Simple Scaling for Variable Metric Updates

Lukšan, Ladislav
1994

INSTITUTE OF COMPUTER SCIENCE

ACADEMY OF SCIENCES OF THE CZECH REPUBLIC

Prague

Simple Scaling for Variable Metric Updates

Ladislav Lukšan, Jan Vlček

Technical Report No. V-611

May 1995

Akademie věd České republiky

ÚSTAV INFORMATIKY A VÝPOČETNÍ TECHNIKY

Institute of Computer Science, Academy of Sciences of the Czech Republic

Pod vodárenskou věží 2, 182 07 Prague 8, Czech Republic

E-mail: ICS@uivt.cas.cz

Fax: (+422) 8585789 Phone: (+422) 846669, (+422) 66051111

Simple Scaling for Variable Metric Updates [1]

L. Lukšan and J. Vlček[2]

Abstract.

In this contribution, we propose an extremely simple scaling strategy which considerably decreases number of function evaluations if variable metric methods from the Broyden class are used for unconstrained minimization of functions with number of variables not extremely small. We confine our attention to BFGS method, since scaling of other variable metric methods has very similar properties. After describing our scaling strategy, we compare six scaling techniques, using an extensive collection of test problems, and present some conclusions.

Key Words Variable metric methods, self scaling, optimization, nonlinear programming.

1. Introduction

Consider the variable metric method whose iteration step has the form

$$x^+ = x + \alpha s \tag{1}$$

where

$$s = -Hg. \tag{2}$$

Here x and x^+ are old and new vectors of variables, respectively. Direction vector s and positive stepsize α are chosen so that

$$F^+ - F \leq \varepsilon_1 \alpha\, s^T g, \tag{3}$$

$$s^T g^+ \geq \varepsilon_2 s^T g, \tag{4}$$

with $0 < \varepsilon_1 < 1/2$ and $\varepsilon_1 < \varepsilon_2 < 1$, where F and F^+ are old and new values of the objective function and g and g^+ are old and new gradients of the objective function, respectively. H is a symmetric positive-definite approximation of the inverse Hessian matrix that is constructed iteratively using the formula

[1]This work was supported by the Czech Republic Grant Agency, Grant No. 201/93/0429.

[2]Researchers, Institute of Computer Science, Academy of Sciences of the Czech Republic, Prague 8, Czech Republic.

$$H^+ = \gamma \left[H + \left(\frac{a}{b} + \frac{1}{\gamma} \right) \frac{1}{b} dd^T - \frac{1}{b} \left(Hyd^T + dy^T H \right) \right]. \tag{5}$$

This is the so-called scaled BFGS method (see Ref.1 for further details). We confine our attention to this particular method, since scaling of other variable metric methods from the Broyden class has very similar properties. It was also confirmed by our extensive experiments. In (5) and in the subsequent text we use the notation

$$y = g^+ - g, \tag{6}$$

$$d = x^+ - x = \alpha s, \tag{7}$$

$$a = y^T H y, \tag{8}$$

$$b = d^T y = \alpha (s^T g^+ - s^T g), \tag{9}$$

$$c = d^T H^{-1} d = -\alpha^2 s^T g, \tag{10}$$

and we suppose that $b > 0$, as it follows from (4) and (9).

2. Basic scaling strategies

The use of the scaling parameter γ in (5) was introduced by Oren and Luenberger in Ref 2. In spite of deep theories to derive the value γ (Refs. 3-4) it was shown, that most suitable choices are

$$\gamma = b/a = y^T d / y^T H y, \tag{11}$$

$$\gamma = c/b = \alpha s^T g / (s^T g - s^T g^+), \tag{12}$$

and, eventually, their geometric mean. The value (11) was used by Shano and Phua in Ref. 5 and the theory connected with the value (12) was proposed by Nocedal and Yuan in Ref. 6. These values are usually combined with the value $\gamma = 1$ using some scaling strategy. The following simple strategies are those which are used most often:

(NS) No scaling. We set $\gamma = 1$ in each iteration.

(PS) Preliminary scaling of Shano and Phua (Ref. 5). We use either (11) or (12) in the first iteration and $\gamma = 1$, otherwise. More precisely, we set $\gamma = \underline{\Gamma}$ or $\gamma = \overline{\Gamma}$ in the first iteration, if the scaling factor does not lie in the interval $[\underline{\Gamma}, \overline{\Gamma}]$.

(AS) Scaling in all iterations. We use either (11) or (12) in all iterations. Again we set $\gamma = \underline{\Gamma}$ or $\gamma = \overline{\Gamma}$ in the first iteration, if the scaling factor does not lie in the interval $[\underline{\Gamma}, \overline{\Gamma}]$.

The values $\underline{\Gamma}$ and $\overline{\Gamma}$ serve as a safeguard (usually $\underline{\Gamma} = 0.01$ and $\overline{\Gamma} = 100$). Besides these simple strategies, more sophisticated ones were recently proposed.

(CS) Controlled scaling of Lukšan (Refs. 7-8). We proceed as in (PS) or (AS) in the first iteration. In other iterations, let α_1 be an initial choice of stepsize (usually $\alpha_1 = 1$) and let $\lambda_1 = s^T g_1 / s^T g$, where $F_1 = F(x + \alpha_1 s)$ and $g_1 = g(x + \alpha_1 s)$. Denote (11) or (12) by $\hat{\gamma}$. Then we set $\gamma = 1$ if at least one of the following conditions holds

(a) $|\lambda_1| \leq \lambda$ and $F_1 \leq F$,
(b) $\hat{\gamma} > 1$ and either $F_1 > F$ or $\lambda_1 < 0$,
(c) $\hat{\gamma} < 1$ and $F_1 \leq F$ and $\lambda_1 > 0$,
(d) $\hat{\gamma} < \underline{\gamma}$ or $\hat{\gamma} > \overline{\gamma}$,

and $\gamma = \hat{\gamma}$ otherwise. The values $\underline{\gamma}$, $\overline{\gamma}$ and λ were obtained experimentally.

(SS) Selective scaling of Contreras and Tapia (Ref. 9). We proceed as in (PS) or (AS) in the first iteration. In other iterations, we compute the so-called centered Oren-Luenberger factor

$$\hat{\gamma} = \frac{b^-/\|d^-\|^2 + c/\|d\|^2}{b^-/\|d^-\|^2 + b/\|d\|^2},$$

where b^- and $\|d^-\|$ are quantities from the previous iteration. Then we set $\gamma = \min(\hat{\gamma}, \overline{\Gamma})$, if $\hat{\gamma} \geq \underline{\gamma} > 1$, and $\gamma = 1$ otherwise.

In this contribution we concentrate on two aims. First, we want to show that the last two scaling strategies can be substantially simplified without loss of their efficiency. We found experimentally that almost all parts of these strategies are irrelevant, the only substantial part is to keep the scaling factor in a suitable interval (part (d) of (CS) or the last decision in (SS)). Second, we want to provide appropriate range for both the values (11) and (12). As a result of our study, we propose the following new strategy.

(IS) Interval scaling. We proceed as in (PS) or (AS) in the first iteration. In other iterations, we use (11) or (12) if $\underline{\gamma} \leq \gamma \leq \overline{\gamma}$, and $\gamma = 1$, otherwise. Here $\underline{\gamma} = 0.8$, $\overline{\gamma} = 8.0$ for the value (11) and $\underline{\gamma} = 0.6$, $\overline{\gamma} = 4.0$ for the value (12).

Our interval scaling strategy is based on two requirements: the initial stepsize acceptability, and the fast ultimate rate of convergence. The first requirement is essential since it helps to decrease the number of function evaluations when starting from points that are far from the solution. On the other hand, in Ref. 6 it was proved

that scaling in all iterations can decrease the ultimate rate of convergence due to oscillation of values of the scaling parameter. This oscillation could be eliminated by setting $\gamma = 1$ in the case, when the value γ lies outside the interval $[\underline{\gamma}, \overline{\gamma}]$. The values $\underline{\gamma}$ and $\overline{\gamma}$ were determined experimentally using extensive computational experiments.

Let us explain our comparison of individual scaling strategies. This comparison was performed using the collection of test problems used in Ref. 8. This collection contains problems 1-15 from Ref. 7 (which are modifications of problems proposed in Ref. 11) and problems 21-30 from Ref. 10. The problems can be utilized with a variable dimension n and we used $n = 20$ in all cases. We chose this dimension, since scaling strategies are inefficient for extremely small problems (like problems 1-18 from Ref. 10). The comparison was carried out using the modular system for universal functional optimization UFO described in Ref.12. In CS we used the same values $\underline{\gamma}$ and $\overline{\gamma}$ as in IS (these values are given above) while in SS we used the original values from Ref. 9. Furthermore we used $\varepsilon_1 = 0.0001$ and $\varepsilon_2 = 0.9$ in (3) and (4). The iterative process was terminated, whenever the gradient norm was less than 10^{-8}. Summary of comparisons of individual scaling strategies is given in Table 1, which contains total number of iterations NIT and total number of function evaluations NFV for summary of problems (results for NS and SS are introduced in the first column even if they do not use scaling factor(11)). Detailed comparison of NS and IS with scaling factor (11) is given in Table 2, which contains the number of iterations NIT, the number of function evaluations NFV and the reached function value F for every test problem. We chose just this comparison to show the fact that our new scaling strategy improves an efficiency of the robust BFGS method.

Table 1: *Comparison of various scaling strategies.*

	Scaling factor (11)		Scaling factor (12)	
Scaling strategy	NIT	NFV	NIT	NFV
NS	3180	4576		
PS	2377	2793	2363	2793
AS	3404	3914	2466	3198
IS	1452	1834	1442	1976
CS	1527	1903	1444	2083
SS	1472	2059		

Table 2: *NS versus IS using the value (11).*

	no scaling (NS)			interval scaling (IS)		
No.	NIT	NFV	F	NIT	NFV	F
1	133	197	4.3E-19	105	137	2.2E-19
2	221	307	1.5E-19	116	139	1.5E-17
3	115	154	2.3E-17	80	88	1.1E-13
4	123	200	3.494213680	68	88	3.494213680
5	64	86	4.1E-16	23	24	1.1E-15
6	87	111	7.9E-16	28	29	1.2E-15
7	34	70	6.392162882	22	23	6.392162882
8	48	98	2.4E-24	26	48	1.6E-22
9	42	65	- 2500.000000	32	51	-2500.000000
10	673	1195	143.5187274	361	505	143.5187274
11	163	208	0.215531757	121	151	0.215531757
12	13	28	1.997866137	12	21	1.997866137
13	1	2	0.000000000	1	2	0.000000000
14	35	51	1.2E-16	57	59	1.5E-18
15	22	42	-8.510866851	17	19	-8.510866851
16	40	52	3.3E-20	33	44	1.8E-20
17	50	54	6.1E-14	54	62	1.4E-17
18	64	86	7.888853E-5	21	28	7.888853E-5
19	1079	1347	3.194840E-3	108	134	3.194840E-3
20	3	6	2.01E-24	3	6	2.1E-24
21	57	63	3.430929E-6	52	58	3.430929E-6
22	30	32	7.3E-22	19	21	3.8E-16
23	40	54	1.1E-16	56	58	1.2E-16
24	7	8	1.1E-18	8	9	1.1E-16
25	36	60	0.356263954	29	30	1.5E-17
TOTAL	3036	4429		1367	1733	

According results presented in tables 1-2 and our additional tests, we can add several comments. First, the robust BFGS method can be improved by a suitable scaling technique, at least for problems contained in our collection. It follows from Table 2. Even if most function evaluations of the NS strategy were taken by solving two problems (10 and 19), the IS strategy gave better results also in the remaining cases (NS strategy was better only for problems 14,17,23,24 and the total number of saved function evaluations for these problems was 21). Failure of NS strategy on problems 10 and 19 slightly distorted Table 1. In fact the NS strategy was better than AS strategy for remaining problems, but other strategies again outperformed the NS one (in the case of remaining 23 problems we needed 2034 function evaluations for NS, 1332 for PS, 2076 for AS, 1195 for IS, 1190 for CS and 1348 for SS). Second, similar properties were observed for other variable metric methods from the Broyden class. In fact, different scaling factors ((11), (12) or their geometric mean) are suitable for different methods, but the recomended intervals $[\underline{\gamma}, \overline{\gamma}]$ are the same for the same

scaling factors.

References

1. FLETCHER, R., *Practical Methods of Optimization*, 2nd Edition, J.Wiley and Sons, Chichester, England, 1987.

2. OREN, S.S., and LUENBERGER, D.G., *Self Scaling Variable Metric (SSVM) Algorithms. Part 1 - Criteria and Sufficient Condition for Scaling a Class of Algorithms. Part 2 - Implementation and Experiments*, Management Science, Vol. 20, pp. 845-874, 1974.

3. OREN, S.S., and SPEDICATO, E., *Optimal Conditioning of Self Scaling Variable Metric Algorithms*, Mathematical Programming, Vol. 10, pp. 70-90, 1976.

4. LUKŠAN, L., *Variationally Derived Scalling and Variable Metric Updates From the Preconvex Part of the Broyden Family*, Journal of Optimization Theory and Applications, Vol. 73, pp. 299-307, 1992.

5. SHANO, D.F., and PHUA, K.J. *Matrix Conditioning and Nonlinear Optimization*, Mathematical Programming, Vol.14, pp. 144-160, 1978

6. J.NOCEDAL, Y.YUAN, Y., *Analysis of a Self-Scaling Quasi-Newton Method*, Mathematical Programming, Vol. 61, pp. 19-37, 1993.

7. LUKŠAN, L., *Computational Experience with Improved Variable Metric Methods for Unconstrained Minimization*, Kybernetika, Vol. 26, pp. 415-431, 1990.

8. LUKŠAN, L., *Computational experience with known variable metric updates*, Journal of Optimization Theory and Applications, Vol. 83, pp. 27-47, 1994.

9. CONTRERAS, C., and TAPIA, R.A., *Sizing the BFGS and DFP Updates: Numerical Study*, Journal of Optimization Theory and Applications, Vol. 78, pp. 93-108, 1993.

10. MORÉ, J.J., GARBOV, B.S., and HILLSTRÖM, K.E., *Testing Unconstrained Optimization Software*, ACM Transactions on Mathematical Software, Vol. 7, pp. 17-41, 1981.

11. CONN, A.R., GOULD, N.I.M., and TOINT, P.L., *Testing a Class of Methods for Solving Minimization Problems with Simple Bounds on the Variables*, Mathematics of Computation, Vol. 50, pp. 399-430, 1998.

12. LUKŠAN, L., ŠIŠKA, M., TŮMA, M., VLČEK, J., and RAMEŠOVÁ, N., *Interactive System for Universal Functional Optimization (UFO), Version 1994*, Research Report No. 599, Institute of Computer Science, Academy of Sciences of the Czech Republic, Prague, Czech Republic, 1994.