A Note on Spectral Properties of Some Gradient Methods

Daniela di Serafino[1,b)], Valeria Ruggiero[2,c)], Gerardo Toraldo[3,a)] and Luca Zanni[4,d)]

[1]*Department of Mathematics and Physics, Second University of Naples, viale A. Lincoln 5, I-81100 Caserta, Italy*
[2]*Department of Mathematics and Computer Science, University of Ferrara, via Saragat 1, I-44122 Ferrara, Italy*
[3]*Department of Mathematics, University of Naples Federico II, Via Cintia 21, I-80126, Naples, Italy*
[4]*Department of Physics, Computer Science and Mathematics, University of Modena and Reggio Emilia, via Campi 213/B, I-41125, Modena, Italy*

[a)]Corresponding author: toraldo@unina.it
[b)]daniela.diserafino@unina2.it
[c)]valeria.ruggiero@unife.it
[d)]luca.zanni@unimore.it

Abstract. Starting from the work by Barzilai and Borwein, the interest for gradient methods has gained a great amount of attention, and efficient low-cost schemes are available nowadays. The acceleration strategies used by these methods are based on the definition of effective steplength updating rules, which capture spectral properties of the Hessian of the objective function. The methods arising from this idea represent effective computational tools, extremely appealing for a variety of large-scale optimization problems arising in applications. In this work we discuss the spectral properties of some recently proposed gradient methods with the aim of providing insight into their computational effectiveness. Numerical experiments supporting and illustrating the theoretical analysis are provided.

INTRODUCTION

Several strategies for accelerating gradient methods have been devised in the last years, stimulated by the seminal work by Barzilai and Borwein [1]. These strategies share the idea of defining steplengths that capture spectral properties of the Hessian of the objective function; based on them, new first-order methods for continuous nonlinear optimization have been designed, which showed to be effective in some practical contexts [2, 3, 4, 5, 6]. However, the convergence results available do not explain the great improvement with respect to the classical Cauchy Steepest Descent (SD) method, and we still do not have a deep understanding of the behaviour of the new methods.

In this work we discuss the spectral properties of some recently proposed steplength rules, with the aim of providing insight into their computational effectiveness. To this purpose, we consider a very simple unconstrained quadratic programming problem, suitable for analyzing the role of the eigenvalues of the Hessian in the behaviour of gradient methods:

$$\min f(x) = \frac{1}{2}x^T Ax - b^T x, \tag{1}$$

where $A \in \mathbb{R}^{n\times n}$ is symmetric positive definite and $b \in \mathbb{R}^n$. The generic gradient method for (1) is defined by the iteration

$$x_{k+1} = x_k - \alpha_k g_k, \tag{2}$$

where $g_k = \nabla f(x_k) = Ax_k - b$, and the steplength $\alpha_k > 0$ is chosen through some predefined rule. For instance, the classical SD and Minimum Residual (MR) methods take the following steplengths, which guarantee monotonicity of the sequences $\{f(x_k)\}$ and $\{\|\nabla f(x_k)\|\}$, respectively:

$$\alpha_k^{\mathrm{SD}} = \operatorname*{argmin}_{\alpha>0} f(x_k - \alpha g_k) = \frac{g_k^T g_k}{g_k^T A g_k}, \qquad \alpha_k^{\mathrm{MR}} = \operatorname*{argmin}_{\alpha>0} \|\nabla f(x_k - \alpha g_k)\| = \frac{g_k^T A g_k}{g_k^T A^2 g_k}. \tag{3}$$

Let $\lambda_1 \geq \lambda_2 \geq \ldots, \lambda_{n-1} \geq \lambda_n$ be the eigenvalues of A, with associated orthonormal eigenvectors $d_1, d_2, \ldots, d_n$. Without loss of generality, henceforth we make the following assumptions:

A1. $\lambda_1 > \lambda_2$ and $\lambda_{n-1} > \lambda_n > 0$;
A2. at the starting point x_0, $\nabla f(x_0) = \sum_{i=1}^{n} \mu_i^0 d_i$, with $\mu_1^0 \neq 0$ and $\mu_n^0 \neq 0$.

Since $g_{k+1} = g_k - \alpha_k A g_k = \prod_{j=0}^{k}(I - \alpha_j A) g_0$, we have

$$g_{k+1} = \sum_{i=1}^{n} \mu_i^{k+1} d_i, \quad \mu_i^{k+1} = \mu_i^0 \prod_{j=0}^{k}(1 - \alpha_j \lambda_i) = \mu_i^k (1 - \alpha_k \lambda_i). \tag{4}$$

STEPLENGTHS AND HESSIAN SPECTRUM

Starting from recurrence (4), the following properties can be deduced:

- if at the k-th iteration $\mu_i^k = 0$ for some i, then $\mu_i^h = 0$ for $h > k$;
- if at the k-th iteration $\alpha_k = 1/\lambda_i$, then $\mu_i^{k+1} = 0$;
- the SD and MR methods have finite termination if and only if at some iteration the gradient is an eigenvector of A;
- $\left|\mu_i^{k+1}\right| < \left|\mu_i^k\right|$ if and only if $\alpha_k < 2/\lambda_i$;
- for $\alpha_k \approx 1/\lambda_j$, $\left|\mu_i^{k+1}\right| > \left|\mu_i^k\right|$ when $i < j$ and $\lambda_i > 2\lambda_j$;
- $\alpha_k^{\mathrm{SD}} = \sum_{i=1}^{n}(\mu_i^k)^2 / \left(\sum_{i=1}^{n}(\mu_i^k)^2 \lambda_i\right)$, $\alpha_k^{\mathrm{MR}} = \sum_{i=1}^{n}(\mu_i^k)^2 \lambda_i / \left(\sum_{i=1}^{n}(\mu_i^k)^2 \lambda_i^2\right)$.

Thus, small steplengths α_k (say close to $1/\lambda_1$) tend to decrease a large number of eigencomponents, with negligible reduction of those corresponding to small eigenvalues. The latter can be significantly reduced by using large values of α_k, but this may end up increasing the eigencomponents corresponding to the dominating eigenvalues, as well as fostering non-monotonic behaviour. Therefore, some balance between large and small steplengths seems to be a key issue in devising effective gradient methods and this basic idea has given rise to novel steplength selection rules, some of which will be described in the sequel.

The spectral properties of the SD method have been deeply investigated [7, 8, 9, 10]. An interesting theoretical result concerning the asymptotic behaviour of this method is reported next [8].

Theorem 1 *Let $\{x_k\}$ be a sequence generated by the SD method. Then*

$$\lim_{k\to\infty} \frac{(\mu_1^k)^2}{\sum_{j=1}^{n}(\mu_j^k)^2} = \begin{cases} \dfrac{c^2}{1+c^2} & \text{if } k \text{ odd}, \\ \dfrac{1}{1+c^2} & \text{if } k \text{ even}, \end{cases} \quad \lim_{k\to\infty} \frac{(\mu_n^k)^2}{\sum_{j=1}^{n}(\mu_j^k)^2} = \begin{cases} \dfrac{1}{1+c^2} & \text{if } k \text{ odd}, \\ \dfrac{c^2}{1+c^2} & \text{if } k \text{ even}, \end{cases} \quad \lim_{k\to\infty} \frac{(\mu_i^k)^2}{\sum_{j=1}^{n}(\mu_j^k)^2} = 0 \quad (1 < i < n)$$

where $c = \lim_{k\to\infty} \mu_n^{2k}/\mu_1^{2k} = -\lim_{k\to\infty} \mu_1^{2k+1}/\mu_n^{2k+1}$.

The main consequence of Theorem 1 is that the SD method eventually performs its search in the 2D space spanned by d_1 and d_n, thus showing the well-known zigzagging behaviour. This is in contrast with the possibility for the sequence $\{1/\alpha_k\}$ to travel in the spectrum of the Hessian, which, according to the previous observations, seems to be a desirable feature for gradient methods. Furthermore, for the Cauchy choice of the steplength it is well known that the method has Q-linear rate of convergence which depends on $\rho = (\lambda_1 - \lambda_n)/(\lambda_1 + \lambda_n)$ [7].

The Barzilai-Borwein (BB) steplength rules are given by:

$$\alpha_k^{\mathrm{BB1}} = \frac{\|s_{k-1}\|^2}{s_{k-1}^T y_{k-1}}, \qquad \alpha_k^{\mathrm{BB2}} = \frac{s_{k-1}^T y_{k-1}}{\|y_{k-1}\|^2},$$

where $s_{k-1} = x_k - x_{k-1}$ and $y_{k-1} = g_k - g_{k-1}$; they were obtained by including some second order information through a secant condition, and can be regarded as quasi-Newton methods with the Hessian approximated by $\frac{1}{\alpha_k} I$.
An interesting property of these rules is that

$$\frac{1}{\lambda_1} \leq \alpha_k^{\mathrm{BB2}} = \frac{{g_{k-1}}^T A g_{k-1}}{{g_{k-1}}^T A^2 g_{k-1}} \leq \alpha_k^{\mathrm{BB1}} = \frac{{g_{k-1}}^T g_{k-1}}{{g_{k-1}}^T A g_{k-1}} \leq \frac{1}{\lambda_n};$$

furthermore, with these rules, both the sequences $\{f(x_k)\}$ and $\{\|\nabla f(x_k)\|\}$ are non-monotonic. For strictly convex quadratic problems the BB methods have R-linear convergence, which does not explain why they are in practice much faster than the SD method. An explanation of this behaviour is the ability of generating sequences $\{1/\alpha_k\}$ sweeping the spectrum of A [11].

Starting from [1], many other gradient methods have been proposed. Several methods, either based on the alternation of Cauchy and BB steplengths or the cyclic use of them (see, e.g., [12, 13, 14]), fit into the framework of Gradient Methods with Retards (GMR) [15]. The convergence rate of these methods is R-linear, but their practical convergence behaviour is superior than the SD one. The approaches based on a prefixed alternation of steplength rules seem to be overcome by the selection rules ABB and $\text{ABB}_{\min}$, proposed in [16] and [17], which use an adaptive switching criterion for alternating the BB1 and BB2 steplengths:

$$\alpha_k^{\text{ABB}} = \begin{cases} \alpha_k^{\text{BB2}} & \text{if } \dfrac{\alpha_k^{\text{BB2}}}{\alpha_k^{\text{BB1}}} < \tau, \\ \alpha_k^{\text{BB1}} & \text{otherwise,} \end{cases} \qquad \alpha_k^{\text{ABB}_{\min}} = \begin{cases} \min\left\{\alpha_j^{\text{BB2}} \,:\, j = \max\{1, k-m\}, .., k\right\}, & \text{if } \dfrac{\alpha_k^{\text{BB2}}}{\alpha_k^{\text{BB1}}} < \tau, \\ \alpha_k^{\text{BB1}}, & \text{otherwise,} \end{cases}$$

where m is a nonnegative integer and $\tau \in (0, 1)$. Following the original Adaptive Barzilai–Borwein (ABB) in [16], the $\text{ABB}_{\min}$ strategy aims at generating a sequence of small steplengths with the BB2 rule so that next value computed by the BB1 rule becomes a suitable approximation of the inverse of some small eigenvalue. The switching criterion is based on the value $\alpha_k^{\text{BB2}} / \alpha_k^{\text{BB1}} = cos^2\theta_{k-1}$, where θ_{k-1} is the angle between g_{k-1} and Ag_{k-1}, and allows to select α_k^{BB1} when g_{k-1} is a sufficiently good approximation of an eigenvector of A [17].

A different approach is behind some recently proposed gradient methods, which alternate SD steplengths with a sequence of constant steplenghts computed by some specific rule that exploits previous SD steplenghts, with the aim of escaping from the two dimensional space in which the SD method tends to eventually reduce its search. The SDA and SDC methods [9, 18] compute their constant steplengths by exploiting the formulas

$$\alpha_k^{\text{A}} = \left(\frac{1}{\alpha_{k-1}^{\text{SD}}} + \frac{1}{\alpha_k^{\text{SD}}}\right)^{-1}, \qquad \alpha_k^{\text{Y}} = 2\left(\sqrt{\left(\frac{1}{\alpha_{k-1}^{\text{SD}}} - \frac{1}{\alpha_k^{\text{SD}}}\right)^2 + 4\frac{\|g_k\|^2}{\left(\alpha_{k-1}^{\text{SD}}\|g_{k-1}\|\right)^2}} + \frac{1}{\alpha_{k-1}^{\text{SD}}} + \frac{1}{\alpha_k^{\text{SD}}}\right)^{-1}.$$

We note that the steplength α_k^{Y}, proposed by Yuan [19] and used in a different algorithmic framework, was determined by imposing finite termination for two-dimensional convex quadratic problems. In [9, 18] the authors prove that the steplengths α_k^{A} and α_k^{Y} are related and share similar asymptotic properties, shown by the following theorem.

Theorem 2 *Let $\left\{\alpha_k^{\text{SD}}\right\}$ be a sequence generated by the SD method. Then the sequences $\left\{\alpha_k^{\text{A}}\right\}$ and $\left\{\alpha_k^{\text{Y}}\right\}$ satisfy*

$$\lim_k \alpha_k^{\text{A}} = \frac{1}{\lambda_1 + \lambda_n}, \qquad \lim_k \alpha_k^{\text{Y}} = \frac{1}{\lambda_1}.$$

The steplengths of the SDA and the SDC methods, α_k^{SDA} and α_k^{SDC}, are defined by the following rule:

$$\alpha_k = \begin{cases} \alpha_k^{\text{SD}} & \text{if } \bmod(k, h+m) < h, \\ \hat{\alpha}_s & \text{otherwise, with } s = \max\{i \le k \,:\, \bmod(i, h+m) = h\}, \end{cases} \tag{5}$$

where $\hat{\alpha}_s = \alpha_s^{\text{A}}$ for SDA and $\hat{\alpha}_s = \alpha_s^{\text{Y}}$ for SDC, and h and m are nonnegative integers with $h \ge 2$. In SDC, the use of a finite sequence of Cauchy steps has a twofold goal: forcing the search in the two-dimensional space spanned by the eigenvectors d_1 and d_n and getting a suitable approximation of the reciprocal of λ_1 through α_k^{Y}, in order to drive toward zero μ_1^k. If the component of the gradient along the eigenvector d_1 were completely removed, a sequence of Cauchy steps followed by constant steps computed with the Yuan rule would drive toward zero the component along the eigenvector d_2, and so on. Thus, the cyclic alternation of steplengths defined by (5) attempts to eliminate the components of the gradient according to the decreasing order of the eigenvalues of A. The SDA method has similar properties; in this case, the selected constant steplength attempts to exploit the tendency of the gradient method with steplength $1/(\lambda_1 + \lambda_n)$ to align the search direction with d_n, i.e., to eliminate the remaining eigencomponents. We also observe that if the Hessian matrix is ill conditioned, $1/(\lambda_1 + \lambda_n) \approx 1/\lambda_1$ and then SDA and SDC are expected to have very close behaviours. As the GMR methods, SDA and SDC have R-linear convergence, but in practice are competitive

with the fastest gradient methods currently available. Furthermore, although the two methods are non-monotonic, a suitable choice of h and m leads to monotonicity in practice.

The alternation of Cauchy steplengths and constant steplengths characterizes also the Cauchy-short steps methods proposed in [10]. The idea is to break the SD cycle by applying either very short or very long steps approximating suitable Hessian eigenvalues. Note that this strategy is also shared by the SDA and SDC methods, although they have been designed by taking a different point of view.

Finally, a different approach aimed at capturing the spectrum of the Hessian is exploited by the limited memory steepest descent method proposed in [20]. The basic idea is to divide the sequence of gradient iterations into groups of $m \geq 1$ iterations, referred to as sweeps, and to compute the steplengths for each sweep as the inverse of the Ritz values of the Hessian matrix, by exploiting the gradients obtained during the previous sweep.

NUMERICAL ILLUSTRATION

In order to illustrate our analysis, we compare some gradient methods on a very simple problem [17] of the form (1), with Hessian matrix

$$A = \mathrm{diag}(\lambda_1, \ldots, \lambda_{10}), \qquad \lambda_i = 111(11 - i) - 110, \qquad i = 1, \ldots, 10,$$

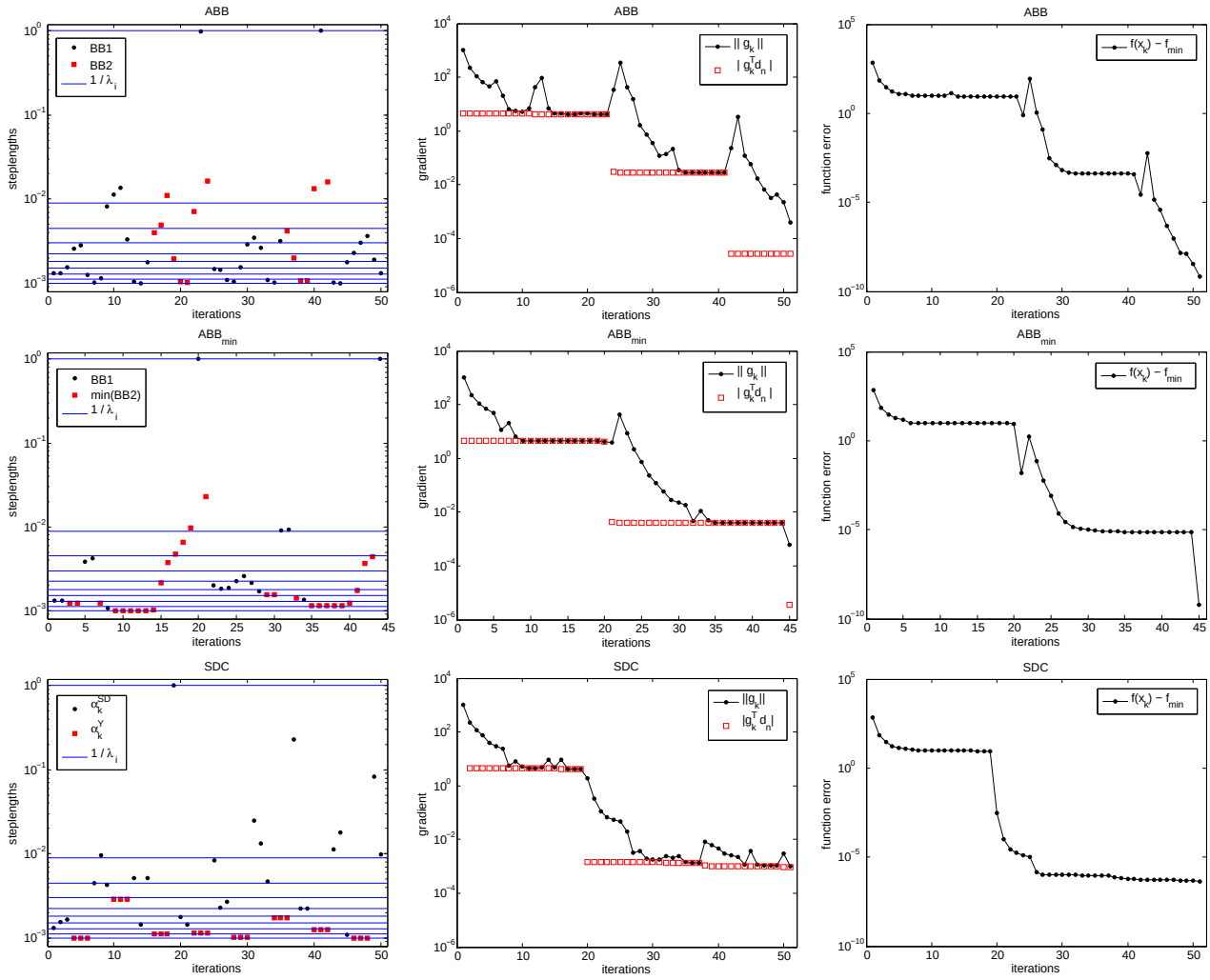

FIGURE 1. History of steplength, gradient norm and component along d_n, and function error ($f(x_k) - f_{min}$), for the ABB (top), $\mathrm{ABB}_{\min}$ (middle) and SDC (bottom) methods.

and random b with entries in $[-10, 10]$. For the sake of space, we only consider ABB ($\tau = 0.15$), $\text{ABB}_{\min}$ ($\tau = 0.8$, $m = 6$) and SDC ($h = 3$, $m = 3$), which are representative of most of the strategies described in the previous section. The starting guess x_0 has been randomly generated too, with entries in $[-1, 1]$. As stop condition we take $\|g_k\| < \|g_0\| \epsilon$, with $\epsilon = 10^{-6}$. We focus on the distribution of the steplengths α_k in the interval $[1/\lambda_1, 1/\lambda_n] = [0.001, 1]$ and on its impact on the convergence behaviour.

We first compare the ABB and $\text{ABB}_{\min}$ methods (see Fig. 1, top and middle). $\text{ABB}_{\min}$ tends to use the BB2 rule many more times, thus taking steplengths that on the average are smaller than ABB. $\text{ABB}_{\min}$ produces very few large steps; this happens twice, at iterations 20 and 44, with a quite remarkable effect in reducing the gradient component along the eigenvector d_n, and more generally along d_i for large i. The long steps appear to produce in the objective function some fluctuation followed by a strong decrease. The general behaviour of ABB is similar, but the non-monotonicity is slightly more noticeable, and this seems to deteriorate the performance of the method. We verified that this behaviour becomes more evident as the accuracy requirement increases. For instance, when $\epsilon = 10^{-8}$, ABB takes almost twice the number of iterations taken by $\text{ABB}_{\min}$.

Figure 1 (bottom) shows that the SDC method has a convergence history close to that of the $\text{ABB}_{\min}$ method. However, as observed in the previous section, SDC has a monotonic behaviour, fostered by Yuan steps that are very short in agreement with Theorem 2. A careful examination shows that the first 18 iterations are able to significantly reduce the gradient components along d_i for small i (this can be deduced from $\|g_k\| \approx |d_n^T g_k|$)), thus allowing the method to adopt a long step (almost equal to $1 = 1/\lambda_n$) at iteration 19, which produces a large decrease in the objective function and a strong reduction of the gradient component along d_n. As for $\text{ABB}_{\min}$, the use of few selected long steps produces remarkable effects on the overall SDC behaviour.

In conclusion, the methods we considered, although based on different strategies, apparently share the ability of using large steplengths in a selective way to overcome the chaotic behaviour of BB, allowing improvement in terms of monotonicity and computational efficiency. This ability can be successfully exploited also in more general contexts of unconstrained/constrained optimization [4, 6, 11, 14, 20], when the large scale of the applications makes the use of effective gradient approaches an unavoidable choice.

ACKNOWLEDGMENTS

This work was partially supported by GNCS-INdAM (Progetti 2016).

REFERENCES

[1] J. Barzilai and J. Borwein, IMA J. Numer. Anal. **8**, 141–148 (1988).
[2] M. Prato, R. Cavicchioli, L. Zanni, P. Boccacci, and M. Bertero, Astron. Astrophys. **539**, A133–(11pp) (2012).
[3] L. Antonelli, V. De Simone, and D. di Serafino, J. Math. Imaging Vis. **54**, 106–116 (2015).
[4] F. Porta, R. Zanella, G. Zanghirati, and L. Zanni, Commun. Nonlinear Sci. Numer. Simul. **21**, 112–127 (2015).
[5] R. De Asmundis, D. di Serafino, and G. Landi, J. Comput. Appl. Math. **302**, 81–93 (2016).
[6] L. Zanni, Computational Management Science **3**, 131–145 (2006).
[7] H. Akaike, Ann. Inst. Stat. Math. Tokyo **11**, 1–16 (1959).
[8] J. Nocedal, A. Sartenaer, and C. Zhu., Comput. Optim. Appl. **22**, 5–35 (2002).
[9] R. De Asmundis, D. di Serafino, F. Riccio, and G. Toraldo, IMA J. Numer. Anal. **33**, 1416–1435 (2013).
[10] C. Gonzaga and R. Schneider, Comput. Optim. Appl. **63**, 523–542 (2016).
[11] R. Fletcher, in *Optimization and Control with Applications*, Applied Optimization, Vol. 96, edited by L. Qi, K. Teo, X. Yang, P. M. Pardalos, and D. Hearn (Springer, US, 2005), pp. 235–256.
[12] M. Raydan and B. F. Svaiter, Comput. Optim. Appl. **21**, 155–167 (2002).
[13] Y.-H. Dai, Optimization **53**, 395–415 (2003).
[14] Y.-H. Dai, W. Hager, K. Schittkowski, and H. Zhang, IMA J. Num. Anal. **26**, 604–627 (2006).
[15] A. Friedlander, J. M. Martínez, B. Molina, and M. Raydan, SIAM J. Numer. Anal. **36**, 275–289 (1999).
[16] B. Zhou, L. Gao, and Y.-H. Dai, Comput. Optim. Appl. **35**, 69–86 (2006).
[17] G. Frassoldati, L. Zanni, and G. Zanghirati, J. Ind. Manag. Optim. **4**, 299–312 (2008).
[18] R. De Asmundis, D. di Serafino, W. Hager, G. Toraldo, and H. Zhang, Comput. Optim. Appl. **59**, 541–563 (2014).
[19] Y. Yuan, J. Comp. Math. **24**, 149–156 (2006).
[20] R. Fletcher, Math. Program., Ser. A **135**, 413–436 (2012).