Annals of the Academy of Romanian Scientists
Series on Science and Technology and Information
Online ISSN **2066 - 8562** Volume **15**, Number 1-2/**2022**

CONJUGATE GRADIENT WITH SUBSPACE MINIMIZATION BASED ON CUBIC REGULARIZATION MODEL OF THE MINIMIZING FUNCTION

Neculai ANDREI[1]

Abstract. *A new algorithm for unconstrained optimization based on the cubic regularization in two dimensional subspace is developed. Different strategies for search direction are also discussed. The stepsize is computed by means of the weak Wolfe line search. Under classical assumptions it is proved that the algorithm is convergent. Intensive numerical experiments with 800 unconstrained optimization test functions with the number of variables in the range [1000 - 10,000] show that the suggested algorithm is more efficient and more robust than the well established conjugate gradient algorithms CG-DESCENT, CONMIN and L-BFGS (m=5). Comparisons of the suggested algorithm versus CG-DESCENT for solving five applications from MINPACK-2 collection, each of them with 40,000 variables, show that CUBIC is 3.35 times faster than CG-DESCENT.*

Keywords: unconstrained optimization, acceleration, regularization, conjugate gradient methods, applications
Mathematics Subject Classification (2000) 49M37. 90C30
DOI https://doi.org/10.56082/annalsarsciinfo.2022.1-2.28

1. Introduction

For solving the unconstrained optimization problem

$$\min f(x), \tag{1}$$

where $f:\mathbb{R}^n \to \mathbb{R}$ is continuously differentiable and bounded from below, besides the well known line-search and trust-region methods, the p-regularization model is constructed by adding a p-th regularization term to the quadratic estimation of f. The idea is to construct and minimize a local quadratic approximation of the minimizing function with a weighted regularization term $(\sigma_k / p)\|x\|^p$, $p > 2$. The most common choice to regularize the quadratic approximation is the p-regularization with $p = 3$, which is known as the *cubic regularization*. The idea of using the cubic regularization into the context of the Newton method first appeared in Griewank (1981) and was later

[1] Dr. Neculai Andrei is full member of Academy of Romanian Scientists, Str. Ilfov, nr. 3, sector 5, București, Romania, Center for Advanced Modeling and Optimization
E-mail: neculaiandrei70@gmail.com

developed by many authors, proving its convergence and complexity (for example see: (Nesterov, & Polyak, 2006), (Cartis, Gould, & Toint, 2011a, 2011b), (Gould, Porcelli, & Toint, 2012), (Bianconcini, Liuzzi, Morini, & Sciandrone, 2013), (Bianconcini, & Sciandrone, 2016), (Hsia, Sheu, & Yuan, 2017)). Griewank proved that any accumulation point of the sequence generated by minimizing the p-regularized subproblem is a second-order critical point of f, i.e., a point $\bar{x} \in \mathbb{R}^n$ satisfying $\nabla f(\bar{x}) = 0$ and $\nabla^2 f(\bar{x})$ semipositive definite. Later, Nesterov and Polyak (2006) proved that the cubic regularization method has a better global iteration complexity bound than the one for the steepest descent method. Based on these results, Cartis, Gould and Toint (2011a, 2011b) proposed an adaptive cubic regularization method for minimizing the function f, where the sequence of the regularization parameter $\{\sigma_k\}$ is dynamically determined and the p-regularized subproblems are inexactly solved. In the adaptive cubic regularization method, the minimizing function f is approximated by the model

$$m_k(d) = f(x_k) + g_k^T d + \frac{1}{2} d^T B_k d + \frac{1}{3} \sigma_k \|d\|^3, \tag{2}$$

where σ_k is a positive parameter (regularization parameter) dynamically updated in a specific way and B_k is an approximation to the Hessian of the objective function. The adaptive cubic regularization method for the unconstrained optimization was further developed by Bianconcini, Liuzzi, Morini and Sciandrone, (2013). The idea was to compute the trial step as a suitable approximate minimizer of the above cubic model of the minimizing function by using the nonmonotone globalization techniques of Grippo and Sciandrone (2002). Another approach was presented by Gould, Porcelli and Toint (2012), who presented new updating strategies for the regularization parameter σ_k based on interpolation techniques, which improved the overall numerical performance of the algorithm. New subspace minimization conjugate gradient methods based on p-regularization models, with $p = 3$ and $p = 4$, were developed by Zhao, Liu and Liu (2019). A complete theory of the p-regularized subproblems for $p > 2$, including the solution of these problems was presented by Hsia, Sheu and Yuan (2017).

This paper develops a variant of the conjugate gradient algorithm with subspace minimization ((Stoer, & Yuan, 1995), (Andrei, 2014), (Li, Liu, & Liu, 2019)) based on the regularization model (Zhao, Liu, & Liu, 2019). Starting with an initial guess x_0 and the initial search direction $d_0 = -g_0$, in our algorithm the next iteration is computed as $x_{k+1} = x_k + \alpha_k d_k$, where the search direction is computed as a linear combination of the current gradient and the previous search direction, while the stepsize α_k is determined by the standard Wolfe line search:

$$f(x_k + \alpha_k d_k) - f(x_k) \le \rho \alpha_k g_k^T d_k, \tag{3}$$

$$g_{k+1}^T d_k \geq \sigma g_k^T d_k , \tag{4}$$

where it is supposed that d_k is a descent direction and the scalar parameters ρ and σ are so that $0 < \rho \leq \sigma < 1.$ The algorithm combines the minimization of a p-regularized model (2) of the minimizing function with the subspace minimization. The main objective is to elaborate numerical algorithms based on the p-regularized model (2) with inexact line searches in which the search direction is a linear combination of the steepest descent direction and the previous search direction. If the minimizing function is close to a quadratic, then a quadratic approximation model in a two-dimensional subspace is minimized to generate the search direction, otherwise a p-regularization model is minimized.

The structure of the paper is as follows. Based on the theoretical developments by (Hsia, Sheu and Yuan (2017), Section 2 presents the $p-$regularized subproblem with a scaled norm and its solution in closed form. Section 3 presents the p-regularized subproblem in two-dimensional subspace, where the search direction is computed as a linear combination of the current gradient and the previous search direction. Using the developments from the previous sections, the strategies for search direction computation are presented in Section 4. The corresponding algorithm and its convergence are shown in Section 5 and Section 6, respectively. Section 7 shows the numerical performances of the algorithm on solving a set of 80 unconstrained optimization problems with different complexities. For each problem from this collection, 10 numerical experiments with an increasing number of variables as $n = 1000, 2000, \ldots, 10000$ have been performed. Hence, 800 problems have been solved in this set of numerical experiments.

2. The p-regularized subproblem with a scaled norm

The general form of the p-regularized subproblem is

$$\min_{x\in\mathbb{R}^n} h(x) = c^T x + \frac{1}{2} x^T Bx + \frac{\sigma}{p}\|x\|^p , \tag{5}$$

where $p > 2,$ $\sigma > 0,$ $c \in \mathbb{R}^n$ and $B \in \mathbb{R}^{n\times n}$ is a symmetric matrix. Because of the regularization term $\sigma\|x\|^p / p$ it follows that $h(x)$ is a coercive function, that is, $\lim_{\|x\|\to\infty} h(x) = +\infty,$ i.e. the p-regularized subproblem can always attain the global minimum, even for non-positive definite $B.$ The solution of this subproblem is given by the following theorem, (Hsia, Sheu and Yuan (2017)).

Theorem 2.1 *For $p > 2$ the point x^* is a global minimizer of (11.76) if and only if*

$$\left(B + \sigma\left\|x^*\right\|^{p-2} I\right)x^* = -c, \qquad B + \sigma\left\|x^*\right\|^{p-2} I \geq 0. \tag{6}$$

Moreover, the l_2 *norms of all global minimizers are equal.* ♦

Another form of the p -regularized subproblem with a scaled norm can be

$$\min_{x\in\mathbb{R}^n} h(x) = c^T x + \frac{1}{2}x^T Bx + \frac{\sigma}{p}\|x\|_A^p, \tag{7}$$

where $A \in \mathbb{R}^{n\times n}$ is a symmetric and positive definite matrix and $\|x\|_A = \sqrt{x^T Ax}$, known as l_A norm. Considering $y = A^{1/2}x$, (7) can be rewritten as

$$\min_{y\in\mathbb{R}^n} h(y) = (A^{-1/2}c)^T y + \frac{1}{2}y^T(A^{-1/2}BA^{-1/2})y + \frac{\sigma}{p}\|y\|^p. \tag{8}$$

From Theorem 2.1 the point y^* is a global minimizer of (8) if and only if

$$\left(A^{-1/2}BA^{-1/2} + \sigma\left\|y^*\right\|^{p-2} I\right)y^* = -A^{-1/2}c, \tag{9a}$$

$$A^{-1/2}BA^{-1/2} + \sigma\left\|y^*\right\|^{p-2} I \geq 0. \tag{9b}$$

The following theorem presents the global solution of the p -regularized subproblem (7) (Andrei (2020)).

Theorem 2.2 *The point* x^* *is a global minimizer of the* p *-regularized subproblem with a scaled norm (7) for* $p > 2$ *if and only if*

$$\left(B + \sigma(z^*)^{p-2}A\right)x^* = -c, \qquad B + \sigma(z^*)^{p-2}A \geq 0, \tag{10}$$

where z^* *is the unique non-negative root of the equation*

$$z^2 - \sum_{i=1}^{n}\frac{\beta_i^2}{(\mu_i + \sigma z^{p-2})^2} = 0. \tag{11}$$

Moreover, the l_A *norms of all global minimizers are equal.*
♦

In the following, let us consider the case in which B is symmetric and positive definite and $A = B$. In this case, since $\sigma > 0$ and $z \geq 0$, it follows that $B + \sigma(z^{p-2})B$ is always a positive definite matrix. Therefore, the global minimizer of the p-regularized subproblem with a scaled norm (7) is unique. In conclusion, from (10) the following theorem is true.

Theorem 2.4 *Let* $B > 0$ *and* $A = B$, *then the point*

$$x^* = \frac{-1}{1+\sigma(z^*)^{p-2}} B^{-1} c \tag{12}$$

is the only global minimizer of (9) for $p > 2$ *where* z^* *is the unique non-negative solution of the equation*

$$\sigma z^{p-1} + z - \sqrt{c^T B^{-1} c} = 0. \tag{13}$$

♦

Concerning the equation (13), observe that for $c = 0$ the equation is $z\left(\sigma z^{p-2}+1\right) = 0.$ Since $\sigma > 0$ it follows that $z^* = 0$ is the unique non-negative solution of (13). On the other hand, for $c \neq 0$ defining the function $\varphi(z) = \sigma z^{p-1} + z - \sqrt{c^T B^{-1} c}$, it is easy to see that $\varphi'(z) = \sigma(p-1)z^{p-2} + 1 > 0$, which proves that $\varphi(z)$ is monotonically increasing. Since $\varphi(0) < 0$ and $\varphi(\sqrt{c^T B^{-1} c}) > 0$, it follows that z^* is the unique positive solution of (13).

3. The p-regularized subproblem in two-dimensional subspace

Consider the quadratic approximation of f in x_{k+1} as

$$\overline{h}_{k+1}(d) = g_{k+1}^T d + \frac{1}{2} d^T B_{k+1} d,$$

where B_{k+1} is a symmetric and positive definite approximation to the Hessian of f in x_{k+1} which satisfies the secant equation $B_{k+1}s_k = y_k$, with $s_k = x_{k+1} - x_k$ and $y_k = g_{k+1} - g_k$. Consider that g_{k+1} and s_k are two linearly independent vectors and define $\Omega_k = \{d_{k+1} : d_{k+1} = \mu_k g_{k+1} + \eta_k s_k\}$, where μ_k and η_k are real scalars. The corresponding p-regularized subproblem is defined as

$$\min_{d_{k+1}\in\Omega_k} h_{k+1}(d_{k+1}) = g_{k+1}^T d_{k+1} + \frac{1}{2} d_{k+1}^T B_{k+1} d_{k+1} + \frac{\sigma_k}{p} \left\| d_{k+1} \right\|_{B_{k+1}}^p, \tag{14}$$

where $\sigma_k > 0$ is the regularized parameter. Having in view that $d_{k+1} \in \Omega_k$ the p-regularized subproblem in the two-dimensional subspace can be expressed as

$$\min_{\mu_k,\eta_k\in\mathbb{R}} \begin{pmatrix} \|g_{k+1}\|^2 \\ g_{k+1}^T s_k \end{pmatrix}^T \begin{pmatrix} \mu_k \\ \eta_k \end{pmatrix} + \frac{1}{2}\begin{pmatrix} \mu_k \\ \eta_k \end{pmatrix}^T M_k \begin{pmatrix} \mu_k \\ \eta_k \end{pmatrix} + \frac{\sigma_k}{p}\left\|\begin{pmatrix} \mu_k \\ \eta_k \end{pmatrix}\right\|_{M_k}^p, \tag{15}$$

where

$$M_k = \begin{bmatrix} \rho_k & g_{k+1}^T y_k \\ y_k^T g_{k+1} & s_k^T y_k \end{bmatrix}, \quad \rho_k = g_{k+1}^T B_{k+1} g_{k+1}. \tag{16}$$

Observe that M_k is a symmetric and positive definite matrix since B_{k+1} is symmetric and positive definite and the vectors g_{k+1} and s_k are linear independent.
By Theorem 2.4, it follows that the unique solution of (15) is

$$\begin{pmatrix} \mu_k^* \\ \eta_k^* \end{pmatrix} = \frac{-1}{1+\sigma_k (z^*)^{p-2}} M_k^{-1} \begin{pmatrix} \|g_{k+1}\|^2 \\ g_{k+1}^T s_k \end{pmatrix}, \tag{17}$$

where z^* is the unique non-negative solution of the equation

$$\sigma_k z^{p-1} + z - \sqrt{\begin{pmatrix} \|g_{k+1}\|^2 \\ g_{k+1}^T s_k \end{pmatrix}^T M_k^{-1} \begin{pmatrix} \|g_{k+1}\|^2 \\ g_{k+1}^T s_k \end{pmatrix}} = 0. \tag{18}$$

Observe that

$$u \equiv \sqrt{\begin{pmatrix} \|g_{k+1}\|^2 \\ g_{k+1}^T s_k \end{pmatrix}^T M_k^{-1} \begin{pmatrix} \|g_{k+1}\|^2 \\ g_{k+1}^T s_k \end{pmatrix}} = \sqrt{\frac{1}{\Delta_k}(\|g_k\|^4 (s_k^T y_k) + \rho_k (g_{k+1}^T s_k))},.$$

where $\Delta_k = \rho_k (s_k^T y_k) - (g_{k+1}^T y_k)^2$ is the determinant of M_k. Therefore, the unique non-negative solution of the equation (18) is:

$$z^* = \frac{2u}{1+\sqrt{1+4\sigma_k u}}. \tag{19}$$

Denote

$$\delta_k = \frac{1}{1+\sigma_k (z^*)^{p-2}}. \tag{20}$$

Therefore, from (17) the solution of the p-regularized subproblem in the two-dimensional subspace (15) is

$$\mu_k^* = \frac{\delta_k}{\Delta_k}\left[(g_{k+1}^T y_k)(g_{k+1}^T s_k) - (s_k^T y_k)\|g_{k+1}\|^2\right], \tag{21a}$$

$$\eta_k^* = \frac{\delta_k}{\Delta_k}\left[(g_{k+1}^T y_k)\|g_{k+1}\|^2 - \rho_k (g_{k+1}^T s_k)\right]. \tag{21b}$$

For the ρ_k computation some procedures are known. One of them, given by Stoer and Yuan (1995), is

$$\rho_k = 2\frac{(g_{k+1}^T y_k)^2}{s_k^T y_k}. \tag{22}$$

Using the Barzilai-Borwein method, another procedure for the ρ_k computation was given by Dai and Kou (2016)

$$\rho_k = \frac{3}{2}\frac{\|y_k\|^2}{s_k^T y_k}\|g_{k+1}\|^2. \tag{23}$$

Another simple way is to let B_{k+1} be a self-scaling memoryless BFGS with parameter τ_k given as

$$\rho_k = g_{k+1}^T\left[\tau_k I - \tau_k \frac{s_k s_k^T}{\|s_k\|^2} + \frac{y_k y_k^T}{s_k^T y_k}\right] g_{k+1}, \tag{24}$$

with $B_k = I\tau_k,$, where τ_k can be chosen as $\tau_k^{OS} = \|y_k\|^2 / y_k^T s_k$, given by Oren and Spedicato (1976), or $\tau_k^{OL} = y_k^T s_k / \|s_k\|^2$ given by Oren and Luenberger (1974).

For the σ_k computation there are a number of procedures. For example Cartis, Gould and Toint (2011a) suggested a procedure based on the trust-region ratio. Another procedure using an interpolation condition was given by Zhao, Liu and Liu (2019). In our algorithm let us define

$$r_k = \frac{f(x_k) - f(x_{k+1})}{f(x_k) - h_{k+1}(s_k)}, \tag{25}$$

which measures the actual decrease in the objective function $f(x_k) - f(x_{k+1})$ versus the predicted model decrease $f(x_k) - h(s_k)$. The regularized parameter σ_k is updated as follows:

$$\sigma_{k+1} = \begin{cases} \max\{\min\{\sigma_k, \|g_{k+1}\|\}, \varepsilon_M\}, & \text{if } r_k > \lambda_2, \\ \sigma_k + \|g_{k+1}\|^2, & \text{if } \lambda_1 \le r_k \le \lambda_2, \\ 3\left|f_k - f_{k+1} + s_k^T g_{k+1} - 0.5 y_k^T s_k\right| / (y_k^T s_k)^{p/2}, & \text{otherwise}, \end{cases} \tag{26}$$

where $\sigma_0 = 1$, ε_M is the relative machine precision, $\lambda_1 = 10^{-5}$ and $\lambda_2 = 0.5$. Of course, this is a suggestion which proved to be successful in our numerical experiments, but some other proposals may be considered as well.

4. Strategies for search direction computation

In our algorithm, if the objective function is close to a quadratic, then a quadratic approximation model in a two-dimensional subspace is used to generate the search direction, otherwise a p-regularization model in a two-dimensional subspace is to be considered. Indeed, to see how the function $f(x)$ is close to a quadratic function on the line segment connecting x_{k-1} and x_k, Yuan (1991) introduced the parameter

$$t_k = \left| \frac{2(f_{k-1} - f_k + g_k^T s_{k-1})}{s_{k-1}^T y_{k-1}} - 1 \right|. \tag{27}$$

On the other hand, the ratio

$$\theta_k = \frac{f_{k-1} - f_k}{0.5 s_{k-1}^T y_{k-1} - g_k^T s_{k-1}} \tag{28}$$

shows the difference between the actual reduction of the function values and the predicted reduction given by the quadratic model.

The strategy for using the quadratic approximation or the p-regularization model of the minimizing function is as follows. If the conditions

$$t_k \le c_1 \quad \text{or} \quad |\theta_k - 1| \le c_2 \tag{29}$$

hold, where c_1 and c_2 are positives constants ($c_1 = 10^{-4}$ and $c_2 = 10^{-5}$), then the function $f(x)$ might be very close to a quadratic on the line segment connecting x_{k-1} and x_k. In this case, for the search direction, the quadratic approximation model in a two-dimensional subspace is selected, which corresponds to (15) with $\sigma_k = 0$. Therefore, in our algorithm the parameters μ_k and η_k which define the search direction d_{k+1} are computed as

$$\mu_k^* = \frac{1}{\Delta_k}\left[(g_{k+1}^T y_k)(g_{k+1}^T s_k) - (s_k^T y_k)\|g_{k+1}\|^2 \right], \tag{30a}$$

$$\eta_k^* = \frac{1}{\Delta_k}\left[(g_{k+1}^T y_k)\|g_{k+1}\|^2 - \rho_k (g_{k+1}^T s_k) \right], \tag{30b}$$

where ρ_k is computed as in (23). On the other hand, if $t_k > c_1$ and $|\theta_k - 1| > c_2$, then the parameters μ_k and η_k which define the search direction d_{k+1} are computed as in (21), where ρ_k and σ_k are computed as in (23) and (26), respectively.

Of course, some other variants of the algorithm may have regard for it, where for the computation of $\rho_k = g_{k+1}^T B_{k+1} g_{k+1}$, (22) or (24) may be used. In our numerical experiments, the variant proposed by Dai and Kou (2016) given by (23) proved to be the most efficient. The other crucial ingredient of our algorithm is the computation of the regularization parameter σ_k. Here, σ_k is computed as in (26), but some other strategies may be implemented. For example, Cartis, Gould and Toint (2011a) proposed a procedure for this parameter computation by analogy with the trust-region method. In such a framework, σ_k could be regarded as the reciprocal of the trust-region radius. Thus, σ_k is increased if insufficient decrease is obtained, but it is decreased or unchanged otherwise. Other procedures for updating the regularization parameter σ_k for minimizing the p-regularization model is discussed by Gould, Porcelli and Toint (2012). However, finding a global minimizer of the $h_{k+1}(.)$ defined by (12) may not be essential in practice. Therefore, the global minimization problem (12) of the p-regularized subproblem may be relaxed by letting d_{k+1} be an approximation to such a minimizer.

5. Algorithm CUBIC

With these developments, taking into consideration the acceleration scheme (Andrei, 2006, 2009), according to the value of the parameter „acceleration" (true or false), the following algorithms CUBIC and CUBICa may be presented. Clearly, CUBICa is the accelerated version of CUBIC.

Algorithm CUBIC / CUBICa

1.	Select a starting point $x_0 \in dom\, f$ and compute: $f_0 = f(x_0)$ and $g_0 = \nabla f(x_0)$. Select $\varepsilon_A > 0$ sufficiently small and positive values $0 < \rho < \sigma < 1$ used in Wolfe line search conditions. Select some positive values for: $c_1, c_2, \lambda_1, \lambda_2$. Set $d_0 = -g_0$ and $k = 0$
2.	Test a criterion for stopping the iterations. If the test is satisfied, then stop; otherwise continue with step 3
3.	Using the Wolfe line search conditions (3) and (34 determine the stepsize α_k.

	Update the variables $x_{k+1} = x_k + \alpha_k d_k$. Compute f_{k+1}, g_{k+1} and $s_k = x_{k+1} - x_k$, $y_k = g_{k+1} - g_k$
4.	If *acceleration* equal true, then a) Compute: $z = x_k + \alpha_k d_k$, $g_z = \nabla f(z)$ and $y_k = g_k - g_z$ b) Compute: $a_k = \alpha_k g_k^T d_k$, and $b_k = -\alpha_k y_k^T d_k$ c) If $\lvert b_k \rvert \geq \varepsilon_A$, then compute $\xi_k = -a_k / b_k$ and update the variables as $x_{k+1} = x_k + \xi_k \alpha_k d_k$. Compute f_{k+1} and g_{k+1}. Compute $s_k = x_{k+1} - x_k$ and $y_k = g_{k+1} - g_k$
5.	If $t_k > c_1$ and $\lvert \theta_k - 1 \rvert > c_2$, then compute z_k and δ_k by (19) and (20) respectively. The search direction is computed as $d_{k+1} = \mu_k g_{k+1} + \eta_k s_k$, where the parameters μ_k and η_k are computed as in (21), with ρ_k and σ_k are computed as in (23) and (26), respectively
6.	If $t_k \leq c_1$ or $\lvert \theta_k - 1 \rvert \leq c_2$, then the search direction is computed as $d_{k+1} = \mu_k g_{k+!} + \eta_k s_k$, where the parameters μ_k and η_k are computed as in (30), with ρ_k computed as in (23)
7.	Restart criterion. If $\lvert g_{k+1}^T g_k \rvert > 0.2 \lVert g_{k+1} \rVert^2$ then set $d_{k+1} = -g_{k+1}$
8.	Consider $k = k+1$ and go to step 2 ♦

This is a variant of the subspace minimization conjugate gradient algorithm based on the cubic regularization model of the unconstrained optimization problem. Some other variants may be generated by selecting different procedures for ρ_k and σ_k computation, as well as for the restarting criterion.

6. Convergence analysis

Assume that:

(i) *The level set* $S = \{x \in \mathbb{R}^n : f(x) \leq f(x_0)\}$ *is bounded, i.e. there exists a constant* $B > 0$ *so that* $\lVert x \rVert \leq B$ *for all* x *in the level set.*

(ii) *In some neighborhood* N *of the level set,* f *is continuously differentiable and its gradient is Lipschitz continuous, i.e. there exists a constant* $L > 0$ *so that*

$$\lVert g(x) - g(y) \rVert \leq L \lVert x - y \rVert, \textit{ for all } x, y \in N.$$

Suppose that the search direction d_{k+1} in the CUBIC algorithm is calculated under the following conditions

$$\xi_1 \leq \frac{s_k^T y_k}{\lVert s_k \rVert^2} \leq \frac{\lVert y_k \rVert^2}{s_k^T y_k} \leq \xi_2, \tag{31}$$

where ξ_1 and ξ_2 are positive constants ($\xi_1 = 10^{-7}$, $\xi_2 = 10^5$). For general nonlinear functions, if (31) holds, then the condition number of the Hessian of the minimizing function might not be very large. In this case, both the quadratic and the p-regularization models may be used.

Proposition 6.1 *Under the conditions* (31) *the search direction* $d_{k+1} = \mu_k g_{k+1} + \eta_k s_k$, *where the parameters* μ_k *and* η_k *are computed as in* (21), *with* ρ_k *and* σ_k *computed as in* (23) *and* (26) *respectively, satisfies the sufficient descent condition* $g_{k+1}^T d_{k+1} \le -\bar{c}\left\|g_{k+1}\right\|^2$, *where* $\bar{c}$ *is a positive constant.*

Proof Since $\sigma_k \ge 0$ and $z^* \ge 0$, it follows that $\delta_k < 1$. Therefore,

$$g_{k+1}^T d_{k+1} \le -\frac{\left\|g_{k+1}\right\|^4}{\Delta_k}\left[(s_k^T y_k) - 2(g_{k+1}^T y_k)\frac{g_{k+1}^T s_k}{\left\|g_{k+1}\right\|^2} + \rho_k\left(\frac{g_{k+1}^T s_k}{\left\|g_{k+1}\right\|^2}\right)^2\right].$$

Denote the term in the square brackets of the above inequality by χ_k and consider it as a function of the variable $g_{k+1}^T s_k / \left\|g_{k+1}\right\|^2$. Now, taking minimization of χ_k it follows that $\chi_k \ge \Delta_k / \rho_k$. Therefore, from (23) and since $s_k^T y_k / \left\|y_k\right\|^2 \ge \xi_2^{-1}$, it follows that

$$g_{k+1}^T d_{k+1} \le -\frac{\left\|g_{k+1}\right\|^4}{\rho_k} = -\frac{2}{3}\frac{s_k^T y_k}{\left\|y_k\right\|^2}\left\|g_{k+1}\right\|^2 \le -\frac{2}{3\xi_2}\left\|g_{k+1}\right\|^2 = -\bar{c}\left\|g_{k+1}\right\|^2,$$

where $\bar{c} = 2/(3\xi_2)$. ♦

Proposition 6.2 *Under the conditions* (31) *the search direction* $d_{k+1} = \mu_k g_{k+1} + \eta_k s_k$, *where the parameters* μ_k *and* η_k *are computed as in* (21), *with* ρ_k *and* σ_k *computed as in* (23) *and* (26) *respectively, satisfies* $\left\|d_{k+1}\right\| \le \tilde{c}\left\|g_{k+1}\right\|$, where $\tilde{c}$ is a positive constant.

Proof Firstly, from (31) and (23) the following lower bound of Δ_k is obtained

$$\Delta_k = \rho_k(s_k^T y_k) - (g_{k+1}^T y_k)^2 = (s_k^T y_k)\left(\rho_k - \frac{(g_{k+1}^T y_k)^2}{(s_k^T y_k)}\right)$$

$$\ge \xi_1\left\|s_k\right\|^2\left(\rho_k - \frac{(g_{k+1}^T y_k)^2}{(s_k^T y_k)}\right) \ge \frac{1}{2}\xi_1\left\|s_k\right\|^2\frac{\left\|y_k\right\|^2}{s_k^T y_k}\left\|g_{k+1}\right\|^2. \qquad (32)$$

From the triangle inequality, the Cauchy-Schwarz inequality, (23), (32) and since $\delta_k < 1$ it follows that

$$\begin{aligned}
\|d_{k+1}\| &= \|\mu_k g_{k+1} + \eta_k s_k\| \\
&\leq \frac{1}{\Delta_k}\left\|((g_{k+1}^T y_k)(g_{k+1}^T s_k) - (s_k^T y_k)\|g_{k+1}\|^2)g_{k+1} + ((g_{k+1}^T y_k)\|g_{k+1}\|^2 - \rho_k (g_{k+1}^T s_k))s_k\right\| \\
&\leq \frac{1}{\Delta_k}\left[\left|(g_{k+1}^T y_k)(g_{k+1}^T s_k) - (s_k^T y_k)\|g_{k+1}\|^2\right|\|g_{k+1}\| + \left|(g_{k+1}^T y_k)\|g_{k+1}\|^2 - \rho_k (g_{k+1}^T s_k)\right|\|s_k\|\right] \\
&\leq \frac{\|g_{k+1}\|^3}{\Delta_k}\left[3\|y_k\|\|s_k\| + \rho_k \frac{\|s_k\|^2}{\|g_{k+1}\|^2}\right] \\
&\leq \frac{2\|g_{k+1}\|(s_k^T y_k)}{\xi_1 \|s_k\|^2 \|y_k\|^2}\left[3\|y_k\|\|s_k\| + \rho_k \frac{\|s_k\|^2}{\|g_{k+1}\|^2}\right] \\
&\leq \|g_{k+1}\|\left[\frac{6(s_k^T y_k)}{\xi_1 \|s_k\|\|y_k\|} + \frac{3}{\xi_1}\right].
\end{aligned}$$

Now, from the Cauchy-Schwarz inequality it follows that

$$\|d_{k+1}\| \leq \frac{9}{\xi_1}\|g_{k+1}\| = \tilde{c}\|g_{k+1}\|,$$

where $\tilde{c} = 9/\xi_1$ is a positive constant. ♦

Theorem 6.1 *Suppose that the assumption (i) and (ii) hold. If the sequence* $\{x_k\}$ *is generated by the algorithm CUBIC, then*

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

Proof Firstly, observe that under the assumption *(i)* and *(ii)* from (4) it follows that

$$\alpha_k \geq \frac{1-\sigma}{L}\frac{\left|g_k^T d_k\right|}{\|d_k\|^2}.$$

Now, from (3),

$$f_{k+1} \leq f_k - \rho\frac{(1-\sigma)}{L}\frac{(g_k^T d_k)^2}{\|d_k\|^2}.$$

From Propositions 6.1 and 6.2, it follows that

$$f_{k+1} \le f_k - \rho \frac{(1-\sigma)\bar{c}^2}{L\tilde{c}^2}\|g_k\|^2.$$

Denote $\omega = \rho(1-\sigma)\bar{c}^2 / (L\tilde{c}^2).$ Therefore,

$$f_{k+1} \le f_k - \omega\|g_k\|^2.$$

By summing this expression over all indices less than or equal to $k,$ it follows that

$$f_{k+1} \le f_0 - \omega\sum_{i=0}^{k}\|g_i\|^2. \tag{33}$$

Since f is bounded from below, it results that $f_0 - f_{k+1}$ is less than some positive constant for all $k.$ Hence, by taking limits in (33) we get

$$\sum_{k=0}^{\infty}\|g_k\|^2 < \infty,$$

which concludes the proof. ♦

7. Numerical results

This section presents the performances of CUBIC (accelerated) for solving 800 unconstrained optimization problems (Andrei, 2018), with the number of variables $n = 1000, 2000, ..., 10000,$ as well as for solving five applications from MINPACK2 collection (Averik, Carter, Moré, & Xue, 1992) each of them with 40,000 variables. The algorithms compared in these numerical experiments find local solutions. Therefore, the comparisons of the algorithms are given in the following context. Let f_i^{ALG1} and f_i^{ALG2} be the optimal value found by ALG1 and ALG2 for problem $i = 1,\ldots,800,$ respectively. We say that, in the particular problem $i,$ the performance of ALG1 was better than the performance of ALG2 if

$$\left|f_i^{ALG1} - f_i^{ALG2}\right| < 10^{-3} \tag{34}$$

and if the number of iterations (#iter), or the number of function-gradient evaluations (#fg), or the CPU time of ALG1 was less than the number of iterations, or the number of function-gradient evaluations, or the CPU time corresponding to ALG2, respectively.

The iterations are stopped if the inequality $\|g_k\|_\infty \le 10^{-6}$ is satisfied, where $\|.\|_\infty$ is the maximum absolute component of a vector. All algorithms implement the standard Wolfe line search (3) and (4), where $\rho = 0.0001$ and $\sigma = 0.8.$ The maximum number of iterations was limited to 2000.

Figure 1 shows the Dolan and Moré performance profiles of CUBIC versus CG-DESCENT (Hager, & Zhang, 2005), DESCON (Andrei, 2023), CONMIN (Shanno, 1983) and LBFGS (m=5) (Liu, & Nocedal, 1989)

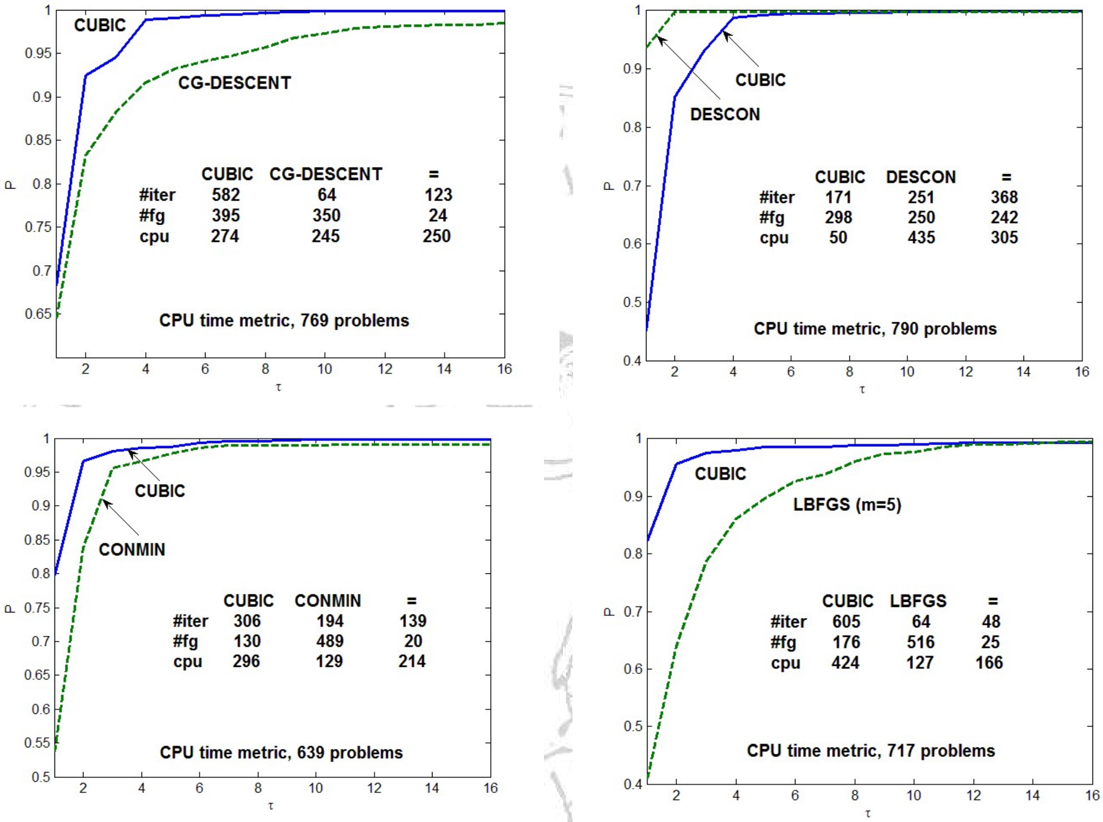

Fig. 1. Performance profiles of CUBIC versus CG-DESCENT, DESCON, CONMIN and LBFGS.

From Figure 1 CUBIC proves to be more efficient and more robust than CG-DESCENT, CONMIN. Note that all these algorithms implement the standard Wolfe line search (3) and (4) with the same values of the parameters ρ and σ.

Table 1 contains the performances of CUBIC and of CG-DESCENT for solving the applications from MINPACK-2, where each application has 40,000 variables, where *#iter* is the number of iterations, *#fg* is the number of functions evaluations and *cpu* is the CPU computing time to get the solution.

Table 1 Performances of CUBIC and CG-DESCENT for solving five applications from the MINPACK-2 collection

	CUBIC			**CG-DESCENT**		
	#iter	*#fg*	*cpu*	*#iter*	*#fg*	*cpu*
A1	241	510	3.16	323	647	9.67
A2	555	1145	8.11	788	1577	31.35
A3	1021	2070	24.39	1043	2088	64.96
A4	299	632	17.92	435	871	81.40
A5	284	588	5.23	286	573	9.89
Total	2400	4945	58.81	2875	5756	197.27

From Table 1 we see that CUBIC is 3.35 times faster than CG-DESCENT.

8. Conclusion

The $p-$regularization method is to construct and minimize a local quadratic approximation of the minimizing function with a weighted regularization term $(\sigma_k / p)\|x\|^p$, $p > 2$. The most used is $p = 3$, known as cubic regularization. In this paper we develop a variant of the conjugate gradient algorithm with subspace minimization based on the regularization model. In this algorithm the search direction is a linear combination of the steepest descent direction and the previous search direction. If the minimizing function is close to a quadratic, then a quadratic approximation model in a two-dimensional subspace is minimized to generate the search direction, otherwise a p-regularization model is minimized. This strategy proved to be very advantageous for enhancing the curvature properties of the minimizing functions. In mild conditions it is shown that the search directions generated by the algorithm satisfy the sufficient descent condition, and the search directions are bounded in norm. Numerical experiments with 800 large scale unconstrained optimization test functions with the number of variables in the range [1000 - 10,000] and with five applications from MINPACK-2 collection, each of them with 40,000 variables prove that CUBIC is more efficient and more robust versus known algorithms like CG-DESCENT, CONMIN and L-BFGS.

REFERENCES

[1.] Griewank, A., (1981). *The modification of Newton's method for unconstrained optimization by bounding cubic terms.* Technical Report NA/12, Department of Applied Mathematics and Theorerical Physics, University of Cambridge.

[2.] Nesterov, Y., & Polyak, B.T., (2006). Cubic regularization of Newton's method and its global performance. *Mathematical Programming, 108*, 177-205.

[3.] Cartis, C., Gould, N.I.M., & Toint, Ph.L., (2011a). Adaptive cubic overestimation methods for unconstrained optimization. Part I: motivation, convergence and numerical results. *Mathematical Programming Series A, 127*, 245–295.

[4.] Cartis, C., Gould, N.I.M., & Toint, Ph.L., (2011b). Adaptive cubic overestimation methods for unconstrained optimization. Part II: worst-case function-evaluation complexity. *Mathematical Programming Series A, 130*, 295–319.

[5.] Gould, N.I.M., Porcelli, M., & Toint, Ph.L., (2012). Updating the regularization parameter in the adaptive cubic regularization algorithm. *Computational Optimization and Applications, 53*, 1-22.

[6.] Bianconcini, T., Liuzzi, G., Morini, B., & Sciandrone, M., (2013). On the use of iterative methods in cubic regularization for unconstrained optimization. *Computational Optimization and Applications, 60*(1), 35-57.

[7.] Bianconcini, T., & Sciandrone, M., (2016). A cubic regularization algorithm for unconstrained optimization using line search and nonmonotone techniques. *Optimization Methods and Software, 31*, 1008-1035.

[8.] Hsia, Y., Sheu, R.L., & Yuan, Y.X. (2017). Theory and application of *p*-regularized subproblems for *p*>2. *Optimization Methods & Software*, *32*(5), 1059–1077.

[9.] Grippo, L., & Sciandrone, M., (2002). Nonmonotone globalization techniques for the Barzilai-Borwein gradient method. *Computational Optimization and Applications, 23*, 143–169.

[10.] Zhao, T., Liu, H., & Liu, Z., (2019). New subspace minimization conjugate gradient methods based on regularization model for unconstrained optimization. *Numerical Algorithms*, Optimization online, OO Digest: April 2020, http://www.optimization-online.org/DB_HTML/2020/04/7720.html

[11.] Stoer, J., & Yuan, Y.X., (1995). A subspace study on conjugate gradient algorithms. *ZAMM – Journl of Applied Mathematics and Mechanics, 75,* 69-77.

[12.] Andrei, N., (2014). An accelerated subspace minimization three-term conjugate gradient algorithm for unconstrained optimization. *Numerical Algorithms, 65*(4), 859-874.

[13.] Li, Y., Liu, Z., & Liu, H., (2019). A subspace minimization conjugate gradient method based on conic model for unconstrained optimization. *Computational and Applied Mathematics,* 38: 16. https://doi.org/10.1007/s40314-019-0779-7

[14.] Andrei, N. (2020). New conjugate gradient algorithms based on self-scaling memoryless Broyden–Fletcher–Goldfarb–Shanno method. *Calcolo* **57,** 17 https://doi.org/10.1007/s10092-020-00365-7

[15.] Dai, Y.H., & Kou, C.X., (2016). A Barzilai-Borwein conjugate gradient method. *Sci. China Math., 59*(8), 1511-1524.

[16.] Oren, S.S., & Spedicato, E., (1976). Optimal conditioning of self-scaling variable metric algorithm. *Mathematical Programming, 10*, 70-90.

[17.] Oren, S.S., & Luenberger, D.G., (1974). Self-scaling variable metric (SSVM) algorithms, part I: criteria and sufficient conditions for scaling a class of algorithms. *Management Science, 20*, 845-862.

[18.] Yuan, Y.X., (1991). A modified BFGS algorithm for unconstrained optimization. *IMA Journal of Numerical Analysis, 11*, 325-332.

[19.] Andrei, N., (2006). An acceleration of gradient descent algorithm with backtracking for unconstrained optimization. *Numerical Algorithms, 42*(1), 63-73.

[20.] Andrei, N., (2009). Acceleration of conjugate gradient algorithms for unconstrained optimization. *Applied Mathematics and Computation, 213*(2), 361-369.

[21.] Andrei, N., (2018). *UOP - A collection of 80 unconstrained optimization test problems*. (Technical Report No. 7/2018, November 17, Research Institute for Informatics, Bucharest, Romania).

[22.] Averick, B.M., Carter, R.G., Moré, J.J., & Xue, G.L., (1992). *The MINPACK-2 test problem collection.* (Mathematics and Computer Science Division, Argonne National Laboratory, 9700 South Cass Avenue, Argonne, Illinois, Preprint MCS-P153-6092, June 1992).

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

[24.] Andrei, N., (2013). Another conjugate gradient algorithm with guaranteed descent and conjugacy conditions for large-scale unconstrained optimization. *Journal of Optimization Theory and Applications, 159*, 159-182.

[25.] Shanno, D.F., (1983). *CONMIN – A Fortran subroutine for minimizing an unconstrained nonlinear scalar valued function of a vector variable x either by the BFGS variable metric algorithm or by a Beale restarted conjugate gradient algorithm*. Private communication, October 17, 1983.

[26.] Liu, D.C., & Nocedal, J., (1989). On the limited-memory BFGS method for large optimization. *Mathematical Programming, 45*, 503-528.

.