Noname manuscript No.
(will be inserted by the editor)

Composite optimization problems via novel proximal gradient algorithms and applications

Pham Thi Hoai[1] · **Nguyen Pham Duy Thai**[2]

Abstract We consider the *composite optimization problems* under convex and nonconvex settings. For the convex case, the *locally Lipschitz* condition is imposed on the gradient of the differentiable convex term. The classical *proximal gradient method* will be studied with our novel *enhanced adaptive* stepsize selection. To obtain the convergence of the proposed algorithm, we establish a sufficient decrease-type inequality associated with our new stepsize choice. This allows us to demonstrate the descent of the objective value from some fixed iteration and yields the sublinear convergence rate of the new method. Especially, in the case of locally strong convexity of the smooth term, our algorithm converges Q-linearly. When the gradient of the smooth term is globally Lipschitz, our method is extended to a wide class of *nonconvex* composite optimization problems where the smooth term can be convex, concave, indefinite quadratic or fractional (with an indefinite quadratic numerator and a positive affine denominator) functions. The experiments for our new method are conducted for seven practical problems with a large number of randomly generated data from small to large sizes. The superior efficiency of our novel proximal gradient algorithms is demonstrated by numerical results (evaluated by performance profile) in comparison with the other state-of-the-art methods.

Keywords proximal gradient method · nonlinear programming · composite optimization problems · locally Lipschitz gradient · lasso problem

Mathematics Subject Classification (2010) 49J40 · 47H04 · 47H10

1 Introduction

1.1 Problem description and motivation

Composite optimization problems (COP) have arisen from many real-life applications, such as machine learning, signal processing, data science, etc, and have received a lot of attention recently, see e.g., [1,5,6,3,4,7,23,18,31,22,9]. The formulation of (COP) considered in this paper is described as follows:

$$\min_{x\in\mathbb{R}^n} F(x) = f(x) + g(x), \tag{P}$$

where f and g are functions satisfying *Assumption 1* below.

✉ Pham Thi Hoai
hoai.phamthi@hust.edu.vn
Nguyen Pham Duy Thai
thai.npd@u.nus.edu

[1] Faculty of Mathematics and Informatics, Hanoi University of Science and Technology, 1 Dai Co Viet Road, Hanoi, Vietnam

[2] Department of Mathematics, National University of Singapore, Singapore

Assumption 1 *(A1)* $g:\mathbb{R}^n\to(-\infty,+\infty]$ *is a proper and closed convex function.*

(A2) $f:\mathbb{R}^n\to(-\infty,+\infty]$ *is proper and closed such that* $dom(f)$ *is convex,* $dom(g)\subset int(dom(f))$ *and* f *is differentiable on* $int(dom(f))$.

(A3) The optimal solution set X^* *of (P) is nonempty and* F_* *stands for the optimal value of (P).*

One of the conventional methods for solving the problem (P) is *proximal gradient* (PG) method introduced by Fukushima and Mine [15] in 1981 and has now become classical. As a matter of fact, the further origin of the proximal gradient method can be traced back to 1970s with the work of Brucks [11] and Passty [28] in the more general setting of forward backward splitting method. The detailed methodology of the PG method can be found in [4,7]. It is observed that the optimal condition for the problem (P) relates to the concept of its stationary points. Specifically, if $x^*\in\text{dom}(g)$ is a local optimal solution of (P) then it should be a *stationary point* of (P) (see e.g., Theorem 3.72 [4]), i.e., for some $t>0$

$$x^*=\text{Prox}_{tg}(x^*-t\nabla f(x^*)),\ \text{(see e.g., Theorem 10.7 [4])}, \tag{1.1}$$

where $\text{Prox}_{tg}(.)$ is the proximal operator and is defined as the unique optimal solution of the minimization problem

$$\text{Prox}_{tg}(y)=\text{argmin}_{x\in\mathbb{R}^n}\left\{g(x)+\frac{1}{2t}\|x-y\|^2\right\}. \tag{1.2}$$

In the convex situation of (P), i.e., f is convex, the set of stationary points of (P) is coincident with X^*. One can see [4] (Theorems 3.72 and 10.7) for more details. Based on the stationary condition (1.1), starting from some $x^0\in\text{dom}(g)$, the well-known PG method to solve problem (P) is designed by generating the sequence $\{x^k\}$ according to the rule

$$x^{k+1}=\text{Prox}_{t_kg}(x^k-t_k\nabla f(x^k)),\ \ k=0,1,2,... \tag{1.3}$$

The PG scheme (1.3) is useful if we can compute $\text{Prox}_{t_kg}(x^k-t_k\nabla f(x^k))$ easily by some explicit formulas. Though, there is a list of functions whose *proximal operator* is analytically computable and that list can be found in [4]; for instances, g is the ℓ_1 norm or the indicator function of a closed convex set $C\subset\mathbb{R}^n$. In formula (1.3), $t_k>0,k=0,1,2,...$ are defined as *stepsizes* which play a crucial role in the proximal gradient scheme. A suitable stepsize selection can be drawn in the two main points: firstly, it should guarantee the convergence of $\{x^k\}$ to some stationary point of Problem (P); secondly, it should navigate x^k to a "good" stationary point. i.e., providing, for example, the low objective value as much as possible with a cheap computational cost. For the class of L_f- smooth function f, i.e.,

$$\|\nabla f(x)-\nabla f(y)\|\le L_f\|x-y\|,\ \forall x,y\in\text{dom}(g),$$

the stepsize t_k in (1.3) can be controlled flexibly by using *constant stepsize* in $\left(0,\frac{2}{L_f}\right)$ or *backtracking line search* rule. Followed by [4] (Theorem 10.21), one gets the *sublinear rate* of convergence, i.e., computational complexity $O(\frac{1}{k})$ of $F(x^k)-F_*$ if f is assumed to be convex and t_k is either in $\left(0,\frac{1}{L_f}\right]$ or taken by backtracking procedure. In the case where f is strongly convex, the convergence rate of $\{x^k\}$ to some $x^*\in X^*$ is proved to be Q-linear. These properties can be seen as the generalization of the convergence results for the gradient descent method solving unconstrained nonlinear optimization problems, i.e., Problem (P) with $g=0$.

Recently, researchers have been concerned about Problem (P) *without the global Lipschitzness assumption on* ∇f, see, e.g., [2,8,18,19,26] because the class of such functions occurs in many applied problems (see e.g., [19,32] and the references therein). In 2017, Bauschke et al. [2] proposed *NoLips Algorithm* that requires Bregman distances-based computation and constant L in the *Lipschitz-like/convexity condition* (LC). One can see [30] to find the role of non-Euclidean proximal

distances of Bregman type in the development and analysis of some typical first order optimization algorithms. If the stepsize is chosen in $(0, \frac{2-\delta}{L})$ then Nolips algorithm is shown in [2] to have the convergence results similar to the ones of the normal PG scheme. Following that, Dragomir et al. [13] give a lower bound to prove that the $O\left(\frac{1}{k}\right)$ convergence rate of the NoLips method is optimal for the class of problems satisfying the relative smoothness assumption. Nevertheless, one knows that there are some restrictions of taking stepsize within $\left(0, \frac{2}{L_f}\right)$ or $\left(0, \frac{2-\delta}{L}\right)$ like: firstly, the process of finding these constants is not easy in general and secondly, if the coefficients L_f or L are large then the constant stepsizes will be very small and that may take long executing time. Another class of ideas to resolve the lack of a globally Lipschitz gradient was proposed Kanzow and Mehlitz [19], Jia et al. [18], and Zhao et al. [33]. While the methods in [33] are designed for convex settings, the approaches proposed in [19] and [18] can be applied to the nonconvex setting of (P) under the Kurdyka–Łojasiewicz condition. Unfortunately, their stepsize choices rely on some backtracking line search procedures, which can make each iteration computationally expensive.

To overcome the drawbacks of stepsize selection based on line search or estimating some unknown constants like L, L_f mentioned above, one try to find an adaptive way to compute stepsizes in a closed form for solving Problem (P) by using proximal gradient scheme (1.3), such that neither estimating constants like $L_f, L, \ldots$ nor backtracking line search procedures are required. Let us review such publications in the literature which concern this topic when problem (P) satisfies Assumption 1, and the objective function f is convex and has a locally Lipschitz gradient.

Related works: In the specific context of the problem (P) with $g = 0$, and f is convex, some gradient descent type algorithms using adaptive stepsize have been proposed recently, for example, AdGD [24] (2019), NGD [17] (2024). All of them use explicit stepsize strategies based on the local curvature of f. In the general setting of Problem (P) with the convex f, Malitsky and Mishchenko [25] have developed their method AdGD [24] to AdPG (Adaptive Proximal Gradient) for solving the convex case of Problem (P) recently. The stepsize of AdPG is defined by

$$t_k = t_{k-1} \min\left\{\sqrt{\frac{2}{3}+\theta_{k-1}}, \frac{1}{\sqrt{\left[\frac{2t_{k-1}^2\|\nabla f(x^k)-\nabla f(x^{k-1})\|^2}{\|x^k-x^{k-1}\|^2}-1\right]^+}}\right\}, \tag{AdPG}$$

where $\theta_0 = \frac{1}{3}, \theta_k = \frac{t_k}{t_{k-1}}, k \geq 1$. And for some $t \in \mathbb{R}$, the notation t^+ stands for $\max\{t, 0\}$. The iterates of AdPG are proved to converge to an optimal solution of (P) with the *sublinear* convergence rate for the best iterates, i.e., the complexity $O(\frac{1}{k})$ of $\min\limits_{1\leq i\leq k}(F(x^i)-F_*)$. In parallel with this work, Latafat et al. [21] proposed adaPGM that has

$$t_k = t_{k-1} \min\left\{\sqrt{1+\frac{t_{k-1}}{t_{k-2}}}, \frac{1}{2\sqrt{\left[t_{k-1}\left(\frac{t_{k-1}\|\nabla f(x^k)-\nabla f(x^{k-1})\|^2-\left\langle\nabla f(x^k)-\nabla f(x^{k-1}, x^k-x^{k-1}\right\rangle}{\|x^k-x^{k-1}\|^2}\right)\right]^+}}\right\}, k \geq 1. \tag{adaPGM}$$

Soon after, adaPGM is generalized to be AdaPG^{q,r} in Latafat et al. [20] with

$$t_k = t_{k-1} \min\left\{\sqrt{\frac{1}{q}+\frac{t_{k-1}}{t_{k-2}}}, \sqrt{\frac{1-r/q}{\left[\frac{t_{k-1}^2\|\nabla f(x^k)-\nabla f(x^{k-1})\|^2+2t_{k-1}(r-1)\left\langle\nabla f(x^k)-\nabla f(x^{k-1}, x^k-x^{k-1}\right\rangle}{\|x^k-x^{k-1}\|^2}-(2r-1)\right]^+}}\right\}, \tag{AdaPG^{q,r}}$$

where $\frac{1}{2} \leq r < p \leq \frac{3+\sqrt{5}}{2}, t_0 = t_{-1} > 0, k \geq 1$. Notably, AdaPG^{q,r} recovers AdPG if $(p,r) = \left(\frac{3}{2}, \frac{3}{4}\right)$ and adaPGM if $(p,r) = \left(1, \frac{1}{2}\right)$ with slight improvements (see [20] for details). The convergence of AdaPG^{q,r} is then established with the sublinear convergence rate for the best iterates. All the algorithms mentioned above concentrate on solving the convex case of (P), the extension of these methods to nonconvex case of (P) remains unsolved.

1.2 Contributions

In this paper, we utilize the idea of adaptive stepsize used in Algorithm 1.1 NGD (proposed by Hoai et al.[17]) for the proximal gradient scheme (1.3) to solve Problem (P) with a locally Lipschitz gradient condition imposed on the smooth term f. Notably, in 2025, motivated by NGD stepsize, Hoai [16] also proposed NPROX - a variant of the proximal gradient algorithm for solving mixed variational inequality problems - a generalization of problem (P). However, the requirements of NPROX include the monotonicity and global Lipschitzness of the related mapping which corresponds to the convex case of (P) with global Lipschitzness of ∇f. Therefore, NPROX cannot handle (P) with either local Lipschitz ∇f or nonconvex f - both of which are considered in this study.
From now on, we refer to our new method as **NPG** when no confusion arises.

Algorithm 1.1 (NGD) for solving problem $\min_{x\in\mathbb{R}^n} f(x)$ (a special case of Problem (P) when $g=0$) with ∇f being locally Lipschitz.

Step 0 (Initialization). Select $\lambda_0 > 0$, $0 < c_1 < c_0 < \frac{1}{2}$, a tolerance $\varepsilon > 0$ and a positive real sequence $\{\varepsilon_k\}$ such that $\sum_{k=0}^{\infty} \varepsilon_k < \infty$. Choose $x^0 \in \mathbb{R}^n$, $x^1 = x^0 - \lambda_0 \nabla f(x^0), \lambda_{-1} = \lambda_0$ and set $k = 1$.

Step 1.

If $$\|\nabla f(x^k) - \nabla f(x^{k-1})\| > \frac{c_0}{\lambda_{k-1}} \|x^k - x^{k-1}\|$$

then
$$\lambda_k = c_1 \frac{\|x^k - x^{k-1}\|}{\|\nabla f(x^k) - \nabla f(x^{k-1})\|}$$

else $\varepsilon'_{k-1} = \varepsilon_{k-1}$

if $\frac{\lambda_{k-1}}{\lambda_{k-2}} < 1$ then $\varepsilon'_{k-1} = \min\left\{\varepsilon_{k-1}, \sqrt{1 + \frac{\lambda_{k-1}}{\lambda_{k-2}}} - 1\right\}$

$\lambda_k = (1 + \varepsilon'_{k-1})\lambda_{k-1}$.

Step 2. Compute $x^{k+1} = x^k - \lambda_k \nabla f(x^k)$.
Step 3. If $\|\nabla f(x^{k+1})\| < \varepsilon$ then STOP
else setting $k := k + 1$, and return to **Step 1**.

Particularly, our main contributions are summarized as follows:

1. Firstly, we extend NGD to the composite model (P). The resulting algorithm is called **NPG1**. When $g = 0$, **NPG1** becomes NGD but with an **enhanced** range for the parameters c_0, c_1, extended by a factor of $\sqrt{2}$.

2. To investigate the convergence of **NPG1**, we establish a *sufficient decrease-type inequality*, which ensures the objective function decreases from some fixed iteration. Consequently, it yields a sublinear convergence rate for the last iterate (from some fixed iteration) when f is convex with a locally Lipschitz gradient, and the Q-linear rate when f is locally strongly convex. This point is the main difference between **NPG1** and the recent methods AdPG [25], adaPGM [21], and AdaPG^{q,r} [20]. In addition, NPG1 also achieves a sublinear convergence rate for the best iterates, similar to the algorithms given in [25,21,20]. Notably, the lack of a descent property in AdPG [25], adaPGM [21], and AdaPG^{q,r} [20] also prevents these methods from being extended to the class of nonconvex composite optimization problems.

3. Secondly, **NPG2** is proposed for a broad class of nonconvex f with globally Lipschitz gradient. In this setting, the range of the coefficients c_0, c_1 is extended to $(0,1)$. Importantly, from some fixed iteration, NPG2 achieves the same standard convergence guarantees as classical proximal gradient methods with constant or backtracking stepsizes in the nonconvex setting; see e.g., [5, Theorem 10.15].

4. Thirdly, **NPG-quad** is designed for indefinite quadratic functions f with a better approximation for t_k according to the local behavior of f. Moreover, c_0, c_1 are chosen in a larger interval $(0,2)$. The convergence results of **NPG-quad** are shown similarly to those of NPG2.

5. Besides the adaptation in computation of the stepsize selection as the existing methods AdPG, adaPGM, AdaPG^{q,r}, a key distinction of our method **NPG** is that, from a fixed iteration onward, its stepsize sequence is guaranteed to **increase** to a positive limit.

6. Comprehensive experiments are conducted to demonstrate the practical effectiveness of our algorithms on a variety of important problems arising in machine learning, optimal transport, signal processing and related fields.

   It is worth noting that, in the context of convex composite optimization problems with locally Lipschitz assumption imposed on the gradient of the differentiable term, NPG1 can be described as an adaptive proximal gradient method that incorporates the descent properties of classical methods. It ensures that, from some fixed iteration onward, the objective value decreases after each step. This property is crucial for extending our method to NPG2 and NPG-quad, which solve nonconvex composite optimization problems with theoretical convergence guarantees.

1.3 **Structure of the paper**

The rest of the paper is organized as follows. After summarizing some necessary preliminaries in Section 2, we propose our new proximal algorithm in Section 3 for solving the convex situation of (P) under the locally Lipschitz condition of ∇f. In Section 4, we consider a nonconvex case of (P) with an other new algorithm. Section 5 presents a special version of proposed method applied for the indefinite quadratic function f. The numerical experiments on a set of practical examples are stated in Section 6. Lastly, the paper is closed by some conclusions in Section 7.

2 Preliminaries

In this section, we recall some necessary fundamental results which are useful to derive our main contributions in the upcoming sections.

Definition 2.1 [4] Consider the problem (P) under Assumption 1.

(i) A point x^* at which f is differentiable is called a **stationary point** of (P) if

$$-\nabla f(x^*) \in \partial g(x^*).$$

(ii) For any $t > 0$, the proximal operator $\mathrm{Prox}_{tg} : \mathbb{R}^n \to \mathbb{R}^n$ and the gradient mapping $G^{f,g}_{1/t} : \mathrm{dom}(f) \to \mathbb{R}^n$ are defined respectively by

$$\mathrm{Prox}_{tg}(y) = \underset{x\in\mathbb{R}^n}{\mathrm{argmin}} \left\{ g(x) + \frac{1}{2t}\|x-y\|^2 \right\};$$

$$G^{f,g}_{1/t}(x) = \frac{x - \mathrm{Prox}_{tg}(x - t\nabla f(x))}{t}.$$

When f, g are clear from the context, we shall use the notation $G_{1/t}$ instead of $G^{f,g}_{1/t}$.

Lemma 2.1 (optimality conditions for Problem (P)**, see e.g., Theorem 3.72 in [4])** *Consider the problem* (P) *under Assumption 1.*

(a) If $x^* \in \mathrm{dom}(g)$ is a local optimal solution of (P), then x^* is a stationary point of (P).

(b) Suppose that f is convex, then x^* is a global optimal solution of (P) if and only if x^* is a stationary point of (P).

The following lemmas are derived directly from Theorems 10.7, 10.9 and Lemma 10.10 in [4].

Lemma 2.2 *Consider the problem* (P) *under Assumption 1 and $t > 0$. Then*

(i) if $g = 0$, we have $G_{1/t}(x) = \nabla f(x)$;

(ii) $x^* \in \mathrm{dom}(g)$ is a stationary point of (P) *if and only if $G_{1/t}(x^*) = 0$.*

Lemma 2.3 *Consider the problem* (P) *under Assumption 1, suppose that $0 < t_1 \leq t_2$, then*

$$\|G_{1/t_1}(x)\| \geq \|G_{1/t_2}(x)\|, \quad \text{for any } x \in \mathrm{dom}(f).$$

Lemma 2.4 *Assuming that f, g satisfy Assumption 1 and furthermore ∇f is Lipschitz continuous with constant L_f. Then, for $t > 0$, we have*

$$\|G_{1/t}(x) - G_{1/t}(y)\| \leq \left(\frac{2}{t} + L_f\right)\|x - y\|, \quad \text{for any } x, y \in \mathrm{dom}(g).$$

Remark 2.1 From the above results, one can use $\|G_{1/t}(x)\|$ as an "optimal measure" for Problem (P) in the sense that "it is always nonnegative and equal to zero if and only if x is a stationary point", see Beck [4] for details.

Lemma 2.5 *Under Assumption 1, the sequence $\{x^k\}$ generated by proximal gradient scheme (1.3) for solving the problem (P) has the following properties:*

(i) there exists $\widetilde{\nabla} g(x^{k+1}) \in \partial g(x^{k+1})$ such that $x^{k+1} = x^k - t_k\left(\nabla f(x^k) + \widetilde{\nabla} g(x^{k+1})\right)$;

(ii) for all $x \in \mathrm{dom}(g)$, we have

$$g(x) - g(x^{k+1}) \geq \left\langle x^{k+1} - x, \nabla f(x^k) + \frac{x^{k+1} - x^k}{t_k} \right\rangle. \tag{2.1}$$

and

$$F(x) - F(x^{k+1}) \geq \langle \nabla f(x^{k+1}) - \nabla f(x^k), x^k - x^{k+1}\rangle + \frac{1}{t_k}\left\|x^{k+1} - x^k\right\|^2 + \frac{1}{t_k}\left\langle x^{k+1} - x^k, x^k - x\right\rangle. \tag{2.2}$$

Proof (i) Since $x^{k+1} \in \underset{x \in \mathbb{R}^n}{\mathrm{argmin}}\left\{g(x) + \frac{1}{2t_k}\left\|x - (x^k - t_k \nabla f(x^k))\right\|^2\right\}$ then

$$0 \in \partial g(x^{k+1}) + \frac{1}{t_k}\left(x^{k+1} - x^k + t_k \nabla f(x^k)\right).$$

Hence there exists $\widetilde{\nabla} g(x^{k+1}) \in \partial g(x^{k+1})$ such that

$$x^{k+1} = x^k - t_k(\nabla f(x^k) + \widetilde{\nabla} g(x^{k+1})). \tag{2.3}$$

(ii) From (i) and the convexity of g we easily get that

$$\begin{aligned} g(x) - g(x^{k+1}) &\geq \left\langle x - x^{k+1}, \widetilde{\nabla} g(x^{k+1})\right\rangle \\ &= \left\langle x^{k+1} - x, \nabla f(x^k) + \frac{x^{k+1} - x^k}{t_k}\right\rangle. \end{aligned}$$

Now, using the convexity of f we derive that

$$\begin{aligned} F(x)-F(x^{k+1}) &= f(x)+g(x)-f(x^{k+1})-g(x^{k+1}) \\ &\geq f(x^k)+\left\langle x-x^k,\nabla f(x^k)\right\rangle+\left\langle x^{k+1}-x,\nabla f(x^k)+\frac{x^{k+1}-x^k}{t_k}\right\rangle-f(x^{k+1}) \\ &= f(x^k)-f(x^{k+1})+\left\langle x^{k+1}-x^k,\nabla f(x^k)\right\rangle+\frac{1}{t_k}\left\langle x^{k+1}-x^k,x^{k+1}-x\right\rangle \\ &\geq \langle \nabla f(x^{k+1})-\nabla f(x^k),x^k-x^{k+1}\rangle+\frac{1}{t_k}\left\|x^{k+1}-x^k\right\|^2+\frac{1}{t_k}\left\langle x^{k+1}-x^k,x^k-x\right\rangle. \end{aligned} \tag{2.4}$$

The following lemma presented in [24] will be useful to analyze the convergence of our proposed method in the next section.

Lemma 2.6 (Lemma 2 [24]) *Let $\{x^k\}\subset\mathbb{R}^n$ be a bounded sequence and its cluster points lie in $X\subset\mathbb{R}^n$. If there exists a nonnegative sequence $\{a_k\}\subset\mathbb{R}_+$ such that*

$$\|x^{k+1}-x\|^2+a_{k+1}\leq\|x^k-x\|^2+a_k,\ \ \forall x\in X, \tag{2.5}$$

then $\{x^k\}$ converges to an element of X.

Proof Assume that there exist two different cluster points $\bar{x}^1,\bar{x}^2$ of $\{x^k\}$. Then, there exist two subsequences $x^{k_i}\to\bar{x}^1$ and $x^{k_j}\to\bar{x}^2$. Given that, the real sequence $\|x^k-x\|^2+a_k$ is lower bounded by zero and nonincreasing, so it converges for any $x\in X$. Let $x=\bar{x}^1$ we have

$$\begin{aligned} \lim_{k\to+\infty}\left(\|x^k-\bar{x}^1\|^2+a_k\right) &= \lim_{i\to+\infty}\left(\|x^{k_i}-\bar{x}^1\|^2+a_{k_i}\right)=\lim_{i\to+\infty}a_{k_i} \\ &= \lim_{j\to+\infty}\left(\|x^{k_j}-\bar{x}^1\|^2+a_{k_j}\right)=\|\bar{x}^2-\bar{x}^1\|^2+\lim_{j\to+\infty}a_{k_j}. \end{aligned}$$

Hence, $\lim_{i\to+\infty}a_{k_i}=\lim_{j\to+\infty}a_{k_j}+\|\bar{x}^2-\bar{x}^1\|^2$. Repeating this with $x=\bar{x}^2$ yields $\lim_{j\to+\infty}a_{k_j}=\lim_{i\to+\infty}a_{k_i}+\|\bar{x}^1-\bar{x}^2\|^2$. Thus, we obtain $\bar{x}^1=\bar{x}^2$, which implies the convergence of $\{x^k\}$.

3 A new proximal gradient algorithm for the convex case of the problem (P) with locally Lipschitz ∇f

It is worth noting that when f has a globally Lipschitz gradient with constant L_f and t_k is chosen as a fixed number in $\left(0,\frac{2}{L_f}\right)$ or by some line search strategy, the common technique establishing the convergence of proximal gradient method (1.3) for solving (P) is related to *the sufficient decrease inequality*, i.e., showing the existence of a positive constant M such that

$$F(x^k)-F(x^{k+1})\geq M\|x^{k+1}-x^k\|^2,\ \ k\geq 0, \qquad \text{(sufficient decrease ineq.)}$$

For the proximal gradient algorithms using adaptive stepsizes solving problem (P) in the literature like AdPG [25], adaPGM[21] and AdaPG^{q,r} [20], the obstacle of the locally Lipschitz gradient condition has been overcome by constructing Lyapunov type functions and then obtain the boundedness of the iterates. Since, on the compact set $T=\overline{conv}\left(\{x^*,x^0,x^1,...\}\right)$ ($x^*\in X^*$ is an optimal solution of (P)) all properties of a function f with locally Lipschitz gradient can be operated as those of a globally Lipschitz gradient function. The convergence of their proposed approaches are then deduced by relying on the interesting techniques different from the usual way based on the sufficient decrease ineq. However, the absence of descent property prevents their algorithms from achieving sublinear convergence rate $O\left(\frac{1}{k}\right)$ of $F(x^k)-F_*$. The convergence rate $O\left(\frac{1}{k}\right)$ of their methods is applied for $\min_{1\leq i\leq k}\{F(x^i)-F_*\}$ when solving convex problem (P) under locally Lipschitz gradient condition. Moreover, extending non-decreasing methods to the nonconvex case of (P) makes proving convergence difficult.

Notably, our stepsize selection **NPG** is not only adapted to the local curvature of f but also controllable by using a pre-selected positive convergent sequence $\{\gamma_k\}$. If $\{\gamma_k\}$ converges to some pre-determined non-negative value, our proposed algorithms based on **NPG** stepsize have all the convergence results as those in [21,20,25]. With a stronger hypothesis requiring the convergence of $\sum_{k=0}^{+\infty}\gamma_k$, our method can establish the sufficient decrease ineq. for (P) from some fixed iteration k^* and therefore could provide the convergence rate $O\left(\frac{1}{k}\right)$ of $F(x^k)-F_*$ with $k>k^*$. We will explore these details in the subsequent parts of the paper.

Firstly, let us introduce the proximal gradient algorithm **NPG1** with our new stepsize selection to solve problem (P) under *Assumption 1* and *Assumption 2* below.

Assumption 2 *f is convex and has a locally Lipschitz gradient.*

Algorithm 3.1 (NPG1)

Step 0. Select $t_0>0$, $0<c_1<c_0<\frac{1}{\sqrt{2}}$, $\theta>0$, a tolerance $\varepsilon>0$ and a nonegative real sequence $\{\gamma_k\}$ such that $\sum_{k=0}^{+\infty}\gamma_k<+\infty$. Choose $x^0\in\mathrm{dom}(g)$, $x^1=\mathrm{Prox}_{t_0 g}(x^0-t_0\nabla f(x^0)), t_{-1}=t_0$ and set $k=1$.

Step 1.

$$\textbf{If}\quad \|\nabla f(x^k)-\nabla f(x^{k-1})\|>\frac{c_0}{t_{k-1}}\|x^k-x^{k-1}\| \tag{3.1}$$

$$\textbf{then}\quad t_k=c_1\frac{\|x^k-x^{k-1}\|}{\|\nabla f(x^k)-\nabla f(x^{k-1})\|} \tag{3.2}$$

$$\textbf{else}\quad \gamma'_{k-1}=\gamma_{k-1}$$

$$\text{if } \frac{t_{k-1}}{t_{k-2}}<\theta \text{ then } \gamma'_{k-1}=\min\left\{\gamma_{k-1},\sqrt{1+\frac{t_{k-1}}{t_{k-2}}}-1\right\} \tag{3.3}$$

$$t_k=(1+\gamma'_{k-1})t_{k-1}. \tag{3.4}$$

Step 2. Compute $x^{k+1}=\mathrm{Prox}_{t_k g}(x^k-t_k\nabla f(x^k))$.

Step 3. If $\|G_{1/t_k}(x^k)\|=\|x^k-x^{k+1}\|/t_k<\varepsilon$ **then** STOP **else** setting $k:=k+1$ and return to **Step 1**.

Remark 3.1 It is observed that, in the case $g=0$, Algorithm 3.1 (NPG1) is similar to NGD [17] but more general in the following three points:

(i) The first is that the bounds of c_0,c_1 are enlarged from $\left(0,\frac{1}{2}\right)$ in NGD to $\left(0,\frac{1}{\sqrt{2}}\right)$ in NPG1.

(ii) The second is in the presence of θ to define (3.3) in NPG1 while this parameter is taken as 1 in NGD.

It is worth noting that the key improvement in enlarging the bounds of c_0,c_1 lies in how we estimate the term $t_k^2\left\|\nabla f(x^k)+\widetilde{\nabla}g(x^k)\right\|^2$ in Lemma 3.3(i). Specifically, when $g=0$, this quantity in NPG1 reduces to $\|x^{k+1}-x^k\|^2$ in NGD [17], where its upper bound was derived using the Cauchy–Schwarz inequality. In contrast, our approach avoids that technique and instead relies solely on the convexity of f and g. As a result, the upper bounds for c_0 and c_1 are improved over those in NGD. The generalization of θ is easy to achieve by refining the technique used in NGD.

Analogous to existing methods we need to prepare some lemmas which will help us to prove the boundedness of $\{x^k\}$ - a key step to overcome the difficulties generated by the locally Lipschitz continuity of ∇f.

Lemma 3.1 *For all $x \in \mathrm{dom}(g)$ we have*

$$\|x^{k+1}-x\|^2 + 2t_k\Big(F(x^k)-F(x)\Big) \le \|x^k-x\|^2 + t_k^2\Big\|\nabla f(x^k)+\widetilde{\nabla} g(x^k)\Big\|^2.$$

Proof From Lemma 2.5 (ii), for all $x \in \mathrm{dom}(g)$

$$\begin{aligned} 2t_k\Big(g(x^{k+1})-g(x)\Big) &\le 2\Big\langle x^{k+1}-x^k+t_k\nabla f(x^k), x-x^{k+1}\Big\rangle \\ &= \|x^k-x\|^2 - \|x^{k+1}-x^k\|^2 - \|x^{k+1}-x\|^2 \\ &\quad +2t_k\Big\langle \nabla f(x^k), x-x^{k+1}\Big\rangle. \end{aligned} \tag{3.5}$$

Using the convexity of f and g, we continue evaluating

$$\begin{aligned} \langle \nabla f(x^k), x-x^{k+1}\rangle &= \langle \nabla f(x^k), x-x^k\rangle + \langle \nabla f(x^k)+\widetilde{\nabla} g(x^k), x^k-x^{k+1}\rangle + \langle \widetilde{\nabla} g(x^k), x^{k+1}-x^k\rangle \\ &\le f(x)-f(x^k)+\Big\langle \nabla f(x^k)+\widetilde{\nabla} g(x^k), x^k-x^{k+1}\Big\rangle + g(x^{k+1})-g(x^k). \end{aligned} \tag{3.6}$$

From (3.5) and (3.6), we derive that

$$\|x^{k+1}-x\|^2 + 2t_k\Big(F(x^k)-F(x)\Big) \le \|x^k-x\|^2 + RH, \tag{3.7}$$

where

$$\begin{aligned} RH &= 2t_k\Big\langle \nabla f(x^k)+\widetilde{\nabla} g(x^k), x^k-x^{k+1}\Big\rangle - \|x^{k+1}-x^k\|^2 \\ &= t_k\Big\langle 2\nabla f(x^k)+2\widetilde{\nabla} g(x^k)-\nabla f(x^k)-\widetilde{\nabla} g(x^{k+1}), x^k-x^{k+1}\Big\rangle \\ &= t_k^2\Big\langle \nabla f(x^k)+2\widetilde{\nabla} g(x^k)-\widetilde{\nabla} g(x^{k+1}), \nabla f(x^k)+\widetilde{\nabla} g(x^{k+1})\Big\rangle \\ &= t_k^2\left(\Big\|\nabla f(x^k)+\widetilde{\nabla} g(x^k)\Big\|^2 - \Big\|\widetilde{\nabla} g(x^{k+1})-\widetilde{\nabla} g(x^k)\Big\|^2\right) \\ &\le t_k^2\Big\|\nabla f(x^k)+\widetilde{\nabla} g(x^k)\Big\|^2. \end{aligned} \tag{3.8}$$

The final conclusion is obtained by (3.7) and (3.8).

Lemma 3.2 *Let $\{t_k\}$ be a sequence of stepsizes generated by Algorithm 3.1 then there exists $k_0 \in \mathbb{N}$ such that*

$$1+\frac{t_k}{t_{k-1}} \ge \frac{t_{k+1}^2}{t_k^2} \ \ \forall k \ge k_0. \tag{3.9}$$

Proof If $\|\nabla f(x^{k+1})-\nabla f(x^k)\| > \frac{c_0}{t_k}\|x^{k+1}-x^k\|$ then $t_{k+1} = \frac{c_1\|x^{k+1}-x^k\|}{\|\nabla f(x^{k+1})-\nabla f(x^k)\|} < \frac{c_1 t_k}{c_0}$ (by (3.2)). Hence $\frac{t_{k+1}}{t_k} < \frac{c_1}{c_0} < 1$ and (3.9) is followed. Conversely, in the case that $\|\nabla f(x^{k+1})-\nabla f(x^k)\| \le \frac{c_0}{t_k}\|x^{k+1}-x^k\|$ then by (3.4), $t_{k+1} = (1+\gamma_k')t_k$ and (3.9) is equivalent to

$$\left(\frac{t_{k+1}}{t_k}\right)^2 = (1+\gamma_k')^2 \le 1+\frac{t_k}{t_{k-1}}. \tag{3.10}$$

Moreover, from (3.3), if $\frac{t_k}{t_{k-1}} \ge \theta$ then $\gamma_k' = \gamma_k$, furthermore since $\sum_{k=0}^{+\infty}\gamma_k < +\infty$, hence there exists k_0 such that

$$\gamma_k' = \gamma_k \le \sqrt{1+\theta}-1 \le \sqrt{1+\frac{t_k}{t_{k-1}}}-1 \quad \forall k \ge k_0. \tag{3.11}$$

For the remaining case $\frac{t_k}{t_{k-1}} < \theta$, we have

$$\gamma_k' = \min\left\{\gamma_k, \sqrt{1+\frac{t_k}{t_{k-1}}}-1\right\} \le \sqrt{1+\frac{t_k}{t_{k-1}}}-1. \tag{3.12}$$

Thus, (3.9) is proved from (3.11) and (3.12).

As mentioned above, the bounded property of the sequence $\{x^k\}$ in the following lemma provides us an important key beyond the challenge of locally Lipschitz continuity of ∇f.

Lemma 3.3 *Let $\{x^k\}$ be a sequence generated by Algorithm 3.1 then the following statements hold*

(i) there exists $k_1 \geq k_0$ such that for all $k \geq k_1$,

$$t_k^2 \left\|\nabla f(x^k) + \widetilde{\nabla} g(x^k)\right\|^2 \leq \frac{1}{2}\|x^k - x^{k-1}\|^2 + \frac{t_k^2}{t_{k-1}}\left(F(x^{k-1}) - F(x^k)\right); \tag{3.13}$$

(ii) $\{x^k\}$ is bounded.

(iii) $\{t_k\}$ is lower bounded by a positive number.

Proof (i) We have the relation

$$t_k^2 \left\|\nabla f(x^k) + \widetilde{\nabla} g(x^k)\right\|^2 = \underbrace{t_k^2 \left\|\nabla f(x^k) - \nabla f(x^{k-1})\right\|^2}_{A} + B, \tag{3.14}$$

where

$$\begin{aligned} B &= 2t_k^2 \left\langle \nabla f(x^k) + \widetilde{\nabla} g(x^k), \nabla f(x^{k-1}) + \widetilde{\nabla} g(x^k)\right\rangle - t_k^2 \left\|\nabla f(x^{k-1}) + \widetilde{\nabla} g(x^k)\right\|^2 \\ &= \frac{t_k^2}{t_{k-1}} \left\langle \nabla f(x^k) + \widetilde{\nabla} g(x^k), x^{k-1} - x^k\right\rangle + \frac{t_k^2}{t_{k-1}} \underbrace{\left\langle \nabla f(x^k) - \nabla f(x^{k-1}), x^{k-1} - x^k\right\rangle}_{\leq 0} \\ &\leq \frac{t_k^2}{t_{k-1}}\left(F(x^{k-1}) - F(x^k)\right). \end{aligned} \tag{3.15}$$

We now prove that there exists $k_1 \geq k_0$ such that

$$A \leq \frac{1}{2}\|x^k - x^{k-1}\|^2 \quad \forall k \geq k_1. \tag{3.16}$$

Indeed, from Algorithm 3.1, if $\left\|\nabla f(x^k) - \nabla f(x^{k-1})\right\| > \frac{c_0}{t_{k-1}}\|x^k - x^{k-1}\|$ then $t_k = \frac{c_1\|x^k - x^{k-1}\|}{\|\nabla f(x^k) - \nabla f(x^{k-1})\|}$ and since $c_1 < \frac{1}{\sqrt{2}}$, we have

$$A = t_k^2\|\nabla f(x^k) - \nabla f(x^{k-1})\|^2 = c_1^2\|x^k - x^{k-1}\|^2 < \frac{1}{2}\|x^k - x^{k-1}\|^2.$$

Conversely, if $\|\nabla f(x^k) - \nabla f(x^{k-1})\| \leq \frac{c_0}{t_{k-1}}\|x^k - x^{k-1}\|$ then

$$t_k = (1 + \gamma'_{k-1})t_{k-1} \leq (1 + \gamma_{k-1})\frac{c_0\|x^k - x^{k-1}\|}{\|\nabla f(x^k) - \nabla f(x^{k-1})\|}$$

which follows

$$t_k^2\|\nabla f(x^k) - \nabla f(x^{k-1})\|^2 \leq (1 + \gamma_{k-1})^2 c_0^2\|x^k - x^{k-1}\|^2. \tag{3.17}$$

The condition $\sum_{k=0}^{+\infty} \gamma_k < +\infty$ indicates that there exists $k_1 \geq k_0$ satisfying

$$\gamma_{k-1} \leq \frac{1}{\sqrt{2}c_0} - 1 \quad \forall k \geq k_1 \ \left(\frac{1}{\sqrt{2}c_0} - 1 > 0 \text{ since } c_0 < \frac{1}{\sqrt{2}}\right), \tag{3.18}$$

which is equivalent to $(1 + \gamma_{k-1})^2 c_0^2 \leq \frac{1}{2}$ for all $k \geq k_1$. From (3.17) we have (3.16). The combination of (3.14), (3.15) and (3.16) indicates (3.13).

(ii) Using Lemma 3.1 with $x = x^*$ and (3.13), for all $k \geq k_1$ we have

$$\begin{aligned}
&\|x^{k+1} - x^*\|^2 + 2t_k\left(F(x^k) - F(x^*)\right) + t_k^2\left\|\nabla f(x^k) + \widetilde{\nabla} g(x^k)\right\|^2 \\
&\leq \|x^k - x^*\|^2 + 2t_k^2\left\|\nabla f(x^k) + \widetilde{\nabla} g(x^k)\right\|^2 \\
&\leq \|x^k - x^*\|^2 + \|x^k - x^{k-1}\|^2 + 2\frac{t_k^2}{t_{k-1}}\left(F(x^{k-1}) - F(x^k)\right). \qquad (3.19)
\end{aligned}$$

Nevertheless,

$$\begin{aligned}
t_k^2\left\|\nabla f(x^k) + \widetilde{\nabla} g(x^k)\right\|^2 &= \left\|t_k\left(\nabla f(x^k) + \widetilde{\nabla} g(x^{k+1})\right) + t_k\left(\widetilde{\nabla} g(x^k) - \widetilde{\nabla} g(x^{k+1})\right)\right\|^2 \\
&= \left\|(x^k - x^{k+1}) + t_k\left(\widetilde{\nabla} g(x^k) - \widetilde{\nabla} g(x^{k+1})\right)\right\|^2 \\
&= \|x^k - x^{k+1}\|^2 + \underbrace{2t_k\left\langle x^k - x^{k+1}, \widetilde{\nabla} g(x^k) - \widetilde{\nabla} g(x^{k+1})\right\rangle}_{\geq 0 \text{ because } g \text{ is convex}} + t_k^2\underbrace{\|\widetilde{\nabla} g(x^k) - \widetilde{\nabla} g(x^{k+1})\|^2}_{\geq 0} \\
&\geq \|x^k - x^{k+1}\|^2. \qquad (3.20)
\end{aligned}$$

Hence, using inequality (3.20) for the left hand side of (3.19) we obtain that

$$\begin{aligned}
&\|x^{k+1} - x^*\|^2 + 2t_k\left(1 + \frac{t_k}{t_{k-1}}\right)\left(F(x^k) - F(x^*)\right) + \|x^k - x^{k+1}\|^2 \\
&\leq \|x^k - x^*\|^2 + \|x^{k-1} - x^k\|^2 + 2\frac{t_k^2}{t_{k-1}}\left(F(x^{k-1}) - F(x^*)\right) \quad \forall k \geq k_1. \qquad (3.21)
\end{aligned}$$

Remember that from Lemma 3.2 we derive $2t_k\left(1 + \frac{t_k}{t_{k-1}}\right) \geq \frac{2t_{k+1}^2}{t_k}\ \forall\ k \geq k_1$. Therefore, by (3.21), for all $k \geq k_1$ we have

$$\begin{aligned}
&\|x^{k+1} - x^*\|^2 + \|x^k - x^{k+1}\|^2 + \frac{2t_{k+1}^2}{t_k}\left(F(x^k) - F(x^*)\right) \\
\leq\ &\|x^k - x^*\|^2 + \|x^{k-1} - x^k\|^2 + \frac{2t_k^2}{t_{k-1}}\left(F(x^{k-1}) - F(x^*)\right). \qquad (3.22)
\end{aligned}$$

This inequality follows that

$$\|x^{k+1} - x^*\|^2 + \|x^k - x^{k+1}\|^2 + \frac{2t_{k+1}^2}{t_k}\left(F(x^k) - F(x^*)\right) \leq \mathscr{R}^2,\ \ \forall k \geq k_1, \qquad (3.23)$$

where

$$\mathscr{R}^2 = \|x^{k_1} - x^*\|^2 + \|x^{k_1-1} - x^{k_1}\|^2 + \frac{2t_{k_1}^2}{t_{k_1-1}}\left(F(x^{k_1-1}) - F(x^*)\right).$$

The relation (3.23) implies the boundedness of $\{x^k\}_{k\geq k_1}$ and hence the boundedness of $\{x^k\}_{k\geq 0}$ is obtained.

(iii) From (ii), the set $T = \overline{conv}\left(\{x^*, x^0, x^1, \ldots\}\right)$ is bounded and closed hence it is compact. From the local Lipschitz continuity of ∇f, it is easy to see that there exists $L_0 > 0$ satisfying $\|\nabla f(x) - \nabla f(y)\| \leq L_0\|x - y\| \quad \forall x, y \in T$. Thereafter, either $t_1 \geq \frac{c_1}{L_0}$ or $t_1 = (1+\gamma_0')t_0 \geq t_0$. The induction process derives that

$$t_k \geq \min\left\{\frac{c_1}{L_0}, t_0\right\} = t_{\min} > 0 \quad \forall k \geq 0. \qquad (3.24)$$

Remark 3.2 From the proof of Lemma 3.3 (eq. (3.11) and (3.18)), we see that the value of k_1 totally depends on the sequence $\{\gamma_k\}$. Particularly, k_1 is the smallest number such that $0 \leq \gamma_k \leq$

$\min\left\{\frac{1}{\sqrt{2}c_0}-1,\sqrt{1+\theta}-1\right\}$ for all $k\geq k_1$. Thus, if the positive sequence $\{\gamma_k\}$ is created such that $0\leq\gamma_k\leq\min\left\{\frac{1}{\sqrt{2}c_0}-1,\sqrt{1+\theta}-1\right\}$ for all $k\geq 1$ then $k_1=1$ and therefore we obtain (3.22) for all $k\geq 1$, i.e.,

$$\|x^{k+1}-x^*\|^2+\|x^k-x^{k+1}\|^2+\frac{2t_{k+1}^2}{t_k}\Big(F(x^k)-F(x^*)\Big)\leq\mathscr{R}^2,\ \ \forall k\geq 1, \tag{3.25}$$

where

$$\mathscr{R}^2=\|x^1-x^*\|^2+\|x^0-x^1\|^2+\frac{2t_1^2}{t_0}\left(F(x^0)-F(x^*)\right).$$

As a consequence, we can establish the convergence results analogous to those given by Malitsky and Mishchenko [25] we immediately obtain all convergence results of NPG1 similar as that of AdPG [25] in solving problem (P) without using the condition $\sum\limits_{k=0}^{+\infty}\gamma_k<+\infty$ as follows.

Theorem 3.1 *Taking $k_1\in\mathbb{N}$ and considering Algorithm 3.1 (NPG1) with the sequence $\{\gamma_k\}$ such that $0\leq\gamma_k\leq\min\left\{\frac{1}{\sqrt{2}c_0}-1,\sqrt{1+\theta}-1\right\}$ for all $k\geq k_1$. Then under Assumptions 1 and 2, the iterates $\{x^k\}$ generated by Algorithm 3.1 (NPG1) satisfies*

(i) $\{x^k\}_{k\geq k_1}$ converges to an optimal solution x^* of Problem (P).

(ii)

$$\min_{k_1\leq i\leq K}(F(x^i)-F_*)\leq\frac{\mathscr{R}^2}{2\sum\limits_{i=k_1}^{K}t_i}\leq\frac{\mathscr{R}^2}{2(K-k_1+1)t_{\min}}. \tag{3.26}$$

Consequently, if $k_1=1$, i.e., we choose γ_k such that $0\leq\gamma_k\leq\min\left\{\frac{1}{\sqrt{2}c_0}-1,\sqrt{1+\theta}-1\right\}$ for all $k\geq 1$ then $\{x^k\}_{k\geq 1}$ converges to an optimal solution x^* of Problem (P) *and*

$$\min_{1\leq i\leq K}(F(x^i)-F_*)\leq\frac{\mathscr{R}^2}{2\sum\limits_{t=1}^{K}t_i}\leq\frac{\mathscr{R}^2}{2Kt_{\min}}. \tag{3.27}$$

The detailed proof of this theorem is similar to that of [25] and is provided in A.1 of the Appendix.

As mentioned previously, we analyze more thoroughly the convergence results of NPG1 in the upcoming parts of the paper by designing a sufficient decrease inequality (in Corollary 3.1) without the global Lipschitz assumption on $\nabla f(x)$. This technique is different from that of [25, 21,20]. To achieve this, we exploit the special property of t_k that it converges to a positive limit t^*. Note that the sequence of stepsizes t_k given by [25,21,20] is not shown to be bounded from above, and this is one of the obstacles to deriving the descent property. For NPG1, we can control the sequence t_k through the sequence γ_k, and if $\sum\limits_{k=1}^{+\infty}\gamma_k$ converges, then t_k will converge as stated in the following lemma.

Lemma 3.4 *Let $\{t_k\}$ be a sequence of stepsizes generated by Algorithm 3.1 then, $\{t_k\}$ is convergent and has a positive limit.*

Proof If we set $r_k=\ln t_{k+1}-\ln t_k$ and $r_k^+=\max\{0,r_k\}\geq 0, r_k^-=-\min\{0,r_k\}\geq 0$, $\forall k\geq 0$ then $r_k=r_k^+-r_k^-$. On the other hand, from Algorithm 3.1, we observe that $0<c_1<c_0<\frac{1}{\sqrt{2}}$, hence both of (3.2) and (3.4) give

$$r_k=\ln\frac{t_{k+1}}{t_k}\leq\ln(1+\gamma_k')\leq\gamma_k'\leq\gamma_k\quad\forall k\geq 0.$$

Thus, $r_k^+ \leq \gamma_k$. Moreover, the series $\sum\limits_{k=0}^{+\infty} \gamma_k$ converges then $\sum\limits_{k=0}^{+\infty} r_k^+ < +\infty$. Noticeably,

$$\ln t_{k+1} - \ln t_0 = \sum_{i=0}^{k} r_i = \sum_{i=0}^{k} (r_i^+ - r_i^-) = \sum_{i=0}^{k} r_i^+ - \sum_{i=0}^{k} r_i^-. \tag{3.28}$$

Hence if the nonnegative series $\sum\limits_{k=0}^{+\infty} r_k^-$ diverges, i.e., $\lim\limits_{k\to+\infty} \sum\limits_{i=0}^{k} r_i^- = +\infty$ then

$$\lim_{k\to+\infty} (\ln t_{k+1}) = -\infty$$

which implies $\lim\limits_{k\to+\infty} t_k = 0$.This result contradicts the assertion (i). Thus, $\sum\limits_{k=0}^{+\infty} r_k^-$ is convergent and therefore $\lim\limits_{k\to+\infty} t_k = t^* \in (0, +\infty)$ (followed by (3.28)).

The result in the following lemma gives us an inequality like Lipschitz gradient continuity but with flexible constants for each pair of x^{k-1} and x^k.

Lemma 3.5 *There exists k^* such that*

$$\|\nabla f(x^k) - \nabla f(x^{k-1})\| \leq \frac{c_0}{t_{k-1}} \|x^k - x^{k-1}\|, \ \ \forall k \geq k^*. \tag{3.29}$$

Proof Assuming that there is a subsequence $\{k_i\} \subset \mathbb{N}$, $k_i \to +\infty$ such that

$$\|\nabla f(x^{k_i}) - \nabla f(x^{k_i-1})\| > \frac{c_0}{t_{k_i-1}} \|x^{k_i} - x^{k_i-1}\|.$$

By Algorithm 3.1, in this case we have

$$\frac{t_{k_i}}{t_{k_i-1}} = \frac{c_1 \|x^{k_i} - x^{k_i-1}\|}{t_{k_i-1} \|\nabla f(x^{k_i}) - \nabla f(x^{k_i-1})\|} < \frac{c_1}{c_0} \quad \forall k_i.$$

However, Lemma 3.4 gives

$$\lim_{k_i\to+\infty} t_{k_i} = \lim_{k_i\to+\infty} t_{k_i-1} = \lim_{k\to+\infty} t_k = t^*.$$

Consequently, $\frac{t^*}{t^*} \leq \frac{c_1}{c_0} < 1$ that is impossible and we obtain the conclusion of the lemma.

Remark 3.3 From Lemma 3.5, we immediately obtain the increasing of the sequence $\{t_k\}_{k\geq k^*}$, i.e., $t_{k^*} \leq t_k \leq t^*$ for all $k \geq k^*$ and therefore $0 < t_{\min} \leq t_k \leq \max\{t_0, ..., t_{k^*-1}, t^*\} = t_{\max}$, for all $k \geq 0$.

The next lemma plays a crucial role in proving the convergence of Algorithm 3.1 (NPG1).

Lemma 3.6 *For any $x \in \mathrm{dom}(g)$, we have*

$$F(x) - F(x^{k+1}) \geq \frac{1 - c_0}{t_k} \|x^{k+1} - x^k\|^2 + \frac{1}{t_k} \langle x^k - x^{k+1}, x - x^k \rangle, \quad \textit{for all } k \geq k^*. \tag{3.30}$$

Proof From Lemma 2.5 (ii) - inequality (2.2),

$$F(x) - F(x^{k+1}) \geq \langle \nabla f(x^{k+1}) - \nabla f(x^k), x^k - x^{k+1} \rangle + \frac{1}{t_k} \left\|x^{k+1} - x^k\right\|^2 + \frac{1}{t_k} \left\langle x^{k+1} - x^k, x^k - x \right\rangle. \tag{3.31}$$

On the other hand, by using Lemma 3.5, we have the evaluation

$$\begin{aligned} \left\langle \nabla f(x^{k+1}) - \nabla f(x^k), x^k - x^{k+1} \right\rangle &\geq - \left\|\nabla f(x^k) - \nabla f(x^{k+1})\right\| \|x^k - x^{k+1}\| \\ &\geq -\frac{c_0}{t_k} \|x^{k+1} - x^k\|^2 \quad \forall k \geq k^*. \end{aligned} \tag{3.32}$$

The proof is completed by utilizing (3.31) and (3.32).

It is observed that if we substitute x by x^k in (3.30) of Lemma 3.6 and using Remark 3.3 (i) we immediately get the following corollary known as a sufficient decrease type inequality.

Corollary 3.1 (sufficient decrease type inequality) *For all $k \geq k^*$ we have*

$$F(x^k) - F(x^{k+1}) \geq \frac{1-c_0}{t_k}\|x^{k+1} - x^k\|^2.$$

Combining with the fact that $t_{k^*} \leq t_k \leq t^*$ for all $k \geq k^*$ we get the following inequality

$$F(x^k) - F(x^{k+1}) \geq \frac{1-c_0}{t^*}\|x^{k+1} - x^k\|^2 \geq 0 \tag{3.33}$$

and as a consequence,

$$F(x^k) - F(x^{k+1}) \geq \frac{1-c_0}{t_k}\|x^{k+1} - x^k\|^2 = (1-c_0)t_k\|G_{1/t_k}(x^k)\|^2 \geq (1-c_0)t_{k^*}\|G_{1/t_k}(x^k)\|^2, \text{ for all } k \geq k^*. \tag{3.34}$$

Now we are ready to establish the convergence properties of Algorithm 3.1 (NPG1) in the following theorem.

Theorem 3.2 (the convergence of NPG1) *Suppose that Problem (P) satisfies Assumptions 1 and 2, k^* is the fixed number taken from Lemma 3.5. Then, the following assertions hold for Algorithm 3.1.*

(i) The sequence $\{F(x^k)\}_{k\geq k^*}$ descends to $\lim_{k\to+\infty} F(x^k) = F_*$.

(ii) The sequence $\{x^k\}$ converges to an optimal solution of Problem (P).

(iii) For any $x^* \in X^*$ and $k \geq k^* + 1$ we have

$$F(x^k) - F_* = F(x^k) - F(x^*) \leq \frac{D}{2t_{k^*}(k-k^*)} = O\left(\frac{1}{k}\right), \tag{3.35}$$

where

$$D = \max\left\{\|x^* - x^{k^*}\|^2, \|x^* - x^{k^*}\|^2 + \frac{t^*(2c_0-1)}{1-c_0}\left(F(x^{k^*}) - F_*\right)\right\}.$$

Proof (i) By (3.33), the sequence $\{F(x^k)\}_{k\geq k^*}$ is decreasing. On the other hand, it is lower bounded by F_* hence converges to $\overline{F} \geq F_*$. Thus, $F(x^k) - F(x^{k+1}) \to 0$. And consequently, the inequality (3.33) follows

$$\lim_{k\to+\infty}\|x^{k+1} - x^k\| = 0. \tag{3.36}$$

Now, replacing x with x^* in (3.30) of Lemma 3.6 to obtain

$$\begin{aligned} 0 \leq F(x^{k+1}) - F(x^*) &\leq -\frac{1-c_0}{t_k}\|x^{k+1} - x^k\|^2 - \frac{1}{t_k}\langle x^k - x^{k+1}, x^* - x^k\rangle \\ &\leq \frac{(c_0-1)\|x^{k+1} - x^k\|^2 + \|x^{k+1} - x^k\|\|x^k - x^*\|}{t_k}, \quad \text{for all } k \geq k^*. \end{aligned} \tag{3.37}$$

However, $\{x^k\}$ is bounded (by Lemma 3.3 (ii)) and $\lim_{k\to+\infty} t_k = t^*$ (from Lemma 3.4) then combining with (3.36) we deduce that the limit of the right hand side of (3.37) is zero as k tending to infinity. Hence, again, by (3.37) we have $\lim_{k\to+\infty} F(x^k) = F_*$.

(ii) Taking into account that the sequence $\{x^k\}$ is bounded then for each cluster point $\bar{x}$ of $\{x^k\}$, we can take a subsequence $\{x^{k_i}\}$ such that $x^{k_i} \to \bar{x}$. On the other hand, the closedness of F (from

Assumption 1) follows its lower semi-continuity and therefore $F(\bar{x}) \leq \lim_{k_i \to \infty} F(x^{k_i}) = F_*$, which implies $\bar{x} \in X^*$.
Setting $a_k = \|x^{k-1} - x^k\|^2 + \frac{2t_k^2}{t_{k-1}} \left(F(x^{k-1}) - F(x^*)\right) \geq 0$ and rewrite (3.22) to be

$$\|x^{k+1} - x^*\|^2 + a_{k+1} \leq \|x^k - x^*\|^2 + a_k, \ \ \forall x^* \in X^*, \ \ k \geq k_1.$$

Moreover, we have just shown that all cluster points of $\{x^k\}$ belong to X^*. Therefore, applying Lemma 2.6, we obtain that $\{x^k\}$ converges to some element of X^*.

(iii) In (3.33), substituting k by j then summing up it from $j = k^*$ to k we derive that

$$F(x^{k^*}) - F(x^{k+1}) \geq \frac{1 - c_0}{t^*} \sum_{j=k^*}^{k} \|x^{j+1} - x^j\|^2. \tag{3.38}$$

This indicates the convergence of $\sum_{j=k^*}^{+\infty} \|x^{j+1} - x^j\|^2$ and

$$\sum_{j=k^*}^{+\infty} \|x^{j+1} - x^j\|^2 \leq \frac{t^*}{1 - c_0} \left(F(x^{k^*}) - F_*\right). \tag{3.39}$$

Applying (3.30) again with $x = x^*$, we obtain that

$$\begin{aligned} F(x^*) - F(x^{j+1}) &\geq \frac{1}{2t_j} \left(\|x^{j+1} - x^j\|^2 + 2\left\langle x^j - x^{j+1}, x^* - x^j \right\rangle\right) + \left(\frac{1}{2} - c_0\right) \frac{\|x^j - x^{j+1}\|^2}{t_j} \\ &\geq \frac{1}{2t_j} \left(\|x^* - x^{j+1}\|^2 - \|x^* - x^j\|^2\right) + \left(\frac{1}{2} - c_0\right) \frac{\|x^j - x^{j+1}\|^2}{t_j} \quad \forall j \geq k^*. \end{aligned} \tag{3.40}$$

On the other hand, Remark 3.3 (i) gives $t_j \geq t_{k^*} \ \forall j \geq k^*$ which helps to infer the following inequality from (3.40)

$$\begin{aligned} &2t_{k^*} \left(F(x^{j+1}) - F(x^*)\right) \leq 2t_j \left(F(x^{j+1}) - F(x^*)\right) \\ &\leq \left(\|x^* - x^j\|^2 - \|x^* - x^{j+1}\|^2\right) + (2c_0 - 1)\|x^j - x^{j+1}\|^2 \quad \forall j \geq k^*. \end{aligned} \tag{3.41}$$

Summing (3.41) side by side for $j = k^*$ to $k + k^* - 1$ $(k \geq 1)$, we get that

$$\begin{aligned} 2t_{k^*} \left(\sum_{j=k^*}^{k+k^*-1} F(x^{j+1}) - kF(x^*)\right) &\leq \left(\|x^* - x^{k^*}\|^2 - \|x^* - x^{k+k^*}\|^2\right) \\ &\quad + (2c_0 - 1) \sum_{j=k^*}^{k+k^*-1} \|x^j - x^{j+1}\|^2 \\ &\leq \|x^* - x^{k^*}\|^2 + (2c_0 - 1) \sum_{j=k^*}^{k+k^*-1} \|x^j - x^{j+1}\|^2 \qquad (3.42) \\ &\leq D, \qquad (3.43) \end{aligned}$$

where, D is defined by (from (3.39))

$$D = \max\left\{\|x^* - x^{k^*}\|^2, \|x^* - x^{k^*}\|^2 + \frac{t^*(2c_0 - 1)}{1 - c_0} \left(F(x^{k^*}) - F_*\right)\right\}.$$

Additionally, the descent of $\{F(x^k)\}_{k \geq k^*}$ induces $\sum_{j=k^*}^{k+k^*-1} F(x^{j+1}) \geq kF(x^{k+k^*})$. Therefore, by (3.43), we have

$$F(x^{k+k^*}) - F(x^*) \leq \frac{1}{2t_{k^*}} \frac{D}{k} \quad \forall k \geq 1,$$

which means that

$$F(x^k) - F(x^*) \leq \frac{D}{2t_{k^*}} \frac{1}{k-k^*} = O\left(\frac{1}{k}\right) \quad \forall k \geq k^* + 1. \tag{3.44}$$

Next, we discuss the complexity bound of Algorithm 3.1 with respect to tolerance ε in the following remark.

Remark 3.4 From (3.34) we derive that for all $k \geq k^*$,

$$F(x^k) - F_* \geq F(x^{k+1}) - F_* + (1-c_0)t_{k^*} \|G_{1/t_k}(x^k)\|^2, \tag{3.45}$$

Taking an integer number $q \geq k^*$ then summing (3.45) from $k = q$ to $2q-1$ yields that

$$F(x^q) - F_* \geq F(x^{2q}) - F_* + (1-c_0)t_{k^*} \sum_{k=q}^{2q-1} \|G_{1/t_k}(x^k)\|^2. \tag{3.46}$$

Now, remember that $F(x^{2q}) - F_* \geq 0$ and combining (3.46) with (3.44) we obtain

$$(1-c_0)t_{k^*} \sum_{k=q}^{2q-1} \|G_{1/t_k}(x^k)\|^2 \leq \frac{D}{2t_{k^*}} \frac{1}{q-k^*}. \tag{3.47}$$

Hence

$$\min_{k=q,\ldots,2q-1} \|G_{1/t_k}(x^k)\|^2 \leq \frac{D}{2(1-c_0)t_{k^*}^2} \frac{1}{q(q-k^*)}. \tag{3.48}$$

Therefore, we have for any $K \geq 2k^* + 1$,

$$\min_{k=0,\ldots,K} \|G_{1/t_k}(x^k)\| \leq \sqrt{\frac{2D}{(1-c_0)t_{k^*}^2} \frac{1}{K(K-2k^*)}} = O\left(\frac{1}{K}\right). \tag{3.49}$$

The final result of this section establishes a stronger convergence rate of Algorithm 3.1 if f is locally strongly convex. The detail is as follows.

Theorem 3.3 *Assuming that $c_0 \leq \frac{1}{2}$ and f is locally strongly convex then under Assumptions 1, 2, the sequence $\{x^k\}$ generated by Algorithm 3.1 satisfies*

$$\|x^{k+1} - x^*\|^2 \leq (1 - \sigma t_{k^*}) \|x^k - x^*\|^2, \ \ \forall k \geq k^*, \tag{3.50}$$

where, $x^* \in X^*$ and $\sigma > 0$ is strong convexity constant of f on the compact set $T = \overline{conv}\left(\{x^*, x^0, x^1, \ldots\}\right)$. Consequently, this result shows the Q-linear convergence rate of $\{x^k\}_{k \geq k^*}$.

Proof The $\sigma-$ strong convexity on T of f implies that

$$f(x) - f(x^k) \geq \langle \nabla f(x^k), x - x^k \rangle + \frac{\sigma}{2} \|x - x^k\|^2, \ \forall x \in T.$$

We update this change and the condition $c_0 \leq \frac{1}{2}$ in the arguments of formulas (3.31) and (3.40) to obtain the following inequality

$$F(x^*) - F(x^{k+1}) \geq \frac{1}{2t_k} \|x^* - x^{k+1}\|^2 + \left(\frac{\sigma}{2} - \frac{1}{2t_k}\right) \|x^* - x^k\|^2,$$

for $x^* \in X^*, k \geq k^*$. Remember that $F(x^*) - F(x^{k+1}) \leq 0$ for all k, hence

$$\frac{1}{2t_k} \|x^* - x^{k+1}\|^2 \leq \left(\frac{1}{2t_k} - \frac{\sigma}{2}\right) \|x^* - x^k\|^2, \ \ k \geq k^*. \tag{3.51}$$

By (3.51), Lemma 3.4(i) and Remark 3.3 (i), we have: for all $k \geq k^*$

$$0 < 1 - \sigma t_k \leq 1 - \sigma t_{k^*} \leq 1 - \sigma t_{\min} < 1,$$

which derives

$$\|x^{k+1} - x^*\|^2 \leq (1 - \sigma t_{k^*}) \|x^k - x^*\|^2, \ \ k \geq k^*.$$

The last inequality demonstrates the Q-linear convergence rate of $\{x^k\}_{k \geq k^*}$.

Remark 3.5 Throughout the convergence analysis, we see that the coefficients c_0, c_1 in NPG1 (Algorithm 3.1) can be selected more flexibly at each iteration, i.e., for iteration k, they are given by c_{0k} and c_{1k} provided that $0 < c_{1k} < c_{0k} < \frac{1}{\sqrt{2}}$.

4 For a class of the nonconvex case of Problem (P)

We now consider Problem (P) satisfying *Assumption 1* and other conditions in *Assumption 3* below

Assumption 3 *(i) f has a globally Lipschitz gradient with constant L_f on* $\mathrm{dom}(g)$.

(ii) For $u, v \in \mathrm{dom}(g)$, the function $h_{uv} : [0,1] \to \mathbb{R}$ defined by

$$h_{uv}(t) = f_t'(u+t(v-u)) = \langle \nabla f(u+t(v-u)), v-u \rangle$$

is quasiconvex.

Example 4.1 Suppose that f is either convex or concave. Then f satisfies *Assumption 3 (ii)*. Indeed, the convexity (concavity, resp.) of f follows the convexity (concavity, resp.) of $f(u+t(v-u))$ on the set $\{t \in \mathbb{R} \mid u+t(v-u) \in \mathrm{dom}(g)\} \supset [0,1]$ (since $\mathrm{dom}(g)$ is convex). As a result, $f_t'(u+t(v-u))$ is increasing (decreasing, resp.) monotone over $[0,1]$ and therefore quasiconvex on that. In the case where f is concave then $F = f + g$ is actually the difference of the two convex functions, or in other words, F belongs to the class of functions that can be represented as the difference of two convex functions (*DC functions*).

Example 4.2 The indefinite quadratic function $f(x) = \frac{1}{2}x^T Ax + b^T x$ (A is a symmetric matrix in $\mathbb{R}^{n\times n}$ and $b \in \mathbb{R}^n$) satisfies *Assumption 3* since it has $L_f = \|A\|$ and $h_{uv}(t) = \langle A(u+t(v-u)) + b, v-u \rangle$ is linear and hence quasiconvex on $[0,1]$ for any $u, v \in \mathrm{dom}(g)$.

Example 4.3 Considering Problem (P) with

(i) $f(x) = \dfrac{x^T Ax + b^T x + c}{p^T x + q}$, where A is a real symmetric matrix in $\mathbb{R}^{n\times n}$ and $b, p \in \mathbb{R}^n$, $c, q \in \mathbb{R}$;

(ii) $g = \iota_C$ (the indicator function of C), where $C \subset \mathbb{R}^n$ is a closed and convex set such that $p^T x + q > 0 \ \forall x \in C$.

Then (P) satisfies Assumption 3 (ii). One can see the detailed proof in Appendix A.2.

From Examples 4.1, 4.2 and 4.3 we see that the class of Problem (P) satisfying *Assumption 1* and *Assumption 3* is nonconvex in general. Subsequently, we propose an other version of Algorithm 3.1 that can be applied for such a kind of problems more effectively.

Similar to the classical analysis under the nonconvex setting of f, we can show that for Algorithm 4.1, the norm of the gradient mapping converges to zero and that all the limit points generated by the algorithm are stationary points of (P). We begin by presenting several preparatory lemmas.

Lemma 4.1 *The sequence $\{t_k\}$ in Algorithm 4.1 satisfies $\inf\limits_{k\geq 0} t_k > 0$ and has a positive limit, i.e., $\lim\limits_{k\to+\infty} t_k = t^* > 0$.*

Proof Similarly to Lemma 3.4 (i), it is clear that $t_k \geq \min\left\{t_0, \frac{c_1}{L_f}\right\} > 0$ for all $k \geq 0$. As a result, $\inf\limits_{k\geq 0} t_k > 0$. The remaining conclusion is shown as Lemma 3.4 (ii).

Algorithm 4.1 (NPG2)

Step 0 (Initialization). Select $t_0 > 0$, $0 < c_1 < c_0 < 1$, $x^0 \in \text{dom}(g)$, a tolerance $\varepsilon > 0$ and a positive real sequence $\{\gamma_k\}$ such that $\sum_{k=0}^{+\infty} \gamma_k < \infty$. Taking $x^1 = \text{Prox}_{t_0 g}(x^0 - t_0 \nabla f(x^0)), t_{-1} = t_0$ and $k = 1$.

Step 1.

$$\begin{aligned}
&\textbf{If} \quad \|\nabla f(x^k) - \nabla f(x^{k-1})\| > \frac{c_0}{t_{k-1}} \|x^k - x^{k-1}\| \\
&\textbf{then} \quad t_k = c_1 \frac{\|x^k - x^{k-1}\|}{\|\nabla f(x^k) - \nabla f(x^{k-1})\|} \qquad (4.1) \\
&\textbf{else} \quad \gamma'_{k-1} = \gamma_{k-1} \\
&\qquad \text{if } \frac{t_{k-1}}{t_{k-2}} < 1 \text{ then } \gamma'_{k-1} = \min\left\{\gamma_{k-1}, \sqrt{1 + \frac{t_{k-1}}{t_{k-2}}} - 1\right\} \qquad (4.2) \\
&\qquad t_k = (1 + \gamma'_{k-1}) t_{k-1}.
\end{aligned}$$

Step 2. Compute $x^{k+1} = \text{Prox}_{t_k g}(x^k - t_k \nabla f(x^k))$.
Step 3. If $\|G_{1/t_k}(x^k)\| = \|x^k - x^{k+1}\|/t_k < \varepsilon$ **then** STOP **else** setting $k := k + 1$ and return to **Step 1**.

Lemma 4.2 *For Algorithm 4.1, there exists $\bar{k}$ such that*

$$\|\nabla f(x^k) - \nabla f(x^{k-1})\| \le \frac{c_0}{t_{k-1}} \|x^k - x^{k-1}\| \ \ \forall k \ge \bar{k}.$$

Consequently, $0 < \inf_{k \ge 0} t_k \le t_{\bar{k}} \le t_k \le t_{k+1} \le t^*$ *for all* $k \ge \bar{k}$.

Proof The proof is similar to that of Lemma 3.5.

The following lemma presents the sufficient decrease type inequality - a key step to obtain the convergence results of our algorithms.

Lemma 4.3 *Assuming that Problem (P) satisfies Assumption 1 and Assumption 3 then the sequence $\{x^k\}$ generated by Algorithm 4.1 has the following property*

$$F(x^k) - F(x^{k+1}) \ge t_{\bar{k}}(1 - c_0)\|G_{1/t^*}(x^k)\|^2, \ \forall k \ge \bar{k}.$$

Proof Invoking the Fundamental Theorem of Calculus, we have

$$\begin{aligned}
f(x^{k+1}) - f(x^k) &= \int_0^1 \left\langle \nabla f(x^k + t(x^{k+1} - x^k)), x^{k+1} - x^k \right\rangle dt \\
&= \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \int_0^1 u_k(t) dt, \ \forall k \ge \bar{k}, \qquad (4.3)
\end{aligned}$$

where

$$\begin{aligned}
u_k(t) &= \langle \nabla f(x^k + t(x^{k+1} - x^k)) - \nabla f(x^k), x^{k+1} - x^k \rangle \\
&= h_{x^k x^{k+1}}(t) - \langle \nabla f(x^k), x^{k+1} - x^k \rangle.
\end{aligned}$$

According to *Assumption 3*, the quasiconvexity of $u_k(t)$ in $[0,1]$ follows that

$$\begin{aligned}
u_k(t) &\le \max\{u_k(0), u_k(1)\} = \max\{0, u_k(1)\} \le |u_k(1)| \\
&= |\langle \nabla f(x^{k+1}) - \nabla f(x^k), x^{k+1} - x^k \rangle|, \ \forall t \in [0,1].
\end{aligned}$$

Thereafter, using Lemma 4.2, we derive that

$$\int_0^1 u_k(t) dt \le \frac{c_0}{t_k} \|x^{k+1} - x^k\|^2, \ \ \forall k \ge \bar{k}. \qquad (4.4)$$

Now, combining (4.3), (4.4) and Lemma 2.5 (ii) with $x = x^{k+1}$ we get that

$$\begin{aligned} F(x^k) - F(x^{k+1}) &= f(x^k) - f(x^{k+1}) + g(x^k) - g(x^{k+1}) \\ &\geq -\left\langle x^{k+1} - x^k, \nabla f(x^k) \right\rangle - \frac{c_0}{t_k}\|x^{k+1} - x^k\|^2 + \left\langle x^{k+1} - x^k, \nabla f(x^k) + \frac{x^{k+1} - x^k}{t_k} \right\rangle \\ &= \frac{1 - c_0}{t_k}\|x^{k+1} - x^k\|^2 \\ &\geq t_{\bar{k}}(1 - c_0)\|G_{1/t_k}(x^k)\|^2 \qquad (4.5) \\ &\geq t_{\bar{k}}(1 - c_0)\|G_{1/t^*}(x^k)\|^2, \quad \forall k \geq \bar{k}, \qquad (4.6) \end{aligned}$$

where the last inequality uses the monotonicity of the gradient mapping (Lemma 2.3) and the fact $t_{\bar{k}} \leq t_k \leq t^*$ from Lemma 4.2.

Theorem 4.1 *Under Assumptions 1 and 3, the following assertions hold for Algorithm 4.1:*

(i) The sequence $\{F(x^k)\}_{k \geq \bar{k}}$ is decreasing and for any $k \geq \bar{k}$, $F(x^{k+1}) < F(x^k)$ unless x^k is a stationary point of Problem (P).

(ii) $\|G_{1/t^*}(x^k)\| \to 0$ as $k \to +\infty$.

(iii)

$$\min_{\bar{k} \leq k \leq K} \|G_{1/t^*}(x^k)\| \leq \min_{\bar{k} \leq k \leq K} \|G_{1/t_k}(x^k)\| \leq \sqrt{\frac{F(x^{\bar{k}}) - F_*}{t_{\bar{k}}(1 - c_0)(K - \bar{k} + 1)}} = O\left(\frac{1}{\sqrt{K}}\right) \quad \forall K \geq \bar{k}.$$

(iv) All limit points of the sequence $\{x^k\}$ are stationary points of Problem (P).

Proof (i) By (4.6) and $c_0 < 1$, it is clear to see that $F(x^k) \geq F(x^{k+1})$ for all $k \geq \bar{k}$. If $F(x^k) = F(x^{k+1})$ then $\|G_{1/t^*}(x^k)\| = 0$ meaning x^k is a stationary point of (P) (by Lemma 2.2).

(ii) Since Problem (P) has a non-empty optimal solution set then the sequence $\{F(x^k)\}_{k \geq \bar{k}}$ is decreasing and lower bounded by F_*, moreover it is non-increasing so it converges. Thus $F(x^k) - F(x^{k+1}) \to 0$. Combined with (4.6), we obtain $\|G_{1/t^*}(x^k)\| \to 0$ as $k \to +\infty$.

(iii) Summing up (4.5) from $\bar{k}$ to K we get that

$$\begin{aligned} F(x^{\bar{k}}) - F(x^{K+1}) \geq t_{\bar{k}}(1 - c_0) \sum_{k=\bar{k}}^{K} \|G_{1/t^*}(x^k)\|^2 &\geq t_{\bar{k}}(1 - c_0)(K - \bar{k} + 1) \min_{\bar{k} \leq k \leq K} \|G_{1/t_k}(x^k)\|^2 \\ &\geq t_{\bar{k}}(1 - c_0)(K - \bar{k} + 1) \min_{\bar{k} \leq k \leq K} \|G_{1/t^*}(x^k)\|^2. \end{aligned}$$

The fact $F(x^{K+1}) \geq F_*$ completes the proof.

(iv) Let $\widehat{x}$ be a limit point of $\{x^k\}_{k \geq 0}$. Then there exists a subsequence $\{x^{k_i}\}$ which converges to $\widehat{x}$. From here for any $k_i > 0$,

$$\|G_{1/t^*}(\widehat{x})\| \leq \|G_{1/t^*}(\widehat{x}) - G_{1/t^*}(x^{k_i})\| + \|G_{1/t^*}(x^{k_i})\| \leq \left(\frac{2}{t^*} + L_f\right)\|\widehat{x} - x^{k_i}\| + \|G_{1/t^*}(x^{k_i})\|, \qquad (4.7)$$

where the last inequality obtained from the Lipschitz continuity of the gradient mapping (Lemma 2.4). Since the right-hand side of (4.7) tends to 0 as $k_i \to +\infty$, we conclude that $G_{1/t^*}(\widehat{x}) = 0$, i.e., $\widehat{x}$ is a stationary point of Problem (P).

Remark 4.1 (i) Remember that $c_0, c_1 \in \left(0, \frac{1}{\sqrt{2}}\right)$ for Algorithm 3.1 (NPG1) but $c_0, c_1 \in (0,1)$ for Algorithm 4.1 (NPG2). This difference comes from the challenge of local Lipschitzness condition imposed on ∇f for NPG1. Intuitively, without the global Lipschitz condition of ∇f, the variation of ∇f with respect to x can be very large if we move a long step (which could be given by larger c_0, c_1) then the restriction of c_0, c_1 in $\left(0, \frac{1}{\sqrt{2}}\right)$ ensures the boundedness of the sequence $\{x^k\}$, thus avoiding the uncontrollable situation of the gradient. Theoretically, both NPG1 and NPG2 need to verify that the lower boundedness of $\{t_k\}$ is a postive number; this can be derived from the global Lipschitzness of ∇f on $T = \overline{conv}\left(\{x^*, x^0, x^1 ...\}\right)$. Therefore, to use the locally Lipschitz gradient of f, NPG1 has to provide the compactness of T that given by c_0, c_1 in $\left(0, \frac{1}{\sqrt{2}}\right)$. However, NPG2 works with the function f satisfying ∇f be globally Lipschitz then c_0, c_1 are just chosen to ensure the descent of the sequence of objective value $\{F(x^k)\}_{k \geq \bar{k}}$ as presented in (4.6).

(ii) Actually, the command (4.2) in Algorithm 4.1 is optional since we do not need it during the proof of the convergence of NPG2. However, through out the numerical experiments, we realize that this step improves the performance of the algorithm.

5 Problem (P) with quadratic function f

In this section, we propose an extension of NPG2 called *NPG-quad* solving Problem (P) with the quadratic function f, i.e., $f(x) = \frac{1}{2}x^T Ax + b^T x$ as described in Example 4.2. The changes compared with NPG2 are in the two points:

(i) Firstly,

$$t_k(\text{ of NPG-quad, in (5.2)}) = \frac{c_1\|x^k - x^{k-1}\|^2}{(x^k - x^{k-1})^T A(x^k - x^{k-1})} \geq \frac{c_1\|x^k - x^{k-1}\|}{\|Ax^k - Ax^{k-1}\|} = t_k(\text{ of NPG2, in (4.1)});$$

(ii) Secondly, c_0, c_1 in $(0,2)$ for NPG-quad while c_0, c_1 in $(0,1)$ for NPG2 . This extension stems from the new formula for t_k, as discussed above, and the special quadratic structure of f, which allows a better evaluation of $F(x^{k+1}) - F(x^k)$ as shown in (5.6) and (5.7).

These points probably make the stepsize of NPG-quad larger and therefore shorten the execution time compared to NPG1 and NPG2 .

Lemma 5.1 *The sequence $\{t_k\}$ generated by Algorithm 5.1 has a positive limit, i.e., $\lim_{k \to +\infty} t_k = t^* > 0$.*

Proof Analogous to former sections, we are easy to have $t_k \geq \min\left\{t_0, \frac{c_1}{\|A\|}\right\} > 0$ for all $k \geq 0$. Therefore, $\inf_{k \geq 0} t_k > 0$. The computation of t_k by (5.2) or (5.4) provides $\ln\left(\frac{t_{k+1}}{t_k}\right) < \ln(1 + \gamma_k)$. The subsequent arguments are akin to the one of Lemma 3.4.

Lemma 5.2 *For Algorithm 5.1, there exists $\tilde{k}$ such that*

$$(x^k - x^{k-1})^T A(x^k - x^{k-1}) \leq c_0 \frac{\|x^k - x^{k-1}\|^2}{t_{k-1}}, \text{ for all } k \geq \tilde{k}. \tag{5.5}$$

Consequently, $0 < \inf_{k \geq 0} t_k \leq t_{\tilde{k}} \leq t_k \leq t_{k+1} \leq t^*$ for all $k \geq \tilde{k}$.

Proof Based on the properties of $\{t_k\}$ in Lemma 5.1 and arguing by contradiction as Lemma 3.5 we have the desired conclusion.

Algorithm 5.1 (NPG-quad)

Step 0 (Initialization). Select $t_0 > 0$, $0 < c_1 < c_0 < 2$, $x^0 \in \mathrm{dom}(g)$, a tolerance $\varepsilon > 0$ and a positive real sequence $\{\gamma_k\}$ such that $\sum_{k=0}^{+\infty} \gamma_k < +\infty$. Taking $x^1 = \mathrm{Prox}_{t_0 g}(x^0 - t_0 \nabla f(x^0)), t_{-1} = t_0$, and $k = 1$.

Step 1.

$$\textbf{If} \quad (x^k - x^{k-1})^T A(x^k - x^{k-1}) > c_0 \frac{\|x^k - x^{k-1}\|^2}{t_{k-1}} \tag{5.1}$$

$$\textbf{then} \quad t_k = \frac{c_1 \|x^k - x^{k-1}\|^2}{(x^k - x^{k-1})^T A(x^k - x^{k-1})} \tag{5.2}$$

$$\textbf{else} \quad \gamma'_{k-1} = \gamma_{k-1}$$

$$\text{if } \frac{t_{k-1}}{t_{k-2}} < 1 \text{ then } \gamma'_{k-1} = \min\left\{\gamma_{k-1}, \sqrt{1 + \frac{t_{k-1}}{t_{k-2}}} - 1\right\} \tag{5.3}$$

$$t_k = (1 + \gamma'_{k-1}) t_{k-1}. \tag{5.4}$$

Step 2. Compute $x^{k+1} = \mathrm{Prox}_{t_k g}(x^k - t_k \nabla f(x^k))$.
Step 3. If $\|G_{1/t_k}(x^k)\| = \|x^k - x^{k+1}\|/t_k < \varepsilon$ **then** STOP **else** setting $k := k+1$ and return to **Step 1**.

Theorem 5.1 *Suppose that Problem (P) satisfies Assumption 1 and f has a quadratic form as in Example 4.2. Let $\{x^k\}$ be the sequence generated by Algorithm 5.1. Then the sequence $\{F(x^k)\}_{k \geq \tilde{k}}$ is nonincreasing and $\|G_{1/t^*}(x^k)\| \xrightarrow{k \to +\infty} 0$. Additionally,*

$$\min_{\tilde{k} \leq k \leq K} \|G_{1/t^*}(x^k)\| \leq \min_{\tilde{k} \leq k \leq K} \|G_{1/t_k}(x^k)\| \leq \sqrt{\frac{F(x^{\tilde{k}}) - F_*}{t_{\tilde{k}}(1 - \frac{c_0}{2})(K - \tilde{k} + 1)}} = O\left(\frac{1}{\sqrt{K}}\right), \ \ \forall K \geq \tilde{k}$$

and any accumulation point of $\{x^k\}$ is a stationary point of (P).

Proof We have

$$\begin{aligned} f(x^{k+1}) - f(x^k) &= \int_0^1 \left\langle \nabla f(x^k + t(x^{k+1} - x^k)), x^{k+1} - x^k \right\rangle dt \\ &= \int_0^1 \left\langle A(x^k + t(x^{k+1} - x^k)) + b, x^{k+1} - x^k \right\rangle dt \\ &= \left\langle A(x^{k+1} - x^k), x^{k+1} - x^k \right\rangle \int_0^1 t dt + \left\langle Ax^k + b, x^{k+1} - x^k \right\rangle \\ &= \frac{1}{2}(x^{k+1} - x^k)^T A(x^{k+1} - x^k) + \left\langle \nabla f(x^k), x^{k+1} - x^k \right\rangle. \end{aligned} \tag{5.6}$$

Now plugging (5.6) in $F(x^k) - F(x^{k+1})$ and using Lemma 2.5 (ii) to obtain

$$\begin{aligned} F(x^k) - F(x^{k+1}) &= f(x^k) - f(x^{k+1}) + g(x^k) - g(x^{k+1}) \\ &\geq -\frac{1}{2}(x^{k+1} - x^k)^T A(x^{k+1} - x^k) - \langle \nabla f(x^k), x^{k+1} - x^k \rangle \\ &\quad + \left\langle x^{k+1} - x^k, \nabla f(x^k) + \frac{x^{k+1} - x^k}{t_k} \right\rangle \\ &= -\frac{1}{2}(x^{k+1} - x^k)^T A(x^{k+1} - x^k) + \frac{1}{t_k}\|x^{k+1} - x^k\|^2. \end{aligned} \tag{5.7}$$

Next, applying Lemma 5.2 for (5.7) we obtain for all $k \geq \tilde{k}$,

$$F(x^k) - F(x^{k+1}) \geq \left(1 - \frac{c_0}{2}\right) \frac{\|x^{k+1} - x^k\|^2}{t_k} = t_k \left(1 - \frac{c_0}{2}\right) \|G_{1/t_k}(x^k)\|^2 \geq t_{\tilde{k}} \left(1 - \frac{c_0}{2}\right) \|G_{1/t_k}(x^k)\|^2 \tag{5.8}$$

$$\geq t_{\tilde{k}} \left(1 - \frac{c_0}{2}\right) \|G_{1/t^*}(x^k)\|^2. \tag{5.9}$$

The remaining arguments are similar to those of Theorem 4.1.

Remark 5.1 If f is a concave quadratic function i.e., A is negative semi-definite then the condition (5.1) is false, hence

- $\tilde{k}$ in Lemma 5.2 should be zero;
- t_k is always defined by formula (5.4) and $\{t_k\}_{k\geq 0}$ is increasing to a finite limit;
- the evaluation (5.8) should be

$$F(x^k) - F(x^{k+1}) \geq \frac{\|x^{k+1} - x^k\|^2}{t_k}, \ \forall k \geq 0. \tag{5.10}$$

6 Numerical experiments

In this section, we investigate the performance of our new stepsize for the proximal gradient scheme by comparing NPG1 (Algorithm 3.1), NPG2 (Algorithm 4.1) and NPG-quad (Algorithm 5.1) with the recent related algorithms including:

- the AdPG proposed by Malitsky and Mishchenko [25] (Algorithm 3 in [25]);
- the AdaPG^{q,r} from Latafat et al. in [20] using $(q,r) = \left(\frac{3}{2}, \frac{3}{4}\right)$;
- the proximal gradient algorithms with stepsize selection based on an improved version of Armijo's backtracking procedure[1], denoted by PG-LS(s,r) where (s,r) equals $(1.1, 0.5)$ or $(1.2, 0.5)$.

To ensure fairness, all of the parameters chosen for PG-LS and AdaPG^{q,r} are the ones with the most stable and efficient empirical performance reported in [25], [20].

For our algorithms, we use the convergent series $\sum_{k=0}^{+\infty} \gamma_k$ defined by

$$\gamma_{k-1} = \frac{0.1(\ln k)^{5.7}}{k^{1.1}}, \quad \forall k \geq 1, \tag{6.1}$$

and $(c_0, c_1) = (0.7, 0.69)$ for NPG1 , $(c_0, c_1) = (0.99, 0.98)$ for NPG2 and NPG-quad .

Discussions on parameters setting.

The parameter choices for the NPG algorithms are guided by a simple yet consistent intuition: choosing values that allow the stepsize to be as large as possible while maintaining algorithmic stability. Specifically:

- The first factor that affects the magnitude of our stepsize is (c_0, c_1). For NPG1 and NPG2, we select (c_0, c_1) to be as close as possible to their theoretical limits - $(0.7, 0.69)$ and $(0.99, 0.98)$, respectively. This maximizes the stepsize without violating convergence guarantees.
- In the case of NPG-quad, despite having a much larger allowable range for (c_0, c_1), we observed that choosing values that are too large can lead to unstable behavior (sometimes resulting in surprisingly fast convergence for problems requiring tens of thousands of iterations, but often leading to poor performance on easier problems that need only a few hundred iterations). To ensure consistent performance across problem scales, we again choose $(0.99, 0.98)$ as a balanced and robust setting.

[1] For $s > 1$, $r < 1$, Armijo's line search in finds the largest $t_k = sr^i t_{k-1}$ for $i = 0, 1, ...$ such that $f(x^{k+1}) \leq f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{1}{2t_k}\|x^{k+1} - x^k\|^2$.

- The second factor that helps increase the stepsize is the sequence γ_k, a natural choice to satisfy our assumption is $\frac{1}{k^p}$ with p close to 1. We could further enlarge this sequence by changing the numerator adaptively, for example [23] proposed $\gamma_k = \frac{w_k}{k^p}$ with w_k chosen adaptively. We instead opt for the explicit form $\gamma_{k-1} = \frac{a(\ln k)^b}{k^p}$ similar to [17]. Stick to our guiding intuition, we choose p close to 1, b large to generate gradually large γ_k and a small to ensure stability. This choice also eliminates the need for computing w_k while increasing the stepsize effectively and leads to empirical improvements.

Experimental setup and details[2]

We conduct experiments on seven representative composite optimization problems, considering various problem sizes and using 10 randomly generated datasets for each instance. The data generation scheme for each problem is described in the corresponding subsections, while detailed dataset statistics are provided in Appendix C. Below we describe the criteria used to evaluate the performance of the algorithms:

1. the number of iterations (*Iter.*);
2. $\|G_{1/t_k}(x^k)\| = \|x^{k+1} - x^k\|/t_k$ (*Res.*);
3. $F(x^k) - F_*$ (*Obj.*), where F_* is computed as the minimum of $F(x^k)$ over all iterations and all tested algorithms;
4. the running time in seconds (*Time(s)*).

For all implemented algorithms, the stopping criterion is either the residual $\|G_{1/t_k}(x^k)\| \leq 1e-06$ or if the maximum of N_{max} iterations is reached. All algorithms use the same initial t_0. In particular, for problems Lasso, Min-length, BCQP, and NMF, we adopt the line search procedure in [25] to determine the initial stepsize. For problems Max-likelyhood, Dual-max-entropy, and BCFP, where stability is crucial, we instead use a small constant stepsize of 0.001.

All experiments were implemented in Python and executed on a computer equipped with a 12th Gen Intel(R) Core(TM) i7-1260P 2.10 GHz processor.

Experimental results

The experimental results are summarized using performance profiles given by Dolan and Jorge [12] (see Appendix B for more details on performance profiles). The metrics used in the performance profiles include the total number of iterations required for an algorithm to converge, the running time, the residual and objective value at the final iteration[3]. The performance profiles for the seven problems are presented in Figures 1, 2, 3, 4, 5, 6, and 7.

To illustrate the convergence behavior of the algorithms, we present iteration–objective and iteration–stepsize plots for selected instances of each problem in Figure 8, 9 in Appendix C. The selected instances are large-scale and relatively difficult (as indicated by the iteration counts), with the random generator seed fixed at 1. For nonconvex problems, where algorithms may converge to different solutions, we choose instances for which all algorithms terminate with similar objective values to ensure comparability.

Across a wide range of test instances, under identical stopping criteria, NPG1, NPG2, and NPG-quad consistently require fewer iterations and shorter runtimes, while typically achieving lower objective values than the competing methods. In particular, the values of $\rho(1)$ in the performance profiles indicate that the NPG algorithms attain the lowest iteration counts and shortest runtimes for the majority of test instances. Moreover, in some cases, the NPG methods require approximately 1.5 to 2 times fewer iterations and runtime compared with other approaches. Notably, for certain nonconvex problems where different methods may converge to different local minima, NPG2 and NPG-quad exhibit a clear advantage in identifying solutions with lower objective values compared to the competing approaches.

[2] Codes are available at https://github.com/hoaiphamthi/NPG-for-composite-optimization-problems.

[3] A value $\varepsilon = 10^{-20}$ is added to the residual and objective values to ensure they are strictly positive.

6.1 Lasso problems

The formulation of Lasso problem is formulated as the ℓ_1 regularized least squares

$$\min_{x\in\mathbb{R}^n} \quad \frac{1}{2}\|Ax-b\|^2+\lambda\|x\|_1, \tag{Lasso}$$

where $A\in\mathbb{R}^{m\times n}, b\in\mathbb{R}^m$. The applications of Lasso can be found in statistic, machine learning, signal processing, see e.g., [3,14]. By using the similar rules in [14], we randomly generate $A\in\mathbb{R}^{m\times n}$ with entries drawn from the normal distribution $\mathcal{N}(0,1)$. We then construct a sparse solution x^* with 5% approximately non-zero entries, drawn from a mixture distribution $\mathcal{N}(0,1)\times B(1,0.05)$ then setting $b=Ax^*+\delta$, where δ is white Gaussian noise with variance 0.01. The regularization term $\lambda=0.01\|A^Tb\|_\infty$. Obviously, Lasso satisfies *Assumptions 1, 2, 3* then both of NPG1 and NPG2 are available for it. Moreover, f is quadratic hence NPG-quad can be applied for solving this problem formally. Figure 1 illustrates the performance of mentioned algorithms for 80 test instances of 8 different sizes. As shown by the figure, NPG-quad shows superior performance by achieving the lowest number of iterations and running time in more than 80% of the datasets, while NPG1 and NPG2 also show decent efficiency by achieving performance of within a factor of 1.5 of the best performance in roughly 80% of the datasets compared to about 10-50% of other methods.

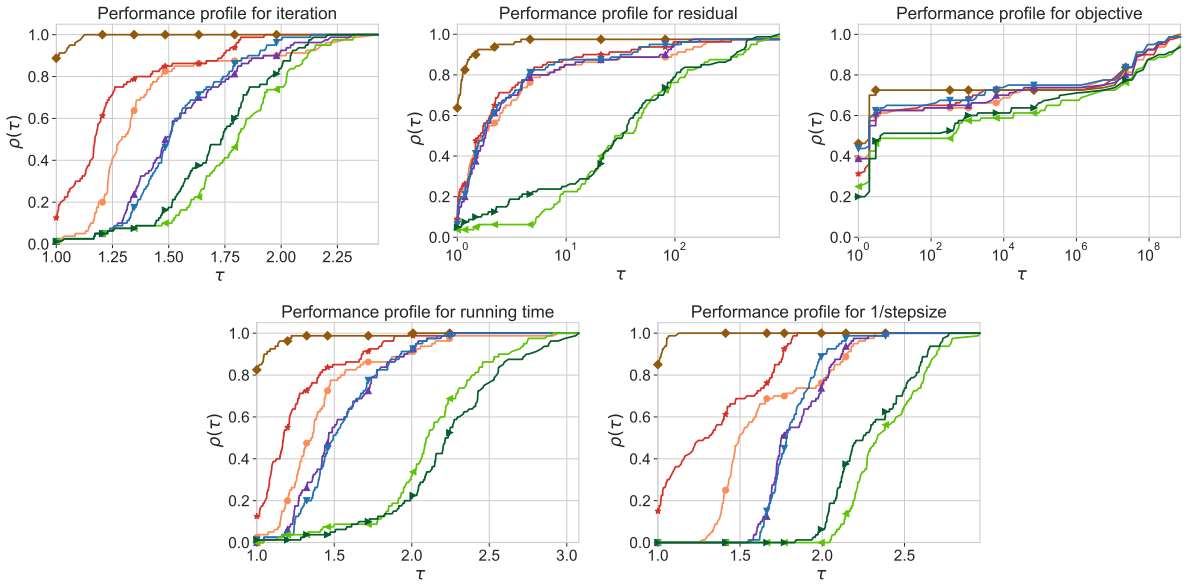

Fig. 1: Performance profiles for Lasso problem on 80 datasets.

6.2 Minimum length piecewise-linear curve subject to equality constraints

We consider an other optimization problem from [10, Example 10.4], where the objective is minimizing the length of the piecewise-linear curve connecting the points $(0,0),(1,x_1),...,(n,x_n)$ such that $Ax=b$, where $A\in\mathbb{R}^{m\times n}, b\in\mathbb{R}^m$. The problem therefore can be formed as

$$\min \sqrt{1+x_1^2}+\sum_{i=1}^{n-1}\sqrt{1+(x_{i+1}-x_i)^2} \quad \text{s.t.} \quad Ax=b. \tag{Min-length}$$

It is seen that Min-length[4] satisfies *Assumptions 1,2,3* and we can use NPG1 and NPG2 to solve it exactly. In the implementation, all members of A are randomly generated by normal distribution

[4] Min-length is a case of problem (P) with $f(x)=\sqrt{1+x_1^2}+\sum_{i=1}^{n-1}\sqrt{1+(x_{i+1}-x_i)^2}$ and $g=\iota_C$ (the indicator function of C) with $C=\{x\in\mathbb{R}^n \mid Ax=b\}$.

$\mathcal{N}(0,1)$. Taking $b = Ax^*$, where $x^* \sim \mathcal{N}(0,1)$. Figure 2 provides the performance profiles of the selected algorithms on 60 datasets of 6 different sizes. Notably, both NPG1 and NPG2 outperform the remaining ones with the big deviation in term of computational time, residual, objective value and the number of iterations.

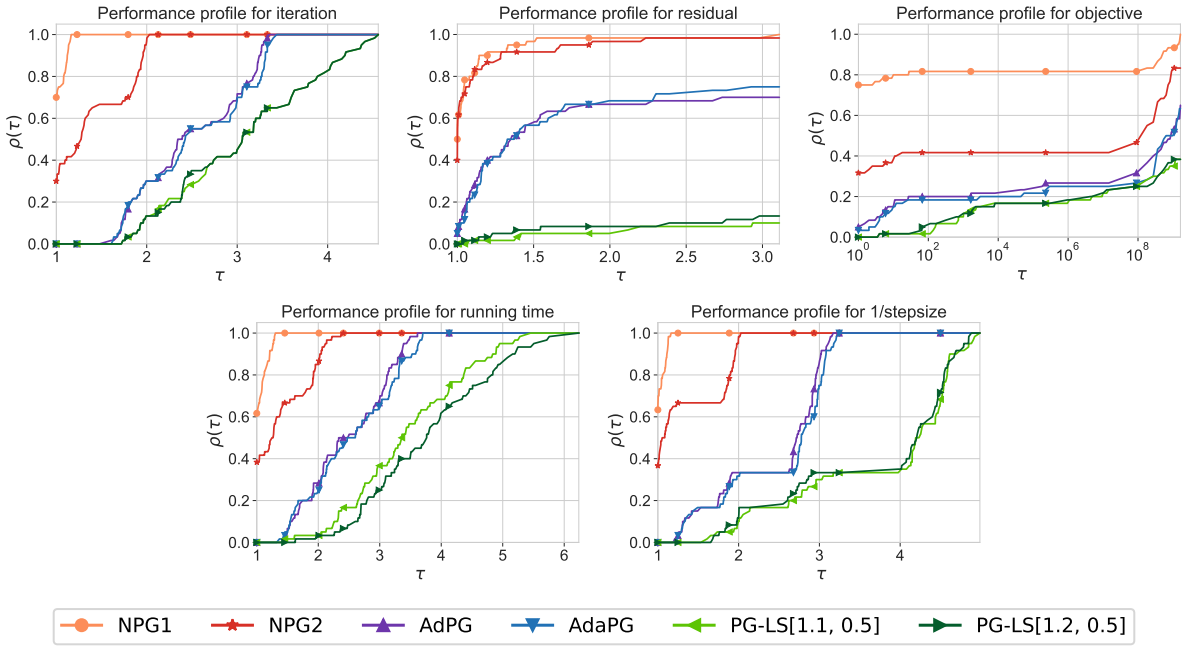

Fig. 2: Performance profiles for Min-length problem on 60 datasets.

6.3 Maximum likelihood estimate of the information matrix

This problem (see [10, Eq. (7.5)]) aims to estimate the inverse of a covariance matrix Y of a multivariate random variable subject to the eigenvalue bounds given some samples of the random variable. The problem can be formulated as

$$\min \quad f(X) = -\log\det(X) + \operatorname{tr}(XY) \quad \text{s.t.} \quad X \in \mathbb{S}_n \ \text{ and } \ lI \preceq X \preceq uI. \qquad \text{(Max-likelyhood)}$$

Here $\mathbb{S}_n$ denotes the space of real symmetric matrices of dimension $n \times n$, and $A \preceq B$ indicates that $B - A$ is positive semi-definite. Observably, Max-likelyhood[5] satisfies *Assumptions 1,2,3* then NPG1 and NPG2 are exact methods to solve Max-likelyhood. The dataset for the implementation is generated analogously to [25] as follows. We initially generate a random vector $y \in \mathbb{R}^n$ with entries from $\mathcal{N}(0,10)$ and $\delta_i \in \mathbb{R}^n$ with entries from $\mathcal{N}(0,1)$, and then set $y_i = y + \delta_i$, $i = 1,\ldots,M$. The covariance matrix of the samples $y_1,\ldots,y_M$ is $Y = \frac{1}{M}\sum_{i=1}^{M} y_i y_i^T$. The performance profiles on 50 datasets of 5 different sizes are presented in Figure 3. It can be seen that for the Max-likelyhood problem, both of NPG1 and NPG2 provide significantly faster computation and smaller objective value compared to the others in a considerable number of datasets.

6.4 Dual of the entropy maximization problems

We consider the entropy maximization problem subject to linear constraints [10, Section 5.1.6] which is

$$\min \sum_{i=1}^{n} x_i \log x_i \quad \text{s.t.} \quad Ax \le b, \ \sum_{i=1}^{n} x_i = 1, \ \text{ and } \ x_i > 0, i = 1,\ldots,n, \qquad (6.2)$$

[5] Max-likelyhood is a case of problem (P) with $f(X) = -\log\det(X) + \operatorname{tr}(XY)$ and $g(X) = \iota_C$ (the indicator function of C) with $C = \{X \in \mathbb{S}_n \mid lI \preceq X \preceq uI\}$.

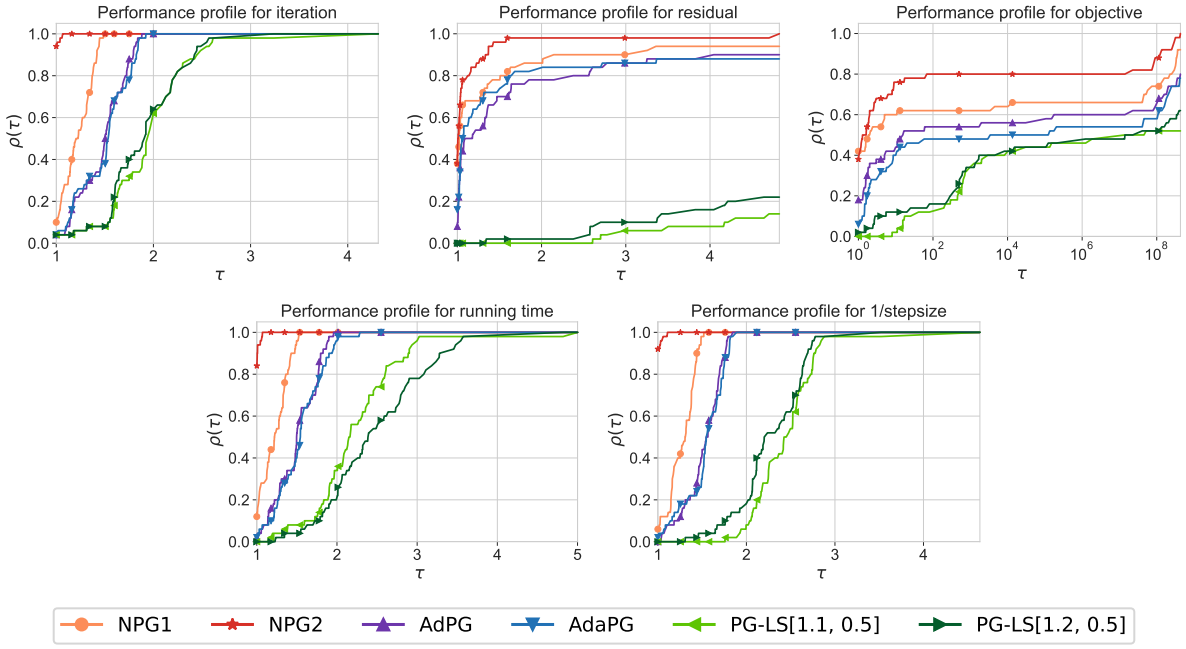

Fig. 3: Performance profiles for Max-likelyhood problem on 50 datasets.

where $A = [a^1, a^2, ..., a^n] \in \mathbb{R}^{m\times n}$, with $a^i \in \mathbb{R}^m$ is the $i-$th column of A and $b \in \mathbb{R}^m$. Its dual problem is

$$\min e^{-\mu-1}\sum_{i=1}^{n} e^{-(a^i)^T\lambda} + b^T\lambda + \mu, \text{ s.t. } \lambda \in \mathbb{R}^m_+, \mu \in \mathbb{R}. \qquad \text{(Dual-max-entropy)}$$

It is observed that Problem Dual-max-entropy[6] matches *Assumptions 1, 2* but *Assumption 3*. Therefore the use of NPG1 is straightforward for it. We still run NPG2 for Dual-max-entropy as a heuristic approach. We use the similar rule of generating data as [25]. Specifically, a $m \times n$ matrix A with entries are generated from $\mathcal{N}(0,1)$, $b = Ax^*$ with a ℓ_1-normalized x^* sampled from the uniform distribution $\mathcal{U}[0.1, 1)$. Results on 40 datasets of 4 different sizes are depicted in Figure 4. For this problem, NPG1 and NPG2 demonstrate descent capability with NPG2 archiving shortest running time in a majority of datasets and still being the algorithm with highest probability of archiving lowest objective value. At the same time, NPG1 shows comparable performance in comparison with other adaptive methods like AdPG and AdaPG^{q,r}.

[6] Dual-max-entropy is a case of problem (P) with $f(\lambda,\mu) = e^{-\mu-1}\sum\limits_{i=1}^{n} e^{-(a^i)^T\lambda} + b^T\lambda + \mu$ and $g(\lambda,\mu) = \iota_C$ (the indicator function of C) with $C = \mathbb{R}^m_+ \times \mathbb{R}$ and ∇f is not globally Lipschitz on C.

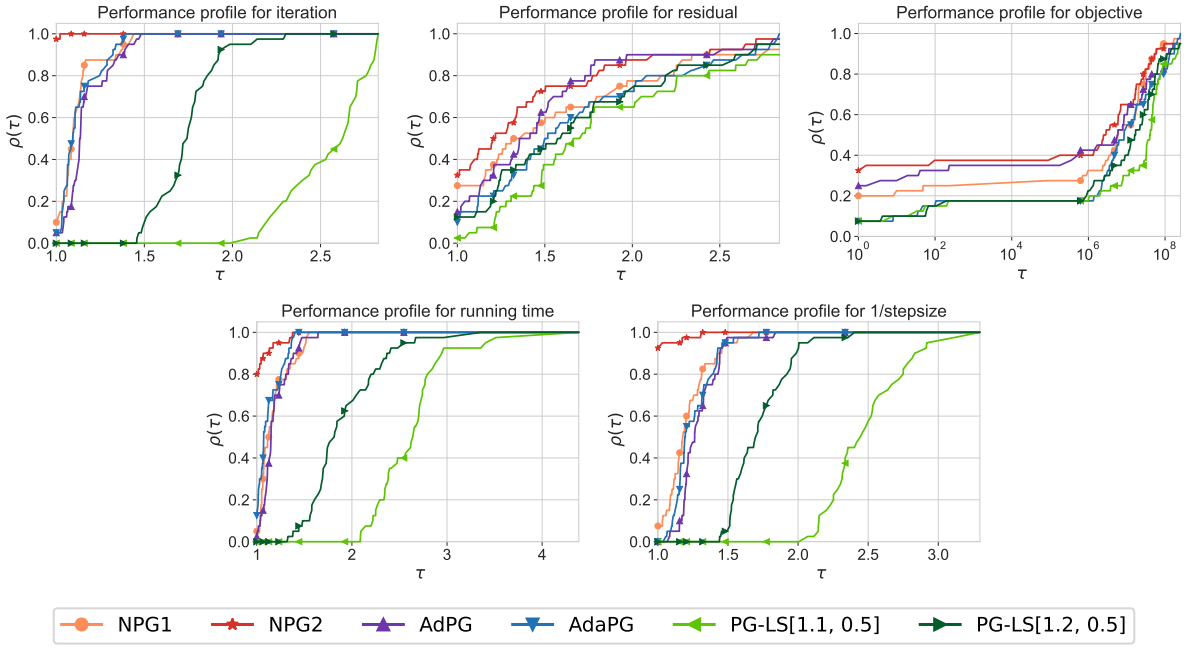

Fig. 4: Performance profiles for Dual-max-entropy problem on 40 datasets.

6.5 Indefinite quadratic programming problem with box constraints

We consider the indefinite quadratic programming problem with box constraints[7].

$$\min_{x\in[-1,1]^n} \quad \frac{1}{2}x^\top Qx + c^\top x. \tag{BCQP}$$

The indefinite quadratic programming problem is nonconvex and NP-hard even when Q has only one negative eigenvalue [27]. Since this is a quadratic programming problem which satisfies Assumptions 1 and 3, it is therefore theoretically guaranteed to use NPG2 and NPG-quad for this problem. We generate $c \in \mathbb{R}^n$ with entries drawn from $\mathcal{N}(0,1)$, and $Q = U^\top \operatorname{diag}(\lambda_1,\ldots,\lambda_n)U$ with $U \in \mathbb{R}^{n\times n}$ has i.i.d. standard normal entries, drawing $\lambda_i \sim \mathcal{U}(-1,r)$ for $i = 1,\ldots,n$. In order to ensure the stability of the experimental results, we control the percentage of negative eigenvalues around 10% to 15% by setting $r = 5, 10$. Results of this experiment on 60 datasets of 6 different sizes are presented in Figure 5. Notably, unlike the previous problems where all algorithms typically converge to solutions with nearly identical objective values, this nonconvex problem often leads different algorithms to solutions with significantly different objective values. In this setting, NPG2 and NPG-quad achieve the lowest objective values on a substantial number of datasets while remaining the two most time-efficient methods.

6.6 Fractional programming problems with box constraints

We consider the box constrained fractional programming problem[8] with a quadratic term on the numerator and positive affine term on the denominator

$$\min_{x\in[0,1]^n} \quad \frac{x^\top Ax + b^\top x + c}{p^\top x + q}. \tag{BCFP}$$

The constants c, q and the elements of the vectors b, p are drawn independently from $\mathcal{U}(1,10)$, and, similar to (BCQP), the matrix A is generated as $A = U^\top \operatorname{diag}(\lambda_1,\ldots,\lambda_n)U$ with $U \in \mathbb{R}^{n\times n}$ having

[7] BCQP is a case of problem (P) with $f(x) = \frac{1}{2}x^\top Qx + c^\top x$ and $g = \iota_C$ (the indicator function of C) with $C = [-1,1]^n$.

[8] BCFP is a case of problem (P) with $f(x) = \frac{x^\top Ax + b^\top x + c}{p^\top x + q}$ and $g = \iota_C$ (the indicator function of C) with $C = [0,1]^n$.

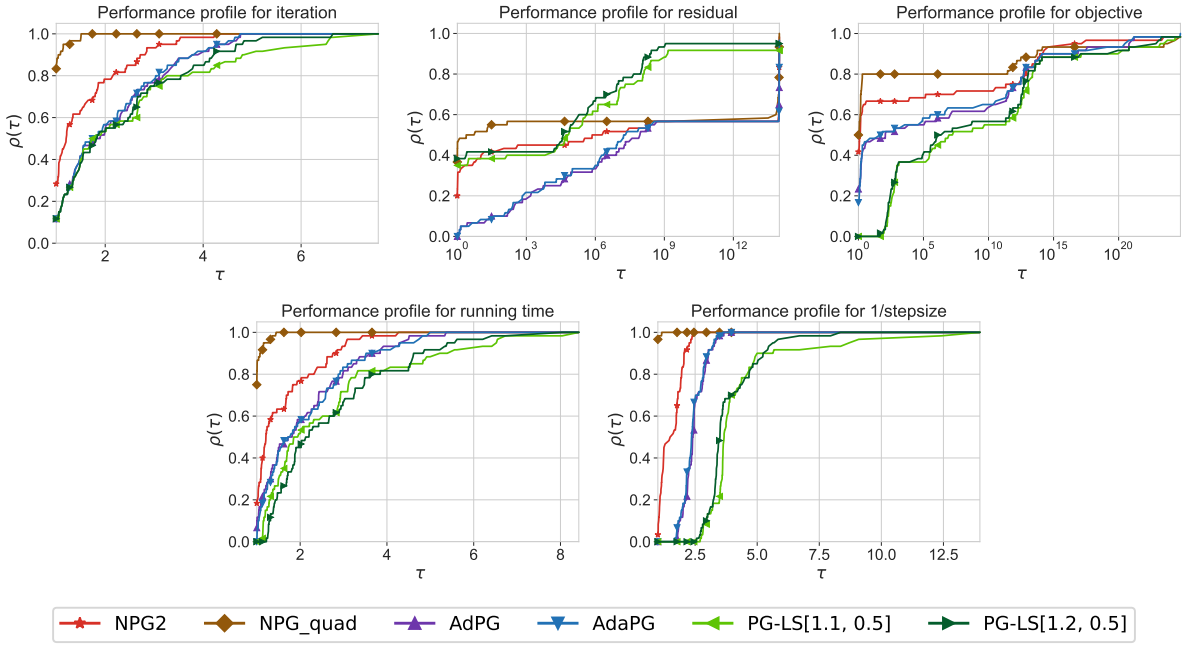

Fig. 5: Performance profiles for BCQP problem on 60 datasets.

i.i.d. standard normal entries and $\lambda_i \sim \mathcal{U}(-1, r)$ for $i = 1, \ldots, n$, where $r \in \{5, 10\}$. Given this setup, problem (BCFP) satisfies Assumptions 1 and 3, and thus NPG2 can be applied. Performance profiles over 60 datasets across six problem sizes are shown in Figure 6. As shown in the figure, NPG2 reaches the stopping criterion in the fewest iterations on almost all datasets. Moreover, it achieves the lowest objective value on about 70% of the datasets, demonstrating its strong ability to rapidly converge to high-quality local minima.

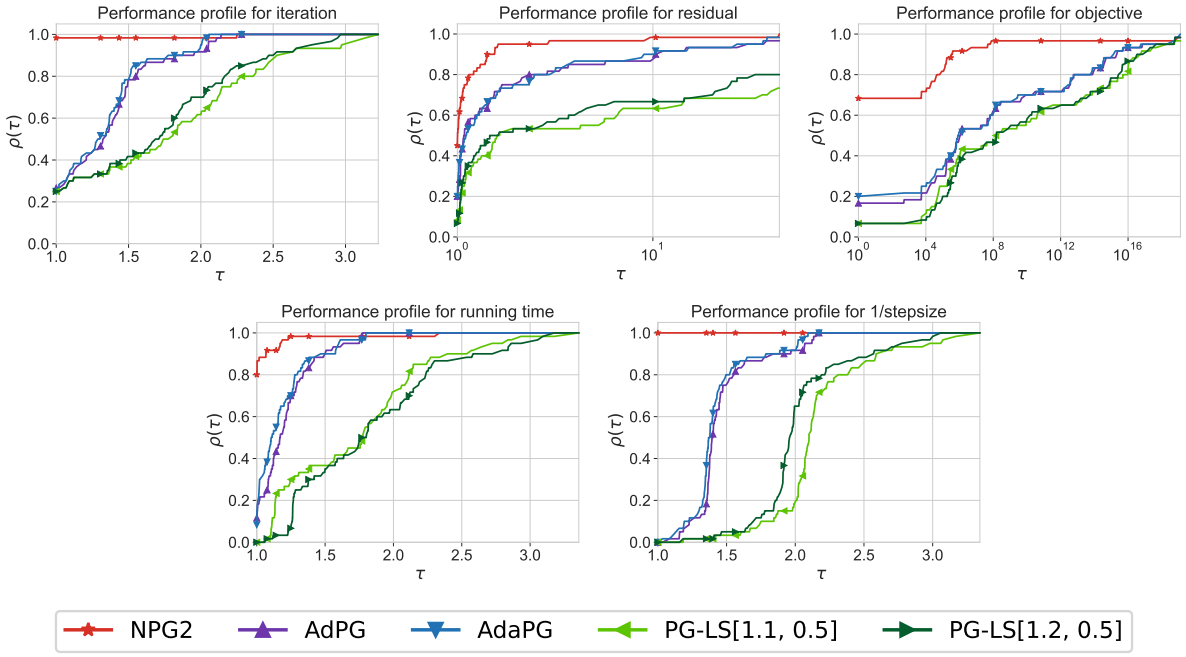

Fig. 6: Performance profiles for BCFP problem on 60 datasets.

6.7 Nonnegative matrix factorization

One of efficient approaches to solve recommendation system problems [29] is based on nonnegative matrix factorization[9]

$$\min f(U,V) = \frac{1}{2}\|UV^T - A\|_F^2, \ \text{ s.t. } U \in {\mathbb{R}_+}^{m\times r}, V \in {\mathbb{R}_+}^{n\times r}, \tag{NMF}$$

where $A \in \mathbb{R}^{m\times n}$ is a low-rank matrix, $\|\cdot\|_F$ stands for Frobenius norm. This problem does not satisfy *Assumption 2* and *Assumption 3*. Therefore our algorithms can be seen as heuristic methods for it. Akin to [25], we create A by multiplying matrices B and $C^\top$, where $B \in {\mathbb{R}_+}^{m\times r}$ and $C \in {\mathbb{R}_+}^{n\times r}$ have entries drawn from a normal distribution $\mathcal{N}(0,1)$. All negative entries of B and C are replaced with zero. The numerical results on 100 datasets of 10 different sizes are reported in Figure 7. It is evident from the figure that NPG1 and NPG2 require no more than 1.2 times the best iteration count and running time on nearly all datasets, whereas this proportion is nearly 0% for the other methods. In fact, each of the other methods achieves this within 1.5 times the best on only about 10% of the datasets. Moreover, the NPG methods attain the lowest objective values on approximately 40% and 60% of the datasets, respectively.

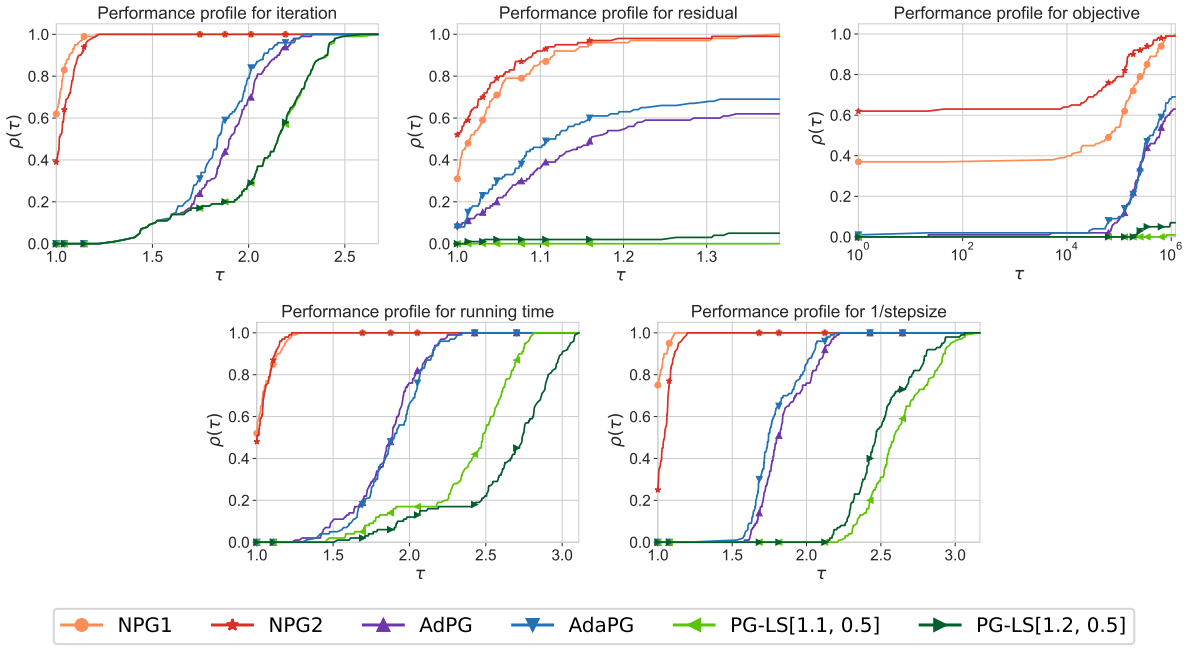

Fig. 7: Performance profiles for NMF problem on 100 datasets.

7 Conclusions

In this paper, we propose an efficient explicit stepsize applied for the proximal gradient (PG) scheme. In particular, Algorithm 3.1 (NPG1) solves the convex situation of the problem (P) under the locally Lipschitz gradient condition imposed on f. The iterates are proven to converge to an optimal solution of (P) with the computational complexity $O\left(\frac{1}{k}\right)$ of $F(x^k) - F_*$ and the Q-linear rate of $\{x^k\}_{k\geq k^*}$ if f has locally strong convexity property (k^* is a fixed number). These convergence results are based on the descent property of our proposed method from some fixed iteration. Moreover, our stepsize is also investigated with a class of nonconvex f satisfying global Lipschitz gradient condition with Algorithm 4.1 (NPG2), where the step length can be bigger. In

[9] NMF is a case of problem (P) with $f(U,V) = \frac{1}{2}\|UV^T - A\|_F^2$ and $g(U,V) = \iota_C$ (the indicator function of C) with $C = {\mathbb{R}_+}^{m\times r} \times {\mathbb{R}_+}^{n\times r}$.

quadratic case of f, Algorithm 5.1 (NPG-quad) is improved significantly in length of stepsizes for solving (P). Basically, our stepsize selection is computed quickly by a closed formulas without line search computation or estimating some constant (like Lipschitz constant of gradient) to ensure the convergence of the PG algorithms. The deep experiments on a variety of test instances with various sizes show the crucial efficiency of the proposed method compared to the recent ones.

Future research may explore several directions, including:

(i) investigating accelerated variants of NPG1 under weaker assumptions on ∇f in order to achieve improved convergence rates, such as $O(1/k^2)$;

(ii) extending NPG2 and NPG-quad to settings where neither convexity nor global Lipschitz continuity of ∇f is assumed.

References

1. M. Ahookhosh, A. Themelis, and P. Patrinos. A bregman forward-backward linesearch algorithm for nonconvex composite optimization: superlinear convergence to nonisolated local minima. *SIAM J. Optim.*, 31(1):653–685, 2021.
2. H.H. Bauschke, J. Bolte, and M. Teboulle. A descent lemma beyond lipschitz gradient continuity: first-order methods revisited and applications. *Math. Oper. Res.*, 42(2):330–348, 2017.
3. A. Beck. *Introduction to Nonlinear Optimization: Theory, Algorithms, and Applications with MATLAB*. Society for Industrial and Applied Mathematics, USA, 2014.
4. A. Beck. *First Order Methods in Optimization*. Society for Industrial and Applied Mathematics, USA, 2017.
5. A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2:183–202, 2009.
6. A. Beck and M. Teboulle. Gradient-based algorithms with applications to signal recovery problems. In D. Palomar and Y.C. Eldar, editors, *Convex Optimization in Signal Processing and Communications*, pages 139–162. Cambridge University Press, Cambridge, 2009.
7. D.P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 3rd edition, 2016.
8. J. Bolte, S. Sabach, M. Teboulle, and Y. Vaisbourd. First order methods beyond convexity and lipschitz gradient continuity with applications to quadratic inverse problems. *SIAM J. Optim.*, 28(3):2131–2151, 2018.
9. S. Bonettini, M. Prato, and S Rebegoldi. A new proximal heavy ball inexact line-search algorithm. *Comput. Optim. Appl.*, 88:1–41, 03 2024.
10. S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, 2004.
11. R.E. Bruck. On the weak convergence of an ergodic iteration for the solution of variational inequalities for monotone operators in hilbert space. *J. Math. Anal. Appl.*, 61:159–164, 1977.
12. E Dolan and J. Moré. Benchmarking optimization software with performance profiles. *Math. Program.*, 91(2):201–213, 2002.
13. R.A. Dragomir, A.B. Taylor, A. d'Aspremont, and J. Bolte. Optimal complexity and certification of bregman first-order methods. *Math. Program.*, 194:41–83, 2022.
14. M.Á.T. Figueiredo, R.D. Nowak, and S.J. Wright. Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems. *IEEE J. Sel. Top. Signal Process.*, 1(4):586–597, 2007.
15. M. Fukushima and H. Mine. A generalized proximal point algorithm for certain non-convex minimization problems. *Int. J. Syst. Sci.*, 12(8):989–1000, 1981.
16. P. T. Hoai. A new proximal gradient algorithm for solving mixed variational inequality problems with a novel explicit stepsize and applications. *Mathematics and Computers in Simulation*, 229:594–610, March 2025.

17. P.T. Hoai, N.T. Vinh, and N.P.H. Chung. A novel stepsize for gradient descent method. *Oper. Res. Lett.*, page 107072, 2024.

18. X. Jia, C. Kanzow, and P. Mehlitz. Convergence analysis of the proximal gradient method in the presence of the kurdyka–Łojasiewicz property without global lipschitz assumptions. *SIAM J. Optim.*, 33(4):3038–3056, 2023.

19. C. Kanzow and P. Mehlitz. Convergence properties of monotone and nonmonotone proximal gradient methods revisited. *J. Optim. Theory Appl.*, 195(2):624–646, 2022.

20. P. Latafat, A. Themelis, and P. Patrinos. On the convergence of adaptive first order methods: proximal gradient and alternating minimization algorithms. In *Proc. Mach. Learn. Res.*, volume 242, pages 197–208, 2024.

21. P. Latafat, A. Themelis, L. Stella, and P. Patrinos. Adaptive proximal algorithms for convex optimization under local lipschitz continuity of the gradient. *Math. Program.*, 2024.

22. Q. Lin, R. Ma, and Y. Xu. Complexity of an inexact proximal-point penalty method for constrained smooth non-convex optimization. *Comput. Optim. Appl.*, 82(1):175–224, 2022.

23. H. Liu, T. Wang, and Z. Liu. Some modified fast iterative shrinkage-thresholding algorithms with a new adaptive non-monotone step-size strategy for nonsmooth and convex minimization problems. *Comput. Optim. Appl.*, 83:651–691, 2022.

24. Y. Malitsky and K. Mishchenko. Adaptive gradient descent without descent. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proc. Mach. Learn. Res.*, pages 6702–6712, 2020.

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

26. A. De Marchi and A. Themelis. Proximal gradient algorithms under local lipschitz gradient continuity. *J. Optim. Theory Appl.*, 194:771–794, 2022.

27. P. M. Pardalos and S. A. Vavasis. Quadratic programming with one negative eigenvalue is NP-hard. *J. Glob. Optim.*, 1(1):15–22, 1991.

28. G.B. Passty. Ergodic convergence to a zero of the sum of monotone operators in hilbert space. *J. Math. Anal. Appl.*, 72:383–390, 1979.

29. P. Symeonidis and A. Zioupos. *Matrix and Tensor Factorization Techniques for Recommender Systems*. Springer Briefs in Computer Science. 2016.

30. M. Teboulle. A simplified view of first order methods for optimization. *Math. Program.*, 170(1):67–96, 2018.

31. A. Themelis, L. Stella, and P. Patrinos. Forward-backward envelope for the sum of two nonconvex functions: further properties and nonmonotone linesearch algorithms. *SIAM J. Optim.*, 28(3):2274–2303, 2018.

32. S.J. Wright, R.D. Nowak, and M.A.T. Figueiredo. Sparse reconstruction by separable approximation. *IEEE Trans. Signal Process.*, 57(7):2479–2493, 2009.

33. X. Zhao, R. Raushan, D. Ghosh, JC. Jao, and M. Qi. Proximal gradient method for convex multiobjective optimization problems without lipschitz continuous gradients. *Comput. Optim. Appl.*, 91:27–66, 2025.

Appendix

A Some missing proofs

A.1 The proof of Theorem 3.1

Proof (i) From Lemma 3.3, f is L_0-smooth on T then

$$f(x^*) - f(x^k) \geq \langle \nabla f(x^k), x^* - x^k \rangle + \frac{1}{2L_0}\|\nabla f(x^k) - \nabla f(x^*)\|^2.$$

Thus, we can update the term $\langle \nabla f(x^k), x^* - x^k \rangle$ in (3.6) with $x = x^*$ and then resulting it in (3.22) as

$$\begin{aligned} &\|x^{k+1} - x^*\|^2 + \|x^k - x^{k+1}\|^2 + \frac{2t_{k+1}^2}{t_k}\Big(F(x^k) - F(x^*)\Big) + \frac{t_k}{L_0}\|\nabla f(x^k) - \nabla f(x^*)\|^2 \\ \leq\ &\|x^k - x^*\|^2 + \|x^{k-1} - x^k\|^2 + \frac{2t_k^2}{t_{k-1}}\Big(F(x^{k-1}) - F(x^*)\Big) \quad \text{for all } k \geq k_1. \end{aligned} \tag{A.1}$$

Now summing up (A.1) from k_1 to k we have that

$$\begin{aligned} &\|x^{k+1} - x^*\|^2 + \|x^k - x^{k+1}\|^2 + \frac{2t_{k+1}^2}{t_k}\Big(F(x^k) - F(x^*)\Big) + \sum_{i=k_1}^{k}\frac{t_i}{L_0}\|\nabla f(x^i) - \nabla f(x^*)\|^2 \\ \leq\ &\|x^{k_1} - x^*\|^2 + \|x^{k_1-1} - x^{k_1}\|^2 + \frac{2t_{k_1}^2}{t_{k_1-1}}\Big(F(x^{k_1-1}) - F(x^*)\Big) \quad \text{for all } k \geq k_1 \end{aligned} \tag{A.2}$$

which follows the convergence of $\sum\limits_{i=k_1}^{+\infty} t_i\|\nabla f(x^i) - \nabla f(x^*)\|^2$ and therefore $\|\nabla f(x^i) - \nabla f(x^*)\| \to 0$.

Moreover, from Lemma 2.5 - inequality (2.1), substituting x by x^k we get that

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

which follows that

$$\frac{1}{2t_k}\|x^{k+1} - x^k\|^2 \leq t_k\|\nabla f(x^k) - \nabla f(x^*)\|^2 + \langle \nabla f(x^*),\, x^k - x^{k+1}\rangle - g(x^{k+1}) + g(x^k). \tag{A.3}$$

Summing up (A.3) from k_1 to $k \geq k_1$ to derive that

$$\begin{aligned} \sum_{i=k_1}^{k}\frac{1}{2t_i}\|x^{i+1} - x^i\|^2 &\leq \sum_{i=k_1}^{k} t_i\|\nabla f(x^i) - \nabla f(x^*)\|^2 + \langle \nabla f(x^*),\, x^{k_1} - x^{k+1}\rangle - g(x^{k+1}) + g(x^{k_1}) \\ &\leq \sum_{i=k_1}^{k} t_i\|\nabla f(x^i) - \nabla f(x^*)\|^2 + \|\nabla f(x^*)\|\|x^{k_1} - x^{k+1}\| + g(x^{k_1}) - F(x^{k+1}) + f(x^{k+1}) \\ &\leq \sum_{i=k_1}^{k} t_i\|\nabla f(x^i) - \nabla f(x^*)\|^2 - F_* + g(x^{k_1}) + \sup_{x\in T}\Big(\|\nabla f(x^*)\|\|x^{k_1} - x\| + f(x)\Big). \end{aligned} \tag{A.4}$$

Remember that T is compact and f is differentiable on T, moreover $\sum\limits_{i=1}^{+\infty} t_i\|\nabla f(x^i) - \nabla f(x^*)\|^2$ is convergent then the right hand side of (A.4) is upper bounded for all $k \geq k_1$. This derives the convergence of $\sum\limits_{i=k_1}^{+\infty}\frac{1}{2t_i}\|x^{i+1} - x^i\|^2$.

Now let $\bar{x}$ is a cluster point of $\{x^k\}$ then there exists a subsequence $\{x^{k_j}\}$ such that $x^{k_j} \to \bar{x}$.

From Lemma 2.5 (ii) - inequality (2.2),

$$F(x^{k_j+1})-F(x^*)\leq\langle\nabla f(x^{k_j+1})-\nabla f(x^{k_j}),x^{k_j+1}-x^{k_j}\rangle-\frac{1}{t_{k_j}}\left\|x^{k_j+1}-x^{k_j}\right\|^2-\frac{1}{t_{k_j}}\left\langle x^{k_j+1}-x^{k_j},x^{k_j}-x^*\right\rangle. \quad (A.5)$$

Tending k_j in (A.5) to infinity we obtain that: (i) the first and the third terms on the right hand side of (A.5) come to zeros because of the convergence of $\{x^{k_j}\}$ and the compactness of T; (ii) the second term on the right hand side of (A.5) comes to zero since $\sum\limits_{i=k_1}^{+\infty}\frac{1}{2t_i}\|x^{i+1}-x^i\|^2$ converges. Therefore, $\lim\limits_{k_j\to\infty}F(x^{k_j})=F_*=F(x^*)$. On the other hand, the closedness of F (from *Assumption 1*) follows its lower semi-continuity and therefore $F(\bar{x})\leq\lim\limits_{k_j\to\infty}F(x^{k_j})=F_*$, which implies $\bar{x}\in X^*$. Next, repeating the arguments in the proof of Theorem 3.2 (ii), we obtain that $\{x^k\}$ converges to some element of X^*.

(ii) Now summing up (3.21) from $k=k_1$ to K we get that,

$$\|x^{K+1}-x^*\|^2+2t_K\left(1+\frac{t_K}{t_{K-1}}\right)\left(F(x^K)-F(x^*)\right)+\|x^K-x^{K+1}\|^2+2\sum_{k=k_1+1}^{K}\left(t_{k-1}+\frac{t_{k-1}^2}{t_{k-2}}-\frac{t_k^2}{t_{k-1}}\right)\left(F(x^{k-1}-F(x^*)\right)$$

$$\leq\|x^{k_1}-x^*\|^2+\|x^{k_1-1}-x^{k_1}\|^2+\frac{2t_{k_1}^2}{t_{k_1-1}}\left(F(x^{k_1-1})-F(x^*)\right)\underset{\text{by(3.23)}}{=}\mathscr{R}^2, \quad (A.6)$$

which follows

$$\left(\frac{t_K^2}{t_{K-1}}+\frac{t_{k_1}^2}{t_{k_1-1}}+2\sum_{k=k_1}^{K}t_k\right)\min_{k_1\leq k\leq K}\left(F(x^k)-F(x^*)\right)\leq\mathscr{R}^2,\ \ \forall K\geq k_1. \quad (A.7)$$

Now, the desired conclusion is obtained.

A.2 The proof of Example 4.3

Proof Indeed, for any $u,v\in C$, $f(u+t(v-u))$ is a fractional function with a quadratic numerator and an affine denominator in t. Thus,

$$h_{uv}(t)=f_t'(u+t(v-u))=\alpha+\frac{\beta}{(p^T(u+t(v-u))+q)^2},$$

where α and β are real constants derived from f and u,v. Hence

$$h_{uv}'(t)=\frac{-2\beta p^T(u-v)}{(p^T(u+t(v-u))+q)^3}$$

which does not change sign for $t\in[0,1]$ since $\left(p^T(u+t(v-u))+q\right)^3>0$ for all $t\in[0,1]$. Thus $h_{uv}(t)$ is monotone over $[0,1]$ and hence quasiconvex in $[0,1]$.

B Details on performance profiles

The numerical performance of the selected algorithms is summarized using performance profiles [12]. Let $\mathscr{P}$ denote the set of problems and $\mathscr{A}$ the set of algorithms. For each problem $p\in\mathscr{P}$ and algorithm $a\in\mathscr{A}$, let $t_{p,a}$ denote the performance measure (e.g., running time or iteration count), where smaller values indicate better performance. For each problem p, the performance ratio of algorithm a is defined as

$$r_{p,a}=\frac{t_{p,a}}{\min\{t_{p,\hat{a}}\mid\hat{a}\in\mathscr{A}\}}.$$

The performance profile of algorithm a is then computed by

$$\rho_a(\tau)=\frac{|\{p\in\mathscr{P}:r_{p,a}\leq\tau\}|}{|\mathscr{P}|}.$$

This number represents the fraction of problems for which algorithm a performs within a factor τ of the best algorithm.

C Additional experimental details and results

The datasets used in the experiments are randomly generated with random seeds ranging from 1 to 10. The parameters describing the size and properties of the datasets are summarized in Table 1. The superscript * denotes the parameter setting used for Figures 8, 9.

Table 1: Data-generation configurations for all problem instances.

Problem	Parameters	Configurations
Lasso	(m,n)	$(512,1024)$, $(512,2048)$, $(512,4096)$, $(1024,2048)$, $(1024,4096)$, $(1024,8192)^*$, $(2048,4096)$, $(2048,8192)$
Max-likelihood	(n,l,u,M)	$(100,0.1,10,50)$, $(100,0.1,10,500)$, $(100,0.1,10,1000)$, $(30,0.1,1000,50)^*$, $(50,0.1,1000,100)$
NMF	(m,n,r)	$(m,n) \in \{(500,1000),(1000,500),(2000,3000),(3000,2000),(3000,3000)^*\}$, $r \in \{20,30^*\}$
Min-length	(m,n)	$(50,5000)$, $(500,5000)$, $(2000,5000)$, $(100,10000)^*$, $(1000,10000)$, $(2000,10000)$
Dual-max-entropy	(m,n)	$(100,500)$, $(500,2000)$, $(2000,4000)$, $(4000,5000)^*$
BCQP	(n,r)	$(1000,5)$, $(2000,5)$, $(5000,5)$, $(1000,10)^*$, $(2000,10)$, $(5000,10)$
BCFP	(n,r)	$(1000,5)$, $(2000,5)$, $(5000,5)$, $(1000,10)$, $(2000,10)$, $(5000,10)^*$

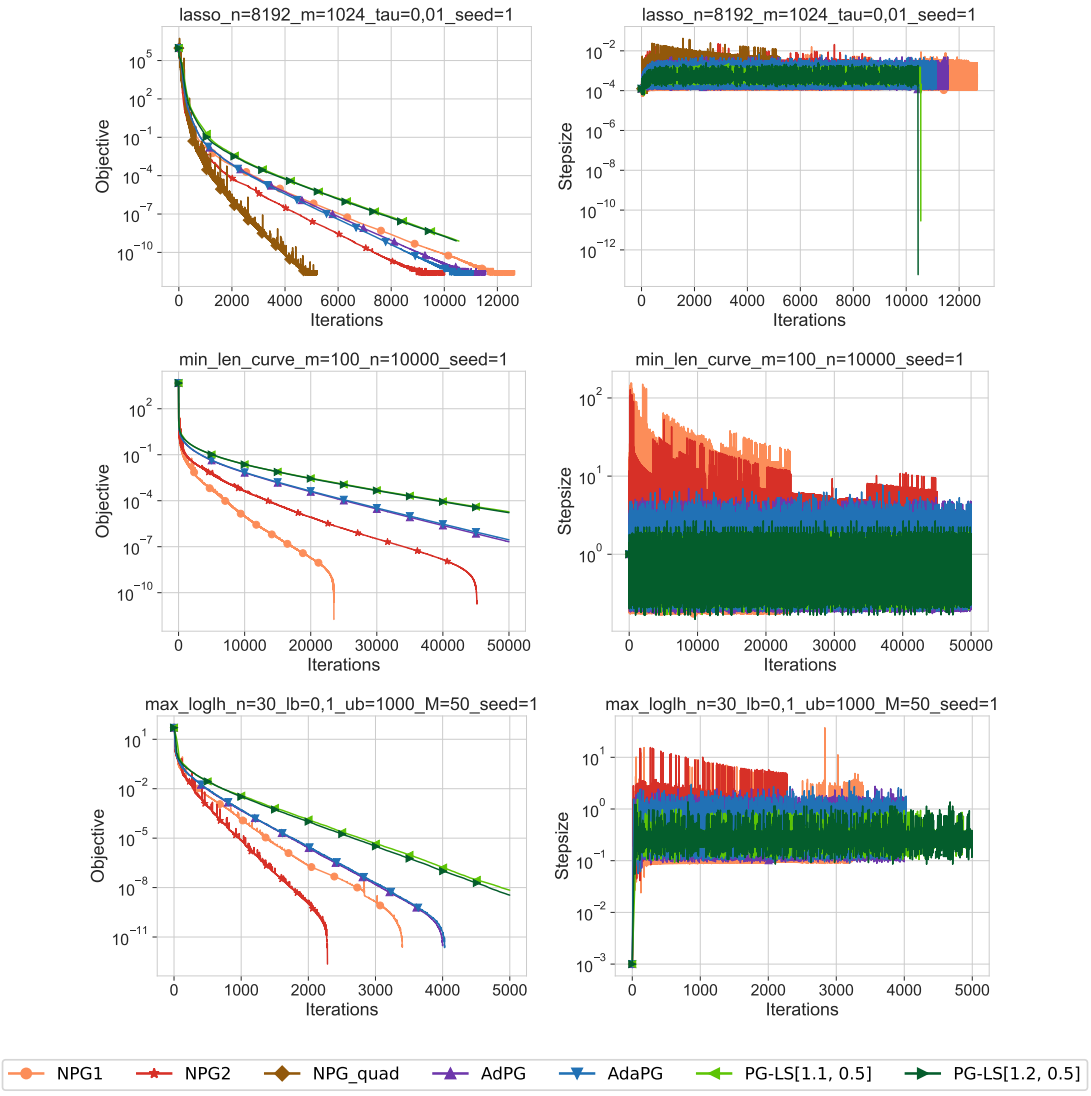

Fig. 8: Objective value and stepsize plots for Experiments Lasso (top), Min-length (middle), and (Max-likelyhood) (bottom).

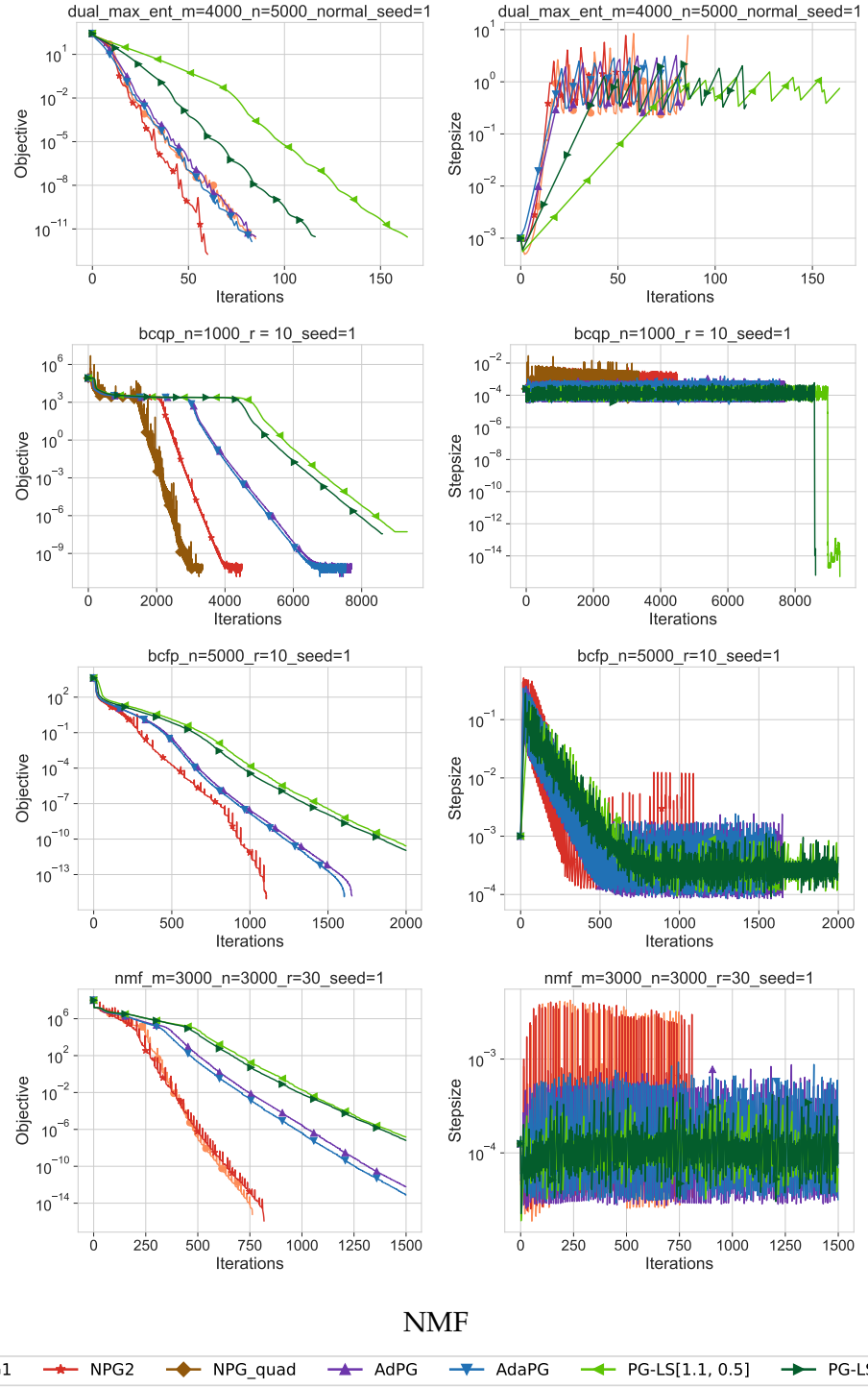

Fig. 9: Objective and stepsize plots for Experiments Dual-max-entropy(first row), BCQP(second row), BCFP(third row), and NMF(fourth row).