Sensitivity of solutions of linear DAE to perturbations of the system matrices

EINDHOVEN UNIVERSITY OF TECHNOLOGY
Department of Mathematics and Computing Science

RANA 97-21
December 1997

Sensitivity of solutions of linear DAE to perturbations of the system matrices

by

R.M.M. Mattheij and P.M.E.J. Wijckmans

Reports on Applied and Numerical Analysis
Department of Mathematics and Computing Science
Eindhoven University of Technology
P.O. Box 513
5600 MB Eindhoven
The Netherlands
ISSN: 0926-4507

Sensitivity of solutions of linear DAE to perturbations of the system matrices

R.M.M. Mattheij [a] and P.M.E.J. Wijckmans [b]

[a] *Department of Mathematics and Computing Science, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands*

[b] *TNO Physics and Electronics Laboratory, P.O. Box 96864, 2509 JG 's-Gravenhage, The Netherlands*

This paper studies the effect of perturbations in the system matrices of linear Differential Algebraic Equations (DAE) onto the solutions. It turns out that these may result in a more complicated perturbation pattern for higher index problems than in the case for (standard) additive perturbations. The analysis which has clear ramifications in nonlinear problems (cf. Newton) is sustained by a number of examples.

Keywords: conditioning, perturbations, DAE

AMS Subject classification: 65L05,65L20

1. Introduction

In an earlier paper [4] we investigated the effect of additive perturbations on the solution of two-deck systems

$$\dot{\mathbf{x}}(t) = \mathbf{A}(t)\mathbf{x}(t) + \mathbf{B}(t)\mathbf{y}(t) + \mathbf{p}(t), \tag{1.1a}$$

$$\mathbf{0} = \mathbf{C}(t)\mathbf{x}(t) + \mathbf{D}(t)\mathbf{y}(t)\mathbf{q}(t), \tag{1.1b}$$

given suitable initial conditions. This problem is in fact equivalent to determining the conditioning constants of this problem; they provide bounds for the solution $\binom{\mathbf{x}}{\mathbf{y}}$ in terms of bounds for the initial values and $\mathbf{p}$, $\mathbf{q}$ and a suitable number of the derivatives of the latter (depending on the index of the problem). One of the conclusions in [4, 6] was that de facto a problem may behave like a higher index problem; in particular when $\mathbf{D}$ is nearly singular the index is effectively 2 or higher.

In the present paper we shall be concerned with the of the solution with respect to perturbations in the system matrices $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ and $\mathbf{D}$. As it will turn out this doesn't necessitate any particular analysis for Ordinary Differential Equations (ODE) and even for index-1 problems it is fairly simple to deal with. However, for higher index problems the matter is more complex. For index-2 problems we shall show that ill-conditioning in the more classical sense, as indicated above, may imply an extreme sort of ill-conditioning when perturbations of the system matrices are taken into account. Although an analysis of the linear problem has a justification in its own right it is also meaningful to understand (and therefore potentially prevent) difficulties in approximation methods where linearization is essential, like Newton's method.

The paper is set up as follows. First we shall consider how the conditioning constants for an ODE are perturbed if we perturb the system matrices, see Section 2. This provides for a general framework, as we shall study so called underlying ODE for index-1 and index-2 DAE. In Section 3 we investigate the sensitivity of the solutions of index-1 DAE. As it turns out this doesn't pose larger instabilities, unless it is "close" to a higher index problem. In Section 4 a general analysis is given for the index-2 case. This is worked out in more detail in Section 5 for various specific matrix perturbations.

2. The ODE case

For ODE it is fairly straightforward to give a bound for the errors in the solutions due to perturbations of the system. In fact there is a direct link to perturbations of the source term (the "standard" regular perturbations). We may proceed as follows. Consider the IVP

$$\mathbf{y}' = \mathbf{A}\mathbf{y} + \mathbf{f}, \quad \text{with } \mathbf{y}(t_0) = \boldsymbol{\beta}, \tag{2.1}$$

on the interval $(t_0, t_1) =: I$. Now let $\boldsymbol{\Phi}$ be a fundamental solution of (2.1). Then we can define two conditioning constants for this problem, which indicate the sensitivity with respect to perturbations of the initial data and source tems respectively (cf. [2]):

$$\kappa_1 = \max_{t \in I} \|\boldsymbol{\Phi}(t)\boldsymbol{\Phi}^{-1}(t_0)\|, \tag{2.2a}$$

(usually $\boldsymbol{\Phi}(t_0) = \mathbf{I}$)

$$\kappa_2 = \sup_{t_0 < t \leq t_1} \int_{t_0}^{t} \|\boldsymbol{\Phi}(t)\boldsymbol{\Phi}^{-1}(s)\| ds. \tag{2.2b}$$

Now consider the following IVP which results from (2.1) by perturbing not only the source term $\mathbf{f}$, but also the system matrix $\mathbf{A}$ (perturbations of the initial value are already "dealt with" by κ_1)

$$\mathbf{z}' = \tilde{\mathbf{A}}\mathbf{z} + \tilde{\mathbf{f}}, \quad \text{with } \mathbf{z}(t_0) = \boldsymbol{\beta}, \tag{2.3}$$

where

$$\tilde{\mathbf{A}} = \mathbf{A} + \delta\mathbf{A}, \text{ and } \tilde{\mathbf{f}} = \mathbf{f} + \delta\mathbf{f}. \tag{2.4}$$

Then (2.3) yields

$$\mathbf{z}' = \mathbf{A}\mathbf{z} + \mathbf{f} + \delta\mathbf{A}\mathbf{z} + \delta\mathbf{f}, \quad \text{with } \mathbf{z}(t_0) = \boldsymbol{\beta}.$$

Let $\tilde{\boldsymbol{\Phi}}$ denote the fundamental solution of(2.3). Like for (2.1) we can define conditioning constants $\tilde{\kappa}_1$ and $\tilde{\kappa}_2$ for (2.2), defined like in (2.2a) and (2.2b) (through replacing $\boldsymbol{\Phi}$ by $\tilde{\boldsymbol{\Phi}}$). Then, the solution of (2.3), is given by

$$\begin{aligned} \mathbf{z}(t) &= \tilde{\boldsymbol{\Phi}}(t)\boldsymbol{\beta} + \int_{t_0}^{t} \tilde{\boldsymbol{\Phi}}(t)\tilde{\boldsymbol{\Phi}}^{-1}(s)\tilde{\mathbf{f}}(s)ds \\ &= \boldsymbol{\Phi}(t)\boldsymbol{\beta} + \int_{t_0}^{t} \boldsymbol{\Phi}(t)\boldsymbol{\Phi}^{-1}(s)(\delta\mathbf{A}(s)\mathbf{z}(s) + \mathbf{f}(s) + \delta\mathbf{f}(s))ds \\ &= \boldsymbol{\Phi}(t)\boldsymbol{\beta} + \int_{t_0}^{t} \boldsymbol{\Phi}(t)\boldsymbol{\Phi}^{-1}(s)(\delta\mathbf{A}(s)\mathbf{z}(s) + \tilde{\mathbf{f}}(s))ds \end{aligned} \tag{2.5}$$

As a consequence,

$$\tilde{\boldsymbol{\Phi}}(t) = \boldsymbol{\Phi}(t) + \int_{t_0}^{t} \boldsymbol{\Phi}(t)\boldsymbol{\Phi}^{-1}(s)\delta\mathbf{A}(s)\tilde{\boldsymbol{\Phi}}(s)ds. \tag{2.6}$$

Let $\mathbf{G}(t,s)$ and $\tilde{\mathbf{G}}(t,s)$ denote the "Green functions" of (2.1) and (2.3), respectively. Then

$$\tilde{\mathbf{G}}(t,s) = \mathbf{G}(t,s) + \boldsymbol{\Phi}(t)\tilde{\mathbf{G}}(t_0,t) - \int_{t_0}^{t} \mathbf{G}(t,s)\delta\mathbf{A}(s)\tilde{\mathbf{G}}(t,s)ds. \tag{2.7}$$

With the definition

$$|\delta\mathbf{A}| \le \varepsilon_{\mathbf{A}},$$

the aforementioned relations make it possible to compare the conditioning constants of the two neighbouring problems (2.1) and (2.3). If $\kappa_2\varepsilon_{\mathbf{A}} < 1$, then

$$\tilde{\kappa}_1 \le \frac{\kappa_1}{1 - \kappa_1\varepsilon_\beta - \kappa_2\varepsilon_{\mathbf{A}}}, \quad \tilde{\kappa}_2 \le \frac{\kappa_1}{1 - \kappa_1\varepsilon_\beta - \kappa_2\varepsilon_{\mathbf{A}}}, \tag{2.8}$$

This means that a well conditioned IVP remains reasonably conditioned if $\mathbf{A}$ and the initial conditions are perturbed only slightly. From relations (2.6) and (2.7) we can conclude that $\tilde{\boldsymbol{\Phi}}$ and $\boldsymbol{\Phi}$ may differ a lot if $\|\delta\mathbf{A}\|$ is large. The same holds for $\tilde{\mathbf{G}}$ and $\mathbf{G}$.

In the sequel we shall derive underlying ODE for index-1 and index-2 problems. Once we have established some sort of bound for the perturbations in the system matrices of the latter we therefore will consider the analysis being completed because of the estimates in this section

3. DAE of index-1

Consider the DAE (1.1). If $\mathbf{D}$ is nonsingular we clearly obtain, by solving $\mathbf{y}$ from the second part,

$$\dot{\mathbf{x}} = (\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})\mathbf{x} + \mathbf{p} - \mathbf{B}\mathbf{D}^{-1}\mathbf{q}. \tag{3.1}$$

We are particularly interested in the case where $\mathbf{D}$ is almost singular. Like in [4, 6] we therefore investigate this by setting $\mathbf{D} = \varepsilon\mathbf{I}$, $\varepsilon > 0$ (which avoids complicated projections on "nearly singular" parts). We then may say that (1.1) has nearly index-2, a clearly ill-conditioned index-1 problem. Hence, the homogeneous part of the underlying ODE (3.1) then reads

$$\dot{\mathbf{z}} = (\mathbf{A} - \varepsilon^{-1}\mathbf{B}\mathbf{C})\mathbf{z}. \tag{3.2}$$

Rather than giving estimates with rigorous bounds we like to show that (3.2) should be regarded as well conditioned with respect to perturbations in the system matrices $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$, provided $\mathbf{CB}$ is a well conditoned matrix. For simplicity all matrices are assumed to be constant. It is not restrictive to even assume then that $\mathbf{B}$ has orthogonal columns (note that also $\mathbf{C}$ cannot be nearly row rank deficient). Let $\mathbf{B}^{\perp}$ denote a matrix $\in \mathbb{R}^{n\times(n-m)}$ such that

$$\mathbf{T} := \left[\mathbf{B}\,\mathbf{B}^{\perp}\right] \tag{3.3}$$

is orthogonal. Then we have

$$\mathbf{T}^{-1}\mathbf{B}\mathbf{C}\mathbf{T} = \begin{bmatrix} \mathbf{CB} & \mathbf{CB}^{\perp} \\ \mathbf{O} & \mathbf{O} \end{bmatrix}. \tag{3.4}$$

If we perturb $\mathbf{B}$ by $\delta\mathbf{B}$, say

$$\tilde{\mathbf{B}} := \mathbf{B} + \delta\mathbf{B}, \tag{3.5}$$

then we can define an orthogonal complement, $\tilde{\mathbf{B}}^{\perp}$ say, and define the matrix (cf. (3.3))

$$\tilde{\mathbf{T}} := \left[\tilde{\mathbf{B}}\,\tilde{\mathbf{B}}^{\perp}\right]. \tag{3.6}$$

The well conditioning of $\mathbf{T}$ implies that

$$\tilde{\mathbf{T}} = \mathbf{T}(\mathbf{I} + \mathbf{E}), \ \|\mathbf{E}\| \text{ small}. \tag{3.7}$$

Here $\|.\|$ is some Höldernorm.
Now consider the perturbed equation (cf. (3.2))

$$\dot{\tilde{\mathbf{z}}} = (\tilde{\mathbf{A}} - \varepsilon^{-1}\tilde{\mathbf{B}}\tilde{\mathbf{C}})\tilde{\mathbf{z}} \tag{3.8}$$

where we have

$$\tilde{\mathbf{A}} := \mathbf{A} + \delta\mathbf{A}, \quad \tilde{\mathbf{B}} := \mathbf{B} + \delta\mathbf{B}, \quad \tilde{\mathbf{C}} := \mathbf{C} + \delta\mathbf{C}, \tag{3.9}$$

for some perturbations $\delta\mathbf{A}$, $\delta\mathbf{B}$, $\delta\mathbf{C}$, respectively. Transformation of the system matrix in (3.8) through $\tilde{\mathbf{T}}$ yields

$$\tilde{\mathbf{T}}^{-1}\tilde{\mathbf{A}}\tilde{\mathbf{T}} - \frac{1}{\varepsilon}\begin{bmatrix} \tilde{\mathbf{C}}\tilde{\mathbf{B}} & \tilde{\mathbf{C}}\tilde{\mathbf{B}}^{\perp} \\ \mathbf{O} & \mathbf{O} \end{bmatrix} =: \mathbf{M}_1 - \frac{1}{\varepsilon}\mathbf{M}_2. \tag{3.10}$$

The second matrix $\mathbf{M}_2$ in (3.10) will have slightly perturbed eigenvalues compared to $\mathbf{CB}$. Indeed, well-conditioning of $\mathbf{CB}$ means that its eigenvalues are $\mathcal{O}(1)$. Because of the factor $\frac{1}{\varepsilon}$ they must be positive to make (3.2) stable (in Lyaponov sense). Hence from classical arguments (like the Theorem of Bauer-Fike, cf. [3]) we conclude that $\frac{1}{\varepsilon}\mathbf{M}_2$ still has eigenvalues of $\mathcal{O}(\frac{1}{\varepsilon})$, being positive moreover. For the first matrix in (3.10), $\mathbf{M}_1$, we can write

$$\mathbf{M}_1 = \tilde{\mathbf{T}}^{-1}\delta\mathbf{A}\tilde{\mathbf{T}} + \tilde{\mathbf{T}}^{-1}\mathbf{A}\tilde{\mathbf{T}} \pm \tilde{\mathbf{T}}^{-1}\mathbf{A}\mathbf{T} \pm \mathbf{T}^{-1}\mathbf{A}\mathbf{T}. \tag{3.11}$$

Hence

$$\begin{aligned} \|\tilde{\mathbf{T}}^{-1}\tilde{\mathbf{A}}\tilde{\mathbf{T}} - \mathbf{T}^{-1}\mathbf{A}\mathbf{T}\| \leq& \|\tilde{\mathbf{T}}^{-1}\|\,\|\tilde{\mathbf{T}}\|\,\|\delta\mathbf{A}\| + \|\tilde{\mathbf{T}}^{-1}\mathbf{A}\|\,\|\tilde{\mathbf{T}} - \mathbf{T}\| \\ &+ \|\tilde{\mathbf{T}}^{-1} - \mathbf{T}^{-1}\|\,\|\mathbf{A}\mathbf{T}\|. \end{aligned} \tag{3.12}$$

Because of (3.7) we see that there exists a constant κ, not much larger than 1, such that

$$\|\tilde{\mathbf{T}}^{-1}\tilde{\mathbf{A}}\tilde{\mathbf{T}} - \mathbf{T}^{-1}\mathbf{A}\mathbf{T}\| \leq \kappa\left[\|\delta\mathbf{A}\| + 2\|\mathbf{E}\|\,\|\mathbf{A}\|\right]. \tag{3.13}$$

We conclude that the "fast" eigenvalues of $\mathbf{A} - \frac{1}{\varepsilon}\mathbf{BC}$ are hardly perturbed by contributions from $\tilde{\mathbf{T}}^{-1}\tilde{\mathbf{A}}\tilde{\mathbf{T}}$, whereas the "slow" ones (which may be found through a suitable Ricatti transformation in (3.10), cf. [2]) are moderate perturbations of the "slow" eigenvalues of $\mathbf{A} - \frac{1}{\varepsilon}\mathbf{BC}$ (i.e. of $\mathbf{A}$ in fact). If the original system (3.2) is stable, then the perturbed system will also be stable for perturbations which are small compared to the magnitude of the eigenvalues. Basically we may now perform an analysis like in Section 2 applied to (3.2) and (3.8). One should also note that it is *essential* that the underlying ODE (3.2) is not close to a problem with index ≥ 3, i.e. that $\mathbf{CB}$ is not (nearly) singular, see the next example (and [5]).

Example 3.1. Consider the DAE (3.1), where

$$\mathbf{A} = \begin{bmatrix} -1 & -1 \\ 0 & 0 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \mathbf{C} = \begin{bmatrix} 0 & 1 \end{bmatrix}, \quad \mathbf{D} = \varepsilon.$$

Let $\tilde{\mathbf{B}} = \left[\begin{smallmatrix}1\\ \delta\end{smallmatrix}\right]$. Then

$$\mathbf{A} - \frac{1}{\varepsilon}\tilde{\mathbf{B}}\mathbf{C} = \begin{bmatrix} -1 & -1 - \frac{1}{\varepsilon} \\ 0 & -\frac{\delta}{\varepsilon} \end{bmatrix}. \tag{3.14}$$

In the unperturbed case ($\delta = 0$) we have eigenvalues -1 and 0. In the perturbed case the latter is changed into $-\frac{\delta}{\varepsilon}$, which can be quite dramatic, in particular if $\delta < 0$. □

4. Index-2 DAE

If the DAE (1.1) has index 2 we have $\mathbf{D} = \mathbf{O}$ and $\mathbf{CB}$ nonsingular. One can easily obtain an underlying ODE by introducing the projection

$$\mathbf{P} := \mathbf{B}(\mathbf{CB})^{-1}\mathbf{C}. \tag{4.1}$$

We now define a new variable $\mathbf{z}$ with

$$\mathbf{z} := (\mathbf{I} - \mathbf{P})\mathbf{x}. \tag{4.2}$$

This implies that $\mathbf{z} = \mathbf{x} + \mathbf{Fq}$, with $\mathbf{F} := \mathbf{B}(\mathbf{CB})^{-1}$ (cf. [1]). Clearly $\mathbf{Cz} = \mathbf{0}$. Using (4.2) the underlying ODE reads

$$\dot{\mathbf{z}} = \big((\mathbf{I} - \mathbf{P})\mathbf{A} - \dot{\mathbf{P}}\big)\mathbf{z} + (\mathbf{I} - \mathbf{P})\big(\mathbf{p} - (\mathbf{AF} - \dot{\mathbf{F}})\mathbf{q}\big) = \hat{\mathbf{A}}\mathbf{z} + (\mathbf{I} - \mathbf{P})\mathbf{g}, \tag{4.3}$$

where $\hat{\mathbf{A}}$ and $\mathbf{g}$ are defined as

$$\hat{\mathbf{A}} := (\mathbf{I} - \mathbf{P})\mathbf{A} - \dot{\mathbf{P}}, \tag{4.4a}$$

and

$$\mathbf{g} := \mathbf{p} - (\mathbf{AF} - \dot{\mathbf{F}})\mathbf{q}, \tag{4.4b}$$

respectively. We remark that $\mathbf{z}$ does not depend on the derivative of $\mathbf{q}$. Let the fundamental solution matrix $\mathbf{Z} \in \mathbb{R}^{n\times n}$ of ODE (4.3) be defined by

$$\dot{\mathbf{Z}} = \hat{\mathbf{A}}\mathbf{Z}, \tag{4.5a}$$

$$\mathbf{Z}(t_0) = \mathbf{I}. \tag{4.5b}$$

With (4.5b), the solution of ODE (4.3) can be expressed as

$$\mathbf{z}(t) = \mathbf{Z}(t)\mathbf{z}(t_0) + \int_{t_0}^{t} \mathbf{Z}(t)\mathbf{Z}^{-1}(s)\big(\mathbf{I} - \mathbf{P}(s)\big)\mathbf{g}(s)ds. \tag{4.6}$$

Since $\mathbf{z} = (\mathbf{I} - \mathbf{P})\mathbf{z}$, we thus find

$$\mathbf{z}(t) = \big(\mathbf{I} - \mathbf{P}(t)\big)\mathbf{Z}(t)\big(\mathbf{I} - \mathbf{P}(t_0)\big)\mathbf{z}_{t_0} + \int_{t_0}^{t} \big(\mathbf{I} - \mathbf{P}(t)\big)\mathbf{Z}(t)\mathbf{Z}^{-1}(s)\big(\mathbf{I} - \mathbf{P}(s)\big)\mathbf{g}(s)ds. \quad (4.7)$$

Slight perturbations of the coefficient matrices **A**, **B**, and **C** yield the following ODE for the corresponding perturbed fundamental solution matrix $\tilde{\mathbf{Z}}$

$$\dot{\tilde{\mathbf{Z}}} = \tilde{\hat{\mathbf{A}}}\tilde{\mathbf{Z}} = (\hat{\mathbf{A}} + \delta\hat{\mathbf{A}})\tilde{\mathbf{Z}}, \quad \tilde{\mathbf{Z}}(t_0) = \mathbf{I}, \quad (4.8)$$

where

$$\delta\hat{\mathbf{A}} \doteq (\mathbf{I} - \mathbf{P})\delta\mathbf{A} - \delta\mathbf{P}\mathbf{A} - \frac{\mathrm{d}}{\mathrm{dt}}(\delta\mathbf{P}).$$

In the constant coefficient case this gives in first order approximation

$$\delta\hat{\mathbf{A}} \doteq (\mathbf{I} - \mathbf{P})\delta\mathbf{A} - \delta\mathbf{P}\mathbf{A}.$$

Therefore, the perturbation $\delta\hat{\mathbf{A}}$ can be large if **A** or **P** are large. Hence (cf. (2.6)),

$$\begin{aligned}\tilde{\mathbf{Z}}(t) &= \mathbf{Z}(t) + \int_{t_0}^{t} \mathbf{Z}(t)\mathbf{Z}^{-1}(s)\delta\hat{\mathbf{A}}(s)\tilde{\mathbf{Z}}(s)ds \\ &= \mathbf{Z}(t)\Big(\mathbf{I} + \int_{t_0}^{t} \mathbf{Z}^{-1}(s)\mathbf{G}(s)\tilde{\mathbf{Z}}(s)ds - \int_{t_0}^{t} \mathbf{Z}^{-1}(s)\mathbf{H}(s)\tilde{\mathbf{Z}}(s)ds\Big). \end{aligned} \quad (4.9)$$

Here, the perturbations **G** and **H** are defined by

$$\mathbf{G}(t) := \big(\mathbf{I} - \mathbf{P}(t)\big)\delta\mathbf{A}(t) \quad (4.10a)$$

and

$$\mathbf{H}(t) := \delta\mathbf{P}(t)\mathbf{A}(t), \quad (4.10b)$$

respectively. The perturbation **G** is of concern if the problem is ill-conditioned in the sense that **P** is skew. The perturbation **H** may be of concern even when the problem is well-conditioned. One should compare (4.9) with (4.7). Since $(\mathbf{I} - \mathbf{P})\mathbf{G} = \mathbf{G}$, the influence of the second term on the right hand side of (4.9) can be considered to be an additive perturbation, i.e. a perturbation caused by the source term **g** in (4.7). Hence, the contribution of this perturbation is already controlled by the additive perturbations. For the perturbation **H** this will not hold in general, since $(\mathbf{I} - \mathbf{P})\mathbf{H} \neq \mathbf{H}$. When **H** is large, $\tilde{\mathbf{Z}}$ may differ substantially from **Z**. Since **H** will generally not be in the subspace defined by range$(\mathbf{I} - \mathbf{P})$, the difference between **Z** and $\tilde{\mathbf{Z}}$ may not be controlled as a source term like in (4.7) implying that the perturbed solution may propagate in the wrong subspace. This implies that the perturbed solution $\tilde{\mathbf{z}}$ may differ very much from the unperturbed solution **z**. From the above, we can draw the following conclusions:

- The perturbation **G** is of concern if the problem is ill-conditioned in the sense that **P** is skew. The stability of the solution can change because of **G**. The fundamental solution of the perturbed problem, however, will propagate in the same subspace as the unperturbed solution.
- The perturbation **H** may be of concern even when the problem is well-conditioned and it may also cause a change in the stability behaviour of the solution. However, in this case the perturbed solution might propagate in the wrong directions.

The following theorem shows the possible dramatic influence of perturbations of **B** and **C** on the stability behaviour of the DAE when **P** is large in norm.

Theorem 4.1. Let (1.1) be a constant coefficient index two DAE. Assume that $\mathbf{CB} = \varepsilon\mathbf{I}$, $(\varepsilon \downarrow 0)$. Let $(\mathbf{I}-\mathbf{P})\mathbf{A}$ have bounded negative nontrivial eigenvalues. Perturbations of the matrices **B** and **C**, say $\bar{\mathbf{B}} = \mathbf{B} + \delta\mathbf{B}$ and $\bar{\mathbf{C}} = \mathbf{C} + \delta\mathbf{C}$, may lead to a shift of order δ/ε in the eigenvalues of $(\mathbf{I}-\mathbf{P})\mathbf{A}$, where $\|\delta\mathbf{B}\| = \mathcal{O}(\delta)$ and likewise for $\|\delta\mathbf{C}\|$.

Proof. The perturbations result in

$$\mathbf{P} = \mathbf{P} + \delta\mathbf{P},$$

where

$$\begin{aligned}\delta\mathbf{P} &= (\mathbf{B}+\delta\mathbf{B})\big((\mathbf{C}+\delta\mathbf{C})(\mathbf{B}+\delta\mathbf{B})\big)^{-1}(\mathbf{C}+\delta\mathbf{C}) - \varepsilon^{-1}\mathbf{BC} \\ &= \varepsilon^{-1}(\mathbf{B}+\delta\mathbf{B})(\mathbf{I}+\varepsilon^{-1}\mathbf{C}\delta\mathbf{B}+\varepsilon^{-1}\delta\mathbf{CB}+\varepsilon^{-1}\delta\mathbf{C}\delta\mathbf{B})^{-1}(\mathbf{C}+\delta\mathbf{C}) - \varepsilon^{-1}\mathbf{BC}.\end{aligned}$$

Let $\delta\mathbf{B}$ and $\delta\mathbf{C}$ be small enough so that

$$\max\{\|\varepsilon^{-1}\mathbf{C}\delta\mathbf{B}\|, \|\varepsilon^{-1}\delta\mathbf{CB}\|, \|\varepsilon^{-1}\delta\mathbf{C}\delta\mathbf{B}\|\} = \eta, \quad (\eta \downarrow 0).$$

Then

$$\begin{aligned}\delta\mathbf{P} &= \varepsilon^{-1}(\mathbf{B}+\delta\mathbf{B})\big(\mathbf{I}-\varepsilon^{-1}\mathbf{C}\delta\mathbf{B}-\varepsilon^{-1}\delta\mathbf{CB}+\mathcal{O}(\eta^2)\big)(\mathbf{C}+\delta\mathbf{C}) - \varepsilon^{-1}\mathbf{BC} \\ &= \varepsilon^{-1}\mathbf{B}\delta\mathbf{C}+\varepsilon^{-1}\delta\mathbf{BC}+\mathcal{O}(\eta/\varepsilon).\end{aligned}$$

□

The theorem above implies that perturbations of the coefficients of the order δ may change the stability of the system dramatically when δ is of the same order of magnitude as ε. However, it is not unreasonable to think of much larger perturbations δ in practice, e.g. if one considers a linearization process for a nonlinear DAE. In the next section we shall consider the effects of the perturbations $\delta\mathbf{A}$, $\delta\mathbf{B}$ and $\delta\mathbf{C}$ separately.

5. Effects of specific perturbations for index-2 DAE

In this section we will consider the influence of perturbations of either one of the coefficient matrices **A**, **B**, or **C**, in order to illustrate the analysis of Section 4.
First, we study the effect of perturbations of **A**; so we consider the perturbed DAE

$$\begin{aligned} \dot{\tilde{\mathbf{x}}}(t) &= \tilde{\mathbf{A}}(t)\tilde{\mathbf{x}}(t) + \mathbf{B}(t)\tilde{\mathbf{y}}(t) + \mathbf{p}(t), \\ \mathbf{0} &= \mathbf{C}(t)\tilde{\mathbf{x}}(t) + \mathbf{q}(t), \end{aligned} \tag{5.1}$$

where $\tilde{\mathbf{A}} = \mathbf{A} + \delta\mathbf{A}$. This leads to the following perturbed underlying ODE

$$\begin{aligned} \dot{\tilde{\mathbf{z}}} &= \big((\mathbf{I}-\mathbf{P})\tilde{\mathbf{A}} - \dot{\mathbf{P}}\big)\tilde{\mathbf{z}} + (\mathbf{I}-\mathbf{P})\mathbf{p} - (\mathbf{I}-\mathbf{P})(\tilde{\mathbf{A}}\mathbf{F} - \dot{\mathbf{F}})\mathbf{q} \\ &= \big(\hat{\mathbf{A}} + (\mathbf{I}-\mathbf{P})\delta\mathbf{A}\big)\tilde{\mathbf{z}} + (\mathbf{I}-\mathbf{P})(\mathbf{g} - \delta\mathbf{A}\mathbf{F}\mathbf{q}). \end{aligned} \tag{5.2}$$

Remark 5.1.

- The system matrix of (5.2) has a perturbation $(\mathbf{I}-\mathbf{P})\delta\mathbf{A}$ as compared to the original state ODE (4.3). This perturbation can be large if $(\mathbf{I}-\mathbf{P})$ is skew. This means that the fundamental solution $\tilde{\mathbf{Z}}$ of (5.2) might exhibit a completely wrong growth behaviour in comparison to $\mathbf{Z}$. For example, it may be possible that a large perturbation $(\mathbf{I}-\mathbf{P})\delta\mathbf{A}$ can destroy the stability of the original DAE. In terms of (kinematic) eigenvalues (cf. [2]) this means that the originally (possibly rather small) negative eigenvalues may result in positive ones.
- Regarding the perturbation $(\mathbf{I}-\mathbf{P})\delta\mathbf{A}\tilde{\mathbf{z}}$ of (5.2) as an inhomogeneity of the original ODE (4.3), it can be seen that this inhomogeneity may be very large if $(\mathbf{I}-\mathbf{P})$ is very large. Hence, a slight perturbation of **A** may cause large inhomogeneities which may lead to a large perturbation of **z**, i.e. $\|\mathbf{z}-\tilde{\mathbf{z}}\|$ is large.
- Finally, observe that not only the homogeneous matrix part, but also the inhomogeneous part of the state ODE has changed because of this perturbation of **A**. So, not only the growth behaviour of **z** can be altered dramatically, but also $(\mathbf{I}-\mathbf{P})\tilde{\mathbf{A}}\mathbf{F}\mathbf{q}$ can be very different from $(\mathbf{I}-\mathbf{P})\mathbf{A}\mathbf{F}\mathbf{q}$.

Example 5.2. Consider the following homogeneous DAE

$$\begin{aligned} \dot{\mathbf{x}} &= \begin{bmatrix} 1/2 & 1 \\ 1 & 1/2 \end{bmatrix}\mathbf{x} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} y, \\ 0 &= [1+\varepsilon, -1]\,\mathbf{x}, \end{aligned}$$

where $\varepsilon = 10^{-4}$. Here,

$$\mathbf{P} = \varepsilon^{-1}\begin{bmatrix} 1+\varepsilon & -1 \\ 1+\varepsilon & -1 \end{bmatrix}, \quad \text{and } (\mathbf{I}-\mathbf{P}) = \varepsilon^{-1}\begin{bmatrix} -1 & 1 \\ -(1+\varepsilon) & 1+\varepsilon \end{bmatrix}.$$

Hence $\|\mathbf{P}\| = \mathcal{O}(\varepsilon^{-1})$ $(\varepsilon \to 0)$. The corresponding state ODE is given by

$$\dot{\mathbf{z}} = \frac{1}{2}\varepsilon^{-1}\begin{bmatrix} 1 & -1 \\ 1+\varepsilon & -(1+\varepsilon) \end{bmatrix}\mathbf{z}.$$

The eigenvalues of $(\mathbf{I}-\mathbf{P})\mathbf{A}$ are 0 and $-\frac{1}{2}$. Let $\delta\mathbf{A} = \mathrm{diag}[\delta a, 0]$. Then the nonzero eigenvalue of $(\mathbf{I}-\mathbf{P})\tilde{\mathbf{A}} = -(\frac{1}{2} + \varepsilon^{-1}\delta a)$, which e.g. equals 9.5 when $\delta a = -10^{-3}$, or, which still equals $1/2$ when $\delta a = -10^{-4}$. In fact, the perturbed DAE is unstable when $\delta a < -\frac{1}{2}\varepsilon$. In this case, the growth behaviour of the state variables is altered in a dramatic way by this slight perturbation of the matrix $\mathbf{A}$. Hence, a slight perturbation of coefficient matrix $\mathbf{A}$ can alter the solution behaviour dramatically. Further, note that the perturbation of the coefficient matrix of the state ODE equals $\delta\hat{\mathbf{A}} = (\mathbf{I}-\mathbf{P})\delta\mathbf{A}$. This implies that this perturbation is a perturbation of a rather harmless kind, as is already explained in (4.9).

□

Next, we consider the influence of perturbations in the matrix coefficient $\mathbf{B}$ on the solution of the DAE, i.e.

$$\begin{aligned} \dot{\tilde{\mathbf{x}}}(t) &= \mathbf{A}(t)\tilde{\mathbf{x}}(t) + \tilde{\mathbf{B}}(t)\tilde{\mathbf{y}}(t) + \mathbf{p}(t), \\ \mathbf{0} &= \mathbf{C}(t)\tilde{\mathbf{x}}(t) + \mathbf{q}(t), \end{aligned} \tag{5.3}$$

where $\tilde{\mathbf{B}} = \mathbf{B} + \delta\mathbf{B}$. The corresponding perturbed underlying ODE can be written as

$$\begin{aligned} \dot{\tilde{\mathbf{z}}} =& \big((\mathbf{I}-\tilde{\mathbf{P}})\mathbf{A} - \dot{\tilde{\mathbf{P}}}\big)\tilde{\mathbf{z}} + (\mathbf{I}-\tilde{\mathbf{P}})\tilde{\mathbf{g}} \\ \doteq& [\hat{\mathbf{A}} - (\mathbf{I}-\mathbf{P})\delta\mathbf{B}(\mathbf{CB})^{-1}\mathbf{CA} - ((\mathbf{I}-\mathbf{P})\delta\mathbf{B}(\mathbf{CB})^{-1}\mathbf{C})^{\dot{}}]\tilde{\mathbf{z}} \\ &+ (\mathbf{I}-\mathbf{P})\big(\mathbf{I} - \delta\mathbf{B}(\mathbf{CB})^{-1}\mathbf{C}\big)\mathbf{g} \\ &- (\mathbf{I}-\mathbf{P})\Big(\mathbf{A}(\mathbf{I}-\mathbf{P})\delta\mathbf{B}(\mathbf{CB})^{-1} - \big((\mathbf{I}-\mathbf{P})\delta\mathbf{B}(\mathbf{CB})^{-1}\big)^{\dot{}}\Big)\mathbf{q}, \end{aligned} \tag{5.4}$$

because $\tilde{\mathbf{P}} \doteq \mathbf{P} + (\mathbf{I}-\mathbf{P})\delta\mathbf{B}(\mathbf{CB})^{-1}\mathbf{C}$.

Remark 5.3.

- As before, the resulting perturbation of the homogeneous part of the state ODE (5.2) might be large if $(\mathbf{I}-\mathbf{P})$ is large. This implies that the growth behaviour of the perturbed problem might be completely different from the growth behaviour of the original problem.

- The fundamental solutions $\mathbf{Z}$ and $\tilde{\mathbf{Z}}$ and the Green functions $\mathbf{G}$ and $\tilde{\mathbf{G}}$ might differ dramatically if $\|\mathbf{P}\|$ is large (cf. 2).
- The perturbations of the imhomogeneous part of the state ODE can be very large if $\|\mathbf{P}\|$ is large, which means that $\|\mathbf{z}-\tilde{\mathbf{z}}\|$ can be very large.

As a consequence, the perturbed solution $\tilde{\mathbf{x}}$ satisfies

$$\begin{aligned}\tilde{\mathbf{x}} &= \tilde{\mathbf{z}} + \tilde{\mathbf{P}}\tilde{\mathbf{x}} \\ &= \tilde{\mathbf{z}} - \mathbf{F}\mathbf{q} - (\mathbf{I}-\mathbf{P})\delta\mathbf{B}(\mathbf{C}\mathbf{B})^{-1}\mathbf{q}. \end{aligned} \tag{5.5}$$

Hence, due to the arguments above $\mathbf{x}$ and $\tilde{\mathbf{x}}$ might differ dramatically.

Example 5.4. Again, consider Example 5.2. Now, the coefficient matrix $\mathbf{B}$ is perturbed slightly, say

$\tilde{\mathbf{B}} := \begin{bmatrix} 1 \\ 1+\delta \end{bmatrix}$. Then

$$(\mathbf{I}-\tilde{\mathbf{P}}) = \mathbf{I} - \tilde{\mathbf{B}}(\mathbf{C}\tilde{\mathbf{B}})^{-1}\mathbf{C} = \frac{1}{\varepsilon-\delta}\begin{bmatrix} -(1+\delta) & 1 \\ -(1+\varepsilon)(1+\delta) & 1+\varepsilon \end{bmatrix}$$

$$\tilde{\tilde{\mathbf{A}}} = (\mathbf{I}-\tilde{\mathbf{P}})\mathbf{A} = \frac{1}{2(\delta-\varepsilon)}\begin{bmatrix} \delta-1 & 1+2\delta \\ (1+\varepsilon)(\delta-1) & (1+\varepsilon)(1+2\delta) \end{bmatrix},$$

with nonzero eigenvalue $\frac{3\delta+\varepsilon+2\varepsilon\delta}{2(\delta-\varepsilon)}$. For $\delta = -5_{10^{-5}}$ this nonzero eigenvalue of the perturbed system is equal to $1.667_{10^{-1}}$, which implies that the originally stable system is perturbed into an unstable one by a small perturbation of the matrix $\mathbf{B}$. Note further, that $(\mathbf{I}-\mathbf{P})\tilde{\tilde{\mathbf{A}}} = \tilde{\tilde{\mathbf{A}}}$. This means that we deal with a perturbation of the state ODE which is of the same kind as the second term of (4.9) and from (4.9) we know that this perturbation may be controlled already. □

Finally, we study the effect of perturbations of the matrix coefficient $\mathbf{C}$ on the solution of the DAE system, i.e. we consider the perturbed problem

$$\begin{aligned}\dot{\tilde{\mathbf{x}}}(t) &= \mathbf{A}(t)\tilde{\mathbf{x}}(t) + \mathbf{B}(t)\tilde{\mathbf{y}}(t) + \mathbf{p}(t), \\ \mathbf{0} &= \tilde{\mathbf{C}}(t)\tilde{\mathbf{x}}(t) + \mathbf{q}(t), \end{aligned} \tag{5.6}$$

where $\tilde{\mathbf{C}} = \mathbf{C} + \boldsymbol{\delta}\mathbf{C}$. In this case one finds for the perturbed projector $\tilde{\mathbf{P}}$

$$\tilde{\mathbf{P}} \doteq \mathbf{P} + \mathbf{F}\delta\mathbf{C}(\mathbf{I} - \mathbf{P}). \tag{5.7}$$

Then, the corresponding perturbed state ODE can be written as

$$\begin{aligned} \dot{\tilde{\mathbf{z}}} =& \big((\mathbf{I} - \tilde{\mathbf{P}})\mathbf{A} - \dot{\tilde{\mathbf{P}}}\big)\tilde{\mathbf{z}} + (\mathbf{I} - \tilde{\mathbf{P}})\tilde{\mathbf{g}} \\ =& [\hat{\mathbf{A}} - \mathbf{F}\delta\mathbf{C}(\mathbf{I} - \mathbf{P})\mathbf{A} - \big(\mathbf{F}\delta\mathbf{C}(\mathbf{I} - \mathbf{P})\dot{\big)}]\tilde{\mathbf{z}} \\ &+ (\mathbf{I} - \mathbf{F}\delta\mathbf{C})(\mathbf{I} - \mathbf{P})\mathbf{g} + (\mathbf{I} - \mathbf{P})\big(\mathbf{A}\mathbf{F}\delta\mathbf{C}\mathbf{F} - (\mathbf{F}\delta\mathbf{C}\mathbf{F}\dot{)}\big)\mathbf{q}. \end{aligned} \tag{5.8}$$

In this case, the same remarks hold as before. Now, we find the following expression for $\tilde{\mathbf{x}}$

$$\begin{aligned} \tilde{\mathbf{x}} &= \tilde{\mathbf{z}} + \tilde{\mathbf{P}}\tilde{\mathbf{x}} \\ &= \tilde{\mathbf{z}} - \mathbf{F}\mathbf{q} - \mathbf{F}\delta\mathbf{C}\mathbf{F}\mathbf{q} \end{aligned} \tag{5.9}$$

and again, the error $\|\mathbf{x} - \tilde{\mathbf{x}}\|$ might become very large. Next

Example 5.5. Consider the DAE

$$\begin{aligned} \dot{\mathbf{x}} &= \begin{bmatrix} \lambda & 0 \\ 0 & \mu \end{bmatrix} \mathbf{x} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} y + \mathbf{p}, \\ 0 &= [1,\, 1]\, \mathbf{x} + q. \end{aligned}$$

The corresponding state ODE is

$$\dot{\mathbf{z}} = \frac{1}{2}\begin{bmatrix} \lambda & -\mu \\ -\lambda & \mu \end{bmatrix}\mathbf{z} + \frac{1}{2}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}\mathbf{p} + \frac{1}{4}(\mu - \lambda)\begin{bmatrix} 1 \\ -1 \end{bmatrix}\mathbf{q}.$$

The eigenvalues of $(\mathbf{I} - \mathbf{P})\mathbf{A}$ are 0 and $\frac{1}{2}(\lambda + \mu)$, respectively. Let $\boldsymbol{\delta}\mathbf{C} = [\delta\varepsilon \;\, 0]$, then

$$(\mathbf{I} - \mathbf{P} - \delta\mathbf{P})\mathbf{A} = \frac{1}{2+\varepsilon}\begin{bmatrix} \lambda & -\mu \\ -\lambda(1+\varepsilon) & \mu(1+\varepsilon) \end{bmatrix},$$

with eigenvalues 0 and $\frac{1}{2+\varepsilon}\big(\lambda + \mu(1+\varepsilon)\big)$, respectively. A typical stable situation (for $\varepsilon = 0$) is $\lambda \approx -\mu$. The perturbed problem has nontrivial eigenvalue $\frac{\lambda\varepsilon}{2+\varepsilon}$. So, we may expect difficulties if $|\lambda| \gg 1$. In particular if $\lambda \gg 1$ the DAE, a slight perturbation of the stable problem, has resulted in a very unstable one. □

In the following important example we consider a nearly index three DAE (cf. [4,6]).

Example 5.6. Consider the homogeneous DAE

$$\dot{\mathbf{x}} = \begin{bmatrix} 1/2 & 2 \\ 1 & 1 \end{bmatrix} \mathbf{x} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} y,$$
$$0 = [1+\varepsilon, -1]\,\mathbf{x},$$

where $\varepsilon = 10^{-4}$. The corresponding state ODE is given by

$$\dot{\mathbf{z}} = (1 + \frac{1}{2}\varepsilon^{-1})\varepsilon^{-1} \begin{bmatrix} 1 & -1 \\ 1+\varepsilon & -(1+\varepsilon) \end{bmatrix} \mathbf{z}.$$

The eigenvalues of $(\mathbf{I}-\mathbf{P})\mathbf{A}$ are 0 and $-(1+\frac{1}{2}\varepsilon^{-1}) \approx -5_{10^3}$, respectively. Hence, the system is extremely stable for $\varepsilon \downarrow 0$. Let $\boldsymbol{\delta}\mathbf{C} = [0\ \delta]$. Then

$$\tilde{\tilde{\mathbf{A}}} = (\mathbf{I}-\tilde{\mathbf{P}})\mathbf{A} = \frac{1}{2(\varepsilon+\delta)} \begin{bmatrix} 1-\delta & 2(\delta-1) \\ (1+\varepsilon) & -2(1+\varepsilon) \end{bmatrix},$$

with eigenvalues 0 and $-\frac{\frac{1}{2}+\varepsilon+\frac{1}{2}\delta}{\varepsilon+\delta}$. For $\delta = -2_{10^{-4}}$, the nonzero eigenvalue of the perturbed coefficient matrix $\tilde{\tilde{\mathbf{A}}}$ is equal to 5_{10^3}. This means that a DAE which was extremely stable is perturbed into an extremely unstable DAE by a slight perturbation of the coefficient matrix $\mathbf{C}$. Here, the direction of the mode corresponding to the nonzero eigenvalue is perturbed into $[1-\delta, 1+\varepsilon]^{\mathrm{T}}$. Hence, the nontrivial solution modes are altered in a dramatic way, viz. the growth of this mode has been changed severely, while simultaneously the direction of this mode is altered. Hence, perturbations of type $\mathbf{H}$ may lead to dramatic results. Note that the matrix pencil $(\tilde{\mathbf{A}}, \tilde{\mathbf{B}})$ of this DAE is regular for $\varepsilon = 0$, implying that the stability constants of this system are bounded (cf. [4, 6]). □

References

[1] U. ASCHER AND L. R. PETZOLD. Projected implicit Runge-Kutta for differential algebraic equations. *SIAM J. Numer. Anal. 28* (1991), 1097–1120.

[2] U. M. ASCHER, R. M. M. MATTHEIJ, AND R. D. RUSSELL. *Numerical Solution of Boundary Value Problems for Ordinary Differential Equations.* SIAM, Philadelphia, 1995.

[3] G. H. GOLUB AND C. F. V. LOAN. *Matrix Computations.* John Hopkins University Press, Baltimore, 1990.

[4] R. R. M. MATTHEIJ AND P. M. E. J. WIJCKMANS. Conditioning of two deck differential algebraic equations. *(submitted).*

[5] G. SÖDERLIND. Remarks on the stability of high-index DAEs with respect to parametric perturbations. *Computing 49* (1992), 303–314.

[6] P. M. E. J. WIJCKMANS. *Conditioning of Differential Algebraic Equations and Numerical Solution of Multibody Dynamics.* PhD thesis, Eindhoven University of Technology, Eindhoven, 1996.

PREVIOUS PUBLICATIONS IN THIS SERIES:

Number	Author(s)	Title	Month
97-13	He Yinnian R.M.M. Mattheij	Optimum Mixed Finite Element Nonlinear Galerkin Method for the Navier-Stokes Equations; I: Error Estimates for Spatial Discretization	September '97
97-14	He Yinnian	Optimum Mixed Finite Element Nonlinear Galerkin Method for the Navier-Stokes Equations; II: Stability Analysis for Time Discretization	September '97
97-15	He Yinnian	Optimum Mixed Finite Element Nonlinear Galerkin Method for the Navier-Stokes Equations; III: Convergence Analysis for Time Discretization	September '97
97-16	A.F.M. ter Elst C.M.P.A. Smulders	Reduced heat kernels on homogeneous spaces	September '97
97-17	H.J.C. Huijberts	Minimal order linear model matching for nonlinear control systems	September '97
97-18	L.C.G.J.M. Habets	System equivalence for AR-systems over rings	November '97
97-19	R.M.M. Mattheij P.M.E.J. Wijckmans	Conditioning of Two Deck Differential Algebraic Equations	November '97
97-20	L.P.H. de Goey J.H.M. ten Thije Boonkkamp	A Comparison of the Integral Analysis and Large Activation Energy Asymptotics for Studying Flame Stretch	November '97
97-21	R.M.M. Mattheij P.M.E.J. Wijckmans	Sensitivity of solutions of linear DAE to perturbations of the system matrices	December '97