Operators and Matrices
Volume 8, Number 3 (2014), 659–682
doi:10.7153/oam-08-36

SELF–CONJUGATE DIFFERENTIAL AND DIFFERENCE OPERATORS ARISING IN THE OPTIMAL CONTROL OF DESCRIPTOR SYSTEMS

VOLKER MEHRMANN AND LENA SCHOLZ

(*Communicated by J. William Helton*)

Abstract. We analyze the structure of the linear differential and difference operators associated with the necessary optimality conditions of optimal control problems for descriptor systems in continuous- and discrete-time. It has been shown in [27] that in continuous-time the associated optimality system is a self-conjugate operator associated with a self-adjoint pair of coefficient matrices and we show that the same is true in the discrete-time setting. We also extend these results to the case of higher order systems. Finally, we discuss how to turn higher order systems with this structure into first order systems with the same structure.

1. Introduction

The *linear quadratic optimal control problem* with constraints that are given by *differential-algebraic equations (DAEs)* has been discussed in several publications [2, 24, 27, 29]. This is the problem of minimizing a cost functional

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

subject to the constraint

$$E\dot{x} = Ax + Bu + f, \quad x(\underline{t}) = \underline{x} \in \mathbb{R}^n, \qquad (2)$$

with coefficient functions $E, A \in C^0(\mathbb{I}, \mathbb{R}^{n,n})$, $W \in C^0(\mathbb{I}, \mathbb{R}^{n,n})$, $B \in C^0(\mathbb{I}, \mathbb{R}^{n,m})$, $S \in C^0(\mathbb{I}, \mathbb{R}^{n,m})$, $R \in C^0(\mathbb{I}, \mathbb{R}^{m,m})$, $f \in C^0(\mathbb{I}, \mathbb{R}^n)$, and $M_e \in \mathbb{R}^{n,n}$, where $R = R^T$, $W = W^T$ and $M_e = M_e^T$. Here, $\mathbb{I} = [\underline{t}, \bar{t}]$ 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.

It has been shown in [24] that in the case that the differential-algebraic equation (2) has some further properties, (i. e., if it is strangeness-free as a behavior system and if the

Mathematics subject classification (2010): 93C05, 93C55, 93C15, 65L80, 49K15, 34H05.

Keywords and phrases: Differential-algebraic equation, self-conjugate difference operator, self-adjoint pair, discrete-time optimal control, necessary optimality condition, congruence transformation, higher order systems.

This research is supported by the *European Research Council*, through ERC Advanced Grant MODSIMCONMP.

Paper OaM-08-36

coefficients are sufficiently smooth), then the necessary optimality condition is given by the boundary value problem

$$\begin{bmatrix} 0 & E & 0 \\ -E^T & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \frac{d}{dt} \begin{bmatrix} \lambda \\ x \\ u \end{bmatrix} = \begin{bmatrix} 0 & A & B \\ A^T + \frac{d}{dt}E^T & W & S \\ B^T & S^T & R \end{bmatrix} \begin{bmatrix} \lambda \\ x \\ u \end{bmatrix} + \begin{bmatrix} f \\ 0 \\ 0 \end{bmatrix}, \tag{3}$$

with boundary conditions $x(\underline{t}) = \underline{x}$, $E(\bar{t})^T \lambda(\bar{t}) - M_e x(\bar{t}) = 0$.

If we denote the associated differential-algebraic equation (3) as $\mathscr{E}\dot{z} = \mathscr{A}z + \tilde{f}$, then the pair of coefficient functions $(\mathscr{E}, \mathscr{A})$ has the property that $\mathscr{E}^T = -\mathscr{E}$ and $\mathscr{A}^T = \mathscr{A} + \dot{\mathscr{E}}$. Such pairs of matrix functions are called *self-adjoint pairs*, since it has been shown in [27] that this is a property that is associated with a linear *self-conjugate* differential-algebraic operator given by $\mathscr{L}_c := \mathscr{E}\dot{z} - \mathscr{A}z$. Note 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 for (3), see [24].

It has also been shown in [25] for strangeness-free DAEs, and in [2, 29] for special DAEs with properly stated leading term, that if one just formally writes down the system (3) regardless of the properties, and if this system is uniquely solvable, then the solution yields the optimal x, u, but may give a different Lagrange multiplier function λ.

REMARK 1. In many practical applications the state x is not directly accessible for measurements or observations and typically an output equation

$$y = Cx + Du + g, \tag{4}$$

with $C \in C^0(\mathbb{I}, \mathbb{R}^{p,n})$, $D \in C^0(\mathbb{I}, \mathbb{R}^{p,m})$, and $g \in C^0(\mathbb{I}, \mathbb{R}^p)$ is added to (2). The cost functional is then typically also stated in terms of the output equation, i. e.,

$$\mathscr{J}(y,u) = \frac{1}{2} y(\bar{t})^T \tilde{M}_e y(\bar{t}) + \frac{1}{2} \int_{\underline{t}}^{\bar{t}} \left(y^T \tilde{W} y + y^T \tilde{S} u + u^T \tilde{S}^T y + u^T \tilde{R} u \right) dt. \tag{5}$$

In this case one can just insert the output equation (4) into the cost functional (5) and obtain a cost functional of the form (1).

A typical approach in practice for the solution of optimal control problems is the *first-discretize-then-optimize* or *direct transcription* approach, where the optimal control problem (1), i. e., the constraint as well as the cost functional are discretized and then classical optimization techniques are applied to the resulting constrained optimization problem, see e. g., [3, 4, 5, 7]. This method is easy to implement and it is also easy to include other constraints like switching or inequality constraints, but, in general, not much can be said about the convergence of the solution of this optimization problem to the optimal solution of the continuous time problem, see [6, 18].

Another viewpoint of the first-discretize-then-optimize approach is that of discrete time optimal control. If we discretize the DAE (2) on a time grid $\underline{t} = t_0 < t_1 < \ldots < t_N = \bar{t}$ with a suitable discretization method [8, 19, 23] and approximate the cost functional

(1) by an appropriate quadrature rule, then we obtain a *discrete-time linear-quadratic optimal control problem* of minimizing

$$\mathscr{J}_d((x_i),(u_i)) = \frac{1}{2}x_N^T M_e x_N + \frac{1}{2}\sum_{j=0}^{N-1}(x_j^T W_j x_j + x_j^T S_j u_j + u_j^T S_j^T x_j + u_j^T R_j u_j), \tag{6}$$

subject to the difference equation

$$E_{i+1}x_{i+1} = A_i x_i + B_i u_i + f_i, \text{ for } i = 0,\dots,N-1 \quad \text{ and } x_0 = \underline{x} \in \mathbb{R}^n, \tag{7}$$

with $E_i, A_i, W_i \in \mathbb{R}^{n,n}$, $B_i, S_i \in \mathbb{R}^{n,m}$, $R_i \in \mathbb{R}^{m,m}$ and $W_i = W_i^T$, $R_i = R_i^T$ for all i and $M_e = M_e^T \in \mathbb{R}^{n,n}$. Note that the matrices in (6) usually do not match to the corresponding matrix functions in (1) at the discrete time points t_i, e. g., usually $E_i \neq E(t_i)$, $A_i \neq A(t_i)$, etc.

Discrete-time optimal control problems of this form also arise when discrete modeling is used right from the start or when the system is obtained by a sampling method, see e. g. [22, 30].

In the following, $x = (x_i)_{i=0}^N$ and $u = (u_i)_{i=0}^N$ will denote sequences of vectors $x_i \in \mathbb{R}^n$ and $u_i \in \mathbb{R}^m$ and we will use the notation

$$\mathbb{R}^n_{0,N} := \{(x_i)_{i=0}^N \mid x_i \in \mathbb{R}^n\}$$

to denote the vector space of sequences in $\mathbb{R}^n$.

The discrete-time optimal control problem (6) can again be seen as a general optimization problem in Banach spaces, such that necessary optimality conditions can be derived in the same way as in [11, 24, 28, 34]. If the constraint equation (7) is strangeness-free, which in the discrete-time case has been defined and analyzed in [9, 10], then we can extend previous results in the constant coefficient case of [34] to show that the necessary optimality condition for $((x_i),(u_i))$ to be an optimal solution is the existence of a sequence of Lagrange multipliers (λ_i) such that $((x_i),(u_i),(\lambda_i))$ satisfy the discrete-time optimality system

$$\begin{aligned} E_{i+1}x_{i+1} &= A_i x_i + B_i u_i + f_i, \\ -E_i^T\lambda_{i-1} &= W_i x_i + S_i u_i - A_i^T\lambda_i, \\ 0 &= S_i^T x_i + R_i u_i - B_i^T\lambda_i, \end{aligned} \tag{8}$$

together with the boundary conditions

$$\begin{aligned} &E_0^+E_0x_0 = \underline{x}, \quad A_0^T\lambda_0 = W_0x_0 + S_0u_0, \\ &E_N^T\lambda_{N-1} = -M_e x_N, \end{aligned} \tag{9}$$

see Section 3.

If we reformulate system (8) as a *second order difference equation* of the form

$$\begin{bmatrix} 0 & E_{i+1} & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}\begin{bmatrix} \lambda_{i+1} \\ x_{i+1} \\ u_{i+1} \end{bmatrix} + \begin{bmatrix} 0 & -A_i & -B_i \\ -A_i^T & W_i & S_i \\ -B_i^T & S_i^T & R_i \end{bmatrix}\begin{bmatrix} \lambda_i \\ x_i \\ u_i \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ E_i^T & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}\begin{bmatrix} \lambda_{i-1} \\ x_{i-1} \\ u_{i-1} \end{bmatrix} = \begin{bmatrix} f_i \\ 0 \\ 0 \end{bmatrix},$$

then again a special structure of the sequences of coefficient matrices (denoted in the following by $((\mathscr{K}_i),(\mathscr{N}_i),(\mathscr{M}_i))$) can be observed, with the middle coefficient being symmetric and the leading and last coefficient being transposes of each other, except that the index is shifted by 1. A triple of matrix sequences with such a structure will be called a *self-adjoint triple of matrix sequences*, see Section 4.

The paper is organized as follows. In Section 2 we recall the main results of the theoretical analysis for DAE optimal control problems as presented in [26] and [27]. In Section 3 necessary optimality conditions for the discrete-time optimal control problem (6) are derived. Then, in Section 4, we investigate self-conjugacy of difference operators and show that the operator associated with the discrete-time boundary value problem (8) fits into this framework. Since we obtain higher order difference equations in the discrete-time case we also discuss the related optimal control problem for higher order systems in Section 5, where also structure preserving first order representations for continuous- as well as discrete-time self-adjoint systems are studied. We close with some concluding remarks in Section 6.

2. Preliminaries

The theoretical basis for DAE optimal control problems has been studied in many different publications, see e. g., [2, 24, 27, 29, 34] and the references therein. We follow the approach in [24, 27] in a behavior setting, see [36], and first summarize some of the main results that are needed in the remainder of the paper.

The behavior approach proceeds by setting

$$\mathscr{E} = \begin{bmatrix} E & 0 \end{bmatrix}, \; \mathscr{A} = \begin{bmatrix} A & B \end{bmatrix}, \; z = \begin{bmatrix} x \\ u \end{bmatrix}$$

and considering the system (2) in the form

$$\mathscr{E}\dot{z} = \mathscr{A}z + f, \tag{10}$$

with initial condition $\begin{bmatrix} I_n & 0 \end{bmatrix} z(\underline{t}) = \underline{x}$. Following the presentation in [24, 27], we assume that the system (10) is already given in *regular strangeness-free form*, meaning that $\mathscr{E}$ and $\mathscr{A}$ are of the form

$$\mathscr{E} = \begin{bmatrix} E_1 & 0 \\ 0 & 0 \end{bmatrix}, \; \mathscr{A} = \begin{bmatrix} A_1 & B_1 \\ A_2 & B_2 \end{bmatrix}, \; f = \begin{bmatrix} f_1 \\ f_2 \end{bmatrix}$$

and satisfy the condition that $\begin{bmatrix} E_1 & 0 \\ A_2 & B_2 \end{bmatrix}$ is pointwise full row rank. A system with these properties can always be obtained using certain regularization techniques. For details, see [23, 24].

Since the use of adjoint equations is only reasonable for regular systems we restrict ourselves to this case. It has been shown in [23] that a regular strangeness-free system (10) has a well-defined *differentiation index* $\nu = 1$ for every sufficiently smooth input function u and every initial condition that is consistent with f and that the chosen input function fixes a unique solution.

For a Banach space formulation of (10), in [24] the Banach spaces $\mathbb{Z} = \mathbb{X} \times \mathbb{U}$ and $\mathbb{Y}$ were defined, where

$$\mathbb{X} = C^1_{E^+E}(\mathbb{I},\mathbb{R}^n) = \left\{x \in C(\mathbb{I},\mathbb{R}^n),\ E^+Ex \in C^1(\mathbb{I},\mathbb{R}^n)\right\}, \quad \mathbb{U} = C(\mathbb{I},\mathbb{R}^m),$$
$$\mathbb{Y} = C(\mathbb{I},\mathbb{R}^n) \times \operatorname{range} E(\underline{t})^T,$$

and E^+ denotes the Moore-Penrose inverse, see e. g. [17], of the matrix function $E = \begin{bmatrix} E_1 \\ 0 \end{bmatrix}$, together with the dual spaces

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

The linear quadratic optimal control problem (1), (2) can then be written as the abstract optimization problem

$$\text{minimize} \quad \frac{1}{2}\mathscr{Q}(z,z) \quad \text{subject to} \quad \mathscr{L}(z) = c, \tag{11}$$

with $z = \begin{bmatrix} x \\ u \end{bmatrix}$ and $c = \begin{bmatrix} f \\ E(\underline{t})^+E(\underline{t})\underline{x} \end{bmatrix}$, where $\mathscr{Q} : \mathbb{Z} \times \mathbb{Z} \to \mathbb{R}$ is a symmetric quadratic form defined by

$$\mathscr{Q}(v,z) = v(\bar{t})^T \begin{bmatrix} M_e & 0 \\ 0 & 0 \end{bmatrix} z(\bar{t}) + \int_{\mathbb{I}} v^T \begin{bmatrix} W & S \\ S^T & R \end{bmatrix} z\,dt,$$

and the linear operators $\mathscr{L} : \mathbb{Z} \to \mathbb{Y}$ and its conjugate $\mathscr{L}^* : \mathbb{Y}^* \to \mathbb{Z}^*$ are given by

$$\mathscr{L}(z) = \left(E\frac{d}{dt}(E^+Ex) - (A + E\frac{d}{dt}(E^+E))x - Bu, E(\underline{t})^+E(\underline{t})x(\underline{t})\right),$$
$$\mathscr{L}^*(\lambda,\gamma) = \left(-E^T\frac{d}{dt}(EE^+\lambda) - (A + EE^+\dot{E})^T\lambda, -B^T\lambda, \gamma - E(\underline{t})^T\lambda(\underline{t}), E(\bar{t})^T\lambda(\bar{t})\right).$$

It has been shown in [27] that with

$$\mathscr{R}(z) = (Wx + Su, S^Tx + Ru, 0, M_ex(\bar{t})) \in \mathbb{Z}^*,$$

and defining the operator

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

the necessary optimality conditions (3) can be written as

$$\mathscr{T}(\Lambda,z) = (c,0) \tag{12}$$

and that the operator $\mathscr{T}$ is self-conjugate. Note that (12) coincides with (3) if we assume sufficient smoothness of the data, see again [24].

REMARK 2. The discussed approach can be easily extended to linear higher order optimal control problems, where one minimizes

$$\mathscr{J}(x,u) = \frac{1}{2}x(\overline{t})^T M_e x(\overline{t}) + \frac{1}{2}\int_{\underline{t}}^{\overline{t}} \left(\sum_{j=0}^{k-1} (x^{(j)})^T W_j x^{(j)} + x^T Su + u^T S^T x + u^T Ru \right) dt, \tag{13a}$$

(with $k > 1$) subject to a constraint given by a k-th order differential-algebraic equation

$$\sum_{j=0}^{k} A_j x^{(j)} + Bu = f, \quad x(\underline{t}) = \underline{x}^0,\ \dot{x}(\underline{t}) = \underline{x}^1,\ \dots,\ x^{(k-1)}(\underline{t}) = \underline{x}^{k-1}. \tag{13b}$$

Here, $W_j = W_j^T \in C^0(\mathbb{I}, \mathbb{R}^{n,n})$ and $A_j \in C^0(\mathbb{I}, \mathbb{R}^{n,n})$ for $j = 0, \dots, k$. If the leading coefficient matrix A_k is pointwise nonsingular, then we can apply the classical procedure to turn (13a) and (13b) into first order systems by introducing new variables $w_i = x^{(i)}$ for $i = 0, \dots, k-1$, see also [35]. The formal necessary optimality conditions for the corresponding first order system then lead to a two-point boundary value problem. With $\lambda = \begin{bmatrix} \lambda_{k-1}^T & \dots & \lambda_0^T \end{bmatrix}^T$ partitioned as $w = \begin{bmatrix} w_0^T & \dots & w_{k-1}^T \end{bmatrix}^T$ we can rewrite the system again as a high order system in (x, μ), where $\mu = \lambda_{k-1}$, yielding a boundary value problem for $2(k-1)$th order equations of the form

$$\begin{aligned} \sum_{j=0}^{k} A_j x^{(j)} + Bu &= f, \\ \sum_{j=0}^{k} (-1)^j \frac{d^j}{dt^j}\left(A_j^T \mu\right) + \sum_{j=0}^{k-1} (-1)^{j-1} \frac{d^j}{dt^j}\left(W_j x^{(j)}\right) - Su &= 0, \\ -B^T \mu + S^T x + Ru &= 0, \end{aligned} \tag{14}$$

with boundary conditions

$$\begin{aligned} & x^{(i)}(\underline{t}) = \underline{x}^i, \quad i = 0, 1, \dots, k-1, \\ & 0 = \sum_{j=0}^{i} (-1)^j \frac{d^j}{dt^j}\left(A_{k-i+j}^T \mu\right) + \sum_{j=0}^{i-1} (-1)^{j+1} \frac{d^j}{dt^j}\left(W_{k-i+j} x^{(k-i+j)}\right)\Big|_{\overline{t}}, \ i = 0, \dots, k-2, \\ & 0 = \sum_{j=0}^{k-1} (-1)^j \frac{d^j}{dt^j}\left(A_{1+j}^T \mu\right) + \sum_{j=0}^{k-2} (-1)^{j+1} \frac{d^j}{dt^j}\left(W_{1+j} x^{(1+j)}\right)\Big|_{\overline{t}} - M_e x(\overline{t}). \end{aligned}$$

In this way, we can always construct an even order boundary value problem and the corresponding DAE operator is formally self-conjugate.

If the weighting matrices W_i are chosen to be zero for all $i > \frac{k-1}{2}$ if k is odd, and for all $i > \frac{k}{2}$ if k is even, then all coefficients in front of derivatives higher than k vanish. For constant coefficient problems (14) reduces to a system with an even matrix tuple of coefficients.

Note that when A_k in (13b) is singular, this approach cannot be applied in a formal straightforward way, because the first order formulation may change the index. In this case first a so-called *trimmed first order formulation* of the higher order system has to be considered, see [13, 39]. Then, for the trimmed first order formulation we can formulate the necessary optimality conditions and reformulation as a higher order boundary value problem leads again to a self-adjoint high order system.

After briefly recalling the results for the continuous-time case, in the next section we prove analogous results in the discrete-time case.

3. Necessary optimality conditions for discrete optimal control problems

In this section we derive necessary optimality conditions for the discrete-time optimal control problem (6) subject to (7). Similar results have been obtained in [34] for systems with constant coefficients and in [28] for system with properly stated leading term of tractability index one.

Again, we may assume without loss of generality that the difference equation (7) is already given in regular strangeness-free form, i. e., using the behavior approach by setting

$$\mathscr{E}_{i+1} = \begin{bmatrix} E_{i+1} & 0 \end{bmatrix}, \ \mathscr{A}_i = \begin{bmatrix} A_i & B_i \end{bmatrix}, \ z_i = \begin{bmatrix} x_i \\ u_i \end{bmatrix},$$

we consider the system (7) in the form

$$\mathscr{E}_{i+1} z_{i+1} = \mathscr{A}_i z_i + f_i, \quad i = 0, \ldots, N-1,$$

with coefficients

$$\mathscr{E}_{i+1} = \begin{bmatrix} E_{1,i+1} & 0 \\ 0 & 0 \end{bmatrix}, \ \mathscr{A}_i = \begin{bmatrix} A_{1,i} & B_{1,i} \\ A_{2,i} & B_{2,i} \end{bmatrix}, \ f_i = \begin{bmatrix} f_{1,i} \\ f_{2,i} \end{bmatrix}$$

that satisfy the condition

$$\begin{bmatrix} E_{1,i+1} & 0 \\ A_{2,i} & B_{2,i} \end{bmatrix} \quad \text{has full row rank for all } i = 0, \ldots, N-1.$$

Numerical methods for the computation of strangeness-free formulations of a discrete-time system (7) have been presented in [9, 10].

To derive the necessary optimality conditions we use the classical approach of appending the constraint equations (7) to the cost term by means of Lagrange multipliers and introducing the discrete functional

$$\begin{aligned} L((x_i),(u_i),(\lambda_i),\delta) = {} & \frac{1}{2} x_N^T M_e x_N + \frac{1}{2} \sum_{j=0}^{N-1} (x_j^T W_j x_j + x_j^T S_j u_j + u_j^T S_j^T x_j + u_j^T R_j u_j) \\ & + \sum_{j=0}^{N-1} (E_{j+1} x_{j+1} - A_j x_j - B_j u_j - f_j)^T \lambda_j + (E_0^+ E_0 x_0 - \underline{x})^T \delta. \end{aligned} \tag{15}$$

Here, as in [24], we apply the projection onto $\operatorname{cokernel} E_0$ for the initial value x_0 in order to meet the consistency requirements for algebraic components.

The necessary conditions for a minimum are given by the requirement that the gradients of L with respect to all unknowns vanish. We have the following gradients

$$\begin{aligned}
\nabla_{\lambda_i} L &= (E_{i+1}x_{i+1} - A_i x_i - B_i u_i - f_i)^T = 0, \quad i = 0,\dots,N-1,\\
\nabla_{x_0} L &= W_0 x_0 + S_0 u_0 - A_0^T \lambda_0 + (E_0^+ E_0)^T \delta = 0,\\
\nabla_{x_i} L &= W_i x_i + S_i u_i + E_i^T \lambda_{i-1} - A_i^T \lambda_i = 0, \quad i = 1,\dots,N-1,\\
\nabla_{x_N} L &= M_e x_N + E_N^T \lambda_{N-1} = 0,\\
\nabla_{u_i} L &= S_i^T x_i + R_i u_i - B_i^T \lambda_i = 0, \quad i = 0,\dots,N-1,\\
\nabla_{\delta} L &= (E_0^+ E_0 x_0 - \underline{x})^T = 0,
\end{aligned}$$

giving the necessary optimality conditions

$$E_{i+1}x_{i+1} - A_i x_i - B_i u_i - f_i = 0, \quad i = 0,\dots,N-1, \tag{16a}$$

$$W_i x_i + S_i u_i + E_i^T \lambda_{i-1} - A_i^T \lambda_i = 0, \quad i = 1,\dots,N-1, \tag{16b}$$

$$S_i^T x_i + R_i u_i - B_i^T \lambda_i = 0, \quad i = 0,\dots,N-1, \tag{16c}$$

together with the boundary conditions

$$W_0 x_0 + S_0 u_0 - A_0^T \lambda_0 + (E_0^+ E_0)^T \delta = 0, \tag{17a}$$

$$M_e x_N + E_N^T \lambda_{N-1} = 0, \tag{17b}$$

$$E_0^+ E_0 x_0 = \underline{x}. \tag{17c}$$

These necessary conditions can be written (in a rather formal way) as a three term recursion of the form

$$\begin{bmatrix} 0 & E_{i+1} & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \lambda_{i+1}\\ x_{i+1}\\ u_{i+1}\\ \delta \end{bmatrix} + \begin{bmatrix} 0 & -A_i & -B_i & 0\\ -A_i^T & W_i & S_i & 0\\ -B_i^T & S_i^T & R_i & 0\\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \lambda_i\\ x_i\\ u_i\\ \delta \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & 0\\ E_i^T & 0 & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \lambda_{i-1}\\ x_{i-1}\\ u_{i-1}\\ \delta \end{bmatrix} = \begin{bmatrix} f_i\\ 0\\ 0\\ 0 \end{bmatrix}, \tag{18}$$

for $i = 1,\dots,N-1$, with boundary conditions

$$\begin{bmatrix} 0 & E_1 & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \lambda_1\\ x_1\\ u_1\\ \delta \end{bmatrix} + \begin{bmatrix} 0 & -A_0 & -B_0 & 0\\ -A_0^T & W_0 & S_0 & (E_0^+E_0)^T\\ -B_0^T & S_0^T & R_0 & 0\\ 0 & E_0^+E_0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \lambda_0\\ x_0\\ u_0\\ \delta \end{bmatrix} = \begin{bmatrix} f_0\\ 0\\ 0\\ \underline{x} \end{bmatrix},$$

and

$$\begin{bmatrix} 0 & 0 & 0 & 0\\ 0 & M_e & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \lambda_N\\ x_N\\ u_N\\ \delta \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & 0\\ E_N^T & 0 & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \lambda_{N-1}\\ x_{N-1}\\ u_{N-1}\\ \delta \end{bmatrix} = \begin{bmatrix} 0\\ 0\\ 0\\ 0 \end{bmatrix}.$$

Here, the additional Lagrange multiplier δ that is used to couple the initial condition to the functional (15), in general, is chosen as $\delta = 0$. Since δ is of no concern in (18), in the following we will omit the last block row and column of (18).

If we look at the structure of the system (18), then we observe that the middle term is symmetric while the leading term is the transpose of the last term with the index shifted by one.

REMARK 3. In a similar fashion we can treat the discrete-time optimal control problem of minimizing

$$\mathscr{J}_d((x_\ell),(u_\ell)) = \frac{1}{2}x_N^T M_e x_N + \frac{1}{2}\sum_{j=0}^{N-1}\left(x_j^T W_j x_j + x_j^T S_j u_j + u_j^T S_j^T x_j + u_j^T R_j u_j\right), \tag{19a}$$

subject to the k-th order discrete-time control system

$$\sum_{i=0}^{k} M_{i+j}^{[i]} x_{i+j} + B_j u_j = f_j, \quad j = 0,1,\ldots,N-k, \tag{19b}$$

with given starting values for $x_0, x_1, \ldots, x_{k-1} \in \mathbb{R}^n$ and coefficient matrices $M_j^{[i]} \in \mathbb{R}^{n,n}$ for $i = 0,\ldots,k$, $B_j \in \mathbb{R}^{n,m}$, $j = 0,\ldots,N$, see e. g. [11] for the constant coefficient case. In this case the Lagrangian takes the form

$$\begin{aligned} L((x_\ell),(u_\ell),(\lambda_\ell),\delta) &= \frac{1}{2}x_N^T M_e x_N + \frac{1}{2}\sum_{j=0}^{N-1}(x_j^T W_j x_j + x_j^T S_j u_j + u_j^T S_j^T x_j + u_j^T R_j u_j) \\ &+ \sum_{j=0}^{N-k}\left(\sum_{i=0}^{k} M_{i+j}^{[i]} x_{i+j} + B_j u_j - f_j\right)^T \lambda_j + \sum_{j=0}^{k-1}\left((M_j^{[j]})^+ M_j^{[j]} x_j - \underline{x}_j\right)^T \delta_j, \end{aligned} \tag{20}$$

and the necessary optimality conditions are given by

$$\begin{aligned} \nabla_{\lambda_\ell} L &= \left(\sum_{i=0}^{k} M_{i+\ell}^{[i]} x_{i+\ell} + B_\ell u_\ell - f_\ell\right)^T = 0, \quad \ell = 0,\ldots,N-k, \\ \nabla_{x_\ell} L &= W_\ell x_\ell + S_\ell u_\ell + \sum_{i=0}^{k-\ell-1} M_\ell^{[i]T} \lambda_{\ell-i} + \left((M_\ell^{[\ell]})^+ M_\ell^{[\ell]}\right)^T \delta_\ell = 0, \quad \ell = 0,\ldots,k-1, \\ \nabla_{x_\ell} L &= W_\ell x_\ell + S_\ell u_\ell + \sum_{i=0}^{k} M_\ell^{[i]T} \lambda_{\ell-i} = 0, \quad \ell = k,\ldots,N-k, \\ \nabla_{x_\ell} L &= W_\ell x_\ell + S_\ell u_\ell + \sum_{i=0}^{k} M_\ell^{[i]T} \lambda_{\ell-i} = 0, \quad \ell = N-k+1,\ldots,N-1, \\ \nabla_{x_N} L &= M_e x_N + \left(M_N^{[k]}\right)^T \lambda_{N-k} = 0, \\ \nabla_{u_\ell} L &= S_\ell^T x_\ell + R_\ell u_\ell + B_\ell^T \lambda_\ell = 0, \quad \ell = 0,\ldots,N-k, \end{aligned}$$

$$\nabla_{u_\ell} L = S_\ell^T x_\ell + R_\ell u_\ell = 0, \quad \ell = N-k+1,\dots,N-1,$$
$$\nabla_{\delta_j} L = \left((M_j^{[j]})^+ M_j^{[j]} x_j - \underline{x}_j\right)^T = 0, \quad j = 0,\dots,k-1.$$

This yields the optimality boundary value problem

$$\begin{bmatrix} 0 & M_{k+\ell}^{[k]} & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \lambda_{\ell+k} \\ x_{\ell+k} \\ u_{\ell+k} \end{bmatrix} + \ldots + \begin{bmatrix} 0 & M_{1+\ell}^{[1]} & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \lambda_{\ell+1} \\ x_{\ell+1} \\ u_{\ell+1} \end{bmatrix} + \begin{bmatrix} 0 & M_\ell^{[0]} & B_\ell \\ (M_\ell^{[0]})^T & W_\ell & S_\ell \\ B_\ell^T & S_\ell^T & R_\ell \end{bmatrix} \begin{bmatrix} \lambda_\ell \\ x_\ell \\ u_\ell \end{bmatrix}$$
$$+ \begin{bmatrix} 0 & 0 & 0 \\ (M_\ell^{[1]})^T & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \lambda_{\ell-1} \\ x_{\ell-1} \\ u_{\ell-1} \end{bmatrix} + \ldots + \begin{bmatrix} 0 & 0 & 0 \\ (M_\ell^{[k]})^T & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \lambda_{\ell-k} \\ x_{\ell-k} \\ u_{\ell-k} \end{bmatrix} = \begin{bmatrix} f_\ell \\ 0 \\ 0 \end{bmatrix},$$

for $\ell = k,\dots,N-k$, together with the corresponding boundary conditions. (Note that, as before, we have omitted the variables δ_j for better readability.) Again, we observe a symmetry of the middle coefficient, while the leading and final coefficients have a transposed structure with shifted indices.

In the following, we will show that the difference operator arising in the optimality system (18) is self-conjugate with respect to suitably chosen dual systems and corresponding Banach spaces.

4. Self-conjugate difference operators

In order to show that the difference operator arising in the optimality system (18) is self-conjugate, we adapt the proof from the continuous-time case in [27] to the discrete-time case. As in the continuous-time case we restrict ourselves to regular and strangeness-free systems. Then, we can rewrite the discrete optimal control problem (6), (7) as

$$\text{minimize} \quad \frac{1}{2}\mathcal{Q}_d((z_i),(z_i)) \quad \text{subject to} \quad \mathcal{L}_d((z_i)) = (c_i),$$

with $(z_i) = \left(\begin{bmatrix} x_i \\ u_i \end{bmatrix}\right)$ and $(c_i) = \left(\begin{bmatrix} f_i \\ \underline{x} \end{bmatrix}\right)$, where $\mathcal{Q}_d : \mathbb{Z}_d \times \mathbb{Z}_d \to \mathbb{R}$ is a discrete symmetric quadratic form defined by

$$\mathcal{Q}_d((v_i),(z_i)) = v_N^T \begin{bmatrix} M_e & 0 \\ 0 & 0 \end{bmatrix} z_N + \sum_{j=0}^{N-1} v_j^T \begin{bmatrix} W_j & S_j \\ S_j^T & R_j \end{bmatrix} z_j$$

and sequence spaces $\mathbb{Z}_d = \mathbb{X}_d \times \mathbb{U}_d$, where $\mathbb{X}_d, \mathbb{U}_d$, $\mathbb{Y}_d$ are given by

$$\mathbb{X}_d = \mathbb{R}^n_{0,N}, \quad \mathbb{U}_d = \mathbb{R}^m_{0,N-1} \quad \text{and} \quad \mathbb{Y}_d = \mathbb{R}^n_{0,N-1} \times \operatorname{range} E_0^T. \tag{21}$$

In view of the results from Section 3, we obtain that the linear difference operator $\mathcal{L}_d : \mathbb{Z}_d \to \mathbb{Y}_d$ for the constraint (7) is given by

$$\mathcal{L}_d((z_i)) = (E_{i+1}x_{i+1} - A_i x_i - B_i u_i, E_0^+ E_0 x_0). \tag{22}$$

In the next step we need to define dual systems $\langle \mathbb{Z}_d, \mathbb{Z}_d^* \rangle$ and $\langle \mathbb{Y}_d, \mathbb{Y}_d^* \rangle$. Then, it follows that the operator $\mathscr{L}_d$ has a unique conjugate operator $\mathscr{L}_d^* : \mathbb{Y}_d^* \to \mathbb{Z}_d^*$ (see also [27]). Keeping in mind the necessary optimality conditions (16), we define the spaces

$$\begin{aligned} \mathbb{Z}_d^* &= \mathbb{R}^n_{1,N-1} \times \mathbb{R}^m_{0,N-1} \times \operatorname{range} E_0^T \times \operatorname{range} E_N^T, \\ \mathbb{Y}_d^* &= \mathbb{R}^n_{0,N-1} \times \operatorname{range} E_0^T, \end{aligned} \tag{23}$$

to obtain the bilinear systems $\langle \mathbb{Z}_d, \mathbb{Z}_d^* \rangle$ and $\langle \mathbb{Y}_d, \mathbb{Y}_d^* \rangle$ with the corresponding bilinear forms

$$\langle (z_i), ((\eta_i), (\vartheta_i), \delta, \varepsilon) \rangle = \sum_{j=1}^{N-1} \eta_j^T x_j + \sum_{j=0}^{N-1} \vartheta_j^T u_j + \delta^T x_0 + \varepsilon^T x_N, \tag{24}$$

for $(z_i) \in \mathbb{Z}_d$, $((\eta_i), (\vartheta_i), \delta, \varepsilon) \in \mathbb{Z}_d^*$, and

$$\langle ((g_i), r), ((\lambda_i), \gamma) \rangle = \sum_{j=0}^{N-1} \lambda_j^T g_j + \gamma^T r, \tag{25}$$

for $((g_i), r) \in \mathbb{Y}_d$, and $((\lambda_i), \gamma) \in \mathbb{Y}_d^*$. In the following, we show that these bilinear systems are dual systems, i. e., the corresponding bilinear forms are non-degenerate.

THEOREM 1. *The bilinear systems $\langle \mathbb{Z}_d, \mathbb{Z}_d^* \rangle$ and $\langle \mathbb{Y}_d, \mathbb{Y}_d^* \rangle$ with sequence spaces as in (21), (23) and corresponding bilinear forms as in (24), (25) are dual systems.*

Proof. Consider the bilinear system $\langle \mathbb{Y}_d, \mathbb{Y}_d^* \rangle$ with its bilinear form given in (25). In the following, we use the standard observation that if $(f_i) \in \mathbb{R}^n_{0,N}$ and $\langle (f_i), (g_i) \rangle = \sum_{j=0}^N f_j^T g_j = 0$ for all $(g_i) \in \mathbb{R}^n_{0,N}$, then $f_i = 0$ for all $i = 0, \ldots, N$. Let $y^* = ((\lambda_i), \gamma) \in \mathbb{Y}_d^*$ be fixed and assume that

$$\langle y, y^* \rangle = \sum_{j=0}^{N-1} \lambda_j^T g_j + \gamma^T r = 0$$

for all $y = ((g_i), r) \in \mathbb{Y}_d$. Choosing $g_i = 0$ for all $i = 0, \ldots, N-1$ and $r = \gamma$ gives $\gamma^T \gamma = 0$, hence $\gamma = 0$. Therefore, $\sum_{j=0}^{N-1} \lambda_j^T g_j = 0$ for all $(g_i) \in \mathbb{R}^n_{0,N}$, and hence $\lambda_j = 0$ for all $j = 0, \ldots, N-1$.

Let $y = ((g_i), r) \in \mathbb{Y}_d$ be fixed and assume that

$$\langle y, y^* \rangle = \sum_{j=0}^{N-1} \lambda_j^T g_j + \gamma^T r = 0$$

for all $y^* = ((\lambda_i), \gamma) \in \mathbb{Y}_d^*$. Choosing $\lambda_i = 0$ for all $i = 0, \ldots, N-1$ and $\gamma = r$ gives $r^T r = 0$, hence $r = 0$. Therefore, $\sum_{j=0}^{N-1} \lambda_j^T g_j = 0$ for all $(\lambda_i) \in \mathbb{R}^n_{0,N-1}$, where $(g_i) \in \mathbb{R}^n_{0,N-1}$ and hence, $g_j = 0$ for $j = 0, \ldots, N-1$.

The proof for $\langle \mathbb{Z}_d, \mathbb{Z}_d^* \rangle$ follows the same lines. $\square$

If $\langle \mathbb{Z}_d, \mathbb{Z}_d^* \rangle$ is a dual system, then we know that the operator $\mathscr{L}_d$ has a unique conjugate operator $\mathscr{L}_d^* : \mathbb{Y}_d^* \to \mathbb{Z}_d^*$ (see also [27]) that is given by

$$\mathscr{L}_d^*((\lambda_i),\gamma) = ((E_i^T\lambda_{i-1} - A_i^T\lambda_i),(-B_i^T\lambda_i),\gamma - A_0^T\lambda_0, E_N^T\lambda_{N-1}). \tag{26}$$

THEOREM 2. *The operator $\mathscr{L}_d^* : \mathbb{Y}_d^* \to \mathbb{Z}_d^*$ defined by (26) is the unique conjugate of $\mathscr{L}_d : \mathbb{Z}_d \to \mathbb{Y}_d$ defined by (22).*

Proof. Let $(z_i) = ((x_i),(u_i)) \in \mathbb{Z}_d$ and $\Lambda = ((\lambda_i),\gamma) \in \mathbb{Y}_d^*$. Using that $E_0^+E_0\gamma = \gamma$ (since $\gamma \in \operatorname{range} E_0^T$ and $E_0^+E_0$ is a projector onto cokernel$(E_0) = \operatorname{range}(E_0^T)$), we have

$$\begin{aligned}
\langle \mathscr{L}_d(z_i),\Lambda\rangle &= \sum_{j=0}^{N-1}\lambda_j^T\left(E_{j+1}x_{j+1} - A_jx_j - B_ju_j\right) + \gamma^T E_0^+E_0x_0\\
&= \sum_{j=1}^{N-1}(\lambda_{j-1}^TE_jx_j - \lambda_j^TA_jx_j) + \lambda_{N-1}^TE_Nx_N - \sum_{j=0}^{N-1}\lambda_j^TB_ju_j - \lambda_0^TA_0x_0 + \gamma^TE_0^+E_0x_0\\
&= \sum_{j=1}^{N-1}(E_j^T\lambda_{j-1} - A_j^T\lambda_j)^Tx_j + \sum_{j=0}^{N-1}(-B_j^T\lambda_j)^Tu_j + (\gamma - A_0^T\lambda_0)^Tx_0 + (E_N^T\lambda_{N-1})^Tx_N\\
&= \langle (z_i),\mathscr{L}_d^*(\Lambda)\rangle. \quad \square
\end{aligned}$$

Finally, we can define an operator $\mathscr{T}_d : \mathbb{Y}_d^* \times \mathbb{Z}_d \to \mathbb{Y}_d \times \mathbb{Z}_d^*$ of the form

$$\mathscr{T}_d(\Lambda,(z_i)) = (\mathscr{L}_d(z_i),\mathscr{L}_d^*(\Lambda) + \mathscr{R}_d(z_i)), \tag{27}$$

with

$$\mathscr{R}_d(z_i) = \big((W_ix_i + S_iu_i),(S_i^Tx_i + R_iu_i),W_0x_0 + S_0u_0,M_ex_N\big) \in \mathbb{Z}_d^*.$$

That means, for $(z_i) = ((x_i),(u_i)) \in \mathbb{Z}_d$ and $\Lambda = ((\lambda_i),\gamma) \in \mathbb{Y}_d^*$ we have

$$\begin{aligned}
\mathscr{T}_d(\Lambda,(z_i)) = \big(&(E_{i+1}x_{i+1} - A_ix_i - B_iu_i), E_0^+E_0x_0,\ (E_i^T\lambda_{i-1} - A_i^T\lambda_i + W_ix_i + S_iu_i),\\
&(-B_i^T\lambda_i + S_i^Tx_i + R_iu_i),\gamma - A_0^T\lambda_0 + W_0x_0 + S_0u_0, E_N^T\lambda_{N-1} + M_ex_N\big)
\end{aligned}$$

and with $\gamma = (E_0^+E_0)^T\delta$ the necessary conditions (16), (17) can be written as

$$\mathscr{T}_d(\Lambda,(z_i)) = ((c_i),0). \tag{28}$$

In order to show that the operator $\mathscr{T}_d$ is self-conjugate we introduce the spaces

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

and set $\mathbb{V}_d^* = \mathbb{W}_d$, $\mathbb{W}_d^* = \mathbb{V}_d$. Then, by construction, we have $\mathscr{T}_d : \mathbb{V}_d \to \mathbb{W}_d$ and also $\mathscr{T}_d : \mathbb{W}_d^* \to \mathbb{V}_d^*$. Obviously, the pairs $\langle \mathbb{V}_d, \mathbb{V}_d^* \rangle$ and $\langle \mathbb{W}_d, \mathbb{W}_d^* \rangle$ are dual systems with respect to the so-called canonical bilinear form

$$\langle (y^*,z),(y,z^*)\rangle = \langle y,y^*\rangle + \langle z,z^*\rangle = \langle (y,z^*),(y^*,z)\rangle. \tag{29}$$

THEOREM 3. *The operator $\mathscr{T}_d$ as defined in (27) is self-conjugate with respect to the canonical bilinear form (29), i. e., we have*

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

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

$$\begin{aligned}\langle \mathscr{T}_d(\Lambda, (z_i)), (\tilde{\Lambda}, (\tilde{z}_i))\rangle &= \langle (\mathscr{L}_d((z_i)), \mathscr{L}_d^*(\Lambda) + \mathscr{R}_d((z_i))), (\tilde{\Lambda}, (\tilde{z}_i))\rangle \\ &= \langle \mathscr{L}_d((z_i)), \tilde{\Lambda}\rangle + \langle (\tilde{z}_i), \mathscr{R}_d((z_i))\rangle + \langle (\tilde{z}_i), \mathscr{L}_d^*(\Lambda)\rangle,\end{aligned}$$

as well as

$$\begin{aligned}\langle (\Lambda, (z_i)), \mathscr{T}_d(\tilde{\Lambda}, (\tilde{z}_i))\rangle &= \langle (\Lambda, (z_i)), (\mathscr{L}_d((\tilde{z}_i)), \mathscr{L}_d^*(\tilde{\Lambda}) + \mathscr{R}_d((\tilde{z}_i)))\rangle \\ &= \langle \mathscr{L}_d((\tilde{z}_i)), \Lambda\rangle + \langle (z_i), \mathscr{R}_d((\tilde{z}_i))\rangle + \langle (z_i), \mathscr{L}_d^*(\tilde{\Lambda})\rangle.\end{aligned}$$

Since $\mathscr{L}_d^*$ is the conjugate of $\mathscr{L}_d$ and because of

$$\langle (\tilde{z}_i), \mathscr{R}_d((z_i))\rangle = Q_d((z_i), (\tilde{z}_i)) = Q_d((\tilde{z}_i), (z_i)) = \langle (z_i), \mathscr{R}_d((\tilde{z}_i))\rangle$$

due to the symmetry of $\mathscr{Q}_d$, the two expressions are equivalent. □

We want to emphasize again, that (28) coincides with (16), (17). In particular, the optimality system (16) can be written as (18) with the corresponding boundary conditions, i. e., as a three-term recursion of the form

$$\mathscr{K}_i v_{i+1} + \mathscr{N}_i v_i + \mathscr{M}_i v_{i-1} = g_i, \quad i = 1, \dots, N-1, \tag{30}$$

with $\mathscr{K}_i, \mathscr{N}_i, \mathscr{M}_i \in \mathbb{R}^{\ell,\ell}$ and inhomogeneity $g_i \in \mathbb{R}^{\ell}$ for all i, together with the boundary conditions

$$\begin{aligned}\mathscr{K}_0 v_1 + \mathscr{N}_0 v_0 &= g_0, \\ \mathscr{N}_N v_N + \mathscr{M}_N v_{N-1} &= g_N.\end{aligned} \tag{31}$$

This observation leads to the following definition.

DEFINITION 1. Let $((\mathscr{K}_i), (\mathscr{N}_i), (\mathscr{M}_i))$ be a triple of matrix sequences in $\mathbb{R}^{\ell,\ell}_{0,N}$ with boundary terms $\mathscr{K}_N = 0$ and $\mathscr{M}_0 = 0$, then the triple $\big((\mathscr{M}_{i+1}^T), (\mathscr{N}_i^T), (\mathscr{K}_{i-1}^T)\big)$ of matrix sequences in $\mathbb{R}^{\ell,\ell}_{0,N}$ with boundary terms $\mathscr{K}_{-1}^T = 0$ and $\mathscr{M}_{N+1}^T = 0$ is called the *adjoint triple* of $((\mathscr{K}_i), (\mathscr{N}_i), (\mathscr{M}_i))$.

We have the following property of adjoint triples.

PROPOSITION 1. *Let $((\mathscr{K}_i), (\mathscr{N}_i), (\mathscr{M}_i))$ have the adjoint triple $((\mathscr{M}_{i+1}^T), (\mathscr{N}_i^T), (\mathscr{K}_{i-1}^T))$. Then, the matrix triple $\big((\mathscr{M}_{i+1}^T), (\mathscr{N}_i^T), (\mathscr{K}_{i-1}^T)\big)$ has an adjoint triple which is given by $((\mathscr{K}_i), (\mathscr{N}_i), (\mathscr{M}_i))$.*

Proof. The adjoint of $\big((\mathscr{M}_{i+1}^T),(\mathscr{N}_i^T),(\mathscr{K}_{i-1}^T)\big)$ is given by

$$\big(((\mathscr{K}_{i-1+1}^T)^T),((\mathscr{N}_i^T)^T),((\mathscr{M}_{i+1-1}^T)^T)\big) = ((\mathscr{K}_i),(\mathscr{N}_i),(\mathscr{M}_i)). \quad \square$$

This observation leads to the definition of self-adjoint triples of matrix sequences.

DEFINITION 2. A triple of matrix sequences $((\mathscr{K}_i),(\mathscr{N}_i),(\mathscr{M}_i))$ in $\mathbb{R}_{0,N}^{\ell,\ell}$, is called *self-adjoint* if the following two conditions are satisfied

$$\mathscr{K}_i = \mathscr{M}_{i+1}^T \quad \text{and} \quad \mathscr{N}_i = \mathscr{N}_i^T \quad \text{for } i = 0,\dots,N \tag{32}$$

with boundary terms $\mathscr{K}_N = \mathscr{M}_{N+1}^T = 0$ and $\mathscr{M}_0 = \mathscr{K}_{-1}^T = 0$.

Note that for a triple of constant matrices, condition (32) reduces to $\mathscr{M} = \mathscr{K}^T$ and $\mathscr{N} = \mathscr{N}^T$, i. e., in this case a self-adjoint triple corresponds to a so-called *palindromic matrix triple* $(\mathscr{M},\mathscr{N},\mathscr{M}^T)$, see [31].

A self-conjugate system of the form

$$\mathscr{M}_{i+1}^T v_{i+1} + \mathscr{N}_i v_i + \mathscr{M}_i v_{i-1} = g_i, \quad i = 1,\dots,N-1, \tag{33}$$

with boundary conditions as in (31) can always be written in the form

$$\begin{bmatrix} \mathscr{N}_0 & \mathscr{M}_1^T & & & & \\ \mathscr{M}_1 & \mathscr{N}_1 & \mathscr{M}_2^T & & & \\ & \mathscr{M}_2 & \mathscr{N}_2 & \mathscr{M}_3^T & & \\ & & \ddots & \ddots & \ddots & \\ & & & \mathscr{M}_{N-1} & \mathscr{N}_{N-1} & \mathscr{M}_N^T \\ & & & & \mathscr{M}_N & \mathscr{N}_N \end{bmatrix} \begin{bmatrix} v_0 \\ v_1 \\ v_2 \\ \vdots \\ v_{N-1} \\ v_N \end{bmatrix} = \begin{bmatrix} g_0 \\ g_1 \\ g_2 \\ \vdots \\ g_{N-1} \\ g_N \end{bmatrix} \tag{34}$$

with symmetric system matrix.

REMARK 4. The described concept of self-conjugate difference operators is in accordance with self-conjugate difference equations given in the form

$$\mathscr{L}_d((x_i)) = (\Delta[P_i \Delta x_{i-1}] + Q_i x_i)_i,$$

where $\Delta x_i := x_{i+1} - x_i$, with $P_i = P_i^T$ and $Q_i = Q_i^T$, see e. g., [1, 21]. Here we have $\mathscr{L}_d^{**} = \mathscr{L}_d$.

REMARK 5. We can also consider linear difference operators of order $k = 2\mu$, $\mu \in \mathbb{N}$ defined by

$$\mathscr{L}_d : \mathbb{V} \to \mathbb{W}, \quad \mathscr{L}_d((x_i)) = \sum_{j=0}^{k} A_j(i) x_{i-\mu+j}, \quad \text{for all } i \in \mathscr{I}, \tag{35}$$

for an index set $\mathscr{I} \subset \mathbb{Z}$, with matrices $A_j(i) \in \mathbb{R}^{n,n}$ for $j = 0,\dots,k$ defined for all i and sequence spaces $\mathbb{V}$ and $\mathbb{W}$ given by

$$\begin{aligned}\mathbb{V} &= \{(x_i)_{i\in\mathscr{I}}, x_i \in \mathbb{R}^n |\, B_j((x_i)) = 0 \text{ for } j = 0,\dots,\mu-1\},\\ \mathbb{W} &= \{(y_i)_{i\in\mathscr{I}}, y_i \in \mathbb{R}^n\}.\end{aligned}$$

With index set $\mathscr{I} = \{-\mu,\dots,N+\mu\}$ the boundary terms are given by

$$\begin{aligned}B_j((x_i)) = \{&A^+_{k-j}(i-\mu+j)A_{k-j}(i-\mu+j)x_i = 0 \text{ for } i = N+1,\dots,N+\mu-j,\\ &A^+_j(i)A_j(i)x_{i-\mu+j} = 0 \text{ for } i = 0,\dots,\mu-1-j\}.\end{aligned}$$

Then, the (formal) adjoint operator $\mathscr{L}^*_d : \mathbb{W}^* \to \mathbb{V}^*$ is given by

$$\mathscr{L}^*_d((y_i)) = \sum_{j=0}^{k} A^T_{k-j}(i-\mu+j)y_{i-\mu+j},$$

with sequence spaces

$$\begin{aligned}\mathbb{V}^* &= \{(x_i)_{i\in\mathscr{I}}, x_i \in \mathbb{R}^n\},\\ \mathbb{W}^* &= \left\{(y_i)_{i\in\mathscr{I}}, y_i \in \mathbb{R}^n \mid B^*_j((y_i)) = 0 \text{ for } j = 0,\dots,\mu-1\right\}\end{aligned}$$

and boundary conditions

$$\begin{aligned}B^*_j((y_i)) = \{&A_{k-j}(i-\mu+j)A^+_{k-j}(i-\mu+j)y_{i-\mu+j} = 0 \text{ for } i = 0,\dots,\mu-1-j,\\ &A_j(i)A^+_j(i)y_i = 0 \text{ for } i = N+1,\dots,N+\mu-j\}.\end{aligned}$$

The difference operator (35) is (formally) self-conjugate if and only if

$$\mathbb{V} = \{(x_i)_{i\in\mathscr{I}}, x_i \in \mathbb{R}^n |\, B_j((x_i)) = B^*_j((x_i)) = 0 \text{ for all } j = 0,\dots,\mu-1\}$$

and

$$A_j(i) = A^T_{k-j}(i+j-\mu) \quad \text{for all } j = 0,\dots,k,\ i \in \mathscr{I}_0 = \{0,\dots,N\}. \tag{36}$$

For constant coefficient systems, the condition of self-conjugacy (36) again reduces to $A_j = A^T_{k-j}$ for $j = 0,\dots,k$ and thus a self-conjugate difference operator is given by a palindromic system

$$\mathscr{L}_d(x) = A_0x_{i-\mu} + A_1x_{i-\mu+1} + \ldots + A_\mu x_i + \ldots + A^T_1 x_{i+\mu-1} + A^T_0 x_{i+\mu}.$$

Following [9, 10] we can simplify matrix sequences associated with the coefficients of difference equations (30) by equivalence transformations that consist of scaling the equation (30) with nonsingular matrices $P_i \in \mathbb{R}^{\ell,\ell}$ and by performing a change of variables $v_i = Q_i y_i$ with nonsingular matrices $Q_i \in \mathbb{R}^{\ell,\ell}$. This gives a transformed difference equation

$$\tilde{K}_i y_{i+1} + \tilde{\mathscr{N}}_i y_i + \tilde{\mathscr{M}}_i y_{i-1} = P_i g_i,$$

with

$$\tilde{\mathscr{K}}_i = P_i \mathscr{K}_i Q_{i+1}, \quad \tilde{\mathscr{N}}_i = P_i \mathscr{N}_i Q_i, \quad \tilde{\mathscr{M}}_i = P_i \mathscr{M}_i Q_{i-1}.$$

Taking a look at the behavior of the adjoint of the triple of matrix sequences under equivalence transformations and assuming that $((\mathscr{K}_i), (\mathscr{N}_i), (\mathscr{M}_i))$ possesses an adjoint triple, we see that $((\tilde{\mathscr{K}}_i), (\tilde{\mathscr{N}}_i), (\tilde{\mathscr{M}}_i))$ possesses an adjoint triple as well, which is given by

$$((\tilde{\mathscr{M}}_{i+1}^T), (\tilde{\mathscr{N}}_i^T), (\tilde{\mathscr{K}}_{i-1}^T)) = ((Q_i^T \mathscr{M}_{i+1}^T P_{i+1}^T), (Q_i^T \mathscr{N}_i^T P_i^T), (Q_i^T \mathscr{K}_{i-1}^T P_{i-1}^T)),$$

i. e., the adjoint triple of the transformed triple is equivalent to the adjoint triple of the original triple.

In order to preserve self-conjugacy of the operator, i. e., self-adjointness of the triple of coefficient sequences, we have to preserve the symmetry of $\mathscr{N}_i$ and, hence, we have to require that $P_i = Q_i^T$, i. e., that the transformation is a *(time-varying) congruence transformation*. We then have the following Lemma.

LEMMA 1. *Consider a self-adjoint triple of matrix sequences* $((\mathscr{K}_i), (\mathscr{N}_i), (\mathscr{M}_i))$ *with* $\mathscr{K}_i, \mathscr{N}_i, \mathscr{M}_i \in \mathbb{R}^{\ell,\ell}$ *and apply a congruence transformation with a sequence of nonsingular* $Q_i \in \mathbb{R}^{\ell,\ell}$*, leading to the triple*

$$((\tilde{\mathscr{K}}_i), (\tilde{\mathscr{N}}_i), (\tilde{\mathscr{M}}_i)) = ((Q_i^T \mathscr{K}_i Q_{i+1}), (Q_i^T \mathscr{N}_i Q_i), (Q_i^T \mathscr{M}_i Q_{i-1})).$$

Then the triple $((\tilde{\mathscr{K}}_i), (\tilde{\mathscr{N}}_i), (\tilde{\mathscr{M}}_i))$ *is again self-adjoint.*

Proof. The condition for $\tilde{\mathscr{N}}_i$ is trivially satisfied and for $\tilde{\mathscr{K}}_i$ and $\tilde{\mathscr{M}}_i$ we get

$$\tilde{\mathscr{K}}_i = Q_i^T \mathscr{K}_i Q_{i+1} = Q_i^T \mathscr{M}_{i+1}^T Q_{i+1} = \tilde{\mathscr{M}}_{i+1}^T. \quad \square$$

In order to understand the solution behavior of linear matrix sequences, one usually computes canonical or condensed forms under the associated equivalence transformation. For constant matrix pairs the general canonical form under equivalence is given by the Kronecker canonical form see, e. g., [16] and the condensed form is the staircase or GUPTRI form [14, 15, 38]. The canonical form under congruence transformations for even pencils has been given in [37] and the condensed form in [12]. For palindromic pencils this form has been derived in [20]. The canonical form for time varying pairs under equivalence has been presented in [10]. For matrix triples such canonical forms in general are not known even for constant triples. Recently, a condensed form which reveals partial information has been presented [13], as well as special structured Smith forms [33, 32]. For systems with variable coefficients such canonical or condensed forms are an open problem.

5. Structure preserving first order formulations

The problem of deriving structure preserving first order formulations for higher order systems has been an active research field in the last years, see e. g. [11, 31]. Since often numerical software is only available for first order systems, it is important to preserve the specific structure of a given problem when it is transformed into an equivalent first order formulation. In this section we discuss first order formulations in the case of systems with self-adjoint coefficient triples.

Consider a linear k-th order differential-algebraic operator of the form

$$\mathscr{L}:\mathbb{Z}\to\mathbb{Y},\quad z\mapsto\mathscr{L}(z)=\sum_{i=0}^{k}A_iz^{(i)}, \tag{37}$$

with a tuple $(A_k,\ldots,A_0)$ of sufficiently smooth coefficient functions $A_i\in C^0(\mathbb{I},\mathbb{R}^{n,n})$ and function spaces

$$\begin{aligned}
&\mathbb{Z}=\{z\in C^0(\mathbb{I},\mathbb{R}^n)\,|A_k^+A_kz\in C^k(\mathbb{I},\mathbb{R}^n),\ B_i(z,\underline{t})=0,\quad i=1,\ldots,k\},\\
&\mathbb{Y}=C^0(\mathbb{I},\mathbb{R}^n),
\end{aligned}$$

with boundary conditions given by

$$B_i(z,\underline{t})=\left\{(A_i^+A_i)^{(\ell)}z^{(i-j-1)}|_{\underline{t}}=0,\ \text{for}\ j=0,\ldots,i-1,\ \ell=0,\ldots,j\right\}.$$

The unique conjugate operator $\mathscr{L}^*:\mathbb{Y}^*\to\mathbb{Z}^*$ is then given by

$$\mathscr{L}^*(y)=\sum_{i=0}^{k}(-1)^i\frac{d^i}{dt^i}(A_i^Ty)=\sum_{i=0}^{k}(-1)^i\sum_{j=0}^{i}\tbinom{i}{j}(A_i^T)^{(j)}y^{(i-j)},$$

with function spaces

$$\begin{aligned}
&\mathbb{Z}^*=C^0(\mathbb{I},\mathbb{R}^n),\\
&\mathbb{Y}^*=\{y\in C^0(\mathbb{I},\mathbb{R}^n)\,|\,A_kA_k^+y\in C^k(\mathbb{I},\mathbb{R}^n),\ B_i^*(y,\overline{t})=0,\quad i=1,\ldots,k\}
\end{aligned}$$

and boundary terms

$$B_i^*(y,\overline{t})=\left\{(A_iA_i^+)^{(\ell)}y^{(j-\ell)}|_{\overline{t}}=0,\ \text{for}\ j=0,\ldots i-1,\ \ell=0\ldots,j\right\}.$$

In this setting, the conditions for self-conjugacy of the operator $\mathscr{L}$ are given by

$$A_\ell=\sum_{i=0}^{k}(-1)^i\tbinom{i}{i-\ell}(A_i^T)^{(i-\ell)}=\sum_{i=\ell}^{k}(-1)^i\tbinom{i}{\ell}(A_i^T)^{(i-\ell)} \tag{38}$$

for $\ell=0,\ldots,k$ (using that $\binom{i}{j}=0$ for $j<0$), defined on a domain

$$\mathbb{Z}=\{z\in C^0(\mathbb{I},\mathbb{R}^n)\,|A_k^+A_kz\in C^k(\mathbb{I},\mathbb{R}^n),\ B_i(z,\underline{t})=B_i^*(z,\overline{t})=0,\quad i=1,\ldots,k\}.$$

In the special case $k=1$, these conditions simplify to

$$A_0 = A_0^T - \dot{A}_1^T, \quad A_1 = -A_1^T$$

with boundary conditions

$$(A_1A_1^+)z|_{\overline{t}} = 0, \quad (A_1^+A_1)z|_{\underline{t}} = 0.$$

For constant coefficient systems the conditions (38) read $A_\ell = (-1)^\ell A_\ell^T$ for $\ell = 0,\dots,k$, i. e., the matrices are alternating symmetric/skew-symmetric. This corresponds to the case of even matrix tuples, see [31, 33].

Note that in contrast to Definition 12, here for simplicity the zero boundary conditions are incorporated into the domains $\mathbb{Z}$ and $\mathbb{Y}^*$.

For these formal self-conjugate operators we obtain the following result.

THEOREM 4. *Any self-conjugate linear k-th order differential operator $\mathscr{L}$ as in (37) with coefficient functions that satisfy the conditions* (38) *can be written in the form*

$$\mathscr{L}(z) = \sum_{\ell=0}^{k} (-1)^\ell \frac{d^\ell}{dt^\ell}\left(A_\ell^T z\right), \tag{39}$$

where the leading coefficient matrix satisfies $A_k = (-1)^k A_k^T$.

Proof. Using the condition for self-conjugacy given in (38), the differential operator can be written as

$$\begin{aligned}
\mathscr{L}z &= \sum_{\ell=0}^{k}\sum_{i=\ell}^{k}(-1)^i \tbinom{i}{\ell}(A_i^{(i-\ell)})^T z^{(\ell)} \\
&= \sum_{\ell=0}^{k}\sum_{j=0}^{\ell}(-1)^\ell \tbinom{\ell}{j}(A_\ell^{(\ell-j)})^T z^{(j)} \\
&= \sum_{\ell=0}^{k}(-1)^\ell \frac{d^\ell}{dt^\ell}\left(A_\ell^T z\right). \quad \Box
\end{aligned}$$

For further investigations it turns out to be useful to split a self-conjugate differential operator into even and odd order parts.

THEOREM 5. *Any self-adjoint linear k-th order differential operator $\mathscr{L}$ as in (37) with coefficient functions that satisfy the conditions* (38) *can be written as a sum of differential operators of the form*

$$\mathscr{L}_{2\nu}(x) = (P_{2\nu}x^{(\nu)})^{(\nu)}, \tag{40a}$$

$$\mathscr{L}_{2\nu-1}(x) = \frac{1}{2}[(Q_{2\nu-1}x^{(\nu-1)})^{(\nu)} + (Q_{2\nu-1}x^{(\nu)})^{(\nu-1)}], \tag{40b}$$

with matrix valued functions $P_{2\nu} = P_{2\nu}^T \in C^\nu(\mathbb{I},\mathbb{R}^{n,n})$, $Q_{2\nu-1} = -Q_{2\nu-1}^T \in C^\nu(\mathbb{I},\mathbb{R}^{n,n})$ for $\nu = 0,\dots,\mu$, where $\mu = \frac{k}{2}$ if k is even and $\mu = \frac{k+1}{2}$ if k is odd.

Proof. We prove the statement by induction. For $k=1$ we have

$$\mathscr{L}(x) = A_1\dot{x} + A_0 x = \frac{1}{2}\frac{d}{dt}(A_1 x) + \frac{1}{2}(A_1\dot{x}) + (A_0 - \frac{1}{2}\dot{A}_1)x$$

with $Q_1 := A_1$ skew-symmetric and $P_0 := A_0 - \frac{1}{2}\dot{A}_1$ symmetric (due to (38)). Similarly, for $k=2$ we have

$$\begin{aligned}\mathscr{L}(x) &= A_2\ddot{x} + A_1\dot{x} + A_0 x\\ &= \frac{d}{dt}(A_2\dot{x}) + (A_1 - \dot{A}_2)\dot{x} + A_0 x\\ &= \frac{d}{dt}(A_2\dot{x}) + \frac{1}{2}\frac{d}{dt}((A_1 - \dot{A}_2)x) + \frac{1}{2}((A_1 - \dot{A}_2)\dot{x}) + (A_0 - \frac{1}{2}(\dot{A}_1 - \ddot{A}_2))x,\end{aligned}$$

with $P_2 := A_2$ and $P_0 := A_0 - \frac{1}{2}(\dot{A}_1 - \ddot{A}_2)$ symmetric, and $Q_1 := A_1 - \dot{A}_2$ skew-symmetric (due to (38)).

Now let $\mathscr{L}(x) = \sum_{i=0}^{k} A_i x^{(i)}$ be a self-conjugate differential operator and assume that $k = 2\mu$ is even. The conditions in (38) imply that $A_k = A_k^T$, and we can write the operator as

$$\begin{aligned}\mathscr{L}(x) &= \frac{d^\mu}{dt^\mu}\left(A_k x^{(\mu)}\right) - \sum_{i=1}^{\mu}\binom{\mu}{i} A_k^{(i)} x^{(k-i)} + A_{k-1}x^{(k-1)} + \ldots + A_1\dot{x} + A_0 x\\ &= \frac{d^\mu}{dt^\mu}\left(A_k x^{(\mu)}\right) + \sum_{i=0}^{k-1}\tilde{A}_i x^{(i)},\end{aligned}$$

with $\tilde{A}_i = A_i - \binom{\mu}{k-i}A_k^{(k-i)}$, for $i = k-\mu,\ldots,k-1$, and $\tilde{A}_i = A_i$ for $i = 0,\ldots,k-\mu-1$. If we subtract from $\mathscr{L}(x)$ the self-conjugate expression

$$\mathscr{L}_{2\mu}(x) = \frac{d^\mu}{dt^\mu}(A_k x^{(\mu)}) = A_k x^k + \sum_{j=1}^{\mu}\binom{\mu}{j}A_k^{(j)}x^{(k-j)},$$

(i. e., $P_k := A_k$), then we obtain again a self-conjugate expression

$$\begin{aligned}\bar{\mathscr{L}}(x) = \mathscr{L}(x) - \mathscr{L}_{2\mu}(x) &= \sum_{i=0}^{k-1} A_i x^{(i)} - \sum_{j=1}^{\mu}\binom{\mu}{j}A_k^{(j)}x^{(k-j)}\\ &= \sum_{i=0}^{\mu-1} A_i x^{(i)} + \sum_{i=\mu}^{k-1}(A_i - \binom{\mu}{k-i}A_k^{(k-i)})x^{(i)}\end{aligned}$$

of odd order $k-1$.

If $k = 2\mu - 1$ is odd, then we have $A_k = -A_k^T$. By subtracting from $\mathscr{L}(x)$ the self-conjugate expression

$$\begin{aligned}\mathscr{L}_{2\mu-1}(x) &= \frac{1}{2}\left[(A_k x^{(\mu-1)})^{(\mu)} + (A_k x^{(\mu)})^{(\mu-1)}\right]\\ &= A_k x^{(k)} + \frac{1}{2}\left[\sum_{j=1}^{\mu}\binom{\mu}{j}A_k^{(j)}x^{(k-j)} + \sum_{j=1}^{\mu-1}\binom{\mu-1}{j}A_k^{(j)}x^{(k-j)}\right],\end{aligned}$$

(i. e., $Q_k := A_k$), then we obtain a self-conjugate expression

$$\bar{\mathscr{L}}(x) = \mathscr{L}(x) - \mathscr{L}_{2\mu-1}(x) = \sum_{i=0}^{k-1} A_i x^{(i)} - \frac{1}{2}\left[\sum_{j=1}^{\mu} \binom{\mu}{j} A_k^{(j)} x^{(k-j)} + \sum_{j=1}^{\mu-1} \binom{\mu-1}{j} A_k^{(j)} x^{(k-j)}\right]$$

of even order $k-1$.

Due to the inductive assumption, a self-adjoint operator of order $k-1$ can be written as a sum of expressions of the form (40a) and (40b). This completes the proof. □

In the following, we say that a kth-order self-adjoint differential operator is in *partitioned form*, if it is given by

$$\mathscr{L}(x) = \begin{cases} \sum_{v=0}^{r} \mathscr{L}_{2v}(x) + \sum_{v=1}^{r} \mathscr{L}_{2v-1}(x), & \text{if } k \text{ is even with } r = \frac{k}{2}, \\ \sum_{v=0}^{r-1} \mathscr{L}_{2v}(x) + \sum_{v=1}^{r} \mathscr{L}_{2v-1}(x), & \text{if } k \text{ is odd with } r = \frac{k+1}{2}. \end{cases} \tag{41}$$

EXAMPLE 1. For a linear second order differential operator of the form

$$M\ddot{x} + C\dot{x} + Kx = f, \tag{42}$$

with coefficient functions $M, C, K \in C(\mathbb{I}, \mathbb{R}^{n,n})$ that are sufficiently smooth and satisfy the conditions

$$M = M^T, \quad C = (2\dot{M} - C)^T, \quad \text{and } K = (\ddot{M} - \dot{C} + K)^T, \text{ for all } t \in \mathbb{I},$$

the formulation (39) is given by

$$\frac{d^2}{dt^2}\left(M^T x\right) - \frac{d}{dt}\left(C^T x\right) + K^T x = f,$$

and the partitioned form (41) by

$$P_0 x + \frac{d}{dt}(P_2 \dot{x}) + \frac{1}{2}\frac{d}{dt}(Q_1 x) + \frac{1}{2} Q_1 \dot{x} = f. \tag{43}$$

with $P_0 := K - \frac{1}{2}(\dot{C} - \ddot{M})$, $P_2 := M$, and, $Q_1 := C - \dot{M}$.

In order to derive structure preserving first order formulations, we assume that the leading matrix A_k is pointwise nonsingular. Otherwise, the task is more complicated, and we have to consider trimmed first order formulations, see [39].

In the second order case (using the notation of Example 1), introducing in (43) the new variable $v = \dot{x}$, we get

$$\frac{d}{dt}(P_2 \dot{x}) = \frac{d}{dt}(P_2 v) = P_2 \dot{v} + \dot{P}_2 v,$$

yielding

$$P_2 \dot{v} + \dot{P}_2 v + P_0 x + \frac{1}{2}\dot{Q}_1 x + Q_1 \dot{x} = f.$$

This gives the first order system

$$\begin{bmatrix} 0 & -P_2 \\ P_2 & Q_1 \end{bmatrix} \begin{bmatrix} \dot v \\ \dot x \end{bmatrix} + \begin{bmatrix} P_2 & 0 \\ \dot P_2 & P_0 + \frac{1}{2}\dot Q_1 \end{bmatrix} \begin{bmatrix} v \\ x \end{bmatrix} = \begin{bmatrix} 0 \\ f \end{bmatrix},$$

or equivalently

$$\underbrace{\begin{bmatrix} 0 & -M \\ M & C - \dot M \end{bmatrix}}_{\mathscr{E}} \begin{bmatrix} \dot v \\ \dot x \end{bmatrix} + \underbrace{\begin{bmatrix} M & 0 \\ \dot M & K \end{bmatrix}}_{\mathscr{A}} \begin{bmatrix} v \\ x \end{bmatrix} = \begin{bmatrix} 0 \\ f \end{bmatrix},$$

with a self-adjoint pair of coefficient functions $(\mathscr{E}, \mathscr{A})$.

Similar, for a third order system in partitioned form

$$P_0 x + \frac{d}{dt}(P_2 \dot x) + \frac{1}{2}\left[\frac{d}{dt}(Q_1 x) + Q_1 \dot x\right] + \frac{1}{2}\left[\frac{d^2}{dt^2}(Q_3 \dot x) + \frac{d}{dt}(Q_3 \ddot x)\right] = f,$$

with nonsingular leading matrix $Q_3 = A_3$, by introducing $v_1 = \dot x$, and $v_2 = \dot v_1 = \ddot x$ we get

$$\frac{d}{dt}(Q_3 \ddot x) = \frac{d}{dt}(Q_3 v_2) = \dot Q_3 v_2 + Q_3 \dot v_2,$$

as well as

$$\begin{aligned} \frac{d^2}{dt^2}(Q_3 \dot x) &= \frac{d^2}{dt^2}(Q_3 v_1) = \frac{d}{dt}(\dot Q_3 v_1 + Q_3 v_2) = \ddot Q_3 v_1 + \dot Q_3 \dot v_1 + \dot Q_3 v_2 + Q_3 \dot v_2, \\ \frac{d}{dt}(P_2 \dot x) &= \frac{d}{dt}(P_2 v_1) = P_2 \dot v_1 + \dot P_2 v_1. \end{aligned}$$

Altogether this yields the first order formulation

$$\begin{bmatrix} 0 & 0 & Q_3 \\ 0 & -Q_3 & -P_2 + \frac{1}{2}\dot Q_3 \\ Q_3 & P_2 + \frac{1}{2}\dot Q_3 & Q_1 \end{bmatrix} \begin{bmatrix} \dot v_2 \\ \dot v_1 \\ \dot x \end{bmatrix} + \begin{bmatrix} 0 & -Q_3 & 0 \\ Q_3 & P_2 - \frac{1}{2}\dot Q_3 & 0 \\ \dot Q_3 & \dot P_2 + \frac{1}{2}\ddot Q_3 & P_0 + \frac{1}{2}\dot Q_1 \end{bmatrix} \begin{bmatrix} v_2 \\ v_1 \\ x \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ f \end{bmatrix},$$

or equivalently, using $Q_3 = A_3$, $P_2 = A_2 - \frac{3}{2}\dot A_3$, $Q_1 = A_1 - \dot A_2 + \ddot A_3$, $P_0 = A_0 - \frac{1}{2}\dot A_1 + \frac{1}{2}\ddot A_2 - \frac{1}{2}\dddot A_3$,

$$\underbrace{\begin{bmatrix} 0 & 0 & A_3 \\ 0 & -A_3 & -A_2 + 2\dot A_3 \\ A_3 & A_2 - \dot A_3 & A_1 - \dot A_2 + \ddot A_3 \end{bmatrix}}_{\mathscr{E}} \begin{bmatrix} \dot v_2 \\ \dot v_1 \\ \dot x \end{bmatrix} + \underbrace{\begin{bmatrix} 0 & -A_3 & 0 \\ A_3 & A_2 - 2\dot A_3 & 0 \\ \dot A_3 & \dot A_2 - \ddot A_3 & A_0 \end{bmatrix}}_{\mathscr{A}} \begin{bmatrix} v_2 \\ v_1 \\ x \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ f \end{bmatrix},$$

and again the matrix pair $(\mathscr{E}, \mathscr{A})$ is self-adjoint, since $A_0, \ldots, A_3$ satisfy the condition (38). It is obvious, but rather technical, how to extend this construction to higher orders $k > 3$.

In the discrete-time case the situation is somehow different. For odd order difference operators there does not exist a self-conjugate operator corresponding to the definition given in Remark 5, since a two-term recursion can never be written in the form (34) with symmetric system matrix.

Nevertheless, we can derive an equivalent two-term recursion with similar structures as in the constant coefficient case, see [11].

EXAMPLE 2. For a second order self-adjoint difference operator with constant coefficients

$$\mathscr{M} x_{i-1} + \mathscr{N} x_i + \mathscr{M}^T x_{i+1} = f_i,$$

by setting $v_i := x_{i+1}$ we have the palindromic first order form

$$\begin{bmatrix} \mathscr{M}^T & \mathscr{N} - \mathscr{M} \\ \mathscr{M}^T & \mathscr{M}^T \end{bmatrix} \begin{bmatrix} v_i \\ x_i \end{bmatrix} + \begin{bmatrix} \mathscr{M} & \mathscr{M} \\ \mathscr{N} - \mathscr{M}^T & \mathscr{M} \end{bmatrix} \begin{bmatrix} v_{i-1} \\ x_{i-1} \end{bmatrix} = \begin{bmatrix} f_i \\ f_i \end{bmatrix},$$

see [31].

Proceeding like this in the case of variable coefficients, for a self-conjugate system (33) we obtain

$$\begin{bmatrix} \mathscr{M}_{i+1}^T & \mathscr{N}_i - \mathscr{M}_i \\ \mathscr{M}_{i+1}^T & \mathscr{M}_{i+1}^T \end{bmatrix} \begin{bmatrix} v_i \\ x_i \end{bmatrix} + \begin{bmatrix} \mathscr{M}_i & \mathscr{M}_i \\ \mathscr{N}_i - \mathscr{M}_{i+1}^T & \mathscr{M}_i \end{bmatrix} \begin{bmatrix} v_{i-1} \\ x_{i-1} \end{bmatrix} = \begin{bmatrix} f_i \\ f_i \end{bmatrix}.$$

For the special case of difference equations from optimal control problems in (18) (omitting the last row and column) by shifting the first block row we get

$$\begin{bmatrix} 0 & E_k & 0 \\ -A_k^T & W_k & S_k \\ -B_k^T & S_k^T & R_k \end{bmatrix} \begin{bmatrix} \lambda_k \\ x_k \\ u_k \end{bmatrix} + \begin{bmatrix} 0 & -A_{k-1} & -B_{k-1} \\ E_k^T & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \lambda_{k-1} \\ x_{k-1} \\ u_{k-1} \end{bmatrix} = \begin{bmatrix} f_{k-1} \\ 0 \\ 0 \end{bmatrix}, \quad k = 0, \dots, N-1,$$

similar to the BVD-pencil structure introduced in [11].

6. Conclusion

We have shown that the necessary optimality conditions for discrete-time linear quadratic control problems with variable coefficients leads to self-conjugate difference operators associated with self-adjoint triples of coefficient functions, thus achieving a similar result as in the continuous time case. We have also extended these results to higher order differential or difference equation constraints and shown how first order reductions can be carried out that lead to first order systems with the same structural properties.

REFERENCES

[1] C. D. AHLBRANDT AND A. C. PETERSON, *Discrete Hamiltonian Systems*, Kluwer Academic Publisher, 1996.

[2] A. BACKES, *Extremalbedingungen für Optimierungs-Probleme mit Algebro-Differentialgleichungen*, PhD thesis, Institut für Mathematik, Humboldt-Universität zu Berlin, Berlin, Germany, 2006.

[3] J. T. BETTS, *Path constrained trajectory optimization using sparse sequential quadratic programming*, J. Guidance Control Dyn., 16: 59–68, 1993.

[4] J. T. BETTS, M. J. CARTER, AND W. P. HUFFMAN, *Software for nonlinear optimization*, Mathematics and Engineering Analysis Library Report MEA-LR-83 R1, Boeing Information and Support Services, Seattle, USA, 1997.

[5] L. T. BIEGLER, *Optimization strategies for complex process models*, Adv. in Chem. Eng., 18: 197–256, 1992.

[6] K. BOBINYEC, S. L. CAMPBELL, AND P. KUNKEL, *Maximally reduced observers for linear time varying DAEs*, In Proc. IEEE Multi-Conference on Systems and Control, Denver, 2011, pages 1373–1378, 2011.

[7] H. G. BOCK, M. M. DIEHL, D. B. LEINEWEBER, AND J. P. SCHLÖDER, *A direct multiple shooting method for real-time optimization of nonlinear DAE processes*, In Nonlinear model predictive control, Ascona, 1998, volume 26 of Progr. Systems Control Theory, pages 245–267, Basel, 2000, Birkhäuser.

[8] 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.

[9] T. BRÜLL, *Existence and uniqueness of solutions of linear variable coefficient discrete-time descriptor systems*, Linear Algebra Appl., 431: 247–265, 2009.

[10] T. BRÜLL, *Explicit solutions of regular linear discrete-time descriptor systems with constant coefficients*, Elec. J. Linear Algebra, 18: 317–338, 2009.

[11] R. BYERS, D. S. MACKEY, V. MEHRMANN, AND H. XU, *Symplectic, BVD, and palindromic approaches to discrete-time control problems*, In P. Petkov and N. Christov, editors, Collection of Papers Dedicated to the 60-th Anniversary of Mihail Konstantinov, pages 81–102, Rodina, Publications, Sofia, Bulgaria, 2009.

[12] R. BYERS, V. MEHRMANN, AND H. XU, *A structured staircase algorithm for skew-symmetric/symmetric pencils*, Electr. Trans. Num. Anal., 26: 1–33, 2007.

[13] R. BYERS, V. MEHRMANN, AND H. XU, *Trimmed linearization for structured matrix polynomials*, Linear Algebra Appl., 429: 2373–2400, 2008.

[14] J. W. DEMMEL AND B. KÅGSTRÖM, *The generalized Schur decomposition of an arbitrary pencil* $\lambda A - B$, Part I, ACM Trans. Math. Software, 19: 160–174, 1993.

[15] J. W. DEMMEL AND B. KÅGSTRÖM, *The generalized Schur decomposition of an arbitrary pencil* $\lambda A - B$, Part II, ACM Trans. Math. Software, 19: 185–201, 1993.

[16] F. R. GANTMACHER, *The Theory of Matrices II*, Chelsea Publishing Company, New York, NY, 1959.

[17] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 3rd edition, 1996.

[18] W. W. HAGER, *Runge-Kutta methods in optimal control and the transformed adjoint system*, Numer. Math., 87: 247–282, 2000.

[19] E. HAIRER AND G. WANNER, *Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems*, Springer-Verlag, Berlin, Germany, 2nd edition, 1996.

[20] R. A. HORN AND V. V. SERGEICHUK, *Canonical forms for complex matrix congruence and* $*$ *congruence*, Linear Algebra Appl., 416: 1010–1032, 2006.

[21] W. G. KELLEY AND A. C. PETERSON, *Difference equations: An introduction with applications*, Academic Press, San Diego, CA, USA, 2001.

[22] V. KUCERA, *The structure and properties of time-optimal discrete linear control*, IEEE Trans. Automatic Control, 16: 375–377, 1971.

[23] P. KUNKEL AND V. MEHRMANN, *Differential-Algebraic Equations. Analysis and Numerical Solution*, EMS Publishing House, Zürich, Switzerland, 2006.

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

[25] P. KUNKEL AND V. MEHRMANN, *Formal adjoints of linear DAE operators and their role in optimal control*, Electr. J. Lin. Alg., 22: 672–693, 2011.

[26] 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.

[27] P. KUNKEL, V. MEHRMANN, AND L. SCHOLZ, *Self-adjoint differential-algebraic equations* Appeared electronically in Mathematics of Control, Signals and Systems, 2013. DOI: 10.1007/s00498-013-0109-3.

P. KUNKEL, V. MEHRMANN AND L. SCHOLZ, *Self-adjoint differential-algebraic equations*. Mathematics of Control, Signals, and Systems, 26 (1): 47–76, 2014.

[28] G. A. KURINA, *Linear-quadratic discrete optimal control problems for descriptor systems in Hilbert space*, J. Dynam. Control Systems, 10: 365–375, 2004.

[29] G. A. KURINA AND R. MÄRZ, *On linear-quadratic optimal control problems for time-varying descriptor systems*, SIAM J. Cont. Optim., 42: 2062–2077, 2004.

[30] H. KWAKERNAAK AND R. SIVAN, *Linear Optimal Control Systems*. Wiley-Interscience, New York, 1972.

[31] D. S. MACKEY, N. MACKEY, C. MEHL, AND V. MEHRMANN, *Structured polynomial eigenvalue problems: Good vibrations from good linearizations*, SIAM J. Matr. Anal. Appl., 28: 1029–1051, 2006.

[32] D. S. MACKEY, N. MACKEY, C. MEHL, AND V. MEHRMANN, *Smith forms of palindromic matrix polynomials*, Electr. J. Lin. Alg., 22: 53–91, 2011.

[33] D. S. MACKEY, N. MACKEY, C. MEHL, AND V. MEHRMANN, *Jordan structures of alternating matrix polynomials*, Linear Algebra Appl., 432: 867–891, 2010.

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

[35] V. MEHRMANN AND D. WATKINS, *Polynomial eigenvalue problems with Hamiltonian structure*, Electr. Trans. Num. Anal., 13: 106–113, 2002.

[36] J. W. POLDERMAN AND J. C. WILLEMS, *Introduction to Mathematical Systems Theory: A Behavioural Approach*, Springer-Verlag, New York, NY, 1998.

[37] R. C. THOMPSON, *Pencils of complex and real symmetric and skew matrices*, Linear Algebra Appl., 147: 323–371, 1991.

[38] P. VAN DOOREN, *The computation of Kronecker's canonical form of a singular pencil*, Linear Algebra Appl., 27: 103–141, 1979.

[39] L. WUNDERLICH, *Analysis and Numerical Solution of Structured and Switched Differential-Algebraic Systems*, Phd thesis, Institut für Mathematik, Technische Universität Berlin, September 2008.

(Received August 2, 2012)

Volker Mehrmann
Institut für Mathematik
MA 4-5, TU Berlin, Straße des 17. Juni 136
10623 Berlin, FRG
e-mail: mehrmann@math.tu-berlin.de

Lena Scholz
Institut für Mathematik
MA 4-5, TU Berlin, Straße des 17. Juni 136
10623 Berlin, FRG
e-mail: lscholz@math.tu-berlin.de