FORMAL ADJOINTS OF LINEAR DAE OPERATORS AND THEIR ROLE IN OPTIMAL CONTROL *

PETER KUNKEL † AND VOLKER MEHRMANN ‡

Abstract. For regular strangeness-free linear differential-algebraic equations (DAEs) the definition of an adjoint DAE is straightforward. This definition can be formally extended to general linear DAEs. In this paper, we analyze the properties of the formal adjoints and their implications in solving linear-quadratic optimal control problems with DAE constraints.

Key words. Differential-algebraic equation, adjoint operator, adjoint pair, formal adjoint pair, optimal control, necessary optimality condition, formal necessary optimality condition.

AMS subject classifications. 93C10, 93C15, 93B52, 65L80, 49K15, 34H05.

1. Introduction.

Consider linear differential-algebraic equations (DAEs) of the form

$$E\dot{x} = Ax + f, \tag{1.1}$$

where (omitting obvious arguments in the functions) $E \in C^0(\mathbb{I}, \mathbb{R}^{n,n})$, $A \in C^0(\mathbb{I}, \mathbb{R}^{n,n})$, and $f \in C^0(\mathbb{I}, \mathbb{R}^n)$. In order to introduce the concept of an adjoint (linear) DAE associated with (1.1), we must formulate (1.1) as an operator equation in appropriate Banach spaces as part of appropriate dual systems, see, e. g., [5]. To obtain a suitable Banach space formulation, we replace (1.1) by a so-called *strangeness-free formulation*

$$\hat{E}\dot{x} = \hat{A}x + \hat{f}, \tag{1.2}$$

where

$$\hat{E} = \begin{bmatrix} \hat{E}_1 \\ 0 \end{bmatrix}, \quad \hat{A} = \begin{bmatrix} \hat{A}_1 \\ \hat{A}_2 \end{bmatrix}, \quad \hat{f} = \begin{bmatrix} \hat{f}_1 \\ \hat{f}_2 \end{bmatrix}$$

with the additional property that

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

is (pointwise) nonsingular, see [7, Sec. 3.4]. Note that this is always possible under suitable regularity assumptions.

In this way, we get an adjoint equation of the form

$$-\hat{E}^T\dot{\lambda} = (\hat{A} + \tfrac{d}{dt}\hat{E})^T\lambda + h, \tag{1.3}$$

where $h \in C^0(\mathbb{I}, \mathbb{R}^n)$ denotes a corresponding inhomogeneity. Accordingly, $(-\hat{E}^T, (\hat{A} + \frac{d}{dt}\hat{E})^T)$ is called the *adjoint pair* of $(\hat{E}, \hat{A})$. Although this motivation is in general not valid for the pair (E, A) of (1.1), see [11, 12], one can formally define $(-E^T, (A + \dot{E})^T)$ as the adjoint pair of (E, A). We therefore call $(-E^T, (A + \dot{E})^T)$ the *formal adjoint* of (E, A).

Adjoint equations typically arise also in the context of linear-quadratic optimal control problems. In the case of DAEs these consist of

$$\mathcal{J}(x, u) = \frac{1}{2}x(\bar{t})^T M x(\bar{t}) + \frac{1}{2}\int_{\underline{t}}^{\bar{t}} (x^T W x + 2x^T S u + u^T R u)\, dt = \min!, \tag{1.4}$$

*Supported through the Research-in-Pairs Program at *Mathematisches Forschungsinstitut Oberwolfach.*

†Mathematisches Institut, Universität Leipzig, Johannisgasse 26, D-04009 Leipzig, Fed Rep. Germany, kunkel@math.uni-leipzig.de. Supported by *Deutsche Forschungsgemeinschaft* under grant no. KU964/7-1.

‡ Institut für Mathematik, MA 4-5, Technische Universität Berlin, D-10623 Berlin, Fed Rep. Germany, mehrmann@math.tu-berlin.de. Supported by *Deutsche Forschungsgemeinschaft* through MATHEON, the DFG Research Center "Mathematics for Key Technologies" in Berlin.

where $W \in C^0(\mathbb{I}, \mathbb{R}^{n,n})$, $S \in C^0(\mathbb{I}, \mathbb{R}^{n,m})$, $R \in C^0(\mathbb{I}, \mathbb{R}^{m,m})$, $M \in \mathbb{R}^{n,n}$, $\mathbb{I} = [\underline{t}, \overline{t}]$, with (pointwise) symmetric W, R, and M, subject to the constraint

$$E\dot{x} = Ax + Bu + f, \quad x(\underline{t}) = \underline{x}, \tag{1.5}$$

where $B \in C^0(\mathbb{I}, \mathbb{R}^{n,m})$. As before, the DAE (1.5) should be replaced by a strangeness-free formulation

$$\hat{E}\dot{x} = \hat{A}x + \hat{B}u + \hat{f}, \tag{1.6}$$

where

$$\hat{E} = \begin{bmatrix} \hat{E}_1 \\ 0 \end{bmatrix}, \quad \hat{A} = \begin{bmatrix} \hat{A}_1 \\ \hat{A}_2 \end{bmatrix}, \quad \hat{B} = \begin{bmatrix} \hat{B}_1 \\ \hat{B}_2 \end{bmatrix}, \quad \hat{f} = \begin{bmatrix} \hat{f}_1 \\ \hat{f}_2 \end{bmatrix}$$

with the additional property that

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

has (pointwise) full row rank. Again, this is possible under suitable regularity assumptions, see [8].

If we replace the DAE in (1.5) by (1.6) in the optimal control problem, then it has been shown in [8] that the corresponding *necessary optimality conditions* for an optimal solution (x, u) state that there exists a Lagrange multiplier λ such that x, u, λ satisfy the boundary value problem

$$\begin{array}{lll} \text{(a)} & \hat{E}\dot{x} = \hat{A}x + \hat{B}u + \hat{f}, & \hat{E}_1(\underline{t})x(\underline{t}) - \hat{E}_1(\underline{t})\underline{x} = 0, \\ \text{(b)} & -\hat{E}^T\dot{\lambda} = Wx + Su + (\hat{A} + \dot{\hat{E}})^T\lambda, & \hat{E}(\overline{t})^T\lambda(\overline{t}) - Mx(\overline{t}) = 0, \\ \text{(c)} & 0 = S^Tx + Ru + \hat{B}^T\lambda, & \end{array} \tag{1.7}$$

provided that the initial condition is consistent according to $\hat{E}_1(\underline{t})^+\hat{E}_1(\underline{t})\underline{x} = \underline{x}$ and that range $M \subseteq$ cokernel $E(\overline{t})$. Here $\hat{E}_1(\underline{t})^+$ denotes the Moore-Penrose inverse of $\hat{E}_1(\underline{t})$, see, e. g. [4]. We should mention here that for this formulation of the necessary conditions we assume sufficient smoothness of the data in order to concentrate on the structure of the equations. We also changed the sign of λ compared with [8], for reasons that will become clear later.

Note that the DAE (1.2) and its adjoint DAE (1.3) with $h = 0$ appear in (1.7) if we omit terms belonging to the cost functional (1.4). Moreover, combining (1.2) and (1.3) yields the pair

$$\left(\begin{bmatrix} 0 & \hat{E} \\ -\hat{E}^T & 0 \end{bmatrix}, \begin{bmatrix} 0 & \hat{A} \\ (\hat{A} + \frac{d}{dt}\hat{E})^T & 0 \end{bmatrix} \right) \tag{1.8}$$

of matrix functions, which is *self-adjoint* in the obvious sense that it equals its adjoint. Finally, the pair

$$\left(\begin{bmatrix} 0 & \hat{E} & 0 \\ -\hat{E}^T & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & \hat{A} & \hat{B} \\ (\hat{A} + \frac{d}{dt}\hat{E})^T & W & S \\ \hat{B}^T & S^T & R \end{bmatrix} \right) \tag{1.9}$$

of matrix functions presenting the coefficient functions in the boundary value problem (1.7) is self-adjoint as well. This self-adjointness is reflected by the self-conjugacy of an associated Banach space operator, see [9].

Analogous to the case of the formal adjoint, one may also consider the so-called *formal necessary conditions*

$$\begin{array}{lll} \text{(a)} & E\dot{x} = Ax + Bu + f, & E(\underline{t})x(\underline{t}) - E(\underline{t})\underline{x} = 0, \\ \text{(b)} & -E^T\dot{\lambda} = Wx + Su + (A + \dot{E})^T\lambda, & E(\overline{t})^T\lambda(\overline{t}) - Mx(\overline{t}) = 0, \\ \text{(c)} & 0 = S^Tx + Ru + B^T\lambda, & \end{array} \tag{1.10}$$

It has been shown, see [1, 10], that if (1.10) is uniquely solvable and the cost functional is positive semidefinite, then surprisingly the part (x, u) of the solution actually is a solution of the optimal control problem.

The aim of this paper is to give more insight into the properties of the formal adjoint and the formal necessary conditions. In particular, we show that if the DAE associated with (E, A) has a well-defined differentiation index ν (see [2] for a definition), then the DAE associated with the formal adjoint pair also has a well-defined differentiation index ν. On the basis of this result, we analyze in detail how the solutions of the formal necessary conditions (1.10) are related to the solutions of the necessary conditions (1.7), which for convenience we address as *true* necessary conditions in the remainder of this paper.

These results also explain the case that the formal necessary conditions fail to have a solution while there is a solution of the true necessary conditions. They also indicate in which way we can modify the formal necessary conditions to have (up to some smoothness requirements) the same solution properties as for the true necessary conditions. We also discuss how these results can be used to numerically solve problems where the DAE in the true necessary conditions is not strangeness-free.

The paper is organized as follows. In Section 2 we introduce the notation and present some preliminary results. Section 3 characterizes the properties of the formal adjoint DAE. These results are then used in Section 4 to analyze the properties of the formal necessary conditions. We finish with some conclusions in Section 5.

2. Preliminaries. To study optimal control problems with DAE constraints as discussed in the introduction, we need to assume some regularity of the pairs of matrix functions under considerations. Since we look at two different pairs, namely (E, A) for the formal adjoint and $([E\,0], [A\,B])$ for the constraint in the optimal control problem, we introduce all assumptions and notation for the second case. We then only need to drop the block which belongs to the variable u to specialize to the first case.

Introducing the so-called behavior formulation, cf. [13], by setting

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

we can write the given DAE (1.5) as

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

Since solutions of DAEs may depend on derivatives of all the data, we follow an idea of [3] and use the so-called derivative array systems

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

where

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

requiring here and in the following that all functions are sufficiently smooth. Moreover, we now turn to the more general situation of complex-valued matrix functions. The main reason for this is that the canonical form we use in the proofs requires complex-valued transformations, see Theorem 2.3 below. Note that all results will contain the real result as special case.

The central regularity assumptions then read as follows.

HYPOTHESIS 2.1. *There exist integers μ, d, and a, such that the pair (M_μ, N_μ) in (2.3) has the following properties:*

1. *For all* $t \in \mathbb{I}$ *we have* $\operatorname{rank} M_\mu(t) = (\mu+1)n - a$. *This implies the existence of a smooth matrix function* Z_2 *of size* $((\mu+1)n, a)$ *and pointwise maximal rank satisfying* $Z_2^H M_\mu = 0$ *on* $\mathbb{I}$.
2. *For all* $t \in \mathbb{I}$ *we have* $\operatorname{rank} Z_2(t)^H N_\mu(t)[I_{n+m}\, 0\, \cdots\, 0]^H = a$. *This implies the existence of a smooth matrix function* T_2 *of size* $(n+m, d)$, $d = n - a$, *and pointwise maximal rank satisfying* $Z_2^H N_\mu [I_{n+m}\, 0\, \cdots\, 0]^H T_2 = 0$ *on* $\mathbb{I}$.
3. *For all* $t \in \mathbb{I}$ *we have* $\operatorname{rank} \mathcal{E}(t) T_2(t) = d$. *This implies the existence of a smooth matrix function* Z_1 *of size* (n, d) *and pointwise maximal rank satisfying* $\operatorname{rank} Z_1^H \mathcal{E} = d$ *on* $\mathbb{I}$.

The strangeness-free formulation in (1.2) then has the coefficients

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

where $V = [\, I_{n+m}\ 0\ \cdots\ 0\,]^H$.

For a linear DAE as in (1.5), scaling of the equation and a change of basis for the unknowns defines an equivalence relation for the pairs of coefficient functions.

DEFINITION 2.2. *Two pairs* $(\mathcal{E}, \mathcal{A})$ *and* $(\tilde{\mathcal{E}}, \tilde{\mathcal{A}})$ *of matrix function* $\mathcal{E}, \mathcal{A}, \tilde{\mathcal{E}}, \tilde{\mathcal{A}} \in C(\mathbb{I}, \mathbb{C}^{n,n+m})$ *are called globally equivalent iff there exist pointwise nonsingular matrix functions* $P \in C(\mathbb{I}, \mathbb{C}^{n,n})$ *and* $Q \in C^1(\mathbb{I}, \mathbb{C}^{n+m,n+m})$ *such that*

$$\tilde{\mathcal{E}} = P\mathcal{E}Q, \quad \tilde{\mathcal{A}} = P\mathcal{A}Q - P\mathcal{E}\dot{Q}. \tag{2.4}$$

We then write

$$(\mathcal{E}, \mathcal{A}) \sim (\tilde{\mathcal{E}}, \tilde{\mathcal{A}}).$$

A suitable canonical form under global equivalence is then given by the following theorem, see [6].

THEOREM 2.3. *Hypothesis 2.1 holds for the pair of matrix functions* $(\mathcal{E}, \mathcal{A})$ *with* $\mathcal{E}, \mathcal{A} \in C(\mathbb{I}, \mathbb{C}^{n,n+m})$ *if and only if*

$$(\mathcal{E}, \mathcal{A}) \sim \left(\begin{bmatrix} I_d & H & 0 \\ 0 & G & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & L \\ 0 & I_a & 0 \end{bmatrix} \right), \tag{2.5}$$

where the matrix functions G, H, L *are of corresponding sizes and* G *has the property that the DAE*

$$G\dot{z}_2 = z_2 + f_2 \tag{2.6}$$

is uniquely solvable for every sufficiently smooth inhomogeneity f_2.

The stated property of G can be shown to be equivalent to the statement that (G, I_a) satisfies Hypothesis 2.1 with the same μ as the given pair $(\mathcal{E}, \mathcal{A})$ and $d = 0$, see again [7]. Note that $m = 0$ in this case.

REMARK 2.4. In the case of $m = 0$, i. e. if the system (2.2) has square coefficients, Hypothesis 2.1 is equivalent to the requirement that the corresponding pair of matrix functions has a well-defined differentiation index ν. In particular, we have

$$\nu = \begin{cases} 0 & \text{for } \mu = 0,\ a = 0, \\ \mu + 1 & \text{otherwise.} \end{cases} \tag{2.7}$$

For details, see [7].

3. Properties of the formal adjoint. In this section, we study the properties of the formal adjoint of a pair of matrix functions, which is defined as follows.

DEFINITION 3.1. *Let* $E \in C^1(\mathbb{I}, \mathbb{C}^{n,n})$ *and* $A \in C(\mathbb{I}, \mathbb{C}^{n,n})$. *The pair* $(-E^H, (A + \dot{E})^H)$ *of matrix functions is called the* formal adjoint *of the pair of matrix functions* (E, A).

This definition can be motivated by the following observation. In the case of the pair $(\hat{E}, \hat{A})$ as in (1.2), we know that $\hat{E}$ has constant rank. We can therefore define the Banach space operators

$$D : \mathbb{X} \to \mathbb{Y}, \quad \mathbb{X} = \{x \in C(\mathbb{I}, \mathbb{C}^n) \mid \hat{E}^+\hat{E}x \in C^1(\mathbb{I}, \mathbb{C}^n),\ (\hat{E}^+\hat{E}x)(\underline{t}) = 0\},\ \mathbb{Y} = C(\mathbb{I}, \mathbb{C}^n),$$

and

$$D^* : \mathbb{Y}^* \to \mathbb{X}^*, \quad \mathbb{Y}^* = \{\lambda \in C(\mathbb{I}, \mathbb{C}^n) \mid \hat{E}\hat{E}^+\lambda \in C^1(\mathbb{I}, \mathbb{C}^n),\ (\hat{E}\hat{E}^+\lambda)(\bar{t}) = 0\},\ \mathbb{X}^* = C(\mathbb{I}, \mathbb{C}^n)$$

via

$$Dx = \hat{E}\tfrac{d}{dt}(\hat{E}^+\hat{E}x) - \hat{A}x - \hat{E}\tfrac{d}{dt}(\hat{E}^+\hat{E})x, \quad D^*\lambda = -\hat{E}^H\tfrac{d}{dt}(\hat{E}\hat{E}^+\lambda) - A^H\lambda - \dot{\hat{E}}(\hat{E}\hat{E}^+)\lambda.$$

Both $\langle \mathbb{X}, \mathbb{X}^* \rangle$ and $\langle \mathbb{Y}, \mathbb{Y}^* \rangle$ form dual systems with respect to the standard scalar product of the Hilbert space $L_2(\mathbb{I}, \mathbb{C}^n)$ considered as corresponding sesquilinear form, see, e. g., [5].

THEOREM 3.2. *The operator D^* is the (unique) conjugate of D.*

Proof. We have that

$$\begin{aligned}\langle Dx, \lambda \rangle &= \int_{\mathbb{I}} \big(\hat{E}\tfrac{d}{dt}(\hat{E}^+\hat{E}x) - \hat{A}x - \hat{E}\tfrac{d}{dt}(\hat{E}^+\hat{E})x \big)^H \lambda \, dt \\ &= \int_{\mathbb{I}} \big(\tfrac{d}{dt}(x^H\hat{E}^+\hat{E})\hat{E}^H\lambda - x^H\hat{A}^H\lambda - x^H\tfrac{d}{dt}(\hat{E}^+\hat{E})\hat{E}^H\lambda \big)\, dt.\end{aligned}$$

Since $\hat{E}^H = \hat{E}^H(\hat{E}^+)^H\hat{E}^H = \hat{E}^H\hat{E}\hat{E}^+$, it follows that

$$\begin{aligned}\langle Dx, \lambda \rangle &= x^H\hat{E}^+\hat{E}\hat{E}^H\hat{E}\hat{E}^+\lambda\,|_{\underline{t}}^{\bar{t}} \\ &\qquad + \int_{\mathbb{I}} \big(- x^H\hat{E}^+\hat{E}\tfrac{d}{dt}(\hat{E}^H\hat{E}\hat{E}^+\lambda) - x^H\hat{A}^H\lambda - x^H\tfrac{d}{dt}(\hat{E}^+\hat{E})\hat{E}^H\lambda \big)\, dt \\ &= \int_{\mathbb{I}} x^H \big(- \dot{\hat{E}}^+\hat{E}\dot{\hat{E}}^H\hat{E}\hat{E}^+\lambda - \hat{E}^+\hat{E}\hat{E}^H\tfrac{d}{dt}(\hat{E}^+\hat{E}\lambda) - \hat{A}^H\lambda - \tfrac{d}{dt}(\hat{E}^+\hat{E})\hat{E}^H\lambda \big)\, dt.\end{aligned}$$

Since $\hat{E}^H = \hat{E}^H(\hat{E}^+)^H\hat{E}^H = \hat{E}^+\hat{E}\hat{E}^H$ and

$$\hat{E}^+\hat{E}\dot{\hat{E}}^H\hat{E}\hat{E}^+ + \tfrac{d}{dt}(\hat{E}^+\hat{E})\hat{E}^H = (\hat{E}^+\hat{E}\dot{\hat{E}}^H + \tfrac{d}{dt}(\hat{E}^+\hat{E})\hat{E}^H)\hat{E}\hat{E}^+ = \tfrac{d}{dt}(\hat{E}^+\hat{E}\hat{E}^H)\hat{E}\hat{E}^+ = \dot{\hat{E}}^H\hat{E}\hat{E}^+,$$

we finally get that

$$\langle Dx, \lambda \rangle = \int_{\mathbb{I}} x^H(-\hat{E}^H\tfrac{d}{dt}(\hat{E}^+\hat{E}\lambda) - \hat{A}^H\lambda - \dot{\hat{E}}^H\hat{E}\hat{E}^+\lambda)\, dt = \langle x, D^*\lambda \rangle.$$

□

The operators D and D^* are defined in such a way that they explicitly exhibit the smoothness requirements contained in the definition of their domains. Supposing sufficient smoothness of $\hat{E}$, x, and λ, the operators can be written as

$$Dx = \hat{E}\dot{x} - \hat{A}x, \quad D^*\lambda = -\hat{E}^H\dot{\lambda} - (\hat{A} + \tfrac{d}{dt}\hat{E})^H\lambda,$$

which then directly suggests Definition 3.1 in the strangeness-free case. Note that a similar argument in the general case is only possible when the matrix function E has constant rank which is equivalent to E^+ being continuous. But this is not required by Hypothesis 2.1, since it is not a necessary property of a regular DAE. This also applies to DAEs with so-called properly stated leading term, see [11, 12].

Theorem 3.2 also shows that the adjoint pair should be defined with a different sign compared to [8]. Note that this extra sign is due to the involved partial integration.

We now present some fundamental properties of the formal adjoint.

THEOREM 3.3. *The formal adjoint of the formal adjoint of a pair of matrix functions is the given pair of matrix functions.*

Proof. Given (E, A) with $E \in C^1(\mathbb{I}, \mathbb{C}^{n,n})$ and $A \in C(\mathbb{I}, \mathbb{C}^{n,n})$, we observe that the formal adjoint $(-E^H, (A + \dot{E})^H)$ satisfies the assumptions of Definition 3.1. Its formal adjoint therefore has the form

$$(-(-E^H)^H, ((A + \dot{E})^H + (-\dot{E}^H))^H) = (E, A + \dot{E} - \dot{E}) = (E, A).$$

□

THEOREM 3.4. *The formal adjoints of two globally equivalent pairs of matrix functions are globally equivalent provided that the involved transformations are sufficiently smooth.*

Proof. Given (E, A) with $E \in C^1(\mathbb{I}, \mathbb{C}^{n,n})$ and $A \in C(\mathbb{I}, \mathbb{C}^{n,n})$, let

$$(\tilde{E}, \tilde{A}) = (PEQ, PAQ - PE\dot{Q})$$

according to (2.4), with the additional requirement that P is continuously differentiable. The formal adjoint of $(\tilde{E}, \tilde{A})$ is then given by

$$\begin{aligned}
&(-(PEQ)^H, (PAQ - PE\dot{Q} + \tfrac{d}{dt}(PEQ))^H) \\
&\quad = (-Q^H E^H P^H, Q^H A^H P^H - \dot{Q}^H E^H P^H + Q^H E^H \dot{P}^H + Q^H \dot{E}^H P^H + \dot{Q}^H E^H P^H) \\
&\quad = (Q^H(-E^H)P^H, Q^H(A + \dot{E})^H P^H - Q^H(-E^H)\dot{P}^H) \sim (-E^H, (A + \dot{E})^H).
\end{aligned}$$

□

An important consequence of Theorem 3.4 is that in the investigation of a pair of matrix functions (E, A) and its formal adjoint $(\tilde{E}, \tilde{A})$, we may assume w.l.o.g. that the pair (E, A) is in the *global canonical form*

$$(E, A) = \left(\begin{bmatrix} I_d & H \\ 0 & G \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & I_a \end{bmatrix} \right), \tag{3.1}$$

and thus, according to Theorem 2.3, the formal adjoint is given by

$$(\tilde{E}, \tilde{A}) = \left(\begin{bmatrix} -I_d & 0 \\ -H^H & -G^H \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ \dot{H}^H & I_a + \dot{G}^H \end{bmatrix} \right), \tag{3.2}$$

provided that Hypothesis 2.1 holds and that the properties under consideration transform covariantly with respect to global equivalence.

The remainder of this section is dedicated to the question whether the formal adjoint pair of a given pair of matrix functions satisfies Hypothesis 2.1 if the given pair does.

THEOREM 3.5. *Let (E, A) have a well-defined differentiation index $\nu \geq 1$ and size d of the differential part. Then the formal adjoint pair $(\tilde{E}, \tilde{A}) = (-E^H, (A + \dot{E})^H)$ also has a well-defined differentiation index, which equals ν, with the same size d of the differential part.*

Proof. Since Hypothesis 2.1 itself transforms covariantly with respect to global equivalence, see [7], we are allowed to assume that we are in the situation of (3.1). Since (E, A) is assumed to have a well-defined differentiation index ν, it satisfies Hypothesis 2.1 with $\mu = \nu - 1$.

The coefficients of the derivative array belonging to (E, A) have the form

$$M_\mu = \left[\begin{array}{cc|cc|cc|cc}
I & H & & & & & & \\
0 & G & & & & & & \\
\hline
0 & \dot{H} & I & H & & & & \\
0 & \dot{G} - I & 0 & G & & & & \\
\hline
\vdots & \vdots & \ddots & & \ddots & & & \\
\vdots & \vdots & & \ddots & & \ddots & & \\
\hline
0 & H^{(\mu)} & \cdots & \cdots & 0 & \mu\dot{H} & I & H \\
0 & G^{(\mu)} & \cdots & \cdots & 0 & \mu\dot{G} - I & 0 & G
\end{array}\right],$$

$$N_\mu = \left[\begin{array}{cccc|cc|cc}
0 & 0 & 0 & 0 & \cdots & \cdots & 0 & 0 \\
0 & I & 0 & 0 & \cdots & \cdots & 0 & 0 \\
\hline
0 & 0 & 0 & 0 & \cdots & \cdots & 0 & 0 \\
0 & 0 & 0 & 0 & \cdots & \cdots & 0 & 0 \\
\hline
\vdots & \vdots & \vdots & \vdots & & & \vdots & \vdots \\
\vdots & \vdots & \vdots & \vdots & & & \vdots & \vdots \\
\hline
0 & 0 & 0 & 0 & \cdots & \cdots & 0 & 0 \\
0 & 0 & 0 & 0 & \cdots & \cdots & 0 & 0
\end{array}\right],$$

so that the quantities of Hypothesis 2.1 are given by

$$Z_2^H = [\,0\ \ Z_{2,0}^H \,|\, 0\ \ Z_{2,1}^H \,|\, \cdots \,|\, 0\ \ Z_{2,\mu}^H\,],$$

where we can choose $Z_{2,0}^H = I$, and by

$$Z_2^H N_\mu V = [\,0\ \ I\,], \quad T_2 = \begin{bmatrix} I \\ 0 \end{bmatrix}, \quad ET_2 = \begin{bmatrix} I \\ 0 \end{bmatrix},$$

see [7]. The coefficients in the derivative array belonging to $(\tilde{E}, \tilde{A})$ have the form

$$\tilde{M}_\mu = \left[\begin{array}{cc|cc|c|cc} -I & 0 & & & & & \\ -H^H & -G^H & & & & & \\ \hline 0 & 0 & -I & 0 & & & \\ -2\dot{H}^H & -2\dot{G}^H - I & -H^H & -G^H & & & \\ \hline \vdots & \vdots & \ddots & & \ddots & & \\ \vdots & \vdots & & \ddots & & \ddots & \\ \hline 0 & 0 & \cdots & \cdots & 0 \quad 0 & & -I \quad 0 \\ -\mu(H^{(\mu)})^H & -\mu(G^{(\mu)})^H & \cdots & \cdots & -\nu\dot{H}^H & -\nu\dot{G}^H - I & -H^T \quad -G^H \end{array}\right],$$

$$\tilde{N}_\mu = \left[\begin{array}{cc|cc|ccc|cc} 0 & 0 & 0 & 0 & \cdots & \cdots & 0 & 0 \\ \dot{H}^H & I + \dot{G}^H & 0 & 0 & \cdots & \cdots & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & \cdots & \cdots & 0 & 0 \\ \ddot{H}^H & \ddot{G}^H & 0 & 0 & \cdots & \cdots & 0 & 0 \\ \hline \vdots & \vdots & \vdots & \vdots & & & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & & & \vdots & \vdots \\ \hline 0 & 0 & 0 & 0 & \cdots & \cdots & 0 & 0 \\ (H^{(\nu)})^H & (G^{(\nu)})^H & 0 & 0 & \cdots & \cdots & 0 & 0 \end{array}\right].$$

Due to the identities in the diagonal of $\tilde{M}_\mu$, possible quantities for Hypothesis 2.1 are

$$\tilde{Z}_2^H = [\,*\ \ Z_{2,0}^H \,|\, *\ \ Z_{2,1}^H \,|\, \cdots \,|\, *\ \ Z_{2,\mu}^H\,],$$

together with

$$\tilde{Z}_2^H \tilde{N}_\mu V = [\,*\ \ I\,], \quad \tilde{T}_2 = \begin{bmatrix} I \\ * \end{bmatrix}, \quad \tilde{E}\tilde{T}_2 = \begin{bmatrix} -I \\ * \end{bmatrix}.$$

Due to the special structure of the canonical form, it is thus sufficient to restrict ourselves to pairs $(E, A) = (G, I)$ and $(\tilde{E}, \tilde{A}) = (-G^H, I + \dot{G}^H)$. In particular, we have to show that $(-G^H, I + \dot{G}^H)$ satisfies Hypothesis 2.1 with $d = 0$.

By assumption, the pair (G, I) satisfies Hypothesis 2.1 with $d = 0$. With the corresponding coefficients in the derivative array (leaving out now the indices for simplicity noting that there is no conflict with the matrix M of (1.4) which does not play any role in the present context)

$$M = \begin{bmatrix} G & & & \\ \dot{G} - I & G & & \\ \vdots & \ddots & \ddots & \\ G^{(\mu)} & \cdots & \mu\dot{G} - I & G \end{bmatrix}, \quad N = \begin{bmatrix} I & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix},$$

the matrix function describing the corange of M is of the form

$$Z^H = \begin{bmatrix} Z_0^H & Z_1^H & \cdots & Z_\mu^H \end{bmatrix},$$

and by a proper scaling we may assume that $Z_0 = I$. To analyze whether Hypothesis 2.1 holds for $(-G^H, I + \dot{G}^H)$, we consider the corresponding derivative array with coefficients

$$\tilde{M} = \begin{bmatrix} -G^H & & & \\ -2\dot{G}^H - I & -G^H & & \\ \vdots & \ddots & \ddots & \\ -\nu(G^{(\mu)})^H & \cdots & -\nu\dot{G}^H - I & -G^H \end{bmatrix}, \quad \tilde{N} = \begin{bmatrix} I + \dot{G}^H & 0 & \cdots & 0 \\ \ddot{G}^H & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ (G^{(\nu)})^H & 0 & \cdots & 0 \end{bmatrix}.$$

In particular, we need to determine the corange of $\tilde{M}$, which is given in the form

$$\tilde{Z}^H = \begin{bmatrix} \tilde{Z}_0^H & \tilde{Z}_1^H & \cdots & \tilde{Z}_\mu^H \end{bmatrix}.$$

We now show that setting

$$\tilde{Z}_i^H = \sum_{l=i}^{\mu} (-1)^l \binom{l}{i} Z_l^{(l-i)}$$

actually yields

$$\tilde{Z}^H \tilde{M} = 0, \quad \tilde{Z}^H \tilde{N} V \text{ pointwise nonsingular.} \tag{3.3}$$

To show this, we first need the following property of Z. By assumption, the DAE

$$G\dot{x} = x + f$$

possesses a unique solution for every sufficiently smooth f. By the construction of Z, this solution is given by the solution of

$$Z^H M \begin{bmatrix} \dot{x} \\ \vdots \\ x^{(\mu+1)} \end{bmatrix} = Z^H N \begin{bmatrix} x \\ \vdots \\ x^{(\mu)} \end{bmatrix} + Z^H g, \quad g = \begin{bmatrix} f \\ \vdots \\ f^{(\mu)} \end{bmatrix}.$$

Since $Z^H M = 0$ and $Z^H N V = I$, this implies that

$$x = -Z^H g.$$

Inserting this into the given DAE gives that

$$G(-\dot{Z}^H g - Z^H \dot{g}) = -Z^H g + f$$

for every sufficiently smooth f. Hence,

$$\sum_{l=0}^{\mu} G\dot{Z}_l^H f^{(l)} + \sum_{l=0}^{\mu} G Z_l^H f^{(l+1)} - \sum_{l=0}^{\mu} Z_l^H f^{(l)} + f = 0,$$

and thus, using $Z_0 = I$, we have that

$$\begin{aligned} (G\dot{Z}_1^H + GZ_0^H - Z_1^H)\dot{f} + (G\dot{Z}_2^H + GZ_1^H - Z_2^H)\ddot{f} + \\ + \cdots + (G\dot{Z}_\mu^H + GZ_{\mu-1}^H - Z_\mu^H) f^{(\mu)} + GZ_\mu^H f^{(\mu+1)} = 0 \end{aligned}$$

for every sufficiently smooth f. Since this can only hold if all coefficients of the derivatives of f vanish, it follows that

$$Z_l = (Z_{l-1} - \dot{Z}_l)G^H, \ l = 1, \ldots, \mu, \quad Z_\mu G^H = 0. \tag{3.4}$$

To show the first part of (3.3), we observe that (with $\delta_{i,j}$ denoting the Kronecker delta)

$$(\tilde{M})_{i,j} = -\binom{i+1}{j+1}(G^{(i-j)})^H - \delta_{i,j+1} I, \quad (\tilde{N})_{i,0} = \delta_{i,0} I + (G^{(i+1)})^H, \quad i, j = 0, \ldots, \mu,$$

for the j-th block of $\tilde{Z}^H\tilde{M}$ we get

$$(\tilde{Z}^H\tilde{M})_j = \sum_{i=j}^{\mu}\Big(\sum_{l=i}^{\mu}(-1)^{l+1}\tbinom{l}{i}Z_l^{(l-i)}\Big)\Big(\tbinom{i+1}{j+1}(G^{(i-j)})^H + \delta_{i,j+1}I\Big).$$

For $j=\mu$, we then obtain that

$$(\tilde{Z}^H\tilde{M})_\mu = (-1)^{\mu+1}Z_\mu G^H = 0,$$

and for $j<\mu$, we have that

$$(\tilde{M})_{i,j} = \sum_{i=j}^{\mu}\Big(\sum_{l=i}^{\mu}(-1)^{l+1}\tbinom{l}{i}Z_l^{(l-i)}\Big)\tbinom{i+1}{j+1}(G^{(i-j)})^H + \sum_{l=j+1}^{\mu}(-1)^{l+1}\tbinom{l}{j+1}Z_l^{(l-j-1)}.$$

Changing the order of summation in the first term and using (3.4) in the second term gives

$$\begin{aligned}(\tilde{M})_{i,j} = &\sum_{l=j}^{\mu}\sum_{i=j}^{\mu}(-1)^{l+1}\tbinom{l}{i}\tbinom{i+1}{j+1}Z_l^{(l-i)}(G^{(i-j)})^H\\ &+\sum_{l=j+1}^{\mu}(-1)^{l+1}\tbinom{l}{j+1}\sum_{k=0}^{l-j-1}\tbinom{l-j-1}{k}\Big(Z_{l-1}^{(l-j-k-1)}+Z_l^{(l-j-k)}\Big)(G^{(k)})^H.\end{aligned}$$

Shifting the summation indices, we get

$$\begin{aligned}(\tilde{M})_{i,j} = &\sum_{l=j}^{\mu}\sum_{i=j}^{\mu}(-1)^{l+1}\tbinom{l}{i}\tbinom{i+1}{j+1}Z_l^{(l-i)}(G^{(i-j)})^H\\ &-\sum_{l=j}^{\mu-1}(-1)^{l+1}\tbinom{l+1}{j+1}\sum_{i=j}^{l}\tbinom{l-j}{i-j}Z_l^{(l-i)}(G^{(i-j)})^H\\ &+\sum_{l=j+1}^{\mu}(-1)^{l+1}\tbinom{l}{j+1}\sum_{i=j}^{l}\tbinom{l-j-1}{i-j}Z_l^{(l-i)}(G^{(i-j)})^H.\end{aligned}\tag{3.5}$$

Observing that

$$\begin{aligned}&\tbinom{\mu}{i}\tbinom{i+1}{j+1}+\tbinom{\mu}{j+1}\tbinom{\mu-j-1}{i-j}\\ &\quad=\frac{\mu!}{i!(\mu-i)!}\frac{(i+1)!}{(j+1)!(i-j)!}+\frac{\mu!}{(j+1)!(\mu-j-1)!}\frac{(\mu-j-1)!}{(i-j)!(\mu-i-1)!}\\ &\quad=\frac{\mu!}{(\mu-i)!(j+1)!(i-j)!}\big((i+1)+(\mu-i)\big)\\ &\quad=\frac{(\mu+1)!}{(j+1)!(\mu-j)!}\frac{(\mu-j)!}{(\mu-i)!(i-j)!}=\tbinom{\mu+1}{j+1}\tbinom{\mu-j}{i-j},\end{aligned}$$

for the terms in (3.5) with $l=\mu$, we get (up to a sign) that

$$\begin{aligned}\sum_{i=j}^{\mu}\Big[\tbinom{\mu}{i}\tbinom{i+1}{j+1}+\tbinom{\mu}{j+1}\tbinom{\mu-j-1}{i-j}\Big]Z_\mu^{(\mu-i)}(G^{(i-j)})^H &= \tbinom{\mu+1}{j+1}\sum_{i=j}^{\mu}\tbinom{\mu-j}{i-j}Z_\mu^{(\mu-i)}(G^{(i-j)})^H\\ &=\tbinom{\mu+1}{j+1}\sum_{k=0}^{\mu-j}\tbinom{\mu-j}{k}Z_\mu^{(\mu-j-k)}(G^{(k)})^H=\tbinom{\mu+1}{j+1}\Big(\frac{d}{dt}\Big)^{\mu-j}(Z_\mu G^H)=0.\end{aligned}$$

For $l=j$, it follows that $i=j$ in (3.5) and the terms sum up to zero because of

$$\tbinom{l}{i}\tbinom{i+1}{j+1}-\tbinom{l+1}{j+1}\tbinom{l-j}{i-j}=\tbinom{j}{j}\tbinom{j+1}{j+1}-\tbinom{j+1}{j+1}\tbinom{0}{0}=0.$$

Since

$$\begin{aligned}\binom{l}{i}\binom{i+1}{j+1} - \binom{l+1}{j+1}\binom{l-j}{i-j} + \binom{l}{j+1}\binom{l-j-1}{i-j}&\\ = \frac{l!}{i!(l-i)!}\frac{(i+1)!}{(j+1)!(i-j)!} - \frac{(l+1)!}{(j+1)!(l-j)!}\frac{(l-j)!}{(i-j)!(l-i)!}&\\ + \frac{l!}{(j+1)!(l-j-1)!}\frac{(l-j-1)!}{(i-j)!(l-i-1)!}&\\ = \frac{l!}{(l-i)!(j+1)!(i-j)!}\big((i+1)-(l+1)+(l-i)\big) = 0,&\end{aligned}$$

also the remaining terms sum up to zero. Hence, we have shown that $\tilde{Z}^H\tilde{M} = 0$ and thus the first part of (3.3).

For the second part of (3.3), we start from

$$\tilde{Z}^H\tilde{N}V = \sum_{i=0}^{\mu}\Big(\sum_{l=i}^{\mu}(-1)^l\tbinom{l}{i}Z_l^{(l-i)}\Big)(G^{(i+1)})^H + \sum_{l=0}^{\mu}(-1)^l Z_l^{(l)}.$$

Changing the order of summation in the first term and using (3.4) in the second term gives

$$\begin{aligned}\tilde{Z}^H\tilde{N}V = \sum_{l=0}^{\mu}\sum_{i=0}^{l}(-1)^l\tbinom{l}{i}Z_l^{(l-i)}(G^{(i+1)})^H + Z_0&\\ + \sum_{l=1}^{\mu}(-1)^l\sum_{k=0}^{l}\tbinom{l}{k}(Z_{l-1}^{(l-k)} + Z_l^{(l-k+1)})(G^{(k)})^H.&\end{aligned}$$

Shifting the summation indices, we get

$$\begin{aligned}\tilde{Z}^H\tilde{N}V = \sum_{l=0}^{\mu}\sum_{i=0}^{l}(-1)^l\tbinom{l}{i}Z_l^{(l-i)}(G^{(i+1)})^H + Z_0&\\ - \sum_{l=0}^{\mu-1}(-1)^l\sum_{k=0}^{l+1}\tbinom{l+1}{k}Z_l^{(l-k+1)}(G^{(k)})^H + \sum_{l=1}^{\mu}(-1)^l\sum_{k=0}^{l}\tbinom{l}{k}Z_l^{(l-k+1)}(G^{(k)})^H.&\end{aligned}$$

For $l \neq 0$ and $l \neq \mu$, the terms for $k = 0$ of the last two sums cancel out, so that we remain with

$$\begin{aligned}\tilde{Z}^H\tilde{N}V = \sum_{l=0}^{\mu}\sum_{i=0}^{l}(-1)^l\tbinom{l}{i}Z_l^{(l-i)}(G^{(i+1)})^H + Z_0 - \dot{Z}_0G^H + (-1)^\mu Z_\mu^{(\mu+1)}G^H&\\ - \sum_{l=0}^{\mu-1}(-1)^l\sum_{i=0}^{l}\tbinom{l+1}{i+1}Z_l^{(l-i)}(G^{(i+1)})^H + \sum_{l=1}^{\mu}(-1)^l\sum_{i=0}^{l-1}\tbinom{l}{i+1}Z_l^{(l-i)}(G^{(i+1)})^H.&\end{aligned} \tag{3.6}$$

Observing that

$$\binom{\mu}{i} + \binom{\mu}{i+1} = \binom{\mu+1}{i+1},$$

for the terms in (3.6) with $l = \mu$, we get (up to a sign) that

$$\begin{aligned}\sum_{i=0}^{\mu-1}\Big[\tbinom{\mu}{i} + \tbinom{\mu}{i+1}\Big]Z_\mu^{(\mu-i)}(G^{(i+1)})^H + Z_\mu(G^{(\mu+1)})^H + Z_\mu^{(\mu+1)}G^H&\\ = \sum_{i=0}^{\mu-1}\tbinom{\mu+1}{i+1}Z_\mu^{(\mu-i)}(G^{(i+1)})^H + Z_\mu(G^{(\mu+1)})^H + Z_\mu^{(\mu+1)}G^H&\\ = \sum_{i=0}^{\mu+1}\tbinom{\mu+1}{i}Z_\mu^{(\mu-i+1)}(G^{(i)})^H = \Big(\frac{d}{dt}\Big)^{\mu+1}(Z_\mu G^H) = 0.&\end{aligned}$$

For $l = 0$, it follows that $i = 0$ in (3.6), and the terms sum up to zero because of

$$\binom{0}{0} - \binom{1}{1} = 0.$$

The same holds for $0 < l < \mu$ and $i = l$ because of

$$\binom{l}{l} - \binom{l+1}{l+1} = 0,$$

and for the remaining terms in the sums because of

$$\binom{l}{i} - \binom{l+1}{i+1} + \binom{l}{i+1} = 0.$$

We therefore end up with

$$\tilde{Z}^H \tilde{N} V = Z_0 - \dot{Z}_0 G^H = I,$$

since $Z_0 = I$. Thus, we have also shown the second part of (3.3). □

If we do not assume that the system (1.1) has a well-defined differentiation index, then the situation becomes more complicated. It is even not clear then, whether the use of an adjoint makes sense in this case, as is demonstrated by the following example.

EXAMPLE 3.6. Consider the pair of constant matrix functions

$$(E, A) = \left(\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \right).$$

The associated DAE with inhomogeneity f then is

$$\dot{x}_1 = f_1, \quad 0 = x_1 + f_2.$$

Obviously, the component x_2 is free, but we need to differentiate the second equation to obtain the consistency condition $f_1 + \dot{f}_2 = 0$. Thus, the strangeness index μ of (E, A) satisfies $\mu = 1$.

The formal adjoint of (E, A) is given by

$$(-E^H, (A + \dot{E})^H) = \left(\begin{bmatrix} -1 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \right).$$

The associated DAE with inhomogeneity h then is

$$-\dot{\lambda}_1 = \lambda_2 + h_1, \quad 0 = h_2.$$

Again, with λ_2 there is a free solution component, but there is no need for differentiating the equations in order to decide on the solution properties of DAE. Thus, we have $\mu = 0$ in this case.

The reason for this observation can be seen in the fact that the bidiagonal blocks in the Kronecker canonical form and their conjugate transposed counterparts do not possess the same strangeness index, see [7].

4. Properties of the formal necessary optimality conditions. In this section, we will investigate the relation between the true necessary conditions (1.7) and the formal necessary conditions (1.10) for the solution (x, u) of the optimal control problem (1.4) with (1.5). The main tool for this analysis will be to transform both to the canonical form (2.5). To show that we are allowed to do so, we first rewrite the formal necessary conditions in terms of a behavior setting. For this, we define

$$\mathcal{E} = [\, E \;\; 0 \,], \quad \mathcal{A} = [\, A \;\; B \,], \quad \mathcal{W} = \begin{bmatrix} W & S \\ S^T & R \end{bmatrix}, \quad z = \begin{bmatrix} x \\ u \end{bmatrix},$$

such that the formal necessary conditions become (ignoring the boundary conditions for the moment)

$$\begin{array}{ll} \text{(a)} & \mathcal{E}\dot{z} = \mathcal{A}z + f, \\ \text{(b)} & -\mathcal{E}^H\dot{\lambda} = (\mathcal{A} + \dot{\mathcal{E}})^H + \mathcal{W}z. \end{array} \tag{4.1}$$

Setting

$$\tilde{\mathcal{E}} = P\mathcal{E}Q, \quad \tilde{\mathcal{A}} = P\mathcal{A}Q - P\mathcal{E}\dot{Q}, \quad \tilde{\mathcal{W}} = Q^H\mathcal{W}Q$$

according to global equivalence (2.4), we have

$$\begin{aligned} &\left(\begin{bmatrix} 0 & \mathcal{E} \\ -\mathcal{E}^H & 0 \end{bmatrix}, \begin{bmatrix} 0 & \mathcal{A} \\ \mathcal{A}^H + \dot{\mathcal{E}}^H & \mathcal{W} \end{bmatrix}\right) \\ &\quad\sim \left(\begin{bmatrix} P & 0 \\ 0 & Q^H \end{bmatrix}\begin{bmatrix} 0 & \mathcal{E} \\ -\mathcal{E}^H & 0 \end{bmatrix}\begin{bmatrix} P^H & 0 \\ 0 & Q \end{bmatrix}, \begin{bmatrix} P & 0 \\ 0 & Q^H \end{bmatrix}\begin{bmatrix} 0 & \mathcal{A} \\ \mathcal{A}^H + \dot{\mathcal{E}}^H & \mathcal{W} \end{bmatrix}\begin{bmatrix} P^H & 0 \\ 0 & Q \end{bmatrix}\right. \\ &\qquad\qquad \left. - \begin{bmatrix} P & 0 \\ 0 & Q^H \end{bmatrix}\begin{bmatrix} 0 & \mathcal{E} \\ -\mathcal{E}^H & 0 \end{bmatrix}\begin{bmatrix} \dot{P}^H & 0 \\ 0 & \dot{Q} \end{bmatrix}\right) \\ &\quad= \left(\begin{bmatrix} 0 & P\mathcal{E}Q \\ -Q^H\mathcal{E}^HP^H & 0 \end{bmatrix}, \begin{bmatrix} 0 & P\mathcal{A}Q - P\mathcal{E}\dot{Q} \\ Q^H(\mathcal{A}^H + \dot{\mathcal{E}}^H)P^H + Q^H\mathcal{E}^H\dot{P}^H & Q^H\mathcal{W}Q \end{bmatrix}\right) \\ &\quad= \left(\begin{bmatrix} 0 & \tilde{\mathcal{E}} \\ -\tilde{\mathcal{E}}^H & 0 \end{bmatrix}\begin{bmatrix} 0 & \tilde{\mathcal{A}} \\ \tilde{\mathcal{A}}^H + \dot{Q}^H\mathcal{E}^HP^H + Q^H\dot{\mathcal{E}}^HP^H + Q^H\mathcal{E}^H\dot{P}^H & \tilde{\mathcal{W}} \end{bmatrix}\right) \\ &\quad= \left(\begin{bmatrix} 0 & \tilde{\mathcal{E}} \\ -\tilde{\mathcal{E}}^H & 0 \end{bmatrix}\begin{bmatrix} 0 & \tilde{\mathcal{A}} \\ \tilde{\mathcal{A}}^H + \dot{\tilde{\mathcal{E}}}^H & \tilde{\mathcal{W}} \end{bmatrix}\right). \end{aligned}$$

Hence, the problem (4.1) transforms covariantly with global equivalence transformations of the pair $(\mathcal{E}, \mathcal{A})$.

On the other hand, the true necessary conditions (1.7) involve the index-reduced DAE (1.2). Defining

$$\hat{\mathcal{E}} = [\,\hat{E}\ 0\,], \quad \hat{\mathcal{A}} = [\,\hat{A}\ \hat{B}\,],$$

the corresponding behavior formulation is given by

$$\begin{array}{ll} \text{(a)} & \hat{\mathcal{E}}\dot{z} = \hat{\mathcal{A}}z + \hat{f}, \\ \text{(b)} & -\hat{\mathcal{E}}^H\dot{\lambda} = (\hat{\mathcal{A}} + \dot{\hat{\mathcal{E}}})^H + \mathcal{W}z. \end{array} \tag{4.2}$$

To show that (4.2) also transforms covariantly with global equivalence transformations involving the same transformations, we must investigate the whole construction of the reduced DAE (1.2).

We start with the original DAE (2.2) and the transformed DAE given by (2.4) and $z = Q\tilde{z}$, $\tilde{f} = Pf$ according to

$$\mathcal{E}\dot{z} = \mathcal{A}z + f, \quad \tilde{\mathcal{E}}\dot{\tilde{z}} = \tilde{\mathcal{A}}z + \tilde{f}.$$

The coefficients of the corresponding derivative arrays are denoted by (M, N) and $(\tilde{M}, \tilde{N})$, respectively, omitting the index μ for simplicity. Then (2.4) implies that

$$\tilde{M} = \Pi M\Theta, \quad \tilde{N} = \Pi N\Theta - \Pi M\Psi,$$

where

$$\begin{aligned} &\Pi_{i,j} = \tbinom{i}{j}P^{(i-j)}, \quad \Theta_{i,j} = \tbinom{i+1}{j+1}Q^{(i-j)}, \\ &\Psi_{i,j} = \begin{cases} Q^{(i+1)} & \text{for } i = 0, \ldots, \mu,\ j = 0, \\ 0 & \text{otherwise,} \end{cases} \end{aligned}$$

see [7, Th. 3.29]. For the index reduction, we follow Hypothesis 2.1 and choose Z_2 such that

$$Z_2^H M = 0.$$

This corresponds to choosing $\tilde{Z}_2$ for the transformed DAE according to

$$\tilde{Z}_2^H = Z_2^H \Pi^{-1}.$$

Hypothesis 2.1 then implies that $Z_2^H NV$ has (pointwise) full row rank, or equivalently that

$$\tilde{Z}_2^H \tilde{N} V = Z_2^H \Pi^{-1}(\Pi N\Theta - \Pi M\Psi)V = Z_2^H N\Theta V = Z_2^H NVQ$$

has (pointwise) full row rank, where we have used the special structure of N, Θ, and V. The choice of T_2 in the next step according to $Z_2^H NVT_2 = 0$, then corresponds to

$$\tilde{T}_2 = Q^{-1}T_2.$$

Hence, $\operatorname{rank} \mathcal{E}T_2 = d$ is equivalent to

$$\operatorname{rank} \tilde{\mathcal{E}}\tilde{T}_2 = \operatorname{rank} P\mathcal{E}QQ^{-1}T_2 = d$$

and the choice of Z_1 so that $Z_1^H \mathcal{E}T_2$ is pointwise nonsingular corresponds to

$$\tilde{Z}_1^H = Z_1^H P^{-1}.$$

Index reduction of $\mathcal{E}\dot{z} = \mathcal{A}z + f$ then gives

$$\begin{aligned} &\text{(a)} \quad Z_1^H \mathcal{E}\dot{z} = Z_1^H \mathcal{A}z + Z_1^H f, \\ &\text{(b)} \quad 0 = Z_2^H NVz + Z_2^T g, \end{aligned} \tag{4.3}$$

whereas index reduction of $\tilde{\mathcal{E}}\dot{\tilde{z}} = \tilde{\mathcal{A}}z + \tilde{f}$ with

$$\tilde{z} = Q^{-1}z, \quad \tilde{f} = Pf, \quad \tilde{g} = \Pi g \tag{4.4}$$

yields

$$\begin{aligned} &\text{(a)} \quad \tilde{Z}_1^H \tilde{\mathcal{E}}\dot{\tilde{z}} = \tilde{Z}_1^H \tilde{\mathcal{A}}\tilde{z} + \tilde{Z}_1^H \tilde{f}, \\ &\text{(b)} \quad 0 = \tilde{Z}_2^H \tilde{N} V\tilde{z} + \tilde{Z}_2^T \tilde{g}. \end{aligned} \tag{4.5}$$

Inserting the transformation into (4.5a) gives

$$Z_1^H P^{-1} P\mathcal{E}Q(Q^{-1}\dot{z} - Q^{-1}\dot{Q}Q^{-1}z) = Z_1^H P^{-1}(P\mathcal{A}Q - P\mathcal{E}\dot{Q})Q^{-1}z + Z_1^H P^{-1}Pf,$$

which is (4.3a). Inserting the transformation into (4.5b) gives

$$0 = Z_2^H \Pi^{-1}(\Pi N\Theta - \Pi M\Psi)VQ^{-1}z + Z_2^H \Pi^{-1}\Pi g = Z_2^H N\Theta VQ^{-1}z + Z_2^H g = Z_2^H NVz + Z_2^H g,$$

which is (4.3b).

Thus, we are allowed to assume that $(\mathcal{E}, \mathcal{A})$ is in the global canonical form (2.5) when dealing with both (4.1) and (4.2). In particular, the formal necessary conditions (4.1) then have the form

$$\begin{aligned} &\text{(a)} \quad \dot{z}_1 + H\dot{z}_2 = Lz_3 + f_1, \\ &\text{(b)} \quad G\dot{z}_2 = z_2 + f_2, \\ &\text{(c)} \quad -\dot{\lambda}_1 = W_{11}z_1 + W_{12}z_2 + W_{13}z_3, \\ &\text{(d)} \quad -H^H\dot{\lambda}_1 - G^H\dot{\lambda}_2 = \lambda_2 + \dot{H}^H\lambda_1 + \dot{G}^H\lambda_2 + W_{21}z_1 + W_{22}z_2 + W_{23}z_3, \\ &\text{(e)} \quad 0 = L^H\lambda_1 + W_{31}z_1 + W_{32}z_2 + W_{33}z_3, \end{aligned} \tag{4.6}$$

whereas the true necessary conditions (1.7) then have the form

$$\begin{aligned} &\text{(a)} \quad \dot{z}_1 + H\dot{z}_2 = Lz_3 + f_1, \\ &\text{(b)} \quad 0 = z_2 + g_2, \\ &\text{(c)} \quad -\dot{\lambda}_1 = W_{11}z_1 + W_{12}z_2 + W_{13}z_3, \\ &\text{(d)} \quad -H^H\dot{\lambda}_1 = \lambda_2 + \dot{H}^H\lambda_1 + W_{21}z_1 + W_{22}z_2 + W_{23}z_3, \\ &\text{(e)} \quad 0 = L^H\lambda_1 + W_{31}z_1 + W_{32}z_2 + W_{33}z_3. \end{aligned} \tag{4.7}$$

Due to the special properties of G given in Theorem 2.3 and the results of index reduction, the parts (4.6b) and (4.7b) fix the same solution z_2. Compare also with the previous section. Thus, (4.6) and (4.7) only differ in the parts (4.6d) and (4.7d). Since these equations determine λ_2 in terms of the other unknowns, both systems yield the same solutions for the other unknowns as long as the correct boundary conditions are incorporated. Observe that (4.6), however, may require more smoothness of the data due to a possible higher index of (4.6d). In particular, we may need derivatives of $\mathcal{W}$.

Of course, the true necessary optimality conditions (1.7) state the correct boundary conditions, which may also be written as

$$\hat{E}(\underline{t})x(\underline{t}) = \hat{E}(\underline{t})\underline{x}, \quad \hat{E}(\overline{t})^H\lambda(\overline{t}) = Mx(\overline{t}) \tag{4.8}$$

with the requirement that range $M \subseteq$ range $\hat{E}(\overline{t})^H$. Note that each boundary condition actually contains only d linear independent conditions due to the rank d of $\hat{E}$. Since the formal necessary conditions (1.10) are not based on index reduction, one is tempted to use the boundary conditions

$$E(\underline{t})x(\underline{t}) = E(\underline{t})\underline{x}, \quad E(\overline{t})^H\lambda(\overline{t}) = Mx(\overline{t}), \tag{4.9}$$

which differ from (4.8) in the case of a higher-index DAE in the constraint. Moreover, the restriction on M is not visible here. Thus, the boundary conditions (4.9) may yield contradictions in the formal necessary conditions. But since they contain the correct boundary conditions, we have the following result, compare also with the sufficient conditions given in [10].

THEOREM 4.1. *Let all data of the given optimal control problem (1.4) and (1.5) be sufficiently smooth and let the formal necessary optimality conditions (1.10) have a solution (x, u, λ). Then, there exist a function η replacing λ such that (x, u, η) solves the true necessary optimality conditions (1.7).*

In summary, the formal optimality conditions may need extra smoothness assumptions and may lead to extra consistency conditions for the boundary values. If, however, these two extra requirements are satisfied, then the resulting solutions x, u are the same for both systems while the Lagrange multiplier λ may be different. This is illustrated by the following example, see [1, 8].

EXAMPLE 4.2. Consider the problem

$$\mathcal{J}(x, u) = \frac{1}{2}\int_0^1 (x_1(t)^2 + u(t)^2)dt = \min!$$

subject to the differential-algebraic system

$$\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} u + \begin{bmatrix} f_1 \\ f_2 \end{bmatrix}.$$

The reduced system (1.2) in this case is the purely algebraic equation

$$0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} u + \begin{bmatrix} f_1 + \dot{f}_2 \\ f_2 \end{bmatrix}.$$

The associated adjoint equation is then

$$0 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix},$$

and no initial conditions are needed. The true necessary optimality conditions (1.7) are completed by the optimality condition

$$0 = u + \lambda_1.$$

A simple calculation yields the solution

$$x_1 = u = -\lambda_1 = -\tfrac{1}{2}(f_1 + \dot{f}_2), \quad x_2 = -f_2, \quad \lambda_2 = 0.$$

If, however, we consider the formal adjoint equation given by

$$-\begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}\begin{bmatrix} \dot{\lambda}_1 \\ \dot{\lambda}_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix}, \quad \lambda_1(1) = 0$$

together with the optimality condition, then we obtain that

$$x_1 = u = -\lambda_1 = -\tfrac{1}{2}(f_1 + \dot{f}_2), \quad x_2 = -f_2, \quad \lambda_2 = \tfrac{1}{2}(\dot{f}_1 + \ddot{f}_2)$$

without using the initial condition $\lambda_1(1) = 0$. Depending on the data, this initial condition may be consistent or not. In view of the correct solution it is obvious that this initial condition should not be present. But this cannot be seen from (1.10). Moreover, the determination of λ_2 requires more smoothness of the inhomogeneity than in (1.7).

REMARK 4.3. We have seen that the formal optimality conditions may lead to inconsistencies and extra smoothness conditions. They may, however, have the following computational advantage. In the numerical solution of the optimal control problem via the solution of the true necessary optimality conditions, the needed coefficients of the reduced DAE are obtained pointwise by the pointwise numerical computation of suitable values of the matrix functions Z_1 and Z_2, see [8].

If, however, the DAE boundary value problem of the true necessary optimality conditions itself possesses a nonvanishing strangeness index, then we cannot perform an index reduction for this DAE via derivative arrays, since the coefficients of the DAE are computed quantities. On the other hand, it is no problem to perform a numerical index reduction for the formal necessary conditions, since these are formulated in terms of original data. This procedure will then yield all algebraic constraints contained in the DAE of the boundary value problem and exhibits the smoothness requirements for the inhomogeneity. Moreover, with the help of the algebraic constraints we can check the consistency of the boundary conditions. In this way, we can adjust (if necessary) the boundary conditions and the smoothness of the inhomogeneity to guarantee the existence of a solution.

If these adjustments only influence the formal Lagrange multiplier, then the resulting x and u from the formal necessary conditions even solve the true optimality system and are thus the desired optimal state and input of the optimal control problem.

5. Conclusion. In this paper we have analyzed the properties of the formal adjoint equation associated with a linear differential-algebraic equation. We have shown how their strangeness indices and solution properties are related and used these results to compare the solutions of the true and formal necessary optimality conditions for optimal control problems with DAE constraints. This analysis resolves some of the open questions in the analysis of these optimal control problems and also indicates how to use the formal necessary optimality conditions in the numerical solution of optimal control problems.

REFERENCES

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

[2] 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 ed., 1996.

[3] S. L. CAMPBELL, *Comment on controlling generalized state-space (descriptor) systems*, Internat. J. Control, 46 (1987), pp. 2229–2230.

[4] S. L. CAMPBELL AND C. D. MEYER, *Generalized Inverses of Linear Transformations*, Pitman, San Francisco, CA, 1979.

[5] H. HEUSER, *Funktionalanalysis*, B. G. Teubner, Stuttgart, 3rd ed., 1992.

[6] P. KUNKEL AND V. MEHRMANN, *Characterization of classes of singular linear differential-algebraic equations*, Electr. J. Lin. Alg., 13 (2005), pp. 359–386.

[7] ———, *Differential-Algebraic Equations. Analysis and Numerical Solution*, EMS Publishing House, Zürich, Switzerland, 2006.

[8] ———, *Optimal control for unstructured nonlinear differential-algebraic equations of arbitrary index*, Math. Control, Signals, Sys., 20 (2008), pp. 227–269.

[9] P. KUNKEL, V. MEHRMANN, AND L. SCHOLZ, *Self-adjoint differential-algebraic equations*, in preparation 2011.

[10] G. A. Kurina and R. März, *On linear-quadratic optimal control problems for time-varying descriptor systems*, SIAM J. Cont. Optim., 42 (2004), pp. 2062–2077.
[11] R. März, *The index of linear differential algebraic equations with properly stated leading terms*, Res. in Math., 42 (2002), pp. 308–338.
[12] ———, *Solvability of linear differential algebraic equations with properly stated leading terms*, Res. in Math., 45 (2004), pp. 88–105.
[13] J. W. Polderman and J. C. Willems, *Introduction to Mathematical Systems Theory: A Behavioural Approach*, Springer-Verlag, New York, NY, 1998.