ACTIVE OPERATOR INFERENCE FOR LEARNING LOW-DIMENSIONAL DYNAMICAL-SYSTEM MODELS FROM NOISY DATA*

WAYNE ISAAC TAN UY†, YUEPENG WANG†, YUXIAO WEN†, AND BENJAMIN PEHERSTORFER†

Abstract. Noise poses a challenge for learning dynamical-system models because already small variations can distort the dynamics described by trajectory data. This work builds on operator inference from scientific machine learning to infer low-dimensional models from high-dimensional state trajectories polluted with noise. The presented analysis shows that, under certain conditions, the inferred operators are unbiased estimators of the well-studied projection-based reduced operators from traditional model reduction. Furthermore, the connection between operator inference and projection-based model reduction enables bounding the mean-squared errors of predictions made with the learned models with respect to traditional reduced models. The analysis also motivates an active operator inference approach that judiciously samples high-dimensional trajectories with the aim of achieving a low mean-squared error by reducing the effect of noise. Numerical experiments with high-dimensional linear and nonlinear state dynamics demonstrate that predictions obtained with active operator inference have orders of magnitude lower mean-squared errors than operator inference with traditional, equidistantly sampled trajectory data.

Key words. scientific machine learning, non-intrusive model reduction, operator inference, design of experiments, reduced models, noise

AMS subject classifications. 65P99, 65Y99, 65F99, 93C05, 93C10, 68T99, 62J05, 60B20

1. Introduction.

Noise poses a challenge for learning dynamical-system models because already small variations can distort the dynamics described by trajectory data. In this work, we build on operator inference [41] from scientific machine learning to derive low-dimensional dynamical-system models from high-dimensional, noisy state trajectories. We introduce a sampling scheme to query the high-dimensional systems for data so that, under certain conditions, in particular if the high-dimensional system dynamics are polynomially nonlinear, the inferred operators are unbiased estimators of the well-studied reduced operators obtained via projection of the governing equations of the high-dimensional systems in classical model reduction [1, 8, 46]. Additionally, we show that the mean-squared error (MSE) of the states predicted with the learned models can be bounded independently of the dimensions of the high-dimensional systems and in terms of the noise-to-signal ratio of the trajectory data. Motivated by the analysis, we propose active operator inference that queries high-dimensional systems in a principled way to generate data with low noise-to-signal ratios, which reduces by a factor of up to three the number of data samples that are required from the high-dimensional systems to make accurate state predictions in our numerical experiments. For the same number of data samples, active operator inference achieves orders of magnitude lower MSEs than traditional, equidistant-in-time sampled trajectory data.

Learning models from data is an active research topic in the field of scientific machine learning. A prominent approach is to fit dynamical-system models to data via

*Submitted to the editors DATE.

Funding: This work was partially supported by US Department of Energy, Office of Advanced Scientific Computing Research, Applied Mathematics Program (Program Manager Dr. Steven Lee), DOE Award DESC0019334, and by the National Science Foundation under Grant DMS-2012250.

†Courant Institute of Mathematical Sciences, New York, NY (wayne.uy@cims.nyu.edu, yw3114@nyu.edu, yw3210@nyu.edu, pehersto@cims.nyu.edu).

dynamic mode decomposition and Koopman-based methods [12, 32, 45, 52, 60, 66]. In another research direction, sparse representations of governing equations are sought with tools from sparse regression and compressive sensing [11, 47, 50, 51]. There is also work on non-intrusive model reduction that learns coefficients of low-dimensional representations from data [22, 23, 26]. If frequency-domain or impulse-response data are available, then data-driven modeling methods from the systems and control community are often used, such as the Loewner approach [2, 3, 6, 21, 27, 33, 37], vector fitting [18, 24], and eigensystem realization [29, 31].

In terms of learning from noisy data, there is the work [58] that establishes probabilistic recovery guarantees via compressive sensing of sparse systems. Noise-robust data-driven discovery of governing equations is considered in [68, 69] using sparse Bayesian regression. A strategy is proposed to subsample the data utilized in solving the regression problem with the goal of reducing the influence of noise on the learned model. A signal-noise decomposition is pursued in [48] in which a neural network is trained to discover the underlying dynamics while simultaneously estimating the noise. In system identification, works such as [9, 13, 35, 55, 56, 64] derive probabilistic error bounds for oftentimes linear models using tools from, e.g., random matrix theory. The effect of the presence of noise and perturbations in frequency-domain data have also been studied in data-driven interpolatory model reduction and Loewner methods [7, 19, 20, 34]. However, except for the interpolatory model reduction methods, which require frequency-domain data, no low-dimensional models are considered in these works. In contrast, our approach based on operator inference and re-projection [39, 41] aims to learn low-dimensional models that are suited for solving outer-loop applications such as design, control, and inverse problems. Operator inference can learn non-Markovian low-dimensional models [61] and it is also a building block for other learning methods such as lift & learn introduced in [43, 57], which comes with a sensitivity analysis with respect to deterministic perturbations in data [42, Chapter 4.3]. In [62], probabilistic *a posteriori* error bounds for operator-inference models are derived for linear models; however, the bounds only hold when data are free of noise. In the following, we exploit the bridge between data-driven modeling with operator inference and traditional model reduction [1, 8, 46] to establish probabilistic guarantees for learning from noisy data and to inform in a principled way which data samples to query from the high-dimensional system to reduce the effect of noise on the MSE of state predictions.

This manuscript is organized as follows. Section 2 discusses preliminaries about learning low-dimensional dynamical-system models from data via operator inference and re-projection. Section 3 describes the sampling and inference problem for learning models from noisy trajectories with the proposed approach. Then, bounds are derived for the MSE of the inferred operators and of the state predictions with respect to projection-based reduced models from traditional model reduction. A design of experiments approach is proposed in Section 4, which leads to active operator inference that selects data samples to reduce the effect of noise on the MSE of state predictions. Numerical results presented in Section 5 are in agreement with the analysis: the results indicate that active operator inference learns low-dimensional models with MSEs that are orders of magnitude more accurate than with an uninformed design of experiments.

2. Preliminaries.

We review operator inference [41] for learning low-dimensional models from data in Section 2.1. Section 2.2 describes operator inference with the re-projection data sampling scheme [39] to recover projection-based reduced models

from data.

2.1. Learning low-dimensional dynamical-system models from data with operator inference. Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_K \in \mathbb{R}^N$ be states at time steps $k = 1, \ldots, K$ that are obtained by exciting a dynamical system

$$\boldsymbol{x}_{k+1} = \boldsymbol{f}(\boldsymbol{x}_k, \boldsymbol{u}_k), \qquad k = 0, \ldots, K-1, \tag{2.1}$$

at the control inputs $\boldsymbol{u}_0, \ldots, \boldsymbol{u}_{K-1} \in \mathbb{R}^p$ and initial condition $\boldsymbol{x}_0 \in \mathbb{R}^N$. Let further $\mathcal{V} \subset \mathbb{R}^N$ be a subspace of the N-dimensional state space $\mathbb{R}^N$. The subspace $\mathcal{V}$ is spanned by the orthonormal columns of the basis matrix $\boldsymbol{V} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n] \in \mathbb{R}^{N \times n}$. For example, the subspace $\mathcal{V}$ can be obtained via principal component analysis (proper orthogonal decomposition) applied to sampled state trajectories using the Euclidean inner product on $\mathbb{R}^N$.

Operator inference introduced in [41] learns low-dimensional dynamical-system models with polynomial nonlinear terms that best fit the temporal evolution of the state in the subspace $\mathcal{V}$ with respect to the Euclidean norm in a least-squares sense. Operator inference first projects the high-dimensional states $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_K$ onto the subspace $\mathcal{V}$ to obtain the projected states $\breve{\boldsymbol{x}}_0, \ldots, \breve{\boldsymbol{x}}_K$ with $\breve{\boldsymbol{x}}_k = \boldsymbol{V}^T \boldsymbol{x}_k \in \mathbb{R}^n$ for $k = 0, \ldots, K$ and then solves the least-squares problem

$$\min_{\widehat{\boldsymbol{A}}_1, \ldots, \widehat{\boldsymbol{A}}_\ell, \widehat{\boldsymbol{B}}} \sum_{k=0}^{K-1} \left\| \sum\nolimits_{j=1}^{\ell} \widehat{\boldsymbol{A}}_j \breve{\boldsymbol{x}}_k^j + \widehat{\boldsymbol{B}} \boldsymbol{u}_k - \breve{\boldsymbol{x}}_{k+1} \right\|_2^2, \tag{2.2}$$

where $\ell \in \mathbb{N}$ is the polynomial order, $\widehat{\boldsymbol{B}} \in \mathbb{R}^{n \times p}$, $\widehat{\boldsymbol{A}}_j \in \mathbb{R}^{n \times n_j}$ with

$$n_j = \binom{n+j-1}{j}, \qquad j = 1, \ldots, \ell,$$

and $\breve{\boldsymbol{x}}_k^j$ is obtained for $k = 0, \ldots, K$ by forming the Kronecker product j times $\breve{\boldsymbol{x}}_k \otimes \cdots \otimes \breve{\boldsymbol{x}}_k$ and retaining only the factors whose components are unique up to permutation [41]. Note that if problem (2.2) is underdetermined, then regularization as proposed in the context of operator inference in, e.g., [38, 49, 57] can be performed to bias the operators towards, e.g., giving a stable low-dimensional model. It is challenging, however, to design regularization terms that lead to a meaningful bias. In the following, we will avoid regularization and instead build on re-projection and sufficient data as discussed in the following section.

2.2. Recovering projection-based reduced models from data with operator inference and re-projection. The re-projection data-sampling scheme introduced in [39] judiciously excites the high-dimensional system (2.1) to generate a re-projected trajectory $\breve{\boldsymbol{Y}} = [\breve{\boldsymbol{y}}_1, \ldots, \breve{\boldsymbol{y}}_K] \in \mathbb{R}^{n \times K}$. The following description follows the version of re-projection described in [43, Section 3.2]. Let $\bar{\boldsymbol{X}} = [\bar{\boldsymbol{x}}_1, \ldots, \bar{\boldsymbol{x}}_K]$ be a matrix where each column contains an N-dimensional vector. For example, in [39, 43], it is proposed to generate $\bar{\boldsymbol{X}}$ by first querying the high-dimensional system (2.1) at an initial condition and inputs to sample the trajectory $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_K]$ and then setting $\bar{\boldsymbol{X}} = \boldsymbol{X}$. Let now $\boldsymbol{U} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_K]$ be an input trajectory and let $\breve{\boldsymbol{X}} = [\breve{\boldsymbol{x}}_1, \ldots, \breve{\boldsymbol{x}}_K]$ be the projected trajectory obtained as $\breve{\boldsymbol{X}} = \boldsymbol{V}^T \bar{\boldsymbol{X}}$ from $\bar{\boldsymbol{X}}$. Re-projection then computes $\boldsymbol{Y} = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_K]$ by querying the high-dimensional system

$$\boldsymbol{y}_k = \boldsymbol{f}(\boldsymbol{V} \breve{\boldsymbol{x}}_k, \boldsymbol{u}_k), \qquad k = 1, \ldots, K,$$

to obtain $\breve{\boldsymbol{Y}} = \boldsymbol{V}^T\boldsymbol{Y}$. The re-projection scheme can be applied to black-box dynamical systems that can be queried at arbitrary initial conditions in $\mathbb{R}^N$ and inputs in $\mathbb{R}^p$.

As shown in [39, 42], if the high-dimensional system (2.1) from which data are sampled has polynomial form, i.e.,

$$\boldsymbol{f}(\boldsymbol{x},\boldsymbol{u}) = \sum_{j=1}^{\ell} \boldsymbol{A}_j\boldsymbol{x}^j + \boldsymbol{B}\boldsymbol{u}\,, \tag{2.3}$$

and if there are sufficiently many data samples, then the solution of the least-squares problem

$$\min_{\widehat{\boldsymbol{A}}_1,\ldots,\widehat{\boldsymbol{A}}_\ell,\widehat{\boldsymbol{B}}} \bar{J}(\widehat{\boldsymbol{A}}_1,\ldots,\widehat{\boldsymbol{A}}_\ell,\widehat{\boldsymbol{B}};\breve{\boldsymbol{X}},\breve{\boldsymbol{Y}},\boldsymbol{U}) \tag{2.4}$$

with objective

$$\bar{J}(\widehat{\boldsymbol{A}}_1,\ldots,\widehat{\boldsymbol{A}}_\ell,\widehat{\boldsymbol{B}};\breve{\boldsymbol{X}},\breve{\boldsymbol{Y}},\boldsymbol{U}) = \sum_{k=1}^{K} \left\| \sum\nolimits_{i=1}^{\ell} \widehat{\boldsymbol{A}}_i\breve{\boldsymbol{x}}_k^j + \widehat{\boldsymbol{B}}\boldsymbol{u}_k - \breve{\boldsymbol{y}}_k \right\|_2^2 \tag{2.5}$$

is unique and coincides with the projected operators

$$\begin{aligned} \widetilde{\boldsymbol{B}} &= \boldsymbol{V}^T\boldsymbol{B}\,, \\ \widetilde{\boldsymbol{A}}_j &= \boldsymbol{V}^T\boldsymbol{A}_j\boldsymbol{S}_j(\boldsymbol{V}\otimes\cdots\otimes\boldsymbol{V})\boldsymbol{R}_j, \qquad j=1,\ldots,\ell, \end{aligned} \tag{2.6}$$

where the matrices $\boldsymbol{S}_j \in \mathbb{R}^{N_j\times N^j}$ and $\boldsymbol{R}_j \in \mathbb{R}^{n^j\times n_j}$ satisfy

$$\boldsymbol{z}^j = \boldsymbol{S}_j(\boldsymbol{z}\otimes\cdots\otimes\boldsymbol{z}), \qquad \widetilde{\boldsymbol{z}}\otimes\cdots\otimes\widetilde{\boldsymbol{z}} = \boldsymbol{R}_j\widetilde{\boldsymbol{z}}^j$$

for all $\boldsymbol{z}\in\mathbb{R}^N, \widetilde{\boldsymbol{z}}\in\mathbb{R}^n$ and $j=1,\ldots,\ell$ and the Kronecker product is applied j times. Notice that the re-projected trajectory $\breve{\boldsymbol{Y}}$ enters in the objective in the least-squares problem (2.4), whereas only the projected trajectory $\breve{\boldsymbol{X}}$ enters in problem (2.2).

In traditional model reduction, see, e.g., [1,8,46], the projected operators $\widetilde{\boldsymbol{A}}_1,\ldots,\widetilde{\boldsymbol{A}}_\ell$, $\widetilde{\boldsymbol{B}}$ are computed directly by computing the matrix-matrix products in the projection step (2.6). Thus, such traditional model reduction methods are intrusive in the sense that they require the high-dimensional operators $\boldsymbol{A}_1,\ldots,\boldsymbol{A}_\ell,\boldsymbol{B}$ either in assembled form or implicitly via matrix-vector products.

3. Learning low-dimensional models from noisy data. This work investigates operator inference and re-projection for learning low-dimensional models of noisy dynamical systems,

$$\boldsymbol{x}_{k+1} = \boldsymbol{f}(\boldsymbol{x}_k,\boldsymbol{u}_k) + \boldsymbol{\xi}_k\,, \qquad k=0,\ldots,K-1\,, \tag{3.1}$$

where $\boldsymbol{\xi}_0,\ldots,\boldsymbol{\xi}_{K-1}$ represent noise. The random vectors $\boldsymbol{\xi}_0,\ldots,\boldsymbol{\xi}_{K-1}$ are independent and each noise vector $\boldsymbol{\xi}_k \sim N(\boldsymbol{0},\sigma^2\boldsymbol{I})$, for $k=0,\ldots,K-1$, is an N-dimensional Gaussian random vector with a diagonal covariance matrix and standard deviation $\sigma>0$ in all directions. Notice that the noise vectors $\boldsymbol{\xi}_0,\ldots,\boldsymbol{\xi}_{K-1}$ are independent; however, the states $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_K$ can be dependent. In the following, for ease of exposition, the noisy high-dimensional system (3.1) can be queried at any initial condition in $\mathbb{R}^N$ with any input in $\mathbb{R}^p$; however, the space of initial conditions and inputs can be restricted to subsets of $\mathbb{R}^N$ and $\mathbb{R}^p$ if necessary.

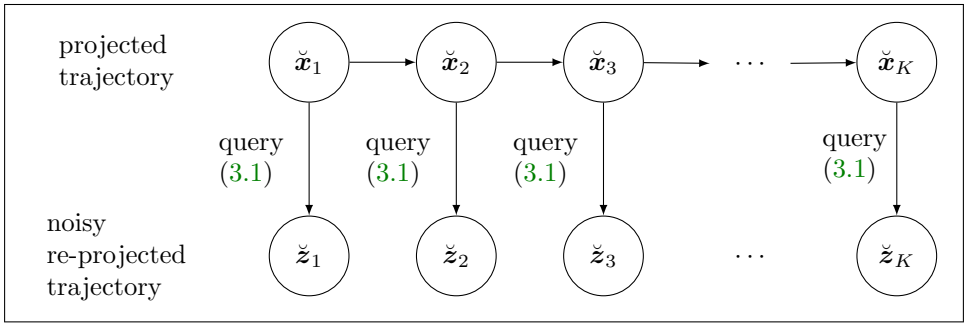


Fig. 1: Applying re-projection to query the noisy high-dimensional system (3.1) leads to unbiased estimators of the projected operators (2.6), which are the very same operators that are obtained with classical, intrusive model reduction.

Section 3.1 applies operator inference and re-projection to learn low-dimensional models from noisy trajectories and derives conditions under which the inferred operators are unbiased estimators of the projection-based reduced operators. The MSE of the learned low-dimensional operators is quantified in terms of the noise-to-signal ratio. In Section 3.2, we derive bounds on the bias and the MSE of the predicted states of the system described by the learned low-dimensional model with the learned operators for linear and polynomially nonlinear dynamics, respectively. The bounds scale with respect to the noise-to-signal ratio.

3.1. Operator inference with re-projection with noisy state trajectories. Let $\bar{\boldsymbol{X}}$ be a dictionary of initial conditions, i.e., a matrix with N-dimensional columns (cf. Section 2.2), and let $\boldsymbol{U} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_K]$ be an input trajectory. The purpose of the columns of $\bar{\boldsymbol{X}}$ is to excite the underlying system in the following data generation scheme to obtain trajectories that are informative for inferring reduced operators. This means that $\bar{\boldsymbol{X}}$ is not necessarily a trajectory of length K obtained from the underlying system. In particular, the columns of $\bar{\boldsymbol{X}}$ are not necessarily random but instead will be deterministic in most of the following. A similar situation is found in frequency-domain system identification, where deterministic input signals such as the chirp and ramp signal are used to excite systems [35, 53]. In contrast, we focus here on the time domain and therefore are interested in columns of $\bar{\boldsymbol{X}}$ to excite the underlying system in an informative way in the sense of a high signal-to-noise ratio. We will make this more precise now. Additionally, later in Section 4, we will provide an active learning approach for designing the columns of $\bar{\boldsymbol{X}}$.

Given $\bar{\boldsymbol{X}}$, we then apply re-projection to obtain $\boldsymbol{Z} = [\boldsymbol{z}_1, \ldots, \boldsymbol{z}_K]$ by querying the noisy high-dimensional system (3.1) as

$$\boldsymbol{z}_k = \boldsymbol{f}(\boldsymbol{V}\breve{\boldsymbol{x}}_k, \boldsymbol{u}_k) + \boldsymbol{\xi}_k\,, \qquad k = 1, \ldots, K \tag{3.2}$$

where the columns of $\boldsymbol{\Xi} = [\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_K]$ are independent random noise vectors defined above and $\breve{\boldsymbol{X}} = [\breve{\boldsymbol{x}}_1, \ldots, \breve{\boldsymbol{x}}_K] = \boldsymbol{V}^T\bar{\boldsymbol{X}}$ is the projection of $\bar{\boldsymbol{X}}$. The noisy re-projected state trajectory is $\breve{\boldsymbol{Z}} = \boldsymbol{V}^T\boldsymbol{Z} = [\breve{\boldsymbol{z}}_1, \ldots, \breve{\boldsymbol{z}}_K]$. Sampling the re-projected trajectory $\breve{\boldsymbol{Z}}$ can be interpreted as sampling bursts of length one of the high-dimensional noisy system from the initial conditions given by the columns of $\breve{\boldsymbol{X}}$.

Remark 3.1. Following the re-projection scheme, the vector $\boldsymbol{V}\breve{\boldsymbol{x}}_{k-1}$ at time step

$k-1$ is used to obtain $\boldsymbol{z}_{k-1}$ and $\boldsymbol{V}\breve{\boldsymbol{x}}_k$ at time step k to obtain $\boldsymbol{z}_k$, instead of using the vector $\boldsymbol{z}_{k-1}$ at time $k-1$ to query the system at time k for $\boldsymbol{z}_k$. This process of generating vectors $\boldsymbol{z}_k$ can be viewed as querying the high-dimensional system (3.1) at $k=1,\ldots,K$ separate, deterministic initial conditions $\boldsymbol{V}\breve{\boldsymbol{x}}_k$ from the columns of $\bar{\boldsymbol{X}}$, which is advantageous because noise remains independent and because it leads to a fixed design regression problem in the following; cf. Remark 3.3. Such a data generation process can be numerically realized by having a dictionary of potential initial conditions, from which K are selected to form the columns of the matrix $\bar{\boldsymbol{X}}$ to query the high-dimensional system and only the state after the first time step of each of the K queried trajectories is used; cf. active operator inference introduced in Section 4. The following analysis bounds the error introduced by the noise added when querying the high-dimensional system via this data generation process.

The corresponding operator-inference problem is

$$\min_{\widehat{\boldsymbol{A}}_1,\ldots,\widehat{\boldsymbol{A}}_\ell,\widehat{\boldsymbol{B}}} \bar{J}(\widehat{\boldsymbol{A}}_1,\ldots,\widehat{\boldsymbol{A}}_\ell,\widehat{\boldsymbol{B}};\breve{\boldsymbol{X}},\breve{\boldsymbol{Z}},\boldsymbol{U}) \tag{3.3}$$

where the noisy re-projected trajectory $\breve{\boldsymbol{Z}}$ enters in the objective (2.5). To analyze the solution of (3.3), it is beneficial to write (3.3) in matrix form as

$$\min_{\boldsymbol{O}} \|\boldsymbol{D}\boldsymbol{O}-\breve{\boldsymbol{Z}}^T\|_F^2\,, \tag{3.4}$$

where the data matrix is $\boldsymbol{D}=[\breve{\boldsymbol{X}}^T,(\breve{\boldsymbol{X}}^2)^T,\ldots,(\breve{\boldsymbol{X}}^\ell)^T,\boldsymbol{U}^T]$ with $\breve{\boldsymbol{X}}^i=[\breve{\boldsymbol{x}}_1^i,\ldots,\breve{\boldsymbol{x}}_K^i]$ for $i=2,\ldots,\ell$. The operators $\widehat{\boldsymbol{A}}_1,\ldots,\widehat{\boldsymbol{A}}_\ell,\widehat{\boldsymbol{B}}$ that we seek are submatrices of $\boldsymbol{O}=[\widehat{\boldsymbol{A}}_1,\ldots,\widehat{\boldsymbol{A}}_\ell,\widehat{\boldsymbol{B}}]^T$. The size of the data matrix $\boldsymbol{D}$ is $K\times M$ with $M=p+\sum_{j=1}^{\ell}n_j$. Correspondingly, the size of $\boldsymbol{O}$ is $M\times n$.

We now characterize the solution of (3.4) with respect to the noise that is added during the re-projection step. Recall that the procedure to generate $\breve{\boldsymbol{Z}}$ is to query the noisy high-dimensional system (3.1) at the columns of the projected trajectory $\breve{\boldsymbol{X}}$, which is deterministic because $\bar{\boldsymbol{X}}$ is deterministic. Thus, the data matrix $\boldsymbol{D}$ in the regression problem (3.4) is deterministic while the noisy re-projected trajectory $\breve{\boldsymbol{Z}}$ is a random matrix.

Following standard results of least-squares regression, the following proposition summarizes that operator inference together with re-projection leads to an unbiased estimator of the projected operators (2.6) whose variance grows linearly with the variance of the noise. Additionally, the upper bound of the MSE of the estimator is controlled by the noise-to-signal ratio $\sigma/s_{\min}(\boldsymbol{D})$, where $s_{\min}(\cdot)$ is the minimum singular value of the matrix argument.

PROPOSITION 3.2. *If $K\geq M$ and $\boldsymbol{D}$ is full rank, then the solution of problem (3.4) is*

$$\widehat{\boldsymbol{O}}=[\widehat{\boldsymbol{A}}_1,\ldots,\widehat{\boldsymbol{A}}_\ell,\widehat{\boldsymbol{B}}]^T=\widetilde{\boldsymbol{O}}+(\boldsymbol{D}^T\boldsymbol{D})^{-1}\boldsymbol{D}^T(\boldsymbol{V}^T\boldsymbol{\Xi})^T\,,$$

where $\widetilde{\boldsymbol{O}}=[\widetilde{\boldsymbol{A}}_1,\ldots,\widetilde{\boldsymbol{A}}_\ell,\widetilde{\boldsymbol{B}}]^T\in\mathbb{R}^{M\times n}$. In particular, the inferred operators are unbiased estimators of the projection-based reduced operators in the sense that $\mathbb{E}[\widehat{\boldsymbol{A}}_j]=\widetilde{\boldsymbol{A}}_j$ for $j=1,\ldots,\ell$ and $\mathbb{E}[\widehat{\boldsymbol{B}}]=\widetilde{\boldsymbol{B}}$. The columns $\widehat{\boldsymbol{o}}_1,\ldots,\widehat{\boldsymbol{o}}_n$ of $\widehat{\boldsymbol{O}}$ are independent random vectors that are distributed as $\widehat{\boldsymbol{o}}_i\sim N(\widetilde{\boldsymbol{o}}_i,\sigma^2(\boldsymbol{D}^T\boldsymbol{D})^{-1})$ for $i=1,\ldots,n$ where $\widetilde{\boldsymbol{o}}_1,\ldots,\widetilde{\boldsymbol{o}}_n\in\mathbb{R}^M$ are the columns of $\widetilde{\boldsymbol{O}}$. In addition, the MSE is bounded as

$$\mathbb{E}[\|\widehat{\boldsymbol{O}}-\widetilde{\boldsymbol{O}}\|_F^2]\leq nM\left(\frac{\sigma}{s_{min}(\boldsymbol{D})}\right)^2. \tag{3.5}$$

Proof. The following are standard arguments from least-squares regression: because the data matrix $\boldsymbol{D}$ is full rank and $K \geq M$, the solution of (3.4) is given by the normal equations

$$\widehat{\boldsymbol{O}} = (\boldsymbol{D}^T\boldsymbol{D})^{-1}\boldsymbol{D}^T\breve{\boldsymbol{Z}}^T = \widetilde{\boldsymbol{O}} + (\boldsymbol{D}^T\boldsymbol{D})^{-1}\boldsymbol{D}^T\breve{\boldsymbol{\Xi}}^T,$$

where $\breve{\boldsymbol{\Xi}} = \boldsymbol{V}^T\boldsymbol{\Xi}$. Since the random vectors $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_K$ have zero mean, the expectation of $\widehat{\boldsymbol{O}}$ is $\mathbb{E}[\widehat{\boldsymbol{O}}] = \widetilde{\boldsymbol{O}}$. Additionally, since $\boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}$ is the identity matrix, the entries of $\breve{\boldsymbol{\Xi}}$ are iid $N(0, \sigma^2)$ random variables which means that the columns of $\breve{\boldsymbol{\Xi}}^T$ are independent Gaussian random vectors of dimension K with an identity covariance matrix scaled by σ^2. Thus, the columns of $\widehat{\boldsymbol{O}}$ are Gaussian with covariance $\sigma^2(\boldsymbol{D}^T\boldsymbol{D})^{-1}$, which leads to the MSE

$$\mathbb{E}[\|\widehat{\boldsymbol{O}} - \widetilde{\boldsymbol{O}}\|_F^2] = \sum_{i=1}^{n}\sum_{j=1}^{M} \mathrm{Var}[\boldsymbol{e}_j^T\widehat{\boldsymbol{o}}_i] = n\,\mathrm{tr}((\boldsymbol{D}^T\boldsymbol{D})^{-1})\sigma^2 \leq Mn\left(\frac{\sigma}{s_{\min}(\boldsymbol{D})}\right)^2,$$

where $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_M$ are the canonical basis vectors of $\mathbb{R}^M$. The first equality follows from the unbiasedness of $\widehat{\boldsymbol{O}}$. ∎

The independence of the columns of the random matrix $\widehat{\boldsymbol{O}}$ leads to the independence of the rows of each of the random matrices $\widehat{\boldsymbol{B}}$ and $\widehat{\boldsymbol{A}}_j$ for $j = 1, \ldots, \ell$. However, since the covariance matrix $\sigma^2(\boldsymbol{D}^T\boldsymbol{D})^{-1}$ of $\widehat{\boldsymbol{o}}_i^T$ is not necessarily block diagonal, the random matrices $\widehat{\boldsymbol{B}}$ and $\widehat{\boldsymbol{A}}_j$ for $j = 1, \ldots, \ell$ are not necessarily independent.

In [42, Chapter 4.3], a sensitivity analysis of lift & learn is presented that just as well applies to operator inference. The analysis leads to bounds with similar right-hand sides as our bound (3.5) on the MSE; however, the analysis in [42] is restricted to deterministic perturbations and no bounds of the error in the state predictions (as in Section 3.2) are presented. Another option is to regularize the least-squares problem underlying operator inference as proposed in [38, 49, 57] to impose a desired bias on the operators; however, because of the assumption of having a full rank data matrix and having data that are generated via re-projection, we can avoid regularization in the following to work with unbiased operator estimators.

Regardless of how $\bar{\boldsymbol{X}}$ is modeled, we pursue the version of the re-projection sampling scheme in [43] instead of that introduced in [39] since it ensures that $\boldsymbol{D}$ and $\breve{\boldsymbol{\Xi}}$ in the normal equations are independent. Indeed, other sampling schemes have been proposed to overcome the dependence between the data matrix and the right-hand side, see [16] for an example in system identification.

Remark 3.3. In the analysis above, we have only focused on the case when $\bar{\boldsymbol{X}}$ is deterministic which can be achieved, for example, by selecting the columns of $\bar{\boldsymbol{X}}$ from a fixed dictionary. This means that (3.4) is a fixed design regression problem since the data matrix $\boldsymbol{D}$ is deterministic [25, 44]. If instead $\bar{\boldsymbol{X}}_r$ is a realization of a trajectory $\bar{\boldsymbol{X}}$ of the noisy system (3.1), then the results above can be interpreted as being conditioned on the realization $\bar{\boldsymbol{X}}_r$, which leads to

$$\mathbb{E}[\|\widehat{\boldsymbol{O}} - \widetilde{\boldsymbol{O}}\|_F^2 \mid \bar{\boldsymbol{X}} = \bar{\boldsymbol{X}}_r] \leq nM\left(\frac{\sigma}{s_{\min}(\boldsymbol{D})}\right)^2,$$

where $\mathbb{E}[\cdot|\cdot]$ denotes the conditional expectation. In a similar way, results can be derived that are conditioned on a realization of $\boldsymbol{V}$ of the basis matrix obtained from trajectories of the noisy system. Unconditional results would rely on being able to

characterize the distribution of the columns of $\bar{\boldsymbol{X}}$ which is challenging due to the nonlinearity of (3.1). For example, in system identification [16], a probabilistic bound on $\|\widehat{\boldsymbol{O}} - \widetilde{\boldsymbol{O}}\|_2$ can be obtained for linear systems provided that the rows of $\boldsymbol{D}$ are Gaussian random vectors.

3.2. Error of predicted states with respect to noise-to-signal ratio. We now consider the random states $\widehat{\boldsymbol{x}}_1, \ldots, \widehat{\boldsymbol{x}}_K$ predicted by the system described by the learned model

$$\widehat{\boldsymbol{x}}_{k+1} = \sum_{j=1}^{\ell} \widehat{\boldsymbol{A}}_j \widehat{\boldsymbol{x}}_k^j + \widehat{\boldsymbol{B}} \boldsymbol{u}_k, \quad k = 0, \ldots, K-1, \tag{3.6}$$

with a deterministic initial state $\widehat{\boldsymbol{x}}_0 \in \mathbb{R}^n$, which potentially is different from the training initial conditions used to generate the re-projected trajectory. Since the operators $\widehat{\boldsymbol{B}}, \widehat{\boldsymbol{A}}_j, j = 1, \ldots, \ell$ are random matrices, $\widehat{\boldsymbol{x}}_k$ is a random vector for $k \geq 1$. In the following, we bound the bias which is the expectation of the difference between the states $\widehat{\boldsymbol{x}}_1, \ldots, \widehat{\boldsymbol{x}}_K$ and the deterministic states $\widetilde{\boldsymbol{x}}_1, \ldots, \widetilde{\boldsymbol{x}}_K$ of the reduced model from intrusive model reduction

$$\widetilde{\boldsymbol{x}}_{k+1} = \sum_{j=1}^{\ell} \widetilde{\boldsymbol{A}}_j \widetilde{\boldsymbol{x}}_k^j + \widetilde{\boldsymbol{B}} \boldsymbol{u}_k, \quad k = 0, \ldots, K-1, \tag{3.7}$$

with the operators $\widetilde{\boldsymbol{B}}, \widetilde{\boldsymbol{A}}_j, j = 1, \ldots, \ell$ defined in (2.6). Bounds for the MSE between the random states $\widehat{\boldsymbol{x}}_1, \ldots, \widehat{\boldsymbol{x}}_K$ and the deterministic states $\widetilde{\boldsymbol{x}}_1, \ldots, \widetilde{\boldsymbol{x}}_K$ are also deduced.

3.2.1. Technical preliminaries. It will be useful to account for the difference between the inferred operators and the operators from intrusive model reduction. Let $\boldsymbol{E}_{\widehat{\boldsymbol{A}}_j}, \boldsymbol{E}_{\widehat{\boldsymbol{B}}}$ be $n \times n_j$ and $n \times p$ random matrices, respectively, such that

$$\widehat{\boldsymbol{A}}_j = \widetilde{\boldsymbol{A}}_j + \boldsymbol{E}_{\widehat{\boldsymbol{A}}_j}, j = 1, \ldots, \ell, \quad \text{and} \quad \widehat{\boldsymbol{B}} = \widetilde{\boldsymbol{B}} + \boldsymbol{E}_{\widehat{\boldsymbol{B}}}.$$

The distribution of the rows of $\boldsymbol{E}_{\widehat{\boldsymbol{A}}_j}, \boldsymbol{E}_{\widehat{\boldsymbol{B}}}$ can be described as follows. Define the selection matrices $\boldsymbol{P}_{\boldsymbol{A}_j} \in \mathbb{R}^{n_j \times M}$ for $j = 1, \ldots, \ell$ and $\boldsymbol{P}_{\boldsymbol{B}} \in \mathbb{R}^{p \times M}$ which satisfy

$$\boldsymbol{P}_{\boldsymbol{A}_j} \widehat{\boldsymbol{O}} = \widehat{\boldsymbol{A}}_j^T \quad \text{and} \quad \boldsymbol{P}_{\boldsymbol{B}} \widehat{\boldsymbol{O}} = \widehat{\boldsymbol{B}}^T.$$

For $i = 1, \ldots, n$, the i-th row of $\boldsymbol{E}_{\widehat{\boldsymbol{A}}_j}$ and $\boldsymbol{E}_{\widehat{\boldsymbol{B}}}$ are zero-mean multivariate Gaussian random vectors with covariance matrices $\sigma^2 \Sigma_{\widehat{\boldsymbol{A}}_j}$ and $\sigma^2 \Sigma_{\widehat{\boldsymbol{B}}}$, respectively, where $\Sigma_{\widehat{\boldsymbol{A}}_j} = \boldsymbol{P}_{\boldsymbol{A}_j} (\boldsymbol{D}^T \boldsymbol{D})^{-1} \boldsymbol{P}_{\boldsymbol{A}_j}^T$ and $\Sigma_{\widehat{\boldsymbol{B}}} = \boldsymbol{P}_{\boldsymbol{B}} (\boldsymbol{D}^T \boldsymbol{D})^{-1} \boldsymbol{P}_{\boldsymbol{B}}^T$. Observe that

$$\|\Sigma_{\widehat{\boldsymbol{B}}}^{1/2}\|_2 = \|\boldsymbol{P}_{\boldsymbol{B}} (\boldsymbol{D}^T \boldsymbol{D})^{-1} \boldsymbol{P}_{\boldsymbol{B}}^T\|_2^{1/2} \leq \|(\boldsymbol{D}^T \boldsymbol{D})^{-1}\|_2^{1/2} = \sqrt{s_{\max}((\boldsymbol{D}^T \boldsymbol{D})^{-1})} = \frac{1}{s_{\min}(\boldsymbol{D})} \tag{3.8}$$

where $s_{\max}(\cdot)$ is the largest singular value of the matrix argument. Analogously, we have

$$\|\Sigma_{\widehat{\boldsymbol{A}}_j}^{1/2}\|_2 \leq \frac{1}{s_{\min}(\boldsymbol{D})}, \quad j = 1, \ldots, \ell. \tag{3.9}$$

The following is a technical lemma derived from [63, Theorem 5.32 and Proposition 5.34] that provides an upper bound for the expected value of the powers of the norm of a Gaussian random matrix, which will be utilized in the calculations below; cf. Appendix A for the proof.

LEMMA 3.4 (see, e.g., Theorem 5.32 and Proposition 5.34 in [63]). *Let $\boldsymbol{G}$ be an $n\times p$ random matrix whose entries are independent standard normal random variables. For $l \in \mathbb{N}$,*

$$\mathbb{E}[\|\boldsymbol{G}\|_2^l] \leq (\sqrt{n} + \sqrt{p} + 2^{1/l}\sqrt{l})^l. \tag{3.10}$$

3.2.2. Error in states for linear systems. In this section, we consider only systems with $\ell = 1$ and therefore drop the subscript in $\boldsymbol{A}_1, \widetilde{\boldsymbol{A}}_1, \widehat{\boldsymbol{A}}_1$. The operator-inference model is $\widehat{\boldsymbol{x}}_{k+1} = \widehat{\boldsymbol{A}}\widehat{\boldsymbol{x}}_k + \widehat{\boldsymbol{B}}\boldsymbol{u}_k$ and the model from intrusive model reduction is $\widetilde{\boldsymbol{x}}_{k+1} = \widetilde{\boldsymbol{A}}\widetilde{\boldsymbol{x}}_k + \widetilde{\boldsymbol{B}}\boldsymbol{u}_k$.

PROPOSITION 3.5. *Let $\widehat{\boldsymbol{x}}_0 = \widetilde{\boldsymbol{x}}_0$. Suppose that the conditions of Proposition 3.2 hold. If the high-dimensional system (3.1) from which data are sampled and the learned low-dimensional model have linear state dependence, for $k \in \mathbb{N}$ with $k \geq 1$, the bias of the state predictions is bounded as*

$$\|\mathbb{E}[\widehat{\boldsymbol{x}}_k - \widetilde{\boldsymbol{x}}_k]\|_2 \leq \sum_{l=2}^{k} C_l \left(\frac{\sigma}{s_{min}(\boldsymbol{D})}\right)^l, \tag{3.11}$$

where $0 < C_2, \ldots, C_k$ are constants that are not functions of σ and $s_{min}(\boldsymbol{D})$. The constants are

$$C_l = (2\sqrt{n} + 2^{1/l}\sqrt{l})^l \left[\binom{k}{l}\|\widetilde{\boldsymbol{A}}\|_2^{k-l}\|\widetilde{\boldsymbol{x}}_0\|_2 + \sum_{i=l}^{k-1}\binom{i}{l}\|\widetilde{\boldsymbol{A}}\|_2^{i-l}\|\widetilde{\boldsymbol{B}}\boldsymbol{u}_{k-1-i}\|_2\right] \tag{3.12}$$
$$+ \sum_{i=l-1}^{k-1}\binom{i}{l-1}\|\widetilde{\boldsymbol{A}}\|_2^{i-l+1}\|\boldsymbol{u}_{k-1-i}\|_2\left(2\sqrt{n} + 2^{\frac{1}{2(i-l+1)}}\sqrt{2(i-l+1)}\right)^{i-l+1}(\sqrt{n}+\sqrt{p}+2),$$

for $l = 2, \ldots, k$.

Proof. Define the $n \times n$ random matrix $\boldsymbol{G}_{\widehat{\boldsymbol{A}}}$ as $\boldsymbol{G}_{\widehat{\boldsymbol{A}}} = \frac{1}{\sigma}\Sigma_{\widehat{\boldsymbol{A}}}^{-1/2}\boldsymbol{E}_{\widehat{\boldsymbol{A}}}^T$ and the $p \times n$ random matrix $\boldsymbol{G}_{\widehat{\boldsymbol{B}}}$ as $\frac{1}{\sigma}\Sigma_{\widehat{\boldsymbol{B}}}^{-1/2}\boldsymbol{E}_{\widehat{\boldsymbol{B}}}^T$. Observe that the entries of $\boldsymbol{G}_{\widehat{\boldsymbol{A}}}, \boldsymbol{G}_{\widehat{\boldsymbol{B}}}$ are independent standard random variables. At time step k, the solution to the reduced system using the inferred operators is

$$\widehat{\boldsymbol{x}}_k = \widehat{\boldsymbol{A}}^k\widetilde{\boldsymbol{x}}_0 + \sum_{i=0}^{k-1}\widehat{\boldsymbol{A}}^i\widehat{\boldsymbol{B}}\boldsymbol{u}_{k-1-i}. \tag{3.13}$$

We now introduce the following notation: Let $\boldsymbol{M}, \boldsymbol{N}$ be square matrices of the same size. For $m, i \in \mathbb{N}$, denote by $\rho_1(\boldsymbol{M}, \boldsymbol{N}; i, m-i), \ldots, \rho_{\binom{m}{i}}(\boldsymbol{M}, \boldsymbol{N}; i, m-i)$ all the $\binom{m}{i}$ possible matrix products with i multiplications of $\boldsymbol{M}$ and $m-i$ multiplications of $\boldsymbol{N}$. For example, if $i = 1, m = 3$ then $\rho_1(\boldsymbol{M}, \boldsymbol{N}; 1, 2) = \boldsymbol{M}\boldsymbol{N}^2, \rho_2(\boldsymbol{M}, \boldsymbol{N}; 1, 2) = \boldsymbol{N}\boldsymbol{M}\boldsymbol{N}$, and $\rho_3(\boldsymbol{M}, \boldsymbol{N}; 1, 2) = \boldsymbol{N}^2\boldsymbol{M}$.

We then have with $\widehat{\boldsymbol{A}} = \widetilde{\boldsymbol{A}} + \boldsymbol{E}_{\widehat{\boldsymbol{A}}}$ that

$$\widehat{\boldsymbol{A}}^k = \sum_{l=0}^{k}\sum_{j=1}^{\binom{k}{l}}\rho_j(\widetilde{\boldsymbol{A}}, \boldsymbol{E}_{\widehat{\boldsymbol{A}}}; k-l, l),$$

which we substitute into (3.13) at time step k, to obtain

$$\begin{aligned}\widehat{\boldsymbol{x}}_k &= \sum_{l=0}^{k}\sum_{j=1}^{\binom{k}{l}}\rho_j(\widetilde{\boldsymbol{A}},\boldsymbol{E}_{\widehat{\boldsymbol{A}}};k-l,l)\widetilde{\boldsymbol{x}}_0+\sum_{i=0}^{k-1}\sum_{l=0}^{i}\sum_{j=1}^{\binom{i}{l}}\rho_j(\widetilde{\boldsymbol{A}},\boldsymbol{E}_{\widehat{\boldsymbol{A}}};i-l,l)\widehat{\boldsymbol{B}}\boldsymbol{u}_{k-1-i}\\ &= \sum_{l=0}^{k}\sum_{j=1}^{\binom{k}{l}}\rho_j(\widetilde{\boldsymbol{A}},\boldsymbol{E}_{\widehat{\boldsymbol{A}}};k-l,l)\widetilde{\boldsymbol{x}}_0+\sum_{l=0}^{k-1}\sum_{i=l}^{k-1}\sum_{j=1}^{\binom{i}{l}}\rho_j(\widetilde{\boldsymbol{A}},\boldsymbol{E}_{\widehat{\boldsymbol{A}}};i-l,l)\widetilde{\boldsymbol{B}}\boldsymbol{u}_{k-1-i}\\ &\quad+\sum_{l=0}^{k-1}\sum_{i=l}^{k-1}\sum_{j=1}^{\binom{i}{l}}\rho_j(\widetilde{\boldsymbol{A}},\boldsymbol{E}_{\widehat{\boldsymbol{A}}};i-l,l)\boldsymbol{E}_{\widehat{\boldsymbol{B}}}\boldsymbol{u}_{k-1-i}\end{aligned}\tag{3.14}$$

where in the second equality, we used $\widehat{\boldsymbol{B}}=\widetilde{\boldsymbol{B}}+\boldsymbol{E}_{\widehat{\boldsymbol{B}}}$ and interchanged the order of the summation for the last two terms. Notice that for the state obtained with intrusive model reduction we have

$$\widetilde{\boldsymbol{x}}_k = \sum_{j=1}^{\binom{k}{0}}\rho_j(\widetilde{\boldsymbol{A}},\boldsymbol{E}_{\widehat{\boldsymbol{A}}};k,0)\widetilde{\boldsymbol{x}}_0+\sum_{i=0}^{k-1}\sum_{j=1}^{\binom{i}{0}}\rho_j(\widetilde{\boldsymbol{A}},\boldsymbol{E}_{\widehat{\boldsymbol{A}}};i,0)\widetilde{\boldsymbol{B}}\boldsymbol{u}_{k-1-i}\tag{3.15}$$

which corresponds to the first 2 terms of (3.14) but with $l=0$ fixed. Thus, (3.15) consists of all terms in (3.14) where the random matrices $\boldsymbol{E}_{\widehat{\boldsymbol{A}}},\boldsymbol{E}_{\widehat{\boldsymbol{B}}}$ are absent. Hence,

$$\begin{aligned}\widehat{\boldsymbol{x}}_k-\widetilde{\boldsymbol{x}}_k &= \sum_{l=1}^{k}\sum_{j=1}^{\binom{k}{l}}\rho_j(\widetilde{\boldsymbol{A}},\boldsymbol{E}_{\widehat{\boldsymbol{A}}};k-l,l)\widetilde{\boldsymbol{x}}_0+\sum_{l=1}^{k-1}\sum_{i=l}^{k-1}\sum_{j=1}^{\binom{i}{l}}\rho_j(\widetilde{\boldsymbol{A}},\boldsymbol{E}_{\widehat{\boldsymbol{A}}};i-l,l)\widetilde{\boldsymbol{B}}\boldsymbol{u}_{k-1-i}\\ &\quad+\sum_{l=0}^{k-1}\sum_{i=l}^{k-1}\sum_{j=1}^{\binom{i}{l}}\rho_j(\widetilde{\boldsymbol{A}},\boldsymbol{E}_{\widehat{\boldsymbol{A}}};i-l,l)\boldsymbol{E}_{\widehat{\boldsymbol{B}}}\boldsymbol{u}_{k-1-i}.\end{aligned}$$

Additionally, when $l=1$, the terms $\mathbb{E}[\rho_j(\widetilde{\boldsymbol{A}},\boldsymbol{E}_{\widehat{\boldsymbol{A}}};k-l,l)]$ and $\mathbb{E}[\rho_j(\widetilde{\boldsymbol{A}},\boldsymbol{E}_{\widehat{\boldsymbol{A}}};i-l,l)]$ are zero because $\boldsymbol{E}_{\widehat{\boldsymbol{A}}}$ has zero mean. Similarly, for $l=0$, the terms $\mathbb{E}[\rho_j(\widetilde{\boldsymbol{A}},\boldsymbol{E}_{\widehat{\boldsymbol{A}}};i,0)\boldsymbol{E}_{\widehat{\boldsymbol{B}}}]$ are zero. This means that

$$\|\mathbb{E}[\widehat{\boldsymbol{x}}_k-\widetilde{\boldsymbol{x}}_k]\|_2\le\tau_1+\tau_2+\tau_3\tag{3.16}$$

where

$$\tau_1=\sum_{l=2}^{k}\sum_{j=1}^{\binom{k}{l}}\|\mathbb{E}[\rho_j(\widetilde{\boldsymbol{A}},\boldsymbol{E}_{\widehat{\boldsymbol{A}}};k-l,l)\widetilde{\boldsymbol{x}}_0]\|_2,\ \tau_2=\sum_{l=2}^{k-1}\sum_{i=l}^{k-1}\sum_{j=1}^{\binom{i}{l}}\|\mathbb{E}[\rho_j(\widetilde{\boldsymbol{A}},\boldsymbol{E}_{\widehat{\boldsymbol{A}}};i-l,l)\widetilde{\boldsymbol{B}}\boldsymbol{u}_{k-1-i}]\|_2,$$

$$\tau_3=\sum_{l=1}^{k-1}\sum_{i=l}^{k-1}\sum_{j=1}^{\binom{i}{l}}\|\mathbb{E}[\rho_j(\widetilde{\boldsymbol{A}},\boldsymbol{E}_{\widehat{\boldsymbol{A}}};i-l,l)\boldsymbol{E}_{\widehat{\boldsymbol{B}}}\boldsymbol{u}_{k-1-i}]\|_2.$$

It remains to bound each of τ_1,τ_2,τ_3. Since

$$\begin{aligned}\|\mathbb{E}[\rho_j(\widetilde{\boldsymbol{A}},\boldsymbol{E}_{\widehat{\boldsymbol{A}}};k-l,l)\widetilde{\boldsymbol{x}}_0]\|_2 &= \|\mathbb{E}[\rho_j(\widetilde{\boldsymbol{A}},\sigma\boldsymbol{G}_{\widehat{\boldsymbol{A}}}^T\Sigma_{\widehat{\boldsymbol{A}}}^{1/2};k-l,l)]\widetilde{\boldsymbol{x}}_0\|_2\\ &\le\sigma^l\|\Sigma_{\widehat{\boldsymbol{A}}}^{1/2}\|_2^l\|\widetilde{\boldsymbol{A}}\|_2^{k-l}\|\widetilde{\boldsymbol{x}}_0\|_2\mathbb{E}[\|\boldsymbol{G}_{\widehat{\boldsymbol{A}}}\|_2^l]\\ &\le\left(\frac{\sigma}{s_{min}(\boldsymbol{D})}\right)^l\|\widetilde{\boldsymbol{A}}\|_2^{k-l}\|\widetilde{\boldsymbol{x}}_0\|_2\mathbb{E}[\|\boldsymbol{G}_{\widehat{\boldsymbol{A}}}\|_2^l],\end{aligned}$$

we obtain

$$\begin{aligned}\tau_1 &\leq \sum_{l=2}^{k} \binom{k}{l} \left(\frac{\sigma}{s_{\min}(\boldsymbol{D})}\right)^l \|\widetilde{\boldsymbol{A}}\|_2^{k-l} \|\widetilde{\boldsymbol{x}}_0\|_2 \mathbb{E}[\|\boldsymbol{G}_{\widehat{\boldsymbol{A}}}|\|_2^l] \\ &\leq \sum_{l=2}^{k} \binom{k}{l} \left(\frac{\sigma}{s_{\min}(\boldsymbol{D})}\right)^l \|\widetilde{\boldsymbol{A}}\|_2^{k-l} \|\widetilde{\boldsymbol{x}}_0\|_2 (2\sqrt{n} + 2^{1/l}\sqrt{l})^l\end{aligned}$$

by applying Lemma 3.4. Likewise,

$$\begin{aligned}\tau_2 &\leq \sum_{l=2}^{k-1} \sum_{i=l}^{k-1} \binom{i}{l} \left(\frac{\sigma}{s_{\min}(\boldsymbol{D})}\right)^l \|\widetilde{\boldsymbol{A}}\|_2^{i-l} \|\widetilde{\boldsymbol{B}}\boldsymbol{u}_{k-1-i}\|_2 \mathbb{E}[\|\boldsymbol{G}_{\widehat{\boldsymbol{A}}}|\|_2^l] \\ &\leq \sum_{l=2}^{k-1} \sum_{i=l}^{k-1} \binom{i}{l} \left(\frac{\sigma}{s_{\min}(\boldsymbol{D})}\right)^l \|\widetilde{\boldsymbol{A}}\|_2^{i-l} \|\widetilde{\boldsymbol{B}}\boldsymbol{u}_{k-1-i}\|_2 (2\sqrt{n} + 2^{1/l}\sqrt{l})^l.\end{aligned}$$

Finally,

$$\begin{aligned}\|\mathbb{E}[\rho_j(\widetilde{\boldsymbol{A}}, \boldsymbol{E}_{\widehat{\boldsymbol{A}}}; i-l, l)\boldsymbol{E}_{\widehat{\boldsymbol{B}}}\boldsymbol{u}_{k-1-i}]\|_2 &\leq \mathbb{E}[\|\rho_j(\widetilde{\boldsymbol{A}}, \sigma\boldsymbol{G}_{\widehat{\boldsymbol{A}}}^T \Sigma_{\widehat{\boldsymbol{A}}}^{1/2}; i-l, l)\sigma\boldsymbol{G}_{\widehat{\boldsymbol{B}}}^T \Sigma_{\widehat{\boldsymbol{B}}}^{1/2} \boldsymbol{u}_{k-1-i}\|_2] \\ &\leq \sigma^{l+1} \|\Sigma_{\widehat{\boldsymbol{A}}}^{1/2}\|_2^l \|\Sigma_{\widehat{\boldsymbol{B}}}^{1/2}\|_2 \|\widetilde{\boldsymbol{A}}\|_2^{i-l} \|\boldsymbol{u}_{k-1-i}\|_2 \mathbb{E}[\|\boldsymbol{G}_{\widehat{\boldsymbol{A}}}\|_2^{i-l} \|\boldsymbol{G}_{\widehat{\boldsymbol{B}}}\|_2] \\ &\leq \left(\frac{\sigma}{s_{\min}(\boldsymbol{D})}\right)^{l+1} \|\widetilde{\boldsymbol{A}}\|_2^{i-l} \|\boldsymbol{u}_{k-1-i}\|_2 \mathbb{E}[\|\boldsymbol{G}_{\widehat{\boldsymbol{A}}}\|_2^{i-l} \|\boldsymbol{G}_{\widehat{\boldsymbol{B}}}\|_2]\end{aligned}$$

so that

$$\begin{aligned}\tau_3 \leq \sum_{l=2}^{k} \sum_{i=l-1}^{k-1} \binom{i}{l-1} \left(\frac{\sigma}{s_{\min}(\boldsymbol{D})}\right)^l \|\widetilde{\boldsymbol{A}}\|_2^{i-l+1} \|\boldsymbol{u}_{k-1-i}\|_2 \\ \left(2\sqrt{n} + 2^{\frac{1}{2(i-l+1)}} \sqrt{2(i-l+1)}\right)^{i-l+1} (\sqrt{n} + \sqrt{p} + 2)\end{aligned} \tag{3.17}$$

because the Cauchy-Schwarz inequality and Lemma 3.4 lead to

$$\begin{aligned}\left|\mathbb{E}[\|\boldsymbol{G}_{\widehat{\boldsymbol{A}}}\|_2^{i-l+1} \|\boldsymbol{G}_{\widehat{\boldsymbol{B}}}\|_2]\right| &\leq \sqrt{\mathbb{E}[\|\boldsymbol{G}_{\widehat{\boldsymbol{A}}}\|_2^{2(i-l+1)}] \mathbb{E}[\|\boldsymbol{G}_{\widehat{\boldsymbol{B}}}\|_2^2]} \\ &\leq \left(2\sqrt{n} + 2^{\frac{1}{2(i-l+1)}} \sqrt{2(i-l+1)}\right)^{i-l+1} (\sqrt{n} + \sqrt{p} + 2).\end{aligned}$$

The result follows by combining the upper bounds for τ_1, τ_2, τ_3. ∎

COROLLARY 3.6. *Let $\widehat{\boldsymbol{x}}_0 = \widetilde{\boldsymbol{x}}_0$. If $\ell = 1$ and high-dimensional system (3.1) from which data are sampled is autonomous, then*

$$\|\mathbb{E}[\widehat{\boldsymbol{x}}_k - \widetilde{\boldsymbol{x}}_k]\|_2 \leq \sum_{l=2}^{k} \binom{k}{l} \left(\frac{\sigma}{s_{min}(\boldsymbol{D})}\right)^l (2\sqrt{n} + 2^{1/l}\sqrt{l})^l \|\widetilde{\boldsymbol{A}}\|_2^{k-l} \|\widetilde{\boldsymbol{x}}_0\|_2, \tag{3.18}$$

for $k \in \mathbb{N}$ with $k \geq 1$.

Proof. The proof follows that of Proposition 3.5 noting that $\widehat{\boldsymbol{B}}, \widetilde{\boldsymbol{B}}$ and hence $\boldsymbol{E}_{\widehat{\boldsymbol{B}}}$ are zero matrices. ∎

Several remarks are in order. As the time step k increases, i.e., as we move forward in time, the bound (3.11) also increases, which is expected because the bias of the state estimators of previous time steps is accumulated. Notice that the bound also depends on n, the dimension of the reduced space, and on p, the dimension of the input. The bound (3.11) further suggests that if the noise-to-signal ratio $\sigma/s_{\min}(\boldsymbol{D}) < 1$, the term associated with $(\sigma/s_{\min}(\boldsymbol{D}))^2$ at time step $k = 2$ dominates the upper bound as $\sigma/s_{\min}(\boldsymbol{D}) \to 0$. Hence, we expect that for $\sigma/s_{\min}(\boldsymbol{D})$ sufficiently small, an order of magnitude decrease in the noise-to-signal ratio yields at least a decrease of 2 orders of magnitude in the bias of the predicted states.

3.2.3. Error of state predictions with polynomially nonlinear systems. We now derive bounds for the bias $\|\mathbb{E}[\widehat{\boldsymbol{x}}_k - \widetilde{\boldsymbol{x}}_k]\|_2$ and the MSE $\mathbb{E}[\|\widehat{\boldsymbol{x}}_k - \widetilde{\boldsymbol{x}}_k\|_2^2]$ of state predictions where data are sampled from polynomially nonlinear systems.

We start by writing the state $\widehat{\boldsymbol{x}}_k$ at time step k (which only involves the initial condition and previous inputs) as a sum of vectors, each of which is formed as a combination of matrix and Kronecker products.

LEMMA 3.7. *The state $\widehat{\boldsymbol{x}}_k$ at time step $k, k \in \mathbb{N}$, of the polynomially nonlinear model (3.6) is*

$$\widehat{\boldsymbol{x}}_k = \sum_{l=0}^{Q_k} \sum_{\substack{j_1,\ldots,j_{\ell+1}\in\mathbb{N} \\ j_1+\cdots+j_{\ell+1}=l}} \zeta_{j_1,\ldots,j_{\ell+1}}(\boldsymbol{E}_{\widehat{\boldsymbol{A}}_1},\ldots,\boldsymbol{E}_{\widehat{\boldsymbol{A}}_\ell},\boldsymbol{E}_{\widehat{\boldsymbol{B}}}), \tag{3.19}$$

where $Q_k = (\ell^k - 1)/(\ell - 1)$ and $\zeta_{j_1,\ldots,j_{\ell+1}}$ is a sum of vectors in $\mathbb{R}^n$, where each term is a combination of matrix and Kronecker products involving the deterministic quantities $\widehat{\boldsymbol{x}}_0, \boldsymbol{u}_0, \ldots, \boldsymbol{u}_{k-1}, \widetilde{\boldsymbol{A}}_1, \ldots, \widetilde{\boldsymbol{A}}_\ell, \widetilde{\boldsymbol{B}}$ and the random matrices $\boldsymbol{E}_{\widehat{\boldsymbol{A}}_1}, \ldots, \boldsymbol{E}_{\widehat{\boldsymbol{A}}_\ell}, \boldsymbol{E}_{\widehat{\boldsymbol{B}}}$. Each term in the sum $\zeta_{j_1,\ldots,j_{\ell+1}}$ consists of j_l multiplications of $\boldsymbol{E}_{\widehat{\boldsymbol{A}}_l}$ for $l = 1, \ldots, \ell$ and $j_{\ell+1}$ multiplications of $\boldsymbol{E}_{\widehat{\boldsymbol{B}}}$.

Proof. We recast the system (3.6) as

$$\widehat{\boldsymbol{x}}_{k+1} = \sum_{j=1}^{\ell} \widehat{\boldsymbol{A}}_j \boldsymbol{S}_j(\widehat{\boldsymbol{x}}_k \otimes \cdots \otimes \widehat{\boldsymbol{x}}_k) + \widehat{\boldsymbol{B}}\boldsymbol{u}_k, \tag{3.20}$$

where $\boldsymbol{S}_j \in \mathbb{R}^{n_j \times n^j}$ is a selection matrix such that $\boldsymbol{S}_j(\widehat{\boldsymbol{x}}_k \otimes \cdots \otimes \widehat{\boldsymbol{x}}_k) = \widehat{\boldsymbol{x}}_k^j$ for $j = 1, \ldots, \ell$. Thus, the state $\widehat{\boldsymbol{x}}_{k+1}$ at time step $k+1$ is the result of recursively applying (3.20) until the right-hand side contains only the initial condition $\widehat{\boldsymbol{x}}_0$ and the inputs $\boldsymbol{u}_0, \ldots, \boldsymbol{u}_{k-1}$. Since $\widehat{\boldsymbol{A}}_j = \widetilde{\boldsymbol{A}}_j + \boldsymbol{E}_{\widehat{\boldsymbol{A}}_j}$ for $j = 1, \ldots, \ell$ and $\widehat{\boldsymbol{B}} = \widetilde{\boldsymbol{B}} + \boldsymbol{E}_{\widehat{\boldsymbol{B}}}$, the right-hand side is a sum of combinations of matrix and Kronecker products involving the deterministic vectors $\widehat{\boldsymbol{x}}_0, \boldsymbol{u}_0, \ldots, \boldsymbol{u}_{k-1}$, the deterministic matrices $\widetilde{\boldsymbol{A}}_1, \ldots, \widetilde{\boldsymbol{A}}_\ell, \widetilde{\boldsymbol{B}}, \boldsymbol{S}_1, \ldots, \boldsymbol{S}_\ell$ and the random matrices $\boldsymbol{E}_{\widehat{\boldsymbol{A}}_1}, \ldots, \boldsymbol{E}_{\widehat{\boldsymbol{A}}_\ell}, \boldsymbol{E}_{\widehat{\boldsymbol{B}}}$. The right-hand side can then be ordered with respect to the number of times there is a multiplication with a random matrix which is represented by the outer sum in (3.19). The outer sum is further partitioned according to how often there is a multiplication involving $\boldsymbol{E}_{\widehat{\boldsymbol{A}}_1}, \ldots, \boldsymbol{E}_{\widehat{\boldsymbol{A}}_\ell}, \boldsymbol{E}_{\widehat{\boldsymbol{B}}}$ with corresponding frequencies of multiplications $j_1, \ldots, j_{\ell+1}$ times. The frequencies $j_1, \ldots, j_{\ell+1}$ serve as indices of the inner sum (3.19).

It remains to show that the state at time step k is obtained using at most Q_k multiplications with a random matrix. We proceed via induction. When $k = 1$,

$$\widehat{\boldsymbol{x}}_1 = \sum_{j=1}^{\ell} \widetilde{\boldsymbol{A}}_j \boldsymbol{S}_j(\widehat{\boldsymbol{x}}_0 \otimes \cdots \otimes \widehat{\boldsymbol{x}}_0) + \widetilde{\boldsymbol{B}}\boldsymbol{u}_0 + \sum_{j=1}^{\ell} \boldsymbol{E}_{\widehat{\boldsymbol{A}}_1} \boldsymbol{S}_j(\widehat{\boldsymbol{x}}_0 \otimes \cdots \otimes \widehat{\boldsymbol{x}}_0) + \boldsymbol{E}_{\widehat{\boldsymbol{B}}}\boldsymbol{u}_0,$$

thereby implying that there is at most one ($Q_1 = 1$) random-matrix multiplication to obtain $\widehat{\boldsymbol{x}}_1$. Suppose that at time step $k = m$, obtaining the state $\widehat{\boldsymbol{x}}_m$ requires at most Q_m random-matrix multiplications. At time step $k = m+1$, the maximum number of random-matrix multiplications is determined by the expression $\boldsymbol{E}_{\widehat{\boldsymbol{A}}_\ell}\boldsymbol{S}_\ell(\widehat{\boldsymbol{x}}_m \otimes \cdots \otimes \widehat{\boldsymbol{x}}_m)$. From the induction step, $\widehat{\boldsymbol{x}}_m$ has at most $Q_m = (\ell^m - 1)/(\ell - 1)$ random matrix multiplications which means that $\widehat{\boldsymbol{x}}_{m+1}$ has at most $Q_m\ell + 1 = (\ell^{m+1} - 1)/(\ell - 1) = Q_{m+1}$ random matrix multiplications due to the ℓ Kronecker products of $\widehat{\boldsymbol{x}}_m$ and the random matrix $\boldsymbol{E}_{\widehat{\boldsymbol{A}}_\ell}$. ∎

The following proposition shows that the bound for the bias $\|\mathbb{E}[\widehat{\boldsymbol{x}}_k - \widetilde{\boldsymbol{x}}_k]\|_2$ of state predictions is still polynomial in terms of the noise-to-signal ratio even when data are sampled from polynomially nonlinear systems and polynomially nonlinear models are learned. In particular, when $\sigma/s_{\min}(\boldsymbol{D}) < 1$ and $\sigma/s_{\min}(\boldsymbol{D}) \to 0$, the behavior of the upper bound is dominated by the term associated with $(\sigma/s_{\min}(\boldsymbol{D}))^2$.

PROPOSITION 3.8. *Let $\widehat{\boldsymbol{x}}_0 = \widetilde{\boldsymbol{x}}_0$. Suppose that the conditions of Proposition 3.2 hold. If the high-dimensional system (3.1) from which data are sampled and the learned low-dimensional model are polynomially nonlinear, for $k \in \mathbb{N}$ with $k \geq 1$, it holds that*

$$\|\mathbb{E}[\widehat{\boldsymbol{x}}_k - \widetilde{\boldsymbol{x}}_k]\|_2 \leq \sum_{l=2}^{Q_k} \bar{C}_l \left(\frac{\sigma}{s_{min}(\boldsymbol{D})}\right)^l$$

for some constants $0 < \bar{C}_l < \infty, l = 2, \ldots, Q_k$, which are not functions of σ and $\boldsymbol{D}$.

Proof. Define the $n_j \times n$ random matrix $\boldsymbol{G}_{\widehat{\boldsymbol{A}}_j}$ as $\boldsymbol{G}_{\widehat{\boldsymbol{A}}_j} = \frac{1}{\sigma}\Sigma_{\widehat{\boldsymbol{A}}_j}^{-1/2}\boldsymbol{E}_{\widehat{\boldsymbol{A}}_j}^T$ for $j = 1, \ldots, \ell$ and the $p \times n$ random matrix $\boldsymbol{G}_{\widehat{\boldsymbol{B}}}$ as $\boldsymbol{G}_{\widehat{\boldsymbol{B}}} = \frac{1}{\sigma}\Sigma_{\widehat{\boldsymbol{B}}}^{-1/2}\boldsymbol{E}_{\widehat{\boldsymbol{B}}}^T$. Observe that the entries of $\boldsymbol{G}_{\widehat{\boldsymbol{A}}_j}$ for $j = 1, \ldots, \ell$ and $\boldsymbol{G}_{\widehat{\boldsymbol{B}}}$ are independent standard normal random variables, however, in general, the random matrices are dependent due to the dependence between $\widehat{\boldsymbol{B}}, \widehat{\boldsymbol{A}}_j, j = 1, \ldots, \ell$. According to Lemma 3.7, the state $\widehat{\boldsymbol{x}}_k$ at time step k is

$$\widehat{\boldsymbol{x}}_k = \sum_{l=0}^{Q_k} \sum_{\substack{j_1,\ldots,j_{\ell+1} \in \mathbb{N} \\ j_1+\cdots+j_{\ell+1}=l}} \zeta_{j_1,\ldots,j_{\ell+1}}(\boldsymbol{E}_{\widehat{\boldsymbol{A}}_1}, \ldots, \boldsymbol{E}_{\widehat{\boldsymbol{A}}_\ell}, \boldsymbol{E}_{\widehat{\boldsymbol{B}}}). \tag{3.21}$$

Since the system is polynomially nonlinear, the state obtained with intrusive model reduction,

$$\widetilde{\boldsymbol{x}}_k = \zeta_{0,\ldots,0}(\boldsymbol{E}_{\widehat{\boldsymbol{A}}_1}, \ldots, \boldsymbol{E}_{\widehat{\boldsymbol{A}}_\ell}, \boldsymbol{E}_{\widehat{\boldsymbol{B}}}) \tag{3.22}$$

is comprised of those terms for which no random matrix is present in the multiplications ($l = 0$). We now isolate the terms in (3.21) in which a single random matrix is involved in the multiplication. By linearity of expectation and using that the random matrices $\boldsymbol{E}_{\widehat{\boldsymbol{B}}}, \boldsymbol{E}_{\widehat{\boldsymbol{A}}_s}, s = 1, \ldots, \ell$ have zero mean,

$$\mathbb{E}\left[\sum_{\substack{j_1,\ldots,j_{\ell+1} \\ j_1+\cdots+j_{\ell+1}=1}} \zeta_{j_1,\ldots,j_{\ell+1}}(\boldsymbol{E}_{\widehat{\boldsymbol{A}}_1}, \ldots, \boldsymbol{E}_{\widehat{\boldsymbol{A}}_\ell}, \boldsymbol{E}_{\widehat{\boldsymbol{B}}})\right] = \boldsymbol{0}.$$

Therefore, with the triangle and Jensen's inequality follows

$$\|\mathbb{E}[\widehat{\boldsymbol{x}}_k - \widetilde{\boldsymbol{x}}_k]\|_2 \leq \sum_{l=2}^{Q_k} \sum_{\substack{j_1,\ldots,j_{\ell+1} \in \mathbb{N} \\ j_1+\cdots+j_{\ell+1}=l}} \mathbb{E}\left[\|\zeta_{j_1,\ldots,j_{\ell+1}}(\boldsymbol{E}_{\widehat{\boldsymbol{A}}_1}, \ldots, \boldsymbol{E}_{\widehat{\boldsymbol{A}}_\ell}, \boldsymbol{E}_{\widehat{\boldsymbol{B}}})\|_2\right]. \tag{3.23}$$

It remains to bound $\mathbb{E}\left[\|\zeta_{j_1,\ldots,j_{\ell+1}}(\boldsymbol{E}_{\widehat{\boldsymbol{A}}_1},\ldots,\boldsymbol{E}_{\widehat{\boldsymbol{A}}_\ell},\boldsymbol{E}_{\widehat{\boldsymbol{B}}})\|_2\right]$. We use that for matrices $\boldsymbol{A},\boldsymbol{B}$ of appropriate dimensions, $\|\boldsymbol{AB}\|_2 \leq \|\boldsymbol{A}\|_2\|\boldsymbol{B}\|_2$ and that $\|\boldsymbol{A}\otimes\boldsymbol{B}\|_2 = \|\boldsymbol{A}\|_2\|\boldsymbol{B}\|_2$. Note also that $\|\boldsymbol{S}_j\|_2 = 1$ for $j = 1,\ldots,\ell$. Recall that $\widetilde{\boldsymbol{x}}_0,\boldsymbol{u}_0,\ldots,\boldsymbol{u}_{k-1}$, $\widetilde{\boldsymbol{A}}_1,\ldots,\widetilde{\boldsymbol{A}}_\ell$, $\widetilde{\boldsymbol{B}}$, $\boldsymbol{S}_1,\ldots,\boldsymbol{S}_\ell$ are deterministic quantities with finite norm. We thus have, for some finite constant $\bar{C}(j_1,\ldots,j_{\ell+1}) > 0$, the bound

(3.24)

$$\begin{aligned}
&\mathbb{E}\left[\|\zeta_{j_1,\ldots,j_{\ell+1}}(\boldsymbol{E}_{\widehat{\boldsymbol{A}}_1},\ldots,\boldsymbol{E}_{\widehat{\boldsymbol{A}}_\ell},\boldsymbol{E}_{\widehat{\boldsymbol{B}}})\|_2\right]\\
&\leq \bar{C}(j_1,\ldots,j_{\ell+1})\mathbb{E}[\|\boldsymbol{E}_{\widehat{\boldsymbol{A}}_1}\|_2^{j_1}\cdots\|\boldsymbol{E}_{\widehat{\boldsymbol{A}}_\ell}\|_2^{j_\ell}\|\boldsymbol{E}_{\widehat{\boldsymbol{B}}}\|_2^{j_{\ell+1}}]\\
&\leq \bar{C}(j_1,\ldots,j_{\ell+1})\sigma^{j_1+\cdots+j_{\ell+1}}\|\Sigma_{\widehat{\boldsymbol{A}}_1}^{1/2}\|_2^{j_1}\cdots\|\Sigma_{\widehat{\boldsymbol{A}}_\ell}^{1/2}\|_2^{j_\ell}\|\Sigma_{\widehat{\boldsymbol{B}}}^{1/2}\|_2^{j_{\ell+1}}\mathbb{E}[\|\boldsymbol{G}_{\widehat{\boldsymbol{A}}_1}\|_2^{j_1}\cdots\|\boldsymbol{G}_{\widehat{\boldsymbol{A}}_\ell}\|_2^{j_\ell}\|\boldsymbol{G}_{\widehat{\boldsymbol{B}}}\|_2^{j_{\ell+1}}]\\
&\leq \bar{C}(j_1,\ldots,j_{\ell+1})\left(\frac{\sigma}{s_{\min}(\boldsymbol{D})}\right)^{j_1+\cdots+j_{\ell+1}}\mathbb{E}[\|\boldsymbol{G}_{\widehat{\boldsymbol{A}}_1}\|_2^{j_1}\cdots\|\boldsymbol{G}_{\widehat{\boldsymbol{A}}_\ell}\|_2^{j_\ell}\|\boldsymbol{G}_{\widehat{\boldsymbol{B}}}\|_2^{j_{\ell+1}}]
\end{aligned}$$

where we utilized (3.8), (3.9).

Recursively applying the Cauchy-Schwarz inequality and invoking concentration inequalities on $\|\boldsymbol{G}_{\widehat{\boldsymbol{B}}}\|_2, \|\boldsymbol{G}_{\widehat{\boldsymbol{A}}_s}\|_2, s = 1,\ldots,\ell$ shows that $\mathbb{E}[\|\boldsymbol{G}_{\widehat{\boldsymbol{A}}_1}\|_2^{j_1}\cdots\|\boldsymbol{G}_{\widehat{\boldsymbol{A}}_\ell}\|_2^{j_\ell}\|\boldsymbol{G}_{\widehat{\boldsymbol{B}}}\|_2^{j_{\ell+1}}]$ is finite. To illustrate this, consider

$$\begin{aligned}
&\left|\mathbb{E}[\|\boldsymbol{G}_{\widehat{\boldsymbol{A}}_1}\|_2^{j_1}\cdots\|\boldsymbol{G}_{\widehat{\boldsymbol{A}}_\ell}\|_2^{j_\ell}\|\boldsymbol{G}_{\widehat{\boldsymbol{B}}}\|_2^{j_{\ell+1}}]\right|\\
&\qquad\leq \sqrt{\mathbb{E}[\|\boldsymbol{G}_{\widehat{\boldsymbol{A}}_1}\|_2^{2j_1}]\mathbb{E}[\|\boldsymbol{G}_{\widehat{\boldsymbol{A}}_2}\|_2^{2j_2}\cdots\|\boldsymbol{G}_{\widehat{\boldsymbol{A}}_\ell}\|_2^{2j_\ell}\|\boldsymbol{G}_{\widehat{\boldsymbol{B}}}\|_2^{2j_{\ell+1}}]}.
\end{aligned} \tag{3.25}$$

By invoking Lemma 3.4, we can obtain a bound for $\mathbb{E}[\|\boldsymbol{G}_{\widehat{\boldsymbol{A}}_1}\|_2^{2j_1}]$. The Cauchy-Schwarz inequality is then applied to $\mathbb{E}[\|\boldsymbol{G}_{\widehat{\boldsymbol{A}}_2}\|_2^{2j_2}\cdots\|\boldsymbol{G}_{\widehat{\boldsymbol{A}}_\ell}\|_2^{2j_\ell}\|\boldsymbol{G}_{\widehat{\boldsymbol{B}}}\|_2^{2j_{\ell+1}}]$ after which concentration inequalities are invoked to bound $\mathbb{E}[\|\boldsymbol{G}_{\widehat{\boldsymbol{A}}_2}\|_2^{4j_2}]$, which is repeated ℓ times until the expected value of products is decomposed into a product of expected values. The proposition then follows from (3.24) by summing over the indices for which $j_1+\cdots+j_{\ell+1} = l$ with $l = 2,\ldots,Q_k$. ∎

We now derive a bound for the MSE of the predicted states, which shows that for polynomially nonlinear systems, the MSE in the asymptotic regime $\sigma/s_{\min}(\boldsymbol{D}) \to 0$ is dominated by $(\sigma/s_{\min}(\boldsymbol{D}))^2$.

PROPOSITION 3.9. *Let $\widehat{\boldsymbol{x}}_0 = \widetilde{\boldsymbol{x}}_0$. Suppose that the conditions of Proposition 3.2 hold. If the high-dimensional system (3.1) from which data are sampled and the learned low-dimensional model are polynomially nonlinear, then*

$$\mathbb{E}[\|\widehat{\boldsymbol{x}}_k - \widetilde{\boldsymbol{x}}_k\|_2^2] \leq \sum_{l=2}^{2Q_k}\widehat{C}_l\left(\frac{\sigma}{s_{min}(\boldsymbol{D})}\right)^l, \qquad 1 \leq k \in \mathbb{N},$$

for some constants $0 < \widehat{C}_l < \infty, l = 2,\ldots,2Q_k$, which are not functions of σ and $\boldsymbol{D}$.

Proof. From (3.21) and (3.22), we obtain

$$\widehat{\boldsymbol{x}}_k - \widetilde{\boldsymbol{x}}_k = \sum_{l=1}^{Q_k}\sum_{\substack{j_1,\ldots,j_{\ell+1}\in\mathbb{N}\\ j_1+\cdots+j_{\ell+1}=l}}\zeta_{j_1,\ldots,j_{\ell+1}}(\boldsymbol{E}_{\widehat{\boldsymbol{A}}_1},\ldots,\boldsymbol{E}_{\widehat{\boldsymbol{A}}_\ell},\boldsymbol{E}_{\widehat{\boldsymbol{B}}}).$$

Thus, for some finite constant $\widehat{C}(j_1, \ldots, j_{\ell+1}) > 0$,

$$
\begin{aligned}
&\|\widehat{\boldsymbol{x}}_k - \widetilde{\boldsymbol{x}}_k\|_2 \\
&\leq \sum_{l=1}^{Q_k} \sum_{\substack{j_1,\ldots,j_{\ell+1} \in \mathbb{N} \\ j_1+\cdots+j_{\ell+1}=l}} \|\zeta_{j_1,\ldots,j_{\ell+1}}(\boldsymbol{E}_{\widehat{\boldsymbol{A}}_1}, \ldots, \boldsymbol{E}_{\widehat{\boldsymbol{A}}_\ell}, \boldsymbol{E}_{\widehat{\boldsymbol{B}}})\|_2 \\
&\leq \sum_{l=1}^{Q_k} \sum_{\substack{j_1,\ldots,j_{\ell+1} \in \mathbb{N} \\ j_1+\cdots+j_{\ell+1}=l}} \widehat{C}(j_1, \ldots, j_{\ell+1}) \|\boldsymbol{E}_{\widehat{\boldsymbol{A}}_1}\|_2^{j_1} \cdots \|\boldsymbol{E}_{\widehat{\boldsymbol{A}}_\ell}\|_2^{j_\ell} \|\boldsymbol{E}_{\widehat{\boldsymbol{B}}}\|_2^{j_{\ell+1}} \\
&\leq \sum_{l=1}^{Q_k} \sum_{\substack{j_1,\ldots,j_{\ell+1} \in \mathbb{N} \\ j_1+\cdots+j_{\ell+1}=l}} \widehat{C}(j_1, \ldots, j_{\ell+1}) \left(\frac{\sigma}{s_{\min}(\boldsymbol{D})}\right)^{j_1+\cdots+j_{\ell+1}} \|\boldsymbol{G}_{\widehat{\boldsymbol{A}}_1}\|_2^{j_1} \cdots \|\boldsymbol{G}_{\widehat{\boldsymbol{A}}_\ell}\|_2^{j_\ell} \|\boldsymbol{G}_{\widehat{\boldsymbol{B}}}\|_2^{j_{\ell+1}}
\end{aligned}
$$

following calculations in (3.24) where $\boldsymbol{G}_{\widehat{\boldsymbol{A}}_1}, \ldots, \boldsymbol{G}_{\widehat{\boldsymbol{A}}_\ell}, \boldsymbol{G}_{\widehat{\boldsymbol{B}}}$ are the same random matrices defined in the proof of Proposition 3.8. Note that in the above inequality, the powers of the noise-to-signal ratio range from $j_1+\cdots+j_{\ell+1} = 1$ to $j_1+\cdots+j_{\ell+1} = Q_k$, whereas in the proof of Proposition 3.8 the inequality (3.23) starts at $j_1+\cdots+j_{\ell+1} = 2$. The conclusion now follows by squaring both sides of the inequality, applying expectation, and performing calculations similar to (3.25) to show that the resulting constants are finite. ∎

4. Active operator inference for selecting training data. This section proposes active operator inference, which selects from a dictionary at which initial condition and inputs to sample the high-dimensional system for generating data with low noise-to-signal ratios. The proposed active operator inference is motivated by the bounds derived in Section 3.2, which show that the noise-to-signal ratio $\sigma/s_{\min}(\boldsymbol{D})$ controls the MSE of the learned operators as well as the bias and the MSE of the state dynamics.

Section 4.1 formalizes the dictionary whose elements are candidates for sampling the high-dimensional system at. The proposed selection of elements of the dictionary is described in Section 4.2 and builds on ideas from selecting points [40] in empirical interpolation [5, 14]. The computational procedure for active operator inference is presented in Section 4.3, which summarizes the proposed workflow for learning low-dimensional models from noisy data.

4.1. Dictionary of candidate states and inputs. Consider a dictionary $\mathcal{D} \in \mathbb{R}^{L \times M}$ of candidate states and inputs given by

$$
\mathcal{D} = [\breve{\boldsymbol{X}}_L^T, (\breve{\boldsymbol{X}}_L^2)^T, \ldots, (\breve{\boldsymbol{X}}_L^\ell)^T, \boldsymbol{U}_L^T],
$$

where $\breve{\boldsymbol{X}}_L \in \mathbb{R}^{n \times L}, \boldsymbol{U}_L \in \mathbb{R}^{p \times L}$ are defined identically as $\breve{\boldsymbol{X}}, \boldsymbol{U}$ in Section 3.1 but with L states and inputs, i.e. $\breve{\boldsymbol{X}}_L = [\breve{\boldsymbol{x}}_1, \ldots, \breve{\boldsymbol{x}}_L]$ and $\boldsymbol{U}_L = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_L]$. Let $\boldsymbol{P}_K \in \{0,1\}^{L \times K}$ be a selection operator that selects $K \leq L$ rows of $\mathcal{D}$ via $\boldsymbol{P}_K^T \mathcal{D} = \boldsymbol{D}$ so that $\boldsymbol{D}$ can serve as a data matrix in the sense of (3.4).

Using all rows of $\mathcal{D}$ to form the data matrix $\boldsymbol{D}$ leads to the minimal noise-to-signal ratio $\sigma/s_{\min}(\boldsymbol{D})$ over all possible selections, which follows by considering the addition of a row as a rank-one update to a matrix together with Weyl's theorem; see [40, Section 5.1] for a proof. Note that $\boldsymbol{D}$ has to have full rank; see Proposition 3.2. However, using all rows of $\mathcal{D}$ to construct the data matrix $\boldsymbol{D}$ in the least squares

problem (3.4) is computationally expensive as the high-dimensional system has to be queried for each of the L initial conditions and inputs. We therefore propose in the following subsection an approach to subselect rows of $\mathcal{D}$ with the aim of having a small noise-to-signal ratio with only a few rows of $\mathcal{D}$.

4.2. A design of experiments approach via oversampled empirical interpolation. Propositions 3.5, 3.8, and 3.9 demonstrate that a low noise-to-signal ratio $\sigma/s_{\min}(\boldsymbol{D})$ is desirable. Since the standard deviation σ is fixed, we propose a design of experiments strategy that forms a data matrix $\boldsymbol{D}$ by selecting rows of $\mathcal{D}$ so that $s_{\min}(\boldsymbol{D})$ is large. To find a selection of rows, we follow the procedure proposed in [40] that pursues an equivalent objective for selecting points for empirical interpolation and gappy proper orthogonal decomposition [5, 14, 17]; see [4, 15, 36, 54] for other design of experiment approaches based on similar linear-algebra concepts. Set $K \geq M$. The method introduced in [40] constructs a selection matrix $\boldsymbol{P}_K$ which selects K rows of $\mathcal{D}$ with the objective of maximizing $s_{\min}(\boldsymbol{P}_K^T\mathcal{D})$. First, the selection matrix $\boldsymbol{P}_M \in \mathbb{R}^{M\times M}$ is initialized with the approach introduced in [17]. Then, new rows of $\mathcal{D}$ are selected in a greedy fashion. To describe the greedy update, suppose we have the selection matrix $\boldsymbol{P}_m$, which selects m rows of $\mathcal{D}$, with $M \leq m < K$. Let the SVD of $\boldsymbol{P}_m^T\mathcal{D}$ be $\boldsymbol{P}_m^T\mathcal{D} = \boldsymbol{\Phi}_m\boldsymbol{\Sigma}_m\boldsymbol{\Psi}_m^T$ where $\boldsymbol{\Phi}_m \in \mathbb{R}^{m\times M}$ is the matrix of left-singular vectors, $\boldsymbol{\Sigma}_m \in \mathbb{R}^{M\times M}$ is the diagonal matrix of singular values $s_1^{(m)}, \ldots, s_M^{(m)}$ in descending order, and $\boldsymbol{\Psi}_m \in \mathbb{R}^{M\times M}$ is the matrix of right-singular vectors. Define the gap $g = (s_{M-1}^{(m)})^2 - (s_M^{(m)})^2$ and set $\bar{\boldsymbol{d}}_+ = \boldsymbol{\Psi}_m^T\boldsymbol{d}_+^T$, where $\boldsymbol{d}_+ \in \mathbb{R}^{1\times M}$ is a candidate row of $\mathcal{D}$ that has not been selected by $\boldsymbol{P}_m$. Further, let $\boldsymbol{e} \in \mathbb{R}^M$ be the canonical basis vector with all entries 0 except for the last component that is set to 1. It is shown in [28], see also the discussion in [40], that

(4.1)
$$s_{\min}(\boldsymbol{P}_{m+1}^T\mathcal{D})^2 - s_{\min}(\boldsymbol{P}_m^T\mathcal{D})^2 \geq \frac{1}{2}\left(g + \|\bar{\boldsymbol{d}}_+\|_2^2 - \sqrt{(g+\|\bar{\boldsymbol{d}}_+\|_2^2)^2 - 4g(\boldsymbol{e}^T\bar{\boldsymbol{d}}_+)^2}\right)$$

which suggests that the new row $\boldsymbol{d}_+$ should be selected that maximizes the lower bound (4.1). This greedy step is then repeated until the desired number of rows K is reached.

Based on the just described greedy scheme, we use a modified greedy update rule, which was proposed in an earlier preprint version of [40]: we choose the new row $\boldsymbol{d}_+$ that maximizes

(4.2)
$$(\boldsymbol{e}^T\bar{\boldsymbol{d}}_+)^2\,,$$

which is obtained by simplifying the lower bound (4.1).

Choosing the new row $\boldsymbol{d}_+$ by maximizing the lower bound (4.1) was tested in [40] for the case where the columns of $\mathcal{D}$ are orthonormal. Since this condition does not necessarily hold in our setting, the lower bound in (4.1) for the greedy update can lead to cancellation errors especially when $4g(\boldsymbol{e}^T\bar{\boldsymbol{d}}_+)^2$ is small, which has been first observed in [65].

We note that QDEIM [17] was also applied in [67] to subselect rows from the data matrix to improve its condition number for operator inference. The approach above is a generalization since it allows one to subselect $K \geq M$ rows instead of being limited to M rows, the number of columns in $\mathcal{D}$, as in [67].

4.3. Active operator inference. The proposed active operator inference approach to learn low-dimensional models from noisy data is summarized in Algorithm 4.1. The inputs of the algorithm are the dictionary $\mathcal{D}$ and the number of times K

Algorithm 4.1 Active operator inference based on QDEIM [17] and oversampling [40]

1: **procedure** AOpInf($\mathcal{D}, K$)
2: Initialize $\boldsymbol{P}_M$ with QDEIM [17]
3: **for** $m = M, \ldots, K-1$ **do** ▷ Follow [40] with criterion (4.2)
4: Compute the SVD of $\boldsymbol{P}_m^T\mathcal{D} = \boldsymbol{\Phi}_m\boldsymbol{\Sigma}_m\boldsymbol{\Psi}_m^T$
5: Find the row $\boldsymbol{d}_+$ of $\mathcal{D}$ not in $\boldsymbol{P}_m^T\mathcal{D}$ such that $\bar{\boldsymbol{d}}_+ = \boldsymbol{\Psi}_m^T\boldsymbol{d}_+^T$ maximizes (4.2)
6: Update $\boldsymbol{P}_m$ to $\boldsymbol{P}_{m+1}$
7: Construct $\boldsymbol{D}$ via $\boldsymbol{P}_K^T\mathcal{D} = \boldsymbol{D}$
8: Perform re-projection as in (3.2) to generate $\breve{\boldsymbol{Z}}$
9: Solve the least-squares problem (3.4) to obtain $\widehat{\boldsymbol{A}}_1, \ldots, \widehat{\boldsymbol{A}}_\ell, \widehat{\boldsymbol{B}}$
return $\widehat{\boldsymbol{A}}_1, \ldots, \widehat{\boldsymbol{A}}_\ell, \widehat{\boldsymbol{B}}$

to query the high-dimensional system, i.e., the number of rows of the data matrix $\boldsymbol{D}$. Line 2 of Algorithm 4.1 initializes the sampling matrix via QDEIM [17] by computing the QR decomposition of $\mathcal{D}^T$ with pivoting. For $m \in \{M, M+1, \ldots, K-1\}$, the SVD of $\boldsymbol{P}_m^T\mathcal{D}$ is obtained in line 4 and the candidate row $\boldsymbol{d}_+$ of $\mathcal{D}$ that maximizes (4.2) is selected in in lines 5–6 to update $\boldsymbol{P}_m$ to $\boldsymbol{P}_{m+1}$. In lines 7–9, re-projection (3.2) is performed using the projected states and the inputs in the data matrix $\boldsymbol{D} = \boldsymbol{P}_K^T\mathcal{D}$ to obtain the re-projected trajectory $\breve{\boldsymbol{Z}}$. The least-squares problem (3.4) is then solved to learn the low-dimensional operators.

Note that the high-dimensional system gets queried only at the subselected rows of the dictionary and thus, at a typically much lower number of initial conditions than the number of rows in the dictionary. This means that the proposed active operator inference approach can be helpful if querying the high-dimensional system is expensive, which makes it intractable to query the high-dimensional system at all initial conditions given in the dictionary $\mathcal{D}$ but tractable to query the system at the few subselected initial conditions from $\mathcal{D}$ that are collected as columns in $\bar{\boldsymbol{X}}$.

5. Numerical experiments. We now numerically demonstrate that the proposed active operator inference leads to predicted states with orders of magnitude lower biases and MSEs than an uninformed equidistant-in-time selection of data samples. Additionally, we demonstrate that the bias and MSE of predicted states decay with the noise-to-signal ratio in agreement with the analysis developed in Section 3.2. Numerical results for a linear state dynamics are shown in Section 5.1 and for quadratic dynamics in Section 5.2. In all experiments within one example, we use the same basis matrix $\boldsymbol{V}$, which ensures consistent comparisons among different noise-to-signal ratios.

5.1. Heat transfer problem for cooling of steel profiles. The model and problem setup are described in Section 5.1.1 and the numerical results are presented in Section 5.1.2.

5.1.1. Model of cooling steel profiles. We describe a mathematical model for the cooling process of steel rail profiles in a rolling mill following [10]. Set $\Omega \subset \mathbb{R}^2$ as the spatial domain and denote by $x(\boldsymbol{\eta}, t)$ the temperature at the spatial point $\boldsymbol{\eta} \in \Omega$

and time $t > 0$. The heat transfer model is

$$\begin{aligned} \frac{\partial x(\boldsymbol{\eta},t)}{\partial t} &= \frac{\lambda}{c\rho}\Delta x(\boldsymbol{\eta},t), \quad (\boldsymbol{\eta},t) \in \Omega \times [0,T], \\ \nabla x(\boldsymbol{\eta},t) \cdot \mathbf{n} &= \begin{cases} \frac{\kappa}{\lambda}(u_j(t) - x(\boldsymbol{\eta},t)) & \text{for} \quad \boldsymbol{\eta} \in \Gamma_j, j = 1,\ldots,7, \\ 0 & \text{for} \quad \boldsymbol{\eta} \in \Gamma_0, \end{cases} \\ x(\boldsymbol{\eta},0) &= 500, \end{aligned} \tag{5.1}$$

where λ is the heat conductivity, c the specific heat capacity, ρ the profile density, κ the heat transfer coefficient, $\Gamma_j, j = 0,\ldots,7$ are segments of the domain boundary $\partial\Omega$ such that $\partial\Omega = \cup_{j=0}^{7}\Gamma_j$ and $u_j(t), j = 1,\ldots,7$ is the external temperature applied to each boundary segment. The domain is visualized in Figure 1 of [10]. The values of the constants are chosen as $\lambda = 26.4, c = 7620, \rho = 654, \kappa = 69.696$.

Equation (5.1) is spatially discretized using the finite element method with linear triangular elements, temporally discretized with implicit Euler with step size $\delta t = 0.01$, and the noise term $\boldsymbol{\xi}_k$ is added to the fully discrete system to yield the high-dimensional system (3.1) with right-hand side function (2.3) with $\ell = 1$, where $\boldsymbol{x}_k \in \mathbb{R}^N, N = 1357$ and $\boldsymbol{u}_k \in \mathbb{R}^p, p = 7$. We utilized the Python[1] code based on the FEniCS Project to generate the computational mesh and the system matrices; see also [10]. More specifically, the semi-discrete system is

$$\boldsymbol{M}\dot{\boldsymbol{x}}(t) = \boldsymbol{G}\boldsymbol{x}(t) + \boldsymbol{H}\boldsymbol{u}(t)$$

where $\boldsymbol{x}(t) \in \mathbb{R}^N, \boldsymbol{u}(t) \in \mathbb{R}^p, \boldsymbol{M}, \boldsymbol{G} \in \mathbb{R}^{N\times N}$, and $\boldsymbol{H} \in \mathbb{R}^{N\times p}$. Using implicit Euler, the fully discrete high-dimensional system has the form

$$\boldsymbol{x}_{k+1} = \boldsymbol{A}\boldsymbol{x}_k + \boldsymbol{B}\boldsymbol{u}_k, \quad k = 0,\ldots,K-1,$$

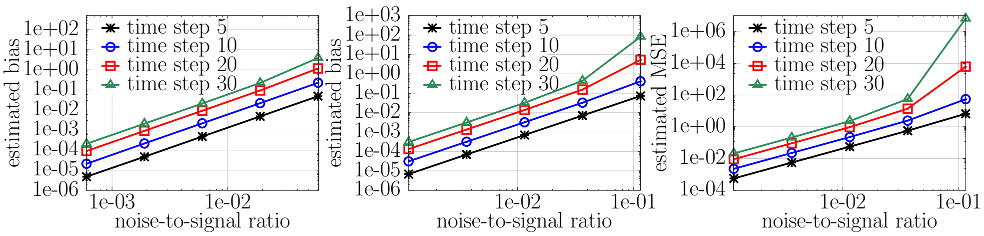
where $\boldsymbol{A} = (\boldsymbol{M} - \delta t\boldsymbol{G})^{-1}\boldsymbol{M}$ and $\boldsymbol{B} = \delta t(\boldsymbol{M} - \delta t\boldsymbol{G})^{-1}\boldsymbol{H}$. The basis matrix $\boldsymbol{V}$ is computed from snapshots $\boldsymbol{x}_k^{\text{basis}}, k = 0,\ldots,L, L = 10000$, of the high-dimensional system driven by the control input $\boldsymbol{u}_k^{\text{basis}}$ whose i-th component, $i = 1,\ldots,7$ is given by $500(1 - \tanh(k\delta t/i^2)) + 250\gamma_{i,k}$. Here, $\gamma_{i,0} = 0$ while $\gamma_{i,k}$ for $k > 0$ is a realization of a uniform random variable on $[0,1]$. The realization of the control trajectory is then fixed and used as the control input $\boldsymbol{u}_k^{\text{basis}}, k = 0,\ldots,L-1$. The projected states $\boldsymbol{V}^T\boldsymbol{x}_k^{\text{basis}}$ and the input $\boldsymbol{u}_k^{\text{basis}}$ for $k = 0,\ldots,L-1$ constitute the 10000 rows of the dictionary $\mathcal{D} \in \mathbb{R}^{L\times(n+p)}$.

Data is obtained by querying the system with noise, which is

$$\boldsymbol{x}_{k+1} = \boldsymbol{A}\boldsymbol{x}_k + \boldsymbol{B}\boldsymbol{u}_k + \boldsymbol{\xi}_k, \quad k = 0,\ldots,K-1, \tag{5.2}$$

with $\boldsymbol{\xi}_k$ being a zero-mean Gaussian vector with diagonal covariance and standard deviation σ that we vary and therefore specify later. The noisy system (5.2) is queried at initial conditions given by the columns of $\bar{\boldsymbol{X}}$, which are selected from the dictionary $\mathcal{D}$. Note that (5.2) has the same form as system (3.1) and that querying (5.2) introduces noise in the observed trajectories. We generate data from the time-discrete system rather than from the underlying time-continuous system to avoid mixing errors from other sources and so to be able to focus on the error induced by noise. We refer to [41], where operator inference has been applied to data from time-continuous systems.

[1] https://gitlab.mpi-magdeburg.mpg.de/models/fenicsrail/-/tree/master/



(a) decay of bias, $n = 7$ (b) decay of bias, $n = 10$ (c) decay of MSE, $n = 10$

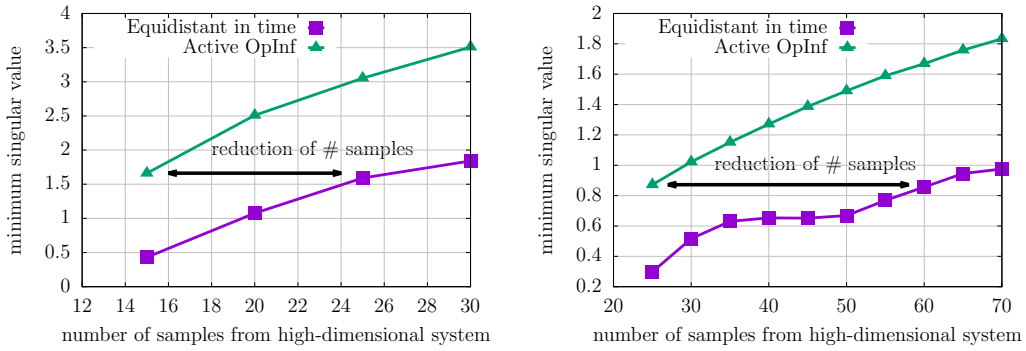
Fig. 2: Cooling of steel profiles (Section 5.1). The estimated bias and MSE decay by 2 orders of magnitude per 1 order of magnitude decrease in the noise-to-signal ratio in the asymptotic regime. The results are in agreement with Propositions 3.5 and 3.9.

In the simulations below, we consider 5 equally spaced values in the logarithm scale for the standard deviation σ of the noise between 1×10^{-3} and 1×10^{-1}. The test control input $\boldsymbol{u}_k^{\text{test}}$ at time step $k \in \mathbb{N}$ has components given by $500(1 - \tanh(k\delta t/i^2))$ for $i = 1, \ldots, 7$.

5.1.2. Results. We learn a low-dimensional model of dimension $n = 7$ and $n = 10$ from noisy data. For $n = 7$, active operator inference is applied to select 15 rows from $\mathcal{D}$, which leads to a data matrix $\boldsymbol{D}$ with $s_{\min}(\boldsymbol{D}) \approx 1.661$. For $n = 10$, 25 rows are selected resulting in $s_{\min}(\boldsymbol{D}) \approx 0.8713$. Denote by $\widehat{\boldsymbol{x}}_k^{\text{test}}$ the predicted state at time step k of the low-dimensional model with inferred operators $\widehat{\boldsymbol{A}}, \widehat{\boldsymbol{B}}$ corresponding to the test input $\boldsymbol{u}_k^{\text{test}}$. Likewise, let $\widetilde{\boldsymbol{x}}_k^{\text{test}}$ be the low-dimensional state from intrusive model reduction for the same input. Recall that $\widetilde{\boldsymbol{x}}_k^{\text{test}}$ is deterministic while $\widehat{\boldsymbol{x}}_k^{\text{test}}$ is a random vector.

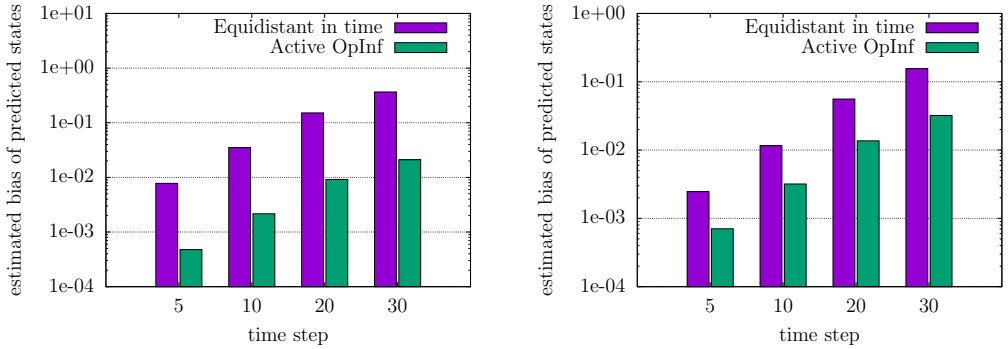
Figure 2 shows a Monte Carlo estimate of the bias $\|\mathbb{E}[\widehat{\boldsymbol{x}}_k^{\text{test}} - \widetilde{\boldsymbol{x}}_k^{\text{test}}]\|_2$ and the MSE $\mathbb{E}[\|\widehat{\boldsymbol{x}}_k^{\text{test}} - \widetilde{\boldsymbol{x}}_k^{\text{test}}\|_2^2]$ as a function of the noise-to-signal ratio $\sigma/s_{\min}(\boldsymbol{D})$ for various time steps k. We use 7.5×10^7 samples to approximate the expected value with Monte Carlo. The plots illustrate that in the asymptotic regime, when $\sigma/s_{\min}(\boldsymbol{D}) \to 0$, an order decrease in the noise-to-signal ratio leads to a decrease of two orders of magnitude in the approximation of the bias and the MSE, which agrees with Propositions 3.5 and 3.9. Notice that for $n = 10$, the behavior of the bias and the MSE for the largest noise value σ is already dominated by constants, rather than the noise-to-signal ratio, which explains the quicker error increase.

We now compare active operator inference, which carefully selects rows of the data matrix $\boldsymbol{D}$ to keep the noise-to-signal ratio low, with a traditional sample selection that queries the high-dimensional system equidistantly in time, i.e., picks columns corresponding to equidistant times from the dictionary $\mathcal{D}$. The high-dimensional system is sampled every $667\delta t$ time units for $n = 7$ and $400\delta t$ time units for $n = 10$. Figure 3 compares the minimum singular value of the data matrix for both approaches over the number of queries to the high-dimensional system. Equidistant sampling requires up to 3 times as many queries to the high-dimensional system to achieve the same noise-to-signal ratio as active operator inference in our experiment. The selection of active operator inference is initialized with QDEIM; see Section 4.1. The QDEIM alone can select only as many rows as there are columns. In case of dimension $n = 7$, there are $M = 14$ columns, in which case QDEIM selects $K = 14$ rows and achieves a minimal singular value of approximately 1.1497. In contrast, with active



(a) dimension $n = 7$ (b) dimension $n = 10$

Fig. 3: Cooling of steel profiles (Section 5.1). To achieve the same noise-to-signal ratio, active operator inference requires almost 3 times fewer queries to the high-dimensional system than a traditional selection of equidistant-in-time samples.

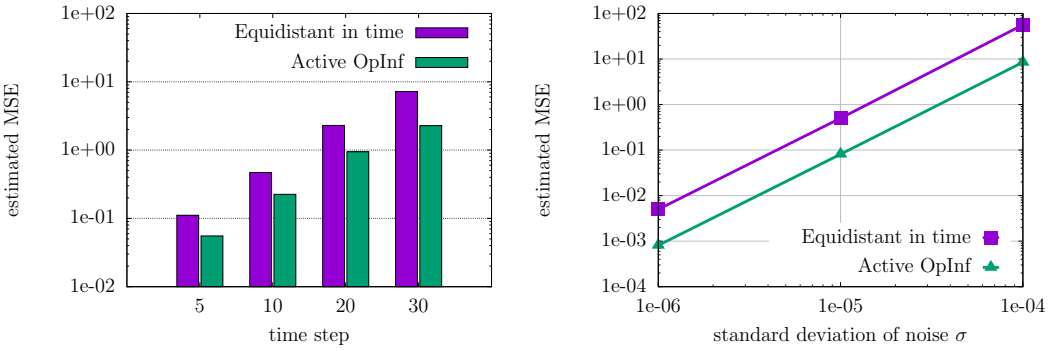


(a) dimension $n = 7$ (b) dimension $n = 10$

Fig. 4: Cooling of steel profiles (Section 5.1). Active operator inference yields predictions which have a lower bias compared to the predictions delivered by sampling equidistantly in time in the dictionary. The reduction in the estimated bias achieved by active operator inference is up to 1.5 orders in magnitude.

operator inference, we can select more than M rows and so increase the minimal singular value. By just selecting $K = M + 1 = 15$ rows, active operator inference achieves roughly 1.661 for the smallest singular value, which can be further increased by increasing K as shown in Figure 3. Similarly, for dimension $n = 10$, QDEIM alone selects $M = 17$ rows and achieves a smallest singular value of about 0.3423, whereas the proposed active operator inference approach increases the smallest singular value to about 0.8713 by selecting $K = 25$ rows.

The estimate of the bias $\|\mathbb{E}[\widehat{\boldsymbol{x}}_k^{\text{test}} - \widetilde{\boldsymbol{x}}_k^{\text{test}}]\|_2$ for $\sigma = 1 \times 10^{-2}$ for the equidistant and active operator inference approach is presented in Figure 4. The results show



(a) equidistant vs Active OpInf, $n = 10$

(b) time step 10000

Fig. 5: Cooling of steel profiles (Section 5.1). The predictions obtained with active operator inference have a lower MSE than those obtained from equidistant-in-time samples.

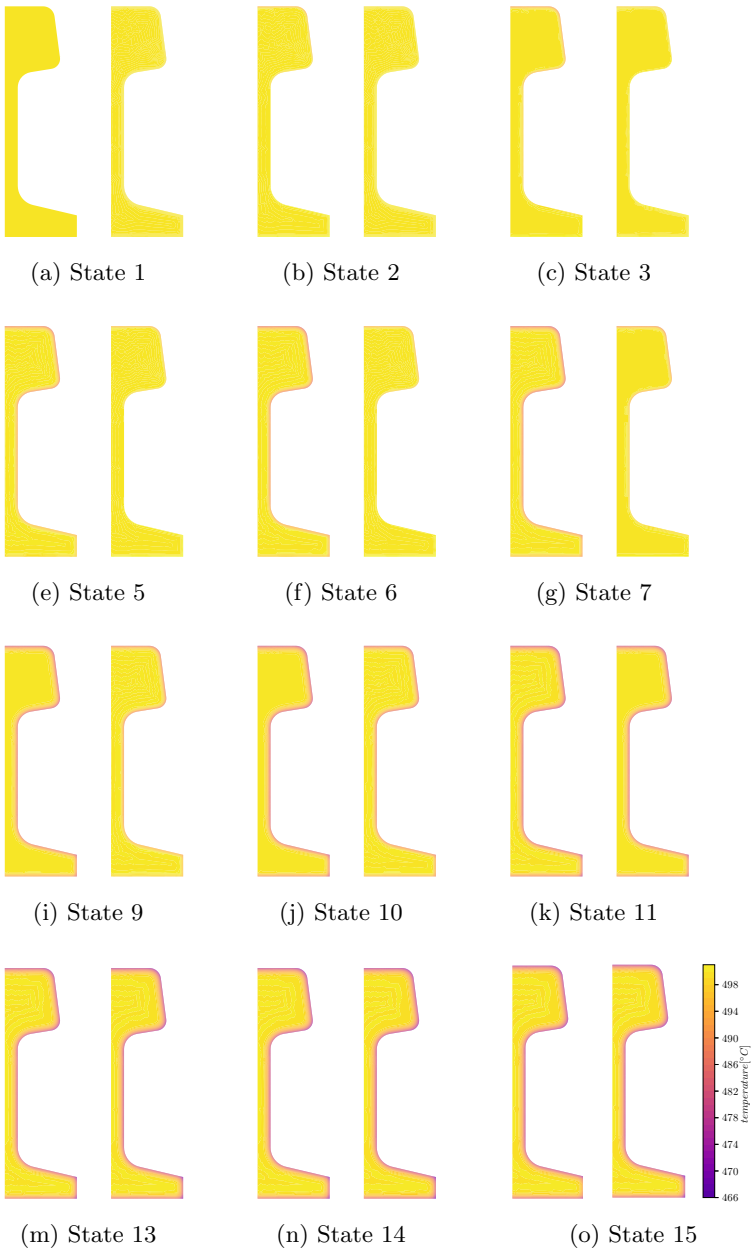
that active operator inference yields a reduction in the estimated bias of up to 1.5 orders of magnitude.

The MSE $\mathbb{E}[\|\widehat{\boldsymbol{x}}_k^{\text{test}} - \widetilde{\boldsymbol{x}}_k^{\text{test}}\|_2^2]$ for equidistant vs. active operator inference is shown in the left panel of Figure 5 for $n = 10$. Lastly, we consider the MSE of the predicted state further in time. The right panel of Figure 5 plots the estimated MSE of the predicted state at 10000 time steps for $n = 10$ using 10 Monte Carlo samples only. An order decay in the noise standard deviation leads to 2 orders decay in the estimated MSE. For fixed σ, the model learned through active operator inference achieves a smaller MSE.

In Figure 6 we visualize the 15 high-dimensional states corresponding to the rows of $\mathcal{D}$ selected according to the design of experiments schemes we compare for $n = 7$. The respective inputs are not shown. By examining the segments of the steel profile boundary with Robin condition, the equidistant scheme tends to select more states with lower temperature at the boundary, many of which correspond to later time steps. In contrast, active operator inference selects more states at the beginning of the cooling process, where there is a stronger variation from one time step to the next.

5.2. Diffusive Lotka-Volterra model for population dynamics of fish species. Section 5.2.1 discusses the model and the problem setup while Section 5.2.2 summarizes the results of the numerical experiments.

5.2.1. Model description. Consider the population dynamics of three species of fish species in the Danube river [30]. At time $t > 0$ and distance η from the mouth of the river, set $x_1(\eta, t), x_2(\eta, t), x_3(\eta, t)$ to be the density of forage fishes, German carp, and predators, respectively. For $\eta \in [0, \pi]$ and $t \in [0, T]$, a diffusive Lotka-Volterra



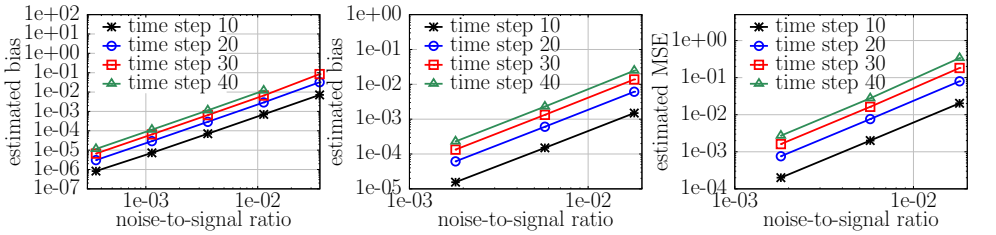
(a) State 1 (b) State 2 (c) State 3 (d) State 4

(e) State 5 (f) State 6 (g) State 7 (h) State 8

(i) State 9 (j) State 10 (k) State 11 (l) State 12

(m) State 13 (n) State 14 (o) State 15

Fig. 6: Cooling of steel profiles (Section 5.1). High-dimensional states selected by sampling equidistant times (left) and by active operator inference (right) from the dictionary for $n = 7$. For equidistant sampling, a majority of the states selected have cooler temperatures at the domain boundary with Robin condition. In contrast, active operator inference selects more states at the beginning of the cooling process, which leads to more accurate models in our experiments.



(a) decay of bias, $n = 12$ (b) decay of bias, $n = 15$ (c) decay of MSE, $n = 15$

Fig. 7: Population dynamics of fish species (Section 5.2). For this quadratic system, an order of magnitude decay in the noise-to-signal ratio causes a two orders of magnitude decay in the estimated bias and MSE, demonstrating the bound of Propositions 3.8 and 3.9.

model that describes the interaction between the species is given by

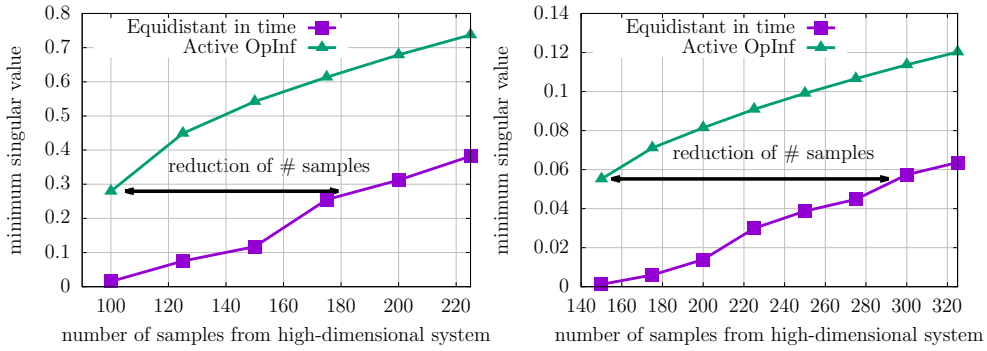
$$\begin{aligned}
\frac{\partial x_1(\eta,t)}{\partial t} &= d_1 \frac{\partial^2 x_1(\eta,t)}{\partial \eta^2} + x_1(a_1 - a_2 x_2 - a_3 x_3) \\
\frac{\partial x_2(\eta,t)}{\partial t} &= d_2 \frac{\partial^2 x_2(\eta,t)}{\partial \eta^2} + x_2(a_4 - a_5 x_3) \\
\frac{\partial x_3(\eta,t)}{\partial t} &= d_3 \frac{\partial^2 x_3(\eta,t)}{\partial \eta^2} + x_3(a_6 x_1 + a_7 x_2 - a_8)
\end{aligned} \tag{5.3}$$

subject to the Neumann boundary condition $\frac{\partial x_i(0,t)}{\partial \eta} = \frac{\partial x_i(\pi,t)}{\partial \eta} = 0$ for $i = 1, 2, 3$. The values of the constants are $a_1 = 1.01, a_2 = 0.93, a_3 = 0.1, a_4 = 0.19, a_5 = 0.2, a_6 = 1, a_7 = 0.05, a_8 = 0.2, d_1 = 0.01, d_2 = 0.03, d_3 = 0.009$.

The differential equation (5.3) is spatially discretized at 100 equidistant points in $[0, \pi]$. To temporally discretize (5.3), we apply the Crank-Nicolson method to the diffusion term using second-order central finite difference scheme [59, Table 3.2.2] and evaluate the nonlinear term explicitly in time with step size $\delta t = 0.01$, resulting in an implicit-explicit scheme. The noise term $\boldsymbol{\xi}_k$ is then added to the fully discrete system which leads to the autonomous system (3.1) in which the right-hand side is given by (2.3) with $\ell = 2$ where $\boldsymbol{x}_k \in \mathbb{R}^N, N = 300$.

Set $x_3^* = a_4/a_5$, $x_2^* = (a_1 a_5 - a_3 a_4)/(a_2 a_5)$, $x_1^* = (a_8 - a_7 x_2^*)/a_6$. Observe that $(x_1, x_2, x_3) = (x_1^*, x_2^*, x_3^*)$ is a spatially homogeneous equilibrium point of (5.3). The basis matrix $\boldsymbol{V}$ is obtained from snapshots $\boldsymbol{x}_k^{\text{basis}}$ of the high-dimensional system initiated at the following 6 conditions $x_{1,i}^{\text{basis}}(\eta, 0) = x_1^* + \gamma_{1i} \sin(6\gamma_{2i}\eta)/10, x_{2,i}^{\text{basis}}(\eta, 0) = x_2^* + \gamma_{3i} \cos(4\gamma_{4i}\eta)/10, x_{3,i}^{\text{basis}}(\eta, 0) = x_3^* + \gamma_{5i} \sin(2\gamma_{6i}\eta)/10, i = 1, \ldots, 6$, where for each i, $\gamma_{1i}, \ldots, \gamma_{6i}$ are realizations of a uniform random variable on $[0, 1]$. The realizations of the initial conditions are then fixed and treated as deterministic quantities. For each initial condition, the high-dimensional system is simulated until $T = 50$ resulting in 30000 elements in $\mathcal{D}$. These initial states represent perturbations around the spatially homogeneous equilibrium. The standard deviations of the noise σ are 5 equidistant values in the logarithm scale between 1×10^{-4} and 1×10^{-2}. For prediction, the initial condition we use is given by $x_1^{\text{test}}(\eta, 0) = x_1^* + \sin(6\eta)/10$, $x_2^{\text{test}}(\eta, 0) = x_2^* + \cos(4\eta)/10$, and $x_3^{\text{test}}(\eta, 0) = x_3^* + \sin(2\eta)/10$.

5.2.2. Results. Active operator inference is applied to select 100 rows for $n = 12$, leading to a data matrix with $s_{\min}(\boldsymbol{D}) \approx 0.2794$. For $n = 15$, 150 rows are selected,



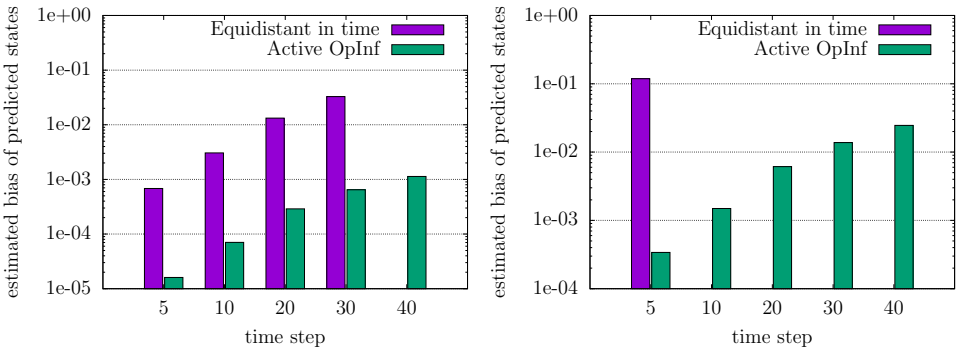
(a) dimension $n = 12$ (b) dimension $n = 15$

Fig. 8: Population dynamics of fish species (Section 5.2). Active operator inference requires up to two times fewer queries to the high-dimensional system for generating data than a traditional equidistant-in-time sampling process.

which results in $s_{\min}(\boldsymbol{D}) \approx 0.0552$. Monte Carlo estimates of the bias and MSE are shown in Figure 7. The number of Monte Carlo samples used is 5×10^7. The plots are consistent with the analysis in Proposition 3.8 and 3.9, particularly for quadratic systems, since we observe that an order decay in the noise-to-signal ratio leads to two orders decay in the estimated bias and MSE. The missing value in Figure 7(a) represents a large bias in $\widehat{\boldsymbol{x}}_k^{\text{test}}$ which we do not plot and is caused by the accumulation of errors in the learned reduced operators over time. It represents a non-asymptotic regime in which the constants in the bias dominate the behavior of the noise-to-signal ratio. In Figure 7(b) and 7(c), results for larger values of σ are not shown in the plot for the same reason.

We now compare active operator inference to a traditional equidistant-in-time sampling from the dictionary. The high-dimensional system is sampled every $300\delta t$ time units for $n = 12$ and $200\delta t$ units for $n = 15$. The minimum singular value of the data matrix resulting from both approaches is compared in Figure 8. In this example, active operator inference reduces the number of times the high-dimensional system is queried by up to roughly a factor of two compared to equidistant sampling to achieve the same minimal singular value of approximately 0.3 ($n = 12$) and 0.06 ($n = 15$). For reference, with QDEIM alone, only M rows can be selected, which means for the minimal singular value that $s_{\min}(\boldsymbol{P}_M^T \mathcal{D}) \approx 0.0943$ for $n = 12$ ($M = 90$) and $s_{\min}(\boldsymbol{P}_M^T \mathcal{D}) \approx 0.0212$ for $n = 15$ ($M = 135$). In contrast, active operator inference with $K = 100$ and $K = 150$ rows, achieves roughly 0.2794 and 0.0552 for dimension $n = 12$ and $n = 15$, respectively.

The estimated bias and the MSE for both approaches at $\sigma = 1 \times 10^{-3}$ is shown in Figure 9 and in the left panel of Figure 10. We also plot the estimated MSE at $T = 50$ using 10 Monte Carlo samples for $n = 15$ in the right panel of Figure 10. Results are not plotted if the corresponding models numerically led to unstable behavior with unbounded errors. Active operator inference provides reasonable numerical predictions in all cases, whereas equidistant sampling quickly leads to models that show unstable behavior. This behavior is amplified for increasing dimension n. Overall, the results indicate that for polynomially nonlinear systems it becomes even more important



(a) dimension $n = 12$ (b) dimension $n = 15$

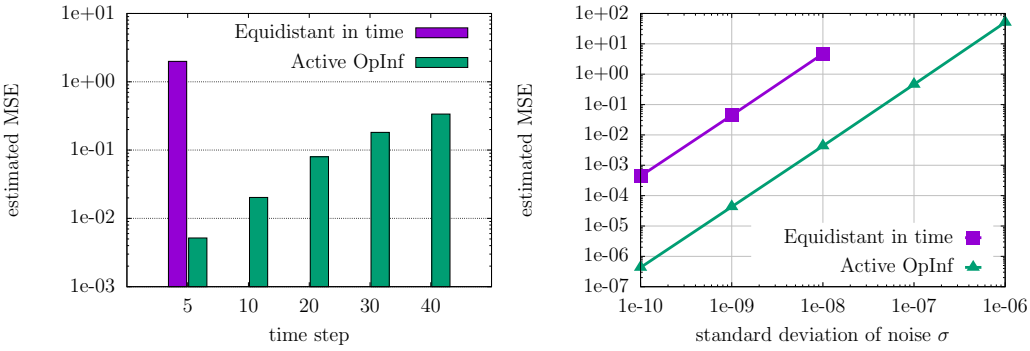
Fig. 9: Population dynamics of fish species (Section 5.2). Selecting the data matrix by sampling equidistant in time quickly leads to numerical instabilities in the learned models while the selection obtained with active operator inference leads to models that show stable and accurate behavior in this example.

than for linear systems to carefully query the high-dimensional system. Indeed, we have empirically observed that active operator inference selects states which exhibit substantial variation across time. For example, Figure 11 plots the high-dimensional states for $x_3(\eta)$ corresponding to the 150 rows in the data matrix selected by equidistant in time sampling and active operator inference for $n = 15$. As can be observed, active operator inference results in states which have a better coverage of the state space.

We now consider the case where the elements of the dictionary $\mathcal{D}$ are generated as a fixed realization obtained by querying (3.1) with $\sigma = 1 \times 10^{-3}$ to obtain $\boldsymbol{x}_k^{\text{basis}}$. Active operator inference and equidistant in time sampling are performed to select the dictionary of initial conditions $\bar{\boldsymbol{X}}$ that are used to construct the data matrix with $s_{\min}(\boldsymbol{D})$ of roughly 0.1107 and 0.0112, respectively. Figure 12 compares the estimated bias and MSE under both approaches for various time points conditioned on $\bar{\boldsymbol{X}}$. The results are consistent with those shown in previous examples in that active operator inference yields stabler models with lower errors.

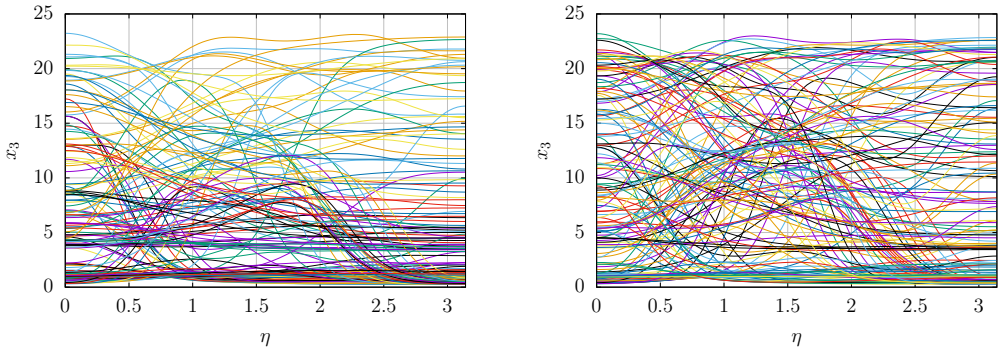
6. Conclusions. In this work, we established probabilistic guarantees on predictions made with low-dimensional models learned from noisy data, which motivated an active data sampling approach to reduce the effect of noise. The key ingredient of the analysis and the numerical approach was building a bridge from data-driven modeling via operator inference and re-projection to classical projection-based model reduction. Thus, the proposed approach can be seen as an example of scientific machine learning that demonstrates the benefits of merging traditional scientific computing concepts such as model reduction with learning methods to effectively leverage data. There are several future research directions in the context of learning reduced models from noisy data such as methods for chaotic systems and systems with non-polynomial nonlinear terms.

Acknowledgements. We are grateful to Jens Saak for providing the code to generate the computational mesh and the system matrices for the heat transfer prob-



(a) equidistant vs Active OpInf, $n = 15$ (b) time step 5000

Fig. 10: Population dynamics of fish species (Section 5.2). Sampling the dictionary at equidistant times results in learned models that become numerically unstable while active operator inference leads to models with orders of magnitude lower MSEs.



(a) Equidistant in time (b) Active operator inference

Fig. 11: Population dynamics of fish species (Section 5.2). Active operator inference tends to select states which exhibit substantial variation across time.

lem on steel profiles. We also thank Jonathan Niles-Weed for directing us to references for deriving upper bounds on moments of the norm of Gaussian random matrices.

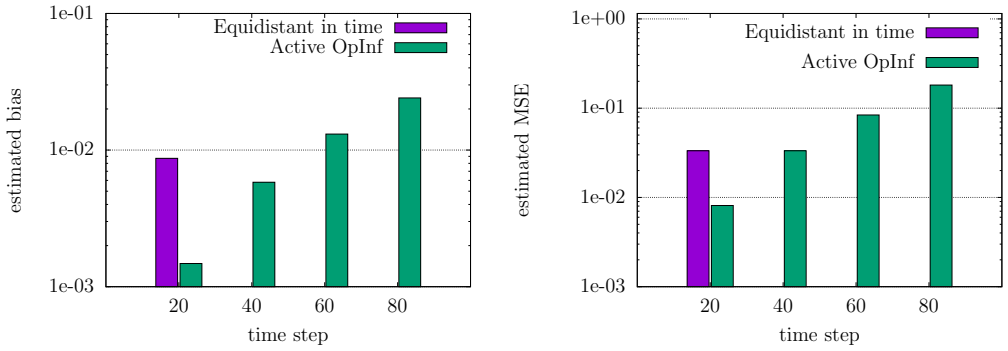
REFERENCES

[1] A. C. Antoulas. *Approximation of Large-Scale Dynamical Systems*. Society for Industrial and Applied Mathematics, 2005.

[2] A. C. Antoulas and B. D. O. Anderson. On the scalar rational interpolation problem. *IMA Journal of Mathematical Control & Information*, 3(2-3):61–88, 1986.

[3] A. C. Antoulas, I. V. Gosea, and A. C. Ionita. Model reduction of bilinear systems in the Loewner framework. *SIAM Journal on Scientific Computing*, 38(5):B889–B916, 2016.

[4] P. Astrid, S. Weiland, K. Willcox, and T. Backx. Missing point estimation in models described by proper orthogonal decomposition. *IEEE Transactions on Automatic Control*,



(a) dimension $n = 15$ (b) dimension $n = 15$

Fig. 12: Population dynamics of fish species (Section 5.2). Equidistant sampling in time results in a numerically unstable model in this example wherein the bias and MSE are conditioned on the dictionary of initial conditions $\bar{\boldsymbol{X}}$ used in re-projection.

53(10):2237–2251, 2008.

[5] M. Barrault, Y. Maday, N. C. Nguyen, and A. T. Patera. An 'empirical interpolation' method: application to efficient reduced-basis discretization of partial differential equations. *Comptes Rendus Mathematique*, 339(9):667–672, 2004.

[6] C. Beattie and S. Gugercin. Realization-independent $\mathcal{H}_2$-approximation. In *Proc. IEEE Conf. Decis. Control*, pages 4953–4958, Maui, HI, USA, 2012.

[7] C. Beattie, S. Gugercin, and S. Wyatt. Inexact solves in interpolatory model reduction. *Linear Algebra and its Applications*, 436(8):2916–2943, 2012. Special Issue dedicated to Danny Sorensen's 65th birthday.

[8] P. Benner, S. Gugercin, and K. Willcox. A survey of projection-based model reduction methods for parametric dynamical systems. *SIAM Review*, 57(4):483–531, 2015.

[9] P. Benner, V. Mehrmann, V. Sima, S. Van Huffel, and A. Varga. Slicot—a subroutine library in systems and control theory. In B. N. Datta, editor, *Applied and Computational Control, Signals, and Circuits: Volume 1*, pages 499–539, Boston, MA, 1999. Birkhäuser Boston.

[10] P. Benner and J. Saak. Linear-quadratic regulator design for optimal cooling of steel profiles. Technical Report SFB393/05-05, Sonderforschungsbereich 393 *Parallele Numerische Simulation für Physik und Kontinuumsmechanik*, TU Chemnitz, D-09107 Chemnitz (Germany), 2005.

[11] S. L. Brunton, J. L. Proctor, and J. N. Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113(15):3932–3937, 2016.

[12] D. Burov, D. Giannakis, K. Manohar, and A. Stuart. Kernel analog forecasting: Multiscale test problems. *Multiscale Modeling & Simulation*, 19(2):1011–1040, 2021.

[13] M. Campi and E. Weyer. Finite sample properties of system identification methods. *IEEE Transactions on Automatic Control*, 47(8):1329–1334, 2002.

[14] S. Chaturantabut and D. C. Sorensen. Nonlinear model reduction via discrete empirical interpolation. *SIAM Journal on Scientific Computing*, 32(5):2737–2764, 2010.

[15] E. Clark, S. L. Brunton, and J. N. Kutz. Multi-fidelity sensor selection: Greedy algorithms to place cheap and expensive sensors with cost constraints. *IEEE Sensors Journal*, 21(1):600–611, 2021.

[16] S. Dean, H. Mania, N. Matni, B. Recht, and S. Tu. On the sample complexity of the linear quadratic regulator. *Foundations of Computational Mathematics*, 20(4):633–679, 2019.

[17] Z. Drmač and S. Gugercin. A new selection operator for the discrete empirical interpolation method—improved a priori error bound and extensions. *SIAM Journal on Scientific Computing*, 38(2):A631–A648, 2016.

[18] Z. Drmač, S. Gugercin, and C. Beattie. Vector fitting for matrix-valued rational approximation. *SIAM Journal on Scientific Computing*, 37(5):A2346–A2379, 2015.

[19] Z. Drmač and B. Peherstorfer. Learning low-dimensional dynamical-system models from noisy frequency-response data with Loewner rational interpolation. In *Realization and Model Reduction of Dynamical Systems: A Festschrift in Honor of the 70th Birthday of Thanos Antoulas*. Springer, 2020.

[20] M. Embree and A. C. Ionita. Pseudospectra of Loewner matrix pencils. *arXiv*, 1910.12153, 2019.

[21] I. V. Gosea and A. C. Antoulas. Data-driven model order reduction of quadratic-bilinear systems. *Numerical Linear Algebra with Applications*, 25(6):e2200, 2018.

[22] M. Guo and J. S. Hesthaven. Reduced order modeling for nonlinear structural analysis using Gaussian process regression. *Computer Methods in Applied Mechanics and Engineering*, 341:807–826, 2018.

[23] M. Guo and J. S. Hesthaven. Data-driven reduced order modeling for time-dependent problems. *Computer Methods in Applied Mechanics and Engineering*, 345:75–99, 2019.

[24] B. Gustavsen and A. Semlyen. Rational approximation of frequency domain responses by vector fitting. *IEEE Transactions on Power Delivery*, 14(3):1052–1061, 1999.

[25] P. C. Hansen, J. Jørgensen, and W. R. B. Lionheart. *Computed Tomography: Algorithms, Insight, and Just Enough Theory*. SIAM, 2021.

[26] J. S. Hesthaven and S. Ubbiali. Non-intrusive reduced order modeling of nonlinear problems using neural networks. *Journal of Computational Physics*, 363:55–78, 2018.

[27] A. C. Ionita and A. C. Antoulas. Data-driven parametrized model reduction in the Loewner framework. *SIAM Journal on Scientific Computing*, 36(3):A984–A1007, 2014.

[28] I. C. F. Ipsen and B. Nadler. Refined perturbation bounds for eigenvalues of hermitian and non-hermitian matrices. *SIAM Journal on Matrix Analysis and Applications*, 31(1):40–53, 2009.

[29] J.-N. Juang and R. S. Pappa. An eigensystem realization algorithm for modal parameter identification and model reduction. *Journal of Guidance, Control, and Dynamics*, 8(5):620–627, 1985.

[30] T. Kmet' and J. Holčík. The diffusive Lotka-Volterra model as applied to the population dynamics of the german carp and predator and prey species in the Danube river basin. *Ecological Modelling*, 74(3-4):277–285, 1994.

[31] B. Kramer and S. Gugercin. Tangential interpolation-based eigensystem realization algorithm for MIMO systems. *Mathematical and Computer Modelling of Dynamical Systems*, 22(4):282–306, 2016.

[32] J. N. Kutz, S. L. Brunton, B. W. Brunton, and J. L. Proctor. *Dynamic mode decomposition: Data-driven modeling of complex systems*. SIAM, 2016.

[33] S. Lefteriu and A. C. Antoulas. A new approach to modeling multiport systems from frequency-domain data. *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, 29(1):14–27, 2010.

[34] S. Lefteriu, A. C. Ionita, and A. C. Antoulas. Modeling systems based on noisy frequency and time domain measurements. In J. C. Willems, S. Hara, Y. Ohta, and H. Fujioka, editors, *Perspectives in Mathematical System Theory, Control, and Signal Processing: A Festschrift in Honor of Yutaka Yamamoto on the Occasion of his 60th Birthday*, pages 365–378. Springer Berlin Heidelberg, 2010.

[35] L. Ljung. *System identification*. Prentice Hall, 1987.

[36] K. Manohar, B. W. Brunton, J. N. Kutz, and S. L. Brunton. Data-driven sparse sensor placement for reconstruction: Demonstrating the benefits of exploiting known patterns. *IEEE Control Systems Magazine*, 38(3):63–86, 2018.

[37] A. J. Mayo and A. C. Antoulas. A framework for the solution of the generalized realization problem. *Linear Algebra and its Applications*, 425(2–3):634–662, 2007.

[38] S. A. McQuarrie, C. Huang, and K. E. Willcox. Data-driven reduced-order models via regularised operator inference for a single-injector combustion process. *Journal of the Royal Society of New Zealand*, 51(2):194–211, Jan. 2021.

[39] B. Peherstorfer. Sampling low-dimensional Markovian dynamics for pre-asymptotically recovering reduced models from data with operator inference. *SIAM Journal on Scientific Computing*, 42:A3489–A3515, 2020.

[40] B. Peherstorfer, Z. Drmač, and S. Gugercin. Stability of discrete empirical interpolation and gappy proper orthogonal decomposition with randomized and deterministic sampling points. *SIAM Journal on Scientific Computing*, 42(5):A2837–A2864, 2020.

[41] B. Peherstorfer and K. Willcox. Data-driven operator inference for nonintrusive projection-based model reduction. *Computer Methods in Applied Mechanics and Engineering*, 306:196–215, 2016.

[42] E. Qian. *A scientific machine learning approach to learning reduced models for nonlinear*

partial differential equations. PhD thesis, Massachusetts Institute of Technology, 2021.

[43] E. Qian, B. Kramer, B. Peherstorfer, and K. Willcox. Lift & learn: Physics-informed machine learning for large-scale nonlinear dynamical systems. *Physica D: Nonlinear Phenomena*, 406:132401, 2020.

[44] S. Rosset and R. J. Tibshirani. From fixed-x to random-x regression: Bias-variance decompositions, covariance penalties, and prediction error estimation. *Journal of the American Statistical Association*, 115(529):138–151, 2020.

[45] C. Rowley, I. Mezić, S. Bagheri, P. Schlatter, and D. Henningson. Spectral analysis of nonlinear flows. *Journal of Fluid Mechanics*, 641:115–127, 2009.

[46] G. Rozza, D. Huynh, and A. T. Patera. Reduced basis approximation and a posteriori error estimation for affinely parametrized elliptic coercive partial differential equations. *Archives of Computational Methods in Engineering*, 15(3):1–47, 2008.

[47] S. H. Rudy, S. L. Brunton, J. L. Proctor, and J. N. Kutz. Data-driven discovery of partial differential equations. *Science Advances*, 3(4), 2017.

[48] S. H. Rudy, J. N. Kutz, and S. L. Brunton. Deep learning of dynamics and signal-noise decomposition with time-stepping constraints. *Journal of Computational Physics*, 396:483–506, 2019.

[49] N. Sawant, B. Kramer, and B. Peherstorfer. Physics-informed regularization and structure preservation for learning stable reduced models from data with operator inference. *arXiv:2107.02597*, 2021.

[50] H. Schaeffer, R. Caflisch, C. D. Hauck, and S. Osher. Sparse dynamics for partial differential equations. *Proceedings of the National Academy of Sciences*, 110(17):6634–6639, 2013.

[51] H. Schaeffer, G. Tran, and R. Ward. Extracting sparse high-dimensional dynamics from limited data. *SIAM Journal on Applied Mathematics*, 78(6):3279–3295, 2018.

[52] P. J. Schmid. Dynamic mode decomposition of numerical and experimental data. *Journal of Fluid Mechanics*, 656:5–28, 2010.

[53] J. Schoukens and L. Ljung. Nonlinear system identification: A user-oriented road map. *IEEE Control Systems Magazine*, 39(6):28–99, 2019.

[54] P. Seshadri, A. Narayan, and S. Mahadevan. Effectively subsampled quadratures for least squares polynomial approximations. *SIAM/ASA Journal on Uncertainty Quantification*, 5(1):1003–1023, 2017.

[55] V. Sima and P. Benner. Fast system identification and model reduction solvers. *IFAC Proceedings Volumes*, 40(13):477–482, 2007. 9th IFAC Workshop on Adaptation and Learning in Control and Signal Processing.

[56] M. Simchowitz, H. Mania, S. Tu, M. I. Jordan, and B. Recht. Learning without mixing: Towards a sharp analysis of linear system identification. In S. Bubeck, V. Perchet, and P. Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 439–473. PMLR, 06–09 Jul 2018.

[57] R. Swischuk, B. Kramer, C. Huang, and K. Willcox. Learning physics-based reduced-order models for a single-injector combustion process. *AIAA Journal*, 58(6):2658–2672, 2020.

[58] G. Tran and R. Ward. Exact recovery of chaotic systems from highly corrupted data. *Multiscale Modeling & Simulation*, 15(3):1108–1129, 2017.

[59] L. N. Trefethen. Finite difference and spectral methods for ordinary and partial differential equations, 1996.

[60] J. H. Tu, C. W. Rowley, D. M. Luchtenburg, S. L. Brunton, and J. N. Kutz. On dynamic mode decomposition: Theory and applications. *Journal of Computational Dynamics*, 1(2):391–421, 2014.

[61] W. I. T. Uy and B. Peherstorfer. Operator inference of non-Markovian terms for learning reduced models from partially observed state trajectories. *Journal of Scientific Computing*, 88(3):91, 2021.

[62] W. I. T. Uy and B. Peherstorfer. Probabilistic error estimation for non-intrusive reduced models learned from data of systems governed by linear parabolic partial differential equations. *ESAIM: Mathematical Modelling and Numerical Analysis (M2AN)*, 55(3):735–761, 2021.

[63] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Y. C. Eldar and G. Kutyniok, editors, *Compressed Sensing*, pages 210–268. Cambridge University Press, 2012.

[64] M. Vidyasagar and R. L. Karandikar. A learning theory approach to system identification and stochastic adaptive control. *Journal of Process Control*, 18(3):421–430, 2008. Festschrift honouring Professor Dale Seborg.

[65] C. R. Wentland, C. Huang, and K. Duraisamy. Investigation of sampling strategies for reduced-order models of rocket combustors. In *AIAA Scitech 2021 Forum*, pages 1–31. AIAA, 2021.

[66] M. O. Williams, I. G. Kevrekidis, and C. W. Rowley. A data–driven approximation of the

Koopman operator: Extending dynamic mode decomposition. *Journal of Nonlinear Science*, 25(6):1307–1346, 2015.

[67] S. Yıldız, P. Goyal, P. Benner, and B. Karasözen. Data-driven learning of reduced-order dynamics for a parametrized shallow water equation, 2020.

[68] S. Zhang and G. Lin. Robust data-driven discovery of governing physical laws with error bars. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 474(2217):20180305, 2018.

[69] S. Zhang and G. Lin. SubTSBR to tackle high noise and outliers for data-driven discovery of differential equations. *Journal of Computational Physics*, 428:109962, 2021.

Appendix A. Proof of Lemma 3.4.

By using the triangle inequality for the norm $\mathbb{E}[|\cdot|^l]^{1/l}$,

$$\begin{aligned}\left(\mathbb{E}[\|\boldsymbol{G}\|_2^l]\right)^{1/l} &= \left(\mathbb{E}\left[\left|\|\boldsymbol{G}\|_2 - \mathbb{E}[\|\boldsymbol{G}\|_2] + \mathbb{E}[\|\boldsymbol{G}\|_2]\right|^l\right]\right)^{1/l} \\ &\leq \left(\mathbb{E}\left[\left|\|\boldsymbol{G}\|_2 - \mathbb{E}[\|\boldsymbol{G}\|_2]\right|^l\right]\right)^{1/l} + \mathbb{E}[\|\boldsymbol{G}\|_2] \\ &\leq \left(\mathbb{E}\left[\left|\|\boldsymbol{G}\|_2 - \mathbb{E}[\|\boldsymbol{G}\|_2]\right|^l\right]\right)^{1/l} + \sqrt{n} + \sqrt{p} \qquad (1.1)\end{aligned}$$

where we have used the bound [63, Theorem 5.32].

Denote by $\Gamma(\cdot)$ the gamma function. Recall that that $\Gamma(x+1) \leq x^x$ for $x \geq 0$ and $\Gamma(x+1) = x\Gamma(x)$. To bound the first term in the right hand side of the inequality (1.1), we proceed as follows. For $t \geq 0$,

$$\begin{aligned}\mathbb{E}\left[\left|\|\boldsymbol{G}\|_2 - \mathbb{E}[\|\boldsymbol{G}\|_2]\right|^l\right] &= l\int_0^\infty t^{l-1} P\left(\left|\|\boldsymbol{G}\|_2 - \mathbb{E}[\|\boldsymbol{G}\|_2]\right| \geq t\right) dt \\ &\leq 2l\int_0^\infty t^{l-1} e^{-t^2/2}\, dt = 2l\int_0^\infty (2u)^{\frac{l-2}{2}} e^{-u}\, du \\ &= l2^{l/2}\int_0^\infty u^{l/2-1} e^{-u}\, du = 2^{l/2+1}\frac{l}{2}\Gamma\left(\frac{l}{2}\right) = 2^{l/2+1}\Gamma\left(\frac{l}{2}+1\right) \\ &\leq 2^{l/2+1}\left(\frac{l}{2}\right)^{l/2} = 2l^{l/2}\end{aligned}$$

where we utilized the concentration inequality [63, Proposition 5.34] and properties of the gamma function mentioned above. This implies that $\left(\mathbb{E}\left[\left|\|\boldsymbol{G}\|_2 - \mathbb{E}[\|\boldsymbol{G}\|_2]\right|^l\right]\right)^{1/l} \leq 2^{1/l}\sqrt{l}$ and the conclusion follows from (1.1).