Journal name
(will be inserted by the editor)

Exploiting damped techniques for nonlinear conjugate gradient methods

Mehiddin Al-Baali · Andrea Caliciotti · Giovanni Fasano · Massimo Roma

May 24, 2017

Abstract In this paper we propose the use of damped techniques within Nonlinear Conjugate Gradient (NCG) methods. Damped techniques were introduced by Powell and recently reproposed by Al-Baali and till now, only applied in the framework of quasi–Newton methods. We extend their use to NCG methods in large scale unconstrained optimization, aiming at possibly improving the efficiency and the robustness of the latter methods, especially when solving difficult problems. We consider both unpreconditioned and Preconditioned NCG (PNCG). In the latter case, we embed damped techniques within a class of preconditioners based on quasi–Newton updates. Our purpose is to possibly provide efficient preconditioners which approximate, in some sense, the inverse of the Hessian matrix, while still preserving information provided by the secant equation or some of its modifications. The results of an extensive numerical experience highlights that the proposed approach is quite promising.

Mehiddin Al-Baali
Department of Mathematics and Statistics
Sultan Qaboos University, P.O. Box 36, Muscat 123, Oman
E-mail: albaali@squ.edu.om

Andrea Caliciotti
Dipartimento di Ingegneria Informatica, Automatica e Gestionale "A. Ruberti"
SAPIENZA, Università di Roma; via Ariosto, 25 – 00185 Roma, Italy
E-mail: caliciotti@dis.uniroma1.it

Giovanni Fasano
Department of Management
University Ca' Foscari of Venice; S. Giobbe, Cannaregio 873 – 30121 Venice, Italy
E-mail: fasano@unive.it

Massimo Roma
Dipartimento di Ingegneria Informatica, Automatica e Gestionale "A. Ruberti"
SAPIENZA, Università di Roma; via Ariosto, 25 – 00185 Roma, Italy
E-mail: roma@dis.uniroma1.it

Keywords Large scale unconstrained optimization, Nonlinear Conjugate Gradient methods, quasi-Newton updates, damped techniques.

1 Introduction

In this paper we consider the large scale unconstrained optimization problem

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

where $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ is a real valued function and the dimension n is large. It is assumed that f is a twice continuously differentiable function and that for a given $x_1 \in \mathbb{R}^n$ the level set $\mathcal{L}_1 = \{x \in \mathbb{R}^n \ : \ f(x) \leq f(x_1)\}$ is compact.

Nowadays, the Nonlinear Conjugate Gradient (NCG) and the Limited Memory quasi–Newton methods are usually considered effective iterative methods for large scale unconstrained optimization. In particular, L–BFGS is typically the method of choice due to its efficiency (see e.g. [27]). However, on nonlinear "difficult" problems where the Hessian matrix is possibly highly ill–conditioned, also quasi–Newton methods may be inefficient. This was already known since 1986 when Powell in [29] analyzed the performance of the BFGS and DFP algorithms in the case of a quadratic function of two variables (see also [7]).

To overcome the latter drawback, in this paper we aim at introducing in the framework of NCG methods a technique originated by Powell in [28], and recently reproposed by Al-Baali in [2] for quasi–Newton methods: the so called "*damped technique*". To the best of our knowledge, the first damped update was defined in [28]. In that paper, Powell deals with SQP Lagrangian BFGS methods for nonlinearly constrained problems. He proposes a modification of BFGS that, to some extent, offsets the lack of positive definiteness in the Hessian of the Lagrangian at the solution. Indeed, due to the presence of negative curvature directions of the Lagrangian function, using BFGS for approximating the Hessian matrix with a positive definite matrix, may be seriously inappropriate (see also [27] Section 18.3). Damped techniques have been recently extended by Al-Baali also to the restricted Broyden class of quasi–Newton methods for unconstrained optimization problems in [2]. The author extends the global and superlinear convergence properties that the Broyden family of methods fulfills for convex functions, to a novel class of methods, namely the D–Broyden class (see also [4] [5] [6] [7] [8]).

We aim at extending the use of damped techniques to both NCG and PNCG methods. To this purpose, the following possibilities can be considered:

- *Modified* methods. In this case a damped technique is only used to modify the scalar (usually denoted by β_k) which characterizes the different NCG methods. The search direction is therefore modified, hence the necessity to ensure the global convergence of the resulting novel NCG method, the damped one.

- *Preconditioned* methods. They are obtained without modifying the original expression of the scalar β_k. Here, the damped techniques are only used for constructing a preconditioner based on quasi–Newton updates. In this case we do not obtain a novel NCG algorithm or focus on a particular NCG method. On the contrary, we have a new methodology for defining preconditioning strategies, to be possibly used within any NCG method for improving its performance.
- *Modified preconditioned* methods. In this case, a damped technique is used both to modify the scalar β_k and to construct a suitable preconditioner for NCG schemes.

We deal with all the three items above, even if the main focus is actually on the second one. Indeed, we believe that, since damped techniques were conceived in the framework of quasi–Newton methods, we expect to inherit their good features when building a preconditioner based on quasi–Newton updates. To this aim, we introduce two different damping strategies, which seem to be suited for our purposes. In particular, we focus on Polak–Ribière (PR) (recently, Polak–Ribière–Polyak (PRP)) method, proving that under reasonable assumptions, the damped and preconditioned version of this method (denoted by D-PR-PNCG) to some extent retains the convergence properties of the (undamped and unpreconditioned) PR method.

We propose to combine damped techniques with preconditioning strategies, aiming at making the resulting D-PR-PNCG method able to efficiently tackle also difficult problems. To this aim, in order to perform extensive numerical results, we consider a novel class of preconditioners based on quasi–Newton updates, which has been recently introduced in [12]. These preconditioners are based on novel low-rank quasi-Newton symmetric updating formulae, resulting as by-product of the NCG method at some previous steps. Hence, the construction of this class of preconditioners is matrix–free and iteratively defined. Their purpose is to approximate, in some sense, the inverse of the Hessian matrix, while still retaining efficiency and preserving information provided by the secant equation or some of its modifications. The rationale behind the idea of adopting a damped strategy in defining preconditioners for NCG methods, relies on the fact that an approximation of the (inverse of the) Hessian matrix by means of a positive definite matrix is required. Therefore, modifying the quasi–Newton updates used for building a preconditioner for NCG methods, in order to prevent the lack of positive definiteness of the Hessian matrix, sounds meaningful. An extensive numerical experience confirmed this fact, showing the possible fruitful use of the damped techniques in constructing preconditioners for NCG, based on quasi–Newton updates. For the sake of completeness, we also report results obtained by using the *modified* methods, showing that these latter do not seem to produce noticeable improvement in terms of efficiency and robustness.

The paper is organized as follows. In Section 2, we recall the PNCG methods and briefly describe the original damped techniques. Moreover, we report the class of preconditioners we adopt. Section 3 describes the novel damped

strategies we propose and some adaptive criteria used. In Section 4 we study the global convergence of one *modified preconditioned* method, the damped Polak–Ribière method (D-PR-PNCG). In Section 5 the results of an extensive numerical testing by using the class of preconditioners proposed in [12] are described. Finally, Section 6 includes some concluding remarks and some guidelines for future works.

As regards the notations, given a sequence of points $\{x_k\}$, $x_k \in \mathbb{R}^n$, we denote by $g_k = g(x_k) = \nabla f(x_k)$ and $f_k = f(x_k)$, the gradient and the function value at x_k, respectively. Similarly, $M_k = M(x_k)$ indicates that the matrix M_k depends on the iterate x_k. Moreover, we use the following standard settings:

$$y_k = g_{k+1} - g_k, \qquad s_k = x_{k+1} - x_k. \tag{1.2}$$

Finally, $A \succ 0$ indicates that the matrix A is positive definite, $\|x\|$ denotes the 2-norm of the vector $x \in \mathbb{R}^n$, and $\lambda_m(A)$, $\lambda_M(A)$ stand for the smallest and the largest eigenvalue of A, respectively.

2 Preliminaries

In this section we report some basics which we will use in the sequel. In particular, first we recall a general scheme of a PNCG algorithm. Afterwards, we report the damped strategies introduced in literature in quasi–Newton frameworks. Finally, we detail the preconditioning strategy we adopt.

2.1 The Preconditioned Nonlinear Conjugate Gradient (PNCG) algorithm

A PNCG algorithm can be outlined in the following standard scheme (see e.g. [30]), where $M_k \succ 0$ denotes the preconditioner at the k–th iteration.

Preconditioned Nonlinear Conjugate Gradient algorithm

Step 1: Set $x_1 \in \mathbb{R}^n$ and $M_1 \succ 0$. Set $p_1 = -M_1 g_1$ and $k = 1$.
If a *stopping criterion* is satisfied then stop.

Step 2: Compute the steplength α_k by using a linesearch procedure, which ensures the *strong Wolfe conditions*, and compute

$$x_{k+1} = x_k + \alpha_k p_k.$$

Step 3: If a *stopping criterion* is satisfied then stop,
else compute a scalar β_k and a new search direction

$$p_{k+1} = -M_{k+1} g_{k+1} + \beta_k p_k, \tag{2.3}$$

set $k := k + 1$ and go to *Step 2*.

By setting $M_k = I$ for all k, the popular (unpreconditioned) NCG method is trivially obtained. In practice, the explicit expression of M_{k+1} that we adopt in (2.3) is specified later on in (2.9). As is well known, several expressions for the parameter β_k have been proposed in literature. We recall the classical ones (Fletcher–Reeves, Polak–Ribière, Hestenes–Stiefel)

$$\beta_k^{\mathrm{FR}} = \frac{g_{k+1}^\top M_{k+1} g_{k+1}}{g_k^\top M_k g_k}, \quad \beta_k^{\mathrm{PR}} = \frac{y_k^\top M_{k+1} g_{k+1}}{g_k^\top M_k g_k}, \quad \beta_k^{\mathrm{HS}} = \frac{y_k^\top M_{k+1} g_{k+1}}{y_k^\top p_k}, \tag{2.4}$$

respectively, and among the many others, the recent proposal by Hager and Zhang [23]

$$\beta_k^{\mathrm{HZ}} = \frac{y_k^\top M_{k+1} g_{k+1}}{p_k^\top y_k} - \Theta_k \frac{y_k^\top M_{k+1} y_k}{p_k^\top y_k} \frac{p_k^\top g_{k+1}}{p_k^\top y_k},$$

where Θ_k is a suitable parameter.

2.2 Basics on damped techniques

Damped techniques were introduced in the framework of quasi–Newton methods and the rationale behind these techniques is the following. As is well known (see e.g. [16], [27]), in dealing with the BFGS update, a crucial issue in order to guarantee positive definiteness of the updated Hessian approximation is the curvature condition

$$s_k^\top y_k > 0. \tag{2.5}$$

If f is strongly convex on an open set containing $\mathcal{L}_1$, then (2.5) holds for any two points x_k and x_{k+1} belonging to $\mathcal{L}_1$ (see, e.g. [9]). In case of nonconvex functions, the satisfaction of condition (2.5) must be ensured by means of the linesearch procedure used for determining the stepsize α_k. Indeed, the satisfaction of (2.5) can be always obtained by a linesearch procedure if the objective function is bounded below on $\mathcal{L}_1$. To this aim the Wolfe conditions (in practice, strong Wolfe conditions) are usually adopted, which ensure condition (2.5). However, if the linesearch is not fairly accurate, the value of $s_k^\top y_k$ may not be sufficiently positive. In addition, if only the backtracking linesearch framework is employed, the curvature condition (2.5) may not hold.

A first possible strategy to cope with this issue is to reinitialize the model Hessian to the identity matrix or skip the update whenever $s_k^\top y_k \leq 0$ (see e.g. Section 4.2.2 of [24]). However, this strategy is usually not recommended, due to the loss of information on the curvature of the function. A more successful strategy is the *damped technique* proposed by Powell in [28], in the context of SQP Lagrangian BFGS method for constrained optimization, for which (2.5) may not hold even when the Wolfe conditions are employed. To overcome this difficulty, the author proposes to modify the difference of the gradients vector y_k in (1.2) before performing the update. Namely, on denoting by B_k

the available positive definite Hessian approximation at k-th iteration of the method, the following modified (damped) vector is used:

$$\widehat{y}_k = \varphi_k y_k + (1-\varphi_k) B_k s_k, \tag{2.6}$$

where φ_k is chosen in $(0,1]$ such that $s_k^\top \widehat{y}_k$ is "sufficiently positive". Namely, given $\sigma \in (0,1]$, the value of φ_k is set as follows:

$$\varphi_k = \begin{cases} \dfrac{\sigma s_k^\top B_k s_k}{s_k^\top B_k s_k - s_k^\top y_k}, & \text{if} \quad s_k^\top y_k < (1-\sigma) s_k^\top B_k s_k, \\ \\ 1, & \text{otherwise}, \end{cases} \tag{2.7}$$

(see also Section 18.3 in [27]). This choice ensures that $s_k^\top \widehat{y}_k = (1-\sigma) s_k^\top B_k s_k$ which is sufficiently positive, since B_k is imposed to be positive definite at each iteration. In [28] the value of $\sigma = 0.8$ is suggested as a "suitable size" to be used in (2.7) (see also [27]); the value of $\sigma = 0.9$ is sometimes used, too (see e.g. [2] [7]).

To the best of our knowledge, Powell's damped technique was never applied to unconstrained optimization problems until Al-Baali used it for improving the performance of the standard BFGS and DFP methods (see [2] [4] [5] [6] [7] [8]). In particular, the author extends the damped technique to the Broyden family of quasi–Newton methods for unconstrained optimization. The choice given in (2.7) is modified so that the damped vector $\widehat{y}_k$ replaces y_k in the quasi-Newton updating formulae whenever the ratio $s_k^\top y_k / s_k^\top B_k s_k$ is sufficiently close to zero or negative (like in the Powell's strategy). This choice enforces both global and superlinear convergence properties of the novel class of methods proposed in [2], namely the *D–Broyden* class. We note that (2.7) does not modify y_k when $s_k^\top y_k / s_k^\top B_k s_k$ is larger than 1. Therefore, Al-Baali also suggests using the modified damped vector (2.6) when the ratio $s_k^\top y_k / s_k^\top B_k s_k$ is large enough by extending the above choice as follows:

$$\varphi_k = \begin{cases} \dfrac{\sigma s_k^\top B_k s_k}{s_k^\top B_k s_k - s_k^\top y_k}, & \text{if} \quad s_k^\top y_k < (1-\sigma) s_k^\top B_k s_k, \\ \\ \dfrac{\hat{\sigma} s_k^\top B_k s_k}{s_k^\top B_k s_k - s_k^\top y_k}, & \text{if } s_k^\top y_k > (1+\hat{\sigma}) s_k^\top B_k s_k, \\ \\ 1, & \text{otherwise}, \end{cases} \tag{2.8}$$

where $\hat{\sigma} \geq 2$. In this paper, we consider the value of $\hat{\sigma} = \infty$ which reduces the above choice to (2.7).

2.3 The class of preconditioners

The preconditioners we adopt belong to the class of preconditioners proposed in [12]. They are based on low–rank (quasi–Newton) updates and they approximate, in some sense, the inverse of the Hessian matrix. It has been shown that their application leads to an improvement of the performance of an NCG algorithm, both in terms of efficiency and robustness. For all the details we refer to [12] and here we just report the expression of the preconditioners along with some comments. The preconditioners can be written as follows:

$$M_{k+1} = \tau_k C_k + \gamma_k v_k v_k^\top + \omega_k \sum_{j=k-m}^{k} \frac{s_j s_j^\top}{y_j^\top s_j}, \tag{2.9}$$

with

$$C_k = \frac{s_k^\top y_k}{\|y_k\|^2} I_n, \qquad v_k = s_k - \tau_k \frac{s_k^\top y_k}{\|y_k\|^2} y_k - \omega_k \sum_{j=k-m}^{k} \frac{s_j^\top y_k}{y_j^\top s_j} s_j,$$

$$\gamma_k = \frac{1}{(1-\tau_k) s_k^\top y_k - \omega_k \displaystyle\sum_{j=k-m}^{k} \frac{(s_j^\top y_k)^2}{y_j^\top s_j}},$$

and where the following parameters are used in practice:

$$m = 4, \quad \omega_k = \frac{\frac{1}{2} s_k^\top y_k}{y_k^\top C_k y_k + \displaystyle\sum_{j=k-m}^{k} \frac{(s_j^\top y_k)^2}{s_j^\top y_j}}, \quad \tau_k = \omega_k, \quad \gamma_k = \frac{2}{s_k^\top y_k}. \tag{2.10}$$

In (2.9) the term $\gamma_k v_k v_k^\top$ represents a rank–1 matrix, while the rightmost term is aimed at building an approximate inverse of the Hessian matrix on a specific subspace. The integer m can be viewed as a "limited memory" parameter, similarly to the L–BFGS method. The matrix C_k, the vector v_k and the parameters τ_k, γ_k and ω_k are such that the preconditioners are positive definite, and satisfy the secant equation at the current iterate, namely $M_{k+1} y_k = s_k$, along with a modified secant equation at some previous iterates (see e.g. [10] [11]). In [12], besides some theoretical properties, the results of an extensive numerical experience is reported, showing that the use of such preconditioners makes PNCG algorithms more efficient and robust than the unpreconditioned ones, on most CUTEst [18] large scale problems.

These preconditioners are also inspired by some recent proposals in the context of Newton–Krylov methods (see [14] [15]), along with some effective preconditioning techniques from the literature of preconditioners for symmetric linear systems, namely the Limited Memory Preconditioners [19].

3 Novel damped strategies

In this section we introduce two novel damped strategies, to be considered within NCG methods, along with an adaptive criterion for deciding if it is worth to replace the undamped vector with the damped one. In the sequel, whenever we consider the preconditioned case, we refer to a positive definite preconditioner based on quasi–Newton updates, which will be denoted by $P_k(y_k, s_k)$ to evidence the current pair (y_k, s_k) used for constructing the quasi–Newton update.

Drawing inspiration from the Al-Baali–Powell proposals briefly described in Section 2, now we aim at defining modifications of the vector y_k which should lead to obtain more efficient and/or robust NCG methods. Once a damped vector $\widehat{y}_k$ is defined, it can be used: *(i)* in the definition of β_k, replacing y_k with $\widehat{y}_k$ (*modified method*); *(ii)* in the definition of the preconditioner replacing $P_k(y_k, s_k)$ with $P_k(\widehat{y}_k, s_k)$. In order to clearly evaluate the effect of the damped techniques, we study the cases *(i)* and *(ii)* separately. Furthermore, we also investigate in this paper the joint modification of both β_k and $P_k(y_k, s_k)$, by means of the damped vector $\widehat{y}_k$ (*modified preconditioned method*). Note that in the unpreconditioned case, a damped strategy obviously may affect the definition of β_k only when y_k explicitly appears in the formula of β_k.

Broadly speaking, in extending the definition of the damped vector $\widehat{y}_k$ introduced in (2.6), our aim is to define a vector $\widehat{y}_k$ as a combination of the original vector y_k and an appropriate vector z_k, namely

$$\widehat{y}_k = \varphi_k y_k + (1 - \varphi_k) z_k, \tag{3.11}$$

such that $s_k^\top \widehat{y}_k$ is sufficiently positive for suited values of $\varphi_k \in (0, 1]$. Of course, a key point of this approach is an appropriate choice of z_k, both in terms of certain gained information and in terms of a good relative scaling of $\widehat{y}_k$. Note that the choice (3.11) is reduced to (2.6) if $z_k = B_k s_k$, which cannot be computed explicitly in the NCG context, being B_k unavailable.

In our first proposal, we set $z_k = \eta_k s_k$, where η_k is a positive scalar, based on approximating B_k by $\eta_k I$. It originates from the idea of using $z_k = A_{k+1} y_k$ in (3.11), where A_{k+1} is a positive definitive approximation of the inverse Hessian, satisfying the *modified secant equation*

$$A_{k+1} y_k = \eta_k s_k.$$

Hence, by using the latter equation, we can define the damped formula

$$\widehat{y}_k^{(1)} = \varphi_k y_k + (1 - \varphi_k) \eta_k s_k, \tag{3.12}$$

which does not require the explicit knowledge of the approximate inverse A_{k+1} of the Hessian matrix. Since (3.12) follows from (2.6) with $B_k s_k$ replaced by

$\eta_k s_k$, we use the same replacement in (2.7) to obtain the following formula

$$\varphi_k = \begin{cases} \dfrac{\sigma\eta_k\|s_k\|^2}{\eta_k\|s_k\|^2 - s_k^\top y_k} & \text{if} \quad s_k^\top y_k < (1-\sigma)\eta_k\|s_k\|^2 \\ \\ 1, & \text{otherwise,} \end{cases} \tag{3.13}$$

where $\eta_k \geq 1$.

Then, in order to set $\varphi_k \neq 1$ only whenever $s_k^\top y_k$ is sufficiently small, we modify (3.13) as

$$\varphi_k = \begin{cases} \dfrac{\sigma\eta_k\|s_k\|^2}{\eta_k\|s_k\|^2 - s_k^\top y_k} & \text{if} \quad s_k^\top y_k < (1-\sigma)\|s_k\|^2 \\ \\ 1, & \text{otherwise.} \end{cases} \tag{3.14}$$

Note that on some iterations, the former formula may modify y_k, while the latter one may not, because the condition in the former formula, i.e.

$$s_k^\top y_k < (1-\sigma)\eta_k\|s_k\|^2, \tag{3.15}$$

can be satisfied for sufficiently large values of η_k. For certain choices of η_k, our numerical experience (which we will describe in Section 5) was carried on adopting (3.14), which showed favourable results avoiding the dependence on the product $(1-\sigma)\eta_k$. Nevertheless, the numerical impact of (3.15) deserves further investigations.

We now give an alternative motivation for choice (3.12) which, in practice, represents a combination of y_k and s_k with the scalar η_k. Moreover, we can derive the novel adaptive criterion used in (3.14) starting from a geometric interpretation of the curvature condition (2.5).

As already mentioned, (see e.g. [9]) if f is strongly convex, the curvature condition (2.5) holds. Roughly speaking, f strongly convex means that its curvature is positive and not too close to zero. Hence, motivated by the former idea of Powell in [28], we intend to define a criterion based on the (local) strong convexity of the function for deciding if a damped vector $\widehat{y}_k$ must be used in place of y_k.

It is well known that if f is strongly convex on a convex set $S \subseteq \mathbb{R}^n$, then there exists $\theta > 0$ such that

$$\left[\nabla f(y) - \nabla f(x)\right]^\top (y - x) \geq \theta\|y - x\|^2, \tag{3.16}$$

for all x and y belonging to S. For $\theta = 0$, we recover the basic inequality characterizing the convexity, namely the curvature condition, provided by the Wolfe line search procedure. For $\theta > 0$, we obtain a strong lower bound in (3.16). Hence, given $\theta > 0$, if we adopt (3.16) as selection criterion, we actually obtain the one used in (3.14) with $\theta = 1 - \sigma$. Therefore, the rationale behind this criterion is the following: whenever $s_k^\top y_k \geq (1-\sigma)\|s_k\|^2 > 0$ and hence

the curvature is "sufficiently positive", there is no need to modify the vector y_k; otherwise the damped vector $\widehat{y}_k$ is considered.

Now, we remark that we are interested in obtaining the vector $\widehat{y}_k$ such that $s_k^\top \widehat{y}_k$ is sufficiently positive, and that an improvement in the curvature condition is obtained, namely

$$s_k^\top \widehat{y}_k^{(1)} \geq s_k^\top y_k. \tag{3.17}$$

Recalling that we are considering in (3.14) the case $s_k^\top y_k < (1-\sigma)\|s_k\|^2$, by substituting the value of φ_k in (3.12), by simple computation we obtain

$$s_k^\top \widehat{y}_k^{(1)} = (1-\sigma)\eta_k \|s_k\|^2. \tag{3.18}$$

Therefore $s_k^\top \widehat{y}_k^{(1)}$ is sufficiently positive for suited values of the parameters σ and η_k. Moreover, we can guarantee that $\widehat{y}_k^{(1)}$ satisfies (3.17) by setting $\eta_k > 1$ whenever $s_k^\top y_k > 0$. On the other hand, if $s_k^\top y_k$ is negative, (3.17) is trivially satisfied by the choice $\widehat{y}_k^{(1)}$.

In our second proposal we set $z_k = -\alpha_k g_k$ in (3.12) to obtain the damped vector

$$\widehat{y}_k^{(2)} = \varphi_k y_k - (1-\varphi_k)\alpha_k g_k \tag{3.19}$$

which arises from the following observation: if $B_k \succ 0$ is an approximation of the Hessian and we consider as search direction $-B_k^{-1} g_k$, it immediately follows that $s_k = x_{k+1} - x_k = -\alpha_k B_k^{-1} g_k$ which implies

$$B_k s_k = -\alpha_k g_k.$$

This allows us to consider the original damped vector (2.6), without computing B_k explicitly, i.e. by replacing $B_k s_k$ with $-\alpha_k g_k$, as defined in (3.19). In this case adapting the Powell's rule in (2.7) (replacing $B_k s_k$ with $-\alpha_k g_k$), it follows that

$$\varphi_k = \begin{cases} \dfrac{\sigma \alpha_k s_k^\top g_k}{\alpha_k s_k^\top g_k + s_k^\top y_k}, & \text{if } s_k^\top y_k < -(1-\sigma)\alpha_k s_k^\top g_k, \\ \\ 1, & \text{otherwise.} \end{cases} \tag{3.20}$$

Substituting the value of φ_k from the first case (i.e. $\varphi_k \neq 1$) into (3.19), we obtain

$$s_k^\top \widehat{y}_k^{(2)} = -\alpha_k (1-\sigma) s_k^\top g_k = -\alpha_k^2 (1-\sigma) p_k^\top g_k > 0,$$

where the last inequality follows since p_k is a descent direction at x_k. Moreover, here we also have that the final steplength computed by the line search procedure plays a keynote role.

Following guidelines adopted to obtain (3.14), formula (3.20) can be changed to define

$$\varphi_k = \begin{cases} \dfrac{\sigma\eta_k\alpha_k s_k^\top g_k}{\eta_k\alpha_k s_k^\top g_k + s_k^\top y_k}, & \text{if } s_k^\top y_k < -(1-\sigma)\alpha_k s_k^\top g_k, \\ \\ 1, & \text{otherwise,} \end{cases} \tag{3.21}$$

where $\eta_k \geq 1$. Furthermore, similar formulae with the three cases in (2.8) can be also defined.

Finally, observe that in our first proposal the conditions (3.14)-(3.18) omit the dependency on any considerations regarding the global convergence of the final damped techniques. In this regard, a further study on the latter issue (see also [1] [2] [8]) seems to be necessary, which will be the object of future research.

4 Convergence properties for preconditioned damped Polak–Ribière (D-PR-PNCG) method

As already recalled, in paper [2] the author extends global convergence properties of the Broyden family of quasi–Newton methods to the damped version of such methods. In a similar fashion, we aim at proving that some global convergence properties of NCG methods still hold in the general case corresponding to the damped and preconditioned version (*modified* PNCG method). Obviously, results for undamped and/or unmodified methods are straightforwardly obtained as particular cases.

As first step of the convergence analysis, in this section our preliminary focus is on the Polak–Ribière (PR) version of the NCG. In particular, here we limit our analysis to consider only the first proposal in (3.12). Note that in this regard, developing convergence properties with the choice $\widehat{y}_k^{(2)}$ needs additional analysis, which is part of our future work.

Using the damped vector $\widehat{y}_k^{(1)}$ we therefore consider the damped preconditioned PR method (namely D-PR-PNCG method):

$$\hat{\beta}_k^{\mathrm{PR}} = \frac{\widehat{y}_k^{(1)\top} M_{k+1} g_{k+1}}{g_k^\top M_k g_k}. \tag{4.22}$$

The resulting D-PR-PNCG method actually is a novel *modified* NCG method. Hence the necessity of ensuring its global convergence properties. To this aim, in this section, we prove that, to some extent, the D-PR-PNCG method enjoys the same properties as the standard (undamped and unpreconditioned) PR method (see e.g. [22]). In particular, we have the following result. We also need the following assumption to prove our final results.

Assumption 1

a) Given the vector $x_1 \in \mathbb{R}^n$ and the function $f \in C^1(\mathbb{R}^n)$, the level set $\mathcal{L}_1 = \{x \in \mathbb{R}^n : f(x) \leq f_1\}$ is compact.

b) There exists an open ball $\mathcal{B}_r := \{x \in \mathbb{R}^n : \|x\| < r\}$ containing $\mathcal{L}_1$ where $f(x)$ is continuously differentiable and its gradient $g(x)$ is Lipschitz continuous. In particular, there exists $L > 0$ such that

$$\|g(x) - g(y)\| \leq L\|x - y\| \qquad \text{for all} \qquad x, y \in \mathcal{B}_r.$$

c) There exist $\lambda > 0$ and $\Lambda > 0$ such that the preconditioner $M(x)$, for any $x \in \mathcal{B}_r$, is positive definite with the smallest [largest] eigenvalue $\lambda_m(M(x))$ [$\lambda_M(M(x))$] satisfying

$$0 < \lambda \leq \lambda_m(M(x)) \leq \lambda_M(M(x)) < \Lambda.$$

Proposition 1 *Let $\{x_k\}$ be an infinite sequence (with $g_k \neq 0$) generated by the D-PR-PNCG method, where the steplength $\alpha_k > 0$ is determined by a linesearch procedure such that, for all k, the following conditions hold:*

(i) $x_k \in \mathcal{L}_1$ for all k;
(ii) $\lim_{k\to+\infty} \dfrac{|g_k^\top p_k|}{\|p_k\|} = 0$;
(iii) $\lim_{k\to+\infty} \alpha_k \|p_k\| = 0$.

If Assumption 1 holds, then

$$\liminf_{k\to+\infty} \|g_k\| = 0$$

and hence there exists at least a stationary limit point of $\{x_k\}$.

Proof: First observe that by the Lipschitz continuity of $g(x)$ and the compactness of $\mathcal{L}_1$, there exists a number $\Gamma > 0$ such that

$$\|g(x)\| \leq \Gamma, \qquad \text{for all} \qquad x \in \mathcal{L}_1. \tag{4.23}$$

Moreover, from *(i)* and the compactness of $\mathcal{L}_1$, the sequence $\{x_k\}$ admits limit points in $\mathcal{L}_1$. Now, by contradiction, assume that there exist $\varepsilon > 0$ and $\bar{k}$ such that

$$\|g_k\| \geq \varepsilon, \qquad \text{for all} \qquad k > \bar{k}. \tag{4.24}$$

By using (4.22)-(4.24) and *(i)*, and recalling that we are considering D-PR-PNCG, we get for any $k \geq \bar{k}$,

$$\begin{aligned}
\|p_{k+1}\| &= \| - M_{k+1}g_{k+1} + \hat{\beta}_k^{\mathrm{PR}} p_k\| \\
&= \left\| -M_{k+1}g_{k+1} + \frac{[\varphi_k y_k + (1-\varphi_k)\eta_k s_k]^\top M_{k+1}g_{k+1}}{g_k^\top M_k g_k} p_k \right\| \\
&\leq \|M_{k+1}g_{k+1}\| + \frac{\|\varphi_k y_k + (1-\varphi_k)\eta_k s_k\| \|M_{k+1}g_{k+1}\|}{\|g_k\| \|M_k g_k\|} \|p_k\| \\
&\leq \Gamma\lambda_M(M_{k+1}) + \Gamma\lambda_M(M_{k+1}) \frac{\|\varphi_k y_k + (1-\varphi_k)\eta_k s_k\|}{\varepsilon^2 \lambda_m(M_k)} \|p_k\|. \qquad (4.25)
\end{aligned}$$

From (4.25), recalling the Lipschitz continuity of $g(x)$ on $\mathcal{L}_1$, we have

$$\begin{aligned}\|\varphi_k y_k + (1-\varphi_k)\eta_k s_k\| &= \|\varphi_k(g_{k+1}-g_k) + (1-\varphi_k)\eta_k(x_{k+1}-x_k)\| \\ &\leq \varphi_k L\|x_{k+1}-x_k\| + (1-\varphi_k)\eta_k\|x_{k+1}-x_k\| \\ &= \|\alpha_k p_k\|(\varphi_k L + (1-\varphi_k)\eta_k). \end{aligned} \tag{4.26}$$

Hence, by using (4.25) we obtain

$$\|p_{k+1}\| \leq \Gamma\lambda_M(M_{k+1}) + \Gamma\lambda_M(M_{k+1})\left(\frac{\varphi_k L + (1-\varphi_k)\eta_k}{\varepsilon^2\lambda_m(M_k)}\right)\|\alpha_k p_k\|\|p_k\|. \tag{4.27}$$

Now, by *(iii)*, given $q \in (0,1)$, we can assume there exists k_1 sufficiently large such that

$$\Gamma\lambda_M(M_{k+1})\left(\frac{\varphi_k L + (1-\varphi_k)\eta_k}{\varepsilon^2\lambda_m(M_k)}\right)\|\alpha_k p_k\| \leq q < 1, \quad \text{for any} \quad k \geq k_1 > \bar{k}. \tag{4.28}$$

Thus, by (4.27)-(4.28) we get

$$\|p_{k+1}\| \leq \Gamma\lambda_M(M_{k+1}) + q\|p_k\|, \qquad \text{for any} \qquad k \geq k_1,$$

and by Lemma 1 in the Appendix

$$\|p_{k+1}\| \leq \frac{\Gamma\lambda_M(M_{k+1})}{1-q} + \left(\|p_{k_1}\| - \frac{\Gamma\lambda_M(M_{k+1})}{1-q}\right)q^{(k+1)-k_1}, \qquad \forall k \geq k_1, \tag{4.29}$$

showing that $\|p_{k+1}\|$ is bounded as $\|p_{k_1}\|$ is bounded. As a consequence, again from *(iii)* we have

$$\lim_{k\to+\infty} \alpha_k\|p_k\|^2 = 0. \tag{4.30}$$

Furthermore, the boundedness of $\|p_k\|$ and *(ii)* yield

$$\lim_{k\to+\infty} |g_k^\top p_k| = 0. \tag{4.31}$$

Since $M_{k+1}g_{k+1} = \hat{\beta}_k^{\mathrm{PR}}p_k - p_{k+1}$, by (4.27) it results

$$\begin{aligned} g_{k+1}^\top M_{k+1}g_{k+1} &= g_{k+1}^\top\hat{\beta}_k^{\mathrm{PR}}p_k - g_{k+1}^\top p_{k+1} \\ &\leq \|g_{k+1}\|\|\hat{\beta}_k^{\mathrm{PR}}p_k\| + |g_{k+1}^\top p_{k+1}| \\ &\leq \frac{\alpha_k\Gamma\lambda_M(M_{k+1})\|g_{k+1}\|\|p_k\|^2(\varphi_k L + (1-\varphi_k)\eta_k)}{\varepsilon^2\lambda_m(M_k)} \\ &\quad + |g_{k+1}^\top p_{k+1}|. \end{aligned} \tag{4.32}$$

By (4.30), (4.31) and the compactness of $\mathcal{L}_1$, taking limits in (4.32) as $k \to +\infty$, we obtain

$$\lim_{k\to+\infty} g_{k+1}^\top M_{k+1}g_{k+1} = 0.$$

Finally, by (c) of Assumption 1

$$\lim_{k\to+\infty} \|g_k\| = 0$$

and this contradicts assumption (4.24). □

As is well known, in the last two decades, several papers have been devoted to define inexact linesearch procedures ensuring *(i)-(iii)* of Proposition 1 or other similar technical conditions. The latter procedures enable to guarantee some global convergence properties for the PR method (see e.g. [21] and the references reported therein). We only mention here, as an example, the approach proposed in [20] where an Armijo–type linesearch method is given with an acceptability criterion of the steplength based on a "parabolic bound". We are certainly aware of the fact that to guarantee the properties *(ii)-(iii)* might not be easy, and requires some additional effort.

However, in the current paper we limit our numerical experience to consider standard linesearch procedures based on the Wolfe conditions, so that we adopt the well known implementation proposed in [26], which finds a steplength such that the strong Wolfe conditions hold with the parameter values of 0.0001 and 0.1. Our choice is motivated by the fact that, in order to avoid a possible bias for the conclusions in our study, we need to accurately discard nonstandard elements in our scheme.

5 Damped preconditioned NCG methods: a numerical experience

In this section we consider the use of the damped vectors defined in (3.12)-(3.19), for constructing a preconditioner based on quasi–Newton updates. Therefore, according to the taxonomy in Section 1, here we consider *unmodified* PNCG methods, where the use of damped techniques only affects the preconditioning strategies and not the value of β_k. Our aim is to perform a numerical assessment when adopting damped techniques within a PNCG method. On the other hand, note that as regards the convergence (and the order of convergence) of PNCG methods, an interesting theoretical result has been proved in [3]. However, it considers the use of an exact linesearch and a strong assumption on the preconditioner is required, namely the preconditioner is assumed to be a "strongly consistent approximation" of the Hessian matrix at the solution. Therefore this result risks to be seldom applied in practice.

The preconditioner we use is an approximate inverse preconditioner belonging to the class proposed in [12] and briefly recalled in Section 2.3. It is based on quasi–Newton updates and thus constructed, at each iteration k of the PNCG method, by adding the contribution of the current pair (y_k, s_k). According to the "limited memory" strategy, it looks backwards by taking into account the most recent m pairs. Since it is iteratively constructed, it is quite simple to introduce an adaptive rule and to choose, at each iteration k, if it convenient to replace y_k with a damped vector $\widehat{y}_k$. If so, the resulting preconditioner $P_k(\widehat{y}_k, s_k)$ is then used in place of $P_k(y_k, s_k)$.

We embedded the latter strategy in the implementation of the PNCG described in [12] (we refer to this paper for all the details). Note that this implementation is based on the standard CG+ code (see [17]), where the preconditioner reported in (2.9) with the parameters in (2.10) is included, and the linesearch technique is the same as that of [26].

In particular, we focused on the *unmodified* preconditioned Polak–Ribière method and performed an extensive numerical testing by considering all the large scale problems available in the CUTEst collection [18], namely 112 problems whose dimension ranges from 1000 to 10000. The stopping criterion is the standard one (see e.g. [25]) which is given by

$$\|g_k\| \leq 10^{-5} \max\{1, \|x_k\|\}.$$

In the sequel our numerical results are reported by using performance profiles [13], both in terms of number of iterations and number of function and gradient evaluations. For each comparison we report two profiles: a standard profile and a *detailed* profile; the latter one differs from the standard one only with respect to the scale of the abscissa axis, which is restricted to values closer to 1. Moreover, we also recall that in the linesearch procedure adopted by the authors in [26], the number of function and gradient evaluations coincide.

We started by considering our first proposal, namely the use of the damped vector $\widehat{y}_k^{(1)}$ in (3.12) combined with the adaptive rule in (3.14). First of all, we needed to tune the choice of the two parameters η_k in (3.12) and σ in (3.14). In Figure 1 (profile) and Figure 2 (*detailed* profile) we report the results obtained for different choices of $\eta_k \in \{2, 3, 4, 5\}$ and by setting $\sigma = 0.8$. Conversely, in Figure 3 (profile) and Figure 4 (*detailed* profile) we report the results obtained for different choices of $\sigma \in \{0.8, 0.6, 0.4, 0.2\}$ and by setting $\eta_k = 4$. By observing the profiles, the values $\eta_k = 4$ and $\sigma = 0.8$ seem to be the best ones, based on our experiments on the above mentioned set of test problems. These latter have been used in the sequel of this section as default values of η_k and σ.

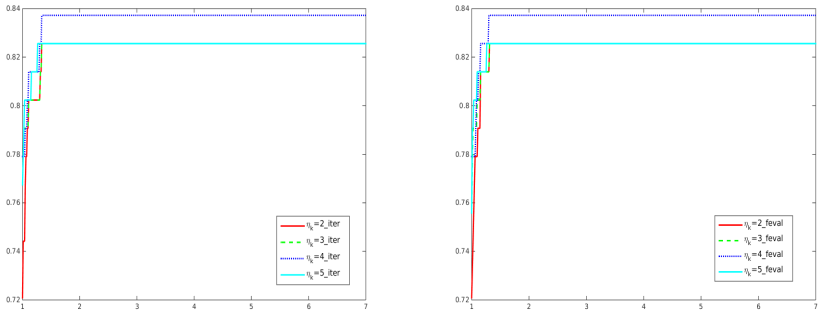

Fig. 1 Comparison among different choices of η_k in (3.12), setting $\sigma = 0.8$ in (3.14). Profiles with respect to # *iterations* (left) and # *function and gradient evaluations* (right).

Figures 5–6 report the results of the comparison between the *unmodified preconditioned* PR method, whose preconditioner is damped according to the formula (3.12), and the standard preconditioned PR method. These profiles

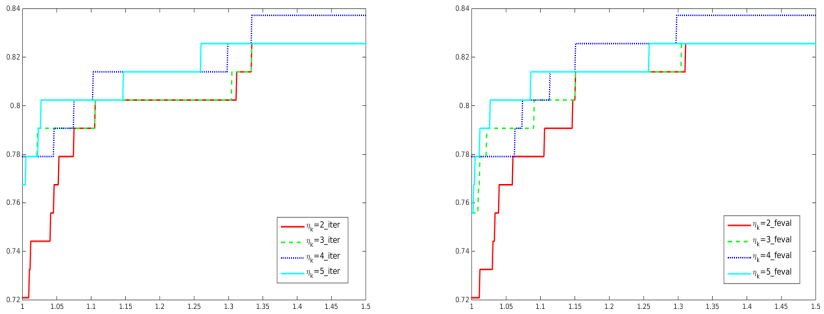

Fig. 2 Comparison among different choices of η_k in (3.12), setting $\sigma = 0.8$ in (3.14). Detailed profiles with respect to # *iterations* (left) and # *function and gradient evaluations* (right).

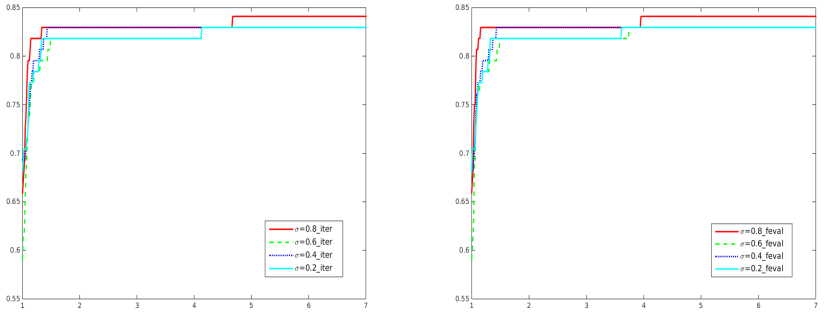

Fig. 3 Comparison among different choices of σ in (3.14), setting $\eta_k = 4$ in (3.12). Profiles with respect to # *iterations* (left) and # *function and gradient evaluations* (right).

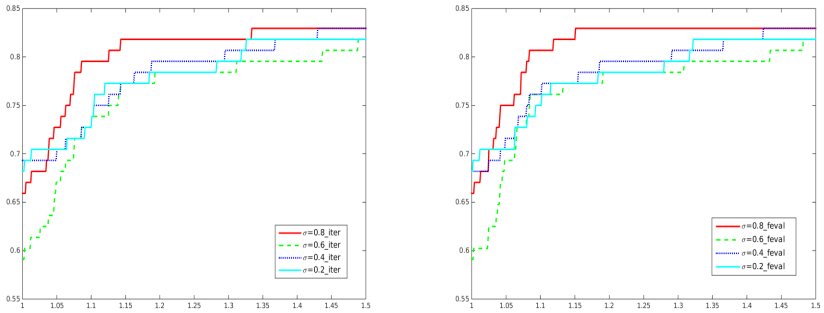

Fig. 4 Comparison among different choices of σ in (3.14), setting $\eta_k = 4$ in (3.12). Detailed profiles with respect to # *iterations* (left) and # *function and gradient evaluations* (right).

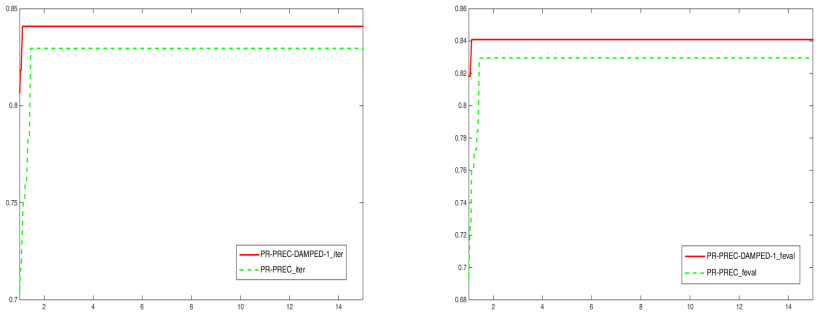

Fig. 5 Comparison between *unmodified preconditioned* PR damped according to (3.12) and the standard preconditioned PR (undamped). Profiles with respect to # *iterations* (left) and # *function and gradient evaluations* (right).

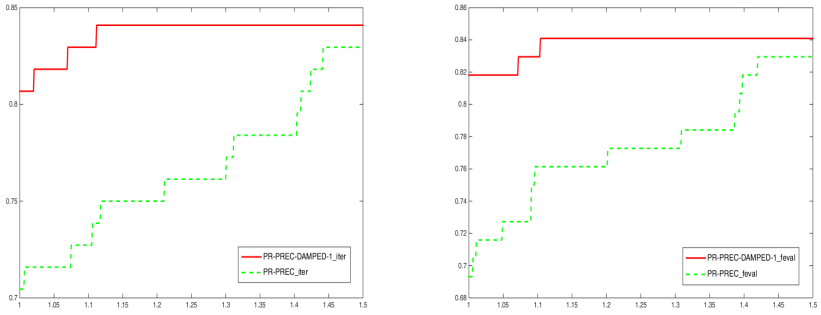

Fig. 6 Comparison between *unmodified preconditioned* PR damped according to (3.12) and the standard preconditioned PR (undamped). Detailed profiles with respect to # *iterations* (left) and # *function and gradient evaluations* (right).

clearly evidence the fruitful use of the first damped strategy both in terms of efficiency and in terms of robustness.

Then we turned to our second proposal, namely the use of the damped vector $\widehat{y}_k^{(2)}$ in (3.19) combined with the original rule in (2.7) for choosing φ_k (with $B_k s_k$ replaced by $-\alpha_k g_k$) with $\sigma = 0.8$. In the Figures 7–8 the comparison between the *unmodified preconditioned* PR method, whose preconditioner is damped according to formula (3.19), and the standard preconditioned PR method is reported. Also in this case the adoption of the damped strategy for computing the preconditioner is very useful.

Finally, we compared the two damped strategies proposed in this paper. The results of this comparison are reported in Figures 9–10. By observing these profiles, the adoption of the first damped strategy seems to be slightly preferable.

It is also worth to highlight that from the detailed complete numerical results we obtained (not all reported in this paper), as expected the damped

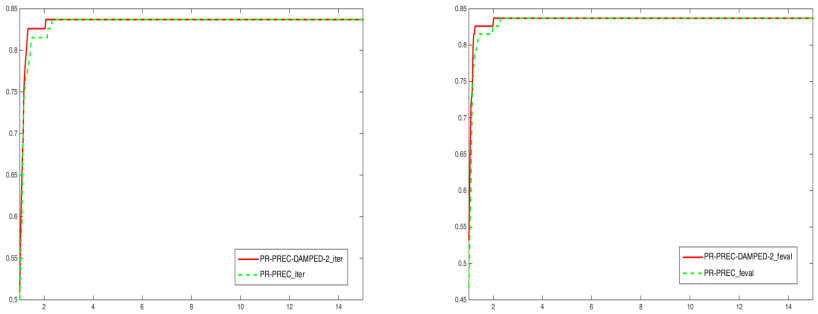

Fig. 7 Comparison between *unmodified preconditioned* PR damped according to (3.19) and the standard preconditioned PR (undamped). Profiles with respect to # *iterations* (left) and # *function and gradient evaluations* (right).

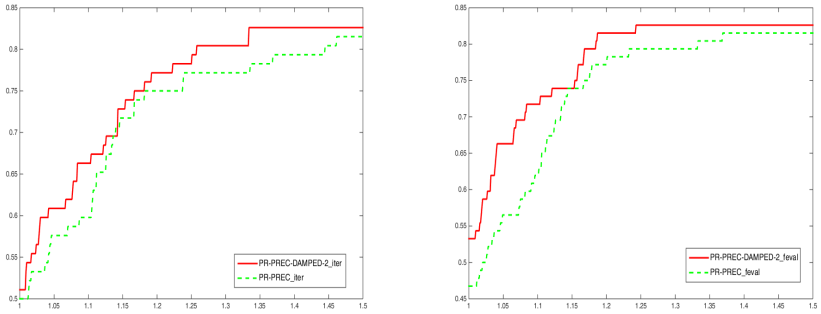

Fig. 8 Comparison between *unmodified preconditioned* PR damped according to (3.19) and the standard preconditioned PR (undamped). Detailed profiles with respect to # *iterations* (left) and # *function and gradient evaluations* (right).

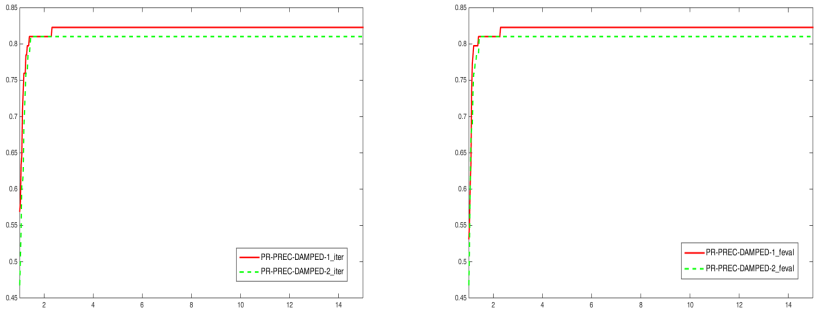

Fig. 9 Comparison between the adoption of the two damped strategies in (3.12) and in (3.19). Profiles with respect to # *iterations* (left) and # *function and gradient evaluations* (right).

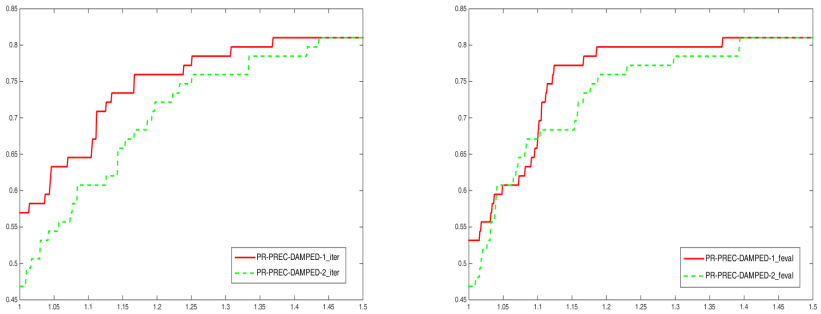

Fig. 10 Comparison between the adoption of the two damped strategies in (3.12) and in (3.19). Detailed profiles with respect to # *iterations* (left) and # *function and gradient evaluations* (right).

strategy occurs in few cases. In particular, when it takes place it enhances either the robustness or the efficiency of the algorithm. In other words, in the case of test problems without "pathologies", correctly the damped strategy is not invoked by the adaptive rule.

On the overall, the results of the numerical experiences reported indicate that the use of a damped strategy can definitely improve the performance of the PR algorithm, at least on the CUTEst problems considered.

So far, the damped strategy was experimented in constructing our quasi–Newton based preconditioner, which is the main focus of this paper. Now, for the sake of completeness, since the theoretical part in Section 4 encompasses the possibility to embed the damped strategy both in the definition of the scalar β_k and in the preconditioner, we urge to perform numerical testing also on the use of $\widehat{\beta}_k^{PR}$ in (4.22). In this regard, note that the use of damped strategy was conceived in the context of quasi–Newton updates, and it is not expected to be successfully exploited in the definition of the scalar β_k used in a NCG/PNCG method. In the sequel we report results obtained by using the damped vector $\widehat{\beta}_k^{PR}$ confirming this claim. In particular, we first consider the unpreconditioned cased and compare the behaviour of the unmodified NCG method with the method which adopts $\widehat{\beta}_k^{PR}$, setting $\widehat{y}_k = \widehat{y}_k^{(1)}$ with the default values of $\sigma = 0.8$ and $\eta_k = 4$. Then, we perform the same comparing in the preconditioned case. Figure 11 and Figure 12 report the performance profiles in terms of number of iterations and number of function/gradient evaluations corresponding to these comparisons. As it can be observed from these profiles, the use of the $\widehat{\beta}_k^{PR}$ does not yield a noteworthy improvement neither in terms of iterations or function evaluations. Nevertheless we also observe that the D-PR-PNCG scheme which also uses $\widehat{\beta}_k^{PR}$ reveals to outperform the standard NCG method. Thus, on the overall, the adoption of the damped strategy within PNCG methods seems to be definitely promising.

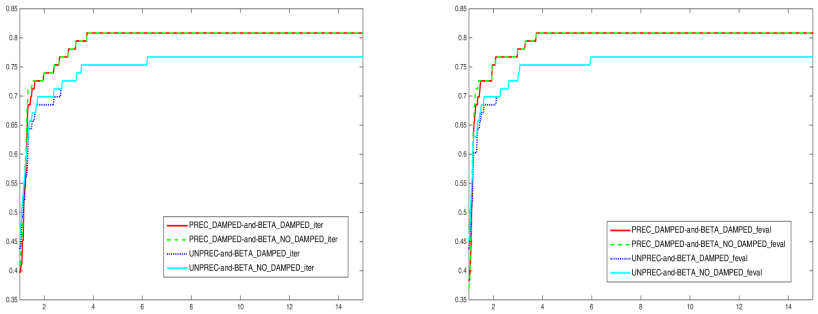

Fig. 11 Comparison between the use $\widehat{\beta}_k^{PR}$ in (4.22) (setting $\widehat{y}_k = \widehat{y}_k^{(1)}$) and β_k^{PR} in (2.4), in both preconditioned and unpreconditioned cases. Profiles with respect to # *iterations* (left) and # *function and gradient evaluations* (right).

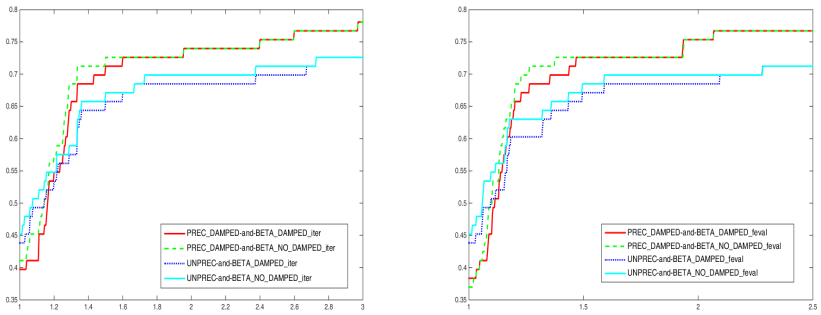

Fig. 12 Comparison between the use $\widehat{\beta}_k^{PR}$ in (4.22) (setting $\widehat{y}_k = \widehat{y}_k^{(1)}$) and β_k^{PR} in (2.4), in both preconditioned and unpreconditioned cases. Detailed profiles with respect to # *iterations* (left) and # *function and gradient evaluations* (right).

6 Conclusions and future works

In this seminal paper we proposed the introduction of damped techniques within the framework of the NCG methods. We drew our inspiration from the damped quasi–Newton methods proposed by Al-Baali and Powell. In particular, by referring to the PR method, we investigated separately two possibilities:

- the use of a damped vector in the definition of the scalar β_k, hence affecting the definition of the search direction and producing a modified NCG/PNCG method;
- the use of a damped vector in the unmodified preconditioned NCG method.

As regards the first one, we proved that some global convergence properties still hold for the modified D-PR-PNCG method, while substantially preserving numerical performance. As concerns the second one, we used the damped strategy for constructing a preconditioner based on quasi–Newton updates to

be used in the PNCG method. The results obtained clearly highlighted the potentialities of this approach.

Of course several other aspects of interest on damped PNCG were not treated in this paper. They range from (but are not limited to) the use of damped techniques to possibility enhance some global convergence properties of the NCG methods, to their more sophisticated use in the construction of a preconditioner (for instance, by introducing a dependence on the iteration k of the parameter $\sigma = \sigma_k$ and a dependence of σ_k and η_k on $\|g_k\|$ or the number of iterations). Considering self–scaling quasi-Newton methods, it might be also useful to consider the choice of $\bar{\eta}_k = \frac{s_k^\top B_k s_k}{s_k^\top y_k} = \frac{-\alpha_k s_k^\top g_k}{s_k^\top y_k}$ and to use $\eta_k = \max(\bar{\eta}_k, 2)$ in the numerical experiences. Moreover, the combined use of damped strategies with other linesearch procedures (different from the standard Wolfe method) is surely of great interest, too. Finally, adopting the test (3.15) in place of the one in (3.14) can be a possible alternative to explore, in order to improve performance.

Appendix

In this appendix, we report a technical result used in the proofs of Proposition 1 (see [22]).

Lemma 1 *Let $\{\xi_k\}$ be a sequence of nonnegative real numbers. Let $\Omega > 0$ and $q \in (0,1)$ and suppose that there exists $k_1 \geq 1$ such that*

$$\xi_k \leq \Omega + q\xi_{k-1}, \qquad \textit{for any} \qquad k \geq k_1.$$

Then,

$$\xi_k \leq \frac{\Omega}{1-q} + \left(\xi_{k_1} - \frac{\Omega}{1-q}\right) q^{k-k_1}, \qquad \textit{for any} \qquad k \geq k_1.$$

Proof: Starting from relation

$$\xi_k \leq \Omega + q\xi_{k-1}, \qquad \text{for any} \qquad k \geq k_1,$$

considering $k - k_1$ iterations we get:

$$\xi_k \leq \Omega \left(\sum_{i=0}^{k-k_1-1} q^i \right) + q^{k-k_1}\xi_{k_1},$$

from which we obtain

$$\xi_k \leq \Omega\Big(\frac{1-q^{k-k_1}}{1-q}\Big) + q^{k-k_1}\xi_{k_1} = \frac{\Omega}{1-q} + \left(\xi_{k_1} - \frac{\Omega}{1-q}\right) q^{k-k_1}.$$

□

References

1. Al-Baali, M.: Descent property and global convergence of the Fletcher-Reeves method with inexact line search. IMA Journal on Numerical Analysis **5**, 121–124 (1985)
2. Al-Baali, M.: Damped techniques for enforcing convergence of quasi–Newton methods. Optimization Methods and Software **29**, 919–936 (2014)
3. Al-Baali, M., Fletcher, R.: On the order of convergence of preconditioned nonlinear conjugate gradient methods. SIAM Journal on Scientific Computing **17**, 658–665 (1996)
4. Al-Baali, M., Grandinetti, L.: Improved damped quasi–Newton methods for unconstrained optimization. Pacific Journal of Optimization (To appear)
5. Al-Baali, M., Grandinetti, L.: On practical modifications of the quasi–Newton BFGS methods. AMO – Advanced Modeling and Optimization **11**, 63–76 (2009)
6. Al-Baali, M., Grandinetti, L., Pisacane, O.: Damped techniques for the limited memory BFGS method for large–scale optimization. Journal of Optimization Theory and Applications **161**, 688–699 (2014)
7. Al-Baali, M., Purnama, A.: Numerical experience with damped quasi–Newton optimization methods when the objective function is quadratic. SQU Journal for Science **17**, 1–11 (2012)
8. Al-Baali, M., Spedicato, E., Maggioni, F.: Broyden's quasi–Newton methods for a nonlinear system of equations and unconstrained optimization: a review and open problems. Optimization Methods and Software **29**, 937–954 (2014)
9. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press, New York (2004)

10. Caliciotti, A., Fasano, G., Roma, M.: Preconditioned nonlinear conjugate gradient methods based on a modified secant equation. Submitted to *Applied Mathematics and Computation*
11. Caliciotti, A., Fasano, G., Roma, M.: Preconditioning strategies for nonlinear conjugate gradient methods, based on quasi-newton updates. In: Y. Sergeyev, D. Kvasov, F. Dell'Accio, M. Mukhametzhanov (eds.) AIP Conference Proceedings, vol. 1776 090007, pp. 1–4. American Institute of Physics (2016)
12. Caliciotti, A., Fasano, G., Roma, M.: Novel preconditioners based on quasi–Newton updates for nonlinear conjugate gradient methods. Optimization Letters **11**, 835–853 (2017)
13. Dolan, E.D., Moré, J.: Benchmarking optimization software with performance profiles. Mathematical Programming **91**, 201–213 (2002)
14. Fasano, G., Roma, M.: Preconditioning Newton-Krylov methods in nonconvex large scale optimization. Computational Optimization and Applications **56**, 253–290 (2013)
15. Fasano, G., Roma, M.: A novel class of approximate inverse preconditioners for large scale positive definite linear systems in optimization. Computational Optimization and Applications **65**, 399–429 (2016)
16. Fletcher, R.: Practical Methods of Optimization. John Wiley and Sons, New York (1987)
17. Gilbert, J., Nocedal, J.: Global convergence properties of conjugate gradient methods for optimization. SIAM Journal on Optimization **2**, 21–42 (1992)
18. Gould, N.I.M., Orban, D., Toint, P.L.: CUTEst: a constrained and unconstrained testing environment with safe threads. Computational Optimization and Applications **60**, 545–557 (2015)
19. Gratton, S., Sartenaer, A., Tshimanga, J.: On a class of limited memory preconditioners for large scale linear systems with multiple right-hand sides. SIAM Journal on Optimization **21**, 912–935 (2011)
20. Grippo, L., Lucidi, S.: A globally convergent version of Polak–Ribière conjugate gradient method. Mathematical Programming **78**, 375–391 (1997)
21. Grippo, L., Lucidi, S.: Convergence conditions, line search algorithms and trust region implementations for the Polak–Ribière conjugate gradient method. Optimization Methods and Software **20**, 71–98 (2005)
22. Grippo, L., Sciandrone, M.: Metodi di ottimizzazione non vincolata. Springer–Verlag Italia, Milan (2011)
23. Hager, W., Zhang, H.: The limited memory conjugate gradient method. SIAM Journal on Optimization **23**, 2150–2168 (2013)
24. Kelley, C.T.: Iterative methods for Optimization. SIAM Frontiers in Applied Mathematics, Philadelphia, PA (1999)
25. Morales, J., Nocedal, J.: Automatic preconditioning by limited memory quasi–Newton updating. SIAM Journal on Optimization **10**, 1079–1096 (2000)
26. Moré, J., Thuente, D.: Line search algorithms with guaranteed sufficient decrease. ACM Transactions on Mathematical Software (TOMS) **20**, 286–307 (1994)
27. Nocedal, J., Wright, S.: Numerical Optimization. Springer–Verlag, New York (2006). Second edition
28. Powell, M.J.D.: Algorithms for nonlinear constraints that use Lagrangian functions. Mathematical Programming **14**, 224–248 (1978)
29. Powell, M.J.D.: How bad are the BFGS and DFP methods when the objective function is quadratic? Mathematical Programming **34**, 34–47 (1986)
30. Pytlak, R.: Conjugate Gradient Algorithms in Nonconvex Optimization. Springer, Berlin (2009)