CompAMa
www.compama.co.usb.ve

Bull. Comput. Appl. Math., Vol.4, No.1, 2016
ISSN 2244-8659

Two extensions of the Dai-Liao method with sufficient descent property based on a penalization scheme

Masoud Fatemi [*], *Saman Babaie-Kafaki* [†]

CompAMa Vol.4, No.1, pp.7-19, 2016 - Accepted April 13, 2016

Abstract

To achieve the good features of the linear conjugate gradient algorithm in a recent extension of the Dai–Liao method, two adaptive choices for parameter of the extended method are proposed based on a penalization approach. It is shown that the suggested parameters guarantee the sufficient descent property independent to the line search and the objective function convexity. Furthermore, they ensure the global convergence of the related algorithm for uniformly convex objective functions. Using a set of unconstrained optimization test problems from the CUTEr library, effectiveness of the suggested choices are numerically compared with two other recently proposed adaptive choices. Results of comparisons show that one of the proposed choices is computationally promising in the sense of the Dolan–Moré performance profile.

Keywords: Unconstrained optimization, conjugate gradient method, sufficient descent property, penalty method, line search.

[*]Department of Mathematics, K. N. Toosi University of Technology, P.O. Box: 16315-1618, Tehran, Iran (smfatemi@kntu.ac.ir).

[†]Department of Mathematics, Faculty of Mathematics, Statistics and Computer Science, Semnan University, P.O. Box: 35195-363, Semnan, Iran (sbk@semnan.ac.ir).

1 Introduction

Unconstrained optimization deals with the problem of minimizing an objective function $f : \mathbb{R}^n \to \mathbb{R}$ with no restriction on its variables, that is

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

Here, we assume that f is continuously differentiable and its gradient is available. The problem (1) not only directly arises in some applications but also indirectly arises in reformulations of constrained optimization problems; often it is practical to replace the constraints of an optimization problem with penalized terms in the objective function and to solve an unconstrained problem.

As practical tools for solving (1), iterative methods generate a series of progressively improved solutions to the problem, starting with an initial estimate and stopping with some predefined criterion. When the dimension n is large, iterative methods which require low memory storage are more encouraging. Among them there are the conjugate gradient (CG) methods with the following iterative formula:

$$x_0 \in \mathbb{R}^n,\ x_{k+1} = x_k + s_k,\ s_k = \alpha_k d_k,\ k = 0, 1, ...,$$

where α_k is a step length to be computed by a line search technique along the search direction d_k defined by

$$d_0 = -g_0,\ d_{k+1} = -g_{k+1} + \beta_k d_k,\ k = 0, 1, ...,$$

in which $g_k = \nabla f(x_k)$ and β_k is a scalar called the CG (update) parameter. The step length α_k is often determined to fulfill the Wolfe conditions, that is

$$f(x_{k+1}) - f(x_k) \leq \delta \alpha_k d_k^T g_k, \tag{2}$$

$$d_k^T g_{k+1} \geq \sigma d_k^T g_k, \tag{3}$$

with the constants δ and σ satisfying $0 < \delta < \sigma < 1$ [1].

Different CG methods mainly correspond to different ways of computing β_k, leading to different numerical performances [2, 3]. In a fundamental attempt to employ the quasi–Newton aspects [1] in the CG methods, Dai and Liao [4] (DL) proposed the following CG parameter:

$$\beta_k^{DL} = \frac{g_{k+1}^T y_k}{d_k^T y_k} - t \frac{g_{k+1}^T s_k}{d_k^T y_k}, \tag{4}$$

where $y_k = g_{k+1} - g_k$ denotes the gradient change, and t is a nonnegative parameter. If $t = 0$ or the exact line search is used, then the method reduces to the Hestenes–Stiefel method [5] (HS).

To the best of our knowledge, computational performance of the DL method is very dependent to the parameter t for which efforts have been made to find appropriate choices. The interested readers can study references [6–11]. Recently, Babaie–Kafaki and Ghanbari [12] employed the three–term extension of the HS method proposed by Zhang et al. [13] and developed a three–term version of the DL method (TTDL) with the search directions $d_0 = -g_0$ and

$$d_{k+1} = -g_{k+1} + \underbrace{\frac{g_{k+1}^T y_k}{d_k^T y_k}}_{\beta_k^{HS}} d_k - t\frac{g_{k+1}^T s_k}{|d_k^T y_k|} d_k - \frac{g_{k+1}^T d_k}{d_k^T y_k} y_k, \ \forall k \geq 0, \tag{5}$$

where t is a real parameter. If $t = 0$, then TTDL reduces to the three–term CG method proposed in [13]. It is worth noting that if $t \geq 0$, then TTDL satisfies the sufficient descent condition [12], that is

$$d_k^T g_k \leq -c||g_k||^2, \ k = 0, 1, ..., \tag{6}$$

where c is a positive constant and $||.||$ stands for the ℓ_2 norm. In [12], based on the standard secant equation [1] an adaptive choice for t has been proposed as follows:

$$t_k = \max\{\xi, 1 - \frac{||y_k||^2}{s_k^T y_k}\}, \tag{7}$$

where ξ is a nonnegative constant. More recently, making search directions of the TTDL method as close as possible to the search directions of the scaled memoryless BFGS method [1] in a least–squares scheme, in [14] a class of one–parameter choices for t has been proposed as follows:

$$t_k = \max\left\{\xi, 1 + \theta_k \frac{||y_k||^2}{s_k^T y_k} - (\theta_k + 1)\frac{s_k^T y_k}{||s_k||^2}\right\}, \tag{8}$$

where ξ is a nonnegative constant and θ_k is the scaling parameter of the memoryless BFGS update. Numerical experiments of [14] showed that memoryless version of the scaling parameter proposed by Oren and Spedicato [15], that is

$$\theta_k = \frac{s_k^T y_k}{||y_k||^2},$$

which considering (8) yields

$$t_k = \max\left\{\xi, 2 - \frac{(s_k^T y_k)^2}{||s_k||^2||y_k||^2} - \frac{s_k^T y_k}{||s_k||^2}\right\}, \tag{9}$$

turns out to be numerically promising.

Here, we employ the penalization strategy of [9,10] and propose two other choices for the parameter t in (5). This work is organised as follows: in Section 2, two adaptive versions of the TTDL method are proposed and then, a global convergence analysis is conducted for uniformly convex objective functions. In Section 3, the methods are numerically compared with the nonlinear CG methods proposed in [12,14], using the Dolan–Moré performance profile [16]. Finally, conclusions are drawn in Section 4.

2 Two adaptive choices for the TTDL parameter

Here, we intend to propose suitable choices for the parameter t in such a way that the TTDL method reflects good features of the linear CG methods as much as possible. Hereafter, we assume that $d_k^T y_k > 0$, as guaranteed by the Wolfe condition (3).

As known, in the linear CG methods a strongly convex quadratic function is successively minimized in at most n steps along some individual directions of a conjugate set. The methods possess the following remarkable properties:

(i) the sufficient descent property (6);

(ii) the conjugacy property, namely, $d_{k+1}^T y_k = 0$;

(iii) the orthogonality property, namely,

$$g_{k+2}^T d_i = 0,\ i = 0, \ldots, k+1.$$

Using the same idea as [9,10], here we combine (i)–(iii) and propose a suitable parameter t by solving the following optimization problem:

$$\min_t g_{k+1}^T d_{k+1} + M\big(|g_{k+2}^T d_{k+1}| + (d_{k+1}^T y_k)^2\big), \tag{10}$$

where M is a penalty parameter. As seen, the first, the second and the third term in (10) measure the validity of (i), (ii) and (iii), respectively. In what

follows, we convert the nonsmooth problem (10) to an equivalent smooth optimization problem.

Assuming d_{k+1} is a descent direction and using the Wolfe conditions (2) and (3), we have

$$g_{k+2}^T d_{k+1} = y_{k+1}^T d_{k+1} + g_{k+1}^T d_{k+1} < y_{k+1}^T d_{k+1},$$

and

$$g_{k+2}^T d_{k+1} \geq \sigma g_{k+1}^T d_{k+1} = -\sigma y_{k+1}^T d_{k+1} + \sigma g_{k+2}^T d_{k+1}.$$

The above inequalities imply that

$$|g_{k+2}^T d_{k+1}| \leq \max\left\{\frac{\sigma}{1-\sigma}, 1\right\} y_{k+1}^T d_{k+1}. \tag{11}$$

Now, we replace the absolute term in (10) by the right–hand side of (11), and propose the following equivalent smooth problem:

$$\min_t g_{k+1}^T d_{k+1} + M\big(y_{k+1}^T d_{k+1} + (d_{k+1}^T y_k)^2\big). \tag{12}$$

In order to solve (12), we should suggest an estimation for $y_{k+1} = g_{k+2} - g_{k+1}$, because g_{k+2} is an unknown vector. Considering

$$\Phi(d_{k+1}) = f_{k+1} + g_{k+1}^T d_{k+1} + \frac{1}{2} d_{k+1}^T B_{k+1} d_{k+1},$$

as a quadratic approximation of the objective function, we get

$$\nabla\Phi(\alpha_{k+1} d_{k+1}) = \alpha_{k+1} B_{k+1} d_{k+1} + g_{k+1},$$

being the gradient of Φ at x_{k+2} which can be considered as an approximation for g_{k+2}. Therefore, we have

$$y_{k+1} = g_{k+2} - g_{k+1} \approx \alpha_{k+1} B_{k+1} d_{k+1}. \tag{13}$$

Now, substituting (13) in (12), we obtain the following easier to solve problem:

$$\min_t g_{k+1}^T d_{k+1} + M\big(\alpha_{k+1} d_{k+1}^T B_{k+1} d_{k+1} + (d_{k+1}^T y_k)^2\big). \tag{14}$$

The problem (14) can be solved by substituting (5) in (14) and differentiating with respect to t. After some algebraic manipulations, the optimal solution of (14) can be given as follows:

$$t_{opt} = \frac{1}{2M(\alpha_{k+1} + s_k^T y_k)} - \frac{||y_k||^2}{s_k^T y_k}, \tag{15}$$

provided that B_{k+1} to be updated using the standard secant equation $B_{k+1}s_k = y_k$ [1]. Next, using (15), we suggest two adaptive choices for the TTDL parameter t, being independent of M.

The t_{opt} parameter given by (15) clearly depends on the choice of the penalty parameter M. A desirable value for M is infinity, because it increases the chance of satisfying (ii)–(iii), forcing TTDL method to reflect the good features of a linear CG method as much as possible. Approaching M to infinity, we obtain

$$t_{opt}^{\infty} = -\frac{||y_k||^2}{s_k^T y_k}. \tag{16}$$

Unfortunately, since $t_{opt}^{\infty} \leq 0$, the TTDL method with $t = t_{opt}^{\infty}$ may not guarantee the descent property. It is reasonably clear from (14) that the descent property is lost when M approaches to infinity.

As mentioned in Section 1, if the TTDL parameter is nonnegative, then the method satisfies the sufficient descent condition (6). It can be seen that $t_{opt} > 0$ if we have

$$M < \frac{s_k^T y_k}{2(\alpha_{k+1} + s_k^T y_k)||y_k||^2}.$$

So, based on the above inequality we suggest the following choice for M:

$$M = \eta \frac{s_k^T y_k}{2(\alpha_{k+1} + s_k^T y_k)||y_k||^2}, \tag{17}$$

where $0 < \eta < 1$ is some fixed constant. To justify the choice (17), note that in vicinity of the optimal solution, $||y_k||$ approaches to zero and consequently, M tends to infinity provided that $\alpha_{k+1}\frac{||y_k||^2}{s_k^T y_k}$ be sufficiently small. In practice, we observe this situation for approximately all of the test problems. Using (17), we get our first choice for t as follows:

$$t_{opt}^{(1)} = (\frac{1}{\eta} - 1)\frac{||y_k||^2}{s_k^T y_k}. \tag{18}$$

In another scheme, we allow the TTDL parameter t to get nonpositive values in the sense of setting $t = t_{opt}^{\infty}$ when the sufficient descent condition (6) is satisfied, increasing the chance of M to be infinity. It can be seen that if

$$||y_k|||g_{k+1}^T s_k| \leq \gamma ||g_{k+1}||(s_k^T y_k), \tag{19}$$

with some fixed constant $\gamma \in (0,1)$, then the TTDL method with $t = t_{opt}^{\infty}$ fulfills the sufficient descent condition (6) with $c = 1-\gamma^2$. Hence, we suggest our second choice for the TTDL parameter t as follows:

$$t_{opt}^{(2)} = \begin{cases} t_{opt}^{\infty}, & \text{if (19) holds,} \\ \\ t_{opt}^{(1)}, & \text{otherwise,} \end{cases} \tag{20}$$

being an adaptive combination of t_{opt}^{∞} and $t_{opt}^{(1)}$.

Here, we discuss global convergence of the TTDL method with the choices (18) and (20). In our analysis, we need to make the following basic assumptions on the objective function.

Assumption 1. *The level set $\mathcal{L} = \{x|\ f(x) \leq f(x_0)\}$ is bounded. Also, in some open convex neighborhood $\mathcal{N}$ of $\mathcal{L}$, f is continuously differentiable and its gradient is Lipschitz continuous; that is, there exists a positive constant L such that*

$$||\nabla f(x) - \nabla f(y)|| \leq L||x-y||,\ \forall x, y \in \mathcal{N}. \tag{21}$$

Now, we can establish the following global convergence theorem, using Theorem 2.1 of [12].

Theorem 1. *Suppose that Assumption 1 holds. Consider the TTDL method in which the parameter t in (5) is computed by (18) or (20), and the step length α_k is determined such that the Wolfe conditions (2) and (3) are satisfied. If the objective function f is uniformly convex on $\mathcal{N}$, then the method converges in the sense that $\lim_{k\to\infty} ||g_k|| = 0$.*

Proof. Note that the choices (7) or (9) for the TTDL parameter t guarantee the the sufficient descent condition (6). So, from the line search condition (2), we have $\{x_k\}_{k\geq 0} \subseteq \mathcal{L}$. Also, uniform convexity of the differentiable function f ensures that there exists a positive constant μ such that

$$s_k^T y_k \geq \mu ||s_k||^2,\ k = 0, 1, ..., \tag{22}$$

(see Theorem 1.3.16 of [1]). Now, from (21) and (22) we get

$$t_{opt}^{(1)} \leq (\frac{1}{\eta} - 1)\frac{L^2}{\mu},$$

and

$$|t_{opt}^{\infty}| \leq \frac{L^2}{\mu}.$$

Thus, both of the choices $t_{opt}^{(1)}$ and $t_{opt}^{(2)}$ are bounded above by a positive constant. Remainder of the proof is similar to the proof of Theorem 2.1 of [12] and here is omitted. □

It is worth noting that inequality (19) may hold in vicinity of the optimal solution under the Lipschitz assumption (21). More exactly, considering (11), (21) and the Cauchy–Schwarz inequality, we have

$$||g_{k+1}||(s_k^T y_k) = O(||s_k||^2),$$

and

$$||y_k|||g_{k+1}^T s_k| \leq \max\left\{\frac{\sigma}{1-\sigma}, 1\right\} ||y_k||(s_k^T y_k) = O(||s_k||^3).$$

Therefore, near the optimal solution whenever $||s_k||$ tends to zero, inequality (19) may be satisfied.

3 Numerical experiments

Here, we present some numerical results obtained by applying MATLAB implementations of the following four versions of the TTDL method:

- TTDL1: the TTDL method with the choice $t_{opt}^{(1)}$ given by (18);
- TTDL2: the TTDL method with the choice $t_{opt}^{(2)}$ given by (20);
- TTDL3: the TTDL method with the choice (7);
- TTDL4: the TTDL method with the choice (9).

The codes were run on a Laptop computer with 2.4 GHz Intel I7–5500 of CPU, 16 GB of RAM and Ubuntu 15.10 Linux operation system. Since CG methods are mainly appropriate for solving large–scale problems, the experiments were performed on a set of 64 unconstrained optimization test problems of the CUTEr collection [17] with default dimensions being at least equal to 1000, as specified in [18].

For all the four methods, we used the approximate Wolfe conditions proposed by Hager and Zhang [19] in the line search procedure, with the same

parameter values as specified in [20]. For TTDL1 and TTDL2, we respectively set $\eta = 0.8$ in (18) and $\gamma = 0.98$ in (20) because of their promising computational results among the different values 0.1, 0.2, 0.3, 0.5, 0.8, 0.9, 0.98. Also, for TTDL3 and TTDL4 we respectively set $\xi = 0.66$ in (7) and $\xi = 0.4$ in (9), as suggested in [12, 14]. In addition, all attempts to solve the test problems were terminated when $||g_k||_\infty < 10^{-6}(1 + |f(x_k)|)$.

Efficiency comparisons were made using the Dolan–Moré performance profile [16] which for every $\omega \geq 1$ yields the proportion $p(\omega)$ of the test problems that each considered algorithmic variant has a performance within a factor of ω of the best. Figures 1 and 2 show the results of comparisons. Note that in Figure 1 total number of function and gradient evaluations is equal to $N_f + 3N_g$ where N_f and N_g respectively stand for the number of function and gradient evaluations [20].

As shown by the figures, although TTDL2 outperforms the other methods both in the perspectives of the total number of function and gradient evaluations and the running time, TTDL1 has an unpromising computational performance in contrast to the other methods, especially with respect to the CPU time. Also, the figures show that TTDL4 is preferable to TTDL3 with respect to the total number of function and gradient evaluations while the method is dominated by TTDL3 with respect to the running time. As a result, the numerical comparisons imply that our hybrid adaptive choice (20) for the TTDL parameter turns out to be practically effective.

4 Conclusions

Two adaptive choices have been proposed for parameter of a three–term version of the Dai–Liao conjugate gradient method (TTDL) by solving a one–dimensional unconstrained optimization problem which simultaneously contains the linear conjugate gradient aspects of descent, conjugacy and orthogonality. As established, the suggested parameter choices ensure the sufficient descent property for general objective functions as well as the global convergence for uniformly convex objective functions. Numerical experiments have been done on a set of CUTEr test problems. The results show that one of the proposed choices lead to an efficient nonlinear conjugate gradient method.

Acknowledgements: This research was supported by Research Councils of K. N. Toosi University of Technology and Semnan University. The authors

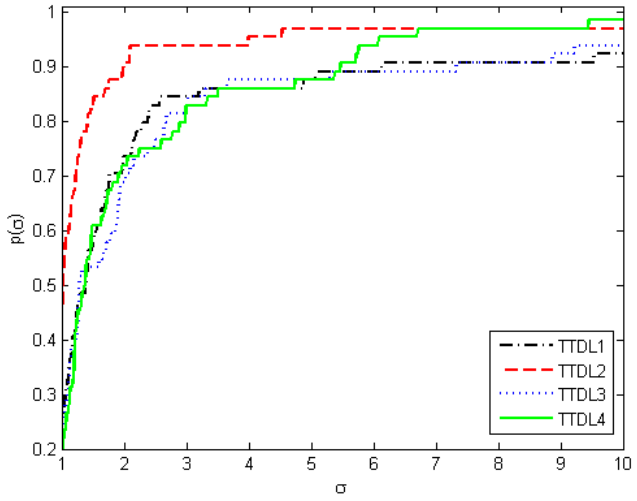

Fig. 1: Total number of function and gradient evaluations performance profiles

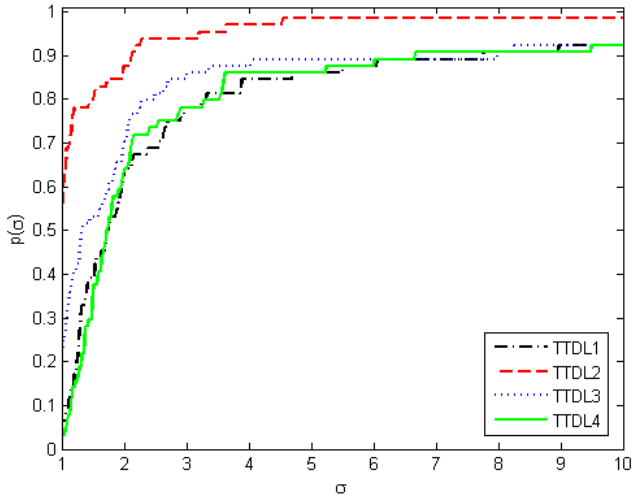

Fig. 2: CPU time performance profiles

are grateful to Professor William W. Hager for providing the line search code.

References

[1] W. Sun and Y.X. Yuan. *Optimization Theory and Methods: Nonlinear Programming*. Springer, New York, 2006. ISBN: 978-0-387-24976-6.

[2] N. Andrei. Numerical comparison of conjugate gradient algorithms for unconstrained optimization. *Stud. Inform. Control*, 16(4):333–352, 2007. https://www.researchgate.net/publication/228811831_Numerical_comparison_of_conjugate_gradient_algorithms_for_unconstrained_optimization.

[3] W.W. Hager and H. Zhang. A survey of nonlinear conjugate gradient methods. *Pac. J. Optim.*, 2(1):35–58, 2006. http://www.ybook.co.jp/online2/oppjo/vol2/p35.html.

[4] Y.H. Dai and L.Z. Liao. New conjugacy conditions and related nonlinear conjugate gradient methods. *Appl. Math. Optim.*, 43(1):87–101, 2001. doi: 10.1007/s002450010019.

[5] M.R. Hestenes and E. Stiefel. Methods of conjugate gradients for solving linear systems. *J. Research Nat. Bur. Standards*, 49(6):409–436, 1952. https://archive.org/details/jresv49n6p409.

[6] S. Babaie–Kafaki and R. Ghanbari. The Dai–Liao nonlinear conjugate gradient method with optimal parameter choices. *European J. Oper. Res.*, 234(3):625–630, 2014. doi:10.1016/j.ejor.2013.11.012.

[7] S. Babaie–Kafaki and R. Ghanbari. Two optimal Dai–Liao conjugate gradient methods. *Optimization*, 64(11):2277–2287, 2015. doi: 10.1080/02331934.2014.938072.

[8] S. Babaie–Kafaki and R. Ghanbari. A descent family of Dai–Liao conjugate gradient methods. *Optim. Methods Softw.*, 29(3):583–591, 2014. doi: 10.1080/10556788.2013.833199.

[9] M. Fatemi. An optimal parameter for Dai–Liao family of conjugate gradient methods. *J. Optim. Theory Appl.*, pages 1–19, 2015. doi: 10.1007/s10957-015-0786-9.

[10] M. Fatemi. A new efficient conjugate gradient method for unconstrained optimization. *J. Comput. Appl. Math.*, page 207–216, 2016. doi: 10.1016/j.cam.2015.12.035.

[11] Y.H. Dai and C.X. Kou. A nonlinear conjugate gradient algorithm with an optimal property and an improved Wolfe line search. *SIAM J. Optim.*, 23(1):296–320, 2013. doi: 10.1137/100813026.

[12] S. Babaie–Kafaki and R. Ghanbari. Two modified three–term conjugate gradient methods with sufficient descent property. *Optim. Lett.*, 8(8):2285–2297, 2014. doi: 10.1007/s11590-014-0736-8.

[13] L. Zhang, W. Zhou, and D.H. Li. Some descent three–term conjugate gradient methods and their global convergence. *Optim. Methods Softw.*, 22(4):697–711, 2007. doi: 10.1093/imanum/drl016.

[14] S. Babaie–Kafaki and R. Ghanbari. A class of descent four–term extension of the Dai–Liao conjugate gradient method based on the scaled memoryless BFGS update. *J. Ind. Manag. Optim.*, 2016, to appear.

[15] S.S. Oren and E. Spedicato. Optimal conditioning of self–scaling variable metric algorithms. *Math. Program.*, 10(1):70–90, 1976. doi: 10.1007/BF01580654.

[16] E.D. Dolan and J.J. Moré. Benchmarking optimization software with performance profiles. *Math. Program.*, 91(2, Ser. A):201–213, 2002. doi: 10.1007/s101070100263.

[17] N.I.M. Gould, D. Orban, and Ph.L. Toint. CUTEr and SifDec: a constrained and unconstrained testing environment, revisited. *ACM Trans. Math. Softw.*, 29(4):373–394, 2003. doi: 10.1145/962437.962439.

[18] S. Babaie–Kafaki and R. Ghanbari. A descent extension of the Polak–Ribière–Polyak conjugate gradient method. *Comput. Math. Appl.*, 68(12):2005–2011, 2014. doi: 10.1155/2014/921364.

[19] W.W. Hager and H. Zhang. A new conjugate gradient method with guaranteed descent and an efficient line search. *SIAM J. Optim.*, 16(1):170–192, 2005. doi: 10.1137/030601880.

[20] W.W. Hager and H. Zhang. Algorithm 851: CG_Descent, a conjugate gradient method with guaranteed descent. *ACM Trans. Math. Softw.*, 32(1):113–137, 2006. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.367.7043.