COMPLEXITY OF PROJECTED GRADIENT METHODS FOR STRONGLY CONVEX OPTIMIZATION WITH HÖLDER CONTINUOUS GRADIENT TERMS*

XIAOJUN CHEN†, C. T. KELLEY‡, AND LEI WANG§

November 11, 2025

Abstract. This paper studies complexity of projected gradient descent methods for strongly convex constrained optimization problems where the objective function has α-Hölder ($0 < \alpha \leq 1$) continuous gradient terms. We first show that with an appropriately fixed stepsize, the complexity bound for finding an approximate minimizer with a distance to the true minimizer less than ε is $O(\log(\varepsilon^{-1})\varepsilon^{2(\alpha-1)/(1+\alpha)})$, which extends the well-known complexity result for $\alpha = 1$. Next we show that the complexity bound can be improved to $O(\log(\varepsilon^{-1})\varepsilon^{2(\alpha-1)/(1+3\alpha)})$ if the stepsize is updated by the universal scheme. We illustrate our complexity results by numerical examples arising from elliptic equations with a non-Lipschitz term.

Key words. Projected gradient descent, complexity, Hölder continuity

MSC codes. 90C25, 65L05, 65Y20

1. Introduction.

Given a closed and convex set $\Omega \subseteq \mathbb{R}^n$, this paper considers the following optimization problem,

$$\min_{\mathbf{u}\in\Omega} \; f(\mathbf{u}) := \frac{1}{m}\sum_{i=1}^{m} f_i(\mathbf{u}), \tag{1.1}$$

where the objective function $f : \mathbb{R}^n \to \mathbb{R}$ satisfies the following assumption.

Assumption 1.1.

1. *The function f is μ-strongly convex with a parameter $\mu > 0$ on Ω, that is,*

$$f(\mathbf{u}) \geq f(\mathbf{v}) + \langle \nabla f(\mathbf{v}), \mathbf{u} - \mathbf{v}\rangle + \frac{\mu}{2}\left\|\mathbf{u} - \mathbf{v}\right\|^2,$$

for all $\mathbf{u}, \mathbf{v} \in \Omega$.

2. *For each $i \in [m] := \{1, 2, \ldots, m\}$, $f_i : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable and the gradient ∇f_i is (globally) Hölder continuous with an exponent $\alpha_i \in (0, 1]$ on Ω, namely, there exists a constant $L_i > 0$ such that*

$$\|\nabla f_i(\mathbf{u}) - \nabla f_i(\mathbf{v})\| \leq L_i \left\|\mathbf{u} - \mathbf{v}\right\|^{\alpha_i}, \tag{1.2}$$

for all $\mathbf{u}, \mathbf{v} \in \Omega$.

Here, $\|\cdot\|$ is the ℓ_2 norm and $\langle\cdot,\cdot\rangle$ is the inner product on $\mathbb{R}^n$. We also denote by $\mathbf{u}^* \in \Omega$ and $f^* = f(\mathbf{u}^*)$ the global minimizer and the optimal value of

---

*Submitted to the editors DATE.

Funding: We would like to acknowledge support for this project from RGC grant JLFS/P-501/24 for the CAS AMSS-PolyU Joint Laboratory in Applied Mathematics and Hong Kong Research Grant Council project PolyU15300024.

†Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong, China (maxjchen@polyu.edu.hk).

‡Department of Mathematics, Box 8205, North Carolina State University, Raleigh, NC 27695-8205, USA (Tim_Kelley@ncsu.edu).

§Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong, China (lei2wang@polyu.edu.hk).

problem **??**, respectively. Let $\Pi_\Omega(\cdot)$ be the projection operator onto the set Ω and $\hat{\alpha} = \min_{i\in[m]} \alpha_i \in (0,1]$.

Suppose that each ∇f_i is Lipschitz continuous, which corresponds to condition **??** with $\alpha_i = 1$ for all $\mathbf{u}, \mathbf{v} \in \Omega$. Then ∇f is also Lipschitz continuous and the associated Lipschitz constant is $L = \sum_{i=1}^m L_i/m$. It is well known that the classical projected gradient descent method

$$\mathbf{u}_{k+1} = \Pi_\Omega\left(\mathbf{u}_k - \tau\nabla f(\mathbf{u}_k)\right), \tag{1.3}$$

with any initial point $\mathbf{u}_0 \in \mathbb{R}^n$ and the stepsize $\tau \in (0, 2/(\mu+L)]$, achieves a linear rate of convergence [**?**, Theorem 2.2.14] as follows,

$$\|\mathbf{u}_k - \mathbf{u}^*\| \le (1-\mu\tau)^k \|\mathbf{u}_0 - \mathbf{u}^*\|.$$

Therefore, for a given $\varepsilon > 0$, method (**??**) is guaranteed to find a point $\mathbf{u}_k \in \Omega$ satisfying $\|\mathbf{u}_k - \mathbf{u}^*\| \le \varepsilon$ after at most $O(\log(\varepsilon^{-1}))$ iterations. Unfortunately, this analysis fails if there exists at least one index $i \in [m]$ such that $\alpha_i < 1$. We explain the failure of the convergence of method (**??**) to $\mathbf{u}^*$ by the following example.

Example 1.1. [**?**, Example 1] Consider the following univariate optimization problem,

$$\min_{x\in\mathbb{R}} \; f(x) = \frac{1}{2}x^2 + \frac{2}{3}|x|^{3/2},$$

which is a special instance of problem **??** with $f_1(x) = x^2/2$, $f_2(x) = 2|x|^{3/2}/3$, and $\Omega = \mathbb{R}$. It is easy to see that the global minimizer is $x^* = 0$. Method **??** with the fixed stepsize $\tau > 0$ starting from $x_0 \neq 0$ reads as follows,

$$x_{k+1} = x_k - \tau\nabla f(x_k) = (1-\tau)\,x_k - \tau\,\mathsf{sign}(x_k)\,|x_k|^{1/2}.$$

A straightforward verification reveals that

$$|x_{k+1}|^2 - |x_k|^2 = -\tau(2-\tau)|x_k|^2 - 2\tau(1-\tau)|x_k|^{3/2} + \tau^2|x_k|.$$

It is evident that, when $|x_k|$ is sufficiently small, the last term in the right-hand side becomes dominant, resulting in that $|x_{k+1}|^2 - |x_k|^2 \ge 0$. Therefore, the distance to the global minimizer ceases to decrease once it achieves a certain level.

Moreover, in [**?**] we show that ∇f is locally $\frac{1}{2}-$Hölder continuous, but not globally Hölder continuous. In fact, from

$$\nabla f(x^* + |h|) - \nabla f(x^*) = |h| + |h|^{\frac{1}{2}} = (|h|^{1-\alpha} + |h|^{\frac{1}{2}-\alpha})|h|^\alpha =: (\hat{L}_1(h) + \hat{L}_2(h))|h|^\alpha,$$

we have $\hat{L}_1(h) \to \infty$ when $\alpha = (0,1)$ and $|h| \to \infty$, while $\hat{L}_2(h) \to \infty$ when $\alpha = 1$ and $|h| \to 0$.

This example demonstrates that a function f expressed as a sum of component functions f_i, each endowed with a Hölder continuous gradient, may itself fail to possess a Hölder continuous gradient. This phenomenon was revisited by Nesterov (see [**?**, Example 1]).

On the other hand, this example satisfies Assumption 1.1 (ii) as

$$|\nabla f_1(x) - \nabla f_1(y)| \le L_1|x-y| \quad \text{and} \quad |\nabla f_2(x) - \nabla f_2(y)| \le L_2|x-y|^{\frac{1}{2}},\ \forall x, y \in \mathbb{R}$$

with $L_1 = L_2 = 1$.

In [?], the authors presented the following bound for method (??)

$$f(\mathbf{u}_k) - f(\mathbf{u}^*) \le K(N) := \frac{L_\alpha \|\mathbf{u}_0 - \mathbf{u}^*\|^{1+\alpha}}{1+\alpha} \left(\frac{2}{N}\right)^{\frac{1+\alpha}{2}},$$

where L_α is the $\alpha-$Hölder Lipschitz constant and $\hat{\mathbf{u}}_N = \sum_{k=1}^N \mathbf{u}_k/N$. In the unconstrained case, (51) in [?] comes to

$$\|\hat{\mathbf{u}}_N - \mathbf{u}^*\|^2 \le \frac{2}{\mu} K(N),$$

which implies that finding an N average of iterations $\hat{\mathbf{u}}_N$ satisfying $\|\hat{\mathbf{u}}_N - \mathbf{u}^*\| \le \epsilon$ requires $O(\epsilon^{-4/(1+\alpha)})$ iterations.

The contribution of this paper is to provide new complexity results of the projected gradient descent method for problem (??) when the objective function is strongly convex, but its gradient is not Lipschitz due to a α-Hölder continuous term with $0 < \alpha < 1$. We first show that with an appropriately fixed stepsize, the complexity bound for finding the global minimizer less than ε is $O(\log(\varepsilon^{-1})\varepsilon^{2(\alpha-1)/(1+\alpha)})$, which extends the well-known complexity result for $\alpha = 1$. Next we show that the complexity bound can be improved to $O(\log(\varepsilon^{-1})\varepsilon^{2(\alpha-1)/(1+3\alpha)})$ if the stepsize is updated at each step using the universal scheme. Our complexity bound is at least $O(\epsilon^{-1})$ lower than (51) in [?]. For example, when $\alpha = 1$, our bound is $O(\log(\epsilon^{-1}))$ but (51) in [?] is $O(\epsilon^{-2})$.

Our study is motivated by elliptic equations with a non-Lipschitz term [?, ?], as well as optimization problems with a ℓ_p-norm ($1 < p < 2$) regularization term [?, ?]. We illustrate our complexity results by two numerical examples arising from elliptic equations with a non-Lipschitz term in Section 5, after we present complexity of projected gradient methods with fixed stepsize and updated stepsize in Sections 2-4, respectively.

2. Vanilla Projected Gradient Descent Method with a Fixed Stepsize.

In this section, we attempt to employ the vanilla projected gradient descent method **??** with a fixed stepsize to solve problem **??**, whose complexity bound is also provided. Example 1.1 illustrates that the projected gradient descent method with a fixed stepsize **??** will experience stagnation before reaching the global minimizer.

To obtain an approximate solution to problem **??**, it is necessary to choose a sufficiently small stepsize τ in the projected gradient descent method **??**, the magnitude of which depends on the desired level of accuracy. Let $M > 0$ be a constant defined as

$$M = \max_{i \in [m]} \left\{ \left[\frac{2(1-\alpha_i)}{\mu(1+\alpha_i)} \right]^{(1-\alpha_i)/(1+\alpha_i)} L_i^{2/(1+\alpha_i)} \right\}. \tag{2.1}$$

We select a specific stepsize $\tau = \varepsilon^{2(1-\hat{\alpha})/(1+\hat{\alpha})}/M$ in the projected gradient descent method, whose complete framework is presented in **??**. Two sequences $\{\mathbf{v}_k\}$ and $\{\mathbf{u}_k\}$ are maintained in **??**, where $\mathbf{v}_k$ is generated by the projected gradient descent method and $\mathbf{u}_k$ corresponds to the iterate achieving the smallest objective function value among the first k iterations.

Our subsequent analysis is based on the inexact oracle [?] derived from the Hölder continuity condition of gradients, which is generalized to problem **??** and demonstrated in the following proposition.

Algorithm 1: Projected Gradient Descent Method (PGDM).

Input: $\varepsilon > 0$.
Initialize $\mathbf{u}_0 = \mathbf{v}_0 \in \Omega$.
Choose the stepsize $\tau = \varepsilon^{2(1-\hat{\alpha})/(1+\hat{\alpha})}/M$.
for $k = 0, 1, 2, \ldots$ **do**
Compute

$$\mathbf{v}_{k+1} = \Pi_\Omega \left(\mathbf{v}_k - \tau \nabla f(\mathbf{v}_k)\right).$$

Set $\mathbf{u}_{k+1} = \begin{cases} \mathbf{v}_{k+1} \text{ if } f(\mathbf{v}_{k+1}) \leq f(\mathbf{u}_k) \\ \mathbf{u}_k \text{ otherwise} \end{cases}$.

Output: $\mathbf{u}_{k+1}$.

PROPOSITION 2.1. *Suppose that* **??** *holds. Let* $\delta > 0$ *and*

$$\rho \geq \max_{i\in[m]} \left\{ \left[\frac{1-\alpha_i}{(1+\alpha_i)\delta}\right]^{(1-\alpha_i)/(1+\alpha_i)} L_i^{2/(1+\alpha_i)} \right\}.$$

Then for all $\mathbf{u}, \mathbf{v} \in \Omega$*, we have*

$$f(\mathbf{v}) \leq f(\mathbf{u}) + \langle \nabla f(\mathbf{u}), \mathbf{v} - \mathbf{u}\rangle + \frac{\rho}{2}\|\mathbf{v} - \mathbf{u}\|^2 + \frac{\delta}{2}.$$

Proof. Since ∇f_i is Hölder continuous with an exponent α_i, we can obtain that

$$f_i(\mathbf{v}) \leq f_i(\mathbf{u}) + \langle \nabla f_i(\mathbf{u}), \mathbf{v} - \mathbf{u}\rangle + \frac{L_i}{1+\alpha_i}\|\mathbf{v} - \mathbf{u}\|^{1+\alpha_i},$$

for all $\mathbf{u}, \mathbf{v} \in \Omega$. Then, for each i, it follows from [**?**, Lemma 2] that

$$f_i(\mathbf{v}) \leq f_i(\mathbf{u}) + \langle \nabla f_i(\mathbf{u}), \mathbf{v} - \mathbf{u}\rangle + \frac{\rho}{2}\|\mathbf{v} - \mathbf{u}\|^2 + \frac{\delta}{2}.$$

Summing the above relationship over $i \in [m]$, we immediately arrive at the assertion of this proposition. The proof is completed. ∎

Now, we are in the position to derive the complexity bound of **??** in the following theorem.

THEOREM 2.2. *Let* $\varepsilon \in (0, 1)$ *be a sufficiently small constant. Then after at most*

$$O\left(\log\left(\frac{1}{\varepsilon}\right)\frac{1}{\varepsilon^{2(1-\hat{\alpha})/(1+\hat{\alpha})}}\right)$$

iterations, **??** *will find an iterate* $\mathbf{u}_k \in \Omega$ *satisfying* $\|\mathbf{u}_k - \mathbf{u}^*\| \leq \varepsilon$*.*

Proof. In view of **??**, we take

$$\rho = \frac{1}{\tau} = \frac{M}{\varepsilon^{2(1-\hat{\alpha})/(1+\hat{\alpha})}} \geq \max_{i\in[m]} \left\{ \left[\frac{2(1-\alpha_i)}{\mu(1+\alpha_i)\varepsilon^2}\right]^{(1-\alpha_i)/(1+\alpha_i)} L_i^{2/(1+\alpha_i)} \right\}.$$

Then it holds that

$$f(\mathbf{v}_{k+1}) \leq f(\mathbf{v}_k) + \langle \nabla f(\mathbf{v}_k), \mathbf{v}_{k+1} - \mathbf{v}_k\rangle + \frac{1}{2\tau}\|\mathbf{v}_{k+1} - \mathbf{v}_k\|^2 + \frac{\mu\varepsilon^2}{4},$$

which, after a suitable rearrangement, can be equivalently written as

$$\langle \nabla f(\mathbf{v}_k), \mathbf{v}_k - \mathbf{v}_{k+1} \rangle \leq f(\mathbf{v}_k) - f(\mathbf{v}_{k+1}) + \frac{\mu\varepsilon^2}{4} + \frac{1}{2\tau} \|\mathbf{v}_{k+1} - \mathbf{v}_k\|^2 . \tag{2.2}$$

Recall that $f^* = f(\mathbf{u}^*)$. By virtue of the strong convexity of f, we can obtain that

$$\langle \nabla f(\mathbf{v}_k), \mathbf{u}^* - \mathbf{v}_k \rangle \leq f^* - f(\mathbf{v}_k) - \frac{\mu}{2} \|\mathbf{v}_k - \mathbf{u}^*\|^2 . \tag{2.3}$$

The optimality condition of the projection problem defining $\mathbf{v}_{k+1}$ yields that

$$\langle \mathbf{v}_{k+1} - \mathbf{v}_k + \tau \nabla f(\mathbf{v}_k), \mathbf{u} - \mathbf{v}_{k+1} \rangle \geq 0,$$

for all $\mathbf{u} \in \Omega$. Upon taking $\mathbf{u} = \mathbf{u}^*$, we have

$$\begin{aligned} \langle \mathbf{v}_{k+1} - \mathbf{v}_k, \mathbf{v}_{k+1} - \mathbf{u}^* \rangle &\leq \tau \langle \nabla f(\mathbf{v}_k), \mathbf{u}^* - \mathbf{v}_{k+1} \rangle \\ &= \tau \langle \nabla f(\mathbf{v}_k), \mathbf{u}^* - \mathbf{v}_k \rangle + \tau \langle \nabla f(\mathbf{v}_k), \mathbf{v}_k - \mathbf{v}_{k+1} \rangle , \end{aligned}$$

which together with **??** and **??** implies that

$$\begin{aligned} \langle \mathbf{v}_{k+1} - \mathbf{v}_k, \mathbf{v}_{k+1} - \mathbf{u}^* \rangle \leq \tau \left(f^* - f(\mathbf{v}_{k+1}) + \frac{\mu\varepsilon^2}{4} \right) - \frac{\mu\tau}{2} \|\mathbf{v}_k - \mathbf{u}^*\|^2 \\ + \frac{1}{2} \|\mathbf{v}_{k+1} - \mathbf{v}_k\|^2 . \end{aligned}$$

Moreover, it can be readily verified that

$$\begin{aligned} \|\mathbf{v}_{k+1} - \mathbf{u}^*\|^2 &= \|\mathbf{v}_{k+1} - \mathbf{v}_k + \mathbf{v}_k - \mathbf{u}^*\|^2 \\ &= \|\mathbf{v}_k - \mathbf{u}^*\|^2 + 2 \langle \mathbf{v}_{k+1} - \mathbf{v}_k, \mathbf{v}_k - \mathbf{u}^* \rangle + \|\mathbf{v}_{k+1} - \mathbf{v}_k\|^2 \\ &= \|\mathbf{v}_k - \mathbf{u}^*\|^2 + 2 \langle \mathbf{v}_{k+1} - \mathbf{v}_k, \mathbf{v}_{k+1} - \mathbf{u}^* \rangle - \|\mathbf{v}_{k+1} - \mathbf{v}_k\|^2 . \end{aligned} \tag{2.4}$$

Collecting the above two relationships together, we arrive at

$$\|\mathbf{v}_{k+1} - \mathbf{u}^*\|^2 \leq (1 - \mu\tau) \|\mathbf{v}_k - \mathbf{u}^*\|^2 + 2\tau \left(f^* - f(\mathbf{v}_{k+1}) + \frac{\mu\varepsilon^2}{4} \right).$$

From the construction of $\mathbf{u}_k$ in **??**, it then follows that $f(\mathbf{v}_l) \geq f(\mathbf{u}_k)$ for all $l \in \{1, 2, \ldots, k\}$. Let $C_k = \sum_{l=1}^{k} (1 - \mu\tau)^{l-1}$ be a constant. Applying the above relationship recursively for k times leads to that

$$\begin{aligned} \|\mathbf{v}_k - \mathbf{u}^*\|^2 \leq\ & (1 - \mu\tau)^k \|\mathbf{u}_0 - \mathbf{u}^*\|^2 + 2\tau \sum_{l=1}^{k} (1 - \mu\tau)^{l-1} \left(f^* - f(\mathbf{v}_l) + \frac{\mu\varepsilon^2}{4} \right) \\ \leq\ & (1 - \mu\tau)^k \|\mathbf{u}_0 - \mathbf{u}^*\|^2 + 2\tau \left(f^* - f(\mathbf{u}_k) + \frac{\mu\varepsilon^2}{4} \right) C_k, \end{aligned}$$

which together with $\|\mathbf{v}_k - \mathbf{u}^*\| \geq 0$ and $C_k \geq 1$ implies that

$$f(\mathbf{u}_k) - f^* \leq \frac{(1 - \mu\tau)^k}{2\tau C_k} \|\mathbf{u}_0 - \mathbf{u}^*\|^2 + \frac{\mu\varepsilon^2}{4} \leq \frac{(1 - \mu\tau)^k}{2\tau} \|\mathbf{u}_0 - \mathbf{u}^*\|^2 + \frac{\mu\varepsilon^2}{4}.$$

According to the strong convexity of f and the optimality condition of problem **??**, we have

$$f(\mathbf{u}_k) - f^* \geq \langle \nabla f(\mathbf{u}^*), \mathbf{u}_k - \mathbf{u}^* \rangle + \frac{\mu}{2} \|\mathbf{u}_k - \mathbf{u}^*\|^2 \geq \frac{\mu}{2} \|\mathbf{u}_k - \mathbf{u}^*\|^2 . \tag{2.5}$$

Hence, it holds that

$$\begin{aligned}\|\mathbf{u}_k-\mathbf{u}^*\|^2 &\le \frac{2}{\mu}\left(f(\mathbf{u}_k)-f^*\right) \le \frac{(1-\mu\tau)^k}{\mu\tau}\|\mathbf{u}_0-\mathbf{u}^*\|^2+\frac{\varepsilon^2}{2}\\ &\le \frac{M\|\mathbf{u}_0-\mathbf{u}^*\|^2}{\mu\varepsilon^{2(1-\hat\alpha)/(1+\hat\alpha)}}\left(1-\frac{\mu}{M}\varepsilon^{2(1-\hat\alpha)/(1+\hat\alpha)}\right)^k+\frac{\varepsilon^2}{2}.\end{aligned}$$

We denote by K_ε^* the smallest iteration number k such that $\|\mathbf{u}_k-\mathbf{u}^*\|\le\varepsilon$. Then solving the inequality $M\|\mathbf{u}_0-\mathbf{u}^*\|^2\varepsilon^{-2(1-\hat\alpha)/(1+\hat\alpha)}(1-\mu\varepsilon^{2(1-\hat\alpha)/(1+\hat\alpha)}/M)^k/\mu\le\varepsilon^2/2$ indicates that

$$\begin{aligned}K_\varepsilon^* &\le \frac{4\log((2M\|\mathbf{u}_0-\mathbf{u}^*\|^2/\mu)^{(1+\hat\alpha)/4}/\varepsilon)}{-\log(1-\mu\varepsilon^{2(1-\hat\alpha)/(1+\hat\alpha)}/M)(1+\hat\alpha)}\\ &\le \frac{4M\log((2M\|\mathbf{u}_0-\mathbf{u}^*\|^2/\mu)^{(1+\hat\alpha)/4}/\varepsilon)}{\mu(1+\hat\alpha)\varepsilon^{2(1-\hat\alpha)/(1+\hat\alpha)}}.\end{aligned}$$

The proof is completed. ∎

?? demonstrates that the iteration complexity of **??** with a fixed stepsize is $O(\log(\varepsilon^{-1})\varepsilon^{-2(1-\hat\alpha)/(1+\hat\alpha)})$ for problem **??**. This complexity result generalizes the classical linear convergence when $m=1$ and $\hat\alpha=1$, which highlights the performance degradation incurred by non-Lipschitz gradients.

3. Universal Primal Gradient Method.

The fixed stepsize τ chosen in **??** depends on the parameters α_i and L_i for all $i\in[m]$, which are often unknown and hard to estimate in practice. To address this issue, we adopt the universal primal gradient method (UPGM) proposed by Nesterov [**?**] to solve problem **??**. This method incorporates a line-search procedure to adaptively determine the stepsize at each iteration, and its overall framework is outlined in **??**.

Next, we establish the iteration complexity of **??**, which remains on the same order as that of the projected gradient descent method with a fixed stepsize.

THEOREM 3.1. *Let $\varepsilon\in(0,1)$ be a sufficiently small constant. Then after at most*

$$O\left(\log\left(\frac{1}{\varepsilon}\right)\frac{1}{\varepsilon^{2(1-\hat\alpha)/(1+\hat\alpha)}}\right)$$

iterations, **??** *will attain an iterate $\mathbf{u}_k\in\Omega$ satisfying that $\|\mathbf{u}_k-\mathbf{u}^*\|\le\varepsilon$.*

Proof. Obviously, there exists $j_k\in\mathbb{N}$ such that

$$2^{j_k}\rho_k \ge \max_{i\in[m]}\left\{\left[\frac{2(1-\alpha_i)}{\mu(1+\alpha_i)\varepsilon^2}\right]^{(1-\alpha_i)/(1+\alpha_i)}L_i^{2/(1+\alpha_i)}\right\}.$$

By invoking the results of **??**, we know that condition **??** is satisfied. Hence, the line-search step in **??** can be terminated after a finite number of trials and the required number of trials j_k satisfies

$$2^{j_k}\rho_k \le 2\max_{i\in[m]}\left\{\left[\frac{2(1-\alpha_i)}{\mu(1+\alpha_i)\varepsilon^2}\right]^{(1-\alpha_i)/(1+\alpha_i)}L_i^{2/(1+\alpha_i)}\right\}\le\frac{2M}{\varepsilon^{2(1-\hat\alpha)/(1+\hat\alpha)}}, \tag{3.2}$$

where $M>0$ is a constant defined in **??**. Moreover, the line-search condition **??**

Algorithm 2: Universal Primal Gradient Method (UPGM).

Input: $\varepsilon > 0$.
Initialize $\mathbf{u}_0 = \mathbf{v}_0 \in \Omega$ and $\rho_0 > 0$.
for $k = 0, 1, 2, \ldots$ **do**
  for $j_k = 0, 1, 2, \ldots$ **do**
    Compute

$$\mathbf{v}_{k+1} = \Pi_\Omega\left(\mathbf{v}_k - \frac{1}{2^{j_k}\rho_k}\nabla f(\mathbf{v}_k)\right).$$

    If $\mathbf{v}_{k+1}$ satisfies the following line-search condition,

$$\begin{aligned} f(\mathbf{v}_{k+1}) \le f(\mathbf{v}_k) &+ \langle \nabla f(\mathbf{v}_k), \mathbf{v}_{k+1} - \mathbf{v}_k\rangle \\ &+ \frac{2^{j_k}\rho_k}{2}\left\|\mathbf{v}_{k+1} - \mathbf{v}_k\right\|^2 + \frac{\mu\varepsilon^2}{4}, \end{aligned} \tag{3.1}$$

    then break.
  Update $\rho_{k+1} = 2^{j_k}\rho_k$.
  Choose $\bar{k} = \max\{k^* \mid k^* \in \arg\min_{l\in\{0,1,\ldots,k\}} f(\mathbf{v}_{l+1})\}$.
  Set $\mathbf{u}_{k+1} = \mathbf{v}_{\bar{k}+1}$.
Output: $\mathbf{u}_{k+1}$.

directly yields that

$$\langle \nabla f(\mathbf{v}_k), \mathbf{v}_k - \mathbf{v}_{k+1}\rangle \le f(\mathbf{v}_k) - f(\mathbf{v}_{k+1}) + \frac{2^{j_k}\rho_k}{2}\left\|\mathbf{v}_{k+1} - \mathbf{v}_k\right\|^2 + \frac{\mu\varepsilon^2}{4}. \tag{3.3}$$

According to the optimality condition of the projection problem defining $\mathbf{v}_{k+1}$, we have

$$\left\langle \mathbf{v}_{k+1} - \mathbf{v}_k + \frac{1}{2^{j_k}\rho_k}\nabla f(\mathbf{v}_k), \mathbf{u}^* - \mathbf{v}_{k+1}\right\rangle \ge 0,$$

which further implies that

$$\begin{aligned} \langle \mathbf{v}_{k+1} - \mathbf{v}_k, \mathbf{v}_{k+1} - \mathbf{u}^*\rangle &\le \frac{1}{2^{j_k}\rho_k}\left\langle \nabla f(\mathbf{v}_k), \mathbf{u}^* - \mathbf{v}_{k+1}\right\rangle \\ &\le \frac{1}{2^{j_k}\rho_k}\left\langle \nabla f(\mathbf{v}_k), \mathbf{u}^* - \mathbf{v}_k\right\rangle + \frac{1}{2^{j_k}\rho_k}\left\langle \nabla f(\mathbf{v}_k), \mathbf{v}_k - \mathbf{v}_{k+1}\right\rangle. \end{aligned}$$

Substituting **??** and **??** into the above relationship leads to that

$$\begin{aligned} \langle \mathbf{v}_{k+1} - \mathbf{v}_k, \mathbf{v}_{k+1} - \mathbf{u}^*\rangle \le{}& \frac{1}{2^{j_k}\rho_k}\left(f^* - f(\mathbf{v}_{k+1}) + \frac{\mu\varepsilon^2}{4}\right) \\ &+ \frac{1}{2}\left\|\mathbf{v}_{k+1} - \mathbf{v}_k\right\|^2 - \frac{\mu}{2^{j_k+1}\rho_k}\left\|\mathbf{v}_k - \mathbf{u}^*\right\|^2, \end{aligned}$$

Thus, it follows from relationship **??** that

$$\begin{aligned}\|\mathbf{v}_{k+1}-\mathbf{u}^*\|^2 &\leq \left(1-\frac{\mu}{2^{j_k}\rho_k}\right)\|\mathbf{v}_k-\mathbf{u}^*\|^2+\frac{2}{2^{j_k}\rho_k}\left(f^*-f(\mathbf{v}_{k+1})+\frac{\mu\varepsilon^2}{4}\right)\\ &\leq \left(1-\frac{\mu\varepsilon^{2(1-\hat{\alpha})/(1+\hat{\alpha})}}{2M}\right)\|\mathbf{v}_k-\mathbf{u}^*\|^2+\frac{2}{\rho_0}\left(f^*-f(\mathbf{v}_{k+1})+\frac{\mu\varepsilon^2}{4}\right),\end{aligned}$$

where the last inequality comes from **??** and $2^{j_k}\rho_k \geq \rho_0$. The remaining part of the proof follows the same line of reasoning as that of **??** and is therefore omitted here for the sake of brevity. □

We end this section by estimating the total number of line-search steps required by **??**.

COROLLARY 3.2. *Let $\varepsilon \in (0,1)$ be a sufficiently small constant. Then* **??** *requires at most*

$$O\left(\log\left(\frac{1}{\varepsilon}\right)\frac{1}{\varepsilon^{2(1-\hat{\alpha})/(1+\hat{\alpha})}}\right)$$

line-search steps for the generated sequence $\{\mathbf{u}_k\}$ to satisfy $\|\mathbf{u}_k-\mathbf{u}^*\| \leq \varepsilon$.

Proof. Let N_k be the total number of line-search steps after k iterations in **??**. From the update rule $\rho_{k+1} = 2^{j_k}\rho_k$, we can obtain that $j_k = \log\rho_{k+1} - \log\rho_k$. Then a straightforward verification reveals that

$$N_k = \sum_{l=0}^{k}(j_l+1) = k+1+\log\rho_{k+1}-\log\rho_0, \tag{3.4}$$

which together with relationship **??** implies that

$$\begin{aligned}N_k &\leq k+1+\log\left(\frac{2M}{\varepsilon^{2(1-\hat{\alpha})/(1+\hat{\alpha})}}\right)-\log\rho_0\\ &\leq k+\frac{2(1-\hat{\alpha})}{1+\hat{\alpha}}\log\left(\frac{1}{\varepsilon}\right)+\log\left(\frac{2M}{\rho_0}\right)+1.\end{aligned}$$

By invoking the results of **??**, we conclude that **??** requires at most $O(\log(\varepsilon^{-1})\varepsilon^{-2(1-\hat{\alpha})/(1+\hat{\alpha})})$ line-search steps, which completes the proof. □

At each iteration of **??**, we evaluate both the function value and the gradient at $\mathbf{v}_k$. In addition, an extra function evaluation at $\mathbf{v}_{k+1,j_k}$ is involved during each line-search step. Therefore, **??** and **??** together reveal that the total number of function and gradient evaluations required by **??** is $O(\log(\varepsilon^{-1})\varepsilon^{-2(1-\hat{\alpha})/(1+\hat{\alpha})})$.

4. Universal Fast Gradient Method. To obtain a sharper complexity bound, we devise in this section a universal fast gradient method (UFGM) tailored to problem **??**. The proposed scheme, summarized in **??**, exhibits slight but essential differences from the algorithm introduced by Nesterov [**?**] to exploit the strong convexity of the objective function.

The following lemma illustrates that the line-search process in **??** is well-defined, which is guaranteed to terminate in a finite number of trials.

LEMMA 4.1. *There exists an integer $j_k \in \mathbb{N}$ such that the line-search condition* **??** *is satisfied in* **??**.

Algorithm 3: Universal Fast Gradient Method (UFGM).

Input: $\varepsilon > 0$.

Initialize $\mathbf{u}_0 = \mathbf{w}_0 \in \Omega$, $\rho_0 \geq \mu$, and $\sigma_0 = 1$.

for $k = 0, 1, 2, \ldots$ **do**

  for $j_k = 0, 1, 2, \ldots$ **do**

    Set $\nu_k = \sqrt{\mu/(2^{j_k}\rho_k)}$, $\theta_k = \nu_k \sigma_k$, and $\sigma_{k+1} = \sigma_k + \theta_k$.

    Compute

$$\mathbf{v}_k = \frac{\sigma_k}{\sigma_k + \theta_k}\mathbf{u}_k + \frac{\theta_k}{\sigma_k + \theta_k}\Pi_\Omega(\mathbf{w}_k), \tag{4.1}$$

    and

$$\mathbf{z}_k = \Pi_\Omega\left(\Pi_\Omega(\mathbf{w}_k) - \frac{\theta_k}{\sigma_k \mu}\nabla f(\mathbf{v}_k)\right). \tag{4.2}$$

    Set

$$\mathbf{u}_{k+1} = \frac{\sigma_k}{\sigma_k + \theta_k}\mathbf{u}_k + \frac{\theta_k}{\sigma_k + \theta_k}\mathbf{z}_k. \tag{4.3}$$

    If $\mathbf{u}_{k+1}$ satisfies the following line-search condition,

$$\begin{aligned} f(\mathbf{u}_{k+1}) \leq f(\mathbf{v}_k) &+ \langle \nabla f(\mathbf{v}_k), \mathbf{u}_{k+1} - \mathbf{v}_k \rangle \\ &+ \frac{\sigma_k^2 \mu}{2\theta_k^2}\left\|\mathbf{u}_{k+1} - \mathbf{v}_k\right\|^2 + \frac{\theta_k \mu \varepsilon^2}{4\sigma_{k+1}}, \end{aligned} \tag{4.4}$$

    then break.

  Set $\rho_{k+1} = 2^{j_k}\rho_k$ and update $\mathbf{w}_{k+1}$ by

$$\mathbf{w}_{k+1} = \frac{1}{\sigma_{k+1}}\left(\sigma_k \mathbf{w}_k + \theta_k \mathbf{v}_k - \frac{\theta_k}{\mu}\nabla f(\mathbf{v}_k)\right). \tag{4.5}$$

Output: $\mathbf{u}_{k+1}$.

Proof. It follows from the definition of θ_k and $\nu_k \leq 1$ that

$$\frac{\theta_k}{\sigma_{k+1}} = \frac{\theta_k}{\sigma_k + \theta_k} = \frac{\nu_k}{1 + \nu_k} \geq \frac{\nu_k}{2},$$

and

$$\frac{\sigma_k^2 \mu}{\theta_k^2} = \frac{\mu}{\nu_k^2} = 2^{j_k}\rho_k.$$

Recall that $\hat{\alpha} = \min_{i\in[m]} \alpha_i \in (0,1]$. Then we have

$$\begin{aligned}\frac{\sigma_k^2\mu}{\theta_k^2}\left[\frac{\theta_k}{\sigma_{k+1}}\right]^{(1-\hat{\alpha})/(1+\hat{\alpha})} &\geq \frac{2^{j_k}\rho_k}{2^{(1-\hat{\alpha})/(1+\hat{\alpha})}}\nu_k^{(1-\hat{\alpha})/(1+\hat{\alpha})}\\ &= \frac{2^{j_k}\rho_k}{2^{(1-\hat{\alpha})/(1+\hat{\alpha})}}\left[\frac{\mu}{2^{j_k}\rho_k}\right]^{(1-\hat{\alpha})/(2(1+\hat{\alpha}))}\\ &= \frac{\mu^{(1-\hat{\alpha})/(2(1+\hat{\alpha}))}}{2^{(1-\hat{\alpha})/(1+\hat{\alpha})}}\left[2^{j_k}\rho_k\right]^{(1+3\hat{\alpha})/(2(1+\hat{\alpha}))},\end{aligned}$$

where the first equality comes from the definition of ν_k. Now it is clear that

$$\frac{\sigma_k^2\mu}{\theta_k^2}\left[\frac{\theta_k}{\sigma_{k+1}}\right]^{(1-\hat{\alpha})/(1+\hat{\alpha})} \to \infty,$$

as $j_k \to \infty$. Thus, there exists $j_k \in \mathbb{N}$ such that

$$\frac{\sigma_k^2\mu}{\theta_k^2}\left[\frac{\theta_k}{\sigma_{k+1}}\right]^{(1-\hat{\alpha})/(1+\hat{\alpha})} \geq \max_{i\in[m]}\left\{\left[\frac{2(1-\alpha_i)}{\mu(1+\alpha_i)\varepsilon^2}\right]^{(1-\alpha_i)/(1+\alpha_i)} L_i^{2/(1+\alpha_i)}\right\}, \tag{4.6}$$

which further implies that

$$\begin{aligned}\frac{\sigma_k^2\mu}{\theta_k^2} &\geq \left[\frac{\sigma_{k+1}}{\theta_k}\right]^{(1-\hat{\alpha})/(1+\hat{\alpha})}\max_{i\in[m]}\left\{\left[\frac{2(1-\alpha_i)}{\mu(1+\alpha_i)\varepsilon^2}\right]^{(1-\alpha_i)/(1+\alpha_i)} L_i^{2/(1+\alpha_i)}\right\}\\ &\geq \max_{i\in[m]}\left\{\left[\frac{2(1-\alpha_i)\sigma_{k+1}}{\mu(1+\alpha_i)\theta_k\varepsilon^2}\right]^{(1-\alpha_i)/(1+\alpha_i)} L_i^{2/(1+\alpha_i)}\right\}.\end{aligned}$$

As a direct consequence of **??**, we can proceed to show that the line-search condition **??** is satisfied, which completes the proof. ∎

LEMMA 4.2. *Let $\{\phi_k\}$ be a sequence of functions defined recursively by*

$$\begin{aligned}\phi_{k+1}(\mathbf{u}) = \phi_k(\mathbf{u}) - \theta_k f^* + \theta_k f(\mathbf{v}_k) + \theta_k\left\langle \nabla f(\mathbf{v}_k), \mathbf{u}-\mathbf{v}_k\right\rangle\\ + \frac{\theta_k\mu}{2}\left\|\mathbf{u}-\mathbf{v}_k\right\|^2,\end{aligned} \tag{4.7}$$

with $\phi_0(\mathbf{u}) = c_0 + \sigma_0\mu\left\|\mathbf{u}-\mathbf{w}_0\right\|^2/2$ for $c_0 = f(\mathbf{u}_0) - f^* - \mu\varepsilon^2/4$, $\sigma_0 = 1$, and $\mathbf{w}_0 \in \Omega$. Then, for all $k \in \mathbb{N}$, the function ϕ_k preserves the following canonical form,

$$\phi_k(\mathbf{u}) = c_k + \frac{\sigma_k\mu}{2}\left\|\mathbf{u}-\mathbf{w}_k\right\|^2, \tag{4.8}$$

where $\{c_k\}$ is a sequence of real numbers and $\{\mathbf{w}_k\}$ is defined recursively by **??**.

Proof. We first prove that $\nabla^2\phi_k = \sigma_k\mu I$ for all $k \in \mathbb{N}$ by induction. It is evident that $\nabla^2\phi_0 = \sigma_0\mu I$. Now we assume that $\nabla^2\phi_k = \sigma_k\mu I$ for some k. Then relationships **??** and $\sigma_{k+1} = \sigma_k + \theta_k$ imply that

$$\nabla^2\phi_{k+1} = \nabla^2\phi_k + \theta_k\mu I = \sigma_k\mu I + \theta_k\mu I = \sigma_{k+1}\mu I.$$

Thus, we know that $\nabla^2\phi_k = \sigma_k\mu I$ for all $k \in \mathbb{N}$, which, in turn, justifies the canonical form of ϕ_k in **??**.

Next, by combining two relationships **??** and **??** together, we can obtain that

$$\phi_{k+1}(\mathbf{u}) = c_k + \frac{\sigma_k \mu}{2} \left\| \mathbf{u} - \mathbf{w}_k \right\|^2 - \theta_k f^* + \theta_k f(\mathbf{v}_k) \\ + \theta_k \left\langle \nabla f(\mathbf{v}_k), \mathbf{u} - \mathbf{v}_k \right\rangle + \frac{\theta_k \mu}{2} \left\| \mathbf{u} - \mathbf{v}_k \right\|^2 .$$

Since $\mathbf{w}_{k+1}$ is a global minimizer of ϕ_{k+1} over $\mathbb{R}^n$, the first-order optimality condition yields that

$$\begin{aligned} 0 = \nabla \phi_{k+1}(\mathbf{w}_{k+1}) &= \sigma_k \mu (\mathbf{w}_{k+1} - \mathbf{w}_k) + \theta_k \nabla f(\mathbf{v}_k) + \theta_k \mu (\mathbf{w}_{k+1} - \mathbf{v}_k) \\ &= \sigma_{k+1} \mu \mathbf{w}_{k+1} - \sigma_k \mu \mathbf{w}_k - \theta_k \mu \mathbf{v}_k + \theta_k \nabla f(\mathbf{v}_k), \end{aligned}$$

from which the closed-form expression of $\mathbf{w}_{k+1}$ in **??** can be derived. The proof is completed. ∎

LEMMA 4.3. *Let $\{\phi_k\}$ be the sequence of functions defined in* **??**. *Then we have*

$$\phi_k(\mathbf{u}) \leq \sigma_k (f(\mathbf{u}) - f^*) + \phi_0(\mathbf{u}), \tag{4.9}$$

for all $\mathbf{u} \in \Omega$ and $k \in \mathbb{N}$.

Proof. We prove that $\{\phi_k\}$ and $\{\sigma_k\}$ satisfy relationship **??** by induction. It is obvious that **??** holds for $k = 0$ since $f(\mathbf{u}) \geq f^*$ for any $\mathbf{u} \in \Omega$. Now we assume that **??** holds for some $k \in \mathbb{N}$. It follows from the strong convexity of f that

$$f(\mathbf{u}) \geq f(\mathbf{v}_k) + \left\langle \nabla f(\mathbf{v}_k), \mathbf{u} - \mathbf{v}_k \right\rangle + \frac{\mu}{2} \left\| \mathbf{u} - \mathbf{v}_k \right\|^2 ,$$

for all $\mathbf{u} \in \Omega$. Then substituting the above relationship into **??** leads to that

$$\begin{aligned} \phi_{k+1}(\mathbf{u}) &\leq \phi_k(\mathbf{u}) - \theta_k f^* + \theta_k f(\mathbf{u}) \\ &\leq \sigma_k (f(\mathbf{u}) - f^*) + \phi_0(\mathbf{u}) + \theta_k (f(\mathbf{u}) - f^*) \\ &= \sigma_{k+1} (f(\mathbf{u}) - f^*) + \phi_0(\mathbf{u}), \end{aligned}$$

which indicates that **??** also holds for $k + 1$. We complete the proof. ∎

LEMMA 4.4. *Let $\{\mathbf{u}_k\}$ be the sequence generated by* **??**. *Then it holds that*

$$\sigma_k \left(f(\mathbf{u}_k) - f^* - \frac{\mu \varepsilon^2}{4} \right) \leq \phi_k^* := \min_{\mathbf{u} \in \Omega} \phi_k(\mathbf{u}), \tag{4.10}$$

for all $k \in \mathbb{N}$.

Proof. We aim to prove the assertion of this lemma by induction. It is clear that **??** holds for $k = 0$ since $\sigma_0 = 1$ and $\phi_0^* = \phi_0(\mathbf{w}_0) = f(\mathbf{u}_0) - f^* - \mu \varepsilon^2 / 4$. Now we assume that **??** holds for some $k \in \mathbb{N}$ and investigate the situation for $k + 1$.

From the canonical form **??**, it follows that ϕ_k is a strongly convex function and $\Pi_\Omega(\mathbf{w}_k) = \arg\min_{\mathbf{u} \in \Omega} \phi_k(\mathbf{u})$. By invoking the result of [**?**, Corollary 2.2.1], we have

$$\begin{aligned} \phi_k(\mathbf{u}) &\geq \phi_k^* + \frac{\sigma_k \mu}{2} \left\| \mathbf{u} - \Pi_\Omega(\mathbf{w}_k) \right\|^2 \\ &\geq \sigma_k \left(f(\mathbf{u}_k) - f^* - \frac{\mu \varepsilon^2}{4} \right) + \frac{\sigma_k \mu}{2} \left\| \mathbf{u} - \Pi_\Omega(\mathbf{w}_k) \right\|^2 , \end{aligned}$$

for all $\mathbf{u} \in \Omega$. Then relationship **??** yields that

$$\begin{aligned}
\phi_{k+1}(\mathbf{u}) \geq \sigma_k \left(f(\mathbf{u}_k) - f^* - \frac{\mu\varepsilon^2}{4} \right) + \frac{\sigma_k \mu}{2} \|\mathbf{u} - \Pi_\Omega(\mathbf{w}_k)\|^2 - \theta_k f^* \\
+ \theta_k f(\mathbf{v}_k) + \theta_k \langle \nabla f(\mathbf{v}_k), \mathbf{u} - \mathbf{v}_k \rangle + \frac{\theta_k \mu}{2} \|\mathbf{u} - \mathbf{v}_k\|^2 \\
\geq \sigma_{k+1} \left(f(\mathbf{v}_k) - f^* \right) - \frac{\sigma_k \mu \varepsilon^2}{4} + \langle \nabla f(\mathbf{v}_k), \sigma_k \mathbf{u}_k - \sigma_{k+1} \mathbf{v}_k \rangle \\
+ \theta_k \langle \nabla f(\mathbf{v}_k), \mathbf{u} \rangle + \frac{\sigma_k \mu}{2} \|\mathbf{u} - \Pi_\Omega(\mathbf{w}_k)\|^2 \\
= \sigma_{k+1} \left(f(\mathbf{v}_k) - f^* \right) - \frac{\sigma_k \mu \varepsilon^2}{4} + \theta_k \langle \nabla f(\mathbf{v}_k), \mathbf{u} - \Pi_\Omega(\mathbf{w}_k) \rangle \\
+ \frac{\sigma_k \mu}{2} \|\mathbf{u} - \Pi_\Omega(\mathbf{w}_k)\|^2,
\end{aligned}$$

where the second inequality comes from the strong convexity of f and $\sigma_{k+1} = \sigma_k + \theta_k$, and the last equality holds due to the definition of $\mathbf{v}_k$ in **??**. According to the definition of $\mathbf{z}_k$ in **??**, we can obtain that

$$\begin{aligned}
&\theta_k \langle \nabla f(\mathbf{v}_k), \mathbf{u} - \Pi_\Omega(\mathbf{w}_k) \rangle + \frac{\sigma_k \mu}{2} \|\mathbf{u} - \Pi_\Omega(\mathbf{w}_k)\|^2 \\
= &\frac{\sigma_k \mu}{2} \left\| \mathbf{u} - \left(\Pi_\Omega(\mathbf{w}_k) - \frac{\theta_k}{\sigma_k \mu} \nabla f(\mathbf{v}_k) \right) \right\|^2 - \frac{\theta_k^2}{2\sigma_k \mu} \|\nabla f(\mathbf{v}_k)\|^2 \\
\geq &\frac{\sigma_k \mu}{2} \left\| \mathbf{z}_k - \left(\Pi_\Omega(\mathbf{w}_k) - \frac{\theta_k}{\sigma_k \mu} \nabla f(\mathbf{v}_k) \right) \right\|^2 - \frac{\theta_k^2}{2\sigma_k \mu} \|\nabla f(\mathbf{v}_k)\|^2 \\
= &\theta_k \langle \nabla f(\mathbf{v}_k), \mathbf{z}_k - \Pi_\Omega(\mathbf{w}_k) \rangle + \frac{\sigma_k \mu}{2} \|\mathbf{z}_k - \Pi_\Omega(\mathbf{w}_k)\|^2
\end{aligned}$$

As a result, it holds that

$$\begin{aligned}
\phi_{k+1}(\mathbf{u}) \geq \sigma_{k+1} \left(f(\mathbf{v}_k) - f^* \right) - \frac{\sigma_k \mu \varepsilon^2}{4} + \theta_k \langle \nabla f(\mathbf{v}_k), \mathbf{z}_k - \Pi_\Omega(\mathbf{w}_k) \rangle \\
+ \frac{\sigma_k \mu}{2} \|\mathbf{z}_k - \Pi_\Omega(\mathbf{w}_k)\|^2,
\end{aligned} \tag{4.11}$$

for all $\mathbf{u} \in \Omega$. From the definitions of $\mathbf{v}_k$ and $\mathbf{u}_{k+1}$ in **??** and **??**, it can be derived that $\theta_k(\mathbf{z}_k - \Pi_\Omega(\mathbf{w}_k)) = \sigma_{k+1}(\mathbf{u}_{k+1} - \mathbf{v}_k)$. Substituting this relationship into (**??**) and taking $\mathbf{u} = \Pi_\Omega(\mathbf{w}_{k+1})$, we arrive at

$$\frac{\phi_{k+1}^*}{\sigma_{k+1}} \geq f(\mathbf{v}_k) - f^* + \langle \nabla f(\mathbf{v}_k), \mathbf{u}_{k+1} - \mathbf{v}_k \rangle + \frac{\sigma_k \sigma_{k+1} \mu}{2\theta_k^2} \|\mathbf{u}_{k+1} - \mathbf{v}_k\|^2 - \frac{\sigma_k \mu \varepsilon^2}{4\sigma_{k+1}}.$$

which together with the line-search condition **??** and $\sigma_{k+1} \geq \sigma_k$ implies that

$$\frac{\phi_{k+1}^*}{\sigma_{k+1}} \geq f(\mathbf{u}_{k+1}) - f^* - \frac{\theta_k \mu \varepsilon^2}{4\sigma_{k+1}} - \frac{\sigma_k \mu \varepsilon^2}{4\sigma_{k+1}} = f(\mathbf{u}_{k+1}) - f^* - \frac{\mu \varepsilon^2}{4}.$$

Therefore, relationship (**??**) also holds for $k+1$. The proof is completed. ∎

COROLLARY 4.5. *Let $\{\phi_k\}$ be the sequence of functions defined in* **??** *and $\{\mathbf{u}_k\}$ be the sequence generated by* **??**. *Then we have*

$$f(\mathbf{u}_k) - f^* \leq \frac{1}{\sigma_k} \phi_0(\mathbf{u}^*) + \frac{\mu \varepsilon^2}{4}, \tag{4.12}$$

for any $k \in \mathbb{N}$.

Proof. Collecting two relationships **??** and **??** together, we can obtain that

$$\begin{aligned}\sigma_k\left(f(\mathbf{u}_k) - f^* - \frac{\mu\varepsilon^2}{4}\right) &\leq \min_{\mathbf{u}\in\Omega} \phi_k(\mathbf{u}) \leq \min_{\mathbf{u}\in\Omega} \{\sigma_k(f(\mathbf{u}) - f^*) + \phi_0(\mathbf{u})\} \\ &\leq \sigma_k(f(\mathbf{u}^*) - f^*) + \phi_0(\mathbf{u}^*) \\ &= \phi_0(\mathbf{u}^*),\end{aligned}$$

which completes the proof. ∎

We proceed to establish the iteration complexity of **??**, as articulated in the theorem below.

THEOREM 4.6. *Let $\varepsilon \in (0, 1)$ be a sufficiently small constant. Then after at most*

$$O\left(\log\left(\frac{1}{\varepsilon}\right)\frac{1}{\varepsilon^{2(1-\hat{\alpha})/(1+3\hat{\alpha})}}\right)$$

iterations, **??** *will reach an iterate $\mathbf{u}_k$ satisfying $\|\mathbf{u}_k - \mathbf{u}^*\| \leq \varepsilon$.*

Proof. In view of relationship **??**, the number of line-search steps j_k in **??** satisfies

$$\begin{aligned}\frac{\sigma_k^2\mu}{\theta_k^2}\left[\frac{\theta_k}{\sigma_{k+1}}\right]^{(1-\hat{\alpha})/(1+\hat{\alpha})} &\leq 2\max_{i\in[m]}\left\{\left[\frac{2(1-\alpha_i)}{\mu(1+\alpha_i)\varepsilon^2}\right]^{(1-\alpha_i)/(1+\alpha_i)} L_i^{2/(1+\alpha_i)}\right\} \\ &\leq \frac{2M}{\varepsilon^{2(1-\hat{\alpha})/(1+\hat{\alpha})}},\end{aligned}$$

where $M > 0$ is a constant defined in **??**. Since $\sigma_{k+1} = \sigma_k + \theta_k = (1 + \nu_k)\sigma_k \leq 2\sigma_k$, we arrive at

$$\frac{\theta_k^2}{\sigma_k^2\mu} \geq \frac{\varepsilon^{2(1-\hat{\alpha})/(1+\hat{\alpha})}}{2M}\left[\frac{\theta_k}{\sigma_{k+1}}\right]^{(1-\hat{\alpha})/(1+\hat{\alpha})} \geq \frac{\varepsilon^{2(1-\hat{\alpha})/(1+\hat{\alpha})}}{2^{2/(1+\hat{\alpha})}M}\left[\frac{\theta_k}{\sigma_k}\right]^{(1-\hat{\alpha})/(1+\hat{\alpha})}. \tag{4.13}$$

Let $\omega > 0$ be a constant defined as

$$\omega = \frac{1}{2^{2/(1+3\hat{\alpha})}}\left[\frac{\mu}{M}\right]^{(1+\hat{\alpha})/(1+3\hat{\alpha})}.$$

Then it follows from relationship **??** that

$$\theta_k \geq \sigma_k\omega\varepsilon^{2(1-\hat{\alpha})/(1+3\hat{\alpha})}, \tag{4.14}$$

which further infers that

$$\sigma_{k+1} = \sigma_k + \theta_k \geq \sigma_k + \sigma_k\omega\varepsilon^{2(1-\hat{\alpha})/(1+3\hat{\alpha})} = \left(1 + \omega\varepsilon^{2(1-\hat{\alpha})/(1+3\hat{\alpha})}\right)\sigma_k.$$

Applying the above inequality for k times recursively yields that

$$\sigma_k \geq \left(1 + \omega\varepsilon^{2(1-\hat{\alpha})/(1+3\hat{\alpha})}\right)^k.$$

As a direct consequence of **??** and **??**, we can show that

$$\begin{aligned}\|\mathbf{u}_k - \mathbf{u}^*\|^2 &\leq \frac{2}{\mu}\left(f(\mathbf{u}_k) - f^*\right) \leq \frac{2}{\mu}\left(\frac{1}{\sigma_k}\phi_0(\mathbf{u}^*) + \frac{\mu\varepsilon^2}{4}\right) \\ &\leq \chi\left(1 + \omega\varepsilon^{2(1-\hat{\alpha})/(1+3\hat{\alpha})}\right)^{-k} + \frac{\varepsilon^2}{2},\end{aligned}$$

where $\chi = 2(f(\mathbf{u}_0) - f^*)/\mu + \|\mathbf{u}_0 - \mathbf{u}^*\|^2 > 0$ is a constant. Let K_ε^* be the smallest iteration number k such that $\|\mathbf{u}_k - \mathbf{u}^*\| \le \varepsilon$. By solving the inequality $\chi(1 + \omega\varepsilon^{2(1-\hat{\alpha})/(1+3\hat{\alpha})})^{-k} \le \varepsilon^2/2$, we have

$$K_\varepsilon^* \le \log\left(\frac{\sqrt{2\chi}}{\varepsilon}\right)\frac{2}{\log\left(1+\omega\varepsilon^{2(1-\hat{\alpha})/(1+3\hat{\alpha})}\right)} \le \log\left(\frac{\sqrt{2\chi}}{\varepsilon}\right)\frac{4}{\omega\varepsilon^{2(1-\hat{\alpha})/(1+3\hat{\alpha})}}.$$

The proof is completed. ∎

Building upon **??**, we further demonstrate that the number of line-search steps required by **??** is also $O(\log(\varepsilon^{-1})\varepsilon^{-2(1-\hat{\alpha})/(1+3\hat{\alpha})})$.

COROLLARY 4.7. *Let $\varepsilon \in (0,1)$ be a sufficiently small constant. Then, to achieve an iterate $\mathbf{u}_k$ satisfying $\|\mathbf{u}_k - \mathbf{u}^*\| \le \varepsilon$,* **??** *requires at most*

$$O\left(\log\left(\frac{1}{\varepsilon}\right)\frac{1}{\varepsilon^{2(1-\hat{\alpha})/(1+3\hat{\alpha})}}\right)$$

line-search steps.

Proof. It follows from relationship **??** that

$$\rho_{k+1} = 2^{j_k}\rho_k = \frac{\sigma_k^2\mu}{\theta_k^2} \le \frac{2^{2/(1+\hat{\alpha})}M}{\varepsilon^{2(1-\hat{\alpha})/(1+\hat{\alpha})}}\left[\frac{\sigma_k}{\theta_k}\right]^{(1-\hat{\alpha})/(1+\hat{\alpha})},$$

which together with **??** implies that

$$\rho_{k+1} \le \frac{2^{2/(1+\hat{\alpha})}M}{\varepsilon^{2(1-\hat{\alpha})/(1+\hat{\alpha})}}\left[\frac{1}{\omega\varepsilon^{2(1-\hat{\alpha})/(1+3\hat{\alpha})}}\right]^{(1-\hat{\alpha})/(1+\hat{\alpha})} = \frac{2^{2/(1+\hat{\alpha})}M}{\omega^{(1-\hat{\alpha})/(1+\hat{\alpha})}\varepsilon^{4(1-\hat{\alpha})/(1+3\hat{\alpha})}}.$$

Let N_k be the total number of line-search steps after k iterations in **??**. In view of **??**, we have

$$\begin{aligned} N_k &\le k + 1 + \log\left(\frac{2^{2/(1+\hat{\alpha})}M}{\omega^{(1-\hat{\alpha})/(1+\hat{\alpha})}\varepsilon^{4(1-\hat{\alpha})/(1+3\hat{\alpha})}}\right) - \log\rho_0 \\ &\le k + \frac{4(1-\hat{\alpha})}{1+3\hat{\alpha}}\log\left(\frac{1}{\varepsilon}\right) + \log\left(\frac{2^{2/(1+\hat{\alpha})}M}{\omega^{(1-\hat{\alpha})/(1+\hat{\alpha})}\rho_0}\right) + 1. \end{aligned}$$

Consequently, **??** indicates that the total number of line-search steps in **??** is at most $O(\log(\varepsilon^{-1})\varepsilon^{-2(1-\hat{\alpha})/(1+3\hat{\alpha})})$, which completes the proof. ∎

5. Numerical Experiments.

Preliminary numerical results are presented in this section to provide additional insights into the performance guarantees of the gradient descent method **??**. We aim to elucidate that the final error attained by the gradient descent method **??** is influenced by both the stepsize τ and the Hölder exponent α. All codes are implemented in MATLAB R2018b on a workstation with dual Intel Xeon Gold 6242R CPU processors (at 3.10 GHz$\times 20 \times 2$) and 510 GB of RAM under Ubuntu 20.04.

5.1. Two-dimensional PDE with a non-Lipschitz term.

Hölder continuous gradients arise naturally in partial differential equations (PDEs) involving non-Lipschitz nonlinearity [?, ?]. In this subsection, we introduce a numerical example from [?]. This problem is to solve the following two-dimensional PDE,

$$\mathcal{F}(u) = -\Delta u + \nu u_+^{1/2} = 0, \tag{5.1}$$

where $\nu > 0$ is a constant and $u_+ = \max\{u, 0\}$. It should be noted that $\mathcal{F}$ is the gradient of the following energy functional,

$$\hat{f}(u) = \frac{1}{2}\|\nabla u\|^2 + \frac{2\nu}{3}\int_D u_+^{3/2}(y)\,\mathrm{d}y.$$

Discretizing **??** with the standard five point difference scheme [**?**] leads to the following nonlinear system,

$$\mathbf{F}(\mathbf{u}) = \mathbf{A}\mathbf{u} + \nu\mathbf{u}_+^{1/2} - \mathbf{b} = 0, \tag{5.2}$$

where $\mathbf{A} \in \mathbb{R}^{n\times n}$ is the discretization of $-\Delta$ with zero boundary conditions, $\mathbf{b} \in \mathbb{R}^n$ encodes the boundary conditions, and $\mathbf{u}_+^{1/2} = \max\{\mathbf{u}, 0\}^{1/2}$ is understood as a component-wise operation. Problem (**??**) is equivalent to optimization problem (**??**) with $\Omega = \mathbb{R}^n$, and

$$f(\mathbf{u}) = \frac{1}{2}(f_1(\mathbf{u}) + f_2(\mathbf{u})) \quad \text{with} \quad f_1(\mathbf{u}) = \mathbf{u}^\top\mathbf{A}\mathbf{u} - 2\mathbf{b}^\top\mathbf{u}, \quad f_2(\mathbf{u}) = \frac{4\nu}{3}\mathbf{e}^\top\mathbf{u}_+^{3/2},$$

where $\mathbf{e} \in \mathbb{R}^n$ is the vector of all ones.

It is clear that ∇f_1 is Lipschitz continuous with the Lipschitz constant $L_1 = \|\mathbf{A}\|$, and ∇f_2 is locally Hölder continuous with $\alpha = 1/2$ and $L_2 = \nu n^{1/4}$ from

$$\|\nabla f_2(\mathbf{u}) - \nabla f_2(\mathbf{v})\| = \nu\left\|\mathbf{u}_+^{1/2} - \mathbf{v}_+^{1/2}\right\| \le \nu n^{1/4}\|\mathbf{u} - \mathbf{v}\|^{1/2},$$

for all $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$. The function f is $\lambda(\mathbf{A})$-strongly convex, where $\lambda(\mathbf{A})$ is the smallest eigenvalue of the symmetric positive definite matrix $\mathbf{A}$.

Other example from elliptic equations with a non-Lipschitz term is given in Section 5.

To evaluate the performance of the gradient descent method **??**, we focus on the following optimization problem inspired by the PDE model introduced in **??**,

$$\min_{\mathbf{u}\in\mathbb{R}^n} f(\mathbf{u}) = \frac{1}{2}\mathbf{u}^\top\mathbf{A}\mathbf{u} + \frac{1}{1+\alpha}\mathbf{e}^\top\mathbf{u}_+^{1+\alpha} - \mathbf{c}^\top\mathbf{u}, \tag{5.3}$$

where $\mathbf{A} \in \mathbb{R}^{n\times n}$ is a symmetric positive definite matrix, $\alpha \in (0,1)$ is a constant, and $\mathbf{c} = \mathbf{A}\mathbf{u}^* + (\mathbf{u}^*)_+^\alpha \in \mathbb{R}^n$ is a vector with $\mathbf{u}^* \in \mathbb{R}^n$. It is evident that the objective function f is strongly convex with $\mu = \lambda_{\min}(\mathbf{A})$ and its gradient ∇f is locally α-Hölder continuous with $\beta = 1 + \lambda_{\max}(\mathbf{A})$ and $\gamma = 1$. Moreover, a straightforward verification reveals that $\mathbf{u}^*$ is the minimizer of problem **??**.

In our numerical experiments, the initial point $\mathbf{u}_0$, the minimizer $\mathbf{u}^*$ and the matrix $\mathbf{A}$ in the test problem **??** are generated randomly, and the vector $\mathbf{c}$ is defined by $\mathbf{u}^*, \mathbf{A}$ and α with the detailed MATLAB code provided as follows.

```
u_0 = randn(n, 1);
u_star = randn(n, 1);
A = randn(n); A = A'*A + eye(n);
c = A*u_star + max(u_star, 0).^alpha;
```

Moreover, the test problem dimension is fixed at $n = 50$, and method **??** is permitted a maximum of 10000 iterations.

In the first experiment, we scrutinize the performance of the gradient descent method **??** under different stepsizes. Specifically, with the parameter α fixed at 0.5, Figures/test_stepsize.pdf

(a) different stepsizes

Figures/test_alpha.pdf

(b) different values of α

FIG. 1. *Numerical performance of gradient descent method* **??** *for problem* **??**.

the algorithm is tested for stepsizes chosen from the set $\{0.01, 0.005, 0.001, 0.0005\}$. The corresponding numerical results, presented in **??**, illustrate the decay of the distance between the iterates and the global minimizer over iterations. It can be observed that a larger stepsize facilitates a more rapid descent in the early stage of iterations, albeit at the expense of a greater asymptotic error. This phenomenon corroborates our theoretical predictions.

In the second experiment, the stepsize τ is fixed at 0.001, while the parameter α is varied over the values $\{0.2, 0.4, 0.6, 0.8\}$. **??** similarly tracks the decay of the distance to the global minimizer over iterations. It is evident that, as the value of α decreases, the final error attained by the algorithm increases under the same stepsize. Therefore, the associated optimization problems become increasingly ill-conditioned and thus more challenging to solve for smaller values of α. These findings offer empirical support for our theoretical analysis.

Example 2 We consider a numerical example motivated by a semi-linear elliptic problem with a constraint on the solution in a certain set [**?**]. Let $D = (0,1)^3$ and

$$\mathcal{H}(u) = -\Delta u + \lambda |u|^{\nu} - |u|^{p-1} u \tag{5.4}$$

on D with the boundary condition $u = 1$ on the boundary ∂D, where $p > 1$, $\nu \in (0,1)$ and $\lambda > p/\nu$ are constants. We consider the variational inequality that is to find $u^* \in [-1,1]$ such that for any $u \in [-1,1]$,

$$\mathcal{H}(u^*)(u - u^*) \geq 0.$$

This problem is equivalent to the nonlinear equation

$$0 = \mathcal{F}(u) := \begin{cases} \mathcal{H}(u) & \text{if} \quad u - \mathcal{H}(u) \in [-1,1], \\ u - 1 & \text{if} \quad u - \mathcal{H}(u) \geq 1, \\ u + 1 & \text{otherwise}. \end{cases} \tag{5.5}$$

Discretizing **??** with the standard five point difference scheme [**?**], problem **??** leads to the following system of nonlinear equations

$$\mathbf{F}(\mathbf{u}) = \mathbf{u} - \Pi_{\mathbf{U}}\Big(\mathbf{u} - \tau(\mathbf{A}\mathbf{u} + \lambda|\mathbf{u}|^{\nu} - |\mathbf{u}|^{p-1}\mathbf{u} - \mathbf{b})\Big) = 0, \tag{5.6}$$

where $\mathbf{U} = [-1,1]^n$, $\tau > 0$ is a constant, $\mathbf{A} \in \mathbb{R}^{n\times n}$ is a symmetric positive definite matrix and $\mathbf{b} \in \mathbb{R}^n$. Note that **??** is the first-order optimal condition of the minimization problem

$$\min_{\mathbf{u}\in[-1,1]^n} f(\mathbf{u}) := \frac{1}{2}\mathbf{u}^\top \mathbf{A}\mathbf{u} + \frac{\lambda}{1+\nu}\mathbf{e}^\top |\mathbf{u}|^{\nu+1} - \frac{1}{1+p}\mathbf{e}^\top \max(\mathbf{u},-\mathbf{u})^{p+1} + \mathbf{b}^\top \mathbf{u}. \tag{5.7}$$

The Hessian matrix of f at $\mathbf{u}$ with $\mathbf{u}_i \neq 0$, $i = 1,\dots,n$ has the form

$$\nabla^2 f(\mathbf{u}) = \mathbf{A} + \lambda\nu|\mathbf{u}|^{\nu-1} - p\mathrm{diag}\Big(\max(-\mathbf{u},\mathbf{u})^{p-1}\Big),$$

Since $\lambda\nu > p$, $\nabla^2 f(\mathbf{u})$ is symmetric positive definite for any $\mathbf{u} \in [-1,1]^n$ with $\mathbf{u}_i \neq 0$, $i = 1,\dots,n$. Hence f is μ-strongly convex in $[-1,1]^n$ with $\mu = \lambda_{\min}(\mathbf{A})$ and the system **??** has a unique solution in $[-1,1]^n$. However, ∇f is not Lipschitz continuous in $[-1,1]^n$.

Let

$$f_1(\mathbf{u}) = \frac{1}{2}\mathbf{u}^\top \mathbf{A}\mathbf{u} + \mathbf{b}^\top \mathbf{u}, f_2(\mathbf{u}) = \frac{\lambda}{1+\nu}\mathbf{e}^\top|\mathbf{u}|^{\nu+1}, f_3(\mathbf{u}) = -\frac{1}{1+p}\mathbf{e}^\top \max(\mathbf{u},-\mathbf{u})^{p+1}$$

This example satisfies Assumption 1.1 (ii) with $L_1 = \lambda_{\max}(\mathbf{A})$, $L_2 = \lambda\nu$, $L_3 = pn^{\frac{1}{2}}, \alpha_1 = \alpha_3 = 1, \alpha_2 = 1-\nu$.

6. Conclusion.

In this paper, we have established a new complexity result for the projected gradient descent method with a fixed stepsize applied to strongly convex optimization problems where the objective function has a α-Hölder continuous gradient term with $\alpha \in (0,1]$. The main conclusion is that, to achieve an approximate minimizer with a distance to the minimizer less than ε, the total number of iterations required by the gradient descent method **??** is $O(\log(\varepsilon^{-1})\varepsilon^{2\alpha-2})$ at the most. This recovers the classical complexity result when $\alpha = 1$ and reveals the additional difficulty imposed by the weaker smoothness of the objective function for $\alpha < 1$. Numerical experiments are conducted to validate our theoretical findings, demonstrating the expected behavior of gradient descent under different stepsizes and Hölder exponents. These results offer new insights into the performance guarantees of the classic gradient descent method for a broader class of optimization problems with non-Lipschitz gradients.