Time-lag in Derivative Convergence for Fixed Point Iterations*

Andreas Griewank and Andrej Ponomarenko

Abstract

In an earlier study it was proven and experimentally confirmed on a 2D Euler code that fixed point iterations can be differentiated to yield first and second order derivatives of implicit functions that are defined by state equations. It was also asserted that the resulting approximations for reduced gradients and Hessians converge with the same R-factor as the underlying fixed point iteration.

A closer look reveals now that nevertheless these derivative values lag behind the functions values in that the ratios of the corresponding errors grow proportional to the iteration counter or its square towards infinity. This rather subtle effect is caused mathematically by the occurrence of nontrivial Jordan blocks associated with degenerate eigenvalues. We elaborate the theory and report its confirmation through numerical experiments.

Résumé

Une étude antérieure a prouvé et vérifié expérimentalement sur un code Euler 2D que les calculs itératifs avec point fixe peuvent être différentiés pour obtenir les dérivées aux premier et deuxième ordres des fonctions implicites définies par des équations d'état. On considérait également que les itérées correspondantes des gradients et Hessiens réduits convergent à la même vitesse que l'itération de point fixe d'origine.

Cette étude plus détaillée révèle néanmoins que ces dérivées convergent avec un certain retard par rapport aux valeurs de la fonction. En effet le rapport des erreurs correspondantes croit vers l'infini proportionnellement au compteur d'itérations ou à son carré. Mathématiquement, cet effet plutôt subtil est causé par l'apparition de blocs de Jordan correspondant à des valeurs propres dégénérées. Nous construisons un modèle théorique de cet effet et nous le validons par des expériences numériques.

1 Introduction and Assumption

The effect to be analyzed arises in the context of design optimization by what has been called piggy-back optimization [2]. Design optimization problems are distinguished from general nonlinear programming problems (NLP) by the fact that the vector of variables x is a priori partitioned into a state vector $y \in Y$ and a set of design variables $u \in U$. For application of this scenario in computational fluidynamics see for example [7], [5], [6], and [4]. Throughout we assume that the "user" has provided an iteration function

$$G : Y \times U \to Y$$

that is contractive with respect to an inner product norm on Y so that for all $u \in U$ and $y, \bar{y} \in Y$

$$\|G(y,u) - G(\bar{y},u)\| \leq \varrho \, \|y - \bar{y}\|.$$

*Supported by the DFG research center "Mathematics for the key technologies" (FZT 86) in Berlin

Here $\varrho < 1$ may vary continuously as a function of the design u and its exact size will usually not be available to a practical algorithm.

As an immediate consequence it follows by the Banach fixed point theorem that for fixed u and any initial $y_0 \in Y$ the sequence $\{y_k\}$ generated by

$$y_{k+1} = G(y_k, u)$$

must converge to the unique fixed point $y_* = y_*(u)$ with $y_* = G(y_*, u)$. In other words, the assumptions made so far ensure that one can obtain for any u a solution $y_*(u)$, a process which one may call "simulate" the underlying system. In a practical simulation the variables u and y will often be restricted to open subsets of the spaces U and Y, respectively.

In order to progress from simulation to design we require more smoothness of G, namely, that it is at least once continuously differentiable in the joint variable vector (y, u). The same assumption will be made for the objective function

$$f : Y \times U \to \mathbb{R},$$

which is meant to be minimized. Provided at least $f \in C^1(Y, U)$, one can obtain in a completely automated fashion the adjoint iteration function

$$\bar{G}(y, \bar{y}, u) \equiv \bar{y} G_y(y, u) + f_y(y, u). \tag{1}$$

Here subscripts denote partial differentiation and $\bar{y}$ like the gradient f_y is considered a row-vector belonging to the dual space of Y, which we identify with the Hilbert space Y itself. Then we have in the induced matrix and operator norm

$$\varrho(u) = \max_{y \in Y} \|G_y(y, u)\| \leq \varrho < 1$$

so that also in the dual norm

$$\|\bar{G}(y, \bar{y}, u) - \bar{G}(y, \hat{y}, y, u)\| \leq \varrho \|\bar{y} - \hat{y}\|$$

for any two row-vectors $\bar{y}, \hat{y} \in \bar{Y} \equiv Y$.

2 Piggy-Back Convergence of Adjoints

Throughout the remainder of this paper we consider u as constant and may therefore omit it occasionally as an argument in analyzing the simultaneous iteration

$$\begin{bmatrix} y_{k+1} \\ \bar{y}_{k+1} \end{bmatrix} = \begin{bmatrix} G(y_k, u) \\ \bar{G}(y_k, \bar{y}_k, u) \end{bmatrix} \tag{2}$$

Even if G is merely C^1 and thus $G_y(y) = G_y(y, u)$ continuous with respect to y it follows from $y_k \to y_*$ that $G_y(y_k, u) \to G_y(y_*, u)$ and hence the adjoint iterates $\bar{y}_k$ converge to $\bar{y}_*$ the unique solution of the adjoint equation

$$\bar{y}_* = \bar{y}_* G_y(y_*, u) + f_y(y_*, u) \tag{3}$$

The vector $\bar{y}_*$ can be used to compute the so called reduced gradient

$$\bar{u}_* = \bar{y}_* G_u(y_*, u) + f_u(y_*, u) \tag{4}$$

This row vector represents the total derivatives of f with respect to u, after the elimination of the state vector y using the implicit function theorem. In order to be more specific about the rate of convergence we assume that G_y and f_y are Lipschitz continuous with respect to y so that for some $\nu > 0$

$$\|G_y(\tilde{y}, u) - G_y(y, u)\| \leq \nu \|\tilde{y} - y\| \geq \|f_y(\tilde{y}, u) - f_y(y, u)\|.$$

Then we find immediately for the discrepancies $\Delta y_k \equiv y_k - y_*$ and $\Delta \tilde{y}_k \equiv \tilde{y}_k - \tilde{y}_*$ that

$$\begin{aligned} \|\Delta \tilde{y}_{k+1}\| &\leq \|\tilde{y}_k G_y(y_k, u) - \tilde{y}_* G_y(y_*, u)\| + \|f_y(y_k, u) - f_y(y_*, u)\| \\ &= \|\Delta \tilde{y}_k G_y(y_k, u) + \tilde{y}_*(G_y(y_k, u) - G_y(y_*, u))\| + \nu \|y_k - y_0\| \\ &\leq \varrho \|\Delta \tilde{y}_k\| + \|\Delta y_k\|(\|\tilde{y}_*\| + 1)\nu \end{aligned}$$

Consequently we have for any weighted error combination

$$\varepsilon_k \equiv \|\Delta y_k\| + \omega \|\Delta \tilde{y}_k\|$$

the recurrence

$$\begin{aligned} \varepsilon_{k+1} &\leq \varrho \|\Delta y_k\| + \omega(\varrho \|\Delta \tilde{y}_k\| + \nu(\|\tilde{y}_*\| + 1)\|\Delta y_k\|) \\ &= (\varrho + \omega\nu(\|\tilde{y}_*\| + 1))\|\Delta y_k\| + \omega\varrho \|\Delta \tilde{y}_k\| \\ &\leq (\varrho + \omega\nu(\|\tilde{y}_*\| + 1))\varepsilon_k \end{aligned}$$

This implies for any $\omega < (1 - \varrho)/(\nu(\|\tilde{y}_*\| + 1))$ the Q-linear convergence result

$$\limsup_{k \to \infty} \varepsilon_{k+1}/\varepsilon_k \leq \varrho + \omega\nu(\|\tilde{y}_*\| + 1) < 1$$

By standard arguments one derives the R-linear convergence results

$$\limsup_{k \to \infty} \sqrt[k]{\|\Delta \tilde{y}_k\|} \leq \lim_{k \to \infty} \sqrt[k]{\varepsilon_k/\omega} \leq \varrho + \omega\nu(\|\tilde{y}_*\| + 1) < 1$$

Taking the infimum over all $\omega > 0$ one finally obtains as in [1]

$$\limsup_{k \to \infty} \sqrt[k]{\|\Delta \tilde{y}_k\|} \leq \varrho \geq \limsup_{k \to \infty} \sqrt[k]{\|\Delta y_k\|}.$$

Here the inequality on the right was just added for comparison. Since these convergence speed cannot be improved under our assumptions (namely G_y has maximal norm ϱ and is Lipschitz continuous with respect to y) one may arrive at the conclusion that the sequences $\{y_k\}$ and $\{\tilde{y}_k\}$ converge essentially at the same speed. In fact this claim has been made repeatedly in the literature and the present author has suffered from the same impression for a long time. On the other hand there has been the persistent notion that the convergence of derivatives is lagging behind those of the underlying fixed point iterates.

3 Relative Convergence Speed of First Adjoints

In the remainder of this paper we require that $Y = \mathbb{R}^n$ and $U = \mathbb{R}^m$ are finite dimensional Euclidean spaces so that all linear operators can be identified with their matrix presentation. Assuming furthermore, that G and f are twice Lipschitz-continuously differentiable, we may rewrite the recurrence (2) as

$$\begin{bmatrix} y_{k+1} \\ \tilde{y}_{k+1} \end{bmatrix} = \begin{bmatrix} G(y_k, u) \\ N_y(y_k, \tilde{y}_k, u) \end{bmatrix} \tag{5}$$

Here we have expressed the G from (3) as the gradient of the function

$$N(y, \bar{y}, u) \equiv \bar{y}\, G(y, u) + f(y, u)$$

with respect to y. Notice that this function N differs from the familiar Lagrange function L of the optimization problem $\mathrm{Min} f(y, u)$ s.t. $G(y, u) - y = 0$ by the shift

$$\bar{y}\, y = N(y, \bar{y}, u) - L(y, \bar{y}, u)$$

Consequently, we have

$$N_{\bar{y}} = L_{\bar{y}} + \bar{y} \quad \text{and} \quad N_{\bar{y}} = L_{\bar{y}} + y \quad \text{but} \quad N_u = L_u$$

and, for the subsequent analysis more importantly, all second derivatives are identical:

$$N_{\bar{y}\bar{y}} = L_{\bar{y}\bar{y}}, N_{\bar{y}u} = L_{\bar{y}u}, N_{uu} = L_{uu}$$

Differentiating (5) we obtain the block-triangular Jacobian

$$J_k \equiv \frac{\partial(y_{k+1}, \bar{y}_{k+1})}{\partial(y_k, \bar{y}_k)} = \begin{bmatrix} G_y(y_k, u) & 0 \\ N_{\bar{y}y}(y_k, \bar{y}_k, u) & G_y^T(y_k, u) \end{bmatrix}$$

Obviously we have the characteristic polynomial

$$\det(J_k - \lambda I) = \det^2(G_y(y_k, u) - \lambda I)$$

which means that J_k has the same eigenvalues as $G_y(y_k, u)$ but all of them with the multiplicity 2. Moreover, at least one of these eigenvalues will be defective and thus generate a Jordan block of dimension grater than 2, unless the quadratic form $v^T N_{\bar{y}y} v$ vanishes for all eigenvectors v of $G_y(y_*, u)$. As a consequence one can deduce a linear-geometric decline in the adjoint error as follows.

Linearizing about the fixed point $(y_*, \bar{y}_*)$ we obtain the Taylor expansion

$$\begin{bmatrix} \Delta y_{k+1} \\ \Delta \bar{y}_{k+1} \end{bmatrix} = \begin{bmatrix} A & 0 \\ B & A^T \end{bmatrix} \begin{bmatrix} \Delta y_k \\ \Delta \bar{y}_k \end{bmatrix} + O(\|\Delta y_k\|^2 + \|\Delta \bar{y}_k\|^2)$$

where $A \equiv G_y(y_*, u)$ and $B \equiv N_{\bar{y}y}(\bar{y}_*, y_*, u)$. From this it follows by induction using the R-linear convergence of $\|\Delta y_k\| + \|\Delta \bar{y}_k\|$ that for any k and $j > 0$

$$\begin{bmatrix} \Delta y_{k+j} \\ \Delta \bar{y}_{k+j} \end{bmatrix} = \begin{bmatrix} A & 0 \\ B & A^T \end{bmatrix}^j \begin{bmatrix} \Delta y_k \\ \Delta \bar{y}_k \end{bmatrix} + O(\|\Delta y_k\|^2 + \|\Delta \bar{y}_k\|^2) \tag{6}$$

Similarly it can be easily verified by induction that

$$J_*^j \equiv \begin{bmatrix} A & 0 \\ B & A^T \end{bmatrix}^j = \begin{bmatrix} A^j & 0 \\ \sum_{i=1}^{j} (A^T)^{i-1} B A^{j-i} & (A^T)^j \end{bmatrix} \tag{7}$$

To simplify the matrix on the bottom left we assume that $A = G_y$ is real diagonalizable so that

$$A = T \Gamma T^{-1} \quad \text{with} \quad \Gamma = \mathrm{diag}\{\gamma_j\}_{j=1}^{n}$$

where

$$\varrho_* \equiv \max_{1 \le j \le n} |\gamma_j| \quad \le \varrho < 1$$

Then we can perform a two stage reduction to obtain the Jordan-like representation

$$\begin{aligned}\begin{bmatrix} A & 0 \\ B & A^T \end{bmatrix} &= \begin{bmatrix} T & 0 \\ 0 & T^{-T} \end{bmatrix} \begin{bmatrix} \Gamma & 0 \\ T^T B T & \Gamma \end{bmatrix} \begin{bmatrix} T^{-1} & 0 \\ 0 & T^T \end{bmatrix} \\ &= \begin{bmatrix} T & 0 \\ 0 & T^{-T} \end{bmatrix} \begin{bmatrix} J & 0 \\ C^T & I \end{bmatrix} \begin{bmatrix} \Gamma & 0 \\ D & \Gamma \end{bmatrix} \begin{bmatrix} I & 0 \\ C & I \end{bmatrix} \begin{bmatrix} T^{-1} & 0 \\ 0 & T^T \end{bmatrix}\end{aligned} \tag{8}$$

Here D is the (real) diagonal of $T^T B T$ and $C = -C^T$ is the antisymmetric solution of the Liapunov equation

$$\Gamma C - C\Gamma = T^T B T - D.$$

It is well known that the linear mapping from C to $\Gamma C - C\Gamma$ has the n^2 eigenvalues $\gamma_i - \gamma_j$ and the eigenvectors $e_i e_j^T$ for $1 \le i, j \le n$, so that the Liapunov equation must be solvable if all eigenvalues of A are distinct, as we will assume for simplicity. Then it follows immediately that the j-th power of J_* is given by

$$\begin{bmatrix} A & 0 \\ B & A^T \end{bmatrix}^j = \begin{bmatrix} T & 0 \\ (CT^{-1})T & T^{-T} \end{bmatrix} \begin{bmatrix} \Gamma^j & 0 \\ j\, D\Gamma^{j-1} & \Gamma^j \end{bmatrix} \begin{bmatrix} T^{-1} & 0 \\ CT^{-1} & T^T \end{bmatrix} \tag{9}$$

Thus we see that unless the diagonal D of $T^T B T$ happens to vanish there might be a pretty strong growth in the adjoint error component $\Delta\bar{y}_k$. To compare it to the original error Δy_k itself we firstly have to analyze its recurrence a bit more carefully. Using the Lipschitz constant ν one finds by standard estimates

$$\|\Delta y_{k+1} - A\Delta y_k\| \le \nu \|\Delta y_k\|^2.$$

Abbreviating

$$\Delta\tilde{y}_{k+1} = T^{-1}\Delta y_{k+1} \quad \text{and} \quad \kappa = \|\mathrm{T}\|\|\mathrm{T}^{-1}\|$$

one has consequently

$$\|\Delta\tilde{y}_{k+1} - \Gamma\Delta\tilde{y}_k\| \le \nu\kappa\|\Delta\tilde{y}_k\|^2.$$

Using this estimate and the assumption that there is only one eigenvalue, say γ_1 with the maximal modulus $\varrho_* = |\gamma_1|$ one can show that the angle between $\tilde{y}_k$ and the first Cartesian basis vector e_1 satisfies a recurrence that has exactly one stable fixed point namely 0. Thus we we have generically

$$\lim_{k\to\infty} \frac{|e_1^T \Delta\tilde{y}_k|}{\|\Delta\tilde{y}_k\|} = 1 \quad \text{and} \quad \lim_{k\to\infty} \frac{\|\Delta\tilde{y}_k\|}{\varrho_*^k} = c_* \in (0, \infty).$$

From (6) and (9) it then follows that

$$T^T \Delta\tilde{y}_{k+j} = [C^T\Gamma^j + jD\Gamma^{j-1} + \Gamma^j C]\Delta\tilde{y}_k + \Gamma^j T^T \Delta\tilde{y}_k + O(\|\Delta y_k\|^2 + \|\Delta\tilde{y}_k\|^2)$$

Assuming finally that the first diagonal element d_1 of D does not vanish we find that the second term in the above expansion dominates as j grows and we get the limit exactly

$$\lim_{k\to\infty} \frac{1}{k} \frac{\|T^T \Delta\tilde{y}_k\|}{\|T^{-1}\Delta y_k\|} = \frac{d_1}{\varrho_*} \tag{10}$$

so that obviously

$$\lim_{k\to\infty} \frac{1}{k} \frac{\|\Delta\tilde{y}_k\|}{\|\Delta y_k\|} \le \kappa \frac{d_1}{\varrho_*}$$

and approximately

$$\frac{\|\Delta\bar{y}_k\|}{\|\Delta\bar{y}_k\|} \quad \sim \quad k$$

Hence we see that the convergence of the adjoint vectors $\bar{y}_k$ really lags behind that of the underlying iterates y_k even though both sequences have the same R-factor ϱ_*.

4 Convergence of Second Order Adjoints

The above analysis can be extended to second derivatives representing products of the projected Hessian with certain direction vectors. More specifically, after picking a direction $\dot{u} \in U$ we may append (2) by the iterations

$$\dot{y}_{k+1} \equiv \dot{G}(y_k, \dot{y}_k, u, \dot{u}) \equiv G_y(y_k, u)\dot{y}_k + G_u(y_k, u)\dot{u} \tag{11}$$

and

$$\begin{aligned} \dot{\bar{y}}_{k+1} \equiv \dot{\bar{G}}(y_k, \bar{y}_k, \dot{y}_k, \dot{\bar{y}}_k, u, \dot{u}) \;&\equiv\; \dot{\bar{y}}_k G_y + \bar{y}_k G_{yy}\dot{y}_k + f_{yy}\dot{y}_k + \bar{y}_k G_{yu}\dot{u} + f_{yu}\dot{u} \\ &=\; \dot{\bar{y}}_k G_y(y_k, u) + N_{yy}(y_k, \bar{y}_k, u)\dot{y}_k + N_{yu}(y_k, \bar{y}_k, u)\dot{u} \end{aligned} \tag{12}$$

where all derivatives of G and f are evaluated at the current argument (y_k, u). Then an analysis along the lines of Section 3 shows that the $\dot{y}_k$ and $\dot{\bar{y}}_k$ also converge R-linearly to respective fixed points $\dot{y}_*$ and $\dot{\bar{y}}_*$ solving

$$\dot{y}_* = \dot{G}(y_*, \dot{y}_*, u, \dot{u}) \quad \text{and} \quad \dot{\bar{y}}_* = \dot{\bar{G}}(y_*, \bar{y}_*, \dot{y}_*, \dot{\bar{y}}_*, u, \dot{u}).$$

The vector $\dot{y}_*$ represents the feasible direction in state space associated with the variation $\dot{u}$ in the design space. The vector $\dot{\bar{y}}_*$ can be used to compute

$$\dot{\bar{v}}_* \;\equiv\; \dot{\bar{y}}_* G_u(y_*, u) + N_{uy}(y_*, \bar{y}_*, u)\dot{y}_* + N_{uu}(y_*, \bar{y}_*, u)\dot{u} \tag{13}$$

which represents the product of the reduced Hessian with the direction $\dot{u}$. To analyze the speed of convergence more carefully let us consider the extended Jacobian

$$\frac{\partial(y_{k+1}, \bar{y}_{k+1}, \dot{y}_{k+1}, \dot{\bar{y}}_{k+1})}{\partial(y_k, \bar{y}_k, \dot{y}_k, \dot{\bar{y}}_k)} = \begin{bmatrix} G_y(y_k, u) & 0 & 0 & 0 \\ N_{yy}(y_k, \bar{y}_k, u) & G_y^T(y_k, u) & 0 & 0 \\ P(y_k, \dot{y}_k, u, \dot{u}) & 0 & G_y(y_k, u) & 0 \\ H(y_k, \bar{y}_k, \dot{y}_k, \dot{\bar{y}}_k, u, \dot{u}) & P(y_k, \dot{y}_k, u, \dot{u})^T & N_{yy}^T(y_k, \bar{y}_k, u) & G_y^T(y_k, u) \end{bmatrix}$$

where

$$\begin{aligned} P(y, \dot{y}, u, \dot{u}) \;&=\; G_{yy}(y, u)\dot{y} + G_{yu}(y, u)\dot{u} \\ H(y, \bar{y}, \dot{y}, \dot{\bar{y}}, u, \dot{u}) \;&=\; \dot{\bar{y}} G_{yy}(y, u) + N_{yyy}(y, \bar{y}, u)\dot{y} + N_{yyu}(y, \bar{y}, u)\dot{u}. \end{aligned}$$

We notice that the matrix H is symmetric, while P is general and the values of these two square matrices at the fixed point $(y_*, \bar{y}_*, \dot{y}_*, \dot{\bar{y}}_*)$ are independent of each other as well as A and B.

We are looking now for estimates of the corresponding discrepancies $\Delta\dot{y}_k = \dot{y}_k - \dot{y}_*$ and $\Delta\dot{\bar{y}}_k \equiv \dot{\bar{y}}_k - \dot{\bar{y}}_*$ in addition to the Δy_k and $\Delta\bar{y}_k$ considered before. Similarly to (6) we obtain the linearization

$$\begin{bmatrix} \Delta y_{k+1} \\ \Delta\bar{y}_{k+1} \\ \Delta\dot{y}_{k+1} \\ \Delta\dot{\bar{y}}_{k+1} \end{bmatrix} = \begin{bmatrix} A & 0 & 0 & 0 \\ B & A^T & 0 & 0 \\ P & 0 & A & 0 \\ H & P^T & B & A^T \end{bmatrix}^j \begin{bmatrix} \Delta y_k \\ \Delta\bar{y}_k \\ \Delta\dot{y}_k \\ \Delta\dot{\bar{y}}_k \end{bmatrix} + O(\|\Delta y_k\|^2 + \|\Delta\bar{y}_k\|^2 + \|\Delta\dot{y}_k\|^2 + \|\Delta\dot{\bar{y}}_k\|^2)$$

Using the same transformation as in (8) the j-th power can be rewritten as following

$$\begin{bmatrix} A & 0 & 0 & 0 \\ B & A^T & 0 & 0 \\ P & 0 & A & 0 \\ H & P^T & B & A^T \end{bmatrix}^j =$$

$$\begin{bmatrix} T & 0 & 0 & 0 \\ T^{-T}C^T & T^{-T} & 0 & 0 \\ 0 & 0 & T & 0 \\ 0 & 0 & T^{-T}C^T & T^{-T} \end{bmatrix} \begin{bmatrix} \Gamma^j & 0 & 0 & 0 \\ jD\Gamma^{j-1} & \Gamma^j & 0 & 0 \\ \hat{P}_j & 0 & \Gamma^j & 0 \\ \hat{H}_j & \hat{P}_j^T & jD\Gamma^{j-1} & \Gamma^j \end{bmatrix} \begin{bmatrix} T^{-1} & 0 & 0 & 0 \\ CT^{-1} & T^T & 0 & 0 \\ 0 & 0 & T^{-1} & 0 \\ 0 & 0 & CT^{-1} & T^T \end{bmatrix}$$

where with $\hat{P} \equiv T^{-1}PT$ and $\hat{H} \equiv T^T HT$

$$\begin{bmatrix} \hat{P}_j & 0 \\ \hat{H}_j & \hat{P}_j^T \end{bmatrix} = \sum_{i=1}^{j} \begin{bmatrix} \Gamma^{i-1} & 0 \\ (i-1)\Gamma^{i-2}D & \Gamma^{i-1} \end{bmatrix} \begin{bmatrix} \hat{P} & 0 \\ \hat{H} & \hat{P}^T \end{bmatrix} \begin{bmatrix} \Gamma^{j-i} & 0 \\ (j-i)\Gamma^{j-i-1}D & \Gamma^{j-i} \end{bmatrix}$$

Here we used the relation (7) once again. Hence we have the expressions

$$\hat{P}_j = \sum_{i=1}^{j} \Gamma^{i-1}\hat{P}\Gamma^{j-i}$$

$$\hat{H}_j = \textstyle\sum_{i=1}^{j}(i-1)\Gamma^{i-2}D\hat{P}\Gamma^{j-i} + \Gamma^{i-1}\hat{H}\Gamma^{j-i} + (j-i)\Gamma^{i-1}P\Gamma^{j-i-1}D$$

Taking norms we obtain for constants c_1 and c_2

$$\|\hat{P}_j\| \le c_1\, j\, \rho_*^{j-1} \quad \text{and} \quad \|\hat{H}_j\| \le c_2\, j^2\, \rho_*^{j-2}$$

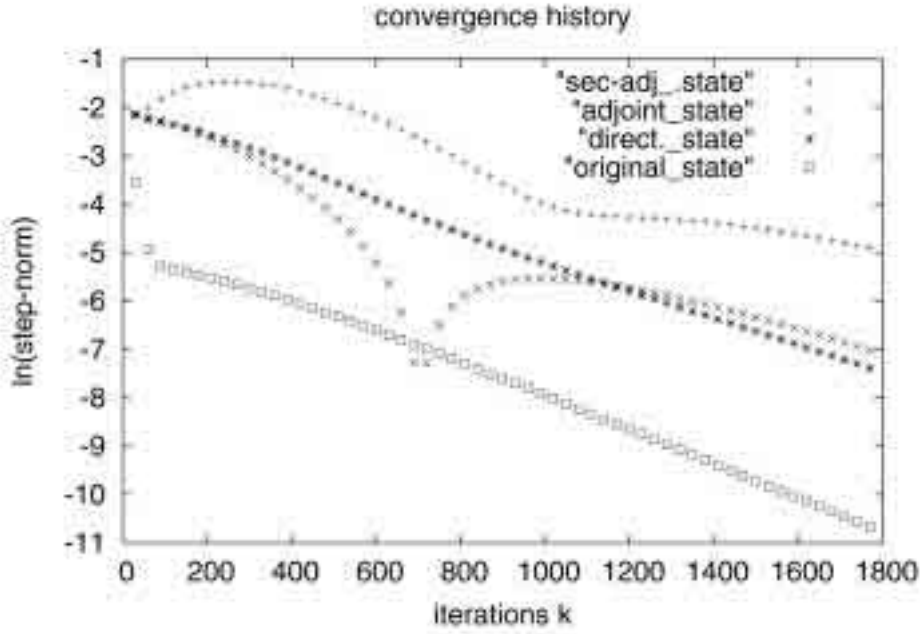
The later inequality is true because Γ has like A the spectral radius $\rho_* < 1$. Thus we can estimate all four error components as follows.

$$\begin{aligned}
\|\Delta y_{k+j}\| &\le \rho_*^j c_{11} \|\Delta y_k\| + O(\|\Delta y_k\|^2) \\
\|\Delta \tilde{y}_{k+j}\| &\le \rho_*^j \left[c_{22}\|\Delta \tilde{y}_k\| + c_{21}\, j\, \|\Delta y_k\|\right] + O(\|\Delta y_k\|^2 + \|\Delta \tilde{y}_k\|^2) \\
\|\Delta \hat{y}_{k+j}\| &\le \rho_*^j \left[c_{33}\|\Delta \hat{y}_k\| + c_{31}\, j\, \|\Delta y_k\|\right] + O(\|\Delta y_k\|^2 + \|\Delta \hat{y}_k\|^2) \\
\|\Delta \hat{\tilde{y}}_{k+j}\| &\le \rho_*^j \left[c_{44}\|\Delta \hat{\tilde{y}}_k\| + c_{41} j^2 \|\Delta y_k\| + c_{42} j (\|\Delta \tilde{y}_k\| + \|\Delta \hat{y}_k\|\right] \\
&+ O(\|\Delta y_k\|^2 + \|\Delta \tilde{y}_k\|^2 + \|\Delta \hat{y}_k\|^2 + \|\Delta \hat{\tilde{y}}_k\|^2)
\end{aligned}$$

Using the assumption (10) one can actually obtain the proportionality relations

$$\|\Delta \hat{y}_k\| \sim k\, \|\Delta y_k\| \sim k\, \rho_*^k \quad \text{and} \quad \|\Delta \hat{\tilde{y}}_k\| \sim k^2 \|\Delta y_k\| \sim k^2\, \rho_*^k$$

This means in particular that the second derivatives lag behind the first derivatives by a factor of order k and thus behind the original iteration by a factor of order k^2. A closer analysis of the propagation constants in the above system of bounds might allow an optimized decision of when to start propagating first and second derivatives. If this is done too early the error $\|\Delta y_j\|$ might still be so large that no reduction in the derivative errors occurs at all.



5 Numerical Results

The following results were obtained on the boundary control problem

$$\Delta_x y(x) + e^{y(x)} = 0 \quad \text{for} \quad x = (x_1, x_2) \in [0,1]^2$$

with the periodic and Dirichlet boundary conditions

$$y(0,\zeta) = y(1,\zeta), \quad y(\zeta,0) = \sin(2\pi\zeta), \quad y(\zeta,1) = u(\zeta) \quad \text{for} \quad \zeta \in [0,1]$$

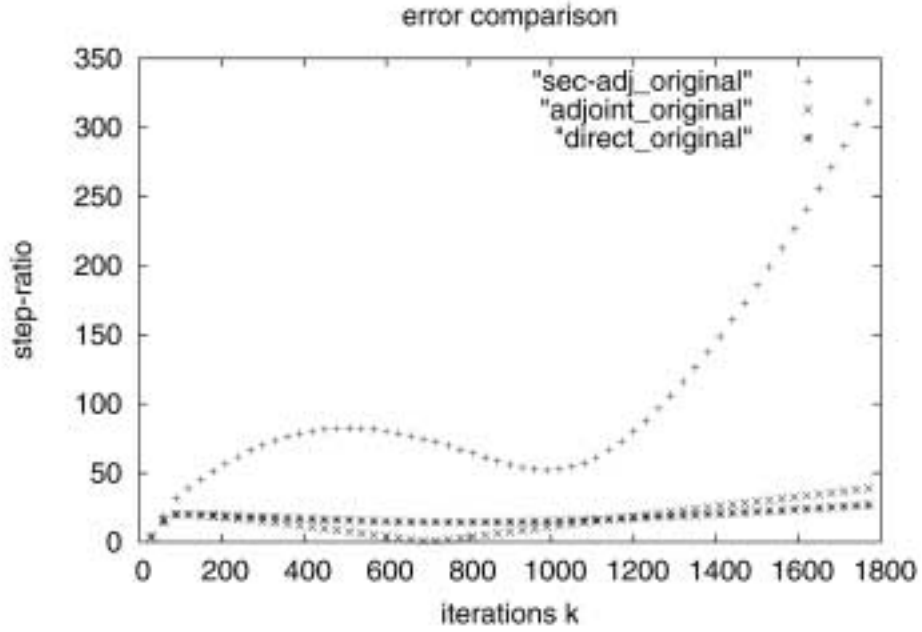
The function u is viewed as a boundary control that can be varied to minimize the objective function

$$f(y,u) = \int_0^1 \left[\left. \frac{\partial y(\eta,\zeta)}{\partial \eta} \right|_{\eta=0} - 4 - \cos(2\pi\zeta) \right]^2 d\zeta + \sigma \int_0^1 \left[u(\zeta)^2 + u'(\zeta)^2 \right] d\zeta$$

In the following calculations we used $\sigma = 0.001$ and set constantly $u(\zeta) = 2.2$. This value is not all that far from the fold point where solutions cease to exist.

We use a central difference discretization with the mesh-width 1/12.0 so that the resulting algebraic system involves 144 equations in as many variables. Since the nonlinearities occur only on the diagonal one can easily implement Jacobi's method to obtain the basic function $G(y,u)$. For this simple example we also coded by hand the corresponding derived functions $\dot{G}, \bar{G}$ and even $\hat{G}$ as defined in (1, 11) and (12), respectively. The results were later confirmed using the automatic differentiation tool ADOL-C [3].

As can be seen in Fig.1 the convergence of the Jacobi method is rather slow with the common R-factor being about $(1 - 1/300)$. The lowest curve represents the natural logarithms of the Euclidean norm ratios $\|y_{k+1} - y_k\| / \|y_1 - y_0\|$, which provide some indication of the norm ratios



$\|\Delta y_k\|/\|\Delta y_0\|$. In view of the very slow convergence this relation need certainly not be very close. Nevertheless the theory is basically confirmed with the first direct and adjoint derivatives $\|\dot{y}_{k+1} - \dot{y}_k\|/\|\dot{y}_1 - \dot{y}_0\|$ and $\|\bar{y}_{k+1} - \bar{y}_k\|/\|\bar{y}_1 - \bar{y}_0\|$ lagging somewhat behind and the second derivatives $\|\ddot{y}_{k+1} - \ddot{y}_k\|/\|\ddot{y}_1 - \ddot{y}_0\|$ coming in last. The ratio between these derivative quantities and the original iterates themselves is plotted in Fig. 2. After an initial transition phase one sees quite clearly a growth proportional to k and k^2 for the first and second derivatives, respectively. While the adjoints were defined as in (3) by the gradient of f, the direct differentiation was performed simultaneously with respect to all components of the discretized u so that the quantity $\dot{u}$ occurring in (11) and (12) was in fact the identity matrix of order 12. Consequently, $\dot{y}_k$ and $\ddot{y}_k$ had also 12 times as many components as the underlying y_k and $\bar{y}_k$, which are of the same size.

6 Summary, Conclusion and Outlook

We studied the convergence behavior of fixed point iterations for derivatives of implicit functions. These recurrences are generated in a completely mechanical fashion from a user supplied contractive fixed point solver for evaluating the implicit function. While the contractivity and thus the asymptotic convergence rate is inherited by the derived solvers there is a certain time lag. This is not really surprising since the equations for the adjoints $\bar{y}$ and those for the feasible directions $\dot{y}$ are dependent on y and both in turn impact the second order adjoint equation for $\ddot{y}$. Mathematically we obtain Jordan blocks of size 2 for the double eigenvalues of the first derivative systems and of size 3 for the quadruple eigenvalues of the second order adjoint system. One does not obtain blocks of size 4 since the $(3, 2)$ sub-block in the big Jacobian system vanishes identically. Otherwise it would connect the two first derivative systems.

Generally if one were to iteratively evaluate derivatives of order d one can expect that the

relative errors compared to those of the underlying function iteration grows like k^q, where k is the iteration counter. In the context of constrained optimization one can expect that the correct values of reduced gradients (4) and Hessians (13) are obtained slower than feasibility so that optimality will be arrived at in the tangential fashion that is familiar from SQP calculations [8, 9, 10]. In fact when the state equation only be solved by a slowly convergent fixed point solver as we have assumed throughout it makes little sense to apply an SQP type algorithms. Instead one will prefer a so-called one-shot optimization strategy [11], where feasibility and optimality is achieved at the same time. We are currently investigating a piggy-back optimization scheme, where a third iteration updating the design variables u on the basis of approximate reduced gradient information is appended to (3).

References

[1] A. Griewank, C.H. Bischof, G.F. Corliss, A. Carle, and K. Williamson. Derivative Convergence of Iterative Equation Solvers. *Optimization Methods and Software*, 2:321–355, 1993.

[2] A. Griewank and C. Faure. Reduced functions, gradients and hessians from fixed point iteration for state equations. *Numerical Algorithms*, 30(2):113–139, 2002.

[3] Andreas Griewank, David Juedes, and Jean Utke. ADOL-C, A Package for the Automatic Differentiation of Algorithms Written in C/C++. *TOMS*, 22(2):131–167, 1996.

[4] M. Hinze and T. Slawig. Adjoint gradients compared to gradients from algorithmic differentiation in instantaneous control of the navier-stokes equations. *Optimization Methods and Software*, 18(3):299–315, 2003.

[5] A. Jameson. Optimum aerodynamic design using cfd and control theory. In *12th AIAA Computational Fluid Dynamics Conference, AIAA Paper 95-1729*, San Diego, CA, 1995. American Institute of Aeronautics and Astronautics.

[6] B. Mohammadi and O. Pironneau. *Applied Shape Optimization for Fluids*. Numerical Mathematics and Scientific Computation. Clarendon Press, Oxford, 2001.

[7] P.A. Newman, G.J.-W. Hou, H.E. Jones, A.C. Taylor, and V.M. Korivi. Observations on computational methodologies for use in large-scale, gradient-based, multidisciplinary design incorporating advanced CFD codes. Technical Memorandum 104206, NASA Langley Research Center, February 1992. AVSCOM Technical Report 92-B-007.

[8] J. Nocedal and S.J. Wright. *Numerical Optimization*, Springer Series in Operation Research. Springer Verlag, New York,...Tokyo, 1999.

[9] E. W. Sachs. Control applications of reduced SQP methods. In R. Bulirsch and D. Kraft, editors, *Computational Optimal Control*, volume 115 of *Int. Series Num. Math.*, pages 89–104. Birkhäuser, 1994.

[10] V. H. Schulz. Solving discretized optimization problems by partially reduced SQP methods. *Computing and Visualization in Science*, 1:83–96, 1998.

[11] S. Ta'asan, G. Kuruvila, and M.D. Salas. Aerodynamic design and optimization in one shot. In *30th AIAA Aerospace Sciences Meeting and Exhibit, AIAA Paper 91-0025*, Reno, Nevada, 1992. American Institute of Aeronautics and Astronautics.