An extended gradient method for smooth strongly convex functions

Xuexue Zhang (✉ zhangxuexue2012@163.com)  
Xidian University  

Sanyang Liu  
Xidian University  

Nannan Zhao  
Chang'an University  

Research Article

Keywords: Gradient method, Strongly convex optimization, Acceleration, Convergence analysis

Posted Date: July 18th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3168818/v1

License: © This work is licensed under a Creative Commons Attribution 4.0 International License.  
Read Full License

Additional Declarations: No competing interests reported.
An extended gradient method for smooth strongly convex functions

Xuexue Zhang\textsuperscript{1,*}, Sanyang Liu\textsuperscript{1†} and Nannan Zhao\textsuperscript{2,3†}

\textsuperscript{1}School of Mathematics and Statistics, Xidian University, Xi’an, 710126, Shaanxi, China.
\textsuperscript{2}School of Science, Chang’an University, Xi’an, 710064, Shaanxi, China.
\textsuperscript{3}School of Civil Engineering, Chang’an University, Xi’an, 710061, Shaanxi, China.

\textsuperscript{*}Corresponding author(s). E-mail(s): zhangxuexue2012@163.com; liusanyang@126.com; nzhao@chd.edu.cn;
\textsuperscript{†}These authors contributed equally to this work.

Abstract

This paper considers an extended gradient method with fixed step size applied to the centralized and decentralized strongly convex and smooth functions. Firstly, the range of step size is derived to guarantee that the function value generated by the extended gradient method converges to the optimal solution at a sublinear rate. Then, we further show that the iterate sequences converge to the optimal solution at a linear rate for the centralized and decentralized problems. Finally, we demonstrate the efficiency of the proposed extended gradient method in numerical experiments.

Keywords: Gradient method, Strongly convex optimization, Acceleration, Convergence analysis

1 Introduction

In this work, we consider the unconstrained optimization problem:

$$\min_{x \in \mathbb{R}^p} f(x).$$  \hspace{1cm} (1)
The cost function $f(x) : \mathbb{R}^p \mapsto \mathbb{R}$ of this problem is a differentiable strongly convex function, and the gradient $\nabla f(x)$ is $L$-Lipschitz continuous. Starting from an initial point $x^0 \in \mathbb{R}^p$ with $\nabla f(x^0) = 0$, the extended gradient method updates the iterates $x^k$ for $k \geq 0$:

$$x^{k+1} = x^k - \alpha (\nabla f(x^k) + \nabla f(x^{k-1})),$$

where the step size $\alpha (\alpha > 0)$ is sufficiently small, and it will be analyzed below to ensure the convergence of the proposed algorithm.

Firstly, we discussed briefly the accelerated scheme of gradient methods, that is, the iterate $x^{k+1}$ updates using the information of the preceding two iterates. Such as the gradient method with extrapolation that has the forms:

$$x^{k+1} = x^k - \alpha_k \nabla f(x^k) + \beta_k (x^k - x^{k-1}),$$

where $\alpha_k$ is the step size and $\beta_k \in [0, 1)$ is a scalar. When $\alpha_k$ and $\beta_k$ are set as the constant scalar $\alpha$ and $\beta$, respectively, the method is equivalent to the heavy ball method [1]. In [2], the linear convergence rate has been established for the heavy ball method applied to smooth, twice continuously differentiable and strongly convex functions. Further, the authors in [3] constructed an example to illustrate that even if the objective functions are strongly convex, the heavy ball method still can not converge when the functions are not twice continuously differentiable. Thus, the property that objective function is twice continuously differentiable is a necessary condition in convergence guarantees of heavy ball method. Based on this fact, for the case of strictly convex quadratic functions, the recent work [4] obtained the iteration complexity of heavy ball method under the assumption that the appropriate upper and lower bounds for eigenvalues of Hessian matrix are known. In particular, several accelerated gradient descent methods with the similar structure of heavy ball method also have been proposed in [5], and they converge linearly to the solution of the smooth and strongly convex functions with the optimal iteration complexity. Compared to the heavy ball method, the above accelerated methods fully utilize the difference of the preceding two iterates. Moreover, the main advantage for solving convex functions in practice is that the accelerated gradient descent methods can converge to the optimum more quickly [6].

On the other hand, the accelerated scheme also can update the variables employing the preceding several gradients. As mentioned in [7], for the saddle point problem, the optimistic gradient descent-ascent (OGDA) method updates the variable via the difference of the preceding two gradients. Besides, for the least mean squares (LMS) estimation of graph signals, the extended LMS algorithm in [8] and proportionate-type graph LMS algorithm in [9] update the variable via the weighted sum of the preceding several gradients. Inspired by the above works, we here propose an extended gradient method, in which the variables are updated along the direction of the sum of the gradients of the preceding two iterates. The main purpose of this work is to analyze the convergence of the extended gradient method for finding the optimal solution of problem (1). We will show that when the step size is less than a given upper bound, the linear convergence rate of the extended gradient method can be guaranteed.

Next, we consider a class of smooth and strongly convex optimization functions on the real vector space with Euclidean norm $\| \cdot \|$ and inner product $\langle \cdot, \cdot \rangle$. 

---

2
Definition 1. A function $f : \mathbb{R}^p \mapsto \mathbb{R}$ is $L$-smooth if the gradient $\nabla f(x)$ is $L$-Lipschitz continuous, that is,

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \forall x, y.$$ 

If $f$ is convex and $L$-smooth it also holds that $f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + \frac{L}{2}\|x - y\|^2.$

Definition 2. A function $f : \mathbb{R}^p \mapsto \mathbb{R}$ is $\mu$-strongly convex if there exists positive $\mu$ such that

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle + \frac{\mu}{2}\|x - y\|^2.$$ 

2 Analysis of the extended gradient method

In this section, in order to establish the convergence results of the extended gradient method, we first obtain an important inequality in Theorem 1, where the upper bound on $\|x^{k+1} - x^*\|$ will be given in terms of the linear combinations of $\|x^k - x^*\|$, $\|x^{k-1} - x^*\|$ and $\|x^{k-2} - x^*\|$. This inequality is vital and will be used in the following analysis.

Theorem 1. Assume that the function $f$ is $\mu$-strongly convex, and the gradient $\nabla f(x)$ is $L$-Lipschitz continuous. Let the step size $\alpha$ satisfy the condition $0 \leq \alpha \leq \frac{\sqrt{1+\frac{\mu}{L}}}{L}$, and $\{x^k\}_{k \in \mathbb{N}}$ be generated by the extended gradient method (2). Then, the followings hold.

(i) The iterates $\{x^k\}$ satisfy the following linear inequality

$$\|x^{k+1} - x^*\|^2 \leq (1 + 2\alpha^2 L^2 - \alpha \mu)\|x^k - x^*\|^2 + (2\alpha^2 L^2 + 2\alpha^3 L^3 - \alpha \mu)\|x^{k-1} - x^*\|^2 + 2\alpha^3 L^3\|x^{k-2} - x^*\|^2,$$

where $x^*$ is the optimal solution.

(ii) $\min_{2 \leq k \leq K} f(x^k)$ converges to the optimal value $f(x^*)$ at the sub-linear rate of $O(\frac{1}{K})$.

Proof. (i) According to the update equation in (2), for any $x \in \mathbb{R}^p$, we have

$$\|x^{k+1} - x\|^2 = \|x^k - \alpha(\nabla f(x^k) + \nabla f(x^{k-1})) - x\|^2$$

$$= \|x^k - x\|^2 + \alpha^2\|\nabla f(x^k) + \nabla f(x^{k-1})\|^2 - 2\alpha\langle \nabla f(x^k) + \nabla f(x^{k-1}), x^k - x \rangle$$

$$= \|x^k - x\|^2 + \alpha^2\Gamma_1 - 2\alpha(\Gamma_2 + \Gamma_3) \tag{3}$$

where $\Gamma_1 := \|\nabla f(x^k) + \nabla f(x^{k-1})\|^2$, $\Gamma_2 := \nabla f(x^k)^T(x^k - x)$, and $\Gamma_3 := \nabla f(x^{k-1})^T(x^k - x)$.

Now, on substituting $x = x^*$, and noting that $\nabla f(x^*) = 0$, We firstly bound the term $\Gamma_1$.

$$\Gamma_1 \leq 2\|\nabla f(x^k)\|^2 + 2\|\nabla f(x^{k-1})\|^2$$

3
\begin{align*}
= 2\|\nabla f(x^k) - \nabla f(x^*)\|^2 + 2\|\nabla f(x^{k-1}) - \nabla f(x^*)\|^2 \\
\leq 2L^2\|x^k - x^*\|^2 + 2L^2\|x^{k-1} - x^*\|^2 \tag{4}
\end{align*}

Secondly, we provide a lower bound of \( \Gamma_2 \) with the strongly convex property of \( f(x) \),
\[
\Gamma_2 = \nabla f(x^k)^T (x^k - x^*) \geq f(x^k) - f(x^*) + \frac{\mu}{2}\|x^k - x^*\|^2. \tag{5}
\]

Finally, we bound the term \( \Gamma_3 \) employing \( \Gamma_1 \) and \( \Gamma_2 \). By the smoothness of \( f \), we can obtain that
\[
f(x^k) \leq f(x^{k-1}) + \nabla f(x^{k-1})^T (x^k - x^{k-1}) + \frac{L}{2}\|x^k - x^{k-1}\|^2.
\]

Add and subtract the inner product \( \nabla f(x^{k-1})^T x^* \) to the right hand side of \( \text{6} \) and rearrange the terms to have
\[
f(x^k) \leq f(x^{k-1}) + \nabla f(x^{k-1})^T (x^* - x^{k-1}) + \frac{L}{2}\|x^k - x^{k-1}\|^2 + \nabla f(x^{k-1})^T (x^k - x^*), \tag{7}
\]

Then by the bound of \( \Gamma_2 \) \( \text{(k is replaced by k - 1)} \), \( \text{7} \) can be rewritten as
\[
f(x^k) \leq f(x^*) - \frac{\mu}{2}\|x^{k-1} - x^*\|^2 + \frac{L}{2}\|x^k - x^{k-1}\|^2 + \nabla f(x^{k-1})^T (x^k - x^*), \tag{8}
\]

wherein, in terms of the update equation \( \text{(2)} \) and the bound of \( \Gamma_1 \) \( \text{(4)} \), the term \( \|x^k - x^{k-1}\|^2 \) in \( \text{8} \) can be further expressed as
\[
\|x^k - x^{k-1}\|^2 = \alpha^2\|\nabla f(x^{k-1}) + \nabla f(x^{k-2})\|^2 \\
\leq 2\alpha^2 L^2\|x^{k-1} - x^*\|^2 + 2\alpha^2 L^2\|x^{k-2} - x^*\|^2. \tag{9}
\]

Therefore, the bound of \( \Gamma_3 \) can be yielded as
\[
\Gamma_3 = \nabla f(x^{k-1})^T (x^k - x^*) \geq f(x^k) - f(x^*) + \left( \frac{\mu}{2} - \alpha^2 L^2 \right)\|x^{k-1} - x^*\|^2 \\
- \alpha^2 L^2\|x^{k-2} - x^*\|^2 \tag{10}
\]

Substituting \( \text{(4)}, \text{(5)} \) and \( \text{(10)} \) (the bounds of \( \Gamma_1, \Gamma_2, \text{ and } \Gamma_3 \)) into \( \text{(3)} \), we have
\[
\|x^{k+1} - x^*\|^2 \leq (1 + 2\alpha^2 L^2 - \alpha\mu)\|x^k - x^*\|^2 + (2\alpha^2 L^2 + 2\alpha^3 L^3 - \alpha\mu)\|x^{k-1} - x^*\|^2 \\
+ 2\alpha^3 L^3\|x^{k-2} - x^*\|^2 - 4\alpha(f(x^k) - f(x^*)). \tag{11}
\]

Considering that \( f(x^k) \geq f(x^*) \) in \( \text{11} \), we complete the proof of (i).

(ii) On rearranging the terms in \( \text{11} \), we have the following inequality
\[
4\alpha(f(x^k) - f(x^*)) \leq (\|x^k - x^*\|^2 - \|x^{k+1} - x^*\|^2)
\]
\[-2\alpha^3 L^3 \left( \|x^{k-1} - x^*\|^2 - \|x^{k-2} - x^*\|^2 \right) + (2\alpha^2 L^2 - \alpha \mu) \|x^k - x^*\|^2 + (2\alpha^2 L^2 + 4\alpha^3 L^3 - \alpha \mu) \|x^{k-1} - x^*\|^2. \]  

(12)

For \(0 \leq \alpha \leq \frac{\sqrt{1+\frac{\mu}{2L}} - 1}{2L}\), we obtain \(2\alpha^2 L^2 + 4\alpha^3 L^3 - \alpha \mu \leq -(2\alpha^2 L^2 - \alpha \mu)\) and thus

\[4\alpha(f(x^k) - f(x^*)) \leq \left( \|x^k - x^*\|^2 - \|x^{k+1} - x^*\|^2 \right) - 2\alpha^3 L^3 \left( \|x^{k-1} - x^*\|^2 - \|x^{k-2} - x^*\|^2 \right) - (2\alpha^2 L^2 - \alpha \mu) \left( \|x^k - x^*\|^2 - \|x^{k-1} - x^*\|^2 \right). \]

(13)

Summing over \(k = 2, \ldots, K\), we have

\[
\sum_{k=2}^{K} 4\alpha \left( f(x^k) - f(x^*) \right) \leq \|x^2 - x^*\|^2 - \|x^{K+1} - x^*\|^2 - 2\alpha^3 L^3 \left( \|x^{K-1} - x^*\|^2 - \|x^0 - x^*\|^2 \right) - (2\alpha^2 L^2 - \alpha \mu) \left( \|x^1 - x^*\|^2 - \|x^K - x^*\|^2 \right).
\]

(14)

Noticed that \(0 \leq \alpha \leq \frac{\sqrt{1+\frac{\mu}{2L}} - 1}{2L}\), then \((2\alpha^2 L^2 - \alpha \mu)\|x^K - x^*\|^2 < 0\). Therefore, the following inequality holds

\[
\min_{2 \leq k \leq K} f(x^k) - f(x^*) \leq \frac{\|x^2 - x^*\|^2 - (2\alpha^2 L^2 - \alpha \mu) \|x^1 - x^*\|^2 + 2\alpha^3 L^3 \|x^0 - x^*\|^2}{4\alpha(K-2)},
\]

(15)

which means that \(\min_{2 \leq k \leq K} f(x^k)\) converges to the optimal value \(f(x^*)\) at the sub-linear rate of \(O\left(\frac{1}{K}\right)\).

The result in (ii) of Theorem 1 shows that the function value generated by extended gradient method converges to the optimal solution of problem (1) at a sublinear rate. Further, in order to derive the linear convergence of the proposed method, we give two important lemmas as follows.

**Lemma 1.** [10] Let \(A \in \mathbb{R}^{n \times n}\) be a nonnegative matrix, and \(\omega \in \mathbb{R}^n\) be a positive vector. If \(A\omega < q\omega\), then \(\rho(A) < q\).

**Lemma 2.** [11] (Gelfand’s formula) It holds that \(\rho(A) = \lim_{k \to \infty} \|A^k\|^\frac{1}{k}\), i.e., the spectral radius of \(A\) gives the asymptotic growth rate of \(\|A^k\|\) : For every \(\epsilon > 0\) there is \(c = c(\epsilon)\) such that \(\|A^k\| \leq c(\rho(A) + \epsilon)^k\), for all \(k\).

Together with (i) of Theorem 1, Lemma 1 and Lemma 2, we give the following Theorem 2. For notational convenience, we define \(z^k := (\|x^k - x^*\|^2, \|x^{k-1} - x^*\|^2, \|x^{k-2} - x^*\|^2)^T\).

**Theorem 2.** The same assumptions hold as in Theorem 1. Let the sequence \(\{x^k\}\) generated by the extended gradient method and \(x^*\) be the optimal solution. Then for
every $\varepsilon > 0$, there is $c = c(\varepsilon)$ such that $\|z^{k+1}\| \leq c(\rho(A) + \varepsilon)^k\|z^0\|$, for all $k$, where $\rho(A) < 1$ is a constant.

Proof. By the result in (i) of Theorem 1, we first readily find that $z^{k+1} \leq A z^k$, here

$$A = \begin{pmatrix} 1 + 2\alpha^2 L^2 - \alpha \mu & 2\alpha^2 L^2 + 2\alpha^3 L^3 - \alpha \mu & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$  

Under the assumption that the spectral radius of matrix $A$ is less than 1 (i.e., $\rho(A) < 1$), the statement of convergence can be obtained via Lemma 2. Thus, it remains to analyze the spectral radius of the matrix $A$ and show that the spectral radius of $A$ is less than 1. Find a positive vector $c = [c_1, c_2, c_3]^T$ satisfying $Ac < c$, then the following inequalities hold

$$(2\alpha^2 L^2 - \alpha \mu)c_1 + (2\alpha^2 L^2 + 2\alpha^3 L^3 - \alpha \mu)c_2 + 2\alpha^3 L^3 c_3 < 0;$$
\[ c_1 < c_2; \]
\[ c_2 < c_3. \]  

To ensure that the first inequality in (16) holds, the necessary condition is that $2\alpha^2 L^2 - \alpha \mu < 0$, which means $\alpha < \frac{\mu}{2\alpha^2 L^2}$. Then, according to the sign of $2\alpha^2 L^2 + 2\alpha^3 L^3 - \alpha \mu$, we further deduce the bounds of $\alpha$ through considering the following two cases.

(i) If $2\alpha^2 L^2 + 2\alpha^3 L^3 - \alpha \mu > 0$, we can derive the first inequality of (16) as

$$(2\alpha^2 L^2 - \alpha \mu)c_1 + (2\alpha^2 L^2 + 2\alpha^3 L^3 - \alpha \mu)c_2 + 2\alpha^3 L^3 c_3 > (4\alpha^2 L^2 + 2\alpha^3 L^3 - 2\alpha \mu)c_1 + 2\alpha^3 L^3 c_3,$$  

due to $0 < c_1 < c_2$. Then, since the left hand side of the (17) is less than 0, we have

$$2\alpha^3 L^3 c_3 < -(4\alpha^2 L^2 + 2\alpha^3 L^3 - 2\alpha \mu)c_1 < -(4\alpha^2 L^2 + 2\alpha^3 L^3 - 2\alpha \mu)c_3,$$  

due to $0 < c_1 < c_3$. Further, it follows from (18) that $-(4\alpha^2 L^2 + 2\alpha^3 L^3 - 2\alpha \mu)c_3 - 2\alpha^3 L^3 c_3 > 0$, namely, $2\alpha^2 L^2 + 2\alpha^3 L^3 - \alpha \mu < 0$, which leads to a contradiction to the assumption.

(ii) If $2\alpha^2 L^2 + 2\alpha^3 L^3 - \alpha \mu \leq 0$, the first inequality of (16) can be written as

$$(2\alpha^2 L^2 - \alpha \mu)c_1 + (2\alpha^2 L^2 + 2\alpha^3 L^3 - \alpha \mu)c_2 + 2\alpha^3 L^3 c_3 > (2\alpha^2 L^2 - \alpha \mu)c_1 + (2\alpha^2 L^2 + 4\alpha^3 L^3 - \alpha \mu)c_3,$$  

due to $0 < c_2 < c_3$. Then, based on the fact that the left hand side of (19) is less than 0, we have

$$(2\alpha^2 L^2 + 4\alpha^3 L^3 - \alpha \mu)c_3 < -(2\alpha^2 L^2 - \alpha \mu)c_1 < -(2\alpha^2 L^2 - \alpha \mu)c_3,$$  

(20)
due to $0 < c_1 < c_3$. On rearranging the terms of (20), we obtain that $2(2\alpha^2L^2 + 2\alpha^3L^3 - \alpha\mu)c_3 < 0$, which is consistent with the assumption. Therefore, the upper bound of step size $\alpha$ can be reduced as

$$\alpha < \sqrt{\frac{1 + \frac{2\mu}{L} - 1}{2L}} = \min \left\{ \frac{\mu}{2L^2}, \sqrt{\frac{1 + \frac{2\mu}{L} - 1}{2L}} \right\}. \quad (21)$$

Finally, we have demonstrated that there exists a positive vector $c = [c_1, c_2, c_3]^T$ such that $Ac < c$, then it is clear $\rho(A) < 1$ according to Lemma 1 for $a < \frac{\sqrt{1 + \frac{2\mu}{L} - 1}}{2L}$.

**Remark 1.** Theorem 2 shows that the linear convergence of the extended gradient method also can be guaranteed for the smooth and strongly convex functions as the classical gradient methods.

### 3 Analysis of the decentralized extended gradient method

#### 3.1 Decentralized extended gradient method

This section develops a decentralized version of the extended gradient method for the following problem:

$$\min_{x \in \mathbb{R}^p} f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x), \quad (22)$$

where the local cost function $f_i$ is $L$-smooth and $\mu$-strongly convex, which is defined over a network consist of $n$ agents. A set of local variables $\{x_i \in \mathbb{R}^p\}_{i=1}^{n}$ is introduced and $x_i$ is assigned to agent $i$. Denote the aggregate variable $X$ and aggregate gradient $\nabla F(X)$ as $X = (x_1, x_2, \ldots, x_n)^T$ and $\nabla F(X) = (\nabla f_1(x_1), \nabla f_2(x_2), \ldots, \nabla f_n(x_n))^T$. The decentralized extended gradient method updates the iterates $X^{k+1}$:

$$X^{k+1} = WX^k - \alpha(\nabla F(X^k) + \nabla F(X^{k-1})), \quad (23)$$

where weight matrix $W$ is symmetric and double stochastic, i.e., $W^T = W$, $W1_n = 1_n$ and $1_n^TW = 1_n^T$, herein $1_n = (1, \ldots, 1) \in \mathbb{R}^p$. Weight matrix $W$ is generated based on the network. When agent $i$ and $j$ are neighbors or $i = j$, $W_{ij} \neq 0$. In decentralized setting, local function $f_i$ is only be accessed by agent $i$, $i = 1, \ldots, n$. And all agents cooperate to optimize the global function $f$ via local computation and communication. The local computation is operated on the individual agent, i.e., the local variable of agent $i$ is updated by the gradient information of the local function $f_i$. The communication is operated over the network, i.e., the local variable is updated by combining the weighted average of its neighbor’s local variable, which enforces the consensus of all local variables.
3.2 convergence result

We need to define the average vector: \( \bar{X}^k = \frac{1}{n} \sum_{i=1}^{n} x_i^k \in \mathbb{R}^{1 \times p} \). In the following lemma, we establish the upper bounds on \( \| \bar{X}^k - (x^*)^\top \| \) and \( \| X^k - 1_n \bar{X}^k \| \) in terms of the linear combinations of their preceding two values.

**Lemma 3.** Let \( \{X^k\}_{k \in \mathbb{N}} \) be the iterates generated by the updates in (23). If \( \xi \in (\alpha - \sqrt{1 - (n-1)\alpha}, \alpha + \sqrt{1 - (n-1)\alpha}) \), then the following relation hold:

\[
z^{k+1} \leq Az^k,
\]

where \( z^k = (\|X^k - 1_n \bar{X}^k\|, \|\bar{X}^k - (x^*)^\top\|, \|X^k - 1_n \bar{X}^{k-1}\|, \|\bar{X}^{k-1} - (x^*)^\top\|)^\top \in \mathbb{R}^4 \) and

\[
A = \left(\begin{array}{cccc}
\delta + \frac{\xi L \sqrt{n(1-\alpha)}}{\alpha - (\xi^2 - 2\alpha \xi + \alpha^2)} & \frac{\xi L \sqrt{n(n-1)}}{\alpha - (\xi^2 - 2\alpha \xi + \alpha^2)} & \frac{\xi L \sqrt{n(n-1)}}{\alpha - (\xi^2 - 2\alpha \xi + \alpha^2)} & \frac{\xi L \sqrt{n(n-1)}}{\alpha - (\xi^2 - 2\alpha \xi + \alpha^2)} \\
\frac{\alpha L}{\sqrt{n}} & 0 & 0 & 0 \\
0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0
\end{array}\right) \in \mathbb{R}^{4 \times 4},
\]

herein \( \lambda = \max\{|1 - \alpha L|, |1 - \alpha \mu|\} \) and \( \delta = \|W - \frac{1_n}{n}\| \).

If \( \frac{1}{\mu + L} < \alpha < \min\{\frac{1}{\xi L \sqrt{n}}, \frac{1}{\alpha - (\xi^2 - 2\alpha \xi + \alpha^2)}\} \) and \( \xi \in (\alpha - \sqrt{1 - (n-1)\alpha}, \alpha + \sqrt{1 - (n-1)\alpha}) \), where \( 0 < \gamma_1 < \gamma_2 < \gamma_3 < \gamma_4 \), \( a = (1 - \delta)\gamma_1 \), \( b = \alpha L \sqrt{n - 1}(\gamma_1 + \sqrt{\gamma_2 + \gamma_3 + \gamma_4}) + 2(1 - \delta)\alpha \gamma_3 \), and \( c = (1 - \delta)(\alpha a^2 - \alpha) \), then \( \rho(A) < 1 \) and thus \( \|X^k - 1_n \bar{X}^k\| \) and \( \|\bar{X}^k - (x^*)^\top\| \) converge to zero at the linear rate of \( O(\rho(A)^k) \).

**Proof.** First, we establish a linear system by using the upper bounds on \( \|X^{k+1} - 1_n \bar{X}^{k+1}\| \) and \( \|\bar{X}^{k+1} - (x^*)^\top\| \). Next, we can find the range of \( \alpha \) such that \( \rho(A) < 1 \). From the update in (23), we derive

\[
\bar{X}^{k+1} = \frac{1}{n} \sum_{i=1}^{n} (WX^k - \alpha (\nabla F(X^k) + \nabla F(X^{k-1}))) = \bar{X}^k - \frac{\alpha}{n} \sum_{i=1}^{n} (\nabla F(X^k) + \nabla F(X^{k-1})).
\]

**Step 1. Bound \( \|X^{k+1} - 1_n \bar{X}^{k+1}\|.**

From the above equation, it follows that

\[
\|X^{k+1} - 1_n \bar{X}^{k+1}\| = \|(I - \frac{1}{n} 1_n 1_n^\top)(WX^k - \alpha (\nabla F(X^k) + \nabla F(X^{k-1})))\|
\leq \|(I - \frac{1}{n} 1_n 1_n^\top)WX^k\| + \|I - \frac{1}{n} 1_n 1_n^\top\| \|\alpha \nabla F(X^k) + \alpha \nabla F(X^{k-1})\|
\]

(25)

Notice that \( (I - \frac{1}{n} 1_n 1_n^\top)W = (W - \frac{1}{n} 1_n 1_n^\top)(I - \frac{1}{n} 1_n 1_n^\top) \) and \( \|I - \frac{1}{n} 1_n 1_n^\top\| = \sqrt{n - 1} \). Therefore, it follows
\[ \|X^{k+1} - 1_n \bar{X}^{k+1}\| \leq \delta \|X^k - 1_n \bar{X}^k\| + \sqrt{n-1} (\|\alpha \nabla F(X^k)\| + \|\alpha \nabla F(X^{k-1})\|), \quad (26) \]

where \( \delta = \|W - \frac{1}{n}1_n 1_n^\top\| < 1 \). For the term \( \|\alpha \nabla F(X^k)\| \), we can further bound as

\[
\|\alpha \nabla F(X^k)\| = \|\alpha \nabla F(X^k) - \frac{\xi}{n} 1_n 1_n^\top \nabla F(X^k) + \frac{\xi}{n} 1_n 1_n^\top \nabla F(X^k) - \xi 1_n \nabla f(\bar{X})^\top \\
+ \xi 1_n \nabla f(\bar{X})^\top - \xi 1_n \nabla f(x^*)^\top \| \\
\leq \|(\alpha I - \frac{\xi}{n} 1_n 1_n^\top) \nabla F(X^k)\| + \xi \|1_n\| \frac{1}{n} \sum_{i=1}^n (\nabla f_i(x^k) - \nabla f_i(\bar{X}^k)) \|^\top \\
+ \xi \|1_n\| \frac{1}{n} \sum_{i=1}^n (\nabla f_i(\bar{X}^k) - \nabla f_i(x^*)) \|^\top \\
\leq (\xi^2 - 2\alpha \xi + n\alpha^2) \|\nabla F(X^k)\| + \xi \sqrt{n} \frac{1}{n} \sum_{i=1}^n \|\nabla f_i(x^k) - \nabla f_i(\bar{X}^k)\| \\
+ \xi \sqrt{n} \frac{1}{n} \sum_{i=1}^n \|\nabla f_i(\bar{X}^k) - \nabla f_i(x^*)\| \\
\leq (\xi^2 - 2\alpha \xi + n\alpha^2) \|\nabla F(X^k)\| + \xi L \sqrt{n} \frac{1}{n} \sum_{i=1}^n \|x_i^k - \bar{X}^k\| \\
+ \xi L \sqrt{n} \|\bar{X}^k - (x^*)^\top\| \\
\leq (\xi^2 - 2\alpha \xi + n\alpha^2) \|\nabla F(X^k)\| + \xi L \sqrt{n} \|\bar{X}^k - (x^*)^\top\| \\
\leq (\xi^2 - 2\alpha \xi + n\alpha^2) \|\nabla F(X^k)\| + \xi L \|X^k - 1_n \bar{X}^k\| + \xi L \sqrt{n} \|\bar{X}^k - (x^*)^\top\|, \quad (27) 
\]

where choosing \( \xi \in (\alpha - \sqrt{1 - (n-1)\alpha}, \alpha + \sqrt{1 - (n-1)\alpha}) \) such that \( \alpha - (\xi^2 - 2\alpha \xi + n\alpha^2) > 0 \). So

\[
\|\nabla F(X^k)\| \leq \frac{1}{\alpha - (\xi^2 - 2\alpha \xi + n\alpha^2)} (\xi L \|X^k - 1_n \bar{X}^k\| + \xi L \sqrt{n} \|\bar{X}^k - (x^*)^\top\|) . \quad (28) 
\]
Similarly, we have
\[
\|\nabla F(X^{k-1})\| \leq \frac{1}{\alpha - (\xi^2 - 2\xi + n\alpha^2)}(\xi L\|X^{k-1} - 1_n\bar{X}^{k-1}\| + \xi L\sqrt{n}\|\bar{X}^{k-1} - (x^*)^\top\|).
\]  
(29)

Substituting (28) and (29) into (26) yields
\[
\|X^{k+1} - 1_n\bar{X}^{k+1}\| \leq (\delta + \frac{\xi L\alpha\sqrt{n - 1}}{\alpha - (\xi^2 - 2\xi + n\alpha^2)})\|X^k - 1_n\bar{X}^k\| \\
+ \frac{\xi L\alpha\sqrt{n - 1}}{\alpha - (\xi^2 - 2\xi + n\alpha^2)}\|X^k - (x^*)^\top\| \\
+ \frac{\xi L\alpha\sqrt{n - 1}}{\alpha - (\xi^2 - 2\xi + n\alpha^2)}\|X^{k-1} - 1_n\bar{X}^{k-1}\| \\
+ \frac{\xi L\alpha\sqrt{n - 1}}{\alpha - (\xi^2 - 2\xi + n\alpha^2)}\|\bar{X}^{k-1} - (x^*)^\top\|.
\]  
(30)

Step 2. Bound \(\|\bar{X} - (x^*)^\top\|\).

\[
\|\bar{X}^{k+1} - (x^*)^\top\| = \|\bar{X}^k - \alpha\frac{1}{n}(\nabla F(X^k) + \nabla F(X^{k-1})) - (x^*)^\top\| \\
= \|\bar{X}^k - \alpha\frac{1}{n}\nabla F(1_n\bar{X}^k) - (x^*)^\top + \alpha\frac{1}{n}(\nabla F(1_n\bar{X}^k) - \nabla F(X^k)) \\
- \alpha\frac{1}{n}\nabla F(X^{k-1})\|
\]
\[
\leq \|\bar{X}^k - \alpha\nabla f(\bar{X}^k) - (x^*)^\top\| + \alpha\|\frac{1}{n}(\nabla F(1_n\bar{X}^k) - \nabla F(X^k))\| \\
+ \alpha\|\frac{1}{n}\nabla F(X^{k-1})\|.
\]  
(31)

Consider the first term in (31), we have
\[
\|\bar{X}^k - \alpha\nabla f(\bar{X}^k) - (x^*)^\top\| \leq \lambda\|\bar{X}^k - (x^*)^\top\|,
\]  
(32)

where \(\lambda = \max\{|1 - \alpha L|, |1 - \alpha\mu|\}\). For the second term and third term in (31), we bound them as
\[
\|\frac{1}{n}\nabla F(1_n\bar{X}^k) - \nabla F(X^k)\| \leq \frac{L}{\sqrt{n}}\|X^k - 1_n\bar{X}^k\|.
\]  
(33)

and
\[
\|\frac{1}{n}\nabla F(X^{k-1})\| = \|\frac{1}{n}\nabla F(X^{k-1}) - \frac{1}{n}\nabla F(1_n\bar{X}^{k-1}) + \frac{1}{n}\nabla F(1_n\bar{X}^{k-1})
\]
\[ -\frac{1}{n} \nabla F(1_n(x^*)^\top) \leq \frac{L}{\sqrt{n}} \|X^{k-1} - 1_n \bar{X}^{k-1}\| + L\|\bar{X}^{k-1} - (x^*)^\top\| \] (34)

due to the smoothness of global function \( f \). Substituting (32)-(34) into (31), we obtain that

\[
\begin{align*}
\|\bar{X}^{k+1} - (x^*)^\top\| & \leq \frac{L\alpha}{\sqrt{n}} \|X^k - 1_n \bar{X}^k\| + \lambda\|X^k - (x^*)^\top\| \\
& \quad + \frac{L\alpha}{\sqrt{n}} \|X^{k-1} - 1_n \bar{X}^{k-1}\| + L\alpha\|\bar{X}^{k-1} - (x^*)^\top\|. 
\end{align*}
\] (35)

The linear system in (24) is obtained by using the results of (30) and (35).

Next, according to lemma1, we give the range of \( \alpha \) and \( \xi \) and a positive vector \( \gamma = (\gamma_1, \gamma_2, \gamma_3, \gamma_4)^\top \) such that \( A\gamma < \gamma \), which is equivalent to the following inequalities

\[
\xi L\alpha \sqrt{n - 1}\gamma_1 + \xi L\alpha \sqrt{n(n - 1)}\gamma_2 + \xi L\alpha \sqrt{n - 1}\gamma_3 + \xi L\alpha \sqrt{n(n - 1)}\gamma_4 < (1 - \delta)(\alpha - (\xi^2 - 2\alpha\xi + n\alpha^2))\gamma_1 \] (36)

\[
\frac{L\alpha}{\sqrt{n}} \gamma_1 + \frac{L\alpha}{\sqrt{n}} \gamma_3 + \alpha L\gamma_4 < (1 - \lambda)\gamma_2 
\] (37)

\[
\gamma_1 < \gamma_3 
\] (38)

\[
\gamma_2 < \gamma_4 
\] (39)

If \( \frac{1}{\sqrt{n}} < \alpha < \frac{2}{L} \), according to the definition that \( \lambda = \max\{|1 - \alpha\mu|, |1 - \alpha L|\} \), we have that \( l - \lambda = 2 - \alpha L > 0 \). Since the step size \( \alpha \) satisfies (37), we require

\[
\frac{1}{\mu + L} < \alpha < \min\left\{\frac{2}{L}, \frac{\sqrt{n}\gamma_2}{\sqrt{n}\gamma_2 + \sqrt{n}\gamma_3 + \sqrt{n}\gamma_4}\right\} 
\] (40)

The range of \( \xi \) follows from (36), that is,

\[
\xi \in \left(\alpha - \sqrt{1 - (n - 1)\alpha}, \min\{\alpha + \sqrt{1 - (n - 1)\alpha}, \frac{\sqrt{b^2 - 4ac}}{2a}\}\right), 
\] (41)

where \( a = (1 - \delta)\gamma_1 \), \( b = \alpha L\sqrt{n - 1}(\gamma_1 + \sqrt{n}\gamma_2 + \gamma_3 + \sqrt{n}\gamma_4) + 2(1 - \delta)\alpha \gamma_1 \) and \( c = (1 - \delta)(n\alpha^2 - \alpha) \). Then, consider that \( \gamma_1, \gamma_2, \gamma_3 \) and \( \gamma_4 \) are positive, we can select proper \( \gamma_1, \gamma_2, \gamma_3 (> \gamma_1) \) and \( \gamma_4 (> \gamma_2) \) to make the set in (41) be nonempty. □
4 Numerical experiments

4.1 Centralized problem

In this subsection, we consider the performance of the extended gradient method when it is applied to the LASSO problem:

\[
\min_{x \in \mathbb{R}^p} f(x) = \frac{1}{2} \|Mx - b\|^2 + \nu \|x\|_1, \tag{42}
\]

where \( M \in \mathbb{R}^{m \times p} \) is a random matrix generated from a uniform distribution \( U(0, 1) \), and \( b = Mx \) where \( x \) also follows the uniform distribution \( U(0, 1) \) with sparsity \( r \). In particular, we set the parameters \( m = 512, p = 1024, r = 0.1 \), and the initial condition \( x^0 \in U(0, 1) \). The termination criterion is set to reach the maximum number of the iteration or \( |f(x^k) - f(x^{k-1})| \leq 10^{-15} \). When regularization parameter \( \nu \) is zero, the LASSO problem becomes the least squares problem. Otherwise, the objective function \( f(x) \) is nonsmooth. Consider that the nonsmooth term \( \|x\|_1 \) is actually the sum of the absolute values of each component of \( x \), we employ the proximal gradient method and approximation method based on Huber function respectively to solve the LASSO problem. The Huber function becomes closer to the absolute value function as the parameter \( \theta \) approaches 0.

\[
h_\theta = \begin{cases} 
\frac{1}{2} \theta^2, & |x_i| < \theta \\
|x_i| - \frac{\theta}{2}, & \text{otherwise}
\end{cases} \tag{43}
\]

In order to study the influence of the previous gradients on search direction of optimization algorithm, we employ the gradient (denoted as \( g_1 \)), the sum of the preceding two gradients (denoted as \( g_2 \)), and the sum of the preceding three gradients (denoted as \( g_3 \)) to construct search directions respectively. Figure 1 shows the performance of gradient descent method (GD) for least square problem, proximal gradient method (PG) and approximation method based on Huber function (AH) for LASSO problem (\( \nu = 0.001 \)) when the above three search directions are employed. We set the parameter \( \theta \) as 0.02. For these algorithms with the three search directions, we set the step sizes as \( \frac{1}{L} \), \( 0.8 \frac{1}{L} \), and \( 0.4 \frac{1}{L} \) respectively. Observe that when the sum of the preceding two gradients (\( g_2 \)) is selected to construct the search direction, the optimization algorithms achieve better acceleration than them with other two search directions (\( g_1 \) and \( g_3 \)).

On the other hand, we compare the extended gradient method (our algorithm) with gradient descent method, heavy ball method ( iPiasco \( [2] \)) and Nesterov accelerated gradient descent method \( [5] \). Specifically, for the gradient descent method, the step size \( \alpha \) is \( \frac{1}{L} \). For the extended gradient method, the step size \( \alpha \) is \( 0.8 \frac{1}{L} \). For the heavy ball method, the step size \( \alpha \) is \( 4 \left( \sqrt{\frac{\nu}{\mu + \sqrt{L}}} \right)^2 \), and extrapolation parameter \( \beta \) is \( \left( \frac{\sqrt{\mu} - \sqrt{\nu}}{\sqrt{\mu + \sqrt{L}}} \right)^2 \). For the Nesterov accelerated gradient descent method, the extrapolation parameter is \( \frac{k}{k+1} \) and step size \( \alpha \) is \( \frac{1}{L} \). Figure 2 depicts the comparison result of all compared methods for the least squares problem. As we notice that the extended gradient method
Fig. 1: Performance of optimization algorithms with three search directions.

Fig. 2: Comparison of different algorithms for the least squares problem.

outperforms the other algorithms in terms of the number of iterations. Moreover, in order to show that the sum of two gradients ($g_2$) can accelerate convergence of algorithms, we use it to construct the new search direction of Nesterov accelerated gradient descent method and iPiasco. We set step size $\alpha$ and $\nu$ as $0.8\frac{1}{T}$ and $0.001$. From Fig. 3, we observe that when using the sum of two gradients to construct the new search direction, both proximal Nesterov accelerated gradient descent method and proximal iPiasco can achieve better acceleration performance.

4.2 Decentralized problem

We illustrate the performance of the decentralized extended gradient method ($DGD_{g_2}$) on the following problem

$$\min_{x \in \mathbb{R}^p} f(x) = \sum_{i=1}^{n} f_i(x), \quad \text{with} \quad f_i \equiv \frac{1}{2} ||M_i x - b_i||^2 + \nu ||x||_1,$$

(44)

where $M_i \in \mathbb{R}^{m \times p}$ is a random matrix generated from a uniform distribution $U(0,1)$ and each column of $M_i$ is normalized to be 1. We set the parameters $m = 512,$
Fig. 3: Performance of Nesterov accelerated gradient descent method and iPiasco with two search directions.

$p = 1024$, $r = 0.1$, $\nu = 0.0001$, and $b_i = M_i x$ where $x$ also follows the uniform distribution $U(0, 1)$ with sparsity $r$. The undirected graph is generated by Erdős-Rényi model [12] where each pair of nodes has the connection probability $p = 0.8$, and the weight matrix $W = \frac{I + M}{2}$, here $M$ is generated with the Metropolis constant edge rules [13]. To demonstrate the effect of previous gradient on search direction of decentralized optimization algorithms, we use the sum of the preceding two gradients ($g_2$) to construct the search direction in accelerated penalty method with consensus (APMC) [14], decentralized gradient descent (DGD) [15] and decentralized stochastic gradient descent (DSGD)[16]. For DGD, DSGD, $DGD_{g2}$ and $DSGD_{g2}$, we tune the optimized step size by hand, which are 0.09, 0.9, 0.058 and 0.58 respectively. For $APMC$ and $APMC_{g2}$, the step sizes are $\frac{1}{L}$ and $0.1 \frac{1}{L}$. The initial condition $X^0$ is $0_{n \times p}$.

Compared to the original $APMC$, DGD and DSGD, we can see from Fig. 4 that the algorithms with the preceding two gradients ($g_2$) to construct search direction outperform than that with the current negative gradient as the search direction.

5 Conclusions

The gradient methods have been widely employed to solve optimization problems. In this paper, we proposed an extended gradient method for smooth and strongly convex functions. Then, we proved the linear convergence of the extended gradient method. Besides, the numerical examples are used to demonstrate the convergence and to validate the efficiency compared to current classical methods. In the future, the acceleration technique used in this work can be applied in more optimization algorithms.

Acknowledgments. The authors gratefully appreciate the anonymous referees for the valuable comments and suggestions.

Funding. This work was supported by the National Natural Science Foundation of China under Grant No. 12271419 (S. Liu), the Natural Science Basic Research
Program of Shaanxi Province, China Grant No. 2023-JC-QN-0009 (N. Zhao), and the Fundamental Research Funds for the Central Universities, CHD300102122113 (N. Zhao).

**Author Contributions.** X.Z. and S.L. designed the work. X.Z. and N.Z. wrote the main manuscript text, and X.Z. prepared all figures. S.L. edited the manuscript. All authors reviewed the manuscript.

**Data availability.** Not applicable

**Declarations**

**Competing interests.** The authors declare that they have no conflict of interest.

**Ethical approval.** Not applicable

**References**


