Accepted Manuscript

An adaptive three-term conjugate gradient method based on self-scaling memoryless BFGS matrix

Shengwei Yao, Liangshuo Ning

PII: S0377-0427(17)30503-4
DOI: https://doi.org/10.1016/j.cam.2017.10.013
Reference: CAM 11342

To appear in: *Journal of Computational and Applied Mathematics*

Received date : 4 June 2017
Revised date : 24 September 2017

Please cite this article as: S. Yao, L. Ning, An adaptive three-term conjugate gradient method based on self-scaling memoryless BFGS matrix, *Journal of Computational and Applied Mathematics* (2017), https://doi.org/10.1016/j.cam.2017.10.013

An adaptive three-term conjugate gradient method based on self-scaling memoryless BFGS matrix ☆

Shengwei Yao[a,b,*], Liangshuo Ning[a]

[a]School of Information and Statistics, Guangxi University of Finance and Economics, Nanning, 530003, China.
[b]Guangxi Key Laboratory Cultivation Base of Cross-border E-commerce Intelligent Information Processing, Guangxi University of Finance and Economics, Nanning, 530003, China

Abstract

Due to its simplicity and low memory requirement, conjugate gradient methods are widely used for solving large-scale unconstrained optimization problems. In this paper, we propose a three-term conjugate gradient method. The search direction is given by a symmetrical Perry matrix, which contains a positive parameter. The value of this parameter is determined by minimizing the distance of this matrix and the self-scaling memoryless BFGS matrix in the Frobenius norm. The sufficient descent property of the generated directions holds independent of line searches. The global convergence of the given method is established under Wolfe line search for general non-convex functions. Numerical experiments show that the proposed method is promising.

Keywords: unconstrained optimization, conjugate gradient method, self-scaling memoryless BFGS matrix, global convergence

2010 MSC: 90C53, 49M37, 65F15

1. Introduction

The method involved in this paper is designed to solve the following unconstrained optimization problem:

$$\min f(x), x \in R^n, \tag{1}$$

☆Supported by: Guangxi Universities Foundation Grant no: KY2015YB268
*Corresponding author
Email address: ysw@gxu.edu.cn (Shengwei Yao)

where $f(x)$ is a continuously differentiable objective function.

The most popular methods for such unconstrained optimization problems, especially for large-scale problems, are first-order methods. Namely, only the gradient of the objective function is used in iterations. Since Newton or quasi-Newton methods require calculating and storing Hessian or approximate Hessian matrix, for a n-dimension problem, it need at least $O(n^2)$ storages and calculations at each iteration. However, first-order methods only need $O(n)$ storages and calculations at each iteration. So, in this paper, we mainly focus on studying the first-order method for solving problem (1). There are several types of first-order method to solve problem (1).

Gradient descent method is the simplest iterative first-order method with the form :

$$x_{k+1} = x_k - \alpha_k g_k, \tag{2}$$

where the step size α_k is either fixed or determined by a line search, and $g_k = \nabla f(x_k)$ is the gradient of the objective function. It requires only vector operations. However, its convergent rate is slowly and it is easy to form a zigzag search path.

Subgradient method [1]has the similar form with gradient descent method:

$$x_{k+1} = x_k + \alpha_k \partial f_k, \tag{3}$$

where $\partial f_k$is the subgradient of the objective function f(x) at x_k.

Nesterov accelerated gradient method [2] is one of well-known accelerated gradient method which accelerates first-order method by foming estimating sequences. Gonzaga and Karas [3] presented a variant of Nesterov's method that adapts to unknown strong convexity. To improve the convergence rate, O'Donoghue and Candés [4] proposed a heuristic for resetting the momentum term to zero.

Th heavy ball method proposed by Polyak [5] adds an adjusted term to the gradient step:

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k) + \beta_k d_k, \tag{4}$$

where $d_k = x_k - x_{k-1}$. The motivation is to avoid the bounce between the walls of narrow 'valleys' on the objective surface which may occur in gradient descent method. However, in the heavy ball method, the objective function $f(x)$ is supposed to be strongly convex and strongly smooth, namely:

strongly convex with the constant μ

$$f(y) \geq f(x) + \nabla f(x)^T(y-x) + \frac{\mu}{2}||y-x||^2, \tag{5}$$

strongly smooth with the constant L

$$f(y) \leq f(x) + \nabla f(x)^T(y-x) + \frac{L}{2}||y-x||^2. \tag{6}$$

And, the step size α_k depends on two constants L and μ.

The conjugate gradient (CG) method can be considered as an instance of the heavy ball method with adaptive step size. However, conjugate gradient method has an advantage that it does not require knowledge of L and μ to determine step size. The iterative formula of conjugate gradient method is given by

$$x_{k+1} = x_k + \alpha_k d_k. \tag{7}$$

$$d_{k+1} = \begin{cases} -g_{k+1}, & for \quad k = 0; \\ -g_{k+1} + \beta_k d_k, & for \quad k \geq 1, \end{cases} \tag{8}$$

where β_k is a scalar called the conjugate gradient(CG) parameter, $\alpha_k > 0$ is the step size obtained by some line searches [6, 7]. Among them, the so-called Wolfe line search [8, 9] requires α_k satisfying

$$f(x_k + \alpha_k d_k) - f(x_k) \leq \rho \alpha_k g_k^T d_k, \tag{9}$$

and

$$g(x_k + \alpha_k d_k)^T d_k \geq \sigma g_k^T d_k, \tag{10}$$

where $0 < \delta < \sigma < 1$. In the convergence analysis and numerical implementations of the conjugate gradient methods, the step size α_k is often computed by the strong Wolfe line search [9] which requires α_k satisfying (9) and

$$|g(x_k + \alpha_k d_k)^T d_k| \leq -\sigma g_k^T d_k. \tag{11}$$

Due to its simplicity and low memory requirement, conjugate gradient methods are widely used in solving large-scale optimization problems. In the past decades, a variety of conjugate gradient methods are developed. There are some well known conjugate gradient methods, such as Fletcher-Reeves (FR) method [10], Hestenes-Stiefel (HS) method [11], Polak-Ribiére-Polyak (PRP) method [12, 13] and Dai-Yuan (DY) method [14].

Recent effors have been made to relate the nonlinear conjugate gradient method to modified conjugacy gradient conditions and quasi-Newton method. The following Dai-Liao [15] conjugacy condition

$$d_{k+1}^T y_k = -t g_{k+1}^T s_k, \tag{12}$$

where $s_k = x_{k+1} - x_k$, $y_k = g_{k+1} - g_k$ and t is a positive parameter, is one of the most interesting conjugacy condition. Based on condition (12), Dai and Liao obtained $\beta_k^{DL(t)}$ as follows

$$\beta_k^{DL}(t) = \frac{g_{k+1}^T y_k}{d_k^T y_k} - t\frac{g_{k+1}^T s_k}{d_k^T y_k} \tag{13}$$

Hager and Zhang [16] presented an another choice for the parameter $t = 2\frac{||y_k||^2}{s_k^T y_k}$ and obtained CG-Descent method by computing the parameter β_k in (8) with

$$\bar{\beta}_k{}^N = \max\{\beta_k^N, \eta_k\}, \tag{14}$$

in which

$$\eta_k = \frac{-1}{||d_k|| \min\{\eta, ||g_k||\}}, \quad \beta_k^N = \frac{g_{k+1}^T y_k}{d_k^T y_k} - 2\frac{||y_k||^2}{d_k^T y_k}\frac{g_{k+1}^T d_k}{d_k^T y_k}, \tag{15}$$

where η is a fixed positive parameter. By seeking the conjugate gradient direction closest to the direction of the scaled memoryless BFGS method, Dai and Kou obtianed CGOPT family [17], in which the parameter $\beta_k(\tau_k)$ is determined by

$$\beta_k(\tau_k) = \frac{y_k^T g_{k+1}}{d_k^T y_k} - \frac{d_k^T g_{k+1}}{d_k^T y_k}\frac{\|y_k\|^2}{d_k^T y_k}(1 + \tau_k \frac{s_k^T y_k}{\|y_k\|^2}) + \frac{d_k^T g_{k+1}}{\|d_k\|^2}. \tag{16}$$

Combining with two types of modified secant equations, Kou [18] proposed an improved CGOPT method. Similar with Dai and Liao's approach, based on modified secant equations proposed in [19, 20, 21], other conjugate gradient methods have also beeen developed by Li, Yabe and Zhou et al. in [22, 23, 24], respectively. Other choices for the paramter t in (13) can also be found in [25, 26, 27].

On the other hand, by using quasi-Newton techniques in conjugate gradient method, some authors considered conjugate gradient method as a special type of quasi-Newton method. Based on this technique, Perry [28] proposed the following formula for computing the parameter β in (8):

$$\beta_k^P = \frac{g_{k+1}^T y_k - g_{k+1}^T s_k}{d_k^T y_k}. \tag{17}$$

By substituting (17) into (8) and applying some simple algebraic manipulation, we can obtain the corresponding Perry's search direction as follows:

$$d_{k+1}^P = -Q_{k+1}^P g_{k+1}, \tag{18}$$

where

$$Q_{k+1}^P = I - \frac{s_k y_k^T}{y_k^T s_k} + \frac{s_k s_k^T}{y_k^T s_k}. \tag{19}$$

From equations (18) and (19), it is obviously that Perry conjugate gradient method can be considered as a special case of quasi-Newton method. In Perry method, the matrix Q_{k+1}^P is used to estimate the approximation of the inverse Hessian matrix of the objective function. From a strictly point of view, Perry's method can not be considered as a quasi-Newton method, since the matrix Q_{k+1}^P is not positive symmetric and does not fullfill secant condtion.

To overcome the asymmetry, combining with Dai-Liao conjugacy condition (12), Babaie-Kafaki and Ghanbari [29] proposed the following matrix A_{k+1}

$$A_{k+1} = I - \frac{1}{2}\frac{s_k y_k^T + y_k s_k^T}{s_k^T y_k} + t\frac{s_k s_k^T}{s_k^T y_k}, \tag{20}$$

to replace the the matrix Q_{k+1}^P in (19). Andrei also presented a symmetric matrix to estimate the inverse Hessien approximation as follows:

$$Q_{k+1}^N = I - \frac{s_k y_k^T - y_k s_k^T}{s_k^T y_k} + t\frac{s_k s_k^T}{s_k^T y_k}. \tag{21}$$

By computing the parameter t in some different manners, Andrei obtained some different conjugate gradient methods [30, 31, 32, 33].

The above results indicate that conjugacy conditions and quasi-Newton techniques can be used to improve the traditional conjugate gradient efficiently. In this paper, we will investigate Perry conjugate gradient method by mininizing the distance between a symmetrical Perry matrix with a positive parameter and the self-scaling memoryless BFGS update in the Frobenius norm.

The structure of the paper is as follows. In the next section, the motivations of this paper will be discussed. After that, the corresponding method will be proposed. In section 3, the global convergence results of the obtained algorithm are established under Wolfe line search. In section 4, the numerical Dolan-Moré performance profile [34] of the proposed algorithm with some will known conjugate gradient algorithms will be shown by using the unconstrained optimization test problems from [35]

2. Motivations and the corresponding Perry conjugate gradient method

2.1. Motivations

Let us simply review the update matrices A_{k+1} and Q_{k+1}^N proposed by Saman and Reza [29] and Andrei [30, 31, 32], the iterative update matrix both can be expressed as the following form:

$$Q_{k+1} = I + R_2^{k+1} + t_k R_1^{k+1}, \tag{22}$$

where R_2^{k+1} and R_1^{k+1} are given rank 2 and rank 1 adjusted matrix respectively.

Motivated by this observation, in this paper, we propose the following symmetric Perry matrix

$$Q_{k+1}^M = I + t_k A_2^{k+1} + A_1^{k+1}, \tag{23}$$

where

$$A_2^{k+1} = -\frac{s_k y_k^T + y_k s_k^T}{s_k^T y_k}, \quad A_1^{k+1} = \frac{s_k s_s^T}{s_k^T y_k}, \tag{24}$$

t_k is a positive parameter to be determined. The search direction is generated by

$$d_{k+1} = -Q_{k+1}^M g_{k+1}, \quad k \geq 1. \tag{25}$$

2.2. The optimal choice for the parameter

In the following, we will discuss some properties of the method formed by equations (23), (24) and (25). And then, the parameter t_k will be determined by minimizing the distance between the matrix Q_{k+1}^M and the self-scaling memoryless BFGS matrix in the Frobenius norm.

It is well known that, BFGS method [36, 37, 38, 39] is one of the most efficient quasi-Newton method. The BFGS update matrix is given by

$$H_{k+1} = H_k - \frac{s_k y_k^T H_k + H_k y_k s_k^T}{y_k^T s_k} + (1 + \frac{y_k^T H_k y_k}{y_k^T s_k})\frac{s_k s_k^T}{y_k^T s_k}. \tag{26}$$

For large-scale problems, the expenditure of storing and computing the matrix H_k is huge. Perry [40] and Shanno [41] proposed the self-scaling memoryless BFGS update matrix by replacing H_k with a scaled identity matrix $\xi_k I$. The corresponding self-scaling memoryless BFGS matrix is given as follows:

$$H_{k+1}^{\xi_k} = \xi_k I - \xi_k \frac{s_k y_k^T + y_k s_k^T}{y_k^T s_k} + (1 + \xi_k \frac{||y_k||^2}{y_k^T s_k})\frac{s_k s_k^T}{y_k^T s_k}. \tag{27}$$

Since the self-scaling memoryless BFGS method is one of the most efficient memoryless quasi-Newton method, and, the matrix Q_{k+1}^M defined by (23) has similar structure with the self-scaling memoryless BFGS matrix $H_{k+1}^{\xi_k}$ given by (27), in this paper, the parameter t_k at each iteration, in (23) is defined by

$$\begin{aligned} t_k^* &= \arg\min_{t>0}\{\|Q_{k+1}^M - H_{k+1}^{\xi_k}\|_F^2\} \\ &= \arg\min_{t>0}\{\|(I + tA_2^{k+1} + A_1^{k+1}) - (\xi_k I - \xi_k \frac{s_k y_k^T + y_k s_k^T}{y_k^T s_k} + (1 + \xi_k \frac{||y_k||^2}{y_k^T s_k})\frac{s_k s_k^T}{y_k^T s_k})\|_F^2\}. \end{aligned} \tag{28}$$

where $\|\cdot\|_F$ is the Frobenius matrix norm. For convenience, we use $D_{k+1}(t)$ to denote $Q_{k+1}^M - H_{k+1}^{\xi_k}$. Since $\|D_{k+1}(t)\|_F^2 = tr(D_{k+1}(t)^T D_{k+1}(t))$, it follows that the problem (28) can be expressed as

$$t_k^* = \arg\min_{t>0}\{tr(D_{k+1}(t)^T D_{k+1}(t))\}. \tag{29}$$

After some directly and tediously computation, we can have

$$tr(D_{k+1}(t)^T D_{k+1}(t)) = (2 + 2a_k)t^2 - 4t + g(a_k, \xi_k), \tag{30}$$

where

$$a_k = \frac{\|s_k\|^2 \|y_k\|^2}{(s_k^T y_k)^2}, \tag{31}$$

$$g(a_k, \xi_k) = (1-\xi_k)^2 n - (a_k^2 - 2)\xi_k^2 - 2a_k\xi_k + 4\xi_k. \tag{32}$$

Based on equation (31), obviously, $a_k \geq 1$ and $(2+2a_k) > 0$, so, the problem (29) has the following unique solution

$$t_k^* = \frac{1}{1+a_k}. \tag{33}$$

It should be noticed that, since $a_k \geq 1$ given by (31), the value of $t_k^* \leq \frac{1}{2}$. From (33), the optimal value t_k^* depends only on a_k instead of ξ_k. It is an amazed result, since this result indicates that whatever the scale parameter ξ_k, the solution of (29) uniquely exists.

On the other hand, the descent property of a given search direction is very important for the convergence analysis. Now, we discuss the descent property of the direction generated by (23), (24) and (25). By substituting ((23), (24)) into (25) and multiplying both sides of (25) with g_{k+1}, we have

$$\begin{aligned} d_{k+1}^T g_{k+1} &= -\|g_{k+1}\|^2 + 2t_k \frac{g_{k+1}^T y_k g_{k+1}^T s_k}{s_k^T y_k} - \frac{(g_{k+1}^T s_k)^2}{s_k^T y_k} \\ &\leq -\|g_{k+1}\|^2 + t_k(\|g_{k+1}\|^2 + (\frac{g_{k+1}^T s_k}{s_k^T y_k})^2 \|y_k\|^2) - \frac{(g_{k+1}^T s_k)^2}{s_k^T y_k} \\ &= -(1-t_k)\|g_{k+1}\|^2 - (\frac{g_{k+1}^T s_k}{s_k^T y_k})^2 [s_k^T y_k - t_k\|y_k\|^2]. \end{aligned} \tag{34}$$

If the Wolfe line search (10) is utilized to compute the step size α_k, we can deduce that $s_k^T y_k > 0$. From equation (34), in order to ensure the sufficient descent property, in this paper, we make the following restriction on t_k

$$t_k \leq \frac{s_k^T y_k}{\|y_k\|^2}. \tag{35}$$

Based on the above discussion, the parameter t_k in this paper is determined by

$$t_k = \min\{\frac{1}{1+a_k}, \frac{s_k^T y_k}{\|y_k\|^2}\}. \tag{36}$$

2.3. The corresponding Perry conjugate gradient method

To facilitate the convergence analysis, we rewrite the method formed by (23), (24) and (25) as a typical three-term conjugate gradient method as follows:

$$d_{k+1} = -g_{k+1} + \beta_k d_k + \delta_k y_k, \tag{37}$$

where

$$\beta_k = \frac{t_k g_{k+1}^T y_k - g_{k+1}^T s_k}{d_k^T y_k}, \tag{38}$$

$$\delta_k = \frac{t_k g_{k+1}^T s_k}{s_k^T y_k}, \tag{39}$$

in which t_k is determined by (36). Furthermore, in order to establish the global convergence for general function, we need to make a nonnegative restriction on β_k as follows:

$$\beta_k^+ = \max\{\beta_k, 0\}. \tag{40}$$

Now, we present the detailed description of the obtianed algorithm for solving unconstrained optimization problems.

Algorithm 1. *New Three-term Perry Algorithm (NTPA):*

- *Step 1: Given $x_1 \in R^n$, $\varepsilon \geq 0$, $t_1 > 0$, set $d_1 = -g_1$, $k = 1$, if $\|g_1\| \leq \varepsilon$, then stop;*
- *Step 2:Compute α_k such that the Wolfe line search conditions (9) and (10) hold;*
- *Step 3: Let $x_{k+1} = x_k + \alpha_k d_k$, $g_{k+1} = g(x_{k+1})$, if $\|g_{k+1}\| \leq \varepsilon$, then stop;*
- *Step 4: Generate d_{k+1} by $d_{k+1} = -g_{k+1} + \beta_k^+ d_k + \delta_k y_k$ in which β_k^+, δ_k and t_k are determined by (40), (39) and (36) respectively.*
- *Step 5: Set $k := k+1$, go to step 2.*

The symbol $\|\cdot\|$ stands for the Euclidean norm in this paper.

3. Convergence analysis

In this section, we investigate the convergence properties of the presented Algorithm 1. In the rest parts of this paper, we assume that $g_k \neq 0$ for all k, otherwise a stationary point has been found. We also make the following basic assumptions on the objective function.

Assumption 1. 1. *$f(x)$ is bounded below on the level set $\Gamma = \{x \in R^n : f(x) \leq f(x_1)\}$, i.e., there exists a positive constant B such that for all $x \in \Gamma$, $\|x\| \leq B$.*

2. *In some neighborhood N of Γ, $f(x)$ is differentiable and its gradient $g(x)$ is Lipschitz continuous, namely, there exists a constant $L > 0$ such that*

$$\|g(x) - g(y)\| \leq L\|x - y\|, \ \ for \ \ all \ \ x, y \in N. \tag{41}$$

Under the above assumptions on $f(x)$, there exists a constant $\Theta \geq 0$ such that $\|g(x)\| \leq \Theta$ for all $x \in \Gamma$.

The descent property of search direction is critical in the convergence analysis for conjugate gradient method. The following proposition shows that the search direction generated by the proposed Algorithm 1 possesses sufficient descent property.

Lemma 1. *Suppose that d_k is generated by (37), (38) and (39) in which t_k is determined by (36), then the sufficient descent property holds for all $k \geq 1$, namely, there exists a positive constant c, such that*

$$-g_k^T d_k \geq c\|g_k\|^2, \quad for \ \ all \ \ k \geq 1. \tag{42}$$

Proof: From equations (37), (38) and (39), we have

$$\begin{aligned} d_{k+1}^T g_{k+1} &= -\|g_{k+1}\|^2 + 2t_k \frac{g_{k+1}^T y_k g_{k+1}^T s_k}{s_k^T y_k} - \frac{(g_{k+1}^T s_k)^2}{s_k^T y_k} \\ &\leq -(1 - t_k)\|g_{k+1}\|^2 - (\frac{g_{k+1}^T s_k}{s_k^T y_k})^2 [s_k^T y_k - t_k \|y_k\|^2]. \end{aligned} \tag{43}$$

Combining the equations (36) and (33), we have $t_k \leq \frac{1}{2}$. Inequation (43) indicates that the sufficient descent condition (42) holds for $c = \frac{1}{2}$.

It should be noticed that, the sufficient descent property of the search direction generated by the proposed Algorithm 1 is independent with the line search scheme, also, the objective function $f(x)$ is only required to be continuously differentiable.

The following lemma shows that, if the objective function satisfies the Assumption 1, and the step size α_k fullfills the Wolfe line search conditions (9) and (10), then for all $k \geq 1$, the step size α_k has a positive lower bound.

Lemma 2. *Suppose that d_k is generated by (37), (38) and (39) in which t_k is determined by (36), $f(x)$ satisfies Assumption 1, if the step size α_k fullfills the Wolfe conditions (9) and (10), then*

$$\alpha_k \geq \frac{(\sigma - 1) g_k^T d_k}{L\|d_k\|^2}, \tag{44}$$

where σ and L are positive constant in (10) and (41) respectively.

Proof: Based on Lemma 1, d_k is a descent direction, namely $d_k^T g_k < 0$. Combining with Lisschitz inequation (41), Wolfe condition (10) deduces

$$(\sigma - 1)g_k^T d_k \leq (g_{k+1} - g_k)^T d_k = y_k^T d_k \leq \|y_k\| \|d_k \mid \leq \alpha_k L \|d_k\|^2.$$

So, (44) holds immediately.

Zoutendijk condition [42] plays an important role in the analysis of global convergence for conjugate gradient method. In the following, we will prove that the proposed Algorithm 1 possesses the Zoutendijk condition.

Lemma 3. *Suppose that d_k is generated by (37), (38) and (39) in which t_k is determined by (36), in which step size α_k fullfills Wolfe conditions (9) and (10), if $f(x)$ satisfies the Assumption 1, then the following so-called Zoutendijk condition holds:*

$$\sum_{k \geq 1} \frac{(g_k^T d_k)^2}{\|d_k\|^2} < \infty. \tag{45}$$

Proof: Wolfe condition (9) means that

$$f(x_k) - f(x_k + \alpha_k d_k) \geq -\rho \alpha_k g_k^T d_k,$$

combining with (44), we have

$$f(x_k) - f(x_k + \alpha_k d_k) \geq \frac{\rho(1-\sigma)(g_k^T d_k)^2}{L\|d_k\|^2}. \tag{46}$$

By summing up both sides of (46), and using the bounded below assumption on $f(x)$, we can have zoutendijk condition (45) immediately.

For uniformly convex functions, i.e. there exists a constant $\mu > 0$ such that

$$(\nabla f(x) - \nabla f(y))^T (x - y) \geq \mu \|x - y\|^2, \tag{47}$$

for all $x, \quad y \in R^n$, we can prove that the norm of the directions $\{\|d_k\|\}$generated by Algorithm 1 is bounded above.

Lemma 4. *Suppose that d_k is generated by (37), (38) and (39) in which t_k is determined by (36), in which the step size α_k is determined by Wolfe line search (9) and (10).*

If the objective function $f(x)$ is uniformly convex, then the norm of $\|d_k\|$ is bounded above, namely, there exists $M > 0$ such that

$$\|d_k\| \le M, \tag{48}$$

holds for all $k \ge 1$.

Proof: Based on Lipschitz condition and uniformly convexity, we have

$$\|y_k\| \le L\|s_k\|, \quad y_k^T s_k \ge \mu\|s_k\|^2. \tag{49}$$

The sufficient descent condition $g_k^T d_k \le -c\|g_k\|^2$ indicates that the sequence $\{x_k\} \in \Gamma = \{x \in R^n : f(x) \le f(x_1)\}$. By Assumption 1, there exists a constant $\Theta \ge 0$ such that $\|g(x_k)\| \le \Theta$ holds for all $k \ge 1$. On the other hand, from the definition of d_{k+1}, (49) and (36), we have

$$\begin{aligned}
\|d_{k+1}\| &= \| - g_{k+1} + \frac{t_k g_{k+1}^T y_k - g_{k+1}^T s_k}{s_k^T y_k} s_k + t_k \frac{g_{k+1}^T s_k}{s_k^T y_k} y_k\| \\
&\le \|g_{k+1}\| + \frac{t_k\|g_{k+1}\|\|y_k\| + \|g_{k+1}\|\|s_k\|}{\mu\|s_k\|^2}\|s_k\| + t_k \frac{\|g_{k+1}\|\|s_k\|}{\mu\|s_k\|^2}\|y_k\| \\
&\le \|g_{k+1}\| + \frac{\frac{1}{2}L+1}{\mu}\|g_{k+1}\| + \frac{L}{2\mu}\|g_{k+1}\| \\
&\le (1 + \frac{L+1}{\mu})\Theta := M.
\end{aligned}$$

With Lemma 4, we can prove the following convergence results for uniformly convex function.

Theorem 1. *Assum that $f(x)$ satisfies assumption 1. Consider the search direction d_k generated by (37), (38) and (39) in which t_k is determined by (36), and α_k is calculated by Wolfe line search. If furthermore, $f(x)$ is uniformly convex, we have*

$$\lim_{k\to\infty} \|g_k\| = 0. \tag{50}$$

Proof: Based on Lemma 4, we have $\|d_k\| \le M$. According to Lemma 1, the sufficient descent condition $-g_k^T d_k \ge c\|g_k\|^2$ holds. By using zoutendijk condition (45) we have

$$\infty > \sum_{k\ge1} \frac{(g_k^T d_k)^2}{\|d_k\|^2} \ge \sum_{k\ge1} \frac{(g_k^T d_k)^2}{M^2} \ge \frac{c^2}{M^2} \sum_{k\ge1} \|g_k\|^2. \tag{51}$$

The above inequation deduces (50).

From Theorem 1, we know that, for uniformly convex function, the global convergence can be established without the nonnegative restriction on β_k given by (38). Now, we will discuss the global convergence properties for general functions. In order to prove the establish the global convergence, we need to make a nonnegative restriction on β_k as $\beta_k^+ = \max\{\beta_k, 0\}$, in which β_k is given by (38).

For general function, we can obtain a weaker convergence result in the sence that

$$\liminf_{k\to\infty} \|g_k\| = 0. \tag{52}$$

For this purpose, we are going to prove this convergence result (52) by contradiction. Suppose that (52) does not hold, which means that there exists a positive constant $\gamma > 0$ such that

$$\|g_k\| > \gamma, \quad for\ \ all\ \ k \geq 1. \tag{53}$$

Lemma 5. *Suppose that $f(x)$ satisfies Assumption 1. Consider the proposed Algorithm 1 in which d_{k+1} is generated by by $d_{k+1} = -g_{k+1} + \beta_k^+ d_k + \delta_k y_k$ in which β_k^+, δ_k and t_k are determined by (40), (39) and (36) respectively, step size α_k is calculated by Wolfe line search satisfying (9) and (10). If (53) holds, then,*

$$\sum_{k\geq 1} \|u_{k+1} - u_k\|^2 < \infty, \tag{54}$$

where $u_{k+1} = \frac{d_{k+1}}{\|d_{k+1}\|}$. Proof. Based on the sufficient descent condition 42, $d_{k+1} = 0$ implies $g_{k+1} = 0$ which contradicts with (53), so, u_{k+1} is well defined. From equation $d_{k+1} = -g_{k+1} + \beta_k^+ d_k + \delta_k y_k$, we have

$$\begin{aligned} \frac{d_{k+1}}{\|d_{k+1}\|} &= \frac{-g_{k+1}}{\|d_{k+1}\|} + \beta_k^+ \frac{d_k}{\|d_{k+1}\|} + \delta_k \frac{y_k}{\|d_{k+1}\|} \\ &= \frac{-g_{k+1} + \delta_k y_k}{\|d_{k+1}\|} + \beta_k^+ \frac{\|d_k\|}{\|d_{k+1}\|} \frac{d_k}{\|d_k\|}. \end{aligned} \tag{55}$$

Rewrite (55) as follows:

$$u_{k+1} = \omega_k + \eta_k u_k, \tag{56}$$

where

$$\omega_k = \frac{-g_{k+1} + \delta_k y_k}{\|d_{k+1}\|}, \tag{57}$$

$$\eta_k = \beta_k^+ \frac{\|d_k\|}{\|d_{k+1}\|} \geq 0. \tag{58}$$

Using the identity $\|u_{k+1}\| = \|u_k\| = 1$ *and (56), we obtain*

$$\|\omega_k\| = \|u_{k+1} - \eta_k u_k\| = \|\eta_k u_{k+1} - u_k\|. \tag{59}$$

Since $\eta_k \geq 0$, *triangle inequality and (59) imply that*

$$\begin{aligned} \|u_{k+1} - u_k\| &\leq \|(1+\eta_k)u_{k+1} - (1+\eta_k)u_k\| \\ &\leq \|u_{k+1} - \eta_k u_k\| + \|\eta_k u_{k+1} - u_k\| \\ &= 2\|\omega_k\|. \end{aligned} \tag{60}$$

By the definition of ω_k, δ_k *and* t_k *substituting (39) into (57), we have*

$$\begin{aligned} \|\omega_k\| &= \frac{\|-g_{k+1} + \delta_k y_k\|}{\|d_{k+1}\|} = \frac{\|-g_{k+1} + t_k \frac{g_{k+1}^T s_k}{s_k^T y_k} y_k\|}{\|d_{k+1}\|} \\ &\leq \frac{\|g_{k+1}\| + \frac{1}{1+a_k} \frac{\|g_{k+1}\|\|s_k\|\|y_k\|}{s_k^T y_k}}{\|d_{k+1}\|}. \end{aligned} \tag{61}$$

By using the definition of a_k *(31), (61) indicates*

$$\begin{aligned} \|\omega_k\| &\leq \frac{\|g_{k+1}\| + \|g_{k+1}\| \frac{(s_k^T y_k)^2}{(s_k^T y_k)^2 + \|s_k\|^2\|y_k\|^2} \frac{\|s_k\|\|y_k\|}{s_k^T y_k}}{\|d_{k+1}\|} \\ &\leq \frac{\|g_{k+1}\|(1 + \frac{\|s_k\|^2\|y_k\|^2}{(s_k^T y_k)^2 + \|s_k\|^2\|y_k\|^2})}{\|d_{k+1}\|} \\ &\leq \frac{2\|g_{k+1}\|}{\|d_{k+1}\|}. \end{aligned} \tag{62}$$

If (53) $\|g_{k+1}\| \geq \gamma$, *from the sufficient descent condition (53), and Zoutendijk condition (45), we have*

$$\infty > \sum_{k\geq 1} \frac{(g_{k+1}^T d_{k+1})^2}{\|d_{k+1}\|^2} \geq \sum_{k\geq 1} \frac{c^2\|g_{k+1}\|^4}{\|d_{k+1}\|^2} \geq \sum_{k\geq 1} \frac{c^2\gamma^2\|g_{k+1}\|^2}{\|d_{k+1}\|^2}. \tag{63}$$

Equations (60), (62) and (63) deduce (54).

The above Lemma 5 shows that the search directions u_{k+1} change slowly, asymptotically. To establish the global convergence for general functions, we need to require, in addition, that β_k be small when the step $s_k = x_{k+1} - x_k$ is small.

This property is firstly formally stated by Gilbert and Nocedal [43], and is widely used in the convergence analysis of the typical two-term conjugate gradient method, namely the method formed by (7) and (8). For three-term conjugate gradient method formed by (7) and (37), similar with Gilbert and Nocedal [43], we present this property as follows.

Property(*) 1. *Consider a method of the form (7) and (37), and suppose that*

$$0 < \gamma \leq \|g_k\| \leq \bar{\gamma}, \tag{64}$$

for all $k \geq 1$. Under this assumption, we say that the method has Property(*) if there exists constants $b > 1$ and $\lambda > 0$ such that for all k

$$|\beta_k| \leq b, \tag{65}$$

and

$$\|s_k\| \leq \lambda \Rightarrow |\beta_k| \leq \frac{1}{2b}. \tag{66}$$

For general conjugate gradient method with the form (7) and (8), it is known that many methods satisfy this property. In the following Lemma, we will show that the proposed three-term conjugate gradient method formed by (7) and (37) also possesses this property.

Lemma 6. *Consider the three-term conjugate gradient method form by (7) and (37), in which β_k, δ_k and t_k are defined by (38), (39) and (36) respectively, if the objective function satisfies Assumption 1 and step size α_k is determined by Wolfe line searches (9) and (10), then the method possesses Property (*).*

Proof. *By Wolfe line search condition (10) and the sufficient descent property 42, we have*

$$d_k^T y_k \geq (\sigma - 1) g_k^T d_k \geq c(1-\sigma)\|g_k\|^2. \tag{67}$$

Combining (67), Assumption 1 and (64) with the definition of β_k given by (37), we have

$$\begin{aligned}|\beta_k| &= |\frac{t_k g_{k+1}^T y_k - g_{k+1}^T s_k}{d_k^T y_k}| \\ &\le \frac{t_k\|g_{k+1}\|\|y_k\| + \|g_{k+1}\|\|s_k\|}{c(1-\sigma)\|g_k\|^2} \\ &\le \frac{\frac{1}{2}\|g_{k+1}\|\|g_{k+1}-g_k\| + \|g_{k+1}\|\|s_k\|}{c(1-\sigma)\|g_k\|^2} \\ &\le \frac{\bar{\gamma}^2 + \bar{\gamma}B}{c(1-\sigma)\gamma^2} := b.\end{aligned} \tag{68}$$

Define

$$\lambda := \frac{c^2(1-\sigma)^2\gamma^4}{2\bar{\gamma}^2(\bar{\gamma}+B)(\frac{L}{2}+1)}. \tag{69}$$

On the other hand, if $\|s_k\| \le \lambda$, from the second inequation of (68) and (69), we obtain

$$\begin{aligned}|\beta_k| &\le \frac{\frac{1}{2}L\|g_{k+1}\|s_k\| + \|g_{k+1}\|\|s_k\|}{c(1-\sigma)\|g_k\|^2} \\ &\le \frac{(\frac{1}{2}L\bar{\gamma}+\bar{\gamma})}{c(1-\sigma)\gamma^2}\|s_k\| \le \frac{(\frac{1}{2}L\bar{\gamma}+\bar{\gamma})}{c(1-\sigma)\gamma^2}\lambda \\ &= \frac{1}{2b}.\end{aligned} \tag{70}$$

Since the proposed three-term conjugate gradient method possesses Property(*), in the next lemma, we will show that if the gradients are bounded away from zero, then a fraction of the steps cannot be too small. Let N denote the set of positive integers. For $\lambda > 0$ let

$$K^\lambda := \{i \in N: \ i \ge 1, \ \|s_i\| > \lambda\}, \tag{71}$$

i.e., the set of integers corresponding to steps that are larger than λ. We will need to discuss groups of Δ consecutive iterates, for this purpose, let

$$K_{k,\Delta}^\lambda := \{i \in N: \ k \le i \le k+\Delta-1, \ \ \|s_i\| > \lambda\}. \tag{72}$$

Let $|K_{k,\Delta}^\lambda|$ denote the number of elements of $K_{k,\Delta}^\lambda$, and $\lfloor\cdot\rfloor$ denote floor operator.

Lemma 7. *Consider the three-term conjugate gradient method form by (7) and (37), in which β_k, δ_k and t_k are defined by (38), (39) and (36) respectively, if the objective function satisfies Assumption 1 and step size α_k is determined by Wolfe line searches*

(9) and (10). If (64) holds, then there exists $\lambda > 0$ such that, for any $\Delta \in N$ and any index k_0, there is a greater index $k \geq k_0$ such that

$$|K_{k,\Delta}^{\lambda}| > \frac{\Delta}{2}. \tag{73}$$

Proof. We prove by contradiction. Suppose that

$$\begin{cases} for\ \ any\ \ \lambda > 0,\ \ there\ \ exists\ \ \Delta \in N\ \ and\ \ k_0\ \ such\ \ that, \\ for\ \ any\ \ k \geq k_0,\ \ we\ \ have\ \ |K_{k,\Delta}^{\lambda}| \leq \frac{\Delta}{2}. \end{cases} \tag{74}$$

Based on Lemma 1 and Lemma 3, we have that the sufficient desent condition (42) and Zoutendijk condition (45) hold. From the definition of δ_k and t_k given by (39) and (36) respectively, we have

$$\begin{aligned} \|\delta_k y_k\| &= |t_k \frac{g_{k+1}^T s_k}{s_k^T y_k}| \|y_k\| \\ &\leq \frac{1}{1+a_k} |\frac{g_{k+1}^T s_k}{s_k^T y_k}| \|y_k\| \\ &= \frac{(s_k^T y_k)^2}{(s_k^T y_k)^2 + \|s_k\|^2 \|y_k\|^2} |\frac{g_{k+1}^T s_k}{s_k^T y_k}| \|y_k\| \\ &\leq \frac{\|y_k\|^2 \|s_k\|^2 \|g_{k+1}\|}{(s_k^T y_k)^2 + \|s_k\|^2 \|y_k\|^2} \leq \|g_{k+1}\|. \end{aligned} \tag{75}$$

Since the proposed method has Property(*), there exists $\lambda > 0$ and $b > 1$ such that (65) and (66) hold for all k. For this λ, let Δ and k_0 given by (74) For any given index $l \geq k_0 + 1$, from the definition of d_{k+1} given by (37), we have

$$\begin{aligned} \|d_{k+1}\|^2 &\leq (\beta_k \|d_k\| + \| - g_{k+1} + \delta_k y_k\|)^2 \\ &\leq 2\beta_k^2 \|d_k\|^2 + 2\| - g_{k+1} + \delta_k y_k\|^2 \\ &\leq 2\beta_k^2 \|d_k\|^2 + 2(2\|g_{k+1}\|^2 + 2\|\delta_k y_k\|^2), \end{aligned} \tag{76}$$

the above inequalities follow from the fact that, for any scalars a and b, we have $2ab \leq a^2 + b^2$, hence $(a+b)^2 \leq 2a^2 + 2b^2$. Equations (75) and (76) indicate that

$$\|d_{k+1}\|^2 \leq 2\beta_k^2 \|d_k\|^2 + 8\|g_{k+1}\|^2. \tag{77}$$

For any given index $l \geq k_0 + 1$, by induction, we have

$$\|d_l\|^2 \leq c_1(1 + 2\beta_{l-1}^2 + 2\beta_{l-1}^2 2\beta_{k-2}^2 + \cdots + 2\beta_{l-1}^2 2\beta_{k-2}^2 \cdots 2\beta_{k_0}^2), \tag{78}$$

where c_1 deponds on $\|d_{k_0-1}\|$, but not on the index l. Let us consider a typical term in (78):

$$2\beta_{l-1}^2 2\beta_{l-2}^2 \cdots 2\beta_k^2, \tag{79}$$

where $k_0 \leq k \leq l-1$. We now divide $2(l-k)$ factors of (78) into groups of 2Δ elements, i.e., if $\Lambda := \lfloor (l-k)/\Delta \rfloor$, then (78) can be divided into Λ or $\Lambda+1$ groups as follows:

$$(2\beta_{l_1}^2 \cdots 2\beta_{k_1}^2), \cdots, (2\beta_{l_\Lambda}^2 \cdots 2\beta_{k_\Lambda}^2), \tag{80}$$

and possibly

$$(2\beta_{l_{\Lambda+1}}^2 \cdots 2\beta_k^2), \tag{81}$$

where $l_i = l-1-(i-1)\Delta$, for $i = 1, 2, \cdots \Lambda+1$, and $k_i = l_{i+1}+1$, for $i = 1, 2, \cdots \Lambda$. Since $k_i \geq k_0$ for all $i = 1, 2, \cdots, \Lambda$, so that we can apply relationship (74) for $k = k_i$. Thus we have

$$p_i := |K_{k_i,\Delta}^{\lambda}| \leq \frac{\Delta}{2}. \tag{82}$$

Which means that in the range $[k_i, k_i+\Delta-1]$ there are p_i indices j such that $\|s_j\| > \lambda$, and $(\Delta - p_i)$ indices with $\|s_j\| \leq \lambda$. Using this fact, (65) and (66), for a typical factor in (80), we have

$$\begin{aligned} 2\beta_{l_i}^2 \cdots 2\beta_{k_i}^2 &\leq 2^{\Delta} b^{2p_i} (\frac{1}{2b})^{2(\Delta-p_i)} \\ &= (2b^2)^{2p_i-\Delta} \leq 1, \end{aligned} \tag{83}$$

since by (82), $2p_i - \Delta \leq 0$ and $2b^2 \geq 1$. So, each of the factors in (80) is less or equal to 1, and so is their product. For the last part given in (81), by simply using (65), we have

$$2\beta_{l_{\Lambda+1}}^2 \cdots 2\beta_k^2 \leq (2b^2)^{\Delta}.$$

So, it is obviously that each term on the right-hand side of (78) is bounded by $(2b^2)^{\Delta}$, and as a resut we obtain

$$\|d_l\|^2 \leq c_2(l-k_0+2), \tag{84}$$

for a certain positive constant c_2 independent of l. (84) shows that $\|d_k\|^2$ grows at most linearly, which also indicates

$$\sum_{k\geq 1} \frac{1}{\|d_k\|^2} = \infty. \tag{85}$$

On the other hand, from Zoutendijk condition (45), sufficient descent condition (42) and (64), we have

$$c\gamma^4 \sum_{k\geq 1} \frac{1}{\|d_k\|^2} \leq c \sum_{k\geq 1} \frac{\|g_k\|^4}{\|d_k\|^2} \leq \sum_{k\geq 1} \frac{(g_k^T d_k)^2}{\|d_k\|^2} \leq \infty.$$

this contraicts (85), concluding the proof.

Theorem 2. *Suppose that $f(x)$ satisfies Assumption 1. Consider the proposed Algorithm 1 in which d_{k+1} is generated by by $d_{k+1} = -g_{k+1} + \beta_k^+ d_k + \delta_k y_k$ in which β_k^+, δ_k and t_k are determined by (40), (39) and (36) respectively, step size α_k is calculated by Wolfe line search satisfying (9) and (10). Then the method converges in the sense (52).*

Proof. We proceed by contradiction. Assume that (52) does not hold, this means that the condition (64) holds. Therefore, the conditions of Lemmas 5, 6 and 7 hold. Combining with Assumption 1, we can obtain a contradiction silimarly to the proof of the Theorem 4.3 in [43].

4. Numerical experiments

In this section, we investigate the numerical performance of the proposed algorithm 1 (NTPA). Based on (37), (38) and (39), the proposed algorithm NTPA can be considered as a special three-term conjugate gradient method which has similar structure with THREECG method [31] and TTCG method [30]. So, in this paper, we will compare the numerical performances of the following different methods: NTPA, THREECG and TTCG methods. THREECG and TTCG methods are proposed by Andrei [31, 30] in which the directions are generated by

$$d_{k+1} = -g_{k+1} + \delta_k s_k - \eta_k y_k,$$
$$\delta_k = \frac{g_{k+1}^T y_k - \omega g_{k+1}^T s_k}{s_k^T y_k}, \quad \eta_k = \frac{g_{k+1}^T s_k}{s_k^T y_k}.$$

In THREECG, $\omega = 1 + \frac{\|y_k\|^2}{s_k^T y_k}$, in TTCG, $\omega = 1 + 2\frac{\|y_k\|^2}{s_k^T y_k}$. In this test, the code was downloaded at https://camo.ici.ro/neculai/THREECG/threecg.for, which was written by Andrei and widely used in conjugate gradient method numerical test. 75 unconstrained test problems are selected for comparison which are in the generalized or extended form in [35]. For each test problem, the numerical experiments are carried out with the number of variables increasing as n=1000, 2000, . . . ,10000 which are the same with [31, 30]. All the default values of the parameters involved in the methods are the same with [31, 30]: The Wolfe line search is implemented with $\rho = 0.0001$ and $\sigma = 0.8$, stopping criterion is $\|g_k\|_\infty \leq 10^{-6}$ and the maximum number of iterations is limited to 10000, etc.

The comparing data contain iterations, function evaluations and CPU time. To approximately assess the performance of different algorithms, we use the performance profile introduced by Dolan and Moré [34] as an evaluated tool.

Dolan and Moré [34] gave a new tool to analyze the efficiency of Algorithms. They introduced the notion of a performance profile as a means to evaluate and compare the performance of the set of solvers S on a test set P. Assuming that there exists n_s solvers and n_p problems, for each problem p and solver s, they defined:

$t_{p,s}$= computing cost required to solve problem p by solver s.

Requiring a baseline for comparisons, they compared the performance on problem p by solver s with the best performance by any solver on this problem; that is, using the performance ratio:

$$r_{p,s} = \frac{t_{p,s}}{\min\{t_{p,s} : s \in S\}}. \tag{86}$$

Then they defined

$$\rho_s(\tau) = \frac{1}{n_p} size\{p \in P : r_{p,s} \leq \tau\}, \tag{87}$$

thus $\rho_s(\tau)$ is the probability for solver s that a performance ratio $r_{p,s}$ is within a factor $\tau \geq 1$ of the best possible ratio. Then function ρ_s is the distribution function for the performance ratio. The performance profile ρ_s is a nondecreasing, piecewise constant function. That is, for subset of the methods being analyzed, we plot the fraction P of the problems for which any given method is within a factor τ of the best.

Figure 1 shows the performance profile with respect to the number of iterations. From Figure 1, we can find that NTPA method solves about 70% of the test problems with the least value of iteration. But with the factor τ increasing, THREECG method outperforms NTPA and TTCG methods. Figure 2 gives the profile with respect to function evaluations. Based on Figure 2, we can also find that NTPA method solves about 73% of all problems with the least value of function evaluations, THREECG

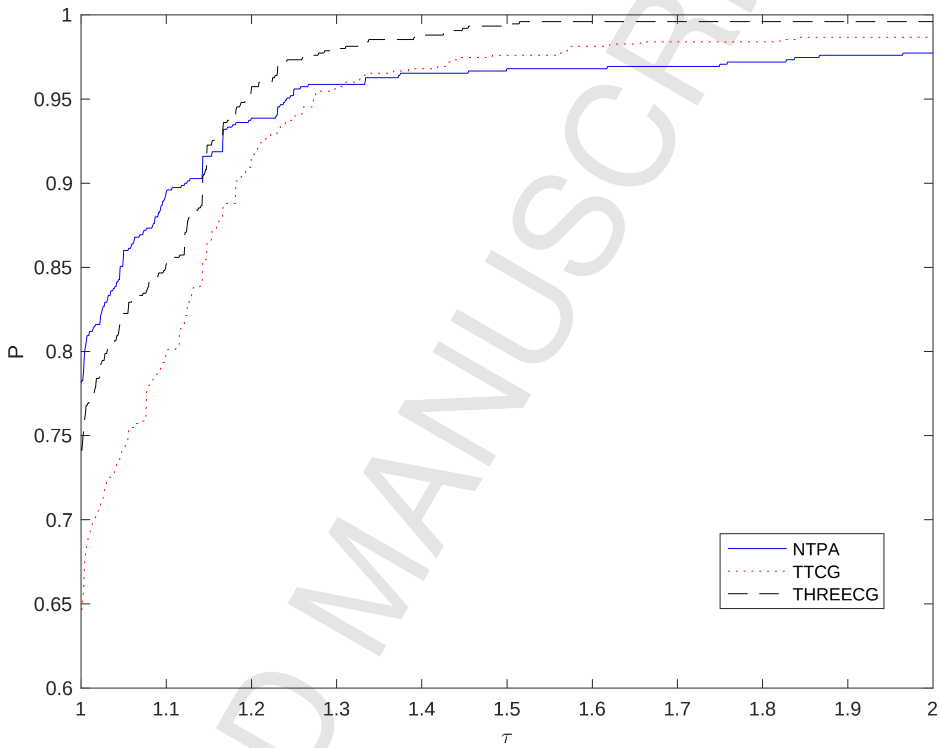

Figure 1: Performance profile based on iterations

method solves about 65% and TTCG method solves about 57% with the least value. Also, with the factor τ increasing, THREECG method outperforms NTPA and TTCG methods. Figures 1 and 2 indicate all three methods perform similarly with respect to the number of iterations and function evaluations. Figure 3 presents the profile with respect to cup time. From Figure 3, NTPA method outperforms THREECG and TTCG methods, which means that NTPA method is very efficient in solving unconstrained optimization problems.

References

[1] Y. Nesterov, Subgradient methods for huge-scale optimization problems, Mathematical Programming 146 (1-2) (2014) 275–297.

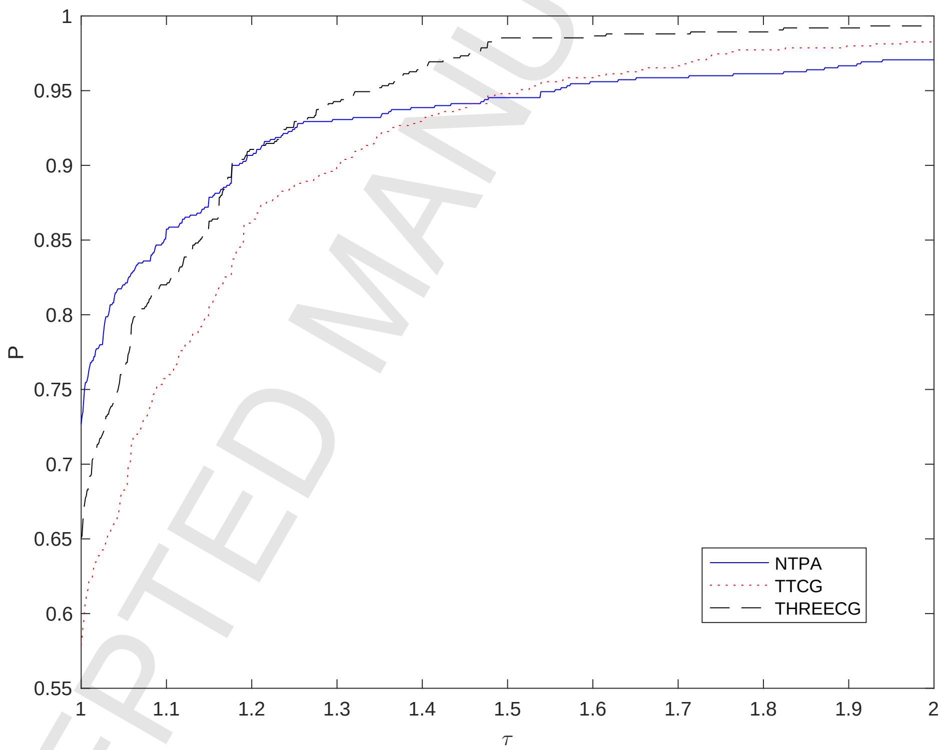

Figure 2: Performance profile based on function evaluations

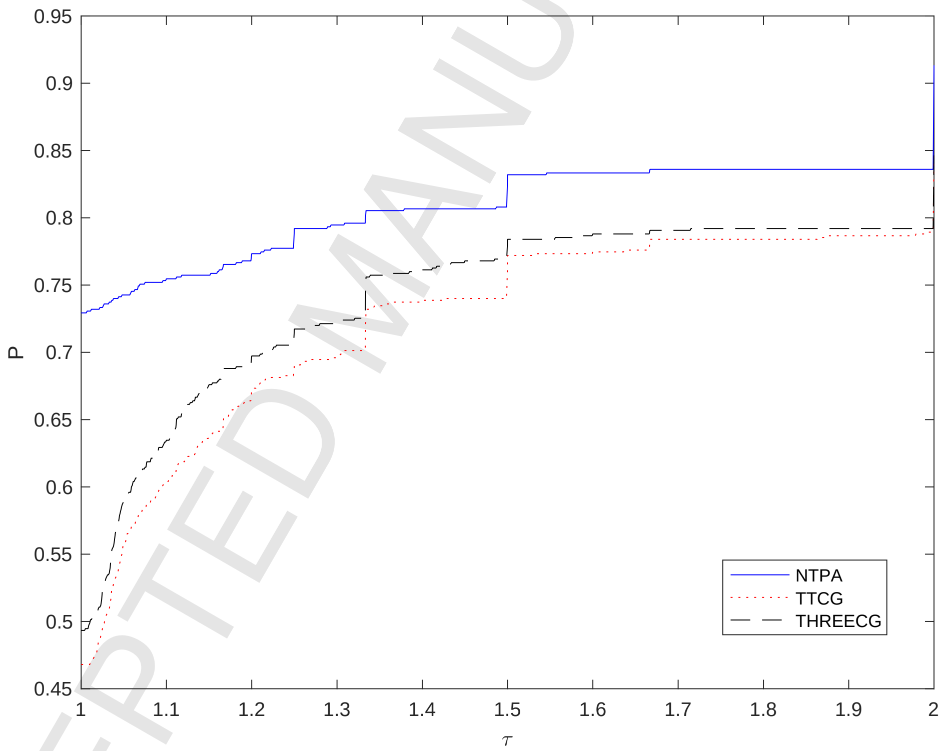

Figure 3: Performance profile based on cpu time

[2] Y. Nesterov, A method of solving a convex programming problem with convergence rate o (1/k2) 27 (2) (1983) 372–376.

[3] C. C. Gonzaga, E. W. Karas, Fine tuning Nesterov's steepest descent algorithm for differentiable convex programming, Mathematical Programming 138 (1-2) (2013) 141–166.

[4] B. O'Donoghue, E. Candés, Adaptive restart for accelerated gradient schemes, Foundations of computational mathematics 15 (3) (2015) 715–732.

[5] B. T. Polyak, Some methods of speeding up the convergence of iteration methods, USSR Computational Mathematics and Mathematical Physics 4 (5) (1964) 1–17.

[6] J. Nocedal, S. Wright, Numerical optimization, Springer Science & Business Media, 2006.

[7] W. Sun, Y.-X. Yuan, Optimization theory and methods: nonlinear programming, Vol. 1, Springer Science & Business Media, 2006.

[8] P. Wolfe, Convergence conditions for ascent methods, SIAM review 11 (2) (1969) 226–235.

[9] P. Wolfe, Convergence conditions for ascent methods. ii: Some corrections, SIAM review 13 (2) (1971) 185–188.

[10] R. Fletcher, C. M. Reeves, Function minimization by conjugate gradients, The computer journal 7 (2) (1964) 149–154.

[11] M. R. Hestenes, E. Stiefel, Methods of conjugate gradients for solving linear systems, Vol. 49, NBS, 1952.

[12] E. Polak, G. Ribiere, Note sur la convergence de méthodes de directions conjuguées, Revue française d'informatique et de recherche opérationnelle, série rouge 3 (1) (1969) 35–43.

[13] B. T. Polyak, The conjugate gradient method in extremal problems, USSR Computational Mathematics and Mathematical Physics 9 (4) (1969) 94–112.

[14] Y.-H. Dai, Y. Yuan, A nonlinear conjugate gradient method with a strong global convergence property, SIAM Journal on Optimization 10 (1) (1999) 177–182.

[15] Y.-H. Dai, L.-Z. Liao, New conjugacy conditions and related nonlinear conjugate gradient methods, Applied Mathematics and Optimization 43 (1) (2001) 87–101.

[16] W. W. Hager, H. Zhang, A new conjugate gradient method with guaranteed descent and an efficient line search, SIAM Journal on Optimization 16 (1) (2005) 170–192.

[17] Y.-H. Dai, C.-X. Kou, A nonlinear conjugate gradient algorithm with an optimal property and an improved wolfe line search, SIAM Journal on Optimization 23 (1) (2013) 296–320.

[18] C. Kou, An improved nonlinear conjugate gradient method with an optimal property, Science China Mathematics 57 (3) (2014) 635–648.

[19] D.-H. Li, M. Fukushima, A modified bfgs method and its global convergence in nonconvex minimization, Journal of Computational and Applied Mathematics 129 (1) (2001) 15–35.

[20] Z. Wei, G. Li, L. Qi, New quasi-newton methods for unconstrained optimization problems, Applied Mathematics and Computation 175 (2) (2006) 1156–1188.

[21] J. Zhang, C. Xu, Properties and numerical performance of quasi-newton methods with modified quasi-newton equations, Journal of Computational and Applied Mathematics 137 (2) (2001) 269–278.

[22] G. Li, C. Tang, Z. Wei, New conjugacy condition and related new conjugate gradient methods for unconstrained optimization, Journal of Computational and Applied Mathematics 202 (2) (2007) 523–539.

[23] H. Yabe, M. Takano, Global convergence properties of nonlinear conjugate gradient methods with modified secant condition, Computational Optimization and Applications 28 (2) (2004) 203–225.

[24] W. Zhou, L. Zhang, A nonlinear conjugate gradient method based on the mbfgs secant condition, Optimisation Methods and Software 21 (5) (2006) 707–714.

[25] S. Babaie-Kafaki, R. Ghanbari, The dai–liao nonlinear conjugate gradient method with optimal parameter choices, European Journal of Operational Research 234 (3) (2014) 625–630.

[26] M. Fatemi, An optimal parameter for dai–liao family of conjugate gradient methods, Journal of Optimization Theory and Applications 169 (2) (2016) 587–605.

[27] S. Babaie-Kafaki, R. Ghanbari, Two optimal dai–liao conjugate gradient methods, Optimization 64 (11) (2015) 2277–2287.

[28] A. Perry, Technical note a modified conjugate gradient algorithm, Operations Research 26 (6) (1978) 1073–1078.

[29] S. Babaie-Kafaki, R. Ghanbari, A descent family of dai–liao conjugate gradient methods, Optimization Methods and Software 29 (3) (2014) 583–591.

[30] N. Andrei, On three term conjugate gradient algorithms for unconstrained optimization, Applied Mathematics and Computation 219 (11) (2013) 6316–6327.

[31] N. Andrei, A simple three-term conjugate gradient algorithm for unconstrained optimization, Journal of Computational and Applied Mathematics 241 (2013) 19–29.

[32] N. Andrei, A new three-term conjugate gradient algorithm for unconstrained optimization, Numerical Algorithms 68 (2) (2015) 305–321.

[33] N. Andrei, Accelerated adaptive Perry conjugate gradient algorithms based on the self-scaling memoryless BFGS update, Journal of Computational and Applied Mathematics, 325 (2017) pp.149-164.

[34] E. D. Dolan, J. J. Moré, Benchmarking optimization software with performance profiles, Mathematical programming 91 (2) (2002) 201–213.

[35] N. Andrei, An unconstrained optimization test functions collection, Adv. Model. Optim 10 (1) (2008) 147–161.

[36] C. G. Broyden, The convergence of a class of double-rank minimization algorithms 1. general considerations, IMA Journal of Applied Mathematics 6 (1) (1970) 76–90.

[37] R. Fletcher, A new approach to variable metric algorithms, The computer journal 13 (3) (1970) 317–322.

[38] D. Goldfarb, A family of variable-metric methods derived by variational means, Mathematics of computation 24 (109) (1970) 23–26.

[39] D. F. Shanno, Conditioning of quasi-newton methods for function minimization, Mathematics of computation 24 (111) (1970) 647–656.

[40] J. Perry, et al., A class of conjugate gradient algorithms with a two-step variable-metric memory, Discussion Papers 269.

[41] D. Shanno, On the convergence of a new conjugate gradient algorithm, SIAM Journal on Numerical Analysis 15 (6) (1978) 1247–1257.

[42] G. Zoutendijk, Nonlinear programming, computational methods, Integer and nonlinear programming 143 (1) (1970) 37–86.

[43] J. C. Gilbert, J. Nocedal, Global convergence properties of conjugate gradient methods for optimization, SIAM Journal on optimization 2 (1) (1992) 21–42.