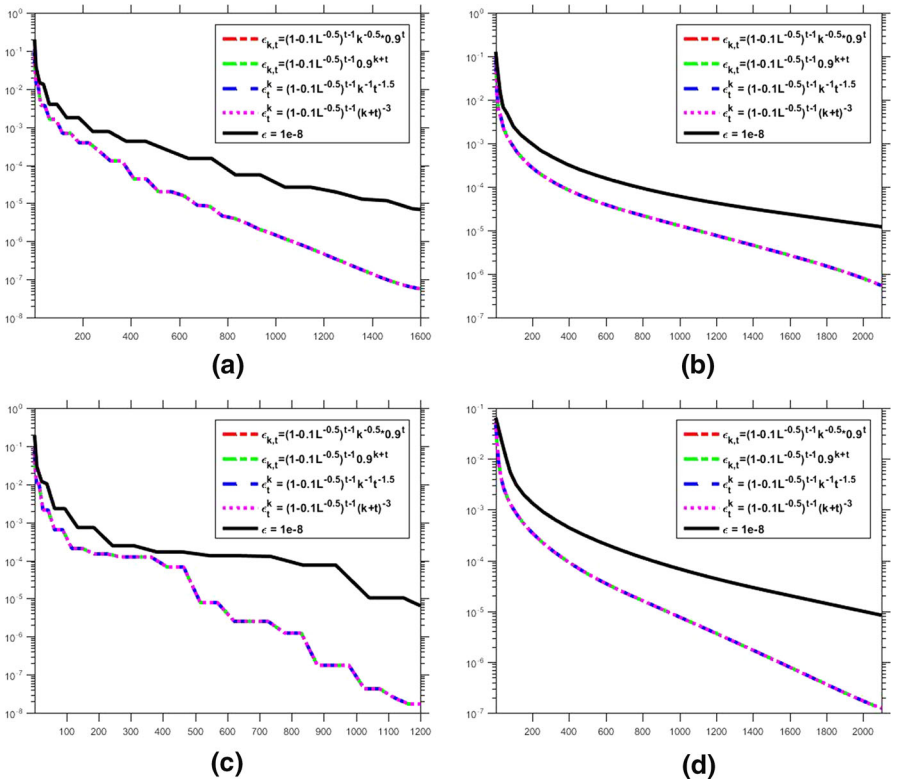

Fig. 3 Objective function value gap $P(\tilde{w}_k) - P(w_*)$ (*vertical axis*) against the number of BCD iterations (*horizontal axis*). In this figure, $\theta_k = \lambda_2 \frac{k+1}{k}(1 + \sqrt{\eta_k}) - 1$ and $\lambda_2 = 0.8$. **a** 9_Tumors, **b** Brain_Tumor1, **c** Leukemia1, **d** SRBCT

In both above two cases the proximal parameter $\eta_k = \max\{0.7/k, 1/\bar{L}\}$ where $\bar{L} = 30$ is an estimate of the Lipschitz constant of ∇f in (1.1).

We have tested four different types of tolerance ε_t^k according to (3.15), (3.17), (3.20) and (3.22) as the following:

(1) $\varepsilon_t^k = (1 - \frac{\sigma}{\sqrt{L_k}})^{t-1} \cdot \alpha^{k+t}$ with $\alpha = 0.9$.
(2) $\varepsilon_t^k = (1 - \frac{\sigma}{\sqrt{L_k}})^{t-1} \cdot \frac{1}{k^\beta} \cdot \alpha^t$ with $\alpha = 0.9$ and $\beta = 0.5$.
(3) $\varepsilon_t^k = (1 - \frac{\sigma}{\sqrt{L_k}})^{t-1} \cdot \frac{1}{k^\beta} \cdot \frac{1}{t^{1+\phi}}$ with $\beta = 1$ and $\phi = 0.5$.
(4) $\varepsilon_t^k = (1 - \frac{\sigma}{\sqrt{L_k}})^{t-1} \cdot \frac{1}{(k+t)^\kappa}$ with $\kappa = 3$.

In all the above cases, we set $\sigma = 0.1$ and $L_k = 1/\eta_k$. All the subproblems (2.10) are solved by the BCD algorithm until the subproblem function value duality gap is below ε_t^k. Then, by [38, Thm. 2.8.7] and [3, Section 4.3], the inexact solution w_t^k obtained by the BCD algorithm is an $(\bar{\varepsilon}_t^k, \hat{\varepsilon}_t^k)$-solution of the problem (2.1) with $\bar{\varepsilon}_t^k + \hat{\varepsilon}_t^k \le \varepsilon_t^k$. Hence, tolerance condition (4.30) on $\bar{\varepsilon}_t^k$ is satisfied.

Firstly, to show the benefits of solving the subproblems inexactly, we also tested the Algorithm 2.1 but with almost exactly solving of the subproblem (2.10) until

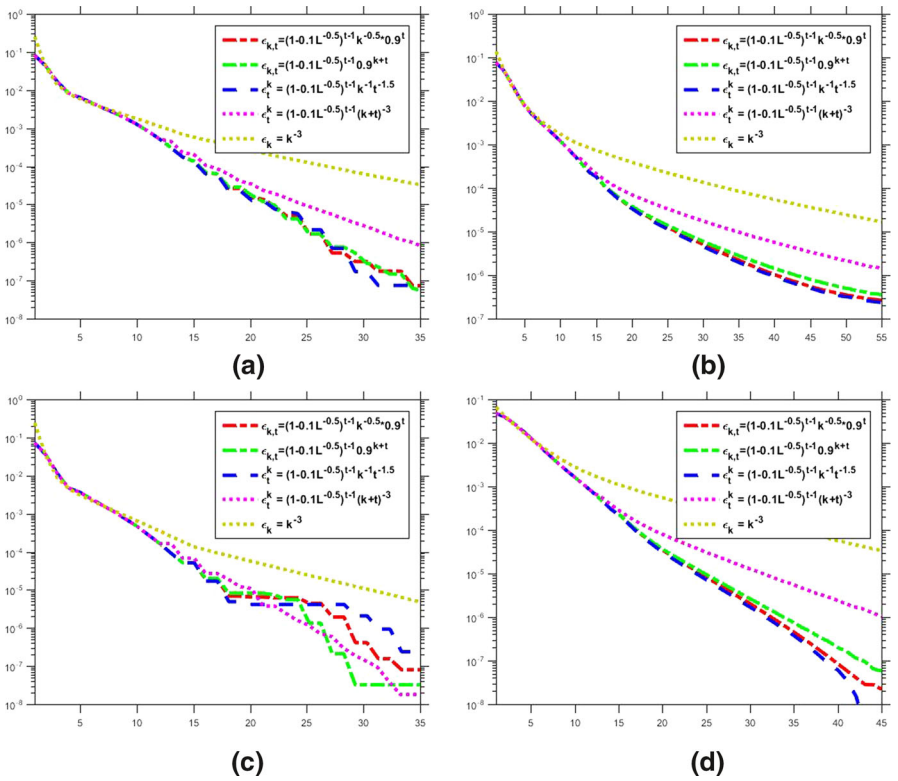

Fig. 4 Objective function value gap $P(w_k) - P(w_*)$ (*vertical axis*) against the number of effective passes (*horizontal axis*). In this figure, $\theta_k = \lambda_1(1 + \sqrt{\eta_k}) - 1$ and $\lambda_1 = 1.001$. $\epsilon_k = k^{-3}$ represents the method proposed in [30]. **a** 9_Tumors, **b** Brain_Tumor1, **c** Leukemia1, **d** SRBCT

its function value duality gap is below 10^{-8}. Note that the only differences of the comparing algorithms are their tolerances for solving the subproblem. Hence, as in [30], in this case we also use the function value gap $P(\tilde{w}_k) - P(w^*)$ against the number of proximal iterations (BCD iterations) as the measure of algorithm performance. In our numerical experiments, the optimal function value $P(w^*)$ is approximated by the minimum objective function value obtained by running all the comparing algorithms until

$$\max_{i=0,1,2} \frac{P(\tilde{w}_{k-i}) - P(\tilde{w}_{k-i-1})}{P(\tilde{w}_{k-i})} \leq 10^{-8}.$$

The numerical results corresponding to different choices of θ_k in cases (1) and (2) are shown in Figs. 2 and 3, respectively. The results clearly show that the inexact methods outperform the the exact method (e.g., the method with fixed tolerance $\epsilon_t^k = 10^{-8}$). Hence, the result confirms our analysis that it is not necessary to solve the subproblems exactly at each iteration. On the other hand, we observe that the new algorithms with different tolerances often behave similar for this set of testing problems, especially at

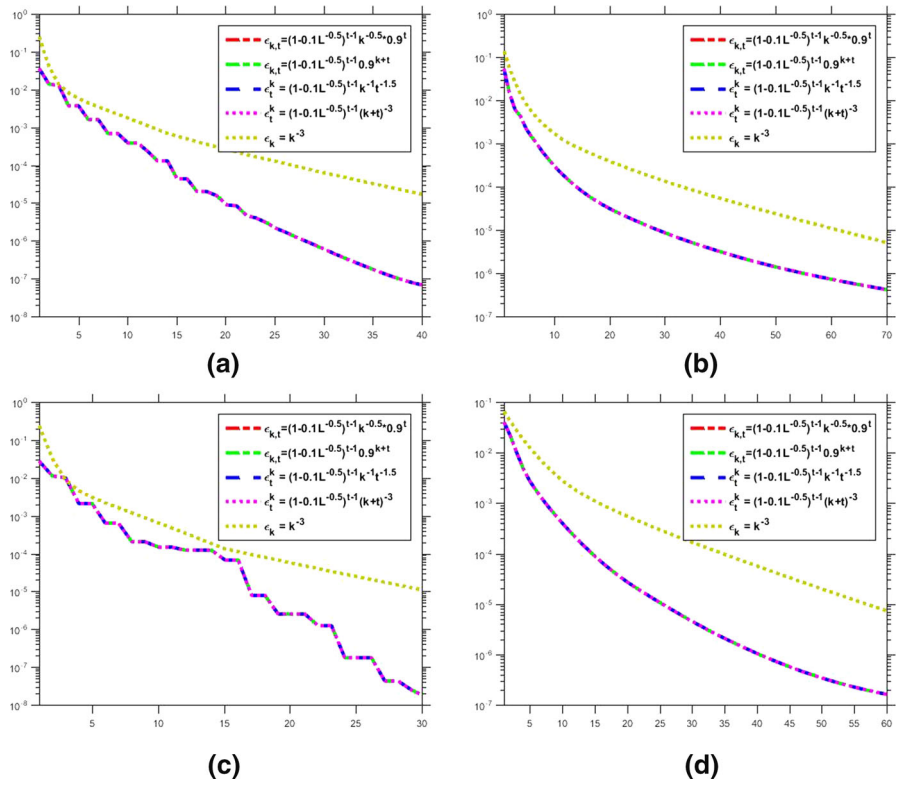

Fig. 5 Objective function value gap $P(w_k) - P(w_*)$ (*vertical axis*) against the number of effective passes (*horizontal axis*). In this figure, $\theta_k = \lambda_2 \frac{k+1}{k}(1+\sqrt{\eta_k}) - 1$ and $\lambda_2 = 0.8$. $\epsilon_k = k^{-3}$ represents the method proposed in [30]. **a** 9_Tumors, **b** Brain_Tumor1, **c** Leukemia1, **d** SRBCT

early iterations, since different tolerances for the subproblem (2.10) are often satisfied in similar number of BCD iterations. Secondly, to show the advantages of IPSG method with variance reduction to the inexact Prox-FG method for solving the problem (1.1)–(1.2), we also compare our new algorithm to the inexact Prox-FG method proposed in [30]. For the inexact Prox-FG method in [30], we use the default tolerance $\epsilon_k = k^{-3}$ suggested in [30], which by our observation also gives the best results of this method for solving our test problems. The comparison numerical results are shown in Figs. 4 and 5 corresponding to different choices of θ_k in cases (1) and (2). Although both the IPSG and the Prox-FG method in [30] are inexact methods, these numerical results indicate by using stochastic gradient with variance reduction techniques in IPSG, the number of effective passes (each effective pass evaluates n component gradients) can be significantly for solving the test problems to the same level of accuracy.

6 Conclusion

In this paper, we propose an inexact proximal stochastic gradient (IPSG) method combining with variance reduction techniques. The method is designed for solving convex composite optimization, whose objective function is an addition of an average of a large number of smooth convex functions and a convex, but possibly nonsmooth, function. Different tolerances for inexactly solving the subproblem are discussed from both global convergence and component gradient complexity point of view. When the smooth part of the objective function is strongly convex, it is shown that under certain subproblem solving tolerance the component gradient complexity of the new IPSG algorithm will be the same as that of the Prox-SVRG method [36], which however requires to solve the subproblems exactly. And when the objective is only convex, instead of strongly convex, with proper subproblem solving tolerance the IPSG method can have almost the same component gradient complexity bound developed in [40] for the SVRG based Prox-SG method. Finally, our preliminary numerical results also confirm the new IPSG method outperforms the corresponding exact Prox-SVRG method and the inexact Prox-FG method for solving the test problems.

References

1. Bauschke, H.H., Combettes, P.L.: A dykstra-like algorithm for two monotone operators. Pac. J. Optim. **4**(3), 383–391 (2008)
2. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sci. **2**(1), 183–202 (2009)
3. Bertsekas, D.P.: Nonlinear Programming, 2nd edn. Athena Scientific, Belmont, Massachusetts (1999)
4. Bertsekas, D.P.: Incremental proximal methods for large scale convex optimization. Math. Program. Ser. B **129**, 163–195 (2011)
5. Bertsekas, D.P.: Incremental gradient, subgradient, and proximal methods for convex optimization: a survey. arXiv: 1507.01030v1, 3 July (2015)
6. Cai, J.-F., Candès, E.J., She, Z.: A singular value thresholding algorithm for matrix completion. SIAM J. Optim. **20**(4), 1956–1982 (2010)
7. Cai, J.-F., Candès, E.J., She, Z.: Fast newton-type methods for total variation regularization. In: ICML (2011)
8. Defazio, A., Bach, F., Lacoste-Julien, S.: Saga: a fast incremental gradient method with support for non-strongly convex composite objectives. In: NIPS, pp. 1646–1654 (2014)
9. Duchi, J., Hazan, E., Singe, Y.: Adaptive subgradient methods for online learning and stochastic optimization. J. Mach. Learn. Res. **12**, 2121–2159 (2015)
10. Elad, M., Aharon, M.: Image denoising via sparse and redundant representations over learned dictionaries. IEEE Trans. Image Process. **54**(12), 3736–3745 (2006)
11. Fadili, J., Peyrè, G.: Total variation projection with first order schemes. IEEE Trans. Image Process. **20**(3), 657–669 (2011)
12. Ghadimi, S., Lan, G.: Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, I: a generic algorithmic framework. SIAM J. Optim. **22**(4), 1469–1492 (2012)
13. Ghadimi, S., Lan, G.: Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, II: shrinking procedures and optimal algorithms. SIAM J. Optim. **23**(4), 2061–2089 (2013)
14. Hovhannisyan, V., Parpas, P., Zafeiriou, S.: Magma: multi-level accelerated gradient mirror descent algorithm for large-scale convex composite minimization. arXiv: 1509.05715v3, July (2016)
15. Jang, K., Sun, D., Toh, K.C.: An inexact accelerated proximal gradient method for large scale linearly constrained convex sdp. SIAM J. Optim. **22**(3), 1042–1064 (2012)
16. Jenatton, R., Mairal, J., Bach, F.R., Obozinski, G.R.: Proximal methods for sparse hierarchical dictionary learning. In: ICML, pp. 487–494 (2010)

17. Johnson R., Zhang, T.: Accelerating stochastic gradient descent using predictive variance reduction. In: NIPS, pp. 315–323 (2013)
18. Kavukcuoglu, K., Ranzato, M., Fergus, R., LeCun, Y.: Learning invariant features through topographic filter maps. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2009)
19. Konečný, J., Liu, J., Richtárik, P., Takáč, M.: Mini-batch semi-stochastic gradient descent in the proximal setting. IEEE J. Sel. Top. Signal Process. **10**(2), 242–255 (2016)
20. Lan, G.: An optimal method for stochastic composite optimization. Math. Program. **133**(1), 365–397 (2012)
21. Ma, S., Goldfarb, D., Chen, L.: Fixed point and bregman iterative methods for matrix rank minimization. Math. Program. **128**(1), 321–353 (2011)
22. Mahoney, M.W., Drineas, P.: CUR matrix decompositions for improved data analysis. Proc. Natl. Acad. Sci. **106**(3), 697 (2009)
23. Mairal, J., Jenatton, R., Obozinski, G., Bach, F.: Convex and network flow optimization for structured sparsity. J. Mach. Learn. Res. **12**, 2681–2720 (2011)
24. Nesterov, Y.E.: Gradient methods for minimizing composite objective function. Math. Program. Ser. B **140**, 341–362 (2013)
25. Nitanda, A.: Stochastic proximal gradient descent with acceleration techniques. In: NIPS, pp. 1574–1582 (2014)
26. Obozinski, G., Taskar, B., Jordan, M.I.: Joint covariate selection and joint subspace selection for multiple classification problems. Stat. Comput. **20**(2), 231–252 (2009)
27. Salzo, S., Villa, S.: Inexact and accelerated proximal point algorithms. J. Convex Anal. **4**(19), 1167–1192 (2012)
28. Scheinberg, K., Tang, X.: Practical inexact proximal quasi-Newton method with global complexity analysis. Math. Program. (2016). doi:10.1007/s10107-016-0997-3
29. Schmidt, M., Le Roux, N., Bach, F.: Minimizing finite sums with the stochastic average gradient. Math. Program. (2016). doi:10.1007/s10107-016-1030-6
30. Schmidt, M., Le Roux, N., Bach, F.: Supplementary material for the paper convergence rates of inexact proximal-gradient methods for convex optimization. In: NIPS (2011)
31. Shalev-Shwartz, S., Zhang, T.: Stochastic dual coordinate ascent methods for regularized loss minimization. J. Mach. Learn. Res. **14**, 567–599 (2013)
32. Tibshirani, R., Hastie, T., Friedman, J.: The elements of statistical learning: data mining, 2nd edn. In: Inference. Springer, New York (2009)
33. Villa, S., Salzo, S., Baldassarre, L., Verri, A.: Accelerated and inexact forward–backward algorithms. SIAM J. Optim. **23**(3), 1607–1633 (2011)
34. Wang, M., Bertsekas, D.P.: Incremental constraint projection methods for variational inequalities. Math. Program. **150**(2), 321–363 (2015)
35. Wang, M., Bertsekas, D.P.: Stochastic first-order methods with random constraint projection. SIAM J. Optim. **26**(1), 681–717 (2016)
36. Xiao, L., Zhang, T.: A proximal stochastic gradient method with progressive variance reduction. SIAM J. Optim. **24**, 2057–2075 (2014)
37. Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. J. R. Stat. Soc. B **68**, 49–67 (2006)
38. Zălinescu, C.: Convex Analysis in General Vector Spaces. World Scientific Publishing Co. Inc., Singapore (2002)
39. Zhang, Y., Lin, X.: Stochastic primal-dual coordinate method for regularized empirical risk minimization. In: ICML (2015)
40. Zhu, Z.A., Yuan, Y.: Univr: A universal variance reduction framework for proximal stochastic gradient method. arXiv: 1506.01972v1, 5 June (2015)