Work Control for Newton Type Coupling

Jürgen Menck

Report 25 February 1999

Technische Universität Hamburg – Harburg
Arbeitsbereich Mathematik, Schwarzenbergstr. 95, D – 21073 Hamburg

Work Control for Newton Type Coupling

Jürgen Menck

Technische Universität Hamburg-Harburg, Arbeitsbereich Mathematik,
Kasernenstraße 12, D-21073 Hamburg, menck@tu-harburg.de,
http://www.tu-harburg.de/mat/

Abstract. For an overview of Coupled Systems, please refer to "Coupling Iterative Subsystem Solvers" [7]. The present paper can be viewed as a related detail study.

We elaborate on the obvious idea of coupling existing subsystem solvers to solve coupled systems of stationary equations. More specifically, we present and analyze a matrix-free iterative method that is inspired by Newton type coupling but aims at efficiently controlled *linear* convergence.

1 Introduction

Complex technical systems are often assembled by coupling well-known subsystems together (cf. [7]). In the present paper, we propose a matrix-free iterative solver for coupled systems that handles all the subsystems by use of their respective solvers and still guarantees local convergence under quite general assumptions; in other words, it knows how to handle the coupling. We suggest a control mechanism that will optimize the efficiency of the algorithm by relating the error reduction to the required computational effort. Details of the actual implementation are given. A more exhaustive version of the present paper has been published as a technical report [8]. It includes the detailed proofs as well as additional explanations and more material on the numerical example.

The proposed algorithm is based on a well-known concept which we will call T(angential) B(lock) N(ewton). Basically, TBN tries to approximate Newton steps for the coupled system while retaining the given block structure. Tony F. Chan ([3], [4]) seems to have been the first to apply the TBN concept to our basic problem. His method, however, differs significantly from ours in some respects, and a proof of convergence required some rather restrictive assumptions. The present paper rather relates to the work of Artlich and Mackens, e. g. [2], although the emphasis on *linear* convergence and work control is a novel feature.

Some of the following material overlaps with [7] due to our wish to keep the papers technically self-contained.

2 Problem

Suppose we are given $k \in \mathbb{N}$ subsystems, each depending on a set $x_i \in \mathbb{R}^{k_i}$ of k_i internal variables and a set $y \in \mathbb{R}^{k_c}$ of common external (or "coupling")

variables. Each system is represented by its respective solver, which is assumed to be an iterative process $x_i^{n+1} := \Phi_i(x_i^n, y)$. (Note that a direct solver will also qualify.) We merge the subsystems into a large system

$$x = \Phi(x, y) , \quad x = (x_1, \ldots, x_k) , \quad \Phi = (\Phi_1, \ldots, \Phi_k) \tag{1}$$

A set of equations $g(x, y) = 0$ will represent the coupling of the units. Setting

$$f(x, y) := x - \Phi(x, y) , \tag{2}$$

we can formulate our problem as a root finding problem:

$$f(x, y) = 0 , \quad g(x, y) = 0 . \tag{3}$$

We assume that g consists of exactly as much equations as y has components, such that the composite system is "square". The aim of our method will be to reduce suitable norms of the residual errors in (3). Thus we end up with an optimization problem,

$$\max(\|f(x, y)\|, \|g(x, y)\|) = \min! . \tag{4}$$

Suppose our starting point lies in the vicinity of a solution $(\hat{x}, \hat{y})$ of (3) at which the joint Jacobian of f and g is regular. Intermediate iterates of our method may move away from the solution a bit but our estimates will make it easy to contain them inside a suitable region U if the starting point is "good enough". We assume that f and g are at least $\mathcal{C}^2$ inside U, that their joint Jacobian remains nonsingular and that Φ is uniformly contractive in the sense that $\|D_x\Phi(x, y)\| \leq q < 1$ holds for all (x, y).

3 Exact Tangential Block-Newton

T(angential) B(lock) N(ewton) (our own name) is basically a blocked Newton's iteration for solving coupled systems of the form (3), cf. [7]. One step consists of a Newton step for f in x,

$$\Delta x := -D_x f(x_n, y_n) f(x_n, y_n) , \tag{5}$$
$$x^+ := x_n + \Delta x , \tag{6}$$

and a Newton step for g along the tangential space of the manifold $M := \{(x, y) \mid f(x, y) = f(x^+, y_n)\}$ at (x^+, y_n),

$$\Delta y := -S^{-1}(x^+, y_n) g(x^+, y_n) , \tag{7}$$
$$x_{n+1} := x^+ - C\Delta y , \tag{8}$$
$$y_{n+1} := y_n + \Delta y . \tag{9}$$

The above matrices C and S are defined as follows:

$$C(x,y) := (D_x f(x,y))^{-1} D_y f(x,y) \tag{10}$$
$$S(x,y) := -D_x g(x,y) C(x,y) + D_y g(x,y) \tag{11}$$

C is the "correction" matrix that generates the tangential directions of M in the sense that they are all of the form $(-C\Delta y, \Delta y)$. The matrix S is the total derivative of g with respect to these directions.

4 Approximate Tangential Block-Newton

Our A(pproximate) TBN has five parameters,

$$0 < \alpha, \beta \leq 1 \ , \quad \kappa_1, \kappa_2 \in \mathbb{N} \ , \quad 0 < \varepsilon_1 \ll 1 \ .$$

α and β are damping parameters, κ_1 denotes the number of successive Φ iterations in the f step, κ_2 denotes the number of Φ iterations used for approximating C, and ε_1 is the error bound for the relative residuum of the linear equation associated with (7). The f step takes the form

$$\Delta x := \Phi^{\kappa_1}(x_n, y_n) - x_n \ , \tag{12}$$
$$x^+ := x_n + \alpha \Delta x \ , \tag{13}$$

and the g step becomes

$$\text{Find } \Delta y \text{ s.t. } \|\tilde{S}(x^+, y_n)\Delta y + g(x^+, y_n)\| \leq \varepsilon_1 \|g(x^+, y_n)\| \ , \tag{14}$$
$$x_{n+1} := x^+ - \beta \tilde{C} \Delta y \ , \tag{15}$$
$$y_{n+1} := y_n + \beta \Delta y \ . \tag{16}$$

(14) can be solved using BiCGStab or some other transpose free iterative solver. $\tilde{C}$ and $\tilde{S}$ indicate approximations of the matrices C and S:

$$\tilde{C}(x,y) := \sum_{i=0}^{\kappa_2} (D_x \Phi(x,y))^i D_y f(x,y) \ , \tag{17}$$
$$\tilde{S}(x,y) := -D_x g(x,y) \tilde{C}(x,y) + D_y g(x,y) \ . \tag{18}$$

$\tilde{C}$ is derived from C by replacing f_x^{-1} by a truncated Neumann series, and $\tilde{S}$ is the resulting approximation of S. Note that the above expressions are only intended for analytical use, *not* for use in the actual implementation.

5 Analysis of the ATBN Step

In the sequel, f_n and g_n will denote the values of f and g at (x_n, y_n), f^+ and g^+ will denote the values at (x^+, y_n), and m_n will be the maximum of the

residua, $m_n = \max(\|f_n\|, \|g_n\|)$. We now define

$$\mu := \sup_{(x,y)\in U} \|D_x g(x,y)\| \,, \tag{19}$$

$$\lambda := \sup_{(x,y)\in U} \|D_y f(x,y)\| \cdot \|\tilde{S}^{-1}(x,y)\| \,. \tag{20}$$

μ is a measure for the sensitivity of $\|g\|$ w. r. t. changes of x. If f and g were decoupled, μ would be 0. Analogously, λ can be viewed as a measure for the sensitivity of $\|f\|$ w. r. t. certain variations associated with the g step. It can be shown that the following estimates hold:

$$\|f^+\| \le ((1-\alpha) + \alpha q^{\kappa_1})\|f_n\| + \alpha \mathcal{O}(\|f_n\|^2) \,, \tag{21}$$

$$\|g^+\| \le \|g_n\| + \alpha\mu \frac{1-q^{\kappa_1}}{1-q}\|f_n\| + \alpha^2 \mathcal{O}(\|f_n\|^2) \,, \tag{22}$$

$$\|f_{n+1}\| \le \|f^+\| + \beta(1+\varepsilon_1) q^{\kappa_2+1}\lambda\|g^+\| + \beta^2 \mathcal{O}(\|g^+\|^2) \,, \tag{23}$$

$$\|g_{n+1}\| \le ((1-\beta) + \beta\varepsilon_1)\|g^+\| + \beta^2 \mathcal{O}(\|g^+\|^2) \,. \tag{24}$$

Thus up to first order terms the undamped f step will reduce $\|f\|$ by a factor of q^{κ_1}, and the undamped g step will reduce $\|g\|$ by a factor of ε_1. The f step may have a negative influence on the norm of g and vice versa, but in the case of the g step the importance of that influence can be lessened by increasing κ_2. For $\kappa_1, \kappa_2 \to \infty$ and $\varepsilon_1 \to 0$ the TBN step will be retrieved.

It can be shown that if for given α, κ_1 and κ_2 the parameter β is chosen appropriately, the following estimate will hold:

$$\frac{m_{n+1}}{m_n} \le \max(q_\alpha, q_{\min}) + \alpha^2 \mathcal{O}(m_n) \,, \quad \text{where} \tag{25}$$

$$q_\alpha = (1-\alpha) + \alpha\left(q^{\kappa_1} + q^{\kappa_2} \cdot 2q\lambda \frac{(1+\mu)^2}{(1-q)^2(1-\varepsilon_1)}\right) \,, \tag{26}$$

$$q_{\min} = \frac{\varepsilon_1}{1-\varepsilon_1}\frac{(1+\mu)^2}{(1-q)^2} \,. \tag{27}$$

Our convergence theorem is a direct consequence of this estimate:

Theorem 1. *(Convergence of ATBN) Let the assumptions of sect. 2 hold true.*

a) ATBN can be made to converge linearly with any given convergence rate $0 < q_{comp.} < 1$. More precisely, there is a neighbourhood of the solution such that for sufficiently large κ_1, κ_2 and sufficiently small ε_1

$$\frac{m_{n+1}}{m_n} \le (1-\alpha) + \alpha q_{comp.} \tag{28}$$

will hold for all $0 \le \alpha \le 1$ and suitably chosen $\beta = \beta(\alpha)$.

b) ATBN can be made to achieve an effective *convergence rate* $q \leq q_{\mathit{eff.}} < 1$ *in the following sense: Assume that the iterative solver always succeeds in solving the modified* g *equation (14) and that the number of iteration steps needed is bounded for any given* ε_1*. Let* κ *denote the number of* Φ *evaluations needed to compute the ATBN step. Then there is a neighbourhood of the solution and a* $q \leq q_{\mathit{eff.}} < 1$ *such that for* $\alpha = 1$ *and suitably chosen* κ_1*,* κ_2*,* ε_1 *and* β *the ATBN step will satisfy*

$$\left(\frac{m_{n+1}}{m_n}\right)^{1/\kappa} \leq q_{\mathit{eff.}} \,. \tag{29}$$

6 Practicalities

The present section is dedicated to the actual implementation of ATBN.

6.1 Matrix Implementation

First of all, we suppose that we attack (14) with a transpose-free Krylov subspace method like GMRES ([10]), BiCGStab ([11]) or TFQMR ([6]). These methods, although hard to analyze in the case of nonsymmetric system matrices, are widely popular and well-respected. By using them, we can restrict ourselves to computing matrix-vector products of $\tilde{C}$ and $\tilde{S}$. We propose to do this via the following differencing schemes, which follow from interpreting the products as directional derivatives:

$$\tilde{C}(x,y)w \simeq \Psi_w^{\kappa_2+1}(0;x,y)\,, \quad \text{where} \tag{30}$$

$$\Psi_w(r;x,y) := \frac{\Phi(x+h_2 r, y) - \Phi(x,y)}{h_2} + \frac{f(x, y+h_1 w) - f(x,y)}{h_1}\,, \tag{31}$$

$$\tilde{S}(x,y)w \simeq \frac{g(x - h_3\tilde{C}(x,y)w, y + h_3 w) - g(x,y)}{h_3}\,. \tag{32}$$

The h_i are supposed to be small numbers suited to approximate the implied directional derivatives. They should be of the order of $\sqrt{\text{macheps}}$, cf. [5].

6.2 Parameters

We will now develop the control mechanism for ATBN's method parameters. In keeping with the previous sections, we shall concentrate on the local convergence behaviour of the method. The control mechanism may be enhanced by including special strategies for certain special situations: For example, the cases $\|f\| \ll \|g\|$ and $\|f\| \gg \|g\|$ can be treated separately by executing only the g or the f step, respectively.

1. $\alpha = 1$ unless complications occur. Details of damping will be postponed to a forthcoming paper.

2. ε_1 is given by the user. κ_1 and κ_2 will adapt to ε_1 but as a rule of thumb, ε_1 should be chosen small but not *too* small to avoid unnecessary costs in computing (14).
3. We supply a formula to choose an optimal β provided all the other parameters are fixed.
4. We eliminate β from the model by inserting this optimal value. Thus we can consider the estimated effective reduction factor (29) as a function of κ_1 and κ_2. We minimize this function.

Some specifics: According to our estimates, the optimal choice of β will be $\beta = \min(1, \beta_*)$, where β_* is the value of β for which the estimates of $\|f_{n+1}\|$ and $\|g_{n+1}\|$ coincide. If a double "+" denotes the values for $\beta = 1$, we have approximately

$$\beta_* = \frac{\|g^+\| - \|f^+\|}{(\|f^{++}\| - \|f^+\|) - (\|g^{++}\| - \|g^+\|)} \,. \tag{33}$$

By using our estimates and supposing β is set as said, we get

$$\|f^+\| \simeq ((1-\alpha) + \alpha q^{\kappa_1})\|f_n\| \,, \tag{34}$$

$$\|g^+\| \simeq \|g_n\| + \alpha\mu \frac{1-q^{\kappa_1}}{1-q}\|f_n\| \,, \tag{35}$$

$$m_{n+1} \simeq \max(\varepsilon_1\|g^+\|, m^*_{n+1}) \,, \quad \text{where} \tag{36}$$

$$m^*_{n+1} := \frac{(1-\varepsilon_1)\|f^+\| + (1+\varepsilon_1)\lambda q^{\kappa_2+1}\|g^+\|}{(1-\varepsilon_1) + (1+\varepsilon_1)\lambda q^{\kappa_2+1}} \,. \tag{37}$$

To supply the constants λ, μ and q, we can exploit the previous step:

$$q \simeq \left(\frac{\|f^+\| - (1-\alpha)\|f_n\|}{\alpha\|f_n\|}\right)^{1/\kappa_1} \,, \tag{38}$$

$$\lambda \simeq \frac{|\,\|f^{++}\| - \|f^+\|\,|}{\beta q^{\kappa_2+1}(1+\varepsilon_1)\|g^+\|} \quad \text{(with the above } q \text{ estimate)}, \tag{39}$$

$$\mu \simeq \frac{|\,\|g^+\| - \|g_n\|\,|}{\alpha\|\Delta x\|} \,. \tag{40}$$

We suppose that the number ℓ of (outer) steps that the linear solver needs to satisfy (14) remains approximately constant, which enables us to estimate the number κ of Φ evaluations belonging to a given pair (κ_1, κ_2). A typical number for a BiCGStab based approach would be

$$\kappa \simeq (3 + \kappa_1) + 2(\ell + 1)(\kappa_2 + 1) \,. \tag{41}$$

Using (36), we can minimize the effective reduction (29) for $1 \le \kappa_1, \kappa_2 \le \kappa_{\max}$, where $\kappa_{\max}$ is some upper bound for the admissible number of Φ iterations in a row.

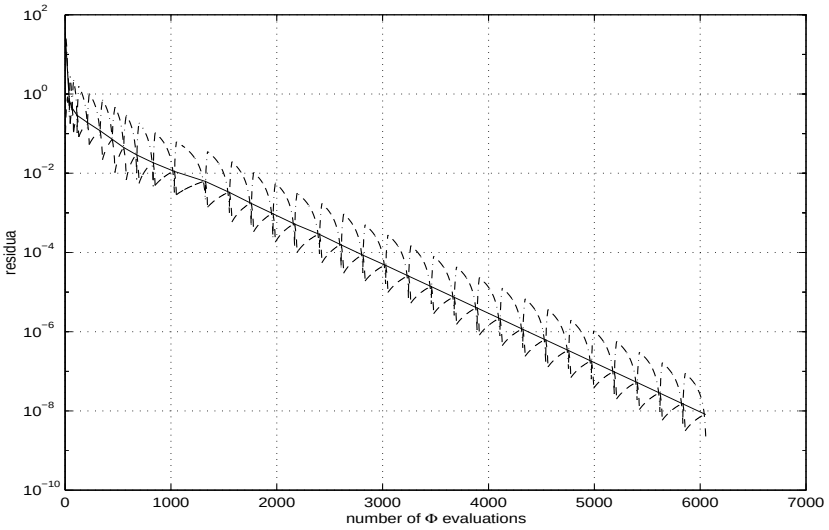
7 A Numerical Example

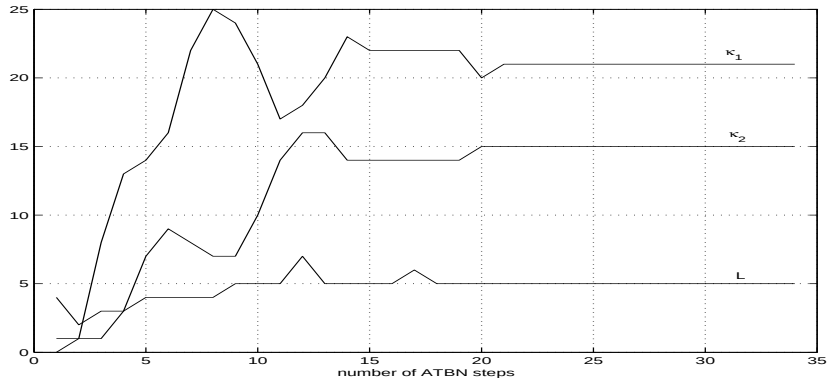
As an example we solve the Bratu problem on a substructured unit square,

$$\begin{aligned} -\Delta u(\xi_1,\xi_2) &= \sigma \exp(u(\xi_1,\xi_2)) \quad \text{for} \quad 0 < \xi_1, \xi_2 < 1 \,, \\ u(\xi_1,\xi_2) &= 0 \qquad\qquad\qquad\quad \text{on the boundary} \,. \end{aligned} \tag{42}$$

In terms of Chemical Engineering, (42) can be viewed as a model problem for zeroth-order exothermic reaction (cf. [1]). We divide the unit square into four identical squares; this substructuring will naturally produce a coupled system if the interior nodes of the subsquares are interpreted as interior variables and the common nodes as coupling variables. We treat the parameter σ (the so-called Thiele modulus) as a further coupling variable and introduce the equation $u(0.5, 0.5) = u_{\max}$ with a prespecified $u_{\max}$ to compensate for it. By separating the diagonal part of the discretized Laplacian from the rest of the equations, we obtain an obvious Jacobi type iteration process Φ.

The figures represent a test run for $\varepsilon_1 = 0.1$ and 225 interior grid points. After a startup phase the method stabilizes itself at $\kappa_1 = 21 \pm 1$ and $\kappa_2 = 15 \pm 1$. As we assumed, the number ℓ of BiCGStab iteration steps remains almost constant after the startup phase. The dashed line in the "residua" picture indicates the values of $\|f\|$, the dash-dotted one the values of $\|g\|$. Both interpolate linearly between three critical phases of each ATBN step: a) before the f step, b) before the g step, c) after a virtual *undamped* g step. The solid line connects the values of $\max(\|f\|, \|g\|)$ before and after the actual composite ATBN steps. It can be used to measure the success of the step: the steeper the slope, the better the step.





Acknowledgements. The author thanks Wolfgang Mackens for his fruitful suggestions and his helpful criticism.

References

1. Aris, R.: *The Mathematical Theory of Diffusion and Reaction in Permeable Catalysts, Vol.1: The Theory of the Steady State*, Clarendon Press: Oxford (1975).
2. Artlich, S. and Mackens, W.: Newton-Coupling of Fixed Point Iterations, in *Numerical Treatment of Coupled Systems*, W. Hackbusch and G. Wittum (eds), Vieweg-Verlag: Braunschweig, Wiesbaden (1995), 1–10.
3. Chan, T. F.: An Approximate Newton Method for Coupled Nonlinear Systems, *SIAM J. Numer. Anal.* **22** (1985), 904–913.
4. Chan, T. F.: An Efficient Modular Algorithm for Coupled Nonlinear Systems, *Springer Lect. Notes in Math.* **1230**, Springer-Verlag: Berlin (1986), 73–85.
5. Dennis, J. E. and Schnabel, R. B.: *Numerical Methods for Unconstrained Optimization and Nonlinear Equations* Prentice-Hall, Inc.: Englewood Cliffs (1983).
6. Freund, R. W.: A Transpose-Free Quasi-Minimal Residual Algorithm for Non-Hermitian Linear Systems, *SIAM J. Sci. Comput.* **14** (1993), 470–482.
7. Mackens, W., Menck, J. and Voß, H.: *Coupling Iterative Subsystem Solvers*, This Volume (1999).
8. Menck, J.: *An Approximate Newton-Like Coupling of Subsystems*, Report 21, Arbeitsbereich Mathematik, TU Hamburg-Harburg (1998).
9. Nachtigal, N. M., Reddy, S. C. and Trefethen, L. N.: How Fast are Nonsymmetric Matrix Iterations?, *SIAM J. Matrix Anal. Appl.* **13** (1992), 778–795.
10. Saad, Y. and Schultz, M. H.: GMRES: A Generalized Minimum Residual Algorithm for Solving Nonsymmetric Linear Systems, *SIAM J. Sci. Statist. Comp.* **7** (1986), 856–869.
11. van der Vorst, H. A.: Bi-CGstab: A Fast and Smoothly Converging Variant of Bi-CG for the Solution of Nonsymmetric Linear Systems, *SIAM J. Sci. Statist. Comp.* **13** (1992), 631–644.