

## SINGULAR VALUE DECAY OF OPERATOR-VALUED DIFFERENTIAL LYAPUNOV AND RICCATI EQUATIONS\*

TONY STILLFJORD†

**Abstract.** We consider operator-valued differential Lyapunov and Riccati equations, where the operators  $B$  and  $C$  may be relatively unbounded with respect to  $A$  (in the standard notation). In this setting, we prove that the singular values of the solutions decay fast under certain conditions. In fact, the decay is exponential in the negative square root if  $A$  generates an analytic semigroup and the range of  $C$  has finite dimension. This extends previous similar results for algebraic equations to the differential case. When the initial condition is zero, we also show that the singular values converge to zero as time goes to zero, with a certain rate that depends on the degree of unboundedness of  $C$ . A fast decay of the singular values corresponds to a low numerical rank, which is a critical feature in large-scale applications. The results reported here provide a theoretical foundation for the observation that, in practice, a low-rank factorization usually exists.

**Key words.** differential Riccati equations, differential Lyapunov equations, operator-valued, infinite dimensional, singular value decay, low rank

**AMS subject classifications.** 47A62, 47A11, 49N10

**DOI.** 10.1137/18M1178815

**1. Introduction.** We consider differential Lyapunov equations (DLEs) and differential Riccati equations (DREs) of the forms

$$(1.1) \quad \dot{P} = A^*P + PA + C^*C, \quad P(0) = G^*G,$$

and

$$(1.2) \quad \dot{P} = A^*P + PA + C^*C - PBB^*P, \quad P(0) = G^*G,$$

respectively. Such equations arise in many different areas, e.g., in optimal/robust control, optimal filtering, spectral factorizations,  $\mathbf{H}_\infty$ -control, differential games, etc. [1, 3, 18, 32].

A typical application for DREs is a linear quadratic regulator (LQR) problem, where one seeks to control the output  $y = Cx$  given the state equation  $\dot{x} = Ax + Bu$  by varying the input  $u$ . In the case of a finite time cost function,

$$J(u) = \int_0^T \|y(t)\|^2 + \|u(t)\|^2 dt + \|Gx(T)\|^2,$$

it is well known that the optimal input function  $u^{\text{opt}}$  is given in state feedback form. In particular,  $u^{\text{opt}}(t) = -B^*P(T-t)x(t)$ , where  $P$  is the solution to the DRE (1.2) [9, 21].

The solution to the DLE, on the other hand, yields the (time-limited) observability Gramian of the corresponding LQR system. It is used in applications such as model order reduction [5, 15] for determining which states  $x$  have negligible effect on the input-output relation  $u \mapsto y$ , and which can therefore safely be discarded from the system [19, 8].

---

\*Received by the editors April 3, 2018; accepted for publication (in revised form) July 31, 2018; published electronically October 9, 2018.

<http://www.siam.org/journals/sicon/56-5/M117881.html>

†Max Planck Institute for Dynamics of Complex Technical Systems, Sandtorstr. 1, DE-39106 Magdeburg, Germany (stillfjord@mpi-magdeburg.mpg.de).

In the continuous case, (1.1) and (1.2) are operator valued. After a spatial discretization they become matrix valued. Approximating their solutions by numerical computations is thus, if done naively, much more expensive than simply approximating, e.g., the corresponding vector-valued equation  $\dot{x} = Ax$ . A standard way to decrease the computational complexity is to utilize structural properties of the solutions. A commonly used such property is that of low numerical rank [23, 20, 38], i.e., a fast (often exponential) decay of the singular values. This allows us to approximate  $P(t) \approx L(t)L(t)^*$ , where  $L(t)$  is of finite rank. In the matrix-valued setting, we would have  $P(t) \in \mathbb{R}^{n \times n}$  and  $L(t) \in \mathbb{R}^{n \times r}$  with  $r \ll n$ .

While there exist results on when such low numerical rank is to be expected for *algebraic* Lyapunov and Riccati equations (i.e., the stationary counterparts of (1.1) and (1.2)) (see, e.g., [2, 34, 4, 6, 31, 7, 16, 29]), the differential case has so far been neglected in the literature.

The aim of this article is to remedy this situation and provide criteria on  $A$ ,  $B$ , and  $C$  that guarantee a certain decay of the singular values  $\{\sigma_k\}_{k=1}^\infty$  of the solutions to (1.1) and (1.2). We consider the operator-valued case, with the standard assumption that  $A$  generates an analytic semigroup. In the LQR setting, this corresponds to the control of abstract parabolic problems (including, for example, heat flows and wave equations with strong damping). We allow relatively unbounded operators  $B$  and  $C$ , which means that we can treat various forms of boundary control and observation. In this setting, we follow the approach suggested in [29] for algebraic equations. There, a decay of the form  $\sigma_k \leq Me^{-\gamma\sqrt{k}}$  was shown, i.e., we cannot expect exponential decay but only exponential in the square root. The main results of the present article demonstrate that this extends to the differential case, under similar assumptions. In the case that  $G = 0$  (and hence  $P(0) = 0$ ), our bounds additionally show that the singular values converge to 0 as  $t \rightarrow 0$  with a rate  $t^{1-2\alpha}$ , where  $\alpha$  is a measure of how unbounded the output operator  $C$  is.

An outline of the article is as follows: In section 2 we specify the abstract framework, state the assumptions on the operators, and recall some resulting properties of the solutions to (1.1) and (1.2). Then in section 3 we use the concept of sinc quadrature to show that certain finite-rank operators approximate the integral  $\int_0^t (Ce^{sA}, Ce^{sA}) ds$  well. Since this is in fact the solution to (1.1) when  $G = 0$ , the main results for DLEs then follow quickly. We generalize these results to DREs in section 4 by factorizing the system using output and input-output mappings. Finally, in section 5, we perform a number of numerical experiments on discretized versions of the equations, which verify the theoretical statements.

**2. Preliminaries.** In the operator-valued case, (1.1) and (1.2) need to be interpreted in an appropriate sense. Here, we mainly follow [21] (see also [10]), and outline the ideas for the DRE (1.2) since all the results carry over to the DLE (1.1) by setting  $B = 0$ . Thus, let  $H$ ,  $Y$ ,  $U$ , and  $Z$  be Hilbert spaces, and let the following operators be given: the (unbounded) state operator  $A : \mathcal{D}(A) \subset H \rightarrow H$ , the input operator  $B : U \rightarrow \mathcal{D}(A^*)'$ , the output operator  $C : \mathcal{D}(A) \rightarrow Y$ , and the final state penalization operator  $G : H \rightarrow Z$ . This corresponds to problems arising from the LQR setting.

By  $A^*$  we mean the adjoint of  $A$  with respect to the inner product on  $H$ , and  $\mathcal{D}(A^*)'$  denotes the dual space of  $\mathcal{D}(A^*)$ , also with respect to the  $H$ -topology. With the proper interpretation (see, e.g., [21]), it is a superset of  $H$ ; in fact, it is the completion of  $H$  in the norm  $\|A^{-1} \cdot\|_H$ . Additionally, for general Hilbert spaces  $X$  and  $Y$  we use the notation  $\mathcal{L}(X, Y)$  to denote the set of linear bounded operators from  $X$  to  $Y$ .

*Remark 2.1.* In order that the notation conforms to the usual evolution equation setting, we have changed the direction of time so that  $P(0) = G^*G$  is the given condition rather than  $P(T) = G^*G$  as in [21]. The only effect of this is to change the signs of all the terms on the right-hand side.

Our main assumption is the following.

*Assumption 2.2.* The operator  $A : \mathcal{D}(A) \subset H \rightarrow H$  is the generator of a strongly continuous analytic semigroup  $e^{tA}$  on  $H$ .

This means that there exists a  $\delta \in (0, \pi/2]$  such that  $z \mapsto e^{zA}$  is analytic on the sector  $\Delta_\delta = \{z \in \mathbb{C} ; z \neq 0, |\arg(z)| < \delta\}$ . Further, there exist constants  $\omega \in \mathbb{R}$  and  $M \geq 0$  such that the fractional powers  $(\omega I - A)^\gamma$  are well-defined, and we have the inequalities  $\|e^{tA}\| \leq Me^{\omega t}$  and  $\|(\omega I - A)^\gamma e^{tA}\| \leq M(1 + t^{-\gamma})e^{\omega t}$ ; see, e.g., [35, section 3.10]. Here,  $\omega < 0$  corresponds to the stable case, but we allow  $\omega > 0$  too. We also note that  $A^*$  is the generator of  $e^{tA^*} = (e^{tA})^*$ .

Further, we allow both  $B$  and  $C$  to be unbounded operators, but not *too* unbounded. In particular, we make the following assumption.

*Assumption 2.3.* The operator  $B : U \rightarrow \mathcal{D}(A^*)'$  is relatively bounded in the sense that there is a  $\beta \in [0, 1)$  such that  $(\omega I - A)^{-\beta}B \in \mathcal{L}(U, H)$ .

*Assumption 2.4.* The operator  $C : \mathcal{D}((\omega I - A)^\alpha) \rightarrow Y$  is relatively bounded in the sense that  $C(\omega I - A)^{-\alpha} \in \mathcal{L}(H, Y)$  for  $0 \leq \alpha < \min(1 - \beta, 1/2)$  with the parameter  $\beta$  from Assumption 2.3.

Finally,  $G$  needs to provide sufficient smoothing to compensate for the roughness of  $B$ .

*Assumption 2.5.* The operator  $G : H \rightarrow Z$  is bounded. If  $\beta \geq 1/2$ , there should also exist a  $\theta \geq \beta - 1/2$  such that  $G(\omega I - A)^\theta : H \rightarrow Z$ .

*Remark 2.6.* In the DLE case, we have  $B = 0$ . Assumption 2.3 is thus always satisfied and there is no extra restriction on  $\alpha$  in Assumption 2.4 except  $\alpha \in [0, 1/2)$ .

*Remark 2.7.* Assumption 2.5 is marginally stronger than the assumption that  $(\omega I - A^*)^\theta G^*G \in \mathcal{L}(H)$ ,  $\theta > 2\beta - 1$ , which is made in [21]. We use Assumption 2.5 for compatibility with results from the Salamon/Weiss/Staffans framework [35], but it can most likely be weakened to the one in [21].

Under Assumptions 2.2 to 2.5, the DRE (1.2) possesses a classical solution  $t \mapsto P(t) \in \mathcal{L}(H)$ ; see, e.g., [21, Theorem 1.2.2.1]. This solution additionally solves the following integral equation for all  $x, y \in H$ , and vice versa:

$$(2.1) \quad \begin{aligned} (P(t)x, y) &= (Ge^{tA}x, Ge^{tA}y) + \int_0^t (Ce^{sA}x, Ce^{sA}y) \, ds \\ &\quad - \int_0^t (B^*P(s)e^{sA}x, B^*P(s)e^{sA}y) \, ds. \end{aligned}$$

Combining Assumptions 2.2 and 2.4 shows that  $Ce^{sA} \in \mathcal{L}(H, Y)$  for  $s > 0$ . This actually holds on every subset of the sector of analyticity  $\Delta_\delta$ , as demonstrated, e.g., in [29]. In particular, for every  $a \in [0, 1)$  there exist positive constants  $M_a$  and  $\omega$  such that  $\|Ce^{zA}\|_{\mathcal{L}(H, Y)} \leq M_a(1 + |z|^{-\alpha})e^{\omega\Re(z)}$  for all  $z \in \Delta_{a\delta}$ . The constants  $M_a$  go to infinity as  $a \rightarrow 1$ , i.e., as we approach the limit of analyticity. However, by simply redefining  $\delta$  as, e.g.,  $\delta/2$  we can always get a uniform estimate. In the following, we

will therefore omit the dependence on  $a$  and write

$$(2.2) \quad \|Ce^{zA}\|_{\mathcal{L}(H,Y)} \leq M(1 + |z|^{-\alpha})e^{\omega\Re(z)}, \quad z \in \Delta_\delta,$$

for two positive constants  $M$  and  $\omega$ . Since  $\alpha < 1/2$ ,  $|z|^{-2\alpha}$  is integrable at 0 and the first integral term in (2.1) is therefore well-defined. That the second integral term is well-defined under Assumptions 2.2 to 2.5 is less straightforward, due to the presence of  $P(s)$  and the fact that  $\beta$  is allowed to take values in  $[1/2, 1)$ . We refer to [21, Chapter 1].

**3. Lyapunov equations.** Let us first consider the Lyapunov case (1.1). Restricting (2.1) by setting  $B = 0$  shows that

$$(3.1) \quad (P(t)x, y) = (Ge^{tA}x, Ge^{tA}y) + \int_0^t (Ce^{sA}x, Ce^{sA}y) ds,$$

which provides a closed-form expression for the solution  $P$ . For  $x, y \in \mathcal{D}(A)$  we denote the integrand by  $F$ ,

$$(3.2) \quad F(z) = (Ce^{zA}x, Ce^{zA}y),$$

and note that in fact  $F : \Delta_\delta \rightarrow \mathbb{C}$ . By (2.2), for all  $x, y \in \mathcal{D}(A)$  we have the bound

$$(3.3) \quad |F(z)| \leq \frac{M^2}{|z|^{2\alpha}} e^{2\omega\Re(z)} \|x\| \|y\|.$$

Our aim is now to approximate the integral  $\int_0^t F(s) ds$  by sinc quadrature, which converges exponentially in the number of quadrature nodes. The basic idea is to map the interval  $(0, t)$  onto the real line, apply the trapezoidal rule, use decay properties of  $F$  at  $\pm\infty$ , and then transform back. The proof uses complex analysis and thus requires us to consider  $(0, t)$  as a subset of a domain in  $\mathbb{C}$  rather than a real interval. In our case, the appropriate mapping is  $\phi_t : \mathbb{C} \rightarrow \mathbb{C}$ ,  $\phi_t(z) = \ln \frac{z}{t-z}$  with inverse  $\psi_t : \mathbb{C} \rightarrow \mathbb{C}$ ,  $\psi_t(w) = \frac{te^w}{e^w + 1}$ . The function  $\phi_t$  maps the eye-shaped domain

$$D_E^d(t) = \left\{ z \in \mathbb{C} ; \left| \arg \left( \frac{z}{t-z} \right) \right| < d \right\},$$

where  $0 < d < \pi/2$ , onto the infinite strip

$$D_S^d(t) = \{w \in \mathbb{C} ; |\Im w| < d\}.$$

Here, of course,  $D_E^d(t) \supset [0, t]$ . See Figure 1 for an illustration of these domains.

The following result is due to Lund and Bowers [25], inspired by [36]. Here, as well as throughout the rest of the paper, we use the letter  $M$  to denote a generic constant that does not depend on  $t$ . It is not necessarily the same  $M$  as in (2.2) and (3.3).

**THEOREM 3.1** (see [25, Theorem 3.8]). *Let  $f$  be an analytic function on  $D_E^d(t)$  that for some  $r \in (0, 1)$  satisfies the condition*

$$(3.4) \quad \int_{\psi_t(u+L)} |f(z)| |dz| = \mathcal{O}(|u|^r), \quad u \rightarrow \pm\infty,$$

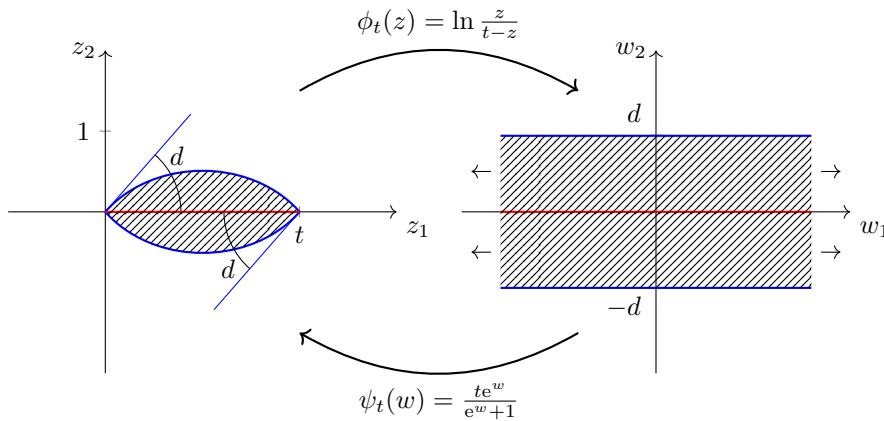


FIG. 1. The transformations  $\phi_t, \psi_t$  and the domains  $D_E^d(t)$  (shaded, left),  $D_S^d(t)$  (shaded, right).

where  $L = \{iv ; |v| \leq d\}$ . Further assume that

$$(3.5) \quad B(f) := \lim_{\gamma \rightarrow \partial D_E^d(t)} \int_{\gamma} |f(z)| |dz| < \infty,$$

where  $\gamma$  denotes any closed simple contour in  $D_E^d(t)$ , and that there are positive constants  $M, \rho$ , and  $\mu$  such that

$$(3.6) \quad \left| \frac{f(z)}{\phi_t'(z)} \right| \leq M \begin{cases} e^{-\rho|\phi_t(z)|} & \forall z \in \psi_t((-\infty, 0)), \\ e^{-\mu|\phi_t(z)|} & \forall z \in \psi_t([0, -\infty)). \end{cases}$$

Choose

$$n = \left\lceil \frac{\rho}{\mu} m + 1 \right\rceil, \quad h = \left( \frac{2\pi d}{\rho m} \right)^{1/2},$$

with  $m$  a nonnegative integer large enough that  $h \leq \frac{2\pi d}{\ln 2}$ , and define the quadrature nodes  $z_k$  and weights  $w_k$  by

$$z_k = \psi_t(kh) = \frac{te^{kh}}{e^{kh} + 1}, \quad w_k = \left( \phi_t'(z_k) \right)^{-1} = \frac{te^{kh}}{(e^{kh} + 1)^2}.$$

Then it holds that

$$\left| \int_0^t f(z) dz - h \sum_{k=-m}^n w_k f(z_k) \right| \leq \left( \frac{M}{\rho} + \frac{M}{\mu} + 2B(f) \right) e^{-(2\pi \rho d m)^{1/2}}.$$

Specifying this theorem to the function  $F$  given in (3.2) leads to the following.

**THEOREM 3.2.** *Let Assumptions 2.2 and 2.4 be satisfied, and let  $h, n, z_k$ , and  $w_k$  be chosen as in Theorem 3.1 with  $d = \delta$ . Then there is a positive constant  $M$ , independent of  $t, x$ , and  $y$ , but dependent on  $\alpha$ , such that*

$$\left| \int_0^t F(z) dz - h \sum_{k=-m}^n w_k F(z_k) \right| \leq M t^{1-2\alpha} e^{-(2\pi(1-2\alpha)\delta m)^{1/2}} \|x\| \|y\|.$$

*Proof.* We verify the conditions of Theorem 3.1. Since the domain  $D_E^\delta(t)$  is a subset of the cone  $\{w \in \mathbb{C} ; |\arg w| \leq \delta\}$  for any  $t > 0$ , the function  $F$  is clearly analytic on  $D_E^\delta(t)$ . Suppose that  $z = \psi_t(u + iv)$ , where  $|v| \leq \delta$ . Then

$$\left| \frac{dz}{dv} \right| = \frac{te^u}{|e^ue^{iv} + 1|^2} \leq t \min(e^u, e^{-u}) \leq t,$$

since  $\delta < \pi/2$  means that  $|e^ue^{iv} + 1| \geq \max(1, e^u)$ . Hence

$$\begin{aligned} \int_{\psi_t(u+L)} |F(z)| |dz| &\leq \int_{-\delta}^\delta \left| F\left(\frac{te^ue^{iv}}{e^ue^{iv} + 1}\right) \right| t \, dv \\ &\leq Mt \int_{-\delta}^\delta \left| \frac{te^ue^{iv}}{e^ue^{iv} + 1} \right|^{-2\alpha} \, dv \\ &\leq 2M\pi t^{1-2\alpha}, \end{aligned}$$

where we have used (3.3) as well as the estimate  $e^{2\omega\Re(z)} \leq \max(1, e^{2\omega T}) \leq M$  in the second step and the inequality  $|e^ue^{iv} + 1| \leq e^u + 1 \leq 2e^u$  in the third step. As this bound is independent of  $u$  and  $1 - 2\alpha > 0$  due to Assumption 2.4, the first condition (3.4) is satisfied.

To check the second condition, we make a change of variables  $w = \eta(z) = \frac{z}{t-z}$ . It is easily seen that  $\eta$  maps the boundary of  $D_E^\delta(t)$  onto the rays  $\{re^{\pm i\delta} ; r \geq 0\}$ , that the inverse is given by  $z = \eta^{-1}(w) = \frac{tw}{1+w}$ , and that the derivative of the inverse is given by  $w \mapsto \frac{t}{(1+w)^2}$ . Denoting the top and bottom parts of  $\partial D_E^\delta(t)$  by  $\partial D_+$  and  $\partial D_-$ , respectively, we thus have  $B(F) = \int_{\partial D_+} |F(z)| |dz| + \int_{\partial D_-} |F(z)| |dz|$ , where

$$\begin{aligned} \int_{\partial D_\pm} |F(z)| |dz| &= \int_0^\infty \left| F\left(\frac{tre^{\pm i\delta}}{1 + re^{\pm i\delta}}\right) \right| t |1 + re^{\pm i\delta}|^{-2} \, dr \\ &\leq M \int_0^\infty \left| \frac{tre^{\pm i\delta}}{1 + re^{\pm i\delta}} \right|^{-2\alpha} t |1 + re^{\pm i\delta}|^{-2} \, dr, \end{aligned}$$

again using (3.3) and bounding the exponential term by  $\max(1, e^{2\omega T})$ . As  $|1 + re^{\pm i\delta}| \geq \max(1, r)$  we get

$$\int_{\partial D_\pm} |F(z)| |dz| \leq t^{1-2\alpha} \left( \int_0^1 r^{-2\alpha} \, dr + \int_1^\infty r^{-2} \, dr \right),$$

so that, in conclusion,

$$B(F) \leq 2t^{1-2\alpha} \left( \frac{1}{1 - 2\alpha} + 1 \right).$$

Finally, we check condition (3.6). A simple computation shows that  $\phi'_t(z) = \frac{t}{z(t-z)}$ . Clearly,  $\psi_t((-\infty, 0)) = (0, t/2) =: \Gamma_1$  and  $\psi_t([0, \infty)) = [t/2, t) =: \Gamma_2$ , which means that on these intervals we have

$$e^{-\rho|\phi_t(z)|} = z^\rho(t-z)^{-\rho} \quad \text{and} \quad e^{-\mu|\phi_t(z)|} = z^{-\mu}(t-z)^\mu.$$

On  $\Gamma_1$ ,  $|t - z| \leq t$ , so by (3.3) we get

$$\begin{aligned} \left| \frac{F(z)}{\phi_t'(z)} \right| &\leq M|z|^{-2\alpha} e^{2\omega\Re(z)} |z| |t - z| t^{-1} \leq M|z|^{1-2\alpha} t^{-1} |t - z|^{2\alpha-1} |t - z|^{2-2\alpha} \\ &\leq M t^{1-2\alpha} |z|^{1-2\alpha} |t - z|^{2\alpha-1}, \end{aligned}$$

i.e., the desired bound holds with  $\rho = 1 - 2\alpha$  and constant  $Mt^{1-2\alpha}$ , where  $M$  is independent of  $t$ . On  $\Gamma_2$ ,  $|z| \leq t$ , and we similarly get

$$\begin{aligned} \left| \frac{F(z)}{\phi_t'(z)} \right| &\leq M|z|^{1-2\alpha} |t - z| t^{-1} \leq M|z|^{-1} |t - z| |z|^{2-2\alpha} t^{-1} \\ &\leq M t^{1-2\alpha} |z|^{-1} |t - z|, \end{aligned}$$

i.e., the desired bound holds with  $\mu = 1$  and constant  $Mt^{1-2\alpha}$ , where  $M$  is again independent of  $t$ . □

We denote the singular values of  $P$  by  $\sigma_k(P)$  and order them in decreasing order. Let us first consider the case when  $G = 0$ .

**THEOREM 3.3.** *Let Assumptions 2.2 and 2.4 be satisfied, with the output space  $Y$  having finite dimension  $\dim Y \geq 1$ . Further assume that  $G = 0$ . Then the singular values of the solution  $P$  to the DLE (3.1) satisfy*

$$\sigma_k(P(t)) \leq M t^{1-2\alpha} e^{-\eta\sqrt{k-2\dim Y}}$$

for  $k \geq 4 \dim Y$ , where  $M$  and  $\eta$  are positive constants independent of  $t$  but dependent on  $\alpha$ .

After our preliminary work, the proof follows almost exactly as in [29].

*Proof.* We have

$$(P(t)x, y) = \int_0^t F(z) \, dz.$$

Now define  $n$ ,  $z_k$ , and  $w_k$  as in Theorem 3.2 and define the approximation  $P_m$  by

$$P_m = h \sum_{k=-m}^n w_k e^{z_k A^*} C^* C e^{z_k A}.$$

Since  $P(t)$  and  $P_m(t)$  are both self-adjoint operators and  $\mathcal{D}(A)$  is dense in  $H$ , by Theorem 3.2 we then get

$$\begin{aligned} \|P(t) - P_m(t)\| &= \sup_{\substack{z \in \mathcal{D}(A) \\ \|z\|=1}} |((P(t) - P_m(t))z, z)| \\ &\leq M t^{1-2\alpha} e^{-(2\pi(1-2\alpha)dm)^{1/2}}. \end{aligned}$$

Now let

$$k_m = (2m + 2) \dim Y.$$

Since  $n \leq m + 1$ , the rank of  $P_m(t)$  is at most  $k_m$ , and we immediately see that we have the bound  $\sigma_{k_m+1}(P(t)) \leq M t^{1-2\alpha} e^{-\eta\sqrt{m}}$  with  $\eta = \sqrt{(2\pi(1-2\alpha)d)}$ . As the

singular values are decreasing, we may rewrite this<sup>1</sup> as

$$\sigma_j \leq \tilde{M} t^{1-2\alpha} e^{-\tilde{\eta} \sqrt{j-2 \dim Y}}$$

for  $j \geq 4 \dim Y$ , with the modified constants  $\tilde{M} = M e^{-\eta(\sqrt{2+1/(2 \dim Y)}-1)}$  and  $\tilde{\eta} = \frac{\eta}{\sqrt{2 \dim Y}}$ .  $\square$

*Remark 3.4.* The theorem is stated for  $k \geq 4 \dim Y$  since this is the maximal rank of the approximant  $P_1(t)$ , which provides the first explicit information we have. As the singular values are decreasing, it is of course possible to scale the constant  $M$  by  $\sigma_1/\sigma_{(4 \dim Y)}$  and show exponential square-root decay for  $k \geq 1$ . However, the bound is then also that much worse in the given interval.

*Remark 3.5.* In the current approach, the factor  $t^{1-2\alpha}$  is desired when  $t$  is small, but this also means that the bound deteriorates when  $t \rightarrow \infty$ . This holds also in the exponentially stable case, i.e., when  $\omega < 0$ , because we cannot bound  $e^{2\omega \Re(z)}$  uniformly on  $(0, t/2)$  by  $e^{-Mt}$  for any positive  $M$ . However, when  $\omega < 0$  the solution to the DLE tends to the solution of the corresponding algebraic Lyapunov equation (ALE)  $0 = A^*P + PA + C^*C$  as  $t \rightarrow \infty$ ; see, e.g., [21, section 2.3] (also for the more general Riccati case). If  $\omega < 0$  and  $t \in [0, T]$ , where  $T$  is very large the bound in Theorem 3.3 is therefore overly pessimistic, and we might instead start from the ALE decay results and consider the small perturbation arising from the difference between the ALE and DLE solutions. The ALE case was considered in [29], which uses the sinc quadrature theory for the infinite interval  $(0, \infty)$  [25, Theorem 3.9] applied to our function  $F(z)$ . (See also [37, Example 4.2.10]). The new integration interval leads to a different choice of transformation  $\phi_t$ , for which it is straightforward to gainfully utilize the  $e^{2\omega \Re(z)}$  term. It results in the exponential square-root decay

$$\left| \int_0^\infty F(z) dz - h \sum_{k=-m}^n F(e^{kh}) e^{kh} \right| \leq M e^{-\sqrt{2\pi\delta\alpha m}}.$$

By (3.3) we have

$$\left| \int_0^T F(z) dz - \int_0^\infty F(z) dz \right| \leq \frac{MT^{-2\alpha} e^{-2\omega T}}{2\omega},$$

and we thus get exponential square-root decay except for a small constant term, if  $T$  is large. We note, however, that if  $T$  is large it might be more worthwhile to consider the ALE with  $T = \infty$  directly, rather than the DLE.

*Remark 3.6.* Similar results are expected to hold in the nonautonomous case, i.e., when  $A$ ,  $B$ , and  $C$  may depend on  $t$ . If the operators  $A(t)$  all generate analytic semigroups with the same domain  $\mathcal{D}(A(t)) = D$  and the map  $t \mapsto A(t) : [0, T] \rightarrow \mathcal{L}(D, H)$  is sufficiently nice (Hölder continuous with  $D$  having the graph norm) then there is an evolution system  $U(t, s)$  satisfying  $\frac{d}{dt}U(t, s) = A(t)U(t, s)$  and  $\|(\omega I - A(t))^\alpha U(t, s)x\| \leq \frac{M}{(t-s)^\alpha}$  for  $x \in D$ . See, e.g., [30, section 5.6]. It can then be

<sup>1</sup>Let  $k = a + bm$  with  $b > 0$ . For  $j = k + 1, \dots, k + 1 + b$  we have  $\sigma_j \leq \sigma_{k+1} \leq M e^{-\eta\sqrt{m}} \leq M e^{-\tilde{\eta}\sqrt{k-a}} \leq M e^{-\tilde{\eta}\sqrt{j-a}} e^{-\tilde{\eta}(\sqrt{k-a}-\sqrt{j-a})}$  with  $\tilde{\eta} = \eta/\sqrt{b}$ . Now,  $\sqrt{k-a} - \sqrt{j-a} \geq \sqrt{k-a} - \sqrt{k+1+b-a} = \sqrt{bm} - \sqrt{b(m+1)+1}$ . The latter function is decreasing with  $m$ , so we get  $\sigma_j \leq M e^{\eta(\sqrt{2+1/b-1})} e^{-\tilde{\eta}\sqrt{j-a}}$ .



verified by differentiation that the function

$$P(t) = \int_0^t U(t, s)^* C(s)^* C(s) U(t, s) \, ds$$

solves the DLE  $\dot{P}(t) = A(t)^* P(t) + P(t) A(t) + C(t)^* C(t)$ ,  $P(0) = 0$ . We can thus follow the same program as in the autonomous case if we can guarantee that  $C(s)(\omega I - A(t))^{-\alpha} \in \mathcal{L}(H, Y)$  with  $\alpha < 1/2$  for  $s$  near  $t$ , since then  $\|C(s)U(t, s)x\| \leq \frac{M}{(t-s)^\alpha}$ . A simple example of when such a condition would hold is when the time dependency is of the form  $A(t) = \kappa(t)\tilde{A}$ ,  $C(t) = \lambda(t)\tilde{C}$ , where  $\tilde{A}$  and  $\tilde{C}$  are fixed operators and the functions  $\kappa, \lambda$  are continuous and bounded away from zero. Then it is clear that if  $\tilde{A}$  and  $\tilde{C}$  satisfy the assumptions for the autonomous case, the above condition is also fulfilled.

A nonzero operator  $G$  makes the situation more delicate. If  $G$  is a finite-rank operator, then the above result is essentially just shifted by  $\text{rank}(G)$ . For consistency, we formulate this in terms of the output space  $Z$ .

**THEOREM 3.7.** *Let Assumptions 2.2, 2.4, and 2.5 be satisfied, with the output spaces  $Y$  and  $Z$  both having finite nonzero dimension. Then the singular values of the solution  $P$  to the DLE (3.1) satisfy*

$$\sigma_k(P(t)) \leq Mt^{1-2\alpha} e^{-\eta\sqrt{k-2\dim Y-\dim Z}}$$

for  $k \geq \max(1, 4\dim Y + \dim Z)$ , where  $M$  and  $\eta$  are positive constants independent of  $t$  but dependent on  $\alpha$ .

*Proof.* This follows by the same procedure as in the proof of Theorem 3.3 after changing the definition of  $P_m$  to

$$P_m = e^{tA^*} G^* G e^{tA} + h \sum_{k=-m}^n w_k e^{z_k A^*} C^* C e^{z_k A}.$$

In this case,  $k_m = \dim Z + (2m + 2)\dim Y$ . □

As an alternative proof, we may make use of the well-known Weyl’s inequality (also known as the Ky Fan inequality): Let  $F_1$  and  $F_2$  be two compact operators on  $H$  with singular values  $\{\sigma_k^1\}_{k=1}^\infty$  and  $\{\sigma_k^2\}_{k=1}^\infty$ , respectively. Denote the singular values of  $F_1 + F_2$  by  $\{\sigma_k\}_{k=1}^\infty$ . Then  $\sigma_{j+k-1} \leq \sigma_j^1 + \sigma_k^2$  for all positive integers  $j$  and  $k$  [14]. If  $\dim Z < \infty$ , then  $G$  and  $e^{tA^*} G^* G e^{tA}$  are both compact operators whose singular values are zero except for the first  $\dim Z$  ones. The operator  $\int_0^t e^{sA^*} C^* C e^{sA} \, ds$  is also compact, since it is the limit of a sequence of finite-rank operators (see the first part of the proof for Theorem 3.3). Hence Weyl’s inequality applies, which shifts the start of the exponential decay by  $\dim Z$ .

Finally, we consider the case where  $G$  is a general operator. To handle the term  $e^{tA^*} G^* G e^{tA}$  we then have to impose stricter requirements on the semigroup  $e^{tA}$  and, by extension, its generator  $A$ . Alternatively, we may require that the singular values of  $G$  decay sufficiently fast.

**THEOREM 3.8.** *Let Assumptions 2.2, 2.4, and 2.5 be satisfied, with the output space  $Y$  having finite dimension  $\dim Y \geq 1$  and  $\dim Z = \infty$ . If the singular values of the solution operator  $e^{tA}$  decay exponentially in the square root,  $\sigma_k(e^{tA}) \leq \tilde{M}e^{-\tilde{\eta}(t)\sqrt{k}}$ ,*

then the singular values of the solution  $P$  to the DLE (3.1) satisfy

$$\sigma_k(P(t)) \leq M \max(1, t^{1-2\alpha}) e^{-\frac{1}{2} \min(\eta, 2\bar{\eta}(t)) \sqrt{k+1-2 \dim Y}}$$

for  $k \geq 6 \dim Y - 1$ , where  $M$  and  $\eta$  are positive constants independent of  $t$  but dependent on  $\alpha$ . If, instead,  $\sigma_k(G) \leq \tilde{M} e^{-\bar{\eta} \sqrt{k}}$ , then the same bound holds but without the time dependence in the exponent.

*Proof.* The extra assumption on  $e^{tA}$  in particular implies that  $e^{tA}$  is compact, and since  $G$  is bounded, the operator  $e^{tA^*} G^* G e^{tA}$  is also compact. Further, the singular values clearly satisfy  $\sigma_k(e^{tA^*} G^* G e^{tA}) \leq \hat{M} e^{-2\bar{\eta} \sqrt{k}}$  for some constant  $\hat{M}$ . We may therefore apply Weyl’s inequality as in the paragraph after the proof of Theorem 3.7. By Theorem 3.3 this directly yields

$$\begin{aligned} \sigma_{2k-2 \dim Y-1}(P(t)) &= \sigma_{k+(k-2 \dim Y)-1}(P(t)) \\ &\leq M t^{1-2\alpha} e^{-\eta \sqrt{k-2 \dim Y}} + \hat{M} e^{-2\bar{\eta}(t) \sqrt{k-2 \dim Y}} \\ &\leq 2 \max \left( M t^{1-2\alpha}, \hat{M} \right) e^{-\min(\eta, 2\bar{\eta}(t)) \sqrt{k-2 \dim Y}}, \end{aligned}$$

and thus

$$\sigma_j(P(t)) \leq 2 \max \left( M t^{1-2\alpha}, \hat{M} \right) e^{-\frac{1}{2} \min(\eta, 2\bar{\eta}(t)) \sqrt{j+1-2 \dim Y}}$$

for all  $j \geq 6 \dim Y - 1$ . For the second case, we note that the assumption implies that

$$\sigma_k \left( e^{tA^*} G^* G e^{tA} \right) \leq \hat{M} e^{-2\bar{\eta} \sqrt{k}}$$

with a different constant  $\hat{M}$ , due to the exponential boundedness of  $e^{tA}$ . We may thus apply Weyl’s inequality in exactly the same way.  $\square$

*Remark 3.9.* When  $A$  is diagonalizable, the assumption on  $e^{tA}$  obviously means that the eigenvalues of  $A$  should go to  $-\infty$  like the negative square root. This assumption is satisfied in many concrete applications. As an example, the Laplacian on  $\Omega \subset \mathbb{R}^d$  with Dirichlet or Neumann boundary conditions has eigenvalues  $\lambda_k(A)$  that decrease as  $\lambda_k(A) = \mathcal{O}(-k^{2/d})$  by Weyl’s law; see, e.g., [12, Chapter VI]. Hence the assumption is satisfied for such problems of up to dimension 4.

**4. Riccati equations.** As in [29], we may extend the Lyapunov results to the Riccati case by using a factorization into output and input-output maps. For this, we will employ the framework of well-posed systems advocated by Salamon [33] and Staffans [35]; see also [27, 40]. As in section 3 we first consider the case of a zero initial condition, then extend this to the finite-rank case and finally to the case of a general  $G$  but with extra requirements on  $A$ .

**THEOREM 4.1.** *Let Assumptions 2.2 to 2.5 be satisfied, with the output spaces  $Y$  and  $Z$  having finite nonzero dimension. Then if  $G = 0$ , the singular values of the solution  $P$  to the DRE (2.1) satisfy*

$$\sigma_k(P(t)) \leq M t^{1-2\alpha} e^{-\eta \sqrt{k-2 \dim Y}}$$

for  $k \geq 4 \dim Y$ . If  $G \neq 0$  but  $\dim Z < \infty$  we instead get

$$\sigma_k(P(t)) \leq M t^{1-2\alpha} e^{-\eta \sqrt{k-2 \dim Y - \dim Z}}$$

for  $k \geq 4 \dim Y + \dim Z$ . If  $\dim Z = \infty$  and  $\sigma_k(e^{tA}) \leq \tilde{M}e^{-\tilde{\eta}(t)\sqrt{k}}$ , then

$$\sigma_{k(P(t))} \leq M \max(1, t^{1-2\alpha}) e^{-\frac{1}{2} \min(\eta, 2\tilde{\eta}(t))\sqrt{k+1-2 \dim Y}}$$

for  $k \geq 6 \dim Y - 1$ . Finally, if  $\dim Z = \infty$  and  $\sigma_k(G) \leq \tilde{M}e^{-\tilde{\eta}\sqrt{k}}$ , then the last bound still holds, but without the time dependency in the exponent. In all the cases above,  $M$  and  $\eta$  are positive constants independent of  $t$  but dependent on  $\alpha$ .

*Remark 4.2.* As in Remark 3.4, we can shift the decay to start at  $k = 1$  by increasing the constant  $M$ , at the expense of a worse bound in the interval given above.

*Proof.* Let the output and input-output mappings  $\mathcal{C}_t$  and  $\mathcal{D}_t$  be given by

$$(\mathcal{C}_t x_0)(s) = Ce^{sA} x_0 \quad \text{and} \quad (\mathcal{D}_t u)(s) = \int_0^s Ce^{(s-\tau)A} Bu(\tau) d\tau.$$

By [35, Theorem 5.7.3], these mappings satisfy  $\mathcal{C}_t \in \mathcal{L}(H, L^2([0, t], Y))$  and  $\mathcal{D}_t \in \mathcal{L}(L^2([0, t], U), L^2([0, t], Y))$ , due to Assumptions 2.3 and 2.4. When  $G = 0$  we can then directly apply the result of Salamon [33, Theorem 5.1], which (in our notation) states that

$$P(t) = \mathcal{C}_t^* (\mathcal{I} + \mathcal{D}_t \mathcal{D}_t^*)^{-1} \mathcal{C}_t.$$

Here,  $\mathcal{I}$  denotes the identity operator on  $L^2([0, t], Y)$ , and the inverse of  $\mathcal{I} + \mathcal{D}_t \mathcal{D}_t^*$  exists as a bounded self-adjoint operator by the Lax–Milgram lemma. A straightforward calculation shows that  $\mathcal{C}_t^*$  is given by  $\mathcal{C}_t^* u = \int_0^t e^{sA^*} C^* u(s) ds$ , and we get

$$\mathcal{C}_t^* \mathcal{C}_t x_0 = \int_0^t e^{sA^*} C^* Ce^{sA} x_0 ds.$$

Thus, in fact, for  $x, y \in \mathcal{D}(A)$  we have  $(\mathcal{C}_t^* \mathcal{C}_t x, y) = F(t)$  with  $F$  defined by (3.2). Hence the singular values of  $\mathcal{C}_t^* \mathcal{C}_t$  decay exponentially in the square root, by exactly the same reasoning as in the proof of Theorem 3.3. Multiplying  $\mathcal{C}_t^* \mathcal{C}_t$  by the bounded operator  $(\mathcal{I} + \mathcal{D} \mathcal{D}^*)^{-1}$  only scales the singular values by the factor  $\|(\mathcal{I} + \mathcal{D} \mathcal{D}^*)^{-1}\|$ , so we have thus proven the first assertion.

The argument in [33, Theorem 5.1] may be extended also to the more general case that  $G \neq 0$ . We instead get

$$P(t) = \mathcal{C}_{G,t}^* \mathcal{C}_{G,t} + \mathcal{C}_t^* \mathcal{C}_t - (\mathcal{C}_{G,t}^* \mathcal{D}_{G,t} + \mathcal{C}_t^* \mathcal{D}_t) (\mathcal{I} + \mathcal{D}_t^* \mathcal{D}_t + \mathcal{D}_{G,t}^* \mathcal{D}_{G,t})^{-1} (\mathcal{D}_{G,t}^* \mathcal{C}_{G,t} + \mathcal{D}_t^* \mathcal{C}_t),$$

where

$$\mathcal{C}_{G,t} x_0 = Ge^{tA} x_0 \quad \text{and} \quad \mathcal{D}_{G,t} u = G \lim_{s \rightarrow t} \int_0^s e^{(s-\tau)A} Bu(\tau) d\tau$$

are the “final-state” versions of the  $\mathcal{C}$  and  $\mathcal{D}$  operators. By [35, Theorem 5.7.3] and [35, Theorem A.3.7(ii)], the input-output operator  $t \mapsto \int_0^t Ge^{(t-s)A} Bu(s) ds$  maps  $u \in L^2([0, t], U)$  into  $C([0, t], Z)$  under Assumption 2.5, and  $\mathcal{D}_{G,t}$  is therefore well-defined.

Recall that the problem is stated on  $t \in [0, T]$ . For any such  $t$ , we define the product space  $X_t = L^2([0, t], Y) \times Z$  with the induced topology

$$\left\| \begin{bmatrix} y \\ z \end{bmatrix} \right\|_{X_t}^2 = \|y\|_{L^2([0,t],Y)}^2 + \|z\|_Z^2.$$

Further let the operators  $\tilde{C}_t : H \rightarrow X_t$  and  $\tilde{D}_t : L^2([0, t], U) \rightarrow X_t$  be defined by

$$\tilde{C}_t = \begin{bmatrix} C_t \\ \mathcal{C}_{G,t} \end{bmatrix} \quad \text{and} \quad \tilde{D}_t = \begin{bmatrix} \mathcal{D}_t \\ \mathcal{D}_{G,t} \end{bmatrix}.$$

Then clearly  $\tilde{C}_t$  and  $\tilde{D}_t$  are linear and bounded with adjoints  $\tilde{C}_t^* : X_t \rightarrow H$  and  $\tilde{D}_t^* : X_t \rightarrow L^2([0, t], U)$  given by

$$\tilde{C}_t^* = [C_t^* \quad C_{G,t}^*] \quad \text{and} \quad \tilde{D}_t^* = [\mathcal{D}_t^* \quad \mathcal{D}_{G,t}^*].$$

It follows that we can factorize the above expression for  $P(t)$  as

$$P(t) = \tilde{C}_t^* (\mathcal{I} + \tilde{D}_t \tilde{D}_t^*)^{-1} \tilde{C}_t.$$

Hence, the singular value decay of  $P(t)$  is the same as that of  $\tilde{C}_t^* \tilde{C}_t \in \mathcal{L}(H)$ , i.e., of  $C_t^* C_t + C_{G,t}^* C_{G,t} = C_t^* C_t + e^{tA^*} G^* G e^{tA}$ . Applying Weyl’s inequality with either the assumption that  $\dim Z < \infty$  or that the singular values of  $e^{tA}$  or  $G$  decay sufficiently fast yields the second, third, and fourth assertions, as in the proofs of Theorems 3.7 and 3.8.  $\square$

*Remark 4.3.* The above theorem extends to the case of a more general cost functional with a coercive weighting term  $\begin{bmatrix} Q & N \\ N^* & R \end{bmatrix}$  in much the same way as [29]. Since  $N = 0$  in most practical applications and  $Q$  and  $R$  may be included in  $C$  and  $B$ , respectively, we choose to omit this from the theorem and proof in order to simplify the notation.

We note that while we have only shown that the given assumptions are sufficient for fast decay of the singular values, we do not claim that they are necessary conditions. Nevertheless, violating one of the assumptions generally either leads to a not well-defined problem or slow decay. See, e.g., [29] for a number of examples in the algebraic setting. As an additional example, consider the advection equation  $\frac{d}{dt} x(t, \xi) = \frac{d}{d\xi} x(t, \xi)$  on  $\xi \in (0, \infty)$ , with  $x(0, \xi) = x_0(\xi)$ . The solution is given by  $x(t, \xi) = x_0(t)$ , i.e., it simply shifts the initial condition to the left. If the output operator  $C$  is the trace of  $x$  at 0, the output map is given by  $C_t x_0 = x_0(\cdot)$ . This means that

$$\|C_t x_0\|_{L^2([0,t],Y)}^2 = \|x_0\|_{L^2([0,t],H)}^2,$$

and  $C_t$  is therefore a partial isometry for any  $t > 0$ . Since  $H$  is infinite dimensional,  $C_t$  has infinitely many singular values that are equal to 1. The solution to the corresponding DLE therefore exhibits no decay of its singular values at all. The main problem here is the lack of analyticity of the operator  $\frac{d}{d\xi}$  (cf. [22, section 8.7A]).

On the other hand, analyticity is sometimes not strictly necessary when  $B$  and  $C$  are bounded operators. This is demonstrated for the algebraic case in [13], which shows that the solution is nuclear. That means that  $\sum_{k=1}^\infty \sigma_k(P) < \infty$ , i.e., the

singular values decay to zero at least as fast as  $1/k$ , but not necessarily as fast as  $e^{-\gamma\sqrt{k}}$ . (These results extend to the differential case.)

**5. Numerical experiments.** To demonstrate the applicability of the bounds proposed in Theorems 3.3, 3.7, 3.8, and 4.1 we have performed a few numerical experiments. In all cases, we consider DRE/DLEs arising from LQR problems with the state and output equations given by

$$(5.1) \quad \dot{x} = Ax + Bu, \quad x(0) = x_0.$$

$$(5.2) \quad y = Cx.$$

The solution  $P$  to the DRE associated with the operators  $A$ ,  $B$ , and  $C$  yields the optimal input function  $u^{\text{opt}}$  in feedback form:  $u^{\text{opt}}(t) = -B^*P(T-t)x(t)$ . It is optimal in the sense that it minimizes the cost functional

$$J(u) = \int_0^T \|y\|_Y^2 + \|u\|_U^2 dt + \|Gx(T)\|_Z^2.$$

The aim is thus to drive the output  $y$  to zero while being mindful of the cost  $\|u\|^2$  of doing so. In the extended case mentioned in Remark 4.3, the weighting factors scale the relative costs of  $y$  and  $u$ , respectively. When  $B = 0$ , the solution to the corresponding DLE yields the observability Gramian, an indicator of which states  $x$  that can be detected by using only the output  $y$ .

In all the following examples we consider the domain  $\Omega = [0, 1]^2$  to be the unit square, with boundary  $\Gamma$ . We further let the state space be  $H = L^2(\Omega)$  except where otherwise noted. We choose  $A = \Delta : \mathcal{D}(A) \subset H \rightarrow H$  to be the Laplacian. Since we will vary the boundary conditions, its domain will change as well. We can, however, always consider it to be generated by the inner product  $a(u, v) = \int_{\Omega} \nabla u \cdot \nabla v$ , where  $u, v \in V = \mathcal{D}((-A)^{1/2})$ . In the case of homogeneous Dirichlet boundary conditions, we have  $\mathcal{D}(A) = H^2 \cap H_0^1(\Omega)$  and  $V = H_0^1(\Omega)$ . We note that Assumption 2.2 is satisfied, with the region of analyticity being the entire right half-plane.

Since we cannot investigate the infinite-dimensional case in finite precision arithmetic, a discretization of the equation is required. For the spatial discretization, we have used the finite element method based on the inner product  $a$ . For a given mesh size  $h$ , we get the finite element space  $V_h \subset V \subset H$  and the approximate solution  $P_h$  is an operator from  $V_h$  to  $V_h$ . We may, however, extend it to an operator on  $H$  by forming  $\mathcal{I}_h P_h \mathcal{P}_h$ , where  $\mathcal{I}_h : V_h \rightarrow H$  denotes the identity operator and  $\mathcal{P}_h : H \rightarrow V_h$  is the  $a$ -orthogonal projection onto the finite element space. For a detailed account of the resulting matrix-valued equations, see, e.g., [26, section 5]. We generate the respective matrices here by using the library FreeFem++ [17] with  $P2$  conforming finite elements unless otherwise noted.

Further, since the discretized DLE/DREs are matrix valued and their solutions are typically dense, it is not feasible to simply transform these into vector-valued ODEs and solve them directly. We use instead the MATLAB package DREsplit<sup>2</sup> developed by the author to compute accurate low-rank approximations to the solutions. The reported singular values are thus not exact, but the integration parameters were chosen in such a way that further refining the temporal discretizations has a negligible effect on the end results. In particular, we used the second-order Strang splitting with 256 time steps. This requires the computation of many matrix exponential actions,

<sup>2</sup>Available from the author via email on request, or from [www.tonystillfjord.net](http://www.tonystillfjord.net).

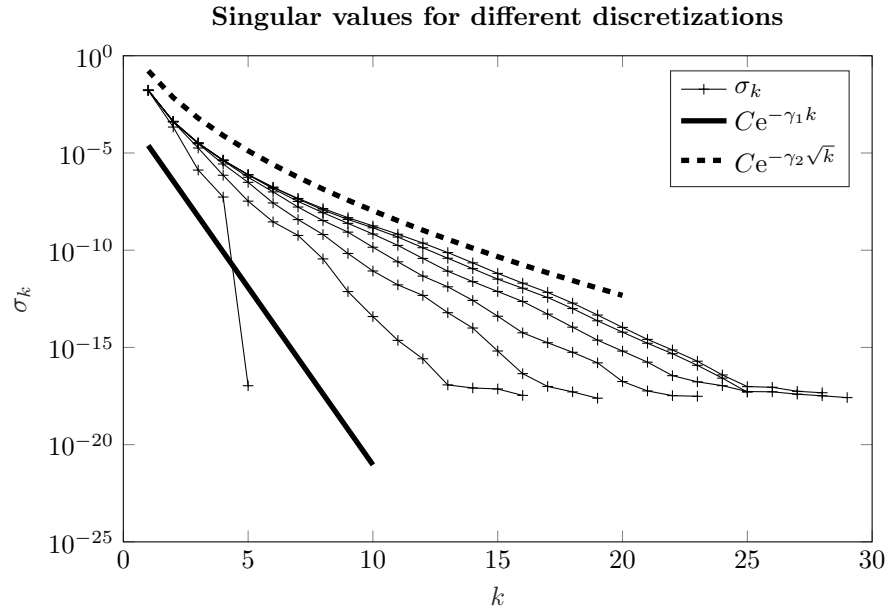


FIG. 2. The singular values of the solutions computed in Example 5.1 at the final time  $T = 0.1$ . They increase monotonically, and thus the lowermost line corresponds to  $N = 9$  while the topmost corresponds to  $N = 65025$ .

and for this a basic block Krylov subspace method with residual norm tolerance  $10^{-4}$  was employed. The relative tolerance for the low-rank approximation was set to the roundoff error level. For further details on the use of splitting schemes in this context, see, e.g., [39] or [38].

With this said, we want to note that the reported results also provide some insight into how the discretized equations converge to their infinite-dimensional counterparts.

*Example 5.1.* We consider first the bounded Lyapunov case by taking the input operator  $B = 0$  and letting the output be the mean of the solution. More specifically, we take  $Y = \mathbb{R}$  and set  $C : H \rightarrow Y$ ,  $Cx = \int_{\Omega} x$ . Then clearly  $\|Cx\|_{\mathbb{R}} \leq \|x\|_H$ , since  $\Omega$  is the unit square. We thus have  $\beta = 0$  and  $\alpha = 0$ . Further setting  $G = 0$  implies that Assumptions 2.3 to 2.5 are satisfied. To complete the specification of  $A$ , we choose homogeneous Dirichlet boundary conditions.

We computed the singular values for a number of different spatial discretizations, starting with a grid that has  $N = 9$  internal nodes and refining this 6 times. Each refinement roughly halves the mesh size and thus roughly quadruples the number of nodes, leading to meshes with  $N = 9, 49, 225, 961, 3969, 16129$ , and  $65025$  internal nodes, respectively. Figure 2 shows the computed singular values of the solutions (the  $\mathcal{L}(H)$ -extended operators, not the matrices) for different spatial discretizations, at the final time  $T = 0.1$ . The curves are ordered in size from bottom to top, i.e., the lowermost curve corresponds to the  $N = 9$  discretization, while the topmost corresponds to the  $N = 65025$  discretization. We observe that while the initial decay is very much exponential in nature, when we refine the discretization the decay worsens and tends to the exponential square-root bound. This is precisely the same behavior as seen in the algebraic case in, e.g., [16].

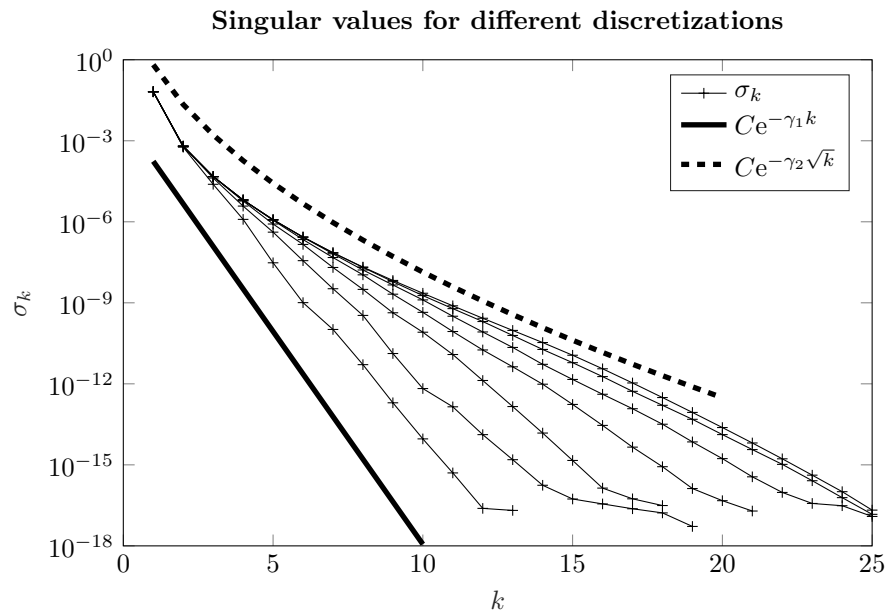


FIG. 3. The singular values of the solutions computed in Example 5.2 at the final time  $T = 0.1$ . They increase monotonically, and thus the lowermost line corresponds to  $N = 20$  while the topmost corresponds to  $N = 65792$ .

*Example 5.2.* In the second example, we change the boundary conditions of  $A$  to be homogeneous Dirichlet on the left edge  $\Gamma_L$  and homogeneous Neumann on the top and bottom edges  $\Gamma_T, \Gamma_B$ . On the right edge,  $\Gamma_R$ , we apply a nonhomogeneous Neumann boundary condition, through which we control the system. That is, we set  $U = \mathbb{R}$  and define  $B : U \rightarrow \mathcal{D}(A^*)'$  by  $Bu = -(AN\mathbb{1})u$ , where the function  $\mathbb{1} \in L^2(\Gamma_R)$  is constant equal to 1 everywhere and  $N : L^2(\Gamma_R) \rightarrow H^{3/2}(\Omega)$  denotes the Neumann operator implicitly defined by  $Nv = w$  if  $Aw = 0$  in  $\Omega$ ,  $\frac{\partial w}{\partial \nu}|_{\Gamma_R} = v$ ,  $w|_{\Gamma_L} = 0$ , and  $\frac{\partial w}{\partial \nu}|_{\Gamma_T \cup \Gamma_B} = 0$ . For further details on this construction, see, e.g., [21, section 3]. That  $N$  maps into  $H^{3/2}(\Omega)$  follows by [24, Thm. 8.3] and shows that  $(-A)^{-\beta}B \in \mathcal{L}(U, H)$  for  $\beta = 1/4 + \epsilon$ ,  $\epsilon > 0$ .

We note that we could equally well take  $U = L^2(\Gamma)$  in the continuous setting and let the input  $u$  vary along the whole edge. However, for the numerics we would then have to also discretize this function, leading to one more layer of complexity.

As the output, we again use the mean of the solution over the whole domain  $\Omega$ , meaning that  $\alpha = 0$ . We discretize the system in the same way as in Example 5.1, but because of the three Neumann edges we now have a slightly higher number of degrees of freedom for each level of discretization. The matrices are in this case of size  $N = 20, 72, 272, 1056, 4160, 16512$ , and  $65792$ , respectively.

Figure 3 shows the computed singular values of the solutions at the final time  $T = 0.1$ . The curves are again ordered in size from coarse (bottom) to fine (top) discretizations. We note that these results are quite similar to the results in Figure 2, i.e., the input operator does not make the situation worse, as predicted by Theorem 4.1.

We have additionally plotted the largest singular value of the finest discretized problem as a function of time in Figure 4. We note that it grows roughly as  $t^1$ , corresponding well to the factor  $t^{1-2\alpha}$  predicted by Theorem 4.1.

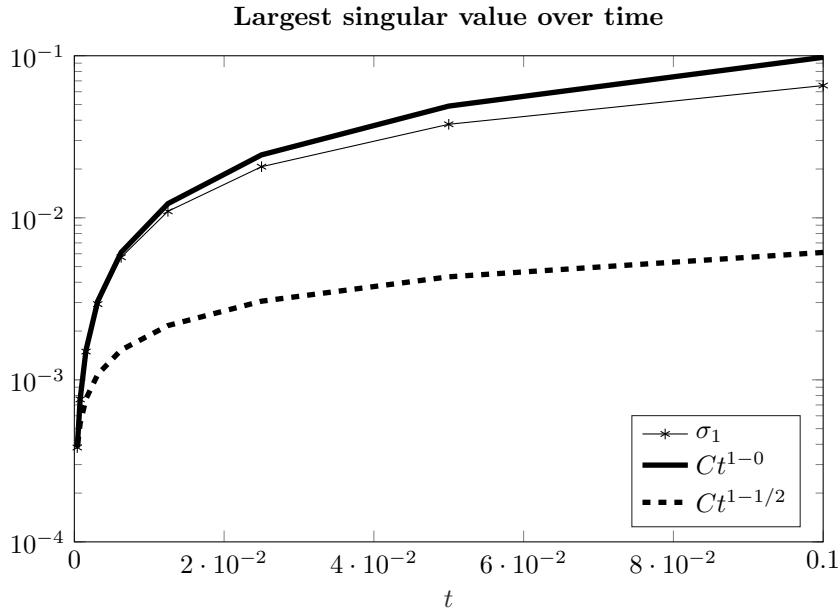


FIG. 4. The largest singular value of the solution with  $N = 65792$  computed in Example 5.2, plotted over time.

*Example 5.3.* Now consider the same setting as in the previous example, but with an unbounded output as well. More precisely, we take  $Y = \mathbb{R}$  and define  $C$  as the integral of the boundary trace over  $\Gamma_T \cap \Gamma_B$ :

$$Cx = \int_{\Gamma_T \cap \Gamma_B} x|_{\Gamma}(s) \, ds.$$

By [24, Theorem 8.3], the map  $x \mapsto x|_{\Gamma}$  belongs to  $\mathcal{L}(H^{1/2}(\Omega), L^2(\Gamma))$  and hence the map  $CA^{-\alpha}$  is bounded for  $\alpha = 1/4 + \epsilon$ ,  $\epsilon > 0$ .

With the same discretizations as in Example 5.2, the behavior of the singular values is similar to when  $C$  was bounded. The decay is, however, noticeably slower, as shown in Figure 5. The effect of a larger  $\alpha$  can also clearly be seen when plotting the singular values for a specific discretization over time. Figure 6 again shows the largest singular value for the finest discretization. We note that in comparison to Figure 4, the increase is now close to  $t^{1/2}$  rather than  $t^1$ . Since  $\alpha = 1/4$ , this is in good agreement with the factor  $t^{1-2\alpha}$  predicted by Theorem 4.1.

*Example 5.4.* Let us now consider a situation when the main assumptions are not satisfied. In particular, let us take the same setup as in Example 5.3 except for the output operator. We now instead take the trace of the normal derivative:

$$Cx = \int_{\Gamma_T \cap \Gamma_B} \left( \frac{\partial x}{\partial \nu} \right) |_{\Gamma}(s) \, ds.$$

Again by [24, Theorem 8.3], the map  $x \mapsto \left( \frac{\partial x}{\partial \nu} \right) |_{\Gamma}$  belongs to  $\mathcal{L}(H^{3/2}(\Omega), L^2(\Gamma))$  and hence the map  $CA^{-\alpha}$  is bounded for  $\alpha = 3/4 + \epsilon$ ,  $\epsilon > 0$ . Since  $\alpha > 1/2$ , Assumption 2.4 is not satisfied, and we can in fact not show the existence of a solution  $P \in \mathcal{L}(H)$ .



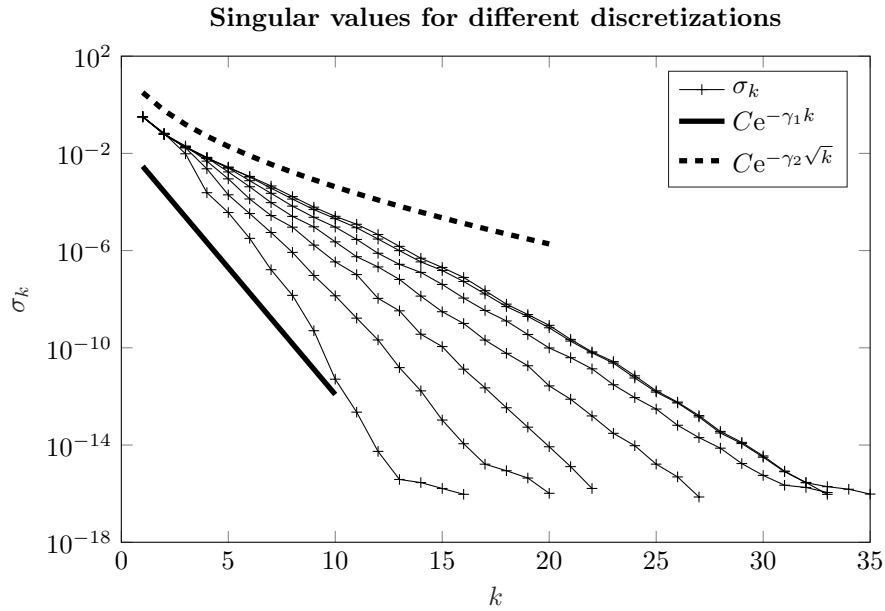


FIG. 5. The singular values of the solutions computed in Example 5.3 at the final time  $T = 0.1$ . They increase monotonically, and thus the lowermost line corresponds to  $N = 20$  while the topmost corresponds to  $N = 65792$ .

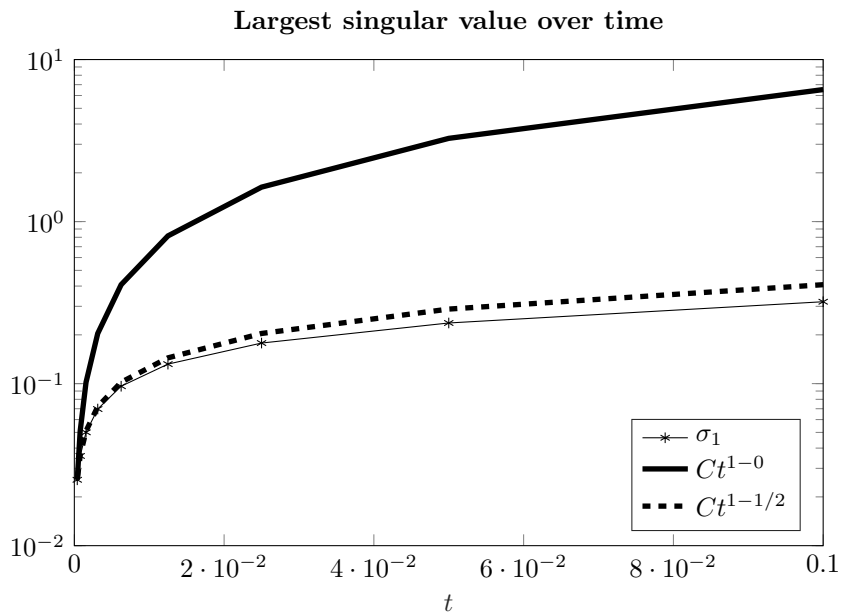


FIG. 6. The largest singular value of the solution with  $N = 65792$  computed in Example 5.3, plotted over time.

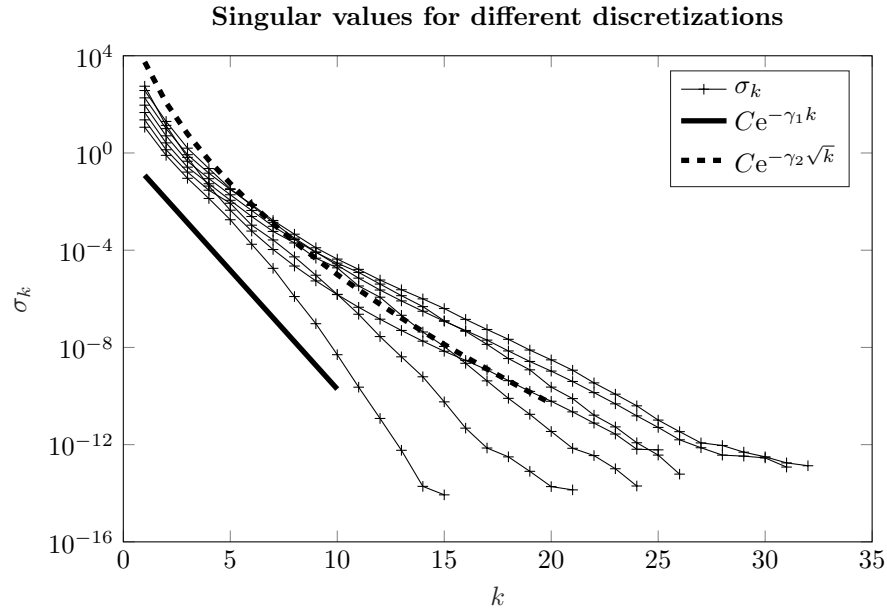


FIG. 7. The singular values of the solutions computed in Example 5.4 at the final time  $T = 0.1$ . They increase (roughly) monotonically, and thus the lowermost line corresponds to  $N = 20$  while the topmost corresponds to  $N = 65792$ . Because the underlying problem is not well-posed, the discretized solutions increase without bound.

This is reflected in the results shown in Figure 7. We have discretized the problem in the same way as previously, and we plot the singular values for the different discretizations like in Figures 2 and 3. In contrast to the previous results, we now see that the singular values keep increasing as we refine the discretization, demonstrating that the singular values of the exact solution are infinite. Thus, while the singular values of a single discretized matrix-valued equation seem to decay exponentially, since the underlying problem is not well-posed these “approximations” are nevertheless worthless.

*Example 5.5.* The situation in the previous example holds when we use  $H = L^2(\Omega)$ . By instead selecting a smaller state space  $H$ , we decrease the value of  $\alpha$ . With  $H = \{x \in H^1(\Omega) ; x|_{\Gamma_L} = 0\}$  and the same operator  $C$  we again get  $\alpha = 1/4 + \epsilon$ . Since we simultaneously increase  $\beta$  by  $1/2$ , we set  $B = 0$  in this example to comply with Assumption 2.3.

We note that we now consider the operator  $A$  as restricted to  $H$  instead of an operator on  $L^2(\Omega)$ . It still generates an analytic semigroup and Assumption 2.2 is satisfied. Since the finite element discretization of the problem is no longer based on  $a(u, v)$  but on the corresponding inner product defined on  $H^1$ , the resulting problem is similar to a biharmonic equation. This imposes extra regularity requirements on the standard conforming finite element spaces, requiring a high number of nodes [11, p. 286]. In order to avoid this, in this example we employ instead the nonconforming Morley elements [28, 11].

The results are shown in Figure 8. We see that since  $\alpha$  is now again less than  $1/2$ , the singular values behave much like in the previous examples.

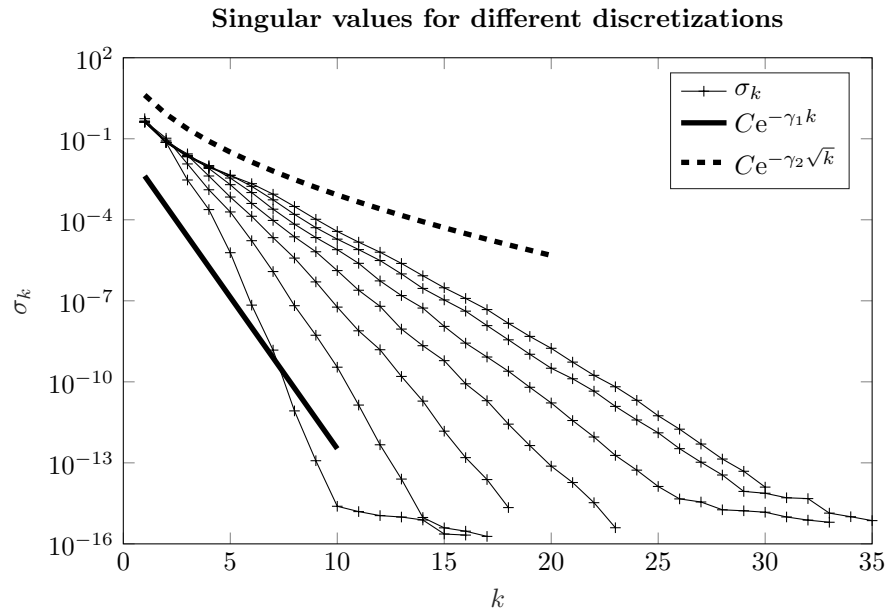


FIG. 8. The singular values of the solutions computed in Example 5.5 at the final time  $T = 0.1$ . They increase monotonically, and thus the lowermost line corresponds to  $N = 20$  while the topmost corresponds to  $N = 65792$ .

**6. Conclusions.** We have proved bounds for the singular values  $\sigma_k$  of the solutions to DLEs and DREs of the form  $\sigma_k \leq M e^{-\gamma\sqrt{k}}$ , extending previous results on algebraic equations to the differential case. This is important, since utilizing the property of low numerical rank is a critical feature in numerical methods for these problems in the large-scale setting. If low numerical rank, i.e., a sufficiently rapid decay of the singular values, cannot be guaranteed, these methods never finish, or fail outright. The current work is thus a step on the way to provide practical criteria for when this is to be expected. We say “a step on the way” because while we have given conditions for when exponential square-root decay is to be expected, we have not indicated how large the constant multiplier in the bound can be. A large value could mean that the numerical rank is too large to be useful in a practical application, even though the decay is  $\mathcal{O}(e^{-\gamma\sqrt{k}})$ . However, the size of this constant depends strongly on the properties of the operators  $A$  and  $C$ , and providing a generally meaningful bound is difficult with current techniques. We therefore leave this question open for future research, but note that the constants arising in our numerical experiments are all of moderate size.

A further interesting unexplored question is how the singular values of the solutions to the spatially discretized matrix-valued problems relate to those of the operator-valued solutions. As noted in the numerical experiments, one often observes exponential decay in the discretized case. When the discretization is refined, the decay rate deteriorates and eventually tends to the exponential square-root bound. The form of this decrease is, however, unclear. While it can be argued that the discretized equations are only steps on the way towards the nondiscretized goal (and the author does argue thus), in practical computations we are of course always in the matrix-valued situation. Analyzing also this case and providing a connection between

the decay rate and the discretization level is therefore both highly interesting and important, but clearly requires a different approach.

**Acknowledgments.** The author is grateful to Mark Opmeer for providing several helpful references. The careful reading and constructive comments from the anonymous referees also led to a greatly improved manuscript.

## REFERENCES

- [1] H. ABOU-KANDIL, G. FREILING, V. IONESCU, AND G. JANK, *Matrix Riccati Equations in Control and Systems Theory*, Birkhäuser, Basel, Switzerland, 2003.
- [2] A. C. ANTOUNAS, D. C. SORENSEN, AND Y. ZHOU, *On the decay rate of Hankel singular values and related issues*, *Systems Control Lett.*, 46 (2002), pp. 323–342, [https://doi.org/10.1016/S0167-6911\(02\)00147-0](https://doi.org/10.1016/S0167-6911(02)00147-0).
- [3] T. BAŞAR AND P. BERNHARD,  *$H^\infty$ -Optimal Control and Related Minimax Design Problems: A Dynamic Game Approach*, *Systems Control Found. Appl.*, 2nd ed., Birkhäuser, Boston, MA, 1995, <https://doi.org/10.1007/978-0-8176-4757-5>.
- [4] J. BAKER, M. EMBREE, AND J. SABINO, *Fast singular value decay for Lyapunov solutions with nonnormal coefficients*, *SIAM J. Matrix Anal. Appl.*, 36 (2015), pp. 656–668, <https://doi.org/10.1137/140993867>.
- [5] U. BAUR, P. BENNER, AND L. FENG, *Model order reduction for linear and nonlinear systems: A system-theoretic perspective*, *Arch. Comput. Methods Eng.*, 21 (2014), pp. 331–358, <https://doi.org/10.1007/s11831-014-9111-2>.
- [6] P. BENNER AND T. BREITEN, *Low rank methods for a class of generalized Lyapunov equations and related issues*, *Numer. Math.*, 124 (2013), pp. 441–470, <https://doi.org/10.1007/s00211-013-0521-0>.
- [7] P. BENNER AND Z. BUJANOVIĆ, *On the solution of large-scale algebraic Riccati equations by using low-dimensional invariant subspaces*, *Linear Algebra Appl.*, 488 (2016), pp. 430–459, <https://doi.org/10.1016/j.laa.2015.09.027>.
- [8] P. BENNER, P. KÜRSCHNER, AND J. SAAK, *Frequency-limited balanced truncation with low-rank approximations*, *SIAM J. Sci. Comput.*, 38 (2016), pp. A471–A499, <https://doi.org/10.1137/15M1030911>.
- [9] P. BENNER, J.-R. LI, AND T. PENZL, *Numerical solution of large-scale Lyapunov equations, Riccati equations, and linear-quadratic optimal control problems*, *Numer. Linear Algebra Appl.*, 15 (2008), pp. 755–777, <https://doi.org/10.1002/nla.622>.
- [10] A. BENSOUSSAN, G. DA PRATO, M. C. DELFOUR, AND S. K. MITTER, *Representation and Control of Infinite Dimensional Systems*, *Systems Control Found. Appl.*, 2nd ed., Birkhäuser, Boston, MA, 2007.
- [11] S. C. BRENNER AND L. R. SCOTT, *The Mathematical Theory of Finite Element Methods*, *Texts Appl. Math.* 15, 3rd ed., Springer, New York, 2008, <https://doi.org/10.1007/978-0-387-75934-0>.
- [12] R. COURANT AND D. HILBERT, *Methods of Mathematical Physics. Vol. I*, Interscience, New York, 1953.
- [13] R. F. CURTAIN AND A. J. SASANE, *Compactness and nuclearity of the Hankel operator and internal stability of infinite-dimensional state linear systems*, *Internat. J. Control*, 74 (2001), pp. 1260–1270, <https://doi.org/10.1080/00207170110061059>.
- [14] K. FAN, *Maximum properties and inequalities for the eigenvalues of completely continuous operators*, *Proc. Natl. Acad. Sci. USA*, 37 (1951), pp. 760–766.
- [15] W. GAWRONSKI AND J.-N. JUANG, *Model reduction in limited time and frequency intervals*, *Internat. J. Systems Sci.*, 21 (1990), pp. 349–376, <https://doi.org/10.1080/00207729008910366>.
- [16] L. GRUBIŠIĆ AND D. KRESSNER, *On the eigenvalue decay of solutions to operator Lyapunov equations*, *Systems Control Lett.*, 73 (2014), pp. 42–47, <https://doi.org/10.1016/j.sysconle.2014.09.006>.
- [17] F. HECHT, *New development in freefem++*, *J. Numer. Math.*, 20 (2012), pp. 251–265.
- [18] A. ICHIKAWA AND H. KATAYAMA, *Remarks on the time-varying  $H_\infty$  Riccati equations*, *Systems Control Lett.*, 37 (1999), pp. 335–345.
- [19] P. KÜRSCHNER, *Balanced truncation model order reduction in limited time intervals for large systems*, *Adv. Comput. Math.*, to appear.
- [20] N. LANG, H. MENA, AND J. SAAK, *On the benefits of the  $LDL^T$  factorization for large-scale differential matrix equation solvers*, *Linear Algebra Appl.*, 480 (2015), pp. 44–71, <https://doi.org/10.1016/j.laa.2015.04.006>.

- [21] I. LASIECKA AND R. TRIGGIANI, *Control Theory for Partial Differential Equations: Continuous and Approximation Theories I. Abstract Parabolic Systems*, Cambridge University Press, Cambridge, 2000.
- [22] I. LASIECKA AND R. TRIGGIANI, *Control theory for partial differential equations: Continuous and approximation theories II. Abstract hyperbolic-like systems over a finite time horizon*, in *Encyclopedia Math. Appl.* 75, Cambridge University Press, Cambridge, 2000, pp. 645–1067.
- [23] J.-R. LI AND J. WHITE, *Low rank solution of Lyapunov equations*, *SIAM J. Matrix Anal. Appl.*, 24 (2002), pp. 260–280, <https://doi.org/10.1137/S0895479801384937>.
- [24] J.-L. LIONS AND E. MAGENES, *Non-homogeneous Boundary Value Problems and Applications. Vol. I*, Grundlehren Math. Wiss. 181, Springer, New York, 1972.
- [25] J. LUND AND K. L. BOWERS, *Sinc Methods for Quadrature and Differential Equations*, Other Titles Appl. Math., SIAM, Philadelphia, 1992, <https://doi.org/10.1137/1.9781611971637>.
- [26] A. MÅLQVIST, A. PERSSON, AND T. STILLEFJORD, *Multiscale differential Riccati equations for linear quadratic regulator problems*, *SIAM J. Sci. Comput.*, 40 (2018), pp. A2406–A2426.
- [27] K. M. MIKKOLA, *Infinite-Dimensional Linear Systems, Optimal Control and Algebraic Riccati Equations*, Dissertation, Helsinki University of Technology, Helsinki, Finland, 2002, <http://lib.tkk.fi/Diss/2002/isbn9512260794/>.
- [28] L. S. D. MORLEY, *The triangular equilibrium element in the solution of plate bending problems*, *Aeronaut. Quart.*, 19 (1968), pp. 149–169.
- [29] M. OPMEER, *Decay of singular values of the Gramians of infinite-dimensional systems*, in *Proceedings 2015 European Control Conference (ECC)*, Linz, Austria, IEEE, Piscataway, NJ, 2015, pp. 1183–1188, <https://doi.org/10.1109/ECC.2015.7330700>.
- [30] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Appl. Math. Sci. 44, Springer, New York, 1983.
- [31] T. PENZL, *Eigenvalue decay bounds for solutions of Lyapunov equations: The symmetric case*, *Systems Control Lett.*, 40 (2000), pp. 139–144, [https://doi.org/10.1016/S0167-6911\(00\)00010-4](https://doi.org/10.1016/S0167-6911(00)00010-4).
- [32] I. R. PETERSEN, V. A. UGRINOVSKII, AND A. V. SAVKIN, *Robust Control Design Using  $H^\infty$  Methods*, Springer, London, 2000.
- [33] D. SALAMON, *Infinite-dimensional linear systems with unbounded control and observation: A functional analytic approach*, *Trans. Amer. Math. Soc.*, 300 (1987), pp. 383–431, <https://doi.org/10.2307/2000351>.
- [34] D. C. SORENSEN AND Y. ZHOU, *Bounds on Eigenvalue Decay Rates and Sensitivity of Solutions to Lyapunov Equations*, Technical report TR02-07, Rice University, Houston, TX, 2002, <https://scholarship.rice.edu/handle/1911/101987>.
- [35] O. STAFFANS, *Well-Posed Linear Systems*, *Encyclopedia Math. Appl.* 103, Cambridge University Press, Cambridge, 2005, <https://doi.org/10.1017/CBO9780511543197>.
- [36] F. STENGER, *Integration Formulae Based on the Trapezoidal Formula*, *J. Inst. Math. Appl.*, 12 (1973), pp. 103–114.
- [37] F. STENGER, *Numerical Methods Based on Sinc and Analytic Functions*, Springer Ser. Comput. Math. 20, Springer, New York, 1993, <https://doi.org/10.1007/978-1-4612-2706-9>.
- [38] T. STILLFJORD, *Low-rank second-order splitting of large-scale differential Riccati equations*, *IEEE Trans. Automat. Control*, 60 (2015), pp. 2791–2796, <https://doi.org/10.1109/TAC.2015.2398889>.
- [39] T. STILLFJORD, *Adaptive high-order splitting schemes for large-scale differential Riccati equations*, *Numer. Algorithms*, 78 (2017), pp. 1129–1151, <https://doi.org/10.1007/s11075-017-0416-8>.
- [40] M. TUCSNAK AND G. WEISS, *Observation and Control for Operator Semigroups*, Birkhäuser Adv. Texts Basler Lehrbücher, Birkhäuser, Basel, 2009, <https://doi.org/10.1007/978-3-7643-8994-9>.