

Alma Mater Studiorum Università di Bologna
Archivio istituzionale della ricerca

Solving Nonlinear Systems of Equations Via Spectral Residual Methods: Stepsize Selection and Applications

This is the final peer-reviewed author's accepted manuscript (postprint) of the following publication:

Published Version:

Meli E., Morini B., Porcelli M., Sgattoni C. (2022). Solving Nonlinear Systems of Equations Via Spectral Residual Methods: Stepsize Selection and Applications. JOURNAL OF SCIENTIFIC COMPUTING, 90(1), 1-41 [10.1007/s10915-021-01690-x].

Availability:

This version is available at: <https://hdl.handle.net/11585/845572> since: 2022-01-14

Published:

DOI: <http://doi.org/10.1007/s10915-021-01690-x>

Terms of use:

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (<https://cris.unibo.it/>).
When citing, please refer to the published version.

(Article begins on next page)

This is the final peer-reviewed accepted manuscript of:

Meli, E., Morini, B., Porcelli, M. et al. Solving Nonlinear Systems of Equations Via Spectral Residual Methods: Stepsize Selection and Applications. J Sci Comput 90, 30 (2022)

The final published version is available online at
<https://dx.doi.org/10.1007/s10915-021-01690-x>

Rights / License:

The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (<https://cris.unibo.it/>)

When citing, please refer to the published version.

SOLVING NONLINEAR SYSTEMS OF EQUATIONS VIA SPECTRAL RESIDUAL METHODS: STEPSIZE SELECTION AND APPLICATIONS

ENRICO MELI^{*}, BENEDETTA MORINI[†], MARGHERITA PORCELLI[‡], CRISTINA SGATTONI[§]

Abstract. Spectral residual methods are derivative-free and low-cost per iteration procedures for solving nonlinear systems of equations. They are generally coupled with a nonmonotone linesearch strategy and compare well with Newton-based methods for large nonlinear systems and sequences of nonlinear systems. The residual vector is used as the search direction and choosing the steplength has a crucial impact on the performance. In this work we address both theoretically and experimentally the steplength selection and provide results on a real application such as a rolling contact problem.

Keywords. Nonlinear systems of equations, spectral gradient methods, steplength selection, approximate norm descent methods

1. Introduction. This work addresses the use of spectral residual methods for solving systems of nonlinear equations

$$F(x) = 0, \tag{1.1}$$

where $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is continuously differentiable. The original proposal of spectral residual methods given in [25] was elaborated in [26] and gave rise to derivative-free iterative procedures for solving (1.1). In fact, given the iterate x_k , the residual vectors $\pm F(x_k)$ are used as search directions in a systematic way and a scalar β_k , inspired by the Barzilai and Borwein method for unconstrained optimization, is used as the first trial stepsize. Similarly to the Barzilai and Borwein method for unconstrained optimization, $\|F\|$ does not decrease monotonically along iterations and the effectiveness of spectral residual methods heavily relies on the steplengths β_k used.

Spectral residual methods belong to the class of Quasi-Newton methods which are particularly attractive when the Jacobian matrix of F is not available analytically or its computation is not relatively easy. Quasi-Newton methods showed to be effective both in the solution of large nonlinear systems and in the solution of sequences of medium-size nonlinear systems generated by model refinement procedures, see e.g., [5, 21, 25, 26, 31, 41]. Specifically, spectral residual methods have received a large attention since they are low-cost per iteration and require a low memory storage, being matrix-free, see e.g., [21, 25–27, 31, 34, 35, 41].

It is well known that the performance of the Barzilai and Borwein method for the unconstrained optimization problem $\min_{x \in \mathbb{R}^n} f(x)$ does not depend on the decrease of f at each iteration but relies on the relationship between the steplengths used and the eigenvalues of the average Hessian matrix of f [3, 15, 36], and consequently on the stepsize selection employed, see e.g., [8–10, 12, 15, 16]. On the other hand, to our knowledge, an analogous study has not been carried out for spectral residual methods. The aim of this paper is to analyze both the relationship between the steplengths β_k and the eigenvalues of average matrices associated to the Jacobian of F , and the impact of the stepsizes on the convergence history. This analysis is addressed from both a theoretical and an experimental point of view.

The main contributions of this work are: the theoretical analysis of the stepsizes proposed in the literature; the impact of the stepsizes on the norm of F when a nonmonotone globalization strategy is used; the analysis of the performance of spectral residual methods coupled with various rules for choosing the steplengths. Inspired by the steplength rules proposed in the literature

^{*}Dipartimento di Ingegneria Industriale, Università degli Studi di Firenze, via S. Marta 3, 50134 Firenze, Email: enrico.meli@unifi.it

[†]Dipartimento di Ingegneria Industriale, Università degli Studi di Firenze, viale G.B. Morgagni 40, 50134 Firenze, Italia. Email: benedetta.morini@unifi.it.

[‡]Dipartimento di Matematica, AM², Università di Bologna, Piazza di Porta San Donato 5, 40126 Bologna, Italia. Email: margherita.porcelli@unibo.it

[§]Dipartimento di Matematica e Informatica “Ulisse Dini”, Università degli Studi di Firenze, viale G.B. Morgagni 67a, 50134 Firenze, Italia. Email: cristina.sgattoni@unifi.it

[¶]Member of the INdAM Research Group GNCS.

^{||}Institute of Information Science and Technologies “A. Faedo”, ISTI-CNR, Via Moruzzi 1 Pisa, Italia.

for unconstrained minimization problems, we propose and extensively test adaptive strategies on sequences of nonlinear systems arising from rolling contact models. These models play a central role in applications, such as rolling bearings and wheel-rail interaction [23, 24], and their solution gives rise to a relevant benchmark test set of nonlinear systems. We show that adaptive rules combining small and large stepsizes are by far more effective than rules based on static choices of β_k .

The paper is organized as follows. Section 2 introduces spectral residual methods. In Section 3 and 4 we provide a theoretical analysis of the steplengths and of the fulfillment of a general nonmonotone linesearch. In Section 5 we introduce the spectral residual method used in our tests and provide a theoretical investigation. The experimental part is developed in Section 6 where we describe several steplength selection strategies, introduce our test set and discuss the numerical results obtained. Some conclusions are presented in Section 7.

1.1. Notations. The symbol $\|\cdot\|$ denotes the Euclidean norm, I denotes the identity matrix, J denotes the Jacobian matrix of F . Given a symmetric matrix M , $\{\lambda_i(M)\}_{i=1}^n$ denotes the set of eigenvalues of M , $\lambda_{\min}(M)$ and $\lambda_{\max}(M)$ denote the minimum and maximum eigenvalue of M respectively, and $\{v_i\}_{i=1}^n$ denotes a set of associated orthonormal eigenvectors. Given a sequence of vectors $\{x_k\}$, for any function f we let $f_k = f(x_k)$.

2. Preliminaries. In the seminal paper [2] Barzilai and Borwein proposed a gradient method for the unconstrained minimization

$$\min_{x \in \mathbb{R}^n} f(x), \quad (2.1)$$

where $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is a differentiable function. Given an initial guess $x_0 \in \mathbb{R}^n$, the Barzilai-Borwein (BB) iteration is defined by

$$x_{k+1} = x_k - \alpha_k \nabla f_k, \quad (2.2)$$

where α_k is a positive steplength inspired by Quasi-Newton methods for unconstrained optimization [11]. In Quasi-Newton methods, the step $p_k = x_{k+1} - x_k$ solves the linear system

$$B_k p_k = -\nabla f_k, \quad (2.3)$$

and B_k , $k \geq 1$, satisfies the secant equation, i.e.,

$$B_k p_{k-1} = z_{k-1}, \quad p_{k-1} = x_k - x_{k-1}, \quad z_{k-1} = \nabla f_k - \nabla f_{k-1}. \quad (2.4)$$

Letting $B_k = \alpha^{-1} I$ and imposing condition (2.4), Barzilai and Borwein derived two steplengths which are the least-square solutions of the following problems:

$$\alpha_{k,1} = \underset{\alpha}{\operatorname{argmin}} \|\alpha^{-1} p_{k-1} - z_{k-1}\|_2^2 = \frac{p_{k-1}^T p_{k-1}}{p_{k-1}^T z_{k-1}}, \quad (2.5)$$

$$\alpha_{k,2} = \underset{\alpha}{\operatorname{argmin}} \|p_{k-1} - \alpha z_{k-1}\|_2^2 = \frac{p_{k-1}^T z_{k-1}}{z_{k-1}^T z_{k-1}}. \quad (2.6)$$

The second least-squares formulation is obtained from the first by symmetry. The steplength α_k in (2.2) is supposed to be positive, bounded away from zero and not too large, i.e., $\alpha_k \in [\alpha_{\min}, \alpha_{\max}]$ for some positive $\alpha_{\min}, \alpha_{\max}$; to this end, one of the two scalars $\alpha_{k,1}, \alpha_{k,2}$ is selected and projected onto $[\alpha_{\min}, \alpha_{\max}]$ if necessary, see e.g., [3, 12, 15].

Choosing $B_k = \alpha^{-1} I$ yields a low-cost iteration while the use of the steplengths $\alpha_{k,1}, \alpha_{k,2}$ yields a considerable improvement in the performance with respect to the classical steepest descent method [2, 15]. Specifically, the performance of the BB method depends on the relationship between $\alpha_{k,1}, \alpha_{k,2}$ and the eigenvalues of the average Hessian matrix $\int_0^1 \nabla^2 f(x_{k-1} + t p_{k-1}) dt$, and an extensive investigation in stepsize selection was made in [8–10, 12, 15, 16]. The BB method is also denoted as *spectral method* and is commonly employed in the solution of large unconstrained

optimization problems (2.1). The behaviour of the sequence $\{f(x_k)\}$ is typically nonmonotone and ruled by linesearch strategies, [15, 17, 38].

The extension of this approach to the solution of nonlinear systems of equations (1.1) was firstly proposed by La Cruz and Raydan in [25]. Here we summarize such a proposal and the issues that were inherited by subsequent procedures for general nonlinear systems [21, 25–27, 31, 34, 41] and for monotone nonlinear systems [1, 29, 30, 32, 40, 44]. Instead of applying the spectral method to the merit function

$$f(x) = \|F(x)\|^2, \quad (2.7)$$

the BB approach is specialized to the Newton equation yielding the so-called *spectral residual method*. Thus, let p_- satisfy the linear system

$$B_k p_- = -F_k, \quad (2.8)$$

and let $B_k = \beta^{-1}I$ satisfy the secant equation

$$B_k p_{k-1} = y_{k-1}, \quad p_{k-1} = x_k - x_{k-1}, \quad y_{k-1} = F_k - F_{k-1}.$$

Reasoning as in BB method, two steplengths are derived:

$$\beta_{k,1} = \frac{p_{k-1}^T p_{k-1}}{p_{k-1}^T y_{k-1}}, \quad (2.9)$$

$$\beta_{k,2} = \frac{p_{k-1}^T y_{k-1}}{y_{k-1}^T y_{k-1}}. \quad (2.10)$$

These scalars may be positive, negative or even null; moreover $\beta_{k,1}$ is not well defined if $p_{k-1}^T y_{k-1} = 0$ and $\beta_{k,2}$ is not well defined if $y_{k-1} = 0$. In practice, the steplength β_k is chosen equal to $\beta_{k,1}$ or to $\beta_{k,2}$ as long as it results to be bounded away from zero and $|\beta_k|$ is not too large, i.e., $|\beta_k| \in [\beta_{\min}, \beta_{\max}]$ for some positive $\beta_{\min}, \beta_{\max}$.

The step resulting from (2.8) turns out to be of the form $p_- = -\beta_k F_k$ but the k th iteration of the spectral residual method employs the residual directions $\pm F_k$ in a systematic way and tests both the steps

$$p_- = -\beta_k F_k \quad \text{and} \quad p_+ = +\beta_k F_k,$$

for acceptance using a suitable linesearch strategy. The use of both directions $\pm F_k$ is motivated by the fact that, contrary to $(-\alpha_k \nabla f_k)$, $\alpha_k > 0$, in (2.2), $(-\beta_k F_k)$ may not be a descent direction for (2.7) at x_k . Letting J be the Jacobian of F , the value $\nabla f_k^T (-\beta_k F_k) = -2\beta_k F_k^T J_k F_k$ could be positive, negative or null but, as long as $F_k^T J_k F_k \neq 0$, either $(-\beta_k F_k)$ or $\beta_k F_k$ is a descent direction for f .

Analogously to the spectral method, the spectral residual method is implemented using non-monotone linesearch strategies. The adaptation of the spectral method to nonlinear systems is low-cost per iteration since the computation of $\beta_{k,1}$ and $\beta_{k,2}$ is inexpensive and quite effective in the solution of medium and large nonlinear systems, see e.g., [21, 25–27, 34, 41].

Unlike the case of spectral method, to our knowledge a systematic analysis of the stepsizes $\beta_{k,1}$ and $\beta_{k,2}$ has not been performed. In a large number of papers, the steplength $\beta_{k,1}$ is used, [25–27, 31, 34]. On the other hand, in [21] it was observed experimentally that alternating $\beta_{k,1}$ and $\beta_{k,2}$ along iterations was beneficial for the performance and in [41] it was observed experimentally that using $\beta_{k,2}$ performed better with respect $\beta_{k,1}$ in terms of robustness. Therefore, in the next two sections we provide: the expression of $\beta_{k,1}$ and $\beta_{k,2}$ in terms of the spectrum of average matrices associated to the Jacobian matrix of F , their mutual relationship, the analysis of their impact on the behaviour of $\|F_k\|$.

The matrices involved in our analysis are the following. Given a square matrix A , we let $A_S = \frac{1}{2}(A + A^T)$ be the symmetric part of A , G_{k-1} be the average matrix associated to the

Jacobian J of F around x_{k-1}

$$G_{k-1} \stackrel{\text{def}}{=} \int_0^1 J(x_{k-1} + t p_{k-1}) dt, \quad (2.11)$$

and $(G_S)_{k-1}$ be the average matrix associated to the symmetric part J_S of J around x_{k-1}

$$(G_S)_{k-1} \stackrel{\text{def}}{=} \int_0^1 J_S(x_{k-1} + t p_{k-1}) dt. \quad (2.12)$$

Moreover, given a symmetric matrix M and a nonzero vector p , we employ the Rayleigh quotient defined as

$$q(M, p) = \frac{p^T M p}{p^T p}, \quad (2.13)$$

and the following property [18, Theorem 8.1-2]

$$\lambda_{\min}(M) \leq q(M, p) \leq \lambda_{\max}(M). \quad (2.14)$$

3. Analysis of the steplengths $\beta_{k,1}$ and $\beta_{k,2}$. We analyze the stepsizes $\beta_{k,1}$ and $\beta_{k,2}$ given in (2.9) and (2.10) making the following assumptions.

ASSUMPTION 3.1. *The scalars $\beta_{k,1}$ and $\beta_{k,2}$ are well defined and nonzero.*

ASSUMPTION 3.2. *Given x and p , F is continuously differentiable in an open convex set $D \subset \mathbb{R}^n$ containing $x + tp$ with $t \in [0, 1]$.*

We note that Assumption 3.1 holds whenever $p_{k-1}^T y_{k-1} \neq 0$.

In the following lemma we analyze the mutual relationship between the stepsizes $\beta_{k,1}$ and $\beta_{k,2}$ and give their characterization in terms of suitable Rayleigh quotients for the average matrices in (2.11) and (2.12). We use repeatedly the property

$$p^T A p = p^T A_S p, \quad (3.1)$$

which holds for any square matrices A , $A_S = \frac{1}{2}(A + A^T)$, and any vector p of suitable dimension.

LEMMA 3.3. *Let Assumption 3.1 hold and Assumption 3.2 hold with $x = x_{k-1}$, $p = p_{k-1} = \pm \beta_{k-1} F_{k-1}$. The steplengths $\beta_{k,1}$, $\beta_{k,2}$ are such that:*

- P1) *they have the same sign and $|\beta_{k,2}| \leq |\beta_{k,1}|$;*
- P2) *either it holds $\beta_{k,1} \leq \beta_{k,2} < 0$ or $0 < \beta_{k,2} \leq \beta_{k,1}$;*
- P3) *they take the form*

$$\beta_{k,1} = \frac{1}{q((G_S)_{k-1}, p_{k-1})} = \frac{1}{q((G_S)_{k-1}, F_{k-1})}, \quad (3.2)$$

and

$$\beta_{k,2} = \frac{q((G_S)_{k-1}, p_{k-1})}{q(G_{k-1}^T G_{k-1}, p_{k-1})} = \frac{q((G_S)_{k-1}, F_{k-1})}{q(G_{k-1}^T G_{k-1}, F_{k-1})}, \quad (3.3)$$

with G_{k-1} , $(G_S)_{k-1}$ and $q(\cdot, \cdot)$ given in (2.11), (2.12) and (2.13), respectively.

Proof. By (2.9) and (2.10), we can write

$$\begin{aligned} \beta_{k,2} &= \frac{p_{k-1}^T p_{k-1}}{p_{k-1}^T y_{k-1}} \frac{(p_{k-1}^T y_{k-1})^2}{(y_{k-1}^T y_{k-1})(p_{k-1}^T p_{k-1})} \\ &= \beta_{k,1} \frac{\|p_{k-1}\|^2 \|y_{k-1}\|^2 \cos^2 \varphi_{k-1}}{\|p_{k-1}\|^2 \|y_{k-1}\|^2} \\ &= \beta_{k,1} \cos^2 \varphi_{k-1}, \end{aligned} \quad (3.4)$$

where φ_{k-1} is the angle between p_{k-1} and y_{k-1} , and P1) follows.

Property P2) follows as well since $\beta_{k,2} \neq 0$ by Assumption 3.1.

As for property P3), by the Mean Value Theorem [11, Lemma 4.1.9] and (2.11) we have

$$y_{k-1} = F_k - F_{k-1} = \int_0^1 J(x_{k-1} + tp_{k-1})p_{k-1} dt = G_{k-1}p_{k-1}.$$

Then using (3.1) and (2.13), $\beta_{k,1}$ takes the form

$$\beta_{k,1} = \frac{p_{k-1}^T p_{k-1}}{p_{k-1}^T G_{k-1} p_{k-1}} = \frac{p_{k-1}^T p_{k-1}}{p_{k-1}^T (G_S)_{k-1} p_{k-1}} = \frac{1}{q((G_S)_{k-1}, p_{k-1})},$$

while $\beta_{k,2}$ takes the form

$$\beta_{k,2} = \frac{p_{k-1}^T G_{k-1} p_{k-1}}{p_{k-1}^T (G_{k-1}^T G_{k-1}) p_{k-1}} \frac{p_{k-1}^T p_{k-1}}{p_{k-1}^T p_{k-1}} = \frac{q((G_S)_{k-1}, p_{k-1})}{q(G_{k-1}^T G_{k-1}, p_{k-1})}.$$

The rightmost equalities in (3.2) and (3.3) easily follow using the form of the step $p_{k-1} = \pm \beta_{k-1} F_{k-1}$. \square

The above characterization P3) yields bounds on the stepsizes $\beta_{k,1}$, $\beta_{k,2}$ in terms of the extreme eigenvalues of the average matrices in (2.11) and (2.12). A relationship between $\beta_{k,1}$ and the eigenvalues of $(G_S)_{k-1}$ was observed in [25, 26, 34] but the following results are not contained in such references.

LEMMA 3.4. *Let Assumption 3.1 hold and Assumption 3.2 hold with $x = x_{k-1}$, $p = p_{k-1}$. Then, the steplengths $\beta_{k,1}$ and $\beta_{k,2}$ are such that:*

- (i) *If the Jacobian J is symmetric and positive definite on the line segment in between x_{k-1} and $x_{k-1} + p_{k-1}$ then $\beta_{k,1}$ and $\beta_{k,2}$ are positive and*

$$\frac{1}{\lambda_{\max}(G_{k-1})} \leq \beta_{k,2} \leq \beta_{k,1} \leq \frac{1}{\lambda_{\min}(G_{k-1})}; \quad (3.5)$$

- (ii) *if $(G_S)_{k-1}$ in (2.12) is positive definite, then $\beta_{k,1}$ and $\beta_{k,2}$ are positive and*

$$\max \left\{ \frac{1}{\lambda_{\max}((G_S)_{k-1})}, \beta_{k,2} \right\} \leq \beta_{k,1} \leq \frac{1}{\lambda_{\min}((G_S)_{k-1})}, \quad (3.6)$$

$$\frac{\lambda_{\min}((G_S)_{k-1})}{\lambda_{\max}(G_{k-1}^T G_{k-1})} \leq \beta_{k,2} \leq \min \left\{ \frac{\lambda_{\max}((G_S)_{k-1})}{\lambda_{\min}(G_{k-1}^T G_{k-1})}, \beta_{k,1} \right\}; \quad (3.7)$$

- (iii) *if $(G_S)_{k-1}$ in (2.12) is indefinite and G_{k-1} in (2.11) is nonsingular, then*

- (iii.1) $\beta_{k,1}$ satisfies either

$$\beta_{k,1} \leq \min \left\{ \frac{1}{\lambda_{\min}((G_S)_{k-1})}, \beta_{k,2} \right\} \quad \text{or} \quad \beta_{k,1} \geq \max \left\{ \frac{1}{\lambda_{\max}((G_S)_{k-1})}, \beta_{k,2} \right\}; \quad (3.8)$$

- (iii.2) $\beta_{k,2}$ satisfies either

$$0 < \beta_{k,2} \leq \min \left\{ \frac{\lambda_{\max}((G_S)_{k-1})}{\lambda_{\min}(G_{k-1}^T G_{k-1})}, \beta_{k,1} \right\}, \quad (3.9)$$

or

$$\max \left\{ \frac{\lambda_{\min}((G_S)_{k-1})}{\lambda_{\max}(G_{k-1}^T G_{k-1})}, \beta_{k,1} \right\} \leq \beta_{k,2} < 0. \quad (3.10)$$

Proof. Consider properties P1), P2) and P3) from Lemma 3.3.

- (i) Steplengths $\beta_{k,1}$ and $\beta_{k,2}$ are positive due to (3.2), (3.3). The rightmost inequality of (3.5) follows from (3.2) and (2.14). The remaining part of (3.5) is proved observing that (3.3) yields

$$\beta_{k,2} = \frac{p_{k-1}^T G_{k-1}^{1/2} G_{k-1}^{1/2} p_{k-1}}{p_{k-1}^T G_{k-1}^{1/2} G_{k-1}^{1/2} p_{k-1}} = \frac{1}{q(G_{k-1}, G_{k-1}^{1/2} p_{k-1})}, \quad (3.11)$$

and using P2) and (2.14).

- (ii) Using (3.2), (2.14) and P2) we get positivity of $\beta_{k,1}$ and (3.6). Consequently, $\beta_{k,2}$ is positive by property P1), and bounds (3.7) can be derived using (3.3), (2.14) and item P2) of Lemma 3.3.
- (iii) If $(G_S)_{k-1}$ is indefinite then its extreme eigenvalues have opposite sign, i.e., $\lambda_{\min}((G_S)_{k-1}) < 0$ and $\lambda_{\max}((G_S)_{k-1}) > 0$. Hence, (3.2), (2.14) and P2) give (3.8). Moreover, since $G_{k-1}^T G_{k-1}$ is symmetric and positive definite, we can use, as before, P1) and (2.14) and get (3.9) and (3.10). \square

REMARK 3.5. Lemma 3.4 easily extends to the case where matrices are negative definite.

Item (ii) of Lemma 3.4 includes the case where F is strictly monotone, i.e., $(F(x) - F(y))^T(x - y) > 0$ for any $x, y \in \mathbb{R}^n$ with $x \neq y$, see e.g. [14].

4. On the impact of the steplength β_k on $\|F_{k+1}\|$. In this section we investigate how the choice of the steplength β_k may affect $\|F_{k+1}\|$. Results are derived using a generic β_k and specialized thereafter for $\beta_{k,1}$ and $\beta_{k,2}$.

The first result concerns the case where J is symmetric and analyzes the residual vector F_{k+1} componentwise. It heavily relies on the existence of a set of orthonormal eigenvectors for the average matrix G_k .

LEMMA 4.1. Suppose that Assumption 3.2 holds with $x = x_k$ and $p = p_k$ and that the Jacobian J is symmetric. Let $p_k = p_- = -\beta_k F_k \neq 0$, $x_{k+1} = x_k + p_k$, $\{\lambda_i(G_k)\}_{i=1}^n$ be the eigenvalues of matrix G_k in (2.11) and $\{v_i\}_{i=1}^n$ be a set of associated orthonormal eigenvectors. Let F_k and F_{k+1} be expressed as

$$F_k = \sum_{i=1}^n \mu_k^i v_i, \quad F_{k+1} = \sum_{i=1}^n \mu_{k+1}^i v_i,$$

where μ_k^i, μ_{k+1}^i , $i = 1, \dots, n$, are scalars. Then

$$F_{k+1} = (I - \beta_k G_k) F_k, \quad (4.1)$$

$$\mu_{k+1}^i = \mu_k^i (1 - \beta_k \lambda_i(G_k)), \quad i = 1, \dots, n. \quad (4.2)$$

Moreover, it holds:

- (a) if $\beta_k \lambda_i(G_k) = 1$, then $|\mu_{k+1}^i| = 0$;
(b) if $0 < \beta_k \lambda_i(G_k) < 2$, then $|\mu_{k+1}^i| < |\mu_k^i|$; otherwise $|\mu_{k+1}^i| \geq |\mu_k^i|$.

Proof. The Mean Value Theorem [11, Lemma 4.1.9] gives

$$F_{k+1} = F_k + \int_0^1 J(x_k + tp_k) p_k dt,$$

and $p_k = -\beta_k F_k$ and (2.11) yield (4.1). Since $\{v_i\}_{i=1}^n$ are orthonormal we have for $i = 1, \dots, n$

$$\begin{aligned} \mu_{k+1}^i &= (v_i)^T F_{k+1} \\ &= (v_i)^T (I - \beta_k G_k) F_k \\ &= \mu_k^i (1 - \beta_k \lambda_i(G_k)), \end{aligned}$$

i.e., equation (4.2). Consequently, Item (a) follows trivially; Item (b) follows noting that $|1 - \beta_k \lambda_i(G_k)| < 1$ if and only if $0 < \beta_k \lambda_i(G_k) < 2$. \square

REMARK 4.2. *Lemma 4.1 trivially extends to the case where $p_k = p_+ = \beta_k F_k$.*

If the nonlinear system (1.1) represents the first-order optimality condition of the quadratic optimization problem (2.1) with $f(x) = \frac{1}{2}x^T A x - b^T x$, A symmetric and positive definite, then the previous lemma reduces to well known results on the behaviour of the gradient method in terms of the spectrum of the Hessian matrix A , see [36]. In fact, the nonlinear residual is $F(x) = Ax - b$ and its Jacobian is constant $J(x) = A$, $\forall x$. Then, (4.2) becomes

$$\mu_{k+1}^i = \mu_k^i (1 - \beta_k \lambda_i(A)) = \mu_0^i \prod_{j=0}^k (1 - \beta_j \lambda_i(A)).$$

As a consequence, a small steplength β_k , i.e., close to $1/\lambda_{\max}(A)$, can significantly reduce the values $|\mu_{k+1}^i|$ corresponding to large eigenvalues $\lambda_i(A)$ while a small reduction is expected for the scalars $|\mu_{k+1}^i|$ corresponding to small eigenvalues $\lambda_i(A)$. On the contrary, a large steplength β_k , i.e., close to $1/\lambda_{\min}(A)$, can significantly reduce the values $|\mu_{k+1}^i|$ corresponding to small eigenvalues $\lambda_i(A)$ while tends to increase the scalar $|\mu_{k+1}^i|$ corresponding to large eigenvalues $\lambda_i(A)$. This offers some intuition for choosing the steplengths by alternating in a balanced way small and large steplengths in order to reduce the eigencomponents, see e.g., [12, p. 178].

On the other hand, if F is a general nonlinear mapping then G_k changes at each iteration and Lemma 4.1 suggests the following guidelines. The first guideline concerns the case where J is positive definite. A nonmonotone behaviour of the sequence $\{\|F_k\|\}$ is expected. By Item (i) of Lemma 3.4, both $\beta_{k,1}$ or $\beta_{k,2}$ are positive and $\beta_k \lambda_i(G_k)$ lies in the interval $\left[\frac{\lambda_i(G_k)}{\lambda_{\max}(G_{k-1})}, \frac{\lambda_i(G_k)}{\lambda_{\min}(G_{k-1})} \right]$ for $i = 1, \dots, n$. Assuming without loss of generality that the eigenvalues are numbered in nondecreasing order, by standard arguments on perturbation theory for the eigenvalues it holds

$$|\lambda_i(G_k) - \lambda_i(G_{k-1})| \leq \|G_k - G_{k-1}\|,$$

$i = 1, \dots, n$, [18, Theorem 8.1-6]. Thus, if the Jacobian is Lipschitz continuous in an open convex set containing $x_{k-1} + tp_{k-1}$ and $x_k + tp_k$ with constant $L_J > 0$, it follows

$$\|G_k - G_{k-1}\| \leq \frac{L_J}{2} (\|p_{k-1}\| + \|p_k\|).$$

Consequently, if $\|p_{k-1}\|$ and/or $\|p_k\|$ are large, by Item (b) no decrease of μ_{k+1}^i may occur. On the contrary, for small values of $\|p_{k-1}\|$ and $\|p_k\|$, as occurs if $\{x_k\}$ is convergent, G_k undergoes small changes with respect to G_{k-1} and the behaviour of μ_{k+1}^i shows similarities with the case above where J is constant and positive definite.

The second guideline concerns the case where J is indefinite and $\lambda_{\min}(G_k) < 0 < \lambda_{\max}(G_k)$. If $\beta_k > 0$, from Item (b) it follows that $|\mu_{k+1}^i|$ corresponding to positive $\lambda_i(G_k)$ are smaller than $|\mu_k^i|$ if $\beta_k \lambda_i(G_k)$ is small enough while all $|\mu_{k+1}^i|$ corresponding to negative eigenvalues increase with respect to $|\mu_k^i|$ and the amplification depends on the magnitude of $\beta_k \lambda_i(G_k)$. If $\beta_k < 0$ similar conclusions hold. In general, a nonmonotone behaviour of the sequence $\{\|F_k\|\}$ is expected but a possibly large increase of $\|F_{k+1}\|$ with respect to $\|F_k\|$ does not occur if $\{|\beta_k \lambda_i(G_k)|\}_{i=1, \dots, n}$ are small or of moderate size. Since a small value of $\{|\beta_k \lambda_i(G_k)|\}_{i=1, \dots, n}$ might be induced by a small value of $|\beta_k|$, the use of $\beta_{k,2}$ might be advisable taking into account that $|\beta_{k,2}| \leq |\beta_{k,1}|$ and $\beta_{k,1}$ can arbitrarily grow in the indefinite case (see Lemma 3.4).

4.1. On the impact of the steplength β_k in the approximate norm descent line-search. In this section we embed the spectral residual method in a general globalization scheme based on the so-called approximate norm descent condition [28]

$$\|F_{k+1}\| \leq (1 + \eta_k) \|F_k\|, \tag{4.3}$$

where $\{\eta_k\}$ is a positive sequence satisfying

$$\sum_{k=0}^{\infty} \eta_k < \eta < \infty. \quad (4.4)$$

Intuitively, large values of η_k allow a highly nonmonotone behaviour of $\|F_k\|$ while small values of η_k promote the decrease of $\|F\|$. Several linesearch strategies in the literature fall in this scheme [19, 28, 31, 34]. The main idea is that, given x_k , the trial steps take the form

$$p_- = -\gamma_k \beta_k F_k \quad \text{or} \quad p_+ = +\gamma_k \beta_k F_k \quad (4.5)$$

where $\gamma_k \in (0, 1]$ and (4.3) is enforced using a backtracking process. The sequence generated is such that $\{\|F_k\|\}$ is convergent [28, Lemma 2.4].

We now analyse the properties of $\|F_{k+1}\|$ as a function of the stepsize γ_k/β_k and determine conditions on γ_k/β_k which enforce (4.3). First of all we observe that by the Mean Value Theorem [11, Lemma 4.1.9] and (4.5) we have

$$F_{k+1} = (I \pm \gamma_k \beta_k G_k) F_k. \quad (4.6)$$

Using this equation we can write

$$\|F_{k+1}\|^2 = \|F_k\|^2 \pm 2\gamma_k \beta_k F_k^T (G_S)_k F_k + \gamma_k^2 \beta_k^2 F_k^T G_k^T G_k F_k, \quad (4.7)$$

and analyze when either $\|F_{k+1}\| < \|F_k\|$ or (4.3) holds.

THEOREM 4.3. *Suppose that Assumption 3.1 holds and Assumption 3.2 holds with $x = x_k$ and $p = p_k$. Suppose $F_k^T J_k F_k \neq 0$ and $F_k^T G_k F_k \neq 0$ with G_k given in (2.11). Let $\Delta = q((G_S)_k, F_k)^2 + (\eta_k^2 + 2\eta_k)q(G_k^T G_k, F_k)$, then*

(1) *If $x_{k+1} = x_k + p_k$, $p_k = p_- = -\gamma_k \beta_k F_k$, $\gamma_k \in (0, 1]$, we have that $\|F_{k+1}\| < \|F_k\|$ when*

$$\beta_k q((G_S)_k, F_k) > 0 \quad \text{and} \quad \gamma_k |\beta_k| < 2 \frac{|q((G_S)_k, F_k)|}{q(G_k^T G_k, F_k)}. \quad (4.8)$$

Condition (4.3) is satisfied when

$$\frac{q((G_S)_k, F_k) - \sqrt{\Delta}}{q(G_k^T G_k, F_k)} \leq \gamma_k \beta_k \leq \frac{q((G_S)_k, F_k) + \sqrt{\Delta}}{q(G_k^T G_k, F_k)}. \quad (4.9)$$

(2) *If $x_{k+1} = x_k + p_k$, $p_k = p_+ = \gamma_k \beta_k F_k$, $\gamma_k \in (0, 1]$, we have that $\|F_{k+1}\| < \|F_k\|$ when*

$$\beta_k q((G_S)_k, F_k) < 0 \quad \text{and} \quad \gamma_k |\beta_k| < 2 \frac{|q((G_S)_k, F_k)|}{q(G_k^T G_k, F_k)} \quad (4.10)$$

Condition (4.3) is satisfied when

$$\frac{-q((G_S)_k, F_k) - \sqrt{\Delta}}{q(G_k^T G_k, F_k)} \leq \gamma_k \beta_k \leq \frac{-q((G_S)_k, F_k) + \sqrt{\Delta}}{q(G_k^T G_k, F_k)}. \quad (4.11)$$

Proof. Concerning Item (1), using (4.6) we get

$$\begin{aligned} \|F_{k+1}\|^2 &= \|(I - \gamma_k \beta_k G_k) F_k\|^2 \\ &= \left(1 - 2\gamma_k \beta_k \frac{F_k^T (G_S)_k F_k}{\|F_k\|^2} + \gamma_k^2 \beta_k^2 \frac{F_k^T G_k^T G_k F_k}{\|F_k\|^2}\right) \|F_k\|^2 \\ &= \left(1 - 2\gamma_k \beta_k q((G_S)_k, F_k) + \gamma_k^2 \beta_k^2 q(G_k^T G_k, F_k)\right) \|F_k\|^2. \end{aligned}$$

Noting that by assumption $q((G_S)_k, F_k) \neq 0$ and $q(G_k^T G_k, F_k) > 0$, $\|F_{k+1}\| < \|F_k\|$ holds if

$$\beta_k q((G_S)_k, F_k) > 0 \quad \text{and} \quad -2\gamma_k \beta_k q((G_S)_k, F_k) + \gamma_k^2 \beta_k^2 q(G_k^T G_k, F_k) < 0,$$

and these conditions can be rewritten as in (4.8). Condition (4.9) follows trivially.

Item (2) follows analogously. From (4.6) and imposing $\|F_{k+1}\| < \|F_k\|$, we get the condition

$$\beta_k q((G_S)_k, F_k) < 0 \quad \text{and} \quad 2\gamma_k \beta_k q((G_S)_k, F_k) + \gamma_k^2 \beta_k^2 q(G_k^T G_k, F_k) < 0$$

which is equivalent to (4.10). Condition (4.11) follows trivially. \square

We remark that, due to the form of G_k and $(G_S)_k$, conditions (4.8)–(4.11) are implicit in $\gamma_k \beta_k$. The above theorem supports testing the two steps (4.5) systematically because at k -th iteration, β_k , $q(J_k, F_k)$ and $q(J_k^T J_k, F_k)$ are given and by continuity of the Jacobian, the Rayleigh quotients $q((G_S)_k, F_k)$ and $q(G_k^T G_k, F_k)$ tend to $q(J_k, F_k)$ and $q(J_k^T J_k, F_k)$ respectively as γ_k tends to zero. Hence, if γ_k is sufficiently small it holds

$$\frac{q(J_k, F_k) - \epsilon}{q(J_k^T J_k, F_k) + \epsilon} \leq \frac{q((G_S)_k, F_k)}{q(G_k^T G_k, F_k)} \leq \frac{q(J_k, F_k) + \epsilon}{q(J_k^T J_k, F_k) - \epsilon},$$

with $0 < \epsilon < \frac{1}{2} \min\{|q(J_k, F_k)|, q(J_k^T J_k, F_k)\}$, and $\frac{q((G_S)_k, F_k)}{q(G_k^T G_k, F_k)}$ has the same sign as $\frac{q(J_k, F_k)}{q(J_k^T J_k, F_k)}$. Consequently, for γ_k sufficiently small, either condition (4.8) or (4.10) is fulfilled. Analogous considerations can be made for conditions (4.9) and (4.11).

As a final comment, the previous theorem suggests that a small $|\beta_k|$ promotes the fulfillment of conditions (4.8), (4.10) or (4.9), (4.11). Again, by Lemma 3.4, the use of $\beta_{k,2}$ may be advisable taking into account that $|\beta_{k,2}| \leq |\beta_{k,1}|$ and that $\beta_{k,1}$ can arbitrarily grow in the indefinite case; taking the steplength equal to $\beta_{k,1}$ may cause a large number of backtracks and an erratic behaviour of $\{\|F_k\|\}$ as long as η_k is sufficiently large.

5. A spectral residual approximate norm descent method. In this section we describe a spectral residual algorithm which implements a line-search along $\pm F_k$ and enforces the approximate norm descent condition (4.3). We also discuss the convergence properties of the method and provide sufficient conditions for the convergence of the sequence $\{\|F_k\|\}$ to zero.

The Projected Approximate Norm Descent (PAND) algorithm was developed in [34] for solving nonlinear systems with convex constraints. In [31, 34], several variants based on Quasi-Newton methods have been proposed. Here we consider the spectral residual implementation for unconstrained nonlinear systems and denote it as Spectral Residual Approximate Norm Descent (SRAND) method.

Given the current iterate x_k , a new iterate x_{k+1} is computed as $x_{k+1} = x_k + p_k$ with p_k given by either $(-\gamma_k \beta_k F_k)$ or $(+\gamma_k \beta_k F_k)$, $\gamma_k \in (0, 1]$. The main phases of SRAND are as follows. First, the scalar β_k is chosen so that $|\beta_k| \in [\beta_{\min}, \beta_{\max}]$. Second, the scalar $\gamma_k \in (0, 1]$ is fixed using a backtracking strategy so that either the linesearch condition

$$\|F(x_k + p_k)\| \leq (1 - \rho(1 + \gamma_k))\|F_k\|, \quad (5.1)$$

holds or the linesearch condition

$$\|F(x_k + p_k)\| \leq (1 + \eta_k - \rho\gamma_k)\|F_k\|, \quad (5.2)$$

holds where $\rho \in (0, 1)$ is quite small [11, 34] and $\{\eta_k\}$ is a positive sequence satisfying (4.4). The linesearch conditions (5.1) and (5.2) are derivative-free; at each iteration, the first condition imposes a sufficient decrease in $\|F\|$ while the second condition allows for an increase of $\|F\|$ depending on the magnitude of η_k . The sufficient decrease condition (5.1) is crucial for establishing results on the convergence of $\{\|F_k\|\}$ to zero and can be accomplished for suitable values of

$\pm \gamma_k \beta_k F_k$ as long as $F_k^T J_k F_k \neq 0$. Trivially, (5.1) implies (5.2) and both imply the approximate norm descent condition (4.3).

The formal description of the SRAND method is reported in Algorithm 5.1 where we deliberately do not specify the form of the stepsize β_k . Termination of Step 2 is guaranteed by Theorem 4.3. The theoretical properties of SRAND are described in [34, Theorem 4.2 and Theorem 4.3] and are summarized in the following theorem.

THEOREM 5.1. *Let the positive sequence $\{\eta_k\}$ satisfy (4.4) and let $\{x_k\}$ be the sequence generated by the SRAND algorithm. Then*

1. *the sequence $\{x_k\}$ is convergent and consequently the sequence $\{\|F_k\|\}$ is convergent;*
2. *the sequence $\{\gamma_k \|F_k\|\}$ is convergent and such that $\lim_{k \rightarrow \infty} \gamma_k \|F_k\| = 0$;*
3. *if (5.1) is satisfied for infinitely many k , then $\lim_{k \rightarrow \infty} \|F_k\| = 0$.*

Algorithm 5.1: The SRAND algorithm

Given $x_0 \in \mathbb{R}^n$, $0 < \beta_{\min} < \beta_{\max}$, $\beta_0 \in [\beta_{\min}, \beta_{\max}]$, $\rho, \sigma \in (0, 1)$, a positive sequence $\{\eta_k\}$ satisfying (4.4).

If $\|F_0\| = 0$ stop.

For $k = 0, 1, 2, \dots$ do

1. Set $\gamma = 1$.
2. Repeat
 - 2.1 Set $p_- = -\gamma \beta_k F_k$ and $p_+ = \gamma \beta_k F_k$.
 - 2.2 If p_- satisfies (5.1), set $p_k = p_-$ and go to Step 3.
 - 2.3 If p_+ satisfies (5.1), set $p_k = p_+$ and go to Step 3.
 - 2.4 If p_- satisfies (5.2), set $p_k = p_-$ and go to Step 3.
 - 2.5 If p_+ satisfies (5.2), set $p_k = p_+$ and go to Step 3.
 - 2.6 Otherwise set $\gamma = \sigma \gamma$.
3. Set $\gamma_k = \gamma$, $x_{k+1} = x_k + p_k$.
4. If $\|F_{k+1}\| = 0$ stop.
5. Choose β_{k+1} such that $|\beta_{k+1}| \in [\beta_{\min}, \beta_{\max}]$.

The above result holds for any choice of the steplength β_k and Item 3. identifies one occurrence where the SRAND algorithm solves problem (1.1), i.e., $\{\|F_k\|\}$ converges to zero. We now complete the theoretical analysis of the SRAND algorithm by providing sufficient conditions that ensure that the sequence $\{\|F_k\|\}$ converges to zero.

We start by recalling a simple result.

LEMMA 5.2. *Suppose that Assumption 3.2 holds. Then for $p_k = \pm \gamma_k \beta_k F_k$, the following equality holds*

$$\|F_{k+1}\|^2 = \left(1 \pm 2\gamma_k \beta_k q((G_S)_k, F_k) \pm 2 \frac{\gamma_k \beta_k}{\|F_k\|^2} \int_0^1 (F(x_k + p_k) - F(x_k))^T J(x_k + tp_k) F_k dt \right) \|F_k\|^2. \quad (5.3)$$

Proof. Assume that $p_k = -\gamma_k \beta_k F_k$. Then,

$$\begin{aligned} \|F_{k+1}\|^2 &= \|F_k\|^2 + 2 \int_0^1 F(x_k + tp_k)^T J(x_k + tp_k) p_k dt \\ &= \|F_k\|^2 - 2\gamma_k \beta_k \int_0^1 F(x_k + tp_k)^T J(x_k + tp_k) F_k dt \\ &= \|F_k\|^2 - 2\gamma_k \beta_k \int_0^1 F(x_k + tp_k)^T J(x_k + tp_k) F_k dt \\ &\quad \pm 2\gamma_k \beta_k \int_0^1 F(x_k)^T J(x_k + tp_k) F_k dt \\ &= \|F_k\|^2 - 2\gamma_k \beta_k F_k^T G_k F_k - 2\gamma_k \beta_k \int_0^1 (F(x_k + p_k) - F(x_k))^T J(x_k + tp_k) F_k dt, \end{aligned}$$

that gives (5.3) using (3.1) and (2.13). The case $p_k = +\gamma_k \beta_k F_k$ is analogous. \square

Under specific assumptions on the Jacobian J , the following two theorems give conditions that ensure $F(x^*) = 0$ where x^* is the limit point of $\{x_k\}$: Theorem 5.3 concerns the cases when $J_S(x^*)$

is positive definite and when J is symmetric too, Theorem 5.4 regards the case when $J_S(x^*)$ is indefinite.

THEOREM 5.3. *Suppose that F is continuously differentiable on \mathbb{R}^n . Let the positive sequence $\{\eta_k\}$ satisfy (4.4) and let $\{x_k\}$ be the sequence generated by the SRAND algorithm. Moreover assume that $J_S(x^*)$ is positive definite at the limit point x^* of $\{x_k\}$. Letting $\sigma_{\max}(J(x^*))$ be the largest singular value of $J(x^*)$, if eventually the following conditions*

$$\nu \geq \beta_k > \frac{\rho}{(1+\epsilon)\sigma_{\max}(J(x^*))} \quad (5.4a) \quad \text{and} \quad \beta_k q((G_S)_k, F_k) > \frac{3}{2}\rho, \quad (5.4b)$$

hold with $\rho \in (0, 1)$ as in (5.1)-(5.2) and for some $\epsilon \in (0, 1)$ and $\nu > 0$, then $F(x^*) = 0$. If β_k is either $\beta_{k,1}$ or $\beta_{k,2}$, only condition (5.4b) has to be satisfied to get $F(x^*) = 0$. Moreover, for some $\omega_1, \omega_2 \in (0, 1)$, sufficient conditions for (5.4b) to hold are

1. if $\beta_k = \beta_{k,1}$ for k large enough:

$$\kappa(J_S(x^*)) < \frac{2\omega_1}{3\rho}; \quad (5.5)$$

2. if $\beta_k = \beta_{k,2}$ for k large enough:

$$\kappa(J_S(x^*)) < \omega_2 \sqrt{\frac{2}{3\rho}}; \quad (5.6)$$

3. if J is symmetric and β_k is either $\beta_{k,1}$ or $\beta_{k,2}$ for k large enough:

$$\kappa(J(x^*)) < \frac{2\omega_1}{3\rho}; \quad (5.7)$$

where $\kappa(\cdot)$ is the 2-norm condition number.

Proof. Since $J_S(x^*)$ is assumed to be positive definite, continuity implies that there exists a scalar $\xi > 0$ sufficiently small such that, for all $y \in \mathcal{B}(x^*, \xi) = \{x \in \mathbb{R}^n : \|x - x^*\| \leq \xi\}$, $J_S(y)$ is positive definite and

$$\lambda_{\min}(J_S(y)) \geq (1 - \epsilon)\lambda_{\min}(J_S(x^*)), \text{ and } \lambda_{\max}(J_S(y)) \leq (1 + \epsilon)\lambda_{\max}(J_S(x^*)), \quad (5.8)$$

with $\epsilon \in (0, 1)$. Moreover, the convergence of the sequence $\{x_k\}$ implies that $x_{k-1} + tp_{k-1}$ and $x_k + tp_k$ both belong to $\mathcal{B}(x^*, \xi)$ for large enough k and all $t \in [0, 1]$. As a consequence, reducing ξ if necessary, we deduce that, for k sufficiently large,

$$\begin{aligned} \min[\lambda_{\min}((G_S)_k), \lambda_{\min}((G_S)_{k-1})] &\geq (1 - \epsilon)\lambda_{\min}(J_S(x^*)), \\ \max[\lambda_{\max}((G_S)_k), \lambda_{\max}((G_S)_{k-1})] &\leq (1 + \epsilon)\lambda_{\max}(J_S(x^*)), \end{aligned}$$

and by (2.14),

$$q((G_S)_k, F_k) \in [\lambda_{\min}((G_S)_k), \lambda_{\max}((G_S)_k)] \subseteq [(1 - \epsilon)\lambda_{\min}(J_S(x^*)), (1 + \epsilon)\lambda_{\max}(J_S(x^*))]. \quad (5.9)$$

Finally, again by continuity, reducing $\xi > 0$ if necessary, for all $y \in \mathcal{B}(x^*, \xi)$ it holds

$$\sigma_{\max}(J(y)) \leq (1 + \epsilon)\sigma_{\max}(J(x^*)), \quad \sigma_{\max}(G_k) \leq (1 + \epsilon)\sigma_{\max}(J(x^*)). \quad (5.10)$$

Now, we consider (5.3) and $p_k = -\gamma_k \beta_k F_k$. From the Mean Value Theorem [11, Lemma 4.1.9], we have that

$$\left| \int_0^1 (F(x_k + tp_k) - F_k)^T J(x_k + tp_k) F_k dt \right| = \left| \int_0^1 \left(\int_0^1 J(x_k + \zeta tp_k) tp_k d\zeta \right) J(x_k + tp_k) F_k dt \right|,$$

$\zeta \in [0, 1]$. Again, for k sufficiently large, $x_k + \zeta tp_k \in \mathcal{B}(x^*, \xi)$ for $t, \zeta \in [0, 1]$. Thus, $p_k = -\gamma_k \beta_k F_k$ and (5.10) imply

$$\begin{aligned} \left| \int_0^1 (F(x_k + tp_k) - F_k)^T J(x_k + tp_k) F_k dt \right| &\leq \int_0^1 t \gamma_k \beta_k \max_{z \in \mathcal{B}(x^*, \xi)} \|J(z)\|^2 \|F_k\|^2 dt \\ &= \frac{1}{2} \gamma_k \beta_k \max_{z \in \mathcal{B}(x^*, \xi)} \sigma_{\max}(J(z))^2 \|F_k\|^2 \\ &\leq \frac{1}{2} \gamma_k \beta_k (1 + \epsilon)^2 \sigma_{\max}(J(x^*))^2 \|F_k\|^2. \end{aligned}$$

Combining this expression with (5.3), we have that for k sufficiently large

$$\begin{aligned} \|F_{k+1}\|^2 &\leq \left(1 - 2\gamma_k \beta_k q((G_S)_k, F_k) + 2 \frac{\gamma_k \beta_k}{\|F_k\|^2} \left| \int_0^1 (F(x_k + tp_k) - F(x_k))^T J(x_k + tp_k) F_k dt \right| \right) \|F_k\|^2 \\ &\leq (1 - 2\gamma_k \beta_k q((G_S)_k, F_k) + \gamma_k^2 \beta_k^2 (1 + \epsilon)^2 \sigma_{\max}(J(x^*))^2) \|F_k\|^2. \end{aligned} \quad (5.11)$$

Thus, for k sufficiently large, the linesearch condition (5.2) is satisfied if

$$1 - 2\gamma_k \beta_k q((G_S)_k, F_k) + \gamma_k^2 \beta_k^2 (1 + \epsilon)^2 \sigma_{\max}(J(x^*))^2 \leq (1 - \rho\gamma)^2,$$

which is equivalent to

$$\delta_2 \gamma^2 + 2\delta_1 \gamma \stackrel{\text{def}}{=} ((1 + \epsilon)^2 \sigma_{\max}(J(x^*))^2 \beta_k^2 - \rho^2) \gamma^2 + 2(\rho - \beta_k q((G_S)_k, F_k)) \gamma \leq 0. \quad (5.12)$$

Clearly (5.4a) implies that $(1 + \epsilon)^2 \sigma_{\max}(J(x^*))^2 \nu^2 \geq \delta_2 > 0$. Moreover, if eventually (5.4b) holds then $\delta_1 < 0$ and (5.12) is satisfied whenever $\gamma \leq \gamma^* = -2\delta_1/\delta_2$. Now, γ_* is uniformly bounded below since $-\delta_1 \geq \frac{1}{2}\rho$, i.e., $\gamma^* \geq \frac{\rho}{\delta_2} \geq \bar{\gamma} \stackrel{\text{def}}{=} \rho/((1 + \epsilon)^2 \sigma_{\max}(J(x^*))^2 \nu^2)$. Then, the mechanism of Step 3.6 of the SRAND algorithm guarantees that, for k sufficiently large, the loop in Step 2 terminates with $\gamma_k \geq \min\{1, \sigma\bar{\gamma}\}$, and $\bar{\gamma}$ independent of k . As a consequence, $\liminf_{k \rightarrow \infty} \gamma_k > 0$ and by Item 2. in Theorem 5.1 we have that $F(x^*) = 0$.

We now show that when β_k is either $\beta_{k,1}$ or $\beta_{k,2}$ for k sufficiently large, only condition (5.4b) has to be satisfied to get $F(x^*) = 0$.

Let $\beta_k = \beta_{k,1}$. Using Item (ii) in Lemma 3.4 and (3.6), we have that β_k is positive and satisfies

$$\frac{1}{(1 + \epsilon)\lambda_{\max}(J_S(x^*))} \leq \beta_k \leq \frac{1}{(1 - \epsilon)\lambda_{\min}(J_S(x^*))}. \quad (5.13)$$

By definition of J_S , $\|J_S(x^*)\| \leq \|J(x^*)\|$, hence $\lambda_{\max}(J_S(x^*)) \leq \sigma_{\max}(J(x^*))$. Therefore (5.4a) is satisfied being $\rho \in (0, 1)$ and setting $\nu = 1/((1 - \epsilon)\lambda_{\min}(J_S(x^*)))$.

Let $\beta_k = \beta_{k,2}$. Since $\beta_{k,2} \leq \beta_{k,1}$, the upper bound in (5.4a) is guaranteed from the discussion above. Moreover from (5.11) and again from $\beta_{k,2} \leq \beta_{k,1}$, the linesearch condition (5.2) is satisfied if

$$\delta_2 \gamma^2 + 2\delta_1 \gamma \stackrel{\text{def}}{=} ((1 + \epsilon)^2 \sigma_{\max}(J(x^*))^2 \beta_{1,k}^2 - \rho^2) \gamma^2 + 2(\rho - \beta_{2,k} q((G_S)_k, F_k)) \gamma \leq 0. \quad (5.14)$$

Following the previous considerations on $\beta_{k,1}$, δ_2 is positive. Further, using (5.4b) and repeating the arguments above on the scalar γ satisfying (5.14), the loop in Step 2 terminates with $\gamma_k \geq \min\{1, \sigma\bar{\gamma}\}$, and $\bar{\gamma}$ independent of k .

To conclude, as for Item 1., if $\beta_{k,1}$ is used eventually then (3.6) and (5.9) give $\beta_k q((G_S)_k, F_k) \geq \frac{\omega_1}{\kappa(J_S(x^*))}$ with $\omega_1 = \frac{1 - \epsilon}{1 + \epsilon}$ and trivially (5.5) implies (5.4b) for all k sufficiently large.

As for Item 2., if $\beta_{k,2}$ is used eventually then (3.7), (5.10) and (5.9) give $\beta_k q((G_S)_k, F_k) \geq \frac{\omega_2^2}{\kappa(J_S(x^*))^2}$ with $\omega_2 = \frac{(1 - \epsilon)\|J_S(x^*)\|}{(1 + \epsilon)\|J(x^*)\|}$, and (5.6) implies (5.4b) for all k sufficiently large.

Concerning Item 3., (5.4b) reads $\beta_k q(G_k, F_k) > \frac{3}{2}\rho$, and by Lemma 3.4 $\beta_{k,1}$ and $\beta_{k,2}$ are positive and

$$\beta_{k,1} \geq \beta_{k,2} \geq \frac{1}{\sigma_{\max}(G_{k-1})} \geq \frac{1}{(1 + \epsilon)\sigma_{\max}(J(x^*))}.$$

Thus, by (5.9) it follows $\beta_k q(G_k, F_k) \geq \frac{\omega_1}{\kappa(J(x^*))}$ and trivially (5.7) implies (5.4b) for all k sufficiently large. \square

We remark that analogous conditions to (5.4) can be derived for the case when $J_S(x^*)$ is negative definite.

THEOREM 5.4. *Suppose that F is continuously differentiable on \mathbb{R}^n . Let the positive sequence $\{\eta_k\}$ satisfy (4.4) and let $\{x_k\}$ be the sequence generated by the SRAND algorithm. Moreover assume that $J_S(x^*)$ is indefinite and $J(x^*)$ is nonsingular at the limit point x^* of $\{x_k\}$. If eventually the following conditions*

$$\nu \geq |\beta_k| > \frac{\rho}{(1+\epsilon)\sigma_{\max}(J(x^*))} \quad (5.15a) \quad \text{and} \quad |\beta_k q((G_S)_k, F_k)| > \frac{3}{2}\rho, \quad (5.15b)$$

hold with $\rho \in (0, 1)$ as in (5.1)-(5.2) and for some $\epsilon \in (0, 1)$ and $\nu > 0$, then $F(x^*) = 0$.

Proof. We observe that for k sufficiently large, the inequalities (5.8)-(5.9) hold for some $\epsilon \in (0, 1)$. Moreover, considering $p_k = \pm \gamma_k \beta_k F_k$ and proceeding as in the proof of Theorem 5.3, we get that for k sufficiently large the following inequality holds

$$\|F_{k+1}\|^2 \leq (1 \pm 2\gamma_k \beta_k q((G_S)_k, F_k) + \gamma_k^2 \beta_k^2 (1+\epsilon)^2 \sigma_{\max}(J(x^*))^2) \|F_k\|^2.$$

Therefore the linesearch condition (5.2) is satisfied if

$$\delta_2 \gamma^2 + 2\delta_1 \gamma \stackrel{\text{def}}{=} ((1+\epsilon)^2 \sigma_{\max}(J(x^*))^2 \beta_k^2 - \rho^2) \gamma^2 + 2(\rho \pm \beta_k q((G_S)_k, F_k)) \gamma \leq 0. \quad (5.16)$$

Clearly (5.15a) implies that $(1+\epsilon)^2 \sigma_{\max}(J(x^*))^2 \nu^2 \geq \delta_2 > 0$.

We now show that (5.15b) implies $F(x^*) = 0$ as in the proof of Theorem 5.3. Let us analyse the case $\beta_k q((G_S)_k, F_k) < 0$ and consider the step $p_k = \gamma_k \beta_k F_k$. Then condition (5.15b) means that $-\beta_k q((G_S)_k, F_k) \geq \frac{3}{2}\rho$, that is $\delta_1 = \rho + \beta_k q((G_S)_k, F_k) < -\frac{1}{2}\rho < 0$. The case $\beta_k q((G_S)_k, F_k) > 0$ is analogous considering the step $p_k = -\gamma_k \beta_k F_k$. Now, repeating the arguments in Theorem 5.3 we conclude that $\liminf_{k \rightarrow \infty} \gamma_k > 0$. \square

6. Numerical experiments. In view of our theoretical analysis and guidelines on the steplength selection, we attempt to tailor Barzilai and Borwein rules for unconstrained optimization to the framework of spectral residual methods for nonlinear systems. In this section we discuss several steplength rules for spectral residual methods and analyze their practical performance using the SRAND algorithm described in Algorithm 5.1. Our test set consists of sequences of nonlinear systems arising in the solution of rail-wheel contact models and is described in details in Section 6.2.

SRAND was implemented in Matlab (MATLAB R2019b) and the experiments were carried out on a Intel Core i7-9700K CPU @ 3.60GHz x 8, 16 GB RAM, 64-bit.

6.1. Steplength rules. We now present six rules for the choice of the steplength in spectral residual methods that will be used in our experiments. Besides the straightforward choice of one of the two steplengths $\beta_{k,1}$, $\beta_{k,2}$, along all iterations, we consider adaptive strategies that suitably combine them and parallel those used for quadratic and nonlinear optimization problems. Below, given a scalar β , $T(\beta)$ is the thresholding rule which projects $|\beta|$ onto the interval $I_\beta \stackrel{\text{def}}{=} [\beta_{\min}, \beta_{\max}]$, i.e.,

$$T(\beta) = \min \left\{ \beta_{\max}, \max \left\{ \beta_{\min}, |\beta| \right\} \right\}. \quad (6.1)$$

BB1 rule. By [21, 25, 27, 34], at each iteration let

$$\beta_k = \begin{cases} \beta_{k,1} & \text{if } |\beta_{k,1}| \in I_\beta \\ T(\beta_{k,1}) & \text{otherwise} \end{cases} \quad (6.2)$$

BB2 rule. At each iteration let

$$\beta_k = \begin{cases} \beta_{k,2} & \text{if } |\beta_{k,2}| \in I_\beta \\ T(\beta_{k,2}) & \text{otherwise} \end{cases} \quad (6.3)$$

ALT rule. Following [8, 21], at each iteration let us alternate between $\beta_{k,1}$ and $\beta_{k,2}$:

$$\beta_k^{\text{ALT}} = \begin{cases} \beta_{k,1} & \text{for } k \text{ odd} \\ \beta_{k,2} & \text{otherwise} \end{cases} \quad (6.4)$$

$$\beta_k = \begin{cases} \beta_k^{\text{ALT}} & \text{if } |\beta_k^{\text{ALT}}| \in I_\beta \\ \beta_{k,1} & \text{if } k \text{ even, } |\beta_{k,1}| \in I_\beta, |\beta_{k,2}| \notin I_\beta \\ \beta_{k,2} & \text{if } k \text{ odd, } |\beta_{k,2}| \in I_\beta, |\beta_{k,1}| \notin I_\beta \\ T(\beta_k^{\text{ALT}}) & \text{otherwise} \end{cases} \quad (6.5)$$

ABB rule. Following [45] and ABB rule in [16], we define the Adaptive Barzilai-Borwein (ABB) rule as follows. Given $\tau \in (0, 1)$, let

$$\beta_k^{\text{ABB}}(\xi_1, \xi_2) = \begin{cases} \xi_2 & \text{if } \frac{\xi_2}{\xi_1} < \tau \\ \xi_1 & \text{otherwise} \end{cases} \quad (6.6)$$

for some given ξ_1, ξ_2 . Then

$$\beta_k = \begin{cases} \beta_k^{\text{ABB}}(\beta_{k,1}, \beta_{k,2}) & \text{if } |\beta_{k,1}|, |\beta_{k,2}| \in I_\beta \\ \beta_{k,1} & \text{if } |\beta_{k,1}| \in I_\beta, |\beta_{k,2}| \notin I_\beta \\ \beta_{k,2} & \text{if } |\beta_{k,2}| \in I_\beta, |\beta_{k,1}| \notin I_\beta \\ \beta_k^{\text{ABB}}(T(\beta_{k,1}), T(\beta_{k,2})) & \text{otherwise} \end{cases} \quad (6.7)$$

Observe that a large value of τ promotes the use of $\beta_{k,2}$ with respect to $\beta_{k,1}$. The rule allows to switch between the steplengths $\beta_{k,1}$ and $\beta_{k,2}$ and was originally motivated by the behaviour of the Barzilai and Borwein method applied to convex and quadratic minimization problem (see [16, 45] and our discussion below Lemma 4.1).

ABBm rule. This rule elaborates the ABBminmin rule given in [16], taking into account that $\beta_{k,2}$ may be negative along iterations. Let m be a nonnegative integer, and

$$\tilde{\beta}_{k,2} = \begin{cases} \beta_{k,2} & \text{if } |\beta_{k,2}| \in I_\beta \\ T(\beta_{k,2}) & \text{otherwise} \end{cases} \quad (6.8)$$

$$j^* = \operatorname{argmin}\{|\tilde{\beta}_{j,2}| : j = \max\{1, k - m\}, \dots, k\}.$$

Given $\tau \in (0, 1)$, we fix β_k as follows

$$\beta_k^{\text{ABBm}}(\xi_1, \xi_2) = \begin{cases} \tilde{\beta}_{j^*,2} & \text{if } \frac{\xi_2}{\xi_1} < \tau \\ \xi_1 & \text{otherwise} \end{cases} \quad (6.9)$$

$$\beta_k = \begin{cases} \beta_k^{\text{ABBm}}(\beta_{k,1}, \beta_{k,2}) & \text{if } |\beta_{k,1}|, |\beta_{k,2}| \in I_\beta \\ \beta_{k,1} & \text{if } |\beta_{k,1}| \in I_\beta, |\beta_{k,2}| \notin I_\beta \\ \beta_{k,2} & \text{if } |\beta_{k,2}| \in I_\beta, |\beta_{k,1}| \notin I_\beta \\ \beta_k^{\text{ABBm}}(T(\beta_{k,1}), T(\beta_{k,2})) & \text{otherwise} \end{cases} \quad (6.10)$$

Again, a large value of τ promotes the use of a step from BB2 rule instead of $\beta_{k,1}$. In case $|\beta_{k,1}|, |\beta_{k,2}| \in I_\beta$ and $\frac{\beta_{k,2}}{\beta_{k,1}} < \tau$, the smallest absolute value $\tilde{\beta}_{j^*,2}$ over the last $m+1$ iterations is selected; taking into account that $\tilde{\beta}_{j,2}$ for $j = \max\{1, k-m\}, \dots, k$ can be negative, the rationale for selecting $\tilde{\beta}_{j^*,2}$ in (6.9) is to mitigate the nonmonotone behavior of the objective function [16]. Consequently, smaller steplengths are expected using the ABBm rule than using the ABB rule.

DABBm rule. Following [4, 6], a dynamic threshold $\tau_k \in (0, 1)$ can be used in place of the prefixed threshold τ in (6.9). Given $\tilde{\beta}_{k,2}$ and j^* in (6.8), we propose the rule defined as

$$\beta_k^{\text{DABBm}}(\xi_1, \xi_2) = \begin{cases} \tilde{\beta}_{j^*,2} & \text{if } \frac{\xi_2}{\xi_1} < \tau_k \\ \xi_1 & \text{otherwise} \end{cases} \quad (6.11)$$

$$\beta_k = \begin{cases} \beta_k^{\text{DABBm}}(\beta_{k,1}, \beta_{k,2}) & \text{if } |\beta_{k,1}|, |\beta_{k,2}| \in I_\beta \\ \beta_{k,1} & \text{if } |\beta_{k,1}| \in I_\beta, |\beta_{k,2}| \notin I_\beta \\ \beta_{k,2} & \text{if } |\beta_{k,2}| \in I_\beta, |\beta_{k,1}| \notin I_\beta \\ \beta_k^{\text{DABBm}}(T(\beta_{k,1}), T(\beta_{k,2})) & \text{otherwise} \end{cases} \quad (6.12)$$

with the dynamic threshold set as

$$\tau_k = \min \left\{ \tau, \|F_k\|^{1/(2+b_t^2)} \right\}, \quad (6.13)$$

$$b_t = \max\{b_j : j = \max\{1, k-w\}, \dots, k\}. \quad (6.14)$$

Here $\tau \in (0, 1)$ is an upper bound on the value of τ_k , w is a nonnegative integer and b_j denotes the number of backtracks performed at iteration j (see Step 2 of Algorithm 5.1). If $\|F_k\|$ is getting small and the number of performed backtracks in the last $w+1$ iterations is small, then (6.13) promotes the use of steplength from BB1 rule, i.e., larger steplengths which can speed convergence to a zero of F . On the other hand, when the number of backtracks performed along previous iterations is large and τ is large, the use of the smaller steplength from BB2 rule is encouraged.

We conclude the discussion on steplength selection, noting that conditions (5.4) and (5.15) for the convergence of $\{x_k\}$ to a solution of problem (1.1) apply to all our rules.

The rules and parameters used in our experiments are summarized in Table 6.1.

Rule	β_k
BB1	β_k in (6.2)
BB2	β_k in (6.3)
ALT	β_k in (6.4), (6.5)
ABB01	β_k in (6.6), (6.7) with $\tau = 0.1$
ABB08	β_k in (6.6), (6.7) with $\tau = 0.8$
ABBm01	β_k in (6.8)-(6.10) with $\tau = 0.1, m = 5$
ABBm08	β_k in (6.8)-(6.10) with $\tau = 0.8, m = 5$
DABBm	β_k in (6.8), (6.11)-(6.14) with $\tau = 0.8, m = 5, w = 20$

TABLE 6.1
Steplength's rules in SRAND implementation.

6.2. Problem set: nonlinear systems arising from rolling contact models. Rolling contact is a fundamental issue in mechanical engineering and plays a central role in many important applications such as rolling bearings and wheel-rail interaction [23, 24]. In order to perform simulations of complex mechanical systems with a good tradeoff between accuracy and efficiency, three working hypotheses are usually made in modelling rolling contact: non-conformal contact,

i.e., the typical dimensions of the contact area are negligible if compared to the curvature radii of the contact body surfaces; planar contact, i.e., the contact area is contained in a plane; half-space contact, i.e., locally, the contact bodies are viewed as three-dimensional half-spaces [23, 24]. In this framework, we focus on the Kalker’s rolling contact model which represents a relevant and general model in contact mechanics.

The solution of Kalker’s rolling contact model can be performed using different approaches. The approach in [42, 43] calls for the solution of constrained optimization problems while the so-called CONTACT algorithm [24] gives rise to sequences of nonlinear systems. Our problem set derives from the application of CONTACT algorithm; here we describe in which phase of the Kalker’s model solution they arise and give some of their features. We refer to Appendix A for a sketch of Kalker’s model, its discretization, and the Kalker’s CONTACT algorithm.

Kalker’s CONTACT algorithm determines the normal pressure, the tangential pressure, the contact area, the adhesion area and the sliding area in the contact between two elastic bodies and relies on the elastic decoupling between the normal contact problem and the tangential contact problem. Such problems are solved separately; first the normal problem is solved via the so-called NORM algorithm, second the tangential problem is solved via the so-called TANG algorithm. Algorithms NORM and TANG are expected to identify the elements in the contact area and in the adhesion-sliding areas, respectively. These algorithms are applied sequentially and repeatedly until the values of the computed pressures undergo a sufficiently small change that suggests their reliable approximation; in general, a few repetitions of NORM and TANG algorithms are required. Each repetition of NORM algorithm calls for the solution of a sequence of linear systems while each repetition of TANG algorithm calls for the solution of a sequence of linear and nonlinear systems. Computationally, the major bottleneck is the numerical solution of the sequence of nonlinear systems generated in the TANG phase. Importantly, each CONTACT iteration requires few repetitions of TANG algorithm but the CONTACT algorithm is performed for several time instances*.

Our tests were made on wheel-rail contact in railway systems. The benchmark vehicle is a driverless subway vehicle, designed by Hitachi Rail on MLA platform (Light Automatic Metro). The vehicle is a fixed-length train composed of four car bodies and five bogies (four motorized and one, the third, trailer), see Figure 6.1. The multibody model has been realized in the Simpack Rail environment [39]. We considered a train route of length 400m including a typical railway curved track characterized by three significant parts: two straight lines (from 0m to 70m and from 233m to 400m), the curve (from 116m to 186m) and two cycloids (from 70m to 116m and from 186m to 233m) which smoothly connect the straight lines and the curve in terms of curvature radius. The radius of the curve is 500m. In this analysis, we focused on the contact between the first vehicle wheel and the rail; since the vehicle length is equal to 45.7m, at the beginning of the dynamic simulation the considered wheel starts in the position 45.7m along the track. We performed a simulation in an interval of 10 seconds using 500 time steps, which amounts to 500 calls to CONTACT algorithm, for train speeds with magnitude v taking the values: $v = 10 \text{ m/s}$ and $v = 16 \text{ m/s}$. Accordingly, during the whole simulation the considered wheel travels along the track a distance equal to 100m and 160m, respectively. The traveling velocities considered give a realistic lateral acceleration along the curve according to the current regulation in force in the railway field.

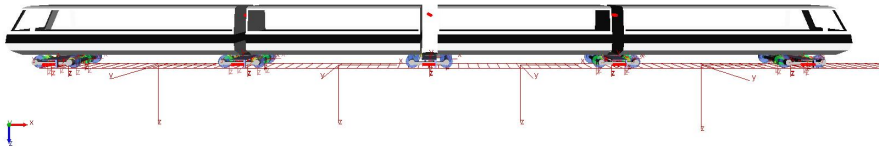


FIG. 6.1. *Multibody model of the benchmark vehicle.*

*In Appendix A see: (A.1) for the form of normal contact problem and tangential contact problem, (A.5) for the form of the nonlinear systems to be solved, Figure A.2 for the flow of Kalker’s CONTACT algorithm.

Two sets of experiments were performed[†]. First, we numerically investigated eight variants of SRAND obtained varying the rules in Table 6.1 on a large number of sequences of nonlinear systems arising from wheel-rail contact in railway systems. Second, we compared the best performing SRAND variant with a standard Newton trust-region method when they are embedded in the CONTACT algorithm.

The set of test problems used in the first part of the experiments was generated implementing the CONTACT algorithm in Matlab and using a standard trust-region Newton method[‡] for solving the arising nonlinear systems. Afterwards, a representative subset of the nonlinear systems was selected to form our problem set. Specifically, six sequences of nonlinear systems generated by the CONTACT algorithm and corresponding to six consecutive time instances for each track section (straight line, cycloid and curve) and for each velocity were selected. Such sequences are representative of the systems arising throughout the whole simulation and allow a fair analysis of SRAND on nonlinear systems from a real application. Table 6.2 summarizes the features of the sequences: magnitude of the train velocity v , section of the route, time instances, number of nonlinear systems in the sequence, dimension n of the systems (proportional to the number of mesh nodes in the potential contact area). A typical feature of the contact model is that n increases as the velocity increases and when the train curves along the route (i.e., the track curvature increases). The total number of systems associated to $v = 10$ m/s and $v = 16$ m/s is 121 and 153 respectively.

$v(m/s)$	Track Section	Time Instances	Number of Systems	n
10	Straight line	100-105	10	156
	Cycloid	300-305	56	897
	Curve	450-455	55	1394
16	Straight line	50-55	8	156
	Cycloid	150-155	63	1120
	Curve	350-355	82	1394

TABLE 6.2
Sequences of nonlinear systems forming the first problem set.

6.3. Numerical results. In this section first we discuss the solution of the sequences of nonlinear systems in Table 6.2 using different stepsize rules within the SRAND algorithm, second we analyze the use of SRAND in the CONTACT algorithm instead of a standard Newton trust-region approach.

SRAND algorithm was implemented as described in Section 6.1 and with parameters

$$\beta_{\min} = 10^{-10}, \quad \beta_{\max} = 10^{10}, \quad \rho = 10^{-4}, \quad \sigma = 0.5, \quad \eta_k = 0.99^k(100 + \|F_0\|^2) \quad \forall k \geq 0,$$

see [34]. The null vector $x_0 = 0$ was chosen as initial guess. A maximum number of iterations and F -evaluations equal to 10^5 was imposed and a maximum number of backtracks equal to 40 was allowed at each iteration. The procedure was declared successful when

$$\|F_k\| \leq 10^{-6}. \quad (6.15)$$

A failure was declared either because the assigned maximum number of iterations or F -evaluations or backtracks is reached, or because $\|F\|$ was not reduced for 50 consecutive iterations.

We now compare the performance of eight variants of the SRAND method in the solution of the sequences of nonlinear systems in Table 6.2. Each variant is obtained selecting one of the stepsize updating rules reported in Table 6.1. Further, in light of the theoretical investigation presented in this work, we analyze in details the results obtained with BB1 and BB2 rule and support the use of rules that switch between the two steplengths.

[†]The data that support the findings of this study are available from the corresponding author upon reasonable request.

[‡]The code in [33] was applied using the default setting and dropping bound constraints on the unknown.

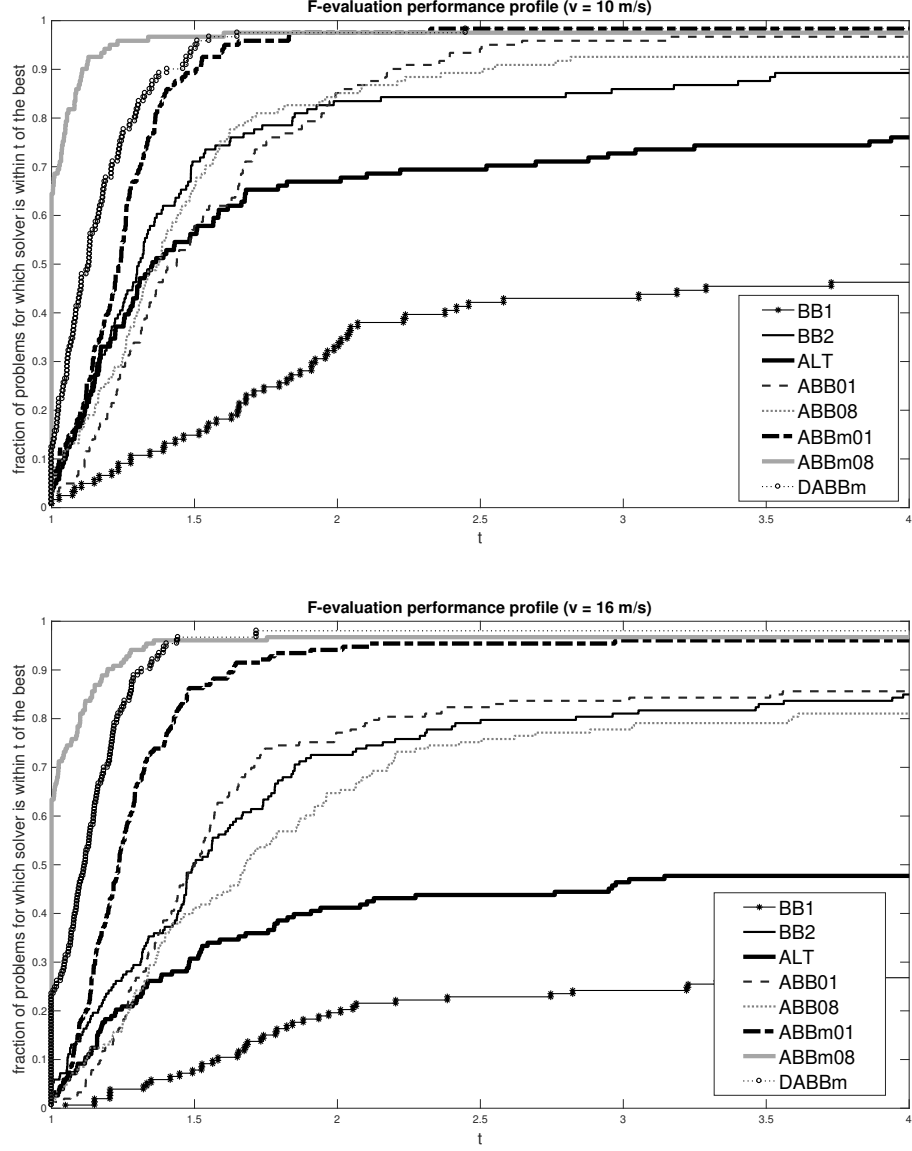


FIG. 6.2. *F*-evaluation performance profiles of SRAND method. Upper: $v = 10 \text{ m/s}$, Lower: $v = 16 \text{ m/s}$.

Figure 6.2 shows the performance profiles [13] in terms of *F*-evaluations employed by the SRAND variants for solving the sequence of systems generated both with $v = 10 \text{ m/s}$ (121 systems) (upper) and with $v = 16 \text{ m/s}$ (153 systems) (lower) and highlights that the choice of the steplength is crucial for both efficiency and robustness. The complete results are reported in Appendix B. We start observing that BB2 rule outperformed BB1 rule; in fact the latter shows the worst behaviour both in terms of efficiency and in terms of number of systems solved. Alternating $\beta_{k,1}$ and $\beta_{k,2}$ in ALT rule without taking into account the magnitude of the two scalars improves performance over BB1 rule but is not competitive with BB2 rule. On the other hand, the variants of SRAND using adaptive strategies are the most robust, i.e., they solve the largest number of problems, and efficient. Specifically, comparing ABB, ABBm and DABBm rules, the most effective steplength selections are ABBm and DABBm. Using ABBm01 rule, 98.3% (2 failures) and 96.1% (6 failures) out of the total number of systems were solved successfully for $v = 10 \text{ m/s}$ and $v = 16 \text{ m/s}$

respectively; using ABBm08 rule, 98.3% (2 failures) and 96.7% (5 failures) of the total number of systems were solved successfully with $v = 10 \text{ m/s}$ and $v = 16 \text{ m/s}$ respectively; using the dynamic selection DABBm, the largest number of systems was solved successfully, i.e., 99.2% (1 failure) and 98% (3 failures) out the total number of systems with $v = 10 \text{ m/s}$ and $v = 16 \text{ m/s}$ respectively. Overall, ABBm08 rule gives rise to the most efficient algorithm for both velocity values and the profile related to BB2 rule is within a factor 2 of it in roughly the 80% and the 70% of the runs for $v = 10 \text{ m/s}$ and $v = 16 \text{ m/s}$, respectively.

Let us now focus on the performance SRAND coupled with BB1 and BB2 rules. As a representative run of our numerical experience reported in Appendix B, we consider the nonlinear system arising with $v = 16 \text{ m/s}$, at time $t = 150$, iteration 2 of the CONTACT algorithm and iteration 2 of the TANG algorithm (system 150.2.2 in Table B.5). In the upper part of Figure 6.3 we display

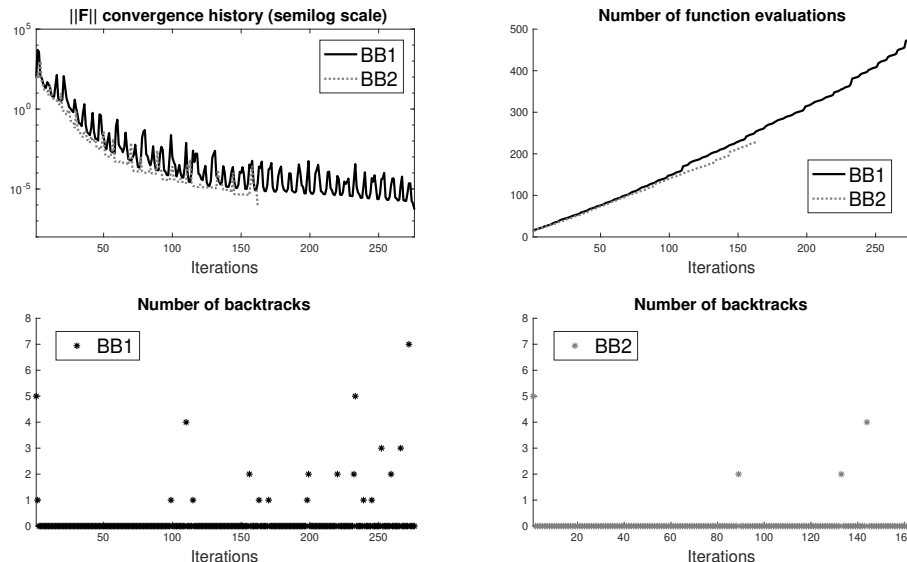


FIG. 6.3. SRAND with BB1 rule vs SRAND with BB2 rule on a single nonlinear system.

$\|F\|$ along iterations and the number of F -evaluations performed. We note that using the step-size $\beta_{k,1}$ causes a highly nonmonotone behavior of $\|F\|$ and such behaviour is not productive for convergence; using BB1 rule 276 iterations and 476 F -evaluations are performed while using BB2 rule 163 iterations and 228 F -evaluations are required. The distinguishing feature of these runs is the high number of backtracks performed using $\beta_{k,1}$ at some iterations, as reported at the bottom part of the figure where the number of backtracks versus iterations is reported for both SRAND variants. This behaviour is in accordance with the analysis in Section 4.1. We know that $\beta_{k,1}$ can be arbitrarily larger than $\beta_{k,2}$ in the indefinite case, hence if $\beta_{k,1}$ is taken as the initial steplength, a large number of backtracks may be necessary to enforce (5.1)-(5.2). Such observation supports the use of $\beta_{k,2}$; the benefit from using shorter steps is further shown by the performance of ABBm over ABB, the former tends to take shorter steps than the latter by exploiting the iteration history and results to be more effective.

We conclude our experimental analysis using a spectral residual method in the CONTACT algorithm. To this purpose, we compare two implementations of CONTACT algorithm which differ only in the nonlinear solver for the nonlinear systems arising in the TANG algorithm. The first implementation (CONTACT-NTR) uses a standard Newton trust-region method and the second one (CONTACT-DABBm) uses SRAND with DABBm which turned out to be the more robust SRAND version in the analysis above (see Figure 6.2). As a standard Newton trust-region method, we used the Matlab code proposed in [33]; default parameters were used and bound constraints on

the unknown were dropped using the setting indicated in the code. The Jacobian matrix of F was approximated by finite differences.

As a preliminary issue, we observe that the Jacobian matrices of F are dense through the iterations; thus they cannot be formed at a low computational cost by finite difference procedures for sparse matrices [7]. We have also observed in the experiments that the Jacobian matrices are nonsymmetric, do not have dominant diagonals and they are not close to diagonal matrices. For example, let us consider the Jacobian matrix of the system corresponding to speed $v = 16 \text{ m/s}$, curve track section, instant $t = 355$, iteration 2 of the CONTACT and iteration 4 of the TANG algorithm (355.2_4 in Table B.6). It has dimension 292×292 and, evaluated at the final iterate computed using ABBm08 rule, 96.18% of its elements are nonzero. The structure of the Jacobian can be observed in Figure 6.4 where the absolute values of its elements are plotted in a logarithmic scale (the surface of the full matrix on the left and a plot of the row 146 on the right). This structure is observed along all the iterations of the nonlinear system solvers and is common to all sequences generated by the CONTACT algorithm.

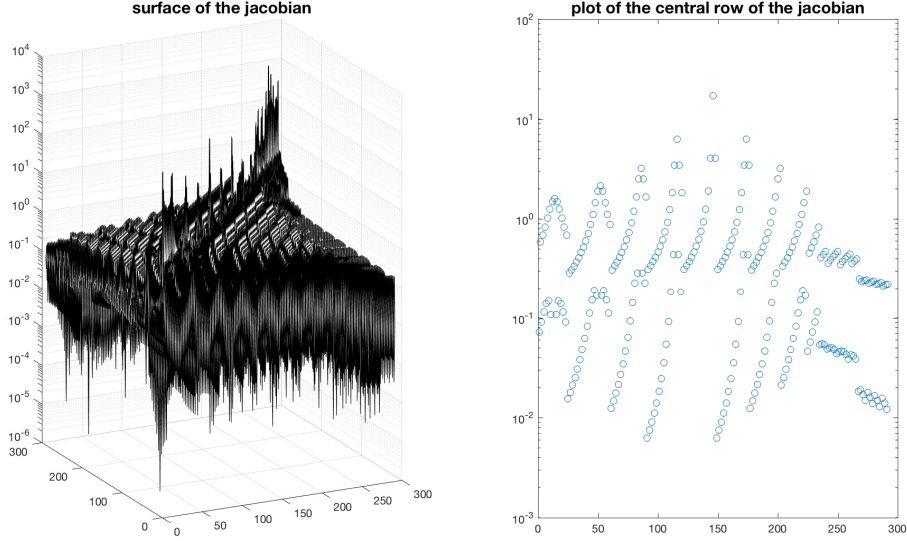


FIG. 6.4. *Jacobian matrix: surface of the full matrix and plot of the central row (base 10 logarithm of the absolute values).*

In our implementation, CONTACT algorithm terminated when the relative error between two successive values of the computed pressures dropped below 10^{-4} or a maximum of 20 alternating cycles between NORM and TANG was reached. Both nonlinear solvers were run until the stopping rule (6.15) is met. We ran CONTACT-NTR and CONTACT-DABBm over the whole track for both velocities, that is we considered the whole sequence of 500 time steps. CONTACT-NTR generated 3759 and 5353 nonlinear systems for $v = 10 \text{ m/s}$ and $v = 16 \text{ m/s}$, respectively and CONTACT-DABBm generated 4496 and 5494 nonlinear systems for the two velocities.

As a first remark, both procedures successfully solved the contact model described above and were reliable and accurate in the numerical simulation of wheel-rail interaction. Secondly, the use of the spectral residual method yields a gain in terms of time with respect to the use of a standard Newton method where finite difference approximation of Jacobian matrices is employed; this feature derives from the fact that spectral residual method is derivative-free and does not ask for the solution of linear systems. Figures 6.5 and 6.6 show the comparison of the two CONTACT implementations in terms of number of F -evaluations (excluding those needed to approximate the Jacobian matrices) and execution elapsed time. From the plots we observe that CONTACT-DABBm takes a larger number of F -evaluations than CONTACT-NTR but it is faster. Over the

whole time interval, CONTACT-DABBm employs 1 hour, 19 mins and 2 hours, 28 mins to solve the generated nonlinear systems with $v = 10 \text{ m/s}$ and $v = 16 \text{ m/s}$, while CONTACT-NTR takes 7 hours and 49 mins and 12 hours and 41 mins, respectively.

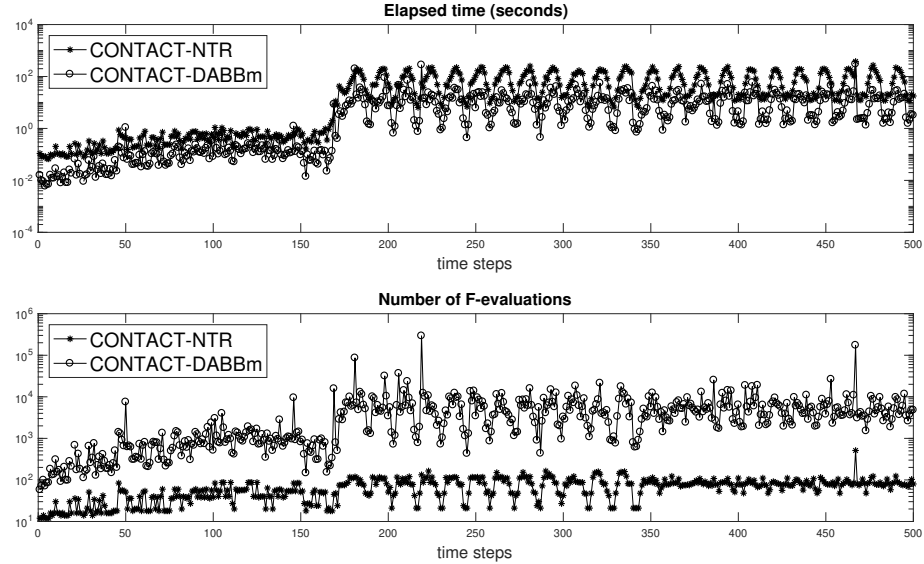


FIG. 6.5. Comparison between CONTACT-DABBm and CONTACT-NTR, $v = 10 \text{ m/s}$: number of F -evaluations and elapsed time in seconds (logarithmic scale).

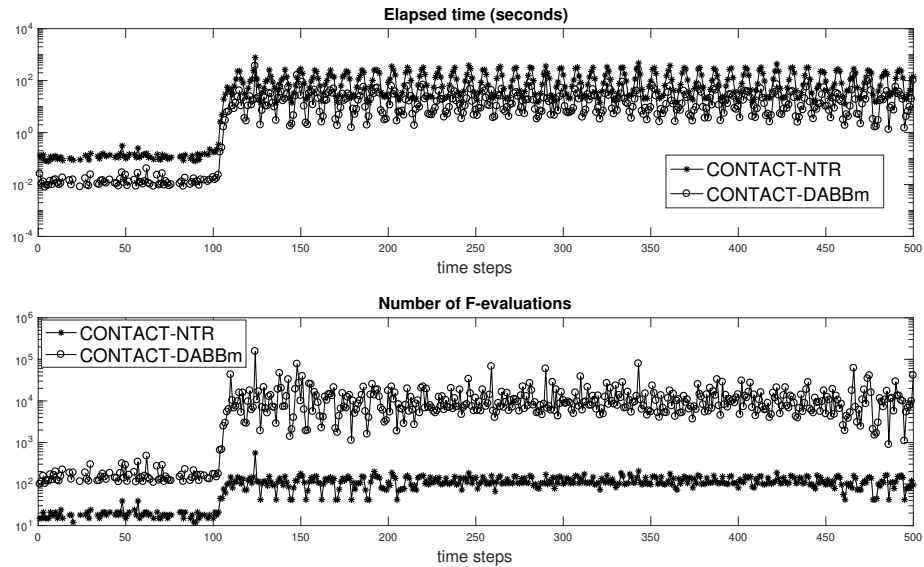


FIG. 6.6. Comparison between CONTACT-DABBm and CONTACT-NTR, $v = 16 \text{ m/s}$: number of F -evaluations and elapsed time in seconds (logarithmic scale).

7. Conclusions.

The numerical behaviour of spectral residual methods for nonlinear systems

is heavily affected by the choice of the steplengths. Although most of the works on this subject make use of the stepsize $\beta_{k,1}$, known results on spectral gradient methods for unconstrained optimization suggest that a suitable combination of the stepsizes $\beta_{k,1}$ and $\beta_{k,2}$ could be beneficial. In this work we analyzed the stepsizes $\beta_{k,1}$ and $\beta_{k,2}$ with respect to the spectrum of average matrices depending on the Jacobian of F and discuss guidelines for their selection. Moreover, we present several practical rules for choosing the steplengths and show the performance of the resulting procedures on sequences of nonlinear systems arising in the solution of a contact wheel-rail model.

Acknowledgments. INdAM-GNCS partially supported the second, the third and the fourth author under Progetti di Ricerca 2019 and 2020.

Declarations

Conflict of interest. The authors declare that they have no conflict of interest.

Funding. Open access funding provided by Università di Bologna within the CRUI-CARE Agreement.

Appendix A. Kalker’s contact model and CONTACT algorithm.

We give an overview of the model and algorithm used to generate our set of nonlinear systems. Let bold letters represent vectors, the subscript T denote a vector with components in the tangential x - y contact plane, the subscript N denote the component of a vector in the normal z contact direction. The contact problem between two elastic bodies [23,24] determines the contact region C inside the potential contact area A_c (usually the interpenetration area between the wheel and rail contact surfaces), its subdivision into adhesion area H and slip area S , and the tangential \mathbf{p}_T and normal p_N pressures such that the following contact conditions are satisfied:

$$\begin{array}{ll}
 \text{normal problem} & \begin{array}{l} \text{in contact } C : \quad e = 0, \quad p_N \geq 0 \\ \text{in exterior } E : \quad p_N = 0, \quad e > 0 \\ C \cup E = A_c, \quad C \cap E = \emptyset \end{array} \\
 \text{tangential problem} & \begin{array}{l} \text{in adhesion } H : \quad \|\mathbf{s}_T\| = 0, \quad \|\mathbf{p}_T\| \leq g \\ \text{in slip } S : \quad \|\mathbf{s}_T\| \neq 0, \quad \mathbf{p}_T = -g \mathbf{s}_T / \|\mathbf{s}_T\| \\ S \cup H = C, \quad S \cap H = \emptyset \end{array}
 \end{array} \tag{A.1}$$

Above, e is the deformed distance between the two bodies and, by definition, it holds $e = 0$ and $p_N \geq 0$ in C . Referring to Figure A.1, the region E where $e > 0$ is called the exterior area and $p_N = 0$ therein. The potential contact area is such that $A_c = C \cup E$. The contact area C is divided into the area of adhesion H where the tangential component \mathbf{s}_T of the slip vanishes, and the area S of slip where \mathbf{s}_T is nonzero. The slip \mathbf{s}_T is the difference between the velocities of two homologous points belonging to deformed wheel and rail surfaces inside the contact area and is a function of the pressures \mathbf{p}_T and p_N , g is the traction bound (Coulomb friction model [23,24]). Overall, the first three equations in (A.1) model the normal contact problem (computation of p_N and of the shapes of the regions C and E), whereas the last three equations describe the tangential contact problem (computation of \mathbf{p}_T , of local slidings \mathbf{s}_T and of the shapes of the regions H and S).

Let us consider the discretization of (A.1). Assuming that the contact patch is entirely contained in a plane, the region within which the potential contact area A_c can be located is easily discretized through a planar quadrilateral mesh, see Figure A.1. The coordinates of the center of each quadrilateral element are denoted $\mathbf{x}_I = (x_{I1}, x_{I2}, 0)$ where the capital index I identifies the specific element, say $I = 1, \dots, N_E$. Also, the standard indices $i = 1, 2, 3$, will indicate the vector components. For any element I and any generic vector $\mathbf{w}_I = (w_{I1}, w_{I2}, w_{I3})$ associated to such mesh element, w_{I1}, w_{I2} are the components in the x - y contact plane and w_{I3} is the component in the normal contact direction z . Namely, $\mathbf{w}_{I,T} = (w_{I1}, w_{I2})$ and w_{I3} are the discrete counterparts of \mathbf{w}_T and w_N , respectively.

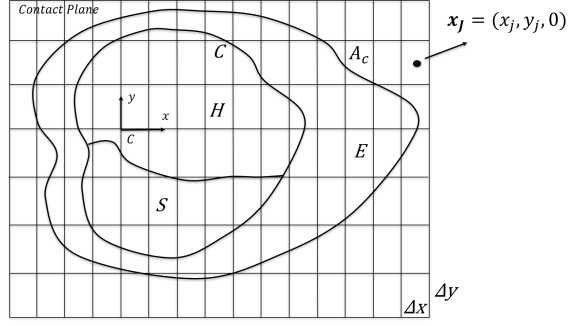


FIG. A.1. Local representation of the discretized contact area.

The discrete values of the elastic deformation \mathbf{u} on the mesh nodes (i.e. the deformation of the elastic bodies in the contact area [23,24]) are defined both at the current time instance t and at the previous time instance t' :

$$\mathbf{u}_I = (u_{Ii}) \text{ at } (\mathbf{x}_I, t), \quad \mathbf{u}'_I = (u'_{Ii}) \text{ at } (\mathbf{x}_I + \mathbf{v}(t - t'), t'), \quad (\text{A.2})$$

where \mathbf{v} is the rolling velocity (i.e. the longitudinal velocity of the wheel) and I is an arbitrary mesh element). Analogously, for the contact pressures \mathbf{p} it holds

$$\mathbf{p}_J = (p_{Jj}) \text{ at } (\mathbf{x}_J, t), \quad \mathbf{p}'_J = (p'_{Jj}) \text{ at } (\mathbf{x}_J + \mathbf{v}(t - t'), t'), \quad (\text{A.3})$$

where J is an arbitrary mesh element. According to the Boundary Element Method Theory [23,24], the discretized displacements \mathbf{u}_I can now be written as a function of the discretized contact pressures \mathbf{p}_J through the discretized version of the problem shape functions, that is

$$u_{Ii} = \sum_{J=1}^{N_E} \sum_{j=1}^3 A_{iJj} p_{Jj}, \quad \text{with } A_{iJj} := B_{iJj}(\mathbf{x}_I),$$

and $B_{iJj}(\mathbf{x}_I)$ are the discrete shape functions of the problem describing the effect of a contact pressure \mathbf{p}_J applied to the element J on displacement \mathbf{u}_I of the node I (see [23,24]). The shape function B_{iJj} usually depends on the problem geometry and the characteristics of the materials. An analogous expression can be derived for u'_{Ii} . The elastic penetration e can be calculated at each node \mathbf{x}_I as

$$e_I = h_I + \sum_J A_{I3J3} p_{J3},$$

where h_I is the discretization of the (known) undeformed distance between the two bodies, see [23,24]. Similarly, the slip \mathbf{s}_T can be discretized by setting

$$\mathbf{s}_{I,T} = \mathbf{c}_{I,T} + (\mathbf{u}_{I,T} - \mathbf{u}'_{I,T})/(t - t'), \quad (\text{A.4})$$

where $\mathbf{c}_{I,T}$ is the discretization of the (given) rigid creep, that is the difference between the velocities of two homologous points belonging to the undeformed wheel and rail surfaces inside the contact area and thought of as rigidly connected to the bodies.

We observe that both \mathbf{u} and \mathbf{s}_T depend linearly on the pressures \mathbf{p} and \mathbf{p}' . Therefore, the discretization of equation $e = 0$ in the norm problem (A.1) yields a linear system in the discretized normal pressures (p_{I3}) while the discretization of the nonlinear equation

$$\mathbf{p}_T = -g \mathbf{s}_T / \|\mathbf{s}_T\|,$$

in the tangential problem yields the nonlinear system

$$\mathbf{s}_{I,T} = -\|\mathbf{s}_{I,T}\| \mathbf{p}_{I,T} / g_I, \quad (\text{A.5})$$

with $\mathbf{p}_{I,T} = (p_{I1}, p_{I2})$ being the unknown[§]. When using the Coulomb-like friction model [23, 24], the friction limit function takes the form $g_I = f_I p_{I3}$, where f_I is a given constant friction value.

The flow of Kalker's CONTACT algorithm is displayed in Figure A.2 [23, 24]. At each time

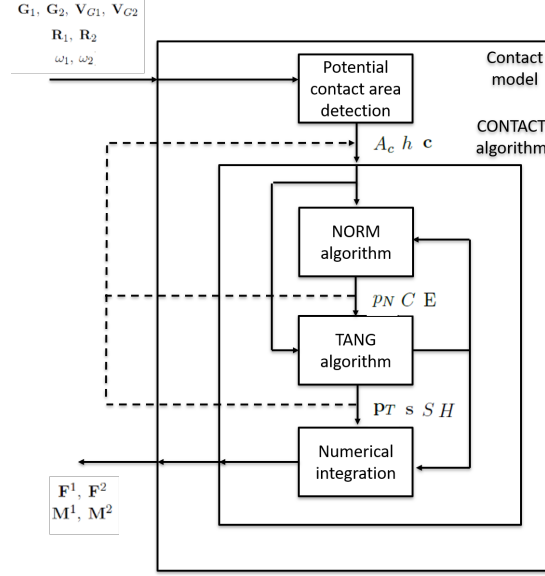


FIG. A.2. The architecture of the Kalker's CONTACT algorithm.

step of time integration, the inputs of the CONTACT algorithm are the potential contact area A_c (usually the interpenetration area between wheel and rail surfaces), the rigid penetration h and the rigid local sliding \mathbf{c}_T (inputs calculated, on turn, from the kinematic variables of the body: position and velocities of the gravity centers $\mathbf{G}_1, \mathbf{G}_2, \mathbf{V}_{G1}, \mathbf{V}_{G2}$, rotation matrices $\mathbf{R}_1, \mathbf{R}_2$ and angular velocities ω_1, ω_2) [23, 24]. All these kinematic quantities are calculated at each time step by the ODE solver of the Simpack Rail multibody environment [39]. NORM algorithm solves the normal contact problem and returns the contact area C , the non-contact area E , the normal contact pressures p_N . Then, TANG algorithm returns the sliding area S , adhesion area H , the tangential contact pressures \mathbf{p}_T and local sliding \mathbf{s}_T . Repetitions of NORM and TANG algorithms are then performed to approximate accurately normal and tangential pressures \mathbf{p}_T, p_N . At the end of CONTACT algorithm, forces and torques exchanged by the contact bodies ($\mathbf{F}^1, \mathbf{F}^2$ and $\mathbf{M}^1, \mathbf{M}^2$) are computed by numerical integration and returned to the time integrator for proceeding in the dynamic simulation of the multibody system.

Appendix B. Complete results. In this section we collect the complete results which gave rise to the performance profiles in Figure 6.2. Results concern two velocities ($v = 10 \text{ m/s}$ in Tables B.1-B.3 and $v = 16 \text{ m/s}$ in Tables B.4-B.6) and the three different track sections (straight line in Tables B.1 and B.4, cycloid in Tables B.2 and B.5 and curve in Tables B.3 and B.6). Given a sequence of nonlinear systems, we label a single system from the sequence as Time_Citer_Titer specifying the instant time (Time), the CONTACT iteration (Citer) and the TANG iteration (Titer). For each SRAND variant applied to a system, we report the number of F -evaluations performed in case of convergence, or, in case of failure, the corresponding flag. We recall from Section 6.3 that a run is successful when $\|F_k\| \leq 10^{-6}$. A failure is declared either because the assigned maximum number of iterations or F -evaluations or backtracks is reached, or because $\|F\|$ was not reduced for 50 consecutive iterations. Such occurrences are denoted as $\mathbf{F}_{it}, \mathbf{F}_{fe}, \mathbf{F}_{bt}, \mathbf{F}_{in}$, respectively.

[§]In the unlikely event $\mathbf{s}_{I,T} = 0$, the system is nonsmooth. We regularize (A.5) replacing the term $\sqrt{s_{I1}^2 + s_{I2}^2}$ with $\sqrt{s_{I1}^2 + s_{I2}^2 + \epsilon}$, for some small positive ϵ .

System	$v = 10 \text{ m/s} - \text{straight line}$							
	BB1	BB2	ALT	ABB		ABBm		DABBm
				$\tau = 0.1$	$\tau = 0.8$	$\tau = 0.1$	$\tau = 0.8$	
101.1.2	69	59	74	75	59	71	57	69
101.2.2	382	148	248	295	205	174	198	220
103.1.2	37	31	35	37	30	37	31	34
103.2.2	37	31	35	37	30	37	31	34
104.1.2	36	36	37	36	38	36	39	38
104.2.2	36	36	37	36	38	36	39	38
105.1.2	39	38	39	39	38	39	39	39
105.1.3	77	69	82	79	70	82	67	74
105.2.2	40	37	39	40	38	40	39	39
105.2.3	74	73	86	75	70	75	67	76

TABLE B.1

Number of function evaluations performed by SRAND variants in the solution of nonlinear systems arising from time 100 to time 105 and corresponding to a straight line with velocity 10 m/s. In the first column we indicate the time step, the CONTACT and the TANG iteration.

REFERENCES

- [1] Awwal, A. M., Kumam, P., Abubakar, A. B., Wakili, A., Pakkaranang, N.: *A new hybrid spectral gradient projection method for monotone system of nonlinear equations with convex constraints*. Thai J. Math. **66-88** (2018).
- [2] Barzilai, J., Borwein, J.: *Two point step gradient methods*. IMA J. Numer. Anal. **8**, 141-148 (1988).
- [3] Birgin, E. G., Martinez, J. M., Raydan, M.: *Spectral Projected Gradient Methods: review and perspectives*. J. Stat. Softw. **60(3)** (2014).
- [4] Bonettini, S., Zanella, R., Zanni, L.: *A scaled gradient projection method for constrained image deblurring*. Inverse Probl. **25(1)**, 015002 (2009).
- [5] Carcasci C., Marini L., Morini B., Porcelli M.: *A new modular procedure for industrial plant simulations and its reliable implementation*. Energy, 94, pp. 380-390 (2016).
- [6] Crisci, S., Ruggiero, V., Zanni, L.: *Steeplength selection in gradient projection methods for box-constrained quadratic programs*. Appl. Math. Comput. **356(1)**, 312-327 (2019).
- [7] Curtis, A.R., Powell, M.J.D., Reid, J.K.: *On the estimation of sparse Jacobian matrices*. IMA J. Appl. Math., **13**, 117-119 (1974).
- [8] Dai, Y. H., Fletcher R.: *Projected Barzilai-Borwein methods for large-scale box-constrained quadratic programming*. Numer. Math. **100**, 21-47 (2005).
- [9] Dai, Y. H., Hager, W., W., Schittkowski, K., Zhang, H.: *The cyclic Barzilai-Borwein method for unconstrained optimization*. IMA J. Numer. Anal. **26(3)**, 604-627 (2006).
- [10] De Asmundis, R., di Serafino, D., Riccio, F., Toraldo, G.: *On spectral properties of steepest descent methods*. IMA J. Numer. Anal. **33(4)**, 1416-1435 (2013).
- [11] Dennis Jr., J. E., Schnabel, R. B.: *Numerical methods for unconstrained optimization and nonlinear equations*. Prentice Hall Series in Computational Mathematics, Prentice Hall, Inc., Englewood Cliffs, NJ (1983).
- [12] di Serafino, D., Ruggiero, V., Toraldo, G., Zanni, L.: *On the steplength selection in gradient methods for unconstrained optimization*. Appl. Math. Comput. **318**, 176-195 (2018).
- [13] Dolan E. D., Moré J. J.: *Benchmarking optimization software with performance profiles*. Math. Programming **91**, 201-213 (2002).
- [14] Facchinei, F., Pang, J.S.: *Finite-Dimensional Variational Inequalities and Complementarity Problems, Volume I*. Springer Series in Operations Research, Springer, New York (2003).
- [15] Fletcher, R.: *On the Barzilai-Borwein method*. Optimization and control with applications, Appl. Optimizat. **96**, 235-256, Springer, New York (2005).
- [16] Frassoldati, G., Zanni, L., Zanghirati, G.: *New adaptive stepsize selections in gradient methods*. J. Ind. Manag. Optim. **4(2)**, 299-312 (2008).
- [17] Glunt, W., Hayden, T., L., Raydan, M.: *Molecular conformations from distance matrices*. J. Comput. Chem. **14(1)**, 114-120 (1993).
- [18] Golub, G. H., Van Loan, C. F.: *Matrix computations*. Johns Hopkins Series in the Mathematical Sciences **3**, Johns Hopkins University Press, Baltimore, MD (1983).
- [19] Gonçalves, M.L.N., Oliveira, F.R.: *On the global convergence of an inexact quasi-Newton conditional gradient method for constrained nonlinear systems* (2018).
- [20] Grippo, L., Lampariello, S., Lucidi, S.: *A nonmonotone linesearch technique for Newton's methods*. SIAM J. Numer. Anal. **23**, 707-716 (1986).
- [21] Grippo, L., Sciandrone, M.: *Nonmonotone derivative-free methods for nonlinear equations*. Comput. Optim.

System	BB1	BB2	ALT	ABB		ABBm		velocity		10 m/s - cycloid	BB1	BB2	ALT	ABB		ABBm		DABm
				$\tau = 0.1$	$\tau = 0.8$	$\tau = 0.1$	$\tau = 0.8$	DABm	System					$\tau = 0.1$	$\tau = 0.8$	$\tau = 0.1$	$\tau = 0.8$	
300.1.2	178	128	137	145	149	174	133	163	303.2.2	F _{Fe}	F _{in}	2196	F _{in}	1486	1413	1111	763	887
300.1.3	513	304	257	296	252	271	230	298	303.2.3	F _{Fe}	1062	7400	1486	1413	911	722	798	
300.1.4	569	402	290	464	350	460	278	299	303.2.4	F _{Fe}	1713	10229	1780	1400	F _{in}	889	1054	
300.2.2	343	203	266	229	194	209	168	204	303.2.5	F _{Fe}	1424	23393	2053	1776	1201	1046	1358	
300.2.3	16421	388	398	406	686	410	330	408	303.3.2	F _{Fe}	926	6424	1352	806	896	814	821	
300.3.2	357	223	248	257	205	225	187	232	303.3.3	F _{Fe}	1318	6285	1508	886	1074	981	896	
300.3.3	1650	385	368	432	530	462	339	499	303.3.4	F _{Fe}	1279	14647	2295	1501	1244	959	1012	
301.1.2	415	281	247	326	325	264	243	248	303.3.5	F _{Fe}	F _{in}	17619	2353	F _{in}	1484	1311	1193	
301.1.3	503	319	351	342	480	280	286	329	304.1.2	39075	962	815	643	504	714	447	491	
301.1.4	582	442	281	380	376	344	291	305	304.1.3	F _{Fe}	711	2891	860	1242	710	607	562	
301.2.2	1127	286	298	271	430	310	284	297	304.1.4	F _{Fe}	1524	3611	966	1423	785	515	752	
301.2.3	630	414	367	388	430	322	313	337	304.2.2	725	366	381	393	416	300	311	317	
301.2.4	758	345	372	408	355	363	319	386	304.2.3	65775	558	648	753	734	577	453	548	
301.3.2	918	357	299	315	350	294	288	326	304.2.4	56953	709	1870	638	920	562	475	523	
301.3.3	750	400	320	473	423	350	305	313	304.3.2	415	421	370	470	431	357	339	325	
301.3.4	440	363	302	352	434	310	301	393	304.3.3	47176	533	2376	616	627	518	411	612	
302.1.2	F _{Fe}	743	3727	993	1022	558	457	495	304.3.4	86605	696	1180	709	603	557	468	488	
302.1.3	F _{Fe}	844	4067	1183	972	1068	670	678	305.1.2	796	270	311	302	323	329	242	364	
302.1.4	F _{Fe}	3546	25810	6171	2529	1735	1267	1342	305.1.3	339	293	270	271	294	288	243	310	
302.2.2	634	444	417	552	539	431	332	376	305.1.4	430	342	301	354	335	307	230	309	
302.2.3	27285	610	508	890	544	502	398	548	305.2.2	F _{Fe}	F _{in}	2434	1401	800	F _{in}	1282	1208	
302.2.4	F _{Fe}	F _{in}	7325	1359	1951	927	853	693	305.2.3	F _{Fe}	1110	2222	1713	1030	950	717	684	
302.3.2	743	426	373	455	438	402	332	361	305.2.4	F _{Fe}	F _{in}	842	1527	846	748	768	648	
302.3.3	39825	739	502	869	616	459	401	463	305.2.5	F _{Fe}	F _{in}	3329	1516	850	1332	573	597	
302.3.4	F _{Fe}	2245	7598	1141	938	1005	660	702	305.3.2	F _{Fe}	980	6755	1524	F _{in}	920	1036	1518	
303.1.2	22687	554	679	502	F _{in}	609	405	460	305.3.3	F _{Fe}	F _{in}	5805	1829	756	694	634	579	
303.1.3	33798	468	684	571	578	461	411	562	305.3.4	F _{Fe}	871	2502	1363	997	857	716	648	
303.1.4	F _{Fe}	965	1163	734	669	653	524	613	305.3.5	F _{Fe}	F _{in}	1786	1286	843	929	702	663	

TABLE B.2

Results for each system of the sequences generated in the cycloid section of the train track with velocity $v = 10$ m/s.

System	velocity 10 m/s - curve									
	BB1	BB2	ALT	ABB		ABBm		DABBm		DABBm
				$\tau = 0.1$	$\tau = 0.8$	$\tau = 0.1$	$\tau = 0.8$	$\tau = 0.1$	$\tau = 0.8$	
450.1.2	386	210	246	251	293	293	211	284	453.1.3	402
450.1.3	623	204	303	285	268	1580	1627	453.1.4	F _{fe}	319
450.2.2	29520	492	457	475	416	320	471	453.2.2	F _{fe}	319
450.2.3	12031	428	433	412	458	309	387	453.2.3	F _{fe}	319
450.3.2	13652	560	403	562	416	379	382	453.2.4	F _{in}	319
450.3.3	11509	464	448	518	493	393	391	453.3.2	F _{fe}	319
451.1.2	681	437	382	520	570	340	397	453.3.3	F _{fe}	319
451.1.3	F _{fe}	1218	4314	999	1564	613	1501	453.3.4	F _{fe}	319
451.1.4	F _{fe}	3805	18920	1790	F _{in}	1083	1334	454.1.2	147	319
451.2.2	324	274	329	264	264	210	250	454.1.3	207	319
451.2.3	F _{fe}	1652	1046	859	1304	691	595	454.1.4	2367	319
451.2.4	F _{fe}	1573	F _{in}	1260	F _{in}	1232	941	454.1.5	861	319
451.3.2	381	253	240	301	243	209	270	454.2.2	237	319
451.3.3	F _{fe}	3141	4232	660	801	606	635	454.2.3	413	319
451.3.4	F _{fe}	F _{in}	F _{in}	F _{in}	F _{in}	936	888	454.2.4	901	319
451.4.2	358	296	321	279	295	213	263	454.3.2	259	319
451.4.3	F _{fe}	2108	901	688	729	597	639	454.3.3	469	319
451.4.4	F _{fe}	F _{in}	12872	1797	F _{in}	905	821	454.3.4	450	319
452.1.2	66785	638	638	548	743	545	522	455.1.2	147	319
452.1.3	71198	701	725	535	789	552	508	455.1.3	212	319
452.1.4	45680	803	521	617	594	470	520	455.1.4	482	319
452.2.2	498	557	887	514	539	301	467	455.2.2	497	319
452.2.3	37679	608	714	474	672	425	454	455.2.3	563	319
452.2.4	40269	718	797	565	790	379	501	455.2.4	F _{fe}	319
452.3.2	31230	433	451	438	517	405	354	455.3.2	341	319
452.3.3	41623	581	634	575	726	400	451	455.3.3	603	319
452.3.4	5592	477	658	572	570	407	470	455.3.4	F _{fe}	319
453.1.2	288	200	257	227	210	190	210			319

TABLE B.3
Results for each system of the sequences generated in the curve segment of the train path with velocity $v = 10$ m/s.

System	velocity 16 m/s - straight line							
	BB1	BB2	ALT	ABB		ABBm		DABBm
				$\tau = 0.1$	$\tau = 0.8$	$\tau = 0.1$	$\tau = 0.8$	
50_1_2	60	45	53	52	47	52	46	49
50_2_2	53	44	51	54	48	54	48	53
50_3_2	53	44	51	48	48	48	48	53
52_2_2	75	78	53	76	75	101	61	91
52_3_2	89	78	53	76	88	112	61	91
55_1_2	65	66	66	83	66	80	62	72
55_2_2	69	79	60	76	61	73	67	71
55_3_2	69	79	60	80	61	73	67	71

TABLE B.4

Number of function evaluations performed by SRAND variants in the solution of nonlinear systems arising from time 50 to time 55 and corresponding to a straight line with velocity 16 m/s. In the first column we indicate the time step, the CONTACT and the TANG iteration.

- Appl. **37**, 297-328 (2007).
- [22] Gu, G. Z., Li, D. H., Qi, L., Zhou, S. Z.: *Descent directions of quasi-Newton methods for symmetric nonlinear equations*. SIAM J. Numer. Anal. **40**, 1763-1774 (2002).
- [23] Kalker, J.: *Three-Dimensional elastic bodies in rolling contact*. Kluwer Academic Print, Delft (1990).
- [24] Kalker, J., Jacobson, B.: *Rolling contact phenomena*. Springer Verlag, Wien (2000).
- [25] La Cruz, W., Raydan, M.: *Nonmonotone spectral methods for large-scale nonlinear systems*. Optim. Method Softw. **18**, 583-599 (2003).
- [26] La Cruz, W., Martinez, J. M., Raydan, M.: *Spectral residual method without gradient information for solving large-scale nonlinear systems of equations*. Math. Comput. **75**, 1429-1448 (2006).
- [27] La Cruz, W.: *A projected derivative-free algorithm for nonlinear equations with convex constraints*. Optim. Method Softw. **29**, 24-41 (2014).
- [28] Li, D.H., Fukushima, M.: *A derivative-free line search and global convergence of Broyden-like method for nonlinear equations*. Optim. Method Softw. **13(3)**, 181-201 (2000).
- [29] Li, Q., Li, D. H.: *A class of derivative-free methods for large-scale nonlinear monotone equations*. IMA J. Numer. Anal. **31**, 1625-1635 (2011).
- [30] Liu, J., Li, S.: *Multivariate spectral dy-type projection method for convex constrained nonlinear monotone equations*. J. Ind. Manag. Optim. **13**, 283-295 (2017).
- [31] Marini, L., Morini, B., Porcelli, M.: *Quasi-Newton methods for constrained nonlinear systems: complexity analysis and applications*. Comput. Optim. Appl. **71**, 147-170 (2018).
- [32] Mohammad, H., Abubakar A., B.: *A positive spectral gradient-like method for large-scale nonlinear monotone equations*. Bull. Comput. Appl. Math. **5**, 99-115 (2017).
- [33] Morini, B., Porcelli, M.: *TRESNEI, a Matlab trust-region solver for systems of nonlinear equalities and inequalities*. Comput. Optim. Appl. **51**, 27-49 (2012).
- [34] Morini, B., Porcelli, M., Toint, P.: *Approximate norm descent methods for constrained nonlinear systems*. Math. Comput. **87**, 1327-1351 (2018).
- [35] Papini A., Porcelli M., Sgattoni C.: *On the global convergence of a new spectral residual algorithm for nonlinear systems of equations*. Boll. Unione Mat. Ital., **14**, 367-378 (2021).
- [36] Raydan, M.: *Convergence properties of the Barzilai and Borwein Gradient Method*. PhD Thesis, Rice University (1991).
- [37] Raydan, M.: *On the Barzilai and Borwein choice of step length for the gradient method*. IMA J. Numer. Anal. **13**, 321-326 (1993).
- [38] Raydan, M.: *The Barzilai and Borwein gradient method for the large scale unconstrained minimization problem*. SIAM J. Optimiz. **7**, 26-33 (1997).
- [39] *Simpack Multibody Simulation Software*. Dassault Systemes GmbH.
- [40] Yu, Z., Lin, J., Sun, J., Xiao, Y., Liu, L., Li, Z.: *Spectral gradient projection method for monotone nonlinear equations with convex constraints*. Appl. Numer. Math. **59**, 2416-2423 (2009).
- [41] Varadhan, R., Gilbert, P. D.: *BB: an R package for solving a large system of nonlinear equations and for optimizing a high-dimensional nonlinear objective function*. J. Stat. Softw. **32 (4)** (2010).
- [42] Vollebregt, E. A. H.: *Refinement of Kalker's rolling contact model*. Bracciali, Proceedings of the 8th International Conference on Contact Mechanics and Wear of Rail-Wheel Systems (CM2009), Firenze, 2009.
- [43] Vollebregt, E. A. H.: *User guide for CONTACT, Rolling and sliding contact with friction*. Technical Report TR09-03, version v15.1.1 (2015).
- [44] Zhang, L., Zhou, W.: *Spectral gradient projection method for solving nonlinear monotone equations*. J. Comput. Appl. Math. **196**, 478-484 (2006).
- [45] Zhou, B., Gao, L., Dai, Y. H.: *Gradient methods with adaptive step-sizes*. Comput. Optim. Appl. **35(1)**, 69-86 (2006).

System	BB1	BB2	ALT	velocity 16 m/s - cycloid				BB1	BB2	ALT	ABB		ABBm	DABBm		DABBm
				$\tau = 0.1$	$\tau = 0.8$	$\tau = 0.1$	$\tau = 0.8$				$\tau = 0.1$	$\tau = 0.8$		$\tau = 0.1$	$\tau = 0.8$	
				ABB	ABB	ABBm	DABBm	System								
150.1.2	985	297	330	366	357	351	278	343	153.1.3	1173	1181	1162	1179	735	568	596
150.1.3	26886	569	512	612	555	487	419	437	153.1.4	991	3881	1003	1590	1044	635	771
150.1.4	F _{fe}	967	3163	653	F _{in}	550	604	617	153.2.2	475	603	688	532	578	396	446
150.1.5	F _{fe}	F _{in}	810	647	1549	614	510	710	153.2.3	1149	3920	1316	1506	843	621	704
150.2.2	476	228	307	295	302	277	216	301	153.2.4	1445	5035	1262	1272	1215	602	784
150.2.3	627	584	404	437	485	377	344	443	153.2.5	772	4023	926	1576	1188	764	725
150.2.4	52373	585	479	494	730	438	391	435	153.3.2	628	754	674	585	489	429	471
150.3.2	F _{fe}	1304	F _{in}	F _{in}	1777	2707	1237	911	153.3.3	770	4768	1187	1882	941	699	860
150.3.3	F _{fe}	2498	F _{in}	F _{in}	F _{in}	2300	1973	1737	153.3.4	1568	4872	923	1161	1173	678	709
150.3.4	F _{fe}	6214	F _{in}	F _{in}	F _{in}	3097	2576	F _{in}	153.3.5	1226	5474	1145	1118	730	688	730
151.1.2	F _{fe}	F _{in}	5095	841	905	664	605	689	154.1.2	776	3124	727	1033	585	534	527
151.1.3	F _{fe}	1114	5312	1421	1144	810	616	829	154.1.3	386	513	467	681	433	310	346
151.1.4	F _{fe}	1454	8154	1630	3755	1125	1139	1046	154.1.4	533	421	539	518	434	404	447
151.1.5	F _{fe}	3590	13111	2610	1435	1231	864	1043	154.2.2	319	312	420	357	341	294	356
151.2.2	F _{fe}	1337	12656	1333	3092	973	864	856	154.2.3	193	220	216	241	238	201	246
151.2.3	F _{fe}	3776	9599	1983	2198	1077	949	961	154.2.4	266	255	255	258	250	228	276
151.2.4	F _{fe}	3013	9073	1867	3551	1409	870	974	154.3.2	403	288	336	394	302	277	354
151.2.5	F _{fe}	5005	18543	1831	3662	1635	1270	1345	154.3.3	248	218	249	276	217	206	233
151.3.2	F _{fe}	F _{in}	7743	F _{in}	3893	F _{in}	939	803	154.3.4	346	278	281	271	267	239	250
151.3.3	F _{fe}	2293	9494	1383	1689	1080	809	982	155.1.2	1161	5470	1151	987	824	718	859
151.3.4	F _{fe}	1235	7622	1416	1884	1075	856	941	155.1.3	F _{in}	31313	4192	4270	1758	1401	1193
151.3.5	F _{fe}	4085	24983	1853	F _{in}	1509	1147	1330	155.1.4	5839	19894	F _{in}	4182	1621	1729	1380
152.1.2	68856	822	1395	742	661	680	473	575	155.1.5	F _{in}	F _{in}	F _{in}	F _{in}	1624	1351	1339
152.1.3	F _{fe}	682	4009	1153	1085	859	648	669	155.2.2	1211	3754	1267	1275	764	651	635
152.1.4	F _{fe}	725	2905	986	1423	799	646	720	155.2.3	F _{in}	F _{in}	2536	F _{in}	1658	1328	1273
152.2.2	21104	604	641	407	681	543	347	399	155.2.4	1623	24770	3690	F _{in}	1626	1461	1427
152.2.3	80349	701	1082	636	845	632	476	610	155.2.5	F _{in}	F _{bt}	F _{in}	F _{in}	1683	1715	1559
152.2.4	F _{fe}	1748	3725	1395	1034	873	590	849	155.3.2	877	6004	990	882	795	567	818
152.3.2	20711	567	601	382	664	453	358	420	155.3.3	F _{in}	23302	1784	F _{in}	F _{in}	1539	1238
152.3.3	75894	966	1098	522	898	639	535	627	155.3.4	2895	32130	1953	F _{in}	1539	1739	1315
152.3.4	F _{fe}	1146	4114	848	1152	744	558	734	155.3.5	F _{in}	F _{in}	6554	F _{in}	F _{in}	F _{in}	F _{in}
153.1.2	1281	408	589	512	495	472	400	397								

TABLE B.5
Results for each system of the sequences generated in the cycloid section of the train track with velocity $v = 16$ m/s.

System	BB1	BB2	ALT	velocity 16 m/s - curve															
				ABB		ABBm		DABm		BB1	BB2	ALT	ABB		ABBm		DABm		
				$\tau = 0.1$	$\tau = 0.8$	$\tau = 0.1$	$\tau = 0.8$												
350.1.2	424	320	308	359	366	297	284	286	352.4.5	F _{Fe}	1132	7322	1252	F _{Fin}	921	F _{Fin}	724		
350.1.3	F _{Fe}	825	5650	826	905	771	540	687	353.1.2	468	357	398	482	342	352	307	357		
350.2.2	308	208	220	244	261	243	197	247	353.1.3	887	640	588	557	441	508	446	456		
350.2.3	F _{Fe}	1322	3384	572	F _{Fin}	501	433	497	353.1.4	F _{Fe}	695	4525	905	1369	781	625	656		
350.2.4	F _{Fe}	F _{Fin}	6845	1204	1523	746	790	718	353.1.5	F _{Fe}	877	4670	793	1551	782	682	764		
350.3.2	311	221	277	264	234	214	188	213	353.2.2	589	357	365	461	398	426	370	386		
350.3.3	76754	F _{Fin}	885	639	666	491	416	481	353.2.3	47619	755	572	913	812	529	459	528		
350.3.4	F _{Fe}	F _{Fin}	6032	675	F _{Fin}	1141	761	647	353.2.4	F _{Fe}	1143	3476	F _{Fin}	857	798	642	687		
350.4.2	271	207	233	229	226	220	201	218	353.2.5	F _{Fe}	1984	8598	1370	1700	F _{Fin}	867	1111		
350.4.3	91233	764	3110	633	829	536	432	526	353.3.2	711	381	394	481	380	408	368	361		
350.4.4	F _{Fe}	1593	6301	722	F _{Fin}	637	F _{Fin}	751	353.3.3	65122	672	600	710	996	604	511	457		
351.1.2	F _{Fe}	1241	1625	920	913	772	597	538	353.3.4	F _{Fe}	837	1623	815	1111	759	588	633		
351.1.3	F _{Fe}	1596	11134	1807	F _{Fin}	1374	1199	1090	353.3.5	F _{Fe}	1250	6524	1233	1350	1110	915	855		
351.1.4	F _{Fe}	2272	20207	1862	F _{Fin}	1555	1217	1240	353.4.2	575	448	505	425	360	350	341	372		
351.2.2	F _{Fe}	1088	F _{Fin}	F _{Fin}	1207	1385	959	1050	353.4.3	57903	732	725	644	469	517	492	533		
351.2.3	F _{Fe}	2428	F _{Fin}	F _{Fin}	F _{Fin}	2185	1567	1825	353.4.4	F _{Fe}	1030	932	873	1055	679	630	669		
351.2.4	F _{Fe}	5683	F _{Fin}	F _{Fin}	F _{Fin}	2421	2064	1636	353.4.5	F _{Fe}	8112	1276	1502	1502	980	904	967		
351.2.5	F _{Fe}	F _{Fin}	F _{Fin}	F _{Fin}	F _{Fin}	3192	2052	2770	354.1.2	313	229	219	320	261	265	187	253		
351.3.2	F _{Fe}	1261	12388	3742	1566	992	1166	876	354.1.3	502	323	369	398	337	318	267	342		
351.3.3	F _{Fe}	2029	F _{Fin}	F _{Fin}	F _{Fin}	F _{Fin}	F _{Fin}	1704	354.1.4	87446	710	4042	610	716	579	536	673		
351.3.4	F _{Fe}	2397	F _{Fin}	F _{Fin}	4270	2105	2074	1630	354.2.2	445	321	348	373	292	289	230	266		
351.3.5	F _{Fe}	F _{Fin}	F _{Fin}	F _{Fin}	F _{Fin}	2833	F _{Fin}	2635	354.2.3	1771	462	359	434	473	355	345	372		
351.4.2	F _{Fe}	1285	F _{Fin}	4846	1378	1262	1313	1028	354.2.4	F _{Fe}	1054	4522	1052	1159	757	649	701		
351.4.3	F _{Fe}	1778	F _{Fin}	F _{Fin}	2581	2073	2144	1764	354.3.2	451	315	295	324	275	259	265	316		
351.4.4	F _{Fe}	F _{Fin}	F _{Fin}	F _{Fin}	F _{Fin}	2848	1794	1763	354.3.3	789	382	392	508	521	409	408	409		
351.4.5	F _{Fe}	F _{Fin}	F _{Fin}	F _{Fin}	F _{Fin}	F _{Fin}	3340	F _{Fin}	354.3.4	F _{Fe}	913	3478	786	921	845	607	665		
352.1.2	F _{Fe}	1794	F _{Fin}	5760	1636	1619	1933	1728	354.4.2	405	323	289	350	308	317	256	295		
352.1.3	F _{Fe}	3141	F _{Fe}	3787	2872	1686	1495	1524	354.4.3	1776	497	363	452	338	399	333	370		
352.1.4	F _{Fe}	F _{Fin}	F _{Fin}	F _{Fin}	F _{Fin}	2334	1657	1721	354.4.4	F _{Fe}	991	4561	830	1141	704	553	634		
352.1.5	F _{Fe}	F _{Fin}	F _{Fin}	F _{Fin}	F _{Fin}	2318	2846	1623	355.1.2	638	226	262	264	292	268	258	266		
352.2.2	72375	676	1359	708	586	643	459	501	355.1.3	527	339	509	348	348	348	286	331		
352.2.3	74955	801	878	794	718	857	481	519	355.1.4	35134	489	1201	464	525	477	382	408		
352.2.4	F _{Fe}	866	5116	1209	1071	837	648	746	355.2.2	346	222	252	246	243	221	194	242		
352.2.5	F _{Fe}	F _{Fin}	12683	1209	F _{Fin}	921	803	909	355.2.3	2303	480	396	402	357	376	261	358		
352.3.2	59157	701	1249	712	652	687	420	589	355.2.4	41075	671	542	511	401	313	355	433		
352.3.3	87628	1116	682	804	611	639	517	517	355.3.2	336	289	249	264	282	194	232	241		
352.3.4	F _{Fe}	808	6379	845	830	726	782	685	355.3.4	639	268	480	340	370	304	291	369		
352.3.5	F _{Fe}	1213	8333	1658	1133	863	697	781	355.3.5	24592	624	753	457	744	448	388	428		
352.4.2	48585	603	818	679	775	668	460	528	355.4.2	363	214	268	226	261	261	203	221		
352.4.3	79649	867	628	720	876	590	470	511	355.4.3	714	463	360	369	343	383	260	314		
352.4.4	F _{Fe}	F _{Fin}	4570	1046	1200	858	708	804	355.4.4	32137	404	700	411	532	562	367	451		

TABLE B.6

Results for each system of the sequences generated in the curve section of the train track with velocity $v = 16$ m/s.