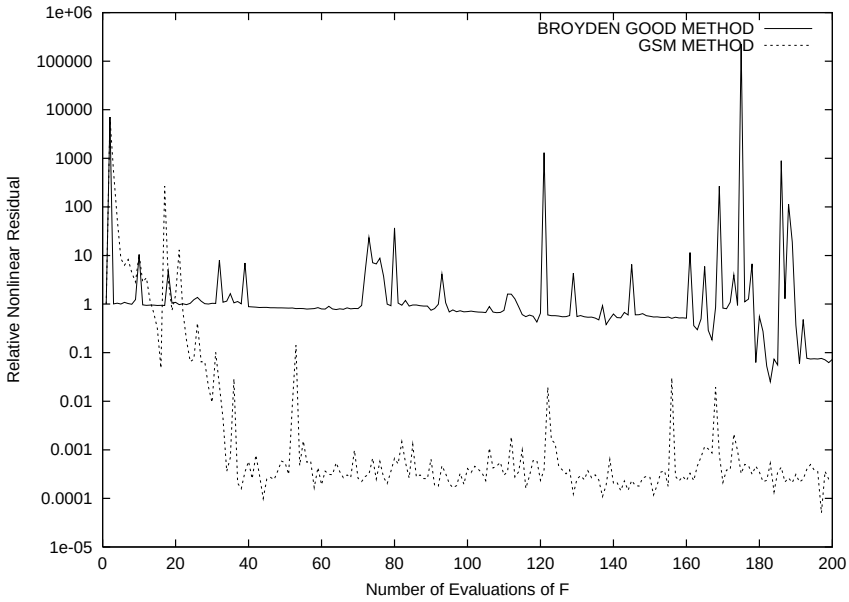


A generalization of secant methods for solving nonlinear systems of equations

Frank Crittin, ROSO-IMA-EPFL
Michel Bierlaire, ROSO-IMA-EPFL

Conference paper STRC 2003
Session Model and Statistics

STRC

3rd Swiss Transport Research Conference

Monte Verità / Ascona, March 19-21, 2003

A generalization of secant methods for solving nonlinear systems of equations

Frank Crittin
Institut de Mathématiques
École Polytechnique Fédérale de Lausanne

Phone: +41 21 693 81 00
Fax: +41 21 693 55 70
e-Mail: frank.crittin@epfl.ch

Michel Bierlaire
Institut de Mathématiques
École Polytechnique Fédérale de Lausanne

Phone: +41 21 693 25 37
Fax: +41 21 693 55 70
e-Mail: michel.bierlaire@epfl.ch

Abstract

Many problems in transportation can be formulated as fixed point problems, or equivalently as a systems of nonlinear equations, as for example multi-user equilibrium problems, planning problems, and consistent anticipatory route guidance generation. These fixed point problems are often characterized by the fact that there is no analytical form of the objective function, the computational cost of evaluating the function is generally high, the dimension is very large and stochasticity is present. Therefore classical algorithms to solve linear systems of equation based on derivatives information or involving line search are unlikely to be attractive in this context.

In this paper we introduce a new class of methods for solving nonlinear systems of equations motivated by the constraints described above. The main idea is to generalize classical secant methods by building the secant model using more than two previous iterates. This approach allows to use all the available information collected through the iterations to construct the model. We prefer a least-square approach to calibrate the secant model, as exact interpolation requires a fixed number of iterates, and may be numerically problematic. We also propose an explicit control of the numerical stability of the method.

We show that our approach can lead to an update formula à la Broyden. In that case, we proof the local convergence of the corresponding quasi-Newton method. Finally, computational comparisons with classical methods, based on performance profiles, highlight a significant improvement in term of robustness and number of function evaluations. We also present preliminary numerical tests showing the robust behavior of our methods in the presence of noisy nonlinear system of equations. Numerical results of this method apply to the consistent anticipatory route guidance will be provided in the other presentation (*"Solving the anticipatory route guidance generation problem using a generalization of secant methods"*.

Keywords

Quasi-Newton methods - population based generation - fixed point problems - 3^{rd} Swiss Transport Research Conference - STRC 2003 - Monte Verità

1 Introduction

Many problems in transportation can be formulated as fixed point problems or equivalently, as shown in the following, as resolutions of systems of nonlinear equations. For example, traffic equilibrium models can be formulated as well as variational inequality, nonlinear complementarity or fixed point problems as demonstrated by Patriksson (1994). A practical modelization is given by Cantarella (1997) who presents a fixed point formulation of multi-mode multi-user equilibrium assignement with elastic demand. Another example as been proposed by Cascetta and Postorino (2001) who suggest fixed point approach to estimate Origin-Destination matrix using traffic counts. Another type of problems is the generation of anticipatory route guidance. Travel guidance refers to information provided to users in an attempt to facilitate their decision. But because traffic information affects drivers behavior, it may invalidate predicted traffic conditions that were used to generate it. This problem has also been formulated as a fixed point problem by Bottom (2000) and a new solution algorithm have been designed by Bierlaire and Crittin (2003) based on an adaptation of methods proposed in this paper.

In transportation fixed point formulations are mainly used to establish theoretical existence and uniqueness results given by the Banach Contraction Principle (Banach, 1922), which also allows the specification of convergent algorithms. But, from a practical point of view, methods proposed to solve fixed point problems are not really efficient compared to algorithms solving nonlinear systems of equations. Unfortunately transportation problems formulated as fixed point problems have generally the following characteristics:

- no analytical form of the objective function,
- high computational costs of evaluating the function,
- large scale problems,
- presence of stochasticity.

Therefore even classical algorithms solving nonlinear systems of equations, often based on derivatives information or involving line search, are unlikely to be attractive in this context. It is the reason why we propose a new class of algorithms designed to take this features into account.

We consider the standard problem of identifying a fixed point of a given mapping. More precisely let $\mathcal{T} : \mathbb{R}^n \rightarrow \mathbb{R}^n$ a continuously differentiable function, the fixed point problem can be expound as finding $x \in \mathbb{R}^n$ such that:

$$\mathcal{T}(x) = x. \tag{1}$$

In the same way we can characterize resolution of a nonlinear system of equations as finding $x \in \mathbb{R}^n$ such that:

$$F(x) = 0 \tag{2}$$

where $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is a continuously differentiable function. The proof of equivalence of this two formulations, *i.e.* if x^* is solution of (1), it is also solution of (2) and reciprocally, is straightforward setting $F(x) = \mathcal{T}(x) - x$. So in the following we will restrict to the resolution of systems of nonlinear equations without loss of generality.

Since Newton, this problem has received a tremendous amount of attention. Newton's method and its many variations are still intensively analyzed and used in practice. The philosophy

of Newton-like methods is to replace the nonlinear function F by a linear model, which approximates F in the neighborhood of the current iterate. The original Newton method invokes Taylor's theorem and uses the first derivative matrix (or Jacobian) to construct the linear model. When the Jacobian is too expensive to evaluate, secant methods build the linear model based on the secant equation. Because secant methods exhibit a decent rate of convergence (q-superlinear), they have been intensively analyzed in the literature.

The secant equation imposes that the linear model perfectly matches the nonlinear function F at two successive iterates. If the number of unknowns n is strictly greater than 1, an infinite number of linear models verify the secant equation. Therefore, each secant method derives a specific update formula which arbitrarily picks one linear model among them. The most common strategies are called "least-change updates" and select the linear model which minimizes the difference between two successive models.

In this paper, we provide a class of algorithms generalizing these methods. They are also based on a linear model. Instead of using only two successive iterates to determine it, we maintain a "population" of iterates. Indeed, in the presence of expensive function evaluation, we want to incorporate all the acquired information about the function to calibrate at best the model. Also, we expect a population-based approach to be more robust in the presence of noise in the function.

An important feature of our method is that we do not impose an exact match between the model and the function. Instead, we use a least-square approach to request that the model matches the function "as much as possible". This class of algorithms exhibits a faster convergence and a greater robustness than quasi-Newton methods for most numerical tests that we have performed (Section 5) at a cost of substantial linear algebra computation. Therefore it is valuable when the cost of evaluating F is high in comparison with the numerical algebra overhead. In this paper, we present the class of algorithms (Section 3) and prove that they are locally convergent (Section 4).

2 Quasi-Newton methods

Quasi-Newton methods consider at each iteration the linear model

$$L_k(x; B_k) = F(x_k) + B_k(x - x_k) \tag{3}$$

which approximates $F(x)$ in the neighborhood of x_k and computes x_{k+1} as a solution of the linear system $L_k(x; B_k) = 0$. Consistently with most of the publications on this topic, quasi-Newton methods can be summarized as methods based on the following iterations:

$$x_{k+1} = x_k - B_k^{-1} F(x_k). \tag{4}$$

followed by the computation of B_{k+1}. Of course the most illustrious quasi-Newton method is the pure Newton method where $B_k = J(x_k) = \nabla F(x_k)^T$ is the Jacobian of F evaluated at x_k, that is a $n \times n$ matrix such that entry (i, j) is $\partial F_i / \partial x_j$. We refer the reader to Dennis and Schnabel (1996) for an extensive analysis of Newton and quasi-Newton methods.

Broyden (1965) proposes a class of quasi-Newton methods based on the *secant equations*, imposing the linear model L_{k+1} to exactly match the nonlinear function at iterates x_k and x_{k+1},

that is

$$\begin{array}{ll} L_{k+1}(x_k) & = F(x_k), \\ L_{k+1}(x_{k+1}) & = F(x_{k+1}). \end{array} \tag{5}$$

Subtracting these two equations and defining $y_k = F(x_{k+1}) - F(x_k)$ and $s_k = x_{k+1} - x_k$ we obtain the classical secant equation:

$$B_{k+1} s_k = y_k \tag{6}$$

Clearly, if the dimension n is strictly greater than 1, there is an infinite number of matrices B_{k+1} satisfying (6). An arbitrary decision must consequently be made. The "least-change secant update" strategy, proposed by Broyden (1965), consists in selecting among the matrices verifying (6) the one minimizing variations between two successive models $L_k(x)$ and $L_{k+1}(x)$. It leads to the following update formula

$$B_{k+1} = B_k + \frac{(y_k - B_k s_k)\, s_k^T}{s_k^T s_k}. \tag{7}$$

This method has been very successful, and has been widely adopted in the field. However, we believe that the philosophy of interpolating the linear model on only two iterates and so forgetting all the information given by previous iterates could be too restrictive, incurring the risk that the arbitrariness introduced by the method plays an overly important role. Therefore, we propose here to use more than two past iterates to build the linear model, expecting a better approximation of the actual tangent model.

This idea has already been considered. Dennis and Schnabel (1996) say that "In fact, multivariate generalizations of the secant method have been proposed ... but none of them seem robust enough for general use." Later, they write "Perhaps the most obvious strategy is to require the model to interpolate $F(x)$ at other past points... One problem is that the directions tend to be linearly dependent or close to it, making the computation of (the approximation matrix) a poorly posed numerical problem".

There are few attempts to generalize this approach in the literature. A first generalization of the secant method is the *sequential secant method* proposed by Wolfe (1959) and discussed by Ortega and Rheinboldt (1970). The idea is to impose exact interpolation of the linear model on $n+1$ iterates instead of 2:

$$L_{k+1}(x_{k-j}) = F(x_{k-j}), \quad j = 0, 1, \ldots, n. \tag{8}$$

or, equivalently,

$$B_{k+1} s_{k-j} = y_{k-j}, \quad j = 0, \ldots, n, \tag{9}$$

where $s_i = x_{k+1} - x_i$, and $y_i = F(x_{k+1}) - F(x_i)$, for all i. If the vectors $s_k, s_{k-1}, \ldots, s_{k-n+1}$ are linearly independent, there exists exactly one matrix B_{k+1} satisfying (9), which is

$$B_{k+1} = \tilde{Y}_{k+1} \tilde{S}_{k+1}^{-1} \tag{10}$$

where $\tilde{Y}_{k+1} = (y_k, y_{k-1}, \ldots, y_{k-n+1})$ and $\tilde{S}_{k+1} = (s_k, s_{k-1}, \ldots, s_{k-n+1})$.

Quoting Ortega and Rheinboldt (1970) "...(sequantial methods) are prone to unstable behavior and ... no satisfactory convergence results can be given". Nevertheless Gragg and Stewart (1976) propose a method which avoid instabilities by working with orthogonal factorizations of the involved matrices. Martinez (1979) gives three implementations of the idea proposed by Gragg and Stewart (1976) and some numerical experiments.

Multi-step quasi-Newton methods have been proposed by Moghrabi (1993), Ford and Moghrabi (1997) and Ford (1999) in the context of nonlinear programming. An interpolating path is built based on previous iterates, and used to produce an alternative secant equation. Interestingly, the best numerical results were obtained with no more than two steps.

We believe that the comments about the poor numerical stability of those methods found in major reference texts such as Dennis and Schnabel (1996) and Ortega and Rheinboldt (1970) have not encouraged more researchers to investigate generalizations of Broyden's method. We provide here such a generalization with robust properties and exhibiting an excellent behavior on numerical examples.

3 Population-based generalization

We propose here a class of methods calibrating a linear model based on several previous iterates. The difference with existing approaches is that we do not impose the linear model to interpolate the function. Instead, we prefer to identify the linear model which is as close as possible to the nonlinear function, in the least-squares sense.

At each iteration, we maintain a finite population of previous iterates. Without loss of generality, we present the method assuming that all previous iterates $x_0, \ldots, x_{k+1}$ are considered. Our method belongs also to the quasi-Newton framework defined by (4), where B_{k+1} is computed as follows.

$$B_{k+1} = \underset{J}{\operatorname{argmin}} \left(\sum_{i=0}^{k} \left\| \omega_{k+1}^i F(x_i) - \omega_{k+1}^i L_{k+1}(x_i; J) \right\|_2^2 + \left\| \Gamma J - \Gamma B_{k+1}^0 \right\|_F^2 \right) \tag{11}$$

where L_{k+1} is defined by (3) and $B_{k+1}^0 \in \mathbb{R}^{n \times n}$ is an a priori approximation of B_{k+1}. The role of the second term is to overcome the under-determination of the least-square problem based on the first term and also control the numerical stability of the method. The matrix Γ contains weights associated with the arbitrary term B_{k+1}^0, and the weights $\omega_{k+1}^i \in \mathbb{R}^+$ are associated with the previous iterates. Equation (11) can be written in matrix form as follows: $B_{k+1} =$

$$\underset{J}{\operatorname{argmin}} \left\| J \begin{pmatrix} S_{k+1} & I_{n\times n} \end{pmatrix} \begin{pmatrix} \Omega & 0_{k\times n} \\ 0_{n\times k} & \Gamma \end{pmatrix} - \begin{pmatrix} Y_{k+1} & B_{k+1}^0 \end{pmatrix} \begin{pmatrix} \Omega & 0 \\ 0 & \Gamma \end{pmatrix} \right\|_F^2$$

where $\Omega \in \mathbb{R}^{k+1}$ is a diagonal matrix with weights ω_{k+1}^i on the diagonal for $i = 0, \cdots, k$. The normal equations of the least-square problem lead to the following formula:

$$B_{k+1} = B_{k+1}^0 + \left(Y_{k+1} - B_{k+1}^0 S_{k+1}\right) \Omega^2 S_{k+1}^T \left(\Gamma^2 + S_{k+1} \Omega^2 S_{k+1}^T\right)^{-1}, \tag{12}$$

where $Y_{k+1} = (y_k, y_{k-1}, \ldots, y_0)$ and $S_{k+1} = (s_k, s_{k-1}, \ldots, s_0)$.

The role of the a priori matrix B_{k+1}^0 is to overcome the possible under-determination of problem (11). For example, choosing $B_{k+1}^0 = B_k$ (similarly to classical Broyden-like methods) exhibits good properties. In that case, (12) becomes an update formula, and local convergence can be proved (see Section 4).

The weights ω_{k+1}^i capture the relative importance of each iterate in the population. Roughly speaking, they should be designed in the lines of the assumptions of Taylor's theorem, that is

assigning more weight to points close to x_{k+1}, and less weight to points which are faraway. The matrix Γ captures the importance of the arbitrary terms defined by B^0_{k+1} for the identification of the linear model. The weights have to be finite, and Γ must be such that

$$\Gamma^2 + S_{k+1}\Omega^2 S^T_{k+1} \tag{13}$$

is safely positive definite. To ensure this property we describe below three possible approaches for choosing Γ^2: the *geometrical approach*, based on specific geometric properties of the population, the *subspace decomposition* approach, decomposing $\mathbb{R}^n$ into the subspace spanned by the columns of S_{k+1} and its orthogonal complement, and the *numerical approach*, designed to guarantee a numerically safe positive definiteness of (13).

The *geometrical approach* assumes that $n+1$ members of the population form a simplex, so that the columns of S_{k+1} span $\mathbb{R}^n$, and (13) is positive definite with $\Gamma = 0$. In that case, (12) becomes

$$B_{k+1} = Y_{k+1}\Omega^2 S^T_{k+1}\left(S_{k+1}\Omega^2 S^T_{k+1}\right)^{-1}. \tag{14}$$

If there are exactly $n+1$ iterates forming a simplex, the geometrical approach is equivalent to the interpolation method proposed by Wolfe (1959), and (14) is exactly (10), as S_{k+1} is square and non singular in that case. This approach have not shown good numerical behavior in practice as mentioned in Section 2. Also, it requires at least $n+1$ iterates, and may not be appropriate for large-scale problems.

The *subspace decomposition* approach is based on the QR decomposition of S_{k+1}. We denote by r the rank of S_{k+1}, with $r \leq n$, and we have $S_{k+1} = QR$, where

$$Q = \begin{pmatrix} Q_1 & Q_2 \end{pmatrix} \tag{15}$$

with Q_1 is $(n \times r)$, Q_2 is $(n \times n - r)$, and R is $(n \times k)$. The r columns of Q_1 form an orthogonal basis of the range of S_{k+1}. We define now Γ such that

$$\Gamma^2 = \begin{pmatrix} 0_{r\times n} \\ Q_2^T \end{pmatrix} \tag{16}$$

that is Q^T where Q_1^T has been replaced by a null matrix. With this construction $\Gamma^2 + S_{k+1}\Omega^2 S^T_{k+1}$ is invertible and the weights associated with the arbitrary matrix B^0_{k+1} are null in the subspace generated by the columns of S_{k+1}. In the case where S_{k+1} spans the entire space then $r = n$ and Γ^2 is a null matrix. Consequently, (12) is equivalent to (14).

With the subspace decomposition approach, the changes of F predicted by B_{k+1} in a direction orthogonal to the range of S_{k+1} is the same as the one predicted by the arbitrary matrix B^0_{k+1}. This idea is exactly the same as the one used by Broyden (1965) to construct his so called *Broyden's Good method*. Actually, the classical Broyden update (7) is a special case of our update formula (12), if $B^0_{k+1} = B_k$ and the population contains just two iterates x_k and x_{k+1}. The secant equation (6) completely defines the linear model in the one-dimensional subspace spanned by $s_k = x_{k+1} - x_k$, while an arbitrary decision is made for the rest of the model. If we define $\omega^k_{k+1} = 1$ and Γ^2 is given by (16), we can write (12) as

$$B_{k+1} = B_k + (y_k - B_k s_k)\, s_k^T \left(\Gamma^2 + s_k s_k^T\right)^{-1}. \tag{17}$$

The equivalence with (7) is due to the following equality

$$s_k^T \left(\Gamma^2 + s_k s_k^T\right)^{-1} = s_k^T \frac{1}{s_k^T s_k}, \tag{18}$$

obtained from the fact that $s_k^T \Gamma^2 = 0$, by (16).

Numerical problems may happen when the columns of S_{k+1} are close to linear dependence. These are the problems already mentioned in the introduction, and reported namely by Ortega and Rheinboldt (1970) and Dennis and Schnabel (1996). Clearly, such problems do not occur when S_{k+1} has exactly one column, which leads to the classical Broyden method.

The *numerical approach* is designed to address both the problem of overcoming the underdetermination, and of guaranteeing numerical stability. It is directly inspired by the modified Cholesky factorization proposed by Schnabel and Eskow (1991). The modified Cholesky factorization of a square matrix A creates a matrix E such that $A + E$ is safely positive definite, while computing its Cholesky factorization. It may namely happen that A has full rank, but with smallest eigenvalue very small with regard to machine precision. In that case, E is non zero despite the fact that A is non singular. We apply this technique with $A = S_{k+1}\Omega^2 S_{k+1}^T$ and $E = \Gamma^2$. So, if the matrix $S_{k+1}\Omega^2 S_{k+1}^T$ is safely positive definite, $\Gamma^2 = 0$ and (12) reduces to (14). If not, the modified Cholesky factorization guarantees that the role of the arbitrary term Γ is minimal.

We conclude this section by emphasizing important advantages of our generalization combined with the *numerical approach*. Firstly, contrarily to interpolation methods, our least-square model allows to use more than p points to identify a model in a subspace of dimension p (where $p \leq n$). This is very important when the objective function is expensive to evaluate. Indeed, we make an efficient use of all the available information about the function to calibrate the secant model. It is namely advantageous compared to Broyden's method, where only two iterates are used to build the model. Secondly, the numerical approach proposed above controls the numerical stability of the model construction process, when a sequence of iterates may be linearly dependent. Finally, the fact that existing methods are special cases of our approach allows to exploit all the theoretical and practical properties already published in the literature, and simplifies their extension to our context. We apply this principle is the local convergence analysis in the next section. The main drawback is the increase in numerical linear algebra as the least-square problem (11) must be solved at each iteration.

4 Local convergence analysis

We show that if Γ^2 is determined by the numerical approach described in Section 3, algorithm (4), where B_{k+1} is defined by (12) in his update form (*i.e.* $B_{k+1}^0 = B_k$), locally converges to a solution of (2) if the following assumptions are verified.

(P1) $F : \mathbb{R}^n \to \mathbb{R}^n$ is continuously differentiable in an open convex set $\mathcal{D}$.

(P2) The system of equations has a solution, that is $\exists x^* \in \mathcal{D}$ such that $F(x^*) = 0$.

(P3) $J(x)$ is Lipschitz continuous at x^* with constant K_{lip}, that is

$$\|J(x) - J(x^*)\| \leq K_{lip}\|x - x^*\| \quad \forall x \in \mathcal{D}. \tag{19}$$

in the neighborhood $\mathcal{D}$.

(P4) $J(x^*)$ is non-singular and there exists $\gamma > 0$ such that $\|J(x^*)^{-1}\| < \gamma$.

Assumptions on the algorithm:

(A1) The algorithm is based on the iteration (4) with x_0 and B_0 as initial guess.

(A2) B_k is generated by (12) with $B_{k+1}^0 = B_k$.

(A3) Γ^2 is computed using the *numerical approach*.

(A4) $\forall i \leq k$, we have $\omega_{k+1}^i \leq M_\omega$ for all k.

(A5) The size of the population $\mathcal{P}$ is bounded above by $M_\mathcal{P}$.

The notation $\|\cdot\|$ is used for the l_2 vector norm $\|x\| = (x^T x)^{\frac{1}{2}}$, and for any matrix norm which is consistent with the l_2 norm in the sense that $\|Ax\| \leq \|A\|\|x\|$ for each $x \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n\times n}$. In particular, the l_2 matrix norm and the Frobenius norm are consistent with the l_2 vector norm. For the sake of simplification, we denote $\omega_{k+1}^i = \omega_i$, $S = S_{k+1}$, $Y = Y_{k+1}$ and $I_p = \{1, \ldots, p\}$. The proof uses some lemma. Lemma 1 and 2 are classical results from the literature. Lemma 3–5 are technical lemma related to our method. Their proofs are provided in the appendix.

Lemma 1 *Let $F : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ be continuously differentiable in the open convex $D \subset \mathbb{R}^n$, $x \in D$, and let J be Lipschitz continuous at x in the neighborhood D with constant K_{lip}. Then for any $u, v \in D$,*

$$\|F(v) - F(u) - J(x)(v-u)\| \leq K_{lip}\frac{\|v-x\| + \|u-x\|}{2}\|v-u\|. \tag{20}$$

Proof. See, for example, Dennis and Schnabel, 1996. □

Lemma 2 *Let $A, C \in \mathbb{R}^{n\times n}$ and assume that A is invertible, with $\|A^{-1}\| \leq \mu$. If $\|A - C\| \leq \beta$ and $\beta\mu < 1$, then C is also invertible and*

$$\left\|C^{-1}\right\| \leq \frac{\mu}{1-\beta\mu}. \tag{21}$$

Proof. This lemma is known as the Banach Perturbation Lemma. (See, for example, Ortega and Rheinboldt, 1970). □

Lemma 3 *If assumptions (A4)-(A5) are verified, then*

$$\|S\Omega^2 S^T\| \leq 2M_\mathcal{P} M_\omega^2 \max_{i\in I_{k+1}} \|x_i - x^*\|^2, \tag{22}$$

$$\|\Omega^2 S^T\| \leq \sqrt{2M_\mathcal{P}} M_\omega^2 \max_{i\in I_{k+1}} \|x_i - x^*\|. \tag{23}$$

where x^* is solution of (2).

Lemma 4 *If assumptions (P1),(P2) and (P3) are verified then:*

$$\|(Y - J(x^*)S)\| \leq \sqrt{2M_\mathcal{P}} K_{lip} \max_{i\in I_{k+1}} \left(\|x_i - x^*\|^2\right) \tag{24}$$

where x^* is solution of (2).

Lemma 5 *If assumption (A3) is verified, then*

$$\| \left(\Gamma^2 + S\Omega^2 S^T\right)^{-1} \| \leq \frac{1}{\tau} \tag{25}$$

where $\tau > 0$.

The parameter τ in Lemma 5 controls the may we perturb $S\Omega^2 S^T$. It guarantees that the smallest eigenvalue of $\left(\Gamma^2 + S\Omega^2 S^T\right)$ is strictly greater than τ and, therefore, safely positive in a finite arithmetic context if τ is properly chosen. Schnabel and Eskow (1991) suggest to choose $\tau = (\text{macheps})^{\frac{1}{3}}$ where macheps is the machine epsilon.

Theorem 6 *Let assumptions (P1) to (P3) hold for the problem and assumptions (A1) to (A5) hold for the algorithm. Then there exist two non-negative constants* α_1 *and* α_2 *such that for each* x_k *and* B_k*:*

$$\begin{aligned} \|B_{k+1} - J(x^*)\| \quad \leq \quad & \left(1 + \alpha_1 \max_{i \in I_{k+1}} \|x_i - x^*\|^2\right) \|B_k - J(x^*)\| \\ + \quad & \alpha_2 \max_{i \in I_{k+1}} \|x_i - x^*\|^3 . \end{aligned} \tag{26}$$

Proof. From the update formula (12), and defining

$$\begin{aligned} T_1 &= I - S\Omega^2 S^T (\Gamma^2 + S\Omega^2 S^T)^{-1} \\ T_2 &= (Y - J(x^*)S)\Omega^2 S^T (\Gamma^2 + S\Omega^2 S^T)^{-1}, \end{aligned}$$

we obtain

$$\begin{aligned} \|B_{k+1} - J(x^*)\| &= \|B_k - J(x^*) + \left[(J(x^*)S - J(x^*)S) + (Y - B_k S)\right] \Omega^2 S^T (\Gamma^2 + S\Omega^2 S^T)^{-1}\| \\ &\leq \|T_1\|\|B_k - J(x^*)\| + \|T_2\|. \end{aligned}$$

From Lemma 3 and 5 we obtain

$$\begin{aligned} \|T_1\| &\leq \|I\| + \|S\Omega^2 S^T\|\|(\Gamma^2 + S\Omega^2 S^T)^{-1}\| && (27) \\ &\leq 1 + \alpha_1 \max_{i \in I_{k+1}} \|x_i - x^*\|^2, && (28) \end{aligned}$$

with $\alpha_1 = \frac{2}{\tau} M_{\mathcal{P}} M_\omega^2 > 0$. We conclude the proof using Lemma 3, 4 and 5 to show that:

$$\begin{aligned} \|T_2\| &\leq \|(Y - J(x^*)S)\|\|\Omega^2 S^T\|\|(\Gamma^2 + S\Omega^2 S^T)^{-1}\| && (29) \\ &\leq \alpha_2 \max_{i \in I_{k+1}} \|x_i - x^*\|^3. && (30) \end{aligned}$$

with $\alpha_2 = \frac{2}{\tau} K_{lip} M_{\mathcal{P}} M_\omega^2 > 0$. □

Theorem 7 *Let assumptions (P1) to (P3) hold for the problem and assumptions (A1) to (A5) hold for the algorithm. If, for each* $r \in]0, 1[$ *there exists* $\varepsilon(r)$ *and* $\delta(r)$ *such that*

$$\|x_0 - x^*\| \leq \varepsilon(r) \tag{31}$$

and

$$\|B_0 - J(x^*)\| \leq \delta(r) \tag{32}$$

then the sequence $x_{k+1} = x_k - B_k^{-1} F(x_k)$ *is well defined and converges* q*-linearly to* x^* *with* q*-factor at most* r*. Furthermore, the sequences* $\{\|B_k\|\}_k$ *and* $\left\{\|B_k^{-1}\|\right\}_k$ *are uniformly bounded.*

Proof. The structure of the demonstration is similar to the proof of Theorem 3.2 in Broyden et al. (1973). We have purposedly skipped some identical technical details.

First choose $\varepsilon(r) = \varepsilon$ and $\delta(r) = \delta$ such that

$$\gamma(1+r)\left(K_{lip}\varepsilon + 2\delta\right) \leq r \tag{33}$$

and

$$\left(2\alpha_1 + \alpha_2 \frac{\epsilon}{1-r}\right) \frac{\epsilon^2}{1-r^2} \leq \delta. \tag{34}$$

We invoke Lemma 2 with $\mu = \gamma$ and $\beta = 2\delta$ to prove that B_0 is non-singular and

$$\|B_0^{-1}\| < \gamma(1+r). \tag{35}$$

Note that assumption $2\delta\gamma < 1$ for Lemma 2 is directly deduced from (33).

The improvement after the first iteration, that is

$$\|x_1 - x^*\| \leq r\|x_0 - x^*\| \tag{36}$$

is independent of the specific update formula and, therefore, is proven in Broyden et al. (1973).

The result for iteration k is proven with an induction argument based on the following recurrence assumptions:

$$\|B_m - J^*\| \leq 2\delta \tag{37}$$
$$\|x_{m+1} - x^*\| \leq r\|x_m - x^*\| \tag{38}$$

for all $m = 1, \ldots, k-1$.
We first prove that $\|B_k - J^*\| \leq 2\delta$ using Theorem 6. From (26) we deduce
$\|B_{m+1} - J(x^*)\| - \|B_m - J(x^*)\|$

$$\begin{aligned} &\leq \alpha_1 \max_{i \in I_{m+1}} \|x_i - x^*\|^2 \|B_m - J(x^*)\| + \alpha_2 \max_{i \in I_{m+1}} \|x_i - x^*\|^3 \\ &\leq \alpha_1 r^{2(m+1)} \varepsilon^2 2\delta + \alpha_2 r^{3(m+1)} \varepsilon^3 \end{aligned} \tag{39}$$

Summing both sides of (39) for m ranging from 0 to $k-1$, we deduce that

$$\|B_k - J(x^*)\| \leq \|B_0 - J(x^*)\| + \left(2\alpha_1\delta + \alpha_2 \frac{\varepsilon}{1-r}\right) \frac{\varepsilon^2}{1-r^2} \tag{40}$$
$$\leq 2\delta, \tag{41}$$

where (41) derives from (32) and (34).

The fact that B_k is invertible and $\|B_k^{-1}\| \leq \gamma(1+r)$ is again a direct application of the Banach Perturbation Lemma 2. Following again Broyden et al. (1973), we can now obtain (38) for $m = k$, concluding the induction proof. □

5 Performance evaluation

We present here a preliminary analysis of the performance of our method, in comparison with classical algorithms. All algorithms and test functions have been implemented with the package Octave (see `http://www.octave.org/`) and computations have been done on a laptop equipped with 1066MHz CPU in double precision. The machine epsilon is about 2.2204e-16.

The numerical experiments were carried out on a set of 30 test functions, seven with fix dimension and 23 with dimension $n = 6, 10, 20, 50, 100$. This set is composed of the four standard nonlinear systems of equations proposed by Dennis and Schnabel (1996) (that is, *Extended Rosenbrock Function, Extended Powell Singular Function, Trigonometric Function, Helical Valley Function*), three functions from Broyden (1965), five functions proposed by Kelley (2002) in his new book on Newton's method (that is, *Arctangent Function, a Simple Two-dimensional Function, Chandrasekhar H-equation, Ornstein -Zernike Equations, Right Preconditioned Convection-Diffusion Equation*), three linear systems of equations (see Appendix) and the test functions given by Spedicato and Huang (1997). For each problem, we have used the starting point proposed in the original paper. In order to challenge even more the algorithms, we have also considered $10x_0$ as another starting point for each problem. The total experimental design is therefore composed of 244 problems. The results include only problems which have been solved at least by one of the considered methods, that is 124 problems.

The algorithms that have been considered are based upon the undamped quasi-Newton method, *i.e.* without any step control or globalization methods. This allows us to compare their speed of convergence, in term of number of function evaluations, and their robustness without introducing a bias due to the step control or the globalization method. All the algorithms have the following structure:

- Given $F : \mathbb{R}^n \to \mathbb{R}^n$, $x_0 \in \mathbb{R}^n$ and $B_0 \in \mathbb{R}^{n \times n}$
- While stopping criteria is not verified:
  - Find s solving $B_k s = -F(x_k)$,
  - Evaluate $F(x_{k+1})$ where $x_{k+1} = x_k + s$,
  - Compute B_{k+1},

with the following characteristics: the initial Jacobian approximation B_0 is the same for all algorithms and equal to the identity matrix. The stopping criteria is a composition of three conditions: small residual, that is $\|F(x_k)\|/\|F(x_0)\| \leq 10e-6$, maximum number of iterations ($k \geq 200$ for problems of size $n \leq 20$ and $k \geq 500$ for problems of size $n = 50$ and $n = 100$), and divergence, diagnosed if $\|F(x_k)\| \geq 10e10$.

Consequently, the algorithms differ only by the method to compute B_{k+1}. We consider four of them:

1. Broyden's Good Method (BGM), using the update (7).
2. Broyden's Bad Method (BBM), also proposed by Broyden (1965). It is based on the following secant equation:
$$s_k = B_{k+1}^{-1} y_k. \tag{42}$$

and directly computes the inverse of B_k:

$$B_{k+1}^{-1} = B_k^{-1} + \frac{\left(s_k - B_k^{-1} y_k\right) y_k^T}{y_k^T y_k}. \tag{43}$$

Broyden (1965) describes this method as "bad", that is numerically unstable. However, we have decided to include it in our tests for the sake of completeness. Moreover, as discussed below, it does not always deserve its name.

3. The Hybrid Method (HMM) proposed by Martinez (1982). At each iteration, the algorithm decides to apply either BGM or BBM. Martinez (2000) observes a systematic improvement of the Hybrid approach with respect to each individual approach. As discussed below, we obtain the same result.

4. Our Generalized Secant Method (GSM), defined by (12) in its update form with $B_{k+1}^0 = B_k$ using the *numerical approach* described in Section 3, with $\tau = (\text{macheps})^{\frac{1}{3}}$. Contrarily to what is presented in the theoretical analysis, we consider only $p = \max(n, 10)$ previous iterates in the population. Indeed, including all previous iterates may generate memory management problems, and anyway does not significantly affect the behavior of the algorithm. The weights are defined as

$$\omega_{k+1}^i = \frac{1}{\|x_{k+1} - x_i\|^2} \qquad \forall i \in I_p \tag{44}$$

The measure of performance is the number of function evaluations to reach convergence. We are presenting the results following the performance profiles analysis method proposed by Dolan and More (2002).

If $f_{p,a}$ is the performance index of algorithm a on problem p, then the *performance ratio* is defined by

$$r_{p,a} = \frac{f_{p,a}}{\min_a \{f_{p,a}\}}, \tag{45}$$

if algorithm a has converged for problem p, and $r_{p,a} = r_{\text{fail}}$ otherwise, where r_{fail} must be strictly larger than any performance ratio (45). For any given threshold π, the overall performance of algorithm a is given by

$$\rho_a(\pi) = \frac{1}{n_p} \Phi_a(\pi) \tag{46}$$

where n_p is the number of problems considered, and $\Phi_a(\pi)$ is the number of problems for which $r_{p,a} \leq \pi$.

In particular, the value $\rho_a(1)$ gives the probability that algorithm a wins over all other algorithms. The value $lim_{\pi \to r_{\text{fail}}} \rho_a(\pi)$ gives the probability that algorithm a solves a problem and, consequently, provides a measure of the robustness of each method.

We first present the performance profile for all algorithms described above on all problems on Figure 1. A zoom for π between 1 and 4 is provided in Figure 2.

The results are very satisfactory for our method. Indeed, we observe that GSM is best on nearly 70% of the problems, and is able to solve more than 90% of the problems. From Figure 2, we note also that when GSM is not the best method, it converges within a factor of 1.5 of the best algorithm for more than 80% of the problems. We also confirm results by Martinez (2000)

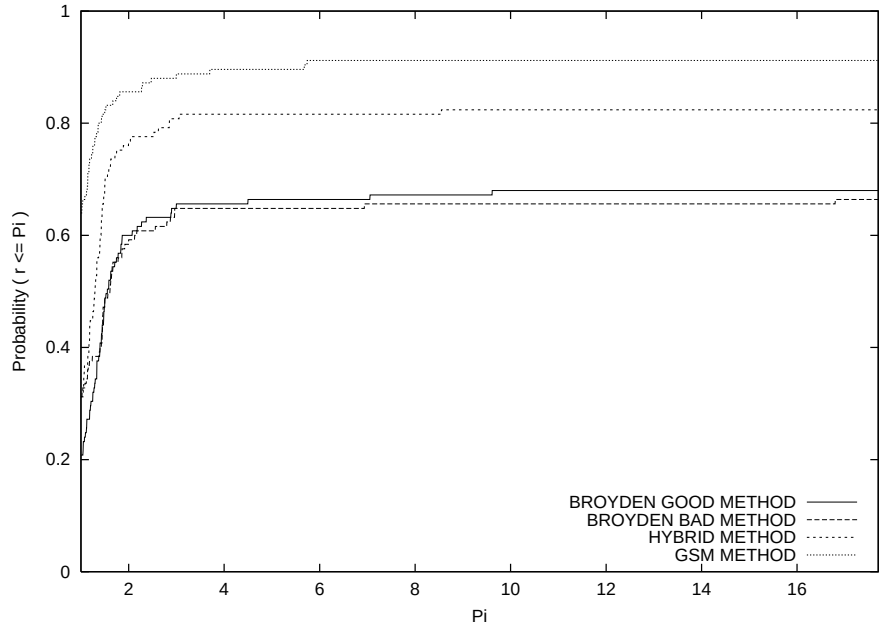


Figure 1: Performance Profile

showing that the Hybrid method is more reliable than BGM and BBM. Indeed, it converges on more than 80% of the problems, while each Broyden method converges only on 60% of the cases. Moreover, HMM wins as often as BBM does, but is more robust, as its performance profile grows faster than the profile for BBM. The relative robustness of BGM and BBM is comparable.

The performance profile analysis strongly depends on the number of methods that are being compared. Therefore, we like to present a comparison between BGM and GSM only, as BGM is probably the most widely used method. The significant improvement provided by our method over Broyden's method is illustrated by Figure 3, based on problems solved by at least one of the two methods, that is 112 problems.

We conclude this section by a preliminary analysis of the behavior of our method in the presence of noise in the function. Indeed, we speculate that the use of a larger sample of iterates to calibrate the secant model smooths the impact of noise on the method.

We consider a generalized noisy nonlinear system of equations described by:

$$G(x) = F_s(x) + \phi(x) \tag{47}$$

where F_s is a smooth nonlinear system of equations and $\phi(x)$ is a random perturbation. Similarly to Choi and Kelley (2000), we assume that the noise decreases near the solution, and we propose

$$\phi(x) \sim N(0, \alpha^2 \|x - x^*\|^2). \tag{48}$$

We consider the problem described in Section 7.4 in the Appendix. We have selected a problem where BGM is better than GSM in the deterministic case. The results for 4 levels of stochasticity are presented in Figure 4. For each value of the parameter α in (48), we plot the relative nonlinear residual, that is $\|G(x_k)\|/\|G(x_0)\|$, against the number of function evaluations.

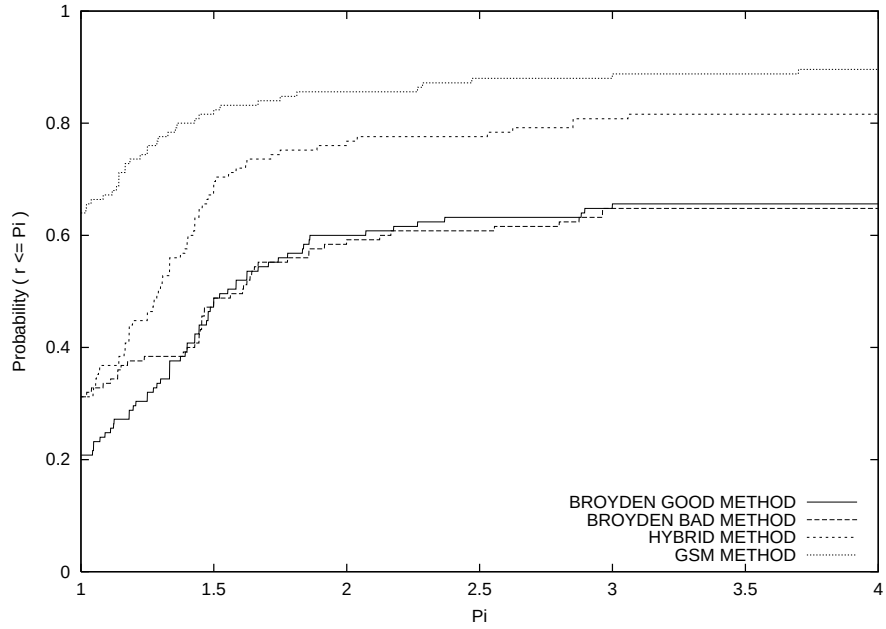


Figure 2: Performance Profile on (1,4)

Figure 4(a) illustrates the deterministic case, with $\phi(x) = 0$, where BGM is slightly better than GSM. When a noise with small variance ($\alpha = 0.001$, Figure 4(b)) is present, GSM decreases the value of the residual pretty quickly, while the descent rate of BGM is much slower. When the variance of the noise increases ($\alpha = 0.05$ in Figure 4(c), and $\alpha = 1$ in Figure 4(d)), the BGM is trapped in higher values of the residual, while GSM achieves a significant decrease. We have performed the same analysis on other problems, and observed a similar behavior, that is a better robustness of GSM when solving a noisy system of equations.

6 Conclusion and perspectives

We have proposed a new family of secant methods, based on the use of more than two iterates to identify the secant model. Contrarily to previous attempts for multi-iterate secant methods, the key ideas of this paper are (i) to use a least-square approach instead of an interpolation method to derive the secant model, and (ii) to explicitly control the numerical stability of the method.

A specific instance of this family of methods provides an update formula. We have provided a proof of the local convergence of a quasi-Newton method based on this update formula. Moreover, we have performed extensive numerical experiences with several algorithms. The results show that our method produces significant improvement in term of robustness and number of function evaluations compared to classical methods. Finally, we have provided preliminary evidences that our method is likely to be more robust in the presence of noise in the function.

The main drawback of our approach is the relatively high cost in numerical linear algebra, as a least-square problem has to be solved at each iteration. This has not been addressed in this paper, as only the number of function evaluations is under consideration, but it may become cumbersome for large problems. Several techniques can be considered to improve the perfor-

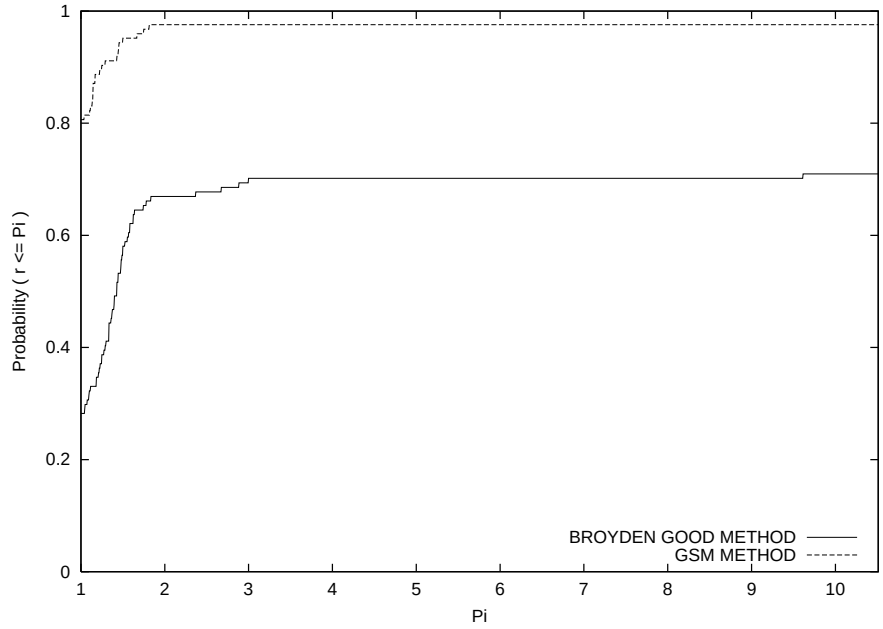


Figure 3: Performance profile – Broyden Good Method and GSM –

mance, like imposing a specific structure to B_{k+1} (e.g. tridiagonal) in order to reduce the size of the least-square problem. Also, techniques inspired by limited memory update can be analyzed in this context as well.

A theoretical analysis of a globally convergent version of our method must also be performed. We also conjecture that the local convergence rate is super-linear. And most importantly, the general behavior of the algorithm for solving noisy functions requires further analysis.

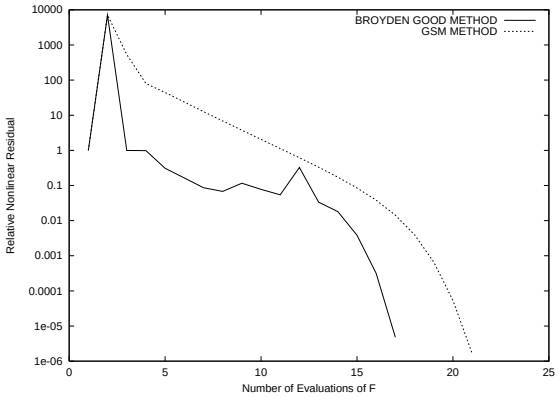
There are several variants of our methods that we plan to analyze in the future. Firstly, following Broyden's idea to derive BBM from (42), an update formula for B_{k+1}^{-1} can easily be derived in the context of our method:

$$B_{k+1}^{-1} = B_k^{-1} + \left(\Gamma^2 + Y_{k+1}\Omega^2 Y_{k+1}^T\right)^{-1} Y_{k+1}^T \Omega^2 \left(S_{k+1} - B_k^{-1} Y_{k+1}\right). \tag{49}$$

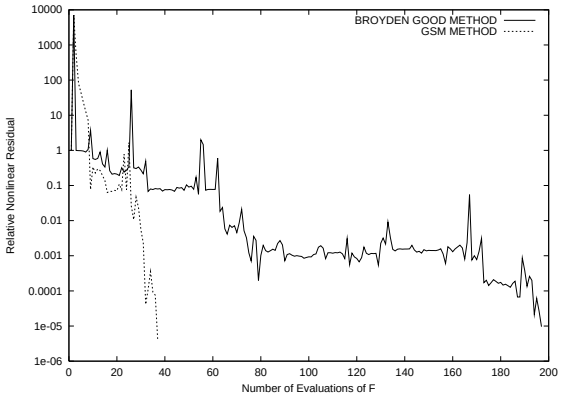
From preliminary tests that we have performed, the "Good" and "Bad" versions of our method compare in a similar way as BGM and BBM.

Secondly, non-update instances of our class of methods can be considered. In that case, the arbitrary matrix B_{k+1}^0 in (11) may be different from B_k. In that case, choosing a matrix independent from k would allow to apply Kalman filtering (Kalman, 1960) to incrementally solve (11) and, consequently, improve the numerical efficiency of the method. For large scale problems, an iterative scheme such as LSQR (Paige and Saunders, 1982) has been considered. LSQR can also improve the efficiency of Kalman filter for the incremental algorithm (see Bierlaire and Crittin, 2001).

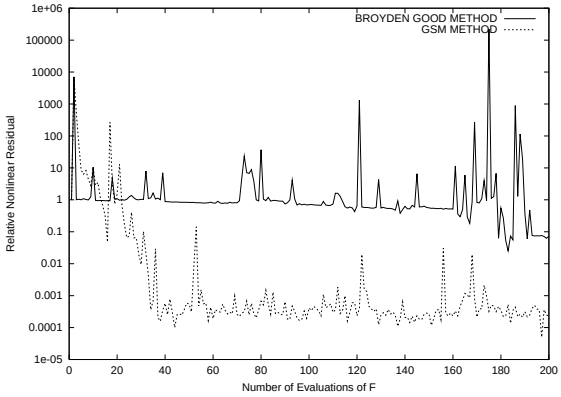
In (Bierlaire and Crittin, 2003) we propose an adaptation of GSM to solve the consistent anticipatory route guidance problem following the propositions described above. Results are impressive as this algorithm clearly outperforms classical fixed point methods. Moreover preliminary results also indicate a very good behavior on nonlinear systems of equations compared to state-



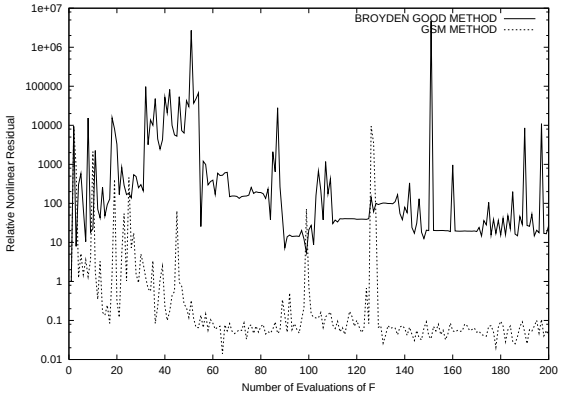
(a) Without noise



(b) Small variance noise



(c) Normal variance noise



(d) Big variance noise

Figure 4: Behavior with stochasticity

of-the-art large scale algorithms.

Finally, the ideas proposed in this paper can easily be tailored to optimization problems, where the symmetry and positive definiteness of the matrix can be explicitly exploited.

7 Appendix

7.1 Proof of Lemma 3

$$\begin{aligned}
\|S\Omega^2 S^T\| &\leq \|S\Omega\|^2 && (50)\\
&\leq \sum_{i=1}^{k} \|\omega_i s_i\|^2 && (51)\\
&\leq k \max_{i\in I_k} \left(|\omega_i|\|s_i\|\right)^2 && (52)\\
&\leq k \max_{i\in I_k} \left(|\omega_i|\|x_{k+1} - \tilde{x} + \tilde{x} - x_i\|\right)^2 && (53)\\
&\leq 2k \max_{i\in I_k} |\omega_i|^2 \max_{i\in I_{k+1}} \|x_i - \tilde{x}\|^2 && (54)\\
&\leq 2M_{\mathcal{P}} M_\omega^2 \max_{i\in I_{k+1}} \|x_i - \tilde{x}\|^2 && (55)
\end{aligned}$$

for all $\tilde{x} \in \mathbb{R}^{n\times n}$, in particular with $\tilde{x} = x^*$ which proves (22).

$$\begin{aligned}
\|\Omega^2 S^T\|^2 &\leq \sum_{i=1}^{k} \|\omega_i^2 s_i\|^2 && (56)\\
&\leq k \max_{i\in I_k} \left(|\omega_i|^2\|s_i\|\right)^2 && (57)\\
&\leq k \max_{i\in I_k} |\omega_i|^4 \max_{i\in I_k} \|x_{k+1} - \tilde{x} + \tilde{x} - x_i\|^2 && (58)\\
&\leq 2k \max_{i\in I_k} |\omega_i|^4 \max_{i\in I_{k+1}} \|x_i - \tilde{x}\|^2 && (59)
\end{aligned}$$

for all $\tilde{x} \in \mathbb{R}^{n\times n}$. We obtain (23) with $\tilde{x} = x^*$:

$$\|\Omega^2 S^T\| \leq \sqrt{2M_{\mathcal{P}}} M_\omega^2 \max_{i\in I_{k+1}} \|x_i - x^*\| \quad (60)$$

7.2 Proof of Lemma 4

Writing explicitly a column of the matrix $A = Y - J(x^*)S$

$$a_{.j} = F(x_{k+1}) - F(x_i) - J(x^*)(x_{k+1} - x_i) \quad (61)$$

with $a_{.j}$ defining the column j of $A = (a_{ij})$.
Using (61) and Lemma 1 we can write:

$$\begin{aligned}
\|Y - J(x^*)S\|^2 &\leq \sum_{j=1}^{k} \|a_{.j}\|^2 && (62)\\
&\leq k \max_{i \in I_k} \|F(x_{k+1}) - F(x_i) - J(x^*)(x_{k+1} - x_i)\|^2 && (63)\\
&\leq k K_{lip}^2 \max_{i \in I_k} \left(\frac{\|x_i - x^*\| - \|x_{k+1} - x^*\|}{2} \|x_{k+1} - x_i\| \right)^2 && (64)\\
&\leq 2k K_{lip}^2 \max_{i \in I_{k+1}} \|x_i - x^*\|^2 \max_{i \in I_{k+1}} \|x_i - x^*\|^2 && (65)
\end{aligned}$$

Taking the square root on both side:

$$\|Y - J(x^*)S\| \leq \sqrt{2M_{\mathcal{P}}} K_{lip} \max_{i \in I_{k+1}} \|x_i - x^*\|^2 \quad (66)$$

7.3 Proof of the Lemma 5

Let $A \in \mathbb{R}^{n \times n}$, we denote by $\lambda_m(A)$ and $\lambda_M(A)$ its smallest and largest eigenvalues, respectively. So we can write using the definition of the l_2 norm:

$$\begin{aligned}
\|(\Gamma^2 + S\Omega^2 S^T)^{-1}\| &= \lambda_M((\Gamma^2 + S\Omega^2 S^T)^{-1}) && (67)\\
&= \frac{1}{\lambda_m(\Gamma^2 + S\Omega^2 S^T)}. && (68)
\end{aligned}$$

Froms assumption (A3), Γ^2 is computed using the modified Cholesky factorization, proposed by Schnabel and Eskow (1991), with parameter τ. Therefore,

$$\lambda_m(\Gamma^2 + S\Omega^2 S^T) \geq \tau, \quad (69)$$

which concludes the proof.

7.4 Description of problem analyze in Figure 4

The considered problem is the following system of equations:

$$f_i = x_i - \frac{\sum_{j=1}^{4} x_j^3 + 1}{8} \qquad i = 1, \ldots, 4 \quad (70)$$

with initial point $x_0 = (1.5, \ldots, 1.5)$. The solution of this system is $x^* = (0.20432, \ldots, 0.20432)$.

7.5 Linear problems in the tests set

We have tested three linear problems of the form $Ax = b$. They have been designed to challenge the tested algorithms.

1. For the first, the matrix A is the Hilbert matrix, and vector b is composed of all ones.

2. The second problem is based on the matrix A such that $a_{ij} = j$ if $i + j = n + 1$, and $a_{ij} = 0$ otherwise. All entries of the right-hand side b are -10. Its structure is designed so that the identitiy matrix is a poor approximation.

3. The third problem is based on a Vandermond matrix $A(v)$ with $v = (-1, -2, \ldots, -n)$. All entries of the right-hand side b are -1.

The starting point for all those problems is $x = (1, \ldots, 1)^T$.

References

Banach, S. (1922). Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales, *Fund. Math.* **3**: 133–181.

Bierlaire, M. and Crittin, F. (2001). An efficient algorithm for real-time estimation and prediction of dynamic OD tables, *Technical Report RO-010808*, Swiss Insitute of Technology, Lausanne, ROSO-DMA-EPFL CH-1015 Lausanne.

Bierlaire, M. and Crittin, F. (2003). Solving the anticipatory route guidance generation problem using a generalization of secant method, *Proceedings of the 3rd Swiss Transportation Research Conference*, Ascona, Switzerland.

Bottom, J. (2000). *Consistent Anticipatory Route Guidance*, PhD thesis, Massachusetts Institute of Technology.

Broyden, C. G. (1965). A class of methods for solving nonlinear simultaneous equations, *Mathematics of Computation* **19**: 577–593.

Broyden, C. G., Dennis, J. E. and Moré, J. J. (1973). On the local and superlinear convergence of quasi-Newton methods, *Journal of the Institute of Mathematics and its Applications* **12**: 233–246.

Cantarella, G. E. (1997). A general fixed-point approach to multimode multi-user equilibrium assignement with elastic demand, *Transportation Science* **31**(2): 107–128.

Cascetta, E. and Postorino, M. (2001). Fixed point appoaches to the estimation of od matrices using traffic counts on congested network, *Transportation Science* **35**: 134–147.

Choi, T. D. and Kelley, C. (2000). Superlinear convergence and implicit filtering, *Siam Journal of Optimization* **10**.

Dennis, J. E. and Schnabel, R. B. (1996). *Numerical methods for unconstrained optimization and nonlinear equations*, Society for Industrial and Applied Mathematics.

Dolan, E. and More, J. (2002). Benchmarking optimization software with performance profiles, *Mathematical Programming, Serie A* **91**: 2.

Ford, J. A. (1999). A survey of multi-step quasi-newton methods, *Proceedings of the International Conference on Scientific Computations*, Beirut, Lebanon.

Ford, J. and Moghrabi, I. (1997). Alternating multi-step quasi-newton methods for unconstrained optimization, *Journal of Computational and Applied Mathematics* **82**: 105–116.

Gragg, W. and Stewart, G. (1976). A stable variant of the secant method for solving nonlinear equations, *SIAM Journal on Numerical Analysis* **13**: 889–903.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems, *J. of Basic Eng., Trans. ASME, Series D* **82**(1): 33–45.

Kelley, C. T. (2002). *Solving nonlinear equations with Newton's method*, Draft stage. http://www4.ncsu.edu/ ctk/newton.html.

Martinez, J. M. (1979). Three new algorithms based on the sequantial secant method, *BIT* **19**: 236–243.

Martinez, J. M. (1982). Sobre dois metodos de broyden, *Matematica Aplicada e Computational* **1**.

Martinez, J. M. (2000). Practical quasi-newton methods for solving nonlinear systems, *Journal of Computational and Applied Mathematics* **124**: 97–122.

Moghrabi, I. (1993). *Multi-step quasi-Newton methods for optimization*, PhD thesis, University of Essex, United Kingdom.

Ortega, J. M. and Rheinboldt, W. C. (1970). *Iterative solution of nonlinear equations in several variables*, Academic Press, New York.

Paige, C. C. and Saunders, M. A. (1982). LSQR: an algorithm for sparse linear equations and sparse least squares, *ACM Transactions on Mathematical Software* **8**: 43–71.

Patriksson, M. (1994). *The Traffic Assignment Problem, Models and Methods*, VSP, Utrecht, NL.

Schnabel, R. B. and Eskow, E. (1991). A new modified Cholesky factorization, *SIAM Journal on Scientific and Statistical Computing* **11**: 1136–1158.

Spedicato, E. and Huang, Z. (1997). Numerical experience with newton-like methods for nonlinear algebraic systems, *Computing* **58**: 69–99.

Wolfe, P. (1959). The secant method for solving nonlinear equations, *Communications ACM* **12**: 12–13.