A Modified Projected Gradient Algorithm for Solving Quasiconvex Programming with Applications

Pham Thi Hoai* Nguyen Duy Hoang† Felipe Lara‡

March 5, 2026

Abstract

We introduce a novel projected gradient algorithm for solving quasiconvex optimization problems over closed convex sets. The key innovation of the algorithm is an adaptive, parameter-free stepsize rule that requires no line search and avoids estimating constants, such as Lipschitz modulus. Unlike recent self-adaptive approach given in [17], which typically produce monotonically non-increasing stepsizes, we propose a rule where the stepsize sequence is proven to be non-decreasing and convergent to a positive limit after finitely many iterations. This property enables consistently longer steps, potentially accelerating convergence significantly. We establish convergence guarantees for the algorithm across various classes of functions, including quasiconvex, pseudoconvex, convex, and strongly convex objective functions. Numerical results on numerous test instances demonstrate the efficiency of the proposed method, showcasing its competitive performance and often superior convergence speed compared to state-of-the-art alternatives.

Keywords: Quasiconvex programming, Projected Gradient method, Constrained optimization, Nonconvex programming, Supervised feature selection

Mathematics Subject Classification: 90C26; 90C30.

*Faculty of Mathematics and Informatics, Hanoi University of Science and Technology, 1 Dai Co Viet Road, Hanoi, VietNam. email: hoai.phamthi@hust.edu.vn; phamhoai051087@gmail.com. Web: hoaiphamthi.github.io, ORCID-ID: 0000-0001-6702-5227

†Faculty of Mathematics and Informatics, Hanoi University of Science and Technology, 1 Dai Co Viet Road, Hanoi, Vietnam. email: 0969742077ndh@gmail.com

‡Instituto de Alta Investigación (IAI), Universidad de Tarapacá, Arica, Chile. E-mail: felipelaraobreque@gmail.com; flarao@academicos.uta.cl. Web: felipelara.cl, Orcid-ID: 0000-0002-9965-0921

1 Introduction

Considering the constrained nonlinear optimization problem

$$\min\{f(x) : x \in C\}, \tag{P}$$

where $C \subseteq \mathbb{R}^n$ is a closed convex set and $f : C \to \mathbb{R}$ is a continuously differentiable function. Throughout the paper, we use the following assumption.

Assumption 1. *Problem (P) has a finite optimal value $f^* > -\infty$, and its optimal solution set X^* is nonempty.*

It is known that a necessary condition for a point $z \in C$ to be a local optimum of Problem (P) is the stationarity condition:

$$\langle \nabla f(z), x - z \rangle \geq 0, \ \forall\, x \in C.$$

This condition is also sufficient when f is convex. Stationarity can be conveniently verified via the fixed-point equation (see, e.g., [2])

$$z = P_C\big(z - s\nabla f(z)\big), \quad \text{for some } s > 0, \tag{1}$$

where $P_C(\cdot)$ denotes the orthogonal projection onto C, defined as

$$P_C(x) = \operatorname*{argmin}_{y \in C} \|y - x\|.$$

Motivated by (1), a classical approach for finding stationary points of (P) is the projected gradient (PG) method. Starting from an initial point x^0, it generates a sequence $\{x^k\}$ via

$$x^{k+1} = P_C\big(x^k - \lambda_k \nabla f(x^k)\big), \quad k = 0, 1, \ldots, \tag{2}$$

where $\lambda_k > 0$ is a stepsize. Standard convergence analysis for (2) typically requires ∇f to be globally Lipschitz continuous on C, that is, there exists $L > 0$ such that

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

The stepsize λ_k can then be chosen as a constant in $(0, 2/L)$ or determined by backtracking line search. However, these strategies suffer from several drawbacks: a large Lipschitz constant L leads to prohibitively small steps; estimating L accurately can be difficult; and line searches incur additional computational cost. These limitations motivate the development of adaptive, parameter-free stepsize rules that do not rely on global constants or backtracking.

In the case when problem (P) is convex, effective adaptive schemes have been proposed, such as those in [16] (AdPG) and [14, 15] (AdaPG). These methods can be applied to a more general setting of (P) that composite problems with locally Lipschitz gradient assumption on the differentiable term.

In the nonconvex case, Thang and Hai [17] introduced GDA algorithm - a self adaptive method for quasiconvex problems with the globally Lipschitz

continuity of the gradient of the objective. Its stepsize sequence is monotonically non-increasing, which may result in conservatively small steps in later iterations. Another approach, the PG-NGD algorithm [8], was designed for nonconvex (P) with objectives satisfying:

(C_1) f is differentiable with a globally Lipschitz gradient on C;

(C_2) For any $u, v \in C$, the directional derivative function

$$g_{uv}(t) = \langle \nabla f(u + t(v - u)), v - u \rangle, \tag{C_2}$$

is quasiconvex on $[0, 1]$.

While condition (C_2) covers convex, concave, and indefinite quadratic functions (see [8]), it excludes general quasiconvex and even strongly quasiconvex functions (cf. Lemmas A.1 and A.2 in Appendix). A notable feature of PG-NGD is that its stepsize sequence becomes non-decreasing after finitely many iterations, thereby avoiding the small stepsize issue and potentially accelerating convergence.

In this paper, we focus on Problem (P) under Assumption 1 and Assumption 2 below.

Assumption 2. *f is quasiconvex and differentiable with a globally Lipschitz gradient on C.*

We will introduce a novel projected gradient algorithm featuring an adaptive stepsize rule. Unlike existing self-adaptive method given by Thang and Hai [17] that yields monotonically decreasing stepsizes, our rule produces a non-decreasing stepsize sequence that converges to a positive limit after finitely many iterations. This property allows the algorithm to take consistently longer steps, which often translates into significantly faster convergence. We establish rigorous convergence guarantees for quasiconvex, pseudoconvex, convex, and strongly convex objectives. Extensive numerical experiments on established benchmarks demonstrate the efficiency of the proposed method, showing competitive performance and frequently superior convergence speed compared to state-of-the-art alternatives.

The remainder of the paper is organized as follows. Section 2 recalls necessary preliminaries. In Section 3, we revisit the PG-NGD method in the context of quasiconvex objectives. Section 4 presents our modified algorithm (MPG-NGD) and its convergence analysis for quasiconvex programming over closed convex sets. Numerical experiments on benchmark and synthetic data are reported in Section 5, illustrating the practical and efficient performance of the proposed method. Finally, Section 6 provides concluding remarks. Some technical proofs are deferred to the Appendix.

2 Preliminaries

During the whole paper, let $C \subseteq \mathbb{R}^n$ be a closed convex set.

Definition 2.1. Let f continuously differentiable over C. Then f is said to be L-smooth or has a global Lipschitz gradient on C with Lipschitz constant $L > 0$ if it satisfies

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

Lemma 2.1. *[2] If f is L-smooth over C, then*

$$f(y) - f(x) - \langle y - x, \nabla f(x)\rangle \leq \frac{L}{2}\|x - y\|^2,\ \forall\ x, y \in C. \tag{3}$$

Definition 2.2. Let f be a continuously differentiable function over C. Then $x^* \in C$ is said to be a *stationary point* of (P) if

$$\langle \nabla f(x^*), x - x^*\rangle \geq 0$$

for every $x \in C$.

Definition 2.3. *The projection of x on C is defined by*

$$P_C(x) = \operatorname{argmin}\{\|y - x\| \mid y \in C\}.$$

Lemma 2.2. *[2, Theorem 9.8] Let $x \in \mathbb{R}^n$. Then $z = P_C(x)$ if and only if*

$$(x - z)^T(y - z) \leq 0,\ \forall\ y \in C. \tag{4}$$

Lemma 2.3. *[2, Theorem 9.2] Let f be a continuously differentiable function on C and x^* be a local minimum of (P). Then x^* is a stationary point of (P).*

Lemma 2.4. *[2, Theorem 9.10] Let f be a continuously differentiable function defined on C and $s > 0$. Then x^* is a stationary point of (P) if and only if*

$$x^* = P_C\big(x^* - s\nabla f(x^*)\big).$$

Definition 2.4. A function $f\colon C \to \mathbb{R}$ is said to be

(i) *convex* on C if for every $\lambda \in [0, 1]$ and $x, y \in C$, we have

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y).$$

It is worth noting that f is convex if and only if the epigraph of f defined by $\operatorname{epi}(f) = \{(w, t) \in \mathbb{R}^n \times \mathbb{R} \mid f(w) \leq t\}$, is a convex set.

(ii) *quasiconvex* if its sublevel set $L_\alpha(f) = \{x \in C \mid f(x) \leq \alpha\}$ is convex for every $\alpha \in \mathbb{R}$. An equivalent condition is that for every $x, y \in C$ and $\lambda \in [0, 1]$,

$$f(\lambda x + (1 - \lambda)y) \leq \max\{f(x), f(y)\}.$$

(iii) *σ-strongly convex* ($\sigma > 0$) on C if for every $\lambda \in [0, 1]$ and $x, y \in C$, we have

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y) - \lambda(1 - \lambda)\frac{\sigma}{2}\|x - y\|^2.$$

If f is continuously differentiable then f is σ-strongly convex if and only if

$$f(x) - f(y) \geq \langle \nabla f(y), x - y\rangle + \frac{\sigma}{2}\|x - y\|^2,\ \forall\ x, y \in C. \tag{5}$$

(iv) *strongly quasiconvex* with modulus $\gamma > 0$ if for every $x, y \in C$ and every $\lambda \in [0,1]$,

$$f(\lambda x + (1-\lambda)y) \leq \max\{f(x), f(y)\} - \frac{\gamma}{2}\lambda(1-\lambda)\|y-x\|^2.$$

(v) If f is continuously differentiable on C, then f is said to be *pseudoconvex* on C if

$$\langle \nabla f(x), y-x\rangle \geq 0 \Longrightarrow f(y) \geq f(x),\ \forall\ x, y \in C.$$

Now, we recall the following first order characterization for differentiable quasiconvex functions due to Arrow and Enthoven [1].

Lemma 2.5. *[1, 6] Suppose that $f\colon C \to \mathbb{R}$ is continuously differentiable. Then f is quasiconvex if and only if for all $x, y \in C$:*

$$f(y) \leq f(x) \implies \langle \nabla f(x), y-x\rangle \leq 0.$$

An extension to the previous characterization is the following result to [19] (see also [13] and [7]).

Lemma 2.6. *[7, Corollary 8] Let h be continuously differentiable on $C \subseteq \mathbb{R}^n$. If h is strongly quasiconvex with modulus $\gamma > 0$, then for every $x, y \in C$ the following implication holds:*

$$h(x) \leq h(y) \implies \langle \nabla h(y),\, y-x\rangle \geq \frac{\gamma}{2}\,\|y-x\|^2. \tag{6}$$

Conversely, if (6) holds, then h is strongly quasiconvex with modulus $\frac{\gamma}{2}$.

Lemma 2.7. *[4, Theorem 2.3.8] Let h and g be functions defined on C. Suppose that h is convex and g is a positive affine function. Defining*

$$z(x) = \frac{h(x)}{g(x)}.$$

Then z is quasiconvex on C.

Lemma 2.8. *[12] Let $f\colon C \to \mathbb{R}$ be a strongly quasiconvex function with constant $\gamma > 0$. Then all lower level sets of f are bounded.*

We also need the concept of quasi-Fejér sequence which will be useful for deriving the convergence results of our new algorithm in the upcoming sections.

Definition 2.5. *[3]* A sequence $\{y^k\}$ is *quasi-Fejér convergent* to a set $U \subseteq \mathbb{R}^n$ if for every $u \in U$ there exists a sequence $\{\epsilon_k\} \subseteq \mathbb{R}_+$ such that $\sum\limits_{k=0}^{+\infty} \epsilon_k < +\infty$ and

$$\|y^{k+1} - u\|^2 \leq \|y^k - u\|^2 + \epsilon_k,\ \forall\ k \geq 0.$$

Lemma 2.9. *[3] If $\{y^k\}$ is quasi-Fejér convergent to a nonempty set $U \subseteq \mathbb{R}^n$, then $\{y^k\}$ is bounded. If furthermore, a cluster point y of $\{y^k\}$ belongs to U, then*

$$\lim_{k\to+\infty} y^k = y.$$

For a further study on generalized convexity and first-order algorithms, we refer to [2, 4, 7, 9, 12, 13] and references therein.

3 Convergence of PG-NGD for minimizing a quasiconvex function over a closed convex set

In this section, we investigate the convergence of PG-NGD given by Hoai et al. [8] for solving (P) under Assumptions 1 and 2. Below is the PG-NGD scheme.

Algorithm 1 (PG-NGD) [8]

Step 0 (Initialization). Select $\lambda_0 > 0$, $0 < \eta_1 < \eta_0 < 1$; a tolerance $\varepsilon > 0$ and a positive real sequence $\{\varepsilon_k\}$ such that $\sum_{k=0}^{+\infty} \varepsilon_k < +\infty$. Choose $x^0 \in C$, $x^1 = P_C(x^0 - \lambda_0 \nabla f(x^0))$, and set $k = 1$.

Step 1. **If** $\|\nabla f(x^k) - \nabla f(x^{k-1})\| > \frac{\eta_0}{\lambda_{k-1}} \|x^k - x^{k-1}\|$
then compute

$$\lambda_k = \eta_1 \frac{\|x^k - x^{k-1}\|}{\|\nabla f(x^k) - \nabla f(x^{k-1})\|}$$

else

$$\lambda_k = (1 + \varepsilon_{k-1})\lambda_{k-1}.$$

Step 2. Compute $x^{k+1} = P_C(x^k - \lambda_k \nabla f(x^k))$.

Step 3. **If** $\dfrac{\|x^{k+1} - x^k\|}{\lambda_k} < \varepsilon$ **then** STOP
else setting $k := k + 1$ and return to **Step 1**.

Remark 1. Our stopping criteria is little different from PG-NGD in Hoai et al. [8]. In particular, the original version in [8] used the condition $\|x^{k+1} - x^k\| < \varepsilon$ but in Algorithm 1 it becomes $\dfrac{\|x^{k+1} - x^k\|}{\lambda_k} < \varepsilon$. According to Lemma 2.4, $\inf_{k \geq 0} \lambda_k > 0$ then these stopping conditions are actually equivalent and both of them will be terminated after finitely number of iterations if we have $\|x^{k+1} - x^k\| \to 0$. But the usage of $\dfrac{\|x^{k+1} - x^k\|}{\lambda_k} < \varepsilon$ brings the consistence with the classical gradient descent algorithm when $C = \mathbb{R}^n$ and the stopping criteria $\|\nabla f(x^k)\| = \dfrac{\|x^{k+1} - x^k\|}{\lambda_k} < \varepsilon$ remains.

It is worth noting that in [8], PG-NGD includes some nice properties of λ_k given in Lemmas 3.2, 3.3 [8] and these ones are derived from the global Lipschitzness of ∇f and the convergence of positive series $\sum_{k=0}^{+\infty} \varepsilon_k$. Hence, under Assumptions 1 and 2, PG-NGD still keeps the results of Lemmas 3.2, 3.3 in [8] as follows.

Lemma 3.1. *[8, Lemma 3.2] If Problem (P) satisfies Assumptions 1 and 2, then for Algorithm 1, we have* $\inf_{k \geq 0} \lambda_k > 0$ *and* $\{\lambda_k\}$ *is convergent.*

Lemma 3.2. *[8, Lemma 3.3] Under Assumptions 1 and 2, Algorithm 1 generates $\{x^k\}$ satisfying that there exists $\hat{k} \in \mathbb{N}$ such that*

$$\|\nabla f(x^k) - \nabla f(x^{k-1})\| \leq \frac{\eta_0}{\lambda_{k-1}}\|x^k - x^{k-1}\|, \ \forall\ k \geq \hat{k}.$$

When condition (C_2) does not satisfy, the quasiconvexity of f still follows the decreasing of $\{f(x^k)\}_{k\geq\hat{k}}$ as presented in the following lemma.

Lemma 3.3. *Suppose that Problem (P) satisfies Assumptions 1 and 2. Then the sequence $\{x^k\}$, generated by Algorithm 1 (PG-NGD), satisfies that*

$$f(x^k) > f(x^{k+1}), \ \forall\ k \geq \hat{k},$$

unless $x^{k+1} = x^k$, i.e., x^k is a stationary point of (P). As a result, if $\|x^{k+1} - x^k\| \neq 0$ for all $k \geq 0$, then the sequence $\{f(x^k)\}_{k\geq\hat{k}}$ is strictly decreasing and has lower bounded by f^*, hence the limit $\lim_{k\to+\infty} f(x^k)$ exists, and thus $\lim_{k\to+\infty} f(x^k) = \hat{f} \geq f^*$.

Proof. Since $x^{k+1} = P_C(x^k - \lambda_k \nabla f(x^k))$ with $\lambda_k > 0$, from Lemma 2.2 we have

$$\begin{aligned} &\langle x^k - \lambda_k \nabla f(x^k) - x^{k+1}, x^k - x^{k+1}\rangle \leq 0, \\ \Longrightarrow\ &\langle \nabla f(x^k), x^{k+1} - x^k\rangle \leq -\frac{1}{\lambda_k}\|x^{k+1} - x^k\|^2. \end{aligned}$$

Hence,

$$\langle \nabla f(x^k) - \nabla f(x^{k+1}), x^{k+1} - x^k\rangle \leq -\frac{1}{\lambda_k}\|x^{k+1} - x^k\|^2 + \langle \nabla f(x^{k+1}), x^k - x^{k+1}\rangle. \tag{7}$$

If there exists $k \geq \hat{k}$ such that $f(x^k) \leq f(x^{k+1})$, then from Lemma 2.5, we obtain that $\langle \nabla f(x^{k+1}), x^k - x^{k+1}\rangle \leq 0$. Therefore, unless $x^k = x^{k+1}$, from (7) and since $0 < \eta_0 < 1$, we have

$$\langle \nabla f(x^k) - \nabla f(x^{k+1}), x^{k+1} - x^k\rangle < -\frac{\eta_0}{\lambda_k}\|x^{k+1} - x^k\|^2. \tag{8}$$

By using Cauchy-Schwarz inequality,

$$\langle \nabla f(x^k) - \nabla f(x^{k+1}), x^{k+1} - x^k\rangle \geq -\|\nabla f(x^k) - \nabla f(x^{k+1})\|.\|x^{k+1} - x^k\|.$$

Furthermore, from Lemma 3.2, we get

$$\begin{aligned} &\|\nabla f(x^{k+1}) - \nabla f(x^k)\| \leq \frac{\eta_0}{\lambda_k}\|x^{k+1} - x^k\| \ \ \forall\ k \geq \hat{k} \\ \Longrightarrow\ &\langle \nabla f(x^k) - \nabla f(x^{k+1}), x^{k+1} - x^k\rangle \geq -\frac{\eta_0}{\lambda_k}\|x^{k+1} - x^k\|^2. \end{aligned} \tag{9}$$

Then, we have a contradiction between (8) and (9), and the result follows. □

Remark 2. This section addresses a fundamental theoretical question: whether the PG-NGD method (Algorithm 1) converges when applied to Problem (P) under Assumptions 1 and 2? Our analysis has yielded partial results. Lemma 3.3 shows that the sequence of function values $\{f(x^k)\}$ decreases strictly to a finite limit $\hat{f} \geq f^*$, provided that the iterates are non-stationary ($\|x^{k+1} - x^k\| \neq 0$). However, $|f(x^k) - f(x^{k+1})| \to 0$ is insufficient to guarantee $\inf_{k \geq \hat{k}} \|x^{k+1} - x^k\| = 0$. Without this, the iterates may oscillate between points with similar function values without approaching a single limit point. Consequently, the convergence of the original PG-NGD algorithm in the context of general quasiconvex minimization is still an open theoretical question.

To address this gap, in the next section, we introduce a modified version of PG-NGD that is designed to ensure convergence of the iterates to a stationary point of Problem (P) under Assumptions 1 and 2.

4 A modified version of PG-NGD for quasiconvex programming over a closed convex set

In this section, we propose the modified version of PG-NGD that is called MPG-NGD as follows

---

Algorithm 2 (MPG-NGD)

Step 0 (Initialization). Select $\lambda_0 > 0$; $0 < \eta_1 < \eta_0 < 1$, a tolerance $\varepsilon > 0$ and a positive real sequence $\{\varepsilon_k\}$ such that $\sum_{k=0}^{+\infty} \varepsilon_k < +\infty$. Choose $x^0 \in C$, $x^1 = P_C(x^0 - \lambda_0 \nabla f(x^0))$, and set $k = 1$.

Step 1. If

$$f(x^k) - f(x^{k-1}) - \langle \nabla f(x^{k-1}), x^k - x^{k-1} \rangle > \frac{\eta_0}{\lambda_{k-1}} \|x^k - x^{k-1}\|^2 \tag{10}$$

then

$$\lambda_k = \frac{\eta_1 \|x^k - x^{k-1}\|^2}{f(x^k) - f(x^{k-1}) - \langle \nabla f(x^{k-1}), x^k - x^{k-1} \rangle} \tag{11}$$

else

$$\lambda_k = (1 + \varepsilon_{k-1}) \lambda_{k-1} \tag{12}$$

Step 2. Compute $x^{k+1} = P_C(x^k - \lambda_k \nabla f(x^k))$.

Step 3. If $\dfrac{\|x^{k+1} - x^k\|}{\lambda_k} < \varepsilon$ **then** STOP

else setting $k := k + 1$ and return to **Step 1**.

The following technical results will be useful in the sequel.

Lemma 4.1. *Suppose that Problem (P) satisfies Assumptions 1 and 2 and the sequence $\{\lambda_k\}$ is generated by Algorithm 2 (MPG-NGD), then we have $\inf_{k \geq 0} \lambda_k > 0$ and $\{\lambda_k\}$ is convergent.*

Proof. We divide the proof into two parts:
(i): Note that, from Assumption 2, f is L-smooth on C, then by Lemma 2.1,

$$f(x^k) - f(x^{k-1}) - \left\langle x^k - x^{k-1}, \nabla f(x^{k-1}) \right\rangle \leq \frac{L}{2}\|x^k - x^{k-1}\|^2, \ \forall\ x, y \in C.$$

Therefore, if condition (10) in Algorithm 2 is true, i.e.,

$$f(x^k) - f(x^{k-1}) - \langle \nabla f(x^{k-1}), x^k - x^{k-1} \rangle > \frac{\eta_0}{\lambda_{k-1}}\|x^k - x^{k-1}\|^2 \geq 0,$$

then

$$\lambda_k = \frac{\eta_1 \|x^k - x^{k-1}\|^2}{f(x^k) - f(x^{k-1}) - \langle \nabla f(x^{k-1}), x^k - x^{k-1} \rangle} \geq \frac{2\eta_1}{L}.$$

The converse case, $\lambda_k = (1 + \varepsilon_{k-1})\lambda_{k-1} \geq \lambda_{k-1}$. Hence $\lambda_k \geq \min\{\lambda_{k-1}, \frac{2\eta_1}{L}\} \geq \min\{\lambda_0, \frac{2\eta_1}{L}\} > 0$ for all $k \geq 0$. It means that $\inf\limits_{k\geq 0} \lambda_k > 0$.

(ii): Let $a_k = \ln \lambda_{k+1} - \ln \lambda_k$ for all $k \geq 0$. Then, $a_k = a_k^+ - a_k^-$, where

$$a_k^+ = \max\{0, a_k\}, \quad a_k^- = -\min\{0, a_k\}.$$

Clearly, $a_k^+ \geq 0$ and $a_k^- \geq 0$ for all $k \geq 0$. From the definition of λ_k in Algorithm 2, if λ_{k+1} is computed by formula (11), i.e.,

$$\lambda_{k+1} = \eta_1 \frac{\|x^{k+1} - x^k\|^2}{|f(x^{k+1}) - f(x^k) - \langle \nabla f(x^k), x^{k+1} - x^k \rangle|} \leq \eta_1 \frac{\lambda_k}{\eta_0} < \lambda_k,$$

then $a_k = \ln\left(\frac{\lambda_{k+1}}{\lambda_k}\right) < 0$. For the other case, i.e., λ_{k+1} is computed by formula (12), we derive that

$$a_k = \ln \frac{\lambda_{k+1}}{\lambda_k} \leq \ln(1 + \varepsilon_k) \leq \varepsilon_k, \ \forall\ k \geq 0.$$

Thus, we obtain $0 \leq a_k^+ \leq \varepsilon_k$ for all $k \geq 0$. Since $\sum\limits_{k=0}^{+\infty} \varepsilon_k$ is convergent then $\sum\limits_{k=0}^{+\infty} a_k^+ < +\infty$. Observing that $\sum\limits_{k=0}^{+\infty} a_k^-$ is a nonnegative series and using the following relation

$$\ln \lambda_{k+1} - \ln \lambda_0 = \sum_{i=0}^{k} a_i = \sum_{i=0}^{k} (a_i^+ - a_i^-) = \sum_{i=0}^{k} a_i^+ - \sum_{i=0}^{k} a_i^-, \tag{13}$$

we assert that: if $\lim\limits_{k\to+\infty} \sum\limits_{i=0}^{k} a_i^- = +\infty$, then

$$\lim_{k\to+\infty} (\ln \lambda_{k+1}) = -\infty \iff \lim_{k\to+\infty} \lambda_k = 0.$$

But we just obtain $\inf\limits_{k\geq 0} \lambda_k > 0$ (from (i)). This contradiction proves the convergence of $\sum\limits_{k=0}^{+\infty} a_k^-$. Finally, from (13) we get $\lim\limits_{k\to+\infty} \lambda_k = \lambda^* < +\infty$. □

Lemma 4.2. *Under Assumptions 1 and 2 and for Algorithm 2, there exists* $\tilde{k} \in \mathbb{N}$ *such that*

$$f(x^k) - f(x^{k-1}) - \langle \nabla f(x^{k-1}), x^k - x^{k-1} \rangle \leq \frac{\eta_0}{\lambda_{k-1}} \|x^k - x^{k-1}\|^2, \ \forall\ k \geq \tilde{k}. \quad (14)$$

Proof. Suppose by contradiction that there exists $\{k_j\}, k_j \to +\infty$ such that

$$f(x^{k_j}) - f(x^{k_j-1}) - \langle \nabla f(x^{k_j-1}), x^{k_j} - x^{k_j-1} \rangle > \frac{\eta_0}{\lambda_{k_j-1}} \|x_j^k - x^{k_j-1}\|^2$$

$$\implies \lambda_{k_j} = \eta_1 \frac{\|x^{k_j} - x^{k_j-1}\|^2}{f(x^{k_j}) - f(x^{k_j-1}) - \langle \nabla f(x^{k_j-1}), x^{k_j} - x^{k_j-1} \rangle},$$

which follows that

$$\lambda_{k_j} < \frac{\eta_1}{\eta_0} \lambda_{k_j-1}.$$

However, from Lemma 4.1 we have $\lim\limits_{k_j \to +\infty} \lambda_{k_j} = \lim\limits_{k_j \to +\infty} \lambda_{k_j-1} = \lim\limits_{k \to +\infty} \lambda_k = \lambda^*$, thus $\dfrac{\lambda^*}{\lambda^*} \leq \dfrac{\eta_1}{\eta_0} < 1$, a contradiction. □

Remark 3. From Lemmas 4.1 and 4.2 we derive that

(i) For all $k \geq 0$, $\lambda_k \geq \lambda_{\min} = \min\left\{\lambda_0, \frac{2\eta_1}{L}\right\}$;

(ii) The sequence $\{\lambda_k\}_{k \geq \tilde{k}}$ is increasing to λ^*, i.e., $\lambda_{\tilde{k}} \leq \lambda_k \leq \lambda_{k+1} \leq \lambda^*$ for all $k \geq \tilde{k}$.

Theorem 4.3. *Assume that the sequence* $\{x^k\}$ *is generated by MPG-NGD (Algorithm 2). Then, under Assumptions 1 and 2, we have*

(i) the sequence $\{f(x^k)\}_{k \geq \tilde{k}}$ *is decreasing to* $\tilde{f} \geq f^*$;

(ii) the sequence $\{x^k\}$ *converges to a stationary point of Problem* (P);

(iii) the computational complexity of Algorithm 2 is $O(\frac{1}{\varepsilon^2})$.

Proof. (i): Since $x^{k+1} = P_C(x^k - \lambda_k \nabla f(x^k))$ with $\lambda_k > 0$, by Lemma 2.2,

$$\langle x^k - \lambda_k \nabla f(x^k) - x^{k+1}, x^k - x^{k+1} \rangle \leq 0 \implies -\langle \nabla f(x^k), x^{k+1} - x^k \rangle \geq \frac{1}{\lambda_k} \|x^{k+1} - x^k\|^2. \quad (15)$$

From Lemma 4.2, we have for any $k \geq \tilde{k}$ that

$$f(x^{k+1}) - f(x^k) - \langle \nabla f(x^k), x^{k+1} - x^k \rangle \leq \frac{\eta_0}{\lambda_k} \|x^{k+1} - x^k\|^2. \quad (16)$$

Hence, for all $k \geq \tilde{k}$,

$$f(x^{k+1}) - f(x^k) \leq \frac{\eta_0}{\lambda_k} \|x^{k+1} - x^k\|^2 + \langle \nabla f(x^k), x^{k+1} - x^k \rangle \leq \frac{(\eta_0 - 1)\|x^{k+1} - x^k\|^2}{\lambda_k}. \quad (17)$$

Since $\eta_0 < 1$, the sequence $\{f(x^k)\}_{k\geq\tilde{k}}$ is decreasing and lower bounded by f^*, hence it converges to $\tilde{f} \geq f^*$.

(ii): From (17) we obtain

$$\|x^{k+1} - x^k\|^2 \leq \lambda_k \frac{f(x^k) - f(x^{k+1})}{1-\eta_0} \leq \lambda^* \frac{f(x^k) - f(x^{k+1})}{1-\eta_0}, \ \forall\ k \geq \tilde{k}. \tag{18}$$

Hence, for $K > \tilde{k}$,

$$\sum_{k=\tilde{k}}^{K-1} \|x^{k+1} - x^k\|^2 \leq \frac{\lambda^*}{1-\eta_0}\left(f(x^{\tilde{k}}) - f(x^K)\right) \underset{\text{by Remark 3}(ii)}{\leq} \frac{\lambda^*}{1-\eta_0}\left(f(x^{\tilde{k}}) - \tilde{f}\right)$$

$$\Longrightarrow \sum_{k=\tilde{k}}^{+\infty} \|x^{k+1} - x^k\|^2 \leq \frac{\lambda^*}{1-\eta_0}\left(f(x^{\tilde{k}}) - \tilde{f}\right), \tag{19}$$

and therefore $\lim_{k\to+\infty} \|x^{k+1} - x^k\| = 0$.

Setting $T = \{z \in C \mid f(z) \leq f(x^k),\ \forall\ k \geq \tilde{k}\}$. Then $X^* \subset T$. Taking some $z \in T$ and using $x^{k+1} = P_C(x^k - \lambda_k \nabla f(x^k))$, it follows from Lemma 2.2 that

$$\begin{aligned}
\|x^{k+1} - z\|^2 &= \|x^k - z\|^2 - \|x^{k+1} - x^k\|^2 + 2\langle x^{k+1} - x^k, x^{k+1} - z\rangle \\
&\leq \|x^k - z\|^2 - \|x^{k+1} - x^k\|^2 + 2\lambda_k\langle \nabla f(x^k), z - x^{k+1}\rangle \\
&= \|x^k - z\|^2 - \|x^{k+1} - x^k\|^2 + 2\lambda_k\langle \nabla f(x^k), z - x^k\rangle + 2\lambda_k\langle \nabla f(x^k), x^k - x^{k+1}\rangle \\
&\leq \|x^k - z\|^2 + 2\lambda_k\langle \nabla f(x^k), z - x^k\rangle + 2\lambda_k\langle \nabla f(x^k), x^k - x^{k+1}\rangle.
\end{aligned} \tag{20}$$

Since $f(z) - f(x^k) \leq 0$, we have $f(x^{k+1}) - f(x^k) \leq 0$ for all $k \geq \tilde{k}$. Moreover, since f is quasiconvex on C, then we get that

$$2\lambda_k\langle \nabla f(x^k), z - x^k\rangle \leq 0,\ \forall\ k \geq \tilde{k}, \tag{21}$$

and

$$2\lambda_k\langle \nabla f(x^k), x^k - x^{k+1}\rangle \geq 0,\ \forall\ k \geq \tilde{k}. \tag{22}$$

Using (20) and (21) we derive that

$$\|x^{k+1} - z\|^2 \leq \|x^k - z\|^2 + 2\lambda_k\langle \nabla f(x^k), x^k - x^{k+1}\rangle. \tag{23}$$

Moreover, by (14), for all $k \geq \tilde{k} - 1$,

$$\begin{aligned}
\langle \nabla f(x^k), x^k - x^{k+1}\rangle &\leq f(x^k) - f(x^{k+1}) + \frac{\eta_0}{\lambda_k}\|x^{k+1} - x^k\|^2 \qquad (24)\\
&\underset{\text{by (17)}}{\leq} f(x^k) - f(x^{k+1}) + \frac{\eta_0}{1-\eta_0}\left(f(x^k) - f(x^{k+1})\right) \\
&= \frac{f(x^k) - f(x^{k+1})}{1-\eta_0}. \qquad (25)
\end{aligned}$$

Hence,

$$2\lambda_k\langle\nabla f(x^k), x^k - x^{k+1}\rangle \leq 2\lambda_k \frac{f(x^k) - f(x^{k+1})}{1-\eta_0} \underset{\text{by Remark 3}(ii)}{\leq} \frac{2\lambda^*}{1-\eta_0}\left(f(x^k) - f(x^{k+1})\right), \tag{26}$$

for all $k \geq \tilde{k} - 1$. As a result, for any $K \geq \tilde{k}$, summing up (26) from $k = \tilde{k}$ to K,

$$\sum_{k=\tilde{k}}^{K} 2\lambda_k\langle\nabla f(x^k), x^k - x^{k+1}\rangle \leq 2\lambda^* \frac{f(x^{\tilde{k}}) - f(x^{K+1})}{1-\eta_0} \leq 2\lambda^* \frac{f(x^{\tilde{k}}) - \tilde{f}}{1-\eta_0}. \tag{27}$$

Consequently, tending K to infinity we obtain

$$\sum_{k=\tilde{k}}^{+\infty} 2\lambda_k\langle\nabla f(x^k), x^k - x^{k+1}\rangle \leq 2\lambda^* \frac{f(x^{\tilde{k}}) - \tilde{f}}{1-\eta_0} < +\infty. \tag{28}$$

Combining (22), (23) with (28) we derive that $\{\|x^k - z\|\}_{k\geq\tilde{k}}$ is quasi-Fejer convergent to T as Definition 2.5. Applying Lemma 2.9, $\{x^k\}$ is bounded, hence it has cluster points.

Now, because of Lemma 2.9, it remains to show that every cluster point of $\{x^k\}$ is a stationary point of (P) and belongs to T. Indeed, let $\overline{x}$ be a cluster point of $\{x^k\}$. Then there exists a subsequence $\{x^{k_i}\} \subset \{x^k\}$ such that $\lim\limits_{k_i\to+\infty} x^{k_i} = \overline{x}$. Take an arbitrary point $x \in C$. Then,

$$\begin{aligned}\|x^{k+1} - x\|^2 &= \|x^k - x\|^2 - \|x^{k+1} - x^k\|^2 + 2\langle x^{k+1} - x^k, x^{k+1} - x\rangle \\ &\leq \|x^k - x\|^2 - \|x^{k+1} - x^k\|^2 + 2\lambda_k\langle\nabla f(x^k), x - x^{k+1}\rangle. \end{aligned}\tag{29}$$

Put $k = k_i$ and $k_i \to +\infty$. Since $\lim\limits_{k\to+\infty}\|x^{k+1} - x^k\| = 0$ and f is continuously differentiable on C, we obtain

$$\langle\nabla f(\overline{x}), x - \overline{x}\rangle \geq 0, \ \forall\ x \in C. \tag{30}$$

Consequently, $\overline{x}$ is a stationary point of (P).

Because f is continuous, $\lim\limits_{k_i\to+\infty} f(x^{k_i}) = f(\overline{x})$. On the other hand, from (i), $\{f(x^k)\}_{k\geq\tilde{k}}$ converges decreasingly to $\tilde{f}$, thus $f(\overline{x}) = \lim\limits_{k\to+\infty} f(x^{k_i}) = \lim\limits_{k\to+\infty} f(x^k) = \tilde{f} \leq f(x^k)$ for all $k \geq \tilde{k}$. Then, $\overline{x} \in T$.

Next, we apply Lemma 2.9 to assert that $\{x^k\}$ converges to its cluster point which is just proved to be a stationary point of (P).

(iii): Using (17) again to derive that

$$\frac{\|x^{k+1}-x^k\|^2}{\lambda_k^2} \leq \frac{f(x^k)-f(x^{k+1})}{(1-\eta_0)\lambda_k} \underset{\text{by Remark 3}(ii)}{\leq} \frac{f(x^k)-f(x^{k+1})}{(1-\eta_0)\lambda_{\tilde{k}}}, \ \forall k \geq \tilde{k}. \tag{31}$$

Summing up (31) from $k=\tilde{k}$ to $K-1$, $K \geq \tilde{k}-1$, we get

$$\sum_{k=\tilde{k}}^{K} \frac{\|x^{k+1}-x^k\|^2}{\lambda_k^2} \leq \frac{f(x^{\tilde{k}})-f(x^K)}{(1-\eta_0)\lambda_{\tilde{k}}} \leq \frac{f(x^{\tilde{k}})-\tilde{f}}{(1-\eta_0)\lambda_{\tilde{k}}}$$

$$\Longrightarrow \min_{j=\tilde{k},\ldots,K-1} \frac{\|x^{k+1}-x^k\|}{\lambda_k} \leq \sqrt{\frac{f(x^{\tilde{k}})-\tilde{f}}{(K-\tilde{k})(1-\eta_0)\lambda_{\tilde{k}}}}.$$

Then, whenever the number of iterations is more than $\tilde{k} + \frac{f(x^{\tilde{k}})-\tilde{f}}{\varepsilon^2(1-\eta_0)\lambda_{\tilde{k}}}$, algorithm MPG-NGD terminates as the stopping criterion $\frac{\|x^{k+1}-x^k\|}{\lambda_k} \leq \varepsilon$ is satisfied. □

When f is pseudoconvex we have a stronger convergence result in the following corollary.

Corollary 4.4. *Under Assumptions 1 and 2, and in addition, f is pseudoconvex, then $\{x^k\}$ generated by MPG-NGD (Algorithm 2) converges to an optimal solution of (P).*

Remember that in [17], the authors established a sublinear convergence rate for the case where $C=\mathbb{R}^n$. In this paper, we extend these results to an arbitrary closed convex set $C \subseteq \mathbb{R}^n$. Specifically, we prove sublinear convergence of the sequence $\{x^k\}_{k\geq\tilde{k}}$ for convex objectives, and linear convergence for strongly convex objectives. These results are formally stated in the following theorem.

Theorem 4.5. *Suppose that Assumptions 1 and 2 hold. Then the sequence $\{x^k\}$, generated by MPG-NGD (Algorithm 2), satisfies that:*

(i) if f is convex, then

$$f(x^k)-f^* \leq O\left(\frac{1}{k}\right), \quad k \geq \tilde{k}; \tag{32}$$

(ii) if f is strongly convex with modulus $\sigma>0$ and $\eta_0 \leq \frac{1}{2}$, then $\{x^k\}$ linearly converges to an optimal solution of (P).

Proof. (i): By Theorem 4.3 and Corollary 4.4, $\{x^k\}$ converges to $x^* \in X^*$ and $\{f(x^k)\}$ converges to $\tilde{f}=f^*$ if f is convex. From (16), we have

$$f(x^{k+1}) \leq f(x^k) + \langle \nabla f(x^k), x^{k+1}-x^k\rangle + \frac{\eta_0}{\lambda_k}\|x^{k+1}-x^k\|^2. \tag{33}$$

Since f is convex, if we take $x^* \in X^*$, then

$$f(x^k) \leq f(x^*) + \langle \nabla f(x^k), x^k - x^* \rangle. \tag{34}$$

Combining (33) with (34) we get that

$$\begin{aligned} f(x^{k+1}) &\leq f(x^*) + \langle \nabla f(x^k), x^k - x^* \rangle + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{\eta_0}{\lambda_k} \|x^{k+1} - x^k\|^2 \\ &\leq f(x^*) + \langle \nabla f(x^k), x^{k+1} - x^* \rangle + \frac{\eta_0}{\lambda_k} \|x^{k+1} - x^k\|^2. \end{aligned} \tag{35}$$

Using Lemma 2.2, $\langle x^k - \lambda_k \nabla f(x^k) - x^{k+1}, x^* - x^{k+1} \rangle \leq 0$, thus

$$\begin{aligned} \langle \nabla f(x^k), x^{k+1} - x^* \rangle &\leq \frac{1}{\lambda_k} \langle x^{k+1} - x^k, x^* - x^{k+1} \rangle \\ &= \frac{\|x^k - x^*\|^2 - \|x^{k+1} - x^*\|^2 - \|x^{k+1} - x^k\|^2}{2\lambda_k}. \end{aligned} \tag{36}$$

Utilizing (35) and (36) we obtain for all $k \geq \tilde{k}$,

$$f(x^{k+1}) \leq f(x^*) + \frac{\|x^k - x^*\|^2 - \|x^{k+1} - x^*\|^2}{2\lambda_k} + \left(\eta_0 - \frac{1}{2}\right) \frac{\|x^{k+1} - x^k\|^2}{\lambda_k}. \tag{37}$$

Now, since $\lambda_k \geq \lambda_{\tilde{k}}$ for all $k \geq \tilde{k}$ (by Remark 3(ii)), it follows from (37) that

$$\begin{aligned} 2\lambda_{\tilde{k}} \left(f(x^{k+1}) - f(x^*)\right) &\leq 2\lambda_k \left(f(x^{k+1}) - f(x^*)\right) \\ &\leq \left(\|x^k - x^*\|^2 - \|x^{k+1} - x^*\|^2\right) + (2\eta_0 - 1) \|x^{k+1} - x^k\|^2. \end{aligned} \tag{38}$$

Summing up (38) from $k = \tilde{k}$ to $K - 1$, we get

$$\begin{aligned} 2\lambda_{\tilde{k}} \sum_{k=\tilde{k}}^{K-1} (f(x^{k+1}) - f(x^*)) &\leq \left(\|x^{\tilde{k}} - x^*\|^2 - \|x^K - x^*\|^2\right) + (2\eta_0 - 1) \sum_{k=\tilde{k}}^{K-1} \|x^{k+1} - x^k\|^2 \\ &\leq \|x^{\tilde{k}} - x^*\|^2 + (2\eta_0 - 1) \sum_{k=\tilde{k}}^{K-1} \|x^{k+1} - x^k\|^2 \\ &\underset{\text{by (19)}}{\leq} \max\left\{\|x^{\tilde{k}} - x^*\|^2, \|x^{\tilde{k}} - x^*\|^2 + \frac{(2\eta_0 - 1)\lambda^*}{1 - \eta_0} \left(f(x^{\tilde{k}}) - \tilde{f}\right)\right\} \\ &= M. \end{aligned} \tag{39}$$

On the other hand, $\{f(x^k)\}_{k \geq \tilde{k}}$ is decreasing, thus

$$\sum_{k=\tilde{k}}^{K-1} \left(f(x^{k+1}) - f(x^*)\right) \geq (K - \tilde{k}) \left(f(x^K) - f(x^*)\right). \tag{40}$$

Finally, using (39) and (40) we infer that

$$f(x^K) - f(x^*) \leq \frac{M}{K - \tilde{k}} \iff f(x^k) - f^* \leq \frac{M}{k - \tilde{k}} = O\left(\frac{1}{k}\right), \ \forall\ k \geq \tilde{k}.$$

(ii): Because of the σ-strong convexity of f, we update (34) by using (5)

$$f(x^k) \leq f(x^*) + \langle \nabla f(x^k), x^k - x^* \rangle - \frac{\sigma}{2}\|x^k - x^*\|^2. \tag{41}$$

Then for all $k \geq \tilde{k}$, the inequality (37) becomes

$$\begin{aligned} f(x^{k+1}) - f(x^*) &\leq \frac{\|x^k - x^*\|^2 - \|x^{k+1} - x^*\|^2}{2\lambda_k} + \left(\eta_0 - \frac{1}{2}\right)\frac{\|x^{k+1} - x^k\|^2}{\lambda_k} - \frac{\sigma}{2}\|x^k - x^*\|^2 \\ &\underset{\eta_0 \leq \frac{1}{2}}{\leq} \left(\frac{1}{2\lambda_k} - \frac{\sigma}{2}\right)\|x^k - x^*\|^2 - \frac{\|x^{k+1} - x^*\|^2}{2\lambda_k}. \end{aligned} \tag{42}$$

On the other hand, $f(x^{k+1}) - f(x^*) \geq 0$ for all $k \geq \tilde{k}$, then it follows from (42) that

$$\|x^{k+1} - x^*\|^2 \leq (1 - \sigma\lambda_k)\|x^k - x^*\|^2 \leq (1 - \sigma\lambda_{\tilde{k}})\|x^k - x^*\|^2, \ \forall\ k \geq \tilde{k},$$

which means $\{x^k\}_{k \geq \tilde{k}}$ converges linearly to x^*. □

5 Numerical experiments

In this section, the experiments were implemented in Python and executed on a personal computer with a AMD Ryzen 7 5800H 3.20 GHz processor and 16.0GB RAM. Codes are available at https://github.com/hoaiphamthi/MPG. We compare MPG-NGD (Algorithm 2) with PG-NGD (Algorithm 1) and related algorithms including Gradient Descent Adaptive (GDA) algorithm [17] and Projected Gradient Descent with Armijo's backtracking line search (PGB) [2] as follows.

Algorithm 3 (GDA) [17]

Step 0 (Initialization). Select $\lambda_0 > 0$ and $\sigma, \kappa \in (0, 1)$. Choose $x^0 \in C$, $x^1 = P_C(x^0 - \lambda_0 \nabla f(x^0))$, and set $k = 1$.

Step 1. If

$$f(x^k) \leq f(x^{k-1}) - \sigma\langle \nabla f(x^{k-1}), x^{k-1} - x^k \rangle$$

then

$$\lambda_k := \lambda_{k-1}$$

else

$$\lambda_k := \kappa\lambda_{k-1}.$$

Step 2. Compute $x^{k+1} = P_C(x^k - \lambda_k \nabla f(x^k))$.

Step 3. If $\frac{\|x^{k+1} - x^k\|}{\lambda_k} < \varepsilon$ **then** STOP, **else** set $k := k + 1$ and return to **Step 1**.

Algorithm 4 (PGB) [2]

Step 0 (Initialization). Select $\lambda_0 > 0$, $c \in (0,1)$, $\beta \in (0,1)$, and a tolerance $\varepsilon > 0$. Choose $x^0 \in C$, $x^1 = P_C(x^0 - \lambda_0 \nabla f(x^0))$, and set $k = 1$.

Step 1. Set $\lambda_k = \lambda_0$.

Step 2. While

$$f(P_C(x^k - \lambda_k \nabla f(x^k))) > f(x^k) - \frac{c}{\lambda_k}\|P_C(x^k - \lambda_k \nabla f(x^k)) - x^k\|^2$$

do $\lambda_k := \beta \lambda_k$
if $\lambda_k \leq 10^{-6}$ **then** BREAK.

Step 3. Compute $x^{k+1} = P_C(x^k - \lambda_k \nabla f(x^k))$.

Step 4. If $\frac{\|x^{k+1} - x^k\|}{\lambda_k} < \varepsilon$ **then** STOP, **else** set $k := k + 1$ and return to **Step 1**.

Parameter setting: All implemented algorithms use the same initial x^0 and λ_0. The stopping criterion is reached if $\dfrac{\|x^{k+1} - x^k\|}{\lambda_k} < \varepsilon = 10^{-6}$, or the number of iterations is over *max_iters = 50000*. Moreover, we set

- $\sigma = 0.1$ and $\kappa = 0.5$ for GDA (these parameters are chosen as suggested in [17]).
- $c = 0.1$ and $\beta = 0.5$ for PGB;
- $\varepsilon_{k-1} = \dfrac{0.1(\ln k)^{5.7}}{k^{1.1}}$ for all $k \geq 1$ and $\eta_0 = 0.45$, $\eta_1 = 0.49$ for PG-NGD and MPG-NGD.

Metrics for comparison include: the number of iterations (Iter.); the running time (Time(s)); the average of the stepsizes over all iterations (Stepsize).

The numerical results are presented by tables and figures.

- Tables of average results over 10 runs are reported corresponding to 10 randomly initial points. The best performances are emphasized by underlining and bold characters, while the worst ones are italicized.
- Figures illustrate the residual ($\|x^{k+1} - x^k\|/\lambda_k$) , the objective gap ($f(x^k) - f^*$), (where f^* denotes the minimum objective value over all iterations and all algorithms for each problem), and the stepsize (λ_k) across the iterations.

Tested instances: we conducted experiments on two problems: Supervised Feature Selection (SFS) (Example 1) with benchmark data and one synthetic example (Example 2) which is quasiconvex but does not satisfy condition (C_2).

Example 1. Supervised feature selection (SFS) is a fundamental problem in machine learning (see, e.g., [10, 11, 20]). Consider a data set with p features $\mathcal{F} = \{F_1, \ldots, F_p\}$ and N samples $\{(x_i, y_i)\}_{i=1}^N$, where $x_i = (x_{i1}, \ldots, x_{ip})^\top \in \mathbb{R}^p$ denotes the feature vector of the i-th sample and $y_i \in \{1, \ldots, m\}$ is the corresponding class label or target value. The goal of SFS is to select a subset of features that are highly relevant to the target $y = (y_1, \ldots, y_N)^\top$ while exhibiting minimal redundancy.

Following Wang et al. [20], feature redundancy is represented by a positive semi-definite matrix $Q \in \mathbb{R}^{p\times p}$, and feature relevance is measured by a vector $\rho = (\rho_1, \ldots, \rho_p)^\top$ with $\rho_i > 0$. The detailed computation of Q and ρ from a data set can be found in Wang et al. [20] and the references therein; in our implementation, the quantities involved in the computation of Q and ρ are estimated using the `NPEET`[1] library. Let $w = (w_1, \ldots, w_p)^\top$ denote the feature weight vector. Minimizing redundancy while maximizing relevance leads to the following fractional programming

$$\begin{aligned} \min \quad & \frac{w^\top Q w}{\rho^\top w} \\ \text{s.t.} \quad & e^\top w = 1, \\ & w \geq 0. \end{aligned} \tag{SFS}$$

It is observed that in Wang et al. [20], (SFS) is recognized as a pseudoconvex programming. However, we can show that it is a convex programming (cf. Lemma A.3 in Appendix). This problem belongs to the class of fractional programming which has received a lot of attention in literature, one can see [18] and the references therein for details. In our experiment, matrix Q and vector ρ are computed form data given by `UCI Machine Learning Repository`[2], which are listed in the following table.

Table 1: The data for Example 1

Dataset	p	N
Contraceptive Method Choice (cmc)	9	1472
Statlog German Credit Data (german_DP)	24	999
Statlog Heart (heart_DP)	12	269
ionosphere	34	350
Musk Version 1 (musk_DP)	167	475
Parkinsons (parkinsons_DP)	22	194
Soybean Large (soybean_DP)	34	264
Breast Cancer Wisconsin Diagnostic (wdbc_DP)	30	568
wine	13	177

The obtained results reported in Table 2 and Figures 1 and 2 indicate that PG-NGD (Algorithm 1) and MPG-NGD (Algorithm 2) achieve the lowest number of

[1] https://github.com/gregversteeg/NPEET
[2] https://archive.ics.uci.edu/datasets

iterations, the fastest execution times for almost cases. Although GDA and PGB also obtain optimal values comparable to those of PG-NGD (Algorithm 1) and MPG-NGD (Algorithm 2), they require significantly longer running times and more iterations. MPG-NGD (Algorithm 2) needs substantially fewer iterations than PG-NGD (Algorithm 1) but for some cases such as *inosophere, musk_DP, wine*, it costs little more running time than PG-NGD. This phenomenon is due to that MPG-NGD requires evaluating $f(x^k)$ at each iteration, whereas PG-NGD does not.

Table 2: The average results for Example 1

Data	**Metrics**	**PG-NGD**	**MPG-NGD**	**GDA**	**PGB**
	Iter.	34.8	**25.7**	*60.3*	33.8
cmc	Time(s)	**0.0022**	**0.0022**	0.0048	0.0055
	Stepsize	2.1758	**3.2707**	*1.5415*	2.5040
	Iter.	18282.6	**10868.1**	*44268.3*	32599.6
german_DP	Time(s)	1.0366	0.8345	3.3125	*11.6635*
	Stepsize	0.1007	**0.1721**	*0.0418*	0.0583
	Iter.	*19.5*	17.8	13	**10**
heart_DP	Time(s)	0.013	0.0015	**0.0011**	0.0021
	Stepsize	*1.01*	1.4417	2.**0536**	1.9318
	Iter.	12.8	**11.8**	*30.1*	*30.1*
ionosphere	Time(s)	**0.0009**	0.0011	0.0026	0.0023
	Stepsize	21.2582	**28.2859**	*10*	*10*
	Iter.	15.0	**11.9**	*24.8*	18.4
musk_DP	Time(s)	**0.0063**	0.0104	0.0205	0.0194
	Stepsize	3.2754	**5.8359**	*0.0048*	5.4471
	Iter.	26.4	**19.5**	*38.1*	31.7
parkinsons_DP	Time(s)	0.0017	**0.0016**	0.0030	0.0120
	Stepsize	*0.3931*	0.5518	0.5819	**0.5987**
	Iter.	66.2	**40**	*135.2*	44.1
soybean_DP	Time(s)	0.0042	**0.0034**	0.0109	0.0099
	Stepsize	0.8346	**1.3135**	*0.4685*	1.3028
	Iter.	19.6	**13.9**	*24.2*	22.6
wdbc_DP	Time(s)	0.0014	**0.0013**	0.0021	*0.0059*
	Stepsize	*0.6459*	0.9193	1.1621	**1.5012**
	Iter.	33.1	**32.7**	56.7	48.7
wine	Time(s)	**0.0025**	0.0030	0.0052	0.0124
	Stepsize	0.9727	**1.3549**	*0.7428*	1.0605

Example 2. We consider the following problem

$$\begin{aligned} \min \quad & f(x) = \frac{2\ell + \sum_{i=1}^{2\ell}(x_i^2 + \sin(x_i)) - a^\top x}{1 + 2\ell + a^\top x} \\ \text{s.t.} \quad & x \in C, \end{aligned} \tag{43}$$

where $\ell \in \mathbb{N}$, $\ell > 1$, $C = \{x \in \mathbb{R}^{2\ell}_+ \mid e^\top x = 2\ell\}$, $a \in \mathbb{R}^{2\ell}$ with components:

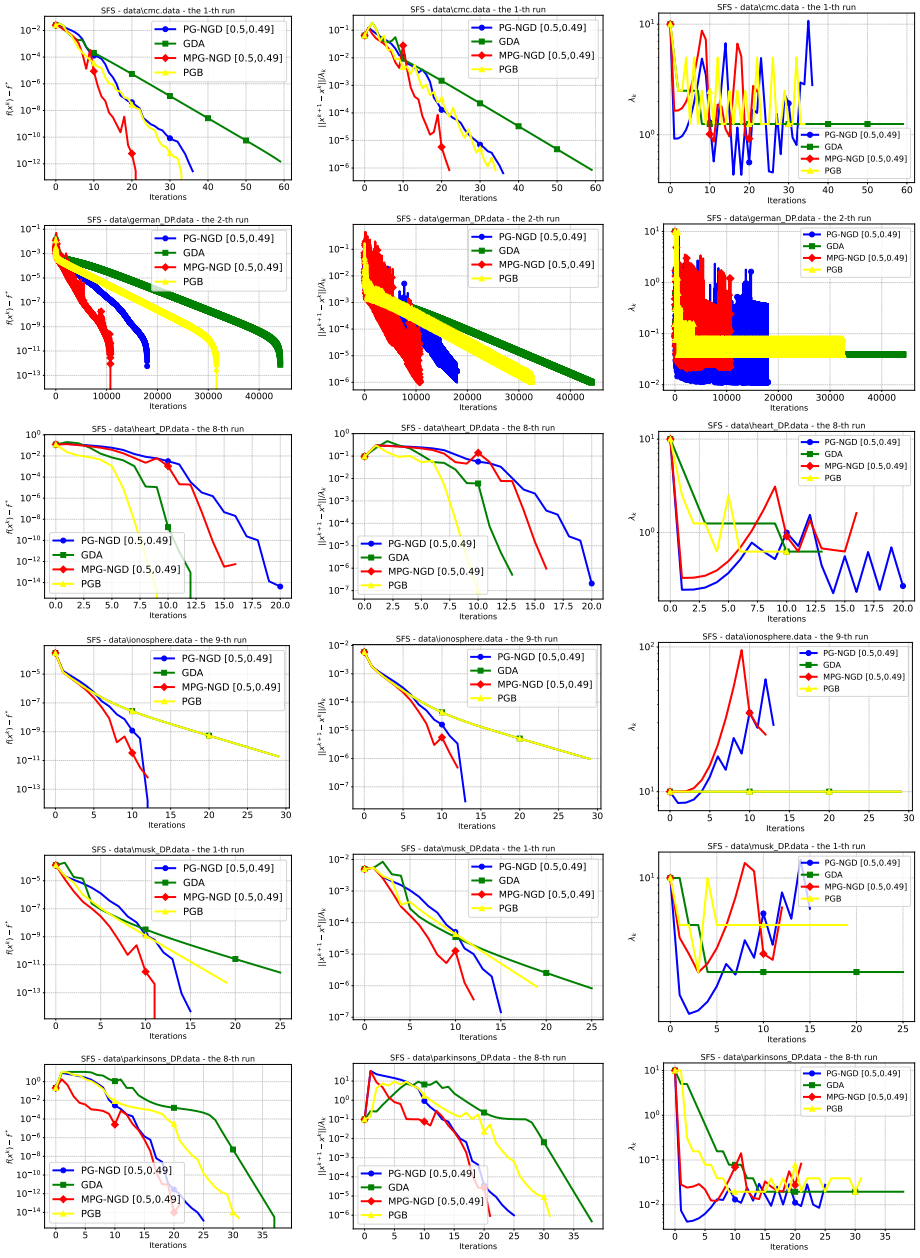

Figure 1: Numerical results for (SFS) with data in Table 1

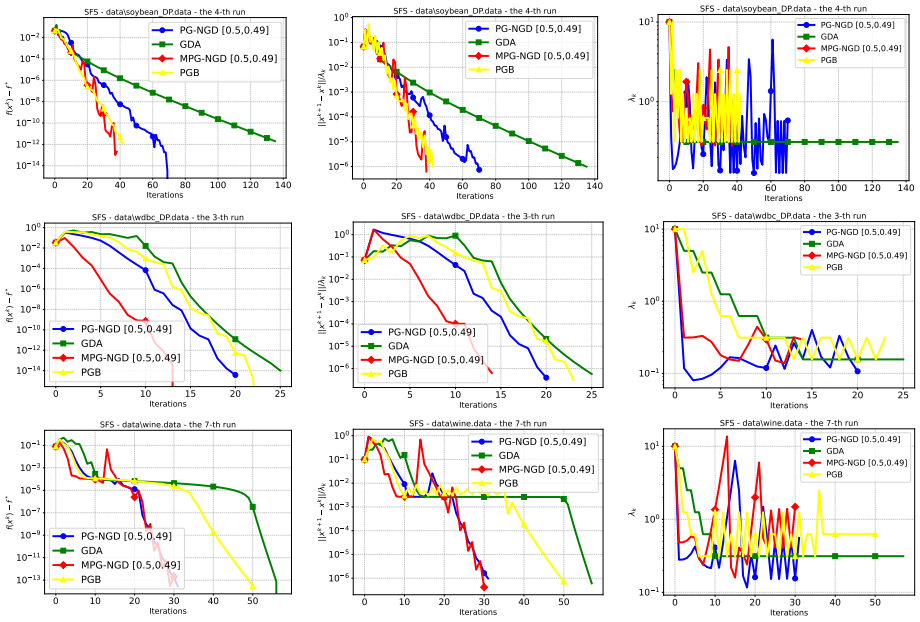

Figure 2: Numerical results for (SFS) with data in Table 1

$a_1 = 1$, a_{2i-1} $(i = 2, ..., \ell)$ is generated randomly by a uniform distribution in $[-1, 1]$, and $a_{2i} = -a_{2i-1}$ for $i = 1, \ldots, \ell$.

It is evident that C is convex and f is quasiconvex (by Lemma 2.7) but does not satisfy condition (C_2) (see Lemma A.4 in Appendix). Table 3 compares PG-NGD (Algorithm 1), MPG-NGD (Algorithm 2), GDA and PGB for solving Example 2 for various sizes ℓ and initial stepsizes λ_0. The numerical results in Table 3 show that MPG-NGD generally performs best overall, followed by PG-NGD, PGB and GDA. These results also indicate that PGB and GDA are sensitive to initial stepsize λ_0, whereas adaptive methods (PG-NGD, MPG-NGD) remain more independently with λ_0.

In summarize, numerical experiments for both Example 1 and Example 2 reveal that MPG-NGD typically uses the fewest number of iterations for almost cases and therefore can speed up the processing time of proximal gradient method significantly. This algorithm also produces the largest average stepsize for most of instances that partially accounts for MPG-NGD's rapid procession.

6 Conclusion

We have contributed to the development of projected gradient methods for nonconvex (quasiconvex) optimization over closed convex sets by introducing a new adaptive algorithm (MPG-NGD). From some fixed iteration, MPG-NGD generates a non-decreasing stepsize sequence that converges to a positive limit,

Table 3: The average results for Example 2

Size		Metrics	PG-NGD	MPG-NGD	GDA	PGB
$\lambda_0 = \ell/2$	$2\ell = 500$	Iter.	15	**12**	*92*	45
		Time(s)	0.0020	**0.0017**	*0.0132*	0.0055
		Stepsize	290.4713	**364.3283**	*63.172*	125
	$2\ell = 1000$	Iter.	15	**12**	*90*	43
		Time(s)	0.0025	**0.0022**	*0.0170*	0.0071
		Stepsize	585.4713	**730.9292**	*126.3736*	250
	$2\ell = 5000$	Iter.	14	**11**	*81*	39
		Time(s)	0.0078	**0.0072**	*0.0535*	0.0222
		Stepsize	2799.9209	**3115.2673**	*632.6220*	1250
	$2\ell = 7000$	Iter.	14	**11**	*79*	38
		Time(s)	0.0108	**0.0103**	*0.0733*	0.0300
		Stepsize	4029.5018	**4360.0438**	*885.9375*	1750
$\lambda_0 = \ell$	$2\ell = 500$	Iter.	14	**10**	*44*	20
		Time(s)	0.0018	**0.0014**	*0.0063*	0.0024
		Stepsize	327.1520	**425.9117**	*127.7778*	250
	$2\ell = 1000$	Iter.	13	**10**	*42*	20
		Time(s)	0.0022	**0.0018**	*0.0075*	0.0030
		Stepsize	637.1711	**854.3854**	*255.8140*	500
	$2\ell = 5000$	Iter.	13	**10**	*38*	18
		Time(s)	0.0071	**0.0064**	*0.0250*	0.0104
		Stepsize	3177.0368	**4269.9434**	*1282.0513*	2500
	$2\ell = 7000$	Iter.	13	**10**	*37*	17
		Time(s)	0.0099	**0.0091**	*0.0345*	0.0139
		Stepsize	4440.7554	**5975.4362**	*1796.0526*	3500
$\lambda_0 = 2\ell$	$2\ell = 500$	Iter.	12	**7**	*19*	8
		Time(s)	0.0016	**0.0011**	*0.0028*	**0.0011**
		Stepsize	378.5017	**510.0116**	*262.5*	500
	$2\ell = 1000$	Iter.	11	**7**	*18*	**7**
		Time(s)	0.0018	**0.0014**	*0.0033*	0.0016
		Stepsize	732.1521	**1080.5062**	*526.3158*	1000
	$2\ell = 5000$	Iter.	11	**7**	*16*	**7**
		Time(s)	0.0063	0.0048	*0.0110*	**0.0045**
		Stepsize	3652.7888	**5394.5878**	*2647.0588*	5000
	$2\ell = 7000$	Iter.	11	**7**	*16*	**7**
		Time(s)	0.0085	0.0066	*0.0154*	**0.0062**
		Stepsize	5105.5929	**7527.9008**	*3705.8824*	7000
$\lambda_0 = 4\ell$	$2\ell = 500$	Iter.	*12*	**8**	**8**	**8**
		Time(s)	*0.0017*	0.0014	**0.0013**	0.0016
		Stepsize	*391.4495*	628.5408	555.5556	**666.6667**
	$2\ell = 1000$	Iter.	*12*	**7**	8	*9*
		Time(s)	0.0020	**0.0013**	0.0016	*0.0021*
		Stepsize	*793.9486*	1299.9517	1111.1111	**1300**
	$2\ell = 5000$	Iter.	*12*	**7**	**7**	8
		Time(s)	*0.0068*	**0.0049**	0.0052	0.0074
		Stepsize	*3962.3667*	6511.2203	5625	**6666.6667**
	$2\ell = 7000$	Iter.	*12*	**6**	7	8
		Time(s)	0.0092	**0.0057**	0.0073	*0.0105*
		Stepsize	*5539.4543*	8953.9939	7875	**9333.3333**

enabling consistently longer steps and higher speed. The algorithm has rigorous convergence guarantees: convergence to stationary points for quasiconvex objectives, to optimal solutions for pseudoconvex objectives, and with sublinear or linear rates for convex and strongly convex objectives, respectively. Numerical experiments on benchmark and synthetic data demonstrate the efficiency and competitive performance of MPG-NGD against state-of-the-art alternatives. Future work may explore the convergence of MPG-NGD, relaxing the global Lipschitz assumption on the gradient of the quasiconvex objective function f.

7 Declarations

7.1 Availability of Supporting Data

The Python codes of all numerical results in Section 5 are available at https://github.com/hoaiphamthi/MPG.

7.2 Author Contributions

All authors contributed equally to the study conception, design and implementation and wrote and corrected the manuscript.

7.3 Competing Interests

There are no conflicts of interest or competing interests related to this manuscript.

7.4 Funding

The author (Felipe Lara) was partially supported by ANID–Chile under project Fondecyt Regular 1241040.

References

[1] K. J. Arrow and A. C. Enthoven. Quasiconcave programming. *Econometrica*, 29(4):779–800, 1961.

[2] A. Beck. *Introduction to Nonlinear Optimization: Theory, Algorithms, and Applications with MATLAB*. Society for Industrial and Applied Mathematics, USA, 2014.

[3] R. Burachik, L. M. Graña Drummond, A. N. Iusem, and B. F. Svaiter. Full convergence of the steepest descent method with inexact line searches. *Optimization*, 59(1):137–146, 2010.

[4] A. Cambini and L. Martein. *Generalized convexity and optimization: Theory and applications*. Springer Berlin, Germany, 1st edition, 2008.

[5] S.-M. Grad, F. Lara, and R. T. Marcavillaca. Strongly quasiconvex functions: What we know (so far). *Journal of Optimization Theory and Applications*, 2025.

[6] H. J. Greenberg and W. P. Pierskalla. A review of quasi-convex functions. *Operations Research*, 19(7), Nov.–Dec. 1971.

[7] N. Hadjisavvas and F. Lara. Characterizations of strongly quasiconvex functions. *arXiv:2509.21580*, 2025.

[8] P.T. Hoai, N.T. Vinh, and N.P.H. Chung. A novel stepsize for gradient descent method. *Operations Research Letters*, 53:107072, 2024.

[9] A. Iusem, F. Lara, R. T. Marcavillaca, and L. H. Yen. A two-steps proximal point algorithm for nonconvex equilibrium problems with applications to fractional programming. *Journal of Global Optimization*, 90(3):755–779, 2024.

[10] K. Kampa, S. Mehta, C. A. Chou, W. A. Chaovalitwongse, and T. J. Grabowski. Sparse optimization in feature selection: application in neuroimaging. *Journal of Global Optimization*, 59:439–457, 2014.

[11] U. M. Khaire and R. Dhanalakshmi. Stability of feature selection algorithm: A review. *Journal of King Saud University - Computer and Information Sciences*, 34(4):1060–1073, 2022.

[12] F. Lara. On strongly quasiconvex functions: Existence results and proximal point algorithms. *Journal of Optimization Theory and Applications*, 192:891–911, 2022.

[13] F. Lara, R. T. Marcavillaca, and P. T. Vuong. Characterizations, dynamical systems and gradient methods for strongly quasiconvex functions. *Journal of Optimization Theory and Applications*, 206:60, 2025.

[14] P. Latafat, A. Themelis, and P. Patrinos. On the convergence of adaptive first order methods: proximal gradient and alternating minimization algorithms. In .A Abate, .M Cannon, .K Margellos, and .A Papachristodoulou, editors, *Proceedings of Machine Learning Research*, volume 242, pages 197–208, UK, 2024.

[15] P. Latafat, A. Themelis, L. Stella, and P. Patrinos. Adaptive proximal algorithms for convex optimization under local lipschitz continuity of the gradient. *Mathematical Programming*, 213:433–471, 2025.

[16] Y. Malitsky and K. Mishchenko. Adaptive proximal gradient method for convex optimization. In *NeurIPS*, 2024.

[17] T. N. Thang and T. N. Hai. Self-adaptive algorithms for quasiconvex programming and applications to machine learning. *Computational and Applied Mathematics*, 43(249), 2024.

[18] H. Tuy, P. T. Thach, and H. Konno. Optimization of polynomial fractional functions. *Journal of Global Optimization*, 29:19–44, 2004.

[19] A. A. Vladimirov, Yu. E. Nesterov, and Ju. N. Chekanov. On uniformly quasi-convex functionals. *Moscow University Computational Mathematics and Cybernetics*, (4):19–30, 1978. Translated from Vestnik Moskovskogo Universiteta, Seriya XV.

[20] Y. Wang, X. Li, and J. Wang. A neurodynamic optimization approach to supervised feature selection via fractional programming. *Neural Networks*, 2021.

Appendix A

Lemma A.1. *With* $f(x) = \left(\frac{1}{5x^2+1} - 1\right)^2$ *and* $C = \mathbb{R}$, *Problem* (P) *satisfies Assumption 1 and Assumption 2. However,* f *does not satisfy condition* (C_2).

Proof. Observing that f is even since

$$f(x) = \left(\frac{1}{5x^2+1} - 1\right)^2 = \frac{25x^4}{(5x^2+1)^2}.$$

Moreover, $f'(x) = \frac{100x^3}{(5x^2+1)^3}$. Hence, $f'(x) \geq 0$ for $x \geq 0$ and $f'(x) \leq 0$ for $x \leq 0$. Therefore, f is increasing on $[0, +\infty)$ and decreasing on $(-\infty, 0]$ and thus quasiconvex by [4, Theorem 2.5.1]. We observe that f' is continuously differentiable and

$$f''(x) = \frac{300x^2(1-5x^2)}{(5x^2+1)^4}$$

is bounded for all $x \in \mathbb{R}$. Therefore, f has a globally Lipschitz gradient. We now show that this function does not satisfy condition (C_2). Based on (C_2), by setting $v = 1$ and $u = 0$, we obtain the function

$$g_{uv}(t) = f'(t) = \frac{100t^3}{(5t^2+1)^3}.$$

We choose $t_1 = 0.1$, $t_2 = 0.9$, and $t_3 = \frac{1}{2}(t_1 + t_2) = 0.5$, and obtain $g_{uv}(t_1) = 0.086$, $g_{uv}(t_2) = 0.566$, and $g_{uv}(t_3) = 1.097$ and hence $g_{uv}(t_3) > \max\{g_{uv}(t_1), g_{uv}(t_2)\}$. Therefore, $g_{uv}(t)$ is not quasiconvex on $[0, 1]$ meaning that condition (C_2) is not satisfied.

□

Lemma A.2. *Problem* (P), *with the objective function* $f(x) = \|x\|^2 + 3\sin^2(\|x\|)$ *over the set* $C = \mathbb{R}^n$, *satisfies Assumption 1 and Assumption 2. Additionally,* f *is strongly quasiconvex but does not satisfy condition* (C_2).

Proof. This function is an extension of the one mentioned in [5]. Clearly, $f(x) > -\infty$ and an optimal solution always exists (satisfying the coercivity condition), hence Assumption 1 is always satisfied.
To show the strong quasiconvexity of f, we intend to prove that f is continuously differentiable and there exists $\gamma > 0$ such that

$$f(x) \leq f(y) \implies \langle \nabla f(y), (y - x) \rangle \geq \frac{\gamma}{2}\|x - y\|^2 \quad \text{for all } x, y \in \mathbb{R}^n.$$

Firstly, we show that f is continuously differentiable on $\mathbb{R}^n$. Obviously, f is continuously differentiable on $\mathbb{R}^n \backslash \{\mathbf{0}\}$ with

$$\nabla f(x) = 2x + 6\sin(\|x\|)\cos(\|x\|)\frac{x}{\|x\|} = 2x + \frac{3\sin(2\|x\|)x}{\|x\|} \text{ for all } x \neq \mathbf{0}. \quad (44)$$

For $x = \mathbf{0}$, we have

$$\begin{aligned}\frac{\partial f}{\partial x_i}(\mathbf{0}) &= \lim_{\Delta x_i \to 0} \frac{f(0,\dots,\Delta x_i,\dots,0) - f(0,\dots,0)}{\Delta x_i} \\ &= \lim_{\Delta x_i \to 0} \frac{(\Delta x_i)^2 + 3\sin^2(|\Delta x_i|)}{\Delta x_i} = 0, \quad i = 1,\dots,n. \end{aligned} \tag{45}$$

From (44) and (45), we obtain that for all $\epsilon > 0$ small enough and $x \neq \mathbf{0}$, $\|x - \mathbf{0}\| < \delta = \dfrac{\epsilon}{8}$,

$$0 \leq \left| \frac{\partial f}{\partial x_i}(x) - \frac{\partial f}{\partial x_i}(\mathbf{0}) \right| = \left| 2x_i + \frac{3x_i \sin(2\|x\|)}{\|x\|} \right| \leq 8|x_i| \leq 8\|x\| \leq \epsilon.$$

This implies $\frac{\partial f}{\partial x_i}(x)$ is continuous at $\mathbf{0}$ for all $i = 1,\dots,n$ meaning that f is continuously differentiable at $\mathbf{0}$ and hence on $\mathbb{R}^n$.

In the following, based on Lemma 2.6 we will show that there exists $\gamma > 0$ such that

$$f(x) \leq f(y) \implies \langle \nabla f(y), (y - x) \rangle \geq \frac{\gamma}{2}\|x - y\|^2 \quad \forall\, x, y \in \mathbb{R}^n.$$

Indeed, considering the function

$$p(z) = z^2 + 3\sin^2(z), \ z \geq 0,$$

whose derivative is

$$p'(z) = 2z + 6\sin(z)\cos(z) = 2z + 3\sin(2z).$$

Clearly, $p'(z) \geq 0$ for all $z \geq 0$, which means that p is strictly increasing on the interval $[0, +\infty)$. Therefore, if $p(z_1) \leq p(z_2)$ with $z_1, z_2 \geq 0$, then $z_1 \leq z_2$. Applying this to the function f, from $f(x) \leq f(y)$, or equivalently $p(\|x\|) \leq p(\|y\|)$, it follows that $\|x\| \leq \|y\|$. We have

$$\langle \nabla f(y), y - x \rangle = \left(2 + \frac{3\sin(2\|y\|)}{\|y\|}\right) \langle y, y - x \rangle = \left(2 + \frac{3\sin(2\|y\|)}{\|y\|}\right) (\|y\|^2 - \langle y, x \rangle).$$

Moreover, combining this with $\|y\| \geq \|x\|$, we observe that $\frac{\sin(2\|y\|)}{\|y\|} > -\frac{1}{2}$ and

$$\|y\|^2 - \langle y, x \rangle = \frac{1}{2}(\|y - x\|^2 + \|y\|^2 - \|x\|^2) \geq \frac{1}{2}\|x - y\|^2.$$

From these inequalities, it follows that

$$\langle \nabla f(y), y - x \rangle \geq \frac{1}{4}\|x - y\|^2 \geq \frac{\gamma}{2}\|x - y\|^2 \quad \text{with } \gamma \in \left(0, \frac{1}{2}\right].$$

Hence, we obtain

$$f(x) \leq f(y) \implies \langle \nabla f(y), y - x \rangle \geq \frac{\gamma}{2}\|x - y\|^2 \quad \text{with } \gamma \in \left(0, \frac{1}{2}\right].$$

Therefore, f is γ- strongly quasiconvex on $\mathbb{R}^n$ with $\gamma \in (0, \frac{1}{4}]$ by Lemma 2.6.

Secondly, we show that the function f satisfies Assumption 2. We observe that ∇f is differentiable on $\mathbb{R}^n$. Indeed, it is clearly differentiable for all $x \neq 0$. For $x = 0$, we have

$$\begin{aligned}\lim_{h\to 0} \frac{\|\nabla f(h+0) - \nabla f(0) - 8h\|}{\|h\|} &= \lim_{h\to 0} \frac{\left\|3h\left(\frac{\sin(2\|h\|)}{\|h\|} - 2\right)\right\|}{\|h\|} \\ &= \lim_{h\to 0} 3\left\|\left(\frac{\sin(2\|h\|)}{\|h\|} - 2\right)\right\| = 0.\end{aligned}$$

Therefore, ∇f is differentiable on $\mathbb{R}^n$, and moreover, $\nabla^2 f(0) = 8I$. Next, we consider the Hessian matrix of f, which is given by

$$\nabla^2 f(x) = \left(2 + \frac{3\sin(2\|x\|)}{\|x\|}\right) I + \frac{3(2\|x\|\cos(2\|x\|) - \sin(2\|x\|))}{\|x\|^3} xx^\top, \quad x \neq 0.$$

Clearly, $\nabla^2 f$ is continuous at $x = 0$. Hence, $\nabla^2 f$ is continuous on $\mathbb{R}^n$, and therefore, ∇f is continuously differentiable. Moreover, we have

$$\begin{aligned}\|\nabla^2 f(x)\| &\leq \left|2 + \frac{3\sin(2\|x\|)}{\|x\|}\right| + \left|6\cos(2\|x\|) - 3\frac{\sin(2\|x\|)}{\|x\|}\right| \\ &\overset{\text{(at least)}}{\leq} 20, \ \forall\, x \in \mathbb{R}^n.\end{aligned}$$

Therefore, $\nabla^2 f(x)$ is bounded, and hence, f has a globally Lipschitz gradient on C.

Thirdly, we prove f does not match the condition (C_2). Indeed, choosing $u = (0, 0, \ldots, 0)^\top$ and $v = (1, 1, \ldots, 1)^\top$ then

$$g_{uv}(t) = nt\left(2 + \frac{3\sin(2nt)}{nt}\right) = 2nt + 3\sin(2nt).$$

We choose $t_1 = \frac{\pi}{2n}$, $t_2 = \frac{\pi}{8n}$ and $t_3 = \frac{1}{3}t_1 + \frac{2}{3}t_2 = \frac{\pi}{4n}$. Then,

$$g_{uv}(t_1) = \pi, \quad g_{uv}(t_2) = \frac{\pi}{4} + \frac{3\sqrt{2}}{2}, \quad g_{uv}(t_3) = \frac{\pi}{2} + 3.$$

Clearly, $g_{uv}(t_3) > \max\{g_{uv}(t_1), g_{uv}(t_2)\}$ for all $n \geq 2$. For $n = 1$, we choose $t_1 = 0.9$, $t_2 = 0.98$, $t_3 = 0.94$, and we also have $g_{uv}(t_3) > \max\{g_{uv}(t_1), g_{uv}(t_2)\}$. Therefore, the function $g_{uv}(t)$ is not quasiconvex on $[0, 1]$, or in other words, it does not satisfy condition (C_2). □

Lemma A.3. *The objective function in Example 1 is convex on its corresponding constraint set.*

Proof. We recall the problem in Example 1 as follows:

$$\begin{aligned} \min \quad & f(w) = \frac{w^\top Q w}{\rho^\top w} \\ \text{subject to} \quad & e^\top w = 1, \\ & w \geq 0. \end{aligned}$$

where Q is a positive semi-definite matrix and $\rho > 0$. The epigraph of f is defined as $\text{epi}(f) = \{(w,t) \in \mathbb{R}^n \times \mathbb{R} \mid f(w) \leq t\}$. We aim to prove that $\text{epi}(f)$ is convex. Indeed, taking $(w_1,t_1),(w_2,t_2) \in \text{epi}(f)$ and $\lambda \in [0,1]$, then defining

$$w_\lambda = \lambda w_1 + (1-\lambda) w_2, \tag{46}$$
$$t_\lambda = \lambda t_1 + (1-\lambda) t_2. \tag{47}$$

By definition of $\text{epi}(f)$, we have

$$f(w_1) \leq t_1 \implies w_1^T Q w_1 \leq t_1\, p^T w_1, \tag{48}$$
$$f(w_2) \leq t_2 \implies w_2^T Q w_2 \leq t_2\, p^T w_2. \tag{49}$$

Now, considering

$$w_\lambda^T Q w_\lambda = (\lambda w_1 + (1-\lambda) w_2)^T Q (\lambda w_1 + (1-\lambda) w_2) \tag{50}$$
$$= \lambda^2 w_1^T Q w_1 + 2\lambda(1-\lambda) w_1^T Q w_2 + (1-\lambda)^2 w_2^T Q w_2. \tag{51}$$

Using (48) and (49), we obtain

$$w_\lambda^T Q w_\lambda \leq \lambda^2 t_1 p^T w_1 + 2\lambda(1-\lambda) w_1^T Q w_2 + (1-\lambda)^2 t_2 p^T w_2. \tag{52}$$

On the other hand,

$$\begin{aligned} t_\lambda\, p^T w_\lambda &= (\lambda t_1 + (1-\lambda) t_2)(\lambda p^T w_1 + (1-\lambda) p^T w_2) \\ &= \lambda^2 t_1 p^T w_1 + \lambda(1-\lambda) t_1 p^T w_2 + \lambda(1-\lambda) t_2 p^T w_1 + (1-\lambda)^2 t_2 p^T w_2. \end{aligned} \tag{53}$$

Subtracting (52) to (53) side by side, we get that

$$\begin{aligned} & w_\lambda^T Q w_\lambda - t_\lambda\, p^T w_\lambda \leq 2\lambda(1-\lambda) w_1^T Q w_2 - \lambda(1-\lambda) t_1 p^T w_2 - \lambda(1-\lambda) t_2 p^T w_1 \\ \iff & w_\lambda^T Q w_\lambda - t_\lambda\, p^T w_\lambda \leq \lambda(1-\lambda)\big(2 w_1^T Q w_2 - t_1 p^T w_2 - t_2 p^T w_1\big). \end{aligned} \tag{54}$$

Since Q is positive semidefinite, by the Cauchy-Schwarz inequality we have

$$w_1^T Q w_2 \leq \sqrt{(w_1^T Q w_1)(w_2^T Q w_2)}.$$

Combining this with (48) and (49), it follows that

$$w_1^T Q w_2 \leq \sqrt{t_1 p^T w_1 \cdot t_2 p^T w_2} \leq \frac{t_1 p^T w_2 + t_2 p^T w_1}{2}.$$

Substituting into (54), we obtain $w_\lambda^T Q w_\lambda \le t_\lambda\, p^T w_\lambda$. Clearly, $p^T w_\lambda > 0$ since the constraint set is convex and w_λ is a convex combination of w_1 and w_2. Therefore, we conclude

$$\frac{w_\lambda^T Q w_\lambda}{p^T w_\lambda} \le t_\lambda, \forall \lambda \in [0,1]$$

which shows that $(w_\lambda, t_\lambda) \in \text{epi}(f)$, or equivalent to, $\text{epi}(f)$ is convex. Hence, this derives the convexity of f on C. □

Lemma A.4. *Problem* (P)*, with the objective function and feasible set identical to those in Example 2, satisfies Assumptions 1 and 2 but f does not satisfy condition* (C_2).

Proof. Setting $n = 2\ell$,

$$N(x) = n + \sum_{i=1}^{n}(x_i^2 + \sin(x_i)) - a^\top x, \quad \text{and} \quad D(x) = 1 + n + a^\top x.$$

We have $D(x) \ge 1 + n - e^T x = 1 > 0$, so $D(x)$ is a positive affine function. Moreover, $N(x)$ is convex since its Hessian is positive definite for all $x \in C$. Hence, by Lemma 2.7, the function $f(x) = \dfrac{N(x)}{D(x)}$ is quasiconvex on the constraint set C. Furthermore, C is compact and f is continuous on C. It follows that problem (P) always admits an optimal solution, and Assumption 1 is satisfied. Moreover, clearly, ∇f is continuously differentiable and $\nabla^2 f$ is continuous on C. Hence, $\|\nabla^2 f\|$ is a continuous function on the compact set C and therefore is bounded on C. Consequently, f has a globally Lipschitz gradient on C. We have

$$\begin{aligned}
\nabla f(x) &= \frac{D(x)\nabla N(x) - N(x)\nabla D(x)}{D^2(x)} \\
&= \frac{D(x)\begin{pmatrix} 2x_1 + \cos(x_1) - a_1 \\ \cdots\cdots\cdots\cdots\cdots \\ 2x_n + \cos(x_n) - a_n \end{pmatrix} - N(x)\begin{pmatrix} a_1 \\ \cdots \\ a_n \end{pmatrix}}{D^2(x)} \\
&= \frac{1}{D^2(x)}\begin{pmatrix} D(x)(2x_1 + \cos(x_1) - a_1) - a_1 N(x) \\ \cdots\cdots\cdots\cdots\cdots \\ D(x)(2x_n + \cos(x_n) - a_n) - a_n N(x) \end{pmatrix}.
\end{aligned}$$

Following the condition (C_2), the function $g_{uv}(t)$ is

$$g_{uv}(t) = \langle \nabla f(u + t(v-u)), v - u\rangle, \quad t \in [0,1], \quad u, v \in C.$$

We choose $u = (n, 0, \ldots, 0)^\top$, $v = (1, 1, \ldots, 1)^\top$. Then,

$$x^0 = u + t(v-u) = (n + t - nt,\, t, \ldots, t), \quad t \in [0,1].$$

At this point, the $D(x)$ and $N(x)$ are given by

$$\begin{aligned} D(x^0) &= 1 + n + a_1(n + t - nt) + a_2 t + a_3 t + \cdots + a_n t \\ &= 1 + n + a_1 n(1 - t) + t \underbrace{(a_1 + a_2 + \cdots + a_{n-1} + a_n)}_{=0} \\ &= 1 + 2n - nt, \end{aligned}$$

and

$$\begin{aligned} N(x^0) &= n + (n + t - nt)^2 + (n - 1)t^2 + \sin(n + t - nt) + (n - 1)\sin(t) \\ &\quad - a_1(n + t - nt) - a_2 t - \cdots - a_n t \\ &= nt + (n + t - nt)^2 + (n - 1)t^2 + \sin(n + t - nt) + (n - 1)\sin(t). \end{aligned}$$

From this we obtain

$$\begin{aligned} g_{uv}(t) &= \frac{1}{D^2(x^0)} \left\langle \begin{pmatrix} (2n + 2t - 2nt + \cos(n + t - nt))D(x^0) - a_1(N(x^0) + D(x^0)) \\ (2t + \cos(t))D(x^0) - a_2(N(x^0) + D(x^0)) \\ \cdots\cdots\cdots\cdots\cdots \\ (2t + \cos(t))D(x^0) - a_n(N(x^0) + D(x^0)) \end{pmatrix}, \begin{pmatrix} 1 - n \\ 1 \\ \cdots \\ 1 \end{pmatrix} \right\rangle \\ &= \frac{1}{D^2(x^0)} \Big[(1 - n)(2n + 2t - 2nt + \cos(n + t - nt))D(x^0) + n(N(x^0) + D(x^0)) \\ &\quad + (n - 1)(2t + \cos(t))D(x^0) - (N(x^0) + D(x^0)) \underbrace{(a_1 + a_2 + \cdots + a_{n-1} + a_n)}_{=0} \Big] \\ &= \frac{1}{D^2(x^0)} \Big[(n - 1)(2nt - 2n + \cos(t) - \cos(n + t - nt))D(x^0) + n(N(x^0) + D(x^0)) \Big] \\ &= \frac{(n - 1)(2nt - 2n + \cos(t) - \cos(n + t - nt)}{D(x^0)} + \frac{n(N(x^0) + D(x^0))}{D^2(x^0)}. \end{aligned}$$

We define

$$\begin{aligned} &P(t) = D(x^0) = 1 + 2n - nt, \\ &H(t) = N(x^0) = nt + (n + t - nt)^2 + (n - 1)t^2 + \sin(n + t - nt) + (n - 1)\sin(t). \end{aligned}$$

Thus, the function $g_{uv}(t)$ can be written as

$$g_{uv}(t) = \frac{(n - 1)(2nt - 2n + \cos(t) - \cos(n + t - nt)}{P(t)} + \frac{n(H(t) + P(t))}{P^2(t)}, \quad t \in [0, 1].$$

We will show that the above problem (P) does not satisfy condition (C_2) by selecting t_1, t_2, and $t_3 = \lambda t_1 + (1 - \lambda)t_2$ with $0 \leq \lambda \leq 1$ such that $g_{uv}(t_3) > \max\{g_{uv}(t_1), g_{uv}(t_2)\}$. In this case, the function $g_{uv}(t)$ is not quasiconvex, which implies that condition (C_2) is not satisfied. We construct

$$t_1 = \frac{n - \frac{\pi}{2} + 2m\pi}{n - 1}, \quad t_2 = \frac{n - \pi + 2m\pi}{n - 1}, \quad t_3 = \frac{1}{2}t_1 + \frac{1}{2}t_2 = \frac{n - \frac{3\pi}{4} + 2m\pi}{n - 1},$$

where $m = -\left\lfloor \frac{n}{2\pi} - \frac{5}{4} \right\rfloor = -\left\lfloor \frac{\ell}{\pi} - \frac{5}{4} \right\rfloor$. In the table below, we present the information that verifies $g_{uv}(t_3) > \max\{g_{uv}(t_1), g_{uv}(t_2)\}$ for some cases of $n = 2\ell$ experimented in Example 2.

Table 4: Comparison of $g_{uv}(t_3)$ and $\max\{g_{uv}(t_1), g_{uv}(t_2)\}$ for Example 2

$n = 2\ell$	m	t_1	t_2	t_3	$g_{uv}(t_1)$	$g_{uv}(t_2)$	$g_{uv}(t_3)$
500	-78	0.01671	0.01357	0.01514	-370.61	-370.52	-370.47
1000	-157	0.01198	0.01041	0.01120	-744.72	-744.62	-744.56
5000	-794	0.00192	0.00160	0.00176	-3745.35	-3745.24	-3745.19
7000	-1112	0.00165	0.00142	0.00154	-5244.86	-5244.76	-5244.71

□