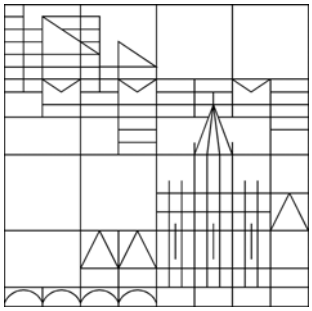

Universität Konstanz

Model reduction techniques with a-posteriori error analysis for linear-quadratic optimal control problems

Georg Vossen
Stefan Volkwein

Konstanzer Schriften in Mathematik

Nr. 298, Februar 2012

ISSN 1430-3558

Konstanzer Online-Publikations-System (KOPS)
URL: http://nbn-resolving.de/urn:nbn:de:bsz:352-184658

Manuscript submitted to
AIMS' Journals
Volume **X**, Number **0X**, XX **200X**

Website: http://AIMsciences.org

pp. **X–XX**

MODEL REDUCTION TECHNIQUES WITH A-POSTERIORI ERROR ANALYSIS FOR LINEAR-QUADRATIC OPTIMAL CONTROL PROBLEMS

Georg Vossen

Chair for Nonlinear Dynamics
Steinbachstr. 15, 52074 Aachen, Germany

Stefan Volkwein

Institut für Mathematik und Statistik
Universität Konstanz, D-78457 Konstanz, Germany

Abstract. The main focus of this paper is on an a-posteriori analysis for different model-order strategies applied to optimal control problems governed by linear parabolic partial differential equations. Based on a perturbation method it is deduced how far the suboptimal control, computed on the basis of the reduced-order model, is from the (unknown) exact one. For the model-order reduction, $\mathcal{H}_{2,\alpha}$-norm optimal model reduction (H2), balanced truncation (BT), and proper orthogonal decomposition (POD) are studied. The proposed approach is based on semi-discretization of the underlying dynamics for the state and the adjoint equations as a large scale linear time-invariant (LTI) system. This system is reduced to a lower-dimensional one using Galerkin (POD) or Petrov-Galerkin (H2, BT) projection. The size of the reduced-order system is iteratively increased until the error in the optimal control, computed with the a-posteriori error estimator, satisfies a given accuracy. The method is illustrated with numerical tests.

1. **Introduction.** Model reduction is a powerful tool widely used for solving partial differential equations (PDEs) or large-scale ordinary differential equations (ODEs) where the latter may arise from semi-discretization of PDEs in space. Many model reduction methods are based on the idea of projecting the state space to a much lower-dimensional one such that the obtained so-called reduced system can be solved much faster. There are several different methods such as Proper Orthogonal Decomposition (POD) [13, 21, 37], Balanced Truncation (BT) (see, e.g. [6, 26]), $\mathcal{H}_2$-norm model reduction (H2) [7, 11, 18, 25], Reduced Basis (RB) [22, 27], Hankel norm approximation [9] and many more out of which the first three, POD, BT and H2, will be the focus of this paper.

Roughly speaking, model reduction methods can be divided into two groups as follows. On the one hand, methods such as BT and H2 compute the reduced system on basis of an approximation of the operator mapping from inputs (i.e. usually time-dependent parameters to control the system) to outputs (i.e. observations of the system). This approach is applicable for linear time-invariant (LTI) large-scale ODE systems which can be obtained by semi-discretization of a PDE. For an overview

2000 *Mathematics Subject Classification.* Primary: 49K20, 90C20; Secondary: 35K10.
Key words and phrases. Model reduction, linear-quadratic optimal control, a-posteriori error.

we refer to [2]. Note that newer approaches for bilinear systems are available [4]. On the other hand, methods such as POD and RB approximate the state and/or adjoint variable itself. This requires solving the system once or even several times for different parameters to obtain the reduced system. However, it provides other advantages such as the possibility of treating nonlinear and time-varying systems.

Nevertheless, in both groups, model reduction is particularly of interest if one seeks a solution of the system for different parameters which can be finite-dimensional (particularly suitable for RB methods) or infinite-dimensional. In many applications, the system has to be solved during a parameter variation or even during an optimization or an optimal control procedure for many parameters, inputs or controls. Hence, even if the full system has to be solved in advance during a so-called off-line phase to compute the reduced system, the decrease of the computational costs for the complete parameter variation or optimization can be very large.

POD is mostly used for approximation in the context of (nonlinear) PDEs for functions depending on time and space. To be more precise, using a POD Galerkin scheme, one seeks a solution as a linear combination of coefficient functions depending on time and basis functions depending on space. Contrary to standard basis functions in a Finite Element Method (FEM), the spatially dependent so-called POD basis functions have a global support and involve some information about the system which is obtained out of the solution at certain time points (called snapshots).

In numerical practice, the snapshots (and hence, the POD basis functions) are generated by first semi-discretizing the system in space using, for instance, finite elements and then a time-integration of the obtained large-scale ODE system. This idea of semi-discretization is hence similar to the procedure employed if methods for LTI systems such as BT or H2 are used for model reduction in the context of PDE simulation.

Hence, from the numerical point of view, the three reduction methods POD, BT and H2 are based on the same idea. The approach involves a semi-discretization of the problem in space (resulting in a high-dimensional time-dependent ODE system) and the projection to a lower dimensional state vector. POD usually involves a Galerkin projection where the projection matrix is generated by snapshots of a solution of the PDE obtained for a certain reference control. The other two methods involve a Petrov-Galerkin projection where the projection matrices are obtained by considering a suitable LTI system with the controls as input and those parts of the state being relevant for the cost functional as output variables. Of course, the topology of the underlying PDE problem (e.g., L^2 or H^1) has to be taken into account when carrying out the projection.

Though POD is an excellent method of model reduction for many time-varying or nonlinear differential equations, it lacks an a priori error analysis that is in some sense uniform in the right-hand side of the underlying partial differential equation, say with respect to the control function. There are results on a priori estimates for POD that depend on certain assumptions on the orthogonal basis generated by the selected snapshots. We refer to Kunisch and Volkwein [19], Sachs and Schu [31], or [34]. However, such estimates will, in general, depend on the control used for generating the snapshots. In [12] an a-priori error analysis is presented for linear-quadratic optimal control problems. If the POD basis is computed utilizing the optimal state and associated adjoint variable, a convergence rate can be shown. But in the actual computation we do not know the optimal solution in advance. In

view of this, we are interested in a-posteriori estimates for assessing the precision of optimal controls for reduced control problems set up by POD. For the reduced-basis method a-posteriori error estimates for linear-quadratic optimal control problems we refer to [10]. An extension to nonlinear problems was recently given in [17].

BT and H2 are very rarely used in the context of optimal control; see [29] for an example. In a recent work [39], BT and H2 have been applied to optimal control problems subject to linear evolution equations. Accuracy of the optimal control has been compared between BT and H2 for different fixed sizes of the reduced system using a first-discretize-then-optimize (FDTO) approach.

There are three goals in this paper. Firstly, BT and H2 model reduction will be combined with a first-optimize-then-discretize (FOTD) approach which requires additional reduction of the adjoint equation. Here, we choose a rather simple method, a gradient descent technique, as the numerical optimization technique since the focus of this paper is on the model reduction techniques. Indeed, there are other techniques such as a primal-dual active-set approaches which are widely used in the context of optimal control problems. Secondly, a-posteriori error analysis is applied for the optimal control obtained for the reduced problem using BT and H2 as model reduction technique. Therefore, a technique developed in [34] for POD is adapted to BT and H2. Let us refer the reader to [33], where the a-posteriori analysis was successfully applied to the reduced-basis method. Thirdly, the three model reduction techniques (BT, H2 and POD) are compared by means of numerical experiments. As an overall goal, we aim for a framework which describes the three techniques in a similar manner providing the possibility of comparison with respect to accuracy and performance. Up to the authors' knowledge, such a comparison can not yet be found in the literature.

The paper is organized as follows. Section 2 introduces the problem class to be considered in this paper and summarizes well-known necessary optimality conditions. The model reduction methods POD, BT and H2 are given in Section 3. Individually, the three methods are well known; however, we will present the computational procedure to obtain the reduced system in one framework with a focus on correlations and differences of the three methods. Section 4 deals with the application of the reduction techniques to optimal control theory using a-posteriori error estimates for the optimal control. Results for numerical test are presented in Section 5.

2. Linear-quadratic parabolic optimal control problems.

In this paper, we will consider optimal control problems subject to parabolic PDEs with the control variable appearing linearly and with cost functionals of tracking type with respect to the state and the control.

2.1. Problem formulation.

The task is to minimize the quadratic cost functional

$$\begin{aligned} J(y,u,v) = &\frac{\lambda_\Omega}{2}\left\|y(T)-y_\Omega\right\|^2_{L^2(\Omega)} + \frac{\lambda_Q}{2}\left\|y-y_Q\right\|^2_{L^2(Q)} + \frac{\lambda_\Sigma}{2}\left\|y-y_\Sigma\right\|^2_{L^2(\Sigma)} \\ &\frac{\lambda_v}{2}\left\|v\right\|^2_{L^2(Q)} + \frac{\lambda_u}{2}\left\|u\right\|^2_{L^2(\Sigma)} \end{aligned} \tag{1}$$

subject to the linear parabolic problem

$$y_t + \mathcal{A}y = \beta_Q v \qquad \text{in } Q = \Omega \times (0,T), \tag{2a}$$
$$(\mathrm{A} \cdot \nabla y) \cdot \nu + \alpha y = \beta_\Sigma u \qquad \text{in } \Sigma = \Gamma \times (0,T), \tag{2b}$$
$$y(0) = y_0 \qquad \text{in } \Omega \tag{2c}$$

and the inequality constraints

$$v(x,t) \in [v_a(x,t), v_b(x,t)] \qquad \text{a.e. in } Q, \tag{2d}$$
$$u(x,t) \in [u_a(x,t), u_b(x,t)] \qquad \text{a.e. in } \Sigma, \tag{2e}$$

where y (depending on $(x,t) \in \bar{Q}$) is the state, v (depending on $(x,t) \in Q$) is the distributed control, u (depending on $(x,t) \in \Sigma$) is the boundary control. Moreover, $\Omega \subset \mathbb{R}^d$ with $d \in \{1,2,3\}$ is the spatial domain with boundary $\Gamma = \partial\Omega$, T is the final time, α is a coefficient, β_Q and β_Σ are coefficient functions (depending on (x,t) in Q and Σ, respectively), λ_Ω, λ_Q, λ_Σ, λ_v, λ_u are nonnegative constants with $\lambda_u + \lambda_v > 0$. Furthermore, y_0 is an initial distribution, ν denotes the normal vector on Γ and $\mathrm{A} = ((a_{ij})) \in L^\infty(\Omega; \mathbb{R}^{d\times d})$ is a given coefficient matrix. The operator $\mathcal{A}$ denotes the usual elliptic operator of the form

$$\mathcal{A}y(x) = -\nabla \cdot \big(\mathrm{A}(x) \cdot \nabla y(x) + \mathrm{b}(x) \cdot \nabla y(x) + c_0 y(x)$$

with time-independent coefficient functions $\mathrm{b} = (b_i) \in L^\infty(\Omega; \mathbb{R}^d)$ and $c_0 \in L^\infty(\Omega)$. A solution to (2) is understood in the weak sense. Notice that problem (1) and (2) can be written in the abstract form considered in [35].

2.2. **Necessary optimality conditions.** Existence and uniqueness are discussed in, e.g., [35] where the following first order necessary optimality conditions can also be found.

Theorem 2.1. *Let $(\bar{y}, \bar{v}, \bar{u})$ be the optimal solution of (1)–(2). Then it satisfies the variational inequalities*

$$\int\int_Q (\beta_Q(x,t)\bar{p}(x,t) + \lambda_v \bar{v}(x,t))\,(v(x,t) - \bar{v}(x,t))\,\mathrm{d}x\,\mathrm{d}t \geq 0 \tag{3a}$$
$$\int\int_Q (\beta_\Sigma(x,t)\bar{p}(x,t) + \lambda_u \bar{u}(x,t))\,(u(x,t) - \bar{u}(x,t))\,\mathrm{d}x\,\mathrm{d}t \geq 0 \tag{3b}$$

for all v and u which satisfy (2d) and (2e), respectively, and where the associated adjoint variable $\bar{p}$ is the solution of the adjoint system

$$-\bar{p}_t + \mathcal{A}^*\bar{p} = \lambda_Q(\bar{y} - y_Q) \qquad \text{in } Q, \tag{4a}$$
$$(\mathrm{A} \cdot \nabla \bar{p}) \cdot \nu + \alpha\bar{p} = \lambda_\Sigma(\bar{y} - y_\Sigma) \qquad \text{in } \Sigma, \tag{4b}$$
$$\bar{p}(T) = \lambda_\Omega(\bar{y}(T) - y_\Omega) \qquad \text{in } \Omega, \tag{4c}$$

where $\mathcal{A}^*$ denotes the adjoint operator of $\mathcal{A}$ satisfying

$$\int_\Omega \big(\mathcal{A}^*\varphi\big)(x)\psi(x)\,dx = \int_\Omega \varphi(x)\big(\mathcal{A}\psi\big)(x)\,dx \quad \text{for all } \varphi, \psi \in L^2(\Omega).$$

.

Second order optimality conditions based on classical results from Maurer and Zowe [24] on optimization in Banach spaces can also be found in [35].

3. **Numerical model reduction by projection.** As mentioned in the introduction, we will focus on the three model reduction methods POD, BT and H2. The methods share the feature that the state is projected on a lower-dimensional space to obtain a reduced system that can be solved much more efficiently. In a first step, the PDE problem is semi-discretized in space to obtain a large-scale LTI ODE system. The second step includes computation of the projection matrices - either out of snapshots obtained from an FEM solution (POD) or out of the transfer function which describes the input/output relation of the LTI system (BT and H2).

3.1. Semi-discretization to an LTI system using finite elements.

3.1.1. *Semi-discretization of the state equations.* Generate a grid on $\bar{\Omega}$ with n (for instance, equidistant) discretization points x_j, $j \in N := \{1, \dots, n\}$. Define basis functions $\varphi_j : \bar{\Omega} \to \mathbb{R}$, $j \in N$; e.g., hat functions such that $\varphi_j(x_i) = \delta_{ji}$ and φ_j is piecewise linear. A solution y of the PDE (2) shall be approximated by y^n with $y^n(x,t) = \sum_{j=1}^n z_j(t)\varphi_j(x)$ and the vector function $z = (z_1, \dots, z_n)^T$ is called the *semi-discretized state*.

Analogously, we approximate the controls by $v^n(x,t) = \sum_{j \in N \setminus J} w_j(t)\varphi_j(x)$ and $u^n(x,t) = \sum_{j \in J} w_{|N\setminus J|+j}(t)\varphi_j(x)$, where $J := \{j \in N : x_j \in \Gamma\}$ and denote $w = (w_1, \dots, w_n)^T$ as the *semi-discretized control*. Of course, we can also use different basis functions for the control variable, e.g., piecewise constant step functions. Note that often some components of the vector function w are irrelevant for computation of the cost functional J. For instance, if $\beta_Q \equiv 0$ (i.e., we have a boundary control problem) it is sufficient to consider only the indices $j \in J$.

Similarly, in many applications, it is not necessary to know the complete semi-discretized state for computation of the (discrete) cost functional, but only a linear combination of the state, for instance, the state at the (or some parts of the) boundary if $\lambda_\Omega = \lambda_Q = 0$. Those linear combinations shall be called *output variables* of the semi-discretized system.

To summarize, we end up with an implicit linear time-invariant dynamical system given by

$$E\dot{z}(t) = Az(t) + Bw(t), \quad z(0) = z_0, \qquad a(t) = Cz(t) \tag{5}$$

with the (semi-discretized) state $z : \mathbb{R} \to \mathbb{R}^n$, the (semi-discretized) control $w : \mathbb{R} \to \mathbb{R}^m$ and the (semi-discretized) output $a : \mathbb{R} \to \mathbb{R}^q$. Here, m is the number of those discretization points "where v and u act" (i.e. where the corresponding factor β_Q or β_Σ, respectively, in (2) does not vanish) and q is the number of linear combinations of state variables "required to compute the cost functional" (i.e. where the corresponding factor λ_Ω, λ_Q or λ_Σ, respectively, in (1) does not vanish). Furthermore, z_0 is the initial vector with components $z_{0,j} = y_0(x_j)$ for $j \in N$ and E, $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{q\times n}$ are constant matrices. Hereby, E, A, B are determined by the type of chosen FEM basis functions. For instance, using hat functions, E and A are tridiagonal for a one-dimensional spatial domain Ω. The matrix E is the (positive definite, symmetric) mass matrix and $-A$ the (positive semi-definite, symmetric) stiffness matrix. In the context of system theory, the matrices E, A, B, C are called system matrices. If $E \neq I_n$ (where I_n is the identity matrix of dimension n), the system is implicit and often called a descriptor system.

Note that many model reduction methods which approximate the input/output operator such as BT and H2 are well-established for explicit systems, i.e. systems

of the form

$$\dot{z}(t) = A_{\text{expl}} z(t) + B_{\text{expl}} w(t), \quad z(0) = z_0, \qquad a(t) = C_{\text{expl}} z(t) \tag{6}$$

which can formally be obtained from (5) by multiplication with E^{-1} from the left since E is positive definite. Hence, $A_{\text{expl}} = E^{-1}A$, $B_{\text{expl}} = E^{-1}B$ and $C_{\text{expl}} = C$ hold; cf., Remark 2.

3.1.2. *Semi-discretization of the adjoint equations.* The structure of the adjoint system (4) is very similar to the structure of the state equations (2). More precisely, the state occurring on the right hand side of (4a) and (4b) can be seen as an *adjoint input*. Furthermore, those parts of the state occurring in the variational inequality (3a)–(3b), can be interpreted as an *adjoint output*.

Hence, to solve the adjoint system, the same procedure can be applied. Semi-discretization provides an implicit LTI system

$$E^{\mathrm{a}} \dot{z}^{\mathrm{a}}(t) = A^{\mathrm{a}} z^{\mathrm{a}}(t) + B^{\mathrm{a}} w^{\mathrm{a}}(t), \quad z^{\mathrm{a}}(T) = z_T^{\mathrm{a}}, \qquad a^{\mathrm{a}}(t) = C^{\mathrm{a}} z^{\mathrm{a}}(t) \tag{7}$$

for the semi-discrete adjoint z^{a}, the semi-discrete adjoint input w^{a} and the semi-discrete adjoint output a^{a}. Hereby, the terminal condition is obtained from condition (4b) where z_T^{a} is the terminal vector with components $z_{T,j}^{\mathrm{a}} = \lambda_\Omega(z_j(T) - y_\Omega)$ for $j \in N$. Analogously, this can formally transformed to an explicit system. Note that one can apply a time transformation $\check{t} = T - t$ to obtain an ODE system with given initial value.

3.1.3. *Basics on LTI systems.* In the following discussions, we need some basic ingredients from system theory. Applying Laplace transformation to the system equations (6) we obtain for $z_0 = 0$ the relation

$$a_{\mathcal{L}}(s) = H(s)\, w_{\mathcal{L}}, \quad H(s) = C(sE - A)^{-1}B, \tag{8}$$

where the subscript $\mathcal{L}$ denotes the corresponding quantity in Laplace space. Equation (8) says that the output depends linearly on the input with the so-called transfer function $H : \mathbb{C} \to \mathbb{R}^{q,m}$ as factor. An important property of an LTI system is that representation (6), also called the *realization of the system*, is not unique. In fact, there are infinitely many other system matrices, namely

$$\tilde{E} = S^{-1}ES, \tilde{A} = S^{-1}AS, \quad \tilde{B} = S^{-1}B, \quad \tilde{C} = CS$$

with an arbitrary regular matrix S (called *state space transformation matrix*) which map all inputs w to the same output a as for the original realization. Note that the transfer functions of these two realizations are identical. As mentioned, we will investigate systems where E is positive definite. Hence, stability can be discussed by means of the eigenvalues of $E^{-1}A$. The system is called α-stable if

$$\alpha > \max\{\Re\,\lambda : \lambda \text{ is an eigenvalue of } E^{-1}A\}. \tag{9}$$

Since $-A$ is positive semi-definite, the system is α-stable for all $\alpha > 0$. If (9) is fulfilled for $\alpha = 0$ (which is the case if $-A$ is positive definite), the system is called *stable*. For stable systems, we finally introduce the generalized Gramians $\mathcal{P}$ and $\mathcal{Q}$ of reachability and observability, respectively, which satisfy the Lyapunov equations

$$A\mathcal{P}E^T + E\mathcal{P}A^T + BB^T = 0, \quad A^T\mathcal{Q}E + E^T\mathcal{Q}A + C^TC = 0.$$

3.1.4. *Time discretization.* To numerically solve the state system (5) or (6) as well as the adjoint system (7), respectively, one additionally has to generate a time grid on $[0,T]$ with k (for simplicity, equidistant) time steps $t_i = T(i-1)/(k-1)$, $i \in K := \{1,\dots,k\}$ resulting in time step size $h_t = T/k$. The full discretized state and control are denoted by $z^i = z(t_i) \in \mathbb{R}^n$ and $w^i = w(t_i) \in \mathbb{R}^m$. The solution can be obtained, for instance, by a standard θ-method

$$E(z^{i+1} - z^i) = h_t\theta(Az^i + Bw^i) + h_t(1-\theta)(Az^{i+1} + Bw^{i+1}) \tag{10}$$

The vector z^i, $i \in K$, is called discrete snapshot at time instance t_i. The same has to be applied to the LTI system obtained from semi-discretization of the adjoint equations.

Remark 1. We note that other semi-discretization techniques such as finite differences or discontinuous Galerkin are also applicable.

3.2. **Projection to a reduced system.** Instead of the large-scale dimensional trajectories z of the system (5), we consider trajectories $\hat{z}$ which evolve in an r-dimensional subspace. We therefore consider a state space transformation partitioned by $S = [Z, S_2]^T$, $S^{-1} = [V, S_1]$ with sub-matrices S_1, $S_2 \in \mathbb{R}^{n\times(n-r)}$ and the so-called projection matrices $Z, V \in \mathbb{R}^{n\times r}$. With this partition, Z and V satisfy $Z^TV = I_r$ (where I_r is the identity matrix of dimension r) which implies that $\Pi := VZ^T$ is a projection. We furthermore partition the state $\tilde{z} = Sz$ by $z = (z_1^T, z_2^T)^T$ with $z_1 \in \mathbb{R}^r$ and $z_2 \in \mathbb{R}^{n-r}$ and obtain, regarding the first r equations in the LTI system,

$$Z^TEV\dot{z_1}(t) = Z^T(AVz_1(t) + S_1z_2(t) + Bw(t)),\ \hat{a}(t) = C(Vz_1(t) + S_1z_2(t)). \tag{11}$$

Note that this formula is exact. Approximation by model reduction comes into play if we assume that S_1z_2 is small and can therefore be neglected. We hence define $\hat{z} \in \mathbb{R}^r$ as solution of the approximated ODE system in (11) where S_1z_2 vanishes; i.e., (6) becomes

$$Z^TEV\dot{\hat{z}}(t) = Z^TAV\hat{z}(t) + Z^TBw(t), \quad \hat{a}(t) = CV\hat{z}(t) \tag{12}$$

with initial condition $\hat{z}(0) = \hat{z}_0 = Z^Tz_0$ which is also a linear time-invariant system with so-called reduced system matrices $\hat{E} := Z^TEV \in \mathbb{R}^{r\times r}$, $\hat{A} := Z^TAV \in \mathbb{R}^{r\times r}$, $\hat{B} := Z^TB \in \mathbb{R}^{r\times m}$ and $\hat{C} := CV \in \mathbb{R}^{q\times r}$.

If $Z = V$, the procedure is called a Galerkin projection. In the general case, we obtain a Petrov-Galerkin projection. POD is usually combined with a Galerkin projection whereas BT and H2 are applied by means of a Petrov-Galerkin projection. The construction of the projection matrices will be summarized in the next paragraph.

Remark 2. POD can be used for implicit systems (indeed for nonlinear problems). BT is also applicable but we note that the standard method in Matlab called `balreal` is implemented only for explicit systems. For extensions to implicit systems, we mention Stykel [32], Saak [30] and references therein. H2 is usually considered for explicit systems. However, the theory and the algorithms can easily be extended to implicit systems (see Section 3.2.3 for details).

This gives rise to two approaches for H2 and BT. On the one hand, the system can firstly be transformed to an explicit one which can be reduced with standard H2 and BT methods. This approach involving *reduction of the explicit* system is called H2-RE or BT-RE, respectively. On the other hand, it is possible to directly

reduce the implicit system and this is referred to as the H2-RI or H2-RI approach. Regarding the first approach, E is a large matrix which can make its inversion numerically infeasible. Nevertheless, since E is sparse its inverse may be computed with moderate numerical effort (or even given analytically; see numerical example). The second approach seems to avoid this problem. However, standard reduction methods are usually more robust for explicit systems and often designed for stable systems. In the numerical tests, we compare POD with both approaches applied to H2 and BT.

3.2.1. *Projection matrix for POD.* The first step is to solve the ODE system in (12) for a certain reference control u_{ref} which may also be chosen as the starting guess for the optimization routine. Following Section 3.1.4, we obtain k discrete snapshots $z^i \in \mathbb{R}^n$ out of which we construct the so-called snapshot matrix $Y^T DY \in \mathbb{R}^{k,k}$ with $Y = [z^1, \dots, z^k] \in \mathbb{R}^{n\times k}$. Here, D stands for E or A (provided A is positive definite). The POD basis functions ψ_j, $j = 1, \dots, k$, in discrete form are given by $\psi_i = Yv_i/\sqrt{\lambda_i}$, where v_i is an ith eigenvector of $Y^T DY$ (associated with the eigenvalue λ_i satisfying $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k \geq 0$. It can be observed that the eigenvalues decay very rapidly such that the idea of model reduction is to use only $r \ll k$ basis functions. The projection matrix is then given by

$$Z = Z_{\mathrm{POD}} = [\psi_1, \dots, \psi_r], \quad V = V_{\mathrm{POD}} = Z_{\mathrm{POD}}.$$

Note that due to construction of Z, we obtain $\hat{E} = I_r$.

We mention again that the original large-scale system has to be solved once for a certain reference u_{ref} control. Hence, the projection matrix Z depends on the choice of this control (see our numerical experiments).

3.2.2. *Projection matrices for Balanced Truncation.* The idea of BT model reduction for stable systems is to firstly find a so-called balanced realization E_{B}, A_{B}, B_{B}, C_{B} characterized so that the corresponding Gramians are diagonal and satisfy $\mathcal{P}_{\mathrm{B}} = \mathcal{Q}_{\mathrm{B}} =: D_{\mathrm{H}}$. The diagonal entries $\sigma_1 \geq \ldots \geq \sigma_n \geq 0$ in the matrix D_{H} are called Hankel singular values of the system. In the second step, less significant parts corresponding to $n - r$ small Hankel singular values of the system are truncated. A (non-unique) form of the projection matrices is given by

$$Z = Z_{\mathrm{BT}} = (\Upsilon_L^T \Upsilon_U D_{\mathrm{H}}^{-1/2})^T, \quad V = V_{\mathrm{BT}} = E\Upsilon_R \Upsilon_V D_{\mathrm{H}}^{-1/2}$$

with singular value decomposition and Cholesky factorizations

$$\Upsilon_L E \Upsilon_R = \Upsilon_U D_{\mathrm{H}} \Upsilon_V, \quad \mathcal{P} = \Upsilon_R \Upsilon_R^T, \quad \mathcal{Q} = \Upsilon_L^T \Upsilon_L.$$

where Υ_L and Υ_R are lower left and upper right triangular matrices, respectively. For explicit systems, there are further transformations which can be found in [2, 5]. Unstable systems are considered in [16]. Note that many methods (also for implicit systems) usually provide $\hat{E} = I_r$.

Contrary to POD, the original system need not be solved in advance. The matrices Z and V do not depend on the choice of any reference control.

3.2.3. *Projection matrices for $\mathcal{H}_{2,\alpha}$-norm optimal model reduction.* For stable systems, the reduced system constructed in such a way that the output of the reduced system approximates the output of the original system optimally w.r.t. the L^2-norm for the impulse input. Equivalently, the transfer function of the reduced

system approximates the transfer function of the original system optimally w.r.t. the $\mathcal{H}_2$-norm given as the special case $\alpha = 0$ of

$$\|H\|_{\mathcal{H}_{2,\alpha}} = \left(\frac{1}{2\pi} \int_{-\infty}^{\infty} \operatorname{tr}\left(H(\alpha + i\omega)^* H(\alpha + i\omega)\right) dw \right)^{1/2} \tag{13}$$

where $\operatorname{tr}(M)$ denotes the trace of a matrix M. For α-stable systems with $\alpha > 0$, one can shift the system by α (i.e., consider a system with perturbed system matrix $A - \alpha I_n$) to obtain a stable system or, equivalently, search the optimal reduced system whose transfer function approximates the original one w.r.t. the $\mathcal{H}_{2,\alpha}$-norm. The choice of this norm is reasonable since many problems comprise L^2 expressions of the state in the cost functional. Correlations between this norm and the error in the optimal control have been investigated in [39].

The projection matrices $V = V_{\mathrm{H2}} = [v_1, \ldots, v_r]$ and $Z = Z_{\mathrm{H2}} = \tilde{Z}(\tilde{Z}^* V)^{-*}$, $\tilde{Z} = [\tilde{z}_1, \ldots, \tilde{z}_r]$ are given by

$$v_j = ((2\alpha - \hat{\lambda}_j^*)E - A)^{-1} B \hat{b}_j, \quad \tilde{z}_j = ((2\alpha - \hat{\lambda}_j^*)E - A^*)^{-1} C^* \hat{c}_j^* \tag{14}$$

for $j = 1, \ldots, r$, where $\hat{\lambda}_j$ are the eigenvalues of $\hat{A}$ and $\hat{B}$, $\hat{C}$ are represented as $\hat{B} = [\hat{b}_1^*, \ldots, \hat{b}_r^*]^*$, $\hat{C} = [\hat{c}_1, \ldots, \hat{c}_r]$. Since $\hat{\lambda}_j$, $\hat{B}$ and $\hat{C}$ are not known, the projection matrices are found by an iterative algorithm. Note that formulae (14) are an extension to the known expressions for explicit systems where $E = I_n$. For systems with $m = q = 1$, the method IRKA [11] is applicable, whereas the method MIRIAm [7] can be used in the multi-dimensional case. Note that the projection matrices and hence the reduced system matrices may have complex entries. However, for systems with $m = q = 1$ it has been shown that Z and V can be transformed so that the reduced system matrices are real [39].

As for BT, the projection matrices are independent on the choice of any reference control.

4. Model reduction for optimal control problems.

4.1. Reduced optimal control problem.

Using semi-discretization as in Section 3, the PDE optimal control problem can be transformed to a large-scale ODE problem which reads as

$$\begin{aligned} &\min \int_0^T f(t, a(t), w(t))\, dt, \qquad a(t) = Cz(t) \\ &\text{subject to } E\dot{z}(t) = Az(t) + Bw(t) \text{ in } (0,T], \quad z(0) = z_0 \quad \text{and} \quad w(t) \in W \end{aligned} \tag{$\mathbf{P}_1$}$$

with some suitable function f in the cost functional depending on the output a induced by the tracking terms for the state and the input w induced by the regularization terms.

For a fixed number $r \ll n$, we can formulate the so-called reduced optimal control problem of order r as follows.

$$\begin{aligned} &\min \int_0^T f(t, \hat{a}(t), w(t))\, dt, \qquad \hat{a}(t) = \hat{C}\hat{z}(t) \\ &\text{subject to } \hat{E}\dot{\hat{z}}(t) = \hat{A}\hat{z}(t) + \hat{B}w(t) \text{ in } (0,T], \quad \hat{z}(0) = \hat{z}_0 \quad \text{and} \quad w(t) \in W. \end{aligned} \tag{$\mathbf{P}_1^r$}$$

Remark 3. Note that the expression of the "reduced problem" is sometimes also used for the problem which is obtained if one considers the cost functional $\hat{J}(u) = J(\mathcal{S}u, u)$ where $\mathcal{S}$ is the control-to-state mapping. Here, the notation is related to the reduced system which is the well-known notation in the context of model reduction.

In [39], the reduced problem is considered for BT and H2 model reduction. Using an FDTO approach, the solution has been investigated for different fixed sizes r and compared with the solution of the original problem for a fine discretization. Here, we will combine model reduction with an FOTD approach which requires model reduction of the adjoint system also. This may lead to additional reduction computations. In simple examples, however, the adjoint equation can have the same structure as the state equation so that the same reduced system may be used. In the numerical tests, a gradient projection method will be used for solving the reduced optimal control problems.

4.2. **A-posteriori error analysis for the control.** Of course, if the reduced system (12) approximates the original system (6) well, we expect the solution $\bar{\tilde{w}}$ of the reduced optimal control problem ($\mathbf{P}_1^r$) to be close to the optimal solution $\bar{w}$ of the original control problem ($\mathbf{P}_1$). Indeed, correlations between the approximation error in the $\mathcal{H}_{2,\alpha}$-norm of the model reduction and the error in the optimal control have been demonstrated for several numerical examples in [39], where BT and H2 were investigated and compared. However, although the error in the optimal control has been observed to be rather small (even for very small numbers of r), it is a-priori not clear how large r should be chosen.$\diamondsuit$

To estimate the error in the optimal control a-posteriori, we will apply a technique also called perturbation method which was first used by Dontchev et al. [8] and Malanowski, Büskens and Maurer [23] for optimal control of ODEs. The method has later been adapted to PDEs [3, 33, 34], where in the latter two references the error of the solution of the reduced problem obtained by POD and/or reduced-basis model reduction was estimated. Here, we apply this technique in the same problem class, but for the different reduction methods, namely BT and H2.

The idea will be summarized here for a class of boundary control problem; the case of a distributed control is analoguous. A non-optimal control $\tilde{u} \neq u$ together with corresponding state $\tilde{y}$ and $\tilde{p}$ for problem (1)–(2) does not satisfy the variational inequality (3b) but the perturbed inequality

$$\int\int_Q (\beta_\Sigma(x,t)\tilde{p}(x,t) + \lambda_u \tilde{u}(x,t) + \zeta(x,t))\,((u(x,t) - \tilde{u}(x,t))\,\mathrm{d}x\,\mathrm{d}t \geq 0$$

for all u satisfying the control constraints (2e) with a perturbation ζ given by

$$\zeta(x,t) = \begin{cases} [\tilde{\zeta}]_- & \text{if } \tilde{u}(x,t) = u_a(x,t), \\ -\tilde{\zeta} & \text{if } \tilde{u}(x,t) \in (u_a(x,t), u_b(x,t)), \\ [\tilde{\zeta}]_+ & \text{if } \tilde{u}(x,t) = u_b(x,t) \end{cases}$$

with $\tilde{\zeta} = \beta_\Sigma(x,t)\tilde{p}(x,t) + \lambda_u \bar{u}(x,t)$ and $[\tilde{\zeta}]_+ := (|\tilde{\zeta}| + \tilde{\zeta})/2$, $[\tilde{\zeta}]- := (|\tilde{\zeta}| - \tilde{\zeta})/2$. The error in the optimal control then satisfies

$$\|\tilde{u} - \bar{u}\|_{L^2(\Sigma)} \leq \frac{1}{\lambda_u} \|\zeta\|_{L^2(\Sigma)}.$$

We stress that it is possible to estimate the error in the control without knowledge of the exact (unknown) control. However, the estimator requires the solution of the

full primal and dual equations (to compute $\tilde{p}(x,t)$); e.g., by standard FEM with fine discretization which we assume to provide $\tilde{y}$ and $\tilde{p}$ with negligible error. Otherwise, the estimator should also include the numerical discretization error for $\tilde{y}$ and $\tilde{p}$.

In [34], an algorithm for solving the optimal control problem up to a desired accuracy was presented. The idea is to iteratively increase the number of POD basis functions for the reduced problem until the estimator ensures the desired accuracy of the optimal control. In each iteration, the reduced control problem was solved by a primal-dual active set strategy, which is locally superlinearly convergent; see, e.g., [14].

Here, we use the same idea of error estimation for the solution of the reduced problem with iteratively increased size r for all three reduction methods POD, BT and H2. In each iteration, the reduced problem is solved by a gradient projection method. Of course, one can also apply a primal-dual active set strategy. However, since we are mostly interested in comparing the errors for different numbers r of basis functions, our main focus in the present paper is not on the optimization method. Starting with $r = 1$ and some starting guess for the optimal control, we will use the obtained optimal control for the reduced problem of size r as the starting guess for the problem of size $r + 1$.

Remark 4. To improve the approximation quality for the POD Galerkin scheme one can combine the change of the number r of basis functions with a change of the POD basis. In [38] a combination of the a-posteriori analysis with Optimality-System POD [20] was presented. Here we combine our a-posteriori error analysis with the following update strategy (compare [1, 28]): Assuming that the main computational effort in determining the POD basis is the solution of the full state and adjoint PDE, one may update the POD basis for each r since these solutions are necessary for the error estimator in each step.

5. Numerical tests: Optimal boundary control of a convection-diffusion equation.

5.1. Problem formulation and optimality conditions.

Consider the following optimal control problem of minimizing the L^2-norm difference of the temperature $y = y(x,t)$ of a rod at the boundary $x = 0$ to a desired temperature $y_\mathrm{d} = y_\mathrm{d}(t)$ involving a regularization term with regularization parameter $\lambda_u > 0$ for the boundary control $u = u(t)$ acting at the boundary $x = 1$:

$$\min J(y,u) := \frac{1}{2}\left(\int_0^T (y(0,t) - y_\mathrm{d}(t))^2 \mathrm{d}t + \lambda_u \int_0^T u(t)^2 \mathrm{d}t\right)$$

subject to

$$\begin{aligned}
y_t(x,t) - \eta y_x(x,t) - y_{xx}(x,t) &= 0 && \text{for } (x,t) \in Q,\\
y_x(0,t) &= 0 && \text{for } t \in (0,T),\\
y_x(1,t) &= u(t) && \text{for } t \in (0,T),\\
y(x,0) &= y_0(x) && \text{for } x \in \Omega = (0,1),\\
u \in U_{ad} &= \{\tilde{u} \in L^2(0,T) \,|\, u_a \le \tilde{u} \le u_b\}
\end{aligned}$$

with $u_a, u_b \in \mathbb{R}$ satisfying $u_a \le u_b$ and some convection parameter $\eta \in \mathbb{R}$. The adjoint equations are given by

$$\begin{aligned} \bar{p}_t(x,t) + \eta \bar{p}_x(x,t) + \bar{p}_{xx}(x,t) &= 0 && \text{for } (x,t) \in Q, \\ \bar{y}(0,t) - y_{\mathrm{d}}(t) + \bar{p}_x(0,t) + \eta \bar{p}(0,t) &= 0 && \text{for } t \in (0,T), \\ \bar{p}_x(1,t) + \eta \bar{p}(1,t) &= 0 && \text{for } t \in (0,T), \\ \bar{p}(x,T) &= 0 && \text{for } x \in \Omega. \end{aligned}$$

The variational inequality comes to

$$\int_0^T (\lambda_u \bar{u}(t) + \bar{p}(1,t))(u(t) - \bar{u}(t)) \mathrm{d}t \ge 0 \quad \text{for all } u \in U_{\mathrm{ad}}.$$

which implies the weak minimum condition

$$\bar{u}(t) = \mathrm{Proj}_{U_{\mathrm{ad}}} \left\{ -\frac{\bar{p}(1,t)}{\lambda_u} \right\}, \quad t \in (0,T), \tag{15}$$

for the optimal control, where $\mathrm{Proj}_{U_{\mathrm{ad}}}$ denotes the projection operator onto the set U_{ad}.

5.2. **Semi-discretization and FEM solution.** Let us semi-discretize the problem using finite elements. For $j = 1 \ldots, n$, we define $x_j := \Delta x(j-1)$, $\Delta x = 1/(n-1)$ and hat functions φ_j by

$$\varphi_j(x) := \begin{cases} \frac{x - x_{j-1}}{\Delta x}, & x \in (x_j - \Delta x, x_j], \ j \ne 1, \\ \frac{x_{j+1} - x}{\Delta x}, & x \in (x_j, x_j + \Delta x), \ j \ne n, \\ 0 & \text{else}, \end{cases}$$

for $j = 1, \ldots, n$. Considering the weak formulation of the problem, we end up by an implicit LTI system for the state in the form (5) with system matrices

$$E = \frac{\Delta x}{6} \begin{pmatrix} 2 & 1 & & & \\ 1 & 4 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 4 & 1 \\ & & & 1 & 2 \end{pmatrix}, \quad A = A_{\mathrm{D}} + A_{\mathrm{C}} \tag{16}$$

$$A_{\mathrm{D}} = \frac{1}{\Delta x} \begin{pmatrix} -1 & 1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 1 \end{pmatrix}, \ A_{\mathrm{C}} = \frac{\eta}{2} \begin{pmatrix} 1 & -1 & & & \\ 1 & 0 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 0 & -1 \\ & & & 1 & -1 \end{pmatrix}, \tag{17}$$

$$B = (0, \ldots, 0, 1)^T, \quad C = (1, 0, \ldots, 0). \tag{18}$$

Hereby, A_{D} is related to the diffusion term in the state equation, and A_{C} is related to the convection term. After the transformation

$$\check{t} = T - t, \quad \check{x} = 1 - x, \tag{19}$$

we obtain a rather similar LTI system for the adjoint system with the same system matrices $E^{\mathrm{a}} = E$, $B^{\mathrm{a}} = B$, $C^{\mathrm{a}} = C$ as for the state LTI system and the matrix $A^{\mathrm{a}} = A + \mathrm{diag}[-\eta, 0, \ldots, 0, \eta]$ which is slightly different for $\eta \ne 0$. The obtained ODE systems are solved with a semi-implicit method for $\theta = 1/2$ in (10).

Note that following the RE method discussed in Remark 2, this can be transformed to an explicit system. Hereby, we deduce from [36] that, indeed, the inverse of E can be given analytically in compact form as

$$\left(E^{-1}\right)_{ij} = \frac{(-1)^{i+j}}{\vartheta}\left(c_1^{n-|i-j|} + c_2^{n-|i-j|} + c_1^{n+2-i-j} + c_2^{n+2-i-j}\right)$$

where $\vartheta = 2(2c_1^{n-1} + 2c_2^{n-1} - c_1^{n-2} - c_2^{n-2})$, $c_1 = 2+\sqrt{3}$, $c_2 = 2-\sqrt{3}$.

5.3. **Non-convective systems.** We first discuss the case $\eta = 0$; i.e., there is no convection in the system. We choose the data and discretizations as

$$y_{\mathrm{d}}(t) \equiv 0.5, \quad \lambda_u = 0.01, \quad -u_a = u_b \equiv 1, \quad y_0(x) \equiv 0, \quad T = 1, \quad n = k = 100. \tag{20}$$

5.3.1. *FEM solution.* We apply a gradient projection method with a starting guess for the control of $u \equiv 0.5$. The stopping criterion is chosen so that (15) is fulfilled in the discrete L^2-sense with error less than $\mathrm{tol}_{\mathrm{G}}$=1e-05 is satisfied after 41 iterations. The obtained optimal control $\bar{u}$ (solid line) and the function $-\bar{p}(1,t)/\lambda_u$ (solid line with dots) appearing in the minimum condition (15) are given in Figure 1 (left). The optimal state $\bar{y}$ at $x = 0$ (solid line) and the desired temperature (dashed line) are given in Figure 1 (right).

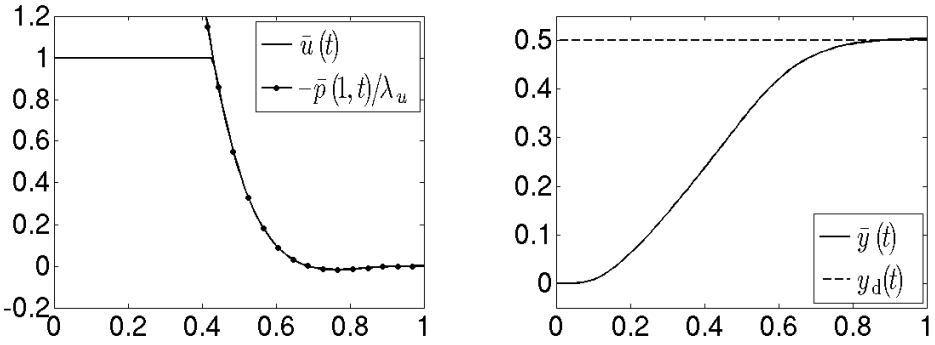

FIGURE 1. Optimal control $\bar{u}(t)$ and function $-\bar{p}(1,t)/\lambda_u$ (left); optimal state $\bar{y}(0,t)$ and desired state $\bar{y}_{\mathrm{d}}$ (right)

The reasonable control strategy suggests to heat the rod with maximal power $\bar{u} = 1$ (i.e., the control constraint is active until the exit time $t_{\mathrm{e}} = 0.4343$) in the beginning (where $\bar{y} \ll y_{\mathrm{d}}$) and, after continuously decreased heating, to keep $\bar{u}$ around zero at the end (where $\bar{y} \approx y_{\mathrm{d}}$) of the time interval. Note that $\bar{u}$ vanishes at time $t_{\mathrm{v}} = 0.6869$, it is minimal at time $t_{\mathrm{m}} = 0.7677$ with $\bar{u}(t_{\mathrm{m}}) = -0.0168$ and then monotonically increases until $t = T$ with $\bar{u}(T) = -1.5517\text{e-}05$.

5.3.2. *Solution using model reduction.* Let us now apply the proposed methods. For each reduced problem, we apply the gradient projection method with a combined Armijo and tri-section method for step size determination. The gradient method stops if the minimum condition is fulfilled in the L^2-sense with error less than $\mathrm{tol}_{\mathrm{G}}$=1e-05 or if the number of iterations exceeds 15 (which does indeed happen for low values of r). Furthermore, the stopping criterion for the outer loop (i.e., the optimization) is that the estimator fulfills $\|\zeta\|_{L^2(0,T)}/\lambda_u <\mathrm{tol}_{\mathrm{E}}$=1e-05 in the discrete sense. Since 0 is the eigenvalue with largest real part for the eigenvector

$[1, \ldots, 1]^T$ of A, we choose $\alpha = 0.1$. The stopping criterion for the IRKA method to construct the reduced system for the H2 technique is given by a relative distance of the largest eigenvalue magnitude of $\hat{A}$ between two iterations less than 1e-08. This results in 10-15 iterations depending on the value of r. For POD, different reference controls are tested: $u_{\mathrm{ref}}(t) = u_1(t) \equiv 0.5$, $u_{\mathrm{ref}}(t) = u_2(t) = t$ and $u_{\mathrm{ref}} = \bar{u}(t)$ where $\bar{u}$ is the optimal control obtained from the FEM solution depicted in Figure 1. The algorithm is tested for POD without and with basis update in each iteration; cf. Remark 4. H2 and BT are tested for the RE and the RI approach; cf. Remark 2. The BT-RE approach is realized with the Matlab routine `balreal` whereas BT-RI is applied using the Generalized Low-rank Cholesky factor ADI iteration (G-LRCF-ADI) by Saak [30]. As mentioned above, the adjoint and the state LTI system are equivalent after transformation (19) resulting in the same reduced system for H2 and BT. However, the sets of POD basis functions differ since the input u for the state is different to the adjoint input $y - y_{\mathrm{d}}$ resulting in different snapshots. As starting guess for the control, we take $u \equiv 0.5$. All tests have been carried out in Matlab using a standard PC.

The results for the three methods (POD, BT and H2) are shown in Tables 1–5. The tables show the values of the control error estimator, the effectivity

$$\eta_{\mathrm{eff}} = \frac{\|\zeta\|_{L^2(0,T)}}{\lambda_u \|\bar{\hat{u}} - \bar{u}\|_{L^2(0,T)}}$$

(where, again, $\bar{u}$ is taken from the FEM solution depicted in Figure 1), the number of gradient method iterations and the relative $\mathcal{H}_{2,\alpha}$-norm error of the reduced system in each iteration step of the algorithm. Hereby, the latter error is computed by means of the corresponding explicit systems. The plots of the obtained solutions are not shown, since there is no visible difference to the solution of the original system.

r	$\|\zeta\|/\lambda_u$	η_{eff}	Iter	MR error	$\|\zeta\|/\lambda_u$	η_{eff}	Iter	MR error
1	2.770e-01	5.146	15	1.453e-01	2.136e-01	3.154	15	1.524e-01
2	1.132e-01	4.751	15	1.958e-02	1.492e-01	5.180	15	1.981e-02
3	1.015e-02	5.616	15	2.642e-03	1.429e-02	5.924	15	2.671e-03
4	1.068e-03	5.926	15	3.559e-04	1.474e-03	6.191	15	3.606e-04
5	1.214e-04	6.373	9	4.798e-05	1.635e-04	6.597	9	4.872e-05
6	1.205e-05	1.841	2	6.476e-06	1.682e-05	2.357	2	6.587e-06
7	1.072e-05	1.735	2	8.756e-07	1.138e-05	1.746	2	8.919e-07
8	3.078e-06	0.549	2	1.189e-07	4.238e-06	0.736	2	1.210e-07

TABLE 1. Results for control error estimation, estimator effectivity, number optimization iterations and relative $\mathcal{H}_{2,\alpha}$ model reduction error for H2-RE (left) and BT-RE (right).

5.3.3. *Discussions.* The algorithm converges for all three reduction methods after 14 iterations or less. As a general observation we find that the method converges faster using H2 and BT than using POD.

The choice of the reference control to construct the POD basis functions seems to have only a marginal effect in the number of iterations for the chosen tolerance accuracy. However, in the first few iterations, the error estimator provides large differences with respect to the choice of the reference control showing that for smaller

r	$\|\zeta\|/\lambda_u$	η_{eff}	Iter	MR error	$\|\zeta\|/\lambda_u$	η_{eff}	Iter	MR error
1	2.852e-01	5.257	15	1.453e-01	1.400e+00	7.633	15	2.611e+00
2	7.054e-03	1.546	15	2.493e-02	4.931e-02	2.481	15	5.822e-02
3	1.577e-04	1.237	15	4.611e-03	5.827e-03	3.631	15	6.227e-03
4	9.526e-06	1.191	11	8.860e-04	6.753e-04	4.172	15	6.843e-04
5					7.938e-05	5.257	9	8.022e-05
6					1.087e-05	1.835	3	9.789e-06
7					8.706e-06	1.497	2	1.232e-06

TABLE 2. Results for control error estimation, estimator effectivity, number optimization iterations and relative $\mathcal{H}_{2,\alpha}$ model reduction error for H2-RI (left) and BT-RI (right).

r	$\|\zeta\|/\lambda_u$	η_{eff}	Iter	MR error	$\|\zeta\|/\lambda_u$	η_{eff}	Iter	MR error
1	1.561e+00	6.745	15	8.487e-01	1.561e+00	6.745	15	8.487e-01
2	1.099e-01	1.897	15	1.255e-01	9.947e-02	1.721	15	1.348e-01
3	1.268e-02	2.551	15	5.248e-02	1.132e-02	2.484	15	5.295e-02
4	3.709e-03	2.044	15	3.554e-02	2.764e-03	1.964	15	3.548e-02
5	2.205e-03	2.106	15	2.261e-02	1.553e-03	2.462	15	2.266e-02
6	8.833e-04	1.525	15	1.344e-02	6.200e-04	1.887	15	1.349e-02
7	5.691e-04	2.187	15	7.625e-03	4.179e-04	2.494	15	7.663e-03
8	1.873e-04	1.538	15	4.188e-03	1.270e-04	1.802	15	4.202e-03
9	1.235e-04	2.320	8	2.352e-03	8.758e-05	2.558	6	2.361e-03
10	3.735e-05	1.529	6	1.405e-03	2.931e-05	1.776	5	1.405e-03
11	2.435e-05	1.737	3	9.671e-04	2.373e-05	2.276	3	9.691e-04
12	1.103e-05	1.231	3	7.367e-04	9.326e-06	1.247	3	7.353e-04
13	1.103e-05	1.231	0	6.089e-04				
14	7.029e-06	1.046	2	5.101e-04				

TABLE 3. Results for control error estimation, estimator effectivity, number optimization iterations and relative $\mathcal{H}_{2,\alpha}$ model reduction error for POD with reference control $u_{\text{ref}}(t) \equiv 0.5$ used for generating snapshots (left: without POD basis update; right: with POD basis update)

tolerance accuracies there is indeed a rather strong dependency on the chosen reference control. Nevertheless, updating the basis after each iteration deletes this dependency.

The choice of α in the H2 reduction is heuristic and numerical tests for different values of α show some dependency on this choice. Furthermore, the accuracy in the number of IRKA iterations affect the form of the reduced system and also the total number of optimization iterations.

The estimator provides an effectivity $\eta_{\text{eff}} \leq 8$. Note that the effectivity turns out to be lower (between 1 and 3) in most of the iterations using POD. Despite this good effectivity for all methods, the error estimator requires two FEM solutions and hence, the main numerical effort in each iteration for 2D or 3D domains.

H2-RE and B2-RE give similar results in each iteration, regarding control error estimation (and hence, number of iterations) as well as the model reduction error.

r	$\|\zeta\|/\lambda_u$	η_{eff}	Iter	MR error	$\|\zeta\|/\lambda_u$	η_{eff}	Iter	MR error
1	3.683e+00	6.720	15	9.380e-01	3.683e+00	6.720	15	9.380e-01
2	1.713e-01	2.812	15	2.168e-01	1.345e-01	1.954	15	1.797e-01
3	7.989e-03	1.988	15	5.320e-02	1.092e-02	2.381	15	5.292e-02
4	9.983e-04	1.795	15	3.459e-02	3.028e-03	2.262	15	3.548e-02
5	4.018e-04	1.545	15	2.596e-02	1.556e-03	2.444	15	2.266e-02
6	1.686e-04	1.285	13	1.526e-02	6.198e-04	1.883	15	1.349e-02
7	8.975e-05	1.377	10	8.442e-03	4.183e-04	2.491	15	7.663e-03
8	4.362e-05	1.348	7	4.552e-03	1.272e-04	1.799	15	4.202e-03
9	2.672e-05	1.585	5	2.513e-03	8.784e-05	2.561	6	2.361e-03
10	1.444e-05	1.266	3	1.470e-03	2.932e-05	1.775	5	1.405e-03
11	9.332e-06	1.069	2	9.941e-04	2.378e-05	2.280	3	9.691e-04
12					9.304e-06	1.243	3	7.353e-04

TABLE 4. Results for control error estimation, estimator effectivity, number optimization iterations and relative $\mathcal{H}_{2,\alpha}$ model reduction error for POD with reference control $u_{\text{ref}}(t) = t$ used for generating snapshots (left: without POD basis update; right: with POD basis update)

r	$\|\zeta\|/\lambda_u$	η_{eff}	Iter	MR error	$\|\zeta\|/\lambda_u$	η_{eff}	Iter	MR error
1	3.005e-01	2.671	15	6.847e-01	3.005e-01	2.671	15	6.847e-01
2	1.061e-01	1.707	15	1.319e-01	7.749e-02	1.575	15	1.280e-01
3	1.205e-02	2.872	15	5.301e-02	1.073e-02	2.211	15	5.297e-02
4	2.341e-03	1.897	15	3.549e-02	2.959e-03	2.161	15	3.548e-02
5	1.558e-03	2.414	15	2.267e-02	1.568e-03	2.491	15	2.266e-02
6	6.200e-04	1.886	15	1.349e-02	6.182e-04	1.888	15	1.349e-02
7	4.181e-04	2.499	15	7.663e-03	4.176e-04	2.498	15	7.663e-03
8	1.268e-04	1.807	15	4.202e-03	1.268e-04	1.805	15	4.202e-03
9	8.765e-05	2.562	6	2.361e-03	8.761e-05	2.562	6	2.361e-03
10	2.934e-05	1.779	5	1.405e-03	2.938e-05	1.784	5	1.405e-03
11	2.378e-05	2.282	3	9.691e-04	2.368e-05	2.269	3	9.691e-04
12	9.244e-06	1.234	3	7.353e-04	9.314e-06	1.244	3	7.353e-04

TABLE 5. Results for control error estimation, estimator effectivity, number optimization iterations and relative $\mathcal{H}_{2,\alpha}$ model reduction error for POD with reference control $u_{\text{ref}}(t) = \bar{u}(t)$ (obtained from FEM for the original problem) used for generating snapshots (left: without POD basis update; right: with POD basis update)

H2-RI however, converges must faster than all other methods whereas BT-RI works only slightly better than BT-RE.

We observe that the relative $\mathcal{H}_{2,\alpha}$-norm error is related to the estimated error in the control in the sense that methods with similar model reduction errors (such as BT-RE and BT-RI or POD with and without update) provide control errors of the same magnitude. Furthermore, the POD reduced system has a larger reduction error than BT-RE and H2-RE and requires more outer loop iterations. However, BT-RI has the largest reduction error in the last iteration (where the control error is low).

Furthermore note that considering a given model reduction error, a corresponding system allows for a lower control error using POD than that obtained by using H2 (which is nearly identical to that obtained by BT).

5.4. **Convective systems.**

5.4.1. *FEM solution.* Let us now consider the case $\eta \neq 0$. As above, we start the discussion with the FEM solution. The data and discretization is chosen as in (20) and the accuracy for the gradient method is $\mathrm{tol}_{\mathrm{G}}$=1e-05. We compute the solution for $\eta = \eta_i$, $i = 1, \ldots, 4$, with $\eta_1 = 0.1$, $\eta_2 = 0.5$, $\eta_3 = -0.1$, $\eta_4 = -0.5$. The solution is depicted in Figure 2. It can be observed that the structure of the optimal control for η_i, $i \neq 4$, is the same as for $\eta = 0$. However, the exit time t_{e}, the vanishing control time t_{v} as well as the time t_{m} of the minimal value and $\bar{u}(t_{\mathrm{m}})$ increase for increasing η. Indeed, for $\eta = \eta_4 = 0.5$, the numerical solution reveals no negative values for the optimal control but $\bar{u}$ monotonically decreases for $t \geq t_{\mathrm{e}}$ with $\bar{u}(T) = 1.3467\text{e-}04$. The state $\bar{y}(0,t)$ approaches the desired state y_{d} faster for smaller values of η revealing that despite of less heating, the convection leads to higher values of $\bar{y}$ at $x = 0$.

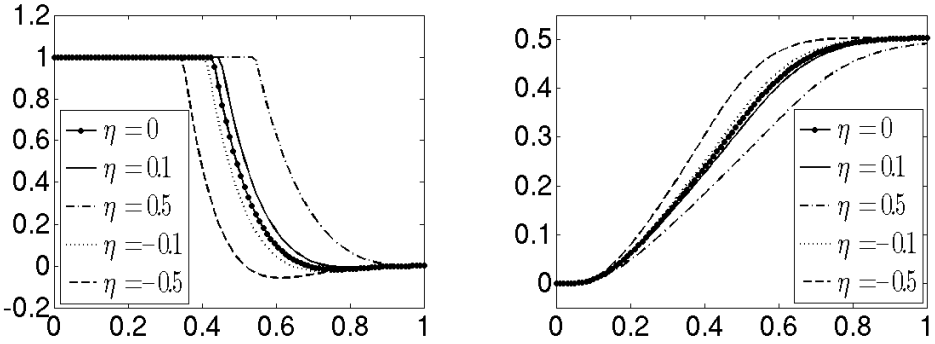

FIGURE 2. Optimal control $\bar{u}(t)$ (left) and state $\bar{y}(0,t)$ (right) for $\eta = \eta_i$,

5.4.2. *Solution using model reduction.* The proposed model reduction techniques are again applied with a tolerance of $\mathrm{tol}_{\mathrm{G}}$=1e-05 and allowing 15 iterations for the gradient method. We choose $\alpha = 0.1$ for the IRKA algorithm. Using a starting guess of $u \equiv 0.5$, we investigate the reference controls $u_{\mathrm{ref}} = u_1(t) \equiv 0.5$, $u_{\mathrm{ref}}(t) = u_2(t) = t$ and $u_{\mathrm{ref}}(t) = \bar{u}(t)$ for POD. BT and H2 are applied for the RE and the RI approach. The number of iterations in the outer loop for each case are depicted in Table 6 where the first, second, third and fourth block (each consisting of five rows) is related to a tolerance of $\mathrm{tol}_{\mathrm{E}}$=1e-02, $\mathrm{tol}_{\mathrm{E}}$=1e-03, $\mathrm{tol}_{\mathrm{E}}$=1e-04 and $\mathrm{tol}_{\mathrm{E}}$=1e-05, respectively.

5.4.3. *Discussions.* The tendencies in the observations are similar to those observed for the $\eta = 0$ case. H2-RE and BT-RE provide similar results. Overall, H2-RI has the smallest number of iterations. BT-RI sometimes works better than BT-RE. However, for strong convection (i.e., $|\eta| = 0.5$), convergence of the model reduction error as well as the value of the error estimator turned out to be so slow such that

η	$\mathrm{tol_E}$	H2		BT		POD					
		RE	RI	RE	RI	no update			with update		
						u_1	u_2	$\bar{u}$	u_1	u_2	$\bar{u}$
0	1e-02	4	2	4	3	4	3	4	4	4	4
0.1	1e-02	3	2	4	6	3	3	4	3	3	3
0.5	1e-02	3	2	3	9	11	11	11	3	3	3
-0.1	1e-02	4	2	4	3	4	4	4	4	4	4
-0.5	1e-02	4	3	4	4	4	3	4	4	4	4
0	1e-03	5	3	5	4	6	4	6	6	6	6
0.1	1e-03	4	3	5	17	14	14	14	6	6	6
0.5	1e-03	4	3	4	14	–	–	–	5	5	5
-0.1	1e-03	5	3	5	4	6	4	10	6	6	6
-0.5	1e-03	5	4	5	4	–	–	–	7	7	7
0	1e-04	6	3	6	5	10	7	9	9	9	9
0.1	1e-04	5	3	6	25	–	–	–	9	9	9
0.5	1e-04	5	4	5	37	–	–	–	8	8	8
-0.1	1e-04	6	4	6	5	–	–	–	9	10	9
-0.5	1e-04	6	4	6	6	–	–	–	10	10	10
0	1e-05	8	4	8	7	14	11	12	12	12	12
0.1	1e-05	6	4	8	–	–	–	–	13	13	13
0.5	1e-05	6	5	6	–	–	–	–	–	–	–
-0.1	1e-05	7	5	7	7	–	–	–	12	12	12
-0.5	1e-05	7	5	7	7	–	–	–	12	12	12

TABLE 6. Number of outer optimization iterations for different values of η and tolerances $\mathrm{tol_E}$ using H2, BT (with RE or RI) and POD(with or without basis updating)

after 40 iterations some small tolerances could not be achieved. This is indicated by a dash in Table 6.

Roughly speaking, H2 and BT converge faster than both versions of POD with and without update. Updating the POD basis brings a certain speed-up in many test cases. In some cases for POD (especially but not exclusively without update), the corresponding eigenvalue for determination of the POD basis vanishes at some saturation value $r = r_{\mathrm{sat}}$. Hence, the algorithm fails to compute the POD basis and small desired tolerances cannot be achieved. In this case, the entry in the table is also given by a dash. Furthermore note that the decay of the error estimator turned out to be not strict at certain iteration numbers for POD.

It can be observed that for increasing values of η, all algorithms converge slower or even fail to converge. Furthermore, the differences in the number of iterations between H2 and BT on the one hand, and POD on the other hand, increase. POD without updating seems to work rather poorly for increasing $|\eta|$. However, updating the basis provides a solution method which can be applied successfully in most test cases.

6. **Conclusions and Outlook.**

This paper deals with numerical solution methods for linear-quadratic parabolic optimal control problems. The focus is on projection based model reduction, in particular Proper Orthogonal Decomposition (POD),

$\mathcal{H}_{2,\alpha}$-norm reduction (H2) and Balanced Truncation (BT) to solve the underlying dynamics efficiently.

We presented a framework for applying the three different model reduction methods based on semi-discretization of the state (and adjoint) equation to an LTI system with occuring control variables (and state variables) as input and relevant states for the cost functional (and for the minimum condition) as output variables. These two LTI systems are reduced to lower-dimensional reduced systems by applying certain Galerkin (for POD) or Petrov-Galerkin (for H2 and BT) projections. The projection matrices are summarized in this paper.

To solve the optimal control problem, an approach from [34] which was used for POD model reduction has been extended to H2 and BT model reduction. The approach involves iterative solutions of certain reduced control problems with increasing dimension r until some desired accuracy in the optimal control is achieved. The latter is verified in an outer loop with a-posteriori error estimation in each iteration. With this general framework, we were able to compare the three model reduction methods in terms of efficiency and performance of the optimal control solution technique by means of numerical tests for a convection-diffusion equation.

Since H2 and BT have so far been used without error estimation and are based on a first-discretize-then-optimize approach (which accounts for only the state equation), this paper makes three contributions. It extends the idea of using H2 and BT model reduction to a first-optimize-then-discretize approach, combines it with a-posteriori error estimation for the optimal control and implements these for comparison with POD which represents as one of the most popular model reduction techniques.

As result of the comparison in the numerical tests, we observe that all three model reduction techniques are well-suited for application. There are, however, some differences such as in the number of iterations of the outer loop. Roughly speaking, H2 and BT provide faster convergence than POD. On the other hand, POD can also be used for nonlinear problems. Furthermore, the reduction can be applied faster than for BT and H2 since the latter involve operations on matrices which depend on the spatial discretization number whereas POD works with matrices which depend on the time discretization number (which is often lower, especially for problems with 2D or 3D domains).

Regarding the performance of the outer loop, H2 and BT provide the possibility of reduction of an implicit or explicit system. The implicit approach H2-RI turned out to be the fastest of all methods and for all test cases whereas BT-RI provided slower convergence and robustness problems in the reduction algorithm for some test cases. The performance of POD depends rather strongly on the chosen reference control to construct the POD basis functions. Furthermore, convection in the system provided poor accuracies in the optimal control. However, by updating the POD basis, we obtain better performance independent on the chosen reference control with increased accuracies in the outer loop. Indeed, convection in the system turned out to decrease efficiency and performance of all solution methods.

In the future, further numerical tests are planned. In particular, it will be interesting to investigate 2D or 3D problems since the differences in the model reduction computational times will be more significant. Furthermore, problems with multidimensional inputs and outputs are of interest. The general framework also gives rise to the idea of applying other reduction techniques to obtain the reduced problem. Further work results from the fact that the error estimator requires solution of the

full state and adjoint system which is the main computational effort, especially for 2D and 3D problems.

REFERENCES

[1] K. Afanasiev and M. Hinze, *Adaptive control of a wake flow using proper orthogonal decomposition*, Lect. Notes Pure Appl. Math., **216** (2001), pp. 317–332.

[2] A.C. Antoulas, "Approximation of large-scale dynamical systems", SIAM, Philadelphia (2005).

[3] N. Arada, E. Casas, and F. Tröltzsch, *Error estimates for the numerical approximation of a semilinear elliptic control problem*, Computational Optimization and Applications , **23** (2002), pp. 201–219.

[4] P. Benner and T. Damm, *Lyapunov Equations, Energy Functionals, and Model Order Reduction of Bilinear and Stochastic Systems*, SIAM Journal on Control and Optimization , **49**, No. 2 (2011), pp. 686–711.

[5] P. Benner and J. Saak, *A Galerkin-Newton-ADI Method for Solving Large-Scale Algebraic Riccati Equations*, SPP1253, http://www.am.uni-erlangen.de/home/spp1253/wiki/index.php/Preprints (2010)

[6] P. Benner and E.S. Quintana-Ortí, *Model reduction based on spectral projection methods*, In Reduction of Large-Scale Systems, P. Benner, V. Mehrmann, D. C. Sorensen (eds.), Lecture Notes in Computational Science and Engineering, **45** (2005), pp. 5-48.

[7] A. Bunse-Gerstner, D. Kubalinska, G. Vossen, and D. Wilczek, *h_2-norm optimal model reduction for large-scale discrete dynamical MIMO systems*, Journal of Computational and Applied Mathematics, **233**, No. 5 (2011), pp. 1202–1216.

[8] A.L. Dontchev, W.W. Hager, A.B. Poore, and B. Yang, *Optimality, stability, and convergence in nonlinear control*, Applied Math. and Optimization , **31** (1995), pp. 297–316.

[9] K. Glover, *All optimal Hankel-norm approximations of linear multi-variable systems and their L_∞ error bounds*, International Journal of Control, **39** (1984), pp. 1115–1193.

[10] M.A. Grepl and M. Kärcher, *Reduced basis a posteriori error bounds for parametrized linear-quadratic elliptic optimal control problems*. C. R. Acad. Sci. Paris, Ser. I, **349** (2011), pp. 873-877.

[11] S. Gugercin, A.C. Antoulas, and C.A. Beattie, *H_2 model reduction for large-scale linear dynamical systems*, SIAM Journal on Matrix Analysis and Applications, **30**, No. 2 (2008), pp. 609–638.

[12] M. Hinze and S. Volkwein, *Error estimates for abstract linear-quadratic optimal control problems using proper orthogonal decomposition*. Computat. Optim. and Appl., **39** (2008), pp. 319-345.

[13] P. Holmes, J.L. Lumley, and G. Berkooz, "Turbulence, Coherent Structures, Dynamical Systems and Symmetry", Cambridge Univ. Press, New York, 1996.

[14] M. Hintermüller, K. Ito, and K. Kunisch. *The primal-dual active set strategy as a semi-smooth Newton method*. SIAM J. Optimization, **13** (2003), pp. 865–888.

[15] C. Joerres, G. Vossen, and M. Herty, *On an inexact gradient method using POD for a parabolic optimal control problem*, submitted, 2011.

[16] E.A. Jonckheere and L.M. Silverman, *A new set of invariants for linear systems – Application to reduced order compensator design* IEEE Trans. Automat. Control , **28**:10 (1983), pp. 953-964.

[17] E. Kammann, F. Tröltzsch, and S. Volkwein, *A method of a-posteriori error estimation with application to proper orthogonal decomposition*, submitted, 2011.

[18] D. Kubalinska, "Optimal interpolation-based model reduction", PhD thesis, University of Bremen, 2008.

[19] K. Kunisch and S. Volkwein, *Galerkin proper orthogonal decomposition methods for parabolic problems*, Numerische Mathematik, **90** (2001), pp. 117–148.

[20] K. Kunisch and S. Volkwein, *Proper orthogonal decomposition for optimality systems*, ESAIM: Mathematical Modelling and Numerical Analysis, **42** (2008), pp. 1-23.

[21] E.N. Lorenz, *Empirical orthogonal functions and statistical weather prediction*, Statistical Forecasting Scientific Rep. 1, Department of Meteorology, Massachusetts Institute of Technology, Cambridge, MA, 1956.

[22] L. Machiels, Y. Maday, I.B. Oliveira, A.T. Patera, and D.V. Rovas, *Output Bounds for Reduced-Basis Approximations of Symmetric Positive Definite Eigenvalue Problems*, CR Acad Sci Paris Series I, **331** (2000), pp. 1531–1548.

[23] K. Malanowski, C. Büskens, and H. Maurer, *Convergence of approximations to nonlinear control problems*, in "Mathematical Programming with Data Perturbation" (eds.: A.V. Fiacco and Marcel Dekker), Inc., New York (1997), pp. 253–284.

[24] H. Maurer and J. Zowe, *First and second-order necessary and sufficient optimality conditions for infinite-dimensional programming problems*, Mathematical Programming, **16**, No. 1, (1979), pp. 98–110.

[25] L. Meier and D. Luenberger, *Approximation of linear constant systems*, IEEE Transactions on Automatic Control, **12**, No. 5 (1967), pp. 585–588.

[26] B.C. Moore, *Principal component analysis in linear systems: controllability, observability and model reduction*, IEEE Trans. Automatic Control , **26**, No. 1 (1981), pp. 17–32.

[27] A.T. Patera and G. Rozza. "Reduced Basis Approximation and A Posteriori Error Estimation for Parametrized Partial Differential Equations", MIT Pappalardo Graduate Monographs in Mechanical Engineering, 2006.

[28] S.S. Ravindran, *Reduced-order adaptive controllers for fluid flows using POD*, SIAM J. Sci. Comput., **15** (2000), pp. 457–478.

[29] J. C. De Los Reyes and T. Stykel, *A balanced truncation based strategy for optimal control of evolution problems*, Optim. Methods Software, **26**(4-5) (2011), pp. 673–694.

[30] J. Saak, "Efficient Numerical Solution of Large Scale Algebraic Matrix Equations in PDE Control and Model Order Reduction", PhD thesis, TU Chemnitz, 2009.

[31] E.W. Sachs and M. Schu, *A priori error estimates for reduced order models in finance*, submitted, 2011.

[32] T. Stykel., *Gramian-based model reduction for descriptor systems. Math.* Control Signals Systems , **16** No. 4 (2004), pp. 297–319.

[33] T. Tonn, K. Urban, and S. Volkwein, *Comparison of the reduced-basis and POD a-posteriori error estimators for an elliptic linear quadratic optimal control problem.* Mathematical and Computer Modelling of Dynamical Systems, Special Issue: Model order reduction of parameterized problems, **17** (2011), pp. 355–369.

[34] F. Tröltzsch and S. Volkwein, *POD a-posteriori error estimates for linear-quadratic optimal control problems*, Computational Optimization and Applications, **44** (2009), 83–115.

[35] F. Tröltzsch. "Optimal Control of Partial Differential Equations. Theory, Methods and Applications", American Math. Society, Providence, volume **112**, 2010.

[36] R. Usmani, *Inversion of a tridiagonal Jacobi matrix*, Linear Algebra Appl. , **212/213** (1994), pp. 413–414.

[37] S. Volkwein, "Model reduction using proper orthogonal decomposition", Lecture Notes, Institute of Mathematics and Statistics, University of Constance, 2011.

[38] S. Volkwein, *Optimality system POD and a-posteriori error analysis for linear-quadratic problems*, submitted, 2011.

[39] G. Vossen, *$\mathcal{H}_{2,\alpha}$-norm optimal model reduction for optimal control problems subject to large-scale dynamical systems with applications to parabolic and hyperbolic evolution equations*, submitted, 2011.

Received xxxx 20xx; revised xxxx 20xx.

E-mail address: Georg.Vossen@nld.rwth-aachen.de
E-mail address: Stefan.Volkwein@uni-konstanz.de