Ladislav Lukšan; Jan Vlček

Nonlinear conjugate gradient methods

In: Jan Chleboun and Petr Přikryl and Karel Segeth and Jakub Šístek and Tomáš Vejchodský (eds.): Programs and Algorithms of Numerical Mathematics, Proceedings of Seminar. Dolní Maxov, June 8-13, 2014. Institute of Mathematics AS CR, Prague, 2015. pp. 130--135.

Persistent URL: http://dml.cz/dmlcz/702674

Programs and Algorithms of Numerical Matematics 17
J. Chleboun, P. Přikryl, K. Segeth, J. Šístek, T. Vejchodský (Eds.)
Institute of Mathematics AS CR, Prague 2015

NONLINEAR CONJUGATE GRADIENT METHODS

Ladislav Lukšan, Jan Vlček

Institute of Computer Science, Academy of Sciences of the Czech Republic
Pod Vodárenskou věží, 182 07 Praha 8
luksan@cs.cas.cz, vlcek@cs.cas.cz

Abstract

Modifications of nonlinear conjugate gradient method are described and tested.

Conjugate gradient method is frequently used to solve the following problems

$$F(x) = \frac{1}{2}(x - x^*)^T A(x - x^*) \to \min, \quad \text{or} \quad A(x - x^*) = 0,$$

where A is a symmetric positive definite matrix. Linear and nonlinear conjugate gradient methods differ in line search and gradient evaluations

Linear CG :	**Nonlinear CG :**
$g_1 = g(x_1),\ \ s_1 = -g_1,$	$F_1 = F(x_1),\ \ g_1 = g(x_1),\ \ s_1 = -g_1,$
For $i = 1,\ 2\ \ \ldots$ **do**	**For** $i = 1,\ 2\ \ \ldots$ **do**
$\alpha_i = \|g_i\|^2 / s_i^T A s_i,$	$\alpha_i > 0$ and $s_i^T g(x_i + \alpha_i s_i) = 0,$
$x_{i+1} = x_i + \alpha_i s_i,$	$x_{i+1} = x_i + \alpha_i s_i,$
$g_{i+1} = g_i + \alpha_i A s_i,$	$F_{i+1} = F(x_{i+1}),\ \ g_{i+1} = g(x_{i+1}),$
If $\|g_{i+1}\| \le \varepsilon \|g_1\|$ **then stop**.	**If** $\|g_{i+1}\| \le \varepsilon \|g_1\|$ **then stop**,
$\beta_i = \|g_{i+1}\|^2 / \|g_i\|^2,$	$\beta_i = \|g_{i+1}\|^2 / \|g_i\|^2,$
$s_{i+1} = -g_{i+1} + \beta_i s_i.$	$s_{i+1} = -g_{i+1} + \beta_i s_i$

($g(x)$ is the gradient of function F at the point x). Nonlinear CG method serves for seeking minima of a general nonlinear function $F(x)$. Instead of the perfect line search with $s_i^T g(x_i + \alpha_i s_i) = 0$, the (generalized) Wolfe conditions

$$F(x_i + \alpha_i s_i) - F_i \le \varepsilon_1 \alpha_i s_i^T g_i, \qquad \varepsilon_2 s_i^T g_i \le s_i^T g(x_i + \alpha_i s_i) \le \varepsilon_3 |s_i^T g_i| \tag{1}$$

are used, where $0 < \varepsilon_1 < 1/2$, $\varepsilon_1 < \varepsilon_2 < 1$ and $\varepsilon_3 \ge 0$. Basic versions of the nonlinear conjugate gradient method use direction vectors $s_1 = -g_1$, $s_{i+1} = -g_{i+1} + \beta_i s_i$, $i \in \mathbb{N}$, with the coefficients

$$\beta_i^{HS} = \frac{y_i^T g_{i+1}}{y_i^T s_i}, \qquad \beta_i^{PR} = \frac{y_i^T g_{i+1}}{g_i^T g_i}, \qquad \beta_i^{LS} = \frac{y_i^T g_{i+1}}{|g_i^T s_i|} \tag{2}$$

(HS–Hestenes and Stiefel, PR–Polak and Ribire, LS–Liu and Storey),

$$\beta_i^{DY} = \frac{g_{i+1}^T g_{i+1}}{y_i^T s_i}, \qquad \beta_i^{FR} = \frac{g_{i+1}^T g_{i+1}}{g_i^T g_i}, \qquad \beta_i^{CD} = \frac{g_{i+1}^T g_{i+1}}{|g_i^T s_i|} \tag{3}$$

(DY–Dai and Yuan, FR–Fletcher and Reeves, CD–conjugate descent), and

$$\beta_i^{HSP} = \frac{p_i^T g_{i+1}}{y_i^T s_i}, \qquad \beta_i^{PRP} = \frac{p_i^T g_{i+1}}{g_i^T g_i}, \qquad \beta_i^{LSP} = \frac{p_i^T g_{i+1}}{|g_i^T s_i|} \tag{4}$$

(Perry's modifications of HS, PR, LS methods). We use the notation $d_i = x_{i+1} - x_i$, $y_i = g_{i+1} - g_i$, $p_i = y_i - d_i$. If $g_{i+1}^T s_i = 0$ (perfect line search) and function F is quadratic, all these methods are identical.

Methods HS, PR, LS are more efficient than DY, FR, CD (since they keep the conjugacy of direction vectors more successfully), but their global convergence cannot be proved without additional modifications. Methods DY, FR, CD are globally convergent (with some limitations concerning the stepsize selection), but they are less efficient than HS, PR, LS methods. The following simple modifications can be used for improving global convergence of HS, PR, LS methods.

$$\begin{aligned} \beta_i^{HS+} &= \max(0, \beta_i^{HS}), & \beta_i^{HSC} &= \max(0, \min(\beta_i^{HS}, \beta_i^{DY})), \\ \beta_i^{PR+} &= \max(0, \beta_i^{PR}), & \beta_i^{PRC} &= \max(0, \min(\beta_i^{PR}, \beta_i^{FR})), \\ \beta_i^{LS+} &= \max(0, \beta_i^{LS}), & \beta_i^{LSC} &= \max(0, \min(\beta_i^{LS}, \beta_i^{CD})). \end{aligned}$$

In this contribution, we will study further modifications of nonlinear CG methods that improve the efficiency of the basic ones. The following modifications, which use direction vectors

$$s_{i+1} = -\left(1 + \beta_i \frac{g_{i+1}^T s_i}{g_{i+1}^T g_{i+1}}\right) g_{i+1} + \beta_i s_i, \tag{5}$$

$$s_{i+1} = -g_{i+1} + \beta_i \left(s_i - \frac{g_{i+1}^T s_i}{g_{i+1}^T y_i}\, y_i\right) \tag{6}$$

guarantee that the direction vectors are descent.

Theorem 1. *Consider the nonlinear CG method with direction vectors $s_1 = -g_1$ and (6) or (6), $i \in \mathbb{N}$. Then $g_i^T s_i = -g_i^T g_i$, $i \in \mathbb{N}$. Let the parameter β_i be given by (2) and the generalized Wolfe conditions (1) with $\varepsilon_2 \le \varepsilon_3 < \infty$ be used. If the objective function F is strongly convex, Lipschitz continuous and bounded from below on $\mathbb{R}^n$, then the nonlinear CG method is uniformly descent and, therefore, globally convergent.*

The following modifications, which use direction vectors

$$s_{i+1} = -\vartheta_i g_{i+1} + \beta_i s_i, \tag{7}$$

where

$$\vartheta_i^{HS} = \vartheta_i^{DY} = \frac{y_i^T s_i}{y_i^T s_i} = 1, \quad \vartheta_i^{PR} = \vartheta_i^{FR} = \frac{y_i^T s_i}{g_i^T g_i}, \quad \vartheta_i^{LS} = \vartheta_i^{CD} = \frac{y_i^T s_i}{|g_i^T s_i|}, \qquad (8)$$

improve the conjugacy of HS, PR, LS methods and guarantee the global convergence of DY, FR, CD methods.

Theorem 2. *Consider the nonlinear CG method with direction vectors $s_1 = -g_1$ and (7), (8), $i \in \mathbb{N}$. Let the parameter β_i be given by (3) and the generalized Wolfe conditions (1) with $0 < \varepsilon_3 < \infty$ be used. If the objective function F is Lipschitz continuous and bounded from below on $\mathbb{R}^n$, then the nonlinear CG method is globally convergent.*

Further nonlinear CG methods with descent property can be obtained using the following lemma, where symbols s_+, g_+ denote s_{i+1}, g_{i+1} and s, z denote s_i, z_i.

Lemma 1. *Let $s_+ = -\vartheta g_+ + \beta s$, where $0 < \underline{\vartheta} \le \vartheta \le \overline{\vartheta}$ and*

$$\beta = g_+^T z - \frac{\lambda}{\vartheta} z^T z g_+^T s. \qquad (9)$$

If $z \in \mathbb{R}^n$ is an arbitrary nonzero vector and $1/4 < \underline{\lambda} \le \lambda \le \overline{\lambda}$, then

$$-\|g_+\|\|s_+\| \le g_+^T s_+ \le -\underline{s}\,\|g_+\|^2, \qquad \underline{s} = \underline{\vartheta}\left(1 - \frac{1}{4\underline{\lambda}}\right) > 0,$$

so that $\|s_+\| \ge \underline{s}\|g_+\|$.

Vector z is chosen in such a way that the first member in (9) corresponds to some basic nonlinear CG method. Let $\vartheta = 1$. Using vectors $z = y/y^T s$, $z = y/g^T g$, $z = y/|g^T s|$ in Lemma 1, we obtain the descent modification of HS, PR, LS methods with

$$\beta^{HSD} = \beta^{HS} - \lambda\frac{y^T y g_+^T s}{(y^T s)^2}, \quad \beta^{PRD} = \beta^{PR} - \lambda\frac{y^T y g_+^T s}{(g^T g)^2}, \quad \beta^{LSD} = \beta^{LS} - \lambda\frac{y^T y g_+^T s}{(g^T s)^2}.$$

Using vectors $z = g_+/y^T s$, $z = g_+/g^T g$, $z = g_+/|g^T s|$ in Lemma 1, we obtain the descent modification of DY, FR, CD methods with

$$\beta^{DYD} = \beta^{DY} - \lambda\frac{g_+^T g_+ g_+^T s}{(y^T s)^2}, \quad \beta^{FRD} = \beta^{FR} - \lambda\frac{g_+^T g_+ g_+^T s}{(g^T g)^2}, \quad \beta^{CDD} = \beta^{CD} - \lambda\frac{g_+^T g_+ g_+^T s}{(g^T s)^2}.$$

Theorem 3. *Consider the nonlinear CG method with direction vectors $s_1 = -g_1$ and $s_{i+1} = -g_{i+1} + \beta_i s_i$, $i \in \mathbb{N}$. Let $\beta_i = \beta_i^{HSD}$ or $\beta_i = \beta_i^{LSD}$ with $1/4 < \underline{\lambda} \le \lambda_i \le \overline{\lambda}$ and the generalized Wolfe conditions (1) with $0 < \varepsilon_3 < \infty$ be used. If the objective function F is uniformly convex, Lipschitz continuous and bounded from below on $\mathbb{R}^n$, then the nonlinear CG method is uniformly descent and, therefore,globally convergent.*

The idea of the proof of the above theorem cannot be used for the PRD method, since the upper bound for $|g_{i+1}^T s_{i+1}|/\|g_{i+1}\|^2$ is unavailable. Setting $\lambda_i = 2$ in the HSD method, we obtain the Hager–Zhang method with

$$\beta_i^{HZ} = \beta_i^{HS} - 2\frac{y_i^T y_i g_{i+1}^T s_i}{(y_i^T s_i)^2}.$$

Setting $\lambda_i = \rho_i y_i^T d_i / y_i^T y_i$ in the HSD method, we obtain the Dai–Liao method with

$$\beta_i^{DL} = \beta_i^{HS} - \rho_i \frac{g_{i+1}^T d_i}{y_i^T s_i}.$$

Further useful nonlinear CG methods can be obtained by the modification of the numerator in (2). In such a way we obtain coefficients

$$\beta_i^{HSM} = \frac{\hat{y}_i^T g_{i+1}}{y_i^T s_i}, \quad \beta_i^{PRM} = \frac{\hat{y}_i^T g_{i+1}}{g_i^T g_i}, \quad \beta_i^{LSM} = \frac{\hat{y}_i^T g_{i+1}}{|g_i^T s_i|}, \tag{10}$$

where

$$\hat{y}_i = g_{i+1} - \frac{\|g_{i+1}\|}{\|g_i\|} g_i \quad \Rightarrow \quad \hat{y}_i^T g_{i+1} = \|g_{i+1}\|^2 - \frac{\|g_{i+1}\|}{\|g_i\|} g_{i+1}^T g_i.$$

Since $|g_{i+1}^T g_i| \le \|g_{i+1}\| \, \|g_i\|$, one has

$$0 \le \|g_{i+1}\|^2 - \frac{\|g_{i+1}\|}{\|g_i\|} g_{i+1}^T g_i \le 2\|g_{i+1}\|^2,$$

or $0 \le \hat{y}_i^T g_{i+1} \le 2\|g_{i+1}\|^2$.

Theorem 4. *Values β_i^{HSM}, β_i^{PRM}, β_i^{LSM} satisfy the inequalities $0 \le \beta_i^{HSM} \le 2\beta_i^{DY}$, $0 \le \beta_i^{PRM} \le 2\beta_i^{FR}$, $0 \le \beta_i^{LSM} \le 2\beta_i^{CD}$. Assume that the strong Wolfe conditions (1) with $0 < \varepsilon_3 = \varepsilon_2$ hold. If $\varepsilon_2 < 1/2$, the LSM method is descent. If $\varepsilon_2 < 1/3$, the HSM method is descent. If $\varepsilon_2 < 1/4$, the PRM method is descent.*

Nonlinear CG methods can be also derived by using variable metric updates. The one step limited memory BFGS method has the form

$$\begin{aligned} s_{i+1} &= -H_{i+1} g_{i+1}, \qquad H_{i+1} y_i = \rho_i d_i, \\ H_{i+1} &= I + \left(\frac{y_i^T y_i}{y_i^T d_i} + \rho_i \right) \frac{d_i d_i^T}{y_i^T d_i} - \frac{y_i d_i^T + d_i y_i^T}{y_i^T d_i}, \end{aligned}$$

where $d_i = x_{i+1} - x_i = \alpha_i s_i$, $y_i = g_{i+1} - g_i$ a $\rho_i > 0$. If $d_i^T g_{i+1} = 0$, we can write

$$\begin{aligned} s_{i+1} &= -g_{i+1} - \left(\frac{y_i^T y_i}{y_i^T d_i} + \rho_i \right) \frac{d_i^T g_{i+1}}{y_i^T d_i} d_i + \frac{d_i^T g_{i+1}}{y_i^T d_i} y_i + \frac{y_i^T g_{i+1}}{y_i^T d_i} d_i \\ &= -g_{i+1} + \frac{y_i^T g_{i+1}}{y_i^T s_i} s_i = -g_{i+1} + \beta_i^{HS} s_i. \end{aligned}$$

Of course, we can omit only selected members containing $d_i^T g_{i+1}$. Setting

$$s_{i+1} = -g_{i+1} + \rho_i \frac{d_i^T g_{i+1}}{y_i^T d_i} d_i - \frac{y_i^T g_{i+1}}{y_i^T d_i} d_i$$

and $\rho_i = 1$, we obtain the HSP method (Perry's modification of the HS method) introduced in (4).

Further interesting nonlinear CG methods can be derived by using the generalized quasi-Newton condition. The generalized QN condition can be written in the form

$$s_{i+1} = -H_{i+1} g_{i+1}, \qquad H_{i+1} y_i = \rho_i d_i \quad \Rightarrow \quad y_i^T s_{i+1} = -\rho_i d_i^T g_{i+1}.$$

Since $s_{i+1} = -g_{i+1} + \beta_i s_i$, this condition is satisfied for $\beta_i = \beta_i^{DL}$, where

$$\beta_i^{DL} = \frac{y_i^T g_{i+1} - \rho_i d_i^T g_{i+1}}{y_i^T s_i} = \beta_i^{HS} - \rho_i \frac{d_i^T g_{i+1}}{y_i^T s_i}.$$

This way leads to the Dai–Liao modifications

$$\beta^{HSDL} = \beta^{HS} - \rho \frac{g_+^T d}{y^T s}, \quad \beta^{PRDL} = \beta^{PR} - \rho \frac{g_+^T d}{g^T g}, \quad \beta^{LSDL} = \beta^{LS} - \rho \frac{g_+^T d}{|g^T s|} \qquad (11)$$

(here $\beta^{HSDL} = \beta^{DL}$).

Nonlinear CG methods can be improved by using restarts. In this case, we set $\beta_i = 0$ if the prescribed condition is not satisfied. The uniform descent condition

$$-g_{i+1}^T s_{i+1} \geq \underline{\eta} \|g_{i+1}\| \|s_{i+1}\|$$

(where, e.g., $\underline{\eta} = 10^{-8}$) guarantees the global convergence of the CG method. The uniform conjugacy condition

$$|y_i^T s_{i+1}| \leq \overline{\eta} \|s_{i+1}\| \|y_i\|$$

(where, e.g., $\overline{\eta} = 5.10^{-2}$) improves efficiency of methods DY, FR, CD and their modifications.

In the tables introduced below, we demonstrate efficiency of selected nonlinear CG methods. The first table contains results obtained using the collection of 73 problems with 10000 variables (TEST 12, `http://camo.ici.ro/neculai/ansoft.htm`). The second table contains results obtained using the collection of 73 problems with 1000 variables (TEST 25, `http://www.cs.cas.cz/luksan/test.html`). In these tables, NIT is the total number of iterations, NFV is the total number of function evaluations and TIME is the total CPU time. At the same time, M denotes basic methods (2), (3), (4), MS denotes modifications (6), MT denotes modifications (6), MI denotes modifications (7), MD denotes modifications based on Lemma 1, MM denotes modifications (10) a MDL denotes modifications (11). Moreover, character + means that value β_i is replaced by $\max(0, \beta_i)$ and character $*$ means that values β^{HS+}, β^{PR+}, β^{LS+} are used in the formulas for β^{HSDL}, β^{PRDL}, β^{LSDL}.

Further details and references can be found in Chapter 3 of the report V1152-14, which is available at `http://www.cs.cas.cz/luksan/lekce4.pdf`.

	Methods of HS type		Methods of PR type		Methods of LS type	
Method	NIT - NFV	TIME	NIT - NFV	TIME	NIT - NFV	TIME
M	73500 - 146562	45.5	97522 - 153458	52.5	90844 - 182707	59.2
M+	64776 - 130153	42.2	99012 - 199048	52.2	109072 - 217871	59.1
MS+	64267 - 127877	39.4	81135 - 162484	46.9	98472 - 197386	54.6
MI+	64776 - 130153	42.2	59242 - 118194	37.5	92908 - 185231	49.8
MT+	56465 - 113023	37.8	66533 - 132821	41.1	69851 - 139348	41.5
MD+	63923 - 128143	42.0	93105 - 187343	51.3	70265 - 140260	41.7
MDL∗	70630 - 138223	46.7	89794 - 180989	50.8	106829 - 214506	65.0
MM	63761 - 127077	38.2	69206 - 139422	40.1	98169 - 196718	48.2
	Methods of DY type		Methods of FR type		Methods of CD type	
Method	NIT - NFV	TIME	NIT - NFV	TIME	NIT - NFV	TIME
M	72624 - 145100	47.1	81152 - 162513	48.0	87805 - 176088	63.7
MS	85372 - 161985	57.9	84886 - 170639	68.1	69839 - 140434	42.2
MI	72624 - 145100	47.1	70155 - 141153	49.1	83105 - 166870	49.6
MT	85249 - 169741	51.1	84001 - 175873	63.8	88634 - 184105	76.1
MD+	84267 - 170722	52.5	82341 - 164020	61.3	75449 - 151144	46.6
	Methods of HSP type		Methods of PRP type		Methods of LSP type	
Method	NIT - NFV	TIME	NIT - NFV	TIME	NIT - NFV	TIME
M	94217 - 189553	99.9	98579 - 195634	52.3	89764 - 168900	55.3
M+	75175 - 150631	46.9	65729 - 132372	40.6	85626 - 164338	48.9
MS+	63356 - 126299	39.6	65561 - 131168	41.6	84874 - 170016	50.0
MI+	75175 - 150631	47.0	66181 - 133055	43.6	68377 - 136899	43.8
MT+	67290 - 136028	48.5	69115 - 138680	44.0	66094 - 132739	44.0
MD+	80467 - 154308	51.4	71019 - 143753	47.6	87721 - 165860	53.5

	Methods of HS type		Methods of PR type		Methods of LS type	
Method	NIT - NFV	TIME	NIT - NFV	TIME	NIT - NFV	TIME
M	182719 - 362799	44.5	193715 - 382239	48.7	186195 - 365074	47.9
M+	181090 - 357804	45.0	194625 - 385349	47.3	171949 - 339448	38.5
MS+	176027 - 348089	44.7	180893 - 356713	45.4	181363 - 357095	46.6
MI+	181090 - 357804	45.0	192212 - 377671	49.0	182165 - 358848	46.9
MT+	179137 - 354722	39.9	166227 - 327249	36.1	172590 - 339757	37.4
MD+	189405 - 372372	48.8	200779 - 394240	49.9	182981 - 361565	45.5
MDL∗	185031 - 366172	45.1	196460 - 583719	51.0	188953 - 373247	45.5
MM	175646 - 346092	45.7	188722 - 373911	47.4	190902 - 376303	46.4
	Methods of HSP type		Methods of PRP type		Methods of LSP type	
Method	NIT - NFV	TIME	NIT - NFV	TIME	NIT - NFV	TIME
M	180076 - 356292	45.4	183742 - 362432	47.0	194143 - 381417	48.1
M+	174388 - 344643	44.3	173541 - 342704	38.3	204033 - 397429	51.3
MS+	185629 - 366182	46.0	185214 - 365439	47.7	181322 - 358282	45.8
MI+	174388 - 344643	44.3	175264 - 346739	44.9	183064 - 361115	45.4
MT+	174902 - 345601	38.8	163751 - 322111	35.8	178082 - 349536	38.7
MD+	190318 - 374073	42.5	191386 - 377971	48.0	183564 - 361783	45.7
MDI+	190318 - 374073	42.5	185624 - 367332	46.1	189522 - 373814	47.1