TECHNISCHE UNIVERSITÄT BERLIN

Self-Adjoint Differential-Algebraic Equations

Peter Kunkel Volker Mehrmann Lena Scholz

Preprint 2011/13

Preprint-Reihe des Instituts für Mathematik
Technische Universität Berlin
http://www.math.tu-berlin.de/preprints

Preprint 2011/13 Juli 2011

Self-Adjoint Differential-Algebraic Equations *

Peter Kunkel† Volker Mehrmann‡ Lena Scholz‡

July 7, 2011

Abstract

Motivated from linear-quadratic optimal control problems for differential-algebraic equations (DAEs), we study the functional analytic properties of the operator associated with the necessary optimality boundary value problem and show that it is associated with a self-conjugate operator and a self-adjoint pair of matrix functions. We then study general self-adjoint pairs of matrix valued functions and derive condensed forms under orthogonal congruence transformations that preserve the self-adjointness. We analyze the relationship between self-adjoint DAEs and Hamiltonian systems with symplectic flows. We also show how to extract self-adjoint and Hamiltonian reduced systems from derivative arrays.

Keywords: Differential-algebraic equation, self-conjugate operator, self-adjoint pair, optimal control, necessary optimality condition, strangeness index, condensed form, congruence transformation, Hamiltonian system, symplectic flow.

AMS(MOS) subject classification: 93C10, 93C15, 93B52, 65L80, 49K15, 34H05.

1 Introduction

In this paper we study a class of structured systems of differential-algebraic equations (DAEs). The main motivation arises from the *linear-quadratic optimal control problem* of minimizing a cost functional

$$\mathcal{J}(x,u) = \frac{1}{2}x(\bar{t})^T M_e x(\bar{t}) + \frac{1}{2}\int_{\underline{t}}^{\bar{t}} \left(x^T W x + x^T S u + u^T S^T x + u^T R u\right)\, dt, \tag{1.1}$$

subject to the constraint

$$E\dot{x} = Ax + Bu + f, \quad x(\underline{t}) = \underline{x} \in \mathbb{R}^n, \tag{1.2}$$

with $E, A \in C^0(\mathbb{I}, \mathbb{R}^{n,n})$, $W \in C^0(\mathbb{I}, \mathbb{R}^{n,n})$, $B \in C^0(\mathbb{I}, \mathbb{R}^{n,m})$, $S \in C^0(\mathbb{I}, \mathbb{R}^{n,m})$, $R \in C^0(\mathbb{I}, \mathbb{R}^{m,m})$, $f \in C^0(\mathbb{I}, \mathbb{R}^n)$ and $M_e \in \mathbb{R}^{n,n}$, where $R = R^T$, $W = W^T$ and $M_e = M_e^T$.

*Partially supported by the Research In Pairs program of *Mathematisches Forschungsinstitut Oberwolfach*, whose hospitality is gratefully acknowledged.

†Mathematisches Institut, Universität Leipzig, Johannisgasse 26, D-04009 Leipzig, Fed. Rep. Germany, kunkel@math.uni-leipzig.de. Partially supported by the *Deutsche Forschungsgemeinschaft* through Project KU964/7-1.

‡Institut für Mathematik, TU Berlin, Str. des 17. Juni 136, D-10623 Berlin, Fed. Rep. Germany, {mehrmann,lscholz}@math.tu-berlin.de. Partially supported by the *Deutsche Forschungsgemeinschaft* through the DFG Research Center MATHEON *Mathematics for key technologies* in Berlin.

Furthermore, $\mathbb{I} = [\underline{t}, \overline{t}]$ is a real time-interval and $C^\ell(\mathbb{I}, \mathbb{R}^{n,m})$ denotes the ℓ-times continuously differentiable functions from the interval $\mathbb{I}$ to the real $n \times m$ matrices. Note that for simplicity we omit the argument t in all matrix and vector valued functions.

Typically in applications, the matrix function

$$\begin{bmatrix} W & S \\ S^T & R \end{bmatrix}$$

and the weight matrix M_e for the final state are pointwise positive semidefinite, but problems where these are indefinite also arise in applications from robust control [4, 24].

If the differential-algebraic equation (1.2) has some further properties, (i. e., if it is strangeness-free as a behavior system and if the coefficients are sufficiently smooth), then it has been shown in [20] that the necessary optimality condition is given by the boundary value problem

$$\begin{bmatrix} 0 & E & 0 \\ -E^T & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \frac{d}{dt} \begin{bmatrix} \lambda \\ x \\ u \end{bmatrix} = \begin{bmatrix} 0 & A & B \\ A^T + \frac{d}{dt}E^T & W & S \\ B^T & S^T & R \end{bmatrix} \begin{bmatrix} \lambda \\ x \\ u \end{bmatrix} + \begin{bmatrix} f \\ 0 \\ 0 \end{bmatrix}, \tag{1.3}$$

with boundary conditions $x(\underline{t}) = \underline{x}$, $E(\overline{t})^T\lambda(\overline{t}) - M_e x(\overline{t}) = 0$. Note that compared to [20] here λ is replaced by $-\lambda$. Since (1.3) is again a differential-algebraic equation, these boundary conditions may not be consistent, which means that there may be restrictions to the value $\underline{x}$ and the weighting matrix M_e that need to be satisfied to guarantee the existence of solutions [20].

If we denote the associated differential-algebraic equation (1.3) as $\mathcal{E}\dot{z} = \mathcal{A}z + \tilde{f}$, then it is an easy calculation to show that the pair $(\mathcal{E}, \mathcal{A})$ of matrix functions has the property that $\mathcal{E}^T = -\mathcal{E}$ and $\mathcal{A}^T = \mathcal{A} + \dot{\mathcal{E}}$. We call pairs of matrix functions with this property *self-adjoint pairs*, since, as we will show below, this is a property that is associated with a linear *self-conjugate* differential-algebraic operator.

Formal adjoint equations (or dual systems) and their role for the solvability of optimal control problems have also been considered in [21], and observability as well as controllability of linear descriptor systems has been previously studied in [9]. Furthermore, self-adjoint differential-algebraic systems and the underlying Hamiltonian subsystem have been studied in [3].

In this paper we will first introduce some preliminary results in Section 2, and then in Section 3 discuss self-conjugate differential-algebraic operators arising in optimal control in an abstract setting. In Section 4 we analyze the structure of the resulting self-adjoint pairs of matrix functions and the associated boundary value problems via condensed forms under congruence transformations using certain constant rank assumptions. Based on these condensed forms we can characterize the consistency of boundary values, as well as the consistency and smoothness requirements for the inhomogeneities, and thus derive altogether the conditions for unique solvability of the system. In Section 5 we then show that the underlying ordinary differential equation of the differential-algebraic equation associated with a self-adjoint pair of matrix functions is a Hamiltonian system and generates a symplectic flow. A global condensed form for self-adjoint DAEs with symplectic flow is derived in Section 6. In Section 7 we then discuss the structure preserving construction of the symplectic flow from derivative arrays. Finally, in Section 8 we show that these results also hold locally for nonlinear optimal control problems and close with some conclusions in Section 9.

2 Preliminaries

In order to treat general linear DAEs and the constraint equation (1.2) in the same framework we introduce the so-called behavior formulation (see [28]) by setting

$$\mathcal{E} = \begin{bmatrix} E & 0 \end{bmatrix}, \quad \mathcal{A} = \begin{bmatrix} A & B \end{bmatrix}, \quad z = \begin{bmatrix} x \\ u \end{bmatrix}, \tag{2.1}$$

such that equations (1.2) can be written as

$$\mathcal{E}\dot{z} = \mathcal{A}z + f. \tag{2.2}$$

with sufficiently smooth $\mathcal{E}, \mathcal{A} \in C^0(\mathbb{I}, \mathbb{R}^{n,n+m})$ and $f \in C^0(\mathbb{I}, \mathbb{R}^n)$. It is well known [5, 11, 19] that the solution of the general differential-algebraic equation (2.2) may depend on derivatives of the coefficient functions $\mathcal{E}, \mathcal{A}$ and the inhomogeneity f.

Since it is generally difficult or even impossible to differentiate data that are numerically computed, an idea due to [8] is to differentiate (2.2) and consider the equation together with its derivatives. In this way, we get so-called *derivative arrays*

$$M_\ell \dot{z}_\ell = N_\ell z_\ell + g_\ell, \tag{2.3}$$

where the coefficient functions form an inflated pair of block matrix functions

$$\begin{aligned} (M_\ell)_{i,j} &= \tbinom{i}{j}\mathcal{E}^{(i-j)} - \tbinom{i}{j+1}\mathcal{A}^{(i-j-1)}, \ i,j = 0,\ldots,\ell, \\ (N_\ell)_{i,j} &= \begin{cases} \mathcal{A}^{(i)} & \text{for } i = 0,\ldots,\ell, \ j = 0, \\ 0 & \text{otherwise}, \end{cases} \\ (z_\ell)_j &= z^{(j)}, \ j = 0,\ldots,\ell, \\ (g_\ell)_i &= f^{(i)}, \ i = 0,\ldots,\ell. \end{aligned} \tag{2.4}$$

Here we have used the convention that $\binom{i}{j} = 0$ for $i < 0$, $j < 0$ or $j > i$.

It is then known, [17, 19], that the following hypothesis is sufficient to characterize the solution behavior.

Hypothesis 2.1 *Consider the system of differential-algebraic equations (2.3). There exist integers μ, a, d, v such that the following properties hold.*

1. *For all $t \in \mathbb{I}$ we have* $\operatorname{rank} M_\mu(t) = (\mu+1)n - a - v$. *This implies the existence of a smooth matrix function Z with orthonormal columns and size $((\mu+1)n, a+v)$ satisfying $Z^T M_\mu = 0$.*

2. *For all $t \in \mathbb{I}$ we have* $\operatorname{rank} Z(t)^T N_\mu(t)[I_{n+m}\, 0 \cdots 0]^T = a$ *and without loss of generality Z can be partitioned as $[Z_2, Z_3]$, with Z_2 of size $((\mu+1)n, a)$ and Z_3 of size $((\mu+1)n, v)$, such that $\hat{\mathcal{A}}_2 = Z_2^T N_\mu [I_{n+m}\, 0 \cdots 0]^T$ has full row rank a and $Z_3^T N_\mu [I_{n+m}\, 0 \cdots 0]^T = 0$. Furthermore, there exists a smooth matrix function T_2 with orthonormal columns and size $(n+m, d)$, $d = n+m-a$ satisfying $\hat{\mathcal{A}}_2 T_2 = 0$.*

3. *For all $t \in \mathbb{I}$ we have that* $\operatorname{rank} \mathcal{E}(t)T_2(t) = d$. *This implies the existence of a smooth matrix function Z_1 with orthonormal columns and size (n, d) so that $\hat{\mathcal{E}}_1 = Z_1^T \mathcal{E}$ has constant rank d.*

If the hypothesis holds, then system (2.2) has the same solution set as the so-called reduced system

$$\begin{bmatrix} \hat{\mathcal{E}}_1 \\ 0 \\ 0 \end{bmatrix} \dot{z} = \begin{bmatrix} \hat{\mathcal{A}}_1 \\ \hat{\mathcal{A}}_2 \\ 0 \end{bmatrix} z + \begin{bmatrix} \hat{f}_1 \\ \hat{f}_2 \\ \hat{f}_3 \end{bmatrix}, \tag{2.5}$$

where $\hat{\mathcal{E}}_1 = Z_1^T \mathcal{E}$, $\hat{\mathcal{A}}_1 = Z_1^T \mathcal{A}$, $\hat{\mathcal{A}}_2 = Z_2^T N_\mu [I_{n+m}\, 0 \cdots 0]^T$, $\hat{f}_1 = Z_1^T f$, and $\hat{f}_i = Z_i^T g_\mu$ for $i = 2, 3$. The block rows have dimensions d, a and v, respectively.

If $v > 0$ and $\hat{f}_3 \neq 0$ then the system has no solution and if $v = 0$ and $m = 0$ (i. e., there are as many equations as unknowns), then every consistent initial condition fixes a unique solution. In the latter case we call the system *regular*. If the system is regular, then from (2.5) we see that an initial condition $z(\underline{t}) = \underline{z}$ for (2.2) is consistent if and only if

$$\hat{\mathcal{A}}_2(\underline{t})\underline{z} + \hat{f}_2(\underline{t}) = 0$$

holds or the second block is void.

The quantity μ in Hypothesis 2.1 is called the *strangeness index* of the DAE system and it is well known [19] that a system with $m = 0$ that satisfies Hypothesis 2.1 with $v = 0$ has a well-defined *differentiation index*, [5]. The differentiation index is commonly used to classify regular DAEs, except for the case of ordinary differential equations it is one less than the strangeness index. Note that the reduced system (2.5) is strangeness-free in the sense that it satisfies Hypothesis 2.1 with $\mu = 0$.

For the DAE (1.2) of the optimal control problem, we require that it satisfies Hypothesis 2.1 with $v = 0$ such that the corresponding reduced system is given by

$$\hat{E}\dot{x} = \hat{A}x + \hat{B}u + \hat{f}, \tag{2.6}$$

where

$$\begin{aligned} \hat{E}_1 = Z_1^T E, \quad \hat{A}_1 = Z_1^T A, \qquad & \hat{B}_1 = Z_1^T B, \qquad & \hat{f}_1 = Z_1^T f, \\ \hat{A}_2 = Z_2^T N_\mu V \begin{bmatrix} I_n \\ 0 \end{bmatrix}, \quad & \hat{B}_2 = Z_2^T N_\mu V \begin{bmatrix} 0 \\ I_m \end{bmatrix}, \quad & \hat{f}_2 = Z_2^T g_\mu, \end{aligned}$$

with $V = \begin{bmatrix} I_{n+m} & 0 & \dots & 0 \end{bmatrix}^T$. Due to construction it satisfies the condition that

$$\begin{bmatrix} \hat{E}_1 & 0 \\ \hat{A}_2 & \hat{B}_2 \end{bmatrix}$$

has (pointwise) full row rank.

3 Self-conjugate differential-algebraic operators

In this section, we present an abstract setting that allows us to interpret the operator behind the boundary value problem as a self-conjugate Banach space operator. We refer to [14] for the general functional framework and the proofs of the following results.

The most general definition of a conjugate operator appears in the context of bilinear systems.

Definition 3.1 *A pair $\langle \mathbb{X}, \mathbb{X}^* \rangle$ of (real) vector spaces equipped with a bilinear form $\langle \cdot, \cdot \rangle$ is called a* bilinear system.

Definition 3.2 *Let $\langle \mathbb{X}, \mathbb{X}^* \rangle$ and $\langle \mathbb{Y}, \mathbb{Y}^* \rangle$ be two bilinear systems and let $D : \mathbb{X} \to \mathbb{Y}$ be a homomorphism. A homomorphism $D^* : \mathbb{Y}^* \to \mathbb{X}^*$ is called* conjugate to D *if and only if*

$$\langle Dx, y^* \rangle = \langle x, D^* y^* \rangle \text{ for all } x \in \mathbb{X},\ y^* \in \mathbb{Y}^*. \tag{3.1}$$

In general, we cannot guarantee that a given homomorphism possesses a conjugate nor that it is unique, if it exists. In order to have at least uniqueness for the conjugate, we need bilinear systems with stronger properties.

Definition 3.3 *A bilinear system $\langle \mathbb{X}, \mathbb{X}^* \rangle$ is called a* dual system *if and only if the bilinear form satisfies*

$$\begin{array}{lll} \langle x, x^* \rangle = 0 \ \text{ for all } x \in \mathbb{X} & \Longleftrightarrow & x^* = 0, \\ \langle x, x^* \rangle = 0 \ \text{ for all } x^* \in \mathbb{X}^* & \Longleftrightarrow & x = 0. \end{array} \tag{3.2}$$

Theorem 3.4 *Let $\langle \mathbb{X}, \mathbb{X}^* \rangle$ and $\langle \mathbb{Y}, \mathbb{Y}^* \rangle$ be two bilinear systems and let $D : \mathbb{X} \to \mathbb{Y}$ be a homomorphism. If $\langle \mathbb{X}, \mathbb{X}^* \rangle$ is a dual system, then D possesses at most one conjugate.*

Since we mainly deal with Banach spaces of continuous functions, the main tool to show that a given bilinear system is a dual system is given by the following well-known result called du Bois-Reymond Lemma, see, e. g., [12, Lemma 3.2], where $C_0^\infty(\mathbb{I}, \mathbb{R}^n)$ denotes the set of functions in $C^\infty(\mathbb{I}, \mathbb{R}^n)$ with compact support.

Theorem 3.5 *Let $f \in C(\mathbb{I}, \mathbb{R}^n)$ with*

$$\langle f, g \rangle = \int_{\mathbb{I}} f^T g \, dt = 0 \quad \text{for all } g \in C_0^\infty(\mathbb{I}, \mathbb{R}^n). \tag{3.3}$$

Then $f \equiv 0$.

With these preparations, following [20], we now write the optimal control problem consisting of (1.1) and (2.6), omitting hats for simplicity, as

$$\frac{1}{2}\mathcal{Q}(z,z) = \min! \quad \text{s. t.} \quad \mathcal{L}(z) = c, \quad z = \begin{bmatrix} x \\ u \end{bmatrix}, \quad c = \begin{bmatrix} f \\ E(\underline{t})^+ E(\underline{t})\underline{x} \end{bmatrix}, \tag{3.4}$$

where $\mathcal{Q} : \mathbb{Z} \times \mathbb{Z} \to \mathbb{R}$ is a (symmetric) quadratic form and $\mathcal{L} : \mathbb{Z} \to \mathbb{Y}$ is a *linear submersion* (i. e., it is Fréchet differentiable with a surjective Fréchet derivative that has a kernel that is continuously projectable), defined by

$$\begin{array}{l} \mathcal{Q}(v,z) = v(\bar{t})^T \begin{bmatrix} M_e & 0 \\ 0 & 0 \end{bmatrix} z(\bar{t}) + \displaystyle\int_{\mathbb{I}} v^T \begin{bmatrix} W & S \\ S^T & R \end{bmatrix} z \, dt, \\ \mathcal{L}(z) = (E\frac{d}{dt}(E^+Ex) - (A + E\frac{d}{dt}(E^+E))x - Bu,\ E(\underline{t})^+E(\underline{t})x(\underline{t})) \end{array} \tag{3.5}$$

with the Banach spaces $\mathbb{Z} = \mathbb{X} \times \mathbb{U}$ and

$$\begin{array}{l} \mathbb{X} = C^1_{E^+E}(\mathbb{I}, \mathbb{R}^n) = \{x \in C(\mathbb{I}, \mathbb{R}^n),\ E^+Ex \in C^1(\mathbb{I}, \mathbb{R}^n)\}, \quad \mathbb{U} = C(\mathbb{I}, \mathbb{R}^m), \\ \mathbb{Y} = C(\mathbb{I}, \mathbb{R}^n) \times \operatorname{range} E(\underline{t})^T. \end{array} \tag{3.6}$$

It should be noted that in contrast to usual convention $\mathbb{Z}$ is not the set of integers. Here E^+ denotes the Moore-Penrose pseudo-inverse of the matrix function E, see [20] for details on

the representation of the DAE operator and the choice of the spaces. In view of the results in [20], we define bilinear systems $\langle \mathbb{Z}, \mathbb{Z}^* \rangle$ and $\langle \mathbb{Y}, \mathbb{Y}^* \rangle$ by introducing the Banach spaces

$$\begin{aligned} \mathbb{Z}^* &= C(\mathbb{I}, \mathbb{R}^n) \times C(\mathbb{I}, \mathbb{R}^m) \times \operatorname{range} E(\underline{t})^T \times \operatorname{range} E(\bar{t})^T, \\ \mathbb{Y}^* &= C^1_{EE^+}(\mathbb{I}, \mathbb{R}^n) \times \operatorname{range} E(\underline{t})^T \end{aligned} \tag{3.7}$$

and corresponding bilinear forms

$$\begin{aligned} \langle z, (\eta, \vartheta, \delta, \varepsilon) \rangle &= \int_{\mathbb{I}} (\eta^T x + \vartheta^T u)\, dt + \delta^T x(\underline{t}) + \varepsilon^T x(\bar{t}), \\ \langle (g, r), (\lambda, \gamma) \rangle &= \int_{\mathbb{I}} \lambda^T g\, dt + \gamma^T r. \end{aligned} \tag{3.8}$$

Theorem 3.6 *The bilinear systems $\langle \mathbb{Z}, \mathbb{Z}^* \rangle$ and $\langle \mathbb{Y}, \mathbb{Y}^* \rangle$ are dual systems.*

Proof. Consider the bilinear system $\langle \mathbb{Y}, \mathbb{Y}^* \rangle$ with its bilinear form given in (3.8).
Let $y^* = (\lambda, \gamma) \in \mathbb{Y}^*$ be fixed and assume that $\langle y, y^* \rangle = 0$ for all $y \in \mathbb{Y}$, i. e.,

$$\int_{\mathbb{I}} \lambda^T g\, dt + \gamma^T r = 0 \text{ for all } (g, r) \in \mathbb{Y}.$$

Choosing $(g, r) = (0, \gamma)$ gives $\gamma^T \gamma = 0$, hence $\gamma = 0$. Therefore,

$$\int_{\mathbb{I}} \lambda^T g\, dt = 0 \text{ for all } g \in C_0^\infty(\mathbb{I}, \mathbb{R}^n) \subseteq C(\mathbb{I}, \mathbb{R}^n),$$

where $\lambda \in C^1_{EE^+}(\mathbb{I}, \mathbb{R}^n) \subseteq C(\mathbb{I}, \mathbb{R}^n)$. Thus, by Theorem 3.5 we have $\lambda = 0$.
Let $y = (g, r) \in \mathbb{Y}$ be fixed and assume that $\langle y, y^* \rangle = 0$ for all $y^* \in \mathbb{Y}^*$, i. e.,

$$\int_{\mathbb{I}} \lambda^T g\, dt + \gamma^T r = 0 \text{ for all } (\lambda, \gamma) \in \mathbb{Y}^*.$$

Choosing $(\lambda, \gamma) = (0, r)$ gives $r^T r = 0$, and hence $r = 0$. Therefore,

$$\int_{\mathbb{I}} \lambda^T g\, dt = 0 \text{ for all } \lambda \in C_0^\infty(\mathbb{I}, \mathbb{R}^n) \subseteq C^1_{EE^+}(\mathbb{I}, \mathbb{R}^n),$$

where $g \in C(\mathbb{I}, \mathbb{R}^n)$. Thus, again by Theorem 3.5 we have $g = 0$.
The proof for $\langle \mathbb{Z}, \mathbb{Z}^* \rangle$ follows the same lines and is therefore omitted. □

In order to bring the necessary conditions (1.3) into this abstract setting, we define the operator $\mathcal{L}^* : \mathbb{Y}^* \to \mathbb{Z}^*$ by

$$\mathcal{L}^*(\lambda, \gamma) = (-E^T \tfrac{d}{dt}(EE^+\lambda) - (A + EE^+\dot{E})^T\lambda, -B^T\lambda, \gamma - E(\underline{t})^T\lambda(\underline{t}), E(\bar{t})^T\lambda(\bar{t})), \tag{3.9}$$

compare again [20]. We can then show that $\mathcal{L}^*$ is conjugate to $\mathcal{L}$.

Theorem 3.7 *The operator $\mathcal{L}^* : \mathbb{Y}^* \to \mathbb{Z}^*$ defined by (3.9) is the (unique) conjugate of $\mathcal{L} : \mathbb{Z} \to \mathbb{Y}$ defined by (3.5).*

Proof. Let $z = (x, u) \in \mathbb{Z}$ and $\Lambda = (\lambda, \gamma) \in \mathbb{Y}^*$. Using that $E(\underline{t})^+ E(\underline{t})\gamma = \gamma$ and

$$E = EE^+E, \quad EE^+\dot{E} = EE^+\dot{E}E^+E + E\tfrac{d}{dt}(E^+E)$$

we have

$$\begin{aligned}
\langle \mathcal{L}(z), \Lambda \rangle &= \int_{\mathbb{I}} \left(\lambda^T \left[E\tfrac{d}{dt}(E^+Ex) - (A + E\tfrac{d}{dt}(E^+E))x - Bu\right]\right) dt + \gamma^T(E^+Ex)(\underline{t}) \\
&= \lambda^T Ex|_{\underline{t}}^{\bar{t}} + \int_{\mathbb{I}} \left(-\tfrac{d}{dt}(\lambda^T EE^+E)E^+Ex - \lambda^T \left[(A + E\tfrac{d}{dt}(E^+E))x - Bu\right]\right) dt + \gamma^T(E^+Ex)(\underline{t}) \\
&= (\lambda^T Ex)(\bar{t}) - (\lambda^T Ex)(\underline{t}) + \gamma^T x(\underline{t}) \\
&\qquad + \int_{\mathbb{I}} \left(-\tfrac{d}{dt}(\lambda^T EE^+)Ex - \lambda^T EE^+\dot{E}E^+Ex - \lambda^T \left[(A + E\tfrac{d}{dt}(E^+E))x - Bu\right]\right) dt \\
&= (\lambda^T Ex)(\bar{t}) - (\lambda^T Ex)(\underline{t}) + \gamma^T x(\underline{t}) \\
&\qquad + \int_{\mathbb{I}} \left(-\tfrac{d}{dt}(\lambda^T EE^+)Ex - \lambda^T(A + EE^+\dot{E})x - \lambda^T Bu)\right) dt = \langle z, \mathcal{L}^*(\Lambda)\rangle.
\end{aligned}$$

□

Finally, defining

$$\mathcal{T} : \mathbb{Y}^* \times \mathbb{Z} \to \mathbb{Y} \times \mathbb{Z}^*, \quad \mathcal{T}(\Lambda, z) = (\mathcal{L}(z), \mathcal{L}^*(\Lambda) - \mathcal{R}(z)), \tag{3.10}$$

with $\mathcal{R} : \mathbb{Z} \to \mathbb{Z}^*$ given by

$$\mathcal{R}(z) = (Wx + Su, S^Tx + Ru, 0, M_e x(\bar{t}))$$

for $z = (x, u) \in \mathbb{Z}$ and $\Lambda = (\lambda, \gamma) \in \mathbb{Y}^*$, we have that

$$\begin{aligned}
\mathcal{T}(\Lambda, z) = \big(&E\tfrac{d}{dt}(E^+Ex) - (A + E\tfrac{d}{dt}(E^+E))x - Bu, E(\underline{t})^+E(\underline{t})x(\underline{t}), \\
&- E^T\tfrac{d}{dt}(EE^+\lambda) - (A + EE^+\dot{E})^T\lambda - Wx - Su, \\
&-B^T\lambda - S^Tx - Ru, \gamma - E(\underline{t})^T\lambda(\underline{t}), E(\bar{t})^T\lambda(\bar{t}) - M_e x(\bar{t})\big)
\end{aligned}$$

and the necessary conditions given by (1.3) and the stated boundary conditions can be written as

$$\mathcal{T}(\Lambda, z) = (c, 0). \tag{3.11}$$

We now show that the operator $\mathcal{T}$ is self-conjugate with respect to suitably chosen dual systems. For this purpose, we introduce the abbreviations

$$\mathbb{V} = \mathbb{Y}^* \times \mathbb{Z}, \quad \mathbb{W} = \mathbb{Y} \times \mathbb{Z}^*,$$

set

$$\mathbb{V}^* = \mathbb{W}, \quad \mathbb{W}^* = \mathbb{V}$$

and introduce the so-called *canonical bilinear forms*

$$\langle (y^*, z), (y, z^*)\rangle = \langle y, y^*\rangle + \langle z, z^*\rangle = \langle (y, z^*), (y^*, z)\rangle.$$

Obviously, the pairs $\langle \mathbb{V}, \mathbb{V}^*\rangle$ and $\langle \mathbb{W}, \mathbb{W}^*\rangle$ become dual systems. By construction, we then not only have $\mathcal{T} : \mathbb{V} \to \mathbb{W}$ but also $\mathcal{T} : \mathbb{W}^* \to \mathbb{V}^*$.

Theorem 3.8 *The operator $\mathcal{T}$ as defined in (3.10) is self-conjugate, i. e., we have*

$$\langle \mathcal{T}(v), \tilde{v}\rangle = \langle v, \mathcal{T}(\tilde{v})\rangle \text{ for all } v, \tilde{v} \in \mathbb{V}. \tag{3.12}$$

Proof. Let $v = (\Lambda, z) \in \mathbb{V}$ and $\tilde{v} = (\tilde{\Lambda}, \tilde{z}) \in \mathbb{V}$. Then

$$\langle \mathcal{T}(\Lambda, z), (\tilde{\Lambda}, \tilde{z}) \rangle = \langle (\mathcal{L}(z), \mathcal{L}^*(\Lambda) - \mathcal{R}(z)), (\tilde{\Lambda}, \tilde{z}) \rangle = \langle \mathcal{L}(z), \tilde{\Lambda} \rangle - \langle \tilde{z}, \mathcal{R}(z) \rangle + \langle \tilde{z}, \mathcal{L}^*(\Lambda) \rangle$$

as well as

$$\langle (\Lambda, z), \mathcal{T}(\tilde{\Lambda}, \tilde{z}) \rangle = \langle (\Lambda, z), (\mathcal{L}(\tilde{z}), \mathcal{L}^*(\tilde{\Lambda}) - \mathcal{R}(\tilde{z})) \rangle = \langle \mathcal{L}(\tilde{z}), \Lambda \rangle - \langle z, \mathcal{R}(\tilde{z}) \rangle + \langle z, \mathcal{L}^*(\tilde{\Lambda}) \rangle.$$

The claim then follows because of

$$\langle \tilde{z}, \mathcal{R}(z) \rangle = \mathcal{Q}(z, \tilde{z}) = \mathcal{Q}(\tilde{z}, z) = \langle z, \mathcal{R}(\tilde{z}) \rangle,$$

using the symmetry of $\mathcal{Q}$. □

Note that the boundary value problem (3.11) coincides with (1.3) together with the stated boundary conditions if we assume sufficient smoothness of the data, see again [20]. In particular, we get the DAE for the Lagrange multiplier λ as

$$\begin{aligned} -E^T\dot{\lambda} &= -\tfrac{d}{dt}(E^T\lambda) + \dot{E}^T\lambda = -\tfrac{d}{dt}(E^TEE^+\lambda) + \dot{E}^T\lambda \\ &= -E^T\tfrac{d}{dt}(EE^+\lambda) - \dot{E}^TEE^+\lambda + \dot{E}^T\lambda \\ &= A^T\lambda + \dot{E}^TEE^+\lambda + Wx + Su - \dot{E}^TEE^+\lambda + \dot{E}^T\lambda \\ &= (A + \dot{E})^T\lambda + Wx + Su. \end{aligned}$$

In view of the observations from the abstract analysis, we introduce the following definition.

Definition 3.9 *A pair $(\mathcal{E}, \mathcal{A})$ of matrix functions, $\mathcal{A} \in C^0(\mathbb{I}, \mathbb{R}^{n,n})$, $\mathcal{E} \in C^1(\mathbb{I}, \mathbb{R}^{n,n})$, is called* self-adjoint *if and only if the following conditions are satisfied*

1. $\mathcal{E}^T = -\mathcal{E}$,
2. $\mathcal{A}^T = \mathcal{A} + \dot{\mathcal{E}}$.

Consider now a self-adjoint pair of sufficiently smooth matrix functions $\mathcal{E}, \mathcal{A} \in C^0(\mathbb{I}, \mathbb{R}^{n,n})$ and an associated DAE

$$\mathcal{E}\dot{z} = \mathcal{A}z + f, \tag{3.13}$$

cf. (2.2), with an inhomogeneity $f \in C^0(\mathbb{I}, \mathbb{R}^n)$ that is also assumed to be sufficiently smooth. Then we can scale the equation with a pointwise nonsingular matrix function $P \in C^0(\mathbb{I}, \mathbb{R}^{n,n})$ and perform a change of variables $z = Qy$ with a pointwise nonsingular matrix function $Q \in C^1(\mathbb{I}, \mathbb{R}^{n,n})$ which gives

$$P\mathcal{E}Q\dot{y} = P\mathcal{A}Qy - P\mathcal{E}\dot{Q}y + Pf. \tag{3.14}$$

We want to discuss transformations that preserve the self-adjointness of the pair. For this, we have to preserve the skew-symmetry of $\mathcal{E}$ and hence we have to require that $P = Q^T$, i. e., that the transformation is a *congruence transformation*. We then have the following lemma.

Lemma 3.10 *Consider a self-adjoint pair $(\mathcal{E}, \mathcal{A})$ of sufficiently smooth matrix functions $\mathcal{E}, \mathcal{A} \in C^0(\mathbb{I}, \mathbb{R}^{n,n})$ and apply a congruence transformation with a pointwise nonsingular $Q \in C^1(\mathbb{I}, \mathbb{R}^{n,n})$, leading to the pair*

$$(\tilde{\mathcal{E}}, \tilde{\mathcal{A}}) = (Q^T\mathcal{E}Q, Q^T\mathcal{A}Q - Q^T\mathcal{E}\dot{Q}).$$

Then the pair $(\tilde{\mathcal{E}}, \tilde{\mathcal{A}})$ is again self-adjoint.

Proof. The condition for $\tilde{\mathcal{E}}$ is trivially satisfied and for $\tilde{\mathcal{A}}$ we get

$$\tilde{\mathcal{A}} + \dot{\tilde{\mathcal{E}}} = Q^T\mathcal{A}Q - Q^T\mathcal{E}\dot{Q} + \dot{Q}^T\mathcal{E}Q + Q^T\dot{\mathcal{E}}Q + Q^T\mathcal{E}\dot{Q} = Q^T\mathcal{A}Q + \dot{Q}^T\mathcal{E}Q + Q^T\dot{\mathcal{E}}Q$$

and

$$\tilde{\mathcal{A}}^T = (Q^T\mathcal{A}Q - Q^T\mathcal{E}\dot{Q})^T = Q^T\mathcal{A}^TQ - \dot{Q}^T\mathcal{E}^TQ = Q^T\mathcal{A}Q + Q^T\dot{\mathcal{E}}Q + \dot{Q}^T\mathcal{E}Q.$$

□

4 Condensed forms for self-adjoint pairs of matrix functions

For matrix pairs $(\mathcal{E}, \mathcal{A})$, with $\mathcal{E}, \mathcal{A} \in \mathbb{R}^{n,n}$, $\mathcal{E} = -\mathcal{E}^T$ and $\mathcal{A} = \mathcal{A}^T$, the canonical form under congruence, i. e., $(Q^T\mathcal{E}Q, Q^T\mathcal{A}Q)$ is well known, see e. g. [29, 30]. If the transformation matrices are restricted to be real orthogonal matrices, then the resulting staircase form has been developed in [7], modifying the staircase form of [31].

We will now extend these results to self-adjoint pairs of matrix functions. To achieve a staircase form, we always have to assume that certain matrix functions have constant rank in the given interval $\mathbb{I}$. If this is not the case, then one can restrict the problem to a smaller interval where this condition holds, and consider the problem piecewise. In the following, we therefore assume that the desired ranks are constant in the complete interval $\mathbb{I}$. Then we can make use of the following theorem which is an extended real version of Theorem 3.9 in [19] originating to [10].

Theorem 4.1 *Let $E \in C^\ell(\mathbb{I}, \mathbb{R}^{m,n})$, $\ell \in \mathbb{N}_0 \cup \{\infty\}$, with* rank $E(t) = r$ *for all $t \in \mathbb{I}$. Then there exist pointwise real orthogonal matrix functions $U \in C^\ell(\mathbb{I}, \mathbb{R}^{m,m})$ and $V \in C^\ell(\mathbb{I}, \mathbb{R}^{n,n})$, such that*

$$U^TEV = \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} \tag{4.1}$$

with pointwise nonsingular $\Sigma \in C^\ell(\mathbb{I}, \mathbb{R}^{r,r})$.

If $E \in \mathcal{C}^\ell(\mathbb{I}, \mathbb{R}^{n,n})$ is symmetric (skew-symmetric), with rank $E(t) = r$ *for all $t \in \mathbb{I}$, then there exists a pointwise real orthogonal matrix function $U \in \mathcal{C}^\ell(\mathbb{I}, \mathbb{R}^{n,n})$ such that*

$$U^TEU = \begin{bmatrix} \Delta & 0 \\ 0 & 0 \end{bmatrix} \tag{4.2}$$

with pointwise nonsingular and symmetric (skew-symmetric) $\Delta \in \mathcal{C}^\ell(\mathbb{I}, \mathbb{R}^{r,r})$.

Based on sequences of factorizations as in Theorem 4.1 we then have the following staircase form.

Theorem 4.2 *Consider a self-adjoint pair $(\mathcal{E}, \mathcal{A})$ of sufficiently smooth matrix functions $\mathcal{E}, \mathcal{A} \in C^0(\mathbb{I}, \mathbb{R}^{n,n})$. Then, under appropriate constant rank conditions, there exists a con-*

gruence transformation with a pointwise orthogonal $U \in C^1(\mathbb{I}, \mathbb{R}^{n,n})$, leading to the pair

$$U^T\mathcal{E}U = \left[\begin{array}{cccc|c|ccc|c|c} \mathcal{E}_{11} & \cdots & \cdots & \mathcal{E}_{1,m} & \mathcal{E}_{1,m+1} & \mathcal{E}_{1,m+2} & \cdots & \mathcal{E}_{1,2m} & 0 & n_1 \\ \vdots & \ddots & & \vdots & \vdots & \vdots & \iddots & \iddots & & \vdots \\ \vdots & & \ddots & \vdots & \vdots & \mathcal{E}_{m-1,m+2} & \iddots & & & \vdots \\ -\mathcal{E}_{1,m}^T & \cdots & \cdots & \mathcal{E}_{m,m} & \mathcal{E}_{m,m+1} & 0 & & & & n_m \\ \hline -\mathcal{E}_{1,m+1}^T & \cdots & \cdots & -\mathcal{E}_{m,m+1}^T & \mathcal{E}_{m+1,m+1} & & & & & l \\ \hline -\mathcal{E}_{1,m+2}^T & \cdots & -\mathcal{E}_{m-1,m+2}^T & 0 & & & & & & q_m \\ \vdots & \iddots & \iddots & & & & & & & \vdots \\ -\mathcal{E}_{1,2m}^T & \iddots & & & & & & & & q_2 \\ 0 & & & & & & & & & q_1 \end{array}\right]$$

$$U^T\mathcal{A}U - U^T\mathcal{E}\dot{U} = \left[\begin{array}{cccc|c|ccc|c|c} \mathcal{A}_{11} & \cdots & \cdots & \mathcal{A}_{1,m} & \mathcal{A}_{1,m+1} & \mathcal{A}_{1,m+2} & \cdots & \cdots & \mathcal{A}_{1,2m+1} & n_1 \\ \vdots & \ddots & & \vdots & \vdots & \vdots & & \iddots & & \vdots \\ \vdots & & \ddots & \vdots & \vdots & \vdots & \iddots & & & \vdots \\ \mathcal{A}_{m,1} & \cdots & \cdots & \mathcal{A}_{m,m} & \mathcal{A}_{m,m+1} & \mathcal{A}_{m,m+2} & & & & n_m \\ \hline \mathcal{A}_{m+1,1} & \cdots & \cdots & \mathcal{A}_{m+1,m} & \mathcal{A}_{m+1,m+1} & & & & & l \\ \hline \mathcal{A}_{m+2,1} & \cdots & \cdots & \mathcal{A}_{m+2,m} & & & & & & q_m \\ \vdots & & \iddots & & & & & & & \vdots \\ \vdots & \iddots & & & & & & & & \vdots \\ \mathcal{A}_{2m+1,1} & & & & & & & & & q_1 \end{array}\right], \tag{4.3}$$

where $q_1 \geq n_1 \geq q_2 \geq n_2 \geq \ldots \geq q_m \geq n_m$,

$$\begin{aligned} \mathcal{E}_{j,2m+1-j} &\in C^0(\mathbb{I}, \mathbb{R}^{n_j,q_{j+1}}), \qquad 1 \leq j \leq m-1, \\ \mathcal{E}_{m+1,m+1} &= \begin{bmatrix} \Delta & 0 \\ 0 & 0 \end{bmatrix}, \quad \Delta = -\Delta^T \in C^0(\mathbb{I}, \mathbb{R}^{2p,2p}), \\ \mathcal{E}_{j,j} &= -\mathcal{E}_{j,j}^T, \qquad j = 1, \ldots, m, \\ \mathcal{A}_{j,2m+2-j} &= \mathcal{A}_{2m+2-j,j}^T = \begin{bmatrix} \Gamma_j & 0 \end{bmatrix} \in C^0(\mathbb{I}, \mathbb{R}^{n_j,q_j}),\ \Gamma_j \in C^0(\mathbb{I}, \mathbb{R}^{n_j,n_j}),\ 1 \leq j \leq m, \\ \mathcal{A}_{m+1,m+1} &= \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{bmatrix},\ \Sigma_{11} = \Sigma_{11}^T + \dot{\Delta}^T \in C^0(\mathbb{I}, \mathbb{R}^{2p,2p}), \\ \Sigma_{22} &= \Sigma_{22}^T \in C^0(\mathbb{I}, \mathbb{R}^{l-2p,l-2p}), \end{aligned}$$

and the blocks Σ_{22} and Δ and Γ_j, $j = 1, \ldots, m$ are pointwise nonsingular. Furthermore, each of the first m block columns (block rows) of the matrix $U^T\mathcal{E}U$ has full column rank (full row rank).

Proof. The proof is an extension of the proof for the matrix case given in [7]. It is described by an explicit, but recursive procedure. Note that some blocks may be void, i. e., they may have zero rows or zero columns or both.

Let $(\mathcal{E}, \mathcal{A})$ be self-adjoint. If $\mathcal{E} = \mathcal{A} = 0$, or if $\mathcal{E}$ is nonsingular, then the pair is trivially in staircase form.

If $\mathcal{E}$ is singular and of constant rank, then determine via Theorem 4.1 a factorization

$$U_1^T \mathcal{E} U_1 = \begin{bmatrix} \Delta & 0 \\ 0 & 0 \end{bmatrix},$$

with U_1 pointwise orthogonal and $\Delta = -\Delta^T$ pointwise nonsingular. Perform a congruence transformation with U_1 to form

$$U_1^T \mathcal{E} U_1 = \begin{bmatrix} \Delta & 0 \\ 0 & 0 \end{bmatrix}, \quad U_1^T \mathcal{A} U_1 - U_1^T \mathcal{E} \dot{U}_1 = \begin{bmatrix} \hat{\mathcal{A}}_{11} & \hat{\mathcal{A}}_{12} \\ \hat{\mathcal{A}}_{21} & \hat{\mathcal{A}}_{22} \end{bmatrix}. \tag{4.4}$$

If $\hat{\mathcal{A}}_{22}$ is pointwise nonsingular, then the staircase form is complete.

If $\hat{\mathcal{A}}_{22}$ is globally singular and has constant rank, then determine via Theorem 4.1 a factorization

$$U_2^T \hat{\mathcal{A}}_{22} U_2 = \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix}$$

with U_2 orthogonal and Σ pointwise nonsingular. This leads to the congruence transformation

$$\begin{aligned}
&\begin{bmatrix} I & 0 \\ 0 & U_2 \end{bmatrix}^T \begin{bmatrix} \Delta & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & U_2 \end{bmatrix} = \begin{bmatrix} \Delta & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \\
&\begin{bmatrix} I & 0 \\ 0 & U_2 \end{bmatrix}^T \begin{bmatrix} \hat{\mathcal{A}}_{11} & \hat{\mathcal{A}}_{12} \\ \hat{\mathcal{A}}_{21} & \hat{\mathcal{A}}_{22} \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & U_2 \end{bmatrix} - \begin{bmatrix} I & 0 \\ 0 & U_2 \end{bmatrix}^T \begin{bmatrix} \Delta & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & \dot{U}_2 \end{bmatrix} \\
&\quad = \begin{bmatrix} \tilde{\mathcal{A}}_{11} & \tilde{\mathcal{A}}_{12} & \tilde{\mathcal{A}}_{13} \\ \tilde{\mathcal{A}}_{21} & \Sigma & 0 \\ \tilde{\mathcal{A}}_{31} & 0 & 0 \end{bmatrix},
\end{aligned} \tag{4.5}$$

with $\tilde{\mathcal{A}}_{21} = \tilde{\mathcal{A}}_{12}^T$ and $\tilde{\mathcal{A}}_{31} = \tilde{\mathcal{A}}_{13}^T$. Under a constant rank assumption for $\tilde{\mathcal{A}}_{13}$, we determine a factorization

$$U_3^T \tilde{\mathcal{A}}_{13} V_3 = \begin{bmatrix} \Gamma & 0 \\ 0 & 0 \end{bmatrix}$$

with U_3 and V_3 pointwise orthogonal and Γ pointwise nonsingular, and perform a congruence

transformation

$$\begin{bmatrix} U_3 & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & V_3 \end{bmatrix}^T \begin{bmatrix} \Delta & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} U_3 & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & V_3 \end{bmatrix} = \begin{bmatrix} \mathcal{E}_{11} & \mathcal{E}_{12} & \mathcal{E}_{13} & 0 & 0 \\ -\mathcal{E}_{12}^T & \mathcal{E}_{22} & 0 & 0 & 0 \\ -\mathcal{E}_{13}^T & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\begin{bmatrix} U_3 & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & V_3 \end{bmatrix}^T \begin{bmatrix} \tilde{\mathcal{A}}_{11} & \tilde{\mathcal{A}}_{12} & \tilde{\mathcal{A}}_{13} \\ \tilde{\mathcal{A}}_{21} & \Sigma & 0 \\ \tilde{\mathcal{A}}_{31} & 0 & 0 \end{bmatrix} \begin{bmatrix} U_3 & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & V_3 \end{bmatrix} \tag{4.6}$$

$$- \begin{bmatrix} U_3 & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & V_3 \end{bmatrix}^T \begin{bmatrix} \Delta & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{U}_3 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \dot{V}_3 \end{bmatrix}$$

$$= \begin{bmatrix} \mathcal{A}_{11} & \mathcal{A}_{12} & \mathcal{A}_{13} & \Gamma & 0 \\ \mathcal{A}_{21} & \mathcal{A}_{22} & \mathcal{A}_{23} & 0 & 0 \\ \mathcal{A}_{31} & \mathcal{A}_{32} & \Sigma & 0 & 0 \\ \Gamma^T & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

where $U_3^T \Delta U_3 = \begin{bmatrix} \mathcal{E}_{11} & \mathcal{E}_{12} \\ -\mathcal{E}_{12}^T & \mathcal{E}_{22} \end{bmatrix}$ is skew-symmetric and $\mathcal{E}_{13} = 0$. The block $\mathcal{E}_{13}$ may fill with nonzero entries later in the process, so we do not distinguish it from other blocks that may be nonzero.

We then recursively apply the same reduction to the central self-adjoint pair

$$\left(\begin{bmatrix} \mathcal{E}_{22} & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} \mathcal{A}_{22} & \mathcal{A}_{23} \\ \mathcal{A}_{32} & \Sigma \end{bmatrix} \right).$$

This corresponds to performing another congruence transformation to (4.6) that modifies block rows and columns 2 and 3, typically changing $\mathcal{E}_{12}$, $\mathcal{E}_{13}$, $\mathcal{A}_{12}$, $\mathcal{A}_{21}$, $\mathcal{A}_{13}$, and $\mathcal{A}_{31}$ along with the central pair. After a finite number of steps of congruence transformations then the pair is still self-adjoint and in the desired staircase form.

The property that each of the first m block columns (block rows) has full rank follows by our construction. □

The orthogonal staircase form allows to characterize many of the properties of the self-adjoint pair and associated differential-algebraic systems. With nonsingular congruence transformations it is possible to reduce the system even further.

Corollary 4.3 *Consider a self-adjoint pair $(\mathcal{E}, \mathcal{A})$ of sufficiently smooth matrix functions $\mathcal{E}, \mathcal{A} \in C^0(\mathbb{I}, \mathbb{R}^{n,n})$. Then, under appropriate constant rank conditions, there exists a congru-*

ence transformation with a pointwise nonsingular $T \in C^1(\mathbb{I}, \mathbb{R}^{n,n})$, leading to the pair

$$T^T\mathcal{E}T = \left[\begin{array}{cccc|c|ccc|c|c} \mathcal{E}_{11} & \cdots & \cdots & \mathcal{E}_{1,m} & \mathcal{E}_{1,m+1} & \mathcal{E}_{1,m+2} & \cdots & \mathcal{E}_{1,2m} & 0 & n_1 \\ \vdots & \ddots & & \vdots & \vdots & \vdots & \iddots & \iddots & & \vdots \\ \vdots & & \ddots & \vdots & \vdots & \mathcal{E}_{m-1,m+2} & \iddots & & & \vdots \\ -\mathcal{E}_{1,m}^T & \cdots & \cdots & \mathcal{E}_{m,m} & \mathcal{E}_{m,m+1} & 0 & & & & n_m \\ \hline -\mathcal{E}_{1,m+1}^T & \cdots & \cdots & -\mathcal{E}_{m,m+1}^T & \mathcal{E}_{m+1,m+1} & & & & & l \\ \hline -\mathcal{E}_{1,m+2}^T & \cdots & -\mathcal{E}_{m-1,m+2}^T & 0 & & & & & & q_m \\ \vdots & \iddots & \iddots & & & & & & & \vdots \\ -\mathcal{E}_{1,2m}^T & \iddots & & & & & & & & q_2 \\ 0 & & & & & & & & & q_1 \end{array}\right]$$

$$T^T\mathcal{A}T - T^T\mathcal{E}\dot{T} = \left[\begin{array}{cccc|c|ccc|c|c} \mathcal{A}_{1,1} & \cdots & \cdots & \mathcal{A}_{1,m} & \mathcal{A}_{1,m+1} & \mathcal{A}_{1,m+2} & \cdots & \cdots & \mathcal{A}_{1,2m+1} & n_1 \\ \vdots & \ddots & & \vdots & \vdots & \vdots & & \iddots & & \vdots \\ \vdots & & \ddots & \vdots & \vdots & \vdots & \iddots & & & \vdots \\ \mathcal{A}_{m,1} & \cdots & \cdots & \mathcal{A}_{m,m} & \mathcal{A}_{m,m+1} & \mathcal{A}_{m,m+2} & & & & n_m \\ \hline \mathcal{A}_{m+1,1} & \cdots & \cdots & \mathcal{A}_{m+1,m} & \mathcal{A}_{m+1,m+1} & & & & & l \\ \hline 0 & \cdots & 0 & \mathcal{A}_{m+2,m} & & & & & & q_m \\ \vdots & & \iddots & & & & & & & \vdots \\ 0 & \iddots & & & & & & & & \vdots \\ \mathcal{A}_{2m+1,1} & & & & & & & & & q_1 \end{array}\right], \tag{4.7}$$

where $q_1 \geq n_1 \geq q_2 \geq n_2 \geq \ldots \geq q_m \geq n_m$,

$$\begin{aligned} \mathcal{E}_{j,2m+1-j} &\in C^0(\mathbb{I}, \mathbb{R}^{n_j, q_{j+1}}), \qquad 1 \leq j \leq m-1, \\ \mathcal{E}_{m+1,m+1} &= \begin{bmatrix} J_p & 0 \\ 0 & 0 \end{bmatrix}, \quad J_p := \begin{bmatrix} 0 & I_p \\ -I_p & 0 \end{bmatrix}, \\ \mathcal{A}_{j,2m+2-j} &= \mathcal{A}_{2m+2-j,j}^T = \begin{bmatrix} I_{n_j} & 0 \end{bmatrix} \in C^0(\mathbb{I}, \mathbb{R}^{n_j, q_j}), \qquad 1 \leq j \leq m, \\ \mathcal{A}_{i,j} &= -\dot{\mathcal{E}}_{i,j}, \quad i = 1, \ldots, m-1,\ j = m+2, \ldots, 2m+1-i, \\ \mathcal{A}_{m+1,m+1} &= \begin{bmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{bmatrix}, \quad \Sigma_{11} = \Sigma_{11}^T \in C^0(\mathbb{I}, \mathbb{R}^{2p,2p}), \quad \Sigma_{22} = \Sigma_{22}^T \in C^0(\mathbb{I}, \mathbb{R}^{l-2p,l-2p}), \end{aligned}$$

and the block Σ_{22} is pointwise nonsingular. Furthermore, each of the first m block columns (block rows) of the matrix $T^T\mathcal{E}T$ has full column rank (full row rank).

Proof. Starting from the staircase form (4.3) we can first perform a congruence transformation

$$(\tilde{\mathcal{E}}, \tilde{\mathcal{A}}) = (T_1^T U^T \mathcal{E} U T_1, T_1^T U^T \mathcal{A} U T_1 - T_1^T U^T \mathcal{E} \tfrac{d}{dt}(U T_1))$$

with $T_1^T = \operatorname{diag}(\Gamma_1^{-1}, \ldots, \Gamma_m^{-1}, L, I_{q_m}, \ldots, I_{q_1})$ where

$$L = \begin{bmatrix} I_{2p} & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I_{2p} \end{bmatrix}.$$

Then, with block-Gauss congruence transformations, we can eliminate all elements above the block anti-diagonal of $\tilde{\mathcal{A}}$ in block-columns $1, \ldots, m$.

Finally, we perform a congruence transformation to the nonsingular first diagonal block Δ in $\mathcal{E}_{m+1,m+1}$. Let

$$Q_1^T \Delta Q_1 = \begin{bmatrix} \Delta_{11} & \Delta_{12} \\ \Delta_{21} & 0 \end{bmatrix}$$

be a block anti-triangular decomposition of Δ, where Δ_{12}, Δ_{21} are invertible. This can be constructed just as in the constant coefficient case, see [6]. Then let

$$Q_2^T = \begin{bmatrix} I & -\frac{1}{2}\Delta_{11}\Delta_{21}^{-1} \\ 0 & -\Delta_{21}^{-1} \end{bmatrix}$$

be partitioned analogously, such that

$$Q_2^T Q_1^T \Delta Q_1 Q_2 = \begin{bmatrix} 0 & I_p \\ -I_p & 0 \end{bmatrix} = J_p.$$

□

Note that neither the orthogonal staircase form (4.3) nor the condensed form (4.7) is a normal from in the algebraic sense, since there is still further refinement possible using congruence transformations. For the purpose of analyzing systems of differential-algebraic equations, however, these condensed forms are sufficient.

Corollary 4.4 *Consider a self-adjoint pair $(\mathcal{E}, \mathcal{A})$ of sufficiently smooth matrix functions $\mathcal{E}, \mathcal{A} \in C^0(\mathbb{I}, \mathbb{R}^{n,n})$ and suppose that appropriate constant rank assumptions hold so that there exists a congruence transformation with a pointwise orthogonal $U \in C^1(\mathbb{I}, \mathbb{R}^{n,n})$ to the staircase form (4.3).*

i) The differential-algebraic equation (3.13) is regular if and only if in the staircase form $n_j = q_j$ for all $j = 1, \ldots, m$.

ii) If $m = 0$ then the DAE is regular and strangeness-free.

iii) If $m > 0$ then $\mu \leq 2m - 1$ differentiations will be necessary to solve the system if $2p = \ell$ and $\mu \leq 2m$ differentiations will be necessary otherwise. If the system is regular, then the inequalities become equalities.

Proof. i) If $q_1 > n_1$ then it is clear that the DAE is nonregular, because then it has a zero row and hence the problem is not solvable for every smooth right hand side. If $n_i = q_i$ for $i = 1, \ldots, \ell - 1$ but $q_\ell > n_\ell$, then we can successively solve the equation from the bottom up in a unique way, until we reach the remaining system with a nonsquare block $\mathcal{A}_{2m+2-\ell,\ell} = \mathcal{A}_{\ell,2m+2-\ell}^T = \begin{bmatrix} \Gamma_\ell & 0 \end{bmatrix}^T$. Then again, the last $q_\ell - n_\ell$ equations associated with this block are not solvable for every smooth right hand side and hence the problem is not regular.

ii) If $m = 0$, then the associated staircase form has the form (4.4) with $\hat{\mathcal{A}}_{22}$ pointwise nonsingular and it is well known already from the unstructured case, see [17, 19], that the associated DAE is regular and strangeness-free.

iii) Using the condensed form (4.7), we can apply backward substitution starting with the last block row. Then we have to differentiate the right hand side at most m times until we reach the middle block. If after backward substitution the middle block contains an algebraic

part, then we continue with at most m further differentiations. If the middle block has no algebraic part, then at most $m-1$ further differentiations are necessary.

In the regular case, using the fact that the first m block columns (block rows) have full rank makes sure that all derivatives actually occur. □

Example 4.5 Consider the DAE

$$\left[\begin{array}{c|cc|c|c} 0 & 0 & 1 & 1 & 0 \\ \hline 0 & 0 & -1 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 \\ \hline -1 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 \end{array}\right] \left[\begin{array}{c} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \\ \dot{x}_5 \end{array}\right] = \left[\begin{array}{c|cc|c|c} 0 & 0 & 0 & 0 & 1 \\ \hline 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 \\ \hline 1 & 0 & 0 & 0 & 0 \end{array}\right] \left[\begin{array}{c} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{array}\right] + \left[\begin{array}{c} f_1 \\ f_2 \\ f_3 \\ f_4 \\ f_5 \end{array}\right],$$

which is in the condensed form (4.7) with $m=2$, $q_1=n_1=1$, $q_2=1$, $n_2=0$, $l=2$, $p=1$. Since $q_2 \neq n_2$ the system is non-regular. To solve the system we first get $x_1=-f_5$, and by substituting $\dot{x}_1=-\dot{f}_5$ the solution for x_2, x_3 can be determined from the Hamiltonian system

$$\left[\begin{array}{cc} 0 & -1 \\ 1 & 0 \end{array}\right] \left[\begin{array}{c} \dot{x}_2 \\ \dot{x}_3 \end{array}\right] = \left[\begin{array}{c} x_2 \\ x_3 \end{array}\right] + \left[\begin{array}{c} f_2 \\ f_3-\dot{f}_5 \end{array}\right].$$

Finally we can solve the differential system

$$\dot{x}_4 = x_2 + x_5 + f_1 + f_2$$

to obtain a solution for x_4. Thus, $\mu=1<3$ differentiations are necessary to solve the system. The component x_5 is undetermined and we have the consistency condition $f_4-\dot{f}_5=0$ for the inhomogeneity.

Example 4.6 Consider the DAE

$$\left[\begin{array}{c|ccc|c} 0 & -1 & 0 & -1 & 0 \\ \hline 1 & 0 & 1 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 \end{array}\right] \left[\begin{array}{c} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \\ \dot{x}_5 \end{array}\right] = \left[\begin{array}{c|ccc|c} 0 & 0 & 0 & 0 & 1 \\ \hline 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ \hline 1 & 0 & 0 & 0 & 0 \end{array}\right] \left[\begin{array}{c} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{array}\right] + \left[\begin{array}{c} f_1 \\ f_2 \\ f_3 \\ f_4 \\ f_5 \end{array}\right],$$

which again is in the condensed form (4.7) with $m=1$, $q_1=n_1=1$, $l=3$, $p=1$. For the solution we get $x_1=-f_5$, and by substituting $\dot{x}_1=-\dot{f}_5$ we get $x_4=-f_4-\dot{f}_5$. The solution for x_2, x_3 can be determined from the Hamiltonian system

$$\left[\begin{array}{cc} 0 & 1 \\ -1 & 0 \end{array}\right] \left[\begin{array}{c} \dot{x}_2 \\ \dot{x}_3 \end{array}\right] = \left[\begin{array}{cc} 0 & 0 \\ 0 & 1 \end{array}\right] \left[\begin{array}{c} x_2 \\ x_3 \end{array}\right] + \left[\begin{array}{c} f_2+\dot{f}_5 \\ f_3 \end{array}\right].$$

and by substituting $\dot{x}_4=-\dot{f}_4-\ddot{f}_5$ we get

$$x_5 = x_3 - f_1 + f_3 + \dot{f}_4 + \ddot{f}_5.$$

Here, $\mu=2=2m$ differentiations are necessary to solve the system.

When the pair $(\mathcal{E}, \mathcal{A})$ is in the condensed form (4.7) and the associated DAE (3.13) is regular, then we can permute and re-arrange the condensed form to the form

$$\left(\begin{bmatrix} \tilde{\mathcal{E}}_{11} & \tilde{\mathcal{E}}_{12} & \tilde{\mathcal{E}}_{13} & \tilde{\mathcal{E}}_{14} \\ -\tilde{\mathcal{E}}_{12}^T & \tilde{\mathcal{E}}_{22} & 0 & 0 \\ -\tilde{\mathcal{E}}_{13}^T & 0 & 0 & 0 \\ -\tilde{\mathcal{E}}_{14}^T & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} \tilde{\mathcal{A}}_{11} & \tilde{\mathcal{A}}_{12} - \dot{\tilde{\mathcal{E}}}_{12} & \tilde{\mathcal{A}}_{13} - \dot{\tilde{\mathcal{E}}}_{13} & I_r - \dot{\tilde{\mathcal{E}}}_{14} \\ \tilde{\mathcal{A}}_{12}^T & \tilde{\mathcal{A}}_{22} & 0 & 0 \\ \tilde{\mathcal{A}}_{13}^T & 0 & \tilde{\mathcal{A}}_{33} & 0 \\ I_r & 0 & 0 & 0 \end{bmatrix} \right), \tag{4.8}$$

where $\tilde{\mathcal{E}}_{22} = J_p$ and $\tilde{\mathcal{A}}_{33}$ are invertible, and $\tilde{\mathcal{E}}_{14}$ is block upper-triangular with square diagonal blocks, which are zero matrices. Performing some further block-Gauss elimination congruence transformation we can eliminate all blocks above $\tilde{\mathcal{A}}_{33}$ and above and to the left of $\tilde{\mathcal{E}}_{22} = J_p$ and obtain the form

$$\left(\begin{bmatrix} \hat{\mathcal{E}}_{11} & 0 & \hat{\mathcal{E}}_{13} & \hat{\mathcal{E}}_{14} \\ 0 & J_p & 0 & 0 \\ -\hat{\mathcal{E}}_{13}^T & 0 & 0 & 0 \\ -\hat{\mathcal{E}}_{14}^T & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} \hat{\mathcal{A}}_{11} & \hat{\mathcal{A}}_{12} - \dot{\hat{\mathcal{E}}}_{12} & -\dot{\hat{\mathcal{E}}}_{13} & I_r - \dot{\hat{\mathcal{E}}}_{14} \\ \hat{\mathcal{A}}_{12}^T & \hat{\mathcal{A}}_{22} & 0 & 0 \\ 0 & 0 & \hat{\mathcal{A}}_{33} & 0 \\ I_r & 0 & 0 & 0 \end{bmatrix} \right). \tag{4.9}$$

One further block permutation (exchanging the first two block rows and columns), partitioning the blocks further, and renaming the blocks, we finally obtain the form

$$\left(\begin{bmatrix} 0 & I_p & 0 & 0 & 0 \\ -I_p & 0 & 0 & 0 & 0 \\ 0 & 0 & \mathcal{E}_{33} & \mathcal{E}_{34} & \mathcal{E}_{35} \\ 0 & 0 & -\mathcal{E}_{34}^T & 0 & 0 \\ 0 & 0 & -\mathcal{E}_{35}^T & 0 & 0 \end{bmatrix}, \begin{bmatrix} \mathcal{A}_{11} & \mathcal{A}_{12} & \mathcal{A}_{13} & 0 & 0 \\ \mathcal{A}_{12}^T & \mathcal{A}_{22} & \mathcal{A}_{23} & 0 & 0 \\ \mathcal{A}_{13}^T & \mathcal{A}_{23}^T & \mathcal{A}_{33} & -\dot{\mathcal{E}}_{34} & I_r - \dot{\mathcal{E}}_{35} \\ 0 & 0 & 0 & \mathcal{A}_{44} & 0 \\ 0 & 0 & I_r & 0 & 0 \end{bmatrix} \right), \tag{4.10}$$

with $\mathcal{A}_{44}$ invertible, and $\mathcal{E}_{35}$ block upper-triangular with square diagonal blocks, which are zero matrices.

If we consider the DAE corresponding to the pair (4.10), then we obtain the following equations.

$$\begin{aligned} \dot{z}_2 &= \mathcal{A}_{11} z_1 + \mathcal{A}_{12} z_2 + \mathcal{A}_{13} z_3 + \tilde{f}_1, \\ -\dot{z}_1 &= \mathcal{A}_{12}^T z_1 + \mathcal{A}_{22} z_2 + \mathcal{A}_{23} z_3 + \tilde{f}_2, \\ \mathcal{E}_{33}\dot{z}_3 + \mathcal{E}_{34}\dot{z}_4 + \mathcal{E}_{35}\dot{z}_5 &= \mathcal{A}_{13}^T z_1 + \mathcal{A}_{23}^T z_2 + \mathcal{A}_{33} z_3 - \dot{\mathcal{E}}_{34} z_4 + (I_r - \dot{\mathcal{E}}_{35}) z_5 + \tilde{f}_3, \\ -\mathcal{E}_{34}^T \dot{z}_3 &= \mathcal{A}_{44} z_4 + \tilde{f}_4, \\ -\mathcal{E}_{35}^T \dot{z}_3 &= z_3 + \tilde{f}_5. \end{aligned} \tag{4.11}$$

From (4.11) we can directly obtain the algebraic constraints that are included in the system which are in the third to fifth equation. These equations determine the consistency conditions for initial or boundary conditions and the smoothness requirements for the inhomogeneities.

5 Self-adjoint DAEs and Hamiltonian systems

It is well known that for the optimal control of ordinary differential equations, i. e., $E = I_n$, with an invertible weight function R, the optimality boundary value problem is associated with a *Hamiltonian system* of differential equations

$$\begin{bmatrix} \dot{x} \\ \dot{\lambda} \end{bmatrix} = \begin{bmatrix} A - BR^{-1}S^T & -BR^{-1}B^T \\ SR^{-1}S^T - W & -(A - BR^{-1}S^T)^T \end{bmatrix} \begin{bmatrix} x \\ \lambda \end{bmatrix} + \begin{bmatrix} f \\ 0 \end{bmatrix}, \tag{5.1}$$

that is obtained by inserting $u = -R^{-1}(B^T\lambda + S^T x)$ and multiplying with J_n^{-1} from the left see [1, 13, 15, 25].

This Hamiltonian system generates a *symplectic flow*, i. e., the fundamental solution Φ satisfies $\Phi^T J_n \Phi = J_n$, see [13].

On the other hand, even in the case of ordinary differential equations, when R is singular, this reduction to a Hamiltonian system is not possible and one typically uses the theory of singular perturbations [26].

In the following, we will analyze whether there is nevertheless a symplectic flow describing the dynamic part of a differential-algebraic equation of the form (3.13).

Lemma 5.1 *Consider a self-adjoint pair $(\mathcal{E}, \mathcal{A})$ of sufficiently smooth matrix functions $\mathcal{E}, \mathcal{A} \in C^0(\mathbb{I}, \mathbb{R}^{n,n})$ and the associated DAE (3.13), where*

$$\mathcal{E} = \begin{bmatrix} 0 & I_p & 0 \\ -I_p & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \ \mathcal{A} = \begin{bmatrix} \mathcal{A}_{11} & \mathcal{A}_{12} & \mathcal{A}_{13} \\ \mathcal{A}_{21} & \mathcal{A}_{22} & \mathcal{A}_{23} \\ \mathcal{A}_{31} & \mathcal{A}_{32} & \mathcal{A}_{33} \end{bmatrix}. \tag{5.2}$$

Suppose that there exists a symmetric matrix $\mathcal{M}_e = \mathcal{M}_e^T \in \mathbb{R}^{p,p}$ such that the Riccati differential equation

$$\dot{P} + P\mathcal{A}_{22}P - \mathcal{A}_{12}P - P\mathcal{A}_{12}^T + \mathcal{A}_{11} = 0, \ P(\bar{t}) = \mathcal{M}_e$$

has a symmetric solution $P \in C^1(\mathbb{I}, \mathbb{R}^{p,p})$. Then there exists a congruence transformation with a pointwise nonsingular $Q \in C^1(\mathbb{I}, \mathbb{R}^{n,n})$, leading to a pair

$$(\tilde{\mathcal{E}}, \tilde{\mathcal{A}}) = (Q^T\mathcal{E}Q, Q^T\mathcal{A}Q - Q^T\mathcal{E}\dot{Q}),$$

with

$$(\tilde{\mathcal{E}}, \tilde{\mathcal{A}}) = \left(\begin{bmatrix} 0 & I_p & 0 \\ -I_p & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & \tilde{\mathcal{A}}_{12} & \tilde{\mathcal{A}}_{13} \\ \tilde{\mathcal{A}}_{21} & \tilde{\mathcal{A}}_{22} & \tilde{\mathcal{A}}_{23} \\ \tilde{\mathcal{A}}_{31} & \tilde{\mathcal{A}}_{32} & \tilde{\mathcal{A}}_{33} \end{bmatrix} \right). \tag{5.3}$$

Proof. With

$$Q^T = \begin{bmatrix} I_p & -P & 0 \\ 0 & I_p & 0 \\ 0 & 0 & I \end{bmatrix}$$

we obtain

$$Q^T\mathcal{E}Q = \begin{bmatrix} 0 & I_p & 0 \\ -I_p & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad Q^T\mathcal{A}Q - Q^T\mathcal{E}\dot{Q} = \begin{bmatrix} \tilde{\mathcal{A}}_{11} & \tilde{\mathcal{A}}_{12} & \tilde{\mathcal{A}}_{13} \\ \tilde{\mathcal{A}}_{21} & \tilde{\mathcal{A}}_{22} & \tilde{\mathcal{A}}_{23} \\ \tilde{\mathcal{A}}_{31} & \tilde{\mathcal{A}}_{32} & \tilde{\mathcal{A}}_{33} \end{bmatrix}$$

with $\tilde{\mathcal{A}}_{11} = \dot{P} + P\mathcal{A}_{22}P - \mathcal{A}_{12}P - P\mathcal{A}_{12}^T + \mathcal{A}_{11} = 0$. □

If the self-adjoint pair is in the form (5.3), then it has exactly the structure of the self-adjoint pair arising from the linear quadratic optimal control problem. If $\mathcal{E}$ has constant rank $r = 2p$ (r has to be even since $\mathcal{E}$ is skew-symmetric), then the form (5.2) is easily achieved as we have seen in the first step of the construction of the condensed form (4.3) which yields the form (4.4). Since in this form the matrix Δ is nonsingular and skew-symmetric, it is congruent to J_p as we have seen in the proof of Corollary 4.3. But in the form (5.3), we cannot decide whether the flow is symplectic, since the matrix $\tilde{\mathcal{A}}_{33}$ may be singular.

Based on the condensed form this decision is possible.

Theorem 5.2 *Consider a self-adjoint pair $(\mathcal{E},\mathcal{A})$ of sufficiently smooth matrix functions $\mathcal{E},\mathcal{A}\in C^0(\mathbb{I},\mathbb{R}^{n,n})$, and an associated DAE system of the form (3.13). If the constant rank conditions that allow the construction of the condensed form (4.3) hold, and the DAE (3.13) is regular then the underlying flow is symplectic.*

Proof. We may assume w.l.o.g. that the pair is in the condensed form (4.10) with the equations (4.11). Then the equations for z_3, z_4, z_5 are the algebraic constraints which can be solved by backward substitution (including differentiation).

The resulting ordinary differential equation can be rewritten as a linear Hamiltonian system

$$\begin{bmatrix} \dot z_1 \\ \dot z_2 \end{bmatrix} = -J_p \begin{bmatrix} \mathcal{A}_{11} & \mathcal{A}_{12} \\ \mathcal{A}_{21} & \mathcal{A}_{22} \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} + \begin{bmatrix} f_1 \\ f_2 \end{bmatrix},$$

that clearly generates a symplectic flow since the coefficient matrix on the right hand side is Hamiltonian. □

Applying Theorem 5.2 to the boundary value problem associated with the linear-quadratic optimal control problem, we thus immediately obtain that the underlying flow (if there is such a flow) is symplectic.

Example 5.3 [7] Consider the optimal control problem to minimize $\frac{1}{2}\int_0^1 x(t)^2\,dt$ subject to $\dot x = u,\ x(0)=1$. Then the necessary optimality condition is given by the boundary value problem

$$\begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \dot\lambda \\ \dot x \\ \dot u \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & -1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \lambda \\ x \\ u \end{bmatrix},\ x(0)=1,\ \lambda(1)=0, \tag{5.4}$$

which is a boundary value problem with a self-adjoint pair associated with a differentiation index 3 DAE which is already in the condensed form (4.8), where the second equation for x_2 is missing and hence there is no flow at all.

Example 5.4 [2] Consider the linear-quadratic control problem (1.1),(1.2) on $\mathbb{I}=[0,1]$ with coefficients

$$E = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix},\quad A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{bmatrix},\quad B = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix},\quad f = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix},$$
$$M_e = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix},\quad W = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix},\quad S = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix},\quad R = 0,$$

and the initial condition $x_1(0)=\alpha,\ x_2(0)=0$.

A simple calculation yields that $u=x_2=\lambda_2=x_3=\lambda_3=0$ combined with the Hamiltonian system $\dot x_1 = 0,\ -\dot\lambda_1 = 0,\ x_1(0)=\alpha,\ -\lambda_1(1)=x_1(1)$.

6 A global condensed form for self-adjoint DAEs

The staircase forms for self-adjoint pairs of matrix function as developed in Section 4 are based on series of constant rank assumptions. Similar to [19, Corollary 3.26], these results

are local in the sense that we may have restricted the interval $\mathbb{I}$. In the following we develop a global condensed form for self-adjoint pairs similar to that of [8] in the unstructured case. To derive this we need the following lemma.

Lemma 6.1 *Let $\mathcal{E} \in \mathbb{R}^{2p,2p}$ with $\mathcal{E} = -\mathcal{E}^T$. Then there exists an orthogonal symplectic matrix $U \in \mathbb{R}^{2p,2p}$ such that*

$$U^T \mathcal{E} U = \begin{bmatrix} 0 & \mathcal{E}_{12} \\ -\mathcal{E}_{12}^T & \mathcal{E}_{22} \end{bmatrix}$$

with $\mathcal{E}_{12} \in \mathbb{R}^{p,p}$, $\mathcal{E}_{22} \in \mathbb{R}^{p,p}$.

Proof. The proof is an immediate consequence of a corresponding factorization for skew-Hamiltonian matrices in [27]. □

Theorem 6.2 *Consider a self-adjoint pair $(\mathcal{E}, \mathcal{A})$ of sufficiently smooth matrix functions $\mathcal{E}, \mathcal{A} \in C^0(\mathbb{I}, \mathbb{R}^{n,n})$, and an associated DAE system of the form (3.13). Suppose that (3.13) has a well-defined differentiation index, and that the underlying flow associated with $2p$ differential equations is symplectic. Then there exists a matrix function $L \in C^1(\mathbb{I}, \mathbb{R}^{n,n})$ such that*

$$\tilde{\mathcal{E}} = L^T \mathcal{E} L = \begin{bmatrix} 0 & \mathcal{E}_{12} & 0 \\ -\mathcal{E}_{12}^T & \mathcal{E}_{22} & 0 \\ 0 & 0 & \mathcal{E}_{33} \end{bmatrix}, \quad \tilde{\mathcal{A}} = L^T \mathcal{A} L - L^T \mathcal{E} \dot{L} = \begin{bmatrix} 0 & -\dot{\mathcal{E}}_{12} & 0 \\ 0 & \mathcal{A}_{22} & \mathcal{A}_{23} \\ 0 & \mathcal{A}_{32} & \mathcal{A}_{33} \end{bmatrix}, \tag{6.1}$$

with $\mathcal{E}_{12}$ pointwise nonsingular, so that z_2 is uniquely determined from

$$\frac{d}{dt}(E_{12} z_2) = f_1,$$

and, furthermore,

$$\mathcal{E}_{33} \dot{z}_3 = \mathcal{A}_{32} z_2 + \mathcal{A}_{33} z_3 + f_3$$

has a unique solution z_3 for every sufficiently smooth inhomogeneity f_3 and given z_2.

Proof. The proof partly follows the lines of the proof of the corresponding result for unstructured pairs of matrix functions given in [8].

If the homogeneous equation

$$\mathcal{E} \dot{z} = \mathcal{A} z$$

has only the trivial solution, then the first two blocks are missing and the claim holds trivially by assumption. In any case, the solution space is finite dimensional. Let $\{\Phi_1, \ldots, \Phi_{2p}\}$ be a basis of the solution space and $\Phi = \begin{bmatrix} \Phi_1 & \cdots & \Phi_{2p} \end{bmatrix}$. Then

$$\operatorname{rank} \Phi(t) = 2p \quad \text{for all } t \in \mathbb{I}.$$

Hence, there exists a smooth, pointwise nonsingular matrix function U with

$$U^T \Phi = \begin{bmatrix} I_{2p} \\ 0 \end{bmatrix} \quad \text{for all } t \in \mathbb{I}.$$

Defining

$$\Phi' = U \begin{bmatrix} 0 \\ I_a \end{bmatrix}$$

with $a = n - 2p$ yields a pointwise nonsingular matrix function $Q = \begin{bmatrix} \Phi & \Phi' \end{bmatrix}$. Since $\mathcal{E}\dot{\Phi} = \mathcal{A}\Phi$, we obtain

$$(\tilde{\mathcal{E}}, \tilde{\mathcal{A}}) = (Q^T \mathcal{E} Q, Q^T \mathcal{A} Q - Q^T \mathcal{E} \dot{Q})$$

with

$$\tilde{\mathcal{E}} = \begin{bmatrix} \Phi^T \mathcal{E} \Phi & \Phi^T \mathcal{E} \Phi' \\ (\Phi')^T \mathcal{E} \Phi & (\Phi')^T \mathcal{E} \Phi' \end{bmatrix} = \begin{bmatrix} \mathcal{E}_{11} & \mathcal{E}_{12} \\ -\mathcal{E}_{12}^T & \mathcal{E}_{22} \end{bmatrix},$$
$$\tilde{\mathcal{A}} = \begin{bmatrix} \Phi^T (\mathcal{A}\Phi - \mathcal{E}\dot{\Phi}) & \Phi^T (\mathcal{A}\Phi' - \mathcal{E}\dot{\Phi}') \\ (\Phi')^T (\mathcal{A}\Phi - \mathcal{E}\dot{\Phi}) & (\Phi')^T (\mathcal{A}\Phi' - \mathcal{E}\dot{\Phi}') \end{bmatrix} = \begin{bmatrix} 0 & \mathcal{A}_{12} \\ 0 & \mathcal{A}_{22} \end{bmatrix},$$

and $\mathcal{E}_{11} = -\mathcal{E}_{11}^T \in C(\mathbb{I}, \mathbb{R}^{2p,2p})$. Here, $\mathcal{E}_1 := \begin{bmatrix} \mathcal{E}_{11} \\ -\mathcal{E}_{12}^T \end{bmatrix}$ has full column rank $2p$. To see this, suppose that $\operatorname{rank} \mathcal{E}_1(\hat{t}) < 2p$ for some $\hat{t} \in \mathbb{I}$. In this case, there would exist a vector $w \neq 0$ with $\mathcal{E}_1(\hat{t})w = 0$. Defining then

$$\tilde{f}(t) = \begin{cases} \frac{1}{t-\hat{t}} \mathcal{E}_1(t) w & \text{for } t \neq \hat{t}, \\ \frac{d}{dt}\left(\mathcal{E}_1(t) w\right) & \text{for } t = \hat{t}, \end{cases}$$

we have a smooth inhomogeneity $\tilde{f}$. The function z given by

$$z(t) = \begin{bmatrix} \log(|t - \hat{t}|) w \\ 0 \end{bmatrix}$$

then solves

$$\tilde{\mathcal{E}}(t)\dot{z} = \tilde{\mathcal{A}}(t) z + \tilde{f}(t)$$

on $\mathbb{I} \backslash \{\hat{t}\}$ in contradiction to the assumption of a well-defined differentiation index, which implies that local solutions can always be extended to a global solution on the entire interval $\mathbb{I}$.

Since $\tilde{\mathcal{E}}$ and $\tilde{\mathcal{A}}$ are obtained by a congruence transformation, the self-adjoint structure of the pair $(\mathcal{E}, \mathcal{A})$ is preserved which directly yields the conditions

$$\dot{\mathcal{E}}_{11} = 0 \quad \text{and} \quad \mathcal{A}_{12} = -\dot{\mathcal{E}}_{12},$$

i. e., the block $\mathcal{E}_{11}$ is constant. Hence, by Lemma 6.1, applied to $\mathcal{E}_{11}$, there exists an orthogonal symplectic $\tilde{U} \in \mathbb{R}^{2p,2p}$ such that

$$\hat{\mathcal{E}} = \begin{bmatrix} \tilde{U}^T & 0 \\ 0 & I \end{bmatrix} \tilde{\mathcal{E}} \begin{bmatrix} \tilde{U} & 0 \\ 0 & I \end{bmatrix} = \begin{bmatrix} 0 & \hat{\mathcal{E}}_{12} & \hat{\mathcal{E}}_{13} \\ -\hat{\mathcal{E}}_{12}^T & \hat{\mathcal{E}}_{22} & \hat{\mathcal{E}}_{23} \\ -\hat{\mathcal{E}}_{13}^T & -\hat{\mathcal{E}}_{23}^T & \hat{\mathcal{E}}_{33} \end{bmatrix},$$
$$\hat{\mathcal{A}} = \begin{bmatrix} \tilde{U}^T & 0 \\ 0 & I \end{bmatrix} \tilde{\mathcal{A}} \begin{bmatrix} \tilde{U} & 0 \\ 0 & I \end{bmatrix} = \begin{bmatrix} 0 & 0 & \hat{\mathcal{A}}_{13} \\ 0 & 0 & \hat{\mathcal{A}}_{23} \\ 0 & 0 & \hat{\mathcal{A}}_{33} \end{bmatrix},$$

with the conditions

$$\hat{\mathcal{A}}_{13} = -\dot{\hat{\mathcal{E}}}_{13}, \quad \hat{\mathcal{A}}_{23} = -\dot{\hat{\mathcal{E}}}_{23}, \quad \dot{\hat{\mathcal{E}}}_{12} = 0, \quad \dot{\hat{\mathcal{E}}}_{22} = 0,$$

and $\left[\begin{array}{cc} \hat{\mathcal{E}}_{12} & \hat{\mathcal{E}}_{13} \end{array}\right]$ has full row rank d. By Theorem 4.1 there exists a smooth, pointwise nonsingular matrix function V such that

$$\left[\begin{array}{cc} \hat{\mathcal{E}}_{12} & \hat{\mathcal{E}}_{13} \end{array}\right] V = \left[\begin{array}{cc} \bar{\mathcal{E}}_{12} & 0 \end{array}\right]$$

and thus

$$\left[\begin{array}{cc} I & 0 \\ 0 & V^T \end{array}\right] \hat{\mathcal{E}} \left[\begin{array}{cc} I & 0 \\ 0 & V \end{array}\right] = \left[\begin{array}{ccc} 0 & \bar{\mathcal{E}}_{12} & 0 \\ -\bar{\mathcal{E}}_{12}^T & \bar{\mathcal{E}}_{22} & \bar{\mathcal{E}}_{23} \\ 0 & -\bar{\mathcal{E}}_{23}^T & \bar{\mathcal{E}}_{33} \end{array}\right]$$

as well as

$$\left[\begin{array}{cc} I & 0 \\ 0 & V^T \end{array}\right] \hat{\mathcal{A}} \left[\begin{array}{cc} I & 0 \\ 0 & V \end{array}\right] - \left[\begin{array}{cc} I & 0 \\ 0 & V^T \end{array}\right] \hat{\mathcal{E}} \left[\begin{array}{cc} 0 & 0 \\ 0 & \dot{V} \end{array}\right] = \left[\begin{array}{ccc} 0 & \bar{\mathcal{A}}_{12} & \bar{\mathcal{A}}_{13} \\ 0 & \bar{\mathcal{A}}_{22} & \bar{\mathcal{A}}_{23} \\ 0 & \bar{\mathcal{A}}_{32} & \bar{\mathcal{A}}_{33} \end{array}\right]$$

where the block $\left[\begin{array}{cc} 0 & \bar{\mathcal{E}}_{12} \\ -\bar{\mathcal{E}}_{12}^T & \bar{\mathcal{E}}_{22} \end{array}\right]$ is invertible. ¿From the self-adjoint structure we have the condition

$$\left[\begin{array}{ccc} 0 & 0 & 0 \\ \bar{\mathcal{A}}_{12}^T & \bar{\mathcal{A}}_{22}^T & \bar{\mathcal{A}}_{32}^T \\ \bar{\mathcal{A}}_{13}^T & \bar{\mathcal{A}}_{23}^T & \bar{\mathcal{A}}_{33}^T \end{array}\right] = \left[\begin{array}{ccc} 0 & \bar{\mathcal{A}}_{12} & \bar{\mathcal{A}}_{13} \\ 0 & \bar{\mathcal{A}}_{22} & \bar{\mathcal{A}}_{23} \\ 0 & \bar{\mathcal{A}}_{32} & \bar{\mathcal{A}}_{33} \end{array}\right] + \left[\begin{array}{ccc} 0 & \dot{\bar{\mathcal{E}}}_{12} & 0 \\ -\dot{\bar{\mathcal{E}}}_{12}^T & \dot{\bar{\mathcal{E}}}_{22} & \dot{\bar{\mathcal{E}}}_{23} \\ 0 & -\dot{\bar{\mathcal{E}}}_{23}^T & \dot{\bar{\mathcal{E}}}_{33} \end{array}\right]$$

yielding that

$$\bar{\mathcal{A}}_{12} = -\dot{\bar{\mathcal{E}}}_{12} \quad \text{and} \quad \bar{\mathcal{A}}_{13} = 0.$$

Finally, defining the matrix

$$W^T = \left[\begin{array}{ccc} I & 0 & 0 \\ 0 & I & 0 \\ \bar{\mathcal{E}}_{23}^T \bar{\mathcal{E}}_{12}^{-1} & 0 & I \end{array}\right]$$

we get

$$W^T \left[\begin{array}{ccc} 0 & \bar{\mathcal{E}}_{12} & 0 \\ -\bar{\mathcal{E}}_{12}^T & \bar{\mathcal{E}}_{22} & \bar{\mathcal{E}}_{23} \\ 0 & -\bar{\mathcal{E}}_{23}^T & \bar{\mathcal{E}}_{33} \end{array}\right] W = \left[\begin{array}{ccc} 0 & \tilde{\mathcal{E}}_{12} & 0 \\ -\tilde{\mathcal{E}}_{12}^T & \tilde{\mathcal{E}}_{22} & 0 \\ 0 & 0 & \tilde{\mathcal{E}}_{33} \end{array}\right],$$

and

$$W^T \left[\begin{array}{ccc} 0 & -\dot{\bar{\mathcal{E}}}_{12} & 0 \\ 0 & \bar{\mathcal{A}}_{22} & \bar{\mathcal{A}}_{23} \\ 0 & \bar{\mathcal{A}}_{32} & \bar{\mathcal{A}}_{33} \end{array}\right] W - W^T \left[\begin{array}{ccc} 0 & \bar{\mathcal{E}}_{12} & 0 \\ -\bar{\mathcal{E}}_{12}^T & \bar{\mathcal{E}}_{22} & \bar{\mathcal{E}}_{23} \\ 0 & -\bar{\mathcal{E}}_{23}^T & \bar{\mathcal{E}}_{33} \end{array}\right] \dot{W} = \left[\begin{array}{ccc} 0 & -\dot{\tilde{\mathcal{E}}}_{12} & 0 \\ 0 & \tilde{\mathcal{A}}_{22} & \tilde{\mathcal{A}}_{23} \\ 0 & \tilde{\mathcal{A}}_{32} & \tilde{\mathcal{A}}_{33} \end{array}\right].$$

Thus, with

$$L = Q \left[\begin{array}{cc} U & 0 \\ 0 & I \end{array}\right] \left[\begin{array}{cc} I & 0 \\ 0 & V \end{array}\right] W$$

we have the form (6.1).

The associated differential-algebraic system can be written as

$$\begin{aligned} \tilde{\mathcal{E}}_{12}\dot{z}_2 + \dot{\tilde{\mathcal{E}}}_{12} z_2 &= \frac{d}{dt}(\tilde{\mathcal{E}}_{12} z_2) = \tilde{f}_1, \\ -\tilde{\mathcal{E}}_{12}^T \dot{z}_1 + \tilde{\mathcal{E}}_{22}\dot{z}_2 &= \tilde{\mathcal{A}}_{22} z_2 + \tilde{\mathcal{A}}_{23} z_3 + \tilde{f}_2, \\ \tilde{\mathcal{E}}_{33}\dot{z}_3 &= \tilde{\mathcal{A}}_{32} z_2 + \tilde{\mathcal{A}}_{33} z_3 + \tilde{f}_3. \end{aligned}$$

If the claim for the third equation does not hold then there exists a sufficiently smooth $\tilde{f}_3$ and z_2 such that the third equation possesses more than one solution. Consequently, the corresponding homogeneous equation $\tilde{\mathcal{E}}_{33}\dot{z}_3 = \tilde{\mathcal{A}}_{33}z_3$ possesses a non-trivial solution space. Together with the $2p$ degrees of freedom for the other two equations which are equivalent to ODEs due to the pointwise nonsingularity of $\tilde{\mathcal{E}}_{12}$ gives a solution space for the homogeneous DAE of dimension at least $2p+1$. But this contradicts the assumption on the dimension of the flow. □

7 Self-adjoint DAEs and derivative arrays

So far we have used global staircase forms to analyze DAE boundary value problems, but this is merely a theoretical result that is used for the analysis. In practice, to avoid the differentiation of numerically computed quantities, one applies a derivative array approach, see [8, 19], and determines a strangeness-free system of equations with the same solution set, where the equations describing the algebraic equations and those describing the dynamical system are separated.

But if one does this for a self-adjoint pair of matrix functions, then unfortunately the self-adjoint structure is destroyed, since the transformations are only applied from the left. It is the purpose of this section to discuss necessary modifications and their numerical costs if we want to retrieve self-adjointness.

Ignoring the structure, from the derivative array of the system (3.13) using Hypothesis 2.1 we obtain matrix functions Z_1, Z_2 and T_2 such that

$$\hat{\mathcal{A}}_2 = Z_2^T N_\mu \begin{bmatrix} I & 0 & \cdots & 0 \end{bmatrix}^T$$

has full row rank a and $\hat{\mathcal{A}}_2 T_2 = 0$. We now consider the overdetermined system

$$\begin{aligned} \mathcal{E}\dot{z} &= \mathcal{A}z + f, \\ 0 &= \hat{\mathcal{A}}_2 z + \hat{f}_2, \end{aligned}$$

where $\hat{f}_2 = Z_2^T g_\mu$. Choosing T_2' such that the matrix $T = \begin{bmatrix} T_2 & T_2' \end{bmatrix}$ is orthogonal we get

$$\begin{aligned} T^T\mathcal{E}T\dot{\tilde{z}} &= T^T\mathcal{A}T\tilde{z} - T^T\mathcal{E}\dot{T}\tilde{z} + T^T f, \\ 0 &= T^T\hat{\mathcal{A}}_2 T\tilde{z} + T^T\hat{f}_2, \end{aligned}$$

with $z = T\tilde{z}$, which yields

$$\begin{bmatrix} \mathcal{E}_{11} & \mathcal{E}_{12} \\ -\mathcal{E}_{12}^T & \mathcal{E}_{22} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{\tilde{z}}_1 \\ \dot{\tilde{z}}_2 \end{bmatrix} = \begin{bmatrix} \mathcal{A}_{11} & \mathcal{A}_{12} \\ \mathcal{A}_{21} & \mathcal{A}_{22} \\ 0 & \hat{\mathcal{A}}_{22} \end{bmatrix} \begin{bmatrix} \tilde{z}_1 \\ \tilde{z}_2 \end{bmatrix} + \begin{bmatrix} \tilde{f}_1 \\ \tilde{f}_2 \\ \tilde{f}_3 \end{bmatrix},$$

with

$$T^T\mathcal{E}T = \begin{bmatrix} \mathcal{E}_{11} & \mathcal{E}_{12} \\ -\mathcal{E}_{12}^T & \mathcal{E}_{22} \end{bmatrix}, \quad T^T\mathcal{A}T - T^T\mathcal{E}\dot{T} = \begin{bmatrix} \mathcal{A}_{11} & \mathcal{A}_{12} \\ \mathcal{A}_{21} & \mathcal{A}_{22} \end{bmatrix}, \quad T^T \begin{bmatrix} f \\ \hat{f}_2 \end{bmatrix} = \begin{bmatrix} \tilde{f}_1 \\ \tilde{f}_2 \\ \tilde{f}_3 \end{bmatrix}$$

and $\mathcal{E}_{11}$ and $\hat{\mathcal{A}}_{22}$ nonsingular. Removing the second block row and eliminating the entries belonging to $\tilde{z}_2$ yields

$$\begin{bmatrix} \mathcal{E}_{11} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{\tilde{z}}_1 \\ \dot{\tilde{z}}_2 \end{bmatrix} = \begin{bmatrix} \mathcal{A}_{11} & 0 \\ 0 & \hat{\mathcal{A}}_{22} \end{bmatrix} \begin{bmatrix} \tilde{z}_1 \\ \tilde{z}_2 \end{bmatrix} + \begin{bmatrix} \bar{f}_1 \\ \bar{f}_2 \end{bmatrix},$$

where

$$\begin{bmatrix} \bar{f}_1 \\ \bar{f}_2 \end{bmatrix} = \begin{bmatrix} \tilde{f}_1 - \mathcal{A}_{12}\hat{\mathcal{A}}_{22}^{-1}\tilde{f}_3 + \mathcal{E}_{12}\frac{d}{dt}(\hat{\mathcal{A}}_{22}^{-1}\tilde{f}_3) \\ \tilde{f}_3 \end{bmatrix}.$$

Note that the values of $\bar{f}_1, \tilde{z}_2$ can be obtained pointwise by solving the corresponding algebraic equations. To obtain the derivatives of $\tilde{z}_2$, we can differentiate the third equation and solve for $\tilde{z}_2$ pointwise. Further, note that $\hat{\mathcal{A}}_{22}$ can be computed in such a way that the matrix function is smooth, see [16].

The first equation, which has the form

$$T_2^T \tfrac{d}{dt}(\mathcal{E}T_2\tilde{z}_1) = T_2^T \mathcal{A}T_2\tilde{z}_1 + \bar{f}_1$$

then gives the subsystem which can be reformulated as a Hamiltonian system and which has a symplectic flow. All quantities can be obtained from the derivative array and by differentiation. Note that the formulation of $\bar{f}_1$ requires the computation of the derivative $\dot{T}$ of the transformation matrix T, for example by means according to [19, Corollary 3.10].

8 Nonlinear DAEs with self-adjoint linearization

In this section we consider the optimality system arising in nonlinear optimal control problems

$$\mathcal{J}(x,u) = \mathcal{M}(x(\bar{t})) + \int_{\underline{t}}^{\bar{t}} \mathcal{K}(t,x(t),u(t))\ dt = \min! \tag{8.1}$$

subject to a constraint

$$F(t,x,u,\dot{x}) = 0 \tag{8.2}$$

and

$$x(\underline{t}) = \underline{x}. \tag{8.3}$$

We assume that $F \in C^0(\mathbb{I} \times \mathbb{D}_x \times \mathbb{D}_u \times \mathbb{D}_{\dot{x}}, \mathbb{R}^l)$ is sufficiently smooth, that $\mathbb{I} = [\underline{t}, \bar{t}] \subseteq \mathbb{R}$ is a (compact) interval, and that $\mathbb{D}_x, \mathbb{D}_{\dot{x}} \subseteq \mathbb{R}^n$, $\mathbb{D}_u \subseteq \mathbb{R}^m$ are open sets.

We will analyze, whether some of the self-adjointness properties can be found in this case as well. We briefly recall the structure of the necessary optimality conditions from [20]. We again use derivative arrays, which take the form

$$F_\ell(t,z,\dot{z},\ldots,z^{(\ell+1)}) = 0, \tag{8.4}$$

with $z = [x^T, u^T]^T$, which stacks the original equation and all its derivatives up to level ℓ in one large system.

Here, partial derivatives of F_ℓ with respect to selected variables p from $(t,z,\dot{z},\ldots,z^{(\ell+1)})$ are denoted by $F_{\ell;p}$. The solution set of the nonlinear algebraic equation associated with the derivative array F_μ for some integer μ is denoted by

$$\mathbb{L}_\mu = \{z_\mu \in \mathbb{I} \times \mathbb{R}^{n+m} \times \mathbb{R}^{n+m} \times \ldots \times \mathbb{R}^{n+m} \mid F_\mu(z_\mu) = 0\} \tag{8.5}$$

and the hypothesis takes the following form, see [19].

Hypothesis 8.1 *Consider the general system of nonlinear differential-algebraic equations (8.2). There exist integers μ, r, a, d, and v such that $\mathbb{L}_\mu$ is not empty and such that for every $z_\mu^0 = (t_0, z_0, \dot{z}_0, \ldots, z_0^{(\mu+1)}) \in \mathbb{L}_\mu$ there exists a (sufficiently small) neighborhood in which the following properties hold:*

1. *The set $\mathbb{L}_\mu \subseteq \mathbb{R}^{(\mu+2)(n+m)+1}$ forms a manifold of dimension $(\mu+2)(n+m)+1-r$.*
2. *We have $\operatorname{rank} F_{\mu;z,\dot{z},\ldots,z^{(\mu+1)}} = r$ on $\mathbb{L}_\mu$.*
3. *We have $\operatorname{corank} F_{\mu;z,\dot{z},\ldots,z^{(\mu+1)}} - \operatorname{corank} F_{\mu-1;z,\dot{z},\ldots,z^{(\mu)}} = v$ on $\mathbb{L}_\mu$, where the corank is the dimension of the corange and the convention is used that $\operatorname{corank} F_{-1;z} = 0$.*
4. *We have $\operatorname{rank} F_{\mu;\dot{z},\ldots,z^{(\mu+1)}} = r - a$ on $\mathbb{L}_\mu$ such that there exist smooth full rank matrix functions Z_2 and T_2 of size $(\mu+1)l \times a$ and $(n+m) \times (n+m-a)$, respectively, satisfying*

$$Z_2^T F_{\mu;\dot{z},\ldots,z^{(\mu+1)}} = 0, \quad \operatorname{rank} Z_2^T F_{\mu;z} = a, \quad Z_2^T F_{\mu;z} T_2 = 0 \tag{8.6}$$

   on $\mathbb{L}_\mu$.
5. *We have $\operatorname{rank} F_{\dot{z}} T_2 = d = l - a - v$ on $\mathbb{L}_\mu$ such that there exists a smooth full rank matrix function Z_1 of size $(n+m) \times d$ satisfying $\operatorname{rank} Z_1^T F_{\dot{z}} T_2 = d$.*

Again, the smallest possible μ for which Hypothesis 8.1 is valid is called the *strangeness index* of (8.2). It has been shown in [18] that Hypothesis 8.1 implies locally (via the implicit function theorem) the existence of a *reduced system* given by

$$\begin{aligned} &\text{(a)}\ \hat{F}_1(t, z_1, z_2, z_3, \dot{z}_1, \dot{z}_2, \dot{z}_3) = 0, \\ &\text{(b)}\qquad\quad \hat{F}_2(t, z_1, z_2, z_3) = 0, \end{aligned} \tag{8.7}$$

with $\hat{F}_1 = Z_1^T F$, where $(z_1, z_2, z_3) \in \mathbb{R}^d \times \mathbb{R}^{n+m-a-d} \times \mathbb{R}^a$ is a suitable splitting of the unknown z. Part 4 of Hypothesis 8.1 guarantees that equation (8.7b) can be solved for z_3 according to $z_3 = \mathcal{R}(t, z_1, z_2)$. Eliminating z_3 and $\dot{z}_3$ in (8.7a) with the help of this relation and its derivative then leads to

$$\hat{F}_1(t, z_1, z_2, \mathcal{R}(t, z_1, z_2), \dot{z}_1, \dot{z}_2, \mathcal{R}_t(t, z_1, z_2) + \mathcal{R}_{z_1}(t, z_1, z_2)\dot{z}_1 + \mathcal{R}_{z_2}(t, z_1, z_2)\dot{z}_2) = 0.$$

By part 5 of Hypothesis 8.1 we may assume without loss of generality that this system can (locally) be solved for $\dot{z}_1$ leading to the system

$$\begin{aligned} \dot{z}_1 &= \mathcal{L}(t, z_1, z_2, \dot{z}_2), \\ z_3 &= \mathcal{R}(t, z_1, z_2). \end{aligned} \tag{8.8}$$

Obviously, in this system, interpreted as a DAE, $z_2 \in C^1(\mathbb{I}, \mathbb{R}^{n+m-a-d})$ can be chosen arbitrarily (at least when staying in the domain of definition of $\mathcal{R}$ and $\mathcal{L}$), while the resulting system has locally a unique solution for z_1 and z_3, provided a consistent initial condition is given. This means that z_2 can be interpreted as a control. The quantity v, which has not been addressed yet, measures the number of equations in the original system that give rise to trivial equations $0 = 0$, i. e., it counts the number of redundancies in the system. Together with a and d it gives a complete classification of the l equations into d differential equations,

a algebraic equations and v trivial equations. Of course, trivial equations can be simply removed without altering the solution set.

If the variable z is a combined vector of states and controls, then, since (8.7) consists of original variables, these can again be split into parts stemming from x and from u. It has been shown in [18, 22], see also [19], how this system then can be treated.

With this preliminaries, it has then been shown in [20] that the necessary optimality conditions are given by the following boundary value problem

$$\begin{aligned}
&\text{(a)} \quad \dot{x}_1 = \mathcal{L}(t, x_1, u), \;\; x_1(\underline{t}) = \underline{x}_1, \\
&\text{(b)} \quad x_2 = \mathcal{R}(t, x_1, u), \\
&\text{(c)} \quad \dot{\lambda}_1 = \mathcal{K}_{x_1}(t, x_1, x_2, u)^T - \mathcal{L}_{x_1}(t, x_1, x_2, u)^T \lambda_1 - \mathcal{R}_{x_1}(t, x_1, u)^T \lambda_1, \\
&\qquad\qquad\qquad\qquad\qquad\qquad \lambda_1(\bar{t}) = -\mathcal{M}_{x_1}(x_1(\bar{t}), x_2(\bar{t}))^T, \\
&\text{(d)} \quad 0 = \mathcal{K}_{x_2}(t, x_1, x_2, u)^T + \lambda_2, \\
&\text{(e)} \quad 0 = \mathcal{K}_u(t, x_1, x_2, u)^T - \mathcal{L}_u(t, x_1, u)^T \lambda_1 - \mathcal{R}_u(t, x_1, u)^T \lambda_2, \\
&\text{(f)} \quad \gamma = \lambda_1(\underline{t}).
\end{aligned} \tag{8.9}$$

Note that the necessary equations are linear with respect to λ. This follows from the general result of Ljusternik [23] where the Lagrangian is a linear form appearing additively in the necessary conditions. Linearizing with respect to the other unknowns yields a linear DAE of the form (1.3) with the replacements

$$\begin{aligned}
E &= \begin{bmatrix} I_d & 0 \\ 0 & 0 \end{bmatrix}, \\
A &= \begin{bmatrix} \mathcal{L}_{x_1}(t, x_1, u) & 0 \\ \mathcal{R}_{x_1}(t, x_1, u) & -I \end{bmatrix}, \; B = \begin{bmatrix} \mathcal{L}_u(t, x_1, u) \\ \mathcal{R}_u(t, x_1, u) \end{bmatrix}, \\
W &= \begin{bmatrix} \mathcal{K}_{x_1,x_1}(t, x_1, x_2, u) & \mathcal{K}_{x_1,x_2}(t, x_1, x_2, u) \\ \mathcal{K}_{x_2,x_1}(t, x_1, x_2, u) & \mathcal{K}_{x_2,x_2}(t, x_1, x_2, u) \end{bmatrix}, \; S = \begin{bmatrix} \mathcal{K}_{x_1,u}(t, x_1, x_2, u) \\ \mathcal{K}_{x_2,u}(t, x_1, x_2, u) \end{bmatrix}, \\
R &= \mathcal{K}_{u,u}(t, x_1, x_2, u).
\end{aligned}$$

Hence, linearization gives a self-adjoint DAE possessing a Hamiltonian subsystem as in the linear time-varying case.

9 Conclusion

We have studied the properties of the necessary optimality systems arising from optimal control problems for differential-algebraic systems. We have shown that the system is self-conjugate with the coefficients forming a self-adjoint pair of matrix functions. We have derived (under some constant rank assumptions) condensed forms under congruence transformations with orthogonal matrix functions and also shown that then there always exists a Hamiltonian subsystem with a symplectic flow. We have discussed that the Hamiltonian subsystem also can be obtained from the derivative array and that similar structures can be achieved locally in the nonlinear case.

References

[1] H. Abou-Khandil, G. Freiling, V. Ionescu, and G. Jank. *Matrix Riccati Equations in Control and Systems Theory.* Birkhäuser, Basel, Switzerland, 2000.

[2] A. Backes. *Optimale Steuerung der linearen DAE im Fall Index 2.* Dissertation, Mathematisch-Naturwissenschaftliche Fakultät, Humboldt-Universität zu Berlin, Berlin, Germany, 2006.

[3] K. Balla and V. H. Linh. Adjoint pairs of differential-algebraic equations and Hamiltonian systems. *Appl. Numer. Math.*, 53:131–148, 2005.

[4] P. Benner, R. Byers, V. Mehrmann, and H. Xu. A robust numerical method for the γ-iteration in $\mathcal{H}_\infty$-control. *Lin. Alg. Appl.*, 425:548–570, 2007.

[5] K. E. Brenan, S. L. Campbell, and L. R. Petzold. *Numerical Solution of Initial-Value Problems in Differential Algebraic Equations.* SIAM Publications, Philadelphia, PA, 2nd edition, 1996.

[6] J. R. Bunch. A note on the stable decomposition of skew-symmetric matrices. *Math. Comp.*, 38:475–479, 1982.

[7] R. Byers, V. Mehrmann, and H. Xu. A structured staircase algorithm for skew-symmetric/symmetric pencils. *Electr. Trans. Num. Anal.*, 26:1–33, 2007.

[8] S. L. Campbell. A general form for solvable linear time varying singular systems of differential equations. *SIAM J. Math. Anal.*, 18:1101–1115, 1987.

[9] S. L. Campbell, N. K. Nichols, and W. J. Terrell. Duality, observability, and controllability for linear time-varying descriptor systems. *Circ. Syst. Signal Process.*, 10:455–470, 1991.

[10] V. Doležal. The existence of a continuous basis of a certain subspace of E_r which depends on a parameter. *Cas. Pro. Pest. Mat.*, 89:466–468, 1964.

[11] E. Griepentrog and R. März. *Differential-Algebraic Equations and their Numerical Treatment.* Teubner Verlag, Leipzig, Germany, 1986.

[12] G. Grubb. *Distributions and Operators.* Springer, New York, 2009.

[13] E. Hairer, C. Lubich, and G. Wanner. *Geometric Numerical Integration. Structure-Preserving Algorithms for Ordinary Differential Equations.* Springer-Verlag, Berlin, Germany, 2002.

[14] H. Heuser. *Funktionalanalysis.* B. G. Teubner, Stuttgart, 3rd edition, 1992.

[15] H. W. Knobloch and H. Kwakernaak. *Lineare Kontrolltheorie.* Springer-Verlag, Berlin, Germany, 1985.

[16] P. Kunkel and V. Mehrmann. Smooth factorizations of matrix valued functions and their derivatives. *Numer. Math.*, 60:115–132, 1991.

[17] P. Kunkel and V. Mehrmann. Canonical forms for linear differential-algebraic equations with variable coefficients. *J. Comput. Appl. Math.*, 56:225–259, 1994.

[18] P. Kunkel and V. Mehrmann. Analysis of over- and underdetermined nonlinear differential-algebraic systems with application to nonlinear control problems. *Math. Control, Signals, Sys.*, 14:233–256, 2001.

[19] P. Kunkel and V. Mehrmann. *Differential-Algebraic Equations. Analysis and Numerical Solution*. EMS Publishing House, Zürich, Switzerland, 2006.

[20] P. Kunkel and V. Mehrmann. Optimal control for unstructured nonlinear differential-algebraic equations of arbitrary index. *Math. Control, Signals, Sys.*, 20:227–269, 2008.

[21] P. Kunkel and V. Mehrmann. Formal adjoints of linear DAE operators and their role in optimal control. Preprint 11/2011, Inst. f. Mathematik, TU Berlin, Berlin, Germany, 2011. `url: http://www.math.tu-berlin.de/preprints/`, submitted for publication.

[22] P. Kunkel, V. Mehrmann, and W. Rath. Analysis and numerical solution of control problems in descriptor form. *Math. Control, Signals, Sys.*, 14:29–61, 2001.

[23] L. Ljusternik. On constrained extrema of functionals. *Math. Sb.*, 41:390–401, 1934. In Russian.

[24] P. Losse, V. Mehrmann, L.K. Poppe, and T. Reis. The modified optimal $\mathcal{H}_\infty$ control problem for descriptor systems. *SIAM J. Cont.*, 2795–2811:47, 2008.

[25] V. Mehrmann. *The Autonomous Linear Quadratic Control Problem*. Springer-Verlag, Berlin, Germany, 1991.

[26] R. E. O'Malley. *Singular Perturbation Methods for Ordinary Differential Equations*. Springer-Verlag, New York, NY, 1991.

[27] C.C. Paige and C.F. Van Loan. A Schur decomposition for Hamiltonian matrices. *Lin. Alg. Appl.*, 14:11–32, 1981.

[28] J. W. Polderman and J. C. Willems. *Introduction to Mathematical Systems Theory: A Behavioural Approach*. Springer-Verlag, New York, NY, 1998.

[29] R. C. Thompson. The characteristic polynomial of a principal submatrix of a Hermitian pencil. *Lin. Alg. Appl.*, 14:135–177, 1976.

[30] R. C. Thompson. Pencils of complex and real symmetric and skew matrices. *Lin. Alg. Appl.*, 147:323–371, 1991.

[31] P. Van Dooren. The computation of Kronecker's canonical form of a singular pencil. *Lin. Alg. Appl.*, 27:103–141, 1979.