SAMPLED-DATA AND DISCRETE-TIME $H_2$ OPTIMAL CONTROL*

H. L. TRENTELMAN† AND A. A. STOORVOGEL‡

Abstract. This paper deals with the sampled-data $H_2$ optimal control problem. Given a linear time-invariant continuous-time system, the problem of minimizing the $H_2$ performance over all sampled-data controllers with a fixed sampling period can be reduced to a pure discrete-time $H_2$ optimal control problem. This discrete-time $H_2$ problem is always singular. Motivated by this, in this paper we give a treatment of the discrete-time $H_2$ optimal control problem in its full generality. The results we obtain are then applied to the singular discrete-time $H_2$ problem arising from the sampled-data $H_2$ problem. In particular, we give conditions for the existence of optimal sampled data controllers. We also show that the $H_2$ performance of a continuous-time controller can always be recovered asymptotically by choosing the sampling period sufficiently small. Finally, we show that the optimal sampled-data $H_2$ performance converges to the continuous-time optimal $H_2$ performance as the sampling period converges to zero.

Key words. sampled-data, lifting technique, discrete-time, $H_2$ optimal control, algebraic Riccati equation, small sampling periods

AMS subject classifications. 93C05, 93C35, 93C60

1. Introduction. Recently, much attention has been paid to $H_2$ and $H_\infty$ optimal control of linear systems using sampled-data control (see [6], [7], [12], [2], [4] and [5], [11], [10], [1], [3], [17], [21]). For a given a continuous-time plant, a sampled-data controller consists of the cascade connection of an A/D converter, a discrete-time controller, and a D/A converter. The A/D device converts the continuous-time measured plant output into a discrete-time signal, which is used as an input for the discrete-time controller. The discrete-time controller generates a discrete-time output signal, which, in turn, is converted into a continuous-time signal that is used as a control input for the continuous-time plant.

Apart from a control input and a measurement output, the plant under consideration has an exogenous input and an output to be controlled. The quality of a controller is given by the performance of the corresponding closed-loop system. This performance measures the influence of the exogenous input on the output to be controlled. In the present paper, we will take the $H_2$ performance of the closed-loop system as performance measure.

In contrast to the $H_\infty$ performance of a sampled-data control system, which in analogy with the pure continuous-time context can simply be defined as the norm of the input/output operator between the exogenous inputs and the outputs to be controlled, it is not clear from the outset how one should define the $H_2$ performance of a sampled-data control system. One definition was proposed in [6]: the $H_2$ performance of the closed-loop system is the number obtained by applying at each input channel a Dirac distribution and by taking the sum of integral squares of the resulting outputs. Of course, this definition exactly mimics the one that is common in the pure continuous-time context.

Received by the editors December 28, 1992; accepted for publication (in revised form) November 24, 1993.

† Mathematics Institute, University of Groningen, P.O. Box 800, 9700 AV Groningen, The Netherlands (h.l.trentelman@math.rug.nl).

‡ Faculty of Mathematics and Computing Science, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands (scoas@win.tue.nl). This research was supported by a fellowship of the Royal Netherlands Academy of Sciences and Arts.
In our opinion, a more natural definition was given independently in [12] and [2]. In these references, the crucial observation is that the closed-loop system resulting from a sampled data controller, albeit time-varying, is in fact a periodic system, with period equal to the sampling period. It is then argued that, instead of applying impulsive inputs at time $t = 0$, one should in fact apply these inputs at all time instances between 0 and the sampling period and take the mean of the integral squares of the resulting outputs. This leads to an $H_2$ performance measure that captures the essential features of a sampled-data closed-loop system more satisfactorily. For a given continuous-time plant, the sampled-data $H_2$ optimal control problem is then to minimize the $H_2$ performance of the closed-loop system over all internally stabilizing sampled-data controllers with a fixed sampling period. It is the latter problem that will be studied in this paper.

It was shown in [12] and [2] (see also [4]) that the sampled-data $H_2$ optimal control problem can be reduced to a pure discrete-time $H_2$ optimal control problem in the following way. First one defines an auxiliary time-invariant discrete-time system (involving the parameters of the original continuous-time plant and the given sampling period). Next, one expresses the sampled-data $H_2$ performance in terms of the ‘normal’ $H_2$ performance of the closed-loop system obtained by interconnecting the auxiliary discrete-time system and the discrete-time controller defining the sampled-data controller. Thus, the sampled-data $H_2$ optimal control problem under consideration is completely resolved once the auxiliary discrete-time $H_2$ problem is. This procedure makes use of the so-called lifting technique (see [20], [1], [3]).

Now it turns out that the auxiliary discrete-time $H_2$ problem obtained in this way is always a singular problem: the direct feedthrough matrix from the exogenous input to the measurement output is always equal to 0. Apart from this, in the auxiliary discrete-time system the direct feedthrough matrix from the control input to the output to be controlled is in general not injective. (Note that, in general, an $H_2$ optimal control problem is called regular if the direct feedthrough matrix from the control input to the output to be controlled is injective, and the direct feedthrough matrix from the exogenous input to the measurement output is surjective. If the problem is not regular it is called singular.) In [12], this difficulty is partly removed by introducing an additional noise on the sampled measured output signal and by assuming the corresponding feedthrough matrix to be surjective.

In the present paper we want to consider the completely general formulation of the sampled-data $H_2$ problem. As a starting point we will take the auxiliary discrete-time $H_2$ problem derived in [12] and [2]. As noted, this problem is inherently singular. To our best knowledge, no resolution of the discrete-time singular $H_2$ optimal is known in the literature. Therefore, a substantial part of this paper is devoted to a study of the completely general discrete-time $H_2$ problem (no assumptions on the direct feedthrough matrices, no assumptions on the absence of zeros on the unit circle). We will describe a complete resolution to this problem, including a characterization of the optimal performance, and necessary and sufficient conditions for the existence of optimal controllers. The expression for the optimal performance is different from the one that might be expected in analogy with the continuous-time case (see [15]). Due to the fact that the role of the imaginary axis is taken over by the unit circle, for the discrete-time $H_2$ performance to be finite it is no longer required that the closed-loop transfer matrix is strictly proper. Intuitively, this enlarges the class of admissible controllers and yields a smaller optimal performance.

We will apply our results on the discrete-time $H_2$ optimal control problem to
the sampled-data $H_2$ problem by simply applying them to the auxiliary discrete-time system derived in [12] and [2]. Our expression for the optimal sampled-data $H_2$ performance will be an immediate consequence of these results. We will, however, also be interested in conditions guaranteeing the existence of optimal sampled-data controllers. Our results on the general discrete-time $H_2$ problem give such conditions in terms of the auxiliary discrete-time system, but we will reformulate these conditions in terms of the original continuous-time plant. Preliminary results in that direction were also found in [12].

Obviously, the sampled-data $H_2$ optimal performance is a function of the sampling period. An important question is: what happens if the sampling period tends to zero. In particular, we will answer the following two questions. First, if we control the original continuous-time plant by a “normal” continuous-time compensator, is it then possible to recover this performance asymptotically by using a sampled-data controller with sufficiently small sampling period? This question was also studied for the $H_{\infty}$ performance and for the $H_2$ performance à la Chen and Francis in [6]. A second, related, question that we will answer is: does the optimal sampled-data $H_2$ performance converge to the optimal continuous-time $H_2$ performance as the sampling period decreases to zero?

The outline of this paper is as follows. In §2 we will define the sampled-data $H_2$ optimal control problem and recall the main results of [12] and [2]. We will also introduce some notation and recall the notions of left-invertibility and right-invertibility of linear systems, zeros, and their most important state space interpretations. In §3 we deal with the discrete-time $H_2$ optimal control problem. In this section we will not yet treat the completely general case but make some assumptions on the absence of zeros on the unit circle. In §4, the results of §3 will be extended to derive a resolution of the general discrete-time $H_2$ optimal control problem. Then, in §5, we return to the sampled-data context and apply the results of §§3 and 4 to the sampled-data $H_2$ optimal control problem. In particular, we will derive conditions in terms of the original continuous-time plant that guarantee the existence of optimal controllers for the sampled-data $H_2$ problem. Finally, in §6 we study the aforementioned questions regarding the behavior of the (optimal) performance as the sampling period tends to zero.

2. Problem formulation. Consider a continuous-time, linear, time-invariant, finite-dimensional plant $\Sigma$. Let $\Sigma$ have inputs $d$ and $u$ and outputs $z$ and $y$, where $d$ is an exogenous input, $u$ is a control input, $z$ is an output to be controlled, and $y$ is a measured output. We want to control $\Sigma$ by means of sampled-data feedback control. We take a fixed $\Delta > 0$, called the sampling period. From the measured output $y$ we obtain a discrete-time signal $\tilde{y} = \{y_k\}$ defined by $y_k := (S_\Delta y)_k$, where $S_\Delta$ denotes the sampling operator defined by $(S_\Delta y)_k := y(k\Delta)$. This discrete-time signal is taken as input for a discrete-time, linear, time-invariant, finite-dimensional compensator $\Gamma_{\text{dis}}$. The latter compensator generates a discrete-time signal $\tilde{u} = \{u_k\}$, which, in turn, yields a (piecewise constant) continuous-time input signal $u$ for the plant by defining $u(t) := (H_\Delta \tilde{u})(t)$, where $H_\Delta$ is the hold operator defined by $(H_\Delta \tilde{u})(t) := u_k$ ($t \in [k\Delta, (k + 1)\Delta)$). This type of feedback control is depicted in Fig. 1.

If we control the system $\Sigma$ by means of a sampled-data controller with sampling period $\Delta$, then the resulting closed-loop system will no longer be time-invariant. In [12] and [2] the following definition of $H_2$ performance in the context of sampled-data control is proposed. First, it is observed that the closed-loop system resulting from a sampled-data controller with sampling period $\Delta$ is always a time-varying, $\Delta$-periodic
system. Then, for $\Delta$-periodic systems the notion of $H_2$ performance is defined as follows. Suppose we have a finite-dimensional, time-varying, $\Delta$-periodic system $\Sigma_{\text{per}}$ described by

\[(2.1)\quad z(t) = \int_0^t G(t, s)d(s)ds.\]

It is argued in [12] and [2] that a natural way to define the $H_2$ performance of (2.1) is

\[(2.2)\quad \|\Sigma_{\text{per}}\|_2^2 := \frac{1}{\Delta} \int_0^\Delta \text{tr} \int_s^\infty G^T(t, s)G(t, s)dt\,ds.\]

Next, if $\Gamma$ is a sampled-data controller with sampling period $\Delta$, the associated performance is defined as $J_{\Sigma, \Delta}(\Gamma) := \|\Sigma \times \Gamma\|_2^2$, the $H_2$ performance of the $(\Delta$-periodic) closed-loop system $\Sigma \times \Gamma$. The sampled-data $H_2$ problem is then to minimize, for a fixed sampling period $\Delta$, the performance criterion $J_{\Sigma, \Delta}(\Gamma)$ over all internally stabilizing sampled-data controllers $\Gamma$ with sampling period $\Delta$. It was shown in [12] and [2] that this problem can be reduced to a discrete-time ‘normal’ $H_2$ optimal control problem. To be specific, let the plant $\Sigma$ be given by the equations

\[(2.3)\quad \begin{align*}
\dot{x}(t) &= Ax(t) + Bu(t) + Ed(t), \\
y(t) &= C_1x(t), \\
z(t) &= C_2x(t) + D_2u(t),
\end{align*}\]

with $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}^m$, $d(t) \in \mathbb{R}^r$, $y(t) \in \mathbb{R}^p$, and $z(t) \in \mathbb{R}^q$. It will be a standing assumption in this paper that $(A, B)$ is stabilizable and that $(C_1, A)$ is detectable, both with respect to $\mathcal{C}^- := \{s \in \mathcal{C} \mid \Re s < 0\}$. Introduce a finite-dimensional linear time-invariant discrete-time system $\Sigma_{\Delta}$:

\[(2.4)\quad \begin{align*}
x_{k+1} &= A_\Delta x_k + B_\Delta u_k + E_\Delta d_k, \\
y_k &= C_1 x_k, \\
z_k &= C_{2, \Delta} x_k + D_{2, \Delta} u_k,
\end{align*}\]

where we define

\[A_\Delta := e^{A\Delta}, \quad B_\Delta := \int_0^\Delta e^{tA}dtB,\]
where $E_{\Delta}$ is any matrix satisfying

$$E_{\Delta}E_{\Delta}^\top = \int_0^\Delta e^{t\Delta} E E^\top e^{t\Delta^\top} \, dt,$$

and where $C_{2,\Delta}$ and $D_{2,\Delta}$ are matrices satisfying

$$(C_{2,\Delta} D_{2,\Delta})^\top (C_{2,\Delta} D_{2,\Delta}) = \int_0^\Delta e^{t\Delta^\top} (C_{2} D_{2})^\top (C_{2} D_{2}) e^{t\Delta} \, dt.$$ 

Here we have denoted

$$\Delta := \begin{pmatrix} A & B \\ 0 & 0 \end{pmatrix}.$$ 

Let $\Delta$ denote the set of sampling periods for which either $(A_\Delta, B_\Delta)$ is not stabilizable or $(C_1, A_\Delta)$ is not detectable, both with respect to the open unit disc $\{z \in \mathbb{C} | |z| < 1\}$. It is well known [13], [8] that if $(A, B)$ is stabilizable and $(C_1, A)$ is detectable, then every bounded subset of $\mathbb{R}^+$ contains only finitely many elements of $\Delta$. We will restrict ourselves to sampling periods that are not in $\Delta$. The plant $E$ is controlled using sampled-data controllers $\Gamma := H_\Delta \Gamma_{\text{dis}} S_{\Delta}$, with $\Gamma_{\text{dis}}$ given by the equations

$$w_{k+1} = Kw_k + Ly_k, \quad u_k = Mw_k + Ny_k.$$ 

Let us denote by $J_{\Sigma_{\Delta}}(\Gamma_{\text{dis}})$ the discrete-time $H_2$ performance of the closed-loop system $\Sigma_{\Delta} \times \Gamma_{\text{dis}}$, i.e., the value $\sum_k \text{tr} (G_k G_k^T)$, where $\{G_k\}$ denotes the pulse response of the closed-loop system. The main result of [12] and [2] is the following:

**Theorem 2.1.** Assume that $\Delta \notin \Delta$. Then there exists a sampled-data controller $\Gamma$ with sampling period $\Delta$ such that the closed-loop system $\Sigma \times \Gamma$ is internally stable. The sampled-data controller $\Gamma = H_\Delta \Gamma_{\text{dis}} S_{\Delta}$ internally stabilizes $\Sigma$ if and only if the discrete-time controller $\Gamma_{\text{dis}}$ internally stabilizes $\Sigma_{\Delta}$. Furthermore, for every such controller we have

$$J_{\Sigma_{\Delta}}(\Gamma) = \frac{1}{\Delta} \int_0^\Delta \int_0^{\Delta-s} \text{tr} \left( C_2 e^{t\Delta} E E^\top e^{t\Delta^\top} C_2^\top \right) \, dt \, ds + \frac{1}{\Delta} J_{\Sigma_{\Delta}}(\Gamma_{\text{dis}}).$$

We shall use this theorem as a starting point and study in this paper the discrete-time $H_2$ optimal control problem for the discrete-time system $E_{\Delta}$ given by (2.4). This $H_2$ problem is inherently singular, due to the fact that the direct feedthrough matrix from the disturbance input to the measured output is always equal to zero.

We conclude this section by introducing some notation and recalling some basic concepts. In this paper, any given continuous-time system $\dot{x} = Ax + Bu, y = Cx + Du$ or discrete-time system $x_{k+1} = Ax_k + Bu_k, y_k = Cx_k + Du_k$ will be denoted simply by $(A, B, C, D)$. It will be clear from the context which interpretation we have in mind. For any such system, the system matrix is defined as the first-order polynomial matrix

$$P(s) = \begin{pmatrix} sI - A & -B \\ C & D \end{pmatrix}.$$ 

If the underlying system is discrete-time, we will rather use the indeterminate $z$ instead of $s$. For a real rational matrix $R$, its normal rank, normrank $R$, is defined as the
rank of \( R \) as a matrix with entries in the field of real rational functions. It is well known that normrank \( R = \max_{\sigma} \text{rank} \ R(\sigma) \). A zero of the system \((A, B, C, D)\) is any complex number \( \lambda \) with the property that \( \text{rank} \ P(\lambda) < \text{normrank} \ P \). The system \((A, B, C, D)\) is called left-invertible (right-invertible) if its transfer matrix \( G(s) = C(sI - A)^{-1}B + D \) is a left-invertible (right-invertible) rational matrix. Assuming that \( A \in \mathbb{R}^{n \times n}, B \in \mathbb{R}^{m \times n}, \) and \( C \in \mathbb{R}^{p \times n} \) we have that \((A, B, C, D)\) is left-invertible (right-invertible) if and only if its system matrix has normal rank \( n + rn \) (\( n + p \)).

If \( M \in \mathbb{R}^{n \times n} \) and \( \mathcal{L} \) is a subspace of \( \mathbb{R}^n \), then \( \langle M \mid \mathcal{L} \rangle \) will denote the smallest \( M \)-invariant subspace containing \( \mathcal{L} \). The largest \( M \)-invariant subspace contained in \( \mathcal{L} \) will be denoted by \( \langle \mathcal{L} \mid M \rangle \). In particular, given \((A, B, C, D)\), the reachable subspace is equal to \( \langle A \mid \text{im} \ B \rangle \) and the unobservable subspace is equal to \( \langle \ker C \mid A \rangle \).

Given the system \((A, B, C, D)\), we define the weakly unobservable subspace \( \mathcal{V} \) to be the smallest subspace \( \mathcal{L} \) of \( \mathbb{R}^n \) with the property that there exists \( F \in \mathbb{R}^{m \times n} \) such that \((A + BF)\mathcal{L} \subset \mathcal{L} \) and \((C + DF)\mathcal{L} = 0 \) (see [14]). In addition, the controllability subspace \( \mathcal{R} \) of \((A, B, C, D)\) is defined as follows:

\[
\mathcal{R} := \langle A + BF \mid \mathcal{V} \cap B \ker D \rangle,
\]

for any \( F \) such that \((A + BF)\mathcal{V} \subset \mathcal{V} \) and \((C + DF)\mathcal{V} = 0 \) (any such \( F \) yields the same \( \mathcal{R} \)). It was shown in [14] that the system \((A, B, C, D)\) is left-invertible if and only if \( \ker B \cap \ker D = 0 \) and \( \mathcal{V} \cap B \ker D = 0 \). Note that \( \mathcal{V} \cap B \ker D = 0 \) if and only if \( \mathcal{R} = 0 \).

Finally, the set of zeros of \((A, B, C, D)\) can be shown to be equal to \( \sigma(A + BF \mid \mathcal{V}/\mathcal{R}) \), for any \( F \) such that \((A + BF)\mathcal{V} \subset \mathcal{V} \) and \((C + DF)\mathcal{V} = 0 \). Here, \( A + BF \mid \mathcal{V}/\mathcal{R} \) is the quotient map of \( A + BF \mid \mathcal{V} \) modulo \( \mathcal{R} \) (see, e.g., [19]).

3. The discrete-time \( H_2 \) problem: No zeros on the unit circle. In this section we shall consider the discrete-time \( H_2 \) problem. Consider the finite-dimensional, linear, time-invariant, discrete-time system \( \Sigma_{\text{dis}} \) given by the equations

\[
\begin{align*}
x_{k+1} &= Ax_k + Bu_k + Ed_k, \\
y_k &= C_1x_k + D_1d_k, \\
z_k &= C_2x_k + D_2u_k.
\end{align*}
\]

There will be no assumptions on the direct feedthrough matrices \( D_1 \) and \( D_2 \). In the present section, however, we will have assumptions on the absence of system zeros on the unit circle in the complex plane: it will be assumed that \((A, B, C_2, D_2)\) and \((A, E, C_1, D_1)\) do not have zeros on the unit circle \( |z| = 1 \). In the next section we will drop these assumptions and treat the completely general case. Of course, it will be a standing assumption that \((A, B)\) is stabilizable and that \((C_1, A)\) is detectable, both with respect to the open unit disc.

We will consider discrete-time controllers \( \Gamma_{\text{dis}} \) given by (2.8). For any internally stabilizing controller \( \Gamma_{\text{dis}} \), let \( J_{\Sigma_{\text{dis}}}^{\star}(\Gamma_{\text{dis}}) \) be its \( H_2 \) performance. Denote by \( J^{\star} \) the optimal performance, i.e., the infimum over all internally stabilizing controllers \( \Gamma_{\text{dis}} \).

For a given matrix \( M \), we will denote by \( M^{\dagger} \) its Moore–Penrose inverse. The solution of the discrete-time \( H_2 \) optimal control problem centers around the following two algebraic Riccati equations:

\[
\begin{align*}
(3.2) \quad P &= A^TPA + C_2^TC_2 - (C_2^TD_2 + A^TPB)(D_2^TD_2 + B^TPB)^{\dagger}(D_2^TC_2 + B^TPA), \\
(3.3) \quad Q &= AQ^TA + EE^T - (AQ^TC_1^T + ED_1^T)(D_1^TD_1^T + C_1Q^TC_1^T)^{\dagger}(D_1^TE^T + C_1QA^T).
\end{align*}
\]
For any real symmetric matrix $P$, we shall denote
\begin{equation}
D_P := (D_2^T D_2 + B^T P B)^{\frac{1}{2}},
\end{equation}
\begin{equation}
C_P := D_P^+(D_2^T C_2 + B^T P A).
\end{equation}

Note that, since for any matrix $M \geq 0$ we have $(M^{\frac{1}{2}} + (M^+)^{\frac{1}{2}}$, we have $D_P C_P = (D_2^T D_2 + B^T P B)^{\frac{1}{2}}(D_2^T C_2 + B^T P A)$. If, in addition, $P$ is a real symmetric solution of (3.2), then $C_P D_P = A^T PA - P + C_2^T C_2$. Note also that $D_P$ is symmetric by definition.

Finally, since $\text{im} (D_2^T C_2 + B^T P A) \subset \text{im} D_P$, we have $D_P C_P = D_2^T C_2 + B^T P A$. (Note that it is a property of the Moore–Penrose inverse that $MM^+$ is the orthogonal projection onto $\text{im} M$.)

The following is a corrected and slightly extended version of a theorem from [14].

A proof can be given along the lines of the proof of [14, Thm. 18].

**Theorem 3.1.** Consider the system $(A, B, C_2, D_2)$ together with the algebraic Riccati equation (3.2). The following two statements are equivalent:

(i) $(A, B)$ is stabilizable and $(A, B, C_2, D_2)$ has no zeros on the unit circle $|z| = 1$,

(ii) Equation (3.2) has a real symmetric solution $P$ with the following property: there exists a matrix $F$ such that
\begin{equation}
|\sigma(A - BD_P^T C_P + B(I - D_P^T D_P)F)| < 1.
\end{equation}

Furthermore, if $P$ satisfies this condition, it is the unique real symmetric solution of (3.2) for which this condition holds. In addition, $P$ is positive semidefinite and is in fact the largest real symmetric solution of (3.2).

Next we consider the dual algebraic Riccati equation (3.3). For any real symmetric matrix $Q$, denote
\begin{equation}
D_Q := (D_1^T D_1 + C_1 Q C_1^T)^{\frac{1}{2}},
\end{equation}
\begin{equation}
E_Q := (A Q C_1^T + E D_1^T) D_Q^+.
\end{equation}

By dualizing the previous theorem, the corresponding result on the Riccati equation (3.3) can be found:

**Theorem 3.2.** Consider the system $(A, E, C_1, D_1)$ together with the algebraic Riccati equation (3.3). The following two statements are equivalent:

(i) $(C_1, A)$ is detectable and $(A, E, C_1, D_1)$ has no zeros on the unit circle $|z| = 1$.

(ii) Equation (3.3) has a real symmetric solution $Q$ with the following property: there exists a matrix $K_1$ such that
\begin{equation}
|\sigma(A - EQ D_Q^+ C_1 + K_1(I - D_Q D_Q^+ C_1))| < 1.
\end{equation}

Furthermore, if $Q$ satisfies this condition, it is the unique real symmetric solution of (3.3) for which this condition holds. In addition, $Q$ is positive semidefinite and is in fact the largest real symmetric solution of (3.3).

In the remainder of the section we will always denote by $P$ and $Q$ the largest real symmetric solution of (3.2) and (3.3), respectively. Now we will state the main result of this section:

**Theorem 3.3.** Consider the system (3.1). Assume that $(A, B)$ is stabilizable and $(C_1, A)$ is detectable. Assume that $(A, B, C_2, D_2)$ and $(A, E, C_1, D_1)$ have no zeros on the unit circle. Then we have the following:
(i) \[ J^* = \text{tr} \left( E^TPE \right) + \text{tr} \left( C_PQC_P^T \right) - \text{tr} \left( (D_P N^* D_Q)(D_P N^* D_Q)^T \right), \]

where \( N^* \) is defined by

\[ N^* := -(D_P^T)^2 (D_P C_P QC_P^T + B_P E_D D^T_P)(D_P^T)^2. \]

(ii) There exists an optimal controller, i.e., an internally stabilizing controller \( \Gamma_{\text{dis}}^* \) such that \( J_{\Sigma_{\text{dis}}} (\Gamma_{\text{dis}}^*) = J^* \). One such optimal controller is given by the following “construction”:

(a) Choose a state feedback matrix \( F \) such that \( |\sigma(A + BF)| < 1 \) and \( C_P + D_PF = 0 \).
(b) Choose an output injection matrix \( G \) such that \( |\sigma(A + GC_1)| < 1 \) and \( E_Q + GD_Q = 0 \).
(c) Define \( \Gamma_{\text{dis}}^* = (K^*, L^*, M^*, N^*) \) by choosing \( N^* \) given by (3.11), and by choosing \( K^* := A + BF + GC_1 - BN^* C_1 \), \( L^* := BN^* - G \), and \( M^* := F - N^* C_1 \).

In the remainder of this section we shall prove this theorem. In addition to the system \( \Sigma_{\text{dis}} \), consider the system \( \Sigma_{\text{dis}, P} \) given by the equations

\[ x_{k+1} = Ax_k + Bu_k + Ed_k, \]
\[ y_k = C_1 x_k + D_1 d_k, \]
\[ z_k = C_P x_k + D_P u_k, \]

with \( P \) the largest real symmetric solution of the algebraic Riccati equation (3.2). The following basic lemma can be proven by a standard completion-of-the-squares argument:

**Lemma 3.4.** For every compensator \( \Gamma_{\text{dis}} = (K, L, M, N) \) we have \( \Gamma_{\text{dis}} \) internally stabilizes \( \Sigma_{\text{dis}} \) if and only if \( \Gamma_{\text{dis}} \) internally stabilizes \( \Sigma_{\text{dis}, P} \). For any such compensator we have

\[ J_{\Sigma_{\text{dis}}} (\Gamma_{\text{dis}}) = \text{tr} \left( E^TPE \right) + 2\text{tr} \left( D_1^T N^T B_P E \right) + J_{\Sigma_{\text{dis}, P}} (\Gamma_{\text{dis}}). \]

In addition to \( \Sigma_{\text{dis}, P} \) we consider the system \( \Sigma_{\text{dis}, P, Q} \) defined by

\[ x_{k+1} = Ax_k + Bu_k + E_Q d_k, \]
\[ y_k = C_1 x_k + D_1 d_k, \]
\[ z_k = C_P x_k + D_P u_k, \]

with \( Q \) the largest real symmetric solution of the dual algebraic Riccati equation (3.3). It is clear that the \( H_2 \) performance of a given compensator \( \Gamma_{\text{dis}} \) applied to \( \Sigma_{\text{dis}} \) is equal to the \( H_2 \) performance of the dual compensator \( \Gamma_{\text{dis}}^* := (K^*, M^*, L^*, N^*) \) applied to the dual system \( \Sigma_{\text{dis}}^T \). By applying Lemma 3.4 to the dual system \( \Sigma_{\text{dis}}^T \) and the dual compensator \( \Gamma_{\text{dis}}^* \) we thus arrive at the following theorem:

**Theorem 3.5.** For every compensator \( \Gamma_{\text{dis}} = (K, L, M, N) \) we have \( \Gamma_{\text{dis}} \) internally stabilizes \( \Sigma_{\text{dis}} \) if and only if \( \Gamma_{\text{dis}} \) internally stabilizes \( \Sigma_{\text{dis}, P, Q} \). For any such compensator we have

\[ J_{\Sigma_{\text{dis}}} (\Gamma_{\text{dis}}) = \text{tr} \left( E^TPE \right) + \text{tr} \left( C_P QC_P^T \right) + 2\text{tr} \left( D_1^T N^T B_P E \right) + J_{\Sigma_{\text{dis}, P, Q}} (\Gamma_{\text{dis}}). \]

Now note that in the above formula the first two terms do not depend on the compensator \( \Gamma_{\text{dis}} \). The remaining three terms do depend on the compensator. Also
note that in the closed-loop system \( \Sigma_{\text{dis},P,Q} \times \Gamma_{\text{dis}} \) the direct feedthrough matrix from the disturbance input to the output to be controlled is equal to \( D_P N D_Q \). As a consequence, \( J_{\Sigma_{\text{dis},P,Q}}(\Gamma_{\text{dis}}) \geq \text{tr} \left( (D_P N D_Q)(D_P N D_Q)^T \right) \), with equality if and only if the transfer matrix \( G_{P,Q,\Gamma_{\text{dis}}}(z) \) of the closed-loop system \( \Sigma_{\text{dis},P,Q} \times \Gamma_{\text{dis}} \) is equal to the constant matrix \( D_P N D_Q \). It thus follows immediately from Theorem 3.5 that

**Lemma 3.6.** For every internally stabilizing compensator \( \Gamma_{\text{dis}} = (K, L, M, N) \) we have

\[
J_{\Sigma_{\text{dis}}}(\Gamma_{\text{dis}}) \geq \text{tr} \left( E^T P E \right) + \text{tr} \left( C_P Q C_P^T \right) + 2 \text{tr} \left( D_P^T N^T B^T P E \right) + 2 \text{tr} \left( C_P Q C_1^T N^T D_P^T \right) + \text{tr} \left( (D_P N D_Q)(D_P N D_Q)^T \right),
\]

with equality if and only if \( G_{P,Q,\Gamma_{\text{dis}}}(z) = D_P N D_Q \).

This lemma shows that, in order to minimize \( J_{\Sigma_{\text{dis}}}(\Gamma_{\text{dis}}) \) over all internally stabilizing compensators, we should do the following:

(i) First minimize the quadratic matrix function

\[
\Phi(N) := 2 \text{tr} \left( D_P^T N^T B^T P E \right) + 2 \text{tr} \left( C_P Q C_1^T N^T D_P^T \right) + \text{tr} \left( (D_P N D_Q)(D_P N D_Q)^T \right),
\]

yielding an optimal \( N^* \).

(ii) Next find a compensator \( \Gamma^*_{\text{dis}} \), described by the quadruple \((K^*, L^*, M^*, N^*)\), that is internally stabilizing and yields \( G_{P,Q,\Gamma^*_{\text{dis}}}(z) = D_P N^* D_Q \), i.e., the closed-loop system \( \Sigma_{\text{dis},P,Q} \times \Gamma^*_{\text{dis}} \) has the constant transfer matrix \( D_P N^* D_Q \).

Indeed, if \( N^* \) minimizes \( \Phi(N) \) and if \( G_{P,Q,\Gamma^*_{\text{dis}}}(z) = D_P N^* D_Q \), then we have

\[
J_{\Sigma_{\text{dis}}}(\Gamma^*_{\text{dis}}) = \text{tr} \left( E^T P E \right) + \text{tr} \left( C_P Q C_P^T \right) + \Phi(N^*),
\]

while for any internally stabilizing compensator \( \Gamma_{\text{dis}} = (K, L, M, N) \) we have

\[
J_{\Sigma_{\text{dis}}}(\Gamma_{\text{dis}}) \geq \text{tr} \left( E^T P E \right) + \text{tr} \left( C_P Q C_P^T \right) + \Phi(N) \geq \text{tr} \left( E^T P E \right) + \text{tr} \left( C_P Q C_P^T \right) + \Phi(N^*).
\]

This clearly implies that

\[
J^* = \text{tr} \left( E^T P E \right) + \text{tr} \left( C_P Q C_P^T \right) + \Phi(N^*)
\]

and that

\[
J_{\Sigma_{\text{dis}}}(\Gamma^*_{\text{dis}}) = J^*.
\]

We will first study the minimization of \( \Phi(N) \).

**Lemma 3.7.** Let \( \Phi(N) \) be defined by (3.15). Define

\[
R^* := D_P^T (D_P C_P Q C_P^T + B^T P E D_Q^T) D_Q^T.
\]

Then

\[
\Phi^* := \min \{ \Phi(N) \mid N \in \mathbb{R}^{m \times p} \} = -\text{tr} \left( R^* R^{*T} \right).
\]

\( N \) minimizes \( \Phi \), i.e., \( \Phi(N) = \Phi^* \), if and only if \( N \) is a solution to the linear equation \( D_P N D_Q = -R^* \). One particular solution of this linear equation is given by \( N^* = -D_P^T R^* D_Q^+ \). We have \( \Phi^* = -\text{tr} \left( (D_P N^* D_Q)(D_P N^* D_Q)^T \right) \).
Proof. Using the facts that
\[
\ker D_Q \subset \ker (D_P C_P Q_1 + D^T P E D^T_1), \\
\im D_P \supset \im (D_P C_P Q_1 + D^T P E D^T_1),
\]
it can be shown by straightforward calculation that
\[
\Phi(N) = -\text{tr} \left( R^* R^{*T} \right) + \text{tr} \left( (D_P N D_Q + R^*)(D_P N D_Q + R^*)^T \right).
\]
The equation \(D_P N D_Q = -R^*\) has a solution since \(\ker D_Q = \ker D_Q^T = \ker D_Q^+ \subset \ker R^*\) and \(\im D_P = \im D_P^+ = \im D_P^+ \supset \im R^*\). Clearly, one particular solution is then given by \(N^* = -D_P^+ R^* D_Q^+\). Finally, the expression for \(\Phi^*\) can be checked in a straightforward manner.

Next we study the question whether, starting with \(N^*\) above, it is possible to find \(K^*, L^*, M^*\) such that the resulting compensator \(\Gamma^*_{\text{dis}} = (K^*, L^*, M^*, N^*)\) yields a closed-loop system \(\Sigma_{\text{dis},P,Q} \times \Gamma^*_{\text{dis}}\) with constant transfer matrix \(D_P N^* D_Q\). We will first prove the following lemma:

**Lemma 3.8.** Assume that \((A, B)\) is stabilizable and that \((A, B, C_2, D_2)\) has no zeros on the unit circle. Let \(P\) be the largest real symmetric solution of the algebraic Riccati equation (3.2). There exists a matrix \(F\) such that

(i) \(|\sigma(A + BF)| < 1\),

(ii) \(C_P + D_PF = 0\).

**Proof.** Let \(F_1\) be such that (3.6) holds, and define \(F = -D_P^+ C_P + (I - D_P^+ D_P)F_1\). Then (i) is satisfied. To prove (ii), note that \(C_P \subset \im D_P^+ = \im D_P\). Consequently, \(-D_P^+ D_P C_P = -C_P\), which proves (ii).

We will also need the dual of this lemma, which reads as follows:

**Lemma 3.9.** Assume that \((C_1, A)\) is detectable and that \((A, E, C_1, D_1)\) has no zeros on the unit circle. Let \(Q\) be the largest real symmetric solution of the dual algebraic Riccati equation (3.3). There exists a matrix \(G\) such that

(i) \(|\sigma(A + G C_1)| < 1\),

(ii) \(E_Q + GD_Q = 0\).

Now we show that by suitable choice of compensator \(\Gamma^*_{\text{dis}}\), the transfer matrix of \(\Sigma_{\text{dis},P,Q} \times \Gamma^*_{\text{dis}}\) can be made equal to any constant matrix product \(M_1M_2\), as long as \(\im D_P \subset \im M_1\) and \(\ker D_Q \subset \ker M_2\).

**Lemma 3.10.** Consider the system (3.1). Assume that \((A, B)\) is stabilizable and \((C_1, A)\) is detectable. Assume that \((A, B, C_2, D_2)\) and \((A, E, C_1, D_1)\) have no zeros on the unit circle. Let \(P\) and \(Q\) be the largest real symmetric solution of the algebraic Riccati equation (3.2) and (3.3), respectively. Then for any pair of matrices \(M_1, M_2\) such that the product \(M_1M_2\) is defined and such that \(\im D_P \subset \im M_1\) and \(\ker D_Q \subset \ker M_2\) there exists an internally stabilizing compensator \(\Gamma_{\text{dis}}\) such that the transfer matrix of \(\Sigma_{\text{dis},P,Q} \times \Gamma_{\text{dis}}\) is equal to the constant \(M_1M_2\).

Specifically, for given \(M_1\) and \(M_2\) let \(F_2\) be a solution of \(M_1 = D_PF_2\) and \(G_2\) be a solution of \(M_2 = -G_2 D_Q\) and take \(F\) such that the conditions in Lemma 3.8 are satisfied and \(G\) such that the conditions of Lemma 3.9 are satisfied. Then the compensator \(\Gamma_{\text{dis}} := (K, L, M, N)\) with \(K := A + BF + GC_1 + BF_2G_2C_1, L := -BF_2G_2 - G, M := F + F_2G_2C_1,\) and \(N := -F_2G_2\) satisfies the requirements.

**Proof.** The equations of the compensator are given by (2.8). Using the specifications of \(K, L, M,\) and \(N\) given above, we find that the error \(e_k := w_k - x_k\) satisfies \(e_{k+1} = (A + GC_1) e_k\). Thus, if \(w_0 = 0\) and \(x_0 = 0\), we have \(x_k = w_k\) for all \(k\). In particular, this implies that \(u_k = Fx_k + F_2M_2w_k\). The output of the closed-loop system
is then equal to $z_k = Cpx_k + Dpu_k = M_1M_2w_k$. This implies that the closed-loop transfer matrix is equal to the constant matrix $M_1M_2$. Finally, the spectrum of the closed-loop system matrix $A_e$ is easily shown to be equal to $\sigma(A + BF) \cup \sigma(A + GC_1)$. This implies that the closed-loop system is internally stable.

Clearly, if in this lemma we take $M_1 = D_P$ and $M_2 = N^*D_Q$, we arrive at an internally stabilizing compensator $\Gamma_{dis}$ such that the closed-loop transfer matrix is equal to the constant matrix $D_PN^*D_Q$. In the formulas for the compensator as given in the lemma, we should then take $F_2 = I$ and $G_2 = -N^*$. The result of Theorem 3.3 follows immediately by combining the above lemmas.

Remark 3.11. For later use we note that Lemma 3.8 also provides a resolution of the discrete-time linear quadratic problem for the case that $(A, B, C_2, D_2)$ has no zeros on the unit circle (see also [14]). Given $x_{k+1} = Ax_k + Bu_k$, the problem is to minimize the cost-functional $J(x_0, u) := \sum_k ||(C_2x_k + D_2u_k)||^2$ over all inputs $u = \{u_k\}$ such that $x_k \to 0$. It was pointed out in [14] that for each such input $u$ we have the completion-of-the-squares formula $J(x_0, u) = x_0^TPx_0 + J_P(x_0, u)$, with $J_P(x_0, u) := \sum_k ||Cpx_k + Dpu_k||^2$. Thus, if we take $F$ satisfying (i) and (ii) of Lemma 3.8, then the input $u_k = Fx_k$ leads to the optimal cost $J^*(x_0) = x_0^TPx_0$. Note that we could also formulate the linear quadratic problem as a minimization over all internally stabilizing feedback laws: minimize the cost-functional $J(x_0, F) := \sum_k ||(C_P + D_PF)x_k||^2$ over all $F \in \mathbb{R}^{m \times n}$ such that $|\sigma(A + BF)| < 1$. By the above argument, any $F$ satisfying (i) and (ii) of Lemma 3.8 is then optimal and the optimal cost is again given by $x_0^TPx_0$.

Remark 3.12. An interesting question is under what conditions the Moore–Penrose inverse $(D_2^TA_2 + B^TPB)^+$ reduces to the inverse $(D_2^TA_2 + B^TPB)^{-1}$, equivalently, under what conditions $D_2^TA_2 + B^TPB$ is positive definite. Using the ideas from [14] it can be shown that if $P$ is a positive semidefinite solution of the algebraic Riccati equation (3.2), then $D_2^TA_2 + B^TPB > 0$ if and only if $(A, B, C_2, D_2)$ is a left-invertible system. Of course, dually, if $Q$ is a positive semidefinite solution of the algebraic Riccati equation (3.3), then $D_1^TA_1 + C_1Q^T > 0$ if and only if the system $(A, E, C_1, D_1)$ is right-invertible. In view of this, it is perhaps more natural to call the discrete-time $H_2$ problem regular if $(A, B, C_2, D_2)$ is a left-invertible system and $(A, E, C_1, D_1)$ is a right-invertible system.

4. The discrete-time $H_2$ problem: The general case. In this section we will extend the results of the previous section and treat the discrete-time $H_2$ problem in its full generality. This means that we will drop the assumption on the absence of zeros on the unit circle that was made in the previous section. First we will prove that also without the assumption that $(A, B, C_2, D_2)$ has no zeros on the unit circle, the Riccati equation (3.2) has a largest real symmetric solution. We will prove that this solution can be obtained as the limit of solutions of algebraic Riccati equations associated with suitable perturbations of the system $(A, B, C_2, D_2)$.

**Theorem 4.1.** If $(A, B)$ is stabilizable, then the Riccati equation (3.2) has a largest real symmetric solution, say $P$. $P$ is positive semidefinite. We have $P = \lim_{\epsilon \to 0} P_\epsilon$, where for $\epsilon > 0$ $P_\epsilon$ is the largest real symmetric solution of the algebraic Riccati equation

$$
\begin{align*}
A^TP_\epsilon A - P_\epsilon + C_2^T C_2 + \epsilon^2 I \\
-(A^TP_\epsilon B + C_2^TD_2)(D_2^TA_2 + B^TP_\epsilon B)^+(B^TP_\epsilon A + D_2^TC_2) &= 0.
\end{align*}
$$

\[ (4.1) \]
Remark 4.2. Note that (4.1) is the Riccati equation associated with the perturbed system \((A, B, (C_2, \varepsilon I), (D_2^\varepsilon 0))\). (Here, \(I\) denotes the \(n\times n\) identity matrix, and \(0\) denotes the \(n\times m\) zero matrix). For \(\varepsilon > 0\), the perturbed system has no zeros. Consequently, the existence of \(P_\varepsilon\) follows from Theorem 3.1.

The idea of the proof of Theorem 4.1 is to show first that the \(P_\varepsilon\) indeed converge to some matrix \(P\) and next to show that \(P\) satisfies (3.2). The difficulty is that in the general case we are considering, the term \(D_2^\varepsilon D_2^\varepsilon + B^TP_\varepsilon B\) need not be invertible, so that we cannot conclude that \((D_2^\varepsilon D_2^\varepsilon + B^TP_\varepsilon B)^+\) converges to \((D_2^0 D_2^0 + B^TP_0 B)^+\). We will show, however, that we can get around this difficulty by considering the so-called linear matrix inequality. Our proof is split up in three lemmas. In the following, let \(J(x_0, u)\) be the cost-functional of the linear quadratic problem, and let \(J^*(x_0)\) be the optimal cost (see Remark 3.11).

**Lemma 4.3.** Let \(P_\varepsilon\) be the largest real symmetric solution of (4.1). There exists a real positive semidefinite matrix \(P\) such that \(P_\varepsilon \downarrow P\) \((\varepsilon \downarrow 0)\). For all \(x_0 \in \mathbb{R}^n\) we have \(J^*(x_0) \leq x_0^TPx_0\).

**Proof.** Let \(J_\varepsilon(x_0, u) := \sum_k \|C_px_k + D_pu_k\|^2 + \varepsilon^2 \|x_k\|^2\), and let \(J_\varepsilon^*(x_0)\) be the infimum of \(J_\varepsilon(x_0, u)\) over all \(u\) such that \(x_k \to 0\). According to Remark 3.11 we have \(J_\varepsilon^*(x_0) = x_0^TP_\varepsilon x_0\). From this interpretation it follows that \(P_\varepsilon\) is monotonically non-increasing as \(\varepsilon \downarrow 0\). Being bounded from below by 0, this yields the existence of a limit \(P\). Obviously, for all \(\varepsilon > 0\) we have \(J^*(x_0) \leq J_\varepsilon^*(x_0) = x_0^TP_\varepsilon x_0\), so \(J^*(x_0) \leq x_0^TP_0 x_0\). Conversely, for all \(\varepsilon > 0\) and for all \(u\) we have \(J_\varepsilon(x_0, u) \geq x_0^TP_\varepsilon x_0\). Taking the limit on both sides this yields \(J(x_0, u) \geq x_0^TPx_0\) for all \(u\). Taking the infimum over \(u\) then yields the converse inequality. \(\square\)

**Lemma 4.4.** \(P\) is the largest real symmetric solution of the linear matrix inequality

\[
\begin{pmatrix}
A^TPA - P + C_2^TC_2 & C_2^TD_2^\varepsilon + A^TP\varepsilon B
\end{pmatrix} \geq 0.
\]

**Proof.** Denote the left-hand side of (4.1) by \(R_\varepsilon(P_\varepsilon)\). Also consider the linear matrix inequality associated with the perturbed system:

\[
M_\varepsilon(P_\varepsilon) := \begin{pmatrix}
A^TP_\varepsilon A - P_\varepsilon + C_2^TC_2 + \varepsilon^2 I & C_2^TD_2^\varepsilon + A^TP_\varepsilon B
\end{pmatrix} \geq 0.
\]

We have \(M_\varepsilon(P_\varepsilon) \geq 0\) if and only if \(R_\varepsilon(P_\varepsilon) \geq 0\). This follows from the fact that the latter is equal to the Schur complement of \(D_2^\varepsilon D_2^\varepsilon + B^TP_\varepsilon B\) in \(M_\varepsilon(P_\varepsilon)\). The Schur complement is defined here with matrix inverse replaced by Moore–Penrose inverse. This can be done because of the fact that \(\ker(D_2^\varepsilon D_2^\varepsilon + B^TP_\varepsilon B) \subset \ker(C_2^TD_2^\varepsilon + A^TP_\varepsilon B)\).

Since \(R_\varepsilon(P_\varepsilon) = 0\), we indeed have \(M_\varepsilon(P_\varepsilon) \geq 0\). Taking the limit \(\varepsilon \downarrow 0\) then yields \(M(P) \geq 0\). To show that \(P\) is the largest real symmetric solution, let \(P_1\) be any real symmetric solution of the linear matrix inequality. Using a standard completion-of-the-squares argument then yields \(J(x_0, u) \geq x_0^TP_1 x_0\) for any \(x_0\) and any \(u\) such that \(x_k \to 0\). Taking the infimum over all such \(u\) yields \(x_0^TPx_0 = J^*(x_0) \geq x_0^TP_1 x_0\). \(\square\)

Now we will show that \(P\) in fact satisfies the algebraic Riccati equation (3.2). Denote

\[
R(P) := A^TPA - P + C_2^TC_2 - (C_2^TD_2^\varepsilon + A^TP\varepsilon B)(D_2^\varepsilon D_2^\varepsilon + B^TP\varepsilon B)^+(D_2^\varepsilon C_2 + B^TP)\]

where
Again, by the fact that \( \ker(D_2^T D_2 + B^T P B) \subset \ker(C_2^T D_2 + A^T P B) \), \( R(P) \) is equal to the Schur complement of \( D_2^T D_2 + B^T P B \) in \( M(P) \). In particular this implies that
\[
\text{rank } M(P) = \text{rank } (D_2^T D_2 + B^T P B) + \text{rank } R(P).
\]
In order to prove that \( R(P) = 0 \) we should therefore prove that \( P \) has the property expressed in the following lemma:

**Lemma 4.5.** \( \text{rank } M(P) = \text{rank } (D_2^T D_2 + B^T P B) \).

**Proof.** Let \( \tilde{C} \) and \( \tilde{D} \) be matrices such that
\[
M(P) = \left( \begin{array}{cc}
\tilde{C} & \tilde{D}
\end{array} \right)^T \left( \begin{array}{cc}
\tilde{C} & \tilde{D}
\end{array} \right).
\]
Again using a standard completion-of-the-squares argument, for any initial state \( x_0 \) and for any input sequence \( u \) such that \( x_k \to 0 \) we have
\[
J(x_0, u) = x_0^T P x_0 + \sum_k \|\tilde{C} x_k + \tilde{D} u_k\|^2 \geq x_0^T x_0 + \|\tilde{C} x_0 + \tilde{D} u_0\|^2.
\]
From Lemma 4.3 we have that \( J^*(x_0) = x_0^T P x_0 \). In particular this implies that the infimum of \( \|\tilde{C} x_0 + \tilde{D} u_0\|^2 \) over all \( u_0 \in \mathbb{R}^m \) is equal to 0. Consequently, for all \( x_0 \) there exists \( u_0 \in \mathbb{R}^m \) such that \( \tilde{C} x_0 + \tilde{D} u_0 = 0 \). This implies \( \text{im } \tilde{C} \subset \text{im } \tilde{D} \) so
\[
\text{rank } M(P) = \text{rank } \left( \begin{array}{cc}
\tilde{C} & \tilde{D}
\end{array} \right) = \text{rank } \tilde{D} = \text{rank } (D_2^T D_2 + B^T P B).
\]

Clearly, the proof of Theorem 4.1 follows by combining these three lemmas. The fact that \( P \) is the largest real symmetric solution of the algebraic Riccati equation follows by noting that any real symmetric solution is also a solution of the linear matrix inequality and by applying Lemma 4.4.

**Remark 4.6.** For later use, note that by combining the above results with Remark 3.11 we obtain that also for the general case the optimal cost \( J^*(x_0) \) of the discrete-time linear quadratic problem associated with the system \( (A, B, C_2, D_2) \) is given by \( J^*(x_0) = x_0^T P x_0 \), with \( P \) the largest real symmetric solution of the Riccati equation (3.2). We will also need the dual result of Theorem 4.1, which is stated below:

**Theorem 4.7.** If \( (C_1, A) \) is detectable, then the Riccati equation (3.3) has a largest real symmetric solution, say \( Q \). \( Q \) is positive semidefinite. We have \( Q = \lim_{\epsilon \to 0} Q_\epsilon \), where for \( \epsilon > 0 \) \( Q_\epsilon \) is the largest real symmetric solution of the algebraic Riccati equation
\[
A Q_\epsilon A^T - Q_\epsilon + E E^T + \epsilon^2 I
\]
(4.3) \( - (AQ_\epsilon C_1^T + ED_1^T)(D_1 D_1^T + C_1 Q_\epsilon C_1^T)^+ (C_1 Q_\epsilon A^T + D_1 E^T) = 0. \)

Now we are in a position to state the main results of this section. It turns out that also for the discrete-time \( H_2 \) problem in its full generality, so without any assumptions on the zeros, the optimal performance \( J^* \) is given by (3.10), with \( P \) and \( Q \) the largest real symmetric solutions of the respective Riccati equations. However, in general no optimal controller will exist. We will, however, derive necessary and sufficient conditions for the existence of an optimal controller. Our first main result deals with the optimal performance.

**Theorem 4.8.** Consider the system (3.1). Assume that \( (A, B) \) is stabilizable and \( (C_1, A) \) is detectable. Then the optimal performance \( J^* \) is given by (3.10), where \( P \) and \( Q \) are the largest real symmetric solutions of (3.2) and (3.3), respectively.
Proof. In addition to the system (3.1), consider its perturbation \( \Sigma_{\text{dis}}^\varepsilon \):

\[
\begin{align*}
    x_{k+1} &= Ax_k + Bu_k + (E \varepsilon I)u_k, \\
    y_k &= C_1x_k + (D_1 0)u_k, \\
    z_k &= (C_2^\varepsilon)x_k + (D_2 0)u_k.
\end{align*}
\]

Let \( J_{\Sigma_{\text{dis}}}^\varepsilon (\Gamma_{\text{dis}}) \) denote the \( H_2 \) performance, and let \( J^\varepsilon \) denote the optimal \( H_2 \) performance. Since, for \( \varepsilon > 0 \), neither \((A, B, (C_1^\varepsilon), (D_2^\varepsilon))\) nor \((A, (E \varepsilon I), C_1, (D_1 0))\) have zeros; we can apply Theorem 3.3 to obtain

\[
J^\varepsilon = \text{tr} \left( (EE^T + \varepsilon^2 I)P_\varepsilon \right) + \text{tr} \left( (A^TP_\varepsilon A - P_\varepsilon + C_2^T C_2 + \varepsilon^2 I)Q_\varepsilon \right) - \text{tr} \left( (D_{P_\varepsilon}^* N_{DQ_\varepsilon}^* D_{Q_\varepsilon}^*)^T \right),
\]

where \( P_\varepsilon \) and \( Q_\varepsilon \) are the largest real symmetric solutions of (4.1) and (4.3), respectively, and where \( D_{P_\varepsilon}^* \), \( N_{DQ_\varepsilon}^* \), and \( D_{Q_\varepsilon}^* \) are defined by (3.4), (3.11), and (3.7), with \( P \) and \( Q \) replaced by \( P_\varepsilon \) and \( Q_\varepsilon \). From Lemma 3.7, recall that

\[
-\text{tr} \left( (D_{P_\varepsilon}^* N_{DQ_\varepsilon}^* D_{Q_\varepsilon}^*)^T \right) \leq (N) \min_{N \in \mathbb{R}} (N),
\]

with

\[
\Phi_\varepsilon (N) := 2\text{tr} \left( \left( \begin{array}{c} D_1 \\ 0 \end{array} \right)^T N^T B^T P_\varepsilon (E \varepsilon I) \right) + 2\text{tr} \left( C_\varepsilon P_\varepsilon C_\varepsilon^T N^T D_{P_\varepsilon} \right) + \text{tr} \left( (D_{P_\varepsilon}^* N_{DQ_\varepsilon}^* D_{Q_\varepsilon}^*)^T \right).
\]

Since \( P \rightarrow P_\varepsilon \) and \( Q \rightarrow Q_\varepsilon \), we see that for every \( N \) we have \( \Phi_\varepsilon (N) \rightarrow \Phi (N) \) \((\varepsilon \downarrow 0)\), where \( \Phi (N) \) is defined by (3.15). Since of course for all \( \varepsilon > 0 \) we have \( J^\varepsilon \leq J^*_\varepsilon \) we see that for all \( \varepsilon > 0 \), for all \( N \) we have

\[
J^* \leq \text{tr} \left( (EE^T + \varepsilon^2 I)P_\varepsilon \right) + \text{tr} \left( (A^TP_\varepsilon A - P_\varepsilon + C_2^T C_2 + \varepsilon^2 I)Q_\varepsilon \right) + \Phi_\varepsilon (N).
\]

Now, letting \( \varepsilon \downarrow 0 \) on the left in this inequality, we find that for all \( N \)

\[
J^* \leq \text{tr} \left( EE^T P \right) + \text{tr} \left( (A^TPA - P + C_2^T C_2)Q \right) + \Phi (N).
\]

Finally, taking the minimum over all \( N \), this yields

\[
J^* \leq \text{tr} \left( EE^T P \right) + \text{tr} \left( C_\varepsilon^T C_\varepsilon P \right) - \text{tr} \left( (D_{P_\varepsilon} N_{DQ_\varepsilon} D_{Q_\varepsilon})^T \right).
\]

To prove the converse inequality note that by using the fact that \( P \) and \( Q \) satisfy (3.2) and (3.3) we can apply a repeated completion-of-the-squares argument as in §3 to obtain that for any internally stabilizing compensator \( \Gamma_{\text{dis}} \) we have

\[
J_{\Sigma_{\text{dis}}}^\varepsilon (\Gamma_{\text{dis}}) \geq \text{tr} \left( EE^T P \right) + \text{tr} \left( C^T Q C^T \right) + \Phi (N^*) \tag{4.5}
\]

Taking the infimum over all such \( \Gamma_{\text{dis}} \) yields the desired inequality. \( \square \)

Next we will study the question: Under what conditions does there exist an optimal controller? Again, let \( P \) and \( Q \) be the largest real symmetric solutions of the respective Riccati equations. Define a system \( \Sigma_{\text{dis},P,Q} \) by (3.14). Again, for any internally stabilizing compensator \( \Gamma_{\text{dis}} = (K, L, M, N) \) we have the inequality (4.5).
As noted in §3, we have equality if \( N = N^* \) and \( \Gamma_{\text{dis}} \) has the property that the closed loop system \( \Sigma_{\text{dis},P,Q} \times \Gamma_{\text{dis}} \) has the constant transfer matrix \( D_P N^* D_Q \). Of course, the latter statement only gives a sufficient condition for a compensator to be optimal. In the following theorem we will give necessary and sufficient conditions for optimality. Let \( R^* \) be as defined in Lemma 3.7.

**Theorem 4.9.** A controller \( \Gamma_{\text{dis}} \) is optimal if and only if \( \Sigma_{\text{dis},P,Q} \times \Gamma_{\text{dis}} \) is internally stable and has constant transfer matrix \( R^* \).

**Proof.** If \( \Gamma_{\text{dis}} = (K, L, M, N) \) is optimal, then we have

\[
J_{\Sigma_{\text{dis}}} (\Gamma_{\text{dis}}) = \text{tr} (E^T P E) + \text{tr} (C_P^T Q C_P) + \Phi^*.
\]

By Lemma 3.6 we also have

\[
J_{\Sigma_{\text{dis}}} (\Gamma_{\text{dis}}) \geq \text{tr} (E^T P E) + \text{tr} (C_P^T Q C_P) + \Phi(N).
\]

This clearly yields \( \Phi(N) = \Phi^* \), i.e., \( N \) minimizes the function \( \Phi \). Again by Lemma 3.6 this implies that \( \Sigma_{\text{dis},P,Q} \times \Gamma_{\text{dis}} \) has the constant transfer matrix \( D_P N D_Q \). However, since \( N \) minimizes \( \Phi \), by Lemma 3.7 we have \( D_P N D_Q = -R^* \). The converse statement is also an immediate consequence of Lemma 3.6.

Our aim is to reformulate these conditions in terms of the original system \( \Sigma_{\text{dis}} \). For any given matrix \( N \in \mathbb{R}^{m \times p} \), consider the system \( \Sigma_{\text{dis},P,Q}^N \) that is obtained by applying to \( \Sigma_{\text{dis},P,Q} \) the static output feedback \( u = N y + v \). This system \( \Sigma_{\text{dis},P,Q}^N \) is described by

\[
\begin{align*}
x_{k+1} &= (A + BNC_1)x_k + Bv_k + (BND_2 + EQ)dk, \\
y_k &= C_1x_k + D_Q dk, \\
z_k &= (C_P + D_P NC_1)x_k + D_P v_k,
\end{align*}
\] (4.6)

Also, for a given compensator \( \Gamma_{\text{dis}} = (K, L, M, N) \), let \( \Gamma_{\text{dis}}^0 = (K, L, M, 0) \) be the compensator with direct feedthrough matrix \( N \) replaced by 0. It is clear that the closed-loop system \( \Sigma_{\text{dis},P,Q}^N \times \Gamma_{\text{dis}} \) has constant transfer matrix \( D_P N D_Q \) if and only if \( \Sigma_{\text{dis},P,Q}^N \times \Gamma_{\text{dis}}^0 \) has transfer matrix equal to \( 0 \). Consequently, an internally stabilizing compensator \( \Gamma_{\text{dis}} = (K, L, M, N) \) is optimal if and only if \( D_P N D_Q = -R^* \) and \( \Sigma_{\text{dis},P,Q}^N \times \Gamma_{\text{dis}}^0 \) has transfer matrix \( 0 \). In other words, in order to find necessary and sufficient conditions for the existence of an optimal controller, we should study the problem of disturbance decoupling with internal stability. This problem has been studied extensively in [16]. One of the main results of [16] gives necessary and sufficient conditions for the existence of an internally stabilizing strictly proper compensator \( \Gamma_{\text{dis}}^0 \) for the system \( \Sigma_{\text{dis}} \) given by (3.1). We will briefly recall this result here. Given \( \Sigma_{\text{dis}} \), let \( V_g \) denote the largest subspace of \( \mathbb{R}^n \) for which there exists \( F \in \mathbb{R}^{m \times n} \) such that \( (A + BF)V_g \subset V_g, |\sigma(A + BF)\mid V_g) < 1, \) and \((C_2 + D_2 F)V_g = 0\). Dually, let \( S_g \) be the smallest subspace of \( \mathbb{R}^n \) for which there exists a matrix \( G \in \mathbb{R}^{n \times p} \) such that \((A + GC_1)S_g \subset S_g, |\sigma(A + GC_1)\mid \mathbb{R}^n/S_g) < 1, \) and \( \text{im} (E + GD_1) \subset S_g \). It was shown in [16, Thm. 2.4] that there exists an internally stabilizing compensator \( \Gamma_{\text{dis}}^0 = (K, L, M, 0) \) such that \( \Sigma_{\text{dis}} \times \Gamma_{\text{dis}}^0 \) has transfer matrix \( 0 \) if and only if the following conditions hold: (i) \((A, B)\) is stabilizable and \((C_1, A)\) is detectable, (ii) the following four subspace inclusions hold: \( \text{im} E \subset V_g, S_g \subset \ker C_2, S_g \subset V_g, \) and \( AS_g \subset V_g \).

Here, we want to apply this result to the system \( \Sigma_{\text{dis},P,Q}^N \) with \( N \) any solution of \( D_P N D_Q = -R^* \). In the following, we will omit some of the details. Using the fact

\[
\text{im} E \subset V_g, S_g \subset \ker C_2, S_g \subset V_g, \text{ and } AS_g \subset V_g.
\]
that \( \text{im} \left( (C_P + D_P NC_1) \right) \subseteq \text{im} \ D_P \), it can be shown that the subspace \( \mathcal{V}_g \) associated with \( \Sigma_{\text{dis},P,Q}^N \) is given by

\[
\mathcal{V}_g = \mathcal{X}_g(A - BD_P^+ C_P) + \langle A - BD_P^+ C_P | B \text{ker} \ D_P \rangle,
\]

where for a given matrix \( M \), \( \mathcal{X}_g(M) \) is the sum of the generalized eigenspaces of \( M \) associated with its eigenvalues in \( |z| < 1 \), and where \( (M | L) \) is the smallest \( M \)-invariant subspace contained in \( L \). It can also be shown, using the fact that \( \text{ker} \ D_Q \subseteq \text{ker} \ (BND_Q + E_Q) \), that

\[
\mathcal{S}_g = \mathcal{X}_b(A - EQD_Q^+ C_1) \cap \langle C_1^{-1} \text{im} \ D_Q | A - EQD_Q^+ C_1 \rangle,
\]

where \( \mathcal{X}_b(M) \) is the sum of the generalized eigenspaces of \( M \) associated with its eigenvalues in \( |z| \geq 1 \) and where \( (L | M) \) is the largest \( M \)-invariant subspace containing \( L \). Using the fact that, from (4.7), \( B \text{ker} \ D_P \subseteq \mathcal{V}_g \), it can be shown that \( \text{im} \ (BND_Q + E_Q) \subseteq \mathcal{V}_g \) if and only if

\[
\text{im} \ (EQ - BD_P^+ R^*) \subseteq \mathcal{V}_g.
\]

Using the fact that, by (4.8), \( \mathcal{S}_g \subseteq C_1^{-1} \text{im} \ D_Q \), it can be shown that \( \mathcal{S}_g \subseteq \text{ker} \ (C_P + D_P NC_1) \) if and only if

\[
\mathcal{S}_g \subseteq \ker \ (C_P - R^* D_Q^+ C_1).
\]

Finally, it can be shown that \( (A + BNC_1)\mathcal{S}_g \subseteq \mathcal{V}_g \) if and only if

\[
(A - BD_P^+ R^* D_Q^+ C_1)\mathcal{S}_g \subseteq \mathcal{V}_g.
\]

Collecting the above facts, we then obtain the following necessary and sufficient conditions for the existence of an optimal controller for the discrete-time \( H_2 \) optimal control problem associated with the system \( \Sigma_{\text{dis}} \):

**Theorem 4.10.** Consider the system (3.1). Assume that \((A, B)\) is stabilizable and \((C_1, A)\) is detectable. Let \( P \) and \( Q \) be the largest real symmetric solution of (3.2) and (3.3), respectively. Let \( \mathcal{V}_g \) and \( \mathcal{S}_g \) be given by (4.7) and (4.8). Then we have: there exists an optimal controller, i.e., an internally stabilizing controller \( \Gamma_{\text{dis}}^* = (K^*, L^*, M^*, N^*) \) such that \( J_{\Sigma_{\text{dis}}} (\Gamma_{\text{dis}}^*) = J^* \), if and only if the four subspace inclusions \( \mathcal{S}_g \subseteq \mathcal{V}_g \), (4.9), (4.10), and (4.11) are satisfied.

5. The sampled-data \( H_2 \) problem. Now we return to the sampled-data \( H_2 \) problem. Consider the continuous-time system \( \Sigma \) given by (2.3), and let \( \Delta \notin \Delta \) be a given sampling period. Let the discrete-time system \( \Sigma_{\Delta} \) be given by (2.4). According to Theorem 2.1, the optimal sampled-data \( H_2 \) performance \( J_{\Sigma_{\Delta}}^* \) is equal to

\[
J_{\Sigma_{\Delta}}^* = \frac{1}{\Delta} \int_0^\Delta \int_0^{\Delta-s} \text{tr} \ (C_2 e^{tA} EE^T e^{tA^T} C_2^T) \ dt \ ds + \frac{1}{\Delta} J_{\Sigma_{\Delta}}^*,
\]

where \( J_{\Sigma_{\Delta}}^* \) is the optimal discrete-time \( H_2 \) performance associated with \( \Sigma_{\Delta} \). According to Theorem 4.8, the optimal performance \( J_{\Sigma_{\Delta}}^* \) can be found in terms of two algebraic Riccati equations associated with \( \Sigma_{\Delta} \). According to Theorem 4.10, an optimal compensator \( \Gamma_{\text{dis},\Delta}^* \) exists if and only if four subspace inclusions involving subspaces associated with the system \( \Sigma_{\Delta} \) are satisfied. According to Theorem 3.3, if the systems \((A_{\Delta}, B_{\Delta}, C_{2,\Delta}, D_{2,\Delta})\) and \((A_{\Delta}, E_{\Delta}, C_1, 0)\) have no zeros on the unit circle, then an
optimal compensator $\Gamma_{\text{dis}, \Delta}$ exists and can be calculated using the “construction” in the statement of Theorem 3.3. The sampled-data controller $\Gamma := H_{\Delta} \Gamma_{\text{dis}, \Delta} S_{\Delta}$ is then optimal for the sampled-data $H_2$ problem under consideration.

In this section we study the following question: what are conditions in terms of the original system $\Sigma$ that guarantee that there exists an optimal compensator for the sampled-data $H_2$ problem? Instead of being completely general, we will study the following question: what are necessary and sufficient conditions in terms of the original system $\Sigma$ such that $(A_{\Delta}, B_{\Delta}, C_{2, \Delta}, D_{2, \Delta})$ and $(A_{\Delta}, E_{\Delta}, C_1, 0)$ have no zeros on the unit circle? In the following, let $R$ be the controllability subspace of the system $(A, B, C_2, D_2)$ (see §2). The main results of this section are the following:

THEOREM 5.1. Consider the system $\Sigma$. Let $\Delta > 0$.

(i) Let $\lambda$ be a zero of $(A_{\Delta}, B_{\Delta}, C_{2, \Delta}, D_{2, \Delta})$, $\lambda \neq 1$. Then there exists an unobservable eigenvalue $\mu$ of $(C_2, A)$ such that $\lambda = e^{\mu \Delta}$.

(ii) If $(A, B, C_2, D_2)$ is left-invertible, then also the converse of (i) holds: if $\mu$ is an unobservable eigenvalue of $(C_2, A)$, then $e^{\mu \Delta}$ is a zero of $(A_{\Delta}, B_{\Delta}, C_{2, \Delta}, D_{2, \Delta})$.

(iii) 1 is a zero of $(A_{\Delta}, B_{\Delta}, C_{2, \Delta}, D_{2, \Delta})$ if and only if at least one of the following two conditions hold:

(a) $0$ is a zero of $(A, B, C_2, D_2)$,
(b) $\frac{\ker C_2}{A}$.

(iv) If $(A, B, C_2, D_2)$ is left-invertible, then 1 is a zero of $(A_{\Delta}, B_{\Delta}, C_{2, \Delta}, D_{2, \Delta})$ if and only if 0 is a zero of $(A, B, C_2, D_2)$.

COROLLARY 5.2. Consider the system $\Sigma$. Let $\Delta > 0$.

(i) If $(C_2, A)$ has no unobservable eigenvalues on the imaginary axis, 0 is not a zero of $(A, B, C_2, D_2)$, and $R \subset \langle \ker C_2 | A \rangle$, then $(A_{\Delta}, B_{\Delta}, C_{2, \Delta}, D_{2, \Delta})$ has no zeros on the unit circle.

(ii) If $(A, B, C_2, D_2)$ is left-invertible, then $(A_{\Delta}, B_{\Delta}, C_{2, \Delta}, D_{2, \Delta})$ has no zeros on the unit circle if and only if $(C_2, A)$ has no unobservable eigenvalues on the imaginary axis and 0 is not a zero of $(A, B, C_2, D_2)$.

THEOREM 5.3. Consider the system $\Sigma$. Let $\Delta > 0$.

(i) Let $\lambda$ be a zero of $(A_{\Delta}, E_{\Delta}, C_1, 0)$. Then there exists an uncontrollable eigenvalue $\mu$ of $(A, E)$ such that $\lambda = e^{\mu \Delta}$.

(ii) If $(A, E, C_1, 0)$ is right-invertible, then also the converse of (i) holds; i.e., if $\mu$ is an uncontrollable eigenvalue of $(A, E)$, then $e^{\mu \Delta}$ is a zero of $(A_{\Delta}, E_{\Delta}, C_1, 0)$.

COROLLARY 5.4. Consider the system $\Sigma$. Let $\Delta > 0$. If $(A, E)$ has no uncontrollable eigenvalues on the imaginary axis, then $(A_{\Delta}, E_{\Delta}, C_1, 0)$ has no zeros on the unit circle. If, in addition, $(A, E, C_1, 0)$ is right-invertible, then also the converse holds: $(A_{\Delta}, E_{\Delta}, C_1, 0)$ has no zeros on the unit circle if and only if $(A, E)$ has no uncontrollable eigenvalues on the imaginary axis.

Note that the conditions on $\Sigma$ obtained in these theorems are independent of the sampling period. In the remainder of this section we shall prove these results.

In order to study the zeros of $(A, B, C_2, D_2)$ and $(A_{\Delta}, B_{\Delta}, C_{2, \Delta}, D_{2, \Delta})$, consider the system matrices of these systems. Let

$$P_{\Delta}(z) := \begin{pmatrix} zI - A_{\Delta} & -B_{\Delta} \\ C_{2, \Delta} & D_{2, \Delta} \end{pmatrix}, \quad P(s) := \begin{pmatrix} sI - A & -B \\ C_2 & D_2 \end{pmatrix}.$$ 

Recall that $\lambda$ is a zero of $(A_{\Delta}, B_{\Delta}, C_{2, \Delta}, D_{2, \Delta})$ if and only if the rank of the complex matrix $P_{\Delta}(\lambda)$ is less than the normal rank of $P_{\Delta}$ (see §2). In order to find out in
which points $\lambda$ this happens, we will study for $\lambda \in \mathcal{C}$ the subspace

$$V_\lambda := \ker P_\Delta(\lambda) \subset \mathcal{C}^{n+m}.$$  

Clearly, for all $\lambda$ we have $\dim V_\lambda = n + m - \text{rank } P_\Delta(\lambda)$. Consequently, for all but finitely many $\lambda$ we have $\dim V_\lambda = d$, where

$$d := n + m - \text{normrank } P_\Delta.$$  

Hence, $\lambda$ is a zero of $(A_{\Delta}, B_{\Delta}, C_{2,\Delta}, D_{2,\Delta})$ if and only if $\dim V_\lambda > d$. In the following lemma we will calculate for each $\lambda$ the subspace $V_\lambda$, its dimension $\dim V_\lambda$, and the number $d$. Denote the unobservable subspace $\ker C_2 \mid A$ by $\mathcal{N}$. Define a subspace $W$ as follows:

(5.3)  

$$W := B^{-1}\mathcal{N} \cap \ker D_2.$$  

**Lemma 5.5.** For every $\lambda \in \mathcal{C}$, $\lambda \neq 1$ we have

(5.4)  

$$V_\lambda = (N \times W) \cap \ker \begin{pmatrix} \lambda I - A_{\Delta} & B_{\Delta} \end{pmatrix},$$  

(5.5)  

$$\dim V_\lambda = \dim \mathcal{N} + \dim W - \dim((\lambda I - A_{\Delta})\mathcal{N} + B_{\Delta}W).$$

For all but finitely many $\lambda$ we have $\dim V_\lambda = d = \dim W$, equivalently, normrank $P_\Delta = n + m - \dim W$. In addition we have

(5.6)  

$$V_1 = \ker \begin{pmatrix} -A & -B \\ C_2 & D_2 \end{pmatrix}.$$  

**Proof.** We will first prove (5.4). We know $(x_0 \ u_0) \in V_\lambda$ if and only if

(5.7)  

$$A_{\Delta} x_0 + B_{\Delta} u_0 = \lambda x_0,$$

(5.8)  

$$C_{2,\Delta} x_0 + D_{2,\Delta} u_0 = 0.$$  

Consider the differential equation $\dot{x}(t) = A_{\Delta} x(t) + B_{\Delta} u_0$, $x(0) = x_0$; and define $z(t) := C_{2} x(t) + D_{2} u_0$. Clearly, $x(\Delta) = A_{\Delta} x_0 + B_{\Delta} u_0$, so (5.7) is equivalent to $x(\Delta) = \lambda x_0$. In turn, this is equivalent to

(5.9)  

$$(\lambda - 1)x_0 = \int_0^\Delta e^{At}(Ax_0 + Bu_0)dt.$$  

Using the definition (2.6) of $C_{2,\Delta}$ and $D_{2,\Delta}$, we see that (5.8) is equivalent to

$$\begin{pmatrix} C_2 & D_2 \end{pmatrix} e^{At} \begin{pmatrix} x_0 \\ u_0 \end{pmatrix} = 0 \text{ for all } t \in [0, \Delta],$$  

which, in turn, is equivalent to $z(t) = 0$ for all $t \in [0, \Delta]$. Obviously,

$$z(t) = C_2 e^{At} x_0 + \left[ C_2 \int_0^t e^{As} Bds + D_2 \right] u_0.$$  

Since $z(t) = 0$ for all $t \in [0, \Delta]$ is satisfied if and only if $z(0) = 0$ and $\dot{z}(t) = 0$ for all $t \in [0, \Delta]$, we find that (5.8) is equivalent to

$$C_2 x_0 + D_2 u_0 = 0 \quad \text{and} \quad C_2 e^{At}(Ax_0 + Bu_0) = 0, \quad t \in [0, \Delta].$$
In other words (5.8) is satisfied if and only if
\[ C_2x_0 + D_2u_0 = 0 \quad \text{and} \quad Ax_0 + Bu_0 \in \mathcal{N}. \]
Now assume that \( \lambda \neq 1 \). Then (5.9) and (5.10) imply that \( x_0 \in \mathcal{N} \subset \ker C_2 \), so \( u_0 \in \ker D_2 \). Also it follows that \( Ax_0 \in \mathcal{N} \), so \( Bu_0 \in \mathcal{N} \) and, in fact, \( u_0 \in \mathcal{W} \). We conclude that, for \( \lambda \neq 1 \), \( \mathcal{V}_\lambda \subset (\mathcal{N} \times \mathcal{W}) \cap \ker \left( \lambda I - A_\Delta \quad B_\Delta \right) \). To prove the converse inclusion, note that \( u_0 \in \mathcal{W} \) implies that \( D_2u_0 = 0 \) and \( B_\Delta u_0 \). If, in addition, \( x_0 \in \mathcal{N} \), then we have \( C_2x_0 + D_2u_0 = 0 \) and \( Ax_0 + Bu_0 \). By the above this is equivalent to (5.8). This completes the proof of (5.4).

To prove (5.5), note that, in general, if \( \mathcal{L} \) is a subspace of some finite-dimensional linear space \( \mathcal{X} \) and if \( T \) is a linear map acting on \( \mathcal{X} \), then we have \( \dim(\mathcal{L} \cap \ker T) = \dim \mathcal{L} - \dim \mathcal{T} \mathcal{L} \). Applying this to the situation at hand, we find that for any \( \lambda \neq 1 \) we have
\[ \dim \mathcal{V}_\lambda = \dim(\mathcal{N} \times \mathcal{W}) - \dim((\lambda I - A_\Delta \quad B_\Delta)(\mathcal{N} \times \mathcal{W})], \]
which immediately yields (5.5).

Next, we will prove the statement on the dimension of \( \mathcal{V}_\lambda \). First note that since \( \mathcal{N} \) is \( A \)-invariant, it is also \( e^{A\mathbf{t}} \)-invariant, for any \( t \). In particular, this implies that \( \mathcal{N} \) is \( A_\Delta \)-invariant and invariant under \( \int_0^\Delta e^{A\mathbf{t}} \mathbf{d}t \). Now assume that \( \lambda \notin \sigma(A_\Delta) \). Then we have \( (\lambda I - A_\Delta)\mathcal{N} = \mathcal{N} \). Also, since \( B\mathcal{W} \subset \mathcal{N} \), we have \( B_\Delta \mathcal{W} \subset \mathcal{N} \). This implies that \( (\lambda I - A_\Delta)\mathcal{N} + B_\Delta \mathcal{W} = \mathcal{N} \). If, in addition, we assume that \( \lambda \neq 1 \), then (5.5) yields \( \dim \mathcal{V}_\lambda = \dim \mathcal{W} \).

Finally, to prove (5.6), recall that (5.7) is equivalent to (5.9). Note that for all \( \Delta > 0 \), \( \int_0^\Delta e^{A\mathbf{t}} \mathbf{d}t \) is a nonsingular matrix (this can be shown using the Jordan form of \( A \)). Thus, for the case that \( \lambda = 1 \) (5.9) is equivalent to \( Ax_0 + Bu_0 = 0 \). Together with the fact that (5.8) is equivalent to (5.10), this proves (5.6).

By applying this lemma, we are now able to prove the statements (i) and (ii) of Theorem 5.1:

**Proof of Theorem 5.1 (i) and (ii).** (i) Assume that \( \lambda \neq 1 \) is a zero of \( (A_\Delta, B_\Delta, C_2, D_2) \). Then we must have \( \dim \mathcal{V}_\lambda > \dim \mathcal{W} \). Using (5.5) this implies
\[ \dim \mathcal{N} > \dim((\lambda I - A_\Delta)\mathcal{N} + B_\Delta \mathcal{W}). \]
As noted in the proof of Lemma 5.5, \( \mathcal{N} \) is \( A_\Delta \)-invariant and \( B_\Delta \mathcal{W} \subset \mathcal{N} \). Consequently,
\[ (\lambda I - A_\Delta)\mathcal{N} + B_\Delta \mathcal{W} \subset \mathcal{N}. \]
Together with the inequality (5.11), this implies that \( (\lambda I - A_\Delta)\mathcal{N} \) is a strict subspace of \( \mathcal{N} \). This implies that the map \( (\lambda I - A_\Delta) \) restricted to \( \mathcal{N} \) is singular. Thus, \( \ker((\lambda I - A_\Delta) \cap \mathcal{N} \neq 0 \). Clearly, this intersection is \( A \)-invariant, so the restriction of \( A \) to this intersection has an eigenvalue, say \( \mu \), with corresponding eigenvector \( p \). This eigenvector satisfies \( A_\Delta p = \lambda p \). Also, since \( Ap = \mu p \), we have \( A_\Delta p = e^{\mu \mathbf{t}} p \), so \( \lambda = e^{\mu \mathbf{t}} \).

Finally, \( p \in \mathcal{N} \subset \ker C_2 \), so \( \mu \) is an unobservable eigenvalue of \((C_2, A)\).

(ii) We claim that if \((A, B, C_2, D_2)\) is left-invertible, then \( \dim \mathcal{W} = 0 \). Indeed, left-invertibility is equivalent to the conditions \( (B_2 \mathbf{d}2) \) is injective and \( \mathcal{V} \cap \mathcal{B} \ker D_2 = 0 \), where \( \mathcal{V} \) denotes the weakly unobservable subspace associated with \((A, B, C_2, D_2)\) (see §2). Assume that \( u_0 \in \mathcal{W} \). Then we have \( D_2u_0 = 0 \) and \( Bu_0 \). Since \( \mathcal{N} \subset \mathcal{V} \), this yields \( Bu_0 = 0 \). Combining this with \( D_2u_0 = 0 \) then leads to \( u_0 = 0 \). This proves our claim. Now let \( \mu \) be an unobservable eigenvalue of \((C_2, A)\). There exists \( x_0 \neq 0 \) such
that \( Ax_0 = \mu x_0 \) and \( C_2 x_0 = 0 \). This yields \( A_\Delta x_0 = \lambda x_0 \), with \( \lambda := e^{\mu \Delta} \). From the
definition of \( C_{2, \Delta} \) it is also easily seen that \( C_{2, \Delta} x_0 = 0 \). Consequently, \( (x_0^T, 0) \in \mathcal{V}_\lambda \), so
dim \( \mathcal{V}_\lambda > 0 = \dim \mathcal{W} \). This implies that \( \lambda \) is a zero of \( (A_{\Delta}, B_{\Delta}, C_{2, \Delta}, D_{2, \Delta}) \).

In order to prove statements (iii) and (iv) of Theorem 5.1, we need the following lemma.

**Lemma 5.6.** Let \( \Delta > 0 \). Then we have

\[
\text{normrank } P_{\Delta} \geq \text{normrank } P,
\]

with equality if and only if \( \mathcal{R} \subset \mathcal{N} \).

**Proof.** For each \( \lambda \not\in \sigma(A) \) define a subspace \( \mathcal{L}_\lambda \) by

\[
\mathcal{L}_\lambda := \left\{ \left( x_0^T, u_0 \right) \mid u_0 \in \mathcal{W}, x_0 = (\lambda I - A)^{-1} B u_0 \right\}.
\]

Clearly, \( \mathcal{L}_\lambda \subset \ker P(\lambda) \) and \( \dim \mathcal{L}_\lambda = \dim \mathcal{W} \). Consequently, for each \( \lambda \not\in \sigma(A) \) we
have \( \dim \mathcal{W} \leq \dim \ker P(\lambda) \). This implies \( \text{normrank } P \leq n + m - \dim \mathcal{W} \). The
inequality (5.12) then follows from Lemma 5.5.

Of course, \( \text{normrank } P_{\Delta} = \text{normrank } P \) if and only if \( \dim \ker P(\lambda) = \dim \mathcal{W} \) for
all but finitely many \( \lambda \), which, in turn, is equivalent to \( \ker P(\lambda) = \mathcal{L}_\lambda \) for all but
finitely many \( \lambda, \lambda \not\in \sigma(A) \). We will prove that the latter statement is equivalent to
\( \mathcal{R} \subset \mathcal{N} \).

Let \( k := \dim \mathcal{R} \), and let \( \lambda_1, \ldots, \lambda_k \) be distinct complex numbers, \( \lambda_i \not\in \sigma(A) \),
such that \( \ker P(\lambda_i) = \mathcal{L}_{\lambda_i} \). There exists \( F \in \mathbb{R}^{m \times n} \) such that \( (A + BF) \mathcal{R} \subset \mathcal{R} \),
\( (C_2 + D_2 F) \mathcal{R} = 0 \), and \( \sigma(A + BF \mid \mathcal{R}) = \{ \lambda_1, \ldots, \lambda_k \} \). Let \( x_1, \ldots, x_k \in \mathcal{R} \) be
corresponding eigenvectors of \( A + BF \mid \mathcal{R} \). Then \( \{x_1, \ldots, x_k\} \) is a basis of \( \mathcal{R} \). We will
prove that \( x_i \in \mathcal{N} \). Indeed, define \( u_i := -Fx_i \). Then \( (x_i^T, u_i) \in \ker P(\lambda_i) = \mathcal{L}_{\lambda_i} \). Since
\( u_i \in \mathcal{W} \), we have \( Bu_i \in \mathcal{N} \), so \( x_i = (\lambda_i I - A)^{-1} B u_i \in \mathcal{N} \) by \( A \)-invariance of \( \mathcal{N} \). We conclude that \( x_i \in \mathcal{N} \), so \( \mathcal{R} \subset \mathcal{N} \).

Conversely, assume that \( \mathcal{R} \subset \mathcal{N} \). It suffices to show that \( \ker P(\lambda) \subset \mathcal{L}_\lambda \) for all
but finitely many \( \lambda \). Let \( \lambda \) be arbitrary, \( \lambda \not\in \sigma(A) \), and \( \lambda \) not a zero of \( (A, B, C_2, D_2) \).
Let \( \left( x_0^T, 0 \right) \in \ker P(\lambda) \). We will prove that \( x_0 \in \mathcal{R} \), so \( x_0 \in \mathcal{N} \). Assume that \( x_0 \neq 0 \). Let \( F \in \mathbb{R}^{m \times n} \) be such that \( Fx_0 = u_0 \). Then we have \( (A + BF)x_0 = \lambda x_0 \) and
\( (C_2 + D_2 F)x_0 = 0 \). This implies \( x_0 \in \mathcal{V} \), the weakly unobservable subspace associated with
the system \( (A, B, C_2, D_2) \). (Indeed, the one-dimensional subspace \( \mathcal{L} \) spanned by the vector \( x_0 \) has the property that \( (A + BF) \mathcal{L} \subset \mathcal{L} \) and \( (C_2 + D_2 F) \mathcal{L} = 0 \) and so
must be contained in \( \mathcal{V} \), the largest subspace for which such \( F \) exists.) By extending
the linear map \( F \) to the whole subspace \( \mathcal{V} \), we obtain that \( (A + BF) \mathcal{V} \subset \mathcal{V} \) and
\( (C_2 + D_2 F) \mathcal{V} = 0 \), so \( \lambda \in \sigma(A + BF \mid \mathcal{V}) \). We have assumed that \( \lambda \) is a zero of
\( (A, B, C_2, D_2) \). This implies \( \lambda \not\in \sigma(A + BF \mid \mathcal{V}) \). But then we must have \( x_0 \in \mathcal{R} \). This implies that \( x_0 \in \mathcal{N} \). Now \( (\lambda I - A)x_0 - Bu_0 = 0 \), so \( Bu_0 \in \mathcal{N} \). This implies that \( u_0 \in \mathcal{W} \). For \( \lambda \not\in \sigma(A) \) this
then yields \( x_0 \in \mathcal{L}_\lambda \). This completes the proof of the lemma.

**Proof of Theorem 5.1 (iii) and (iv).** (iii) We will prove that \( 1 \) is **not** a zero of the system \( (A_{\Delta}, B_{\Delta}, C_{2, \Delta}, D_{2, \Delta}) \) if and only if \( 1 \) is not a zero of \( (A, B, C_2, D_2) \) and
\( \text{normrank } P = \text{normrank } P_{\Delta} \). Clearly, \( 1 \) is not a zero of \( (A_{\Delta}, B_{\Delta}, C_{2, \Delta}, D_{2, \Delta}) \) if and
only if \( \dim \mathcal{V}_1 = n + m - \text{normrank } P_{\Delta} \). By (5.6) we have \( \dim \mathcal{V}_1 = n + m - \text{rank } P(0) \geq
n + m - \text{normrank } P \), with strict inequality if and only if \( 1 \) is a zero of \( (A, B, C_2, D_2) \). Combining these facts proves our claim. The proof of (iii) is then completed by
applying Lemma 5.6.
(iv) If \((A, B, C_2, D_2)\) is left-invertible, then \(\mathcal{R} = 0\). In that case condition (5.2) is never satisfied. □

In order to study the zeros of \((A_\Delta, E_\Delta, C_1, 0)\), consider the system matrix of this system. Let

\[ Q_\Delta(z) := \begin{pmatrix} zI - A_\Delta & -E_\Delta \\ C_1 & 0 \end{pmatrix}. \]

As before, \(\lambda\) is a zero of \((A_\Delta, E_\Delta, C_1, 0)\) if and only if the rank of the complex matrix \(Q_\Delta(\lambda)\) is less than the normal rank of \(Q_\Delta\) (see §2). In order to find out in which points \(\lambda\) this happens, we will study for \(\lambda \in \mathcal{C}\) the subspace

\[ \mathcal{W}_\lambda := (\text{im } Q_\Delta(\lambda))^\perp \subset \mathbb{C}^{n+p}. \]

For all \(\lambda\) we have \(\dim \mathcal{W}_\lambda = n + p - \text{rank } Q_\Delta(\lambda)\). Consequently, for all but finitely many \(\lambda\) we have \(\dim \mathcal{W}_\lambda = d_1\), where

\[ d_1 := n + p - \text{normrank } Q_\Delta. \]

Hence, \(\lambda\) is a zero of \((A_\Delta, E_\Delta, C_1, 0)\) if and only if \(\dim \mathcal{W}_\lambda > d_1\). The following lemma calculates for each \(\lambda\) the subspace \(\mathcal{W}_\lambda\), its dimension \(\dim \mathcal{W}_\lambda\), and the number \(d_1\).

**Lemma 5.7.** Let \(A > 0\). Then we have

\[ \mathcal{W}_\lambda = (M^\perp \times (C^T)^{-1}M^\perp) \cap \ker \left( \begin{pmatrix} \lambda I - A_\Delta^T & C_1^T \end{pmatrix} \right), \]

(5.13) \(\dim \mathcal{W}_\lambda = \dim M^\perp + \dim((C^T)^{-1}M^\perp - \dim((\lambda I - A_\Delta^T)M^\perp + C_1^T(C^T)^{-1}M^\perp). \)

For all but finitely many \(\lambda\) we have \(\dim \mathcal{W}_\lambda = d_1 = \dim((C^T)^{-1}M^\perp, \text{equivalently},\)

\[ \text{normrank } Q_\Delta = n + p - \dim((C^T)^{-1}M^\perp. \]

**Proof.** By definition, \(\begin{pmatrix} x_0 \\ y_0 \end{pmatrix} \in \mathcal{W}_\lambda\) if and only if

\[ (\lambda I - A_\Delta^\lambda)x_0 + C_1^Tx_0 + C_1^Ty_0 = 0 \quad \text{and} \quad x_0^T E_\Delta = 0. \]

Since, by definition, \(\text{im } E_\Delta = \mathcal{M}\), we see that it suffices to show that (5.14) implies \(\begin{pmatrix} x_0 \\ y_0 \end{pmatrix} \in \mathcal{W}_\lambda\). From the fact that \(\mathcal{M}^\perp\) is \(A^T\)-invariant it follows that \(\mathcal{M}^\perp\) is \(A_\Delta^\perp\)-invariant, so \(C_1^T y_0 \in \mathcal{M}^\perp\). The statement (5.13) on the dimension of \(\mathcal{W}_\lambda\) follows in the same way as the corresponding statement in the previous lemma.

Now let \(\lambda\) be any complex number such that \(\lambda \not\in \sigma(A_\Delta^\perp).\) Since \(\mathcal{M}^\perp\) is \(A_\Delta^T\)-invariant, we then have \((\lambda I - A_\Delta^\lambda)M^\perp = M^\perp.\) Also we have \((C^T)^{-1}M^\perp \subset M^\perp\) (no equality!). Thus, for such \(\lambda\) we have \(\dim \mathcal{W}_\lambda = \dim((C^T)^{-1}M^\perp. \)

We are now ready to prove Theorem 5.3.

**Proof of Theorem 5.3.** Let \(\lambda\) be a zero of \((A_\Delta, E_\Delta, C_1, 0)\). Then we have \(\dim \mathcal{W}_\lambda > \dim((C^T)^{-1}M^\perp\). Consequently, by (5.13), \(\dim M^\perp > \dim((\lambda I - A_\Delta^\lambda)M^\perp + C_1^T(C^T)^{-1}M^\perp)\). In particular, this implies that \((\lambda I - A_\Delta^\lambda)\mathcal{M}^\perp\) is a strict subspace of \(\mathcal{M}^\perp\), so \(\ker(\lambda I - A_\Delta^\lambda) \cap \mathcal{M}^\perp \neq 0.\) This subspace is \(A^T\)-invariant, so there exist \(\mu\) and
Assume that $(A, E, C_1, 0)$ is right-invertible. Let

$$Q(s) := \begin{pmatrix} sI - A & -E \\ C_1 & 0 \end{pmatrix}$$

be the system matrix. We have normrank $Q = n + p$. We claim that also normrank $Q_\Delta = n + p$. Indeed, assume that $y_0 \neq 0$ is an element of $(C_1^\top)^{-1}M$. For $\lambda \notin \sigma(A^\top)$, define $x_0 := -(\lambda I - A^\top)^{-1}C_1^\top y_0$. Then $x_0 \in M$ and we have $(x_0^\top \ y_0^\top)Q(\lambda) = (0 \ 0)$. Thus, for all but finitely many $\lambda$ we have rank $Q(\lambda) < n + p$, which is a contradiction. Hence we must have $(C_1^\top)^{-1}M = 0$.

It follows that $\lambda$ is a zero if and only if $W_\lambda \neq 0$. Assume that $\mu$ is an uncontrollable eigenvalue of $(A, E)$. Then there exists $x_0 \neq 0, x_0 \in M$, such that $x_0^\top A = \mu x_0$. Define $\lambda := e^{\mu \Delta}$. Then we have $x_0^\top A^\Delta = 0$ and $x_0^\top (\lambda I - A^\Delta) = 0$. It follows that

$$\begin{pmatrix} x_0 \\ 0 \end{pmatrix} \in W_\lambda,$$

so $\lambda$ is a zero of $(A^\Delta, E^\Delta, C_1, 0)$.

**6. Performance recovery and convergence of optimal performance.** In this section we study the connection between the `ordinary' continuous-time $H_2$ problem and the sampled-data $H_2$ problem. In particular, we are interested in the following questions:

- Suppose that we control the system $\Sigma$ by means of an internally stabilizing continuous-time compensator $\Gamma_{\con}$, yielding continuous-time $H_2$ performance $J_\Sigma(\Gamma_{\con})$. Is it possible to recover this performance asymptotically by using a sampled-data controller with sufficiently small sampling period? More precisely, is it true that for all $\epsilon > 0$ there exists $\Delta > 0$ and an internally stabilizing sampled-data controller $\Gamma$ with sampling-period $\Delta$ such that $|J_\Sigma(\Gamma_{\con}) - J_{\Sigma,\Delta}(\Gamma)| < \epsilon$?

- Does the optimal sampled-data $H_2$ performance converge to the optimal continuous-time $H_2$ performance as the sampling period $\Delta$ decreases to zero? More precisely, suppose that $J_{\Sigma,\con}^*$ is the optimal continuous-time $H_2$ performance associated with the system $\Sigma$ and, as before, denote the optimal sampled-data $H_2$ performance by $J_{\Sigma,\Delta}^*$. Is it true that $\lim_{\Delta \to 0} J_{\Sigma,\Delta}^* = J_{\Sigma,\con}^*$?

The first question above was studied before in [6, Thm. 4] using a different definition of $H_2$ performance and for the $H_\infty$ performance criterion [6, Thm. 5]. In this section we will show that both questions have an affirmative answer.

Let $\Sigma$ be given by (2.2). If the system $\Sigma$ is controlled by a continuous-time compensator $\Gamma_{\con}$ given by the equations

\begin{align}
\dot{w}(t) &= \bar{K}w(t) + \bar{L}y(t), \\
u(t) &= \bar{M}w(t) + \bar{N}y(t),
\end{align}

with $w(t) \in \mathbb{R}^\ell$, then the associated closed-loop system $\Sigma \times \Gamma_{\con}$ is given by

$$\dot{x}_e(t) = A_ex_e(t) + E_ex_y(t),$$

with

$$A_e = \begin{pmatrix} A + B\bar{N}C_1 & B\bar{M} \\ \bar{L}C_1 & \bar{K} \end{pmatrix}, \quad E_e := \begin{pmatrix} E \\ 0 \end{pmatrix}, \quad C_e := \begin{pmatrix} C_2 + D_2\bar{N}C_1 & D_2\bar{M} \end{pmatrix}.$$
If $\Gamma_{\text{con}}$ is internally stabilizing, i.e., $\sigma(A_e) \subset \mathcal{C}^-$, then the $H_2$ performance of the closed-loop system $\Sigma \times \Gamma_{\text{con}}$ is equal to

$$J_{\Sigma}(\Gamma_{\text{con}}) = \text{tr} \left( E_e P_e E_e^T \right),$$

where $P_e$ is the unique solution of the Lyapunov equation

$$A_e^T P_e + P_e A_e + C_e^T C_e = 0. \tag{6.2}$$

On the other hand, if the system $\Sigma$ is controlled by the sampled-data controller $\Gamma = H_\Delta \Gamma_{\text{dis}} S_\Delta$, with $\Gamma_{\text{dis}}$ given by (2.8), then the discrete-time closed-loop system $\Sigma_{\Delta} \times \Gamma_{\text{dis}}$ is given by the equations

$$x_{e,k+1} = A_{e,\Delta} x_{e,k} + E_{e,\Delta} y_k,$$
$$z_k = C_{e,\Delta} x_{e,k},$$

with

$$A_{e,\Delta} = \begin{pmatrix} A_\Delta + B_\Delta N C_1 & B_\Delta M \\ L C_1 & K \end{pmatrix}, \quad E_{e,\Delta} := \begin{pmatrix} E_\Delta \\ 0 \end{pmatrix},$$
$$C_{e,\Delta} := \begin{pmatrix} C_{2,\Delta} + D_{2,\Delta} N C_1 \\ D_{2,\Delta} M \end{pmatrix}. \tag{6.3}$$

If $\Gamma$ is internally stabilizing, equivalently $\sigma(A_e) \subset \mathcal{C}^-$, then the $H_2$ performance of the closed-loop system $\Sigma \times \Gamma$ is given by

$$J_{\Sigma,\Delta}(\Gamma) = \frac{1}{\Delta} \int_0^\Delta \int_0^{\Delta-s} \text{tr} \left( C_2 e^{tA} E E^T e^{tA^T} C_2^T \right) dt \, ds + \frac{1}{\Delta} \text{tr} \left( E_{e,\Delta} P_{e,\Delta} E_{e,\Delta}^T \right), \tag{6.4}$$

where $P_{e,\Delta}$ is the unique solution of the Lyapunov equation

$$A_{e,\Delta}^T P_{e,\Delta} A_{e,\Delta} - P_{e,\Delta} + C_{e,\Delta}^T C_{e,\Delta} = 0. \tag{6.5}$$

The following theorem shows that our first question above indeed has an affirmative answer:

**Theorem 6.1.** Let $\Gamma_{\text{con}}$ be an internally stabilizing continuous-time compensator. For any $\Delta > 0$ define a discrete-time controller $\Gamma_{\text{dis}}$ by $\Gamma_{\text{dis}} := S_\Delta \Gamma_{\text{con}} H_\Delta$, and let $\Gamma_\Delta := H_\Delta \Gamma_{\text{dis}} S_\Delta$ be the corresponding sampled-data controller with sampling period $\Delta$. Then we have that there exists $\Delta_1 > 0$ such that for all $\Delta \not\in \Delta$ with $0 < \Delta < \Delta_1$, $\Gamma_\Delta$ is internally stabilizing. Furthermore,

$$J_{\Sigma,\Delta}(\Gamma_\Delta) \to J_{\Sigma}(\Gamma_{\text{con}}) \quad (\Delta \downarrow 0).$$

**Proof.** It is easily verified that $\Gamma_{\text{dis}} := S_\Delta \Gamma_{\text{con}} H_\Delta$ is described by the equations

$$w_{k+1} = K_\Delta w_k + L_\Delta y_k,$$
$$u_k = M w_k + N y_k,$$

with $K_\Delta := e^{K_\Delta}$ and $L_\Delta := \int_0^\Delta e^{K_t} dt \tilde{L}$. Thus we have

$$A_{e,\Delta} = \begin{pmatrix} A_\Delta + B_\Delta N C_1 & B_\Delta M \\ L C_1 & K_\Delta \end{pmatrix}. \tag{6.6}$$
Note that $A_{e,\Delta} \to I$, the $(n+\ell) \times (n+\ell)$ identity matrix, and that $\frac{1}{\Delta} (A_{e,\Delta} - I) \to A_e (\Delta \downarrow 0)$. Now we will first show that for $\Delta$ sufficiently small we have $|\sigma(A_{e,\Delta})| < 1$. Since $A_e$ is stable, there exists $Q > 0$ such that $A_e^T Q + QA_e < 0$. Now note that
\[
\frac{1}{\Delta}(A_{e,\Delta}^T Q A_{e,\Delta} - Q) = \frac{1}{\Delta}(A_{e,\Delta}^T - I)QA_{e,\Delta} + Q \frac{1}{\Delta}(A_{e,\Delta} - I).
\]
Since the right-hand term converges to $A_e^T Q + QA_e < 0$, for $\Delta$ sufficiently small we have $A_{e,\Delta}^T Q A_{e,\Delta} - Q < 0$. This implies that $A_{e,\Delta}$ is stable.

Next we show the convergence of the $H_2$ performance. For $\Delta$ sufficiently small we have $|\sigma(A_{e,\Delta})| < 1$, so the $H_2$ performance is given by (6.3), with $P_{e,\Delta}$ given by the Lyapunov equation (6.4). We shall prove that $P_{e,\Delta} \to P_e$, the unique solution of (6.2). For any $\Delta$ sufficiently small define a linear map $m_\Delta : \mathbb{R}^{n \times n} \to \mathbb{R}^{n \times n}$ by
\[
m_\Delta(X) := \frac{1}{\Delta} A_{e,\Delta}^T X A_{e,\Delta} - \frac{1}{\Delta} X.
\]
Also define a linear map $m : \mathbb{R}^{n \times n} \to \mathbb{R}^{n \times n}$ by
\[
m(X) := A^T X + X A_e.
\]
Note that $m$ and $m_\Delta$ are all bijections. We can rewrite $m_\Delta$ as
\[
m_\Delta(X) = \frac{1}{\Delta} (A_{e,\Delta}^T - I)XA_{e,\Delta} + X \frac{1}{\Delta} (A_{e,\Delta} - I).
\]
Recall that $A_{e,\Delta} \to I$ and $\frac{1}{\Delta} (A_{e,\Delta} - I) \to A_e$. Thus we see that $m_\Delta \to m (\Delta \downarrow 0)$. Consequently, also $m_\Delta^{-1} \to m^{-1} (\Delta \downarrow 0)$. Obviously, $P_{e,\Delta} = m_\Delta^{-1}(-\frac{1}{\Delta} C_{e,\Delta}^T C_{e,\Delta})$. In addition, it follows from (2.6) that $\frac{1}{\Delta} C_{e,\Delta}^T C_{e,\Delta} \to C_e^T C_e$. This implies that $P_{e,\Delta} \to m^{-1}(C_e^T C_e)$, which, in turn, is equal to $P_e$. By (2.5) we see that $\frac{1}{\Delta} E_{e,\Delta} E_{e,\Delta}^T \to E_e E_e^T$.

Combining these facts we find that
\[
\frac{1}{\Delta} \text{tr} (E_{e,\Delta} E_{e,\Delta}^T P_{e,\Delta}) \to \text{tr} (E_e E_e^T P_e).
\]
Finally, it is immediate that
\[
\frac{1}{\Delta} \int_0^\Delta \int_0^{\Delta - t} \text{tr} \left( C_1 e^{tA} E E^T e^{A^T} C_1^T \right) dt \, ds \to 0, \quad \Delta \downarrow 0,
\]
which completes the proof of the theorem. \qed

Now we turn to the second question posed above. In order to be able to answer this question, it is useful to consider this question first for the linear quadratic problem.

For this, consider the system $\dot{x}(t) = A x(t) + B u(t)$, $z(t) = C_2 x(t) + D_2 u(t)$. Assume that $(A, B)$ is stabilizable. For a given static state feedback control law $u = F x$ and initial state $x_0$, the output function is denoted by $z_{F, x_0}$. The linear quadratic problem is to minimize for each $x_0$ the cost-functional $J(x_0, F) := \int_0^\infty \| z_{F, x_0}(t) \|^2 dt$ over all $F \in \mathbb{R}^{m \times n}$ such that $\sigma(A + BF) \subset C^-$. It is well known (see [9], [18]) that for each $x_0$ the optimal cost
\[
J^*(x_0) := \inf \{ J(x_0, F) \mid F \text{ s.t. } \sigma(A + BF) \subset C^- \} = x_0^T P x_0,
\]
where $P$ is the largest real symmetric solution of the linear matrix inequality
\[
\begin{pmatrix}
A^T P + P A^T + C_1^T C_2 & P B + C_1^T D_2 \\
B^T P + D_1^T C_2 & D_1^T D_2
\end{pmatrix} \geq 0.
\]
We want to compare this "normal" linear quadratic problem with its sampled-data version.

In the following, take a fixed sampling period $\Delta > 0$. The sampled-data version of the linear quadratic problem is to do the minimization over all stabilizing sampled-data static state feedback laws. More precisely, for a given $F \in \mathbb{R}^{m \times n}$ define the sampled-data state feedback control law $u = F_\Delta x$ by $u(t) := Fx(k\Delta)$ ($t \in [k\Delta, (k+1)\Delta)$, $k = 0, 1, 2, \ldots$, or with a slight abuse of notation: $F_\Delta = H_\Delta F S_\Delta$. For a given $F_\Delta$ and initial state $x_0$, denote the output by $z_{F_\Delta,x_0}$. Define the sampled-data cost functional in the obvious way, and denote it by $J(x_0, F_\Delta)$. The control law $F_\Delta$ is called internally stabilizing if for each initial state the controlled state trajectory $x(t)$ converges to $0$ as $t \to \infty$. The sampled-data linear quadratic problem is to minimize for each $x_0$ $J(x_0, F_\Delta)$ over all internally stabilizing control laws $F_\Delta$. Let

$$J^*_\Delta(x_0) := \inf \{ J(x_0, F_\Delta) \mid F_\Delta \text{ is internally stabilizing} \}$$

be the optimal cost. If no internally stabilizing $F_\Delta$ exists, we define $J^*_\Delta(x_0) := \infty$ for all $x_0$. We will briefly explain here how the sampled-data linear quadratic can be resolved. First, note that for any $F_\Delta \in \mathbb{R}^{m \times n}$ we have

$$J(x_0, F_\Delta) = \sum_{k=0}^{\infty} \int_{k\Delta}^{(k+1)\Delta} \| z_{F_\Delta,x_0}(t) \|^2 dt.$$  

Secondly, note that for all $t \in [k\Delta, (k+1)\Delta)$ we have $\dot{x}(t) = Ax(t) + Bu(t)$, $z_{F_\Delta,x_0}(t) = C_2x(t) + D_2u(t)$, with $u(t) = Fx(k\Delta)$. Hence, on the interval $[k\Delta, (k+1)\Delta)$, $x$ and $u$ satisfy

$$\begin{pmatrix} \dot{x} \\ \dot{u} \end{pmatrix} =  \begin{pmatrix} A & B \\ 0 & 0 \end{pmatrix} \begin{pmatrix} x \\ u \end{pmatrix},$$

with $u(k\Delta) = Fx(k\Delta)$. Consequently,

$$\begin{pmatrix} x(t) \\ u(t) \end{pmatrix} = e^{A(t-k\Delta)} \begin{pmatrix} x(k\Delta) \\ Fx(k\Delta) \end{pmatrix},$$

for $t \in [k\Delta, (k+1)\Delta)$, with $A$ defined by (2.7). Using this, it follows immediately from (2.6) that for $t \in [k\Delta, (k+1)\Delta)$ we have $\| z_{F_\Delta,x_0}(t) \|^2 = \| C_2x(k\Delta) + D_2Fz_{F_\Delta,x_0}(k\Delta) \|^2$. Obviously, $x(k\Delta)$ evolves according to $x((k+1)\Delta) = A_\Delta x(k\Delta) + B_\Delta Fx(k\Delta)$. Hence we see that if $F_\Delta = H_\Delta F S_\Delta$, then $J(x_0, F_\Delta) = \sum_{k=0}^{\infty} \| (C_2 + D_2F)x_k \|^2$, with $x_{k+1} = (A_\Delta + B_\Delta F)x_k$. It is also easily seen that $F_\Delta$ is internally stabilizing if and only if $|\sigma(A_\Delta + B_\Delta F)| < 1$. Hence, $J^*_\Delta(x_0) < \infty$ for all $x_0$ if and only if $(A_\Delta, B_\Delta)$ is stabilizable.

Consequently, we can make the following conclusion: the sampled-data linear quadratic problem under consideration is equivalent to the "normal" discrete-time linear quadratic problem of minimizing, for the system $x_{k+1} = A_\Delta x_k + B_\Delta u_k$, the cost functional $J_{\text{dis}}(x_0, F) := \sum_{k=0}^{\infty} \| (C_2 + D_2F)x_k \|^2$ over all $F \in \mathbb{R}^{m \times n}$ such that $|\sigma(A_\Delta + B_\Delta F)| < 1$. The latter problem was discussed in §3, remark (3.11) and §4, remark (4.6). By applying these results to the situation under consideration we can find a characterization of the optimal cost $J^*_\Delta(x_0)$ of the sampled-data linear quadratic problem:

**Lemma 6.2.** Let $\Delta > 0$ be such that $(A_\Delta, B_\Delta)$ is stabilizable. Then for each $x_0$ we have

$$J^*_\Delta(x_0) = x_0^T P_\Delta x_0,$$
where $P_\Delta$ is the largest real symmetric solution of the algebraic Riccati equation

\begin{equation}
A_\Delta^TP_\Delta A_\Delta - P_\Delta + C_2^T_\Delta C_2_\Delta = 0.
\end{equation}

Now we will show that as $\Delta \downarrow 0$ the largest real symmetric solution $P_\Delta$ of (6.6) converges to $P$, the largest real symmetric solution of (6.5). We will prove this by proving that for each $x_0$ we have $J_\Delta^*(x_0) \to J^*(x_0)$. Note that if $(A, B)$ is stabilizable, then for $\Delta > 0$ sufficiently small we have that $(A_\Delta, B_\Delta)$ is stabilizable.

**Lemma 6.3.** Assume that $(A, B)$ is stabilizable. Then there exists $\Delta_1 > 0$ such that for all $0 < \Delta < \Delta_1$, for all $x_0$ we have $J_\Delta^*(x_0) < \infty$. For all $x_0$ we have $\lim_{\Delta \downarrow 0} J_\Delta^*(x_0) = J^*(x_0)$. Also, for all $0 < \Delta < \Delta_1$, $P_\Delta$ exists and we have $\lim_{\Delta \downarrow 0} P_\Delta = P$.

**Proof.** First of all note that for each sampling period $\Delta$ we have $J_\Delta^*(x_0) \geq J^*(x_0)$ for all $x_0$. This can be shown using that, in fact, for each $x_0$,

\[
J^*(x_0) = \inf \left\{ \int_0^\infty \|C_2x(t) + D_2u(t)\|^2 dt \mid u \text{ is such that } \lim_{t \to \infty} x(t) = 0 \right\}.
\]

Hence, by taking $u$ to be generated by the internally stabilizing sampled-data control law $F_\Delta$, it follows that $J(x_0, F_\Delta) \geq J^*(x_0)$.

Now, let $\delta > 0$. Let $F$ be such that $\sigma(A + BF) \subset C^-$ and $J(x_0, F) < J^*(x_0) + \delta$. Clearly, $J(x_0, F) = x_0^TLx_0$, where $L$ is the unique solution of the Lyapunov equation

\[
(A + BF)^TL + L(A + BF) + (C_2 + D_2F)^T(C_2 + D_2F) = 0.
\]

Now consider the sampled-data control law $F_\Delta = H_\Delta FS_\Delta$. By previous arguments, $J(x_0, F_\Delta) = x_0^TLx_0$, where $L_\Delta$ is the unique solution of the Lyapunov equation

\[
(A_\Delta + B_\Delta F)^TL_\Delta(A_\Delta + B_\Delta F) - L_\Delta + (C_2_\Delta + D_2_\Delta F)^T(C_2_\Delta + D_2_\Delta F) = 0.
\]

Note that $A_\Delta + B_\Delta F \to I$, $\frac{1}{\Delta}(A_\Delta + B_\Delta F - I) \to A$, and $\frac{1}{\Delta}(C_2_\Delta + D_2_\Delta F)^T(C_2_\Delta + D_2_\Delta F) \to (C_2 + D_2F)^T(C_2 + D_2F)$ as $\Delta \downarrow 0$. Using a completely similar argument as in the proof of Theorem 6.1 we derive from this that $L_\Delta \to L$, which implies $J(x_0, F_\Delta) \to J(x_0, F)$. Of course, we also have $J^*(x_0) \leq J_\Delta^*(x_0) \leq J(x_0, F_\Delta)$. Combining this with $J(x_0, F) < J^*(x_0) + \frac{\delta}{2}$, we find that for $\delta$ sufficiently small we have $J^*(x_0) \leq J_\Delta^*(x_0) \leq J^*(x_0) + \delta$. Since $\delta$ was arbitrary, this proves the claim. The second statement in the formulation of the theorem is then immediate. \qed

Let $J_{\Sigma, con}^*$ be the optimal continuous-time $H_2$ performance, i.e., the infimum of $J_{\Sigma}^*(F_{\text{con}})$ over all internally stabilizing continuous-time compensators (6.1). It was shown in [15] that if $(A, B)$ is stabilizable and $(C_1, A)$ is detectable, then

\begin{equation}
J_{\Sigma, con}^* = \text{tr} \left( EE^TP + tr \left( (A^TP + PA + C_1^TC_2)Q \right) \right),
\end{equation}

where $P$ is the largest real symmetric solution of the linear matrix inequality (6.5) and $Q$ is the largest real symmetric solution of the dual linear matrix inequality

\begin{equation}
\begin{pmatrix}
AQ + QA^T + EE^T & C_1^TQ \\
C_1Q & 0
\end{pmatrix} \geq 0.
\end{equation}
Let $J_{\Sigma,\Delta}^{*}$ be the optimal sampled-data $H_2$ performance. If $\Delta \in \Delta$, then we define $J_{\Sigma,\Delta}^{*} := +\infty$. Our next theorem gives an affirmative answer to the second question posed in the introduction to this section.

**Theorem 6.4.** Let $(A, B)$ be stabilizable and $(C_1, A)$ be detectable. Then there exists $\Delta_1$ such that for all $0 < \Delta < \Delta_1$, $J_{\Sigma,\Delta}^{*} < \infty$. We have $\lim_{\Delta \to 0} J_{\Sigma,\Delta}^{*} = J_{\Sigma,\text{con}}^{*}$.

In the remainder of this section we will prove this theorem. First, recall the expression (5.1) for $J_{\Sigma,\Delta}^{*}$. Denote the first term in (5.1) by $I(\Delta)$. Then, under the conditions that $(A, B)$ is stabilizable and $(C_1, A)$ is detectable, we know that for $\Delta \notin \Delta$

\[(6.9) \quad J_{\Sigma,\Delta}^{*} = I(\Delta) + \frac{1}{\Delta} \text{tr} (E_\Delta E_\Delta^* P_\Delta) + \frac{1}{\Delta} \text{tr} ((A_\Delta^T P_\Delta A_\Delta - P_\Delta + C_{2,\Delta}^T C_{2,\Delta})Q_\Delta) - \frac{1}{\Delta} \text{tr} ((D_{P_\Delta} N_{\Delta}^* D_{Q_\Delta})(D_{P_\Delta} N_{\Delta}^* D_{Q_\Delta})^T),\]

where $P_\Delta$ is the largest real symmetric solution of (6.6), $Q_\Delta$ is the largest real symmetric solution of the dual Riccati equation

\[(6.10) \quad A_\Delta Q_\Delta A_\Delta^T - Q_\Delta + E_\Delta E_\Delta^T + A_\Delta Q_\Delta C_1^T(C_1 Q_\Delta C_1^T)^T + C_1 Q_\Delta A_\Delta = 0,\]

and

\[N_{\Delta}^* = -D_{P_\Delta}(D_{P_\Delta}^T)^2 D_{P_\Delta} C_{P_\Delta} Q_\Delta C_1^T(D_{Q_\Delta}^T)^2 D_{Q_\Delta}^T.\]

Here, $C_{P_\Delta}$, $D_{P_\Delta}$, and $D_{Q_\Delta}$ are defined by (3.5), (3.4), and (3.7), respectively, with $P = P_\Delta$ and $Q = Q_\Delta$. We will prove that $J_{\Sigma,\Delta}^{*} \to J_{\Sigma,\text{con}}^{*}$ by analyzing the asymptotic behavior of the four terms appearing in (6.9) separately:

- It is immediate that the first term, $I(\Delta)$, converges to 0 as $\Delta \to 0$.
- From (2.5) it follows that $\frac{1}{\Delta} E_\Delta E_\Delta^* \to EE^T$. Since also $P_\Delta \to P$, we conclude that the second term, $\frac{1}{\Delta} \text{tr} (E_\Delta E_\Delta^* P_\Delta)$, converges to $\text{tr} (EE^T P)$.
- To prove convergence of the third term, first note that $Q_\Delta \to Q$. This follows immediately by dualizing Lemma 6.3. Next, as before, rewrite

\[(6.11) \quad \frac{1}{\Delta} \text{tr} (A_\Delta^T P_\Delta A_\Delta - P_\Delta + C_{2,\Delta}^T C_{2,\Delta})Q_\Delta) = \frac{1}{\Delta} (A_\Delta^T - I)P_\Delta A_\Delta + P_\Delta \frac{1}{\Delta} (A_\Delta - I) + \frac{1}{\Delta} C_{2,\Delta}^T C_{2,\Delta}.\]

Since $\frac{1}{\Delta} (A_\Delta - I) \to A$, $A_\Delta \to I$, and $\frac{1}{\Delta} C_{2,\Delta}^T C_{2,\Delta} \to C_2^T C_2$, we conclude that the third term in (6.9) converges to $\text{tr} (A^T P + PA + C_2^T C_2)$.

- In order to complete the proof of Theorem 6.4, we should hence prove that the fourth term in (6.9) converges to 0 as $\Delta \to 0$. This is done in the following lemma:

**Lemma 6.5.** $\frac{1}{\Delta} \text{tr} ((D_{P_\Delta} N_{\Delta}^* D_{Q_\Delta})(D_{P_\Delta} N_{\Delta}^* D_{Q_\Delta})^T) \to 0$ as $\Delta \to 0$.

**Proof.** Rewrite the fourth term in (6.9) as $\frac{1}{\Delta} \|D_{P_\Delta} N_{\Delta}^* D_{Q_\Delta}\|^2$, where for any matrix $M$, $\|M\|$ denotes the Frobenius norm tr $(MM^T)$. Note that if $M$ is a given matrix, then $M^T M$ and $M M^T$ are orthogonal projectors, so consequently $\|M M^T\| = \|M^T M\| = \text{rank} (M)$. In particular, this implies that if $M$ is a $n \times n$ matrix, then $\|M M^T\| = \|M^T M\| \leq n$. Now make the following estimates:

\[\frac{1}{\Delta} \|D_{P_\Delta} N_{\Delta}^* D_{Q_\Delta}\|^2 \]
As noted before, $C_p^T P_\Delta = A_2^T P_\Delta A_2 - P_\Delta + C_1^T Q_\Delta C_1$, so $\frac{1}{\Delta} \|C_{p_\Delta}\|^2 \to \text{tr} (A^T P + PA + C_2^T C_2)$. On the other hand, by noting that $Q_\Delta$ satisfies the Riccati equation (6.10), where $A_\Delta = e^{A_\Delta}$ is invertible, we see that

$$\|Q_\Delta C_1^T D_{Q_\Delta}^+\|^2 = \text{tr} (Q_\Delta C_1^T (C_1 Q_\Delta C_1)^+ C_1 Q_\Delta)$$

$$= \text{tr} (Q_\Delta - A_\Delta^{-1} Q_\Delta A_\Delta^T + A_\Delta^{-1} E_\Delta E_\Delta^T A_\Delta^T).$$

Since $Q_\Delta \to Q$, $A_\Delta^{-1} \to I$ and $E_\Delta E_\Delta^T \to 0$, the latter converges to zero as $\Delta \downarrow 0$. \hfill \Box

REFERENCES