A novel stepsize for gradient descent method

Pham Thi Hoai[1,*], Nguyen The Vinh[2], Nguyen Phung Hai Chung[3]

Abstract. We propose a novel adaptive stepsize for the gradient descent scheme to solve unconstrained nonlinear optimization problems. With the convex and smooth objective satisfying locally Lipschitz gradient we obtain the complexity $O(\frac{1}{k})$ of $f(x^k) - f_*$ at most. By using the idea of the new stepsize, we propose another new algorithm based on the projected gradient for solving a class of nonconvex optimization problems over a closed convex set. The computational experiments show the efficiency of the new method.

Keywords. convex programming, gradient descent method, nonlinear programming, projected gradient method, constrained optimization problem

Mathematics Subject Classification (2010). 90C25, 90C06, 65K10

1. Introduction

The gradient descent (GD) algorithm is a standard algorithm with a rich history. It has a lot of applications for many modern real-life problems such as machine learning, deep learning, data science, etc. Although the idea of this method appeared a long time ago by Cauchy (1847) and became classical [14], it has received a lot of attention recently (see e.g., [3, 8, 11, 13, 16, 17, 18, 19] and references therein). This algorithm considers solving an important class of optimization problems

$$\min\{f(x) : x \in \mathbb{R}^n\}, \tag{P}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is smooth. Throughout the paper, we assume that problem (P) has optimal value $f_* > -\infty$ and $X^* \neq \emptyset$ is the optimal solution set of (P). Starting at some given point x^0, GD constructs a sequence $\{x^k\}$ by the following formula:

$$x^{k+1} = x^k - \lambda_k \nabla f(x^k). \tag{1}$$

The traditional condition imposed on f to ensure the convergence of gradient descent algorithm is the global Lipschitzness of ∇f, i.e., there exists $L > 0$ such that

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \ \forall x, y \in \mathbb{R}^n.$$

One can use some strategies for choosing the stepsize λ_k in (1) to control the performance of GD algorithm. Deterministic stepsize selection criterion fall under three main strategies:

(i) The first one takes a *constant stepsize*. One know that in order to ensure the convergence of GD scheme, the constant stepsize should belong to $(0, \frac{2}{L})$, see e.g [4, 5, 20]. The method's main advantage is its simple implementation. The convergence of the obtained GD works for differentiable objectives with global Lipschitz gradient property. Moreover, if f is convex one get the complexity $O(1/k)$ of $f(x^k) - f_*$. Nevertheless, it is not easy to compute or estimate the constant L of f in general. If the estimation is inexact, GD may not converge. Besides this difficulty, from the practical point of view, L may be large and defining the small stepsize. This may affect the speed of GD.

[*] Corresponding author

[1] Department of Applied Mathematics, School of Applied Mathematics and Informatics, Hanoi University of Science and Technology, 1 Dai Co Viet Road, Hanoi, Vietnam. *Email address:* hoai.phamthi@hust.edu.vn

[2] Department of Mathematics, University of Transport and Communications, 3 Cau Giay Street, Hanoi, Vietnam. *Email address:* thevinhbn@utc.edu.vn

[3] Hanoi University of Science and Technology, 1 Dai Co Viet Road, Hanoi, Vietnam. *Email address:* hchung1997@gmail.com

(ii) The second one is *line search method* based on some rules such as the exact minimization rule or backtracking rules. It is known that the formulation of the exact minimization rule is very brief but its usefulness is just suitable for some specific classes of objective functions, like strongly convex quadratic functions; see [10] for more details. Conversely, the backtracking procedure is more generic, as it can be used for any kind of smooth objective function. The first works considering this topic are [1] and [9]. One can also find recent papers concerning analogous problems such as [7, 24] and references therein. If f is convex, the complexity of GD with backtracking stepsize is $O(1/k)$ for $f(x^k) - f_*$. However, this method may cause the expensive cost due to the backtracking computation and the stepsize may be very small at the large enough iterations.

(iii) The third one is an *adaptive method* that does not require estimating the Lipschitz constant L or backtracking calculation. For instance, the Polyak's stepsize [22] takes $\lambda_k = \frac{f(x^k)-f_*}{\|\nabla f(x^k)\|^2}$ or Barzilai-Borwein's stepsize [2] is $\lambda_k = \frac{\langle x^k - x^{k-1}, \nabla f(x^k) - \nabla f(x^{k-1})\rangle}{\|\nabla f(x^k) - \nabla f(x^{k-1})\|^2}$. In fact, there are some kinds of optimization problems we know their optimal values in advance. For example, $f_* = 0$ in the split feasibility problem considered in [27]. Therefore we can easily apply Polyak's stepsize for these ones. Nevertheless, we do not know f_* in general. We should estimate this value to ensure the convergence of GD. With Barzilai -Borwein's method, the convergence of GD works for only limited class of objectives. In particular, "it may not converge, even when the objective function is strongly convex" [6]. Note that the Polyak's stepsize for GD has linear convergence rate for strongly convex objectives [23].

Recently, Y. Malitsky and K. Mishchenko [17] suggested selecting an adaptive stepsize as "a certain approximation of the inverse local Lipchitz constant"

$$\lambda_k = \min\left\{\sqrt{1+\theta_{k-1}}\lambda_{k-1}, \frac{\|x^k - x^{k-1}\|}{2\|\nabla f(x^k) - \nabla f(x^{k-1})\|}\right\}, k \geq 1,$$

where $\theta_0 = +\infty, \theta_k = \frac{\lambda_k}{\lambda_{k-1}}, k \geq 1$. The proposed algorithm is called Adaptive Gradient Descent (AdGD). It is shown that this algorithm involves not only nice theoretical properties but also good practical ability. In particular, AdGD is convergent for locally Lipschitz gradient and convex objective functions. Besides showing the convergence of $\{x^k\}$ to an optimal solution of problem (P), the authors proved that $f(\hat{x}^k) - f(x^*) \leq O(\frac{1}{k})$, where $x^* \in X^*$ and $\hat{x}^k$ is an ergodic vector obtained from $\{x^k\}$

$$\hat{x}^k = \frac{\lambda_k(1+\theta_k)x^k + \sum_{i=1}^{k-1}\left(\lambda_i(1+\theta_i) - \lambda_{i+1}\theta_{i+1}\right)x^i}{\lambda_k(1+\theta_k) + \sum_{i=1}^{k-1}\left(\lambda_i(1+\theta_i) - \lambda_{i+1}\theta_{i+1}\right)}, k \geq 1.$$

Additionally, if f is locally strongly convex, the complexity to get $\|x^k - x^*\| < \varepsilon$ is $O(\frac{L}{\mu}\log\frac{1}{\varepsilon})$, with L, μ the locally smoothness and strong convexity constants of f. However, there are some remaining questions related to AdGD: 1. Besides getting the complexity result of $f(\hat{x}^k) - f_*$, can we give an evaluation of complexity for $f(x^k) - f_*$? 2. Whether the sequence of stepsize can be monotone? 3. Is AdGD convergent for nonconvex objectives even adding the global Lipschitzness of ∇f? 4. Can we extend this stepsize for a nonlinear optimization problem over a closed convex set?

Contributions: Motivated by the above questions, in this paper we propose a new adaptive stepsize for GD scheme that includes the following interesting features:

(i) The convergence of gradient descent algorithm with our new stepsize ($\{x^k\}$ converges to an optimal solution of (P)) is obtained for **smooth, locally Lipschitz gradient and convex objectives**.

(ii) We show the complexity computation $O(\frac{1}{k})$ **for** $f(x^k) - f_*$. Moreover, **the sequence of our stepsize** $\{\lambda_k\}_{k\geq\bar{k}}$ **is increasing to a limit** λ^* **(**$\bar{k}$ **is a finite number)**. If f satisfies an additional condition that locally strongly convex we obtain **the linear convergent rate of** $\{x^k\}$.

(iii) If f is global Lipschitz gradient, our new stepsize is extensible to the gradient projection algorithm for solving a class of **nonconvex, nonlinear optimization problems over a closed and convex set.** Consequently, it is applicable for a class of **nonconvex case of problem (P).**

The rest of the paper is organized as follows. In Section 2, we propose a GD algorithm with our novel stepsize for solving the convex case of (P). We analyze and prove the convergence of this algorithm for two situations of f as presented above. A new version of the projected gradient algorithm using our stepsize is proposed in Section 3. Numerical experiments for benchmark problems and synthetic data in Section 4 show the efficiency of our method compared to AdGD and GD with constant stepsize. The paper is finally closed with some conclusions in Section 5.

2. A new adaptive stepsize for GD scheme

Below is our new gradient algorithm (NGD) for solving (P) under the following assumptions:
Assumption 1: Problem (P) has a nonempty optimal solution set X^* and the optimal value $f_* > -\infty$.
Assumption 2: f is smooth, convex and locally Lipschitz gradient.

Algorithm 2.1 (NGD)

```
Step 0 (Initialization). Select $\lambda_0 > 0$, $0 < \eta_1 < \eta_0 < \frac{1}{2}$ 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
        $\|\nabla f(x^k) - \nabla f(x^{k-1})\| > \frac{\eta_0}{\lambda_{k-1}} \|x^k - x^{k-1}\|$    (2)
    then
        $\lambda_k = \eta_1 \frac{\|x^k - x^{k-1}\|}{\|\nabla f(x^k) - \nabla f(x^{k-1})\|}$    (3)
    else
        $\varepsilon'_{k-1} = \varepsilon_{k-1}$    (4)
        if $\frac{\lambda_{k-1}}{\lambda_{k-2}} < 1$ then update $\varepsilon'_{k-1} = \min\{\varepsilon_{k-1}, \sqrt{1 + \frac{\lambda_{k-1}}{\lambda_{k-2}}} - 1\}$    (5)
        $\lambda_k = (1 + \varepsilon'_{k-1})\lambda_{k-1}.$    (6)
Step 2. Compute $x^{k+1} = x^k - \lambda_k \nabla f(x^k)$.
Step 3. If $\|\nabla f(x^{k+1})\| < \epsilon$ then STOP
        else setting $k := k + 1$, and return to Step 1.
```

Utilizing well-known tools including Cauchy-Schwarz, convexity inequalities and potential functions [26], the authors of AdGD [17] eliminates the challenge of global Lipschitz gradient. In order to get the convergence of AdGD, they prove that the sequence $\{x^k\}$ and $\{\lambda_k\}$ of AdGD are bounded and lower bounded by a positive number, respectively. For Algorithm 2.1, we also provide the same properties in Lemma 2.1. Inspired by [15] we use a given convergent positive series $\sum_{k=0}^{\infty} \varepsilon_k$ and condition (2) to control the stepsize proposed in Algorithm 2.1. We then obtain further properties of NGD. In particular,

(i) $\{\lambda_k\}$ is convergent to λ^*; (Lemma 2.2)

(ii) there exists $\bar{k}$ such that for all $k \geq \bar{k}$ we have

$$\|\nabla f(x^k) - \nabla f(x^{k-1})\| \leq \frac{\eta_0}{\lambda_{k-1}} \|x^k - x^{k-1}\|, \quad \text{(in Lemma 2.3)}$$

and

$$\lambda_{k+1} > \lambda_k \quad \text{(in Remark 2.2).}$$

These above properties allow us to prove the convergence of Algorithm 2.1 in Theorems 2.1 and 2.2 via the simple and familiar technique (usually applied for line search procedures) of verifying the inequality

$$f(x^k) - f(x^{k+1}) \geq M\|\nabla f(x^k)\|^2$$

for some $M > 0$ and $k \geq \overline{k}$. This inequality shows the decreasing monotonic of $\{f(x^k)\}_{k\geq\overline{k}}$ (obtained by NGD). Remember that, in AdGD [17] one has not known about the descent property of $\{f(x^k)\}$.

Lemma 2.1. *Let $\{x^k\}$ and $\{\lambda_k\}$ be sequences generated by Algorithm 2.1. Then the two statements below hold:*

(i) $\{x^k\}$ is bounded,

(ii) $\{\lambda_k\}$ is lower bounded by a positive number.

Proof. Taking $x^* \in X^*$, it is easy to see that

$$\begin{aligned} \|x^{k+1} - x^*\|^2 &= \|x^{k+1} - x^k\|^2 + 2\langle x^{k+1} - x^k, x^k - x^*\rangle + \|x^k - x^*\|^2 \\ &= \|x^{k+1} - x^k\|^2 + 2\lambda_k\langle \nabla f(x^k), x^* - x^k\rangle + \|x^k - x^*\|^2. \end{aligned} \tag{7}$$

Since f is convex we have

$$2\lambda_k\langle \nabla f(x^k), x^* - x^k\rangle \leq 2\lambda_k(f_* - f(x^k)). \tag{8}$$

Let us rewrite

$$\|x^{k+1} - x^k\|^2 = A + B - \|x^{k+1} - x^k\|^2, \tag{9}$$

where

$$A = 2\lambda_k\langle \nabla f(x^k) - \nabla f(x^{k-1}), x^k - x^{k+1}\rangle, \quad B = 2\lambda_k\langle \nabla f(x^{k-1}), x^k - x^{k+1}\rangle.$$

From Cauchy Schwarz inequality we have

$$A \leq 2\lambda_k\|\nabla f(x^k) - \nabla f(x^{k-1})\|\|x^k - x^{k+1}\|. \tag{10}$$

Consider the two possible cases:

1. If condition (2) is satisfied then $\lambda_k = \eta_1 \frac{\|x^k - x^{k-1}\|}{\|\nabla f(x^k) - \nabla f(x^{k-1})\|}$. Inequality (10) follows that

$$A \leq 2\eta_1\|x^k - x^{k-1}\|\|x^k - x^{k+1}\|. \tag{11}$$

2. If condition (2) is not true, i.e.,

$$\begin{cases} \|\nabla f(x^k) - \nabla f(x^{k-1})\| \leq \frac{\eta_0}{\lambda_{k-1}}\|x^k - x^{k-1}\|, \\ \lambda_k = (1 + \varepsilon'_{k-1})\lambda_{k-1} \leq (1 + \varepsilon_{k-1})\lambda_{k-1}. \end{cases}$$

From (10), we get

$$A \leq \frac{2\eta_0\lambda_k}{\lambda_{k-1}}\|x^k - x^{k-1}\|\|x^k - x^{k+1}\| \leq 2\eta_0(1 + \varepsilon_{k-1})\|x^k - x^{k-1}\|\|x^k - x^{k+1}\|. \tag{12}$$

Since $\eta_0 < \frac{1}{2}$ and $\sum\limits_{k=0}^{\infty} \varepsilon_k$ is convergent then there exists k_1 such that

$$2\eta_0(1 + \varepsilon_{k-1}) < 1 \quad \forall k \geq k_1. \tag{13}$$

Combining (11), (12), (13) and the fact that $\eta_1 < \frac{1}{2}$, we obtain

$$A < \|x^k - x^{k-1}\| \ \|x^k - x^{k+1}\| \leq \frac{1}{2}\|x^k - x^{k-1}\|^2 + \frac{1}{2}\|x^{k+1} - x^k\|^2 \ \forall k \geq k_1. \tag{14}$$

Using the convexity of f we evaluate

$$B = \frac{2\lambda_k^2}{\lambda_{k-1}}\langle x^{k-1} - x^k, \nabla f(x^k)\rangle \leq \frac{2\lambda_k^2}{\lambda_{k-1}}(f(x^{k-1}) - f(x^k)). \tag{15}$$

From (9), (14), (15), we infer that

$$\|x^{k+1} - x^k\|^2 \leq \frac{1}{2}\|x^k - x^{k-1}\|^2 - \frac{1}{2}\|x^{k+1} - x^k\|^2 + \frac{2\lambda_k^2}{\lambda_{k-1}}(f(x^{k-1}) - f(x^k)) \quad \forall k \geq k_1. \tag{16}$$

Plug (16) in (7) we deduce that

$$\begin{aligned} \|x^{k+1} - x^*\|^2 \leq \frac{1}{2}\|x^k - x^{k-1}\|^2 - \frac{1}{2}\|x^{k+1} - x^k\|^2 + 2\lambda_k(f_* - f(x^k)) \\ + \frac{2\lambda_k^2}{\lambda_{k-1}}(f(x^{k-1}) - f(x^k)) + \|x^k - x^*\|^2 \quad \forall k \geq k_1 \end{aligned}$$

$$\begin{aligned} \iff \|x^{k+1} - x^*\|^2 + \frac{1}{2}\|x^{k+1} - x^k\|^2 + 2\lambda_k\left(1 + \frac{\lambda_k}{\lambda_{k-1}}\right)(f(x^k) - f_*) \\ \leq \|x^k - x^*\|^2 + \frac{1}{2}\|x^k - x^{k-1}\|^2 + 2\frac{\lambda_k^2}{\lambda_{k-1}}(f(x^{k-1}) - f_*) \quad \forall k \geq k_1. \end{aligned} \tag{17}$$

Next, remember that since $\sum\limits_{k=0}^{\infty} \varepsilon_k$ is convergent, then there exists k_2 such that

$$\varepsilon_k \leq \sqrt{2} - 1 \quad \forall k \geq k_2. \tag{18}$$

We now show that

$$\frac{\lambda_{k+1}^2}{\lambda_k^2} \leq 1 + \frac{\lambda_k}{\lambda_{k-1}}, \quad \forall k \geq k_2. \tag{19}$$

Indeed, for each $k \geq k_2$ there are two possible cases:

1. If $\|\nabla f(x^{k+1}) - \nabla f(x^k)\| > \frac{\eta_0}{\lambda_k}\|x^{k+1} - x^k\|$ then by (2), $\lambda_{k+1} = \eta_1\frac{\|x^{k+1}-x^k\|}{\|\nabla f(x^{k+1})-\nabla f(x^k)\|} < \frac{\eta_1}{\eta_0}\lambda_k$, that means $\frac{\lambda_{k+1}}{\lambda_k} < \frac{\eta_1}{\eta_0} < 1$ and hence (19) is proved.

2. Otherwise, $\lambda_{k+1} = (1 + \varepsilon'_k)\lambda_k$ then (19) is equivalent to

$$(1 + \varepsilon'_k)^2 \leq 1 + \frac{\lambda_k}{\lambda_{k-1}} \iff \varepsilon'_k \leq \sqrt{1 + \frac{\lambda_k}{\lambda_{k-1}}} - 1. \tag{20}$$

   Note that if $\frac{\lambda_k}{\lambda_{k-1}} < 1$ then by (5) $\varepsilon'_k = \min\{\varepsilon_k, \sqrt{1 + \frac{\lambda_k}{\lambda_{k-1}}} - 1\}$, this follows (20) obviously. Otherwise, if $\frac{\lambda_k}{\lambda_{k-1}} \geq 1$ then (20) is right due to the way we choose k_2 since $\sqrt{1 + \frac{\lambda_k}{\lambda_{k-1}}} - 1 \geq \sqrt{2} - 1 \geq \varepsilon_k \geq \varepsilon'_k$ for all $k \geq k_2$.

Take $k_3 = \max\{k_1, k_2\}$, now, summing up (17) from $k = k_3 + 1$ to $\ell, \ell \geq k_3 + 1$, we get

$$\begin{aligned} \|x^{\ell+1} - x^*\|^2 + \frac{1}{2}\|x^{\ell+1} - x^\ell\|^2 + 2\lambda_\ell\left(1 + \frac{\lambda_\ell}{\lambda_{\ell-1}}\right)(f(x^\ell) - f_*) + \\ + 2\sum_{k=k_3}^{\ell-1}\left(\lambda_k\left(1 + \frac{\lambda_k}{\lambda_{k-1}}\right) - \frac{\lambda_{k+1}^2}{\lambda_k}\right)(f(x^k) - f_*) \\ \leq \|x^{k_3+1} - x^*\|^2 + \frac{1}{2}\|x^{k_3+1} - x^{k_3}\|^2 + 2\frac{\lambda_{k_3+1}^2}{\lambda_{k_3}}(f(x^{k_3}) - f_*) = K \quad \forall \ell \geq k_3 + 1. \end{aligned} \tag{21}$$

From (19) we have

$$\sum_{k=k_3}^{\ell-1}\left(\lambda_k\left(1+\frac{\lambda_k}{\lambda_{k-1}}\right)-\frac{\lambda_{k+1}^2}{\lambda_k}\right)(f(x^k)-f_*)\geq 0 \quad \forall \ell \geq k_3+1.$$

Combining with (21) we obtain

$$\|x^{\ell+1}-x^*\|^2 \leq K \quad \forall \ell \geq k_3+1.$$

We now conclude that $\{x^k\}$ is bounded.
Since $\{x^k\}$ is bounded we have $S=\overline{conv}\{x^*,x^0,x^1,...\}$ is compact. Because f is locally Lipschitz gradient then there exists $L_1>0$ such that

$$\|\nabla f(x)-\nabla f(y)\| \leq L_1\|x-y\| \quad \forall x,y \in S.$$

For $k=1$ if (2) is satisfied then $\lambda_1 \geq \frac{\eta_1}{L_1}$, otherwise $\lambda_1=(1+\varepsilon_0')\lambda_0 \geq \lambda_0$. By induction, we get that

$$\lambda_k \geq \min\{\frac{\eta_1}{L_1},\lambda_0\}=\gamma>0 \quad \forall k \geq 0. \tag{22}$$

□

Remark 2.1. From the proof of Lemma 2.1 we observe that:

(i) We can take $k_3=1$ if inequalities (13) and (18) are satisfied from 1. In particular, $2\eta_0(1+\varepsilon_{k-1})<1$ and $\varepsilon_k \leq \sqrt{2}-1 \quad \forall k \geq 1$. We control that easily by choosing the suitable series $\sum\limits_{k=0}^{\infty}\varepsilon_k$.

(ii) We also easily obtain some convergence results of Algorithm 2.1 that are similar to AdGD by using the analogous arguments in [17]. For instance, the complexity $O(\frac{1}{k})$ of $f(\hat{x}^k)-f_*$. Indeed, for $k=k_3,...,\ell-1$ we set $r_k=\lambda_k\left(1+\frac{\lambda_k}{\lambda_{k-1}}\right)-\frac{\lambda_{k+1}^2}{\lambda_k}$ and $r_\ell=\lambda_\ell\left(1+\frac{\lambda_\ell}{\lambda_{\ell-1}}\right)$. By (22) we have $\sum\limits_{k=k_3}^{\ell}r_k=\sum\limits_{k=k_3}^{\ell}\lambda_k+\dfrac{\lambda_{k_3}^2}{\lambda_{k_3-1}} \geq (\ell-k_3)\gamma$. Moreover, from (21) we get that $\sum\limits_{k=k_3}^{\ell}r_k(f(x^k)-f_*) \leq \frac{K}{2}$. Next, putting $\hat{x}^\ell=\left(\sum\limits_{k=k_3}^{\ell}r_kx^k\right)/\left(\sum\limits_{k=k_3}^{\ell}r_k\right)$. By the convexity of f, it follows that

$$f\left(\hat{x}^\ell\right)-f_* \leq \frac{\sum\limits_{k=k_3}^{\ell}r_k(f(x^k)-f_*)}{\sum\limits_{k=k_3}^{\ell}r_k} \leq \frac{K}{2\sum\limits_{k=k_3}^{\ell}r_k} \leq \frac{K}{2\gamma(\ell-k_3)}=O\left(\frac{1}{\ell}\right), \quad \forall \ell \geq k_3+1. \tag{23}$$

The next result provides a nice property of the sequence $\{\lambda_k\}$.

Lemma 2.2. *The sequence of stepsize $\{\lambda_k\}$ generated by Algorithm 2.1 is convergent.*

Proof. Let $a_k=\ln\lambda_{k+1}-\ln\lambda_k \quad \forall k \geq 0$, we have $a_k=a_k^+-a_k^-$, where

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

Then $a_k^+ \geq 0$ and $a_k^- \geq 0 \quad \forall k \geq 0$.
From the definition of λ_k in Algorithm 2.1, we derive that

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

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

$$\ln\lambda_{k+1}-\ln\lambda_0=\sum_{i=0}^{k}a_i=\sum_{i=0}^{k}(a_i^+-a_i^-)=\sum_{i=0}^{k}a_i^+-\sum_{i=0}^{k}a_i^-, \tag{24}$$

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

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

But the result of Lemma 2.1 gives $\inf\limits_{k\geq 0}\lambda_k > 0$. This contradiction proves the convergence of $\sum\limits_{k=0}^{\infty} a_k^-$. Finally, from (24) we get the desired conclusion that $\lim\limits_{k\to+\infty}\lambda_k = \lambda^* < +\infty$. □

Lemma 2.3. *There exists a fixed number $\overline{k}$ such that*

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

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

$$\|\nabla f(x^{k_j}) - \nabla f(x^{k_j-1})\| > \frac{\eta_0}{\lambda_{k_j-1}}\|x^{k_j} - x^{k_j-1}\|.$$

For this case

$$\lambda_{k_j} = \eta_1\frac{\|x^{k_j} - x^{k_j-1}\|}{\|\nabla f(x^{k_j}) - \nabla f(x^{k_j-1})\|}.$$

Consequently,

$$\frac{\eta_1\|x^{k_j} - x^{k_j-1}\|}{\lambda_{k_j}} = \|\nabla f(x^{k_j}) - \nabla f(x^{k_j-1})\| > \frac{\eta_0}{\lambda_{k_j-1}}\|x^{k_j} - x^{k_j-1}\|,$$

i.e.,

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

On the other hand, from Lemma 2.2 we have

$$\lim_{k_j\to+\infty}\lambda_{k_j} = \lim_{k_j\to+\infty}\lambda_{k_j-1} = \lim_{k\to+\infty}\lambda_k = \lambda^*, \tag{25}$$

hence we deduce that

$$\frac{\lambda^*}{\lambda^*} \leq \frac{\eta_1}{\eta_0} < 1.$$

It is a contradiction and we finish the proof. □

Remark 2.2. From Lemma 2.3 we note that from $\overline{k}$ (obtained by Lemma 2.3):

$$\lambda_{\overline{k}} \leq \lambda_k < \lambda_{k+1} \leq \lambda^* = \lim_{k\to+\infty}\lambda_k, k \geq \overline{k}.$$

Lemma 2.4. *For any $x \in \mathbb{R}^n$ we have*

$$f(x) - f(x^{k+1}) \geq \frac{1-\eta_0}{\lambda_k}\|x^{k+1} - x^k\|^2 + \frac{1}{\lambda_k}\langle x^k - x^{k+1}, x - x^k\rangle,\ \forall k \geq \overline{k}. \tag{26}$$

Proof. By the convexity of f, we have

$$\begin{aligned} f(x) - f(x^{k+1}) =& f(x) - f(x^k) + f(x^k) - f(x^{k+1}) \\ \geq& \langle\nabla f(x^k), x - x^k\rangle + \langle\nabla f(x^{k+1}), x^k - x^{k+1}\rangle \\ =& \frac{1}{\lambda_k}\langle x^k - x^{k+1}, x - x^k\rangle + \langle\nabla f(x^{k+1}) - \nabla f(x^k), x^k - x^{k+1}\rangle\langle\nabla f(x^k), x^k - x^{k+1}\rangle. \end{aligned} \tag{27}$$

Since

$$\begin{aligned}\langle \nabla f(x^{k+1}) - \nabla f(x^k), x^k - x^{k+1}\rangle &= -\langle \nabla f(x^k) - \nabla f(x^{k+1}), x^k - x^{k+1}\rangle \\ &\geq -\|\nabla f(x^k) - \nabla f(x^{k+1})\|\|x^k - x^{k+1}\| \\ &\geq -\frac{\eta_0}{\lambda_k}\|x^{k+1} - x^k\|^2 \quad \forall k \geq \overline{k}\end{aligned} \tag{28}$$

and

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

Combining (27), (28) and (29) we get the desired conclusion (26). □

Remark 2.3. If we replace x by x^k in (26) then we obtain the decreasing of $\{f(x^k)\}_{k\geq\overline{k}}$ since

$$f(x^k) - f(x^{k+1}) \geq \frac{1-\eta_0}{\lambda_k}\|x^{k+1} - x^k\|^2 = (1-\eta_0)\lambda_k\|\nabla f(x^k)\|^2 \geq (1-\eta_0)\lambda_{\overline{k}}\|\nabla f(x^k)\|^2, \forall k \geq \overline{k}. \tag{30}$$

Theorem 2.1. *Suppose that problem (P) satisfies Assumptions 1 and 2. Then the sequence $\{x^k\}$ generated by Algorithm 2.1 (NGD) converges to an optimal solution of problem (P) and for any $x^* \in X^*$ we have*

$$f(x^k) - f_* = f(x^k) - f(x^*) \leq \frac{1-\eta_0}{\lambda_{\overline{k}}}\frac{\|x^* - x^{\overline{k}}\|^2}{k - \overline{k}} = O\left(\frac{1}{k}\right) \quad \forall k \geq \overline{k} + 1. \tag{31}$$

Proof. Since $f(x^*) - f(x^{i+1}) \leq 0$ then from (26), replacing x by x^* we get

$$\langle x^i - x^{i+1}, x^* - x^i\rangle \leq 0, \ \forall i \geq \overline{k} \tag{32}$$

and

$$\begin{aligned}f(x^*) - f(x^{i+1}) &\geq \frac{1-\eta_0}{\lambda_i}\left(\|x^{i+1} - x^i\|^2 + 2\langle x^i - x^{i+1}, x^* - x^i\rangle\right) + \underbrace{\frac{2\eta_0 - 1}{\lambda_i}}_{<0}\underbrace{\langle x^i - x^{i+1}, x^* - x^i\rangle}_{\leq 0(\text{ by } (32))} \\ &\geq \frac{1-\eta_0}{\lambda_i}\left(\|x^* - x^{i+1}\|^2 - \|x^* - x^i\|^2\right) \quad \forall i \geq \overline{k}.\end{aligned} \tag{33}$$

By (33) we see that $\|x^* - x^{i+1}\|^2 - \|x^* - x^i\|^2 \leq 0 \ \forall i \geq \overline{k}$ and using the fact that $\lambda_i \geq \lambda_{\overline{k}} \ \forall i \geq \overline{k}$, we have

$$f(x^*) - f(x^{i+1}) \geq \frac{1-\eta_0}{\lambda_{\overline{k}}}\left(\|x^* - x^{i+1}\|^2 - \|x^* - x^i\|^2\right) \quad \forall i \geq \overline{k}. \tag{34}$$

Telescoping inequality (34) from $i = \overline{k}$ to $k + \overline{k} - 1 \, (k \geq 1)$, we deduce

$$kf(x^*) - \sum_{i=\overline{k}}^{k+\overline{k}-1} f(x^{i+1}) \geq \frac{1-\eta_0}{\lambda_{\overline{k}}}\left(\|x^* - x^{k+\overline{k}}\|^2 - \|x^* - x^{\overline{k}}\|^2\right). \tag{35}$$

Moreover, the decreasing of $\{f(x^k)\}_{k\geq\overline{k}}$ obtained in Remark 2.3 follows

$$-kf(x^{k+\overline{k}}) \geq -\sum_{i=\overline{k}}^{k+\overline{k}-1} f(x^{i+1}). \tag{36}$$

Then combining (35) with (36), we arrive at

$$f(x^{k+\overline{k}}) - f(x^*) \leq \frac{1-\eta_0}{\lambda_{\overline{k}}}\frac{\|x^* - x^{\overline{k}}\|^2}{k} \quad \forall k \geq 1,$$

i.e.,

$$f(x^k) - f(x^*) \leq \frac{1-\eta_0}{\lambda_{\overline{k}}} \frac{\|x^* - x^{\overline{k}}\|^2}{k - \overline{k}} = O\left(\frac{1}{k}\right) \quad \forall k \geq \overline{k} + 1.$$

Hence, we get

$$\lim_{k \to +\infty} f(x^k) = f_*.$$

Now, remember that f is continuous and the sequence $\{x^k\}$ is bounded. Consequently, we confirm that every cluster point of $\{x^k\}$ belongs to X^*. Take $\overline{x}$ as an arbitrary cluster point of $\{x^k\}$ means that there exists a subsequence $\{x^{k_j}\}$ such that $x^{k_j} \to \overline{x}$. Moreover, because $\overline{x} \in X^*$, from (33) we obtain the decreasing monotonicity of $\{\|\overline{x} - x^k\|\}_{k \geq \overline{k}}$. It follows that the sequence $\{\|\overline{x} - x^k\|\}_{k \geq \overline{k}}$ is convergent because it is decreasing and bounded by zero. Finally, we get the desired conclusion since

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

□

Next, we will prove a stronger result for Algorithm 2.1 if f is locally strongly convex. The details are in the following theorem.

Theorem 2.2. *Suppose that problem (P) satisfies Assumptions 1 and 2. In addition, f is locally strongly convex. Then the sequence $\{x^k\}$ generated by Algorithm 2.1 satisfies*

(i)

$$\|x^{k+1} - x^*\|^2 \leq \left(1 - \frac{\sigma\lambda_{\overline{k}}}{2(1-\eta_0)}\right)\|x^k - x^*\|^2, \quad \forall k \geq \overline{k} \tag{37}$$

(ii)

$$f(x^{k+1}) - f(x^*) \leq \left(\frac{1-\eta_0}{\lambda_{\overline{k}}} - \frac{\sigma}{2}\right)\left(1 - \frac{\sigma\lambda_{\overline{k}}}{2(1-\eta_0)}\right)^{k+1-\overline{k}} \|x^{\overline{k}} - x^*\|^2, \quad \forall k \geq \overline{k}, \tag{38}$$

where $\sigma > 0$ is strong convexity constant of f on the compact set $S = \overline{conv}\{x^*, x^0, x^1, ...\}$.

Proof. (i) f is σ-strongly convex over S meaning that

$$\|\nabla f(x^k) - \nabla f(x^{k-1})\| \geq \sigma\|x^k - x^{k-1}\|. \tag{39}$$

From Lemma 2.3 it is easy to see that $\sigma \leq \frac{\eta_0}{\lambda_{k-1}}$, $\forall k \geq \overline{k}$, hence $\sigma\lambda_{\overline{k}} \leq \sigma\lambda_k \leq \eta_0 < \frac{1}{2}, \forall k \geq \overline{k}$.

Because f is $\sigma-$ strongly convex on S then

$$f(x) - f(x^k) \geq \langle \nabla f(x^k), x - x^k \rangle + \frac{\sigma}{2}\|x - x^k\|^2, \ \forall x \in S. \tag{40}$$

Now, using the second line in formula (27) of Lemma 2.4 we obtain that

$$f(x) - f(x^{k+1}) \geq \frac{1-\eta_0}{\lambda_k}\|x^{k+1} - x^k\|^2 + \frac{\sigma}{2}\|x - x^k\|^2 + \frac{1}{\lambda_k}\langle x^k - x^{k+1}, x - x^k \rangle \quad \forall x \in S, k \geq \overline{k}. \tag{41}$$

Let $x = x^*$ in (41), we derive that

$$\langle x^k - x^{k+1}, x^* - x^k \rangle \leq 0 \ \forall k \geq \overline{k}. \tag{42}$$

Additionally, for all $x \in S, k \geq \overline{k}$,

$$\begin{aligned} f(x^*) - f(x^{k+1}) &\geq \frac{1-\eta_0}{\lambda_k}\left(\|x^{k+1} - x^*\|^2 - 2\langle x^k - x^{k+1}, x^* - x^k\rangle - \|x^* - x^k\|^2\right) + \frac{\sigma}{2}\|x^* - x^k\|^2 + \\ &\quad + \frac{1}{\lambda_k}\langle x^k - x^{k+1}, x^* - x^k\rangle \\ &\geq \frac{1-\eta_0}{\lambda_k}\|x^* - x^{k+1}\|^2 + \left(\frac{\sigma}{2} - \frac{1-\eta_0}{\lambda_k}\right)\|x^* - x^k\|^2 + \underbrace{\frac{2\eta_0 - 1}{\lambda_k}\langle x^k - x^{k+1}, x^* - x^k\rangle}_{\geq 0\ \left(\text{by (42) and } \eta_0 < \frac{1}{2}\right).} \end{aligned} \tag{43}$$

Remember that $f(x^*) - f(x^{k+1}) \leq 0\ \forall k$ then we have

$$\frac{1-\eta_0}{\lambda_k}\|x^* - x^{k+1}\|^2 \leq \left(\frac{1-\eta_0}{\lambda_k} - \frac{\sigma}{2}\right)\|x^* - x^k\|^2, \quad k \geq \overline{k}. \tag{44}$$

establishing

$$\|x^{k+1} - x^*\|^2 \leq \left(1 - \frac{\sigma\lambda_k}{2(1-\eta_0)}\right)\|x^k - x^*\|^2 \leq \left(1 - \frac{\sigma\lambda_{\overline{k}}}{2(1-\eta_0)}\right)\|x^k - x^*\|^2, \quad k \geq \overline{k}.$$

(ii) From (43), we also get

$$\begin{aligned} f(x^{k+1}) - f(x^*) &\leq -\frac{1-\eta_0}{\lambda_k}\|x^* - x^{k+1}\|^2 - \left(\frac{\sigma}{2} - \frac{1-\eta_0}{\lambda_k}\right)\|x^* - x^k\|^2 \\ &\leq \left(\frac{1-\eta_0}{\lambda_k} - \frac{\sigma}{2}\right)\|x^* - x^k\|^2 \leq \left(\frac{1-\eta_0}{\lambda_{\overline{k}}} - \frac{\sigma}{2}\right)\|x^* - x^k\|^2 \\ &\leq \left(\frac{1-\eta_0}{\lambda_{\overline{k}}} - \frac{\sigma}{2}\right)\left(1 - \frac{\sigma\lambda_{\overline{k}}}{2(1-\eta_0)}\right)^{k+1-\overline{k}}\|x^{\overline{k}} - x^*\|^2, \quad k \geq \overline{k}. \end{aligned}$$

□

Remark 2.4. Under the assumptions of Theorem 2.2, we end this section by several remarkable points as follows:

(i) The assertion (i) of Theorem 2.2 shows the linear convergence rate of $\{x^k\}_{k \geq \overline{k}}$.

(ii) By Theorem 2.2 (ii) and taking into account that $\log(1-x) < -x$ for $x \in (0,1)$, we will get $f(x^k) - f_* \leq \varepsilon$ if

$$k \geq \overline{k} + \frac{2(1-\eta_0)}{\sigma\lambda_{\overline{k}}}\log\left(\frac{1}{\varepsilon}\right) + \frac{2(1-\eta_0)}{\sigma\lambda_{\overline{k}}}\log\left(\frac{2 - 2\eta_0 - \sigma\lambda_{\overline{k}}}{2\lambda_{\overline{k}}}\|x^{\overline{k}} - x^*\|^2\right). \tag{45}$$

3. A new projected gradient descent algorithm for solving a class of nonconvex optimization over a closed convex set

In this section, we propose a new algorithm for solving a class of nonconvex optimization over a closed convex set that is

$$\min_{x \in C} f(x), \tag{P_1}$$

where $C \subset \mathbb{R}^n$ is a nonempty closed convex set and f satisfies the followings

(C_1) f is smooth and globally gradient Lipschitz with constant L on C,

(C_2) For $u, v \in C$, the function $g_{uv} : \mathbb{R} \to \mathbb{R}$ defined by

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

is quasiconvex on $[0,1]$.

We remember that the necessary optimality condition for a local optimal solution z of problem (P_1) is the stationarity condition that

$$\langle \nabla f(z), x - z \rangle \geq 0, \text{ for all } x \in C.$$

This is also sufficient if f is convex. The stationarity can be also verified via the simple equality

$$z = P_C(z - s\nabla f(z)), \quad \text{for some } s > 0. \tag{46}$$

The above condition (46) is ensured for the stationarity of z for any $s > 0$. One can see [4] for more details. Motivated by (46) it is known that a traditional method for finding a stationary point of the problem (P_1) is the projected gradient (PG) algorithm that calculates the iterative sequence

$$x^{k+1} = P_C(x^k - \lambda_k \nabla f(x^k)), \tag{47}$$

where P_C is the orthogonal projection onto C defined by

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

Obviously, the projected gradient algorithm becomes GD as $C = \mathbb{R}^n$. Similar to GD, one can control the performance of the PG algorithm through stepsize λ_k. For the general case, the function f is global L-Lipschitz gradient, the gradient projection method is convergent if we choose constant stepsize $\lambda_k \in (0, \frac{2}{L})$ or by backtracking procedure, see [4] for more details. However, as presented, it will be saved the computational effort (to estimate L or implement the backtracking procedure) if we apply some adaptive stepsize for PG like the one used in Algorithm 2.1 or AdGD. This suggests us to build a new algorithm based on the projected gradient with our new stepsize for solving the problem (P_1).

Before presenting the new algorithm in detail, let us discuss the assumptions (C_1) and (C_2) of f. It should be noted that if f is convex then the condition (C_2) is always satisfied but the converse is not true. For instance, the 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 $b \in \mathbb{R}^n$) has

$$g_{uv}(t) = \langle A(u + t(v - u)) + b, v - u \rangle$$

which is linear (with respect to variable t) and hence quasiconvex on $\mathbb{R}$ for all $u, v \in \mathbb{R}^n$ although f may be nonconvex for non-semipostive definite matrix A. Therefore, functions satisfying (C_2) are nonconvex in general. Now, without the convexity of f we cannot obtain the boundedness of $\{x^k\}$ or the lower boundedness of $\{\lambda_k\}$ by using similar arguments like in Lemma 2.1. However, the condition (C_1) is helpful to prove the existence of limitation of $\{\lambda_k\}$. Below is the projected gradient algorithm based on our new stepsize.

Algorithm 3.1 (PG-NGD)

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

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

$$\lambda_k = \eta_1 \frac{\|x^k - x^{k-1}\|}{\|\nabla f(x^k) - \nabla f(x^{k-1})\|} \tag{48}$$

else

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

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

To prove the convergence of Algorithm 3.1 we have to confirm some similar results as the previous section.

Lemma 3.1. *For Algorithm 3.1, we have* $\inf\limits_{k\geq 0} \lambda_k > 0$ *and* $\{\lambda_k\}$ *is convergent.*

Proof. By induction, it is easy to show that $\lambda_k \geq \min\{\lambda_0, \frac{\eta_1}{L}\} > 0$ for all $k \geq 0$. Hence $\inf\limits_{k\geq 0} \lambda_k > 0$. The remaining assertion is proved by analogous arguments as Lemma 2.2. □

Lemma 3.2. *For Algorithm 3.1, there exists $\hat{k}$ such that*

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

Proof. The same as the proof of Lemma 2.3. □

Lemma 3.3. *Let f be a function that satisfies the conditions (C_1) and (C_2). Then the sequence $\{x^k\}$ generated by Algorithm 3.1 has the following property*

$$f(x^k) - f(x^{k+1}) \geq \frac{1-\eta_0}{\lambda_k}\|x^{k+1} - x^k\|^2, \ \forall k \geq \hat{k}.$$

Proof. By using the Fundamental Theorem of Calculus, for any $k \geq \hat{k}$ we have

$$\begin{aligned} f(x^{k+1}) - f(x^k) &= \int_0^1 \langle \nabla f(x^k + t(x^{k+1} - x^k)), x^{k+1} - x^k\rangle dt \\ &= \langle \nabla f(x^k), x^{k+1} - x^k\rangle + \int_0^1 \langle \nabla f(x^k + t(x^{k+1} - x^k)) - \nabla f(x^k), x^{k+1} - x^k\rangle dt \\ &= \langle \nabla f(x^k), x^{k+1} - x^k\rangle + \int_0^1 h_k(t)dt. \end{aligned} \tag{49}$$

On the other hand,

$$h_k(t) = \langle \nabla f(x^k + t(x^{k+1} - x^k)) - \nabla f(x^k), x^{k+1} - x^k\rangle = g_{x^k x^{k+1}}(t) - \langle \nabla f(x^k), x^{k+1} - x^k\rangle$$

is quasiconvex in $[0, 1]$ therefore, for all $t \in [0, 1]$,

$$h_k(t) \leq \max\{h_k(0), h_k(1)\} = \max\{0, h_k(1)\} \leq |h_k(1)| = |\langle \nabla f(x^{k+1}) - \nabla f(x^k), x^{k+1} - x^k\rangle|. \tag{50}$$

Next, combining with Lemma 3.2, we infer that

$$\int_0^1 h_k(t)dt \leq \frac{\eta_0}{\lambda_k}\|x^{k+1} - x^k\|^2, \ \ \forall k \geq \hat{k}. \tag{51}$$

Moreover, $x^{k+1} = P_C(x^k - \lambda_k \nabla f(x^k))$ then by Theorem 9.8 in [4] we have

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

from which we obtain

$$\langle \nabla f(x^k), x^{k+1} - x^k\rangle \leq -\frac{1}{\lambda_k}\|x^{k+1} - x^k\|^2. \tag{52}$$

It follows from (49), (51) and (52) that

$$f(x^k) - f(x^{k+1}) \geq \frac{1-\eta_0}{\lambda_k}\|x^{k+1} - x^k\|^2 \ \ \forall k \geq \hat{k}. \tag{53}$$

□

The following theorem gives the convergence of Algorithm 3.1 for solving the problem (P_1).

Theorem 3.1. *Suppose that f satisfies the conditions (C_1) and (C_2). Then we have the following assertions for Algorithm 3.1:*

(i) The sequence $\{f(x^k)\}_{k\geq\hat{k}}$ is non-decreasing and $f(x^{k+1}) < f(x^k)$ unless x^k is a stationary point of problem (P_1) for any $k \geq \hat{k}$.

(ii) If f is lower bounded on C then $f(x^k) - f(x^{k+1}) \to 0$ and $\|x^k - x^{k+1}\| \to 0$.

Proof. (i) Since $\eta_0 < 1$ then from (53) we obtain $f(x^k) \geq f(x^{k+1})$ for all $k \geq \hat{k}$ and the equality holds for only case $x^{k+1} = x^k = P_C(x^k - \lambda_k \nabla f(x^k))$ or x^k is a stationary point of (P_1).

(ii) From (i) the sequence $\{f(x^k)\}_{k \geq \hat{k}}$ is nondecreasing and lower bounded on C then it has a finite limitation. Therefore $f(x^k) - f(x^{k+1}) \to 0$ and $\|x^k - x^{k+1}\| \to 0$ as a consequence.

□

4. Numerical experiments

Firstly, we test the performance of our new stepsize for the GD scheme by comparing Algorithm 2.1 (NGD) with GD and AdGD for some benchmark problems provided by [17]. Note that we use the original python code for GD and AdGD from [17][4]. We report the details in Section 4.1.

Secondly, we do a preliminary computational test for Algorithm 3.1 by comparing it with the two related algorithms including projected gradient with constant stepsize $1/L$ (PG-GD) and projected gradient with stepsize of AdGD (PG-AdGD). The tested data for this part is synthetic and will be described in detail in Section 4.2.

To implement Algorithm 2.1 and Algorithm 3.1, it is necessary to choose the suitable parameters $\lambda_0, \eta_0, \eta_1, \varepsilon_k$ for each kind of tested instance. In particular, we take the convergent series defined by

$$\varepsilon_{k-1} = \frac{\alpha(\ln k)^\beta}{k^{1.1}}, \ \alpha > 0, \beta \geq 0, k \geq 1. \tag{54}$$

All the mentioned algorithms were coded in Python.

4.1. Aglorithm 2.1 (NGD)

We reuse some benchmarks in [17] including logistic regression, matrix factorization, and cubic regularization for testing. The interested readers can see [17] to find more details about the description of these problems. We use the same notations as [17] for the reported results.

For the **logistic regression**, the loss function is defined by $\frac{1}{n}\sum\limits_{i=1}^{n} \log(1 + \exp(-b_i a_i^T x)) + \frac{\gamma}{2}\|x\|^2$, where $(a_i, b_i) \in \mathbb{R}^d \times \mathbb{R}, i = 1, ..., n$ are observations and $\gamma > 0$ is a l_2 regularization parameter. In this case, the objective is strongly convex. The tested datasets include "covtype", "w8a" and "mushrooms" from libsvm[5] library. The results with $(\lambda_0, \eta_0, \eta_1, \alpha, \beta) = (1e-06, 0.2, 0.15, 0.9, 5)$ are presented in Fig. 1.

The second tested problem is **matrix factorization** that is common in recommendation systems [25]. The data is a matrix $A \in \mathbb{R}^{m \times n}$ and $r < \min\{m, n\}$; we need to find $U \in \mathbb{R}^{m \times r}$ and $V \in \mathbb{R}^{r \times n}$ minimizing the nonconvex objective $f(U, V) = \|A - UV\|_F^2$. Similar to [17], we use MoviLens 100K dataset [12] and do the experiments with several values of $r = 10, 20, 30$. The results with $(\lambda_0, \eta_0, \eta_1, \alpha, \beta) = (1e-05, 0.49, 0.48, 75, 0)$ are shown in Fig. 2.

The last benchmark is solving the subproblem obtained by **cubic regularization** of the Newton method. A modified Newton step $x^{k+1}$is defined by $T_M(x^k)$ where $T_M(x^k)$ is a global optimal solution of $\min\limits_{x \in \mathbb{R}^d} F(x) = \langle \nabla f(x^k), x - x^k \rangle + \frac{1}{2}\langle \nabla^2 f(x^k)(x - x^k), x - x^k \rangle + \frac{M}{6}\|x - x^k\|^3$. See [20, 21] for more details. Analogous to [17] we set $x^k = 0$; f is logistic loss of the "covtype" dataset and $M = 10, 20, 100$. The results with $(\lambda_0, \eta_0, \eta_1, \alpha, \beta) = (1e-04, 0.499, 0.49, 2, 4)$ are in Fig. 3.

From Fig. 1, 2 and 3 we see that our NGD provides better performance than the others. Especially for the nonconvex instances, the deviation between our method and the remaining ones are really significant. NGD makes the objective value decreasing rapidly after fewer iterations than GD and AdGD. It is not surprising that GD is the most expensive method among all.

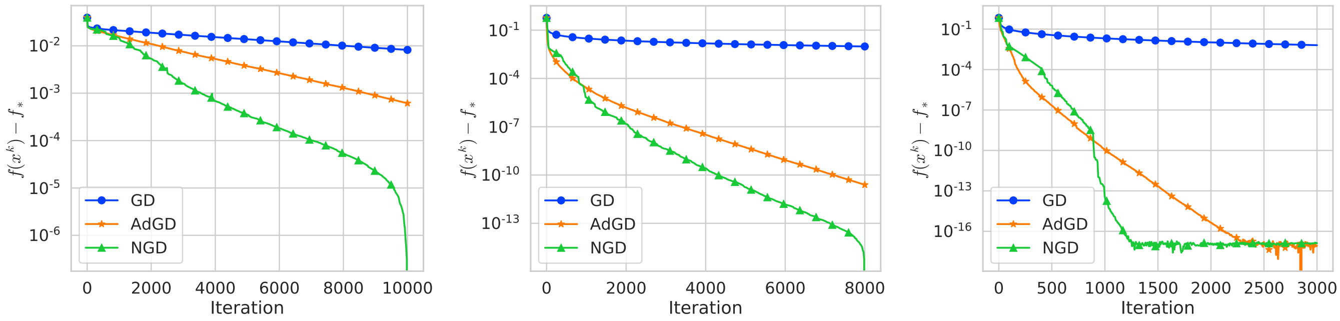
4.2. Algorithm 3.1 (PG-NGD)

In this section, we implement Algorithm 3.1 (PG-NGD) for solving the following problem

$$\min\Big\{f(x) = \frac{1}{2}x^T Ax + b^T x : x \in C\Big\}, \tag{55}$$

[4] https://github.com/ymalitsky/adaptive_gd

[5] https://www.csie.ntu.edu.tw/ cjlin/libsvmtools/datasets/

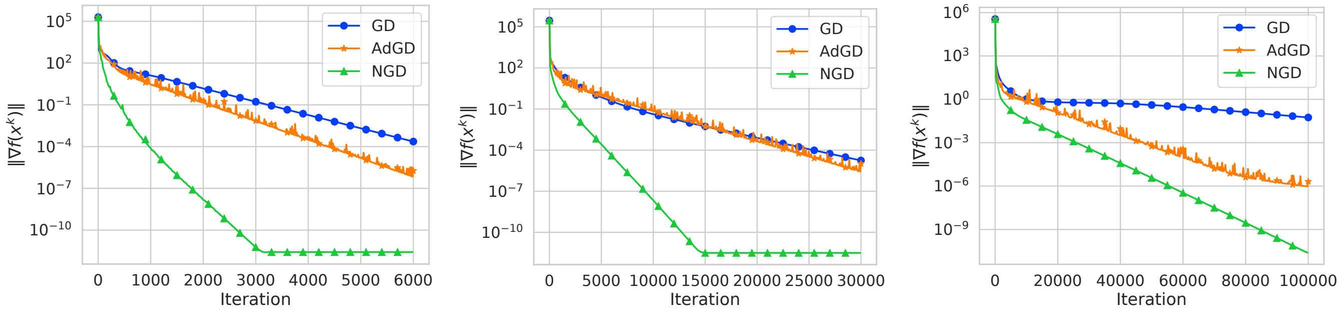


(a) Covtype

(b) W8a

(c) Mushrooms

Fig. 1: Results for logistic regression.



(a) $r = 10$

(b) $r = 20$

(c) $r = 30$

Fig. 2: Results for matrix factorization. The objective is nonconvex.

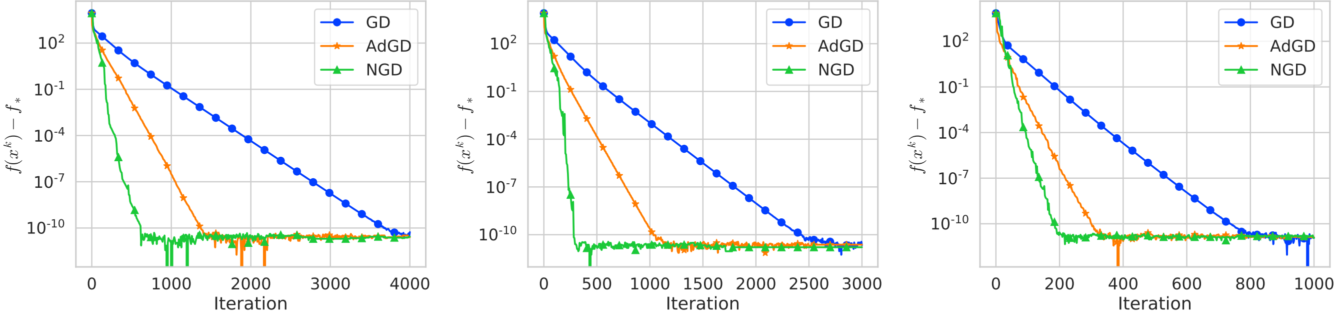
where $A \in R^{n \times n}$ is a real symmetric matrix which is created by $M + M^T$ with M is a matrix that has all entries generated randomly in $[-1, 1]$ by uniform distribution; $b \in R^n$ is a vector that is generated randomly in $[-1, 1]$ by uniform distribution; C can be a box $[l, u] = [-1, 1] \subset \mathbb{R}^n$ or a simplex $\triangle_n = \{x \in \mathbb{R}^n_+ \mid \sum_{i=1}^n x_n = 10\}$; the starting point x^0 is generated radomly in $[0, 1]$ by uniform distribution and the stopping criteria is $\|x^{k+1} - x^k\| \leq 1e - 08$. We set $n = 1000, 5000, 10000$ and report the results with $(\lambda_0, \eta_0, \eta_1, \alpha, \beta) = (1e - 04, 0.5, 0.45, 100, 3)$ in Fig. 4 and 5 for box and simplex constraint sets, respectively. It is shown that PG-NGD achieves the required tolerance very quickly for all the cases.

5. Conclusion

In this paper, we propose a new adaptive stepsize for the GD scheme. Under the locally Lipschitzness of the gradient of the convex objective we obtain the complexity $O(\frac{1}{k})$ of $f(x^k) - f_*$. In addition, we show that the locally strong convexity of f follows the linear convergence of the corresponding GD. Specifically, our stepsize can be applied for projected gradient scheme to solve a class of nonconvex optimization problems over a closed convex set. The sequence of our new stepsize is proved increasing from some fixed iteration. Future research include accelerated and stochastic versions of GD with our new stepsize, as well as convergence of general nonconvex problem without global Lipschitz assumption on the gradient.

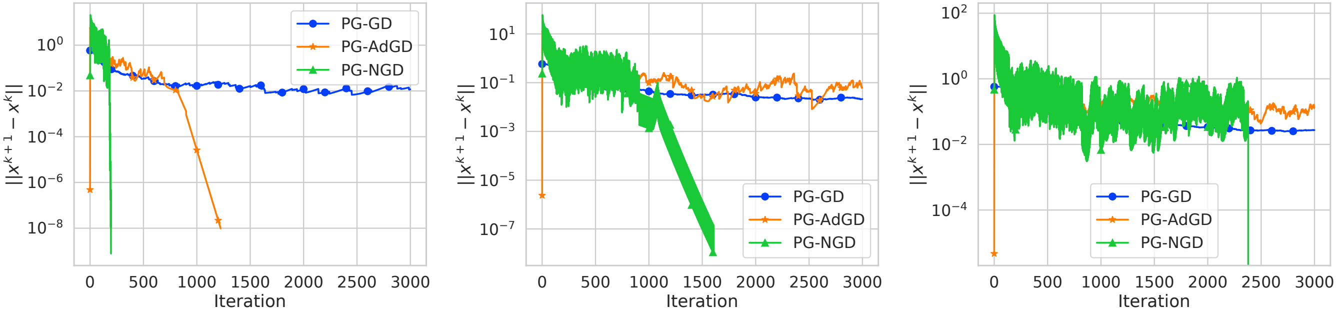
Acknowledgement

The authors wish to thank the anonymous referees for their useful comments which helped them to improve the paper greatly. This research is funded by Hanoi University of Science and Technology (HUST) under project number T2023-PC-081.



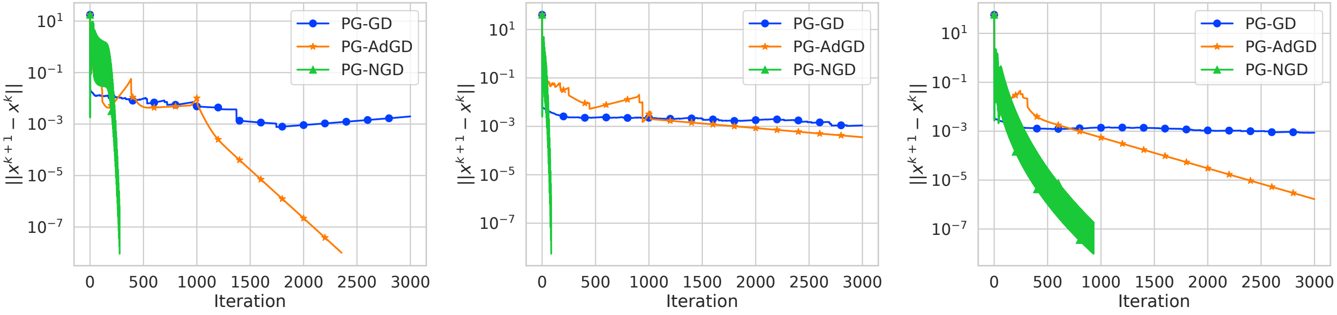
(a) $M = 10$ **(b)** $M = 20$ **(c)** $M = 100$

Fig. 3: Results for solving a subproblem from cubic regularization.



(a) $n = 1000$ **(b)** $n = 5000$ **(c)** $n = 10000$

Fig. 4: Results for nonconvex quadratic programming over the box $[-1, 1]$.



(a) $n = 1000$ **(b)** $n = 5000$ **(c)** $n = 10000$

Fig. 5: Results for nonconvex quadratic programming over Δ_n.

References

[1] L. Armijo, Minimization of functions having lipschitz continuous frst partial derivatives, Pac. J. Math. 16 (1966) 1–3.

[2] J. Barzilai, J.M. Borwein, Two-point step size gradient methods, IMA J. Numer. Anal. 8 (1988) 141–148.

[3] H. H. Bauschke, J. Bolte, M. Teboulle, A descent lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications, Math. Oper. Res. 42 (2016) 330–348.

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

[5] D.P. Bertsekas, Nonlinear programming, 3rd Edition, Athena Scientific, 2016.

[6] O. Burdakov, Y.H. Dai, N. Huang, Stabilized barzilai-borwein method, J. Comput. Math. 37 (2019) 916-936.

[7] J.Y.B. Cruz, T.T.A. Nghia, On the convergence of the forward-backward splitting method with linesearches, Optim. Methods Softw. 31 (2016) 1209-1238.

[8] J. Duchi, E. Hazan, Y. Singer, Adaptive subgradient methods for online learning and stochastic optimization, J. Mach. Learn. Res. 12 (2011) 2121-2159.

[9] A.A. Goldstein, Cauchy's method of minimization, Numer. Math. 4 (1962) 146-150.

[10] O. Güler, Foundations of optimization, Springer, 2010.

[11] N. Hallak, M. Teboulle, A non-euclidean gradient descent method with sketching for unconstrained matrix minimization, Oper. Res. Lett. 47 (2019) 421-426.

[12] F.M. Harper, J.A. Konstan, The movielens datasets: History and context, ACM Trans. Interact. Intell. Syst. 5 (2016) 1-19.

[13] D.P. Kingma, L.J. Ba, Adam: A method for stochastic optimization, ICLR (Poster) 2015.

[14] C. Lemaréchal, Cauchy and the gradient method, Doc. Math. Extra Vol., Optimization Stories, 251-254 (2012).

[15] H. Liu, T. Wang, Z. Liu, Some modifed fast iterative shrinkage thresholding algorithms with a new adaptive non-monotone stepsize strategy for nonsmooth and convex minimization problems, Comput. Optim. Appl. 83 (2022) 651-691.

[16] C.J. Maddison, D. Paulin, Y.W. Teh, A. Doucet, Dual space preconditioning for gradient descent, SIAM J. Optim. 31 (2021) 991-1016.

[17] Y. Malitsky, K. Mishchenko, Adaptive gradient descent without descent, ICML 119 (2020) 6702-6712.

[18] H.B. McMahan, M. Streeter, Adaptive bound optimization for online convex optimization, Proceedings of the 23rd Annual Conference on Learning Theory (COLT) 2010.

[19] A.S. Nemirovsky, D.B. Yudin, Problem complexity and method efficiency in optimization, John Wiley Sons, Inc., New York, 1983.

[20] Y. Nesterov, Lectures on convex optimization, 2nd Edition, Springer, 2018.

[21] Y. Nesterov, B.T. Polyak, Cubic regularization of newton method and its global performance, Math. Program. 108 (2006) 177-205.

[22] B.T. Polyak, Minimization of nonsmooth functionals, Zh. Vychisl. Mat. Mat. Fiz., 9(3) (1969) 509-521.

[23] M. Prazeres, A.M. Oberman, Stochastic gradient descent with polyak's learning rate, J. Sci. Comput. 89 (2021) 1-16.

[24] S. Salzo, The variable metric forward-backward splitting algorithm under mild differentiability assumptions, SIAM J. Optim. 27 (2017) 2153-2181.

[25] P. Symeonidis and A. Zioupos, Matrix and Tensor Factorization Techniques for Recommender Systems, Springer Briefs in Computer Science, 2016.

[26] A. Taylor, F. Bach, Stochastic first-order methods: non-asymptotic and computer-aided analyses via potential functions, Proceedings of Machine Learning Research 99 (2019), 1-59.

[27] F. Wang, Polyak's gradient method for split feasibility problem constrained by level sets, Numer. Algor. 77 (2018) 925-938.