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

Composite optimization models via proximal gradient method with increasing adaptive stepsizes

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

Received: date / Accepted: date

Abstract We first consider the convex composite optimization models with locally Lipschitz condition imposed on the gradient of the differentiable term. The classical method which is proximal gradient will be studied with our new strategy of stepsize selection. Our proposed stepsize can be computed conveniently by explicit forms. The sequence of our stepsizes is proved to be increasing to a finite positive limitation. The PG method with our stepsize selection is shown to be decreasing and convergent with the complexity computation $O\left(\frac{1}{k}\right)$ for $F(x^k) - F_*$. This rate is strengthened to be *Q-linear* if f is added the locally strong convexity property. To the best of our knowledge, for proximal algorithm using an *adaptive* stepsize selection solving convex composite optimization models without globally Lipschitz gradient condition of the smooth term, there has been no method with such convergent properties so far. In addition, we show that our algorithm can be extended for solving a class of nonconvex composite model as complementing the global Lipschitz condition on ∇f. The significant efficiency of our proposed algorithms is expressed by numerical results for a numerous of applicable test problems.

Keywords proximal gradient method · nonlinear programming · composite optimization model · locally Lipschitz gradient · Lasso problem

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

1 Introduction

Composite optimization models (COM) are 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., [1,3,4,5,6,7,11,25,21,29,32,12,9,24,20,26]. The formulation of (COM) considered in this paper can be 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.
Assumption 1:

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

✉ Pham Thi Hoai
hoai.phamthi@hust.edu.vn
Nguyen Pham Duy Thai
duythai09092002@gmail.com

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

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

(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 problem (P) is *proximal gradient* (PG) introduced by Fukushima and Mine [18] in 1981 and has become now classical. The detail methodology of the PG method can be found in Beck [6,7]. It is observed that the optimal conditions for problem (P) relates to the concept of its stationary points. Specifically, if $x^*\in\mathrm{int}(\mathrm{dom}(f))$ is a local optimal solution of (P) then it should be a *stationary point* of (P), i.e., for some $t>0$

$$x^*=\mathrm{Prox}_{tg}(x^*), \tag{1.1}$$

where $\mathrm{Prox}_{tg}(x^*)$ is defined as the unique optimal solution of the minimization problem

$$\min_{x\in\mathbb{R}^n}\left\{g(x)+\frac{1}{2t}\|x-(x^*-t\nabla f(x^*))\|^2\right\}. \tag{1.2}$$

In the convex situation of (P), i.e., f is convex, the set of stationary points of (P) are coincident with X^*. One can see [6] (Theorem 3.72, 10.7) for more details. Based on the mentioned stationary condition, starting from some $x^0\in\mathrm{int}(\mathrm{dom}(f))$, the well-known PG method to solve problem (P) is designed by generating the sequence $\{x^k\}$ according to the rule

$$x^{k+1}=\mathrm{Prox}_{t_kg}(x^k),\ \ k=0,1,2,..., \tag{1.3}$$

where

$$\mathrm{Prox}_{t_kg}(x^k):=\operatorname*{argmin}_{x\in\mathbb{R}^n}\left\{g(x)+\frac{1}{2t_k}\left\|x-(x^k-t_k\nabla f(x^k))\right\|^2\right\}. \tag{1.4}$$

As a matter of fact, the PG scheme (1.3) is very useful if we can compute $\mathrm{Prox}_{t_kg}(x^k)$ easily by some explicit formulas. There is a list of such functions that can be found in [6]; for instances, g is ℓ_1 norm or the indicator function of a closed convex set $C\subset\mathbb{R}^n$. In (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 also navigate x^k to a good stationary point (that provides, for example, the objective value as low as possible) with a cheap cost. For the class of L_f- smooth function f, i.e.,

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

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 [6] (Theorem 10.21), one get the complexity computation $O(\frac{1}{k})$ of $F(x^k)-F_*$ if f is assumed to be convex and for the strongly convex case of f, the convergence rate of $\{x^k\}$ to some $x^*\in X^*$ is proved to be Q-linear. These important properties can be seen as the generalization of the results for the gradient descent method solving unconstrained nonlinear optimization problems, i.e., problem (P) with $g=0$.

Recently, researchers have concerned *problem (P) without the global Lipschitzness assumption on* ∇f, see, e.g., [2,8,21,23,13,14,15] since the class of such functions occurs in many applied problems, see e.g., [23,22,33] and the references therein. In 2017, Bauske 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 [31] to find the role of non-Euclidean proximal distances of Bregman type in the development and analysis of some typical first order optimization algorithms. The stepsize selection of NoLips is then chosen in $(0,\frac{2-\delta}{L})$. This algorithm is shown in [2] to have the convergent results similar to the ones of the normal PG scheme. Following that, Dragomir et al. [16] give a lower bound to prove that the $O(\frac{1}{k})$ convergence rate of the NoLips

method is optimal for the class of problems satisfying the relative smoothness assumption. The other recent results on the convergence of PG method without globally Lipschitz assumption have been studied in Kanzow and Mehlitz [23] and then Jia et al. [21]. Their proposed method can be applied for the nonconvex setting of (P) with the presence of Kurdyka–Łojasiewicz condition. The stepsize choice is based on backtracking line-search procedure. Nevertheless, one know that there are some restrictions of taking stepsize within $(0, \frac{2}{L_f})$ or $(0, \frac{2-\delta}{L})$ like: firstly, the process of finding these constants are not easy in general and secondly, if they are large then such stepsizes will be very small that may take long running time for executing algorithms. Analogously, the backtracking computation for stepsize selection probably consumes expensive cost and also may cause the stepsize to gradually decrease to a tiny number.

To overcome the mentioned drawbacks above, an interesting question should be considered is: *"Under Assumption 1 and f satisfying convex and locally Lipschitz gradient, is there an efficient way to find stepsizes explicitly for PG scheme solving problem (P) such that we do not need neither estimating constants like $L_f, L, ...$ nor backtracking line-search procedures?"* In the specific context of problem (P) with $g = 0$, such an algorithm named AdGD (Adaptive Gradient Descent) was proposed by Malitsky and Mishchenko [28] in 2019 for solving unconstrained convex optimization problems satisfying locally Lipschitz gradient. Continuing this research direction, Hoai et al. [19] proposed NGD algorithm that uses an explicit stepsize strategy based on the local curvature of f.

Contributions: In this paper, we give a positive answer for the question presented above. In particular, by utilizing the idea of adaptive stepsize in NGD [19] with PG scheme (1.3) we propose new proximal gradient algorithms for solving problem (P) with locally Lipschitz gradient condition imposed on the smooth term. More precisely, under *Assumption 1* and f is a convex function satisfying local Lipschitz gradient, we address the following properties for PG algorithm with our new stepsize selection:

- our proposed stepsize is quickly computed by explicit forms without the requirement of estimating any constant (for guaranteeing the convergence) as well as backtracking calculation;
- our proposed method is proved to be decreasing from some fixed iteration;
- the complexity computation of $F(x^k) - F_*$ is $O(\frac{1}{k})$;
- in the case of locally strongly convexity of f, we get the Q-linear rate of the iterates;
- the sequence of our proposed stepsize is increasing to a positive number;
- the range of step length of our proposed stepsize is proved to be bigger than NGD if $g = 0$.

It is worth noting that without global Lipschitz gradient continuity of f, these above convergent results are often obtained with standard strategies of choosing stepsize like fixed stepsize within a given interval (e.g., $(0, \frac{2-\delta}{L})$ with constant L satisfies Lipschitz-like/convexity condition (LC) for NoLips algorithm) or line-search procedures. However, for *PG scheme using an adaptive stepsize* selection, to the best of our knowledge, there has been no method with such convergent properties so far. Moreover, we show that our method can be extended to apply for a class of nonconvex of (P) if ∇f satisfies global Lipschitz continuity. As a byproduct, one special version solving problem (P) is designed in the case f is an indefinite quadratic form with the capability of enlarging stepsize. We also implement our new algorithms in comparison with the recent ones for a numerous of test instances to figure out the crucial efficiency of the new method.

The rest of the paper is structured 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 locally Lipschitz condition of ∇f. In the sequel, we consider a nonconvex case of (P) with an other new algorithm. Section 5 presents a particular 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.

Through out this paper, for any $z \in\in \text{int}(\text{dom}(f))$, and $t > 0$, we use the definition of the proximity as follows

$$\text{Prox}_{tg}(z) := \underset{x\in\mathbb{R}^n}{\text{argmin}} \left\{ g(x) + \frac{1}{2t_k} \|x - (z - t\nabla f(z))\|^2 \right\}.$$

Lemma 2.1 *Under Assumption 1, the sequence $\{x^k\}$ generated by proximal gradient scheme (1.3) for solving problem (P) has the following properties:*

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

(ii) for all $x \in \text{int}(\text{dom}(f))$, we have

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

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

$$x^{k+1} = x^k - t_k (\nabla f(x^k) + \overline{\partial} g(x^{k+1})). \tag{2.2}$$

(ii) From (i) and the convexity of g we are easy to get that

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

Lemma 2.2 (Lemma 2 in [28]) *Let $\{x^k\} \subset \mathbb{R}^n$ be a bounded sequence where its cluster points in $X \subset \mathbb{R}^n$ and the real sequence $\{a_k\} \subset \mathbb{R}_+$. If*

$$\|x^{k+1} - x\|^2 + a_{k+1} \leq \|x^k - x\|^2 + a_k, \ \ \forall x \in X, \tag{2.3}$$

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

3 A new proximal gradient algorithm for the problem (P) with f being convex and locally Lipschitz gradient

In this section, we propose a new proximal gradient algorithm for solving problem (P) satisfying *Assumption 1* and *Assumption 2* below.

Assumption 2: f is convex and locally Lipschitz gradient.

Algorithm 3.1 (NPG1)

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

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

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

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

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

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

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

Lemma 3.1 *For all $x \in \text{int}(\text{dom}(f))$ we have*

$$\|x^{k+1} - x\|^2 + 2t_k \left(F(x^k) - F(x)\right) \leq \|x^k - x\|^2 + t_k^2 \left\|\nabla f(x^k) + \overline{\partial} g(x^k)\right\|^2.$$

Proof From Lemma 2.1 (ii), for all $x \in \text{int}(\text{dom}(f))$

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

Using the convexity of f and g, we continue evaluating

$$\begin{aligned} \langle \nabla f(x^k), x - x^{k+1}\rangle &= \langle \nabla f(x^k), x - x^k\rangle + \langle \nabla f(x^k) + \overline{\partial} g(x^k), x^k - x^{k+1}\rangle + \langle \overline{\partial} g(x^k), x^{k+1} - x^k\rangle \\ &\leq f(x) - f(x^k) + \left\langle \nabla f(x^k) + \overline{\partial} g(x^k), x^k - x^{k+1}\right\rangle + g(x^{k+1}) - g(x^k). \end{aligned} \tag{3.6}$$

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

$$\|x^{k+1} - x\|^2 + 2t_k \left(F(x^k) - F(x)\right) \leq \|x^k - x\|^2 + R, \tag{3.7}$$

where

$$\begin{aligned} R &= 2t_k \left\langle \nabla f(x^k) + \overline{\partial} g(x^k), x^k - x^{k+1}\right\rangle - \|x^{k+1} - x^k\|^2 \\ &= t_k \left\langle 2\nabla f(x^k) + 2\overline{\partial} g(x^k) - \nabla f(x^k) - \overline{\partial} g(x^{k+1}), x^k - x^{k+1}\right\rangle \\ &= t_k^2 \left\langle \nabla f(x^k) + 2\overline{\partial} g(x^k) - \overline{\partial} g(x^{k+1}), \nabla f(x^k) + \overline{\partial} g(x^{k+1})\right\rangle \\ &= t_k^2 \left(\left\|\nabla f(x^k) + \overline{\partial} g(x^k)\right\|^2 - \left\|\overline{\partial} g(x^{k+1}) - \overline{\partial} g(x^k)\right\|^2\right) \\ &\leq t_k^2 \left\|\nabla f(x^k) + \overline{\partial} g(x^k)\right\|^2. \end{aligned} \tag{3.8}$$

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*

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

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 $\|\nabla f(x^{k+1}) - \nabla f(x^k)\| \leq \frac{c_0}{t_k}\|x^{k+1} - x^k\|$ then by (3.4), $t_{k+1} = (1+\gamma'_k)t_k$ and (3.9) is equivalent to

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

Moreover, from (3.3), if $\frac{t_k}{t_{k-1}} \geq 1$ then $\gamma'_k = \gamma_k$ and because $\sum_{k=0}^{+\infty} \gamma_k < +\infty$, there is k_0 such that

$$\gamma'_k = \gamma_k \leq \sqrt{2} - 1 \leq \sqrt{1 + \frac{t_k}{t_{k-1}}} - 1 \quad \forall k \geq k_0. \tag{3.11}$$

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

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

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

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 $k \geq k_1$,

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

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

Proof (i) We have the relation

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

where

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

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 $\|\nabla f(x^k)-\nabla f(x^{k-1})\| \le \frac{c_0}{t_{k-1}}\|x^k-x^{k-1}\|$ then

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

which follows

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

The convergence of $\sum\limits_{k=0}^{+\infty}\gamma_k$ indicates that there exists $k_1 \ge k_0$ satisfying

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

which is equivalent to $(1+\gamma_{k-1})^2c_0^2 \le \frac{1}{2}$ for all $k \ge k_1$. 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 $k \ge k_1$ we have

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

Nevertheless,

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

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

$$\begin{aligned}
&\|x^{k+1}-x^*\|^2 + 2t_k\left(1+\frac{t_k}{t_{k-1}}\right)\left(F(x^k)-F(x^*)\right) + \|x^k-x^{k+1}\|^2 \\
&\le \|x^k-x^*\|^2 + \|x^{k-1}-x^k\|^2 + 2\frac{t_k^2}{t_{k-1}}\left(F(x^{k-1})-F(x^*)\right).
\end{aligned} \tag{3.21}$$

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

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

This inequality follows that

$$\|x^{k+1}-x^*\|^2 + \|x^k-x^{k+1}\|^2 + \frac{2t_{k+1}^2}{t_k}\left(F(x^k)-F(x^*)\right) \le K, \tag{3.23}$$

where

$$K = \|x^{k_1} - x^*\|^2 + \|x^{k_1-1} - x^{k_1}\|^2 + \frac{2t_{k_1}^2}{t_{k_1-1}}\Big(F(x^{k_1-1}) - F(x^*)\Big).$$

The relation (3.23) implies the boundedness of $\{x^k\}$.

Remark 3.1 From the proof of Lemma 3.3 (eq. (3.11) and (3.18)), we see that if the convergent positive series $\sum\limits_{k=0}^{+\infty}\gamma_k$ is created such that $\gamma_k \le \min\Big\{\frac{1}{\sqrt{2}c_0} - 1, \sqrt{2} - 1\Big\}$ for all $k \ge 1$ then $k_1 = 1$ and therefore we obtain (3.22) for any $k \ge 1$.

The bounded property of the sequence $\{x^k\}$ in Lemma 3.3 provides us an important key to beyond the challenge of the usual condition imposed on the gradient of f that the globally Lipschitz continuity of ∇f. In the upcoming lemma, we start deploying the locally Lipschitz of ∇f to obtain several typical characteristics of the sequence of our new stepsize.

Lemma 3.4 *Let $\{t_k\}$ be a sequence of stepsizes generated by Algorithm 3.1. Then*

(i) $\{t_k\}$ is lower bounded by a positive number;

(ii) $\{t_k\}$ is convergent and has a positive limitation.

Proof (i) By Lemma 3.3 the set $T = \overline{conv}\{x^*, x^0, x^1, ...\}$ is closed and compact. From the local Lipschitz continuity of ∇f, it is easy to see that there exists $L_0 > 0$ satisfying $\|\nabla f(x) - \nabla f(y)\| \le L_0\|x - y\| \quad \forall x, y \in T$. Thereafter, either $t_1 \ge \frac{c_1}{L_0}$ or $t_1 = (1 + \gamma_0')t_0 \ge t_0$. The induction process derives that

$$t_k \ge \min\{\frac{c_1}{L_0}, t_0\} = \eta > 0 \quad \forall k \ge 0. \tag{3.24}$$

(ii) If we set $r_k = \ln t_{k+1} - \ln t_k$ and $r_k^+ = \max\{0, r_k\} \ge 0, r_k^- = -\min\{0, r_k\} \ge 0, \forall k \ge 0$ 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

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

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

$$\ln t_{k+1} - \ln t_0 = \sum_{i=0}^{k} r_i = \sum_{i=0}^{k}(r_i^+ - r_i^-) = \sum_{i=0}^{k} r_i^+ - \sum_{i=0}^{k} r_i^-. \tag{3.25}$$

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 is contradict with 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 (3.25)).

Lemma 3.5 *There exists k^* such that*

$$\|\nabla f(x^k) - \nabla f(x^{k-1})\| \le \frac{c_0}{t_{k-1}}\|x^k - x^{k-1}\|, \ \ \forall k \ge k^*. \tag{3.26}$$

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

Remark 3.2 From Lemma 3.5, we immediately obtain the increasing of the sequence $\{t_k\}_{k\ge k^*}$ and $0 < \eta < t_k \le \max\{t_0, ..., t_{k^*-1}, t^*\} = t_{max},\ k \ge 0$.

Lemma 3.6 *For any $x \in \text{int}(\text{dom}(f))$, we have*

$$F(x) - F(x^{k+1}) \ge \frac{1-c_0}{t_k}\|x^{k+1} - x^k\|^2 + \frac{1}{t_k}\langle x^k - x^{k+1}, x - x^k\rangle, \quad \textit{for all } k \ge k^*. \tag{3.27}$$

Proof Because of the convexity of f and Lemma 2.1 (ii) we have

$$\begin{aligned}
&F(x) - F(x^{k+1}) = f(x) + g(x) - f(x^{k+1}) - g(x^{k+1}) \\
&\ge f(x^k) + \left\langle x - x^k, \nabla f(x^k)\right\rangle + \left\langle x^{k+1} - x, \nabla f(x^k) + \frac{x^{k+1} - x^k}{t_k}\right\rangle - f(x^{k+1}) \\
&= f(x^k) - f(x^{k+1}) + \left\langle x^{k+1} - x^k, \nabla f(x^k)\right\rangle + \frac{1}{t_k}\left\langle x^{k+1} - x^k, x^{k+1} - x\right\rangle \\
&\ge \langle \nabla f(x^{k+1}) - \nabla f(x^k), x^k - x^{k+1}\rangle + \frac{1}{t_k}\left\|x^{k+1} - x^k\right\|^2 + \frac{1}{t_k}\left\langle x^{k+1} - x^k, x^k - x\right\rangle
\end{aligned} \tag{3.28}$$

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

$$\begin{aligned}
\left\langle \nabla f(x^{k+1}) - \nabla f(x^k), x^k - x^{k+1}\right\rangle &\ge -\left\|\nabla f(x^k) - \nabla f(x^{k+1})\right\|\|x^k - x^{k+1}\| \\
&\ge -\frac{c_0}{t_k}\|x^{k+1} - x^k\|^2 \quad \forall k \ge k^*.
\end{aligned} \tag{3.29}$$

The proof is completed by utilizing (3.28) and (3.29).

The convergent properties of Algorithm 3.1 are given in the following theorem.

Theorem 3.1 *Suppose that problem (P) satisfies Assumptions 1 and 2. Then the following assertions hold for Algorithm 3.1.*

(i) The sequence $\{F(x^k)\}_{k\ge k^*}$ 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 $k \ge k^* + 1$ we have

$$F(x^k) - F_* = F(x^k) - F(x^*) \le \frac{D}{2t_{k^*}(k - k^*)} = O\left(\frac{1}{k}\right), \tag{3.30}$$

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) Substituting x by x^k in (3.27) of Lemma 3.6 we get that

$$F(x^k) - F(x^{k+1}) \geq \frac{1-c_0}{t_k}\|x^{k+1}-x^k\|^2 \geq \frac{1-c_0}{t^*}\|x^{k+1}-x^k\|^2 \geq 0, \quad \text{for all } k \geq k^*. \tag{3.31}$$

By (3.31), the sequence $\{F(x^k)\}_{k\geq k^*}$ is decreasing. On the other hand, it is lower bounded by F_* hence converges to $\bar{F} \geq F_*$. Thus, $F(x^k) - F(x^{k+1}) \to 0$. And consequently, the inequality (3.31) follows

$$\lim_{k\to+\infty} \|x^{k+1}-x^k\| = 0. \tag{3.32}$$

Now, replacing x with x^* in (3.27) of Lemma 3.6 to obtain

$$\begin{aligned} 0 \leq F(x^{k+1}) - F(x^*) &\leq -\frac{1-c_0}{t_k}\|x^{k+1}-x^k\|^2 - \frac{1}{t_k}\langle x^k - x^{k+1}, x^* - x^k\rangle \\ &\leq \frac{(c_0-1)\|x^{k+1}-x^k\|^2 + \|x^{k+1}-x^k\|\|x^k-x^*\|}{t_k}, \quad \text{for all } k \geq k^*. \end{aligned} \tag{3.33}$$

However, $\{x^k\}$ is bounded (by Lemma 3.3(ii)) and $\lim\limits_{k\to+\infty} t_k = t^*$ (from Lemma 3.4) then combining with (3.32) we deduce that the limitation of the right hand side of (3.33) is zero as k tending to infinity. Hence, again, by (3.33) we have $\lim\limits_{k\to+\infty} F(x^k) = F_*$.

(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-continuous and therefore $F(\bar{x}) \leq \lim\limits_{k_i\to\infty} F(x^{k_i}) = F_*$, 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 2.2 we obtain that $\{x^k\}$ converges to some element of X^*.

(iii) In (3.31), substituting k by j then summing up it from $j = k^*$ to k we derive that

$$F(x^{k^*}) - F(x^{k+1}) \geq \frac{1-c_0}{t^*}\sum_{j=k^*}^{k}\|x^{j+1}-x^j\|^2. \tag{3.34}$$

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

$$\sum_{j=k^*}^{+\infty}\|x^{j+1}-x^j\|^2 \leq \frac{t^*}{1-c_0}\left(F(x^{k^*}) - F_*\right). \tag{3.35}$$

Applying (3.27) again, we obtain that

$$\begin{aligned} F(x^*) - F(x^{j+1}) &\geq \frac{1}{2t_j}\left(\|x^{j+1}-x^j\|^2 + 2\left\langle x^j - x^{j+1}, x^* - x^j\right\rangle\right) + \left(\frac{1}{2} - c_0\right)\frac{\|x^j - x^{j+1}\|^2}{t_j} \\ &\geq \frac{1}{2t_j}\left(\|x^* - x^{j+1}\|^2 - \|x^* - x^j\|^2\right) + \left(\frac{1}{2} - c_0\right)\frac{\|x^j - x^{j+1}\|^2}{t_j} \quad \forall j \geq k^*. \end{aligned} \tag{3.36}$$

On the other hand, Remark 3.2 gives $t_j \geq t_{k^*}\ \forall j \geq k^*$ which helps to infer the following inequality from (3.36)

$$\begin{aligned} &2t_{k^*}\left(F(x^{j+1}) - F(x^*)\right) \leq 2t_j\left(F(x^{j+1}) - F(x^*)\right) \\ &\leq \left(\|x^* - x^j\|^2 - \|x^* - x^{j+1}\|^2\right) + (2c_0 - 1)\|x^j - x^{j+1}\|^2 \quad \forall j \geq k^*. \end{aligned} \tag{3.37}$$

Summing (3.37) side by side for $j = k^*$ to $k + k^* - 1$ $(k \geq 1)$, we get that

$$2t_{k^*}\left(\sum_{j=k^*}^{k+k^*-1} F(x^{j+1}) - kF(x^*)\right) \leq \left(\|x^* - x^{k^*}\|^2 - \|x^* - x^{k+k^*}\|^2\right) + \\ + (2c_0 - 1)\sum_{j=k^*}^{k+k^*-1} \|x^j - x^{j+1}\|^2 \\ \leq D, \tag{3.38}$$

where, (from (3.35))D is defined by

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

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 (3.38), we have

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

which means that $F(x^k) - F(x^*) \leq \dfrac{D}{2t_{k^*}}\dfrac{1}{k - k^*} = O\left(\dfrac{1}{k}\right) \quad \forall k \geq k^* + 1.$

Next, we prove a stronger convergent result of Algorithm 3.1 if f is locally strongly convex. The details is the following.

Theorem 3.2 *Assuming that $c_0 \leq \frac{1}{2}$ and problem (P) satisfies Assumption 1, Assumption 2. Additionally, f is locally strongly convex then the sequence $\{x^k\}$ generated by Algorithm 3.1 satisfies*

$$\|x^{k+1} - x^*\|^2 \leq (1 - \sigma t_{k^*})\|x^k - x^*\|^2, \ \ \forall k \geq k^*, \tag{3.39}$$

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

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 argument of formula (3.28) and (3.36) 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 all $x^* \in X^*, k \geq k^*$, Remember that $F(x^*) - F(x^{k+1}) \leq 0 \ \forall k$ hence

$$\frac{1}{2t_k}\|x^* - x^{k+1}\|^2 \leq \left(\frac{1}{2t_k} - \frac{\sigma}{2}\right)\|x^* - x^k\|^2, \ \ k \geq k^*. \tag{3.40}$$

By (3.40), Lemma 3.4(i) and Remark 3.2, we have: $\forall k \geq k^*$

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

which derives

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

The last inequality aims the Q-linear convergence rate of $\{x^k\}$.

Remark 3.3 **(Comparison with the related work)**

(i) It is observed that, in the case $g = 0$, NPG1 becomes NGD [19] with the bigger range of c_0, c_1. In particular, for NGD, $c_0, c_1 \in \left(0, \frac{1}{2}\right)$ but for NPG1, $c_0, c_1 \in \left(0, \frac{1}{\sqrt{2}}\right)$.

(ii) It is worth noting that very recently, Malitsky and Mishchenko [27] has developed their method AdGD [28] to be AdPG (Adaptive Proximal Gradient) for solving problem (P) with the convex f satifying locally Lipschitz gradient assumption. The stepsize is defined by

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

where $\theta_0 = \frac{1}{3}, \theta_k = \frac{t_k}{t_{k-1}}, k \geq 1$ and $[t]_+ = \max\{t, 0\}$ for $t \in \mathbb{R}$. The iterates of AdPG is proved to converge to an optimal solution of (P) with the complexity $O(\frac{1}{k})$ of $\min\limits_{1\leq i\leq k}(F(x^i) - F_*)$. However, the lack of descent property of AdPG deduces two obstacles

(a) the first one is in producing the convergent result $O(\frac{1}{k})$ of $F(x^k) - F_*$ and the Q-linear rate of $\{x^k\}$ generated by AdPG in the case f assumed to be locally strongly convex. This restriction can be seen as one of open questions mentioned in [27]. Fortunately, as presented above, our proposed method (NPG1) in this paper is able to fill all these gaps. Moreover, the lack of descent property of AdPG

(b) the second one is in the capability of extending to the nonconvex case of (P). However, with NPG1 stepsize, in the upcoming section, we extend it to work for a class of nonconvex composite optimization models.

4 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 is globally Lipschitz gradient with constant L_f on $\text{int}(\text{dom}(f))$.

(ii) For $u, v \in \text{int}(\text{dom}(f))$, 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 $\{t \in \mathbb{R} \mid u+t(v-u) \in \text{int}(\text{dom}(f))\} \supset [0,1]$ (since $\text{int}(\text{dom}(f))$ 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, f is a concave function then $F = f + g$ is actually the difference of the two convex functions, or in other words, F belongs to the class of *dc functions*.

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

From Example 4.1 and 4.2, 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.

Algorithm 4.1 (NPG2)

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

Step 1.

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

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

The convergence of Algorithm 4.1 is established after some lemmas analogous to the ones of Section 3.

Lemma 4.1 *The sequence $\{t_k\}$ in Algorithm 4.1 satisfies $\inf_{k\geq 0} t_k > 0$ and has a positive limitation.*

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

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

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

Proof The proof is the same as in Lemma 3.5.

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*

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

Proof Invoking the Fundamental Theorem of Calculus, we have

$$\begin{aligned}
f(x^{k+1}) - f(x^k) &= \int_0^1 \left\langle \nabla f(x^k + t(x^{k+1} - x^k)), x^{k+1} - x^k \right\rangle dt \\
&= \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \int_0^1 u_k(t) dt, \ \forall k \geq \bar{k}
\end{aligned} \tag{4.2}$$

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

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

Thereafter, using Lemma 4.2, we derive that

$$\int_0^1 u_k(t)dt \leq \frac{c_0}{t_k}\|x^{k+1}-x^k\|^2, \ \ \forall k \geq \bar{k}. \tag{4.3}$$

Now, combining (4.2), (4.3) and Lemma 2.1(ii) with $x = x^{k+1}$ we get that

$$\begin{aligned} F(x^k) - F(x^{k+1}) &= f(x^k) - f(x^{k+1}) + g(x^k) - g(x^{k+1}) \\ &\geq -\left\langle x^{k+1}-x^k, \nabla f(x^k)\right\rangle - \frac{c_0}{t_k}\|x^{k+1}-x^k\|^2 + \\ &\quad + \left\langle x^{k+1}-x^k, \nabla f(x^k) + \frac{x^{k+1}-x^k}{t_k}\right\rangle \\ &= \frac{1-c_0}{t_k}\|x^{k+1}-x^k\|^2 \ \ \forall k \geq \bar{k}. \end{aligned} \tag{4.4}$$

The following theorem gives the convergence of Algorithm 4.1 for solving the problem (P).

Theorem 4.1 *Under Assumption 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) $F(x^k) - F(x^{k+1}) \to 0$ and $\sum\limits_{k=0}^{+\infty}\|x^{k+1}-x^k\|$ is convergent.

Proof (i) By (4.4) 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 $x^{k+1} = x^k = \text{Prox}_{t_k g}(x^k)$ meaning x^k is a stationary point of (P).

(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_*. This follows the existence of a finite limitation $\hat{F}$ of $\{F(x^k)\}_{k\geq\bar{k}}$ ($\hat{F} \geq F_*$). It means that $F(x^k) - F(x^{k+1}) \to 0$. Moreover, by Lemma 4.1 we have $\{t_k\}_{k\geq\bar{k}}$ increasing to $\lim\limits_{k\to+\infty} t_k = t^*$. On the other hand, inequality (4.4) indicates that $\|x^{k+1}-x^k\|^2 \leq \frac{t_k}{1-c_0}(F(x^k) - F(x^{k+1})) \leq \frac{t^*}{1-c_0}(F(x^k) - F(x^{k+1}))$ for all $k \geq \bar{k}$. Therefore $\sum\limits_{k=\bar{k}}^{+\infty}\|x^k - x^{k+1}\| \leq F(x^{\bar{k}}) - \hat{F}$ that follows the desired conclusion.

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).

(ii) Actually, the command (4.1) in Algorithm 4.1 is optional since we do not need it during the proof of the convergence of NPG2.

5 Problem (P) with quadratic function f

In this section, we propose an extension of Algorithm 4.1 called *NPG-quad* solving problem (P) with quadratic function f, i.e., $f(x) = \frac{1}{2}x^TAx + b^Tx$ as described in Example 4.2. With the range of c_0, c_1 in $(0,2)$, the stepsize in NPG-quad can be bigger than the previous ones. This probably makes the execution time of NPG-quad shorter.

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 $x^1 = P_{t_0 g}(x^0), t_{-1} = t_0$, 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 $x^{k+1} = P_{t_k g}(x^k)$.
Step 3. If $\|x^{k+1} - x^k\| < \varepsilon$ **then** STOP **else** setting $k := k + 1$ and return to **Step 1**.

Lemma 5.1 *The sequence $\{t_k\}$ generated by Algorithm 5.1 has a positive limitation.*

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 (ii).

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

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

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.

Theorem 5.1 *Supposing problem (P) satisfies Assumption 1 and f has quadratic form as in Example 4.2. For $\{x^k\}$ generated by Algorithm 5.1, the sequence $\{F(x^k)\}_{k\geq\tilde{k}}$ is decreasing to a limitation $\tilde{F} \geq F_*$ and $\sum_{k=0}^{+\infty} \|x^{k+1} - x^k\|$ is convergent.*

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 2.1(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 $k \geq \tilde{k}$,

$$F(x^k) - F(x^{k+1}) \geq \left(1 - \frac{c_0}{2}\right) \frac{\|x^{k+1} - x^k\|^2}{t_k}. \tag{5.8}$$

The remaining arguments are similar as 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 limitation;
- the evaluation (5.8) should be

$$F(x^k) - F(x^{k+1}) \geq \frac{\|x^{k+1} - x^k\|^2}{t_k}, \ \forall k \geq 0. \tag{5.9}$$

6 Numerical experiments

In this section, we investigate the performance of our new stepsize for the proximal gradient scheme by comparing our Algorithms 3.1(NPG1), 4.1 (NPG2) and 5.1 (NPG-quad) with: 1. the AdPG proposed by Malitksy and Mischenko [27], 2. the proximal gradient algorithms ProxGD(s,r) with stepsize selection based on an improved version of Armijo's backtracking procedure, i.e., For $s > 1$, $r < 1$, Armijo's linesearch in finds the largest $t_k = sr^i t_{k-1}$ for $i = 0, 1, ...$ such that $f(x^{k+1}) \leq f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{1}{2t_k}\|x^{k+1} - x^k\|^2$.. In the implementation we put $(s,r) = (1.1, 0.5)$ or $(1.2, 0.5)$. The chosen parameters for ProxGD are taken as the two most effective sets from the observation on the numerical results provided in [27]. For our algorithms, we use the convergent series $\sum\limits_{k=0}^{+\infty} \gamma_k$ defined by

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

and setting $(c_0, c_1) = (0.7, 0.69)$ for NPG1, $(c_0, c_1) = (0.99, 0.98)$ for NPG2 and NPG-quad. For all implemented algorithms, the stopping criterion is either the residual $\|x^{k+1} - x^k\| \leq 1e-06$ or the number of iterations over N_{max}.

We conduct experiments on five typical optimization problems with various sizes for each one. The average results on 10 randomly generated data for each size of considered problems with respect to

(i) the number of iterations (*Iter.*),

(ii) $\|x^{k+1} - x^k\|$ (*Res.*),

(iii) $F(x^k) - F_*$ (*Obj.*), where F_* is computed as the minimum of $F(x^k)$ over all iterations and all tested algorithms,

(iv) running time in seconds (*Time(s)*).

The details are reported on Tables 1, 2, 3, 4, 5. We emphasize the best results among all by bold characters and the worst results by italic type. We also choose one arbitrary data for each kind of problems to illustrate the performance by Figures 1, 2, 3, 4, 5.

All experiments were implemented in Python and executed on a personal computer equipped with a 12th Gen Intel(R) Core(TM) i7-1260P 2.10 GHz processor, RAM 16.0 GB. For details see our data repository at https://github.com/hoaiphamthi/NPG-for-composite-models

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., [5,11,17]. By using the similar rules in [17], 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 $\lambda = 0.01\|A^T b\|_\infty$. 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 also. Figure 1 illustrates the performance of mentioned algorithms for one of randomly generated data with $m = 2048$, $n = 8192$. The obtained average results in Table 1 show the best performance of NPG-quad for almost dimensions of Lasso.

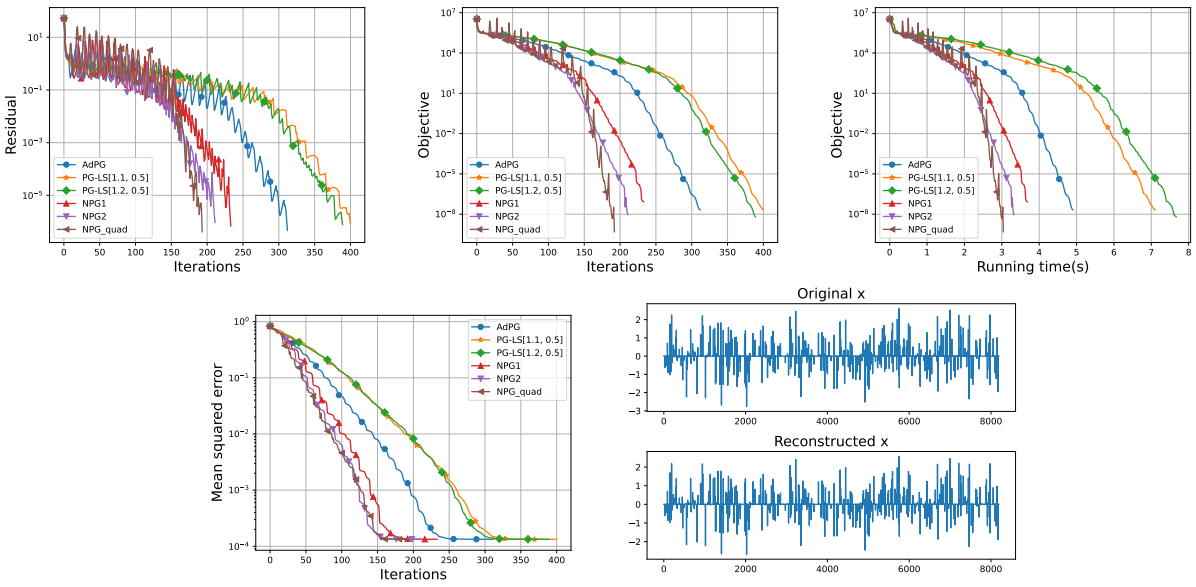

Fig. 1: Illustration for one of randomly generated data of Lasso with size $m = 2048, n = 8192$.

Size		Metrics	Average of all datasets					
m	n		AdPG	PG-LS (1.1, 0.5)	PG-LS (1.2, 0.5)	NPG1	NPG2	NPG-quad
512	1024	Iter.	114,4	*146,7*	138,4	92,1	85,4	**79,7**
		Res.	7,95E-07	6,76E-07	6,29E-07	*7,99E-07*	7,16E-07	**6,25E-07**
		Obj.	1,07E-10	7,05E-11	7,52E-11	*3,45E-10*	1,05E-10	**1,03E-11**
		Time(s)	0,041622	*0,060324*	0,057674	0,029767	0,027424	**0,025698**
512	2048	Iter.	307,7	*402,9*	381,5	235,7	**197,6**	204,7
		Res.	7,26E-07	8,05E-07	*8,55E-07*	6,1E-07	7,29E-07	**4,25E-07**
		Obj.	*8,71E-09*	2,71E-09	4,35E-09	7,23E-09	4,99E-09	**8,57E-11**
		Time(s)	0,133219	0,188929	*0,211354*	0,110316	**0,092419**	0,096476
512	4096	Iter.	5923,4	*8311,4*	8269,7	5690	4534,1	**3066,5**
		Res.	9,65E-07	9,68E-07	9,43E-07	*9,8E-07*	9,73E-07	**9,22E-07**
		Obj.	6,5E-06	1,2E-06	5,69E-07	*9,81E-06*	5,56E-06	**6,86E-08**
		Time(s)	5,106856	8,503539	*9,265539*	5,189247	4,11577	**2,790751**
1024	2048	Iter.	118,8	*153,6*	144,8	102	90,9	**89,6**
		Res.	*8,18E-07*	6,45E-07	5,9E-07	7,94E-07	7,82E-07	**5,68E-07**
		Obj.	3,23E-10	1,97E-10	1,55E-10	*9,11E-10*	2,64E-10	**3,34E-11**
		Time(s)	0,091836	0,127233	*0,138282*	0,081234	**0,073686**	0,076868
1024	4096	Iter.	282,6	*366,6*	342,2	221,7	**187,7**	188,8
		Res.	7,57E-07	*9,1E-07*	7,5E-07	7,24E-07	7,46E-07	**5,91E-07**
		Obj.	1,13E-08	4,27E-09	6E-09	*1,89E-08*	1,11E-08	**9,4E-11**
		Time(s)	0,900778	*1,335362*	1,334489	0,690471	**0,581451**	0,588497
1024	8192	Iter.	5422,5	*7953*	7839,9	5431,7	4345,8	**2967,5**
		Res.	**9,42E-07**	9,7E-07	9,45E-07	9,61E-07	*9,78E-07*	9,43E-07
		Obj.	1,76E-05	2,34E-06	1,65E-06	*1,84E-05*	1,14E-05	**4,27E-07**
		Time(s)	41,86462	69,53798	*75,38844*	41,31976	33,15261	**23,23451**
2048	4096	Iter.	107	*135,6*	129,3	97,5	86,6	**79,2**
		Res.	*7,76E-07*	7,48E-07	7,19E-07	7,43E-07	7,57E-07	**5,46E-07**
		Obj.	4,13E-10	5,13E-10	3,07E-10	*1,37E-09*	3,69E-10	**1,16E-10**
		Time(s)	0,905618	1,328207	*1,350697*	0,79346	0,706674	**0,646555**
2048	8192	Iter.	289,1	*380,7*	361,1	226,8	199,6	**183,5**
		Res.	7,52E-07	7,85E-07	*8,42E-07*	7,67E-07	7,11E-07	**5,15E-07**
		Obj.	*3,93E-08*	1,31E-08	1,33E-08	3,65E-08	2,58E-08	**5,18E-10**
		Time(s)	4,878866	6,889515	*7,239163*	3,60845	3,178414	**2,932337**

Table 1: Average results for Lasso problem ($N_{max} = 15000$).

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 $(0,0),(1,x_1),...,(n,x_n)$ while satisfying the equality constraint $Ax = b$, the problem can be formed as

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

where $A \in \mathbb{R}^{m\times n}, b \in \mathbb{R}^m$. It is seen that Min-length[1] satisfies *Assumption 1,2,3* and we can use NPG1 and NPG2 to solve it. In the implementation, all members of A are randomly generated by normal distribution $\mathcal{N}(0,1)$. Taking $b = Ax^*$, where $x^* \sim \mathcal{N}(0,1)$. Figure 2 provides the line graphs of one randomly generated data with $m = 2000, n = 10000$. Table 2 includes the average computation results for various sizes of Min-length. Notably, both NPG1 and NPG2 outperform the remaining ones with the big deviation in term of computational time, residual, objective

[1] 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 $g(x) = \imath_C$ (the indicator function of C) with $C = \{x \in \mathbb{R}^n \mid Ax = b\}$.

value and the number of iterations. The speed of NPG1 can be seen as the best among all for Min-length.

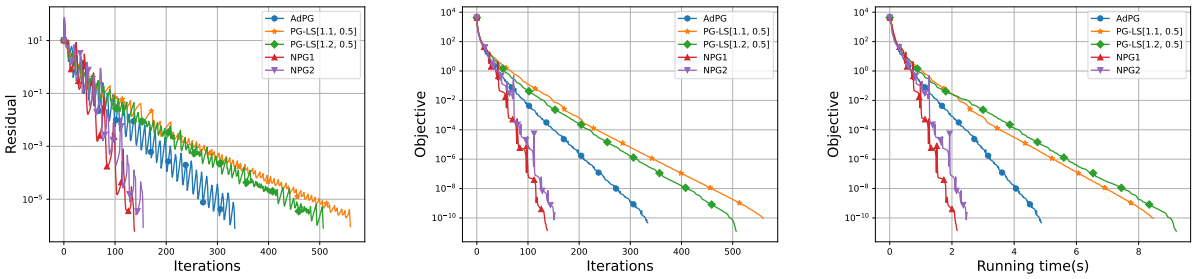

Fig. 2: Illustrations for one of randomly generated data of Min-length with $m = 2000, n = 10000$.

Size		Metrics	Average of all datasets				
m	n		AdPG	PG-LS (1.1, 0.5)	PG-LS (1.2, 0.5)	NPG1	NPG2
50	5000	Iter.	45399,2	*50000*	*50000*	**14476,8**	30012,6
		Res.	3,72E-06	*1,63E-05*	1,51E-05	**9,92E-07**	9,94E-07
		Obj.	9,35E-08	*7,17E-06*	6,71E-06	2,68E-08	**0**
		Time(s)	15,65872	19,51051	*21,40545*	**4,981068**	10,22737
500	5000	Iter.	1035,1	1623,9	*1631,4*	357,2	**328,4**
		Res.	*9,44E-07*	8,82E-07	8,68E-07	7,73E-07	**7,67E-07**
		Obj.	*3,1E-10*	1,88E-10	1,51E-10	1,94E-10	**7,77E-11**
		Time(s)	1,632963	3,080268	*3,386024*	0,587674	**0,533654**
2000	5000	Iter.	120,4	*165,1*	163,7	**73,9**	87,9
		Res.	**6,07E-07**	6,82E-07	*7,87E-07*	6,75E-07	6,28E-07
		Obj.	1,36E-11	*1,41E-11*	1,33E-11	**2,91E-12**	1,14E-11
		Time(s)	1,130271	1,666739	*1,711799*	**0,604635**	0,718454
100	10000	Iter.	49008,7	*50000*	*50000*	**17646,5**	36450,4
		Res.	8,29E-06	3,84E-05	*3,97E-05*	**9,85E-07**	9,88E-07
		Obj.	3,7E-07	*2,68E-05*	2,49E-05	3,87E-08	**0**
		Time(s)	36,15325	42,87928	*47,1404*	**13,09231**	27,43948
1000	10000	Iter.	1052,9	*1614,2*	1609,5	367,4	**354,2**
		Res.	*9,47E-07*	**6,35E-07**	7,61E-07	7,29E-07	7,71E-07
		Obj.	3,6E-10	*3,86E-10*	3,22E-10	1,37E-10	**7,55E-11**
		Time(s)	8,05511	13,69401	*15,06101*	2,742484	**2,656093**
2000	10000	Iter.	330,1	*526*	500,3	**140**	181,3
		Res.	*8,38E-07*	6,99E-07	5,91E-07	7,34E-07	**5,88E-07**
		Obj.	*1,17E-10*	9,79E-11	1,09E-10	4,27E-11	**2,55E-12**
		Time(s)	5,022146	8,726909	*9,14353*	**2,041686**	2,64932

Table 2: Average results for Min-length problem ($N_{max} = 50000$).

6.3 Dual of the entropy maximization problems

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

$$\min \sum_{i=1}^{n} x_i \log x_i \quad \text{s.t.} \quad Ax \leq b, \ \sum_{i=1}^{n} x_i = 1, \ \text{and} \ x_i > 0, i = 1, ..., n, \tag{6.1}$$

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

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

It is observed that Problem Dual-max-entropy[2] matches *Assumption 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 [27]. 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 are depicted in Table 3 and Figure 3. It is shown that the performance of NPG2 significant efficiency compared to the remaining ones.

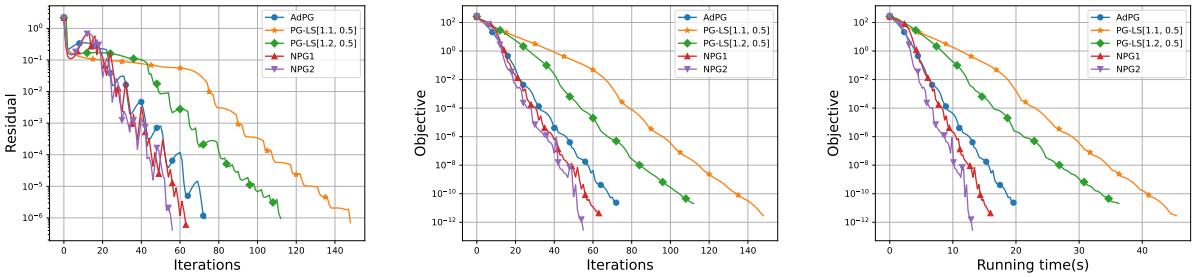

Fig. 3: Illustration for one of randomly generated data of Dual-max-entropy with $m = 4000, n = 5000$.

Size		Metrics	Average of all datasets				
m	n		AdPG	PG-LS (1.1, 0.5)	PG-LS (1.2, 0.5)	NPG1	NPG2
100	500	Iter.	32,7	*80*	51,1	30,6	**29,1**
		Res.	**4,69E-07**	*5,9E-07*	4,72E-07	5,67E-07	4,75E-07
		Obj.	3,85E-14	1,19E-13	1,04E-13	*3,1E-13*	**1,57E-14**
		Time(s)	0,02434	*0,040698*	0,027038	0,013777	**0,013373**
500	2000	Iter.	35,3	*83,7*	54,8	33,4	**31,9**
		Res.	7,44E-07	*7,9E-07*	5,96E-07	6,3E-07	**4,97E-07**
		Obj.	2,08E-13	*7,62E-13*	1,49E-13	7,41E-13	**4,89E-14**
		Time(s)	0,492897	*1,27927*	0,886613	0,496571	**0,466446**
2000	4000	Iter.	50,1	*102,1*	70,2	47,5	**45,9**
		Res.	5,68E-07	*8,26E-07*	7,53E-07	5,02E-07	**4,8E-07**
		Obj.	**2,18E-14**	8,17E-13	7,68E-13	*1,68E-12*	6,67E-13
		Time(s)	5,598936	*12,29137*	9,333802	5,594557	**5,436425**
4000	5000	Iter.	79,6	*151,7*	116,1	73,1	**60,4**
		Res.	6,28E-07	*7,63E-07*	7,42E-07	6,56E-07	**4,28E-07**
		Obj.	6,27E-12	4,53E-12	*6,53E-12*	2,94E-12	**1,39E-12**
		Time(s)	21,63522	*46,93658*	38,94247	20,23978	**16,44809**

Table 3: Average results for Dual-max-entropy problem ($N_{max} = 200$).

6.4 Maximum likelihood estimate of the information matrix

This problem (see [10, Equation (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

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

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

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[3] satisfies *Assumption 1,2,3* then NPG1 and NPG2 are exact methods to solve Max-likelyhood. The dataset for the implementation is generated analogously to [27] 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 obtained results are shown in Table 4 and Figure 4. It is seen that for Max-likelyhood, both of NPG1 and NPG2 provide better results compared to the others with the big deviation. And most of cases NPG2 performs best.

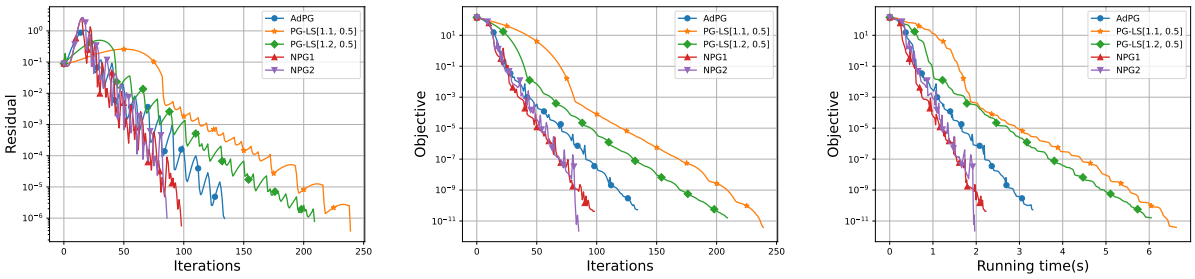

Fig. 4: Illustrations for one of randomly generated data of Max-likelyhood with $n = 100, l = 0.1, u = 10, M = 500$.

Size	Metrics	Average of all datasets				
n, l, u, M		AdPG	PG-LS (1.1, 0.5)	PG-LS (1.2, 0.5)	NPG1	NPG2
100, 0.1, 10, 50	Iter.	1661,5	*2439*	2364,5	1259,7	**1171,8**
	Res.	*9,58E-07*	8,68E-07	9,16E-07	9,21E-07	**8,59E-07**
	Obj.	4,27E-09	1,94E-09	2,74E-09	*6,45E-09*	**8,4E-10**
	Time(s)	44,42071	74,11472	*82,07483*	32,88484	**28,012**
100, 0.1, 10, 500	Iter.	133,7	*219,2*	197,7	103,5	**93,6**
	Res.	7,15E-07	6,76E-07	*7,42E-07*	**5,66E-07**	6,45E-07
	Obj.	2,69E-11	1,29E-11	*2,93E-11*	1,7E-11	**7,07E-12**
	Time(s)	3,48568	*6,391397*	6,273465	2,579751	**2,252374**
100, 0.1, 10, 1000	Iter.	57,9	*103,9*	83,8	58	**49,7**
	Res.	5,69E-07	4,91E-07	**4,44E-07**	*7,56E-07*	6,19E-07
	Obj.	3,46E-12	*5,79E-12*	4,9E-12	**8,64E-13**	2,1E-12
	Time(s)	1,53195	*2,907321*	2,620872	1,525524	**1,305989**
30, 0.1, 1000, 50	Iter.	5210,2	*7612,8*	7518,9	4684,2	**3295,8**
	Res.	9,69E-07	*2,05E-06*	1,86E-06	**9,3E-07**	9,49E-07
	Obj.	4,28E-09	*1,34E-07*	1,2E-07	4,69E-09	**1,54E-09**
	Time(s)	6,992612	11,99245	*12,3711*	6,013213	**4,227009**
50, 0.1, 1000, 100	Iter.	1644,2	*2589,4*	2545,9	1193,8	**954,1**
	Res.	*9,4E-07*	**8,62E-07**	9,01E-07	8,7E-07	8,67E-07
	Obj.	8,07E-10	4,87E-10	3,91E-10	*1,35E-09*	**1,52E-10**
	Time(s)	10,78987	19,67659	*20,37514*	7,702965	**6,14909**

Table 4: Average results for Max-likelyhood problem ($N_{max} = 20000$).

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

6.5 Nonnegative matrix factorization

One of efficient approaches to solve recommendation system problems [30] is based on nonnegative matrix factorization[4]

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

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 [27], 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 computational results are reported in Table 5 and illustrated by Figure 5. For this problem, NPG1 and NPG2 are alternative the most effective method in comparison with the remaining ones.

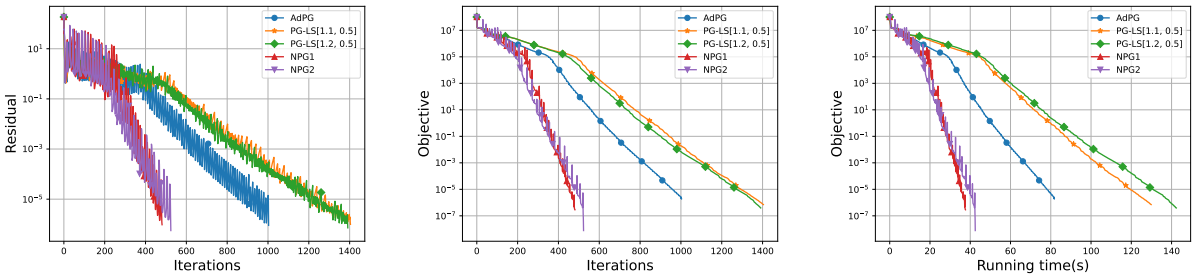

Fig. 5: Illustrations for one of randomly generated data of NMF problem with $m = 3000, n = 3000, r = 30$.

7 Conclusions

In this paper, we propose an efficient explicit stepsize applied for the proximal gradient (PG) scheme. In particular, NPG1 solves the convex situation of problem (P) under locally Lipschitz gradient condition imposed on f. The iterates is proved to converge to an optimal solution of (P) with the complexity computation $O(\frac{1}{k})$ of $F(x^k) - F_*$ and the Q-linear rate if f has local strong convexity property. These convergence results are based on the descent of our proposed method. Moreover, the extensions of NPG1 that NPG2 and NPG-quad are also designed for (P) in case of nonconvex f. Our stepsize selection is computed quickly by a closed formulas without linesearch computation or estimating some constant (like Lipschitz constant of gradient) to ensure the convergence of the PG algorithm. Moreover, the increasing of the sequence of our stepsizes from some fixed iteration opens the ability to speed up the corresponding 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 includes deploying our adaptive stepsize for the composite models in the absence of both convexity and global Lipschitz gradient assumptions on f.

References

1. Ahookhosh, M., Themelis, A., Patrinos, P.: A Bregman forward-backward linesearch algorithm for nonconvex composite optimization: superlinear convergence to nonisolated local minima. SIAM J. Optim. 31(1), pp. 653-685 (2021)

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

Size			Metrics	Average of all datasets				
m	r	n		AdPG	PG-LS (1.1, 0.5)	PG-LS (1.2, 0.5)	NPG1	NPG2
500	20	1000	Iter.	537	*801,4*	746,3	**302,6**	308,5
			Res.	*9,07E-07*	8,8E-07	8,84E-07	7,93E-07	**6,6E-07**
			Obj.	*1,93E-07*	6,04E-08	8,27E-08	1,04E-07	**1,37E-08**
			Time(s)	5,941108	*9,572103*	9,001703	**2,949437**	2,95718
1000	20	500	Iter.	543,9	*777,7*	751,7	**300,5**	309,9
			Res.	8,42E-07	*9,04E-07*	8,7E-07	8,57E-07	**7,76E-07**
			Obj.	*1,44E-07*	7,8E-08	5,78E-08	**2,05E-08**	2,85E-08
			Time(s)	4,319812	7,194566	*7,430679*	2,572676	**2,431692**
2000	20	3000	Iter.	506,7	*731,9*	699,9	**301**	302,6
			Res.	8,33E-07	*9,29E-07*	9,01E-07	**7,56E-07**	7,69E-07
			Obj.	*4,86E-07*	2,13E-07	1,65E-07	1,49E-07	**1,44E-07**
			Time(s)	31,53794	51,34798	*55,23237*	**18,86799**	19,00453
3000	20	2000	Iter.	509,8	*716,2*	672,7	**290,1**	305
			Res.	8,26E-07	8,26E-07	*8,82E-07*	8,15E-07	**6,7E-07**
			Obj.	*4,56E-07*	1,08E-07	2,11E-07	2,65E-07	**6,11E-08**
			Time(s)	34,89519	56,01866	*57,75551*	**19,84947**	21,09172
3000	20	3000	Iter.	498,1	*701*	671,9	**275,3**	276,9
			Res.	**7,95E-07**	8,39E-07	*8,97E-07*	8,74E-07	8,11E-07
			Obj.	*4,63E-07*	1,49E-07	2,44E-07	2,08E-07	**5,89E-08**
			Time(s)	43,91157	69,55736	*74,0461*	**24,11349**	24,23001
500	30	1000	Iter.	982,7	*1493,6*	1422,9	633,6	**598,5**
			Res.	*9,38E-07*	8,85E-07	9,01E-07	**8,37E-07**	8,84E-07
			Obj.	*4,76E-07*	1,85E-07	1,38E-07	4,04E-07	**6,43E-08**
			Time(s)	9,063069	16,18621	*16,60831*	5,608065	**5,171566**
1000	30	500	Iter.	1026,1	*1502,3*	1430,2	603,3	**587,3**
			Res.	*9E-07*	8,93E-07	8,57E-07	**7,87E-07**	8,63E-07
			Obj.	*4,28E-07*	1,78E-07	1,09E-07	2,44E-07	**3,35E-08**
			Time(s)	7,197391	12,64285	*13,08158*	4,361314	**3,594271**
2000	30	3000	Iter.	876,2	*1247,9*	1200,2	**435,5**	467,2
			Res.	8,75E-07	8,78E-07	8,77E-07	*8,94E-07*	**7,64E-07**
			Obj.	*1,49E-06*	2,88E-07	3,06E-07	3,27E-07	**1,1E-07**
			Time(s)	56,17322	96,50053	*115,0343*	**33,76644**	35,68324
3000	30	2000	Iter.	907,4	*1280*	1247,7	**439,6**	469,3
			Res.	8,95E-07	*9,1E-07*	9,06E-07	**7,71E-07**	8,06E-07
			Obj.	*1,47E-06*	5,77E-07	6,12E-07	4,89E-07	**1,3E-07**
			Time(s)	76,98802	117,8925	*125,9915*	**35,08839**	37,14086
3000	30	3000	Iter.	914,1	*1303,2*	1252,5	**457,9**	504
			Res.	8,81E-07	8,8E-07	*8,89E-07*	8,74E-07	**7,46E-07**
			Obj.	*1,7E-06*	4,86E-07	7,59E-07	**1,84E-07**	3,96E-07
			Time(s)	94,52072	*157,0848*	150,6168	**43,47812**	48,21538

Table 5: Average results for NMF problem ($N_{max} = 5000$).

2. Bauschke, H.H., Bolte, J.,Teboulle, M. : A descent lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications. Mathematics of Operations Research, 42(2), pp. 330-348 (2017)

3. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problem. SIAM J. Imaging Sci. 2, pp. 183–202 (2009)

4. Beck, A., Teboulle, M.: Gradient-based algorithms with applications to signal recovery problems. In: Palomar, D., Eldar, Y.C. (eds.) Convex Optimization in Signal Processing and Communications, pp. 139–162. Cambridge University Press, Cambridge (2009)

5. Beck, A.: Introduction to Nonlinear Optimization: Theory, Algorithms, and Applications with MATLAB. Society for Industrial and Applied Mathematics, USA (2014)

6. Beck, A.: First order methods in optimization, Society for Industrial and Applied Mathematics, USA (2017)
7. Bertsekas, D.P., Nonlinear programming, 3rd Edition, Athena Scientific (2016)
8. Bolte, J., Sabach, S., Teboulle, M., Vaisbourd, Y.: First order methods beyond convexity and Lipschitz gradient continuity with applications to quadratic inverse problems. SIAM J. Optim. 28(3), pp. 2131-2151 (2018)
9. Birgin, E.G., Martínez, J.M., Raydan, M.: Nonmonotone spectral projected gradient methods on convex sets. SIAM J. Optim. 10(4), pp. 1196-1211 (2000)
10. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge: Cambridge University Press, 2004
11. Combettes, P.L., Pesquet, J.-C.: Proximal splitting methods in signal processing. In: Bauschke, H.H., Burachik, R.S., Combettes, P.L., Elser, V., Luke, D.R., Wolkowicz, H. (eds.) Fixed-Point Algorithms for Inverse Problems in Science and Engineering, pp. 185–212. Springer, New York (2011)
12. Combettes, P.L., Wajs, V.R.: Signal recovery by proximal forward-backward splitting. Multiscale Model. Simul. 4(4), pp. 1168–1200 (2005)
13. Cohen, E., Hallak, N., and Teboulle, M.: Dynamic alternating direction of multipliers for nonconvex minimization with nonlinear functional equality constraints. Journal of Optimization Theory and Applications, 193, pp. 324–353 (2022)
14. Chen, X., Lu, Z., Pong, T.K.: Penalty methods for a class of non-Lipschitz optimization problems. SIAM J. Optim. 26(3), pp. 1465–1492 (2016)
15. De Marchi, A., Themelis, A.: Proximal Gradient Algorithms Under Local Lipschitz Gradient Continuity. J. Optim Theory Appl 194, pp. 771–794 (2022)
16. Dragomir, R.A., Taylor, A.B., d'Aspremont, A., Bolte, J.: Optimal complexity and certification of Bregman first-order methods. Math. Program. 194, pp. 41-83 (2022)
17. Figueiredo, MÁrio A.T., Nowak, R.D., Wright, S.J.: Gradient Projection for Sparse Reconstruction: Application to Compressed Sensing and Other Inverse Problems. IEEE Journal of Selected Topics in Signal Processing Vol. 1, No. 4, pp. 586-597 (2007)
18. Fukushima, M., Mine, H.: A generalized proximal point algorithm for certain non-convex minimization problems. Int. J. Syst. Sci. 12(8), pp. 989-1000 (1981)
19. Hoai, P.T., Vinh, N.T., Chung, N.P.H.: A novel stepsize for gradient descent method. Operations Research Letters, 107072 (2024) https://doi.org/10.1016/j.orl.2024.107072.
20. Iusem, A.N.: On the convergence properties of the projected gradient method for convex optimization. Comput. Appl. Math. 22, pp. 37-52 (2003)
21. Jia, X., Kanzow, C., Mehlitz, P.: Convergence Analysis of the Proximal Gradient Method in the Presence of the Kurdyka–Łojasiewicz Property Without Global Lipschitz Assumptions. SIAM Journal on Optimization, Vol. 33, No. 4, pp. 3038–3056 (2023)
22. Jia, X., Kanzow, C., Mehlitz, P., Wachsmuth, G.: An augmented Lagrangian method for optimization problems with structured geometric constraints. Math. Program. 199, pp. 1365–1415 (2023)
23. Kanzow, C., Mehlitz, P.: Convergence properties of monotone and nonmonotone proximal gradient methods revisited. Journal of Optimization Theory and Applications. 195(2), pp.624–646 (2022)
24. Li, G., Pong, T.K.: Global convergence of splitting methods for nonconvex composite optimization. SIAM J. Opt. 25(4), pp. 2434–2460 (2015)
25. Liu, H., Wang, T., Liu, Z.: Some modifed fast iterative shrinkage thresholding algorithms with a new adaptive non-monotone step- size strategy for nonsmooth and convex minimization problems, Comput. Optim. Appl. 83, pp. 651-691 (2022)
26. Lee, J.D., Sun, Y., Saunders, M. A.: Proximal Newton-type methods for minimizing composite functions, SIAM J. Optim., 24, pp. 1420-1443 (2014)
27. Malitsky, Y., Mishchenko, K.: Adaptive proximal gradient method for convex optimization https://arxiv.org/pdf/2308.02261.pdf
28. Malitsky, Y., Mishchenko, K.: Adaptive gradient descent without descent, ICML 119, pp. 6702-6712 (2020)
29. Parikh, N., Boyd, S.: Proximal algorithms. Found. Trends Optim. 1(3), pp. 127–239 (2014)
30. Symeonidis, P., Zioupos, A.: Matrix and Tensor Factorization Techniques for Recommender Systems. Springer Briefs in Computer Science (2016)

31. Teboulle, M.: A simplified view of first order methods for optimization. Math. Program. 170(1), pp. 67-96 (2018)
32. Themelis, A., Stella, L., Patrinos, P.: Forward-backward envelope for the sum of two nonconvex functions: further properties and nonmonotone linesearch algorithms. SIAM J. Optim. 28(3), pp. 2274- 2303 (2018)
33. Wright, S.J., Nowak, R.D., Figueiredo, M.A.T.: Sparse reconstruction by separable approximation. IEEE Trans. Signal Process. 57(7), pp. 2479-2493 (2009)