D. Wittwar[a] · A. Schmidt[a] · B. Haasdonk[a]

Reduced Basis Approximation for the Discrete-time Parametric Algebraic Riccati Equation

[a] Institute for Applied Analysis and Numerical Simulation, University of Stuttgart,
Pfaffenwaldring 57, 70569 Stuttgart/ Germany
{dominik.wittwar,andreas.schmidt,bernard.haasdonk}@mathematik.uni-stuttgart.de
www.agh.ians.uni-stuttgart.de

Abstract In this work we study the application of the reduced basis (RB) approximation method for parametric discrete-time algebraic Riccati equations (DARE). The DARE is very challenging to solve in large dimensions and parametric problems for large-scale applications are therefore often infeasible. We thus propose to apply the low-rank factor greedy (LRFG) algorithm to build a suitable low-dimensional subspace for the model reduction approach. Furthermore, we perform a rigorous error estimation, including an effectivity analysis and show how the RB-DARE procedure can be implemented efficiently. Numerical examples for an application in feedback control prove the benefits, in particular excellent speedups and reliability of the error estimators.

Keywords Reduced basis method, large-scale matrix equations, parametric discrete-time algebraic Riccati equation

REDUCED BASIS APPROXIMATION FOR THE DISCRETE-TIME PARAMETRIC ALGEBRAIC RICCATI EQUATION

D. WITTWAR*, A. SCHMIDT*, AND B. HAASDONK*

Abstract. In this work we study the application of the reduced basis (RB) approximation method for parametric discrete-time algebraic Riccati equations (DARE). The DARE is very challenging to solve in large dimensions and parametric problems for large-scale applications are therefore often infeasible. We thus propose to apply the low-rank factor greedy (LRFG) algorithm to build a suitable low-dimensional subspace for the model reduction approach. Furthermore, we perform a rigorous error estimation, including an effectivity analysis and show how the RB-DARE procedure can be implemented efficiently. Numerical examples for an application in feedback control prove the benefits, in particular excellent speedups and reliability of the error estimators.

Key words. Reduced basis method, large-scale matrix equations, parametric discrete-time algebraic Riccati equation

AMS subject classifications. 49N05, 34K35, 65G99

1. Introduction. In this paper we consider the application of the reduced basis (RB) technique to the parametric discrete-time algebraic Riccati equation (DARE), see [15, 16]. Such equations arise frequently in the context of discrete-time systems theory. Popular examples comprise the optimal feedback control and state estimation of linear and time-invariant systems, which are both standard tasks in systems theory, cf. [1, 14]. Often those systems depend on one or several parameters and multiple queries for different parameters are required, e.g. parameter studies, real-time parameter adaptions or statistical analysis. Furthermore, many realistic models are derived from partial differential equations (PDEs), which yield large spatially discretized models. In such cases, and especially in the additional presence of parameters, the solution of the above mentioned problems can be cumbersome, expensive and easily become infeasible. RB methods have proven to be an efficient tool for rapidly solving parameter dependent PDEs, cf. [4, 9, 11, 12]. Therefore, we are interested in developing a RB technique for the DARE. In particular, we continue in the direction of RB methods for parametric matrix-equations, see also [22, 21]. In contrast to the previous two references, we cover a non polynomial, i.e., rational matrix-equation which requires considerable modifications in the analysis.

This paper is structured as follows: We begin with a brief introduction to the theoretical background of the DARE and the reduced model in Section 2. In the subsequent sections we then address the offline/online computational strategy and a-posteriori error analysis. The following numerical examples in Section 5 show the speed-up that can be obtained when using the proposed scheme and illustrate the performance of the error bounds. We conclude with some remarks and an outlook in the final section.

*Institute for Applied Analysis and Numerical Simulation, University of Stuttgart, Germany

2. The RB-DARE method.

2.1. Problem setting. In this article we consider the parametric discrete(-time) algebraic Riccati equation:

(1)
$$E(\mu)^T X(\mu) E(\mu) = A(\mu)^T X(\mu) A(\mu) + C(\mu)^T Q(\mu) C(\mu)$$
$$- A(\mu)^T X(\mu) B(\mu) \left(R(\mu) + B(\mu)^T X(\mu) B(\mu) \right)^{-1} B(\mu)^T X(\mu) A(\mu).$$

The parameter dependent matrices $A(\mu), E(\mu) \in \mathbb{R}^{n \times n}$, $B(\mu) \in \mathbb{R}^{n \times m}$, $C(\mu) \in \mathbb{R}^{p \times n}$, $Q(\mu) \in \mathbb{R}^{p \times p}$ and $R(\mu) \in \mathbb{R}^{m \times m}$, where $E(\mu)$ is invertible, are given and the solution matrix $X(\mu) \in \mathbb{R}^{n \times n}$ is sought. Furthermore, we assume that $Q(\mu)$ and $R(\mu)$ are symmetric positive semi-definite and symmetric positive definite, respectively, which ensures the invertibility of $R(\mu)$. Additionally, we consider cases were $m, p \ll n$, i.e., the input and output dimension are small compared to the dimension of the system, see also Remark 1. The parameter vector μ stems from a bounded set $\mathcal{P} \subset \mathbb{R}^d$, which is called the set of all admissible parameters. In the following we omit the parameter dependency in the notation of the system matrices to provide a better reading experience. However, all definitions and statements should be understood to hold for all $\mu \in \mathcal{P}$.

Equation (1) is a nonlinear rational matrix-valued equation, and it can have multiple solutions among which only one is typically of interest: We call a solution $X \in \mathbb{R}^{n \times n}$ a stabilizing solution of the DARE, when it is symmetric, positive semidefinite and stabilizing in the sense that (E, A_X) is stable, which means that all eigenvalues λ_i of the generalized eigenvalue problem $A_X x = \lambda E x$ satisfy $|\lambda_i| < 1$. Here and in the following, $A_X \in \mathbb{R}^{n \times n}$ denotes the so-called closed loop matrix

(2) $$A_X := A - B K_X, \text{ with } K_X := \left(R + B^T X B \right)^{-1} B^T X A,$$

where K_X denotes the feedback-gain matrix, see also Remark 1. In general neither the existence nor the uniqueness of solutions to (1) are guaranteed and additional restrictions on the system matrices must be posed. One possible set of restrictions is the stabilizability and detectability of the matrices $E, A, B, Q^{1/2} C$, where $Q^{1/2}$ denotes a symmetric positive semi-definite square root of Q. We call the tuple (E, A, B) stabilizable, if there exists a matrix $K \in \mathbb{R}^{m \times n}$ such that $(E, A - BK)$ is stable. Likewise, the tuple $(E, A, Q^{1/2} C)$ is called detectable, if the tuple $(E^T, A^T, C^T Q^{1/2})$ is stable. We summarize the results in the following proposition:

PROPOSITION 1 (Sufficient conditions for the existence of a unique stabilizing solution). *Let (E, A, B) be a stabilizable tuple of matrices and $(E, A, Q^{1/2} C)$ be a detectable tuple of matrices. Then there exists a unique stabilizing solution X to the DARE* (1).

Proof. See for example [17]. □

Although nonsymmetric solutions to (1) can exist, we are usually only interested in the subset of symmetric solutions $X \in \mathcal{D}_n \subset \mathcal{S}_n$, where $\mathcal{S}_n := \{ S \in \mathbb{R}^{n \times n} : S = S^T \}$ denotes the set of symmetric $n \times n$ matrices and $\mathcal{D}_n := \{ S \in \mathcal{S}_n : R + B^T S B \text{ is regular} \}$ denotes the open subset of all viable matrices such that (1) and (2) are well defined. We now define the residual of the DARE (1) $\mathcal{R} : \mathcal{D}_n \to \mathcal{S}_n$ by

(3) $$\mathcal{R}(X) := E^T X E - A^T X A - C^T Q C + A^T X B \left(R + B^T X B \right)^{-1} B^T X A.$$

We can conclude, that X is a solution to the DARE iff $\mathcal{R}(X) = 0$, a property which is used to approximate solutions to the DARE by applying Newton based procedures to $\mathcal{R}(X)$ in order to find its zeros, see for example [3, 17].

An interesting application of the DARE is the feedback stabilization of linear and time-invariant discrete-time systems, what we will briefly recall in the following remark.

REMARK 1. *Consider the following linear quadratic (LQ) optimal control problem*

$$(4) \qquad \min_{u \in \ell^2(\mathbb{R}^m)} \sum_{k=0}^{\infty} y_k^T Q y_k + u_k^T R u_k$$

$$(5) \qquad s.t. \begin{cases} Ex_{k+1} &= Ax_k + Bu_k, \quad k \geq 0 \\ y_k &= Cx_k, \quad\quad\;\; k \geq 0 \\ x_0 &= x^0 \in \mathbb{R}^n. \end{cases}$$

With the set $\ell^2(\mathbb{R}^m)$ we denote the set of all square-integrable sequences in $\mathbb{R}^m$ and we abbreviate $u = (u_0, u_1, \dots)$. Systems of the form (5) arise for example after spatial and temporal discretization of linear PDEs. The control problem (4)–(5) is called linear quadratic regulator (LQR) problem, and has many important applications, for example in engineering problems. The solution to the problem can be calculated explicitly and has the form $u_k := \Phi(x_k)$, with $\Phi(x) := -K_X x$ and $X \in \mathbb{R}^{n \times n}$ is the unique stabilizing solution to the DARE (1), c.f. [17]. Since the control input u_k is directly depending on the state x_k, this type of control is also called feedback control.

Finally, we want to comment on some notation used throughout this article: By $\|\cdot\|$ we mean the Eucliden 2-norm for vectors and the induced norm for operators (matrices). The Frobenius norm of a matrix X is denoted as $\|X\|_F = \sqrt{\sum_{i,j} X_{ij}^2}$, where X_{ij} denotes the entry in the i-th row and the j-th column in the matrix X. By range(V) we denote the linear space spanned by the columns of V. For two symmetric matrices $A, B \in \mathbb{R}^{n \times n}$ we write $A \succeq B$ if $A - B$ is positive semi-definite and $A \succ B$ if $A - B$ is positive definite. Furthermore, for $N \in \mathbb{N}$ let $I_N \in \mathbb{R}^{N \times N}$ denote the identity matrix and $B_\alpha(X)$ the open ball with radius α around X.

2.2. Reduced problem. In oder to derive a reduced problem, we make use of a method based on projections which has already been succesfully applied to different matrix equations. In [13, 20] a method for solving large scale Lyapunov equations is presented, which is based on the projection on a Krylov subspace. The method we present in the following is based on a Petrov-Galerkin projection, i.e., the residual has to satisfy an orthogonality condition, which is an often used criterion in the RB framework, see e.g. [9, 19]. Therefore, we consider a pair of matrices $W, V \in \mathbb{R}^{n \times N}$ of rectangular shape with $N \ll n$ which are assumed to be biorthogonal, i.e. $W^T V = I_N$. We approximate the solution to the DARE in the space of all matrices whose colums are contained in range(W), i.e. the approximation $\hat{X}$ lives (due to symmetry) in the space $\mathcal{W} := \{S \in \mathcal{S}_n : S = W S_N W^T, \text{where } S_N \in \mathcal{S}_N\}$. In an analogous fashion we define the space $\mathcal{V}$. We now impose the aforementioned Petrov-Galerkin condition on the residual $\mathcal{R}(\hat{X})$ with regard to the subspace $\mathcal{V}$, i.e., $\mathcal{R}(\hat{X})$ is perpendicular to $\mathcal{V}$. Therefore, we endow the Hilbert space $\mathcal{S}_n$ with the trace inner product $\langle S, T \rangle := \text{tr}(ST)$ for $S, T \in \mathcal{S}_n$. Making use of the symmetry of $\mathcal{R}(\hat{X})$ and some properties of the trace operator, the Petrov-Galerkin condition can be rewritten as

$$(6) \qquad V^T \mathcal{R}(\hat{X}) V = 0.$$

Inserting the representation $\hat{X} := W X_N W^T$ as an element of $\mathcal{W}$ we arrive at

(7)
$$E_N^T X_N E_N = A_N^T X_N A_N + C_N^T Q C_N$$
$$- A_N^T X_N B_N \left(R + B_N^T X_N B_N \right)^{-1} B_N^T X_N A_N.$$

Equation (7) constitutes a DARE where the system matrices are given as

$$A_N := W^T A V \in \mathbb{R}^{N \times N}, \qquad E_N := W^T E V \in \mathbb{R}^{N \times N},$$
$$B_N := W^T B \in \mathbb{R}^{N \times m}, \qquad C_N := C V \in \mathbb{R}^{p \times N}.$$

We denote (7) as the reduced DARE as it is low dimensional. Its unique stabilizing solution X_N, if it does exist, can be calculated rapidly. Additionally, we denote X_N as the reduced solution and $\hat{X} := W X_N W^T$ as the RB approximation of X. For given X_N we can then compute a low rank factor $\hat{Z}$ of the approximation $\hat{X}$, i.e. $\hat{X} = \hat{Z} \hat{Z}^T$, by employing the eigenvalue decomposition of $X_N = U_N S_N U_N^T$ and by setting

$$\hat{Z} := W U_N S_N^{1/2} \in \mathbb{R}^{n \times N}.$$

This is done in order to avoid computing and storing the dense high dimensional matrix $\hat{X}$. The first property we want to show is inherent to most RB approximation methods: the reproduction of solutions, i.e. $X \in \mathcal{W}$ implies $X = \hat{X}$.

PROPOSITION 2 (Reproduction of solution). *Let $W, V \in \mathbb{R}^{n \times N}$ be given. Let the symmetric positive semidefinite matrix $X \in \mathbb{R}^{n \times n}$ be the unique stabilizing solution to the DARE (1) and assume $\mathrm{range}(X) \subset \mathrm{range}(W)$. Furthermore, assume that for the reduced system (7) the triplets (E_N, A_N, B_N) and $(E_N, A_N, Q^{1/2} C_N)$ are stabilizable and detectable, respectively. Then it follows*

$$X = W X_N W^T = \hat{X},$$

where X_N denotes the stabilizing solution of the reduced DARE.

Proof. Since X is symmetric and $\mathrm{range}(X) \subset \mathrm{range}(W)$ holds by assumption, there exists a unique symmetric positive semidefinite matrix $G \in \mathbb{R}^{N \times N}$, such that $X = W G W^T$. By making use of the fact, that X is a solution to the DARE (1) we get

$$E^T W G W^T E = A^T W G W^T A + C^T Q C$$
$$- A^T W G W^T B \left(R + B^T W G W^T B \right)^{-1} B^T W G W^T A.$$

Multiplying from left with V^T and V from right yields

$$E_N^T G E_N = A_N^T G A_N + C_N^T Q C_N - A_N^T G B_N \left(R + B_N^T G B_N \right)^{-1} B_N^T G A_N.$$

Therefore, G is a symmetric positive semidefinite solution to the reduced DARE (7) and we can conclude, due to the assumption of stabilizability and detectability of the reduced matrices, that G is the unique stabilizing solution of the reduced DARE, i.e. $G = X_N$. It now follows, that $\hat{X} = W X_N W^T = W G W^T = X$. $\square$

2.3. Basis construction. It remains to show how a reduced basis for the DARE can be constructed. The key ingredient, which on the one hand allows the rapid calculation of solutions of large-scale DAREs, and on the other hand opens the way towards

an efficient approximation with an RB-scheme, is the so-called low-rank structure in the solution matrices $X(\mu) \in \mathbb{R}^n$. The low-rank property here means a rapid decay in the eigenvalues of the symmetric matrix $X(\mu)$. A theoretical foundation for this observation can be found in [18, 24] where fast decaying bounds for the eigenvalues of solutions of low rank Lyapunov equations are given. Since a solution $X(\mu)$ of the DARE (1) solves the discrete Lyapunov equation

$$(8) \qquad E(\mu)^T X(\mu) E(\mu) - A_{X(\mu)}^T X(\mu) A_{X(\mu)} = F(X(\mu))^T F(X(\mu)),$$

where $F(X(\mu))^T := \begin{bmatrix} C(\mu)^T Q(\mu)^{1/2} & K_{X(\mu)}^T R(\mu)^{1/2} \end{bmatrix} \in \mathbb{R}^{n \times (p+m)}$, the rapid decay in the eigenvalues carries over. For a derivation of Equation (8) we refer to the appendix.

This low-rank structure allows for an efficient representation of the solution $X(\mu)$ by rectangular matrices $Z(\mu) \in \mathbb{R}^{n \times k}$ such that $X(\mu) \approx Z(\mu) Z(\mu)^T$ with only very little error. All large-scale solvers for DAREs are based on this approximation, and in many real-world scenarios, especially for PDE-constrained optimal control problems, this is typically fulfilled. The low-rank structure can furthermore be employed to construct a basis for the RB-approximation of the DARE. The following algorithm was originally developed for the approximation of solutions to the algebraic Riccati equation (ARE), the continuous-time counterpart to the DARE, see also [21]. The pseudocode of the so-called low-rank factor greedy (LRFG) algorithm is given in Algorithm 1.

Algorithm 1 Low-rank factor greedy algorithm for basis construction

---

Require: Orthonormal initial basis W_0, training set $\mathcal{P}_{train} \subset \mathcal{P}$, tolerance $\varepsilon_{\text{Greedy}}$, POD tolerance ε_{POD}, error indicator $\Delta(W, \mu)$
  Set $W := W_0$
  while $\max\limits_{\mu \in \mathcal{P}_{train}} \Delta(W, \mu) > \varepsilon_{\text{Greedy}}$ **do**
    $\mu^* := \arg \max\limits_{\mu \in \mathcal{P}_{train}} \Delta(W, \mu)$
    Compute the low rank factor $Z(\mu^*)$
    Compute $Z_\perp := (I_n - WW^T) Z(\mu^*)$
    Compute $\bar{Z} = \text{POD}(Z_\perp, \varepsilon_{\text{POD}})$
    Set $W := (W, \bar{Z})$
  end while
  return $W, V := W$

The algorithm works in a typical greedy fashion: starting from an initial basis W_0, it picks out the worst-approximated element by evaluating an error indicator over a finite training set $\mathcal{P}_{\text{train}}$ and chooses the parameter $\mu^* \in \mathcal{P}_{\text{train}}$ that maximizes the error indicator $\Delta(W, \mu)$. The high-dimensional DARE is solved and the solution, or better say parts of it, are added to the basis. This extension procedure is performed in two steps: First, only the part which is perpendicular to the current basis is considered in $Z_\perp$. Then a subsequent reduction of the remaining part is performed by applying a POD to $Z_\perp$ with a prescribed tolerance ε_{POD}, see [21, 25] for details. The basis can furthermore be orthogonalized with respect to the mass matrix $E(\mu)$, where one parameter μ is chosen, such that $W^T E(\mu) V = I_N$, where N is the size of the reduced basis. This can speed up the online calculation of the reduced DARE for certain solvers, such as the builtin MATLAB standard solver.

REMARK 2. *We note that Algorithm 1 is very similar to the POD-Greedy procedure, which is used for the RB approximation of time-dependent problems. In fact, by replacing the low rank factors $Z(\mu^*)$ with the matrix $[x(t_1), \ldots, x(t_{N_T})]$, consisting of full-dimensional solution snapshots at time instances $t_1, \ldots, t_{N_T}$, the algorithms are identical. We refer to [12] for details about the POD-Greedy and to [10] for convergence statements.*

3. Offline-Online-decomposition. One of the central properties to allow for an efficient offline-/online-decomposition is the fast computation of the parameter dependent matrices. This can be achieved by assuming that all matrices can be decomposed in a parameter separable fashion, i.e. there exist parameter dependent coefficient functions and parameter independent components such that

$$A(\mu) = \sum_{q_A=1}^{Q_A} \Theta_A^{q_A}(\mu) A_{q_A}, \quad E(\mu) = \sum_{q_E=1}^{Q_E} \Theta_E^{q_E}(\mu) E_{q_E},$$

$$B(\mu) = \sum_{q_B=1}^{Q_B} \Theta_B^{q_B}(\mu) B_{q_B}, \quad C(\mu) = \sum_{q_C=1}^{Q_C} \Theta_C^{q_C}(\mu) C_{q_C},$$

where a decomposition for $R(\mu)$ and $Q(\mu)$ is not required because of their small sizes. In cases where the above form is not given or does not exist, techniques like the empirical interpolation method can be applied to obtain approximations with a parameter separable structure, see [2, 7]. This property is inherited by the matrices of the reduced problem (7) which can then be assembled rapidly. For example, it holds

$$A_N(\mu) = W^T A(\mu) V = W^T \left(\sum_{q_A=1}^{Q_A} \Theta_A^{q_A}(\mu) A_{q_A} \right) V$$

$$= \sum_{q_A=1}^{Q_A} \Theta_A^{q_A}(\mu) W^T A_{q_A} V = \sum_{q_A=1}^{Q_A} \Theta_A^{q_A}(\mu) A_{N,q_A}.$$

Furthermore, parts of the residual norm and quantities neccessary for the a-posteriori error bounds can be precalculated, as we will see in the next section. All computations in the online step are now independent of the number of degrees of freedom (DOFs) n and only depend on the dimension of the reduced system N, the number of inputs m, the number of outputs p and the number of components in the parameter separable decomposition Q_A, Q_B, Q_C, Q_E. Therefore, these should be preferably small.

4. Error estimation. In this section we want to derive an error estimator for the error $e := \|X - \hat{X}\|$. We therefore apply a more general approximation theory, see [6], which we extend to include an efficiency bound:

THEOREM 1. *Let $(H_1, \|\cdot\|_{H_1})$ and $(H_2, \|\cdot\|_{H_2})$ be two Banach spaces. Furthermore, let $L(H_1, H_2)$ and $L(H_2, H_1)$ denote the space of all linear maps from H_1 to H_2 and H_2 to H_1 respectively. Let $F : H_1 \to H_2$ be continuously differentiable and $v \in H_1$ such that $DF|_v \in L(H_1, H_2)$ is regular. We set*

$$\varepsilon := \|F(v)\|_{H_2}, \quad \gamma := \|DF|_v^{-1}\|_{L(H_2,H_1)}$$

and

$$L(\alpha) := \sup_{x \in B_\alpha(v)} \|DF|_v - DF|_x\|_{L(H_1,H_2)}.$$

6

If the validity criterion

(9) $$\beta := 2\gamma L(2\gamma\varepsilon) \le 1$$

is met, then there exists a unique $u \in B_{2\gamma\varepsilon}(v)$ which satisfies

(10) $$F(u) = 0$$

and for which the following inequalities holds

(11) $$\|u - v\|_{H_1} \le \frac{\gamma\varepsilon}{1 - \beta/2} \le \left(\frac{\beta}{2 - \beta} + \frac{2\gamma}{2 - \beta} \|DF|_v\|_{L(H_1, H_2)} \right) \|u - v\|_{H_1}.$$

Proof. As mentionded above, a proof for the existence of a unique u which satisfies (10) and for which the first inequality in (11) holds can be found in [6]. For the second inequality in (11) we can apply the mean value theorem for Fréchet differentiable operators to F and get

(12) $$\varepsilon = \|F(v)\|_{H_2} = \|F(v) - F(u)\|_{H_2} \le \sup_{x \in B_{2\gamma\varepsilon}(v)} \|DF|_x\|_{L(H_1, H_2)} \|v - u\|_{H_1}.$$

Since the validity criterion (9) is met, it holds $L(2\gamma\varepsilon) = \frac{\beta}{2\gamma}$. From the definition of L we can conclude

$$\sup_{x \in B_{2\gamma\varepsilon}(v)} \|DF|_x\|_{L(H_1, H_2)} \le L(2\gamma\varepsilon) + \|DF|_v\|_{L(H_1, H_2)} \le \frac{\beta}{2\gamma} + \|DF|_v\|_{L(H_1, H_2)}.$$

If we substitute our new found bound in (12) and multiply both sides with $\frac{\gamma}{1 - \beta/2}$, we get

$$\frac{\gamma\varepsilon}{1 - \beta/2} \le \left(\frac{\beta}{2 - \beta} + \frac{2\gamma}{2 - \beta} \|DF|_v\|_{L(H_1, H_2)} \right) \|u - v\|_{H_1}. \qquad \square$$

We now apply Theorem 1 to our reduced basis approximation $\hat{X}$. At first, we need to calculate the Fréchet differential of the residual $\mathcal{R}$ at $\hat{X}$. Therefore, we make use of the identity

(13) $$D\mathcal{R}|_X(S) = \frac{\mathrm{d}}{\mathrm{d}t} \mathcal{R}(X + tS)|_{t=0}$$

and the fact, that the resiudal can be written as a product and sum of Fréchet differentiable operators

(14) $$\mathcal{R}(X) = \mathcal{R}_0(X) + \mathcal{R}_1(X)^T \mathcal{R}_2(X) \mathcal{R}_1(X),$$

where

$$\begin{aligned}
\mathcal{R}_0(X) &:= E^T X E - A^T X A - C^T Q C, \\
\mathcal{R}_1(X) &:= B^T X A, \\
\mathcal{R}_2(X) &:= \left(R + B^T X B \right)^{-1}.
\end{aligned}$$

Since both $\mathcal{R}_0$ and $\mathcal{R}_1$ are affine linear their Fréchet differentials are given by $D\mathcal{R}_0|_X(S) = E^T SE - A^T SA$ and $D\mathcal{R}_1|_X(S) = \mathcal{R}_1(S) = B^T SA$. The Fréchet differential of $\mathcal{R}_2$ can be computed by expanding $\mathcal{R}_2(X + tS)$ in form of Neumann series:

$$
\begin{aligned}
\mathcal{R}_2(X + tS) &= \left(R + B^T XB + tB^T SB\right)^{-1} \\
&= \left(I_m + t\mathcal{R}_2(X)B^T SB\right)^{-1} \mathcal{R}_2(X) \\
&= \left(\sum_{k=0}^{\infty} (-1)^k t^k \left(\mathcal{R}_2(X)B^T SB\right)^k\right) \mathcal{R}_2(X) \\
&= \mathcal{R}_2(X) - t\mathcal{R}_2(X)B^T SB\mathcal{R}_2(X) + \mathcal{O}(t^2) \quad (t \to 0).
\end{aligned}
$$

Applying the identity (13) results in $D\mathcal{R}_2|_X(S) = -\mathcal{R}_2(X)B^T SB\mathcal{R}_2(X)$. Lastly, by using (2) we can write the closed loop matrix as $A_X = A - B\mathcal{R}_2(X)\mathcal{R}_1(X)$ which leads to

$$
\begin{aligned}
D\mathcal{R}|_{\hat{X}}(S) &= D\mathcal{R}_0|_X(S) + D\mathcal{R}_1|_X(S)^T \mathcal{R}_2(X)\mathcal{R}_1(X) + \mathcal{R}_1(X)^T \mathcal{R}_2(X)D\mathcal{R}_1|_X(S) \\
&\quad - \mathcal{R}_1(X)^T \mathcal{R}_2(X)B^T SB\mathcal{R}_2(X)\mathcal{R}_1(X) \\
&= E^T XE - A_X^T XA_X.
\end{aligned}
$$

In the following corollary we now apply Theorem 1 to the DARE to get a first error bound:

COROLLARY 1 (Analytical error bound). *Let $\hat{X}$ be a symmetric positive semidefinite approximation to a solution of the DARE. Furthermore, assume that $(E, A_{\hat{X}})$ is stable and denote by $\mathcal{L}_{\hat{X}} := D\mathcal{R}|_{\hat{X}}$ the discrete Lyapunov operator*

$$
\mathcal{L}_{\hat{X}}(S) = E^T SE - (A_{\hat{X}})^T SA_{\hat{X}}
$$

Set $\gamma := \|\mathcal{L}_{\hat{X}}^{-1}\|$, $\varepsilon := \|\mathcal{R}(\hat{X})\|$ and define

$$
(15) \qquad L(\alpha) := \sup_{Y \in B_\alpha(\hat{X})} \|\mathcal{L}_Y - \mathcal{L}_{\hat{X}}\|,
$$

where $\bar{B}_\alpha(\hat{X}) \subset \mathbb{R}^{n \times n}$ denotes the closed ball with radius α around $\hat{X}$. If the criterion

$$
\beta := 2\gamma L(2\gamma\varepsilon) \leq 1
$$

is met and

$$
(16) \qquad R + B^T \hat{X}B \succ 2\gamma\varepsilon B^T B
$$

holds, then there exists a unique $X^* \in B_{2\gamma\varepsilon}(\hat{X})$ which solves the DARE and which satisfies

$$
\|X^* - \hat{X}\| \leq \frac{\gamma\varepsilon}{1 - \beta/2} \leq \left(\frac{\beta}{2 - \beta} + \frac{2\gamma}{2 - \beta}\|\mathcal{L}_{\hat{X}}\|\right) \|X^* - \hat{X}\|
$$

The additional condition (16) is necessary to guarantee that $\mathcal{R}$ is well defined on $B_{2\gamma\varepsilon}(\hat{X})$. To see this let $Y \in B_{2\gamma\varepsilon}(\hat{X})$, i.e. $\|Y - \hat{X}\| \leq 2\gamma\varepsilon$, then $Y \succeq X - 2\gamma\varepsilon I_n$. With (16) we get $R + B^T YB \succeq R + B^T \hat{X}B - 2\gamma\varepsilon B^T B \succeq 0$ and therefore $Y \in \mathcal{D}_n$.

Since the above corollary only ensures that our approximation is close to a solution X^* of the DARE but not necessarily the stabilizing solution which we want to approximate, the following proposition lays foundation to derive a criterion with which we can verify whether X^* is stabilizing:

PROPOSITION 3 (Stability criterion). *If in addition to the assumptions of* Corollary 1 *the following inequality holds*

$$(17) \qquad \|A_{X^*} - A_{\hat{X}}\| < \sqrt{\|A_{\hat{X}}\|^2 + 1/\gamma} - \|A_{\hat{X}}\|,$$

then X^* is the unique symmetric positive semidefinite and stabilizing solution of the DARE.

Proof. The proof we present here mimics the one given in [23] for the case $E = I$. We therefore prove the following more general result from which (17) can be concluded: Let $A, E \in \mathbb{R}^{n \times n}$, such that (E, A) is stable and let $\mathcal{L}(S) = E^T S E - A^T S A$. If ΔA is a small perturbation in A for which

$$(18) \qquad \|\Delta A\| < \sqrt{\|A\|^2 + 1/\gamma} - \|A\|$$

holds, where $\gamma = \|\mathcal{L}^{-1}\|$, then $(E, A + \Delta A)$ is stable. We define

$$\delta := \min\{\|\Delta A\| : \Delta A \in \mathbb{C}^{n \times n}, \max_{1 \le j \le n} |\lambda_j(E, A + \Delta A)| = 1\}.$$

Let $\Delta A^* \in \mathbb{C}^{n \times n}$ such that $\delta = \|\Delta A^*\|$ and w.l.o.g. $\lambda_1 = \arg \max_{1 \le j \le n} |\lambda_j(E, A + \Delta A^*)|$ and let v_1 denote an eigenvector of unit length to the eigenvalue λ_1 of the adjoint eigenvalue problem $\lambda_1 E^T v_1 = (A + \Delta A^*)^T v_1$. Then $W := v_1 v_1^*$ satisfies

$$E^T W E - (A + \Delta A^*)^T W (A + \Delta A^*) = E^T W E - \lambda_1 E^T v_1 (\lambda_1 E^T v_1)^* = 0$$

which is equivalent to

$$\mathcal{L}(W) = A^T W \Delta A^* + (\Delta A^*)^T W A + (\Delta A^*)^T W (\Delta A^*).$$

Applying the inverse operator $\mathcal{L}^{-1}$ and taking the norm leads to

$$1 \le \gamma \left(2\|A\|\delta + \delta^2 \right).$$

Solving for δ results in $\sqrt{\|A\|^2 + 1/\gamma} - \|A\| \le \delta$ which gives us (18). Finally, by setting $A = A_{\hat{X}}$ and $\Delta A = A_{X^*} - A_{\hat{X}}$ we can conclude (17). □

COROLLARY 2 (Computable error bound). *Let the assumptions of Corollary 1 hold. Furthermore, assume upper bounds $\gamma \le \gamma_N$ and $\varepsilon \le \varepsilon_N$ are available. It follows:*

(a) *If $R + B^T \hat{X} B \succ \alpha B^T B$, then $L(\alpha)$ is bounded by $L_N(\alpha)$, where $L_N(\alpha)$ is given by*

$$L_N(\alpha) = \alpha \left\| B \left(R + B^T \hat{X} B - \alpha B^T B \right)^{-1} B^T \right\|$$
$$\cdot \left(2 + \alpha \left\| \cdot B^T \left(R + B^T \hat{X} B - \alpha B^T B \right)^{-1} B \right\| \right) \|A_{\hat{X}}\|^2.$$

(b) *If*

$$\beta_N := 2\gamma_N L_N(2\gamma_N \varepsilon_N) \le 1$$

and

$$R + B^T \hat{X} B \succ 2\gamma_N \varepsilon_N B^T B,$$

then there exists a unique $X \in B_{2\gamma_N \varepsilon_N}(\hat{X})$, which solves the DARE and for which the following error and effectivity bound holds

$$\|X - \hat{X}\| \le \Delta_X := \frac{\gamma_N \varepsilon_N}{1 - \beta_N/2} \le \left(\frac{\beta_N}{2 - \beta_N} + \frac{2\gamma_N}{2 - \beta_N} \|\mathcal{L}_{\hat{X}}\| \right) \|X - \hat{X}\|.$$

(c) If in addition to 2.

$$\alpha_N \left\| B \left(R + B^T \hat{X} B - \alpha_N B^T B \right)^{-1} B^T \right\| \|A_{\hat{X}}\| < \sqrt{\|A_{\hat{X}}\|^2 + 1/\gamma_N} - \|A_{\hat{X}}\|,$$

where $\alpha_N = 2\gamma_N \varepsilon_N$, then X is the unique symmetric positive semi-definite stabilizing solution of the DARE.

Proof. For part (a) let $X, Y \in \mathcal{D}_n$ and $S \in \mathcal{S}_n$. Then it holds

$$\begin{aligned}
\mathcal{L}_Y(S) - \mathcal{L}_X(S) &= A_Y^T S A_Y - A_X^T S A_X \\
&= \left(A_Y^T - A_X^T \right) S A_Y - A_X^T S \left(A_X - A_Y \right)
\end{aligned}$$

and therefore

(19) $$\|\mathcal{L}_Y - \mathcal{L}_X\| \le \|A_Y - A_X\| \left(\|A_Y\| + \|A_X\| \right).$$

Taking a closer look at how the individual closed-loop matrices relate to one another and using the identity

$$R_X^{-1} - R_Y^{-1} = R_Y^{-1}(R_Y - R_X)R_X^{-1} = R_Y^{-1}B^T(Y - X)BR_X^{-1},$$

where we denote $R_Y := R + B^T Y B$, we obtain

$$\begin{aligned}
A_X - A_Y &= A - BR_X^{-1}B^T X A - \left(A - BR_Y^{-1}B^T Y A \right) \\
&= B \left(R_Y^{-1}B^T Y - R_X^{-1}B^T X \right) A \\
&= B \left(R_Y^{-1}B^T(Y - X) - \left(R_X^{-1} - R_Y^{-1} \right) B^T X \right) A \\
&= B \left(R_Y^{-1}B^T(Y - X) - R_Y^{-1}B^T(Y - X)BR_X^{-1}B^T X \right) A \\
&= BR_Y^{-1}B^T(Y - X) \left(I_n - BR_X^{-1}B^T X \right) A \\
&= BR_Y^{-1}B^T(Y - X)A_X.
\end{aligned}$$

We can now derive the following upper bounds for the matrix norms

(20) $$\|A_X - A_Y\| \le \|B \left(R + B^T Y B \right)^{-1} B^T\| \|Y - X\| \|A_X\|,$$

(21) $$\|A_Y\| \le \left(1 + \|B \left(R + B^T Y B \right)^{-1} B^T\| \|Y - X\| \right) \|A_X\|.$$

In the case of $X = \hat{X}$ it holds $\|Y - \hat{X}\| \le \alpha$, which implies $Y \succeq \hat{X} - \alpha I_n$ and therefore

(22) $$\|B \left(R + B^T Y B \right)^{-1} B^T\| \le \|B \left(R + B^T \hat{X} B - \alpha B^T B \right)^{-1} B^T\|.$$

Combining (19),(20), (21) and (22) we get the proposed upper bound $L(\alpha) \le L_N(\alpha)$.

For part (b) we note, that since

$$\beta = 2\gamma L(2\gamma\varepsilon) \le 2\gamma_N L(2\gamma_N \varepsilon_N) = \beta_N$$

the first inequality in Theorem 1 still holds.

Part (c) is a direct consequence of Proposition 3. In the proof of part one we have already seen that the left hand side poses an upper bound to $\|A_Y - A_{\hat{X}}\|$. On the other hand we can conclude from the alternative representation

$$\sqrt{\|A_{\hat{X}}\|^2 + 1/\gamma} - \|A_{\hat{X}}\| = \frac{1/\gamma}{\sqrt{\|A_{\hat{X}}\|^2 + 1/\gamma} + \|A_{\hat{X}}\|}$$

that the right hand side is decreasing in both γ and $\|A_{\hat{X}}\|$ and therefore

$$\sqrt{\|A_{\hat{X}}\|^2 + 1/\gamma_N} - \|A_{\hat{X}}\| \leq \sqrt{\|A_{\hat{X}}\|^2 + 1/\gamma} - \|A_{\hat{X}}\|.$$

Thus, X is stabilizing by Proposition 3 if part three holds. □

4.1. About the inverse operator norm $\|\mathcal{L}_{\hat{X}}^{-1}\|$. The efficient calculation of the constant $\gamma = \|\mathcal{L}_{\hat{X}}^{-1}\|$ or an upper bound $\gamma \leq \gamma_N$ is of utmost importance for the calculation of the stability criterion and the error bound. Therefore, we will dedicate this section to a more detailed analysis concerning this operator, including rigorous and nonrigorous approximations to γ allow the online-efficient calculation of the error bound.

THEOREM 2. *Let $A_{\hat{X}}, E \in \mathbb{R}^{n \times n}$ where E is regular and $(E, A_{\hat{X}})$ is stable. Let $\mathcal{L}_{\hat{X}} : \mathcal{S}_n \to \mathcal{S}_n$ denote the discrete Lyapunov operator, defined as $\mathcal{L}_{\hat{X}}(S) := E^T SE - A_X^T S A_X$ for $S \in \mathbb{R}^{n \times n}$. Then the following holds:*

1. *$\mathcal{L}_{\hat{X}}$ is invertible and the inverse operator $\mathcal{L}_{\hat{X}}^{-1}$ is given via*

$$\text{(23)} \qquad \mathcal{L}_{\hat{X}}^{-1}(S) = E^{-T} \left(\sum_{k=0}^{\infty} (A_{\hat{X}}^T E^{-T})^k S (E^{-1} A_{\hat{X}})^k \right) E^{-1}.$$

2. *Let $H \in \mathcal{S}_n$ be the unique solution to $\mathcal{L}_{\hat{X}}(H) = I_n$, then the operator norm $\gamma = \|\mathcal{L}_{\hat{X}}^{-1}\|$ of the inverse operator is given by*

$$\text{(24)} \qquad \gamma = \|\mathcal{L}_{\hat{X}}^{-1}\| = \|H\|.$$

Proof. For the first property we note, that since (E, A) is stable, there exists a norm $\|\cdot\|_\star$, such that $\|E^{-1}A\|_\star < 1$ and the series (23) is absolutely convergent with respect to $\|\cdot\|_\star$. One can now easily check, that the right hand side in (23) is the inverse. For the second result, we refer to [8], where the proof is carried out for the case $E = I_n$. The generalization towards systems with non-trivial mass matrices, however, is straightforward. □

Theorem 2 shows how the true norm of the inverse operator can be calculated explicitly: One has to solve a discrete Lyapunov equation $E^T H E - A_{\hat{X}}^T H A_{\hat{X}} = I_n$ and calculate the norm of the solution $\|H\|$. As this might be feasible for small systems (say $n < 500$), it is not possible to solve this Lyapunov equation efficiently for large-scale systems. This is due to the lack of any low-rank structure in the equation, which renders efficient algorithms for solving large and sparse Lyapunov equations inefficient or unapplicable. We thus have to find suitable approximations or efficient techniques to obtain upper bounds $\gamma \leq \gamma_N$. A very simple bound can be found by exploiting the series-representation of the solutions to the Lyapunov equation:

COROLLARY 3 (Upper bound). *Let the assumptions of Theorem 2 hold. If* $\|E^{-1}A_{\hat{X}}\| < 1$ *then*

$$\gamma_N = \frac{\|E^{-1}\|^2}{1 - \|E^{-1}A_{\hat{X}}\|^2}$$

is an upper bound to γ.

Proof. The proof is based on the representation of the solution H of $\mathcal{L}_{\hat{X}}(H) = I_n$. It holds

$$\gamma = \|H\| \leq \|E^{-1}\|^2 \sum_{k=0}^{\infty} \|E^{-1}A_{\hat{X}}\|^{2k} = \frac{\|E^{-1}\|^2}{1 - \|E^{-1}A_{\hat{X}}\|^2}. \qquad \square$$

The above bound in Corollary 3 can be very pessimistic, especially when the closed-loop system $Ex^{k+1} = A_{\hat{X}}x^k$ has a weak damping in the sense that the system norm $\|E^{-1}A_{\hat{X}}\|$ is close to 1.

4.1.1. Approximation by power iteration. If we again recall the identity $\gamma = \|H\|$, where H solves the Lyapunov equation $\mathcal{L}_{\hat{X}}(H) = I_n$, we see that we are only interested in the largest eigenvalue of the symmetric matrix $H = \mathcal{L}_{\hat{X}}^{-1}(I_n)$. This opens a different way to approximate γ efficiently, without solving the (very expensive) Lyapunov equation: We can apply the power iteration method (PI), which only relies on products of the form Hx. The PI approximates the largest eigenvalue and eigenvector by subsequent multiplications of a vector with H. Given an suitable start vector $x_0 \in \mathbb{R}^n$ with $\|x_0\| = 1$, we proceed in the following way:

$$\tilde{x}_{k+1} := Hx_k, \quad \lambda_k := \tilde{x}_{k+1}^T x_k, \quad x_{k+1} := \tilde{x}_k / \|\tilde{x}_k\|, \quad k = 1, 2, \ldots.$$

It is known, that this algorithm converges to the dominant eigenvalue of the matrix H, whenever the initial vector is not orthogonal to the eigenspace spanned by the dominant eigenvector. Thus, the goal is to calculate the matrix-vector product Hx efficiently. For that purpose, we make use of the series representation:

$$Hx = E^{-T}\left[\sum_{k=0}^{\infty}(A_{\hat{X}}^T E^{-T})^k (E^{-1}A_{\hat{X}})^k\right]E^{-1}x.$$

Each summand in the series can be calculate efficiently: First we define $y_0 := E^{-1}x$, and subsequently calculate $y_k := (E^{-1}A_{\hat{X}})y_{k-1}$ for $k = 1, 2, \ldots$. Furthermore, we set $h_k := (A_{\hat{X}}^T E^{-T})^k y_k$ and $H_k := E^{-1}\sum_{i=0}^{k} h_i$. It then holds $Hx = \lim_{k\to\infty} H_k$. We truncate the series after ℓ terms, where ℓ is determined by the criterion $\|h_k\|/\|H_k\| \leq$ tol, thus limiting the number of terms by neglecting terms with only low increment in the result. Note that this is always possible, since the closed-loop system $(E, A_{\hat{X}})$ is stable. Note that the procedure can be implemented efficiently when using techniques like LU-decomposition of E to avoid explicit inversion of E and by applying parallelization techniques. Numerical examples of the performance of this procedure are given in Section 5.

4.1.2. Approximation via projection. Another way to obtain estimates of γ is to once again apply a projection to the Lyapunov Equation

$$(25) \qquad \mathcal{L}_{\hat{X}}(H) = E^T HE - A_{\hat{X}}^T HA_{\hat{X}} = I_n$$

to get an approximation $\hat{H}$ of H. It is very unlikely that any low-rank approximation is able to accurately represent the full solution H, since (25) has no low-rank structure that can be exploited. However, recalling Theorem 2 we can expect, that

$$\hat{\gamma} := \|\hat{H}\| \approx \|H\| = \|\mathcal{L}_{\hat{X}}^{-1}\| = \gamma,$$

is a good approximation whenever the dominant eigenvalue of H is captured well by the approximation $\hat{H}$. We therefore assume, that $\hat{H}$ can be written as $\hat{V} H_M \hat{V}^T$, where $\hat{V} \in \mathbb{R}^{n \times M}$ has orthonormal columns and $H_M \in \mathcal{S}_M$. This leads to the reduced Lyapunov Equation

$$E_M^T H_M E_M - A_{\hat{X},M}^T H_M A_{\hat{X},M} = I_M,$$

where $E_M := \tilde{V}^T E \tilde{V}$ and $A_{\hat{X},M} = \tilde{V}^T A_{\hat{X}} \tilde{V}$. As seen in Subsection 4.1.1, we can employ the power iteration method to approximate the largest eigenvalue of H and a corresponding eigenvector v. A suitable basis $\tilde{V}$ can now be constructed from dominant eigenvectors $\{v(\mu_i)\}_{i=1}^M$ where a training set $\mathcal{P}_M := \{\mu_1, \dots, \mu_M\} \subset \mathcal{P}$ can be chosen arbitrarily. This method is online efficient, since the above computations can be done during the offline phase. We present numerical studies for this procedure in Section 5.

4.2. Online efficient norm calculation. We now shift our focus to the online-efficient calculation of the norms $\left\| B \left(R + B^T \hat{X} B - \alpha B^T B \right)^{-1} B^T \right\|$, $\|A_{\hat{X}}\|$ and $\varepsilon = \|\mathcal{R}(\hat{X})\|$, or upper bounds to these quantities.

First, we take a closer look at

$$H(\hat{X}) := B \left(R + B^T \hat{X} B - \alpha B^T B \right)^{-1} B^T$$

which is symmetric and therefore satisfies $\|H(\hat{X})\| = \lambda_{max}(H(\hat{X}))$. Since the spectral radius of a product of matrices remains identical, when the order is reversed, i.e. it holds

$$\|H(\hat{X})\| = \lambda_{max}(H(\hat{X})) = \lambda_{max}\left(\left(R + B^T \hat{X} B - \alpha B^T B \right)^{-1} B^T B \right),$$

we can compute the norm $\|H(\hat{X})\|$ by determining the maximum eigenvalue of the low dimensional matrix $\left(R + B^T \hat{X} B - \alpha B^T B \right)^{-1} B^T B$.

For $A_{\hat{X}}$ and $\mathcal{R}(\hat{X})$ instead of the spectral norm $\|\cdot\|$ we compute the Frobenius norm $\|\cdot\|_F$, as proposed in [21], which states an upper bound for the Euclidean norm. It holds for all $A \in \mathbb{R}^{n \times n}$

$$(26) \qquad \|A\|_2 \le \|A\|_F \le \sqrt{\mathrm{rank}(A)} \|A\|_2.$$

If we apply (26) to the residual $\mathcal{R}(\hat{X})$ we get

$$\mathrm{rank}(\mathcal{R}(\hat{X})) = \mathrm{rank}\left(E^T \hat{X} E - \left(A^T - A^T \hat{X} B \left(R + B^T \hat{X} B \right)^{-1} B^T \right) \hat{X} A - C^T Q C \right)$$

$$\le \mathrm{rank}(E^T \hat{X} E) + \mathrm{rank}(A_{\hat{X}}^T \hat{X} A) + \mathrm{rank}(C^T Q C)$$

$$\le N + N + p = 2N + p.$$

Therefore the Frobenius norm overestimates the spectral norm by a factor at most $\sqrt{2N+p}$. To make use of the parameter separability of the system matrices we calculate $\|\mathcal{R}(\hat{X})\|_F$, via the identity $\|\mathcal{R}(\hat{X})\|_F^2 = \text{tr}(\mathcal{R}(\hat{X})^T\mathcal{R}(\hat{X}))$ which results in

$$\begin{aligned}
(27) \quad \|\mathcal{R}(\hat{X})\|_F^2 = &\ \text{tr}(E^T\hat{X}EE^T\hat{X}E) + \text{tr}(A^T\hat{X}AA^T\hat{X}A) + \text{tr}(C^TQCC^TQC) \\
&+ \text{tr}(A^T\hat{X}BR_{\hat{X}}^{-1}B^T\hat{X}AA^T\hat{X}BR_{\hat{X}}^{-1}B^T\hat{X}A) \\
&+ 2\,\text{tr}(C^TQCA^T\hat{X}A) - 2\,\text{tr}(E^T\hat{X}EC^TQC) \\
&+ 2\,\text{tr}(A^T\hat{X}BR_{\hat{X}}^{-1}B^T\hat{X}AE^T\hat{X}E) \\
&- 2\,\text{tr}(E^T\hat{X}EA^T\hat{X}A) - 2\,\text{tr}(A^T\hat{X}BR_{\hat{X}}^{-1}B^TR_{\hat{X}}^{-1}\hat{X}AA^T\hat{X}A) \\
&- 2\,\text{tr}(C^TQCA^T\hat{X}BR_{\hat{X}}^{-1}B^T\hat{X}A),
\end{aligned}$$

where we used the abreviation $R_{\hat{X}} = \left(R + B^T\hat{X}B\right)$. Substituting $\hat{X} = WX_NW^T$ and applying the identity $\text{tr}(ST) = \text{tr}(TS)$ we can restructure the matrices in each term of (27) in such a way, that the argument of the trace operator can be written as a product of low dimensional matrices most of which posses a parameter separable decomposition. For example,

$$\text{tr}(E^T\hat{X}EE^T\hat{X}E) = \text{tr}(X_N\underline{W^TEE^TW}X_NW^TEE^TW),$$

where

$$W^TE(\mu)E^T(\mu)W = \sum_{i,j=1}^{Q_E} \Theta_E^i(\mu)\Theta_E^j(\mu)\underbrace{W^TE_iE_j^TW}_{=:M_{i,j}}.$$

In a similar fashion the matrix $R_{\hat{X}} = R + B^T\hat{X}B = R + B_N^TX_NB_N$ can be computed fast and the following inversion is comparably cheap because $R_{\hat{X}} \in \mathbb{R}^{m \times m}$ is low dimensional. Overall, the complexity of the calculation of the Frobenius norm during an online simulation only depends on the dimension N of the reduced system and the number of terms in the parameter separable extension Q_A, Q_B, Q_C, Q_E. The norm $\|A_{\hat{X}}\|$ can be calculated in an analogous fashion.

5. Numerical Examples. We now investigate the proposed application of the LRFG algorithm to the DARE. All examples were calculated on a computer with 4 dual-core Intel Core i7-6700 CPUs with 3.40Ghz each, 16 GB RAM and were implemented in the MATLAB toolbox RBmatlab, where we used version 1.16.09. Wherever possible, we made use of parallelization techniques built into MATLAB. The MATLAB version used for all examples is R2015b. The full dimensional DAREs are solved by using a Newton iteration, where the Lyapunov-equation in each iteration is solved by using a code that was thankfully provided by Patrick Kürschner, see also [5].

The model under consideration is an advection-diffusion equation on the unit square $\Omega := [0,1]^2$ with homogeneous Dirichlet boundary conditions on all edges and distributed control input on the subdomain $\Omega_B := [0.6, 0.8] \times [0.4, 0.6]^2$. The PDE for this example is given by

$$\partial_t w(t,\xi;\mu) - \mu_{\text{diff}}\Delta w(t,\xi;\mu) + \mu_{\text{adv}}\partial_x w(t,\xi;\mu) = \mathbf{1}_{\Omega_B}(\xi)u(t), \quad t \geq 0, \xi \in \Omega,$$

together with the boundary condition $w(t,\xi;\mu) = 0$ on the boundary $\Gamma := \partial\Omega$ and zero initial conditions $w(0,\cdot;\mu) = 0$. The function $\mathbf{1}_{\Omega_B}$ denotes the indicator function of

Ω_B. We furthermore define a measurement output as $s(t; \mu) := \frac{1}{|\Omega_C|} \int_{\Omega_C} w(t, \xi; \mu) \mathrm{d}\xi$ for $t \geq 0$ where the measurement domain is chosen as $\Omega_C := [0, 0.1] \times [0.2, 0.8]$, see Figure 1 for a picture of the setup. The control objective for this problem is to let the output $s(t; \mu)$ follow a prescribed reference trajectory $r(t)$. We obtain a fully discrete difference equation for this model after semidiscretization of the PDE by using finite differences in space on an equidistant grid in both dimensions, where an upwind scheme is used for the discretization of the advection term and the boundary values are not included in the solution vector, followed by the application of an implicit Euler scheme with timestep Δt for the temporal discretization. We end up with the following discrete-time LTI system:

$$E(\mu)x_{k+1}(\mu) = x_k(\mu) + \Delta t B u_k, \quad k \geq 0,$$
$$y_k(\mu) = C x_k(\mu), \quad k \geq 0,$$
$$x_0 = 0.$$

The mass matrix stems from the implicit discretization and takes the form $E(\mu) := (I_n - \mu_{\mathrm{diff}} \Delta t A_1 - \mu_{\mathrm{adv}} \Delta t A_2)$, where the matrix A_1 describes the diffusive part and A_2 adds the advection part to the equation. The matrix B is the discretized counterpart of the indicator function and the output is spatially discretized by using a rectangular quadrature rule, which results in the linear equation for $y_k(\mu)$.

In order to formulate the tracking-control task stated above, we employ a linear-quadratic regulator technique with an additional integral-action for the tracking: We first discretize the reference trajectory as $r_k := r(k\Delta t)$ for $k \geq 0$. We then add an additional artifical state $\tilde{x}_k$ that sums up (integrates) the error between the desired trajectory and the measurement output $\tilde{x}_{k+1} := \tilde{x}_k + \Delta t(y_k - r_k)$. The augmented system can then be written as

$$\begin{pmatrix} E(\mu) & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_{k+1} \\ \tilde{x}_{k+1} \end{pmatrix} = \begin{pmatrix} I & 0 \\ \Delta t C & 1 \end{pmatrix} \begin{pmatrix} x_k \\ \tilde{x}_k \end{pmatrix} + \begin{pmatrix} \Delta t B \\ 0 \end{pmatrix} u_k - \begin{pmatrix} 0 \\ \Delta t \end{pmatrix} r_k,$$
$$y_k(\mu) = C x_k.$$

We introduce the cost functional

$$J(\mathbf{x}, \mathbf{u}; \mu) := \sum_{k=0}^{\infty} \mu_{tracking} \tilde{x}_{k+1}^2 + 0.1 u_k^2,$$

which, when minimized, forces $\tilde{x}_k$ to zero and hence results in the tracking $C x_k \to r_k$ for $k \to \infty$ whenever r_k is constant in time. The parameter $\mu_{tracking}$ can be adjusted to alter the tracking quality/speed: Larger values lead to a faster tracking but involve higher control costs. To sum up, the problem consists of three parameters, and the parameter domain is chosen as

$$\mu = (\mu_{\mathrm{diff}}, \mu_{\mathrm{adv}}, \mu_{tracking})^T \in \mathcal{P} := [0.05, 0.2] \times [0, 10] \times [1, 10].$$

The full discretization in the following examples results in a $n = 1601$ dimensional system with one input $m = 1$ and one output $p = 1$. An example of the full state and the controlled output $y_k(\mu)$ for two different parameters is provided in Figure 2(left). The reference output trajectory in this case is a rectangular signal that changes sign at $t = 5$ which both outputs nicely follow. From Figure 2(right) we can conclude that the exact output y_k and the approximated output for dimension $N = 6$ of the reduced problem are basically indistinguishable.
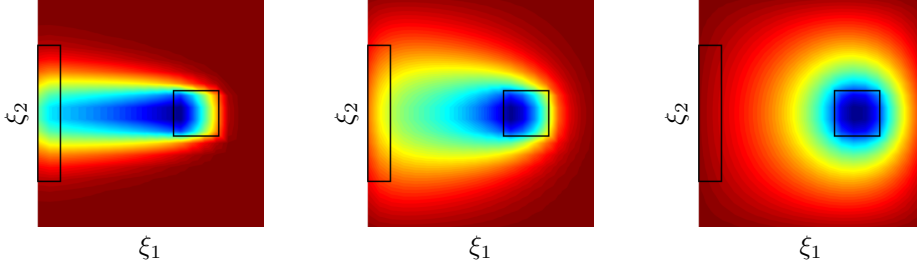
FIG. 1. *Full 2D state plots for different parameter vectors (from left to right):* $\mu = (0.05, 5, 10)^T$, $(0.05, 1, 10)^T$, $(0.2, 0, 10)^T$ *at* $t = 10$.
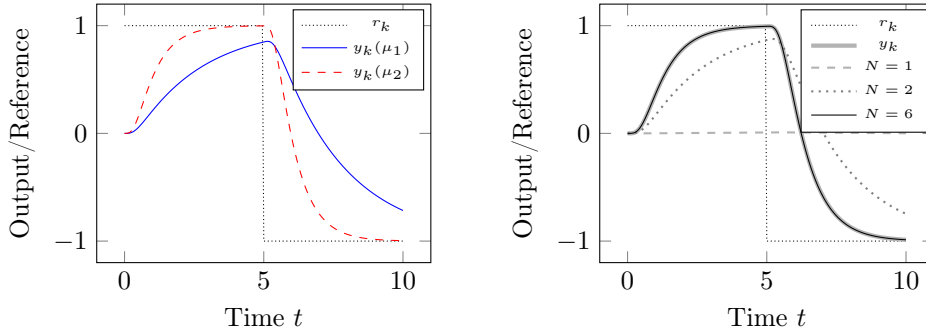

FIG. 2. *Reference trajectory and two (true) output plots for parameters* $\mu_1 = (0.2, 1, 10)^T$, $\mu_2 = (0.2, 1, 10)^T$ *(left). Reference trajectory, output* $y_k(\mu)$ *(right) for* $\mu = (0.1, 1, 1)^T$ *and outputs for increasing dimension* N *of the reduced problem.*

As a first test we investigate the basis generation procedure for this example. For this purpose, we discretize the parameter domain into 343 equidistant points as nodes from a uniform grid in all three dimensions. We set the desired greedy tolerance to $\varepsilon_{\text{Greedy}} := 10^{-4}$ and run the basis generation Algorithm 1 with different POD tolerances. In Figure 3, a qualitative comparison is given in terms of the decay of the error indicator. The marks in the plot indicate the points where full solutions are calculated during the run of the greedy algorithm. The numerical results confirm the expected behavior: The smallest basis is being constructed for $\varepsilon_{\text{POD}} = 0$ with a size of $N = 43$. Setting the POD tolerance to zero means that in each iteration only the dominant mode is added to the basis, which naturally results in a very compressed basis, but requires the highest computational effort since many high dimensional DARE solutions must be calculated: In this case 40, which means that 3 parameters got revisited during the greedy iterations. Increasing the POD tolerance to $\varepsilon_{\text{POD}} > 0.99$ leads to larger bases for this example, but allows a faster basis generation since fewer full solutions are required. Clearly, there is a tradeoff between the desired basis size and the number of full solutions that one can calculate offline. In all of the following examples we choose the smallest basis, i.e. the $N = 43$ dimensional basis which was constructed for $\varepsilon_{\text{POD}} = 0$. We furthermore define another parameter set $\mathcal{P}_{\text{test}}$ for all tests, consisting of 10 randomly chosen elements from $\mathcal{P}$, distinct to the training set $\mathcal{P}_{\text{train}}$.

Next, we analyze the true error which is induced by the RB approximation and the corresponding error estimation from Section 4. In Figure 3 we show the true
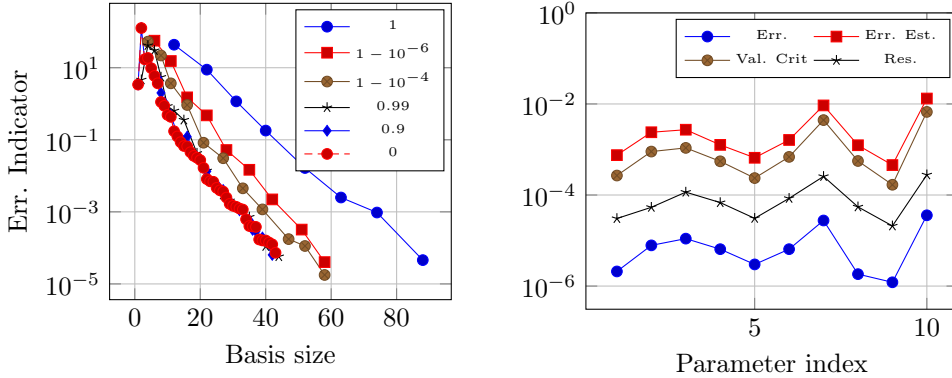
16

Fig. 3. *Error indicator decay for different inner tolerances ε_{POD} on the same training set (left). True error, residual norm, validity criterion and error estimator (right) for all test parameters.*

absolute error, the error estimator and the residual as well as the validity criterion. Note that, although its calculation is really expensive, we choose the true value for the constants $\gamma(\mu)$ in order to get rigorous results and validate the bounds. We will later show that by using the approaches presented in this work, the constant can be estimated much faster. Figure 3 reveals that the error estimator resamples the true behavior of the error very accurately, although it is overestimated by a factor of $\approx 10^2$. However, as Figure 2 (right) indicates, the outputs are indistinguishable which proves the quality of the RB approximation for this application. In all cases, the validity criterion is below 1, and thus all results are rigorous.

The crucial ingredient in the error estimation is clearly the constant $\gamma(\mu)$. Theorem 2 shows that it can be calculated by solving a large-scale discrete Lyapunov equation, which clearly is infeasible online. We thus proposed two methods to overcome this issue and to enable real-time simulations: The first method is an efficient implementation of the power iteration (PI) algorithm to approximate the largest eigenvalue of the Lyapunov-solution H. We perform this procedure for all elements in the test set. Table 1 shows the mean relative approximation error $\text{mean}_{\mu \in \mathcal{P}_{\text{test}}} \frac{|\gamma(\mu) - \gamma_{\text{PI}}(\mu)|}{\gamma(\mu)}$ between the true value $\gamma(\mu)$ and the PI-approximation $\gamma_{\text{PI}}(\mu)$ as well as the corresponding calculation times for different tolerances ε_{PI}. We see that even in relativly low dimensions, the direct calculation easily becomes infeasible. In really high dimensions, say $n \gg 10.000$ this is infeasible. However, we can use the power-iteration algorithm to approximate $\gamma(\mu)$ up to any desired tolerance in (offline-)reasonable time, as the results in Table 1 indicate.

TABLE 1

Comparison of the calculation times for γ when using the full-approach (Lyapunov-equation solve) and the power-iteration method for different tolerances ε_{PI}.

	n=101		n=401		n=1601	
	t[s]	Err	t[s]	Err	t[s]	Err
Full	0.02	-	0.58	-	65.29	-
ε_{PI}=1e-2	0.02	1.40e-02	0.06	2.52e-02	0.48	5.07e-02
ε_{PI}=1e-3	0.03	1.91e-03	0.13	2.72e-03	1.37	5.34e-03
ε_{PI}=1e-4	0.06	1.86e-04	0.23	2.63e-04	3.02	5.71e-04

As explained in Subsection 4.1.1, $\gamma(\mu)$ can furthermore be approximated by projecting the Lyapunov-equation onto a suitable subspace, spanned by dominant eigenvectors of its solution. Those eigenvectors can be obtained "cheaply" by using the power-iteration technique as before. Table 2 gives an overview of the results. With moderate offline costs, very accurate approximations can be obtained online. Compared to the huge costs that are required for one full calcualtion of $\gamma(\mu)$, this is a tremendous improvement.

TABLE 2

Results for the projection-based calculation of $\gamma(\mu)$. The table shows the calculation times for the offline step, average online calculation time and the mean, minimum and maximum relative error over the test set $\mathcal{P}_{test}$ for different offline training sizes.

	8	27	64	125
Offline time [s]	130.7669	257.8988	463.0604	761.6805
Avg. Online time [s]	0.0114	0.0106	0.0100	0.0092
mean	0.0577	0.0028	0.0049	0.0014
min	0.0081	0.0001	0.0001	0.0000
max	0.1265	0.0064	0.0148	0.0053

Finally, we examine the calculation times of the overall procedure in Table 3. One online simulation for this example takes only a fraction of a second, including the calculation of the residual norm and the 2-norm of the closed-loop system. Altogether, we reach speedup factors in the magnitude of several thousands. We note however that the implementation of the large-scale DARE solver used in out experiments is certainly not the most efficient implementation possible.

TABLE 3

Calculation times in seconds for one full-dimensional solution, the calculation of the reduced solution, its residual norm and the closed-loop norm.

	P	P_N	$\|\mathcal{R}(\hat{X})\|_F$	$\|A_{\hat{X}}\|$
mean	18.6486	0.0054	0.0005	0.0116
max	21.7083	0.0074	0.0008	0.0738
min	16.5895	0.0043	0.0004	0.0041

6. Conclusion. In this article we developed the RB-DARE framework for the rapid approximation of solutions to the DARE. We showed how the whole reduction procedure can be implemented efficiently, and how the resulting approximation can be certified by a-posteriori error bounds. The bounds can be calculated quickly, given rapid approximations of the relevant constants are available. The presented application in the field of LQR control is only one example of many possible scenarios, including for example discrete-time $\mathcal{H}_\infty$, $\mathcal{H}_2$ control or Kalman-filtering.

References.

[1] M. ATHANS, *The role and use of the stochastic linear-quadratic-Gaussian problem in control system design*, IEEE Transactions on Automatic Control, 16 (1971),

pp. 529–552, doi:10.1109/tac.1971.1099818, http://dx.doi.org/10.1109/tac.1971.1099818.

[2] M. Barrault, Y. Maday, N. Nguyen, and A. Patera, *An 'empirical interpolation' method: application to efficient reduced-basis discretization of partial differential equations*, CR Acad Sci Paris Series I, 339 (2004), pp. 667–672.

[3] H. Benner, P. & Fassbender, *On the numerical solution of large-scale sparse discrete-time Riccati equations*, Advances in Computational Mathematics, (2011).

[4] P. Benner, S. Gugercin, and K. Willcox, *A survey of projection-based model reduction methods for parametric dynamical systems*, SIAM Rev., 57 (2015), pp. 483–531, doi:10.1137/130932715, http://dx.doi.org/10.1137/130932715.

[5] P. Benner and P. Kürschner, *Computing Real Low-Rank Solutions of Sylvester Equations by the Factored ADI Method*, Computers & Mathematics with Application, 67 (2014).

[6] G. Caloz and J. Rappaz, *Numerical analysis for nonlinear and bifurcation problems*, in Techniques of Scientific Computing (Part 2), vol. 5 of Handbook of Numerical Analysis, Elsevier, 1997, pp. 487 – 637, doi:http://dx.doi.org/10.1016/S1570-8659(97)80004-X, http://www.sciencedirect.com/science/article/pii/S157086599780004X.

[7] M. Drohmann, B. Haasdonk, and M. Ohlberger, *Reduced basis approximation for nonlinear parametrized evolution equations based on empirical operator interpolation*, SIAM J. Sci. Comput., 34 (2012), pp. A937–A969, doi:10.1137/10081157X, arXiv:http://epubs.siam.org/doi/pdf/10.1137/10081157X.

[8] P. Gahinet, A. Laub, C. Kenney, and G. Hewer, *Sensitivity of the stable discrete-time Lyapunov equation*, IEEE Transactions on Automatic Control, 35 (1990), pp. 1209–1217, doi:10.1109/9.59806, http://dx.doi.org/10.1109/9.59806.

[9] M. Grepl and A. Patera, *A posteriori error bounds for reduced-basis approximations of parametrized parabolic partial differential equations*, ESAIM: Mathematical Modelling and Numerical Analysis, 39 (2005), pp. 157–181, doi:10.1051/m2an:2005006, http://dx.doi.org/10.1051/m2an:2005006.

[10] B. Haasdonk, *Convergence rates of the POD–Greedy method*, ESAIM: Mathematical Modelling and Numerical Analysis, 47 (2013), pp. 859–873, doi:10.1051/m2an/2012045, http://dx.doi.org/10.1051/m2an/2012045.

[11] B. Haasdonk, *Reduced basis methods for parametrized PDEs – a tutorial introduction for stationary and instationary problems*, SimTech preprint, IANS, University of Stuttgart, Germany, 2014, http://www.simtech.uni-stuttgart.de/publikationen/prints.php?ID=938. Chapter in P. Benner, A. Cohen, M. Ohlberger and K. Willcox (eds.): "Model Reduction and Approximation: Theory and Algorithms", SIAM, Philadelphia, 2017.

[12] B. Haasdonk and M. Ohlberger, *Reduced basis method for finite volume approximations of parametrized linear evolution equations*, ESAIM: M2AN, 42 (2008), pp. 277–302, doi:10.1051/m2an:2008001, http://dx.doi.org/10.1051/m2an:2008001.

[13] I. Jaimoukha and E. Kasenally, *Krylov subspace methods for solving large Lyapunov equations*, SIAM Journal on Numerical Analysis, 31 (1994), pp. 227–251, doi:10.1137/0731012, http://dx.doi.org/10.1137/0731012.

[14] R. Kalman, *Contributions to the theory of optimal control*, Boletin de la Sociedad Matematica Mexicana, 5 (1960), pp. 102–119.

[15] H. W. Knobloch and H. Kwakernaak, *Lineare Kontrolltheorie*, Springer Berlin Heidelberg, 1985, doi:10.1007/978-3-642-69884-2, http://dx.doi.org/10.1007/978-3-642-69884-2.

[16] H. Kwakernaak and R. Sivan, *Linear optimal control systems*, vol. 1, Wiley-interscience New York, 1972.

[17] P. Lancaster and L. Rodman, *Algebraic Riccati equations*, Oxford University Press, 1995.

[18] T. Penzl, *Eigenvalue decay bounds for solutions of Lyapunov equations: the symmetric case*, Systems & Control Letters, 40 (2000), pp. 139–144, doi:http://dx.doi.org/10.1016/S0167-6911(00)00010-4, http://www.sciencedirect.com/science/article/pii/S0167691100000104.

[19] C. Prud'homme, D. Rovas, K. Veroy, L. Machiels, Y. Maday, A. Patera, and G. Turinici, *Reliable real-time solution of parametrized partial differential equations: Reduced-basis output bound methods*, J. Fluids Engineering, 124 (2002), pp. 70–80, doi:10.1115/1.1448332.

[20] Y. Saad and X. Gv, *Numerical solution of large Lyapunov equations*, in Signal Processing, Scattering and Operator Theory, and Numerical Methods, Proc. MTNS-89, Birkhauser, 1990, pp. 503–511.

[21] A. Schmidt and B. Haasdonk, *Reduced basis approximation of large scale algebraic Riccati equations*, tech. report, University of Stuttgart, 2016, http://www.simtech.uni-stuttgart.de/publikationen/prints.php?ID=1480. Accepted for publication in ESAIM: COCV.

[22] T. Stykel, *Numerical solution and perturbation theory for generalized Lyapunov equations*, Linear Algebra and its Applications, 349 (2002), pp. 155–185, doi:http://dx.doi.org/10.1016/S0024-3795(02)00255-0.

[23] J.-G. Sun, *Perturbation theory for algebraic Riccati equations*, SIAM J. Matrix Anal. Appl., 19 (1998), pp. 39–65.

[24] N. Truhar and K. Veselić, *Bounds on the trace of a solution to the Lyapunov equation with a general stable matrix*, Systems & Control Letters, 56 (2007), pp. 493–503, doi:http://dx.doi.org/10.1016/j.sysconle.2007.02.003, http://www.sciencedirect.com/science/article/pii/S0167691107000217.

[25] S. Volkwein, *Model reduction using proper orthogonal decomposition*, lecture notes, University of Konstanz, 2012.

Appendix A. Proof of equality (8)**.** It is sufficient to show that the residual $\mathcal{R}(X)$ has an alternative representation via

$$\mathcal{R}(X) = E^T X E - A_X^T X A_X - F(X)^T F(X),$$

where $F(X)^T := \begin{bmatrix} C^T Q^{1/2} & K_X^T R^{1/2} \end{bmatrix}$. Therefore, let $X \in \mathcal{D}_n$ and $R_X := R + B^T X B$. it holds, that

$$
\begin{aligned}
(28) \qquad K_X^T B^T X A_X &= K_X^T B^T X \left(A - B K_X \right) \\
&= K_X^T B^T X A - K_X^T B^T X B K_X \\
&= K_X^T B^T X A - K_X^T \left(R + B^T X B - R \right) K_X \\
&= K_X^T B^T X A - K_X^T \left(R_X - R \right) R_X^{-1} B^T X A \\
&= K_X^T B^T X A - K_X^T B^T X A + K_X^T R K_X \\
&= K_X^T R K_X.
\end{aligned}
$$

From which we conclude

$$
\begin{aligned}
\mathcal{R}(X) &= X - A^T X A + A^T X B K_X - C^T Q C \\
&= X - A^T X (A - B K_X) - C^T Q C - K_X^T B^T X A_X + K_X^T B^T X A_X \\
&= X - (A^T - K_X^T B^T) X A_X - C^T Q C - K_X^T B^T X A_X \\
&\overset{(28)}{=} X - (A - B K_X)^T X A_X - \left(C^T Q C + K_X^T R K_X \right) \\
&= X - A_X^T X A_X - F(X)^T F(X).
\end{aligned}
$$