A spectral conjugate gradient method for unconstrained optimization *

Ernesto G. Birgin † José Mario Martínez ‡

May 18, 1998

Abstract

A family of scaled conjugate-gradient algorithms for large-scale unconstrained minimization is defined. The Perry, the Polak-Ribière and the Fletcher-Reeves formulae are compared using a spectral scaling derived from Raydan's spectral gradient optimization method. The best combination of formula, scaling and initial choice of steplength is compared against classical algorithms using a classical set of problems. An additional comparison involving an estimation problem in Optics is presented.

Keywords. Unconstrained minimization, spectral gradient method, conjugate gradients.
AMS: 49M07, 49M10, 90C06, 65K.

1 Introduction

In a recent paper [8] Raydan introduced the spectral gradient method (SGM) for potentially large-scale unconstrained optimization. The main feature of this method is that only gradient directions are used at each line search and a nonmonotone strategy guarantees global convergence. Surprisingly, this method outperforms sophisticated conjugate gradient algorithms in many problems. The numerical results in [2, 5, 6, 8] and others suggested us that

* Abreviated: Spectral conjugate gradient method

† Department of Applied Mathematics, IMECC-UNICAMP, University of Campinas, CP 6065, 13081-970 Campinas SP, Brazil. This author was supported by FAPESP (Grant 95-2452-6). e-mail: ernesto@ime.unicamp.br

‡ Department of Applied Mathematics, IMECC-UNICAMP, University of Campinas, CP 6065, 13081-970 Campinas SP, Brazil. This author was supported by FAPESP (Grant 90-3724-6), CNPq and FAEP-UNICAMP. e-mail: martinez@ime.unicamp.br

spectral gradient and conjugate gradient ideas could be combined in such a way that even more efficient algorithms could be obtained.

Assume that $f : I\!R^n \to I\!R$ has continuous partial derivatives, and denote $g(x) = \nabla f(x)$. In the minimization methods considered in this paper iterates are obtained by means of

$$x_{k+1} = x_k + \alpha_k d_k,$$

and

$$d_{k+1} = -\theta_k g_{k+1} + \beta_k s_k \tag{1}$$

for $k = 0, 1, 2, \ldots$, where g_k denotes $g(x_k)$, $x_0 \in I\!R^n$ is arbitrary and

$$d_0 = -\theta_0 g_0.$$

Suppose that x_k and x_{k+1} are two consecutive points. Denote, also, $s_k = x_{k+1} - x_k = \alpha_k d_k$ and $y_k = g_{k+1} - g_k$. Suppose, for a moment, that f is quadratic and $H \equiv \nabla^2 f(x)$ is positive definite. This implies that $y_k \neq 0$. Therefore, the true minimizer x_* satisfies

$$x_* = x_{k+1} + d_*$$

where

$$H d_* = -g_{k+1}.$$

Pre-multiplying by s_k^T, this gives

$$s_k^T H d_* = -s_k^T g_{k+1},$$

which implies that

$$y_k^T d_* = -s_k^T g_{k+1}.$$

Therefore, the hyperplane

$$\mathcal{H}_k \equiv \{d \in I\!R^n \ \mid \ y_k^T d = -s_k^T g_{k+1}\}$$

contains the optimum increment d_* which gives $x_* = x_{k+1} + d_*$. Observe that the null direction $d = 0$ belongs to $\mathcal{H}$ only if $s_k^T g_{k+1} = 0$ which is not our assumption at all.

So, it is natural to impose that the search direction d_{k+1}, which is going to be computed by an algorithm designed to minimize f, satisfy

$$d_{k+1} \in \mathcal{H}_k. \tag{2}$$

So, by (1),

$$\beta_k = \frac{(\theta_k y_k - s_k)^T g_{k+1}}{s_k^T y_k}. \tag{3}$$

For $\theta_k = 1$ this formula was introduced by Perry in [7]. If we assume that $s_j^T g_{j+1} = 0$, $j = 0, 1, \ldots, k$, we obtain

$$\beta_k = \frac{\theta_k y_k^T g_{k+1}}{\alpha_k \theta_{k-1} g_k^T g_k}. \tag{4}$$

If $\theta_k = \theta_{k-1} = 1$ this is the classical Polak-Ribière formula. Finally, assuming also that the successive gradients are orthogonal, we obtain the generalization of Fletcher-Reeves formula:

$$\beta_k = \frac{\theta_k g_{k+1}^T g_{k+1}}{\alpha_k \theta_{k-1} g_k^T g_k}. \tag{5}$$

In this paper, motivated by the success of the spectral gradient method, we decided to compare the classical choice $\theta_k = 1$ with the spectral gradient choice:

$$\theta_k = s_k^T s_k / s_k^T y_k. \tag{6}$$

In fact the directions $d_k = -\theta_k g_k$ are the ones used by Raydan in his spectral gradient method. The parameter θ_k given by (6) is the inverse of the Rayleigh quotient

$$s_k^T [\int_0^1 \nabla^2 f(x_k + t s_k) dt] s_k / s_k^T s_k$$

which, of course, lies between the largest and the smallest eigenvalue of the Hessian average $\int_0^1 \nabla^2 f(x_k + t s_k) dt$.

Moreover, after some numerical experimentation, we observed that the initial trial choice for the steplength α_k is also a very important parameter that influentiates the algorithmic behavior. So, we decided to test also two different alternatives for this choice.

This paper is organized as follows. In Section 2 we present the model algorithm, giving all the essential features of its implementation. In Section 3 we use the set of test problems given in [8] to answer the following questions:

1. Is the choice (6) better than $\theta_k \equiv 1$?
2. Which is the best choice for β_k, among (3), (4) and (5)?
3. Which is the best initial choice for the steplength?

In Section 4 we compare the new algorithm against CONMIN (a popular conjugate-gradient code based on [9, 10]) and the spectral gradient method of Raydan, using the same test functions of Section 3. In Section 5 we compare the new algorithm against the spectral gradient method using a real-life estimation problem in Optics. Conclusions are given in Section 6.

2 The algorithm

Keeping in mind the definitions of g_k, s_k and y_k given in the Introduction, we define the Scaled Conjugate Gradient method as follows:

Algorithm SCG
Assume that $x_0 \in I\!R^n$, $0 < \sigma < \gamma < 1$. Define $d_0 = -g_0$ and set $k \leftarrow 0$.

Step 1: If $g_k = 0$, terminate the execution of the algorithm.

Step 2: Compute (trying first $\alpha = \bar{\alpha}(k, d_k, d_{k-1}, \alpha_{k-1})$) $\alpha > 0$ such that

$$f(x_k + \alpha d_k) \leq f(x_k) + \sigma \alpha g_k^T d_k \tag{7}$$

and

$$g(x_k + \alpha d_k)^T d_k \geq \gamma g_k^T d_k. \tag{8}$$

Define $\alpha_k = \alpha$ and

$$x_{k+1} = x_k + \alpha_k d_k.$$

Step 3: Compute θ_k by (6) (or $\theta_k = 1$) and β_k by (3), (4) or (5).
Define

$$d = -\theta_k g_{k+1} + \beta_k s_k. \tag{9}$$

If

$$d^T g_{k+1} \leq -10^{-3} \|d\| \|g_{k+1}\| \tag{10}$$

define $d_{k+1} = d$. Otherwise, define

$$d_{k+1} = -\theta_k g_{k+1}.$$

Step 4: Set $k \leftarrow k + 1$ and go to Step 2.

It is well known (see [3, 4]) that a steplength α satisfying (7, 8) always exists if f is bounded below along the direction d_k. We assume that we have an algorithm that either computes α with those conditions or detects that

f is unbounded below. In this case, we say that SCG breaks at iteration k. In practice, we adopted the one-dimensional line search used in CONMIN (see [10]) for computing α.

The search direction d computed by (9) can fail to be a descent direction. This is the reason that motivated several modifications of Perry's formula in [9]. When the angle between d and $-g_{k+1}$ is not acute enough we "restart" the algorithm with the spectral gradient direction $-\theta_k g_{k+1}$. More sophisticated reasons for restarting have been proposed in the literature, but we are interested on the performance of an algorithm that uses this naive criterion, associated to the spectral gradient choice for restarts. Of course, the coefficient θ_k is always well defined and positive, since (8) implies that $s_k^T y_k > 0$.

3 Discussion of alternatives

In this section we use the test problems considered by [8] to answer the questions formulated in the Introduction. With this purpose, we consider the algorithm SCG with $\sigma = 10^{-4}$ and $\gamma = 0.5$.

So, for each choice of β_k (among Perry (3), Polak-Ribière (4) and Fletcher-Reeves (5)) we have four methods:

M1: θ_k is computed by (6) and the initial choice of α is

$$\bar{\alpha}(k, d_k, d_{k-1}, \alpha_{k-1}) = \begin{cases} 1, & \text{if } k = 0 \\ \alpha_{k-1}\|d_{k-1}\|_2/\|d_k\|_2, & \text{otherwise;} \end{cases} \tag{11}$$

M2: θ_k is computed by (6) and $\bar{\alpha}(k, d_k, d_{k-1}, \alpha_{k-1}) \equiv 1$;

M3: $\theta_k \equiv 1$ and the initial α is computed as in (11);

M4: $\theta_k \equiv 1$ and $\bar{\alpha}(k, d_k, d_{k-1}, \alpha_{k-1}) \equiv 1$.

In the following tables, we display the performance of the algorithms described above. For each algorithm we state the number of function-gradient evaluations and the functional value achieved at the approximate solution found. For terminating the executions, we used, as in [8], the criterion

$$\|g(x_k)\|_2 \leq 10^{-6} \max\{1, |f(x_k)|\}.$$

All the experiments were run in a SPARCstation Sun Ultra 1, with an UltraSPARC 64 bits processor, 167-MHz clock and 128-MBytes of RAM

memory. SGM and CONMIN codes are in Fortran and were compiled with f77 compiler (SC 1.0 Fortran v1.4). The other algorithms are in the C/C++ language and were compiled with the g++ compiler (GNU project C and C++ compiler v2.7). In all cases we used the optimization compiler option -O4.

Table 3 corresponds to the four alternatives of formula (3). We show, for each problem, the number of function and gradient evaluations and the optimal functional value found. Let f_i be the optimal functional value found by method M_i and f_j the optimal functional value found by M_j. We say that, in a particular problem, the performance of M_i was better than the performance of M_j if $f_i \leq f_j - 10^{-3}$ or if $|f_i - f_j| < 10^{-3}$ and the number of function-gradient evaluations of M_i was less than the number of function-gradient evaluation of M_j. We say that "M_i beat M_j $k1 - k2$" if the performance of M_i was better than the performance of M_j in k_1 problems whereas the performance of M_j was better than the performance of M_i in k_2 problems.

Problem		M1		M2		M3		M4	
		FGE	$f(x)$	FGE	$f(x)$	FGE	$f(x)$	FGE	$f(x)$
1	100	11	0.0000E+00	8	0.0000E+00	11	0.0000E+00	8	0.0000E+00
1	1000	11	0.0000E+00	8	0.0000E+00	11	0.0000E+00	8	0.0000E+00
1	10000	11	0.0000E+00	60	0.0000E+00	11	0.0000E+00	60	0.0000E+00
2	100	63	5.0500E+02	79	5.0500E+02	63	5.0500E+02	83	5.0500E+02
2	500	85	1.2525E+04	124	1.2525E+04	95	1.2525E+04	140	1.2525E+04
2	1000	96	5.0050E+04	140	5.0050E+04	90	5.0050E+04	186	5.0050E+04
3	100	11	8.7540E−22	7	1.4665E−24	11	8.7540E−22	7	7.8758E−25
3	1000	9	5.2302E−20	6	2.6191E−22	9	5.2302E−20	6	5.4108E−20
3	10000	43	0.0000E+00	18	0.0000E+00	43	0.0000E+00	13	0.0000E+00
4	100	94	1.8410E−06	117	1.8410E−06	98	1.8410E−06	83	1.8410E−06
4	1000	84	2.3338E−07	121	2.1479E−07	86	2.4019E−07	79	2.4705E−07
4	10000	88	2.2553E−08	114	2.2104E−08	94	2.2561E−08	81	2.2369E−08
5	100	55	3.0248E−15	41	4.7261E−15	56	9.0436E−16	99	7.7355E−15
5	1000	106	1.4078E+00	140	7.1253E−01	70	4.8690E−15	190	3.9707E−01
5	3000	95	3.9707E−01	54	8.0458E−15	113	3.9707E−01	193	3.9707E−01
6	100	59	1.2882E−10	93	2.1347E−10	120	1.7172E−10	385	2.7329E−10
6	1000	221	9.7949E−11	449	2.2576E−10	306	6.3862E−10	1794	7.1612E−10
6	10000	753	1.5824E−10	2669	2.4077E−10	765	3.1394E−10	7879	1.4202E−10
7	100	54	7.1299E−24	110	3.7756E−25	49	1.7519E−20	118	3.8932E−15
7	1000	58	4.2731E−18	144	5.6455E−25	50	1.1695E−16	113	2.8596E−27
7	10000	61	2.2113E−21	91	1.2007E−16	53	2.1468E−17	134	6.4250E−23
8	100	151	9.0249E−04	166	9.0249E−04	110	9.0249E−04	255	9.0249E−04
8	1000	107	9.6868E−03	110	9.6862E−03	81	9.6862E−03	579	9.6862E−03
8	10000	96	9.9002E−02	79	9.9002E−02	98	9.9002E−02	650	9.9002E−02
9	100	180	2.6633E−15	272	2.5811E−15	188	1.5503E−15	278	2.7837E−15
9	1000	659	9.1087E−15	927	1.1960E−15	681	1.2204E−15	934	1.1820E−15
10	100	29	1.0583E−19	22	6.2419E−21	63	5.6617E−14	291	5.7001E−19
10	1000	83	1.6030E−15	204	9.3560E−19	132	1.4792E−15	107	4.8044E+93
11	100	168	3.8005E−10	691	4.5490E−11	131	1.6455E−09	263	2.8000E−09
11	1000	366	1.9973E−09	190	2.0085E−09	117	5.6132E−09	826	2.4485E−10
12	100	727	1.0000E+00	707	1.0000E+00	694	1.0000E+00	3950	1.0000E+00
12	500	1899	1.0000E+00	3414	1.0000E+00	2081	1.0000E+00	17602	1.0000E+00
13	100	32	1.0909E+02	27	1.0909E+02	39	1.0909E+02	45	1.0909E+02
13	1000	31	1.1082E+03	22	1.1082E+03	34	1.1082E+03	48	1.1082E+03
13	10000	23	1.1099E+04	19	1.1099E+04	36	1.1099E+04	49	1.1099E+04
14	100	85	1.1965E+04	129	1.1965E+04	65	1.1965E+04	361	1.1965E+04
14	1000	43	1.2147E+05	77	1.2147E+05	121	1.2147E+05	236	1.2147E+05
14	10000	41	1.2165E+06	96	1.2165E+06	38	1.2165E+06	525	1.2165E+06
15	100	116	6.6990E−16	76	3.7810E+02	87	3.7810E+02	339	7.8770E+00
15	1000	106	3.1328E−15	229	7.0812E−15	77	3.9306E+03	362	3.9228E+03

Table 1: Performance of Perry.

Problem		M1		M2		M3		M4	
		FGE	$f(x)$	FGE	$f(x)$	FGE	$f(x)$	FGE	$f(x)$
1	100	13	0.0000E+00	11	0.0000E+00	13	0.0000E+00	11	0.0000E+00
1	1000	13	0.0000E+00	11	0.0000E+00	13	0.0000E+00	166	0.0000E+00
1	10000	13	0.0000E+00	218	0.0000E+00	13	0.0000E+00	166	0.0000E+00
2	100	68	5.0500E+02	79	5.0500E+02	68	5.0500E+02	91	5.0500E+02
2	500	108	1.2525E+04	123	1.2525E+04	108	1.2525E+04	163	1.2525E+04
2	1000	121	5.0050E+04	140	5.0050E+04	122	5.0050E+04	185	5.0050E+04
3	100	11	8.7540E−22	8	4.7974E−19	11	8.7540E−22	95	1.3644E−19
3	1000	9	5.2302E−20	8	3.5315E−21	9	5.2302E−20	89	1.3304E−22
3	10000	43	0.0000E+00	45	0.0000E+00	43	0.0000E+00	216	0.0000E+00
4	100	114	2.4054E−06	123	1.8410E−06	111	2.4054E−06	97	1.8410E−06
4	1000	98	2.3339E−07	108	2.2664E−07	98	2.3339E−07	111	2.1427E−07
4	10000	110	2.2265E−08	122	2.1983E−08	107	2.2265E−08	104	2.2680E−08
5	100	51	9.3293E−15	52	1.1363E−14	51	9.3293E−15	116	5.4871E−15
5	1000	117	7.1253E−01	99	7.1253E−01	115	7.1253E−01	231	7.1253E−01
5	3000	110	3.9707E−01	64	3.8591E−15	112	3.9707E−01	193	3.9707E−01
6	100	75	8.1586E−11	109	4.0205E−10	75	8.1586E−11	361	1.8057E−10
6	1000	271	2.7962E−10	522	3.9881E−10	268	5.1372E−10	1714	8.0629E−11
6	10000	1192	8.5903E−11	3159	5.7281E−10	1064	1.8114E−10	8308	7.1085E−10
7	100	59	2.7563E−16	117	2.0270E−16	59	2.7563E−16	121	1.1097E−22
7	1000	82	3.1513E−15	105	4.5191E−17	79	5.5793E−18	108	1.3089E−15
7	10000	52	8.7075E−17	89	1.6665E−25	52	8.7057E−17	129	4.2441E−18
8	100	183	9.0249E−04	160	9.0249E−04	193	9.0249E−04	359	9.0249E−04
8	1000	155	9.6862E−03	157	9.6862E−03	150	9.6862E−03	626	9.6862E−03
8	10000	100	9.9002E−02	158	9.9002E−02	104	9.9002E−02	840	9.9002E−02
9	100	224	5.5834E−15	261	1.9083E−15	223	1.0253E−14	242	3.7251E−15
9	1000	823	5.3509E−15	936	1.1286E−15	976	9.5998E−15	824	1.6607E−15
10	100	29	1.0512E−19	47	4.4518E−21	29	1.0511E−19	969	1.4175E−25
10	1000	43	2.1838E−23	485	4.9804E−22	43	2.3554E−23	107	1.2014E+76
11	100	329	5.3346E−10	436	4.1324E−09	394	8.2489E−11	310	4.2044E−10
11	1000	1089	4.7035E−09	346	2.4502E−09	607	5.1556E−10	320	2.3313E−09
12	100	840	1.0000E+00	1087	1.0000E+00	914	1.0000E+00	5145	1.0000E+00
12	500	3004	1.0000E+00	4031	1.0000E+00	3062	1.0000E+00	24685	1.0000E+00
13	100	39	1.0909E+02	24	1.0909E+02	39	1.0909E+02	34	1.0909E+02
13	1000	36	1.1082E+03	22	1.1082E+03	36	1.1082E+03	47	1.1082E+03
13	10000	30	1.1099E+04	28	1.1099E+04	30	1.1099E+04	32	1.1099E+04
14	100	76	1.1965E+04	171	1.1965E+04	76	1.1965E+04	261	1.1965E+04
14	1000	62	1.2147E+05	83	1.2147E+05	62	1.2147E+05	376	1.2147E+05
14	10000	62	1.2165E+06	85	1.2165E+06	62	1.2165E+06	296	1.2165E+06
15	100	65	3.8597E+02	80	3.7810E+02	65	3.8597E+02	366	3.9379E+00
15	1000	77	3.9267E+03	70	3.9267E+03	77	3.9267E+03	334	3.9228E+03

Table 2: Performance of Polak-Ribière.

Problem		M1		M2		M3		M4	
		FGE	$f(x)$	FGE	$f(x)$	FGE	$f(x)$	FGE	$f(x)$
1	100	13	1.4211E−14	12	0.0000E+00	13	1.4211E−14	13	2.8422E−13
1	1000	65	0.0000E+00	64	0.0000E+00	65	0.0000E+00	65	0.0000E+00
1	10000	117	0.0000E+00	167	0.0000E+00	117	0.0000E+00	116	0.0000E+00
2	100	87	5.0500E+02	129	5.0500E+02	87	5.0500E+02	104	5.0500E+02
2	500	135	1.2525E+04	203	1.2525E+04	135	1.2525E+04	169	1.2525E+04
2	1000	150	5.0050E+04	235	5.0050E+04	150	5.0050E+04	326	5.0050E+04
3	100	11	8.7540E−22	9	1.6231E−19	11	8.7540E−22	99	1.9409E−17
3	1000	9	5.2302E−20	8	3.1504E−20	9	5.2302E−20	89	1.3690E−22
3	10000	43	0.0000E+00	34	0.0000E+00	43	0.0000E+00	317	0.0000E+00
4	100	9237	2.0511E−06	514	1.8410E−06	9064	2.0431E−06	706	1.8410E−06
4	1000	546	2.3349E−07	478	2.2725E−07	490	2.3349E−07	1066	2.2725E−07
4	10000	260	2.1433E−08	474	1.4611E−08	225	2.1434E−08	1026	2.1369E−08
5	100	94	6.1146E−15	88	6.5000E−15	94	6.1146E−15	112	7.3569E−15
5	1000	355	3.9707E−01	483	3.9707E−01	349	3.9707E−01	2482	3.9707E−01
5	3000	361	1.4132E−14	332	1.4085E−14	368	1.3015E−14	243	3.9707E−01
6	100	72	7.3343E−11	83	3.9906E−10	72	7.3343E−11	335	2.8331E−10
6	1000	181	1.5526E−10	611	7.1882E−11	181	1.5532E−10	1764	1.6422E−10
6	10000	712	7.5201E−11	4835	5.7198E−11	754	6.4120E−11	12168	4.3333E−11
7	100	226	3.4249E−14	308	4.7522E−13	210	2.7708E−13	2549	6.3509E−14
7	1000	150	1.6903E−16	2203	8.5476E−14	166	4.9431E−13	4796	2.2730E−25
7	10000	175	1.2559E−14	3254	3.0481E−13	169	2.1257E−14	2305	6.8998E−13
8	100	1651	9.0249E−04	413	9.0249E−04	1636	9.0249E−04	1480	9.0249E−04
8	1000	738	9.6862E−03	246	9.6862E−03	751	9.6862E−03	582	9.6862E−03
8	10000	524	9.9002E−02	3027	9.9002E−02	478	9.9002E−02	1913	9.9002E−02
9	100	531	2.5973E−15	14001	1.2942E−03	531	2.5971E−15	14492	2.5580E−01
9	1000	5050	5.8067E−16	14001	7.2011E+00	4410	4.6703E−16	14217	1.8909E+01
10	100	29	1.0511E−19	55	5.3174E−19	29	1.0512E−19	887	5.0363E−21
10	1000	43	9.7570E−24	503	1.6185E−22	43	1.2717E−23	107	7.1871E+93
11	100	299	4.0649E−10	497	9.4581E−11	423	5.3053E−10	372	5.3573E−10
11	1000	522	1.3232E−09	4286	5.7282E−10	562	3.0749E−09	1550	5.0101E−10
12	100	9584	1.0065E+02	20633	1.0000E+00	9584	1.0065E+02	135737	1.0000E+00
12	500	9991	2.9167E+02	60913	9.0475E+01	9991	2.9167E+02	135008	2.0857E+02
13	100	50	1.0909E+02	42	1.0909E+02	50	1.0909E+02	50	1.0909E+02
13	1000	40	1.1082E+03	52	1.1082E+03	40	1.1082E+03	28	1.1082E+03
13	10000	33	1.1099E+04	19	1.1099E+04	33	1.1099E+04	64	1.1099E+04
14	100	65	1.1965E+04	5430	1.1965E+04	65	1.1965E+04	3115	1.1965E+04
14	1000	22	1.2147E+05	2676	1.2147E+05	22	1.2147E+05	552	1.2147E+05
14	10000	31	1.2165E+06	419	1.2165E+06	31	1.2165E+06	352	1.2165E+06
15	100	5551	3.9379E+00	794	3.7810E+02	9042	3.9379E+00	379	7.8770E+00
15	1000	207	7.8770E+00	709	3.9391E+00	86	7.8770E+00	2677	7.8770E+00

Table 3: Performance of Fletcher-Reeves.

The conclusions of the experiments displayed above are:

For Perry's formula (Table 1)

- M1 beat M3 $21 - 12$;
- M1 beat M2 $26 - 14$;
- M1 beat M4 $31 - 9$;
- M3 beat M2 $23 - 17$;
- M3 beat M4 $31 - 9$;
- M2 beat M4 $27 - 8$.

For Polak-Ribière formula (Table 2)

- M3 beat M1 $9 - 8$;
- M3 beat M2 $26 - 14$;
- M3 beat M4 $32 - 8$;
- M1 beat M2 $27 - 13$;
- M1 beat M4 $33 - 7$;
- M2 beat M4 $31 - 8$.

For Fletcher-Reeves formula (Table 3)

- M3 beat M1 $10 - 7$;
- M3 beat M2 $24 - 16$;
- M3 beat M4 $29 - 8$;
- M1 beat M2 $24 - 16$;
- M1 beat M4 $30 - 7$;
- M2 beat M4 $28 - 12$.

Now, comparing the best alternatives of each conjugate-gradient formula, we conclude that:

- Perry (M1) beat Polak-Ribière (M3) $30 - 6$.
- Perry (M1) beat Fletcher-Reeves (M3) $29 - 7$.
- Polak-Ribière (M3) beat Fletcher-Reeves (M3) $23 - 11$.

4 Comparisons with CONMIN and SGM

The experiments in Section 3 seem to indicate that the best scaled conjugate-gradient formula is Perry's (3) with the spectral choice (6) of θ_k and the initial choice (11) of the steplength. So, we compared this method against CONMIN [9] and the spectral gradient method of Raydan. We used the original (Fortran) codes of SGM and CONMIN. SGM was used with the parameters recommended by Raydan [8].

The results are given in Table 4. Using the same criteria described above, we see that Perry's M1 beat CONMIN $20-19$. The comparison against Raydan's SGM must take into account that this method evaluates, sometimes, the function at points where the gradient is not evaluated. Counting only gradient evaluations, Perry's M1 was better than SGM 20 times and SGM was better than Perry's M1 in the same number of problems. Counting also function evaluations, Perry's M1 beat SGM $21-19$.

		SGM				CONMIN			Perry(M1)		
Problem		IT	FE	GE	$f(x)$	IT	FGE	$f(x)$	IT	FGE	$f(x)$
1	100	8	8	9	0.0000E+00	15	38	1.6058E−11	5	11	0.0000e+00
1	1000	8	8	9	-1.1369E−13	15	38	1.5154E−10	5	11	0.0000e+00
1	10000	8	8	9	1.8190E−12	15	38	1.4734E−09	5	11	0.0000e+00
2	100	52	57	53	5.0500E+02	40	81	5.0500E+02	45	63	5.0500e+02
2	500	74	80	75	1.2525E+04	63	127	1.2525E+04	67	85	1.2525e+04
2	1000	82	91	83	5.0050E+04	71	145	5.0050E+04	75	96	5.0050e+04
3	100	3	3	4	1.8795E−23	3	7	1.4066E−07	5	11	8.7540e−22
3	1000	4	4	5	1.3346E−23	15	38	8.1381E−18	4	9	5.2302e−20
3	10000	53	56	54	0.0000E+00	14	38	4.6837E−20	5	43	0.0000e+00
4	100	76	80	77	2.4054E−06	51	108	1.8410E−06	62	94	1.8410e−06
4	1000	91	104	92	2.2558E−07	53	112	2.2664E−07	56	84	2.3338e−07
4	10000	89	99	90	2.1659E−08	59	126	2.2674E−08	60	88	2.2553e−08
5	100	34	34	35	1.1369E−14	33	67	3.0081E−14	29	55	3.0248e−15
5	1000	40	40	41	4.3612E−15	81	169	3.9707E−01	72	106	1.4078e+00
5	3000	44	45	45	2.0021E−14	35	71	1.8684E−14	62	95	3.9707e−01
6	100	106	111	107	5.5889E−10	49	99	6.7141E−10	42	59	1.2882e−10
6	1000	296	364	297	1.4567E−09	158	320	1.3921E−10	169	221	9.7949e−11
6	10000	1351	1751	1352	1.0295E−09	464	937	5.3219E−11	565	753	1.5824e−10
7	100	69	91	70	3.4615E−17	19	47	2.9286E−12	29	54	7.1299e−24
7	1000	93	118	94	1.4427E−20	30	73	1.4110E−15	28	58	4.2731e−18
7	10000	70	92	71	1.9663E−17	28	69	1.4479E−14	28	61	2.2113e−21
8	100	48	49	49	9.0249E−04	27	65	9.0249E−04	73	151	9.0249e−04
8	1000	57	57	58	9.6862E−03	25	55	9.6862E−03	47	107	9.6862e−03
8	10000	70	70	71	9.9001E−02	1	3	1.1114E+23	37	96	9.9002e−02
9	100	167	191	168	2.4820E−16	80	161	4.9988E−15	141	180	2.6633e−15
9	1000	878	1152	879	1.6416E−14	306	613	6.1288E−16	520	659	9.1087e−15
10	100	38	38	39	3.1061E−29	13	29	2.8874E−18	11	29	1.0583e−19
10	1000	66	68	67	1.6362E−25	27	62	1.4308E−20	20	83	1.6030e−15
11	100	740	988	741	1.1325E−09	47	95	1.0019E−09	98	168	3.8005e−10
11	1000	1345	1851	1346	7.9283E−09	43	87	2.0417E−09	196	366	1.9973e−09
12	100	1429	1886	1430	1.0000E+00	254	516	1.0000E+00	536	727	1.0000e+00
12	500	4452	5896	4453	1.0000E+00	1082	2180	1.0000E+00	1522	1899	1.0000e+00
13	100	26	26	27	1.0909E+02	13	27	1.0909E+02	16	32	1.0909e+02
13	1000	23	23	24	1.1082E+03	11	23	1.1082E+03	16	31	1.1082e+03
13	10000	21	21	22	1.1099E+04	9	19	1.1099E+04	11	23	1.1099e+04
14	100	438	587	439	1.1965E+04	13	27	1.1965E+04	48	85	1.1965e+04
14	1000	288	391	289	1.2147E+05	12	25	1.2147E+05	22	43	1.2147e+05
14	10000	119	154	120	1.2165E+06	11	23	1.2165E+06	21	41	1.2165e+06
15	100	81	84	82	3.8597E+02	25	53	3.7810E+02	64	116	6.6990e−16
15	1000	80	87	81	7.8770E+00	34	69	3.9267E+03	59	106	3.1328e−15

Table 4: Performance of SGM, CONMIN and Perry-M1.

5 A parameter estimation problem in Optics

In recent works, the spectral gradient method has been successfully used for a hard inverse problem that consists on the estimation of optical parameters of thin films using transmission data. See [1, 2].

The transmission T of a thin absorbing film on a transparent substrate

is given by

$$T = \frac{Ax}{B - Cx + Dx^2}, \tag{12}$$

where

$$A = 16s(r^2 + k^2), \tag{13}$$

$$B = [(r+1)^2 + k^2][(r+1)(r+s^2) + k^2], \tag{14}$$

$$C = [(r^2 - 1 + k^2)(r^2 - s^2 + k^2) - 2k^2(s^2+1)]2\cos\varphi$$
$$-k[2(r^2 - s^2 + k^2) + (s^2+1)(r^2 - 1 + k^2)]2\sin\varphi, \tag{15}$$

$$D = [(r-1)^2 + k^2][(r-1)(r-s^2) + k^2], \tag{16}$$

$$\varphi = 4\pi r d/\lambda, \quad x = exp(-\alpha d), \quad \alpha = 4\pi k/\lambda. \tag{17}$$

In formulae (13)–(17) the following notation is used:

(a) λ is the wavelength;

(b) $s \equiv s(\lambda)$ is the refractive index of the substrate;

(c) $r \equiv r(\lambda)$ is the refractive index of the film;

(d) $k \equiv k(\lambda)$ is the attenuation coefficient of the film;

(e) d is the thickness of the film.

We assume that $s(\lambda)$ and d are known, a set of experimental data $(\lambda_i, T^{obs}(\lambda_i)), i = 1, \ldots, N$, where $(\lambda_{min} \leq \lambda_i < \lambda_{i+1} \leq \lambda_{max}$ for all $i = 1, \ldots, N-1)$, is given and we wish to estimate $r(\lambda)$ and $k(\lambda)$. At a first glance, this problem is underdetermined. In fact, given λ, the following equation must hold:

$$T(\lambda, s(\lambda), d, r(\lambda), k(\lambda)) = T^{obs}(\lambda). \tag{18}$$

Equation (18) has two unknowns $r(\lambda)$ and $k(\lambda)$ and, therefore, in general, its set of solutions is a curve in the two-dimensional $(r(\lambda), k(\lambda))$ space. However, physical constraints reduce drastically the range of variability of the unknowns $r(\lambda), k(\lambda)$. For a class of films studied in [1], these physical constraints are:

$$r(\lambda_{max}) \geq 1, \quad k(\lambda_{max}) \geq 0, \tag{19}$$

$$r'(\lambda_{max}) \leq 0, \quad k'(\lambda_{max}) \leq 0, \tag{20}$$

$$r''(\lambda) \geq 0 \quad \text{for all } \lambda \in [\lambda_{min}, \lambda_{max}], \tag{21}$$

$$k''(\lambda) \geq 0 \quad \text{for all } \lambda \in [\lambda_{min}, \lambda_{max}]. \tag{22}$$

So, the continous least squares solution of the estimation problem is the solution $(r(\lambda), k(\lambda))$ of

$$\text{Minimize} \int_{\lambda_{min}}^{\lambda_{max}} |T(\lambda, s(\lambda), d, r(\lambda), k(\lambda)) - T^{obs}(\lambda)|^2 d\lambda \tag{23}$$

subject to the constraints (19)–(22).

In [2] the authors defined

$$r(\lambda_{max}) = 1 + u^2, \quad k(\lambda_{max}) = v^2, \tag{24}$$

$$r'(\lambda_{max}) = -u_1^2, \quad k'(\lambda_{max}) = -v_1^2, \tag{25}$$

$$r''(\lambda) = w(\lambda)^2 \quad \text{for all } \lambda \in [\lambda_{min}, \lambda_{max}], \tag{26}$$

$$k''(\lambda) = z(\lambda)^2 \quad \text{for all } \lambda \in [\lambda_{min}, \lambda_{max}]. \tag{27}$$

In the real-life situation, in which data are given for a set of N equally spaced points on the interval $[\lambda_{min}, \lambda_{max}]$, we define

$$h = (\lambda_{max} - \lambda_{min})/(N-1)$$

and

$$\lambda_i = \lambda_{min} + (i-1)h, \quad i = 1, \ldots, N.$$

The observed value of the transmission at λ_i will be called T_i^{obs}. Moreover, we use the notation r_i, k_i, w_i, z_i in the obvious way:

$$r_i = r(\lambda_i), \quad k_i = k(\lambda_i),$$

$$w_i = w(\lambda_{i+1}), \quad z_i = z(\lambda_{i+1}),$$

for $i = 1, \ldots, N$. Discretization of (24-27) gives:

$$r_N = 1 + u^2, \quad v_N = v^2, \tag{28}$$

$$r_{N-1} = r_N + u_1^2 h, \quad k_{N-1} = k_N + v_1^2 h, \tag{29}$$

$$r_i = w_i^2 h^2 + 2r_{i+1} - r_{i+2}, \quad i = 1, \ldots, N-2, \tag{30}$$

$$k_i = z_i^2 h^2 + 2k_{i+1} - k_{i+2} \quad i = 1, \ldots, N-2. \tag{31}$$

Finally, the objective function of (23) is approximated by a sum of squares, giving the optimization problem

$$\text{Minimize} \sum_{i=1}^{N} [T(\lambda_i, s(\lambda_i), d, r_i, k_i) - T_i^{obs}]^2. \tag{32}$$

Since r_i and k_i depend on u, u_1, v, v_1, w, z through (28–31), problem (32) takes the form

$$\text{Minimize } f(u, u_1, v, v_1, w_1, \ldots, w_{N-2}, z_1, \ldots, z_{N-2}). \tag{33}$$

In the experiments we used the films considered in [2], with 100 data points:

Film 1: Simulation of an amorphous germanium thin film deposited on a glass substrate with $d = 118$nm. $\lambda_{min} = 600$nm, $\lambda_{max} = 2000$nm;

Film 2: Identical to Film 1 with $d = 782$nm. $\lambda_{min} = 1000$nm, $\lambda_{max} = 2000$nm;

Film 3: Simulation of an amorphous germanium thin film deposited on a crystalline silicon substrate with $d = 147$nm. $\lambda_{min} = 1250$nm, $\lambda_{max} = 2500$nm;

Film 4: Identical to Film 3 with $d = 640$nm. $\lambda_{min} = 640$nm, $\lambda_{max} = 1250$nm;

Film 5: Simulation of an hydrogenated amorphous silicon thin film deposited onto glass with $d = 624$nm. $\lambda_{min} = 600$nm, $\lambda_{max} = 1600$nm.

As initial estimate of $k(\lambda)$ we used a piecewise linear function whose values are 0.1 at the smallest wavelength of the spectrum, 0.01 at $\lambda_{min} + 0.2(\lambda_{max} - \lambda_{min})$, and 10^{-10} and λ_{max}. The initial estimate of $r(\lambda)$ was a linear function varying between 5 (at λ_{min}) and 3 (at λ_{max}). The physically acceptable results of the estimation procedure were obtained in [2] using 30000 iterations of Raydan's spectral gradient method. Here we used as stopping criterion for SCG (Perry-M1) the inequality $f(x^k) < f_{Raydan}$ where f_{Raydan} is the minimum value reached by SGM. In Table 5 we give the results. IT means number of iterations, FE is functional evaluations and FGE function-gradient evaluations. Observe that SCG arrives to the solution of SGM using between one third and one half the computer time used by the spectral gradient method.

	SGM				SCG			
Problem	IT	FGE	Time	f_{Raydan}	IT	FGE	Time	$f(x^k)$
1	30000	35825	45.0"	6.929605E−07	3605	6184	7.7"	6.926210E−07
2	30000	35568	45.8"	2.203053E−07	6798	11092	14.1"	2.201913E−07
3	30000	38113	47.9"	6.224862E−06	7344	13471	17.5"	6.224860E−06
4	30000	35687	44.6"	1.365270E−06	10356	17938	22.3"	1.365184E−06
5	30000	36290	46.3"	2.120976E−07	7611	13205	16.5"	2.066100E−07

Table 5: Optics problems.

6 Final remarks

In the classical paper [9], Perry's basic idea was modified in order to overcome the lack of positive definiteness of the matrix that, implicitly, defines the search direction. As a result, the algorithmic framework of CONMIN was obtained. In this paper we followed a different direction, motivated by the necessity of preserving the nice geometrical properties of Perry's direction. On one hand, we observed that scaling the gradient by means of the spectral parameter of [8] is worthwhile and, on the other hand, we detected that the initial choice of the steplength is a crucial parameter that influentiates the practical behavior of the method. In this way, Perry's algorithm clearly outperforms Polak-Ribière and Fletcher-Reeves and is competitive with CONMIN and Raydan's [8] method.

Moreover, as observed by Raydan, the spectral gradient method certainly needs preconditioning in ill-conditioned problems in a more dramatic way that conjugate-gradient methods do. This is the reason why, in the hard inverse problem studied in Section 5, SGM is outperformed by the M1 version of Perry's method.

References

[1] Chambouleyron I. , Martínez J. M., Moretti A. C., Mulato M. (1997) The retrieval of the optical constants and the thickness of thin films from transmission spectra. *Applied Optics* 36: 8238-8247.

[2] Birgin E. G., Chambouleyron I., Martínez J. M. (1998) Estimation of optical constants of thin films using unconstrained optimization, Technical Report 25–98, Instituto de Matemática, Universidade Estadual de Campinas, 13081-970 Campinas SP, Brazil.

[3] Dennis Jr. J. E., Schnabel R. B. (1983) *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice-Hall, Englewood Cliffs.

[4] Fletcher R. (1987) *Practical Methods of Optimization* (2nd edition), John Wiley and Sons, Chichester, New York, Brisbane, Toronto and Singapore.

[5] Glunt W., Hayden T. L., Raydan M. (1993) Molecular conformations from distance matrices. *Journal of Computational Chemistry* 14: 114-120.

[6] Luengo F., Raydan M., Glunt W., Hayden T. L. (1996) Preconditioned spectral gradient method for unconstrained optimization problems, Technical Report 96-08, Escuela de Computación, Facultad de Ciencias, Universidad Central de Venezuela, 47002 Caracas 1041-A, Venezuela.

[7] Perry A. (1978) A modified conjugate gradient algorithm. *Operations Research* 26: 1073–1078.

[8] Raydan M. (1997) The Barzilain and Borwein gradient method for the large unconstrained minimization problem. *SIAM Journal on Optimization* 7: 26-33.

[9] Shanno D. F. (1978) Conjugate gradient methods with inexact searches. *Mathematics of Operations Research* 3: 244–256.

[10] Shanno D. F., Phua K. H. (1976) Minimization of unconstrained multivariate functions. *ACM Transactions on Mathematical Software* 2: 87–94.