Int. J. Nonlinear Anal. Appl. 15 (2024) 3, 11–20
ISSN: 2008-6822 (electronic)
http://dx.doi.org/10.22075/ijnaa.2019.16973.1899

A non-monotone Hestenes-Stiefel conjugate gradient algorithm for nonsmooth convex optimization

Ahmad Abouyee Mehrizi, Reza Ghanbari*

Faculty of Mathematical Sciences, Department of Applied Mathematics, Ferdowsi University of Mashhad, Mashhad, Iran

(Communicated by Saman Babaie-Kafaki)

Abstract

Here, we propose a practical method for solving nonsmooth convex problems by using conjugate gradient-type methods. The conjugate gradient method is one of the most remarkable methods to solve smooth and large-scale optimization problems. As a result of this fact, We present a modified HS conjugate gradient method. In the case that we have a nonsmooth convex problem, by the Moreau-Yosida regularization, we convert the nonsmooth objective function to a smooth function and then we use our method, by making use of a nonmonotone line search, for solving a nonsmooth convex optimization problem. We prove that our algorithm converges to an optimal solution under standard condition. Our algorithm inherits the performance of HS conjugate gradient method.

Keywords: Nonsmooth convex optimization, Conjugate gradient method, nonmonotone line search, Global convergence
2020 MSC: Primary 26A25; Secondary 39B62

1 Introduction

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a nondifferentiable convex function, consider the following unconstrained optimization problem

$$\min_{x \in \mathbb{R}^n} f(x). \tag{1.1}$$

Problem (1.1) appears in many applications such as economics [41], engineering [34, 35] including power unit commitment problem [8] and continuous casting of steel [31], Image Restoration [25], data analysis [9, 48] including cluster analysis [21, 23] and data classification problems [3, 12], machine learning [24], and optimal control [30, 33].

There are different methods that have been developed to solve problem (1.1). One of the simplest techniques for solving problem (1.1) is the subgradient methods [37, 40, 6]. Subgradient method is exactly the same with steepest descent when objective function f is smooth. Another popular and practical method to solve problem (1.1) is Bundle method and its various modifications use quadratic programming subproblem to find search directions [27, 44, 45]. Recently, Karmitsa [26] proposed a diagonal Bundle method for nonsmooth sparse optimization problems. Another class of methods for solving nonsmooth problems is Derivative free methods. Bagirov et al. [7] developed a new derivative free method by introducing the notion of a discrete gradient.

*Corresponding author
Email addresses: ahmadaboiy@gmail.com (Ahmad Abouyee Mehrizi), rghanbari@um.ac.ir (Reza Ghanbari)

Received: January 2019 *Accepted:* July 2019

Smoothing technique [13, 18, 38, 39], is another practical methods to solve problem (1.1) that makes use of methods for solving smooth problems such as quasi Newton methods, used for small and medium scale [16, 32, 49, 51], and conjugate gradient (CG) methods, which are really suitable for solving large scale problems [1, 5, 17, 20, 36, 46, 53, 52], or trust region methods [47, 19]. In this regard, Burke et al. [10] made a gradient sampling with the assumption that there is an open dense subset D so that objective function is continuously differentiable on D. Another popular smoothing technique is adopting Moreau-Yosida regularization to convert problem (1.1) into a smooth problem (see section 2). This technique has some remarkable results in [14, 28, 29, 50]. In this paper, by using Moreau-Yosida regularization, we represent a new CG method to solve problem (1.1). The algorithm is based on HS method to inherit a significant property of this method, namely, preventing a sequence of tiny step from happening. Moreover, the generated search direction in each iterate satisfies the sufficiently descent property. Also, the search direction belongs to a trust region, hence we can prove the global convergence of the algorithm.

The rest of the paper is organized as follows. In section 2, we refer to the concept of Moreau-Yosida regularization and its relation with problem (1.1). In Section 3, we briefly review some different classes of CG method, and a technique to find the step length inexactly. Then, we introduce our method. We prove the global convergence of the proposed method in Section 4. Finally, Section 5 contains some conclusions about our idea.

2 Requirements of smoothing a nonsmooth convex optimization problem

In this section, at first, some results in convex analysis developed by Moreau-Yosida regularization are required. Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function, maybe nondifferentiable, then $F : \mathbb{R}^n \to \mathbb{R}$ is the so-called Moreau-Yosida regularization of f defined by

$$F(x) = \min_{z\in\mathbb{R}^n}\{f(z) + \frac{1}{2\lambda}||z-x||^2\}, \tag{2.1}$$

where λ is a positive parameter and $||\cdot||$ is l_2 norm. Let $p(x) = \operatorname{argmin}_{z\in\mathbb{R}^n}\theta(z)$ where

$$\theta(z) = f(z) + \frac{1}{2\lambda}||z-x||^2. \tag{2.2}$$

Since $\theta(z)$ is strongly convex, so $p(x)$ is well defined and unique. The following Proposition proves that problem (1.1) is equivalent to following problem

$$\min_{x\in\mathbb{R}^n} F(x). \tag{2.3}$$

Proposition 2.1. [22] the following statement are equivalent

(*i*) x^* minimizes f;

(*ii*) $f(x^*) = F(x^*)$

(*iii*) $g(x^*) = 0$, where

$$g(x^*) = \nabla F(x^*) = \frac{x^* - p(x^*)}{\lambda}, \tag{2.4}$$

is the gradient of F.

Remark 2.2. There are some significant properties of the Moreau-Yosida regularization function F as follows

(*i*) F is finite valued.

(*ii*) The gradient mapping $g : \mathbb{R}^n \to \mathbb{R}$ is globally Lipschitz continuous with the constant Lipschitz $\frac{1}{\lambda}$, namely,

$$||g(x) - g(y)|| \leq \frac{1}{\lambda}||x-y||, \qquad x, y \in \mathbb{R}^n. \tag{2.5}$$

(*iii*) x solves the problem (1.1) iff $\nabla F(x) = 0$, that is, $p(x) = x$.

By the definition of $p(x)$, $F(x)$ can be rewritten as follows

$$F(x) = f(p(x)) + \frac{1}{2\lambda}||p(x)-x||^2. \tag{2.6}$$

From (2.4) and (2.6) to compute $F(x)$ and $g(x)$, we need the value of $p(x)$. Nevertheless, computing $p(x)$, as the minimizer of θ, is really tough or sometimes impossible to obtain. So, in practical way, we can approximate the value of $p(x)$, $F(x)$ and $g(x)$ for every $x \in \mathbb{R}^n$. In a sense, for each x and any $\varepsilon > 0$ there is a vector $p^\alpha(x,\varepsilon) \in \mathbb{R}^n$, where $p^\alpha(x,\varepsilon) \in \mathbb{R}^n$ means the approximation of $p(x)$, so that (by Completeness Principle)

$$f(p^\alpha(x,\varepsilon)) + \frac{1}{2\lambda}||p^\alpha(x,\varepsilon) - x||^2 \leq F(x) + \varepsilon. \tag{2.7}$$

Thus, when ε is small, approximating $F(x)$ and $g(x)$ can be written as follows

$$F^\alpha(x,\varepsilon) = f(p^\alpha(x,\varepsilon)) + \frac{1}{2\lambda}||p^\alpha(x,\varepsilon) - x||^2, \tag{2.8}$$

and

$$g^\alpha(x,\varepsilon) = \frac{x - p^\alpha(x,\varepsilon)}{\lambda}, \tag{2.9}$$

respectively. There are some practical algorithms to approximate $p(x)$ for a nonsmooth convex function [4, 15]. Fukushima and Qi [18] proved that using $p^\alpha(x,\varepsilon)$ for approximating the value of $F(x)$ and $g(x)$ when ε is small enough is possible. The following Proposition says this fact.

Proposition 2.3. [18] Suppose $p^\alpha(x,\varepsilon)$ is a vector satisfying in (2.8), and $F^\alpha(x,\varepsilon)$ and $g^\alpha(x,\varepsilon)$ are defined by (2.8) and (2.9), respectively. Then we have

$$F(x) \leq F^\alpha(x,\varepsilon) \leq F(x) + \varepsilon, \tag{2.10}$$

$$||p^\alpha(x,\varepsilon) - p(x)|| \leq \sqrt{2\lambda\varepsilon}, \tag{2.11}$$

$$||g^\alpha(x,\varepsilon) - g(x)|| \leq \sqrt{\frac{2\varepsilon}{\lambda}}. \tag{2.12}$$

Based on Moreau-Yosida regularization, many authors have represented various algorithms to solve the problem (1.1) so far. In this regard, some methods such as CG [14, 28, 50], quasi Newton [11, 29, 42] and trust region [29, 43] were used to solve problem (1.1).

As mentioned before, CG methods are another effective methods for solving smooth unconstrained optimization problem especially for large scale problems due to simple computations and low storage. Recently, Li [28] represented modified PRP method with Moreau-Yosida regularization. He used the approximation of F and g to solve the problem (1.1). Zhang et al. [50] proposed a tree term modified PRP method by approximating F and g. Cheng [14] represented a two term modified PRP. In the next section, we review former proposed CG methods briefly. Then, we introduce our CG method to solve the problem (1.1).

3 Nonmonotone conjugate gradient methods to solve problem (1.1)

In the case that f is smooth, there have been many attempts to solve a smooth unconstrained optimization problem. CG is one the most popular methods for solving the problem (1.1) especially when n is large, and has the following form

$$x_{k+1} = x_k + \alpha_k d_k, \tag{3.1}$$

$$d_k = \begin{cases} -g_k, & if \quad k = 0, \\ -g_k + \beta_k d_{k-1} & \forall \quad k \geq 1, \end{cases} \tag{3.2}$$

where α_k is a step length, β_k is a scalar. Well-known formulas for β_k are the Hestense-stiefel (HS), Fletcher-Reeves (FR), Polak-Ribiére (PR), Polak-Ribiére-Polyak and Dai-Liao (DL), and Dai-Yuan (DY) which are, respectively, given by (see [17, 20, 53])

$$\beta_k^{HS} = \frac{g_k^\intercal y_{k-1}}{d_{k-1}^\intercal y_{k-1}}, \qquad \beta_k^{FR} = \frac{||g_k||^2}{||g_{k-1}||^2}, \qquad \beta_k^{PRP} = \frac{g_k^\intercal y_{k-1}}{||g_{k-1}||^2},$$

$$\beta_k^{DL} = \frac{g_k^\intercal (y_{k-1} - ts_{k-1})}{d_{k-1}^\intercal y_{k-1}}, \qquad \beta_k^{DY} = \frac{||g_k||^2}{d_{k-1}^\intercal y_{k-1}}, \tag{3.3}$$

where y_{k-1} and s_{k-1} are defined by

$$y_{k-1} = g_k - g_{k-1}, \qquad s_{k-1} = x_k - x_{k-1}.$$

Some recent research aims at generating a search direction satisfying the descent condition $g_k^\intercal d_k < 0$ for all k or the sufficient descent condition; i.e., there is a positive constant r such that

$$g_k^\intercal d_k \leq -r||g_k||^2 \qquad \forall\ k. \tag{3.4}$$

Despite the fact that the original form of CG method is defined by (3.2), there exist some other forms which are really effective to solve the problem (2.3). For instance, Zhang et al. [52] proposed a three term PR method whose the search direction was generated as follows

$$d_k = \begin{cases} -g_k, & if \quad k = 0, \\ -g_k + \beta_k^{PR} d_{k-1} - \theta_k^{(1)} y_{k-1}, & if \quad k \geq 1, \end{cases} \tag{3.5}$$

where $\theta_k^{(1)} = \frac{g_k^\intercal d_{k-1}}{||g_{k-1}||^2}$. In this respect, Narushima et al. [36], Al-Baal et al. [1] and Sugiki et al. [46] proposed three term CG method. Furthermore, There is another form of CG method, Zhang et al. [53] represented a modified FR conjugate gradient method to solve the problem (1.1) as follows:

$$d_k = \begin{cases} -g_k & if \quad k = 0, \\ -\theta_k^{(2)} g_k + \beta_k^{FR} d_{k-1} & if \quad k > 0, \end{cases} \tag{3.6}$$

where $\theta_k^{(2)} = \frac{d_{k-1}^\intercal y_{k-1}}{||g_{k-1}||^2}$. Note the search direction (3.6) can be written by $d_k = \theta_k^{(2)}(-g_k + \beta_k^{DY} d_{k-1})$, and thus, it can be regarded as a scaled DY method.

Now, after this brief review, we introduce our CG method to solve (1.1), in the following subsection

3.1 New HS method for nonsmooth unconstrained convex optimization problem

Hager et. al [20] mentioned that some CG methods which have parameter β_k with $g_k^\intercal y_k$ in the numerator such as HS and PRP methods are better than the performance of methods with $||g_k||^2$ in the numerator of β_k. On the other hand, methods with $g_k^\intercal y_k$ in the numerator suffer from global convergence against methods that have $||g_k||^2$ in the numerator. By the way, we build nonmonotone HS method to solve nondifferentiable unconstrained problem as follows by inspiring what Zhang et al. did in [53], which has a global convergence. Suppose

$$d_k = \begin{cases} -g^\alpha(x_k, \varepsilon_k) & if \quad k = 0, \\ -\theta_k g^\alpha(x_k, \varepsilon_k) + \beta_k^{HS} d_{k-1} & if \quad k > 0, \end{cases} \tag{3.7}$$

where

$$\theta_k = 1 + \frac{g^\alpha(x_k, \varepsilon_k)^\intercal y_{k-1} \times g^\alpha(x_k, \varepsilon_k)^\intercal d_{k-1}}{||g^\alpha(x_k, \varepsilon_k)||^2 (d_{k-1}^\intercal y_{k-1})}, \tag{3.8}$$

here,

$$\beta_k^{HS} = \frac{g^\alpha(x_k, \varepsilon_k)^\intercal y_k}{d_k^\intercal y_k}, \qquad y_k = g^\alpha(x_k, \varepsilon_k) - g^\alpha(x_{k-1}, \varepsilon_{k-1}).$$

By the definition of d_k, one can proof easily that for any $k \geq 0$

$$g^\alpha(x_k, \varepsilon_k)^\intercal d_k = -||g^\alpha(x_k, \varepsilon_k)||^2. \tag{3.9}$$

Consequently, vector d_k is a descent direction of F at x_k. After selecting a descent direction, the main step of the algorithm is choosing an appropriate step length because of the convergence and implementation of CG methods. Amini et al. [2] represented a modified nonmonotone strategy as follows

$$f(x_k + \alpha_k d_k) \leq R_k + \delta \alpha_k g_k^\intercal d_k, \tag{3.10}$$

where

$$R_k = \eta_k f_{l_k} + (1 - \eta_k) f_k, \tag{3.11}$$

where $0 < \delta < 1$, $0 \leq \eta_{\min} \leq \eta_{\max} \leq 1$, $\eta_k \in [\eta_{\min}, \eta_{\max}]$ and f_{l_k} is defined as follows

$$f_{l(k)} = \max_{0 \leq j \leq m_k} \{f_{k-j}\}, \qquad k = 0, 1, 2, \ldots, \tag{3.12}$$

where $m_0 = 0$ and $0 \leq m_k \leq \min\{m_{k-1} + 1, M\}$ with $M \geq 0$. It is clear that condition (3.10) is exactly Armijo condition when constant M is equal to zero, namely,

$$f(x_k + \alpha_k d_k) \leq f(x_k) + \delta \alpha_k g_k^{\intercal} d_k. \tag{3.13}$$

We apply the condition (3.10) to find a appropriate step length. Now, the algorithm of nonmonotone HS method is stated as follows

Algorithm 1 Nonmonotone CG Algorithm

```
Step 0. Given $x_0 \in \mathbb{R}^n$, $0 < \rho < 1$, $0 < \delta < \frac{1}{2}$, $s > 0$, $0 \leq \eta_{\min} \leq \eta_0 \leq \eta_{\max}$, $M \geq 0$, $\lambda > 0$ and $0 < \epsilon < 1$. Let $m_0 = 0$,
    $R_0 = F^{\alpha}(x_0, \varepsilon_0)$, $d_0 = -g^{\alpha}(x_0, \varepsilon_0)$, and $k = 0$.
Step 1. If $||g^{\alpha}(x_k, \varepsilon_k)|| < \varepsilon$, then stop.
Step 2. Compute a scalar $\varepsilon_{k+1}$ so that $0 < \varepsilon_{k+1} < \varepsilon_k$ and $\sum_{k=0}^{\infty} \varepsilon_k < +\infty$. Also, compute the step size $\alpha_k$ by (3.10)
    based on backtracking process ($\alpha_k = \rho^i s, \quad i \in \{0, 1, 2, \ldots\}$).
Step 3. Let $x_{k+1} = x_k + \alpha_k d_k$.
Step 4. Compute the search direction $d_{k+1}$ by (3.7).
Step 5. Let $k := k + 1$, and go to Step 1.
```

To show that Algorithm 1 is well defined, we proof d_k is a descent direction of f at x_k in Lemma 4.2.

4 Global convergence

In order to establish the global convergence property, we make the following standard assumption for objective function.

Assumption A

(i) The level set $\Omega = \{x \in \mathbb{R}^n |\ F(x) \leq F(x_0) + \sum_{k=0}^{\infty} \varepsilon_k\}$ is bounded.

(ii) In some neighbourhood N of Ω, the gradient of F is Lipschitz continuous, that is, there exist a constant $K > 0$ so that

$$||g(x_1) - g(x_2)|| \leq K||x_1 - x_2|| \qquad \forall x_1, x_2 \in N. \tag{4.1}$$

(iii) The sequence ε_k converges to zero.

(iv) F is bounded from below.

Remark 4.1. We can get from Assumption A(ii), that there is a constant $\gamma > 0$, such that

$$||g(x)|| \leq \gamma, \qquad \forall x \in \Omega. \tag{4.2}$$

The following lemma indicates that our purposed d_k belongs to trust region automatically. It is also the fundamental part of our global convergence proof.

Lemma 4.2. Let d_k be generated by (3.7), if there is constant $\mu > 0$ so that,

$$(g^{\alpha}(x, \varepsilon) - g^{\alpha}(\bar{x}, \bar{\varepsilon}))^{\intercal}(x - \bar{x}) \geq \mu||x - \bar{x}|| \qquad x, \bar{x} \in \Omega, \tag{4.3}$$

then

$$||d_k|| \leq (1 + 2\frac{K}{\mu})||g^{\alpha}(x_k, \varepsilon_k)|| \tag{4.4}$$

Proof . From (4.3) and $\alpha_{k-1}d_{k-1} = x_k - x_{k-1}$ we have $d_{k-1}^{\top}y_{k-1} \geq \mu\alpha_{k-1}||d_{k-1}||^2$. On the other hand, according to Cauchy-Schwarz inequality and Assumption A (ii)

$$\begin{aligned}
||d_k|| &\leq |1 + \frac{g^{\alpha}(x_k,\varepsilon_k)^{\top}y_{k-1} \times g^{\alpha}(x_k,\varepsilon_k)^{\top}d_{k-1}}{||g_k||^2(d_{k-1}^{\top}y_{k-1})}|||g^{\alpha}(x_k,\varepsilon_k)|| + |\frac{g^{\alpha}(x_k,\varepsilon_k)^{\top}y_{k-1}}{d_{k-1}^{\top}y_{k-1}}|||d_{k-1}|| \\
&\leq ||g^{\alpha}(x_k,\varepsilon_k)|| + \frac{K\alpha_{k-1}||g^{\alpha}(x_k,\varepsilon_k)||^2||d_{k-1}||^2}{\alpha_{k-1}\mu||g^{\alpha}(x_k,\varepsilon_k)||^2||d_{k-1}||^2}||g^{\alpha}(x_k,\varepsilon_k)|| + \frac{K\alpha_{k-1}||g^{\alpha}(x_k,\varepsilon_k)||}{\alpha_{k-1}\mu||d_{k-1}||^2}||d_{k-1}||^2 \\
&= (1 + 2\frac{K}{\mu})||g^{\alpha}(x_k,\varepsilon_k)||. \qquad (4.5)
\end{aligned}$$

□

The structure of the step length proposed by Amini et al. [2] shows, that by considering Assumption A, if (x_k, ε_k) is generated by Algorithm 1, we have $F^{\alpha}(x_k, \varepsilon_k) \leq R_k$ for each iterate of Algorithm 1. Furthermore, there exists α_k satisfying in (3.10).

Next lemma is a fundamental part of our global convergence proof.

Lemma 4.3. Let Assumption A be satisfied. If $\{(x_k, \varepsilon_k)\}$ is the sequence generated by Algorithm 1 and $\varepsilon_k = o(\alpha_k^2||d_k||^2)$. then, there is $P > 0$ so that

$$\forall k; \quad \alpha_k \geq P \qquad (4.6)$$

Proof . Let α_k be satisfied in (3.10). Suppose $\liminf_{k\to\infty}\alpha_k = 0$. By considering $\alpha_k' = \rho^{-1}\alpha_k$, where $0 < \rho < 1$ is a constant, we have

$$F(x_k + \alpha_k'd_k, \varepsilon_k) - R_k > \delta\alpha_k'g^{\alpha}(x_k, \alpha_k)^{\top}d_k. \qquad (4.7)$$

Now, according to the structure of the step length in (3.10), $F(x_k, \varepsilon_k) \leq R_k$, and (4.7) we can write

$$F(x_k + \alpha_k'd_k, \varepsilon_k) - F(x_k, \varepsilon_k) \geq F(x_k + \alpha_k'd_k, \varepsilon_k) - R_k > \delta\alpha_k'g^{\alpha}(x_k, \varepsilon_k)^{\top}d_k. \qquad (4.8)$$

By (2.10), (4.7) and Tylor Series, we have

$$\begin{aligned}
\delta\alpha_k'g^{\alpha}(x_k,\varepsilon_k)^{\top}d_k &< F(x_k + \alpha_k'd_k, \varepsilon_{k+1}) - F(x_k, \varepsilon_k) \\
&\leq F(x_k + \alpha_k'd_k) - F(x_k) + \varepsilon_{k+1} \\
&= \alpha_k'd_k^{\top}g(x_k) + \frac{1}{2}(\alpha_k')^2d_k^{\top}\nabla^2F(\eta_k)d_k + \varepsilon_{k+1}, \qquad (4.9)
\end{aligned}$$

where η_k is on the line segment connecting x_k and x_{k+1}. Also, according to (4.2) we have $d_k^{\top}\nabla^2F(\eta_k)d_k \leq \gamma||d_k||^2$. So, we can write

$$\delta\alpha_k'g^{\alpha}(x_k,\varepsilon_k)^{\top}d_k < \alpha_k'd_k^{\top}g(x_k) + \frac{\gamma}{2}(\alpha_k')^2||d_k||^2 + \varepsilon_{k+1}. \qquad (4.10)$$

Now, we can rewrite (4.10) as follows

$$\frac{\delta g^{\alpha}(x_k,\varepsilon_k)^{\top}d_k - d_k^{\top}g(x_k)}{\alpha_k'} < \frac{\gamma}{2}||d_k||^2 + \frac{\varepsilon_{k+1}}{(\alpha_k')^2},$$

so we have

$$\begin{aligned}
\rho^{-1}\alpha_k = \alpha_k' &> 2\frac{(g^{\alpha}(x_k,\varepsilon_k) - g(x_k))^{\top}d_k - (1-\delta)g^{\alpha}(x_k,\varepsilon_k)^{\top}d_k - \frac{\varepsilon_{k+1}}{\alpha_k'}}{\gamma||d_k||^2} \\
&\geq \frac{(1-\delta)||g^{\alpha}(x_k,\varepsilon_k)||^2 - \sqrt{\frac{2\varepsilon_k}{\lambda}}||d_k|| - \frac{\varepsilon_k}{\alpha_k'}}{\gamma||d_k||^2} \\
&= \frac{(1-\delta)||g^{\alpha}(x_k,\varepsilon_k)||^2}{\gamma||d_k||^2} - \frac{\frac{o(\alpha_k)}{\sqrt{\lambda}} - o(\alpha_k)}{\gamma}, \qquad (4.11)
\end{aligned}$$

in which, the second inequality is held by (3.9), (2.12) and $\varepsilon_k > \varepsilon_{k+1}$. Hence, according to (4.4), we can result that

$$\rho^{-1}\alpha_k > \frac{(1-\delta)}{(1+\frac{K}{\mu})^2\gamma} - \frac{\frac{o(\alpha_k)}{\sqrt{\lambda}} - o(\alpha_k)}{\gamma}. \qquad (4.12)$$

By dividing The sides of the inequality by α_k, as $k \to \infty$, we have

$$\rho^{-1} \geq \lim_{k\to\infty} \frac{1-\delta}{(1+\frac{K}{\mu})^2\gamma} \times \frac{1}{\alpha_k} - \sqrt{\frac{1}{\lambda}} - \frac{1}{\gamma} = +\infty.$$

But, this is obviously wrong. So, inequality (4.6) holds. □

Theorem 4.4 (Global Convergence). Assume the conditions of Lemma 4.3 hold. Then, $\lim_{k\to\infty} ||g_k|| = 0$ and any accumulation point of $\{x_k\}$ is an optimal solution of problem (1.1)

Proof . At the first, we prove that

$$\lim_{k\to\infty} ||g^\alpha(x_k, \varepsilon_k)|| = 0. \tag{4.13}$$

In order to result (4.13), we assume that there is a scalar $L > 0$ and $N \in \mathbb{N}$ so that

$$||g^\alpha(x_k, \varepsilon_k)|| \geq L, \qquad \forall k \geq N. \tag{4.14}$$

by considering (3.9), the right side of (3.10), and (4.6), we have

$$\sum_{k=1}^{\infty} -\delta\alpha_k g^\alpha(x_k, \varepsilon_k)^\intercal d_k \geq \sum_{k=1}^{\infty} \delta P||g^\alpha(x_k, \varepsilon_k)||^2, \tag{4.15}$$

so, according to (4.14) and (4.15), we can result

$$\sum_{k=1}^{\infty} -\delta\alpha_k g^\alpha(x_k, \varepsilon_k)^\intercal d_k \geq \sum_{k=N}^{\infty} \delta PL = +\infty. \tag{4.16}$$

On the other hand, we get by the condition of step length choice, (3.10),

$$\begin{aligned} -\delta\alpha_k g^\alpha(x_k, \varepsilon_k)^\intercal d_k &\leq R_k - F^\alpha(x_k + \alpha_k d_k, \varepsilon_{k+1}) \\ &= \eta_k \max_{0\leq j\leq m(k)} \{F^\alpha(x_{k-j}, \varepsilon_{k-j})\} + (1-\eta_k)F^\alpha(x_k, \varepsilon_k) - F^\alpha(x_k + \alpha_k d_k, \varepsilon_{k+1}), \end{aligned} \tag{4.17}$$

by considering that d_k ,generated by Algorithm 1, is a descent direction, $g^\alpha(x_k, \varepsilon_k)^\intercal d_k = -||g^\alpha(x_k, \varepsilon_k)||^2$, we have the sequence $\{F^\alpha(x_k, \varepsilon_k)\}$ is strongly decreasing or equivalently $F^\alpha(x_{k+1}, \varepsilon_{k+1}) > F^\alpha(x_k, \varepsilon_k)$ for every $k \in \mathbb{N}$. So, by the condition of step length choice and (4.17), we have

$$\sum_{k=M}^{\infty} -\delta\alpha_k g^\alpha(x_k, \varepsilon_k)^\intercal d_k \leq \sum_{k=M}^{\infty} F^\alpha(x_k, \varepsilon_k) - F^\alpha(x_k + \alpha_k d_k, \varepsilon_{k+1}) + \eta_k \sum_{k=M}^{\infty} F^\alpha(x_{k-M}, \varepsilon_{k-M}) - F^\alpha(x_k, \varepsilon_k), \tag{4.18}$$

where on the right side of (4.18) there are two Telescoping series. By Assumption A we get

$$\sum_{k=M}^{\infty} -\delta\alpha_k g^\alpha(x_k, \varepsilon_k)^\intercal d_k < +\infty. \tag{4.19}$$

But this inequality is in contradiction with (4.16), by getting $Q = \max\{M, N\}$. So, we conclude that

$$||g^\alpha(x_k, \varepsilon_k)|| \to 0 \tag{4.20}$$

as $k \to \infty$. Now, let $\bar{x}$ be an accumulation point of $\{x_k\}$. Hence, there is a subsequence $\{x_{k_n}\}$ so that

$$\lim_{k_n\to\infty} x_k = \bar{x}. \tag{4.21}$$

By the definition of F, we conclude $g(x_k) = \frac{(x_k - p(x_k))}{\lambda}$. Thus, according to (4.20) and (4.21), the equality

$$\bar{x} = p(\bar{x})$$

holds. Moreover, $\bar{x}$ is an optimal solution of problem (1.1). □

5 Conclusions and future works

We presented a modified HS conjugate gradient method for solving nonsmooth convex optimization by making use of the Moreau-Yosida regularization to convert nonsmooth objective function to a smooth function. After converting, we applied our method to solve new problem. Our method produced a descent direction in each iteration. The global convergence was established under standard conditions. another feature of this method was using nonmonotone line search [2] to have better implementation.

Acknowledgement

The authors thank the Research Council of Ferdowsi University of Mashhad for supporting this work

References

[1] M. Al-Baali, Y. Narushima, and H. Yabe, *A family of three-term conjugate gradient methods with sufficient descent property for unconstrained optimization*, Comput. Optim. Appl. **60** (2015), 89–110.

[2] K. Amini, M. Ahookhosh, and H. Nosratipour, *An inexact line search approach using modified nonmonotone strategy for unconstrained optimization*, Numer. Algorithms **66** (2014), 49–78.

[3] A. Astorino, A. Fuduli, and E. Gorgone, *Nonsmoothness in classification problems*, Optim. Metho. Software **23** (2008), 675–688.

[4] A. Auslender, *Numerical methods for nondifferentiable convex optimization*, Math. Programm. **30** (1987), 102–126.

[5] S. Babaie-Kafaki and R. Ghanbari, *The Dai–Liao nonlinear conjugate gradient method with optimal parameter choices*, Eur. J. Oper. Res. **234** (2014), 625–630.

[6] A.M. Bagirov, L. Jin, N. Karmitsa, A.Al. Nuaimat, and N. Sultanova, *Subgradient method for nonconvex nonsmooth optimization*, J. Optim. Theory Appl. **157** (2013), 416–435.

[7] A.M. Bagirov, B. Karasözen, and M. Sezer, *Discrete gradient method: Derivative-free method for nonsmooth optimization*, J. Optim. Theory Appl. **137** (2007), 317–334.

[8] A. Bagirov, N. Karmitsa, and M.M. Mäkelä, *Introduction to Nonsmooth Optimization: Theory, Practice and Software*, Springer International Publishing, 2014.

[9] P.S. Bradley, U.M. Fayyad, and O.L. Mangasarian, *Mathematical programming for data mining: Formulations and challenges*, Inf. J. Comput. **11** (1999), 217–238.

[10] J.V. Burke, A.S. Lewis, and M.L. Overton, *A robust gradient sampling algorithm for nonsmooth, nonconvex optimization*, SIAM J. Optim. **15** (2005), 751–779.

[11] J.V. Burke and M. Qian, *On the superlinear convergence of the variable metric proximal point-algorithm using Broyden and BFGS matrix secant updating*, Math. Programm. **88** (2000), 157–181.

[12] E. Carrizosa and D.R. Morales, *Supervised classification and mathematical optimization*, Comput. Oper. Res. **40** (2013), 150–165.

[13] X. Chen and M. Fukushima, *Proximal quasi-Newton methods for nondifferentiable convex optimization*, Math. Programm. **85** (1999), 313–334.

[14] W.Y. Cheng, *A two term PRP based descent method*, Numer. Funct. Anal. Optim. **28** (2007), 1217–1230.

[15] A. Conn, N. Gould, and P. Toint, *Trust Region Methods*, Society for Industrial and Applied Mathematics, 2000.

[16] Y.H. Dai, *Convergence properties of the BFGS algorithm*, SIAM J. Optim. **13** (2002), 693–701.

[17] Y.H. Dai and L.Z. Liao, *New conjugacy conditions and related nonlinear conjugate gradient methods*, Appl. Math. Optim. **43** (2001), 87–101.

[18] M. Fukushima and L. Qi, *A globally and superlinearly convergent algorithm for nonsmooth convex minimization*, SIAM J. Optim. **6** (1996), 1106–1120.

[19] N.I.M. Gould, C. Sainvitu, and P.L. Toint, *A filter-trust-region method for unconstrained optimization*, SIAM J. Optim. **16** (2005), 341–357.

[20] W.W. Hager and H. Zang, *A survey of the nonlinear conjugate gradient methods*, Pacific J. Optim. **2** (2006), 35–58.

[21] P. Hansen and B. Jaumard, *Cluster analysis and mathematical programming*, Math. Programm. **79** (1997), 191–215.

[22] J.B. Hiriart-Urruty and C. Lemarechal, *Convex Analysis and Minimization Algorithms*, Springer, Berlin, 1993.

[23] A. Jain, M. Murty, and P. Flynn, *Data clustering: A review*, ACM Comput. Surveys **31** (1999), 264–323.

[24] T. Kärkkäinen and E. Heikkola, *Robust formulations for training multilayer perceptrons*, Neural Comput. Appl. **16** (2004), 837–862.

[25] T. Kärkkäinen and K. Majava, *Semi-adaptive, convex optimization methodology for image denoising*, IEEE Proc.-Vision Image Signal Process. **152** (2005), 553–560.

[26] N. Karmitsa, *Diagonal bundle method for nonsmooth Sparse optimization*, J. Optim. Theory Appl. **166** (2015), 889–905.

[27] N. Karmitsa and M.M. Mäkelä, *Adaptive limited memory bundle method for bound constrained large-scale nonsmooth optimization*, Optimization **59** (2010), 945–962.

[28] Q. Li, *Conjugate gradient type methods for the nondifferentiable convex minimization*, Optim. Lett. **7** (2013), 533–545.

[29] S. Lu, Z. Wei, and L. Li, *A trust region algorithm with adaptive cubic regularization methods for nonsmooth convex minimization*, Comput. Optim. Appl. **51** (2012), no. 2, 551-573.

[30] M.M. Mäkelä and T. Männikkö, *Numerical solution of nonsmooth optimal control problems with an application to the continuous casting process*, Adv. Math. Sci. Appl. **4** (1994), 491–515.

[31] M. M. Mäkelä, T. Männikkö, and H. Schramm, H *Application of nonsmooth optimization methods to continuous casting of steel, DFG-Schwerpunktprogramm "Anwendungsbezogene Optimierung und Steuerung"*, Report 421, Universität Bayreuth, 1993.

[32] M. Mehiddin Al-Baali, E. Spedicato, and F. Maggioni, *Broyden's quasi-Newton methods for a nonlinear system of equations and unconstrained optimization: A review and open problems*, Optim. Meth. Software **29** (2014), 937–954.

[33] K. Miettinen, M.M. Mäkelä, and T. Männikkö, *Optimal control of continuous casting by nondifferentiable multiobjective optimization*, Comput. Optim. Appl. **11** (1998), 177–194.

[34] E.S. Mistakidis and G.E. Stavroulakis, *Nonconvex Optimization in Mechanics: Smooth and Nonsmooth Algorithms, Heuristics and Engineering Applications by the F. E. M.*, Kluwer Academic Publishers, Dordrecht, 1998.

[35] J.J. Moreau, P.D. Panagiotopoulos, and G. Strang, *Topics in Nonsmooth Mechanics*, Birkhäuser Verlag, Basel, 1988.

[36] Y. Narushima, H. Yabe, and J.A. Ford, *A three-term conjugate gradient method with sufficient descent property for unconstrained optimization*, SIAM J. Optim. **21** (2011), 212-230.

[37] A. Nedić and A. Ozdaglar, *Subgradient methods for saddle-point problems*, J. Optim. Theory Appl. **142** (2009), 205–228.

[38] Y.U. Nesterov, *Excessive gap technique in nonsmooth convex minimization*, SIAM J. Optim. **16** (2005), 235-249.

[39] Y.U. Nesterov, *Smooth minimization of nonsmooth functions*, Math. Programm. **103** (2005), 127–152.

[40] Y.U. Nesterov, *Primal-dual subgradient methods for convex problems*, Math. Programm. **120** (2009), 221–259.

[41] J. Outrata, M. Kočvara, and J. Zowe, *Nonsmooth Approach to Optimization Problems with Equilibrium Constraints: Theory, Applications and Numerical Results*, Kluwer Academic Publisher, Dordrecht, 1998.

[42] A.I. Rauf and M. Fukushima, *Global convergent BFGS method for nonsmooth convex optimization*, J. Optim.

Theory Appl. **104** (2000), 539–558.

[43] N. Sagara and M. Fukushima, *A trust region method for nonsmooth convex optimization*, J. Ind. Manag. Optim. **1** (2005), 171–180.

[44] C. Sagastizábal and M. Solodov, *An infeasible bundle method for nonsmooth convex constrained optimization without a penalty function or a filter*, SIAM J. Optim. **16** (2005), 146–169.

[45] H. Schramm and J. Zowe, *A version of the bundle idea for minimizing a nonsmooth function: conceptual idea, convergence analysis, numerical results*, SIAM J. Optim. 2 (1992), 121–152.

[46] K. Sugiki, Y. Narushima, and H. Yabe, *Globally convergent three-term conjugate gradient methods that use secant conditions and generate descent search directions for unconstrained optimization*, J. Optim. Theory Appl. **153** (2012), 733–757.

[47] W. Sun, *Nonmonotone trust region method for solving optimization problems*, Appl. Math. Comput. **156** (2004), 159–174.

[48] I.H. Witten and E. Frank, *Data mining: Practical Machine Learning Tools and Techniques*, 2nd ed., Elsevier Inc., Amsterdam, 2005.

[49] H. Yabe, H. Ogasawara, and M. Yoshino, *Local and superlinear convergence of quasi-Newton methods based on modified secant conditions*, J. Comput. Appl. Math. **205** (2007), 617–632.

[50] G. Yuan, Z. Weia, and G.A. Li, *Modified Polak–Ribiére–Polyak conjugate gradient algorithm for nonsmooth convex programs*, J. Comput. Appl. Math. **255** (2014), 86–96.

[51] J.Z. Zhang, N.Y. Deng, and L.H. Chen, *New Quasi-Newton equation and related methods for unconstrained optimization*, J. Optimi. Theory Appl. **102** (1999), 147–167.

[52] L. Zhang, W. Zhou, and D.H. Li, *A descent modified Polak-Ribiére-Polyak conjugate gradient method and its global convergence*, IMA J. Numer. Anal. **26** (2006), 629–640.

[53] L. Zhang, W. Zhou, and D.H. Li, *Global convergence of a modified Fletcher–Reeves conjugate gradient method with Armijo-type line search*, Numer. Math. **104** (2006), 561–572.