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

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

**Abstract.** This paper studies the complexity of projected gradient descent methods for a class of strongly convex constrained optimization problems where the objective function is expressed as a summation of $m$ component functions, each possessing a gradient that is Hölder continuous with an exponent $\alpha_i \in (0,1]$. Under this formulation, the gradient of the objective function may fail to be globally Hölder continuous, thereby existing complexity results inapplicable to this class of problems. Our theoretical analysis reveals that, in this setting, the complexity of projected gradient methods is determined by $\hat{\alpha} = \min_{i\in\{1,\dots,m\}} \alpha_i$. We first prove that, with an appropriately fixed stepsize, the complexity bound for finding an approximate minimizer with a distance to the true minimizer less than $\varepsilon$ is $O(\log(\varepsilon^{-1})\varepsilon^{2(\hat{\alpha}-1)/(1+\hat{\alpha})})$, which extends the well-known complexity result for $\hat{\alpha} = 1$. Next we show that the complexity bound can be improved to $O(\log(\varepsilon^{-1})\varepsilon^{2(\hat{\alpha}-1)/(1+3\hat{\alpha})})$ 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,

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

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

Assumption 1.1.

*(i) The function $f$ is $\mu$-strongly convex with a parameter $\mu > 0$ on $\Omega$, that is,*

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

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

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

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

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

---

*Submitted to the editors 15 January, 2026.

**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).

Here, $\|\cdot\|$ is the $\ell_2$ 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 problem (1.1), respectively.

Suppose that each $\nabla f_i$ is Lipschitz continuous, which corresponds to condition (1.2) with $\alpha_i = 1$ for all $\mathbf{u}, \mathbf{v} \in \Omega$. Then $\nabla f$ is also Lipschitz continuous and the associated Lipschitz constant is $L = \sum_{i=1}^m L_i/m$. Let $\Pi_\Omega(\cdot)$ be the projection operator onto the set $\Omega$. 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 [11, Theorem 2.2.14] as follows,

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

Therefore, for a given $\varepsilon > 0$, method (1.3) is guaranteed to find a point $\mathbf{u}_k \in \Omega$ satisfying $\|\mathbf{u}_k - \mathbf{u}^*\| \leq \varepsilon$ 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 (1.3) to $\mathbf{u}^*$ by the following example.

*Example* 1.2. [6, Example 1] Consider the following univariate optimization problem on $\Omega = \mathbb{R}$,

$$\min_{x \in \mathbb{R}} \ f(x) = \frac{1}{2}x^2 + \frac{2}{3}|x|^{3/2}, \tag{1.4}$$

which is a special instance of problem (1.1) with $f_1(x) = x^2/2$ and $f_2(x) = 2|x|^{3/2}/3$. It is easy to see that the global minimizer is $x^* = 0$. Method (1.3) with the fixed stepsize $\tau > 0$ starting from $x_0 \neq 0$ proceeds as follows,

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

where $\mathsf{sign}(x) = 1$ if $x > 0$, $0$ if $x = 0$, and $-1$ otherwise. 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 $|x_{k+1}|^2 - |x_k|^2 \geq 0$. Therefore, the distance to the global minimizer ceases to decrease once it achieves a certain level.

Moreover, in [6] we show that $\nabla f$ is locally, but not globally, Hölder continuous. In fact, from

$$|\nabla f(|h|) - \nabla f(0)| = |h| + |h|^{1/2} = \left(|h|^{1-\alpha} + |h|^{1/2-\alpha}\right)|h|^\alpha,$$

we can obtain that, $|h|^{1-\alpha} \to \infty$ when $\alpha \in (0, 1)$ and $|h| \to \infty$, while $|h|^{1/2-\alpha} \to \infty$ when $\alpha = 1$ and $|h| \to 0$. Therefore, $\nabla f$ cannot be globally Hölder continuous for all $\alpha \in (0, 1]$.

On the other hand, problem (1.4) satisfies all the conditions in Assumption 1.1. It is clear that $f$ is strongly convex. In addition, we have

$$|\nabla f_1(x) - \nabla f_1(y)| = |x - y|,$$

and

$$|\nabla f_2(x) - \nabla f_2(y)| = \left|\mathsf{sign}(x)\,|x|^{1/2} - \mathsf{sign}(y)\,|y|^{1/2}\right| \leq \sqrt{2}\,|x-y|^{1/2},$$

for all $x, y \in \mathbb{R}$.

This simple example demonstrates that, in problem (1.1), 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, initially observed in our previous work [6], was later revisited and further highlighted by Nesterov (see [12, Example 1]).

Since $\nabla f$ may not be globally Hölder continuous, most existing complexity results are inapplicable to problem (1.1). For the special case where $m = 1$, namely, $\nabla f$ is globally Hölder continuous with an exponent $\alpha \in (0, 1]$, Devolder et al. [7] presented the following bound for method (1.3),

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

where $L_\alpha$ is the Hölder constant and $\hat{\mathbf{u}}_N = \sum_{k=1}^N \mathbf{u}_k/N$. In the strongly convex case, (51) in [7] comes to

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

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

The contribution of this paper is to provide new complexity results of the projected gradient descent methods for problem (1.1), which are dictated by the parameter $\hat{\alpha} = \min_{i\in[m]} \alpha_i \in (0, 1]$. We first show that, with an appropriately fixed stepsize, the complexity bound for finding an iterate with a distance to the global minimizer less than $\varepsilon$ is $O(\log(\varepsilon^{-1})\varepsilon^{2(\hat{\alpha}-1)/(1+\hat{\alpha})})$, which extends the well-known complexity result for $\hat{\alpha} = 1$. Next, we demonstrate that this complexity bound can be improved to $O(\log(\varepsilon^{-1})\varepsilon^{2(\hat{\alpha}-1)/(1+3\hat{\alpha})})$ if the stepsize is updated at each iteration using the universal scheme. Even in the special case where $m = 1$, our complexity bound is at least $O(\varepsilon^{-1})$ lower than (51) in [7]. For example, when $\hat{\alpha} = 1/2$, our bound is $O(\log(\varepsilon^{-1})\varepsilon^{-2/5})$ but (51) in [7] is $O(\varepsilon^{-8/3})$.

Our study is motivated by elliptic equations with a non-Lipschitz term [3, 14], complementarity problems [1, 13], and optimization problems with an $\ell_p$-norm ($1 < p < 2$) regularization term [2, 5]. 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 stepsizes and updated stepsizes in sections 2 to 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 (1.3) with a fixed stepsize to solve problem (1.1), whose complexity bound is also provided. Example 1.2 illustrates that the projected gradient descent method (1.3) with a fixed stepsize will experience stagnation before reaching the global minimizer.

To obtain an approximate solution to problem (1.1), it is necessary to choose a sufficiently small stepsize $\tau$ in the projected gradient descent method (1.3), the magnitude of which depends on the desired level of accuracy. Let $M > 0$ be a constant defined as

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

with the convention $0^0 = 1$. 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 Algorithm 1. Two sequences $\{\mathbf{v}_k\}$ and $\{\mathbf{u}_k\}$ are maintained in Algorithm 1, 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.

**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

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

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

Our subsequent analysis is based on the inexact oracle [7] derived from the Hölder continuity condition of gradients, which is generalized to problem (1.1) and demonstrated in the following proposition.

PROPOSITION 2.1. *Let $\delta > 0$ and*

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

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

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

*Proof.* Since $\nabla f_i$ is Hölder continuous with an exponent $\alpha_i$, we can obtain from [15, Lemma 1] that

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

for all $\mathbf{u}, \mathbf{v} \in \Omega$. Then, for each $i$, it follows from [10, Lemma 2] that

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

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

Now, we are able to derive the complexity bound of Algorithm 1 in the following theorem.

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

$$O\left(\log\left(\frac{M^{(1+\hat{\alpha})/4}}{\varepsilon}\right)\frac{M}{\varepsilon^{2(1-\hat{\alpha})/(1+\hat{\alpha})}}\right)$$

*iterations, Algorithm 1 will find an iterate $\mathbf{u}_k \in \Omega$ satisfying $\|\mathbf{u}_k - \mathbf{u}^*\| \le \varepsilon$.*

*Proof.* In view of Proposition 2.1, we take

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

Then it holds that

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

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

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

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

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

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

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

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

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

which together with (2.2) and (2.3) implies that

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

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

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

From the construction of $\mathbf{u}_k$ in Algorithm 1, it then follows that $f(\mathbf{v}_l) \ge f(\mathbf{u}_k)$ for all $l \in \{1,2,\dots,k\}$. 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

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

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

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

According to the strong convexity of $f$ and the optimality condition of problem (1.1), we have

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

Hence, it holds that

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

We denote by $K_\varepsilon^*$ 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. ∎

Theorem 2.2 demonstrates that the iteration complexity of Algorithm 1 with a fixed stepsize is $O(\log(\varepsilon^{-1})\varepsilon^{2(\hat\alpha-1)/(1+\hat\alpha)})$ for problem (1.1). This complexity result generalizes the classical linear convergence when $\hat\alpha = 1$, which highlights the performance degradation incurred by non-Lipschitz gradients.

## 3. Universal Primal Gradient Method.

The fixed stepsize $\tau$ chosen in Algorithm 1 depends on the parameters $\alpha_i$ 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 [10] to solve problem (1.1). This

**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,

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

    **then** break.
  Update $\rho_{k+1} = 2^{j_k}\rho_k$.
  Set

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

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

method incorporates a line-search procedure to adaptively determine the stepsize at each iteration, and its overall framework is outlined in Algorithm 2.

Next, we establish the iteration complexity of Algorithm 2, 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*

$$O\left( \log\left( \frac{M^{(1+\hat{\alpha})/4}}{\varepsilon} \right) \frac{M}{\varepsilon^{2(1-\hat{\alpha})/(1+\hat{\alpha})}} \right)$$

*iterations, Algorithm 2 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 Proposition 2.1, we know that condition (3.1) is satisfied. Hence, the line-search step in Algorithm 2 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 (2.1). Moreover, the line-search condition (3.1) directly yields that

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

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

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

which further implies that

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

Substituting (2.3) and (3.3) into the above relationship leads to that

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

Thus, it follows from relationship (2.4) 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 (3.2) and $2^{j_k}\rho_k \geq \rho_0$. The remaining part of the proof follows the same line of reasoning as that of Theorem 2.2 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 Algorithm 2.

COROLLARY 3.2. *Let $\varepsilon \in (0, 1)$ be a sufficiently small constant. Then Algorithm 2 requires at most*

$$O\left( \log\left( \frac{M^{(1+\hat{\alpha})/4}}{\varepsilon} \right) \frac{M}{\varepsilon^{2(1-\hat{\alpha})/(1+\hat{\alpha})}} \right)$$

*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 Algorithm 2. 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 (3.2) 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 Theorem 3.1, we conclude that Algorithm 2 requires at most $O(\log(\varepsilon^{-1})\varepsilon^{2(\hat{\alpha}-1)/(1+\hat{\alpha})})$ line-search steps, which completes the proof. ∎

At each iteration of Algorithm 2, 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, Theorem 3.1 and Corollary 3.2 together reveal that the total number of function and gradient evaluations required by Algorithm 2 is $O(\log(\varepsilon^{-1})\varepsilon^{2(\hat{\alpha}-1)/(1+\hat{\alpha})})$.

## 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 (1.1). The proposed scheme, summarized in Algorithm 3, exhibits slight but essential differences from the algorithm introduced by Nesterov [10] to exploit the strong convexity of the objective function.

The following lemma illustrates that the line-search process in (4.4) 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* (4.4) *is satisfied in Algorithm* 3.

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

$$
\eta_k = \frac{\nu_k}{1+\nu_k} \geq \frac{\nu_k}{2}, \quad \text{and} \quad \frac{\mu}{\nu_k^2} = 2^{j_k}\rho_k.
$$

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

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

where the first equality comes from the definition of $\nu_k$. Now it is clear that

$$
\frac{\mu}{\nu_k^2}\eta_k^{(1-\hat{\alpha})/(1+\hat{\alpha})} \to \infty,
$$

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

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

**Algorithm 3:** Universal Fast Gradient Method (UFGM).

**Input:** $\varepsilon > 0$.
Initialize $\mathbf{u}_0 = \mathbf{w}_0 \in \Omega$ and $\rho_0 \geq \mu$.
**for** $k = 0, 1, 2, \ldots$ **do**
  **for** $j_k = 0, 1, 2, \ldots$ **do**
    Set $\nu_k = \sqrt{\mu/(2^{j_k}\rho_k)}$ and $\eta_k = \nu_k/(1+\nu_k)$.
    Compute

$$\mathbf{v}_k = (1-\eta_k)\mathbf{u}_k + \eta_k \Pi_\Omega(\mathbf{w}_k), \tag{4.1}$$

    and

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

    Set

$$\mathbf{u}_{k+1} = (1-\eta_k)\mathbf{u}_k + \eta_k \mathbf{z}_k. \tag{4.3}$$

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

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

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

$$\mathbf{w}_{k+1} = (1-\eta_k)\mathbf{w}_k + \eta_k \mathbf{v}_k - \frac{\eta_k}{\mu}\nabla f(\mathbf{v}_k). \tag{4.5}$$

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

which further implies that

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

As a direct consequence of Proposition 2.1, we can proceed to show that the line-search condition (4.4) is satisfied, which completes the proof. ∎

*Remark* 4.2. When the parameters of problem (1.1) are fully specified, Algorithm 3 may alternatively be implemented with a fixed stepsize. Recall that $M > 0$ is a constant defined in (2.1). By invoking the result of Lemma 4.1, we can fix

$$\nu_k = 2\left[\frac{\mu}{4M}\right]^{(1+\hat\alpha)/(1+3\hat\alpha)} \varepsilon^{2(1-\hat\alpha)/(1+3\hat\alpha)},$$

and dispense with the parameter $\rho_k$ and the line-search procedure in (4.4). Under this choice, Algorithm 3 continues to enjoy the same iteration complexity established later.

We now introduce the estimating sequences associated with Algorithm 3, which play a crucial role in our subsequent analysis.

LEMMA 4.3. *Let $\{\sigma_k\}$ be a sequence of positive constants defined recursively by*

$$\sigma_{k+1} = (1+\nu_k)\sigma_k, \tag{4.7}$$

*with $\sigma_0 = 1$. And let $\{\phi_k\}$ be a sequence of functions defined recursively by*

$$\begin{aligned}\phi_{k+1}(\mathbf{u}) = \phi_k(\mathbf{u}) - \nu_k\sigma_k f^* + \nu_k\sigma_k f(\mathbf{v}_k) + \nu_k\sigma_k \langle \nabla f(\mathbf{v}_k), \mathbf{u}-\mathbf{v}_k\rangle \\ + \frac{\nu_k\sigma_k\mu}{2}\|\mathbf{u}-\mathbf{v}_k\|^2,\end{aligned} \tag{4.8}$$

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

$$\phi_k(\mathbf{u}) = c_k + \frac{\sigma_k\mu}{2}\|\mathbf{u}-\mathbf{w}_k\|^2, \tag{4.9}$$

*where $\{c_k\}$ is a sequence of real numbers and $\{\mathbf{w}_k\}$ is defined recursively by* (4.5)*.*

*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 (4.7) and (4.8) imply that

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

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 $\phi_k$ in (4.9).

Next, by combining two relationships (4.8) and (4.9) together, we can obtain that

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

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

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

from which the closed-form expression of $\mathbf{w}_{k+1}$ in (4.5) can be derived. The proof is completed. ∎

The following lemma characterizes the relationship between the objective function of problem (1.1) and the estimating sequences.

LEMMA 4.4. *Let $\sigma_k$ and $\{\phi_k\}$ be the sequences defined in Lemma* 4.3*. Then we have*

$$\phi_k(\mathbf{u}) \le \sigma_k(f(\mathbf{u}) - f^*) + \phi_0(\mathbf{u}), \tag{4.10}$$

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

*Proof.* We prove that $\{\phi_k\}$ and $\{\sigma_k\}$ satisfy relationship (4.10) by induction. It is obvious that (4.10) holds for $k=0$ since $f(\mathbf{u}) \geq f^*$ for any $\mathbf{u} \in \Omega$. Now we assume that (4.10) holds for some $k \in \mathbb{N}$. It follows from the strong convexity of $f$ that

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

for all $\mathbf{u} \in \Omega$. Then substituting the above relationship into (4.8) leads to that

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

which indicates that (4.10) also holds for $k+1$. We complete the proof. ∎

Next, we proceed to show that the function value error of Algorithm 3 is controlled by the estimating sequences.

Proposition 4.5. *Let $\{\sigma_k\}$ and $\{\phi_k\}$ be the sequences defined in Lemma 4.3. Then the sequence $\{\mathbf{u}_k\}$ generated by Algorithm 3 satisfies*

$$f(\mathbf{u}_k) - f^* \leq \frac{1}{\sigma_k}\phi_0(\mathbf{u}^*) + \frac{\mu\varepsilon^2}{4}, \tag{4.11}$$

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

*Proof.* Let $\phi_k^* := \min_{\mathbf{u}\in\Omega} \phi_k(\mathbf{u})$. We first prove by induction that

$$\sigma_k \left( f(\mathbf{u}_k) - f^* - \frac{\mu\varepsilon^2}{4} \right) \leq \phi_k^*, \tag{4.12}$$

for any $k \in \mathbb{N}$. It is clear that (4.12) 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 (4.12) holds for some $k \in \mathbb{N}$ and investigate the situation for $k+1$.

From the canonical form (4.9), it follows that $\phi_k$ 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 [11, Corollary 2.2.1], we have

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

for all $\mathbf{u} \in \Omega$. Then relationship (4.8) yields that

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

where the second inequality comes from the strong convexity of $f$ and (4.7), and the last equality holds due to the definition of $\mathbf{v}_k$ in (4.1). According to the definition of $\mathbf{z}_k$ in (4.2), we can obtain that

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

As a result, it holds that

$$\begin{aligned}
\phi_{k+1}(\mathbf{u}) \geq \sigma_{k+1}\left(f(\mathbf{v}_k) - f^*\right) - \frac{\sigma_k\mu\varepsilon^2}{4} + \nu_k\sigma_k\langle \nabla f(\mathbf{v}_k), \mathbf{z}_k - \Pi_\Omega(\mathbf{w}_k)\rangle \\
+ \frac{\sigma_k\mu}{2}\|\mathbf{z}_k - \Pi_\Omega(\mathbf{w}_k)\|^2,
\end{aligned} \tag{4.13}$$

for all $\mathbf{u} \in \Omega$. From the definitions of $\mathbf{v}_k$ and $\mathbf{u}_{k+1}$ in (4.1) and (4.3), it can be derived that $\mathbf{z}_k - \Pi_\Omega(\mathbf{w}_k) = (\mathbf{u}_{k+1} - \mathbf{v}_k)/\eta_k$. Substituting this relationship into (4.13) and taking $\mathbf{u} = \Pi_\Omega(\mathbf{w}_{k+1})$, we arrive at

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

which together with the line-search condition (4.4) implies that

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

Therefore, relationship (4.12) also holds for $k+1$.

Finally, by collecting two relationships (4.10) and (4.12) 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. ∎

With the above preparatory results in place, we are now in a position to establish the iteration complexity of Algorithm 3, as articulated in the theorem below.

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

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

*iterations, Algorithm 3 will reach an iterate $\mathbf{u}_k$ satisfying $\|\mathbf{u}_k - \mathbf{u}^*\| \leq \varepsilon$.*

*Proof.* In view of relationship (4.6), the number of line-search steps $j_k$ in (4.4) satisfies

$$\frac{\mu}{\nu_k^2}\eta_k^{(1-\hat{\alpha})/(1+\hat{\alpha})} \leq 2\max_{i\in[m]}\left\{\left[\frac{2(1-\alpha_i)}{\mu(1+\alpha_i)\varepsilon^2}\right]^{(1-\alpha_i)/(1+\alpha_i)} L_i^{2/(1+\alpha_i)}\right\} \leq \frac{2M}{\varepsilon^{2(1-\hat{\alpha})/(1+\hat{\alpha})}},$$

where $M > 0$ is a constant defined in (2.1). Since $\eta_k = \nu_k/(1+\nu_k) \geq \nu_k/2$, we arrive at

$$\frac{\nu_k^2}{\mu} \geq \frac{\varepsilon^{2(1-\hat{\alpha})/(1+\hat{\alpha})}}{2M}\eta_k^{(1-\hat{\alpha})/(1+\hat{\alpha})} \geq \frac{\varepsilon^{2(1-\hat{\alpha})/(1+\hat{\alpha})}}{2^{2/(1+\hat{\alpha})}M}\nu_k^{(1-\hat{\alpha})/(1+\hat{\alpha})}. \tag{4.14}$$

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 (4.14) that

$$\nu_k \geq \omega\varepsilon^{2(1-\hat{\alpha})/(1+3\hat{\alpha})}, \tag{4.15}$$

which further infers that

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

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 (2.5) and (4.11), 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 $K_\varepsilon^*$ be the smallest iteration number $k$ such that $\|\mathbf{u}_k-\mathbf{u}^*\| \leq \varepsilon$. By solving the inequality $\chi(1+\omega\varepsilon^{2(1-\hat{\alpha})/(1+3\hat{\alpha})})^{-k} \leq \varepsilon^2/2$, we have

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

The proof is completed. ∎

The complexity bound established in Theorem 4.6 is markedly lower than those presented in Theorems 2.2 and 3.1, thereby highlighting the acceleration effect attained by Algorithm 3. Finally, we demonstrate that the number of line-search steps required by Algorithm 3 is also $O(\log(\varepsilon^{-1})\varepsilon^{2(\hat{\alpha}-1)/(1+3\hat{\alpha})})$.

COROLLARY 4.7. *Let $\varepsilon \in (0,1)$ be a sufficiently small constant. Then, to achieve an iterate $\mathbf{u}_k$ satisfying $\|\mathbf{u}_k-\mathbf{u}^*\| \leq \varepsilon$, Algorithm 3 requires at most*

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

*line-search steps.*

*Proof.* It follows from relationship (4.14) that

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

which together with (4.15) 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 Algorithm 3. In view of (3.4), 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, Theorem 4.6 indicates that the total number of line-search steps in Algorithm 3 is at most $O(\log(\varepsilon^{-1})\varepsilon^{2(\hat{\alpha}-1)/(1+3\hat{\alpha})})$, which completes the proof. ∎

*Remark* 4.8. By an analogous argument, we can also prove that Algorithm 3 requires at most $O(\log(\varepsilon^{-1})\varepsilon^{(\hat{\alpha}-1)/(1+3\hat{\alpha})})$ iterations to generate an iterate $\mathbf{u}_k$ such that $f(\mathbf{u}_k) - f^* \le \varepsilon$ for problem (1.1). Very recently, Doikov [8] has shown that, in the case $m = 2$, where $f_1$ is a convex function with a Hölder continuous gradient and $f_2(\mathbf{u}) = \|\mathbf{u}\|^2$, the lower complexity bound for first-order methods is precisely $O(\log(\varepsilon^{-1})\varepsilon^{(\hat{\alpha}-1)/(1+3\hat{\alpha})})$ in terms of function value accuracy. This finding confirms that Algorithm 3 achieves the optimal iteration complexity.

## 5. Numerical Experiments.

Preliminary numerical results are presented in this section to provide additional insights into the performance guarantees of the algorithms proposed in this paper. We aim to elucidate that the final error attained by the algorithm is influenced by both the stepsize and the Hölder exponent. The numerical experiments are conducted using Julia [4] (version 1.12) on an Apple Macintosh Mini with an M2 processor, 8 performance cores, and 32GB of memory. We have placed the Julia codes in the GitHub repository (https://github.com/ctkelley/Grad_Des_CKW.jl) with instructions for reproducing the figures.

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

Hölder continuous gradients arise naturally in partial differential equations (PDEs) involving non-Lipschitz nonlinearity [3, 14]. In this subsection, we introduce a numerical example from [3]. This problem is to solve the following two-dimensional PDE,

$$\mathcal{F}(u) = -\Delta u + \gamma u_+^{\alpha} = 0, \tag{5.1}$$

where $\alpha \in (0,1)$, $\gamma > 0$ is a constant and $u_+ = \max\{u, 0\}$. Discretizing (5.1) with the standard five point difference scheme [9] leads to the following nonlinear system,

$$\mathbf{F}(\mathbf{u}) = \mathbf{A}\mathbf{u} + \gamma\mathbf{u}_+^{\alpha} - \mathbf{b} = 0, \tag{5.2}$$

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 $\mathbf{u}_+^{\alpha} = \max\{\mathbf{u}, 0\}^{\alpha}$ is understood as a component-wise operation.

We now modify the above problem to enable direct computation of errors in the iterations. To this end, we follow [13, Example 4.4] and take as the exact solution the function

$$u^*(x,y) = \left(\frac{3r-1}{2}\right)^2 \max\left\{0, r - \frac{1}{3}\right\},$$

where $r = \sqrt{x^2+y^2}$. We enforce the following boundary conditions,

$$u(x,1) = u^*(x,1),\ u(x,0) = u^*(x,0),\ u(1,y) = u^*(1,y),\ u(0,y) = u^*(0,y),$$

for $0 < x, y < 1$. And these conditions are encoded into $\mathbf{b}$. Then our modified equation is

$$\mathbf{F}(\mathbf{u}) - \mathbf{c}^* = 0, \tag{5.3}$$

where $\mathbf{c}^* = \mathbf{F}(\mathbf{u}^*)$. The nonlinear system (5.3) corresponds to the optimality condition of the following problem,

$$\min_{\mathbf{u}\in\mathbb{R}^n} \ f(\mathbf{u}) = \frac{1}{2}\mathbf{u}^\top \mathbf{A}\mathbf{u} + \frac{\gamma}{1+\alpha}\mathbf{e}^\top \mathbf{u}_+^{1+\alpha} - (\mathbf{b}+\mathbf{c}^*)^\top \mathbf{u}, \tag{5.4}$$

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

The optimization model (5.4) is a special instance of problem (1.1) with $\Omega = \mathbb{R}^n$, $m = 2$,

$$f_1(\mathbf{u}) = \mathbf{u}^\top \mathbf{A}\mathbf{u} - 2(\mathbf{b}+\mathbf{c}^*)^\top \mathbf{u}, \quad \text{and} \quad f_2(\mathbf{u}) = \frac{2\gamma}{1+\alpha}\mathbf{e}^\top \mathbf{u}_+^{1+\alpha}.$$

It is clear that, $\nabla f_1$ is Lipschitz continuous with the corresponding Lipschitz constant $L_1 = 2\|\mathbf{A}\|$, and $\nabla f_2$ is Hölder continuous with the Hölder exponent $\alpha$ and $L_2 = 2\gamma$ from

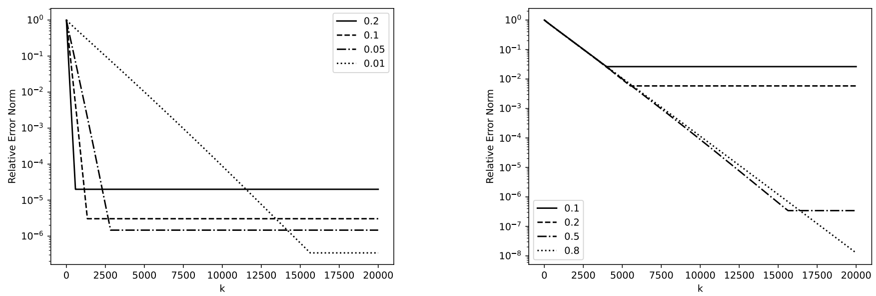
$$\|\nabla f_2(\mathbf{u}) - \nabla f_2(\mathbf{v})\| = 2\gamma \left\|\mathbf{u}_+^\alpha - \mathbf{v}_+^\alpha\right\| \le 2\gamma \|\mathbf{u}-\mathbf{v}\|^\alpha,$$

for all $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$. Moreover, the function $f = (f_1+f_2)/2$ is $\lambda(\mathbf{A})$-strongly convex, where $\lambda(\mathbf{A})$ is the smallest eigenvalue of the symmetric positive definite matrix $\mathbf{A}$. Let $\mathbf{u}^*$ be the vector obtained by evaluating $u^*$ at the interior grid points. Then $\mathbf{u}^*$ serves as the unique global minimizer of problem (5.4).

In the subsequent experiments, we use the solution of $\mathbf{A}\mathbf{u}_0 = \mathbf{b}$ as the initial iterate. This is the discretization of Laplace's equation with the boundary conditions. In this way, we ensure that the entire iteration satisfies the boundary conditions. Unless otherwise specified, we set the spatial mesh width as $h = 2^{-4}$ in this subsection. The dimension of the discretized problem is $n = (h^{-1}-1)^2$.

**5.1.1. Numerical results of Algorithm 1.** In the first experiment, we scrutinize the performance of Algorithm 1 under different stepsizes for problem (5.4) with $\alpha = 0.5$ and $\gamma = 0.5$. Specifically, Algorithm 1 is tested for stepsizes of the form $\tau = \tau_0 h^2$, where $\tau_0$ is taken from the set $\{0.2, 0.1, 0.05, 0.01\}$. The corresponding numerical results, presented in Figure 1(a), 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.

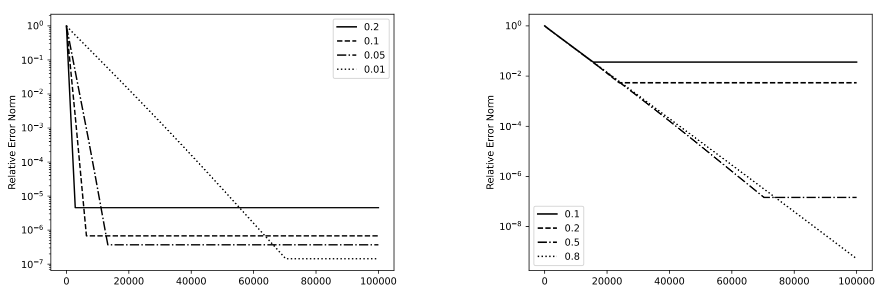
In the second experiment, we vary the Hölder exponent $\alpha$ over the values in $\{0.1, 0.2, 0.5, 0.8\}$, while fixing $\tau_0 = 0.01$. Figure 1(b) similarly tracks the decay of the distance to the global minimizer over iterations. It is evident that, as the value of $\alpha$ decreases, the final error attained by Algorithm 1 increases under the same stepsize. Therefore, the associated optimization problems become increasingly ill-conditioned and thus more challenging to solve for smaller values of $\alpha$. These findings offer empirical support for our theoretical analysis.




(a) different values of $\tau_0$ (b) different values of $\alpha$

FIG. 1. *Numerical performance of Algorithm* 1 *for problem* (5.4) *with* $h = 2^{-4}$.

We now repeat the experiment with $h = 2^{-5}$, so we reduce the mesh width by a factor of two and increase the norm of $\mathbf{A}$ by a factor of four. As one would expect the stepsize must decrease by a factor of four for stability.




(a) different values of $\tau_0$ (b) different values of $\alpha$

FIG. 2. *Numerical performance of Algorithm* 1 *for problem* (5.4) *with* $h = 2^{-5}$.

**5.1.2. Numerical results of Algorithm 2.** We repeat the study in subsection 5.1.1 for Algorithm 2 by varying the values of the Hölder exponent $\alpha$. We set $\varepsilon = 10^{-6}$ and $\mu = 2\pi^2$ in Algorithm 2, which is a lower estimate for the smallest eigenvalue of $\mathbf{A}$. The stepsize is initialized to $0.1h^2$ in the line-search procedure. The corresponding numerical results are depicted in Figure 3. Comparing Figure 3 to Figure 2(b) shows the benefits of the line-search procedure in Algorithm 2, which does not need to manually adjust the value of $\tau_0$ to converge for a given value of $\varepsilon$.

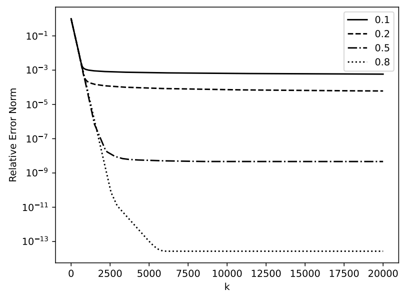
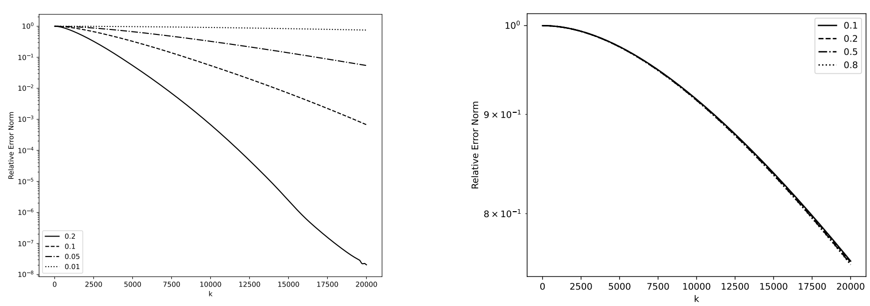



FIG. 3. *Numerical performance of Algorithm* 2 *for problem* (5.4) *with different values of* $\alpha$*.*

#### 5.1.3. Numerical results of Algorithm 3.

We report the numerical performance of Algorithm 3 on two experiments. Guided by the observation in Remark 4.2, we test Algorithm 3 with a fixed stepsize $\nu = \tau_0 h^2$. In the first example, we use the values for $\tau_0$ from Figure 1. In this way we can directly compare the performance of Algorithm 3 with that of Algorithm 1. The corresponding results, shown in Figure 4, are poor. The reason for this is that we are not exploiting the ability of Algorithm 3 to use larger stepsizes. In the second example, we consider larger values for $\tau_0$ in Figure 5(a) and set $\tau_0 = 20$ in Figure 5(b). The convergence is much better in all cases. The hardest case ($\alpha = 0.1$) has very irregular convergence in the terminal phase of iterations.




(a) different values of $\tau_0$ (b) different values of $\alpha$

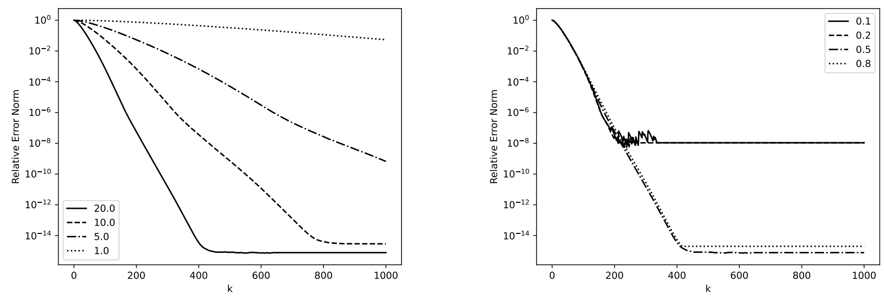
FIG. 4. *Numerical performance of Algorithm* 3 *for problem* (5.4) *with smaller stepsizes.*

#### 5.1.4. Stepsize and termination.

It is useful to look at the values of stepsizes from Remark 4.2. We note that for problem (5.4), $M = O(h^{-2})$. We are using $\hat{\alpha} = \alpha$ and neglecting constants in the estimate. We tabulate in Table 1 the value of

$$\nu = h^{2p_1} \varepsilon^{p_2} \tag{5.5}$$

where $p_1 = (1+\alpha)/(1+3\alpha)$ and $p_2 = 2(1-\alpha)/(1+3\alpha)$. Contrasting the values of $\nu$ in Table 1 to the value of $20h^2 \approx 0.08$, we can see that the stepsize estimate from




(a) different values of $\tau_0$

(b) different values of $\alpha$

FIG. 5. *Numerical performance of Algorithm* 3 *for problem* (5.4) *with larger stepsizes.*

(5.5) is very pessimistic. For smaller values of $\alpha$, the predicted stepsize is too small to be useful in practice.

TABLE 1
*Representative values of* $\nu$.

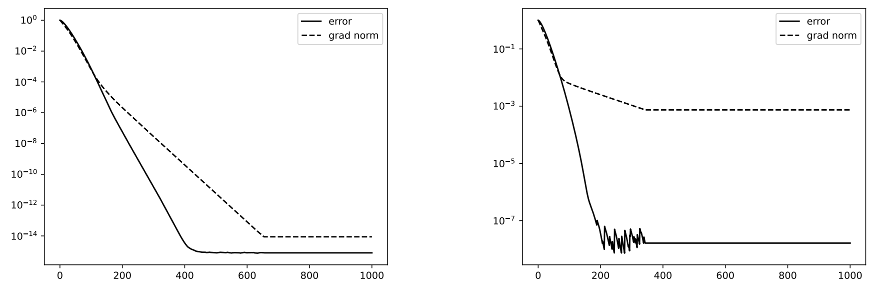
| $\alpha\backslash\varepsilon$ | 1.00e-02 | 1.00e-03 | 1.00e-05 | 1.00e-08 |
|---|---|---|---|---|
| 0.1 | 1.56e-05 | 6.43e-07 | 1.09e-09 | 7.68e-14 |
| 0.2 | 1.56e-04 | 1.56e-05 | 1.56e-07 | 1.56e-10 |
| 0.5 | 5.69e-03 | 2.26e-03 | 3.59e-04 | 2.26e-05 |
| 0.8 | 3.09e-02 | 2.36e-02 | 1.37e-02 | 6.08e-03 |

Next, we consider the complexity bound $O(\log(\varepsilon^{-1})M^{p_1}\varepsilon^{-p_2})$. In Table 2 we present the predicted number of iterations. The estimates are pessimistic except for the larger values of $\alpha$ when compared to the findings we report in Figure 5.

TABLE 2
*Representative iteration numbers.*

| $\alpha\backslash\varepsilon$ | 1.00e-02 | 1.00e-03 | 1.00e-05 | 1.00e-08 |
|---|---|---|---|---|
| 0.1 | 4.26e+05 | 1.55e+07 | 1.52e+10 | 3.46e+14 |
| 0.2 | 4.25e+04 | 6.38e+05 | 1.06e+08 | 1.70e+11 |
| 0.5 | 1.17e+03 | 4.40e+03 | 4.63e+04 | 1.17e+06 |
| 0.8 | 2.15e+02 | 4.23e+02 | 1.21e+03 | 4.37e+03 |

Finally, we consider termination of the iteration. In problem (5.4), we know the exact solution and can evaluate the algorithms in terms of the error. In practice we cannot do that and must use the gradient norm as a surrogate for the error. While this is standard for smooth optimization, it could be a problem when the gradient is not Lipschitz continuous. We illustrate this in Figure 6, where we compare the gradient norm with the error for the case $\tau_0 = 20$ using Algorithm 3. The numerical results in Figure 6 indicate that, when the gradient norm stops decreasing, the error has also stopped decreasing. However, the gradient norm is larger than the error norm, especially when the error is small, which is consistent with Hölder continuity.




(a) $\alpha = 0.5$ (b) $\alpha = 0.1$

FIG. 6. *Gradient and error norms for problem* (5.4).

**5.2. Semi-linear elliptic problem with a constraint.** We consider a second numerical example motivated by a semi-linear elliptic problem with a constraint on the solution in a certain set [14]. Let

$$\mathcal{H}(u) = -\Delta u + \delta |u|^{\alpha} \mathsf{sign}(u) - |u|^{p-1} u, \tag{5.6}$$

on $D = (0,1)^2$ with the boundary condition $u(x,y) = 0.5 - \sin(x)\sin(y)$ on $\partial D$. Here, $\alpha \in (0,1)$, $p > 1$, and $\delta > p/\alpha$ are three constants. We consider the variational inequality that is to find $u^* \in [-1,1]$ such that

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

for any $u \in [-1,1]$. This problem is equivalent to the following nonlinear equation,

$$0 = \mathcal{F}(u) := \begin{cases} \mathcal{H}(u), & \text{if } u - \mathcal{H}(u) \in [-1,1], \\ u - 1, & \text{if } u - \mathcal{H}(u) \geq 1, \\ u + 1, & \text{otherwise.} \end{cases} \tag{5.7}$$

By discretizing (5.6) with the standard five point difference scheme [9], problem (5.7) leads to the following system of nonlinear equations,

$$0 = \mathbf{F}(\mathbf{u}) := \frac{1}{\theta}\left(\mathbf{u} - \Pi_{\mathbf{U}}\left(\mathbf{u} - \theta\left(\mathbf{A}\mathbf{u} + \delta\,|\mathbf{u}|^{\alpha}\,\mathsf{sign}(\mathbf{u}) - |\mathbf{u}|^{p-1}\,\mathbf{u} - \mathbf{b}\right)\right)\right), \tag{5.8}$$

where $\mathbf{U} = [-1,1]^n$, $\theta > 0$ is a constant, $\mathbf{A} \in \mathbb{R}^{n\times n}$ is a symmetric positive definite matrix, and $\mathbf{b} \in \mathbb{R}^n$ encodes the boundary conditions. Note that (5.8) is the optimality condition of the following problem,

$$\min_{\mathbf{u}\in\mathbf{U}} \; f(\mathbf{u}) := \frac{1}{2}\mathbf{u}^\top \mathbf{A}\mathbf{u} + \frac{\delta}{1+\alpha}\mathbf{e}^\top |\mathbf{u}|^{1+\alpha} - \frac{1}{1+p}\mathbf{e}^\top |\mathbf{u}|^{1+p} - \mathbf{b}^\top \mathbf{u}. \tag{5.9}$$

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

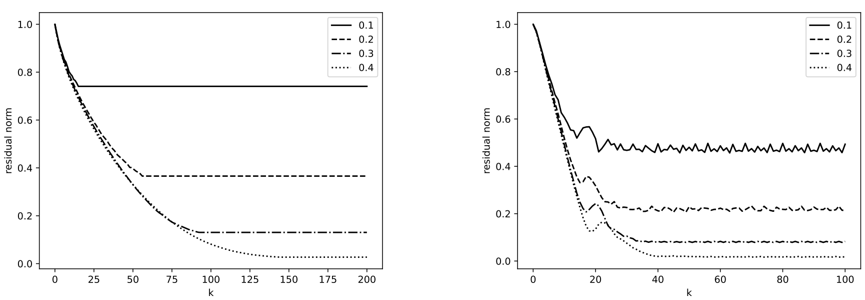
$$\nabla^2 f(\mathbf{u}) = \mathbf{A} + \delta\alpha \mathsf{Diag}\left(|\mathbf{u}|^{\alpha-1}\right) - p\mathsf{Diag}\left(|\mathbf{u}|^{p-1}\right),$$

Since $\delta > p/\alpha$, $\nabla^2 f(\mathbf{u})$ is symmetric positive definite for any $\mathbf{u} \in \mathbf{U}$ with $\mathbf{u}_i \neq 0$ $(i = 1, \dots, n)$. Hence, the function $f$ is $\mu$-strongly convex in $\mathbf{U}$ with $\mu = \lambda(\mathbf{A})$ and the system (5.8) has a unique solution in $\mathbf{U}$. The optimization model (5.9) is a special instance of problem (1.1) with $\Omega = \mathbf{U}$, $m = 2$,

$$f_1(\mathbf{u}) = \mathbf{u}^\top \mathbf{A}\mathbf{u} - 2\mathbf{b}^\top \mathbf{u} - \frac{2}{1+p}\mathbf{e}^\top |\mathbf{u}|^{1+p}, \text{ and } f_2(\mathbf{u}) = \frac{2\delta}{1+\alpha}\mathbf{e}^\top |\mathbf{u}|^{1+\alpha}.$$

It is clear that Assumption 1.1 (ii) holds with $\alpha_1 = 1$, $L_1 = 2\|\mathbf{A}\| + 2p$, $\alpha_2 = \alpha$, and $L_2 = 2\delta\alpha$.

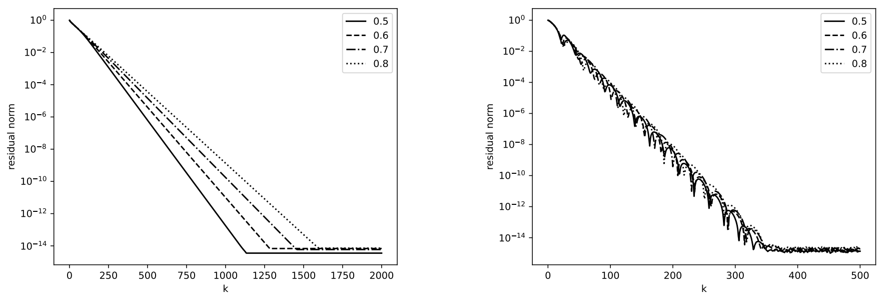
In this example, we do not have an analytic solution and we only plot the residual norm $\|\mathbf{F}(\mathbf{u})\|$ with $\theta$ being the stepsize. We compare the performance of Algorithm 1 and Algorithm 3 on problem (5.9) with $p = 1.5$ and $\delta = 20$. The stepsizes of Algorithm 1 and Algorithm 3 are set to $0.1h^2$ and $20h^2$, respectively. Figure 7 and Figure 8 illustrate the performance of two algorithms for $\alpha \in \{0.1, 0.2, 0.3, 0.4\}$ and $\alpha \in \{0.5, 0.6, 0.7, 0.8\}$, respectively. As was the case for subsection 5.1, problems where the exponent $\alpha$ for the non-Lipschitz term in the gradients is small are difficult. In particular, one cannot drive the residual to a small value. For larger values of $\alpha$, both algorithms demonstrate strong performance. Furthermore, Algorithm 3 exhibits a faster convergence rate, benefiting from the use of a larger stepsize.




(a) Algorithm 1 (b) Algorithm 3

FIG. 7. *Numerical performance of Algorithm* 1 *and Algorithm* 3 *for problem* (5.9) *with smaller values of* $\alpha$.

**6. Conclusion.** In this paper, we consider a class of strongly convex constrained optimization problems of the form (1.1). Example 1.2 shows that although each component function $f_i$ of the objective function $f$ admits a Hölder continuous gradient with an component $\alpha_i \in (0, 1]$, the gradient of $f$ is not necessarily Hölder continuous. To establish the iteration complexity of the projected gradient descent methods for this class of problems, we use the parameter $\hat{\alpha} = \min_{i\in[m]} \alpha_i$ to determine the complexity bound. Algorithm 1 is a new version of projected gradient method for problem (1.1) with an appropriately fixed stepsize. Theorem 2.2 shows that Algorithm 1 can find an iterate in the feasible set $\Omega$ with a distance to the global minimizer less than $\varepsilon$ at most $O(\log(\varepsilon^{-1})\varepsilon^{2(\hat{\alpha}-1)/(1+\hat{\alpha})})$ iterations. This recovers the classical complexity result when $\hat{\alpha} = 1$ and reveals the additional difficulty imposed by the weaker smoothness of the objective function for $\hat{\alpha} < 1$. Algorithm 2 is a modification of Algorithm 1 for problems where the parameters $\alpha_i$ and $L_i$ are difficult to estimate for the stepsize. In Algorithm 3, the stepsize is updated by the universal scheme at each iteration, which improves the complexity bound to $O(\log(\varepsilon^{-1})\varepsilon^{2(\hat{\alpha}-1)/(1+3\hat{\alpha})})$. Numerical




(a) Algorithm 1 (b) Algorithm 3

Fig. 8. *Numerical performance of Algorithm* 1 *and Algorithm* 3 *for problem* (5.9) *with larger values of* $\alpha$*.*

experiments are conducted to validate our theoretical findings, demonstrating the expected behavior of projected gradient descent methods under different stepsizes and Hölder exponents. These results offer new insights into the performance guarantees of the classic projected gradient descent methods for a broader class of optimization problems with non-Lipschitz gradients.

## REFERENCES

[1] G. Alefeld and X. Chen, *A regularized projection method for complementarity problems with non-Lipschitzian functions*, Math. Comput., 77 (2008), pp. 379–395.

[2] J.-C. Baritaux, K. Hassler, and M. Unser, *An efficient numerical method for general $L_p$ regularization in fluorescence molecular tomography*, IEEE Trans. Med. Imaging, 29 (2010), pp. 1075–1087.

[3] J. W. Barrett and R. M. Shanahan, *Finite element approximation of a model reaction-diffusion problem with a non-Lipschitz nonlinearity*, Numer. Math., 59 (1991), pp. 217–242.

[4] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah, *Julia: A fresh approach to numerical computing*, SIAM Rev., 59 (2017), pp. 65–98.

[5] L. S. Borges, F. S. V. Bazán, and L. Bedin, *A projection-based algorithm for $\ell_2$-$\ell_p$ Tikhonov regularization*, Math. Methods Appl. Sci., 41 (2018), pp. 5919–5938.

[6] X. Chen, C. T. Kelley, and L. Wang, *A new complexity result for strongly convex optimization with locally $\alpha$-Hölder continuous gradients*, arXiv:2505.03506v1, (2025).

[7] O. Devolder, F. Glineur, and Y. Nesterov, *First-order methods of smooth convex optimization with inexact oracle*, Math. Program., 146 (2014), pp. 37–75.

[8] N. Doikov, *Lower complexity bounds for minimizing regularized functions*, Optim. Lett., (2025), pp. 1–20.

[9] R. J. LeVeque, *Finite Difference Methods for Ordinary and Partial Differential Equations: Steady-State and Time-Dependent Problems*, Society for Industrial and Applied Mathematics, 2007.

[10] Y. Nesterov, *Universal gradient methods for convex optimization problems*, Math. Program., 152 (2015), pp. 381–404.

[11] Y. Nesterov, *Lectures on Convex Optimization*, Springer, 2018.

[12] Y. Nesterov, *Universal complexity bounds for universal gradient methods in nonlinear optimization*, arXiv:2509.20902, (2025).

[13] X. Qu, W. Bian, and X. Chen, *An extra gradient Anderson-accelerated algorithm for pseudomonotone variational inequalities*, Math. Comput., (2025).

[14] M. Tang, *Uniqueness of bound states to $\Delta u - u + |u|^{p-1}u = 0$ in $\mathbb{R}^n$, $n \geq 3$*, Invent. Math., (2025), pp. 1–47.

[15] M. Yashtini, *On the global convergence rate of the gradient descent method for functions with Hölder continuous gradients*, Optim. Lett., 10 (2016), pp. 1361–1370.