Managing the drift-off in numerical index-2 differential algebraic equations by projected defect corrections

Roswitha März, Berlin

Abstract

When integrating index-2 differential-algebraic equations, the given constraint may be failed to be met due to the integration method itself and also due to numerical defects in the realization. This so-called drift-off gives rise to bad instabilities. In 1991 Ascher and Petzold proposed to manage the drift-off caused by symmetric implicit Runge-Kutta methods in Hessenberg systems by means of backprojections onto the constraint. In the present paper, this nice idea is generalized and analyzed in some detail for general index-2 differential-algebraic equations and, in particular, for quasilinear equations $a(x,t)x' + g(x,t) = 0$, as they arise in applications. Now the constraint under consideration is only implicitly given and the backprojection turns out to be rather a projected defect correction.

1 Introduction

In order to apply symmetric discretization problems reasonably also in case of index-2 differential-algebraic equations (DAEs)

$$x_1'(t) + b_1(x_1(t), x_2(t), t) \;=\; 0, \tag{1.1}$$

$$b_2(x_1(t), t) \;=\; 0 \tag{1.2}$$

Ascher and Petzold (1991) proposed to complete standard methods by an additional backward projection onto the constraint given by equation (1.2). An already known approximation $x_{l,1}, x_{l,2}$ of the true solution value $x_1(t_l), x_2(t_l)$ is replaced by a new approximation $x_{l,1}^{\mathrm{new}}, x_{l,2}^{\mathrm{new}}$ such that the condition

$$b_2(x_{l,1}^{\mathrm{new}}, t_l) = 0 \tag{1.3}$$

is satisfied. This is accomplished by means of the ansatz

$$x_{l,1}^{\mathrm{new}} \;=\; x_{l,1} + \frac{\partial}{\partial x_2} b_1\left(x_{l,1}^{\mathrm{new}}, x_{l,2}^{\mathrm{new}}, t_l\right)\lambda, \tag{1.4}$$

$$x_{l,2}^{\mathrm{new}} \;=\; x_{l,2}. \tag{1.5}$$

The defect $b_2(x_{l,1}, t_l)$ represents the drift off from the constraint manifold given by equation (1.2). Clearly, with the backprojection one aims at reducing this drift off.
The index-2 property of (1.1),(1.2) guarantees that λ and x_l^{new} are locally uniquely determined by (1.3)-(1.5) (cf. Hairer and Wanner (1991), p. 513) provided that the initial

approximations are sufficiently accurate.
One step of the Newton method applied to the nonlinear system (1.3)-(1.5) for $x_l^{\text{new}}, \lambda$ starting with the initial guess $x_l^{\text{new}(0)} := x_l, \lambda^{(0)} := 0$ gives

$$x_{l,1}^{\text{new}(1)} = x_{l,1} - F_l \, b_2(x_{l,1}, t_l), \tag{1.6}$$

where we denote $\frac{\partial b_1}{\partial x_2} \left(\frac{\partial b_2}{\partial x_1} \frac{\partial b_1}{\partial x_2} \right)^{-1} (x_{l,1}, x_{l,2}, t_l) =: F_l$. Furthermore, if we take into account that $F_l \frac{\partial b_2}{\partial x_1}(x_{l,1}, t_l) =: H_l$ is a projector, then (1.6) turns out to be a kind of projected defect correction. Namely, since $H_l F_l = F_l$ holds true, formula (1.6) means in more detail

$$\begin{aligned} H_l x_{l,1}^{\text{new}(1)} &= H_l x_{l,1} - F_l \, b_2(x_{l,1}, t_l), \\ (I - H_l) x_{l,1}^{\text{new}(1)} &= (I - H_l) x_{l,1}. \end{aligned}$$

Thus, only the particular component related to the projector H_l is affected.
The resulting projected Runge-Kutta methods have proved their value in lots of cases. The idea of a backward projection has also been used for the numerical integration of regular ordinary differential equations (ODEs) and index-1 DAEs with great success for maintaining given invariants numerically (cf. Shampine (1986), Schulz, Bock and Steinbach (1996), Eich (1992)).
In Gear (1986), it was shown that the overdetermined system of a regular ODE with an invariant

$$\begin{aligned} x_1'(t) + \Phi(x_1(t), t) &= 0, \\ \Psi(x_1, (t)) &= 0 \end{aligned}$$

is equivalent, in some sense, to the special index-2 DAE

$$\begin{aligned} x_1'(t) + \Phi(x_1(t), t) + \frac{\partial}{\partial x_1} \Psi(x_1(t))^T x_2(t) &= 0, \\ \Psi(x_1(t)) &= 0, \end{aligned}$$

which makes the close relationship between the projected integration method for index-2 DAEs and the backprojection onto invariants more transparent.

The present paper pursues two different objectives. On the one hand, the procedure (1.6) of the projected defect correction shall be applied to DAEs that are not of the special Hessenberg form (1.1), (1.2). For more general index-2 DAEs, like those occurring in the circuit simulation when using the charge oriented modified nodal analysis (e.g. März and Tischendorf (1996)), the structure is not given as explicitly as in case of Hessenberg systems. The component that has to be improved and which is an analogue to $H_l x_{l,1}$, as well as the critical part of the defect are implicitly contained in the system and have to be found out first.
For quasilinear index-2 DAEs

$$a(x(t), t) x'(t) + b(x(t), t) = 0, \tag{1.7}$$

for instance, we propose and investigate a generalization of the correction formulas (1.5), (1.6) that looks like

$$x_l^{\text{new}(1)} := x_l - P Q_{1,l} \;\; G_{2,l}^{-1} \, b(x_l, t_l) \tag{1.8}$$

with a scaling matrix $G_{2,l}$ and a special projector $PQ_{1,l}$.
The term $PQ_{1l}G_{2l}^{-1}\,b(x_l,t_l)$ turns out to represent a critical particular component of the defect caused by the given approximation x_l in the derivative free part of the DAE (1.7). In other words, it represents the drift off from a certain constraint implicitly contained in (1.7).
Recall once more that the remarkable thing about Hessenberg form DAEs (1.1), (1.2) is just the explicitly given relevant constraint (1.2).

There are well-known integration methods like the BDF that satisfy a priori the constraint (1.2). Thus, theoretically, applying e.g. the BDF to integrate (1.1), (1.2), there is no drift off caused by the integration method. However, in practical computations we are again confronted with nontrivial defects. Now they represent the residuals of the Newton iterations when solving the nonlinear systems arising per integration step.
The accuracy of numerical approximations generated e.g. by a BDF-code is known to be seriously influenced by the critical defect components amplified by h_l^{-1}, where h_l denotes the current stepsize. This is why the nonlinear equations arising per integration step have to be solved sufficiently accurately. Whether an integration code applied to an index-2 DAE performs reliably or not depends strongly on the tolerances for the Newton iterations. The code fails in both cases if tolerances are too precise or too coarse.
The second objective in the present paper is to treat the backprojection as a usual defect correction step within the more critical subsystem of the nonlinear equation system to be solved. As we know, this subsystem is given only implicitly, but can be figured out by projections. Solving this subsystem more precisely by means of defect corrections, we may expect the code to perform well also if the tolerances for the whole nonlinear system are weakened. In the consequence, the integration code gains more reliability and robustness.

The paper is organized as follows:
In Section 2 we develop a projected defect correction method for general implicit index-2 DAE's and show in what sense (1.3)-(1.5) and (1.6) are thus generalized. A detailed error analysis is presented in Section 3, where particularly simple relations result for quasilinear DAEs (1.7).
In Section 4 we describe a method for the numerical calculation of the special projector Q_{1l}.
Section 5 deals with relations between the projected defect correction and the selective projected methods proposed in Ascher and Spiteri (1994) for the case of semi-explicit index-2 systems.
An experiment in applying the projected defect correction for the improvement of the reliability of the BDF is reported in Section 6.
No daubt, the proposed projected defect corrections allow, as in the case of Hessenberg systems, the use of projected Runge-Kutta methods etc. without having to transform the DAE itself anytime. Here, the effect cannot but be the same as already described in Ascher and Petzold (1991).

2 A general projected defect correction formula

Consider the DAE

$$f(x'(t), x(t), t) = 0, \tag{2.1}$$

where $f : \mathcal{G}_f \times \mathcal{I}_f \subseteq I\!\!R^m \times I\!\!R^m \times I\!\!R \to I\!\!R^m$ is continuous together with its partial Jacobians $f'_{x'}, f'_x : \mathcal{G}_f \times \mathcal{I}_f \to L(I\!\!R^m)$.
The nullspace of the leading Jacobian $\ker f'_{x'}(x', x, t)$ is assumed to be constant. Denote

$$N := \ker f'_{x'}(x', x, t),$$

and introduce projector matrices $Q, P \in L(I\!\!R^m)$ such that $Q^2 = Q$, $imQ = N$, $P := I - Q$.

In case of the Hessenberg system (1.1), (1.2) the leading nullspace N is simply

$$N = \left\{ \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \in I\!\!R^m : z_1 = 0 \right\},$$

further we may choose $Q = diag\ (0, I)$.

Let $x_*(.) : \mathcal{I} \to I\!\!R^m$ denote the solution of (2.1) which is to be approximated. The DAE (2.1) is supposed to be index-2 tractable in a certain neighbourhood $\mathcal{G} \times \mathcal{I}$ of the graph

$$\Gamma := \{((Px_*)'(t), x_*(t), t) : t \in \mathcal{I}\}.$$

More precisely, this means by definition (e.g. März (1995)) that the intersection $N \cap S(x', x, t)$ has a constant but nonzero dimension on $\mathcal{G} \times \mathcal{I}$, and the subspaces $N_1(x', x, t)$, $S_1(x', x, t)$ intersect trivially on $\mathcal{G} \times \mathcal{I}$, i.e.,

$$N_1(x', x, t) \cap S_1(x', x, t) = \{0\}, \quad (x', x, t) \in \mathcal{G} \times \mathcal{I}. \tag{2.2}$$

There, the following denotations are used:

$$\begin{aligned}
S(x', x, t) &:= \{z \in I\!\!R^m : f'_x(x', x, t)z \in im\ f'_{x'}(x', x, t)\}, \\
S_1(x', x, t) &:= \{z \in I\!\!R^m : f'_x(x', x, t)Pz \in im\ G_1(x', x, t)\}, \\
G_1(x', x, t) &:= f'_{x'}(x', x, t) + f'_x(x', x, t)Q, \\
N_1(x', x, t) &:= \ker G_1(x', x, t).
\end{aligned}$$

Recall that the nullspace $N_1(x', x, t)$ has the same dimension as the intersection $N \cap S(x', x, t)$. Moreover, there is a uniquely determined projector matrix function $Q_1 : \mathcal{G} \times \mathcal{I} \to L(I\!\!R^m)$ such that $Q_1(x', x, t)$ projects onto $N_1(x', x, t)$ along $S_1(x', x, t)$.
The matrix

$$G_2(x', x, t) := G_1(x', x, t) + f'_x(x', x, t)PQ_1(x', x, t)$$

remains nonsingular on $\mathcal{G} \times \mathcal{I}$.
Clearly, all matrices given here depend continuously on their arguments.

Now let $t_l \in \mathcal{I}$ denote the current time and let x_l', x_l be the given approximations of $(Px_*)'(t_l)$, $x_*(t_l)$, which cause the defect

$$\delta_l := f(x_l', x_l, t_l) \neq 0. \tag{2.3}$$

It should be emphasized once more that it does not matter at all how x_l', x_l are computed. Usually, x_l' represents a finite difference, typically $P\frac{1}{h_l}(x_l - x_{l-1})$.
We aim at a new, improved approximation x_l^{new} such that the critical part of the defect δ_l vanishes. However, what is the critical part of δ_l and which component of the approximations correspond to it?
For shortness, let us write : $Q_{1l} := Q_1(x_l', x_l, t_l)$, $P_{1l} := I - Q_{1l}$, $G_{2l} := G_2(x_l', x_l, t_l)$. We refer to Section 4 below for the numerical computation of Q_{1l}.
By means of a careful error analysis of integration methods it is shown (e.g. Tischendorf (1996), März (1992)) that $PQ_{1l}G_{2l}^{-1}\delta_l$ represents the critical part of the defect. On the other hand, there is a close relationship between that part of the defect and the error of the PQ_{1l}-component of the solution.
Consider the ansatz

$$x_l^{\text{new}} = x_l + PQ_1 z, \tag{2.4}$$

$$PQ_{1l}G_{2l}^{-1} f(x_l', x_l^{\text{new}}, t_l) = 0, \tag{2.5}$$

$$z = PQ_{1,l} z. \tag{2.6}$$

In Section 3 it will be shown that (2.4) - (2.6) determine x_l^{new}, z locally uniquely provided that the approximations we start with are suffciently close. With the initial guess $x_l^{\text{new}(0)} = x_l, z^{(0)} = 0$ one Newton iteration for the system

$$x_l^{\text{new}} - x_l - PQ_{1l} z = 0$$

$$PQ_{1l}G_{2l}^{-1} f(x_l', x_l^{\text{new}}, t_l) + (I - PQ_{1l})z = 0$$

yields

$$x_l^N := x_l^{\text{new}(1)} = x_l - PQ_{1l}G_{2l}^{-1}\delta_l, \tag{2.7}$$

which is the projected defect correction formula we are looking for. In Section 4 it is described how to realize (2.7) numerically.

In the special case of linear DAEs

$$A(t)x'(t) + B(t)x(t) - q(t) = 0 \tag{2.8}$$

we have, on the one hand,

$$A(t_l)(Px_*)'(t_l) + B(t_l)x_*(t_l) - q(t_l) = 0,$$

but, on the other hand,

$$A(t_l)x_l' + B(t_l)x_l - q(t_l) = \delta_l.$$

Taking into account the given relations

$$A(t_l) = G_{1l}P = G_{2l}P_{1l}P,$$

$$PQ_{1l}G_{2l}^{-1}A(t_l) = PQ_{1l}G_{2l}^{-1}G_{2l}P_{1l}P = PQ_{1l}P_{1l}P = 0$$

$$Q_{1l} = Q_{1l}G_{2l}^{-1}B(t_l) = Q_{1l}G_{2l}^{-1}B(t_l)P$$

we find

$$PQ_{1l}x_l = PQ_{1l}G_{2l}^{-1}q(t_l) + PQ_{1l}G_{2l}^{-1}\delta_l$$

$$PQ_{1l}x_*(t_l) = PQ_{1l}G_{2l}^{-1}q(t_l)$$

thus $PQ_{1l}x_l - PQ_{1l}G_{2l}^{-1}\delta_l = PQ_{1l}x_*(t_l)$. Consequently, the defect correction formula (2.7) means for linear DAEs

$$\begin{aligned} x_l^N &= (I - PQ_{1l})x_l + PQ_{1l}x_l - PQ_{1l}G_{2l}^{-1}\delta_l \\ &= (I - PQ_{1l})x_l + PQ_{1l}x_*(t_l), \end{aligned}$$

i.e., the PQ_{1l}-component of the approximation is replaced by the corresponding component of the true solution.
Note that in the linear case it holds that

$$x_l^{\text{new}} = x_l^N.$$

Next we show the projected defect correction formula (2.7) to generalize (1.6). For that, we return to the Hessenberg DAE (1.1), (1.2) and form $Q = diag(0, I)$,

$$PQ_{1l} = \begin{pmatrix} H_l & 0 \\ 0 & 0 \end{pmatrix}, \quad PQ_{1l}G_{2l}^{-1} = \begin{pmatrix} 0 & F_l \\ 0 & 0 \end{pmatrix}.$$

Recall that H_l represents a projector. It projects onto $im\frac{\partial b_1}{\partial x_2}(x_{l,1}, x_{l,2}, t_l)$ along $\ker\frac{\partial b_2}{\partial x_1}(x_{l,1}, t_l)$.
In this case, formula (2.7) reads

$$x_{l,1}^N = x_{l,1} - F_l b_2(x_{l,1}, t_l), \tag{2.9}$$

$$x_{l,2}^N = x_{l,2}, \tag{2.10}$$

that is, again we have the projected defect correction formula (1.6) for Hessenberg form DAEs .

3 Error Analysis

Now we have to ask whether x_l^{new} is well-defined by the nonlinear equation system (2.4)-(2.6) and how it is related to the true solution value $x_*(t_l)$. At the same time we try to give an estimate for the first Newton-iteration x_l^N.

Let the DAE (1.2) be index-2 tractable in the neighbourhood $\mathcal{G} \times \mathcal{I}$ of the graph Γ of the true solution as supposed in Section 2.
As above, $t_l \in \mathcal{I}$ and $(x_l', x_l) \in \mathcal{G}$ denote the current time and the given approximation to be improved, respectively. Q_{1l}, Q_{2l} etc. are the matrices used in Section 2.
Next introduce the auxiliary function

$$\tilde{f}(x', x) := PQ_{1l}G_{2l}^{-1} f(x', x, t_l), \quad (x', x) \in \mathcal{G}, \tag{3.1}$$

that is continuously differentiable due to our general assumptions agreed upon above. Of course, $\tilde{f}$ depends on l, but we drop that index.

When investigating the solvability of DAEs with index 2, functions like $\tilde{f}$ are supposed to be at least in C^2 (e.g. Tischendorf (1996)). From this point of view, it seems to be very natural that we now assume the Lipschitz-conditions

$$|\tilde{f}_x'(x', x) - \tilde{f}_x'(\bar{x}', \bar{x})| \leq L_1|x - \bar{x}| + L_2|x' - \bar{x}'|, \tag{3.2}$$

$$|\tilde{f}_{x'}'(x', x) - \tilde{f}_{x'}'(\bar{x}', \bar{x})| \leq L_3|x - \bar{x}| + L_4|x' - \bar{x}'|, \quad (x', x), (\bar{x}', \bar{x}) \in \mathcal{G} \tag{3.3}$$

to be satisfied.

It should be noted that our construction leads to the relation

$$PQ_{1l} = PQ_{1l}G_{2l}^{-1} f_x'(x_l', x_l, t_l) = \tilde{f}_x'(x_l', x_l),$$

which will be used later.
Furthermore, if we denote the critical part of the defect shortly by $\tilde{\delta}_l$, we have

$$\tilde{\delta}_l := PQ_{1l}G_{2l}^{-1}\delta_l = PQ_{1l}G_{2l}^{-1} f(x_l', x_l, t_l) = \tilde{f}(x_l', x_l).$$

In applications, the main interest is directed to special quasi-linear DAEs

$$a(x, t)x'(t) + b(x(t), t) = 0. \tag{3.4}$$

For those equations the auxiliary function $\tilde{f}$ becomes invariant of its first argument if the image space of the leading coefficient matrix $a(x, t)$ does not rotate with changing x. This fact is figured out by the following lemma. Things become essentially simpler in that case.

Lemma 3.1 *Let $f(x', x, t) \equiv a(x, t)x' + b(x, t)$ and let im $a(x, t)$ be independent of x.*

(i) Then it holds that

$$\tilde{f}(x', x) \equiv PQ_{1l}G_{2l}^{-1}b(x, t_l).$$

(ii) The subspaces $S(x', x, t), N_1(x', x, t), S_1(x', x, t)$ as well as the matrix $G_1(x', x, t)$ do not depend on their first argument x' at all.

(iii) The projector $Q_1(x', x, t)$ onto $N_1(x', x, t)$ along $S_1(x', x, t)$ is invariant of x'.

(iv) The expression $Q_1G_2^{-1}$ does not vary with x', and $Q_1G_2^{-1} = Q_1(G_1 + b'_xPQ_1)^{-1}$ holds true.

Proof:
Since *im* $a(x, t)$ does not vary with x, it follows that

$$a'_x(x, t)w \in im\ a(x, t), \quad w \in I\!R^m,$$

holds true. Since Q is assumed to remain constant, the matrix function G_1 simplifies to $G_1 = a + b'_xQ$. Moreover, now we have

$$S_1(x', x, t) = \{z \in I\!R^m : b'_x(x, t)Pz \in im(a(x, t) + b'_x(x, t)Q)\}.$$

By this, assertion (ii) becomes straight forward. Further, (ii) implies (iii) due to the uniqueness of that projection. Further, the matrix function $\tilde{G}_2 := G_1 + b'_xPQ_1$ is nonsingular and assertion (iv) follows immediately. To show (i), consider the relation

$$im\ a(x_l, t_l) \in \ker PQ_{1l}G_{2l}^{-1}$$

given by our construction. Namely, it holds that $Q_{1l}G_{2l}^{-1}f'_{x'}(x'_l, x_l, t_l) = Q_{1l}G_{2l}^{-1}G_{2l}P_{1l}P = 0$, thus $Q_{1l}G_{2l}^{-1}a(x_l, t_l) = 0$.
Because of *im* $a(x, t_l) \equiv$ *im* $a(x_l, t_l)$ we finally obtain $\tilde{f}(x', x) \equiv PQ_{1l}G_{2l}^{-1}(a(x, t_l)x' + b(x, t_l)) \equiv PQ_{1l}G_{2l}^{-1}b(x, t_l)$. □

Corollary 3.2 *Lemma 3.1 implies $L_2 = L_3 = L_4 = 0$ in the conditions (3.1), (3.2).*

Now turn to the nonlinear system (2.4)-(2.6) that can equivalently be rewritten as

$$x_l^{\text{new}} = x_l + z \tag{3.5}$$

$$PQ_{1l}G_{2l}^{-1}f(x'_l, x_l + z, t_l) + (I - PQ_{1l})z = 0. \tag{3.6}$$

Hence, we may solve equation (3.6) for z and then determine x_l^{new} by means of (3.5).

Theorem 3.3 *Given sufficiently accurate approximations x_l', x_l of $(Px_*)'(t_l), x_*(t_l)$.*

(i) Then there is a radius $\varrho > 0$ such that equation (3.6) has a unique solution $z_* \in \bar{B}(0, \varrho)$.

(ii) For $x_l^{\mathrm{new}} := x_l + z_*$ the error estimation

$$|PQ_{1l}(x_l^{\mathrm{new}} - x_*(t_l))| \leq M_1|x_l - x_*(t_l)|^2 + M_2|x_l' - (Px_*)'(t_l)|^2$$

is valid with certain constants M_1, M_2.

Proof:
Introduce the function

$$F(z) := \tilde{f}(x_l', x_l + z) + (I - PQ_{1l})z, \ z \in \bar{B}(0, \varrho)$$

choosing $\varrho > 0$ small enough to realize $L_1\varrho < 1$ and $\{(x_l', x_l + z) : |z| \leq \varrho\} \subset \mathcal{G}$.
Solving equation (3.6) means in fact to look for a zero of F.
A priori, the function F belongs to C^1. Compute

$$\begin{aligned} F'(z) &= \tilde{f}_x'(x_l', x_l + z) + (I - PQ_{1l}) \\ &= PQ_{1l}G_{2l}^{-1}(f_x'(x_l', x_l + z, t_l) - f_x'(x_l', x_l, t_l)) + PQ_{1l} + (I - PQ_{1l}) \\ &= I + r(z), \end{aligned}$$

where $r(z) := \tilde{f}_x'(x_l', x_l + z) - \tilde{f}_x'(x_l', x_l, t_l)$, $|r(z)| \leq L_1|z|$, $z \in \bar{B}(0, \varrho)$.
Further, the Lipschitz condition

$$|F'(z) - F'(\bar{z})| \leq L_1|z - \bar{z}|, \quad z, \bar{z} \in \bar{B}(0, \varrho)$$

results immediately, and the Jacobian $F'(z)$ remains nonsingular on $\bar{B}(0, \varrho)$.

Next we show that Banach's Fixed Point Theorem applies to the map

$$\mathfrak{A}z := z - F(z), \ z \in \bar{B}(0, \varrho) \cap im\ PQ_{1l} =: \mathcal{M}(\varrho).$$

For $z, \bar{z} \in \mathcal{M}(\varrho)$ we derive

$$|\mathfrak{A}z - \mathfrak{A}\bar{z}| \leq \int_0^1 |I - F'(sz + (1 - s)\bar{z})| \,\mathrm{ds}\, |z - \bar{z}| \leq L_1\varrho|z - \bar{z}|$$

and

$$\begin{aligned} \mathfrak{A}z &= z - \tilde{f}(x_l', x_l + z) \\ &= z - (\tilde{f}(x_l', x_l + z) - \tilde{f}(x_l', x_l)) - \tilde{\delta}_l \\ &= z - \textstyle\int_0^1 \tilde{f}_x'(x_l', x_l + sz)\mathrm{ds}\ z - \tilde{\delta}_l \\ &= z - \textstyle\int_0^1 \{\tilde{f}_x'(x_l', x_l + sz) - \tilde{f}_x'(x_l', x_l)\mathrm{ds}\ z - PQ_{1l}z - \tilde{\delta}_l, \end{aligned}$$

thus

$$|\mathfrak{A}z| \leq \frac{1}{2}L_1|z|^2 + |\tilde{\delta}_l| \leq \frac{1}{2}\varrho + |\tilde{\delta}_l|.$$

It becomes clear that, actually, $\mathfrak{A}$ maps $\mathcal{M}(\varrho)$ into itself contractively if the given approximations are sufficiently accurate to satisfy the condition

$$|\tilde{\delta}_l| \leq \frac{1}{2}\varrho, \tag{3.7}$$

i.e., assertion (i) is proved.

Next, consider the resulting error (cf. (3.5))

$$x_l^{\text{new}} - x_*(t_l) = x_l + z_* - x_*(t_l),$$

where $F(z_*) = 0$ as well as x_l and $x_*(t_l)$ are close to each other so that $x_*(t_l) \in \bar{B}(x_l, \varrho)$. Temporarily, we abbreviate $x_* := x_*(t_l), x'_* := (Px_*)'(t_l)$. Recall that $x_l^{\text{new}} - x_l = z_*, PQ_{1l}z_* = z_*$, $F(x_l^{\text{new}} - x_l) = 0$ and derive

$$\begin{aligned} F(x_* - x_l) &= F(x_* - x_l) - F(x_l^{\text{new}} - x_l) \\ &= \textstyle\int_0^1 F'(sx_* + (1-s)x_l^{\text{new}} - x_l)\mathrm{ds}(x_* - x_l^{\text{new}}). \end{aligned}$$

Since the matrix

$$E_l := \int_0^1 F'(sx_* + (1-s)x_l^{\text{new}} - x_l)\mathrm{ds} \tag{3.8}$$

satisfies the inequality $|E_l - I| \leq L_1\varrho < 1$, it is invertible such that we obtain the error representation

$$x_* - x_l^{\text{new}} = E_l^{-1}F(x_* - x_l). \tag{3.9}$$

Due to its construction, E_l has the properties

$$PQ_{1l}(E_l - I) = E_l - I, \quad (I - PQ_{1l})E_l = I - PQ_{1l}$$

which lead to

$$\begin{aligned} (I - PQ_{1l})E_l^{-1} &= (I - PQ_{1l}) \\ PQ_{1l}E_l^{-1} &= E_l^{-1}PQ_{1l} + E_l^{-1}(I - E_l)(I - PQ_{1l}). \end{aligned}$$

Together with (3.9) this implies

$$(I - PQ_{1l})(x_* - x_l^{\text{new}}) = (I - PQ_{1l})(x_* - x_l) \tag{3.10}$$

and

$$PQ_{1l}(x_* - x_l^{\text{new}}) = E_l^{-1}\tilde{f}(x'_l, x_*) + E_l^{-1}(I - E_l)(I - PQ_{1l})(x_* - x_l). \tag{3.11}$$

The first relation (3.10) confirms that exactly the PQ_{1l}-component of the given approximation x_l is changed. Let us take a closer look at the two terms of the right-hand side in formula (3.11). We should consider in some detail that E_l depends on x_l^{new}.

We have

$$\begin{aligned}\tilde{f}(x_l', x_*) &= \tilde{f}(x_l', x_*) - \tilde{f}(x_*', x_*) = \\ &= \int_0^1 \tilde{f}_{x'}'(sx_l' + (1-s)x_*', x_*)\mathrm{ds}(x_l' - x_*') \\ &= \int_0^1 \{\tilde{f}_{x'}'(sx_l' + (1-s)x_*', x_*) - \tilde{f}_{x'}'(x_l', x_l)\}\mathrm{ds}(x_l' - x_*'),\end{aligned}$$

hence

$$|\tilde{f}(x_l', x_*)| \leq \frac{1}{2}L_4|x_l' - x_*'|^2 + L_3|x_l - x_*|\ |x_l' - x_*'|. \tag{3.12}$$

On the other hand, from

$$(E_l - I)(I - PQ_{1l}) = \int_0^1 \{\tilde{f}_x'(x_l', sx_* + (1-s)x_l^{\mathrm{new}}) - \tilde{f}_x'(x_*', x_*)\}(I - PQ_{1l})$$

we derive

$$\begin{aligned}|(E_l - I)(I - PQ_{1l})(x_* - x_l)| &\leq \{\tfrac{1}{2}L_1|x_l^{\mathrm{new}} - x_*| + L_2|x_l' - x_*'|\}|(I - PQ_{1l})(x_* - x_l)| \\ &\leq \{\tfrac{1}{2}L_1|(I - PQ_{1l})(x_l - x_*)| + L_2|x_l' - x_*'|\}|(I - PQ_{1l})(x_* - x_l)| + \\ &\quad + \tfrac{1}{2}L_1\gamma\varrho|PQ_1(x_l^{\mathrm{new}} - x_*)|\end{aligned} \tag{3.13}$$

where $\gamma \leq |I - PQ_{1l}|$ denotes a constant such that $|(I - PQ_{1l})(x_* - x_l)| \leq \gamma\varrho$.
Using the bound $|E_l^{-1}| \leq (1 - L_1\varrho)^{-1}$ and inserting (3.12), (3.13), into (3.11) we find

$$\begin{aligned}|PQ_{1l}(x_l^{\mathrm{new}} - x_*)| &\leq (1 - L_1\varrho)^{-1}\{\tfrac{1}{2}L_4|x_l' - x_*'|^2 + L_3|x_l - x_*||x_l' - x_*'| + \\ &\quad + \tfrac{1}{2}L_1|(I - PQ_{1l})(x_l - x_*)|^2 + L_2|x_l' - x_*'||(I - PQ_{1l})(x_* - x_l)|\} + \\ &\quad + (1 - L_1\varrho)^{-1}L_1\gamma\varrho \cdot \tfrac{1}{2}|PQ_{1l}(x_l^{\mathrm{new}} - x_*)|.\end{aligned} \tag{3.14}$$

Finally, supposing further ϱ to be small enough to satisfy $\varrho L_1(1 + \frac{\gamma}{2}) < 1$, we obtain the estimation offered in assertion (ii) with

$$\begin{aligned}M_1 &:= \tfrac{1}{2}\left(1 - \varrho L_1(1 + \tfrac{\gamma}{2})\right)^{-1}(L_1|I - PQ_{1l}|^2 + L_2|I - PQ_{1l}| + L_3) \\ M_2 &:= \tfrac{1}{2}\left(1 - \varrho L_1(1 + \tfrac{\gamma}{2})\right)^{-1}(L_2 + L_3 + L_4).\end{aligned}$$

□

Corollary 3.4 *In case of Lemma 3.1 assertion (ii) of Theorem 3.3 simplifies to*

$$|PQ_{1l}(x_l^{\mathrm{new}} - x_*(t_l))| \leq \frac{1}{2}L_1\left(1 - \varrho L_1(1 + \frac{\gamma}{2})\right)^{-1}|(I - PQ_{1l})(x_l - x_*)|^2. \tag{3.15}$$

Proof:
The inequality (3.15) is a direct consequence of (3.14). □

Remarks

1. Starting a Newton iteration for solving $F(z) = 0$ with the initial guess $z^{(0)} \in \bar{B}(0,\varrho)$, we find

$$
\begin{aligned}
z^{(1)} - z_* &= z^{(0)} - F'(z^{(0)})^{-1}F(z^{(0)}) - z_* \\
&= F'(z^{(0)})^{-1}\int_0^1 \{F'(z^{(0)}) - F'(sz^{(0)} + (1-s)z_*)\}\mathrm{d}s(z^{(0)} - z_*),
\end{aligned}
$$

$|z^{(1)} - z_*| \leq (1 - L_1\varrho)^{-1} \cdot \frac{1}{2}L_1|z^{(0)} - z_*|^2$.
This shows $\bar{B}(0,\varrho)$ to belong to the region where the Newton iterations converge quadratically.
For $z^{(0)} = 0$, $F'(0) = I$ it results that

$$z^{(1)} = -F(0) = -\tilde{\delta}_l.$$

2. Summarizing the conditions for x_l', x_l to be sufficiently accurate, we have standard restrictions $L_1\varrho(1 + \frac{\gamma}{2}) < 1$ (which implies $L_1\varrho < 1$) and $|x_*(t_l) - x_l| \leq \varrho$, but also $|\tilde{\delta}_l| \leq \frac{1}{2}\varrho$.
In special cases the size of $\tilde{\delta}_l$ is fully governed by x_l (cf. Lemma 3.1) and the P-component of x_l (Hessenberg form DAEs), respectively.

3. Of course, also the simple iterations $z^{[0]} := 0$, $z^{[j+1]} := z^{[j]} - F(z^{[j]})$ remain in $\mathcal{M}(\varrho)$ and converge to z_*.

Lemma 3.5 *It holds that*

$$\tilde{\delta}_l = PQ_{1l}(x_l - x_*(t_l)) + R_l \tag{3.16}$$

with

$$
\begin{aligned}
|R_l| \leq\ & \left(\tfrac{1}{2}L_1 + \tfrac{1}{4}(L_2 + L_3)\right)|x_l - x_*(t_l)|^2 + \\
& + \left(\tfrac{1}{2}L_4 + \tfrac{1}{4}(L_2 + L_3)\right)|x_l' - (Px_*)'(t_l)|^2
\end{aligned}
$$

Proof:
Use once more the abbreviations $x_* := x_*(t_l)$, $x_*' := (Px_*)'(t_l)$. Compute

$$
\begin{aligned}
\tilde{\delta}_l &= \tilde{f}(x_l', x_l) = \tilde{f}(x_l', x_l) - \tilde{f}(x_*', x_*) \\
&= \textstyle\int_0^1 \Big\{\tilde{f}_{x'}'(sx_l' + (1-s)x_*', sx_l + (1-s)x_*)(x_*' - x_l') \\
&\quad + \tilde{f}_x'(sx_l' + (1-s)x_*', sx_l + (1-s)x_*)(x_* - x_l)\Big\}\,\mathrm{ds} \\
&= \textstyle\int_0^1 \Big\{\tilde{f}_{x'}'(sx_l' + (1-s)x_*', sx_l + (1-s)x_*) - \tilde{f}_{x'}'(x_l', x_l)\Big\}\,\mathrm{ds}(x_*' - x_l') \\
&\quad + \textstyle\int_0^1 \Big\{\tilde{f}_x'(sx_l' + (1-s)x_*', sx_l + (1-s)x_*) - \tilde{f}_x'(x_l', x_l)\Big\}\,\mathrm{ds}(x_* - x_l) \\
&\qquad + PQ_{1l}(x_* - x_l) \\
&= R_l + PQ_{1l}(x_* - x_l)
\end{aligned}
$$

and estimate

$$|R_l| \le \tfrac{1}{2}L_4|x'_* - x'_l|^2 \;+\; \tfrac{1}{2}(L_3 + L_2)|x'_* - x'_l||x_* - x_l| \\ \;+\; \tfrac{1}{2}L_1|x_* - x_l|^2.$$

□

As far as the error $x_l^N - x_*(t_l)$ is concerned, Lemma 3.5 indicates what it looks like. Hence, the next theorem is obvious.

Theorem 3.6 *For the first Newton iteration x_l^N given in (2.7) it holds that*

$$(I - PQ_{1l})(x_l^N - x_*(t_l)) = (I - PQ_{1l})(x_l - x_*(t_l))$$

and

$$|PQ_{1l}(x_l^N - x_*(t_l))| \le |R_l|.$$

Corollary 3.7 *For quasi-linear DAEs (3.4) in case of Lemma 3.1 it holds that*

$$|PQ_{1l}(x_l^N - x_*(t_l))| \le \frac{1}{2}L_1|x_l - x_*(t_l)|^2.$$

It should be mentioned that there are (Freude (1995)) certain similar convergence results as described in Theorem 3.3 concerning a more special class of index-2 DAEs, where formally the projector $Q_1((Px_*)'(t_l), x_*(t_l), t_l)$ and the corresponding matrices are used instead of $Q_{1l} = Q_1(x'_l, x_l, t_l)$ etc. in our context. Then, continuity arguments are applied to justify the practical use of Q_{1l}.

Another straightforward generalization of (1.3)-(1.5) would use the new approximation x_l^{new} also for determining the back-projection. In this case, instead of (2.4)-(2.6), one has to solve the nonlinear system

$$x_l^{\text{new}} - x_l - PQ_1(x'_l, x_l^{\text{new}}, t_l)z = 0, \tag{3.17}$$

$$(PQ_1G_2^{-1}f)(x'_l, x_l^{\text{new}}, t_l) + (I - PQ_1(x'_l, x_l^{\text{new}}, t_l))z = 0, \tag{3.18}$$

which is even practically more expensive. One Newton step with the initial guess

$$x_l^{\text{new}(0)} := x_l, \quad z^{(0)} := 0$$

leads to

$$x_l^{\text{new}(1)} = x_l + PQ_{1l}z^{(1)}$$

$$(PQ_1G_2^{-1}f)'_x(x'_l, x_l, t_l)(x_l^{\text{new}(1)} - x_l) + (I - PQ_{1l})z^{(1)} = -\tilde{\delta}_l.$$

Considering

$$\begin{aligned}(PQ_1G_2^{-1}f)'_x(x'_l, x_l, t_l) &= PQ_{1l}G_{2l}^{-1}f'_x(x'_l, x_l, t_l) + \\ &\quad +(PQ_1G_2^{-1})'_x(x'_l, x_l, t_l)f(x'_l, x_l, t_l) = \\ &= PQ_{1l} + (PQ_1G_2^{-1})'_x|_l\delta_l\end{aligned}$$

we know the Jacobian of (3.17), (3.18) to become nonsingular at the initial guess provided that the term $(PQ_1G_2^{-1})'_x|_l\delta_l$ is small enough, which can be ensured for sufficiently accurate approximations x'_l, x_l such that δ_l becomes small.
Hence, the first Newton iteration yields

$$x_l^{\text{new}(1)} = x_l + PQ_{1l}z^{(1)}, \tag{3.19}$$

$$PQ_{1l}(x_l^{\text{new}(1)} - x_l) + (I - PQ_{1l})z^{(1)} = -\tilde{\delta}_l - (PQ_1G_2^{-1})'_x|_l\delta_l PQ_{1l}z^{(1)}, \tag{3.20}$$

that is, $PQ_{1l}z^{(1)}$ in (3.19) is now the solution of the linear equation

$$PQ_{1l}z^{(1)} = -\tilde{\delta}_l - (PQ_1G_2^{-1})'_x|_l\delta_l PQ_{1l}z^{(1)}. \tag{3.21}$$

From this point of view, the projected defect correction formula (2.7) represents an incomplete realization of (3.19), (3.20) that corresponds to a single simple iteration step in the linear equation (3.21) to be solved for $z^{(1)}$.
Finally, emphasize once more that the matrix $(PQ_1G_2^{-1})'_x\delta_l$ seems not to be available in practice in general.

4 Computing the projector Q_{1l} numerically

To realize the projected defect correction (2.4)-(2.6) or (2.7), the projector value $Q_{1l} := Q_1(x'_l, x_l, t_l)$, that is, the projector onto the nullspace $N_1(x'_l, x_l, t_l)$ along the associated subspace $S_1(x'_l, x_l, t_l)$, is needed. Both subspaces are determined by the matrix pair $\{C_l, D_l\}$ via

$$\begin{aligned} C_l &:= f'_{x'}(x'_l, x_l, t_l) + f'_x(x'_l, x_l, t_l)Q, \qquad (4.1)\\ D_l &:= f'_x(x'_l, x_l, t_l)P, \\ N_1(x'_l, x_l, t_l) &= \ker C_l, \\ S_1(x'_l, x_l, t_l) &= \{z \in I\!R^m : D_l z \in im\ C_l\}. \qquad (4.2) \end{aligned}$$

The projector Q used to form C_l, D_l is often known a priori by the DAE structure. Typically, for semi-explicit DAEs, we simply have $Q = diag(0, I)$. If Q is not given in advance, it can be provided easily, say by a Householder decomposition. Note that any projector onto N can be used for Q, i.e., there is no need to have the orthoprojector.

Looking at (4.2), (4.2) we know the projector Q_{1l} to be nothing else but the canonical projector of the matrix pair $\{C_l, D_l\}$ (cf. Griepentrog and März (1986), page 201). Namely, because of the index-2 condition (2.2) for the DAE (2.1), the matrix pair $\{C_l, D_l\}$ has index 1. Hence, the canonical projector of $\{C_l, D_l\}$ is represented by

$$V_l(C_l + D_lV_l)^{-1}D_l, \tag{4.3}$$

where V_l denotes any projector onto the nullspace of C_l. Recall that the matrix $C_l + D_lV_l$ remains nonsingular due to the index-1 property of the pair $\{C_l, D_l\}$.

In the following we describe how formula (4.3) is realized in Freude (1995) as well as in our numerical experiment reported below. For more clarity we drop the index l.
First, we transform C via a Householder decomposition into

$$HC\Pi = \begin{pmatrix} R_{11} & R_{12} \\ 0 & 0 \end{pmatrix} \tag{4.4}$$

such that R_{11} is a nonsingular block. $H, \Pi \in L(\mathbb{R}^m)$ are an orthogonal matrix and a permutation matrix, respectively.
Since $\Pi \begin{pmatrix} 0 & -\ R_{11}^{-1}R_{12} \\ 0 & I \end{pmatrix} \Pi^{-1} =: V$ is a projector onto $\ker C$, we may use it to determine the canonical projector $V(C+DV)^{-1}D$ we are looking for.
For this purpose transform

$$HD\Pi =: \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}$$

and then compute

$$C + DV = H^T H(C+DV)\Pi\Pi^{-1} = H^T \begin{pmatrix} R_{11} & T_{12} \\ 0 & T_{22} \end{pmatrix} \Pi^{-1}$$

with

$$\begin{array}{lllll} T_{12} := & R_{12} & - & S_{11}R_{11}^{-1}R_{12} & + \ S_{12} \\ T_{22} := & & - & S_{21}R_{11}^{-1}R_{12} & + \ S_{22}. \end{array}$$

The block T_{22} is nonsingular. We have

$$(C+DV)^{-1} = \Pi \begin{pmatrix} R_{11}^{-1} & R_{11}^{-1}T_{12}T_{22}^{-1} \\ 0 & T_{22}^{-1} \end{pmatrix} H.$$

Further, we obtain the canonical projector as

$$Q_{1l} = V(C+DV)^{-1}D = \Pi \begin{pmatrix} -R_{11}^{-1}R_{12}T_{22}^{-1}S_{21} & -R_{11}^{-1}R_{12}T_{22}^{-1}S_{22} \\ T_{22}^{-1}S_{21} & T_{22}^{-1}S_{22} \end{pmatrix} \Pi^{-1}.$$

It becomes clear that, in order to compute Q_{1l}, we have to realize basically the decomposition (4.4) and to invert the blocks R_{11}, T_{22}.

Before starting the procedure, a standard scaling of C is recommended such that each row has a max-norm equal to one. Of course, D has to be scaled by the same diagonal matrix, too.

For very sensitive problems we have good experience with a more expensive scaling and preconditioning, respectively. Having carried out the decomposition (4.4) and having computed $(C+DV)^{-1}$, we use this matrix for scaling the pair $\{C, D\}$ to $\{(C+DV)^{-1}C, (C+$

$DV)^{-1}D\} =: \{\tilde{C}, \tilde{D}\}$ and restart the procedure for $\{\tilde{C}, \tilde{D}\}$. Note that in this preconditioning any projector $\bar{V} \in I\!R^m$ onto $\ker C$ can be used instead of V.
For the particular case of Hessenberg form DAEs (1.1), (1.2) we have, with $Q = diag(0, I)$, $B_{ik} := \frac{\partial b_i}{\partial x_k}$, at the very beginning

$$C = \begin{pmatrix} I & B_{12} \\ 0 & 0 \end{pmatrix}$$

such that the decomposition (4.4) is trivially given and $H = I, \Pi = I, R_{11} = I, R_{12} = B_{12}$.
Then it holds that

$$D = \begin{pmatrix} B_{11} & 0 \\ B_{12} & 0 \end{pmatrix} = S$$

i.e. $S_{11} = B_{11}, S_{21} = B_{12}, S_{12} = 0, S_{22} = 0$.
This yields

$$T_{12} = B_{12} - B_{11}B_{12}, \quad T_{22} = -B_{21}B_{12}$$

and finally

$$Q_1 = \begin{pmatrix} B_{12}(B_{21}B_{12})^{-1}B_{12} & 0 \\ -(B_{21}B_{12})^{-1}B_{12} & 0 \end{pmatrix}.$$

Again, things become much simpler in case of Hessenberg form DAEs.

5 The case of semi-explicit DAEs and selective projected methods

In this section we deal with linear DAEs

$$Ax' + Bx = q \tag{5.1}$$

whose coefficients have the special form

$$A = \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}, \tag{5.2}$$

that is, with semi-explicit DAEs. For this kind of equations, the respective subspaces and matrices from Section 2 are much simpler. Pointwise for all $t \in J$ we have (the argument t is dropped)

$$\begin{aligned}
N &= \left\{ \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} : z_1 = 0 \right\}, S = \left\{ \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} : B_{21}z_1 + B_{22}z_2 = 0 \right\}, \\
G_1 &= \begin{pmatrix} I & B_{12} \\ 0 & B_{22} \end{pmatrix} \quad \text{with} \quad Q = \begin{pmatrix} 0 & 0 \\ 0 & I \end{pmatrix}, \\
N_1 &= \left\{ \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} : z_1 + B_{12}z_2 = 0, \ B_{22}z_2 = 0 \right\}, \\
S_1 &= \left\{ \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} : B_{21}z_1 \in im B_{22} \right\},
\end{aligned}$$

such that

$$\begin{aligned} N \cap S &= \left\{ \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} : z_1 = 0, \quad z_2 \in \ker B_{22} \right\}, \\ N_1 \cap S_1 &= \left\{ \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} : z_1 = -B_{12} z_2, \quad z_2 \in \ker B_{22}, \quad B_{21} B_{12} z_2 \in im B_{22} \right\}. \end{aligned}$$

By definition, index-2 tractability means that B_{22} is singular but has constant rank, and that the relations $z_2 \in \ker B_{22}$, $B_{21}B_{12}z_2 \in imB_{22}$ imply $z_2 = 0$. Hence, the following index-2 criterion for (5.1) results immediately.

Lemma 5.1 *The DAE (5.1), (5.2) is index-2 tractable if and only if the matrix pair* $\{B_{22}, -B_{21}B_{12}\}$ *represents a regular index-1 matrix pencil, or equivalently, if the matrix* $B_{22} - B_{21}B_{12}Q_B$ *remains nonsingular for any projector function* Q_B *onto* $\ker B_{22}$.

Next we assume the index-2 conditions to be satisfied and denote

$$\begin{aligned} H &:= -B_{12} Q_B (B_{22} - B_{21} B_{12} Q_B)^{-1} B_{21} = B_{12} K \\ K &:= -Q_B (B_{22} - B_{21} B_{12} Q_B)^{-1} B_{21}. \end{aligned} \tag{5.3}$$

It is easily checked that KB_{12} is a projector function which projects onto $\ker B_{22}$ along $\{z_2 : -B_{21}B_{12}z_2 \in imB_{22}\}$. Then also $H^2 = H$ may be proved, i.e. H is a projector function, too. For a better understanding of this fact we recall once more the Hessenberg form case where we have, in particular,

$$B_{22} = 0, \ Q_B = I, \ H = B_{12}(B_{21}B_{12})^{-1}B_{21}, \ KB_{12} = I.$$

Return to the general semi-explicit case and construct the projector functions

$$Q_1 = \begin{pmatrix} H & 0 \\ -K & 0 \end{pmatrix}, \quad PQ_1 = \begin{pmatrix} H & 0 \\ 0 & 0 \end{pmatrix}.$$

Q_1 may be realized to project onto N_1 along S_1 in fact. Further, derive

$$G_2 = G_1 + BPQ_1 = \begin{pmatrix} I + B_{11}H & B_{12} \\ B_{21}H & B_{22} \end{pmatrix}$$

and

$$PQ_1 G_2^{-1} = \begin{pmatrix} 0 & F \\ 0 & 0 \end{pmatrix} \tag{5.4}$$

with $F := -B_{12}Q_B(B_{22} - B_{21}B_{12}Q_B)^{-1}$.
Obviously, the corresponding things for Hessenberg form DAEs are generalized now in a very straightforward way.
The projected defect correction formula (2.7) for a semi-explicit system reads in detail

$$x_{l,1}^{\text{new}} = x_{l,1} - F_l \delta_{l,2}, \tag{5.5}$$

$$x_{l,2}^{\text{new}} = x_{l,2}. \tag{5.6}$$

Ascher and Spiteri (1994) propose to treat semi-explicit index-2 DAEs (5.1), (5.2) via a pointwise singular value decomposition of B_{22}, i.e.,

$$B_{22} = U \sum V^T, \tag{5.7}$$

where U, V are orthogonal matrices and $\sum = diag(S, 0)$ with S a nonsingular diagonal matrix of size $\mu < m - rankA$.
Let $U = (U_1 U_2)$, $V = (V_1 V_2)$, where U_1 and V_1 consist of the first μ columns of U and V, respectively. Using the decomposition (5.7), the backprojection step is proposed to be realized in the same way as for Hessenberg systems, but now with $B_{12}V_2$ and $U_2^T B_{21}$ instead of B_{12} and B_{21}. This procedure is called a selective projected method. The back-projection is carried out only onto the constraint manifold given by $U_2^T B_{21} x_1 - U_2^T q_2 = 0$, which corresponds in fact to the so-called *pure index-2 part*.

If such a singular value decomposition (5.7) is available, we may choose

$$Q_B = V \begin{pmatrix} 0 & 0 \\ 0 & I \end{pmatrix} V^T \tag{5.8}$$

so that F simplifies further to

$$F = B_{12} V \begin{pmatrix} 0 & 0 \\ 0 & N^{-1} \end{pmatrix} U^T = B_{12} V_2 N^{-1} U_2^T, \tag{5.9}$$

where $N := U_2^T B_{21} B_{12} V_2$.
The resulting projected defect correction formula (5.5), (5.6) is exactly the same one as we would obtain by the selective backprojection in Ascher and Spiteri (1994). However, note that in our context there is no need for an orthoprojector Q_B in (5.4) at all.
Of course, if the singular value decomposition (5.7) is given, the choice (5.8) is natural. But, fortunately, one can do without the more expensive decomposition (5.7).

Let us mention that (5.7), (5.8) imply the representation

$$B_{22} - B_{21} B_{12} Q_B = U \begin{pmatrix} S & -U_1^T B_{21} B_{12} V_2 \\ 0 & -U_2^T B_{21} B_{12} V_2 \end{pmatrix} V^T,$$

thus the following assertion becomes true.

Corollary 5.2 *Given the decomposition (5.7). Then the DAE (5.1), (5.2) is index-2 tractable iff the matrix $N = U_2^T B_{21} B_{12} V_2$ remains nonsingular.*

Hence, the corresponding index-2 condition used in Ascher and Spiteri (1994) is confirmed once more.

6 Projected defect correction in the BDF - an experiment

The BDF is well-known to belong to the class of numerical integration formulas that do not require a backward projection theoretically, since the relevant defect disappears in theory. Nevertheless, we are confronted with nontrivial defects in all practical computations. We apply the projected defect correction after each standard BDF-step to suppress the critical defect additionally. As the following experiment shows, this may improve the performance of the numerical integration considerably.
For more transparency we choose a small DAE that is even in Hessenberg form for carrying out our experiments.
Consider the system (Freude (1995))

$$\begin{aligned}
x_1'(t) - x_4(t) + x_5(t) &= 0, && (6.1)\\
x_2'(t) + 2(x_4(t)x_5(t))^{\frac{1}{2}} &= 0, && (6.2)\\
x_3'(t) - 5 &= 0, && (6.3)\\
25\sin(\arcsin(x_1(t)^3) - 75\sin\left(\frac{1}{375}x_3(t)^3\right) + 100\sin^3\left(\frac{1}{3}t^3\right) &= 0, && (6.4)\\
2x_1(t)x_2(t) - \sin\left(\frac{2}{5}x_3(t)\right) &= 0, && (6.5)
\end{aligned}$$

which has the solution of moderate size and behaviour

$$x_*(t) = \begin{pmatrix} \sin(t) \\ \cos(t) \\ 5t \\ \cos^2(\frac{1}{2}t) \\ \sin^2(\frac{1}{2}t) \end{pmatrix}, \quad t \in [0.1, 0.5].$$

The DAE (6.1) - (6.5) is index-2 tractable around the graph of this solution. In particular, the corollaries 3.4 and 3.7 apply. We simply have

$$Q = \begin{pmatrix} 0 &&&& \\ & 0 &&& \\ && 0 && \\ &&& 1 & \\ &&&& 1 \end{pmatrix}, \quad P = \begin{pmatrix} 1 &&&& \\ & 1 &&& \\ && 1 && \\ &&& 0 & \\ &&&& 0 \end{pmatrix},$$

i.e., the first three components x_1, x_2, x_3 represent Px, whereas the last two ones represent the nullspace component Qx. The critical defect part is the one arising in the equations (6.4), (6.5).

For various given tolerances TOL required for the numerical solution we have used the backward Euler and the two-step BDF with various tolerances NTOL to control the Newton iterations in the nonlinear systems to be solved per integration step. All computations

were realized by means of the BDF-implementation *dae2sol.for* proposed in Tischendorf (1992).
The four tables below contain the following information:

TOL :	required tolerances for the numerical DAE solution
NTOL :	tolerances to control the Newton iteration
PDC? :	is a projected defect correction realized or not
Q :	absolute error of the nullspace components at the end of the integration interval
δ ?	critical defect at the interval end
STEPS :	number of accepted integration steps when applying the standard error control of *dae2sol*, which is related to the P-components only
STEPS* :	number of accepted steps when controlling all error components
Q^* :	error of the nullspace components when controlling all components

Table 1: Backward Euler method, TOL $= 10^{-4}$

NTOL	PDC?	Q	δ	STEPS	STEPS*	Q^*
1.0-4	no	-	-	-	-	-
	yes	1.9-3	1.0-10	79	483	1.5-4
1.0-5	no	-	-	-	-	-
	yes	1.9-3	4.3-14	79	471	2.3-4
1.0-6	no	-	-	-	-	-
	yes	1.9-3	2.8-14	79	472	1.3-4
1.0-7	no	-	-	-	-	-
	yes	1.9-3	2.8-14	79	472	2.9-4
1.0-8	no	-	-	-	-	-
	yes	3.2-3	5.7-14	81	474	2.8-4
1.0-9	no	1.9-3	2.0-8	79	472	2.2-4
	yes	1.9-3	2.4-13	79	472	2.2-4
1.0-10	no	4.0-3	9.6-9	81	466	5.6-5
	yes	4.0-3	1.4-14	81	466	5.6-5
1.0-11	no	-	-	-	-	-
	yes	-	-	-	-	-

Table 2: Backward Euler method, TOL $= 10^{-5}$

NTOL	PDC?	Q	δ	STEPS	STEPS*	Q^*
1.0-5	no	-	-	-	-	-
	yes	6.8-4	2.8-13	235	1554	7.0-5
1.0-6	no	-	-	-	-	-
	yes	6.8-4	1.7-13	235	2064	1.1-5
1.0-7	no	-	-	-	-	-
	yes	6.8-4	1.3-13	235	1917	5.1-5
1.0-8	no	6.8-4	1.8-10	235	2055	1.0-5
	yes	6.8-4	1.3-13	235	1910	1.4-5

Table 3: Two-step BDF, TOL $= 10^{-4}$

NTOL	PDC?	Q	δ	STEPS	STEPS*	Q^*
1.0-4	no	-	-	-	-	-
	yes	5.2-4	4.3-14	28	76	3.4-4
1.0-5	no	-	-	-	-	-
	yes	5.3-4	1.4-14	28	39	8.1-4
1.0-6	no	-	-	-	-	-
	yes	5.3-4	1.4-14	28	40	7.9-4
1.0-7	no	-	-	-	-	-
	yes	5.3-4	1.4-14	28	39	8.2-4
1.0-8	no	-	-	-	-	-
	yes	7.2-4	8.5-14	30	61	7.6-4
1.0-9	no	5.3-4	5.7-9	28	59	7.6-4
	yes	5.3-4	4.3-14	28	59	7.6-4

Table 4: Two-step BDF, TOL $= 10^{-5}$

NTOL	PDC?	Q	δ	STEPS	STEPS*	Q^*
1.0-5	no	-	-	-	-	-
	yes	1.4-6	2.9-16	53		3.8-5
1.0-6	no	-	-	-	-	-
	yes	1.4-6	7.1-14	53	300	4.5-6
1.0-7	no	-	-	-	-	-
	yes	1.4-6	2.1-13	53	858	1.0-6
1.0-8	no	1.4-6	5.1-11	53	857	1.2-6
	yes	1.4-6	9.9-14	53	854	1.1-5

Note that similar results were obtained by the three-step BDF and so on.
For the projected defect correction we used formula (2.7). The needed projections are computed as described in Section 4. The same experiments were realized also with 5 Newton iterations applied to the defect correction system (2.4)-(2.6) instead of using formula (2.7) in Freude (1995), but no significant differences were observed.

We do not quote the errors in the P-components. All of them satisfy the given tolerances very well.
The typical situation is shown in Table 1. For very fine tolerances NTOL of about 10^{-11} the integration fails completely, independently whether the projected defect correction step is carried out or not. For coarser tolerances NTOL the integration without projected defect correction fails, but it performs well if the projected defect correction is realized. Further, there is a (sometimes very tight) range of tolerances NTOL such that both versions, with and without defect correction, work. Then, the results provided are essentially the same except for the critical defect that is smaller in the corrected version. Hence, the BDF integration completed with projected defect correction steps shows a much more robust and reliable performance.

Acknowledgement:

My thanks are due to my colleague M. Beer for carrying out the computations as well as to C. Tischendorf and R. Lamour for stimulating discussions.

References

U.M. Ascher and L.R. Petzold (1991): Projected implicit Runge-Kutta methods for differential-algebraic equations. SIAM J. Numer. Anal. 28 (4), 1097-1120

E. Hairer and G. Wanner (1991): Solving ordinary differential equations II. Springer-Verlag Berlin Heidelberg

L.F. Shampine (1986): Conservation laws and the numerical solution of ODEs. Comp. and Maths. with Applications 12, 1287-1296

E. Eich: Projizierende Mehrschrittverfahren zur numerischen Lösung von Bewegungsgleichungen Technischer Mehrkörpersysteme mit Zwangsbedingungen und Unstetigkeiten. Fortschrittsberichte VDI, Reihe 18, Nr. 109, VDI-Verlag, Düsseldorf, Germany, 1992

V.H. Schulz, H.G. Bock and M.C. Steinbach (1996): Exploiting invariants in the numerical solution of multipoint boundary value problems for DAEs. SIAM J. Sci. Comp., to appear

C.W. Gear (1986): Maintaining solution invariants in the numerical solution of ODEs. SIAM J. Sci. Stat. Comp. 7, 734-743

R. März (1995): On linear differential-algebraic equations and linearizations. Applied Numerical Mathematics 18, 267-292

R. März (1992): Numerical methods for differential-algebraic equations. Acta Numerica 1992, 141-198

C. Tischendorf (1996): Solution of index-2 differential-algebraic equations and its application in circuit simulation. Dissertation, Humboldt-Universität Berlin, Institut für Mathematik

R. März and C. Tischendorf (1996) : Recent results in solving index-2 differential-algebraic equations in circuit simulation. SIAM J. Sci. Comp., to appear

S. Freude (1995): Projizierende Defektkorrektur für Algebro-Differentialgleichungen mit dem Index 2. Diplomarbeit, Humboldt-Universität Berlin, Institut für Mathematik

E. Griepentrog and R. März (1986): Differential-algebraic equations and their numerical treatment. Teubner-Texte zur Mathematik 88, BSB B.G. Teubner Verlagsgesellschaft, Leipzig

U.M. Ascher and R.J. Spiteri (1994): Collocation software for boundary value differential-algebraic equations. SIAM J. Sci. Comp. 4(1994) 15, 938-952

C. Tischendorf (1992): Die BDF für nichtlineare Algebro-Differentialgleichungen vom Index 2. Diplomarbeit, Humboldt-Universität Berlin, Institut für Mathematik