Jan Vlček; Ladislav Lukšan

Modifications of the limited-memory BFGS method based on the idea of conjugate directions

In: Jan Chleboun and Karel Segeth and Jakub Šístek and Tomáš Vejchodský (eds.): Programs and Algorithms of Numerical Mathematics, Proceedings of Seminar. Dolní Maxov, June 3-8, 2012. Institute of Mathematics AS CR, Prague, 2013. pp. 209–214.

Persistent URL: http://dml.cz/dmlcz/702729

Programs and Algorithms of Numerical Matematics 16
J. Chleboun, K. Segeth, J. Šístek, T. Vejchodský (Eds.)
Institute of Mathematics AS CR, Prague 2013

MODIFICATIONS OF THE LIMITED-MEMORY BFGS METHOD BASED ON THE IDEA OF CONJUGATE DIRECTIONS

Jan Vlček[1], Ladislav Lukšan[1,2]

[1] Institute of Computer Science, Academy of Sciences of the Czech Republic
Pod vodárenskou věží 2, 182 07 Prague 8, Czech Republic
vlcek@cs.cas.cz, luksan@cs.cas.cz
[2] Technical University of Liberec, Hálkova 6, 461 17 Liberec, Czech Republic

Abstract

Simple modifications of the limited-memory BFGS method (L-BFGS) for large scale unconstrained optimization are considered, which consist in corrections of the used difference vectors (derived from the idea of conjugate directions), utilizing information from the preceding iteration. For quadratic objective functions, the improvement of convergence is the best one in some sense and all stored difference vectors are conjugate for unit stepsizes. The algorithm is globally convergent for convex sufficiently smooth functions. Numerical experiments indicate that the new method often improves the L-BFGS method significantly.

1. Introduction

We propose some modifications of the L-BFGS method (see [5], [10]) for large scale unconstrained minimization of the differentiable function $f : \mathcal{R}^N \to \mathcal{R}$. Similarly as in the multi-step quasi-Newton methods (see e.g. [9]), we utilize information from the preceding iteration. However, while the multi-step methods derive the corrections of the difference vectors from various interpolation methods, our approach is based on the idea of conjugate directions (see e.g. [4, 11]).

The L-BFGS method belongs to the variable metric (VM) or quasi-Newton line search methods, see [4], [8]. They start with an initial point $x_0 \in \mathcal{R}^N$ and generate iterations $x_{k+1} \in \mathcal{R}^N$ by the process $x_{k+1} = x_k + t_k d_k$, $k \ge 0$, where d_k is the direction vector and $t_k > 0$ is a stepsize, usually chosen in such a way that

$$f_{k+1} - f_k \le \varepsilon_1 t_k g_k^T d_k, \qquad g_{k+1}^T d_k \ge \varepsilon_2 g_k^T d_k, \tag{1}$$

$k \ge 0$, where $0 < \varepsilon_1 < 1/2$, $\varepsilon_1 < \varepsilon_2 < 1$, $f_k = f(x_k)$, $g_k = \nabla f(x_k)$ and $d_k = -H_k g_k$ with a symmetric positive definite matrix H_k; usually $H_0 = I$ and H_{k+1} is obtained from H_k by a VM update to satisfy the quasi-Newton condition $H_{k+1} y_k = s_k$ (see [4, 8]), where $s_k = x_{k+1} - x_k = t_k d_k$ and $y_k = g_{k+1} - g_k$, $k \ge 0$.

Among VM methods, the BFGS method belongs to the most efficient; the update formula can be written in the form (note that $b_k > 0$ for $g_k \ne 0$ by (1))

$$H_{k+1} = (1/b_k) s_k s_k^T + V_k H_k V_k^T, \quad b_k = s_k^T y_k, \quad V_k = I - (1/b_k) s_k y_k^T,$$

$k \geq 0$, see [4, 8, 11], on which the L-BFGS method – a limited-memory adaptation of the BFGS method – is based. Instead of an $N \times N$ matrix H_k, only the last $\tilde{m}+1$ couples $\{s_j, y_j\}_{j=k-\tilde{m}}^{k}$ are stored, where $\tilde{m} = \min(k, m-1)$ and $m \geq 1$ is a given parameter. The direction vector is computed by the Strang recurrences, see [10], and still satisfies $d_{k+1} = -H_{k+1} g_{k+1}$, $k \geq 0$, but matrix H_{k+1} is not formed explicitly.

Here we will investigate such corrections of vectors s_k, y_k which provide conjugacy of consecutive corrected vectors. Thus we will define corrected quantities $\bar{s}_k$, $\bar{y}_k$, $\bar{b}_k$ and $\bar{V}_k$, $k \geq 0$, by $\bar{s}_0 = s_0$, $\bar{y}_0 = y_0$, $\bar{b}_0 = b_0$, $\bar{V}_0 = V_0$ and

$$\bar{s}_k = s_k - \alpha_k \bar{s}_{k-1}, \quad \bar{y}_k = y_k - \beta_k \bar{y}_{k-1}, \quad \bar{b}_k = \bar{s}_k^T \bar{y}_k, \quad \bar{V}_k = I - (1/\bar{b}_k) \bar{s}_k \bar{y}_k^T, \tag{2}$$

$k > 0$, with such α_k, $\beta_k \in \mathcal{R}$ that $\bar{b}_k > 0$. Correspondingly, we will use a direction vector $d_k = -\bar{H}_k g_k$, $k \geq 0$, where $\bar{H}_0 = I$ and symmetric positive definite matrix

$$\begin{aligned} \bar{H}_{k+1} \;=\;& (s_k^T y_k / |y_k|^2)\, \bar{V}_k \cdots \bar{V}_{k-\tilde{m}}\, \bar{V}_{k-\tilde{m}}^T \cdots \bar{V}_k^T \\ &+\; (1/\bar{b}_{k-\tilde{m}})\, \bar{V}_k \cdots \bar{V}_{k-\tilde{m}+1}\, \bar{s}_{k-\tilde{m}} \bar{s}_{k-\tilde{m}}^T\, \bar{V}_{k-\tilde{m}+1}^T \cdots \bar{V}_k^T \\ &+\; \cdots \;+\; (1/\bar{b}_{k-1})\, \bar{V}_k\, \bar{s}_{k-1} \bar{s}_{k-1}^T\, \bar{V}_k^T \;+\; (1/\bar{b}_k)\, \bar{s}_k \bar{s}_k^T, \; k \geq 0, \end{aligned} \tag{3}$$

satisfies the quasi-Newton condition $\bar{H}_{k+1} \bar{y}_k = \bar{s}_k$ and is obtained by the repeated BFGS update of $(s_k^T y_k / |y_k|^2) I$ with corrected vectors. We denote $\bar{B}_k = \bar{H}_k^{-1}$, $k \geq 0$.

In Section 2 we investigate the standard BFGS update with corrected vectors

$$\bar{H}_+ = (1/\bar{b}) \bar{s} \bar{s}^T + \bar{V} \bar{H} \bar{V}^T, \quad \bar{b} = \bar{s}^T \bar{y}, \quad \bar{V} = I - (1/\bar{b}) \bar{s} \bar{y}^T, \tag{4}$$

(in the simplified form) of any symmetric positive definite matrix $\bar{H}$ with corrected difference vectors $\bar{s} = s - \alpha \bar{s}_-$, $\bar{y} = y - \beta \bar{y}_-$ and discuss the choice of parameters α, β. In Section 3 we focus on quadratic functions and show optimality of our choice of parameters and conjugacy and other properties for unit stepsizes. Application to limited-memory methods and the corresponding algorithm are described in Section 4, global convergence of the algorithm is established in Section 5 and numerical results are reported in Section 6. Details and proofs of assertions can be found in [13].

2. The BFGS update with corrected vectors

The following lemma enables us to distinguish roles of products $\bar{s}^T \bar{y}_-$, $\bar{s}_-^T \bar{y}$ and shows that, under some assumptions, the conjugacy of difference vectors $\bar{s}$, $\bar{s}_-$ with respect to matrices $\bar{B} = \bar{H}^{-1}$, $\bar{B}_+ = \bar{H}_+^{-1}$ is equivalent to the satisfaction of condition $\bar{H}_+ \bar{y}_- = \bar{s}_-$. Note that condition $\bar{H} \bar{y}_- = \bar{s}_-$ represents the quasi-Newton condition from the preceding update, which is satisfied for $m > 1$, see [13].

Lemma 1. *Let $\bar{H}$ be any symmetric positive definite matrix with $\bar{H} \bar{y}_- = \bar{s}_-$, matrix $\bar{H}_+$ be given by (4) with $\bar{b} > 0$ and $\Delta_1 = (\bar{H}_+ \bar{y}_- - \bar{s}_-)^T \bar{B}_+ (\bar{H}_+ \bar{y}_- - \bar{s}_-)$. Then*

$$\Delta_1 = \left[(\bar{s}_-^T \bar{y} - \bar{s}^T \bar{y}_-)^2 + \omega (\bar{s}^T \bar{y}_-)^2 \right] / \bar{b}, \tag{5}$$

where $\omega \geq 0$, *with* $\omega = 0$ *only in case of dependency of vectors* $\bar{s}$, $\bar{H}\bar{y}$. *If vectors* $\bar{s}$, $\bar{H}\bar{y}$ *are linearly independent then* $\bar{H}_+$ *satisfies* $\bar{H}_+\bar{y}_- = \bar{s}_-$ *if and only if vectors* $\bar{s}$, $\bar{s}_-$ *are conjugate with respect to matrices* $\bar{B}$, $\bar{B}_+$.

Since value ω could be large, we can see from relation (5) that mainly value $\bar{s}^T\bar{y}_-$ should be close to zero, to have Δ_1 small. Therefore we prefer the choice $\alpha = s^T\bar{y}_-/\bar{b}_-$, for which $\bar{s}^T\bar{y}_- = 0$. Similarly, the basic choice of β is $\beta_Z = \bar{s}_-^T y/\bar{b}_-$, which yields $\bar{s}_-^T\bar{y} = 0$ (thus $\bar{H}_+\bar{y}_- = \bar{s}_-$ by $\Delta_1 = 0$) and has some interesting properties.

Theorem 2. *Let* $\bar{H}$ *be any symmetric positive definite matrix with* $\bar{H}\bar{y}_- = \bar{s}_-$ *and matrix* $\bar{H}_+$ *be given by (4) with* $\bar{b} > 0$. *If* $\alpha = s^T\bar{y}_-/\bar{b}_-$ *then* $\bar{s}^T\bar{y}_- = 0$, $\bar{b} = b - \alpha\,\bar{s}_-^T y$ *and both value* $\bar{a}$ *and the condition number of matrix* $\bar{H}^{1/2}\bar{B}_+\bar{H}^{1/2}$ *as functions of* β *are minimized by the choice* $\beta = \bar{s}_-^T y/\bar{b}_-$.

Satisfaction of condition $\bar{H}_+\bar{y}_- = \bar{s}_-$ also guarantees that matrix $\bar{H}_+$ is closer to $\bar{H}$ than to $\bar{H}_-$ in some sense, as we can see from Theorem 3 with $\bar{H}_-$, $\bar{H}$, $\bar{s}_-$, $\bar{y}_-$ instead of $\bar{H}$, $\bar{H}_+$, $\bar{s}$, $\bar{y}$ and $\tilde{G} = \bar{H}_+^{-1}$ ($\|.\|_F$ denotes the Frobenius matrix norm).

Theorem 3. *Let* $\bar{H}$ *be any symmetric positive definite matrix, matrix* $\bar{H}_+$ *be given by (4) with* $\bar{b} > 0$, $\tilde{G}$ *be any symmetric positive definite matrix satisfying* $\tilde{G}\bar{s} = \bar{y}$, $W_+ = \tilde{G}^{1/2}\bar{H}_+\tilde{G}^{1/2}$ *and* $W = \tilde{G}^{1/2}\bar{H}\tilde{G}^{1/2}$. *Then*

$$\|I - W_+\|_F^2 - \|I - W\|_F^2 = -\|W_+ - W\|_F^2 \leq -\left(\bar{a}/\bar{b} - 1\right)^2. \tag{6}$$

The following lemma indicates that β should also be near to α, to have $|\bar{H}_+y - s|$ small. E.g. the choice $\beta = \pm\sqrt{\beta_Z\alpha}$ has interesting properties.

Lemma 4. *Let* $\bar{H}$ *be any symmetric positive definite matrix with* $\bar{H}\bar{y}_- = \bar{s}_-$ *and matrix* $\bar{H}_+$ *be given by (4) with* $\bar{b} > 0$. *If* $\alpha = s^T\bar{y}_-/\bar{b}_-$ *then* $\Delta_1 = (\bar{s}_-^T\bar{y})^2/\bar{b}$ *and*

$$(\bar{H}_+y - s)^T\bar{B}_+(\bar{H}_+y - s) = \bar{b}_-[(\beta - \alpha)^2 + (\beta - \beta_Z)^2(s^T\bar{y}_-)^2/(\bar{b}\,\bar{b}_-)]. \tag{7}$$

Moreover, if $\beta^2 = s^T\bar{y}_-\,\bar{s}_-^T y/\bar{b}_-^2$, *then* $y^T(\bar{H}_+y - s) = 0$.

3. Results for quadratic functions

In this section we suppose that f is a quadratic function with a symmetric positive definite matrix G and that $\beta = \alpha$, which is a natural choice, if we want to have $\bar{y} = G\bar{s}$, similarly as for non-corrected vectors. Here we consider only the G-conjugacy of vectors.

The conjugacy of $\bar{s}$, $\bar{s}_-$ can be achieved by the choice $\alpha = s^T\bar{y}_-/\bar{b}_- = \bar{s}_-^T y/\bar{b}_-$ by (2). The following theorem shows that this choice is the best in some sense.

Theorem 5. *Let* $\hat{\alpha} = s^T\bar{y}_-/\bar{b}_- = \bar{s}_-^T y/\bar{b}_-$, $\bar{H}$ *be any symmetric positive definite matrix with* $\bar{H}\bar{y}_- = \bar{s}_-$, $\bar{H}_+$ *be given by (4) with* $\beta = \alpha$ *and let* f *be a quadratic function* $f(x) = \frac{1}{2}(x - x^*)^T G(x - x^*)$, $x^* \in \mathcal{R}^N$, *with a symmetric positive definite matrix* G. *If vectors* s, $\bar{s}_-$ *are linearly independent, then* $\bar{b} > 0$ *and the choice* $\alpha = \hat{\alpha}$ *implies* $\bar{H}_+y = s$ *and minimizes the values* $\bar{b}$, $\|G^{1/2}\bar{H}_+G^{1/2} - I\|_F$ *as functions of* α.

The L-BFGS method with exact line searches generates conjugate directions vectors and preserves $\tilde{m}$ previous quasi-Newton conditions, see e.g. [10]. Similarly for update (4) with unit stepsizes we get that all stored vectors $\bar{s}_k$ are conjugate and $\tilde{m}$ previous quasi-Newton conditions are preserved, if every stepsize is unit.

Theorem 6. *Let $x_0 \in \mathcal{R}^N$, $x^* \in \mathcal{R}^N$, $\bar{k} > 0$, $m \geq 1$, f be the quadratic function $f(x) = \frac{1}{2}(x-x^*)^T G(x-x^*)$ with a symmetric positive definite matrix G, and let for $0 \leq k \leq \bar{k}$ iterations $x_{k+1} = x_k + s_k$ be generated by $s_k = -t_k \bar{H}_k g_k$, $g_k = \nabla f(x_k)$, $t_k > 0$, with matrices $\bar{H}_k$ defined in the following way: $\bar{H}_0 = I$ and matrices $\bar{H}_{k+1}$, $0 \leq k < \bar{k}$, are given by (3), where $\tilde{m} = \min(k, m-1)$, $y_k = g_{k+1} - g_k$, and quantities $\bar{s}_j$, $\bar{y}_j$, $\bar{V}_j$ and $\bar{b}_j$, $j \geq 0$, are formally defined by $\bar{s}_0 = s_0$, $\bar{y}_0 = y_0$, $\bar{s}_{j+1} = s_{j+1} - \alpha_{j+1}\bar{s}_j$, $\bar{y}_{j+1} = y_{j+1} - \alpha_{j+1}\bar{y}_j$, $\alpha_{j+1} = s_{j+1}^T \bar{y}_j / \bar{b}_j$, $\bar{V}_j = I - (1/\bar{b}_j)\bar{s}_j \bar{y}_j^T$, $\bar{b}_j = \bar{s}_j^T \bar{y}_j$.*

If every generated vector s_k is linearly independent of $\bar{s}_{k-1}$, $0 < k \leq \bar{k}$, then the method is well defined. Moreover, if $t_{k+1} = 1$ for some k, $0 \leq k < \bar{k}$, it holds

$$\bar{H}_{k+i}\bar{y}_k = \bar{s}_k, \quad \bar{s}_k^T G \bar{s}_{k+i} = 0, \quad \bar{s}_k^T g_{k+i+1} = 0, \quad 1 \leq i \leq \min(\tilde{m}+1, \bar{k}-k). \tag{8}$$

4. Application to limited-memory methods

From the theory in Section 3 we can deduce that we should use the corrected difference vectors whenever objective function is close to a quadratic function. As measure of deviation from a quadratic function at points x_{k-1}, x_k, x_{k+1}, e.g. value $|s_k^T y_{k-1} - s_{k-1}^T y_k|$ could serve (zero for quadratic functions), $k > 0$; we use value $|s_k^T \bar{y}_{k-1} - \bar{s}_{k-1}^T y_k| = \bar{b}_{k-1}|\alpha_k - \beta_k|$, which gives very similar results. We do not correct if it is greater than $\bar{b}_{k-1}^2 / b_k$, if $(s_k^T \bar{y}_{k-1}).(\bar{s}_{k-1}^T y_k) \leq 0$ or if $\bar{b}_k \leq 10^{-6} b_k$.

Value $\beta_k = \mathrm{sgn}(\alpha_k)\sqrt{\theta_k / \bar{b}_{k-1}}$, corresponding to the choice in Lemma 4, appears to be suitable if value $\bar{b}_k$ is sufficiently large with respect to b_k (we use condition $\bar{b}_k > 10^{-2} b_k$). This choice satisfies $|\beta_k| < \sqrt{b_k / \bar{b}_{k-1}}$; it is a reason why we use this value β_k also in case that $|\bar{s}_{k-1}^T y_k / \bar{b}_{k-1}| > 2\sqrt{b_k / \bar{b}_{k-1}}$ to prove global convergence.

Global convergence can be easily established (in a similar way as for the L-BFGS method, see [5]), if $|\bar{s}_k|/|s_k| \leq \Delta$ and $|\bar{y}_k|/|y_k| \leq \Delta$, $k > 0$, where $\Delta > 1$ is a constant. If this condition is not satisfied, it suffices to replace the oldest saved vectors $\bar{s}_{k-\tilde{m}}$, $\bar{y}_{k-\tilde{m}}$ e.g. by s_k, y_k. Note that in our numerical experiments with $N = 1000$, value $|\bar{y}_k|/|y_k|$ was rarely greater than 10 and value $|\bar{s}_k|/|s_k|$ greater than 50.

We now state the method in details. For simplicity, we omit stopping criteria.

Algorithm 4.1

Data: The number $m \geq 1$ of VM updates per iteration, line search parameters ε_1, ε_2, $0 < \varepsilon_1 < 1/2$, $\varepsilon_1 < \varepsilon_2 < 1$, and correction parameter $\Delta > 1$.

Step 0: Initiation. Choose starting point $x_0 \in \mathcal{R}^N$, define starting matrix $\bar{H}_0^0 = I$ and direction vector $d_0 = -\nabla f(x_0)$ and initiate iteration counter k to zero.

Step 1: Line search. Compute $x_{k+1} = x_k + t_k d_k$, where t_k satisfies (1), $s_k = x_{k+1} - x_k$, $g_{k+1} = \nabla f(x_{k+1})$, $y_k = g_{k+1} - g_k$ and $b_k = s_k^T y_k$. If $k = 0$ set $\bar{s}_k = s_k$, $\bar{y}_k = y_k$ and go to Step 4.

Step 2: Correction preparation. Set $\alpha_k = s_k^T \bar{y}_{k-1}/\bar{b}_{k-1}$, $\beta_k = \bar{s}_{k-1}^T y_k/\bar{b}_{k-1}$. If $\alpha_k\beta_k \le 0$ or $\bar{b}_k \le 10^{-6} b_k$ or $|\alpha_k - \beta_k| \ge \bar{b}_{k-1}/b_k$, set $\alpha_k = \beta_k = 0$ and go to Step 3. If $|\beta_k| > 2\sqrt{b_k/\bar{b}_{k-1}}$ or $\bar{b}_k > 10^{-2} b_k$, replace β_k by $\beta_k\sqrt{\alpha_k/\beta_k}$.

Step 3: Correction. Set $\bar{s}_k = s_k - \alpha_k \bar{s}_{k-1}$, $\bar{y}_k = y_k - \beta_k \bar{y}_{k-1}$.

Step 4: Update definition. Set $\tilde{m} = \min(k, m-1)$, $\bar{b}_k = \bar{s}_k^T \bar{y}_k$ and define $\bar{V}_k = I - (1/\bar{b}_k)\bar{s}_k \bar{y}_k^T$. If $|\bar{s}_{k-\tilde{m}}|/|s_{k-\tilde{m}}| > \Delta$ or $|\bar{y}_{k-\tilde{m}}|/|y_{k-\tilde{m}}| > \Delta$, set $\bar{s}_{k-\tilde{m}} = s_k$, $\bar{y}_{k-\tilde{m}} = y_k$ and $\bar{b}_{k-\tilde{m}} = b_k$. Define $\bar{H}_{k+1}$ by (3).

Step 5: Direction vector. Compute $d_{k+1} = -\bar{H}_{k+1} g_{k+1}$ by the Strang recurrences, set $k := k+1$ and go to Step 1.

5. Global convergence

Assumption 7. *The objective function $f : \mathcal{R}^N \to \mathcal{R}$ is bounded from below and uniformly convex with bounded second-order derivatives (i.e. $0 < \underline{G} \le \underline{\lambda}(G(x)) \le \overline{\lambda}(G(x)) \le \overline{G} < \infty$, $x \in \mathcal{R}^N$, where $\underline{\lambda}(G(x))$ and $\overline{\lambda}(G(x))$ are the lowest and the greatest eigenvalues of the Hessian matrix $G(x)$).*

Theorem 8. *If objective function f satisfies Assumption 7, Algorithm 4.1 generates a sequence $\{g_k\}$ that either satisfies $\lim_{k\to\infty}|g_k|=0$ or terminates with $g_k=0$ for some k.*

6. Numerical results

In this section, we demonstrate the influence of vector corrections on the number of evaluations (NFE) and computational time, using the following collections of test problems: Test 11 from [7] (55 chosen problems), which are modified problems from CUTE collection [2] with N ranging from 1000 to 5000, test from [1], termed Test 12 here, 73 problems, $N = 5000$, Test 25 from [6] (68 chosen problems), $N = 10000$.

Table 1 contains results for the following limited-memory methods: L-BFGS, see [10], method from [12] (Algorithm 3.1 with $\sigma = 0.4$) and new Algorithm 4.1. We have used $m = 5$, $\Delta = 100$, the final precision $\|g(x^\star)\|_\infty \le 10^{-6}$, $\varepsilon_1 = 10^{-4}$ and $\varepsilon_2 = 0.8$.

Method	Test 11		Test 12		Test 25	
	NFE	Time	NFE	Time	NFE	Time
L-BFGS	80539	32.50	43648	46.17	462104	519.40
Alg. 3.1 in [12]	80328	34.52	43182	56.67	512880	649.15
Algorithm 4.1	64395	30.20	34472	37.57	296321	381.08

Table 1. Comparison of the selected methods.

For a better demonstration of both the efficiency and the reliability, we compare selected optimization methods for Test 25 by using performance profiles introduced in [3]. The value of $\pi_M(\tau)$ at $\tau = 0$ gives the percentage of test problems for which the method M is the best and the value for τ large enough is the percentage of test problems that method M can solve. The relative efficiency and reliability of each method can be directly seen from the performance profiles: the higher is the particular curve the better is the corresponding method.

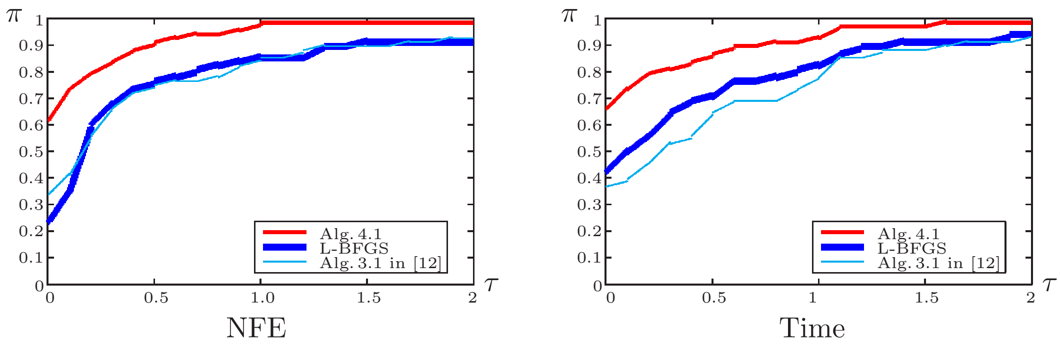

Figure 1: (Test 25, $m = 5$, 68 problems, N=10 000)

Acknowledgements

This work was supported by RVO: 67985807.

References

[1] Andrei, N.: An unconstrained optimization test functions collection. Advanced Modeling and Optimization **10** (2008), 147–161.

[2] Bongartz, I., Conn, A. R., Gould, N., and Toint, P. L.: CUTE: constrained and unconstrained testing environment. ACM Trans. Math. Software **21** (1995), 123–160.

[3] Dolan, E. D. and Moré, J. J.: Benchmarking optimization software with performance profiles. Math. Prog. **91** (2002) 201–213.

[4] Fletcher, R.: *Practical Methods of Optimization*. John Wiley & Sons, Chichester, 1987.

[5] Liu, D. C. and Nocedal, J.: On the limited memory BFGS method for large scale optimization. Math. Prog. **45** (1989) 503–528.

[6] Lukšan, L., Matonoha, C., and Vlček, J.: Sparse test problems for unconstrained optimization. Report V-1064, ICS AS CR, Prague, 2010.

[7] Lukšan L., Matonoha C., Vlček J.: Modified CUTE problems for sparse unconstrained optimization. Report V-1081, ICS AS CR, Prague, 2010.

[8] Lukšan, L. and Spedicato, E.: Variable metric methods for unconstrained optimization and nonlinear least squares. J. Comput. Appl. Math. **124** (2000) 61–95.

[9] Moughrabi, I. A.: New implicit multistep quasi-Newton methods. Numerical Analysis and Applications **2** (2009) 154–164.

[10] Nocedal, J.: Updating quasi-Newton matrices with limited storage. Math. Comp. **35** (1980) 773–782.

[11] Nocedal, J. and Wright, S. J.: *Numerical optimization*. Springer-Verlag, New York, 1999.

[12] Vlček, J., Lukšan, L.: Limited-memory variable metric methods that use quantities from the preceding iteration. Proc. of Semin. PANM 15, D. Maxov, 2010.

[13] Vlček, J. and Lukšan, L.: A conjugate directions approach to improve the limited-memory BFGS method. Appl. Math. Comput. **219** (2012), 800–809.