

A SIMPLIFIED NEWTON METHOD TO GENERATE SNAPSHOTS FOR POD MODELS OF SEMILINEAR OPTIMAL CONTROL PROBLEMS *

PAUL MANNS[†] AND STEFAN ULBRICH[‡]

Abstract. In PDE-constrained optimization, proper orthogonal decomposition (POD) provides a surrogate model of a (potentially expensive) PDE discretization, on which optimization iterations are executed. Because POD models usually provide good approximation quality only locally, they have to be updated during optimization. Updating the POD model is usually expensive, however, and therefore often impossible in a model-predictive control (MPC) context. Thus, reduced models of mediocre quality might be accepted. We take the view of a simplified Newton method for solving semilinear evolution equations to derive an algorithm that can serve as an *offline phase* to produce a POD model. Approaches that build the POD model with *impulse response snapshots* can be regarded as the first Newton step in this context.

In particular, POD models that are based on impulse response snapshots are extended by adding a second simplified Newton step. This procedure improves the approximation quality of the POD model significantly by introducing a moderate amount of extra computational costs during optimization or the MPC loop. We illustrate our findings with an example satisfying our assumptions.

Key words. Proper Orthogonal Decomposition, Snapshot Generation, Simplified Newton Method

AMS subject classifications. 65M60,35K20

1. Introduction. Proper Orthogonal Decomposition (POD) is a well-known method to derive low-dimensional reduced-order models of dynamical systems. In the field of optimization of partial differential equations (PDEs), POD is employed as a snapshot-based model order reduction technique to replace expensive finite-element method (FEM) solves of a discretized PDE by computationally cheap surrogates in the optimization iterations; see, for example, [1, 4, 10, 26–28, 33, 38]. Because the control inputs change during the optimization, the quality of the reduced-order model usually deteriorates, and an update or recomputation may become necessary [1, 4]. We propose a POD model that provides increased accuracy for varying controls compared with common snapshot-based approaches.

We summarize the rationale of POD and refer to [19, 28] for details. For a given solution y of the evolution equation, let $\text{span}\{y(t) \mid 0 \leq t \leq T\} \subset H$ be the subspace of interest, where H is a Hilbert space. Now assume that the span of a set of vectors $\{v_1, \dots, v_n\} \subset H$ approximates $\text{span}\{y(t) \mid 0 \leq t \leq T\}$ well for trajectories y of our interest. Then we can compute a (reduced) basis of length $k \leq n$, which minimizes

*Submitted August 4, 2021

Funding: This work was supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Scientific Discovery through the Advanced Computing (SciDAC) Program through the FASTMath Institute under Contract No. DE-AC02-06CH11357. Stefan Ulbrich received support by the German Research Foundation (DFG) within the Collaborative Research Center TRR 154 Project-ID 239904186 - TRR 154 "Mathematical Modelling, Simulation and Optimization using the Example of Gas Networks", project A02, and by the DFG within the Collaborative Research Center SFB 1194 Project-ID 265191195 - SFB 1194 "Interaction between Transport and Wetting Processes", project B04.

[†]Mathematics and Computer Science Division, Argonne National Laboratory, Lemont, IL (pmanns@anl.gov).

[‡]Department of Mathematics, TU Darmstadt, Darmstadt, Germany (ulbrich@mathematik.tu-darmstadt.de).

the squared reconstruction error of the vectors v_1, \dots, v_n , by solving

$$\min_{\psi^1, \dots, \psi^k} \frac{1}{2} \sum_{\ell=1}^n \left\| v_\ell - \sum_{i=1}^k (\psi^i, v_\ell)_H \psi^i \right\|_H^2 \quad \text{s.t.} \quad (\psi^i, \psi^j)_H = \delta_{ij} \text{ for } 1 \leq i \leq j \leq k.$$

Such vectors v_1, \dots, v_n are called *snapshots*. We can solve the least-squares problem with the help of a singular-value decomposition (SVD) of the correlation matrix

$$K = ((v_i, v_j)_H)_{1 \leq i, j \leq n}$$

and a suitable transformation into H yielding n basis vectors (see [28, Sec. 3]). The resulting reduced basis vectors ψ^1, \dots, ψ^n in H and the associated singular values $\lambda_1, \dots, \lambda_n$ that are given by the aforementioned SVD satisfy

$$\sum_{\ell=1}^n \left\| v_\ell - \sum_{i=1}^k (\psi^i, v_\ell)_H \psi^i \right\|_H^2 = \sum_{j=k+1}^n \lambda_j,$$

(see [28, Sec. 3]), which allows for a trade-off between reconstruction accuracy and the number of basis vectors. The computations are faster for fewer basis vectors k . If linear system solves are the bottleneck in the numerical computations, the steps in the optimization procedure that employ the POD model have a complexity of $\mathcal{O}(k^3)$.

The selection of snapshots is crucial when building the POD model. The state iterates move through the state space during optimization, reducing the approximation quality of POD models that were computed for snapshots in different regions. Therefore, the locations of good snapshots may be unknown when starting an optimization procedure, and different strategies have been developed to handle this situation. Hinze and Volkwein [23] optimize the POD model until convergence, compute additional snapshots, and compute a new model from the increased snapshot set. Sachs et al. [4, 10, 33] integrate the update of a POD model in a trust-region globalization strategy. Schmidt et al. [36] optimize the POD model until convergence and compute a new model from information at the final iterate. Bott [12] uses error estimators in a multilevel sequential quadratic programming (SQP) method to trigger model updates. Gubisch and Volkwein [21] increase the number of basis vectors during the optimization iterations.

These adaptive strategies require expensive *offline phases* that update the model and are succeeded by cheap *online phases* until the next *offline phase*. However, this approach may be difficult in the context of model predictive control (MPC), where there might not be enough time or compute resources available for multiple offline phases. Ghiglieri and Ulbrich [18] present an MPC problem, for which they combine uncontrolled and impulse response snapshots—which can both be computed ahead—and keep the POD model fixed during the whole MPC loop. Their article provides a useful insight: the impulse response snapshots are a fundamental solution of the linearized PDE, thereby incorporating properties of convolution representations into the snapshot ensemble. This is the starting point for our investigations in this work.

Convolution representations using impulse responses or *Green's functions* are a common tool for analyzing dynamical systems. Bai and Skoogh [6] consider the Volterra series representation of bilinear dynamical systems. They construct reduced models that match a desired number of moments of the transfer functions of the kernels of the Volterra series. Gu [20] states that Volterra series-based approaches may suffer from bad approximation quality outside a small region around the expansion

point. To alleviate this problem, he proposes to reformulate the polynomial nonlinear system into so-called quadratic-linear differential algebraic equations with larger system size. Then, reduced models are constructed to match a desired number of moments of the transfer functions of the reformulated system. Flagg et al. [16] and Benner et al. [8, 9] derive optimality conditions of the corresponding approximation problems for bilinear and quadratic bilinear systems. For example, in [9], a truncated \mathcal{H}_2 -norm, which includes the first three summands of the Volterra series, of a quadratic bilinear system is minimized. Importantly, the optimality conditions do not depend on any input data of the system and can be satisfied approximately by the system matrices produced by an efficient iterative algorithm.

We propose a novel approach to improve the approximation quality and increase the region of good approximation quality. Assume we want to solve $E(y) = 0$, that is, compute the state for a fixed control. Then we can compute an approximation $y^{(1)} = \bar{y} + d^{(1)}$ of y , where $d^{(1)}$ is the solution of one step of Newton's method,

$$E_y(\bar{y})d^{(1)} = -E(\bar{y}),$$

and E_y denotes the derivative of E with respect to y . Impulse response snapshots yield a high approximation quality of the linear subspace, in which $d^{(1)}$ lives. However, the quality may be poor outside a small neighborhood of the Taylor expansion point \bar{y} . Now, we carry out a simplified second step of the Newton method,

$$E_y(\bar{y})d^{(2)} = -E(\bar{y} + d^{(1)}).$$

The step is *simplified* because we reuse the linearization and update only the right-hand side. We approximate the subspace containing $d^{(2)}$ also by means of additional impulse response snapshots. Using *simplified Newton steps* still yields local convergence and is used in SQP methods as *second-order corrections* [17, 39].

1.1. Contribution. We formalize the described methodology for a class of semilinear evolution equations with linear control inputs. We characterize the orbits of $d^{(1)}$ and $d^{(2)}$ and prove bounds on corresponding POD approximation errors. We show how this allows us to compute the enriched POD bases by solving suitable impulse response problems as well as how the latter occur in time discretizations. We provide computational results that demonstrate the improved approximation properties for a semilinear evolution PDE and a tracking-type optimal control problem (OCP) constrained by it that fit into our framework of assumptions. Furthermore, we outline how Mixed-Integer Optimal Control Problems (MIOCPs) may benefit from the use of the proposed method.

1.2. Structure of the paper. In [section 2](#) we recall the simplified Newton method and its local convergence properties. In [section 3](#) we analyze $d^{(1)}$ and $d^{(2)}$, their orbits and the POD approximation errors. [Section 4](#) transfers the resulting approximation errors to a Galerkin ansatz for the PDE on the POD model. [Section 5](#) states an algorithm that executes the two investigated Newton steps to compute a combined enriched POD basis. [Section 6](#) presents computational results. In [section 7](#) we outline how the proposed method can be used to solve relaxations of MIOCPs. We give concluding remarks in [section 8](#).

2. Simplified Newton method. We begin by stating the simplified Newton method, which is defined for initial vectors $y^{(0)} \in Y$ and operators $F : Y \rightarrow Z$ satisfying [Assumption 2.1](#).

ASSUMPTION 2.1. Let Y, Z be Banach spaces, let $F : Y \rightarrow Z$ be continuously Fréchet differentiable on an open convex neighborhood D of $y^{(0)} \in Y$, and let $F'(y^{(0)}) \in \mathcal{L}(Y, Z)$ be invertible.

Algorithm 2.1 Simplified Newton method

Require: F satisfying [Assumption 2.1](#), $y^{(0)} \in Y$
for $k = 1, \dots$ **do**
 $d^{(k)} \leftarrow \text{SOLVE}(F'(y^{(0)})d^{(k)} = -F(y^{(k-1)}))$
 $y^{(k)} \leftarrow y^{(k-1)} + d^{(k)}$
end for

The following local convergence result is, for example, shown in [24, Sec. 4.2].

PROPOSITION 2.2. Let $y^* \in Y$ be such that $F(y^*) = 0$, let F be continuously differentiable in an open neighborhood of y^* , and let $F'(y^*) \in \mathcal{L}(Y, Z)$ be invertible. Then there exists $\delta > 0$ such that for all $y^{(0)} \in B_\delta(y^*)$, the iterates $(y^{(k)})_{k \in \mathbb{N}}$ produced by [Algorithm 2.1](#) satisfy $\|y^{(k+1)} - y^*\|_Y \leq c \|y^{(k)} - y^*\|_Y$ for some $0 < c < 1$.

Now, we state the approximation of the zero of the state equation achieved by the simplified Newton iteration.

PROPOSITION 2.3. Let [Assumption 2.1](#) and the Lipschitz condition

$$\|F'(y^{(0)})^{-1}(F'(y) - F'(y^{(0)}))\|_{\mathcal{L}(Y, Y)} \leq \omega_0 \|y - y^{(0)}\|_Y \quad \forall y \in D$$

hold for some $\omega_0 > 0$ with $h_0 := \omega_0 \|d^{(1)}\|_Y < 0.5$. Let $\overline{B_r(y^{(0)})} \subset D$ for $r = (1 - \sqrt{1 - 2h_0})/\omega_0$. Then there exists $0 < c < 1$ such that $\|d^{(k+1)}\|_Y \leq c \|d^{(k)}\|_Y$, and for all iterations k it holds that

$$\|F(y^{(k)})\|_Z = \mathcal{O}(\|d^{(k+1)}\|_Y), \quad \text{and} \quad \|F(y^{(k)})\|_Z \leq \|F'(y^{(0)})\|_{\mathcal{L}(Y, Z)} c^k \|d^{(1)}\|_Y.$$

Proof. The iteration reads $-F(y^{(k)}) = F'(y^{(0)})d^{(k+1)}$. This implies $\|F(y^{(k)})\|_Z = \mathcal{O}(\|d^{(k+1)}\|_Y)$. The constant c and the estimate follow from [13, Thm 2.5]. \square

3. Application to evolution equations. We analyze iterations $k = 1$ and $k = 2$ of [Algorithm 2.1](#) for a class of evolution equations. The state equation is $E(y, u) = 0$, where $E : Y \times U \rightarrow Z$ satisfies the following assumption.

ASSUMPTION 3.1. Let U, Y, Z be Banach spaces, and let $E : Y \times U \rightarrow Z$ be continuously Fréchet differentiable and linear with respect to u . Moreover, for all $(\bar{y}, \bar{u}) \in Y \times U$ let $E_y(\bar{y}, \bar{u}) \in \mathcal{L}(Y, Z)$ be invertible.

Now fix some $(\bar{y}, \bar{u}) \in Y \times U$. We will later choose \bar{y} constant in time; see [Assumption 3.4](#). Moreover, often it makes sense to choose (\bar{y}, \bar{u}) as a steady-state solution, that is, $E(\bar{y}, \bar{u}) = 0$, but this is not required. Let some control $u \in U$ be given. In order to compute a solution of $E(y, u) = 0$, the first two steps of [Algorithm 2.1](#) are

$$(3.1) \quad E_y(\bar{y}, \bar{u})d^{(1)} = -E(\bar{y}, u), \quad y^{(1)} := \bar{y} + d^{(1)}, \quad \text{and}$$

$$(3.2) \quad E_y(\bar{y}, \bar{u})d^{(2)} = -E(y^{(1)}, u), \quad y^{(2)} := y^{(1)} + d^{(2)}.$$

In the following, the operator equation $E(y, u) = 0$ represents a semilinear parabolic problem of the form

$$(3.3) \quad \partial_t y(t) - Ay(t) + N(y(t)) = Fu(t), \quad y(0) = y_0,$$

where A is an elliptic spatial operator, N is a nonlinear term of lower order, and F is a control operator. An appropriate setting to ensure [Assumption 3.1](#) will be given below for particular examples.

3.1. Guiding example. The following semilinear initial boundary value problem (IBVP) serves as our guiding example throughout the remainder of the article.

$$(3.4) \quad \begin{cases} \partial_t y(t) - a\Delta y(t) + by(t)^3 - Fu(t) = 0 & \text{on } (0, T) \times \Omega, \\ y = 0 & \text{on } (0, T) \times \partial\Omega, \\ y(0) = y_0 & \text{on } \Omega. \end{cases}$$

Here, $a, b > 0$, Δ denotes the Dirichlet Laplacian, and $\Omega \subset \mathbb{R}^d$, $d \in \{1, 2, 3\}$, is an open domain that is convex or of class C^2 . We set $V := H_0^1(\Omega) \cap H^2(\Omega)$, $W := H_0^1(\Omega)$, $H := L^2(\Omega)$, and $\mathcal{H} := \{y \in L^2(0, T; V) \mid \partial_t y \in L^2(0, T; H)\}$. We work with the following data and spaces:

$$(3.5) \quad \begin{aligned} F &\in H, \quad y_0 \in W, \quad U = L^2(0, T), \quad Z = L^2(0, T; H) \times W, \\ Y &= \{y \in L^\infty(0, T; W) \cap \mathcal{H} \mid y(0) \in W\}, \\ \|y\|_Y &= \|y\|_{L^\infty(0, T; W)} + \|y\|_{\mathcal{H}} + \|y(0)\|_W. \end{aligned}$$

Note that Y is well defined because of the continuous embedding $\mathcal{H} \hookrightarrow C([0, T]; W)$; see [Appendix A](#). We have the following existence and uniqueness result.

PROPOSITION 3.2. *Consider the setting (3.5). Then for any $u \in L^2(0, T)$ there exists a unique solution $y \in Y \hookrightarrow C([0, T]; W)$ of (3.4).*

Proof. We apply [7, Prop. 5.1] with $\beta = \partial g$, where $g : \mathbb{R} \rightarrow \mathbb{R}$, $g(s) = bs^4/4$, and $f = Fu \in L^2(0, T; H)$. Then $\beta : \mathbb{R} \rightarrow \mathbb{R}$, $\beta(s) = bs^3$ is continuous and monotonically nondecreasing, and thus the induced graph is maximally monotone, where the domain satisfies $\overline{D(\beta)} = \mathbb{R}$ and thus satisfies the assumptions of [7, Prop. 5.1]. Since the embedding $W \hookrightarrow L^6(\Omega)$ is continuous for $d = 1, 2, 3$, we have $g(y_0) \in L^1(\Omega)$. Now [7, Prop. 5.1] and the continuous embedding $\mathcal{H} \hookrightarrow C([0, T]; W)$ yield the assertion.

A formal calculation shows that the following problems have to be solved for the two simplified Newton steps given in (3.1), and (3.2):

$$(3.6) \quad \begin{cases} (\partial_t - a\Delta + 3b\bar{y}^2)d^{(1)} &= -(\partial_t \bar{y} - a\Delta \bar{y} + b\bar{y}^3 - Fu), \\ d^{(1)}(0) &= y_0 - \bar{y}(0), \quad d^{(1)}|_{(0, T) \times \partial\Omega} = 0 \end{cases}$$

$$(3.7) \quad \begin{cases} (\partial_t - a\Delta + 3b\bar{y}^2)d^{(2)} &= -(b(y^{(1)})^3 - b\bar{y}^3 - 3b\bar{y}^2 d^{(1)}), \\ d^{(2)}(0) &= 0, \quad d^{(2)}|_{(0, T) \times \partial\Omega} = 0. \end{cases}$$

To make this rigorous, we introduce the operator E :

$$(3.8) \quad E : Y \times U \rightarrow Z, \quad E(y, u) := \begin{pmatrix} \partial_t y - a\Delta y + by^3 - Fu \\ y(0) - y_0 \end{pmatrix},$$

and show that it satisfies [Assumption 3.1](#).

PROPOSITION 3.3. *Let E be given by (3.8). Then for any $\bar{u} \in U$ the equation $E(\bar{y}, \bar{u}) = 0$ has a unique solution $\bar{y} \in Y$. Moreover, E is continuously Fréchet differentiable. For $(y, u) \in Y \times U$ and any $(v, w) \in Y \times U$ it holds that*

$$E_y(y, u)v = \begin{pmatrix} \partial_t v - a\Delta v + 3by^2v \\ v(0) \end{pmatrix}, \quad \text{and} \quad E_u(y, u)w = \begin{pmatrix} -Fw \\ 0 \end{pmatrix}.$$

Moreover, $E_y(y, u) \in \mathcal{L}(Y, Z)$ has a bounded inverse.

Proof. By the definition of Y, U, Z all linear parts of (3.8) are in $\mathcal{L}(Y, Z)$. Moreover, the mapping $N : y \in Y \mapsto y^3 \in L^2(0, T; H)$ is continuously Fréchet differentiable. In fact, the trilinear form $B : (y_1, y_2, y_3) \in Y \times Y \times Y \mapsto y_1 y_2 y_3 \in L^2(0, T; H)$ is bounded because

$$\begin{aligned} \|y_1(t)y_2(t)y_3(t)\|_H &\leq \|y_1(t)\|_{L^6(\Omega)} \|y_2(t)\|_{L^6(\Omega)} \|y_3(t)\|_{L^6(\Omega)} \\ &\leq C_1^3 \|y_1(t)\|_W \|y_2(t)\|_W \|y_3(t)\|_W \end{aligned}$$

for some $C_1 > 0$ and thus

$$\|y_1 y_2 y_3\|_{L^2(0, T; H)} \leq \sqrt{T} \|y_1 y_2 y_3\|_{L^\infty(0, T; H)} \leq \sqrt{T} C^3 \|y_1\|_Y \|y_2\|_Y \|y_3\|_Y.$$

Hence, B is infinitely many times continuously Fréchet differentiable, and by the chain rule also $N(y) = B(y, y, y)$ with derivative $v \in Y \mapsto 3B(y, y, v) = 3y^2 v \in L^2(0, T; H)$.

Furthermore, since $3by^2 \in L^\infty(0, T; L^3(\Omega))$, we have for $v_1, v_2 \in W$

$$\int_{\Omega} v_1 3by(t)^2 v_2 dx \leq 3bC_2^2 \|y\|_Y^2 \|v_1\|_{L^3(\Omega)} \|v_2\|_{L^3(\Omega)} \leq 3bC_2^2 C_3^2 \|y\|_Y^2 \|v_1\|_W \|v_2\|_W$$

for some $C_2, C_3 > 0$. Hence,

$$a(t; v_1, v_2) = (\nabla v_1, \nabla v_2)_{H^a} + \int_{\Omega} v_1 3by(t)^2 v_2 dx$$

defines uniformly in t a bounded and coercive bilinear form on $W \times W$. Standard parabolic theory yields a unique solution of $E_y(y, u)v = z$ for all $z \in Z$ with $v \in \mathcal{W} := \{w \in L^2(0, T; W) \mid \partial_t w \in L^2(0, T; H^{-1}(\Omega))\}$ with $\|v\|_{\mathcal{W}} \leq C \|z\|_Z$ for a constant C independent of z . Then $3by^2 v \in L^2(0, T; H)$ and thus $v \in Y$ by [Proposition 3.2](#). Hence, $E_y(y, u)^{-1} \in \mathcal{L}(Z, \mathcal{W})$ and $E_y(y, u)^{-1} : Z \rightarrow Y$. Now $E_y(y, u)^{-1} \in \mathcal{L}(Z, Y)$ follows from the closed graph theorem. Alternatively, one can apply standard parabolic regularity theory; see, for example, [15, 7.1, Thm. 5]. \square

Hence, for the setting (3.5) we have justified that the formally derived simplified Newton steps in (3.1) and (3.2) are well defined and have the desired regularities.

Since $by^3 \in L^2(0, T; H)$ for $y \in Y$, the uniqueness of the mild solution of $E(y, u) = 0$ and a bootstrapping argument imply that y can be represented by the variation of constants formula

$$(3.9) \quad y(t) = S(t)y_0 + \int_0^t S(t-s)(Fu(s) - by^3(s)) ds,$$

where $(S(t))_{t \geq 0}$ denotes the strongly continuous semigroup generated by the Dirichlet Laplacian (scaled by $a > 0$) on H . Similarly, the solution $v \in Y$ of $E_y(y, u)v = (z, 0)$ for $(z, 0) \in Z$ can be represented by

$$(3.10) \quad v(t) = \int_0^t S(t-s)(z(s) - 3by^2(s)v(s)) ds.$$

3.2. General framework. This section provides our standing assumptions on the considered IBVPs (3.3). The guiding example presented above meets the assumptions. We associate with (3.3) an operator E ,

$$(3.11) \quad E : Y \times U \rightarrow Z, \quad E(y, u) = \begin{pmatrix} \partial_t y - Ay + N(y) - Fu \\ y(0) - y_0 \end{pmatrix},$$

where boundary conditions are included in the definition of Y . We will work under the following assumption.

ASSUMPTION 3.4. Let $V \hookrightarrow W \hookrightarrow H$ be Hilbert spaces with dense imbeddings. Assume that with $y_0 \in W$, $F \in H$, $U = L^2(0, T)$ and appropriate spaces $Y \hookrightarrow L^2(0, T; V) \cap C([0, T], H)$, $Z \hookrightarrow L^2(0, T; H) \times W$ the operator E defined in (3.11) satisfies Assumption 3.1. Moreover, for any time-independent state $\bar{y} \in Y$ the operator $B := A - N_y(\bar{y}) : D(B) \rightarrow H$ generates a strongly continuous semigroup $(T(t))_{t \geq 0}$.

Remark 3.5.

- If $A : D(A) \rightarrow H$ generates a strongly continuous semigroup and $C := -N_y(\bar{y}) \in \mathcal{L}(H, H)$, then $B = A + C$ generates a strongly continuous semigroup on H ; see, for example, Corollary 3.5.6 in [3]. Thus $(T(t))_{t \geq 0}$ is well defined.

However, any setting where $A + C$ generates a strongly continuous semigroup is allowed. This is, for example, the case for the perturbation $C = \bar{y} \cdot \nabla$ (Oseen semigroup) if A is the Stokes operator; see [31].

- By Assumption 3.4, for any time-independent state $\bar{y} \in Y$ and all $\bar{u} \in U$, $(z, v_0) \in Z$ the linearized equation $E_y(\bar{y}, \bar{u})v = (z, v_0)$ has a unique solution $v \in Y$. Semigroup theory allows one to represent v as the unique mild solution

$$(3.12) \quad v(t) = T(t)v_0 + \int_0^t T(t-s)z(s) ds.$$

- It makes sense to consider linearizations at stationary states \bar{y} . For example, since $a > 0$ and $b > 0$ in our guiding example we may expect a damping behavior toward a stationary state. Moreover, in many applications a stabilization around a stationary state by optimal control is relevant.

3.3. A convolution formula for the first Newton step. We investigate the first simplified Newton step and derive a convolution formula for $d^{(1)}(t)$. We fix a linearization point $\bar{y} \in Y$ that is constant in time, that is, $\bar{y}(t) = \bar{y}_0$ for some $\bar{y}_0 \in V$.

The first Newton step (3.1) for (3.11) with $C = -N_y(\bar{y})$ is

$$(3.13) \quad \begin{pmatrix} (\partial_t - A - C)d^{(1)} \\ d^{(1)}(0) \end{pmatrix} = \begin{pmatrix} A\bar{y} - N(\bar{y}) + Fu \\ y_0 - \bar{y}_0 \end{pmatrix}.$$

We will show that the solution $d^{(1)} \in Y$ of (3.13) can be computed from the solutions of the following problems:

$$(3.14) \quad \begin{pmatrix} (\partial_t - A - C)v \\ v(0) \end{pmatrix} = \begin{pmatrix} A\bar{y} - N(\bar{y}) \\ y_0 - \bar{y}_0 \end{pmatrix}, \text{ and}$$

$$(3.15) \quad \begin{pmatrix} (\partial_t - A - C)w \\ w(0) \end{pmatrix} = \begin{pmatrix} 0 \\ F \end{pmatrix}.$$

Because of the structure of (3.15), we call $w \in C([0, T]; H)$ the impulse response for the right-hand side (impulse) F . Now the following holds.

LEMMA 3.6. Let Assumption 3.4 hold. Let $v \in Y$ solve (3.14), and let $w \in C([0, T]; H)$ solve (3.15). Then the solution $d^{(1)} \in Y$ of (3.13) is given by

$$d^{(1)}(t) = v(t) + \int_0^t w(t-s)u(s) ds.$$

If $F \in W$, then we have $w \in Y$.

Proof. By **Assumption 3.4** the problem (3.13) has a unique solution $d^{(1)} \in Y$ that can also be represented as a mild solution (see (3.12)):

$$d^{(1)}(t) = T(t)(y_0 - \bar{y}_0) + \int_0^t T(t-s)(A\bar{y} - N(\bar{y}) + Fu(s)) ds.$$

Now set $v(t) = T(t)(y_0 - \bar{y}_0) + \int_0^t T(t-s)(A\bar{y} - N(\bar{y})) ds$ and $w(t) = T(t)F$. Then v is the mild solution of (3.14), w is the mild solution of (3.15), and the claimed representation of $d^{(1)}(t)$ follows. Moreover, if $F \in W$, then **Assumption 3.4** implies that w solving (3.15) is in Y . \square

3.4. A discrete convolution formula for the first Newton step. We consider now a θ -scheme for time discretization that comprises the implicit Euler scheme ($\theta = 1$) and the Crank–Nicolson scheme ($\theta = 1/2$).

Let $0 = t_0 < \dots < T_K = T$, $\Delta t = t_{k+1} - t_k$ be a uniform time grid, and $(u_k)_{k \in \{0, \dots, K-1\}}$ be an interval-wise constant discretization of the control u . We approximate (3.3) by

$$\frac{y_{k+1} - y_k}{\Delta t} - A(\theta y_{k+1} + (1-\theta)y_k) + \theta N(y_{k+1}) + (1-\theta)N(y_k) + Fu_k, \quad 0 \leq k < K,$$

where $\theta \in [\frac{1}{2}, 1]$. Then the discrete analogue of (3.13) is

$$(3.16) \quad \begin{aligned} \frac{d_{k+1}^{(1)} - d_k^{(1)}}{\Delta t} - (A+C)(\theta d_{k+1}^{(1)} + (1-\theta)d_k^{(1)}) &= A\bar{y} - N(\bar{y}) + Fu_k, \quad 0 \leq k < K, \\ d_0^{(1)} &= y_0 - \bar{y}_0, \end{aligned}$$

and the one of (3.14) is

$$(3.17) \quad \begin{aligned} \frac{v_{k+1} - v_k}{\Delta t} - (A+C)(\theta v_{k+1} + (1-\theta)v_k) &= A\bar{y} - N(\bar{y}), \quad 0 \leq k < K, \\ v_0 &= y_0 - \bar{y}_0. \end{aligned}$$

Now consider the following discretization of (3.15):

$$(3.18) \quad \begin{aligned} \frac{w_{k+1} - w_k}{\Delta t} - (A+C)(\theta w_{k+1} + (1-\theta)w_k) &= 0, \quad 0 \leq k < K, \\ (I + \Delta t(1-\theta)(A+C))w_0 &= F. \end{aligned}$$

Then we obtain the following discrete convolution formula for the θ -scheme.

PROPOSITION 3.7. *Let (v_k) and (w_k) solve (3.17) and (3.18). Then the first Newton step $(d_k^{(1)})$ for the θ -scheme (3.16) can be represented by*

$$d_k^{(1)} = v_k + \Delta t \sum_{j=0}^{k-1} w_{k-j} u_j,$$

where we use the convention that the sum vanishes for $k = 0$.

Proof. Set $e_k := \Delta t \sum_{j=0}^{k-1} w_{k-j} u_j$. Then $e_0 = 0$, and by using (3.18) we have

$$\begin{aligned} \frac{e_{k+1} - e_k}{\Delta t} &= \sum_{j=0}^k (w_{k+1-j} - w_{k-j}) u_j + w_0 u_k \\ &= (A + C) \Delta t \sum_{j=0}^k (\theta w_{k+1-j} + (1 - \theta) w_{k-j}) u_j + w_0 u_k \\ &= (A + C) (\theta e_{k+1} + (1 - \theta) e_k) + (I + (A + C) \Delta t (1 - \theta)) w_0 u_k \\ &= (A + C) (\theta e_{k+1} + (1 - \theta) e_k) + F u_k. \end{aligned}$$

Now by superposition $v_k + e_k$ satisfies (3.16) as asserted. \square

3.5. Subspace characterization and approximation of the first Newton step. We use the convolution formula to characterize the orbit of the first Newton step for arbitrary controls $u \in U$. This will be exploited to obtain the reduced basis for the POD model proposed in this article. To state the result precisely, we need further notation. For a function $f \in L^p(0, T; X)$ for some Banach space X we write

$$f([0, T]) := \bigcap_{N \subset [0, T], \lambda(N) = 0} \overline{\{f(t) \mid t \in [0, T] \setminus N\}}^X,$$

where λ denotes the Lebesgue measure and we call $f([0, T])$ the *essential range* of f .

THEOREM 3.8. *Let \bar{y} be a given linearization point of E , v solve (3.14), w solve (3.15), $u \in U$ be arbitrary, and $d^{(1)}$ solve (3.13). Then*

$$d^{(1)}(t) - v(t) \in \overline{\text{span } w([0, T])}^W.$$

The discrete analogs $(d_k^{(1)})$, (v_k) , and (w_k) for the θ -scheme (3.16), (3.17), and (3.18) satisfy $d_k^{(1)} - v_k \in \text{span}\{w_1, \dots, w_k\}$.

Proof. The trajectories y_1 , \bar{y} and z are mild solutions and continuous accordingly. Thus, the pointwise evaluation makes sense. Employing Lemma 3.6, we observe that

$$d^{(1)}(t) - v(t) = \int_0^t w(t-s) u(s) \, ds = \int_0^t w(s) u(t-s) \, ds$$

with $u(t-s) \in \mathbb{R}$. Because $w \in L^2(0, T; W)$ and $u \in L^2(0, T)$, it follows that the integrand in the convolution formula above is in $L^1(0, T; W)$ for all $t \in [0, T]$. Thus, a vector-valued version of the mean value theorem for Bochner integrals (see [14, Cor. II.8]) (after replacing all *for all* statements in its proof by *for almost all*) yields

$$d^{(1)}(t) - v(t) \in \overline{t \text{ conv } f_t([0, t])}^W,$$

where $f_t \in L^1(0, t; W)$ with $f_t(s) := w(s) u(t-s)$ for a.a. $s \in [0, t]$ and all $t \in [0, T]$. Because u is \mathbb{R} -valued, it follows that $f_t([0, t]) \subset \text{span } w([0, T])$, which closes the argument. The span for the discrete trajectories follows from Proposition 3.7. \square

To approximate $d^{(1)}$, we consider a POD approximation of w in $L^2(0, T; W)$ of rank $n \in \mathbb{N}$. That is, we seek to bound the approximation error

$$(3.19) \quad \begin{aligned} \min_{\psi^1, \dots, \psi^n} \frac{1}{2} \int_0^T \left\| w(t) - \sum_{i=1}^n (\psi^i, w(t))_W \psi^i \right\|_W^2 dt \\ \text{s.t. } (\psi^i, \psi^j)_W = \delta_{ij} \text{ for all } 1 \leq i \leq j \leq n. \end{aligned}$$

To this end, we adapt the POD approximation from [28, Sec.3]. Let the operator $\mathcal{Y} : L^2(0, T; \mathbb{R}) \rightarrow W$ be defined as $\mathcal{Y}\varphi := \int_0^T \varphi(t)w(t) dt$. Its adjoint $\mathcal{Y}^* : W \rightarrow L^2(0, T; \mathbb{R})$ is $(\mathcal{Y}^*f)(t) = (f, w(t))_W$ for a.a. $t \in (0, T)$. Defining $\mathcal{R} := \mathcal{Y}\mathcal{Y}^*$ yields

$$\mathcal{R}z = \int_0^T (z, w(t))_W w(t) dt.$$

Then we can characterize the POD approximation by means of the spectrum of \mathcal{R} .

PROPOSITION 3.9. *Let the assumptions of [Theorem 3.8](#) hold. Then there exists an orthonormal basis $(\psi^i)_{i \in \mathbb{N}}$ of W and $(\lambda_i)_{i \in \mathbb{N}} \subset [0, \infty)$ such that $\mathcal{R}\psi^i = \lambda_i\psi^i$ for all $i \in \mathbb{N}$ and $\lambda_i \rightarrow 0$. Moreover, it follows that*

$$\int_0^T \|w(t)\|_W^2 dt = \sum_{i=1}^{\infty} \lambda_i$$

and for all $n \in \mathbb{N}$ it holds that

$$\int_0^T \left\| w(t) - \sum_{i=1}^n (\psi^i, w(t))_W \psi^i \right\|_W^2 dt = \sum_{i=n+1}^{\infty} \lambda_i.$$

Proof. This follows from the analysis in Section 3 of [28], in particular the Hilbert–Schmidt theorem applied to \mathcal{R} , with the choice $X = W$ if we are able to show that the mapping \mathcal{Y}^* is compact (see the 2nd paragraph on page 498 in [28]) for $w \in L^2(0, T; W)$. Let $f \in B$ and $B \subset W$ be bounded, that is, $C := \sup_{f \in B} \|f\|_W < \infty$. We obtain

$$\begin{aligned} \sup_{f \in B} |(\mathcal{Y}^*f)(t) - (\mathcal{Y}^*f)(t+h)| &= \sup_{f \in B} |(f, w(t) - w(t+h))_W| \\ &\leq C \|w(t) - w(t+h)\|_W \end{aligned}$$

for a.a. $t, t+h \in (0, T)$ using the Cauchy–Schwarz inequality. Because $\int_0^{T-h} \|w(t) - w(t+h)\|_W^2 dt \rightarrow 0$ for $h \rightarrow 0$, it holds that

$$\sup_{f \in B} \int_0^{T-h} |(\mathcal{Y}^*f)(t) - (\mathcal{Y}^*f)(t+h)|^2 dt \leq C^2 \int_0^{T-h} \|w(t) - w(t+h)\|_W^2 dt \rightarrow 0$$

for $h \rightarrow 0$, which shows equicontinuity of \mathcal{Y}^* with respect to $L^2(0, T; \mathbb{R})$. We can hence apply the Riesz–Kolmogorov compactness theorem [32, 37] to deduce that $\mathcal{Y}^*(B)$ is a compact set, which implies that \mathcal{Y}^* is a compact operator. \square

To use this approximation in the remainder, we introduce the following notation. Let ψ^1, \dots, ψ^n be an orthonormal subset of W . Then for $f \in L^2(0, T; W)$, we define the (pointwise a.e.) orthogonal projection

$$\Pi_\psi f(t) := \sum_{i=1}^n (\psi^i, f(t))_W \psi^i.$$

The argument above does not depend on the function w , and the function v can be approximated analogously. Therefore, we consider a joint reduced basis for w and v in the remainder. We consider the projection $\Pi_\psi d^{(1)}$ of $d^{(1)}$ on the reduced basis

$$\Pi_\psi d^{(1)}(t) = \sum_{i=1}^n \left(\psi_i, v(t) + \int_0^t w(s)u(t-s) ds \right)_W \psi_i.$$

We denote the corresponding approximation of the first Newton step as

$$y_\psi^{(1)} := \bar{y} + \Pi_\psi d^{(1)}$$

and denote the projection error, which can be driven to zero by [Proposition 3.9](#), as

$$e_1 := \|d^{(1)} - \Pi_\psi d^{(1)}\|_{L^2(0,T;W)}.$$

We summarize the resulting approximation quality $y_\psi^{(1)}$ below.

COROLLARY 3.10. *Let the assumptions of [Theorem 3.8](#) hold. Let $y \in Y$ solve [\(3.3\)](#). Then*

$$\|y - y_\psi^{(1)}\|_{L^2(0,T;W)} \leq \|y - y^{(1)}\|_{L^2(0,T;W)} + e_1.$$

Let $(\psi^i)_{i \in \mathbb{N}}$ and $(\lambda_i)_{i \in \mathbb{N}}$ be as in [Proposition 3.9](#). Then

$$\|y - y_\psi^{(1)}\|_{L^2(0,T;W)} \leq \|y - y^{(1)}\|_{L^2(0,T;W)} + \sum_{i=n+1}^{\infty} \lambda_i.$$

Proof. The claim follows from $y^{(1)} = \bar{y} + d^{(1)}$, [Lemma 3.6](#), and [Proposition 3.9](#) applied to $d^{(1)}$ instead of w . \square

Remark 3.11. This means that the error of approximating y with a projection of the first Newton step to the reduced space is bounded by the sum of the error of the Newton step and the POD approximation error of v and the impulse response w , which are both independent of the control input.

3.6. Subspace characterization of the second simplified Newton step.

We consider again the fixed linearization point $\bar{y} \in Y$, where \bar{y} is constant in time, that is, $\bar{y}(t) = \bar{y}_0$ for some $\bar{y}_0 \in V$. We recall the second simplified Newton step

$$E_y(\bar{y}, \bar{u})d^{(2)} = -E(y^{(1)}, u), \quad y^{(2)} := y^{(1)} + d^{(2)}.$$

Applying this to the nonlinear operator E defined in [\(3.11\)](#), we obtain

$$(3.20) \quad \begin{pmatrix} (\partial_t - A - C)d^{(2)} \\ d^{(2)}(0) \end{pmatrix} = \begin{pmatrix} N(\bar{y}) - Cd^{(1)} - N(y^{(1)}) \\ 0 \end{pmatrix},$$

which follows after inserting that the first Newton step $d^{(1)}$ solves [\(3.13\)](#) and the fact $\partial_t \bar{y} = 0$ into the definition of E .

LEMMA 3.12. *Let [Assumption 3.4](#) hold. Then the solution $d^{(2)} \in Y$ of [\(3.20\)](#) is given by*

$$(3.21) \quad d^{(2)}(t) = \int_0^t T(t-s) \left(N(\bar{y}) - Cd^{(1)} - N(y^{(1)}) \right) ds.$$

Proof. [Assumption 3.4](#) implies that [\(3.20\)](#) has a unique solution $d^{(2)} \in Y$ that is a mild solution and can be represented with the variation of constants formula. \square

Next, we assume that an orthonormal subset $\psi^1, \dots, \psi^M \subset W$ is given, and we aim to characterize the solution $d^{(2)}[\psi]$ of [\(3.20\)](#) for the case that $d^{(1)}$ and $y^{(1)}$ have been replaced by the approximations $y_\psi^{(1)}$ and $\Pi_\psi d^{(1)}$ obtained in [subsection 3.5](#).

We restrict our analysis to the case that N is the superposition operator of a polynomial with degree $p \in \mathbb{N}$. The set of monomials $\{1, x, \dots, x^p\}$ constitutes a basis of the polynomials, which implies

$$N(y_\psi^{(1)})(t) \in \bigcup_{i=1}^p \text{span} \{ \Pi_{j=1}^i b_j \mid b_1, \dots, b_i \in \{\bar{y}, \psi^1, \dots, \psi^n\} \} =: \mathcal{C},$$

where we further require that $\mathcal{C} \subset H$ and deduce that there are orthonormal vectors c_1, \dots, c_m —e.g. obtained by Gram–Schmidt orthonormalization—such that we may write $\mathcal{C} = \text{span}\{c_1, \dots, c_m\}$. Note that $\mathcal{C} \subset H$ is satisfied for our guiding example because of the continuous embedding $W \hookrightarrow L^6(\Omega)$. In particular we obtain

$$N(y_\psi^{(1)})(t) = \sum_{j=1}^m \left(c^j, N(y_\psi^{(1)})(t) \right)_H c^j$$

for all $t \in [0, T]$.

We consider (3.20) with the approximations $\Pi_\psi d^{(1)}$ and $y_\psi^{(1)}$ substituted for $d^{(1)}$ and $y^{(1)}$. Then the representation (3.21) and the subspaces $\text{span}\{\psi^1, \dots, \psi^n\}$ and \mathcal{C} give rise to the initial value problems

$$(3.22) \quad \left\{ \begin{array}{l} \left((\partial_t - A - C)\beta^i \right) = \begin{pmatrix} 0 \\ -C\psi^i \end{pmatrix} \text{ for } i \in \{1, \dots, n\}, \\ \beta^i(0) \end{array} \right.$$

$$(3.23) \quad \left\{ \begin{array}{l} \left((\partial_t - A - C)\gamma^j \right) = \begin{pmatrix} 0 \\ -c_j \end{pmatrix} \text{ for } j \in \{1, \dots, m\}. \\ \gamma^j(0) \end{array} \right.$$

Similar to the first Newton step, we can now characterize the subspace that contains the orbit of the second simplified Newton step if $d^{(1)}$ has already been reduced by means of a POD approximation.

THEOREM 3.13. *Let \bar{y} be a given linearization point of E . Let N be the superposition operator of a polynomial of such that $N \in C(W, H)$. Let $\{\psi^1, \dots, \psi^n\}$, \mathcal{C} , (3.22), and (3.23) be as introduced above. Let $d^{(2)}[\psi]$ solve (3.20) with the approximations $\Pi_\psi d^{(1)}$ and $y_\psi^{(1)}$ substituted for $d^{(1)}$ and $y^{(1)}$. Then for all $t \in [0, T]$ we have*

$$d^{(2)}[\psi](t) = r(t) + b(t) + c(t)$$

with

$$\begin{aligned} r(t) &= \int_0^t T(t-s)N(\bar{y}) \, ds, \\ b(t) &\in \overline{\text{span} \left\{ \bigcup_{i=1}^n \beta^i([0, T]) \right\}}^W, \text{ and} \\ c(t) &\in \overline{\text{span} \left\{ \bigcup_{j=1}^m \gamma^j([0, T]) \right\}}^W, \end{aligned}$$

where $\beta^i([0, T])$ and $\gamma^j([0, T])$ denote the essential ranges of β^i and γ^j .

Proof. With the analysis above we have that (3.22) and (3.23) admit unique solutions $\beta_i, \gamma_j \in L^2(0, T; W)$ for $i \in \{1, \dots, n\}$, and $j \in \{1, \dots, m\}$. Lemma 3.12 implies that $d^{(2)}[\psi](t) = r(t) + b(t) + c(t)$ holds with r as claimed:

$$b(t) = \int_0^t \sum_{i=1}^n \beta^i(s) u^i(t-s) ds, \quad \text{and} \quad c(t) = \int_0^t \sum_{j=1}^m \gamma^j(s) v^j(t-s) ds.$$

Repeating the argument from the proof of Theorem 3.8, we obtain

$$b(t) \in t \operatorname{conv} \left\{ \overline{\bigcup_{i=1}^n f_t^i([0, t])} \right\}^W \quad \text{and} \quad c(t) \in t \operatorname{conv} \left\{ \overline{\bigcup_{j=1}^m g_t^j([0, t])} \right\}^W,$$

where $f_t^i(s) := \beta^i(s) u^i(t-s)$ and $g_t^j(s) := \gamma^j(s) v^j(t-s)$ for a.a. $s \in [0, t]$ and all $t \in [0, T]$. \square

Remark 3.14. \mathcal{C} is generated by the sets of k -combinations (for $k = 1, \dots, p+1$) of (basis) vectors $\bar{y}, \psi^1, \dots, \psi^n$, which grows excessively with n and p . Therefore it may be advisable to reduce the basis c^1, \dots, c^m with POD as well.

3.7. Subspace approximation of the second simplified Newton step. We consider ψ^1, \dots, ψ^n and e_1 as in subsection 3.5. The error estimates below depend on the approximation error of the nonlinear operator N at $y^{(1)}$, which we define as

$$\ell(e_1, y^{(1)}) := \|N(y^{(1)}) - N(y_\psi^{(1)})\|_{L^2(0, T; H)}.$$

We briefly show how an estimate on $\ell(e_1, y^{(1)})$ can be derived for our guiding example.

Example 3.15. We consider the POD approximation of $d^{(1)}$ analyzed in subsection 3.5 and N defined as $N(\eta) := \eta^3$. For brevity of the presentation, we assume that $\bar{y} \in \operatorname{span}\{\psi^1, \dots, \psi^n\}$, and we define $y := y^{(1)}$ and $y_\psi := y_\psi^{(1)}$.

For a.a. $t \in [0, T]$, we obtain

$$\begin{aligned} \|y(t)^3 - y_\psi(t)^3\|_{L^2} &\leq \|y(t)^2 + y(t)y_\psi(t) + y_\psi(t)^2\|_{L^3} \|y(t) - y_\psi(t)\|_{L^6} \\ &\leq 3 \|y(t)\|_{H_0^1}^2 \|y(t) - y_\psi(t)\|_{H_0^1}, \end{aligned}$$

where Hölder's inequality yields the first inequality. The second inequality follows from the fact that $y_\psi(t) = \Pi_\psi y(t)$ and thus $\|y_\psi(t)\|_{H_0^1} \leq \|y(t)\|_{H_0^1}$ and the embedding $H_0^1(\Omega) \hookrightarrow L^6(\Omega)$. We integrate over both sides and use Hölder's inequality to obtain

$$\begin{aligned} \int_0^T \|y(t)^3 - y_\psi(t)^3\|_{L^2}^2 dt &\leq \int_0^T 3 \|y(t)\|_{H_0^1}^4 \|y(t) - y_\psi(t)\|_{H_0^1}^2 dt \\ &\leq 3 \|y\|_{C([0, T]; H_0^1)}^4 \int_0^T \|y(t) - y_\psi(t)\|_{H_0^1}^2 dt \\ &\leq 3 \|y\|_{C([0, T]; H_0^1)}^4 e_1^2, \end{aligned}$$

which yields the estimate $\ell(e_1, y^{(1)}) \leq \sqrt{3} \|y^{(1)}\|_{C([0, T]; H_0^1)}^2 e_1$. If the input u is, for example, bound constrained or L^2 regularized in an optimal control setting, then this implies that $\ell(e_1, y^{(1)})$ is uniformly bounded by a multiple of e_1 .

To derive an approximation of the second Newton step, we again restrict ourselves to the case that N is the superposition operator of a polynomial such that $N \in C(W, H)$. Taking on our comments in [Remark 3.14](#), we apply the argument of [Proposition 3.9](#) to $N(y_\psi^{(1)})$ (to the set \mathcal{C}). Thus there exist orthonormal vectors $\phi^1, \dots, \phi^m \in H$ such that the approximation error of the (pointwise a.e.) orthogonal projection

$$e_2 := \|N(y_\psi^{(1)}) - \Pi_\phi N(y_\psi^{(1)})\|_{L^2(0,T;H)}$$

can be made arbitrarily small, where

$$\Pi_\phi N(y_\psi^{(1)})(t) = \sum_{j=1}^m (\phi^j, N(y_\psi^{(1)})(t))_H \phi^j.$$

We define the second simplified Newton step that is based on the approximations $\Pi_\psi d^{(1)}$ and $\Pi_\phi N(y_\psi^{(1)})$

$$d^{(2)}[\psi, \phi](t) := \int_0^t T(t-s) \left(N(\bar{y}) - C \Pi_\psi d^{(1)} - \Pi_\phi N(y_\psi^{(1)}) \right) ds.$$

LEMMA 3.16. *Let the assumptions of [Theorem 3.8](#) hold. Let N be the superposition operator of a polynomial such that $N \in C(W, H)$. Then there exists $\kappa_1 > 0$, independent of $(\phi^i)_i$ and $(\psi^j)_j$, such that*

$$\|d^{(2)} - d^{(2)}[\psi, \phi]\|_{L^2(0,T;W)} \leq \kappa_1 (e_1 + e_2 + \ell(e_1, y^{(1)})).$$

Proof. The functions $d^{(2)}$ and $d^{(2)}[\psi, \phi]$ are unique solutions of [\(3.20\)](#) (where the right-hand side is changed appropriately in the case of $d^{(2)}[\psi, \phi]$). Parabolic regularity theory gives the estimate

$$\begin{aligned} \|d^{(2)} - d^{(2)}[\psi, \phi]\|_{L^2(0,T;W)} &\leq \\ &\kappa_2 \|C d^{(1)} + N(y^{(1)}) - C \Pi_\psi d^{(1)} - \Pi_\phi N(y_\psi^{(1)})\|_{L^2(0,T;H)} \end{aligned}$$

for some $\kappa_2 > 0$. The boundedness of C gives the estimate $\|C d^{(1)} - C d_\psi^{(1)}\|_{L^2} \leq \kappa_3 e_1$, where $\kappa_2 > 0$ is the operator norm of C . The insertion of a zero and the triangle inequality yield

$$\|N(y^{(1)}) - \Pi_\phi N(y_\psi^{(1)})\|_{L^2(0,T;H)} \leq e_2 + \ell(e_1, y^{(1)}).$$

Thus, the claim holds with the choice $\kappa_1 := \kappa_2 \max\{\kappa_3, 1\}$. \square

We define

$$y_{\phi\psi}^{(2)} := \bar{y} + \Pi_\psi d^{(1)} + d^{(2)}[\psi, \phi],$$

where we reapply the argument of [Proposition 3.9](#) and obtain a POD approximation of $d^{(2)}[\psi, \phi]$ with basis vectors $\{\theta^1, \dots, \theta^k\} \subset W$. We define the approximation error

$$e_3 := \|\Pi_\theta d^{(2)}[\phi, \psi] - d^{(2)}[\phi, \psi]\|_{L^2(0,T;W)}$$

and

$$y_{\phi\psi\theta}^{(2)} := \bar{y} + \Pi_\psi d^{(1)} + \Pi_\theta d^{(2)}[\psi, \phi].$$

We are ready to prove our main approximation result.

THEOREM 3.17. *Let the assumptions of [Theorem 3.8](#) hold. Let N be the superposition operator of a polynomial such that $N \in C(W, H)$. Let $y \in Y$ solve [\(3.3\)](#). Then there exists $\kappa_1 > 0$, independent of $(\phi^i)_i$, $(\psi^j)_j$, and $(\theta^\ell)_\ell$, such that*

$$\|y - y_{\phi\psi}^{(2)}\|_{L^2(0,T;W)} \leq \|y - y^{(2)}\|_{L^2(0,T;W)} + (1 + \kappa_1)e_1 + \kappa_1(\ell(e_1, y^{(1)}) + e_2)$$

and

$$\|y - y_{\phi\psi\theta}^{(2)}\|_{L^2(0,T;W)} \leq \|y - y^{(2)}\|_{L^2(0,T;W)} + (1 + \kappa_1)e_1 + \kappa_1(\ell(e_1, y^{(1)}) + e_2) + e_3.$$

Moreover, let $(\psi^i)_{i \in \mathbb{N}}$ and $(\lambda_i)_{i \in \mathbb{N}}$ be as in [Proposition 3.9](#), let $(\phi^j)_{j \in \mathbb{N}}$ and $(\mu_j)_{j \in \mathbb{N}}$ be an orthonormal basis of eigenvectors and corresponding eigenvalues of a POD approximation of $N(y_\psi^{(1)})$, and let $(\theta^\ell)_{\ell \in \mathbb{N}}$ and $(\nu_\ell)_{\ell \in \mathbb{N}}$ be an orthonormal basis of eigenvectors and corresponding eigenvalues of a POD approximation of $d^{(2)}[\phi, \psi]$. Then

$$\begin{aligned} \|y - y_{\phi\psi}^{(2)}\|_{L^2(0,T;W)} \leq \\ \|y - y^{(2)}\|_{L^2(0,T;W)} + (1 + \kappa_1) \sum_{i=n+1}^{\infty} \lambda_i + \kappa_1 \sum_{j=m+1}^{\infty} \mu_j + \kappa_1 \ell(e_1, y^{(1)}), \end{aligned}$$

and

$$\begin{aligned} \|y - y_{\phi\psi\theta}^{(2)}\|_{L^2(0,T;W)} \leq \\ \|y - y^{(2)}\|_{L^2(0,T;W)} + (1 + \kappa_1) \sum_{i=n+1}^{\infty} \lambda_i + \kappa_1 \sum_{j=m+1}^{\infty} \mu_j + \kappa_1 \ell(e_1, y^{(1)}) + \sum_{\ell=k+1}^{\infty} \nu_\ell. \end{aligned}$$

Proof. The first and second estimates follow from the estimates in [subsection 3.5](#) and [Lemma 3.16](#). The third and fourth estimates follow from [Proposition 3.9](#) and the fact that the proof of [Proposition 3.9](#) can be replayed for a POD approximation of $N(y_\psi^{(1)})$ in the space $L^2(0, T; H)$ with basis $(\phi^j)_{j \in \mathbb{N}}$ and eigenvalues $(\mu_j)_{j \in \mathbb{N}}$, which gives $e_2 \leq \sum_{j=m+1}^{\infty} \mu_j$. An analogous argument gives $e_3 \leq \sum_{\ell=k+1}^{\infty} \mu_\ell$. \square

Remark 3.18. This means that the error of approximating y with a POD approximation of both Newton steps can be bounded by the sum of the error of the Newton steps and four terms. Two of them are the POD approximation errors of the first Newton step $d^{(1)}$ and the term $N(y_\psi^{(1)})$. The third term relates the POD approximation error of $d^{(1)}$ to the corresponding error between $N(y_\psi^{(1)})$ and $N(y^{(1)})$ in $L^2(0, T; H)$. As we have seen in [Example 3.15](#), this error may depend on the unknown quantity $\|y^{(1)}\|_Y$, and additional assumptions such as restrictions of the control input may be necessary to ensure boundedness of $\|y^{(1)}\|_Y$. The last term is the POD approximation error of $d^{(2)}[\phi, \psi]$. For this POD approximation, the snapshots can again be collected from impulse responses by using the characterization developed in [Theorem 3.13](#).

3.8. Discretization of the second simplified Newton step. We consider the θ -scheme for time discretization that we have used in [subsection 3.4](#) already. Again, let $0 = t_0 < \dots < T_K = T$, $\Delta t = t_{k+1} - t_k$ be a uniform time grid, and let $\theta \in [\frac{1}{2}, 1]$. Moreover, for $i \in \{1, \dots, n\}$ and $j \in \{1, \dots, m\}$ let $(u_k^i)_{k \in \{0, \dots, K-1\}}$ and $(v_k^j)_{k \in \{0, \dots, K-1\}}$ be interval-wise constant discretizations of u^i and v^j .

Then the discrete analogue of (3.20) is

$$(3.24) \quad \begin{aligned} \frac{d_{k+1}^{(2)} - d_k^{(2)}}{\Delta t} - (A + C)(\theta d_{k+1}^{(2)} + (1 - \theta)d_k^{(2)}) &= N(\bar{y}) - \sum_{i=1}^n u_k^i C \psi^i \\ &\quad - \sum_{j=1}^m v_k^j c_j, \quad 0 \leq k < K, \\ d_0^{(2)} &= 0. \end{aligned}$$

Those of (3.22) and (3.23) are

$$(3.25) \quad \begin{aligned} \frac{\beta_{k+1}^i - \beta_k^i}{\Delta t} - (A + C)(\theta \beta_{k+1}^i + (1 - \theta)\beta_k^i) &= 0, \quad 0 \leq k < K, \\ (I + \Delta t(1 - \theta)(A + C))\beta_0^i &= -C\psi^i \end{aligned}$$

and

$$(3.26) \quad \begin{aligned} \frac{\gamma_{k+1}^j - \gamma_k^j}{\Delta t} - (A + C)(\theta \gamma_{k+1}^j + (1 - \theta)\gamma_k^j) &= 0, \quad 0 \leq k < K, \\ (I + \Delta t(1 - \theta)(A + C))\gamma_0^j &= c_j. \end{aligned}$$

The analog of (3.20) with $d^{(2)} = 0$ and $N(y^{(1)}) = 0$ is

$$(3.27) \quad \begin{aligned} \frac{r_{k+1} - r_k}{\Delta t} - (A + C)(\theta r_{k+1} + (1 - \theta)r_k) &= N(\bar{y}), \quad 0 \leq k < K, \\ r_0 &= 0. \end{aligned}$$

We obtain the following discrete convolution formula for the θ -scheme.

PROPOSITION 3.19. *Consider the first simplified Newton step ($d_k^{(1)}$) for the θ -scheme (3.16). Let (β_k^i) and (γ_k^j) be the solutions of (3.25) and (3.26), respectively. Then $(d_k^{(2)})$ can be represented by the discrete convolution formula*

$$d_k^{(2)} = r_k + \Delta t \sum_{i=1}^n \sum_{\ell=0}^{k-1} \beta_{k-\ell}^i u_\ell^i + \Delta t \sum_{j=1}^m \sum_{\ell=0}^{k-1} \gamma_{k-\ell}^j v_\ell^j,$$

where we use the convention that the sum vanishes for $k = 0$. Consequently,

$$d_k^{(2)} - r_k \in \text{span} \left\{ \bigcup_{i=1}^n \{\beta_1^i, \dots, \beta_k^i\} \cup \bigcup_{j=1}^m \{\gamma_1^j, \dots, \gamma_k^j\} \right\}.$$

Proof. We define $e_k^i := \Delta t \sum_{\ell=0}^{k-1} \beta_{k-\ell}^i u_\ell^i$ and $f_k^j := \Delta t \sum_{\ell=0}^{k-1} \gamma_{k-\ell}^j v_\ell^j$. Then $e_0^i = 0$, $f_0^j = 0$, and from (3.25) and (3.26) we obtain—analogously to Proposition 3.7—that

$$\frac{e_{k+1}^i - e_k^i}{\Delta t} = \sum_{\ell=0}^k (\beta_{k+1-\ell}^i - \beta_{k-\ell}^i) u_\ell^i + \beta_0^i u_k^i = (A + C)(\theta e_{k+1}^i - (1 - \theta)e_k^i) - C b_i u_k^i$$

and

$$\frac{f_{k+1}^j - f_k^j}{\Delta t} = \sum_{\ell=0}^k (\gamma_{k+1-\ell}^j - \gamma_{k-\ell}^j) v_\ell^j + \beta_0^j v_k^j = (A + C)(\theta f_{k+1}^j - (1 - \theta)f_k^j) + c_j v_k^j.$$

By superposition $r_k + \sum_{i=1}^n e_k^i + \sum_{j=1}^m f_k^j$ satisfies (3.24) as asserted. The last claim follows by inspection. \square

4. Galerkin ansatz. We derive error estimates of a Galerkin ansatz with the POD basis vectors to approximate the space W . To this end, we consider the bilinear forms $a : W \times W \rightarrow \mathbb{R}$ and $c : W \times W \rightarrow \mathbb{R}$ that arise from the linear operators A and C in the general setting of [subsection 3.2](#).

4.1. Error bound for Newton steps on POD model. We consider $W_\psi := \text{span}\{\psi^1, \dots, \psi^n\} \subset W$. Let $d_\psi^{(1)}$ solve [\(3.13\)](#) on W_ψ ; that is,

$$(N1) \quad \begin{aligned} (\partial_t d_\psi^{(1)}, v_\psi)_H + a(d_\psi^{(1)}, v_\psi) + c(d_\psi^{(1)}, v_\psi) - (Fu, v_\psi)_H &= 0, \\ (d_\psi^{(1)}(0) - \bar{y} - y_0, v_\psi)_H &= 0 \end{aligned}$$

for all $v_\psi \in W_\psi$. Moreover, we consider the subspace $W_\theta := \text{span}\{\theta^1, \dots, \theta^k\}$. Let $d_\theta^{(2)}$ solve the second simplified Newton step [\(3.20\)](#) on W_θ ; that is,

$$(N2) \quad \begin{aligned} (\partial_t d_\theta^{(2)}, v_\theta)_H + a(d_\theta^{(2)}, v_\theta) + c(d_\theta^{(2)}, v_\theta) - (r, v_\theta)_H - c(d_\psi^{(1)}, v_\theta) &= 0, \\ (d_\theta^{(2)}(0), v_\theta)_H &= 0 \end{aligned}$$

for all $v_\theta \in W_\theta$, where $r = N(\bar{y}) - \Pi_\phi N(y_\psi^{(1)})$.

THEOREM 4.1. *Let $a + c$ be a coercive bilinear form on W . Let $(\psi^i)_i$, $(\phi^j)_j$, and $(\theta^\ell)_\ell$ be as in [subsections 3.5](#) and [3.7](#). Then there exist $\kappa_2, \kappa_3 > 0$ such that*

$$\begin{aligned} \|y - (\bar{y} + d_\psi^{(1)} + d_\theta^{(2)})\|_{L^2(0,T;W)} &\leq \\ \|y - y^{(2)}\|_{L^2(0,T;W)} + \kappa_2(1 + \kappa_1\kappa_3)e_1 + \kappa_1\kappa_3(\ell(e_1, y^{(1)}) + e_2) + \kappa_3e_3. \end{aligned}$$

Proof. The coercivity of $a + c$ allows us to obtain the error bounds

$$\begin{aligned} \|d^{(1)} - d_\psi^{(1)}\|_{L^2(0,T;W)} &\leq \kappa_2 \|d^{(1)} - \Pi_\psi d^{(1)}\|_{L^2(0,T;W)}, \text{ and} \\ \|d^{(2)}[\phi, \psi] - d_\theta^{(2)}\|_{L^2(0,T;W)} &\leq \kappa_3 \|d^{(2)}[\phi, \psi] - \Pi_\theta d^{(2)}[\phi, \psi]\|_{L^2(0,T;W)} \end{aligned}$$

for $\kappa_2, \kappa_3 > 0$ by following, for example, the proof of Theorem 2.3 in [\[11\]](#). Then the claim follows after combining these error bounds with the triangle inequality

$$\|d^{(2)} - d_\theta^{(2)}\|_{L^2(0,T;W)} \leq \|d^{(2)} - d^{(2)}[\phi, \psi]\|_{L^2(0,T;W)} + \|d^{(2)}[\phi, \psi] - d_\theta^{(2)}\|_{L^2(0,T;W)},$$

the bound from [Lemma 3.16](#), and the POD approximation errors e_1 and e_3 . \square

4.2. Galerkin approximation error for the nonlinear equation. Let $a : W \times W \rightarrow \mathbb{R}$ be a continuous and coercive bilinear form, and let $N : W \rightarrow H$ be a polynomial. We consider the variational formulations of [\(3.3\)](#) on W ,

$$(Q) \quad (\partial_t y, v)_H + a(y, v) + (N(y), v)_H - (Fu, v)_H = 0, \quad (y(0) - y_0, v) = 0$$

for all $v \in W$, and on $W_\varrho := \text{span}\{\varrho^1, \dots, \varrho^k\} \subset W$,

$$(Q_\varrho) \quad (\partial_t y_\varrho, v_\varrho)_H + a(y_\varrho, v_\varrho) + (N(y_\varrho), v_\varrho)_H - (Fu, v_\varrho)_H = 0, \quad (y_\varrho(0) - y_0, v_\varrho) = 0$$

for all $v_\varrho \in W_\varrho$. Let y solve [\(Q\)](#), and let y_ϱ solve [\(Q_\varrho\)](#). We estimate $\|y_\varrho - \Pi_\varrho y\|_H$ below.

THEOREM 4.2. *Let the nonlinearity N satisfy the estimate*

$$(4.1) \quad \|N(y) - N(y_\varrho)\|_H \leq \bar{\ell}(\|y\|_W + \|y_\varrho\|_W)\|y - y_\varrho\|_W$$

for some monotone function $\bar{\ell}: [0, \infty) \rightarrow [0, \infty)$. Then it holds for $t \in [0, T]$ that

$$\|(y_\varrho - \Pi_\varrho y)(t)\|_H^2 \leq c_1 c_2 \left(\|(y_\varrho - \Pi_\varrho y)(0)\|_H^2 + \|y - \Pi_\varrho y\|_{L^2(0,t;W)}^2 \right),$$

where $c_1 > 0$ is an independent constant and $c_2 = e^{2T\bar{\ell}(\|y\|_{L^\infty(0,T;W)} + \|y_\varrho\|_{L^\infty(0,T;W)})^2}$.

Proof. We follow the ideas of [11, Thm 2.3] and observe that $(\partial_t(y - \Pi_\varrho y), y_\varrho - \Pi_\varrho y)_H = 0$. Combining this with the choice $y_\varrho - \Pi_\varrho y$ for the test functions in (Q $_\varrho$) and (Q) and following the steps in [11, Thm 2.3], we have that

$$\begin{aligned} \frac{1}{2} \partial_t (y_\varrho - \Pi_\varrho y, y_\varrho - \Pi_\varrho y) + \frac{1}{2} a(y_\varrho - \Pi_\varrho y, y_\varrho - \Pi_\varrho y) &\leq \\ \frac{1}{2} a(y - \Pi_\varrho y, y - \Pi_\varrho y) + (N(y) - N(y_\varrho), y_\varrho - \Pi_\varrho y)_H. \end{aligned}$$

The estimate (4.1) and the triangle inequality yield

$$\|N(y) - N(y_\varrho)\|_H \leq \bar{\ell}(\|y\|_W + \|y_\varrho\|_W)(\|y - \Pi_\varrho y\|_W + \|y_\varrho - \Pi_\varrho y\|_W).$$

We apply the Cauchy–Schwarz inequality to $(N(y) - N(y_\varrho), y_\varrho - \Pi_\varrho y)_H$, insert the estimate above, and apply the inequality $ab \leq \frac{1}{2}a^2 + \frac{1}{2}b^2$ suitably to obtain

$$\begin{aligned} (N(y) - N(y_\varrho), y_\varrho - \Pi_\varrho y)_H &\leq \frac{1}{4} a(y - \Pi_\varrho y, y - \Pi_\varrho y) \\ &\quad + \frac{1}{4} a(y_\varrho - \Pi_\varrho y, y_\varrho - \Pi_\varrho y) + \frac{2}{\alpha^2} \bar{\ell}(\|y\|_W + \|y_\varrho\|_W)^2 \|y_\varrho - \Pi_\varrho y\|_H^2, \end{aligned}$$

where $a(v, v) \geq \alpha \|v\|_W^2$ by coercivity. Then the bilinearity and coercivity of a yield

$$\begin{aligned} \frac{1}{2} \partial_t (y_\varrho - \Pi_\varrho y, y_\varrho - \Pi_\varrho y)_H + \frac{\alpha}{4} \|y_\varrho - \Pi_\varrho y\|_W^2 &\leq \\ \frac{3}{4} \|y - \Pi_\varrho y\|_W^2 + \frac{2}{\alpha^2} \bar{\ell}(\|y\|_W + \|y_\varrho\|_W)^2 \|y_\varrho - \Pi_\varrho y\|_H^2. \end{aligned}$$

Making the dependency on t explicit and rearranging, we obtain

$$\begin{aligned} \partial_t \frac{1}{2} \|(y_\varrho - \Pi_\varrho y)(t)\|_H^2 &\leq \frac{3}{4} \|(y - \Pi_\varrho y)(t)\|_W^2 \\ &\quad + \frac{2}{\alpha^2} \bar{\ell}(\|y(t)\|_W + \|y_\varrho(t)\|_W)^2 \|(y_\varrho - \Pi_\varrho y)(t)\|_H^2 - \partial_t \frac{\alpha}{4} \|y_\varrho - \Pi_\varrho y\|_{L^2(0,t;W)}^2. \end{aligned}$$

We scale by 2 and apply the Gronwall lemma to obtain

$$\begin{aligned} \|(y_\varrho - \Pi_\varrho y)(t)\|_H^2 &\leq \\ c_2 \left(\|(y_\varrho - \Pi_\varrho y)(0)\|_H^2 + \frac{3}{2} \int_0^t \|(y - \Pi_\varrho y)(s)\|_W^2 ds - \frac{\alpha}{2} \|(y_\varrho - \Pi_\varrho y)(t)\|_W^2 \right) \end{aligned}$$

We use the estimate $\|y_\varrho - \Pi_\varrho y\|_H^2 \leq \beta \|y_\varrho - \Pi_\varrho y\|_W^2$ for some $\beta > 0$ and the fact that $c_2 \geq 1$ to deduce that there exists $c_1 > 0$ such that

$$\|(y_\varrho - \Pi_\varrho y)(t)\|_H^2 \leq c_1 c_2 \left(\|(y_\varrho - \Pi_\varrho y)(0)\|_H^2 + \|y - \Pi_\varrho y\|_{L^2(0,t;W)}^2 \right). \quad \square$$

5. Augmented POD basis computation. Having established the theoretical framework above, we argue for the following augmentation of the common POD basis computation procedure. We compute v and collect impulse response snapshots $w(t)$ using F as initial value. Then, we reduce the collected set with POD and obtain a reduced basis $\mathcal{B}^{(1)}$ of $d^{(1)}$, cf. [Theorem 3.8](#). We compute a basis of a linear subspace \mathcal{C} , in which $N(y_{\mathcal{B}^{(1)}}^{(1)})$ takes its values, cf. [subsection 3.6](#). This step depends on the nonlinearity N . For $N(y) = by^3$ in our guiding example, we have

$$\mathcal{C} = \text{span} \{ (\bar{y} + \psi^i)(\bar{y} + \psi^j)(\bar{y} + \psi^k) \mid \forall \text{ combinations } i, j, k \}$$

for $\mathcal{B}^{(1)} = \{\psi^1, \dots, \psi^n\}$. Now, we compute impulse responses for the right-hand side of the second Newton step given in [Lemma 3.12](#) by means of impulse response snapshots, cf. [Theorem 3.13](#).

After collecting and reducing the snapshots, we obtain the basis $\mathcal{B}^{(2)}$ of $d^{(2)}$. Because $y^{(2)} = \bar{y} + d^{(1)} + d^{(2)}$, we can compute a basis and reduced FEM operators for the second Newton iterate $y^{(2)}$ by applying POD to the set $\{\bar{y}\} \cup \mathcal{B}^{(1)} \cup \mathcal{B}^{(2)}$. We summarize this procedure in [Algorithm 5.1](#).

Algorithm 5.1 Two-step Newton-based POD Computation

Require: IBVP solution operator SOLVE, POD basis computation POD

Require: Linearization point \bar{y}

- 1: $(v_k)_k \leftarrow \text{SOLVE (3.17)}$
 - 2: $(w_k)_k \leftarrow \text{SOLVE (3.18)}$
 - 3: $\mathcal{B}^{(1)} \leftarrow \text{POD}((v_k)_k \cup (w_k)_k)$
 - 4: $\{\phi^1, \dots, \phi^m\} \leftarrow \text{compute basis of } N(y^{(1)}) \text{ from } \mathcal{B}^{(1)}$
 - 5: $(r_k)_k \leftarrow \text{SOLVE (3.27)}$
 - 6: **for** $i = 1$ **to** n **do**
 - 7: $(\beta_k^i)_k \leftarrow \text{SOLVE (3.25)}$
 - 8: **end for**
 - 9: **for** $j = 1$ **to** m **do**
 - 10: $(\gamma_k^j)_k \leftarrow \text{SOLVE (3.26)}$
 - 11: **end for**
 - 12: $\mathcal{B}^{(2)} \leftarrow \text{POD} \left((r_k)_k \cup \bigcup_{i=1}^n (\beta_k^i)_k \cup \bigcup_{j=1}^m (\gamma_k^j)_k \right)$
 - 13: $\mathcal{B}^{(12)} \leftarrow \text{POD} (\{\bar{y}\} \cup \mathcal{B}^{(1)} \cup \mathcal{B}^{(2)})$
 - 14: **return** $\mathcal{B}^{(12)}$
-

6. Computational results. We demonstrate our findings by means of a numerical implementation of the guiding example from [subsection 3.1](#).

6.1. Setup. We have chosen $a = 0.01$ and $b = 3$ as parameters for the PDE and its linearizations. Regarding the time domain, we have used an equidistant grid consisting of 65 intervals. The time stepping has been realized with the help of the backward Euler method. Regarding the spatial domain, we have used finite elements of quadratic order on a triangulation of an L-shaped domain. For the linearization point (\bar{y}, \bar{u}) , we have set $\bar{u} \equiv 2$ as well as $\partial_t \bar{y} \equiv 0$ and computed the resulting \bar{y} to solve (3.4). Regarding the error or difference computations between state vectors, we note that we have always used the H^1 -norm for the spatial domain. The same applies for the POD computations.

Table 1: Relative approximation error between y and the Newton step approximations $y^{(1)}$, and $y^{(2)}$ (FEM model) as well as the POD approximations $y_{\mathcal{B}^{(1)}}$ and $y_{\mathcal{B}^{(12)}}$.

$\frac{\ y-y^{(1)}\ }{\ y\ }$	$\frac{\ y-y^{(2)}\ }{\ y\ }$	$\frac{\ y-y_{\mathcal{B}^{(1)}}\ }{\ y\ }$	$\frac{\ y-y_{\mathcal{B}^{(12)}}\ }{\ y\ }$
8.8660×10^{-4}	6.3124×10^{-5}	1.6310×10^{-4}	1.3320×10^{-6}

6.2. Approximation with two simplified Newton steps. We compare the solution y of (3.4) to $y^{(1)} = \bar{y} + d^{(1)}$ and $y^{(2)} = \bar{y} + d^{(1)} + d^{(2)}$ for a given test control u , which is displayed in Figure 1b, and given $\bar{y} \equiv y_0$. We have computed $y^{(1)}$ and $y^{(2)}$ with the help of the linearizations of (3.4) described in section 3. The relative difference between $y^{(2)}$ and y is more than one order of magnitude smaller than the relative difference between $y^{(1)}$ and y . We have computed $\mathcal{B}^{(1)}$ and $\mathcal{B}^{(12)}$ by means of Algorithm 5.1. Consequently, (3.4) has been solved by using the reduced spaces, i.e. reduced versions of the operators, yielding solutions $y_{\mathcal{B}^{(1)}}$, $y_{\mathcal{B}^{(12)}}$. We are interested in their ability to approximate y . We observe that the relative approximation error of $y_{\mathcal{B}^{(12)}}$ is two orders of magnitude smaller than that of $y_{\mathcal{B}^{(1)}}$.

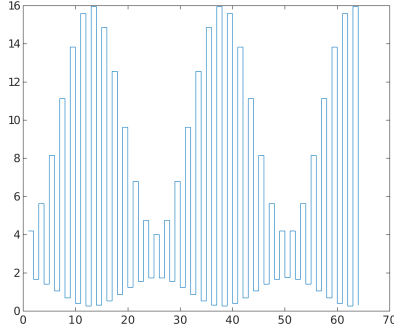
We note that the dimension of the discrete state vectors y using the FEM matrices was 2037, 14 for $y_{\mathcal{B}^{(1)}}$ and 1168 for $y_{\mathcal{B}^{(12)}}$. The exact results of these four computations are given in Table 1. The high number of basis vectors in $\mathcal{B}^{(12)}$ is due to the fact that we have included every basis vector of $\mathcal{B}^{(2)}$ except for those with a singular value smaller than 10^{-8} , the cutoff value of the SVD, into $\mathcal{B}^{(12)}$. It is interesting what happens when we do not use all of them and drop those corresponding very small singular values. This situation is investigated in the context of an OCP in the next subsection.

6.3. Application to an optimal control problem (OCP). We have solved the following a tracking-type OCP with given desired state y_d and Tikhonov regularization parameter $\gamma = 10^{-7}$:

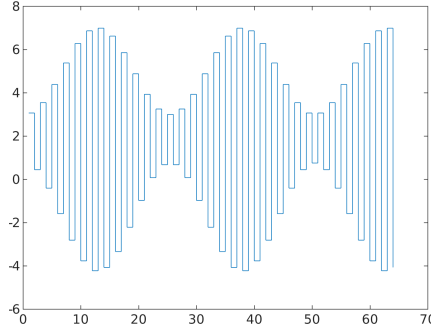
$$\min_{y,u} \frac{1}{2} \|y - y_d\|_Z^2 + \frac{\gamma}{2} \|u\|_{L^2}^2 \quad \text{s.t.} \quad \partial_t y - a\Delta y + by^3 = Fu, \quad y|_{\partial\Omega} = 0.$$

A reduced objective approach has been chosen to obtain an unconstrained OCP. The optimization routine has been initialized with $u \equiv 0$. The target state y_d is the solution of the state equation for the control input visualized in Figure 1a. We have solved the IBVP with FEM, $\mathcal{B}^{(1)}$, and $\mathcal{B}^{(12)}$. Regarding $\mathcal{B}^{(12)}$, we have run the computations for different sizes $B_2 = |\mathcal{B}^{(2)}|$. Specifically, we have successively increased the number of basis vectors in $\mathcal{B}^{(2)}$ following a descending order of the corresponding singular values. The experiment has been run on two spatial grids with different mesh sizes.

On the coarse grid, the state vector of the FEM discretization has 2,469 entries while the state vector of the one-step POD $\mathfrak{B}^{(1)}$ has 12 entries. On the fine grid, the state vector of FEM discretization has 5,597 entries while the state vector of the discretization using the one-step POD $\mathcal{B}^{(1)}$ has 13 entries. In both cases, vectors from $\mathcal{B}^{(2)}$ were included in $\mathcal{B}^{(12)}$ until the corresponding singular value fell below 10^{-8} . For both grids, the additional basis vectors yield more accurate optimized objective values compared with the FEM discretization. Adding 10 basis vectors yields a drop of the relative error in the objective value from 2.56×10^{-2} to 1.25×10^{-4} while the computation time increases from 229 s to 306 s, compared with 9733 s for the FEM solution. The number of optimization iterations stays almost constant: 44 iterations



(a) Reference control to compute y_d in subsection 6.3.



(b) Test control input for the experiment in subsection 6.2.

Figure 1: Control inputs for computational test cases.

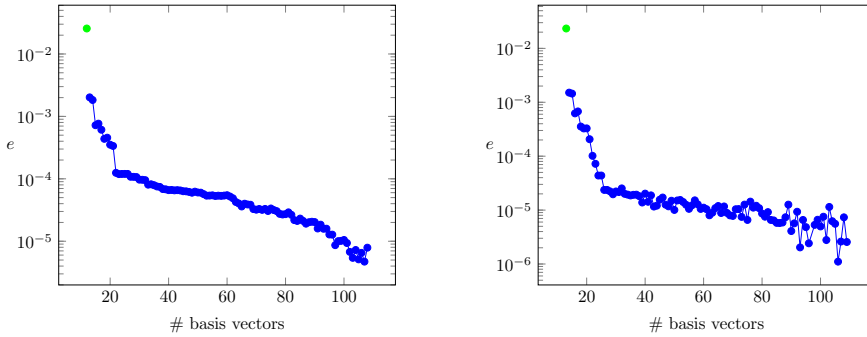


Figure 2: Relative objective error (vertical axes) against size of $\mathcal{B}^{(12)}$ (horizontal axes) for coarse (left) and fine (right) grid. The relative objective error for the $\mathcal{B}^{(1)}$ model is marked in green.

are used on the FEM model, 43 on the $\mathcal{B}^{(1)}$ model, and 44 on all $\mathcal{B}^{(12)}$ models. Similarly, on the fine grid, adding 10 basis vectors resulted in a drop in the relative error in the objective value from 2.34×10^{-2} to 7.22×10^{-5} while the computation time increased from 428 s to 555 s, compared with 41 739 s for the FEM solution. In all cases, the optimization consumes 40 iterations.

Two figures illustrate our results. Figure 2 shows how the relative objective error decreases for an increasing basis $\mathcal{B}^{(12)}$. Figure 3 shows the running time of the OCP solves with the $\mathcal{B}^{(12)}$ model for an increasing number of basis vectors.

7. Application in mixed-integer optimal control. We briefly outline how the presented model reduction can help to solve relaxations of MIOCPs. Employing Sager’s convexification technique [34,35] to control problems constrained by semilinear evolution equations with discrete-valued control inputs, one obtains state equations of the form

$$\partial_t y - Ay = \sum_{i=1}^M \omega_i f_i(y), \quad y(0) = y_0$$

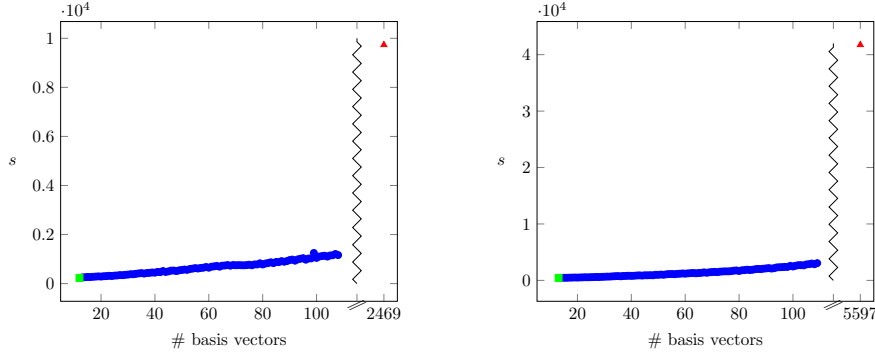


Figure 3: Running time of the OCP (vertical axes) against the number of basis vectors (horizontal axes) for coarse (left) and fine (right) grid for $\mathcal{B}^{(1)}$ (green squares), increasing $\mathcal{B}^{(12)}$ (blue dots), and FEM (red triangle) models.

with $\omega_i \in L^\infty((0, T), \mathbb{R}^M)$ and $\omega_i(t) \in \{0, 1\}^M$ and $\sum_{i=1}^M \omega_i(t) = 1$ for a.a. $t \in [0, T]$; see [22, 30]. The ω_i may be regarded as activations of the different right-hand sides f_i . Following the ideas in [35], one can approach the OCP by first solving a relaxation in which the constraint $\omega_i(t) \in \{0, 1\}^M$ is relaxed to $\omega_i(t) \in [0, 1]^M$ and then computing a binary-valued approximation of the relaxed activation, a procedure that is known as *combinatorial integral decomposition*; see [25]. Using the proposed method for snapshot generation, one can obtain improved reduced models for the semilinear equations

$$\partial_t y - Ay = \omega_i f_i(y), \quad y(0) = y_0$$

for $i \in \{1, \dots, M\}$ and in particular reduced bases for the terms

$$(7.1) \quad \int_0^t S(t-s) f_i(y(s)) \omega_i(s) ds, \quad i \in \{1, \dots, M\}$$

if $(S(t))_{t \geq 0}$ denotes the semigroup generated by A . By the variation of constants formula, we have

$$y(t) = S(t)y_0 + \sum_{i=1}^M \int_0^t S(t-s) f_i(y(s)) \omega_i(s) ds.$$

Consequently, by combining the bases for the terms in (7.1) using a POD computation as in Algorithm 5.1, one obtains an efficient approximation of the solution operator of the semilinear equation in the relaxed problem.

If many relaxations have to be solved, for example in a branch-and-bound procedure, high-quality surrogate models are even more important. In [5], POD models are used for a linear parabolic equation in a branch-and-bound procedure. We envision efficient treatment of semilinear equations in this context using the proposed method.

8. Conclusion. We have developed an algorithm to compute POD models for a class of semilinear evolution equations using the approximation properties of simplified Newton steps on the state equation. The computational results validate the theoretical findings. Furthermore, we have solved a tracking-type OCP constrained by a semilinear PDE from the investigated class on an FEM model, the one-step POD

model and a sequence of increasingly augmented POD models. A moderate number of additional basis vectors improves the approximation of the optimization on the FEM model significantly compared with the one-step POD model.

Thus, if one is willing to spend the expensive *offline phase* for snapshot generation in [Algorithm 5.1](#), for example because many similar OCPs have to be solved in an MPC context or to solve relaxations of MIOCPs, one can trade in a moderate loss in speed-up of the reduced model for a much better capture of the result. For example, in our computational setup on the fine grid, we have achieved an improvement of the relative objective error by a factor of 200 at the cost of approximately halving the speed-up when including 30 additional basis vectors of $\mathfrak{B}^{(2)}$ into $\mathfrak{B}^{(12)}$.

Appendix A. The continuous embedding $\mathcal{H} \hookrightarrow C([0, T], H_0^1(\Omega))$.

For existence of solutions of the semilinear equation (3.4) in [Proposition 3.2](#), we refer to [7, Prop. 5.1]. Considering the results therein, a regularity of the solution in the space $C([0, T]; H_0^1(\Omega))$ seems to be out of reach. Moreover, the application of the vector-valued embedding theorem [29, Thm 8.60] with the choices $Y = H^2(\Omega) \cap H_0^1(\Omega)$ and $H = H_0^1(\Omega)$ seems to require a simultaneous identification of both Hilbert spaces $L^2(\Omega)$ and $H_0^1(\Omega)$ with their respective topological dual spaces.

However, we may substitute the identification of $H_0^1(\Omega) \cong H^{-1}(\Omega)$ with the multidimensional integration by parts formula that arises from the divergence theorem and otherwise follow the proof of [29, Thm 8.60] using $L^2(\Omega)$ instead of V^* . This approach allows us to use only the continuous embeddings $V \hookrightarrow H_0^1(\Omega) \hookrightarrow L^2(\Omega)$. For completeness, we sketch the modified proof below. Note that the assumed boundary regularity that Ω is convex or of class C^2 (see [subsection 3.1](#)) is sufficient for this argument.

PROPOSITION A.1. *Consider $\mathcal{H} = \{u \in L^2(0, T; V) \mid \partial_t u \in L^2(0, T; L^2(\Omega))\}$ with $\|u\|_{\mathcal{H}} = \|u\|_{L^2(0, T; V)} + \|\partial_t u\|_{L^2(0, T; L^2(\Omega))}$ for $u \in \mathcal{H}$. Then the continuous embedding $\mathcal{H} \hookrightarrow C([0, T], H_0^1(\Omega))$ holds because there exists $C > 0$ such that*

$$\sup_{t \in [0, T]} \|u(t)\|_{H_0^1(\Omega)} \leq C (\|u\|_{L^2(0, T; V)} + \|\partial_t u\|_{L^2(0, T; L^2(\Omega))})$$

holds for all $u \in \mathcal{H}$.

Proof. Let $u \in \mathcal{H}$. We use extension by reflection to extend the function u to the interval $(-\beta, T + \beta)$ for some $\beta > 0$. We smooth u with a family of standard mollifiers $(\varphi_\varepsilon)_{\varepsilon > 0}$ that are compactly supported in $(-\beta, T + \beta)$ and define $u_\varepsilon := u * \varphi_\varepsilon$. Then we obtain $u_\varepsilon \rightarrow u \in L^2((0, T), V)$ and $\partial_t u_\varepsilon \rightarrow \partial_t u \in L^2((0, T), L^2(\Omega))$. We highlight that for the convergence $\partial_t u_\varepsilon \rightarrow \partial_t u$ it is important that the mollification of the derivative is the derivative of the mollification. A cutoff argument to prove this works only by virtue of the extension to the interval $(-\beta, T + \beta)$, and we cannot extend it using absolute continuity because this is essentially what is to be shown.

Now, the mollification gives that $u_\varepsilon, \partial_t u_\varepsilon \in C_c^\infty(\mathbb{R}, V)$ and thus $u_\varepsilon, \partial_t u_\varepsilon \in C_c^\infty(\mathbb{R}, H_0^1(\Omega))$. As in [29, (8.32)], we obtain for $x, x_0 \in [0, T]$ that

$$\|u_\varepsilon(x)\|_{H_0^1(\Omega)}^2 = \|u_\varepsilon(x_0)\|_{H_0^1(\Omega)}^2 + 2 \int_{x_0}^x (\partial_t u_\varepsilon(s), u_\varepsilon(s))_{H_0^1(\Omega)} ds,$$

where $(\cdot, \cdot)_{H_0^1(\Omega)}$ is the usual inner product on $H_0^1(\Omega)$. In particular, we can write

$$\|u_\varepsilon(x)\|_{H_0^1(\Omega)}^2 = \|u_\varepsilon(x_0)\|_{H_0^1(\Omega)}^2 + 2 \int_{x_0}^x \int_{\Omega} \nabla \partial_t u_\varepsilon(s)^T \nabla u_\varepsilon(s) d\omega ds,$$

which allows us to apply multidimensional integration by parts that follows from the divergence theorem to deduce

$$\|u_\varepsilon(x)\|_{H_0^1(\Omega)}^2 = \|u_\varepsilon(x_0)\|_{H_0^1(\Omega)}^2 + 2 \int_{x_0}^x \int_{\partial\Omega} \partial_t u_\varepsilon(s) \nabla u_\varepsilon(s) \cdot d\sigma - \int_{\Omega} \partial_t u_\varepsilon(s) \Delta u_\varepsilon(s) \, d\omega \, ds.$$

Since $\partial_t u_\varepsilon(s) \in H_0^1(\Omega)$ for all $s \in [x_0, x]$ by virtue of the mollification, we obtain

$$\|u_\varepsilon(x)\|_{H_0^1(\Omega)}^2 = \|u_\varepsilon(x_0)\|_{H_0^1(\Omega)}^2 - 2 \int_{x_0}^x \int_{\Omega} \partial_t u_\varepsilon(s) \Delta u_\varepsilon(s) \, d\omega \, ds,$$

which implies

$$\|u_\varepsilon(x)\|_{H_0^1(\Omega)}^2 = \|u_\varepsilon(x_0)\|_{H_0^1(\Omega)}^2 + 2 \|\partial_t u_\varepsilon\|_{L^2(0,T;L^2(\Omega))} \|u_\varepsilon\|_{L^2(0,T;V)}$$

by virtue of Hölder's inequality and $V \hookrightarrow L^2(\Omega)$.

Now the remainder of the proof of [29, Thm 8.60] applies if the dual space of V is replaced with $L^2(\Omega)$ and the duality pairing $\langle \partial_t u(s), u(s) \rangle_{V^*,V}$ is replaced with $\int_{\Omega} \partial_t u(s) \Delta u(s) \, d\omega$. \square

Remark A.2. The fact that $\partial_t u_\varepsilon(s)$ is in $H_0^1(\Omega)$, which follows from $u(s) \in H_0^1(\Omega)$, seems to be crucial for the proof of [Proposition A.1](#). However, there is also an abstract argument based on interpolation spaces. In particular, one may combine [2, Thm 4.10.2] (choices $E_0 = H^2(\Omega)$, $E_1 = L^2(\Omega)$, $p = 2$) with the continuous embedding $H_0^1(\Omega) \hookrightarrow H^1(\Omega)$ to obtain that $\mathcal{H} \hookrightarrow C([0, T], H_0^1(\Omega))$, where the fact that $u(s) \in H_0^1(\Omega)$ seems to be irrelevant.

REFERENCES

- [1] Afanasiev, K., Hinze, M.: Adaptive control of a wake flow using proper orthogonal decomposition. *Lecture Notes in Pure and Applied Mathematics* pp. 317–332 (2001)
- [2] Amann, H.: *Linear and Quasilinear Parabolic Problems*, vol. 1. Springer (1995)
- [3] Arendt, W., Batty, C.J., Hieber, M., Neubrander, F.: *Vector-valued Laplace transforms and Cauchy problems*, vol. 96. Springer Science & Business Media (2011)
- [4] Arian, E., Fahl, M., Sachs, E.W.: Trust-region proper orthogonal decomposition for flow control. Tech. rep., Institute for Computer Applications In Science and Engineering, Hampton VA (2000)
- [5] Bachmann, F., Beermann, D., Lu, J., Volkwein, S.: POD-based mixed-integer optimal control of the heat equation. *Journal of Scientific Computing* pp. 1–28 (2019)
- [6] Bai, Z., Skoogh, D.: A projection method for model reduction of bilinear dynamical systems. *Linear algebra and its applications* **415**(2-3), 406–425 (2006)
- [7] Barbu, V.: *Nonlinear differential equations of monotone types in Banach spaces*. Springer Science & Business Media (2010)
- [8] Benner, P., Breiten, T.: Interpolation-based \mathcal{H}_2 -model reduction of bilinear control systems. *SIAM Journal on Matrix Analysis and Applications* **33**(3), 859–885 (2012)
- [9] Benner, P., Goyal, P., Gugercin, S.: \mathcal{H}_2 -quasi-optimal model order reduction for quadratic-bilinear control systems. *SIAM Journal on Matrix Analysis and Applications* **39**(2), 983–1032 (2018)
- [10] Bergmann, M., Cordier, L.: Optimal control of the cylinder wake in the laminar regime by trust-region methods and POD reduced-order models. *Journal of Computational Physics* **227**(16), 7813–7840 (2008)
- [11] Bernardi, C., Raugel, G.: A conforming finite element method for the time-dependent Navier–Stokes equations. *SIAM Journal on Numerical Analysis* **22**(3), 455–473 (1985)
- [12] Bott, S.M.: Adaptive SQP method with reduced order models for optimal control problems with constraints on the state applied to the Navier-Stokes equations. Ph.D. thesis, TU Darmstadt (2015)
- [13] Deuffhard, P.: *Newton methods for nonlinear problems: affine invariance and adaptive algorithms*, vol. 35. Springer Science & Business Media (2011)

- [14] Diestel, J., Uhl, J.J.: Vector measures. 15 (1977). DOI <http://dx.doi.org/10.1090/surv/015>
- [15] Evans, L.C.: Partial Differential Equations, vol. 322. American Mathematical Society (1998)
- [16] Flagg, G., Gugercin, S.: Multipoint Volterra series interpolation and \mathcal{H}_2 optimal model reduction of bilinear systems. *SIAM Journal on Matrix Analysis and Applications* **36**(2), 549–579 (2015)
- [17] Fletcher, R., Leyffer, S.: Nonlinear programming without a penalty function. *Mathematical Programming* **91**(2), 239–269 (2002)
- [18] Ghiglieri, J., Ulbrich, S.: Optimal flow control based on POD and MPC and an application to the cancellation of Tollmien–Schlichting waves. *Optimization Methods and Software* **29**(5), 1042–1074 (2014)
- [19] Gräßle, C., Hinze, M., Volkwein, S.: Model order reduction by proper orthogonal decomposition. In: P. Benner, W. Schilders, S. Grivet-Talocia, A. Quarteroni, G. Rozza, L. Miguel Silveira (eds.) *Model Order Reduction: Volume 2: Snapshot-Based Methods and Algorithms*. De Gruyter (2020)
- [20] Gu, C.: QLMOR: A new projection-based approach for nonlinear model order reduction. In: 2009 IEEE/ACM International Conference on Computer-Aided Design-Digest of Technical Papers, pp. 389–396. IEEE (2009)
- [21] Gubisch, M., Volkwein, S.: Proper orthogonal decomposition for linear-quadratic optimal control. *Model Reduction and Approximation: Theory and Algorithms* **15**, 1 (2017)
- [22] Hante, F.M., Sager, S.: Relaxation methods for mixed-integer optimal control of partial differential equations. *Computational Optimization and Applications* **55**(1), 197–225 (2013)
- [23] Hinze, M., Volkwein, S.: Proper orthogonal decomposition surrogate models for nonlinear dynamical systems: Error estimates and suboptimal control. In: *Dimension reduction of large-scale systems*, pp. 261–306. Springer (2005)
- [24] Jay, L.O.: Inexact simplified Newton iterations for implicit Runge-Kutta methods. *SIAM Journal on Numerical Analysis* **38**(4), 1369–1388 (2000)
- [25] Jung, M.N., Reinelt, G., Sager, S.: The Lagrangian relaxation for the combinatorial integral approximation problem. *Optimization Methods and Software* **30**(1), 54–80 (2015)
- [26] Kunisch, K., Volkwein, S.: Control of the Burger’s equation by a reduced-order approach using proper orthogonal decomposition. *Journal of Optimization Theory and Applications* **102**(2), 345–371 (1999)
- [27] Kunisch, K., Volkwein, S.: Galerkin proper orthogonal decomposition methods for parabolic problems. *Numerische Mathematik* **90**(1), 117–148 (2001)
- [28] Kunisch, K., Volkwein, S.: Galerkin proper orthogonal decomposition methods for a general equation in fluid dynamics. *SIAM Journal on Numerical Analysis* **40**(2), 492–515 (2002)
- [29] Leoni, G.: *A First Course in Sobolev Spaces*. American Mathematical Soc. (2017)
- [30] Manns, P., Kirches, C.: Improved regularity assumptions for partial outer convexification of mixed-integer PDE-constrained optimization problems. *ESAIM: Control, Optimisation and Calculus of Variations* **26**(32) (2020)
- [31] Miyakawa, T.: On nonstationary solutions of the Navier–Stokes equations in an exterior domain. *Hiroshima Mathematical Journal* **12**(1), 115–140 (1982)
- [32] Riesz, M.: Sur les ensembles compacts de fonctions sommables. *Acta Szeged Sect. Math* **6**, 136–142 (1933)
- [33] Sachs, E.W., Volkwein, S.: POD-Galerkin approximations in PDE-constrained optimization. *GAMM-Mitteilungen* **33**(2), 194–208 (2010)
- [34] Sager, S.: *Numerical methods for mixed-integer optimal control problems*. Der andere Verlag Tönning, Lübeck, Marburg (2005)
- [35] Sager, S., Bock, H., Diehl, M.: The integer approximation error in mixed-integer optimal control. *Mathematical Programming, Series A* **133**(1–2), 1–23 (2012)
- [36] Schmidt, A., Potschka, A., Korkel, S., Bock, H.G.: Derivative-extended POD reduced-order modeling for parameter estimation. *SIAM Journal on Scientific Computing* **35**(6), A2696–A2717 (2013)
- [37] Simon, J.: Compact sets in the space $L^p(O, T; B)$. *Annali di Matematica pura ed applicata* **146**(1), 65–96 (1986)
- [38] Volkwein, S.: Optimal control of a phase-field model using proper orthogonal decomposition. *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik: Applied Mathematics and Mechanics* **81**(2), 83–97 (2001)
- [39] Wächter, A., Biegler, L.T.: Line search filter methods for nonlinear programming: Local convergence. *SIAM Journal on Optimization* **16**(1), 32–48 (2005)

The submitted manuscript has been created by UChicago Argonne, LLC, Operator of Argonne National Laboratory (“Argonne”). Argonne, a U.S. Department of Energy Office of Science laboratory, is operated under Contract No. DE-AC02-06CH11357. The U.S. Government retains for itself, and others acting on its behalf, a paid-up nonexclusive, irrevocable worldwide license in said article to reproduce, prepare derivative works, distribute copies to the public, and perform publicly and display publicly, by or on behalf of the Government. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan <http://energy.gov/downloads/doe-public-access-plan>.