Preconditioning strategies for nonlinear conjugate gradient methods, based on quasi-Newton updates

Caliciotti Andrea' , Fasano Giovanni' , and Roma Massimo'

Citation: **1776**, 090007 (2016); doi: 10.1063/1.4965371
View online: http://dx.doi.org/10.1063/1.4965371
View Table of Contents: http://aip.scitation.org/toc/apc/1776/1
Published by the American Institute of Physics

Preconditioning Strategies for Nonlinear Conjugate Gradient Methods, Based on Quasi-Newton Updates

Caliciotti Andrea[1,b)], Fasano Giovanni[2,c)] and Roma Massimo[3,a)]

[1]*Dipartimento di Ingegneria Informatica, Automatica e Gestionale, SAPIENZA - Università di Roma, via Ariosto 25 00185 Roma, Italy*
[2]*Department of Management, University Ca'Foscari of Venice, S.Giobbe Cannaregio 873, 30121, Venice, Italy*
[3]*Dipartimento di Ingegneria Informatica, Automatica e Gestionale, SAPIENZA - Università di Roma, via Ariosto 25 00185 Roma, Italy*

[a)]Corresponding author: roma@dis.uniroma1.it
[b)]caliciotti@dis.uniroma1.it
[c)]fasano@unive.it

Abstract. This paper reports two proposals of possible preconditioners for the Nonlinear Conjugate Gradient (NCG) method, in large scale unconstrained optimization. On one hand, the common idea of our preconditioners is inspired to L-BFGS quasi–Newton updates, on the other hand we aim at explicitly approximating in some sense the inverse of the Hessian matrix. Since we deal with large scale optimization problems, we propose matrix–free approaches where the preconditioners are built using symmetric low–rank updating formulae. Our distinctive new contributions rely on using information on the objective function collected as by-product of the NCG, at previous iterations. Broadly speaking, our first approach exploits the secant equation, in order to impose interpolation conditions on the objective function. In the second proposal we adopt and *ad hoc* modified–secant approach, in order to possibly guarantee some additional theoretical properties.

INTRODUCTION

In this paper we consider the large scale unconstrained optimization problem

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

where $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ is twice continuously differentiable and n is large. Without loss of generality, we assume here that an NCG iterative scheme is used to solve (1), starting from the point $x_0 \in \mathbb{R}^n$, such that the level set

$$\Omega_0 = \{x \in \mathbb{R}^n \ : \ f(x) \leq f(x_0)\}$$

is compact. There is possibly no need to remark the amount of real applications where the model (1) naturally arises. In this regard, though unconstrained optimization is surely by now a mature research area, there is yet room for improvements when tough highly nonlinear problems are considered (see also [1]).

Effective iterative methods for large scale unconstrained optimization are undoubtedly the NCG method and Limited Memory quasi–Newton methods, being L-BFGS often the method of choice due to its efficiency (see again [1]). Nevertheless, on highly nonlinear problems where the Hessian matrix is indefinite [2] and ill-conditioning easily arises, also quasi–Newton methods may become inefficient, showing the importance of further research on this relevant topic. Our perspective starts from considering some keynote technicalities of L-BFGS, whose update can be suitably reformulated in order to evidence some of its features. Then, we explicitly attempt to replicate in our two proposals the idea behind the latter features, in the light of implicitly pursuing an approximation of the Hessian matrix of the function. To this purpose we directly inherit the rationale behind [3] and [4], where in different contexts the information on the objective function is collected by the Conjugate Gradient (CG) method, and is then used as by-product to build efficient general preconditioners.

Numerical Computations: Theory and Algorithms (NUMTA–2016)
AIP Conf. Proc. 1776, 090007-1–090007-4; doi: 10.1063/1.4965371
Published by AIP Publishing. 978-0-7354-1438-9/$30.00

Here we similarly focus on the NCG method, in order to provide the necessary information to build our preconditioners. Then, a Preconditioned Nonlinear Conjugate Gradient (PNCG) method is adopted for the solution of (1). In particular, we recall that the NCG method generates the sequence of iterates $\{x_k\}$, based on the recursion

$$x_{k+1} = x_k + \alpha_k p_k, \qquad p_k = -\nabla f(x_k) + \beta_k p_{k-1},$$

where $\{p_k\}$ is a sequence of search directions, α_k is a suitable steplength obtained by a proper linesearch procedure based on Wolfe conditions [1]. As well known, different values of β_k and α_k may give rise to different algorithms (see also [5] for a survey).

Regardless of the techniques used to select the parameters α_k and β_k, preconditioning strategies remain a key aspect for increasing the efficiency of NCG methods, especially on ill–conditioned problems. Finally, observe that our proposals owe much of their interest to the close connection between BFGS and NCG, as stated in [6]. The latter consideration is better detailed and fully exploited in the next section, in order to show the role played by some specific dyads in L–BFGS update.

L–BFGS UPDATE AND APPROXIMATE INVERSE HESSIANS

Here we review information on L–BFGS, in order to detail how our proposals tend to approximate information on the inverse Hessian matrix of the objective function in (1). As well known the search direction using L–BFGS is generated at step k as

$$p_k = -H_k \nabla f(x_k),$$

where $H_k \in \mathbb{R}^{n\times n}$ is updated according with

$$H_{k+1} = V_k^T H_k V_k + \rho_k s_k s_k^T, \qquad \rho_k = \frac{1}{y_k^T s_k}, \qquad V_k = I_n - \rho_k y_k s_k^T,$$

and the n-dimensional vectors y_k, s_k are computed as

$$s_k = x_{k+1} - x_k = \alpha_k p_k, \qquad y_k = \nabla f(x_{k+1}) - \nabla f(x_k).$$

Two relevant reasons for the successful application of L–BFGS in the literature is surely due to the following distinctive properties:

(i) starting from a positive definite matrix H_k, the update H_{k+1} is the unique positive definite matrix which solves the subproblem (here $\|\cdot\|_F$ indicates the Frobenius norm)

$$\begin{aligned} \min_H \ & \|H - H_k\|_F \\ s.t. \ & H = H^T \\ & s_k = H y_k; \end{aligned}$$

(ii) when the objective function $f(x)$ is quadratic, with $f(x) = \frac{1}{2}x^T A x - b^T x$, $A \in \mathbb{R}^{n\times n}$ and $b \in \mathbb{R}^n$, then after some computation the L–BFGS update becomes explicitly

$$H_{k+1} = V_k^T H_k V_k + \frac{s_k s_k^T}{y_k^T s_k} = V_k^T V_{k-1}^T \cdots V_1^T H_0 V_1 \cdots V_{k-1} V_k + \sum_{i=1}^{k} \frac{s_i s_i^T}{s_i^T A s_i}.$$

Hence, on one hand the L–BFGS update H_{k+1} is well-scaled, by means of using the Frobenius norm; then, it satisfies some interpolation conditions summarized by the secant equation $s_k = H_{k+1} y_k$. Finally, the rank-k update

$$\sum_{i=1}^{k} \frac{s_i s_i^T}{s_i^T A s_i}$$

in (ii) can be seen, in some sense, as an *approximate* inverse Hessian matrix, since (see [6]) for $f(x) = \frac{1}{2}x^T A x - b^T x$ and $k = n$, *exploiting the conjugacy* among $s_1, \ldots, s_n$ we obtain

$$A^{-1} = \sum_{i=1}^{n} \frac{s_i s_i^T}{s_i^T A s_i}.$$

Similarly to [2], [3] and [4] the above conclusions suggest the possibility to use the NCG in order to convey information on the inverse Hessian matrix, using the search directions $\{p_k\}$ and the vectors $\{y_k\}$. The next section details some basic aspects of our two proposals, following the guidelines of the above considerations.

Note that in [2] the preconditioner is built starting from the identity matrix, then adding the sum of specific dyads obtained from the NCG at different iterations. On the contrary, our second proposal in the present paper sets the new preconditioner starting from the previous one, and using information by the NCG just from the current iteration.

OUR TWO PROPOSALS FOR PRECONDITIONER UPDATES, BASED ON L–BFGS

For the sake of brevity here we simply report our two approaches, which provide preconditioners to the NCG method, aiming at approximating the inverse Hessian matrix of the objective function $f(x)$. The first proposal is based on the recurrence (with H_0 positive definite)

$$H_{k+1} = H_k + \gamma_k v_k v_k^T + \omega_k \frac{p_k p_k^T}{y_k^T p_k}, \qquad \gamma_k, \omega_k \in \mathbb{R} \setminus \{0\}, \qquad v_k \in \mathbb{R}^n, \tag{2}$$

where p_k is the search direction calculated by the NCG at step k, and the vector v_k is computed such that the secant condition $H_{k+1} y_k = s_k$ (similarly to SR1 quasi–Newton updates) is imposed. After some easy computation we obtain for v_k the expression $v_k = s_k - H_k y_k - \omega_k p_k$. Moreover, the next result holds for the update (2) (the proof is omitted for the sake of brevity).

Proposition 1 *Assume that f is the quadratic function $f(x) = \frac{1}{2} x^T A x - b^T x$, where $A \in \mathbb{R}^{n \times n}$ is symmetric and $b \in \mathbb{R}^n$. Suppose that k steps of the (unpreconditioned) CG are performed, in order to detect the stationary point (if any) of the function f, and that the vectors $p_1, \ldots, p_k$ are generated. Then, the matrix H_{k+1} in (2) satisfies the secant equations*

$$H_{k+1} y_j = s_j, \qquad j = 1, \ldots, k, \tag{3}$$

provided that the nonzero coefficients γ_j, ω_j, $j = 1, \ldots, k$ are computed such that

$$\gamma_j = \frac{1}{s_j^T y_j - y_j^T H_j y_j - \omega_j p_j^T y_j}, \qquad \omega_j \neq \frac{s_j^T y_j - y_j^T H_j y_j}{p_j^T y_j}, \qquad j = 1, \ldots, k.$$

Finally, $H_{n+1} = A^{-1}$.

The Proposition 1 specifically highlights in which sense the update (2) *tends to approximate* the inverse Hessian matrix in the quadratic case; moreover, the same formula can be adopted also for general nonlinear (and possibly nonconvex) functions. The main inconvenience of (2) is that the matrix H_{k+1} might be rarely indefinite, i.e. in some specific cases the positive definiteness of H_{k+1} may be hardly imposed. A possible remedy to the latter drawback is pursued by replacing (2) with the update (see also [2])

$$H_{k+1} = \tau_k C_k + \gamma_k v_k v_k^T + \omega_k \sum_{j=k-m}^{k} \frac{s_j s_j^T}{y_j^T s_j}, \tag{4}$$

where $0 \leq m \leq k-1$ represents a *memory of the preconditioner*, $\gamma_k, \omega_k \geq 0$, $\tau_k > 0$, $C_k \in \mathbb{R}^{n \times n}$ is symmetric positive definite and $v_k \in \mathbb{R}^n$. The vector v_k is computed following guidelines similar to those adopted for (2), while C_k satisfies at step k

$$C_k = \sigma_k I_n, \qquad \sigma_k \in \mathbb{R},$$

and σ_k solves the least squares subproblem

$$\min_{\sigma} \|(\sigma I_n) y_k - s_k\|^2 ,$$

yielding $C_k = s_k^T y_k / \|y_k\|^2 I_n$. We remark that now a suitable choice of $\gamma_k > 0$ and $\omega_k > 0$ always exists such that in (4) we obtain that H_{k+1} is positive definite.

Our second proposal is based on an updating formula which is more similar to L-BFGS, since it adopts the rank-2 update (H_0 positive definite)

$$H_{k+1} = \delta_k H_k + \gamma_k v_k v_k^T + \omega_k \frac{p_k p_k^T}{y_k^T p_k}, \qquad \gamma_k, \omega_k \in \mathbb{R} \setminus \{0\}, \qquad v_k \in \mathbb{R}^n. \tag{5}$$

However, now the matrix H_{k+1} is also assumed to satisfy the *modified secant equations*

$$\begin{cases} H_{k+1} y_j = \delta_j s_j, \qquad \delta_j > 0, \qquad \text{for all} \qquad j < k, \\ H_{k+1} y_k = s_k. \end{cases} \tag{6}$$

We incidentally observe that the latter appealing property is satisfied by all the updates of the Broyden class, with $\delta_j = 1$, for any $j \geq 1$, provided that the linesearch adopted is exact (see e.g. [1]). Following the idea in Proposition 1, the next interesting result can be proved for the update (5).

Proposition 2 *Let f be a nonlinear twice continuously differentiable function. Suppose that the NCG method is used to minimize the function f. Suppose that at current step k, H_k is positive definite and set*

$$\begin{aligned} 0 &< \delta_k = (1 - \varepsilon_k) \frac{s_k^T y_k}{y_k^T H_k y_k}, \\ 0 &< \omega_k < \varepsilon_k \alpha_k, \\ 0 &< \gamma_k = \frac{1}{(\varepsilon_k \alpha_k - \omega_k) p_k^T y_k}. \end{aligned}$$

with $\varepsilon_k \in (0, 1)$. Then, H_{k+1} is positive definite and satisfies the modified secant equations

$$\begin{cases} H_{k+1} y_j = \delta_j s_j, \qquad \delta_j > 0, \qquad \text{for all} \qquad j < k, \\ H_{k+1} y_k = s_k. \end{cases}$$

Since H_0 is positive definite, this proposition ensures that H_k, for $k \geq 1$, is positive definite, thus overcoming the drawback of (2). The prize to pay is that now the secant equations in (3) are not satisfied since they are replaced by the weaker conditions (6).

The effectiveness and the robustness highlighted by the use of the preconditioners proposed suggest promising guidelines for further investigations on both the approaches.

ACKNOWLEDGMENTS

G. Fasano thanks the National Research Council - Marine Technology Research Institute (CNR-INSEAN), Italy, for the support received. The work of G. Fasano is partially supported by the Italian Flagship Project RITMARE, coordinated by the Italian National Research Council (CNR) and funded by the Italian Ministry of Education, within the National Research Program 2012-2016.

REFERENCES

[1] J. Nocedal and S. Wright, *Numerical Optimization - second edition* (Springer, New York, 2006).
[2] A. Caliciotti, G. Fasano, and M. Roma, Optimization Letters (2016), 10.1007/s11590-016-1060-2.
[3] G. Fasano and M. Roma, Computational Optimization and Applications **38**, 81–104 (2007).
[4] G. Fasano and M. Roma, Computational Optimization and Applications **56**, 253–290 (2013).
[5] W. Hager and H. Zhang, Pacific Journal of Optimization **2**, 35–58 (2006).
[6] L. Nazareth, SIAM Journal on Numerical Analysis **16**, 794–800 (1979).