Mathematical Modelling and Numerical Analysis Will be set by the publisher
Modélisation Mathématique et Analyse Numérique

ANOVA APPROXIMATION FOR HIGH-DIMENSIONAL CONTROL SYSTEMS

Melanie Kircheis[1], Daniel Potts[2], Manuel Schaller[2] and Karl Worthmann[3]

Abstract. We study the data-driven approximation of control-affine systems, where the drift and each control vector field exhibit a low-dimensional structure. While the ANOVA (ANalysis Of VAriance) decomposition is tailored to low-dimensional structures, a straightforward implementation does not exploit control affinity, rendering the approximation computationally expensive. We propose a coupled least squares approach for given state-control samples, whose ansatz is affine linear in the control by construction and can be realized by a preprocessing of the ANOVA index sets. Based on a Marcinkiewicz–Zygmund inequality for the resulting ansatz space, we prove a probabilistic L_2-error bound with explicitly given constants. Separate error control on drift and input vector fields renders the approach flexible to, e. g., parameter changes. Finally, we numerically illustrate the identification of the coupling structure via global sensitivity indices on discretizations of a nonlinear advection-diffusion equation, the shallow water equations and the viscous Burgers equation.

2020 Mathematics Subject Classification. 41A25, 41A63, 65D15, 65Txx, 93C55.

The dates will be set by the publisher.

1. Introduction

Whenever a first-principles model of a dynamical system is unavailable, too complex, or too costly to evaluate, a surrogate model can be learned from measured data. Such a model can then be used for prediction and controller design, e. g., within model predictive control [26]. In this work, we consider discrete-time control-affine systems

$$\boldsymbol{x}^+ = F(\boldsymbol{x}) + G(\boldsymbol{x})\boldsymbol{u}$$

with state $\boldsymbol{x}$, successor state $\boldsymbol{x}^+ \in \mathbb{R}^d$, and control input $\boldsymbol{u} \in \mathbb{R}^m$. Such control-affine structures naturally arise, e. g., from (time-discretized) mechanical, chemical, or fluid-dynamical models as well as from spatial semi-discretizations of controlled partial differential equations [5, 11]. Our approach relies on two structural properties. First, the dynamics are affine linear in the control. Second, as is the case in many applications, each component of F and G depends on a few state variables only and couples them through low-order interactions. For instance, in finite-difference or finite-volume discretizations of PDEs, each component couples only with its spatial neighbours. Identifying this coupling structure from data reduces the computational effort, and the learned model becomes interpretable and amenable to structure-exploiting controller design.

Keywords and phrases: control systems; fast Fourier methods; high-dimensional approximation; data-driven method; ANOVA

[1] kircheis@hs-mittweida.de, Faculty of Applied Computer Sciences and Biosciences, University of Applied Sciences Mittweida

[2] {potts, manuel.schaller}@math.tu-chemnitz.de, Faculty of Mathematics, Chemnitz University of Technology

[3] karl.worthmann@tu-ilmenau.de, Optimization-based control group, Institute of Mathematics, Technische Universität Ilmenau

© EDP Sciences, SMAI 2026

The ANOVA (ANalysis Of VAriance) decomposition is tailored to such low-dimensional structures, since it splits high-dimensional functions into terms depending on subsets of the variables only and quantifies the importance of each term by global sensitivity indices. For control systems, however, the ANOVA approximation is not directly applicable. On the one hand, F and G can, in general, not be sampled separately. On the other hand, a naive ANOVA approximation of the joint map $(\boldsymbol{x}, \boldsymbol{u}) \mapsto H(\boldsymbol{x}, \boldsymbol{u}) = F(\boldsymbol{x}) + G(\boldsymbol{x})\boldsymbol{u}$ would spend considerable computational effort on interactions of the control variables which are excluded a priori by the control-affine structure. In this paper, we propose an approach which takes the control-affine structure into account and works with given samples $(\boldsymbol{x}^i, \boldsymbol{u}^i, H(\boldsymbol{x}^i, \boldsymbol{u}^i))$, i. e., it does not require particular control values to be chosen by the user. Rather, it encodes the control-affine structure into the approximation ansatz itself. To this end, the ANOVA index sets and the basis in the control variables are chosen a priori. This preprocessing step ensures that the ansatz is affine linear in $\boldsymbol{u}$, so that spurious control interactions cannot arise by construction.

We briefly comment on related work. Data-driven identification of dynamical systems under structural assumptions has been studied, for example, using sparse regression approaches such as SINDy [4] and its extension to control systems [3]. In these methods, sparsity is enforced on the coefficient vector over a fixed dictionary. In contrast, our approach identifies sparsity in the *variable couplings*, i. e., in the ANOVA terms. Koopman-based methods, in particular extended dynamic mode decomposition [32], likewise exploit the control-affine structure by means of bilinear surrogate models. This is possible because the Koopman *generator* inherits the control affinity of the dynamics [21]. For an error analysis of kernel-based EDMD we refer the reader to [15] and, for an extension to control systems, to [2]. Finally, this work builds upon the ANOVA and high-dimensional model representation (HDMR) literature on sensitivity analysis [25, 30] and on the ANOVA approximation approach [23, 28]. To the best of our knowledge, a combination of the ANOVA approximation with the control-affine structure as proposed here has not been considered so far.

Contribution and outline. In the preparatory Section 2, we briefly recall the ANOVA decomposition, its data-driven implementation and the corresponding error analysis for the Chebyshev basis. The contribution of this work is Section 3, where we make this machinery applicable to control-affine systems by means of a coupled least squares approach for given state-control samples. In Section 3.1, we characterize its limit for infinitely many samples as an orthogonal projection in Proposition 3.4. In Section 3.2, we establish a Marcinkiewicz–Zygmund inequality for the resulting ansatz space in Proposition 3.7, i. e., a norm equivalence between the discrete and the continuous norm which holds uniformly on the ansatz space with high probability. Based on it, we prove a probabilistic L_2-error bound with explicit constants in Theorem 3.8, which also yields error bounds for the drift and the input map separately, see Remark 3.10. Section 3.3 extends these results to multivariate controls, see Corollary 3.13. In Section 4, we illustrate the approach numerically, before we conclude in Section 5. Appendix A describes how the coupled approach may be implemented within existing ANOVA software by a suitable preprocessing of the index sets, see Algorithm A.1.

2. The ANOVA decomposition of scalar-valued functions

In this part, we recall the ANOVA decomposition and refer the reader to [6,16,18,25] for a detailed exposition. Let $\boldsymbol{v} \subseteq [d] \coloneqq \{1, \dots, d\}$ denote a subset of indices. For high-dimensional functions $f\colon \mathbb{R}^d \to \mathbb{R}$, we utilize additive decompositions of the form

$$f(\boldsymbol{x}) = f_\varnothing + \sum_{j=1}^{d} f_{\{j\}}(x_j) + \sum_{1\le i<j\le d} f_{\{i,j\}}(x_i, x_j) + \cdots + f_{[d]}(x_1, \dots, x_d) =: \sum_{\boldsymbol{v}\subseteq[d]} f_{\boldsymbol{v}}(\boldsymbol{x}_{\boldsymbol{v}}), \tag{1}$$

where the term $f_{\boldsymbol{v}}$ corresponding to $\boldsymbol{v}$ only depends on x_j with $j \in \boldsymbol{v}$, or, equivalently, on $\boldsymbol{x}_{\boldsymbol{v}} \coloneqq (x_j)_{j\in\boldsymbol{v}}$ using the inherent ordering of $[d]$. Clearly, every function $f\colon \mathbb{R}^d \to \mathbb{R}$ may be decomposed in such a way by summing over 2^d terms with $\boldsymbol{v}$ in the power set of $[d]$. In this work, however, we are interested in the case where the sum above contains significantly fewer summands and the cardinality $|\boldsymbol{v}|$ of the involved index sets $\boldsymbol{v}$ is small in comparison to the dimension d. Such functions are said to have a *low-dimensional structure*.

Example 2.1. *For $d = 4$ consider the function*

$$f(\boldsymbol{x}) = 1 + x_1^2 + x_1 x_2^4 + x_3 x_4 - \sin(x_4). \tag{2}$$

Clearly, this function can be represented using the five summands

$$f_\varnothing = 1, \quad f_{\{1\}}(x_1) = x_1^2, \quad f_{\{1,2\}}(x_1, x_2) = x_1 x_2^4, \quad f_{\{3,4\}}(x_3, x_4) = x_3 x_4, \quad f_{\{4\}}(x_4) = -\sin(x_4). \tag{3}$$

This yields significantly less summands than the number $2^4 = 16$ of possible terms, which corresponds to the cardinality of the power set of $[4] = \{1, 2, 3, 4\}$, since each subset $\boldsymbol{v}$ may contribute a term $f_{\boldsymbol{v}}$ in (1). Moreover, in (2), each non-trivial term $f_{\boldsymbol{v}}$, $\boldsymbol{v} \subseteq [d]$, depends on two arguments at most.

The representation (1) is not unique in general. It becomes unique if one additionally requires the terms $f_{\boldsymbol{v}}$ to have mean zero and to be pairwise L_2-orthogonal, see, e. g., [16, Theorem 2.1 and Example 2.2]. This is the ANOVA decomposition defined below. These conditions characterize the $f_{\boldsymbol{v}}$ through L_2 inner products over the domain $\Omega \subseteq \mathbb{R}^d$, i. e., through integrals that are however inaccessible when f is only known through pointwise samples. In Section 2.2, we therefore replace them by a data-driven least squares approximation over a subspace $\operatorname{span}\{\varphi_1, \dots, \varphi_N\}$ spanned by linearly independent functions.

2.1. **Definition**

In this part, we gather the notational conventions needed for the definition of the ANOVA decomposition. For a bounded set $\Omega \subset \mathbb{R}^d$ and a probability density $\omega \colon \Omega \to (0, \infty)$ we write $L_2(\Omega, \omega)$ for the weighted Lebesgue space of real-valued functions with the inner product $\langle f, g \rangle_{L_2(\Omega,\omega)} \coloneqq \int_\Omega f g \, \omega \, \mathrm{d}\boldsymbol{x}$ and the induced norm. We consider tensor product domains and densities

$$\Omega \coloneqq \bigtimes_{j=1}^{d} \Omega_j, \quad \omega(\boldsymbol{x}) \coloneqq \prod_{j=1}^{d} \omega_j(x_j), \quad \int_{\Omega_j} \omega_j(x_j) \, \mathrm{d}x_j = 1, \quad j = 1, \dots, d, \tag{4}$$

and write $\Omega_{\boldsymbol{v}} \coloneqq \bigtimes_{j \in \boldsymbol{v}} \Omega_j$ and $\omega_{\boldsymbol{v}}(\boldsymbol{x}_{\boldsymbol{v}}) \coloneqq \prod_{j \in \boldsymbol{v}} \omega_j(x_j)$ for $\boldsymbol{v} \subseteq [d]$. By separability, $L_2(\Omega, \omega)$ has an orthonormal tensor product basis $\varphi_{\boldsymbol{k}} \coloneqq \prod_{j=1}^d \varphi_{k_j}(x_j)$, $\boldsymbol{k} \in \mathbb{N}_0^d$, built from orthonormal bases of the $L_2(\Omega_j, \omega_j)$. Hence, every $f \in L_2(\Omega, \omega)$ admits the expansion

$$f(\boldsymbol{x}) = \sum_{\boldsymbol{k} \in \mathbb{N}_0^d} c_{\boldsymbol{k}}(f) \, \varphi_{\boldsymbol{k}}(\boldsymbol{x}), \qquad c_{\boldsymbol{k}}(f) \coloneqq \langle f, \varphi_{\boldsymbol{k}} \rangle_{L_2(\Omega,\omega)}, \tag{5}$$

where we may assume $\varphi_{\boldsymbol{0}} \equiv 1$ without loss of generality. In the error analysis and in the numerical experiments, we work with the *Chebyshev basis*[1]

$$\begin{aligned} &\varphi_{k_j}^{\mathrm{cheb}}(x_j) \coloneqq \sqrt{2}^{\min\{1,k_j\}} \cos(k_j \arccos(x_j)), \ k_j \in \mathbb{N}_0, \\ &\text{with} \ \ \Omega_j = [-1, 1], \ \omega_j^{\mathrm{cheb}}(x_j) = \frac{1}{\pi \sqrt{1 - x_j^2}}, \quad j \in [d]. \end{aligned} \tag{6}$$

With these conventions in place, we can now state the ANOVA decomposition. We follow [16], where it is derived together with the properties collected below.

[1] A general interval $[a, b]$ is mapped to $[-1, 1]$ via $x \mapsto \frac{2(x-a)}{b-a} - 1$, so restricting to $\Omega_j = [-1, 1]$ in (6) is no loss of generality.

Definition 2.2. *For a function $f \in L_2(\Omega,\omega)$ and a subset of indices $\boldsymbol{v} \subseteq [d]$, the* ANOVA (ANalysis Of VAriance) terms *are recursively defined by*

$$f_\varnothing \coloneqq \int_\Omega f(\boldsymbol{x})\,\omega(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}, \qquad f_{\boldsymbol{v}}(\boldsymbol{x}_{\boldsymbol{v}}) \coloneqq \int_{\Omega_{\boldsymbol{v}^c}} f(\boldsymbol{x})\,\omega_{\boldsymbol{v}^c}(\boldsymbol{x}_{\boldsymbol{v}^c})\,\mathrm{d}\boldsymbol{x}_{\boldsymbol{v}^c} - \sum_{\boldsymbol{z}\subsetneq\boldsymbol{v}} f_{\boldsymbol{z}}(\boldsymbol{x}_{\boldsymbol{z}}), \tag{7}$$

where the complement of $\boldsymbol{v}$ is defined as $\boldsymbol{v}^c \coloneqq [d]\setminus\boldsymbol{v}$. The corresponding expansion of the form (1) *is called the* ANOVA decomposition *and it fulfills the zero-mean property*

$$\int_{\Omega_{\boldsymbol{v}}} f_{\boldsymbol{v}}(\boldsymbol{x}_{\boldsymbol{v}})\,\omega_{\boldsymbol{v}}(\boldsymbol{x}_{\boldsymbol{v}})\,\mathrm{d}\boldsymbol{x}_{\boldsymbol{v}} = 0, \quad \boldsymbol{v}\neq\varnothing, \tag{8}$$

the annihilation property

$$\int_{\Omega_j} f_{\boldsymbol{v}}(\boldsymbol{x}_{\boldsymbol{v}})\,\omega_j(x_j)\,\mathrm{d}x_j = 0, \quad j\in\boldsymbol{v}\neq\varnothing,$$

and thus $L_2(\Omega,\omega)$-orthogonality

$$\langle f_{\boldsymbol{v}}, f_{\boldsymbol{z}}\rangle_{L_2(\Omega,\omega)} = 0, \quad \boldsymbol{v}\neq\boldsymbol{z}\subseteq[d]. \tag{9}$$

The following example illustrates that the ANOVA terms need not coincide with the naive representation of a function.

Example 2.3. *We illustrate Definition 2.2 by computing the ANOVA terms of the function* (2) *from Example 2.1 in the weighted Chebyshev space $L_2([-1,1]^4,\omega^{\mathrm{cheb}})$. Evaluating* (7) *yields the constant term $f_\varnothing = \frac{3}{2}$, the one-dimensional terms*

$$f_{\{1\}}(x_1) = x_1^2 + \tfrac{3}{8}x_1 - \tfrac{1}{2}, \qquad f_{\{4\}}(x_4) = -\sin(x_4), \qquad f_{\{2\}} = f_{\{3\}} = 0,$$

and the two-dimensional terms

$$f_{\{1,2\}}(x_1,x_2) = x_1x_2^4 - \tfrac{3}{8}x_1, \qquad f_{\{3,4\}}(x_3,x_4) = x_3x_4.$$

All remaining pair terms as well as all terms of order three and four vanish. Hence, the decomposition is of the expected form

$$f(\boldsymbol{x}) = f_\varnothing + f_{\{1\}}(x_1) + f_{\{1,2\}}(x_1,x_2) + f_{\{3,4\}}(x_3,x_4) + f_{\{4\}}(x_4). \tag{10}$$

Note, however, that the ANOVA terms do not coincide with the representation (3)*. The zero-mean and orthogonality constraints enforce additional lower-order corrections such as the summand $\frac{3}{8}x_1$ appearing in $f_{\{1\}}$ and $f_{\{1,2\}}$. Thus, even if a function has an obvious low-dimensional structure, the number of ANOVA terms required to represent it is not known a priori.*

Using the basis expansion (5), the ANOVA terms (7) may also be written as

$$f_{\boldsymbol{v}}(\boldsymbol{x}_{\boldsymbol{v}}) = \sum_{\substack{\boldsymbol{k}\in\mathbb{N}_0^d\\ \operatorname{supp}\boldsymbol{k}=\boldsymbol{v}}} c_{\boldsymbol{k}}(f)\,\varphi_{\boldsymbol{k}}(\boldsymbol{x}_{\boldsymbol{v}}), \quad \boldsymbol{v}\subseteq[d], \tag{11}$$

where $\operatorname{supp}\boldsymbol{k} \coloneqq \{j\in[d]\colon k_j\neq 0\}$, cf. [22]. Note that for $\operatorname{supp}\boldsymbol{k} = \boldsymbol{v}$ the tensor-product basis function $\varphi_{\boldsymbol{k}}(\boldsymbol{x}) = \prod_{j=1}^d \varphi_{k_j}(x_j)$ satisfies $\varphi_{k_j}\equiv 1$ for all $j\notin\boldsymbol{v}$, since $\varphi_0\equiv 1$. Therefore, $\varphi_{\boldsymbol{k}}$ depends only on $\boldsymbol{x}_{\boldsymbol{v}}$,

which justifies writing $\varphi_{\boldsymbol{k}}(\boldsymbol{x}_{\boldsymbol{v}})$. The same convention is used in Appendix A for indices with a non-zero control component. In other words, the representation (11) exhibits the ANOVA decomposition as a partition of the space of basis coefficients $c_{\boldsymbol{k}}(f)$. Therein, the term for $\boldsymbol{v} = \varnothing$ collects the single coefficient $c_{\boldsymbol{0}}(f)$, the term for a singleton $\boldsymbol{v} = \{j\}$ collects all coefficients on the respective axis except the 0-th one, and so on. Figure 1 illustrates this partition for the index set $\mathbb{N}_0^d$ underlying the bases considered here. Every coefficient of (5) is thus attributed to exactly one ANOVA term.

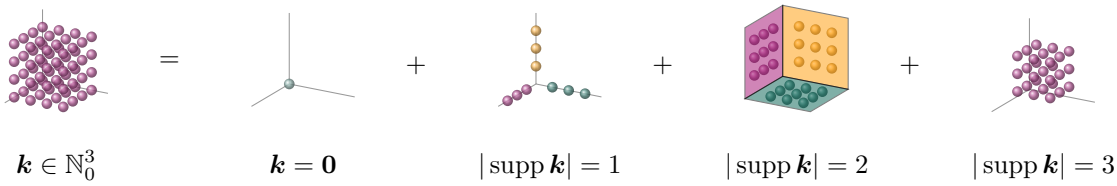


FIGURE 1. Visualization of the partition in the space of basis coefficients, as provided by the ANOVA decomposition, for a 3-dimensional function in the setting $\boldsymbol{k} \in \mathbb{N}_0^d$ underlying the Chebyshev basis (6).

The ANOVA decomposition also quantifies how much an individual term contributes to f, which will later allow us to discard the irrelevant ones. With the variance of f and of an ANOVA term,

$$\sigma^2(f) \coloneqq \|f\|^2_{L_2(\Omega,\omega)} - f_\varnothing^2, \qquad \sigma^2(f_{\boldsymbol{v}}) \coloneqq \|f_{\boldsymbol{v}}\|^2_{L_2(\Omega_{\boldsymbol{v}},\omega_{\boldsymbol{v}})}, \quad \boldsymbol{v} \subseteq [d], \tag{12}$$

the zero-mean property (8) and the orthogonality (9) yield the variance decomposition $\sigma^2(f) = \sum_{\varnothing \neq \boldsymbol{v} \subseteq [d]} \sigma^2(f_{\boldsymbol{v}})$, which explains the name "analysis of variance". The share of the variance explained by the interaction of the variables indexed by $\boldsymbol{v}$ is measured by the *global sensitivity indices*, or *Sobol indices*,

$$S_{\boldsymbol{v}}(f) \coloneqq \frac{\sigma^2(f_{\boldsymbol{v}})}{\sigma^2(f)} = \frac{\sum_{\boldsymbol{k} \in \mathbb{N}_0^d,\, \operatorname{supp} \boldsymbol{k} = \boldsymbol{v}} |c_{\boldsymbol{k}}(f)|^2}{\sum_{\boldsymbol{k} \in \mathbb{N}_0^d,\, \boldsymbol{k} \neq \boldsymbol{0}} |c_{\boldsymbol{k}}(f)|^2} \in [0,1], \quad \varnothing \neq \boldsymbol{v} \subseteq [d], \tag{13}$$

see [30], where the second expression follows from the basis representations (5) and (11).

2.2. **Efficient and data-driven implementation**

Although (1) carries valuable information about a function, its defining relations (7) are not amenable to computation. On the one hand, the number of terms grows exponentially in the dimension d. On the other hand, each term is defined through integrals, which are inaccessible when f is only known through samples. To enable an efficient and purely data-driven implementation, we therefore replace these relations by a chain of approximations, cf. [23]. We first *truncate* the decomposition to its relevant low-dimensional terms, then *project* each term onto a finite basis, *estimate* the resulting coefficients from samples, and finally *post-process* the approximation to quantify the importance of the individual ANOVA terms. These four steps are detailed in the following.

2.2.1. *Truncation*

Since it is unknown a priori which of the 2^d terms in (1) are relevant, we first truncate the decomposition, that is, we remove terms $f_{\boldsymbol{v}}$. To this end, we invoke the sparsity-of-effects principle [19], according to which many phenomena are well described by a few low-order interactions. More precisely, we introduce the *superposition dimension* $q \in \mathbb{N}$, $q \ll d$, and truncate (1) as

$$f(\boldsymbol{x}) \approx \sum_{|\boldsymbol{v}| \leq q} f_{\boldsymbol{v}}(\boldsymbol{x}_{\boldsymbol{v}}). \tag{14}$$

In many real-world problems the superposition dimension is indeed low, i. e., small choices $q \in \{2, 3\}$ already suffice for a good approximation, see [6, 9, 16, 22].

Now that we have reduced the number of ANOVA terms to be determined, we turn to their actual computation. Both representations (7) and (11) involve integrals and are thus not directly amenable to computation. We build on the latter and render it computable in two steps, the first of which is the following projection.

2.2.2. *Projection*

The representation (11) of an ANOVA term still involves infinitely many basis coefficients. Therefore, we introduce another truncation step, that is, for every subset $\boldsymbol{v} \subseteq [d]$ we define the index set

$$\mathcal{I}_{\boldsymbol{N}^{\boldsymbol{v}}} \coloneqq \left\{ \boldsymbol{k} \in \mathbb{N}_0^d \colon \operatorname{supp} \boldsymbol{k} = \boldsymbol{v},\, \boldsymbol{k}_{\boldsymbol{v}} \in \bigtimes_{j=1}^{|\boldsymbol{v}|} \{1, \dots, N_j^{\boldsymbol{v}} - 1\} \right\}, \qquad \boldsymbol{N}^{\boldsymbol{v}} \coloneqq \left(N_j^{\boldsymbol{v}}\right)_{j=1}^{|\boldsymbol{v}|} \in (2\mathbb{N})^{|\boldsymbol{v}|}. \tag{15}$$

In other words, for $\boldsymbol{v} = \varnothing$ we keep the single coefficient $c_{\boldsymbol{0}}(f)$ and for all $\boldsymbol{v} \neq \varnothing$ we consider only the first $N_j^{\boldsymbol{v}} - 1$ coefficients of (11) in every direction j that is present in $\boldsymbol{v}$. For instance, when considering $\boldsymbol{v} = \{1, 2\}$ with $\boldsymbol{N}^{\boldsymbol{v}} = (6, 4)^\top \in (2\mathbb{N})^2$ for our 4-dimensional example function (2), then the index set $\mathcal{I}_{\boldsymbol{N}^{\boldsymbol{v}}}$ contains indices $\boldsymbol{k} = (k_1, k_2, 0, 0)^\top$ with $k_1 \in \{1, 2, \dots, 5\}$ and $k_2 \in \{1, 2, 3\}$. Hence, the approximation of the ANOVA terms (7) is given by

$$f_{\boldsymbol{v}}(\boldsymbol{x}_{\boldsymbol{v}}) \approx \sum_{\boldsymbol{k} \in \mathcal{I}_{\boldsymbol{N}^{\boldsymbol{v}}}} c_{\boldsymbol{k}}(f)\, \varphi_{\boldsymbol{k}}(\boldsymbol{x}_{\boldsymbol{v}}), \quad \boldsymbol{v} \subseteq [d].$$

Using the (disjoint) union

$$\mathcal{I}_{\boldsymbol{N}} \coloneqq \bigcup_{|\boldsymbol{v}| \leq q} \mathcal{I}_{\boldsymbol{N}^{\boldsymbol{v}}} \tag{16}$$

of the index sets (15) we approximate the function f in (5) as

$$f(\boldsymbol{x}) \approx P_{\boldsymbol{N}} f(\boldsymbol{x}) \coloneqq \sum_{\boldsymbol{k} \in \mathcal{I}_{\boldsymbol{N}}} c_{\boldsymbol{k}}(f)\, \varphi_{\boldsymbol{k}}(\boldsymbol{x}), \tag{17}$$

i. e., by the orthogonal projection of f onto the subspace $\operatorname{span}\{\varphi_{\boldsymbol{k}}\}_{\boldsymbol{k} \in \mathcal{I}_{\boldsymbol{N}}} \subset L_2(\Omega, \omega)$ employing the finite dictionary $\mathcal{D} \coloneqq \{\varphi_{\boldsymbol{k}}\}_{\boldsymbol{k} \in \mathcal{I}_{\boldsymbol{N}}} \subset \{\varphi_{\boldsymbol{k}}\}_{\boldsymbol{k} \in \mathbb{N}_0^d}$.

2.2.3. *Estimation*

The above approximation (17) of the function f still requires the integrals $c_{\boldsymbol{k}}(f) = \langle f, \varphi_{\boldsymbol{k}} \rangle_{L_2(\Omega,\omega)}$, $\boldsymbol{k} \in \mathcal{I}_{\boldsymbol{N}}$, from (5), which are inaccessible in a purely data-driven setting. We therefore keep the ansatz (17) and determine its coefficients from data. To this end, assume we are given i.i.d. sample points $\{\boldsymbol{x}^1, \dots, \boldsymbol{x}^M\} \subset \Omega$, $M \in \mathbb{N}$, drawn according to the density ω, as well as the corresponding function evaluations $\{f(\boldsymbol{x}^1), \dots, f(\boldsymbol{x}^M)\} \subset \mathbb{R}$. Then we may approximate the basis coefficients $c_{\boldsymbol{k}}(f)$, $\boldsymbol{k} \in \mathcal{I}_{\boldsymbol{N}}$, in (17) by solving the least squares problem

$$\min_{\boldsymbol{c} \in \mathbb{R}^{|\mathcal{I}_{\boldsymbol{N}}|}} \|\boldsymbol{A}\boldsymbol{c} - \boldsymbol{f}\|_2^2, \tag{18}$$

where

$$\boldsymbol{A} \coloneqq \left(\varphi_{\boldsymbol{k}}(\boldsymbol{x}^i)\right)_{i=1,\, \boldsymbol{k} \in \mathcal{I}_{\boldsymbol{N}}}^{M} \in \mathbb{R}^{M \times |\mathcal{I}_{\boldsymbol{N}}|}, \qquad \text{and} \qquad \boldsymbol{f} \coloneqq \left(f(\boldsymbol{x}^i)\right)_{i=1}^{M} \in \mathbb{R}^M.$$

We compute the solution of (18) via the normal equations of the first kind

$$\boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{c} = \boldsymbol{A}^\top \boldsymbol{f}. \tag{19}$$

If $\boldsymbol{A}$ has full column rank $|\mathcal{I}_{\boldsymbol{N}}|$, the matrix $\boldsymbol{A}^\top \boldsymbol{A}$ is invertible and the unique solution of the least squares problem (18) is given by

$$(\tilde{c}_{\boldsymbol{k}})_{\boldsymbol{k}\in\mathcal{I}_{\boldsymbol{N}}} = \tilde{\boldsymbol{c}} \coloneqq (\boldsymbol{A}^\top \boldsymbol{A})^{-1} \boldsymbol{A}^\top \boldsymbol{f}. \tag{20}$$

Full column rank of $\boldsymbol{A}$ requires the overdetermined setting $M \geq |\mathcal{I}_{\boldsymbol{N}}|$. Moreover, if the $\boldsymbol{x}^i$ are drawn i.i.d. according to ω and the oversampling factor $M/|\mathcal{I}_{\boldsymbol{N}}|$ is large enough, $\boldsymbol{A}$ is well-conditioned with high probability, see [28, Theorem 4.2] and [13]. In the underdetermined case, regularization techniques are available [28], which we do not pursue here. In high dimensions, (20) is computed iteratively by the LSQR algorithm [20], which never forms $\boldsymbol{A}$ explicitly but only requires multiplication with $\boldsymbol{A}$ and $\boldsymbol{A}^\top$. We use the implementation [27], which provides fast algorithms for these products.

2.2.4. *Post-processing and sensitivities*

It remains to assess the importance of the individual ANOVA terms. We approximate the global sensitivity indices (13) using the estimated coefficients (20) and the index sets (15), (16) as

$$S_{\boldsymbol{v}}(f) \approx S_{\boldsymbol{v}}(\tilde{f}) = \frac{\sum_{\boldsymbol{k}\in\mathcal{I}_{\boldsymbol{N}^{\boldsymbol{v}}}} |\tilde{c}_{\boldsymbol{k}}|^2}{\sum_{\boldsymbol{k}\in\mathcal{I}_{\boldsymbol{N}}\setminus\{\boldsymbol{0}\}} |\tilde{c}_{\boldsymbol{k}}|^2}, \quad \varnothing \neq \boldsymbol{v} \subseteq [d], \tag{21}$$

with the intermediate approximation $\tilde{f}$ built from the $\tilde{c}_{\boldsymbol{k}}$ of (20). This is accurate whenever $f_{\boldsymbol{v}} \approx \tilde{f}_{\boldsymbol{v}}$, since $|\sigma^2(f_{\boldsymbol{v}}) - \sigma^2(\tilde{f}_{\boldsymbol{v}})| \leq \|f_{\boldsymbol{v}} - \tilde{f}_{\boldsymbol{v}}\| \, (2\|f_{\boldsymbol{v}}\| + \|f_{\boldsymbol{v}} - \tilde{f}_{\boldsymbol{v}}\|)$ in $L_2(\Omega_{\boldsymbol{v}}, \omega_{\boldsymbol{v}})$, cf. [24, (5.3)]. Given a threshold $\varepsilon \in (0,1)$, the relevant ANOVA terms are collected in

$$V \coloneqq \{\varnothing\} \cup \left\{ \boldsymbol{v} \subseteq [d] \colon 0 < |\boldsymbol{v}| \leq q, \ S_{\boldsymbol{v}}(\tilde{f}) > \varepsilon \right\}, \tag{22}$$

where $\boldsymbol{v} = \varnothing$ is added manually, since (13) carries no information about the constant term. Restricting (16) to these terms gives

$$\mathcal{I}(V) \coloneqq \bigcup_{\boldsymbol{v}\in V} \mathcal{I}_{\boldsymbol{N}^{\boldsymbol{v}}}, \tag{23}$$

and we solve the correspondingly truncated least squares problem

$$\min_{\boldsymbol{c}\in\mathbb{R}^{|\mathcal{I}(V)|}} \|\boldsymbol{A}_V \boldsymbol{c} - \boldsymbol{f}\|_2^2 \tag{24}$$

with the truncated matrix $\boldsymbol{A}_V \coloneqq \big(\varphi_{\boldsymbol{k}}(\boldsymbol{x}^i)\big)_{i=1,\,\boldsymbol{k}\in\mathcal{I}(V)}^{M} \in \mathbb{R}^{M\times|\mathcal{I}(V)|}$ to improve the fit of (20), and set

$$f^\star(\boldsymbol{x}) \coloneqq \sum_{\boldsymbol{k}\in\mathcal{I}(V)} c_{\boldsymbol{k}}^\star \varphi_{\boldsymbol{k}}(\boldsymbol{x}) \tag{25}$$

as our approximation of f. The resulting workflow is summarized as follows.

Algorithm 2.4 (ANOVA approximation, cf. [22, 23, 27, 29])**.**

Input:	$d \in \mathbb{N}$	*dimension*
	$q \in \mathbb{N}$, $q \ll d$	*superposition dimension*
	$\{\boldsymbol{x}^1, \dots, \boldsymbol{x}^M\} \subset \Omega$, $M \in \mathbb{N}$	*i.i.d. sample points drawn according to the density* ω
	$\boldsymbol{f} = \big(f(\boldsymbol{x}^i)\big)_{i=1}^M$	*function evaluations at the sample points*
	$\varepsilon \in (0,1)$	*threshold parameter*
	$r > 0$	*confidence parameter, cf. Theorem 2.6*

(1) *For all subsets with* $|\boldsymbol{v}| \le q$ *choose* $\boldsymbol{N}^{\boldsymbol{v}} = \left(N_j^{\boldsymbol{v}}\right)_{j=1}^{|\boldsymbol{v}|} \in (2\mathbb{N})^{|\boldsymbol{v}|}$ *in* (15) *such that the index set* $\mathcal{I}_{\boldsymbol{N}}$ *in* (16) *satisfies* $|\mathcal{I}_{\boldsymbol{N}}| \le \frac{1}{10\cdot 2^{\alpha}(r+1)} \cdot \frac{M}{\log M}$ *with the given* $r > 0$*, where* $\alpha \coloneqq \max_{\boldsymbol{k}\in\mathcal{I}_{\boldsymbol{N}}} |\operatorname{supp} \boldsymbol{k}| \le q$.

(2) *Compute the estimated coefficients* $\tilde{\boldsymbol{c}} = (\tilde{c}_{\boldsymbol{k}})_{\boldsymbol{k}\in\mathcal{I}_{\boldsymbol{N}}}$ *as solution to the least squares problem* (18) *using the LSQR algorithm.*

(3) *Compute the approximate global sensitivity indices*

$$S_{\boldsymbol{v}}(\tilde{f}) = \frac{\sum_{\boldsymbol{k}\in\mathcal{I}_{\boldsymbol{N}^{\boldsymbol{v}}}} |\tilde{c}_{\boldsymbol{k}}|^2}{\sum_{\boldsymbol{k}\in\mathcal{I}_{\boldsymbol{N}}\setminus\{\boldsymbol{0}\}} |\tilde{c}_{\boldsymbol{k}}|^2}, \quad \varnothing \ne \boldsymbol{v} \subseteq [d],$$

see (21).

(4) *Determine the set*

$$V = \{\varnothing\} \cup \left\{\boldsymbol{v} \subseteq [d] \colon 0 < |\boldsymbol{v}| \le q,\ S_{\boldsymbol{v}}(\tilde{f}) > \varepsilon\right\}$$

of relevant ANOVA terms, see (22)*, and set* $\mathcal{I}(V) = \bigcup_{\boldsymbol{v}\in V} \mathcal{I}_{\boldsymbol{N}^{\boldsymbol{v}}}$*, see* (23).

(5) *Compute the coefficients* $\boldsymbol{c}^{\star} = (c_{\boldsymbol{k}}^{\star})_{\boldsymbol{k}\in\mathcal{I}(V)}$ *as solution to the truncated least squares problem* (24) *using the LSQR algorithm.*

Output: *coefficients* $(c_{\boldsymbol{k}}^{\star})_{\boldsymbol{k}\in\mathcal{I}(V)} \in \mathbb{R}^{|\mathcal{I}(V)|}$ *of the approximation* $f^{\star} = \sum_{\boldsymbol{k}\in\mathcal{I}(V)} c_{\boldsymbol{k}}^{\star}\varphi_{\boldsymbol{k}} \approx f$*, see* (25)

Algorithm 2.4 combines the truncation, projection, estimation, and post-processing steps introduced above into a single data-driven ANOVA approximation. Before analyzing its approximation error, we comment on two ways in which the procedure can be refined in practice.

Remark 2.5 (Possible improvements of the algorithm)**.** *Note that in steps 4 and 5 of Algorithm 2.4, the index sets* (15) *may be further adjusted. Since the truncation by* V *in* (22) *typically renders* $\mathcal{I}(V)$ *in* (23) *considerably smaller than* $\mathcal{I}_{\boldsymbol{N}}$ *in* (16)*, the condition* $M \ge |\mathcal{I}(V)|$ *now suffices. This allows for larger choices of the corresponding bandwidth parameters* $\boldsymbol{N}^{\boldsymbol{v}}$ *for the least squares problem* (24)*. In doing so, the additional basis coefficients are spent on the relevant terms* $\boldsymbol{v} \in V$ *only. Note, however, that* $M \ge |\mathcal{I}(V)|$ *only guarantees that the truncated least squares problem is not underdetermined. The error bound of Theorem 2.6 below requires the strictly stronger sampling condition* $|\mathcal{I}(V)| \le \frac{1}{10\cdot 2^{\alpha}(r+1)} \cdot \frac{M}{\log M}$*. If the bandwidths are enlarged beyond this condition, the resulting approximation* $f^{\star}$ *is still computable, but it is no longer covered by the theory.*

Moreover, it is also possible to reiterate steps 3–5 of Algorithm 2.4 until no further truncation of the ANOVA decomposition can be achieved, i. e., until the set V *in* (22) *no longer changes.*

2.3. **Error Analysis**

In this part, we recall the approximation error of the ANOVA approximation, restricting our discussion to the Chebyshev basis (6) for the sake of clarity of presentation. In this case, we define the Sobolev-type space of dominating mixed smoothness

$$H^s_{\mathrm{mix}}([-1,1]^d) \coloneqq \left\{f \colon [-1,1]^d \to \mathbb{R} \mid \|f\|^2_{H^s_{\mathrm{mix}}([-1,1]^d)} < \infty\right\}, \quad s > 0$$

with the norm

$$\|f\|^2_{H^s_{\mathrm{mix}}([-1,1]^d)} \coloneqq \sum_{\boldsymbol{k}\in\mathbb{N}_0^d} |c_{\boldsymbol{k}}(f)|^2 \prod_{j=1}^{d} \left(1 + |k_j|^2\right)^s, \tag{26}$$

where the coefficients $c_{\boldsymbol{k}}(f)$ are given as in (5), and we refer the reader to [1, Corollary 2.2] and [10, p. 9] for details.

The following result, which we recall from [28], yields a worst-case bound for the error of the ANOVA approximation in Algorithm 2.4.

Theorem 2.6 ([28, Theorem 4.10])**.** *Let $M > 3$ sample points drawn i.i.d. w.r.t. the Chebyshev density ω^{cheb} in (6) be given such that for the index set $\mathcal{I}(V)$ in (23) we have*

$$|\mathcal{I}(V)| \leq \frac{M}{10 \cdot 2^{\alpha}(r+1)\log M}$$

for some $r > 0$ and $\alpha \coloneqq \max_{\boldsymbol{k}\in\mathcal{I}(V)} |\operatorname{supp}\boldsymbol{k}|$. Then for all $f \in H^s_{\mathrm{mix}}([-1,1]^d)$, $s > \frac{1}{2}$, the solution $f^\star$ in (25) of the least squares problem (24) satisfies the estimate

$$\|f - f^\star\|^2_{L_2([-1,1]^d,\omega^{\mathrm{cheb}})} \leq 5\|f\|^2_{H^s_{\mathrm{mix}}([-1,1]^d)} \cdot \max\left\{ \frac{2^d(1+\sqrt{5})^2}{2^\alpha \cdot 5|\mathcal{I}(V)|} \sum_{\boldsymbol{k}\in\mathbb{N}_0^d\setminus\mathcal{I}(V)} \frac{1}{\beta(\boldsymbol{k})}, \frac{1}{\inf_{\boldsymbol{k}\in\mathbb{N}_0^d\setminus\mathcal{I}(V)}\beta(\boldsymbol{k})} \right\}$$

with probability at least $1 - (2^{3/4}+1)M^{-r}$, where

$$\beta(\boldsymbol{k}) \coloneqq \prod_{j=1}^{d} \left(1 + |k_j|^2\right)^s \tag{27}$$

denotes the weight function of $H^s_{\mathrm{mix}}([-1,1]^d)$, cf. (26).

The error bound of Theorem 2.6 is governed by the chosen index set $\mathcal{I}(V)$ and holds with a probability depending on the coupled parameters r and M. Before relating it to the individual approximation steps, we point out a limitation in the high-dimensional regime. The first argument of the maximum carries the explicit factor 2^d, and the sum $\sum_{\boldsymbol{k}\notin\mathcal{I}(V)} 1/\beta(\boldsymbol{k})$ is exponential in d as well, since $\sum_{\boldsymbol{k}\in\mathbb{N}_0^d} 1/\beta(\boldsymbol{k}) = C_s^d$ with $C_s \coloneqq \sum_{n\in\mathbb{N}_0}(1+n^2)^{-s} > 1$. The second argument, in contrast, is free of any explicit dimension dependence. Theorem 2.6 is therefore informative for moderate d and large smoothness s (note $C_s \to 1$ as $s \to \infty$), while for large d it is a qualitative rather than a quantitative statement.

Remark 2.7 (Error terms and choice of the parameters)**.** *Writing $P_{\mathcal{I}(V)}$ for the orthogonal projection*

$$P_{\mathcal{I}(V)}f(\boldsymbol{x}) \coloneqq \sum_{\boldsymbol{k}\in\mathcal{I}(V)} \langle f, \varphi_{\boldsymbol{k}}\rangle_{L_2(\Omega,\omega)}\, \varphi_{\boldsymbol{k}}(\boldsymbol{x}) \tag{28}$$

onto $\operatorname{span}\{\varphi_{\boldsymbol{k}}\}_{\boldsymbol{k}\in\mathcal{I}(V)}$, cf. (17), the error of the approximation $f^\star$ from (25) splits as

$$\|f^\star - f\|^2_{L_2(\Omega,\omega)} = \underbrace{\|f^\star - P_{\mathcal{I}(V)}f\|^2_{L_2(\Omega,\omega)}}_{\text{estimation error}} + \underbrace{\|P_{\mathcal{I}(V)}f - f\|^2_{L_2(\Omega,\omega)}}_{\text{truncation and projection error}} .$$

Therein, equality holds by the theorem of Pythagoras. Both $f^\star$ and $P_{\mathcal{I}(V)}f$ lie in $\operatorname{span}\{\varphi_{\boldsymbol{k}}\}_{\boldsymbol{k}\in\mathcal{I}(V)}$, whereas $f - P_{\mathcal{I}(V)}f$ is orthogonal to that subspace, such that the two summands of $f^\star - f = (f^\star - P_{\mathcal{I}(V)}f) + (P_{\mathcal{I}(V)}f - f)$ are orthogonal. The two approximations differ only in how their coefficients are obtained, using the empirical problem (24) instead of the exact inner products. Hence, the first term measures the error introduced by the regression step, while the second term is the error of the best-approximation in the subspace.

Both sources appear in Theorem 2.6, though not as a trade-off between the two arguments of the maximum. By (26), the second argument is the best-approximation bound

$$\|f - P_{\mathcal{I}(V)}f\|^2_{L_2(\Omega,\omega)} \leq \|f\|^2_{H^s_{\mathrm{mix}}([-1,1]^d)} / \inf_{\boldsymbol{k}\notin\mathcal{I}(V)} \beta(\boldsymbol{k}),$$

whereas the first carries the constants of the sampling step. Both decrease monotonically as $\mathcal{I}(V)$ grows and tend to zero for $\mathcal{I}(V) \uparrow \mathbb{N}_0^d$, and neither contains M. The sample size enters only through the admissibility condition $|\mathcal{I}(V)| \leq M/(10 \cdot 2^\alpha (r+1)\log M)$. Consequently, enlarging $\mathcal{I}(V)$ improves the bound, and for a given M it is optimal to choose $\mathcal{I}(V)$ as large as that condition permits. The projection error itself vanishes for $\mathcal{I}(V) = \mathbb{N}_0^d$ and can be reduced either by admitting more basis coefficients per ANOVA term, cf. (15)*, or by admitting more ANOVA terms. For specific $\mathcal{I}(V)$, the two quantities of Theorem 2.6 are evaluated explicitly in [28, pp. 126–127]. The estimation error vanishes as $M \to \infty$ as due to the normal equations* (19) *and for i.i.d. samples, Monte-Carlo integration gives $\frac{1}{M}\boldsymbol{A}_V^\top \boldsymbol{A}_V \to \boldsymbol{I}$ and $\frac{1}{M}\boldsymbol{A}_V^\top \boldsymbol{f} \to (\langle f, \varphi_{\boldsymbol{k}}\rangle)_{\boldsymbol{k}\in\mathcal{I}(V)}$, so that* (19) *converges to the coefficients of* (28)*.*

For finite M, the parameters are coupled through the two constraints

$$c_1 \leq \frac{M}{(r+1)\log M}, \quad c_1 \coloneqq 10 \cdot 2^\alpha |\mathcal{I}(V)|, \tag{29}$$

$$1 - c_2 M^{-r} \geq p, \quad c_2 \coloneqq 2^{3/4} + 1, \tag{30}$$

for a prescribed target certainty $p \in (0,1)$. Here $r > 0$ is a purely theoretical parameter, while M is the actual sampling effort. Since the right-hand side of (29) *increases monotonically for $M > 3$, the smallest admissible M is optimal, so* (29) *may be assumed active. This gives $r = M/(c_1 \log M) - 1$ and, inserted into* (30)*, the condition $\log((1-p)/c_2) \geq \log M - M/c_1$ determining M. These constraints couple target accuracy (through c_1, since a smaller error bound requires a larger $\mathcal{I}(V)$), certainty p and sample size M. Fixing any one of them yields a monotone trade-off between the remaining two. For fixed c_1, for instance, the achievable certainty $p(M) = 1 - c_2\, M^{\,1-M/(c_1 \log M)}$ increases monotonically in M with $p(M) \to 1$. Conversely, for fixed M a larger $\mathcal{I}(V)$ lowers the error bound but requires a smaller r, which in turn leads to a smaller certainty.*

3. Control systems

In this section, we present our main contribution, i. e., a tailored and efficient ANOVA approximation for control-affine systems whose right-hand side exhibits a low-dimensional coupling structure. Such systems arise from spatial discretizations of controlled partial differential equations, where the locality of the differential operators induces sparse interactions between the state variables. Consequently, each component couples only with its spatial neighbours. We will make the ANOVA machinery of Section 2 applicable to this setting while leveraging the control-affine structure.

Control-affine dynamical systems are of the form

$$\boldsymbol{x}^+ = F(\boldsymbol{x}) + G(\boldsymbol{x})\boldsymbol{u}.$$

Therein, $\boldsymbol{x} \in \mathbb{R}^d$ is the state, the control $\boldsymbol{u}$ evolves in a compact control constraint set $\mathbb{U} \subset \mathbb{R}^m$, and the maps $F\colon \mathbb{R}^d \to \mathbb{R}^d$ and $G\colon \mathbb{R}^d \to \mathbb{R}^{d\times m}$ with $d, m \in \mathbb{N}$ describe the drift and the input, respectively. Note that the number of actuators is usually much smaller than the number of states, i. e., $m \ll d$. We now define $H\colon \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^d$ via

$$H(\boldsymbol{x}, \boldsymbol{u}) \coloneqq \begin{bmatrix} F(\boldsymbol{x}) & G(\boldsymbol{x}) \end{bmatrix} \begin{bmatrix} 1 \\ \boldsymbol{u} \end{bmatrix} = F(\boldsymbol{x}) + G(\boldsymbol{x})\boldsymbol{u} \tag{31}$$

and assume that F and G are of low-dimensional structure. More precisely, we require that, for suitable index sets, we have

$$F_j(\boldsymbol{x}) = \sum_{\boldsymbol{v}\in V_{j,0}} f_{\boldsymbol{v}}^j(\boldsymbol{x}_{\boldsymbol{v}}), \qquad G_{j\ell}(\boldsymbol{x}) = \sum_{\boldsymbol{v}\in V_{j,\ell}} g_{\boldsymbol{v}}^{j\ell}(\boldsymbol{x}_{\boldsymbol{v}}), \quad j = 1,\dots,d,\ \ell = 1,\dots,m, \tag{32}$$

where we denote the functions F and G entry-wise as $F(\boldsymbol{x}) = (F_j(\boldsymbol{x}))_{j=1}^{d}$ and $G(\boldsymbol{x}) = (G_{j\ell}(\boldsymbol{x}))_{j=1,\ell=1}^{d,\,m}$, with $V_{j,\ell} \subset 2^{[d]}$ being index sets with small cardinality for all $j = 1, \dots, d$, and $\ell = 0, \dots, m$. Moreover, we denote by $G_\ell \coloneqq (G_{j\ell})_{j=1}^{d}$ the ℓ-th column of G, i. e., the ℓ-th control vector field. Our task is then to reconstruct this structure from given samples. We illustrate the class of systems (31) with two examples, namely a small benchmark oscillator and a higher-dimensional system stemming from a discretized partial differential equation.

Example 3.1. *Consider the Euler approximation of a controlled Duffing oscillator*

$$\boldsymbol{x}^+ = \boldsymbol{x} + \Delta t \begin{pmatrix} x_2 \\ x_1 - 3x_1^3 u \end{pmatrix} = \begin{pmatrix} x_1 + \Delta t x_2 \\ x_2 + \Delta t(x_1 - 3x_1^3 u) \end{pmatrix} \tag{33}$$

which corresponds to the system (31) *with* $d = 2$, $m = 1$, *and*

$$F(\boldsymbol{x}) = \begin{pmatrix} x_1 + \Delta t x_2 \\ x_2 + \Delta t x_1 \end{pmatrix}, \qquad G(\boldsymbol{x}) = \begin{pmatrix} 0 \\ -3\Delta t x_1^3 \end{pmatrix}.$$

Clearly, these functions F and G are of low-dimensional structure compared to the small state space dimension of this problem, since only one-dimensional terms occur, but no constant or two-dimensional terms. In the notation of (32) *we have the index sets $V_{1,0} = V_{2,0} = \{\{1\}, \{2\}\}$, $V_{1,1} = \varnothing$, as $G_{11} \equiv 0$ contributes no terms at all, and $V_{2,1} = \{\{1\}\}$.*

The next example is higher-dimensional and closer to the applications targeted in Section 4, where we will illustrate the proposed method on discretizations of partial differential equations, which naturally lead to high-dimensional problems with a low-dimensional coupling structure.

Example 3.2 (Controlled viscous Burgers equation)**.** *Consider the controlled viscous Burgers equation*

$$\frac{\partial}{\partial t} x = D\,\frac{\partial^2}{\partial \omega^2} x - x\,\frac{\partial}{\partial \omega} x + \gamma(\omega)\, x\, U_1 + b(\omega)\, U_2, \qquad \omega \in (0, L),\ t > 0,$$

with viscosity $D > 0$ and two scalar controls $\boldsymbol{u} = (U_1, U_2)^\top$ acting through nonnegative actuator profiles $\gamma, b : (0, L) \to \mathbb{R}$. We discretize the spatial domain by an equidistant grid $\omega_i \coloneqq i\,\Delta\omega$, $i = 1, \dots, d$, with grid size $\Delta\omega \coloneqq L/d > 0$. All quantities indexed by i denote evaluations at the grid point ω_i, i. e., $X_i \approx x(\omega_i, \cdot)$, $\gamma_i \coloneqq \gamma(\omega_i)$, and $b_i \coloneqq b(\omega_i)$. Using a finite-difference approximation of the spatial derivatives on this grid and an explicit Euler method with time step $\Delta t > 0$, we obtain a system of the form (31) *with $m = 2$ and*

$$\begin{aligned} F_i(X) &= X_i + \Delta t \left(D\,\frac{X_{i+1} - 2X_i + X_{i-1}}{\Delta\omega^2} - X_i\,\frac{X_{i+1} - X_{i-1}}{2\Delta\omega} \right), \\ G_{i,1}(X) &= \Delta t\, \gamma_i\, X_i, \qquad G_{i,2}(X) = \Delta t\, b_i, \qquad i = 1, \dots, d, \end{aligned}$$

at each spatial point i. Here, we understand the indices cyclically, i. e., $X_0 \coloneqq X_d$ and $X_{d+1} \coloneqq X_1$, corresponding to periodic boundary conditions. Again, F and G are of low-dimensional structure. The diffusion term contributes the one-dimensional terms of node i and of its two neighbours, while the advection term contributes the two adjacent pair interactions. The input map is state-affine in the first control and constant in the second. In the notation of (32) *this yields, for every $i = 1, \dots, d$, the index sets*

$$V_{i,0} = \{\{i-1\}, \{i\}, \{i+1\}, \{i-1, i\}, \{i, i+1\}\}, \quad V_{i,1} = \{\{i\}\}, \quad V_{i,2} = \{\varnothing\},$$

where $V_{i,1} = \varnothing$ for $\gamma_i = 0$ and $V_{i,2} = \varnothing$ for $b_i = 0$. Hence, although the number of discretization points and hence the dimension d may be large, every interaction set satisfies $|\boldsymbol{v}| \le 2$. We remark this low-dimensional structure is also present in the context of other boundary conditions, e. g., of Dirichlet or Neumann type, only modify the stencil at the two boundary nodes $i \in \{1, d\}$.

In general, the vector field F and the columns of G cannot be sampled separately, such that an off-the-shelf application of the ANOVA approximation of Section 2 to the individual functions in (32) is not possible. Therefore, we aim for a coupled reconstruction of the dynamics map (31). However, applying the ANOVA approximation as a black box to this coupled map does not leverage the structure, since ignoring the control-affine structure would spend considerable computational effort on interactions of the control components that are excluded a priori.

In many data-driven scenarios, one merely observes the system under some given excitation, i. e., one is provided with samples $(\boldsymbol{x}^i, \boldsymbol{u}^i, H(\boldsymbol{x}^i, \boldsymbol{u}^i))$. These pairs of samples correspond to state, the employed action, and the resulting successor state. In the following, we therefore develop an approach that works with i.i.d. sampling points $(\boldsymbol{x}^i, \boldsymbol{u}^i)$ drawn from a tensor-product density on $\Omega \times \mathbb{U}$ and computes an approximation by coupled least squares leveraging the control-affine structure. To this end, we assume that both the control domain $\mathbb{U} = \bigtimes_{\ell=1}^m \mathbb{U}_\ell$ and the sampling density $\widetilde{\omega} = \prod_{\ell=1}^m \omega_{d+\ell}$ are of tensor-product structure. The main idea is to encode the control-affine structure of (31) directly into the approximation ansatz. To this end, we employ a dictionary consisting of basis functions in $\boldsymbol{x}$ only, augmented by the same basis functions multiplied by the components u_ℓ of the control. Thereby, the ansatz is affine linear in $\boldsymbol{u}$ by construction, and no spurious interactions between the components of $\boldsymbol{u}$ have to be computed. We first discuss the case of a univariate control in Section 3.1, including the resulting coupled least squares problem and its interpretation as an orthogonal projection. In Section 3.2, we prove an error bound for finitely many samples in the spirit of the bound for uncontrolled scalar functions recalled in Theorem 2.6. Afterwards, we generalize the ansatz and the error bound to multivariate controls in Section 3.3 and refer the reader to Appendix A for an implementation of the resulting approximation via preprocessed ANOVA index sets.

3.1. **Coupled least squares approximation for univariate control**

We first examine the case $m = 1$ in (31), i. e., single-input systems. In this case, the system to be approximated reads

$$\begin{pmatrix} H_1(\boldsymbol{x}, u) \\ \vdots \\ H_d(\boldsymbol{x}, u) \end{pmatrix} = H(\boldsymbol{x}, u) = F(\boldsymbol{x}) + G(\boldsymbol{x})u, \quad u \in \mathbb{R}.$$

Since the ANOVA decomposition of Section 2 is defined for scalar-valued functions, we proceed rowwise, i. e., our objective is to approximate each row of (31) separately by means of the ANOVA approximation of Section 2.2. Consistent with the structure of the map H, we choose an ansatz of the form

$$H_j(\boldsymbol{x}, u) \approx \sum_{\boldsymbol{k} \in \mathcal{I}(V_1)} c^1_{\boldsymbol{k}} \varphi^1_{\boldsymbol{k}}(\boldsymbol{x}) + \sum_{\boldsymbol{s} \in \mathcal{I}(V_2)} c^2_{\boldsymbol{s}} \varphi^2_{\boldsymbol{s}}(\boldsymbol{x}) u, \quad j = 1, \dots, d, \tag{34}$$

where $\{\varphi^1_{\boldsymbol{k}}\}_{\boldsymbol{k} \in \mathcal{I}(V_1)}$ and $\{\varphi^2_{\boldsymbol{s}}\}_{\boldsymbol{s} \in \mathcal{I}(V_2)}$ are two possibly different orthonormal systems. To compute the coefficients, we sample system (31) at i.i.d. data points $(\boldsymbol{x}^i, u^i) \in \Omega \times \mathbb{U} \subset \mathbb{R}^{d+1}$, $i = 1, \dots, M$, drawn according to the tensor product density $(\omega \cdot \omega_{d+1})(\boldsymbol{x}, u) \coloneqq \omega(\boldsymbol{x})\, \omega_{d+1}(u)$ with $M \in \mathbb{N}$ and ω_{d+1} being the density corresponding to $\mathbb{U} \subset \mathbb{R}$. Throughout this section, we abbreviate the first two moments of the control distribution by

$$\eta \coloneqq \langle 1, u \rangle_{L_2(\mathbb{U}, \omega_{d+1})} = \int_{\mathbb{U}} u\, \omega_{d+1}(u)\, \mathrm{d}u, \qquad \nu \coloneqq \|u\|^2_{L_2(\mathbb{U}, \omega_{d+1})} = \int_{\mathbb{U}} u^2\, \omega_{d+1}(u)\, \mathrm{d}u. \tag{35}$$

Here, η is the mean and ν the second moment of a control drawn according to ω_{d+1}, such that $\nu - \eta^2$ is its variance. Note that $\|1\|^2_{L_2(\mathbb{U}, \omega_{d+1})} = \int_{\mathbb{U}} \omega_{d+1}(u)\, \mathrm{d}u = 1$, since ω_{d+1} is a probability density, such that expanding the square yields

$$0 \leq \|u - \eta\|^2_{L_2(\mathbb{U}, \omega_{d+1})} = \nu - 2\eta \langle 1, u \rangle_{L_2(\mathbb{U}, \omega_{d+1})} + \eta^2 \|1\|^2_{L_2(\mathbb{U}, \omega_{d+1})} = \nu - \eta^2. \tag{36}$$

We therefore have $\eta^2 \leq \nu$, and equality holds if and only if $\|u - \eta\|_{L_2(\mathbb{U},\omega_{d+1})} = 0$, i. e., if and only if u equals its mean η ω_{d+1}-almost everywhere. Negating this equality case shows that the following three conditions are equivalent:

(i) The control u is not ω_{d+1}-almost everywhere constant.
(ii) The variance $\nu - \eta^2$ of the control distribution is positive.
(iii) The control u attains at least two distinct values on sets of positive measure.

Indeed, (i) is the negation of the equality case just discussed, (ii) is equivalent to (i) by (36), and (iii) restates (i) for a real-valued random variable. We call a control distribution with these properties *non-degenerate*. For later reference, we state this property as an assumption.

Assumption 3.3 (Non-degenerate control distribution)**.** *The control distribution is non-degenerate, i. e., $\nu - \eta^2 > 0$ with η and ν from* (35).

This assumption will be used repeatedly below. It guarantees that the control samples carry enough information to separate the contributions of F_j and G_j, whereas these contributions are indistinguishable for an almost everywhere constant control.

For each $j = 1, \dots, d$, we define the vectors $\boldsymbol{h} \coloneqq \big(H_j(\boldsymbol{x}^i, u^i)\big)_{i=1}^M \in \mathbb{R}^M$, $\boldsymbol{c}^1 \coloneqq \big(c^1_{\boldsymbol{k}}\big)_{\boldsymbol{k}\in\mathcal{I}(V_1)} \in \mathbb{R}^{|\mathcal{I}(V_1)|}$, and $\boldsymbol{c}^2 \coloneqq \big(c^2_{\boldsymbol{s}}\big)_{\boldsymbol{s}\in\mathcal{I}(V_2)} \in \mathbb{R}^{|\mathcal{I}(V_2)|}$, as well as the matrices

$$\boldsymbol{A}_1 \coloneqq \big(\varphi^1_{\boldsymbol{k}}(\boldsymbol{x}^i)\big)_{i=1,\,\boldsymbol{k}\in\mathcal{I}(V_1)}^M \in \mathbb{R}^{M\times|\mathcal{I}(V_1)|}, \qquad \boldsymbol{A}_2 \coloneqq \big(\varphi^2_{\boldsymbol{s}}(\boldsymbol{x}^i)\big)_{i=1,\,\boldsymbol{s}\in\mathcal{I}(V_2)}^M \in \mathbb{R}^{M\times|\mathcal{I}(V_2)|}, \tag{37}$$

and

$$\boldsymbol{U} \coloneqq \operatorname{diag}(u^1, \dots, u^M) \in \mathbb{R}^{M\times M}. \tag{38}$$

Then, for fixed $j = 1, \dots, d$, the approximation (34) can be written in matrix-vector notation as

$$\boldsymbol{h} \approx \begin{pmatrix}\boldsymbol{A}_1 & \boldsymbol{U}\boldsymbol{A}_2\end{pmatrix}\begin{pmatrix}\boldsymbol{c}^1\\ \boldsymbol{c}^2\end{pmatrix}. \tag{39}$$

The coefficients $\boldsymbol{c}^1, \boldsymbol{c}^2$ are then computed analogously to (18) by solving the least squares problem

$$\min_{(\tilde{\boldsymbol{c}}^1,\tilde{\boldsymbol{c}}^2)^\top\in\mathbb{R}^{|\mathcal{I}(V_1)|+|\mathcal{I}(V_2)|}} \left\| \begin{pmatrix}\boldsymbol{A}_1 & \boldsymbol{U}\boldsymbol{A}_2\end{pmatrix}\begin{pmatrix}\tilde{\boldsymbol{c}}^1\\ \tilde{\boldsymbol{c}}^2\end{pmatrix} - \boldsymbol{h}\right\|_2^2. \tag{40}$$

For the least squares solution to be computable by the normal equations

$$\begin{pmatrix}\boldsymbol{A}_1^\top\\ \boldsymbol{A}_2^\top\boldsymbol{U}\end{pmatrix}\begin{pmatrix}\boldsymbol{A}_1 & \boldsymbol{U}\boldsymbol{A}_2\end{pmatrix}\begin{pmatrix}\tilde{\boldsymbol{c}}^1\\ \tilde{\boldsymbol{c}}^2\end{pmatrix} = \begin{pmatrix}\boldsymbol{A}_1^\top\\ \boldsymbol{A}_2^\top\boldsymbol{U}\end{pmatrix}\boldsymbol{h}, \tag{41}$$

we need to ensure that the system matrix $\begin{pmatrix}\boldsymbol{A}_1 & \boldsymbol{U}\boldsymbol{A}_2\end{pmatrix} \in \mathbb{R}^{M\times(|\mathcal{I}(V_1)|+|\mathcal{I}(V_2)|)}$ has full column rank. To this end, we necessarily require $M \geq |\mathcal{I}(V_1)| + |\mathcal{I}(V_2)|$, i. e., sufficiently many sample points $\boldsymbol{x}^i$, $i = 1, \dots, M$. In addition, we may assume that $\boldsymbol{A}_1, \boldsymbol{A}_2$ have full column rank, see Section 2.2. Whether the composed matrix has full column rank depends on the sampled control values u^i and thus on their underlying sampling distribution. For instance, for $\varphi^1_{\boldsymbol{0}} = \varphi^2_{\boldsymbol{0}} \equiv 1$ and constant control samples $u^1 = \cdots = u^M = c \neq 0$ the matrix $\begin{pmatrix}\boldsymbol{A}_1 & \boldsymbol{U}\boldsymbol{A}_2\end{pmatrix} = \begin{pmatrix}\boldsymbol{1} & c\boldsymbol{1}\end{pmatrix}$ is rank deficient. However, by Proposition 3.4 below the scaled Gram matrix converges to a positive-definite limit under Assumption 3.3. Thus, for almost every realization the system matrix has full column rank for all sufficiently large M. A quantitative version of this statement, which holds with high probability, is given in Proposition 3.7 below.

Before turning to the error analysis, we clarify the main objective of the coupled least squares problem (40). To this end, we consider the infinite-data limit. Roughly speaking and after scaling by $\frac{1}{M}$, the entries of the normal equations (41) are Monte-Carlo estimates of the corresponding inner products in $L_2(\Omega \times \mathbb{U}, \omega \cdot \omega_{d+1})$. Hence, the scaled discrete system converges to a continuous one as the number M of samples grows. The following proposition identifies this limit system as the one characterizing the orthogonal projection of H_j onto the span of a particular dictionary corresponding to the ansatz space of (34). It also states that this system is uniquely solvable and shows that the coefficients computed from finitely many samples converge to those of this projection. Consequently, the approach of this section is an empirical least squares approximation with respect to this dictionary. This interpretation is used throughout the remainder of the section. It may be used as a guidance for the choice of the control density (see Remark 3.5), allows the comparison with a separate ANOVA approximation of F and G in Remark 3.6, and leads to the multivariate analogue in Section 3.3.

Proposition 3.4. *Let $(\boldsymbol{x}^i, u^i)$, $i = 1, \dots, M$, be i.i.d. samples drawn according to the density $\omega \cdot \omega_{d+1}$, let Assumption 3.3 hold, and let F_j, $G_j \in L_2(\Omega, \omega)$. Then, abbreviating by*

$$\widetilde{\mathcal{D}} = \{\varphi^1_{\boldsymbol{k}} \colon \boldsymbol{k} \in \mathcal{I}(V_1)\} \cup \{\varphi^2_{\boldsymbol{s}} u \colon \boldsymbol{s} \in \mathcal{I}(V_2)\}$$

the dictionary of the ansatz (34), the following assertions hold.

- *(i) After scaling by $\frac{1}{M}$, the normal equations (41) converge for $M \to \infty$, entrywise and almost surely, to a limit system.*
- *(ii) This limit system is the Gram system of $\widetilde{\mathcal{D}}$ in $L_2(\Omega \times \mathbb{U}, \omega \cdot \omega_{d+1})$. It is uniquely solvable, and its solution consists of the coefficients of the orthogonal projection $P_{\widetilde{\mathcal{D}}} H_j$ of H_j onto $\operatorname{span} \widetilde{\mathcal{D}}$.*
- *(iii) For almost every realization of the samples, the scaled normal equations are uniquely solvable for all sufficiently large M, and their solutions converge to the coefficients of $P_{\widetilde{\mathcal{D}}} H_j$.*

Proof. We prove the three assertions in the order in which they are stated.

Convergence of the sampled inner products. The normal equations (41) may also be written as

$$\begin{pmatrix} \boldsymbol{A}_1^\top \boldsymbol{A}_1 & \boldsymbol{A}_1^\top \boldsymbol{U} \boldsymbol{A}_2 \\ \boldsymbol{A}_2^\top \boldsymbol{U} \boldsymbol{A}_1 & \boldsymbol{A}_2^\top \boldsymbol{U}^2 \boldsymbol{A}_2 \end{pmatrix} \begin{pmatrix} \tilde{\boldsymbol{c}}^1 \\ \tilde{\boldsymbol{c}}^2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{A}_1^\top \boldsymbol{h} \\ \boldsymbol{A}_2^\top \boldsymbol{U} \boldsymbol{h} \end{pmatrix}. \tag{42}$$

Then the strong law of large numbers yields for i.i.d. data points $(\boldsymbol{x}^i, u^i) \in \Omega \times \mathbb{U} \subset \mathbb{R}^{d+1}$, $i = 1, \dots, M$, drawn according to the density $\omega \cdot \omega_{d+1}$ that, after scaling, the entries of the block matrix converge almost surely to

$$\lim_{M \to \infty} \frac{1}{M} [\boldsymbol{A}_1^\top \boldsymbol{A}_1]_{\boldsymbol{k},\boldsymbol{n}} = \lim_{M \to \infty} \frac{1}{M} \sum_{i=1}^{M} \varphi^1_{\boldsymbol{k}}(\boldsymbol{x}^i) \varphi^1_{\boldsymbol{n}}(\boldsymbol{x}^i) = \langle \varphi^1_{\boldsymbol{k}}, \varphi^1_{\boldsymbol{n}} \rangle_{L_2(\Omega,\omega)} = \delta_{\boldsymbol{k},\boldsymbol{n}}, \quad \boldsymbol{k}, \boldsymbol{n} \in \mathcal{I}(V_1),$$

$$\lim_{M \to \infty} \frac{1}{M} [\boldsymbol{A}_1^\top \boldsymbol{U} \boldsymbol{A}_2]_{\boldsymbol{k},\boldsymbol{s}} = \lim_{M \to \infty} \frac{1}{M} \sum_{i=1}^{M} \varphi^1_{\boldsymbol{k}}(\boldsymbol{x}^i) u^i \varphi^2_{\boldsymbol{s}}(\boldsymbol{x}^i) = \langle \varphi^1_{\boldsymbol{k}}, u \varphi^2_{\boldsymbol{s}} \rangle_{L_2(\Omega \times \mathbb{U}, \omega \cdot \omega_{d+1})}, \quad \boldsymbol{k} \in \mathcal{I}(V_1), \boldsymbol{s} \in \mathcal{I}(V_2),$$

$$\lim_{M \to \infty} \frac{1}{M} [\boldsymbol{A}_2^\top \boldsymbol{U}^2 \boldsymbol{A}_2]_{\boldsymbol{r},\boldsymbol{s}} = \lim_{M \to \infty} \frac{1}{M} \sum_{i=1}^{M} u^i \varphi^2_{\boldsymbol{r}}(\boldsymbol{x}^i) u^i \varphi^2_{\boldsymbol{s}}(\boldsymbol{x}^i) = \langle u \varphi^2_{\boldsymbol{r}}, u \varphi^2_{\boldsymbol{s}} \rangle_{L_2(\Omega \times \mathbb{U}, \omega \cdot \omega_{d+1})}, \quad \boldsymbol{r}, \boldsymbol{s} \in \mathcal{I}(V_2).$$

Here, the strong law of large numbers is applicable since all involved random variables are integrable. Indeed, the products of two basis functions are integrable by the Cauchy–Schwarz inequality, the additional factors u^i are bounded by the compactness of $\mathbb{U}$, and $H_j \in L_2(\Omega \times \mathbb{U}, \omega \cdot \omega_{d+1})$. Moreover, all integrands appearing above are absolutely integrable. By Tonelli's theorem, the Cauchy–Schwarz inequality and the orthonormality of the two systems we have

$$\int_{\Omega \times \mathbb{U}} |\varphi^1_{\boldsymbol{k}}(\boldsymbol{x}) \varphi^2_{\boldsymbol{s}}(\boldsymbol{x})\, u| \, (\omega \cdot \omega_{d+1})(\boldsymbol{x}, u) \, \mathrm{d}(\boldsymbol{x}, u) \leq \|\varphi^1_{\boldsymbol{k}}\|_{L_2(\Omega,\omega)} \|\varphi^2_{\boldsymbol{s}}\|_{L_2(\Omega,\omega)} \max_{u \in \mathbb{U}} |u| < \infty.$$

The remaining integrands can be treated analogously, where we used the compactness of $\mathbb{U}$. Hence, Fubini's theorem yields

$$\begin{aligned}
\langle \varphi^1_{\boldsymbol{k}}, u\varphi^2_{\boldsymbol{s}} \rangle_{L_2(\Omega\times\mathbb{U},\omega\cdot\omega_{d+1})} &= \int_{\Omega\times\mathbb{U}} \varphi^1_{\boldsymbol{k}}(\boldsymbol{x})\varphi^2_{\boldsymbol{s}}(\boldsymbol{x}) u \,(\omega\cdot\omega_{d+1})(\boldsymbol{x},u)\,\mathrm{d}(\boldsymbol{x},u) \\
&= \int_\Omega \int_{\mathbb{U}} \varphi^1_{\boldsymbol{k}}(\boldsymbol{x})\varphi^2_{\boldsymbol{s}}(\boldsymbol{x}) u \,\omega(\boldsymbol{x})\,\omega_{d+1}(u)\,\mathrm{d}u\,\mathrm{d}\boldsymbol{x} \\
&= \langle \varphi^1_{\boldsymbol{k}}, \varphi^2_{\boldsymbol{s}} \rangle_{L_2(\Omega,\omega)} \langle 1, u \rangle_{L_2(\mathbb{U},\omega_{d+1})}, \qquad (43) \\
\langle u\varphi^2_{\boldsymbol{r}}, u\varphi^2_{\boldsymbol{s}} \rangle_{L_2(\Omega\times\mathbb{U},\omega\cdot\omega_{d+1})} &= \int_{\Omega\times\mathbb{U}} \varphi^2_{\boldsymbol{r}}(\boldsymbol{x})\varphi^2_{\boldsymbol{s}}(\boldsymbol{x}) u^2 \,(\omega\cdot\omega_{d+1})(\boldsymbol{x},u)\,\mathrm{d}(\boldsymbol{x},u) \\
&= \int_\Omega \int_{\mathbb{U}} \varphi^2_{\boldsymbol{r}}(\boldsymbol{x})\varphi^2_{\boldsymbol{s}}(\boldsymbol{x}) u^2 \,\omega(\boldsymbol{x})\,\omega_{d+1}(u)\,\mathrm{d}u\,\mathrm{d}\boldsymbol{x} \\
&= \langle \varphi^2_{\boldsymbol{r}}, \varphi^2_{\boldsymbol{s}} \rangle_{L_2(\Omega,\omega)} \|u\|^2_{L_2(\mathbb{U},\omega_{d+1})}. \qquad (44)
\end{aligned}$$

For the right-hand side we obtain similarly, now using in addition that $F_j, G_j \in L_2(\Omega,\omega)$,

$$\begin{aligned}
\lim_{M\to\infty} \frac{1}{M} [\boldsymbol{A}_1^\top \boldsymbol{h}]_{\boldsymbol{k}} &= \lim_{M\to\infty} \frac{1}{M} \sum_{i=1}^M \varphi^1_{\boldsymbol{k}}(\boldsymbol{x}^i) H_j(\boldsymbol{x}^i, u^i) = \langle H_j, \varphi^1_{\boldsymbol{k}} \rangle_{L_2(\Omega\times\mathbb{U},\omega\cdot\omega_{d+1})} \\
&= \langle F_j, \varphi^1_{\boldsymbol{k}} \rangle_{L_2(\Omega,\omega)} + \langle G_j, \varphi^1_{\boldsymbol{k}} \rangle_{L_2(\Omega,\omega)} \langle 1, u \rangle_{L_2(\mathbb{U},\omega_{d+1})}
\end{aligned}$$

as well as

$$\begin{aligned}
\lim_{M\to\infty} \frac{1}{M} [\boldsymbol{A}_2^\top \boldsymbol{U}\boldsymbol{h}]_{\boldsymbol{s}} &= \lim_{M\to\infty} \frac{1}{M} \sum_{i=1}^M u^i \varphi^2_{\boldsymbol{s}}(\boldsymbol{x}^i) H_j(\boldsymbol{x}^i, u^i) = \langle H_j, u\varphi^2_{\boldsymbol{s}} \rangle_{L_2(\Omega\times\mathbb{U},\omega\cdot\omega_{d+1})} \\
&= \langle F_j, \varphi^2_{\boldsymbol{s}} \rangle_{L_2(\Omega,\omega)} \langle 1, u \rangle_{L_2(\mathbb{U},\omega_{d+1})} + \langle G_j, \varphi^2_{\boldsymbol{s}} \rangle_{L_2(\Omega,\omega)} \|u\|^2_{L_2(\mathbb{U},\omega_{d+1})}.
\end{aligned}$$

Thus, by (43) and (44), for $M \to \infty$ the linear system (42) converges to

$$\begin{aligned}
&\begin{pmatrix} \boldsymbol{I}_{|\mathcal{I}(V_1)|} & \langle 1, u \rangle_{L_2(\mathbb{U},\omega_{d+1})} \cdot \boldsymbol{\Phi} \\ \langle 1, u \rangle_{L_2(\mathbb{U},\omega_{d+1})} \cdot \boldsymbol{\Phi}^\top & \|u\|^2_{L_2(\mathbb{U},\omega_{d+1})} \cdot \boldsymbol{I}_{|\mathcal{I}(V_2)|} \end{pmatrix} \begin{pmatrix} \tilde{\boldsymbol{c}}^1 \\ \tilde{\boldsymbol{c}}^2 \end{pmatrix} \\
&\quad = \begin{pmatrix} \big(\langle F_j, \varphi^1_{\boldsymbol{k}} \rangle_{L_2(\Omega,\omega)} + \langle G_j, \varphi^1_{\boldsymbol{k}} \rangle_{L_2(\Omega,\omega)} \langle 1, u \rangle_{L_2(\mathbb{U},\omega_{d+1})} \big)_{\boldsymbol{k}\in\mathcal{I}(V_1)} \\ \Big(\langle F_j, \varphi^2_{\boldsymbol{s}} \rangle_{L_2(\Omega,\omega)} \langle 1, u \rangle_{L_2(\mathbb{U},\omega_{d+1})} + \langle G_j, \varphi^2_{\boldsymbol{s}} \rangle_{L_2(\Omega,\omega)} \|u\|^2_{L_2(\mathbb{U},\omega_{d+1})} \Big)_{\boldsymbol{s}\in\mathcal{I}(V_2)} \end{pmatrix} \qquad (45)
\end{aligned}$$

with $\boldsymbol{\Phi} \coloneqq \big(\langle \varphi^1_{\boldsymbol{k}}, \varphi^2_{\boldsymbol{s}} \rangle_{L_2(\Omega,\omega)} \big)_{\boldsymbol{k}\in\mathcal{I}(V_1),\, \boldsymbol{s}\in\mathcal{I}(V_2)}$, which shows (i).

Identification with the orthogonal projection. We consider the dictionary

$$\widetilde{\mathcal{D}} \coloneqq \{\psi_{\boldsymbol{n}} \colon \boldsymbol{n} \in \mathcal{I}\} = \{\varphi^1_{\boldsymbol{k}} \colon \boldsymbol{k} \in \mathcal{I}(V_1)\} \cup \{\varphi^2_{\boldsymbol{s}} u \colon \boldsymbol{s} \in \mathcal{I}(V_2)\}, \quad \mathcal{I} \coloneqq \mathcal{I}(V_1) \cup \mathcal{I}(V_2), \qquad (46)$$

where the unions are to be understood as disjoint unions, since an index $\boldsymbol{k} \in \mathcal{I}(V_1) \cap \mathcal{I}(V_2)$ refers to the two different functions $\varphi^1_{\boldsymbol{k}}$ and $\varphi^2_{\boldsymbol{k}} u$. The corresponding orthogonal projection of a function $f \in L_2(\Omega\times\mathbb{U}, \omega\cdot\omega_{d+1})$ onto $\operatorname{span}\widetilde{\mathcal{D}}$ is given by

$$P_{\widetilde{\mathcal{D}}} f(\boldsymbol{x}, u) \coloneqq \sum_{\boldsymbol{n}\in\mathcal{I}} c_{\boldsymbol{n}} \psi_{\boldsymbol{n}}(\boldsymbol{x}, u),$$

where the coefficients are given as the solution of the linear system

$$(\langle \psi_{\boldsymbol{p}}, \psi_{\boldsymbol{n}} \rangle_{L_2(\Omega\times\mathbb{U},\omega\cdot\omega_{d+1})})_{\boldsymbol{p},\boldsymbol{n}\in\mathcal{I}} \cdot (c_{\boldsymbol{n}})_{\boldsymbol{n}\in\mathcal{I}} = (\langle f, \psi_{\boldsymbol{p}} \rangle_{L_2(\Omega\times\mathbb{U},\omega\cdot\omega_{d+1})})_{\boldsymbol{p}\in\mathcal{I}}. \tag{47}$$

By the structure of $\widetilde{\mathcal{D}}$, every element of $\operatorname{span}\widetilde{\mathcal{D}}$ is of the form $\sum_{\boldsymbol{k}\in\mathcal{I}(V_1)} c^1_{\boldsymbol{k}}\varphi^1_{\boldsymbol{k}}(\boldsymbol{x}) + \sum_{\boldsymbol{s}\in\mathcal{I}(V_2)} c^2_{\boldsymbol{s}}\varphi^2_{\boldsymbol{s}}(\boldsymbol{x})\, u$, i. e., precisely of the form of our approximation (34). Moreover, ordering the dictionary according to this splitting and choosing $f = H_j$, the Gram system (47) coincides with the limit system (45), since its entries $\langle \psi_{\boldsymbol{p}}, \psi_{\boldsymbol{n}} \rangle$ and $\langle H_j, \psi_{\boldsymbol{p}} \rangle$ are the limits computed in the first step. For functions not depending on u, the inner products of $L_2(\Omega\times\mathbb{U}, \omega\cdot\omega_{d+1})$ and of $L_2(\Omega,\omega)$ agree, since ω_{d+1} is a probability density. Recall that a Gram matrix is positive definite if and only if the underlying family is linearly independent. It therefore remains to verify the latter, which holds by Assumption 3.3. If

$$\sum_{\boldsymbol{k}\in\mathcal{I}(V_1)} a_{\boldsymbol{k}}\varphi^1_{\boldsymbol{k}}(\boldsymbol{x}) + (\sum_{\boldsymbol{s}\in\mathcal{I}(V_2)} b_{\boldsymbol{s}}\varphi^2_{\boldsymbol{s}}(\boldsymbol{x}))u = 0$$

for $(\omega\cdot\omega_{d+1})$-almost all $(\boldsymbol{x}, u)$, then, since u attains at least two distinct values on a set of positive measure, the left-hand side is an affine function of u vanishing at two distinct points for almost every $\boldsymbol{x}$. Hence, both sums vanish separately. The orthonormality of the two systems then yields $a_{\boldsymbol{k}} = b_{\boldsymbol{s}} = 0$.[2] Altogether, the limit system (45) and the projection system (47) for $f = H_j$ are one and the same system, and it is invertible. Hence, both systems have the same unique solution. Viewing the index set $\mathcal{I}$ in (46) as the disjoint union of $\mathcal{I}(V_1)$ and $\mathcal{I}(V_2)$, this solution is given by the pair $(\tilde{\boldsymbol{c}}^1, \tilde{\boldsymbol{c}}^2)$ and, equivalently, by the coefficient vector $(c_{\boldsymbol{n}})_{\boldsymbol{n}\in\mathcal{I}}$ of $P_{\widetilde{\mathcal{D}}}H_j$. Thus, the two coefficient representations coincide, which proves (ii).

Convergence of the solutions. The determinant depends continuously on the entries of a matrix, such that the almost sure convergence of the scaled system matrix to the invertible limit matrix implies that, for almost every realization, the scaled system matrix is invertible for all sufficiently large M. The threshold M from which on this holds depends on the realization, so this argument does not yield a quantitative statement, cf. Remark 3.5 and Proposition 3.7 below. Since matrix inversion is continuous as well, the solutions $(\tilde{\boldsymbol{c}}^1, \tilde{\boldsymbol{c}}^2)$ of (40) then converge almost surely to the coefficients of $P_{\widetilde{\mathcal{D}}}H_j$, which shows (iii). □

The following remark quantifies this solvability in terms of the control distribution.

Remark 3.5 (Quantitative solvability and conditioning)**.** *The unique solvability of the limit system* (45) *was obtained in the proof of Proposition 3.4 from the linear independence of the dictionary, which is a purely qualitative statement. It may be quantified in terms of the first two moments* (35) *of the control distribution alone. To this end, we abbreviate* $g \coloneqq \sum_{\boldsymbol{k}\in\mathcal{I}(V_1)} a_{\boldsymbol{k}}\varphi^1_{\boldsymbol{k}}$ *and* $h \coloneqq \sum_{\boldsymbol{s}\in\mathcal{I}(V_2)} y_{\boldsymbol{s}}\varphi^2_{\boldsymbol{s}}$ *for coefficient vectors* $\boldsymbol{a}, \boldsymbol{y}$*, so that* $\|\boldsymbol{a}\|_2 = \|g\|_{L_2(\Omega,\omega)}$ *and* $\|\boldsymbol{y}\|_2 = \|h\|_{L_2(\Omega,\omega)}$ *by the orthonormality of the two systems. Denoting the system matrix of* (45) *by* $\mathcal{G}$ *and recalling from the proof of Proposition 3.4 that it is the Gram matrix of the dictionary* (46)*, its quadratic form reads*

$$(\boldsymbol{a},\boldsymbol{y})^\top \mathcal{G}\,(\boldsymbol{a},\boldsymbol{y}) = \|g + hu\|^2_{L_2(\Omega\times\mathbb{U},\omega\cdot\omega_{d+1})} = \|g\|^2_{L_2(\Omega,\omega)} + 2\eta\,\langle g, h\rangle_{L_2(\Omega,\omega)} + \nu\,\|h\|^2_{L_2(\Omega,\omega)}.$$

By the Cauchy–Schwarz inequality, $|\langle g, h\rangle_{L_2(\Omega,\omega)}| \le \|g\|_{L_2(\Omega,\omega)}\|h\|_{L_2(\Omega,\omega)}$*, such that the right-hand side is bounded from below and from above by the quadratic forms of the symmetric matrices* $\left(\begin{smallmatrix} 1 & \mp|\eta| \\ \mp|\eta| & \nu \end{smallmatrix}\right)$ *evaluated at* $(\|g\|_{L_2(\Omega,\omega)}, \|h\|_{L_2(\Omega,\omega)})$*. Both matrices have trace* $1+\nu$ *and determinant* $\nu - \eta^2$*, hence the same eigenvalues, and we obtain*

$$\lambda_-\,\boldsymbol{I} \preceq \mathcal{G} \preceq \lambda_+\,\boldsymbol{I}, \qquad \lambda_\pm \coloneqq \frac{(1+\nu) \pm \sqrt{(1-\nu)^2 + 4\eta^2}}{2}, \qquad \lambda_-\lambda_+ = \nu - \eta^2. \tag{48}$$

[2]Note that for an ω_{d+1}-almost everywhere constant control the two parts of the dictionary may overlap and this argument breaks down, in accordance with Assumption 3.3.

Therein, $\nu - \eta^2$ is the variance of the control, cf. (35), such that $\mathcal{G}$ is positive definite if and only if the control distribution is non-degenerate, and its smallest eigenvalue degenerates as the control variance tends to zero. Note that (48) holds for arbitrary basis systems φ^1 and φ^2, and that both bounds are attained for $\{\varphi^1_{\boldsymbol{k}}\}_{\boldsymbol{k}\in\mathcal{I}(V_1)} = \{\varphi^2_{\boldsymbol{s}}\}_{\boldsymbol{s}\in\mathcal{I}(V_2)}$, where $\mathcal{G} = \left(\begin{smallmatrix} 1 & \eta \\ \eta & \nu \end{smallmatrix}\right) \otimes \boldsymbol{I}_{|\mathcal{I}(V_1)|}$. Moreover, the condition number of $\mathcal{G}$ is bounded by the fraction λ_+/λ_-, which for fixed ν is minimal for $\eta = 0$. If in addition $\nu = 1$, then $\lambda_\pm = 1$, the limit matrix is the identity and the least squares problem is asymptotically perfectly conditioned. This provides a practical guideline for the choice of the control density ω_{d+1}.

The next remark identifies the situation in which the coupled reconstruction coincides with separate ANOVA approximations of the maps F and G.

Remark 3.6. *If one has access to the samples*

$$\boldsymbol{f} \coloneqq \big(F_j(\boldsymbol{x}^i)\big)_{i=1}^M \quad \text{and} \quad \boldsymbol{g} \coloneqq \big(G_j(\boldsymbol{x}^i)\big)_{i=1}^M, \tag{49}$$

one could apply the ANOVA approximation separately to the functions F and G. Therein, one would again use the matrices $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ in (37) and seek coefficients $\boldsymbol{c}^{1,\star}$ and $\boldsymbol{c}^{2,\star}$ with

$$\boldsymbol{c}^{1,\star} = \operatorname*{argmin}_{\boldsymbol{c}\in\mathbb{R}^{|\mathcal{I}(V_1)|}} \|\boldsymbol{A}_1\boldsymbol{c} - \boldsymbol{f}\|_2^2, \qquad \boldsymbol{c}^{2,\star} = \operatorname*{argmin}_{\boldsymbol{c}\in\mathbb{R}^{|\mathcal{I}(V_2)|}} \|\boldsymbol{A}_2\boldsymbol{c} - \boldsymbol{g}\|_2^2, \tag{50}$$

cf. Section 2.2. In the limit $M \to \infty$ they would result from the normal equations as

$$\boldsymbol{I}_{|\mathcal{I}(V_1)|} \cdot \boldsymbol{c}^{1,\star} = \big(\langle F_j, \varphi^1_{\boldsymbol{k}}\rangle_{L_2(\Omega,\omega)}\big)_{\boldsymbol{k}\in\mathcal{I}(V_1)}, \qquad \boldsymbol{I}_{|\mathcal{I}(V_2)|} \cdot \boldsymbol{c}^{2,\star} = \big(\langle G_j, \varphi^2_{\boldsymbol{s}}\rangle_{L_2(\Omega,\omega)}\big)_{\boldsymbol{s}\in\mathcal{I}(V_2)},$$

or combined from the system

$$\begin{pmatrix} \boldsymbol{I}_{|\mathcal{I}(V_1)|} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{|\mathcal{I}(V_2)|} \end{pmatrix} \begin{pmatrix} \boldsymbol{c}^{1,\star} \\ \boldsymbol{c}^{2,\star} \end{pmatrix} = \begin{pmatrix} \big(\langle F_j, \varphi^1_{\boldsymbol{k}}\rangle_{L_2(\Omega,\omega)}\big)_{\boldsymbol{k}\in\mathcal{I}(V_1)} \\ \big(\langle G_j, \varphi^2_{\boldsymbol{s}}\rangle_{L_2(\Omega,\omega)}\big)_{\boldsymbol{s}\in\mathcal{I}(V_2)} \end{pmatrix}.$$

Inserting this equality into the left-hand side of the system (45), we obtain

$$\begin{pmatrix} \big(\langle F_j, \varphi^1_{\boldsymbol{k}}\rangle_{L_2(\Omega,\omega)}\big)_{\boldsymbol{k}\in\mathcal{I}(V_1)} + \langle 1, u\rangle_{L_2(\mathbb{U},\omega_{d+1})} \cdot \boldsymbol{\Phi}\big(\langle G_j, \varphi^2_{\boldsymbol{s}}\rangle_{L_2(\Omega,\omega)}\big)_{\boldsymbol{s}\in\mathcal{I}(V_2)} \\ \langle 1, u\rangle_{L_2(\mathbb{U},\omega_{d+1})} \cdot \boldsymbol{\Phi}^\top\big(\langle F_j, \varphi^1_{\boldsymbol{k}}\rangle_{L_2(\Omega,\omega)}\big)_{\boldsymbol{k}\in\mathcal{I}(V_1)} + \|u\|^2_{L_2(\mathbb{U},\omega_{d+1})} \big(\langle G_j, \varphi^2_{\boldsymbol{s}}\rangle_{L_2(\Omega,\omega)}\big)_{\boldsymbol{s}\in\mathcal{I}(V_2)} \end{pmatrix}.$$

In other words, the coupled reconstruction approach above coincides with the separate application of the ANOVA approximation, if

$$\begin{aligned} \boldsymbol{\Phi}\big(\langle G_j, \varphi^2_{\boldsymbol{s}}\rangle_{L_2(\Omega,\omega)}\big)_{\boldsymbol{s}\in\mathcal{I}(V_2)} &= \left(\sum_{\boldsymbol{s}\in\mathcal{I}(V_2)} \langle \varphi^1_{\boldsymbol{k}}, \varphi^2_{\boldsymbol{s}}\rangle_{L_2(\Omega,\omega)} \langle G_j, \varphi^2_{\boldsymbol{s}}\rangle_{L_2(\Omega,\omega)}\right)_{\boldsymbol{k}\in\mathcal{I}(V_1)} \\ &= \big(\langle G_j, \varphi^1_{\boldsymbol{k}}\rangle_{L_2(\Omega,\omega)}\big)_{\boldsymbol{k}\in\mathcal{I}(V_1)} \end{aligned}$$

and

$$\begin{aligned} \boldsymbol{\Phi}^\top\big(\langle F_j, \varphi^1_{\boldsymbol{k}}\rangle_{L_2(\Omega,\omega)}\big)_{\boldsymbol{k}\in\mathcal{I}(V_1)} &= \left(\sum_{\boldsymbol{k}\in\mathcal{I}(V_1)} \langle \varphi^1_{\boldsymbol{k}}, \varphi^2_{\boldsymbol{s}}\rangle_{L_2(\Omega,\omega)} \langle F_j, \varphi^1_{\boldsymbol{k}}\rangle_{L_2(\Omega,\omega)}\right)_{\boldsymbol{s}\in\mathcal{I}(V_2)} \\ &= \big(\langle F_j, \varphi^2_{\boldsymbol{s}}\rangle_{L_2(\Omega,\omega)}\big)_{\boldsymbol{s}\in\mathcal{I}(V_2)}. \end{aligned}$$

Both conditions are fulfilled if $\{\varphi^1_{\boldsymbol{k}}\}_{\boldsymbol{k}\in\mathcal{I}(V_1)} = \{\varphi^2_{\boldsymbol{s}}\}_{\boldsymbol{s}\in\mathcal{I}(V_2)}$*, as in this case* $\boldsymbol{\Phi} = \big(\langle\varphi^1_{\boldsymbol{k}}, \varphi^2_{\boldsymbol{s}}\rangle_{L_2(\Omega,\omega)}\big)_{\boldsymbol{k}\in\mathcal{I}(V_1),\,\boldsymbol{s}\in\mathcal{I}(V_2)}$ *is the identity matrix. Two refinements of this conclusion are in order. First, in the two conditions above the common factor* $\langle 1, u\rangle_{L_2(\mathbb{U},\omega_{d+1})}$ *has been cancelled. If the control density has zero mean, i. e.,* $\langle 1, u\rangle_{L_2(\mathbb{U},\omega_{d+1})} = 0$*, the off-diagonal blocks in* (45) *vanish and the coupled approach coincides in the limit with the separate ANOVA approximation for* arbitrary *basis systems. This provides yet another argument in favour of zero-mean control densities. Second, for* $\langle 1, u\rangle_{L_2(\mathbb{U},\omega_{d+1})} \neq 0$*, if the two conditions are required to hold for all* F_j, G_j*, they are equivalent to* $\operatorname{span}\{\varphi^1_{\boldsymbol{k}}\}_{\boldsymbol{k}\in\mathcal{I}(V_1)} = \operatorname{span}\{\varphi^2_{\boldsymbol{s}}\}_{\boldsymbol{s}\in\mathcal{I}(V_2)}$*. The systems need not be identical, since any orthonormal basis of the same subspace suffices.*

3.2. **Error analysis**

We now turn to the error analysis of the coupled approach for finitely many samples. The empirical least squares problem (40) controls the residual of H_j at the sampling points, whereas we aim to bound the L_2-error of the resulting approximation. The link between these two quantities is a norm equivalence between the discrete and the continuous norm. For a *fixed* bounded function f, such an equivalence holds with high probability by Hoeffding's inequality [12], since $\frac{1}{M}\sum_{i=1}^{M} |f(\boldsymbol{z}^i)|^2$ is the sample mean of the i.i.d. random variables $|f(\boldsymbol{z}^i)|^2$ with expectation $\|f\|^2_{L_2(\Omega\times\mathbb{U},\omega\cdot\omega_{d+1})}$. The residual $H_j - \tilde{H}_j$, however, depends on the samples through the computed coefficients, such that a statement for a fixed function does not suffice. Instead, we establish a norm equivalence which holds *uniformly* on the ansatz space, i. e., a Marcinkiewicz–Zygmund inequality for the dictionary $\widetilde{\mathcal{D}}$ under random sampling, cf. [7, 8], where such norm equivalences form the core of the stability analysis of least squares approximation. In the proof of the error bound, it is applied to the component of the residual in $\operatorname{span}\widetilde{\mathcal{D}}$, whereas the remaining component is a fixed function to which Hoeffding's inequality applies.

In the following, we fix three conventions. First, the index sets $\mathcal{I}(V_1)$ and $\mathcal{I}(V_2)$ are understood as *prescribed*, deterministic sets, i. e., we assume that the truncation step of Algorithm 2.4 is not carried out but that the index sets are built from the exact structure (32) instead. We refer the reader to Remark 3.9 for a discussion of the data-dependent case. Second, we use the Chebyshev basis (6) in both systems, i. e., $\varphi^1 = \varphi^2 = \varphi^{\mathrm{cheb}}$ with possibly different index sets $\mathcal{I}(V_1)$ and $\mathcal{I}(V_2)$, we sample according to $\omega^{\mathrm{cheb}}\cdot\omega_{d+1}$ with a probability density ω_{d+1} on the compact set $\mathbb{U}\subset\mathbb{R}$, and we let Assumption 3.3 hold, such that $\lambda_- > 0$ in (48). Third, the estimates are stated for one fixed row $j\in[d]$. A statement holding simultaneously for all $j = 1,\dots,d$ follows from a union bound over the d rows, which multiplies the failure probability by d. We abbreviate

$$n_1 \coloneqq |\mathcal{I}(V_1)|, \qquad n_2 \coloneqq |\mathcal{I}(V_2)|, \qquad u_{\max} \coloneqq \max_{u\in\mathbb{U}} |u|, \qquad \alpha \coloneqq \max_{\boldsymbol{k}\in\mathcal{I}(V_1)\cup\mathcal{I}(V_2)} |\operatorname{supp}\boldsymbol{k}|,$$

and denote by $\mathcal{G}_M \coloneqq \frac{1}{M}\boldsymbol{B}^\top\boldsymbol{B}$ with $\boldsymbol{B} \coloneqq (\boldsymbol{A}_1 \;\; \boldsymbol{U}\boldsymbol{A}_2) \in \mathbb{R}^{M\times(n_1+n_2)}$ the scaled system matrix of the normal equations (42), whose limit for $M\to\infty$ is the Gram matrix $\mathcal{G}$ of $\widetilde{\mathcal{D}}$ by Proposition 3.4. For $g = \sum_{\boldsymbol{k}\in\mathcal{I}(V_1)} a_{\boldsymbol{k}}\varphi_{\boldsymbol{k}} + u\sum_{\boldsymbol{s}\in\mathcal{I}(V_2)} y_{\boldsymbol{s}}\varphi_{\boldsymbol{s}} \in \operatorname{span}\widetilde{\mathcal{D}}$ with coefficient vector $\boldsymbol{c} = (\boldsymbol{a}, \boldsymbol{y})^\top$ we have

$$\|g\|^2_{L_2(\Omega\times\mathbb{U},\omega^{\mathrm{cheb}}\cdot\omega_{d+1})} = \boldsymbol{c}^\top\mathcal{G}\,\boldsymbol{c} \qquad \text{and} \qquad \frac{1}{M}\sum_{i=1}^{M} |g(\boldsymbol{x}^i, u^i)|^2 = \frac{1}{M}\|\boldsymbol{B}\boldsymbol{c}\|_2^2 = \boldsymbol{c}^\top\mathcal{G}_M\,\boldsymbol{c}. \tag{51}$$

Proposition 3.7 (Marcinkiewicz–Zygmund inequality for the coupled dictionary)**.** *Let* $\delta\in(0,1)$*,* $r > 0$ *and* $M > 3$ *satisfy*

$$n_1 + u_{\max}^2\, n_2 \;\le\; \frac{c_\delta\,\lambda_-\,M}{2^\alpha\,(r+1)\log M}, \qquad c_\delta \coloneqq (1+\delta)\log(1+\delta) - \delta, \tag{52}$$

where n_1*,* n_2*,* $u_{\max}$ *and* α *are given above, and where*

$$\lambda_- = \frac{(1+\nu) - \sqrt{(1-\nu)^2 + 4\eta^2}}{2} > 0$$

is the lower bound on the spectrum of the Gram matrix $\mathcal{G}$ from (48) *with the moments η, ν of the control distribution from* (35)*. Then, with probability at least $1 - 2M^{-r}$,*

$$(1-\delta)\,\|g\|^2_{L_2(\Omega\times\mathbb{U},\omega^{\mathrm{cheb}}\cdot\omega_{d+1})} \;\le\; \frac{1}{M}\sum_{i=1}^{M} |g(\boldsymbol{x}^i,u^i)|^2 \;\le\; (1+\delta)\,\|g\|^2_{L_2(\Omega\times\mathbb{U},\omega^{\mathrm{cheb}}\cdot\omega_{d+1})} \qquad \text{for all } g\in\operatorname{span}\widetilde{\mathcal{D}}, \tag{53}$$

or, equivalently, $(1-\delta)\,\mathcal{G} \preceq \mathcal{G}_M \preceq (1+\delta)\,\mathcal{G}$.

Proof. Let $\boldsymbol{v}(\boldsymbol{z}) \in \mathbb{R}^{n_1+n_2}$ denote the vector of all elements of $\widetilde{\mathcal{D}}$ evaluated at $\boldsymbol{z} = (\boldsymbol{x}, u)$, such that $\mathcal{G} = \mathbb{E}[\boldsymbol{v}(\boldsymbol{z})\boldsymbol{v}(\boldsymbol{z})^\top]$ and $\mathcal{G}_M = \frac{1}{M}\sum_{i=1}^M \boldsymbol{v}(\boldsymbol{z}^i)\boldsymbol{v}(\boldsymbol{z}^i)^\top$. The matrices $\boldsymbol{X}_i \coloneqq \mathcal{G}^{-1/2}\boldsymbol{v}(\boldsymbol{z}^i)\boldsymbol{v}(\boldsymbol{z}^i)^\top\mathcal{G}^{-1/2}$ are i.i.d. and symmetric positive semidefinite and satisfy $\mathbb{E}[\boldsymbol{X}_i] = \mathcal{G}^{-1/2}\,\mathbb{E}[\boldsymbol{v}(\boldsymbol{z}^i)\boldsymbol{v}(\boldsymbol{z}^i)^\top]\,\mathcal{G}^{-1/2} = \boldsymbol{I}$ as well as

$$\|\boldsymbol{X}_i\|_2 = \|\mathcal{G}^{-1/2}\boldsymbol{v}(\boldsymbol{z}^i)\|_2^2 \le K \coloneqq \sup_{\boldsymbol{z}\in\Omega\times\mathbb{U}} \|\mathcal{G}^{-1/2}\boldsymbol{v}(\boldsymbol{z})\|_2^2.$$

Hence, the matrix Chernoff inequality [31, Theorem 1.1] yields, with $n \coloneqq n_1 + n_2$,

$$\mathbb{P}(\lambda_{\min}(\sum_{i=1}^M \boldsymbol{X}_i) \le (1-\delta)M) \le n\,\mathrm{e}^{-M\tilde{c}_\delta/K}, \qquad \mathbb{P}(\lambda_{\max}(\sum_{i=1}^M \boldsymbol{X}_i) \ge (1+\delta)M) \le n\,\mathrm{e}^{-Mc_\delta/K},$$

where $\tilde{c}_\delta \coloneqq \delta + (1-\delta)\log(1-\delta) \ge c_\delta$. Since $\mathcal{G}^{-1/2}\mathcal{G}_M\mathcal{G}^{-1/2} = \frac{1}{M}\sum_{i=1}^M \boldsymbol{X}_i$ is symmetric, the event

$$E_\delta \coloneqq \{\|\mathcal{G}^{-1/2}\mathcal{G}_M\mathcal{G}^{-1/2} - \boldsymbol{I}\|_2 > \delta\}$$

occurs if and only if the smallest eigenvalue of this matrix is smaller than $1-\delta$ or its largest eigenvalue is larger than $1+\delta$. Since the probability of a union of two events is at most the sum of their probabilities, we get $\mathbb{P}(E_\delta) \le 2n\,\mathrm{e}^{-Mc_\delta/K}$.

It remains to bound $K = \sup_{\boldsymbol{z}\in\Omega\times\mathbb{U}} \|\mathcal{G}^{-1/2}\boldsymbol{v}(\boldsymbol{z})\|_2^2$. By (48), $\|\mathcal{G}^{-1/2}\boldsymbol{v}(\boldsymbol{z})\|_2^2 = \boldsymbol{v}(\boldsymbol{z})^\top\mathcal{G}^{-1}\boldsymbol{v}(\boldsymbol{z}) \le \|\boldsymbol{v}(\boldsymbol{z})\|_2^2/\lambda_-$. Moreover, the Chebyshev basis (6) satisfies $|\varphi_{\boldsymbol{k}}(\boldsymbol{x})|^2 \le 2^{|\operatorname{supp}\boldsymbol{k}|} \le 2^\alpha$ for all $\boldsymbol{x}\in[-1,1]^d$ and $\boldsymbol{k}\in\mathcal{I}(V_1)\cup\mathcal{I}(V_2)$, such that

$$\|\boldsymbol{v}(\boldsymbol{z})\|_2^2 = \sum_{\boldsymbol{k}\in\mathcal{I}(V_1)} |\varphi_{\boldsymbol{k}}(\boldsymbol{x})|^2 + u^2 \sum_{\boldsymbol{s}\in\mathcal{I}(V_2)} |\varphi_{\boldsymbol{s}}(\boldsymbol{x})|^2 \le 2^\alpha(n_1 + u_{\max}^2\, n_2).$$

Hence, $K \le 2^\alpha(n_1 + u_{\max}^2 n_2)/\lambda_-$, and (52) yields $K \le c_\delta M/((r+1)\log M)$, i. e., $\mathrm{e}^{-Mc_\delta/K} \le M^{-(r+1)}$. Since $n = \operatorname{tr}(\boldsymbol{I}) = \mathbb{E}[\operatorname{tr}\boldsymbol{X}_i] = \mathbb{E}\|\mathcal{G}^{-1/2}\boldsymbol{v}(\boldsymbol{z}^i)\|_2^2 \le K \le M$ (note that $c_\delta < 1$ and $(r+1)\log M > 1$ for $M > 3$), we obtain $\mathbb{P}(E_\delta) \le 2nM^{-(r+1)} \le 2M^{-r}$. On the complement of E_δ, we have $(1-\delta)\boldsymbol{I} \preceq \mathcal{G}^{-1/2}\mathcal{G}_M\mathcal{G}^{-1/2} \preceq (1+\delta)\boldsymbol{I}$, which is equivalent to $(1-\delta)\mathcal{G} \preceq \mathcal{G}_M \preceq (1+\delta)\mathcal{G}$ and, by (51), to (53). □

We now may prove an error bound for the ANOVA-based approximation tailored to control-affine systems. For an index set $\mathcal{I} \subset \mathbb{N}_0^d$ we abbreviate the two tail quantities of the weight β from (27) by

$$\tau_{\mathcal{I}} \coloneqq \sup_{\boldsymbol{k}\in\mathbb{N}_0^d\setminus\mathcal{I}} \frac{1}{\beta(\boldsymbol{k})} \qquad \text{and} \qquad \rho_{\mathcal{I}} \coloneqq \sum_{\boldsymbol{k}\in\mathbb{N}_0^d\setminus\mathcal{I}} \frac{1}{\beta(\boldsymbol{k})},$$

where $\rho_{\mathcal{I}} < \infty$ for $s > \frac{1}{2}$, cf. the discussion following Theorem 2.6.

Theorem 3.8 (Error bound for the coupled approach)**.** *Let $m = 1$, let $F_j, G_j \in H^s_{\mathrm{mix}}([-1,1]^d)$ with $s > \frac{1}{2}$, and let $\delta \in (0,1)$, $r > 0$ and $M > 3$ satisfy* (52)*. Let $\tilde{H}_j$ be the approximation* (34) *with the coefficients obtained*

from the least squares problem (40)*, and abbreviate* $\|\cdot\|_{H^s_{\mathrm{mix}}} \coloneqq \|\cdot\|_{H^s_{\mathrm{mix}}([-1,1]^d)}$*. Then, with probability at least* $1-3M^{-r}$*,*

$$\begin{aligned}\|H_j-\tilde H_j\|^2_{L_2([-1,1]^d\times\mathbb{U},\omega^{\mathrm{cheb}}\cdot\omega_{d+1})} &\le \frac{2(2-\delta)}{1-\delta}\big(\|F_j\|_{H^s_{\mathrm{mix}}}\tau^{1/2}_{\mathcal{I}(V_1)}+\sqrt{\nu}\,\|G_j\|_{H^s_{\mathrm{mix}}}\tau^{1/2}_{\mathcal{I}(V_2)}\big)^2\\ &\quad+\frac{2^{d+1}}{1-\delta}\sqrt{\frac{r\log M}{2M}}\big(\|F_j\|_{H^s_{\mathrm{mix}}}\rho^{1/2}_{\mathcal{I}(V_1)}+u_{\max}\,\|G_j\|_{H^s_{\mathrm{mix}}}\rho^{1/2}_{\mathcal{I}(V_2)}\big)^2.\end{aligned} \tag{54}$$

Proof. The proof consists of four steps. We write $\|\cdot\|_{L_2}$ for the norm of $L_2([-1,1]^d\times\mathbb{U},\omega^{\mathrm{cheb}}\cdot\omega_{d+1})$, $\|\cdot\|_C$ for the supremum norm on $[-1,1]^d\times\mathbb{U}$, and $\|w\|_M^2 \coloneqq \frac{1}{M}\sum_{i=1}^M |w(\boldsymbol{x}^i,u^i)|^2$ for the squared discrete seminorm induced by the samples.

Step 1. With the orthogonal projections (28), we set

$$g \coloneqq P_{\mathcal{I}(V_1)}F_j + u\,P_{\mathcal{I}(V_2)}G_j \in \operatorname{span}\widetilde{\mathcal{D}} \qquad\text{and}\qquad e \coloneqq H_j - g = (F_j - P_{\mathcal{I}(V_1)}F_j) + u\,(G_j - P_{\mathcal{I}(V_2)}G_j).$$

Note that g and e do not depend on the samples. For $f\in H^s_{\mathrm{mix}}([-1,1]^d)$ and an index set $\mathcal{I}$, Parseval's identity and (26) yield

$$\|f-P_{\mathcal{I}}f\|^2_{L_2([-1,1]^d,\omega^{\mathrm{cheb}})} = \sum_{\boldsymbol{k}\notin\mathcal{I}}|c_{\boldsymbol{k}}(f)|^2 \le \tau_{\mathcal{I}}\,\|f\|^2_{H^s_{\mathrm{mix}}([-1,1]^d)},$$

and the Cauchy–Schwarz inequality with the weights $\beta(\boldsymbol{k})^{1/2}\beta(\boldsymbol{k})^{-1/2}$ gives $\sum_{\boldsymbol{k}\notin\mathcal{I}}|c_{\boldsymbol{k}}(f)| \le \rho_{\mathcal{I}}^{1/2}\|f\|_{H^s_{\mathrm{mix}}([-1,1]^d)}$. In particular, $\sum_{\boldsymbol{k}\in\mathbb{N}_0^d}|c_{\boldsymbol{k}}(f)|<\infty$ for $s>\frac12$, such that the Chebyshev series of f converges absolutely and uniformly, f coincides with a continuous function (which is the representative being sampled), and $f-P_{\mathcal{I}}f = \sum_{\boldsymbol{k}\notin\mathcal{I}}c_{\boldsymbol{k}}(f)\varphi_{\boldsymbol{k}}$ holds pointwise. Since $|\varphi_{\boldsymbol{k}}(\boldsymbol{x})| \le 2^{|\operatorname{supp}\boldsymbol{k}|/2} \le 2^{d/2}$ for all $\boldsymbol{x}\in[-1,1]^d$ by (6), the triangle inequality yields

$$\|f-P_{\mathcal{I}}f\|_{C([-1,1]^d)} \le 2^{d/2}\sum_{\boldsymbol{k}\notin\mathcal{I}}|c_{\boldsymbol{k}}(f)| \le 2^{d/2}\,\rho_{\mathcal{I}}^{1/2}\,\|f\|_{H^s_{\mathrm{mix}}([-1,1]^d)}.$$

Moreover, for $a,b\in L_2([-1,1]^d,\omega^{\mathrm{cheb}})$ we have $\|a+ub\|_{L_2}^2 = \|a\|^2+2\eta\langle a,b\rangle+\nu\|b\|^2 \le (\|a\|+\sqrt{\nu}\|b\|)^2$ with the norm and inner product of $L_2([-1,1]^d,\omega^{\mathrm{cheb}})$, since $\eta^2\le\nu$ by the Cauchy–Schwarz inequality. Applying these three estimates to e yields

$$\|e\|_{L_2} \le \|F_j\|_{H^s_{\mathrm{mix}}}\tau^{1/2}_{\mathcal{I}(V_1)}+\sqrt{\nu}\,\|G_j\|_{H^s_{\mathrm{mix}}}\tau^{1/2}_{\mathcal{I}(V_2)}, \qquad \|e\|_C \le 2^{d/2}(\|F_j\|_{H^s_{\mathrm{mix}}}\rho^{1/2}_{\mathcal{I}(V_1)}+u_{\max}\|G_j\|_{H^s_{\mathrm{mix}}}\rho^{1/2}_{\mathcal{I}(V_2)}). \tag{55}$$

Step 2. On the event where (53) holds, $\mathcal{G}_M$ is positive definite, hence $\boldsymbol{B}$ has full column rank and (40) has a unique solution $\boldsymbol{c}^\star$. Let $\boldsymbol{\Pi}\in\mathbb{R}^{M\times M}$ denote the orthogonal projector onto the range of $\boldsymbol{B}$, let $\boldsymbol{c}_g$ be the coefficient vector of g, and let $\boldsymbol{e} \coloneqq (e(\boldsymbol{x}^i,u^i))_{i=1}^M$. Then $\boldsymbol{B}\boldsymbol{c}^\star = \boldsymbol{\Pi}\boldsymbol{h}$ with the sample vector $\boldsymbol{h} = (H_j(\boldsymbol{x}^i,u^i))_{i=1}^M$ from (40) and $\boldsymbol{B}\boldsymbol{c}_g = \boldsymbol{\Pi}\boldsymbol{B}\boldsymbol{c}_g$, such that $\boldsymbol{B}(\boldsymbol{c}^\star-\boldsymbol{c}_g) = \boldsymbol{\Pi}(\boldsymbol{h}-\boldsymbol{B}\boldsymbol{c}_g) = \boldsymbol{\Pi}\boldsymbol{e}$ and therefore

$$\|\tilde H_j - g\|_M = M^{-1/2}\|\boldsymbol{B}(\boldsymbol{c}^\star-\boldsymbol{c}_g)\|_2 = M^{-1/2}\|\boldsymbol{\Pi}\boldsymbol{e}\|_2 \le M^{-1/2}\|\boldsymbol{e}\|_2 = \|e\|_M.$$

As $\tilde H_j - g\in\operatorname{span}\widetilde{\mathcal{D}}$, the lower bound in (53) gives $\|\tilde H_j-g\|_{L_2}\le(1-\delta)^{-1/2}\|\tilde H_j-g\|_M$, and the triangle inequality yields

$$\begin{aligned}&\|H_j-\tilde H_j\|_{L_2} \le \|e\|_{L_2}+\|g-\tilde H_j\|_{L_2} \le \|e\|_{L_2}+(1-\delta)^{-1/2}\|e\|_M \qquad\text{and hence}\\ &\|H_j-\tilde H_j\|^2_{L_2} \le 2\|e\|^2_{L_2}+\frac{2}{1-\delta}\|e\|^2_M.\end{aligned} \tag{56}$$

Step 3. The quantity $\|e\|_M^2$ is the sample mean of the i.i.d. random variables $|e(\boldsymbol{x}^i, u^i)|^2 \in [0, \|e\|_C^2]$ with expectation $\|e\|_{L_2}^2$. Hence, Hoeffding's inequality [12] yields $\mathbb{P}(\|e\|_M^2 - \|e\|_{L_2}^2 \geq t) \leq \exp(-2Mt^2/\|e\|_C^4)$ for every $t > 0$, and the choice $t \coloneqq \|e\|_C^2 \sqrt{r \log M/(2M)}$ gives

$$\|e\|_M^2 \leq \|e\|_{L_2}^2 + \|e\|_C^2 \sqrt{\frac{r \log M}{2M}} \tag{57}$$

with probability at least $1 - M^{-r}$ (if $\|e\|_C = 0$, there is nothing to show).

Step 4: Conclusion. By Boole's inequality, (53) and (57) hold simultaneously with probability at least $1 - 2M^{-r} - M^{-r}$. Inserting (57) into (56) and using $2 + \frac{2}{1-\delta} = \frac{2(2-\delta)}{1-\delta}$, we obtain

$$\|H_j - \tilde{H}_j\|_{L_2}^2 \leq \frac{2(2-\delta)}{1-\delta} \|e\|_{L_2}^2 + \frac{2}{1-\delta} \|e\|_C^2 \sqrt{\frac{r \log M}{2M}},$$

and (55) yields the assertion. □

We give a short interpretation of the estimate (54), which features two central components. The first term is independent of M and bounds the squared distance $\|e\|_{L_2}^2$ of H_j to the comparison function g, i. e., the truncation and projection error of Remark 2.7 in the coupled setting. Note that g consists of the L_2-projections of F_j and G_j, i. e., of their separate ANOVA approximations in the limit $M \to \infty$, cf. Remark 3.6. For $\eta = 0$, the Gram matrix $\mathcal{G}$ is block diagonal and g coincides with the orthogonal projection $P_{\widetilde{\mathcal{D}}} H_j$ of Proposition 3.4, such that $\|e\|_{L_2}$ is the best-approximation error of H_j in $\operatorname{span} \widetilde{\mathcal{D}}$ itself. The factor $2(2-\delta)/(1-\delta)$ stems from the binomial estimate in (56) and from the lower bound in (53). The first term can only be reduced by enlarging the index sets, and the factor $\sqrt{\nu}$ weights the contribution of G_j with the second moment of the control. The second term is the estimation error and stems from replacing the L_2-norm of the fixed remainder e by its discrete counterpart, decays like $(\log M/M)^{1/2}$ for fixed index sets, and is proportional to the squared supremum norm of e. This is why it involves the ℓ_1-tails $\rho_{\mathcal{I}}$ instead of the ℓ_∞-tails $\tau_{\mathcal{I}}$ and carries the factor 2^d.

We briefly comment on the choice of the index sets in the above theorem.

Remark 3.9 (Deterministic versus data-dependent index sets)**.** *Theorem 3.8 (and likewise Corollary 3.13 below) is formulated for index sets that are fixed* before *the data are drawn. This is consistent with the sampling condition, which is a statement about the random matrix $\boldsymbol{B}$ associated with the prescribed dictionary $\widetilde{\mathcal{D}}$, cf. Proposition 3.7 and [13, 28]. In Algorithm 2.4, however, the set V of relevant terms in* (22) *is determined from the global sensitivity indices of a first approximation, i. e., from the same data that are subsequently used for the final least squares fit. Strictly speaking, the resulting index sets are then random and depend on the samples, so that the cited concentration results are not directly applicable and the theorem does not cover this variant. Hence, we briefly discuss two possible remedies that are standard in the ANOVA literature. Either the truncation step is skipped and the structure* (32) *is prescribed, in which case the theorem applies as stated. Alternatively, the data are split into two independent batches. The first one is used for the detection of the active sets, while the second, conditionally independent batch is used for the final fit. In this case, the theorem applies conditionally on the first batch.*

The above results are statements about the combined map H_j. They also yield bounds for the drift and the input map separately, which we make precise in the following remark.

Remark 3.10 (Separation of drift and input map)**.** *First note that computation in Remark 3.5 does not use that the functions g and h therein belong to the spans of the two systems. Hence, for all $a, b \in L_2(\Omega, \omega)$, we have*

$$\|a + b\,u\|_{L_2(\Omega \times \mathbb{U}, \omega \cdot \omega_{d+1})}^2 = \|a\|_{L_2(\Omega,\omega)}^2 + 2\eta \langle a, b \rangle_{L_2(\Omega,\omega)} + \nu \|b\|_{L_2(\Omega,\omega)}^2 \geq \lambda_-(\|a\|_{L_2(\Omega,\omega)}^2 + \|b\|_{L_2(\Omega,\omega)}^2).$$

Choosing $a = F_j - \tilde{F}_j$ *and* $b = G_j - \tilde{G}_j$ *with* $\tilde{F}_j \coloneqq \sum_{\boldsymbol{k}\in\mathcal{I}(V_1)} c^1_{\boldsymbol{k}}\varphi^1_{\boldsymbol{k}}$ *and* $\tilde{G}_j \coloneqq \sum_{\boldsymbol{s}\in\mathcal{I}(V_2)} c^2_{\boldsymbol{s}}\varphi^2_{\boldsymbol{s}}$*. yields*

$$\|F_j - \tilde{F}_j\|^2_{L_2(\Omega,\omega)} + \|G_j - \tilde{G}_j\|^2_{L_2(\Omega,\omega)} \leq \frac{1}{\lambda_-}\,\|H_j - \tilde{H}_j\|^2_{L_2(\Omega\times\mathbb{U},\omega\cdot\omega_{d+1})}.$$

Consequently, Theorem 3.8 yields error bounds for F_j *and* G_j *individually, with the same probability and the additional factor* $1/\lambda_-$*. In this sense, the coupled ansatz separates the drift from the input map for finitely many samples under Assumption 3.3.*

3.3. **Multivariate control**

For $m > 1$ note that the rows of (31) can be written as

$$H_j(\boldsymbol{x},\boldsymbol{u}) = F_j(\boldsymbol{x}) + \sum_{\ell=1}^{m} G_{j\ell}(\boldsymbol{x})u_\ell, \quad j = 1,\dots,d,$$

with $\boldsymbol{u} = (u_\ell)_{\ell=1}^m \in \mathbb{R}^m$. Thus, we generalize the above idea of (34) as

$$H_j(\boldsymbol{x},\boldsymbol{u}) \approx \sum_{\boldsymbol{k}\in\mathcal{I}(V_1)} c^1_{\boldsymbol{k}}\varphi^1_{\boldsymbol{k}}(\boldsymbol{x}) + \sum_{\ell=1}^{m}\left(\sum_{\boldsymbol{s}\in\mathcal{I}(V_{\ell+1})} c^{\ell+1}_{\boldsymbol{s}}\varphi^{\ell+1}_{\boldsymbol{s}}(\boldsymbol{x})\right)u_\ell, \quad j = 1,\dots,d. \tag{58}$$

Again we sample the whole system (31) at i.i.d. data points $(\boldsymbol{x}^i,\boldsymbol{u}^i) \in \Omega\times\mathbb{U} \subset \mathbb{R}^{d+m}$, $i = 1,\dots,M$, drawn according to the tensor product density $\omega\cdot\widetilde{\omega}$ with $M \in \mathbb{N}$ and $\widetilde{\omega} \coloneqq \prod_{\ell=1}^m \omega_{d+\ell}$ being the density corresponding to $\mathbb{U} = \bigtimes_{\ell=1}^m \mathbb{U}_\ell \subset \mathbb{R}^m$. Analogously to (35), we abbreviate the first two moments of the components of the control by

$$\eta_\ell \coloneqq \langle 1, u_\ell\rangle_{L_2(\mathbb{U}_\ell,\omega_{d+\ell})}, \qquad \nu_\ell \coloneqq \|u_\ell\|^2_{L_2(\mathbb{U}_\ell,\omega_{d+\ell})}, \quad \ell = 1,\dots,m, \tag{59}$$

and replace Assumption 3.3 by its multivariate counterpart.

Assumption 3.11 (Non-degenerate control distribution, multivariate case)**.** *The distribution of each component* u_ℓ *of the control is non-degenerate, i. e.,* $\nu_\ell - \eta_\ell^2 > 0$ *for* $\ell = 1,\dots,m$*.*

Introduce the vectors

$$\boldsymbol{c}^{\ell+1} \coloneqq \left(c^{\ell+1}_{\boldsymbol{k}}\right)_{\boldsymbol{k}\in\mathcal{I}(V_{\ell+1})} \in \mathbb{R}^{|\mathcal{I}(V_{\ell+1})|}, \quad \ell = 0,\dots,m,$$

and matrices

$$\boldsymbol{A}_{\ell+1} \coloneqq \left(\varphi^{\ell+1}_{\boldsymbol{k}}(\boldsymbol{x}^i)\right)_{i=1,\,\boldsymbol{k}\in\mathcal{I}(V_{\ell+1})}^{M} \in \mathbb{R}^{M\times|\mathcal{I}(V_{\ell+1})|}, \quad \ell = 0,\dots,m, \tag{60}$$

as well as

$$\boldsymbol{U}_\ell \coloneqq \operatorname{diag}(u^1_\ell,\dots,u^M_\ell) \in \mathbb{R}^{M\times M}, \quad \ell = 1,\dots,m. \tag{61}$$

Then, the same technique as pursued for the single-input case $m = 1$ applies to the individual components of $\boldsymbol{u}^i$. For fixed $j = 1,\dots,d$, we obtain the matrix-vector form

$$\boldsymbol{h} = \left(H_j(\boldsymbol{x}^i,\boldsymbol{u}^i)\right)_{i=1}^M \approx \begin{pmatrix}\boldsymbol{A}_1 & \boldsymbol{U}_1\boldsymbol{A}_2 & \dots & \boldsymbol{U}_m\boldsymbol{A}_{m+1}\end{pmatrix}\begin{pmatrix}\boldsymbol{c}^1\\ \vdots\\ \boldsymbol{c}^{m+1}\end{pmatrix}.$$

The least squares solution to this problem is again computed by means of the normal equations. Thus, we need to ensure that the system matrix has full column rank. Since we may assume that all $\boldsymbol{A}_{\ell+1}$ have full column rank by Section 2.2, we require an overdetermined system with $M \geq |\mathcal{I}(V_1)| + \cdots + |\mathcal{I}(V_{m+1})|$. Again, the scaled Gram matrix converges almost surely, entrywise, to the positive definite limit matrix of the following remark, provided that Assumption 3.11 holds. Since the limit matrix is invertible, the event of full column rank has asymptotic probability one. A quantitative statement, which holds with high probability, follows from the multivariate analogue of Proposition 3.7, see the proof of Corollary 3.13 below. Concerning the conditioning, zero-mean and normalized control densities again yield the identity as limit matrix, cf. Section 3.1.

Remark 3.12. *In analogy to the proof of Proposition 3.4, the normal equations converge after scaling by* $\frac{1}{M}$ *for* $M \to \infty$ *to the limit system*

$$\begin{pmatrix} \boldsymbol{I}_{|\mathcal{I}(V_1)|} & \eta_1 \cdot \boldsymbol{\Phi}_{\{1,2\}} & \dots & \eta_m \cdot \boldsymbol{\Phi}_{\{1,m+1\}} \\ \eta_1 \cdot \boldsymbol{\Phi}_{\{1,2\}}^\top & \nu_1 \cdot \boldsymbol{I}_{|\mathcal{I}(V_2)|} & \dots & \eta_1\eta_m \cdot \boldsymbol{\Phi}_{\{2,m+1\}} \\ \vdots & \vdots & \ddots & \vdots \\ \eta_m \cdot \boldsymbol{\Phi}_{\{1,m+1\}}^\top & \eta_1\eta_m \cdot \boldsymbol{\Phi}_{\{2,m+1\}}^\top & \dots & \nu_m \cdot \boldsymbol{I}_{|\mathcal{I}(V_{m+1})|} \end{pmatrix} \begin{pmatrix} \boldsymbol{c}^1 \\ \boldsymbol{c}^2 \\ \vdots \\ \boldsymbol{c}^{m+1} \end{pmatrix}$$

$$= \begin{pmatrix} (\langle F_j, \varphi_{\boldsymbol{k}}^1 \rangle_{L_2(\Omega,\omega)} + \sum_{n=1}^m \eta_n \langle G_{jn}, \varphi_{\boldsymbol{k}}^1 \rangle_{L_2(\Omega,\omega)})_{\boldsymbol{k} \in \mathcal{I}(V_1)} \\ (\eta_1 \langle F_j, \varphi_{\boldsymbol{s}}^2 \rangle_{L_2(\Omega,\omega)} + \nu_1 \langle G_{j1}, \varphi_{\boldsymbol{s}}^2 \rangle_{L_2(\Omega,\omega)} + \eta_1 \sum_{n \neq 1} \eta_n \langle G_{jn}, \varphi_{\boldsymbol{s}}^2 \rangle_{L_2(\Omega,\omega)})_{\boldsymbol{s} \in \mathcal{I}(V_2)} \\ \vdots \\ (\eta_m \langle F_j, \varphi_{\boldsymbol{s}}^{m+1} \rangle_{L_2(\Omega,\omega)} + \nu_m \langle G_{jm}, \varphi_{\boldsymbol{s}}^{m+1} \rangle_{L_2(\Omega,\omega)} + \eta_m \sum_{n \neq m} \eta_n \langle G_{jn}, \varphi_{\boldsymbol{s}}^{m+1} \rangle_{L_2(\Omega,\omega)})_{\boldsymbol{s} \in \mathcal{I}(V_{m+1})} \end{pmatrix}$$

with the matrices

$$\boldsymbol{\Phi}_{\{\ell+1,n+1\}} \coloneqq \left(\langle \varphi_{\boldsymbol{k}}^{\ell+1}, \varphi_{\boldsymbol{s}}^{n+1} \rangle_{L_2(\Omega,\omega)} \right)_{\boldsymbol{k} \in \mathcal{I}(V_{\ell+1}),\, \boldsymbol{s} \in \mathcal{I}(V_{n+1})}, \quad 0 \leq \ell < n \leq m,$$

and the moments η_ℓ, ν_ℓ *from* (59). *This limit system again coincides with the linear system characterizing the orthogonal projection of* H_j *onto the span of the corresponding dictionary, cf.* (45) *and the projection interpretation given there. Here, we used* $H_j = F_j + \sum_{n=1}^m G_{jn} u_n$ *and the independence of the components of* $\boldsymbol{u}$, *i. e.,* $\langle u_n, u_\ell \rangle = \eta_n \eta_\ell$ *for* $n \neq \ell$. *For* $m = 1$, *the system indeed reduces to* (45). *Concerning solvability, the interpretation as an orthogonal projection carries over from the proof of Proposition 3.4. The limit system is the Gram system of the dictionary* $\{\varphi_{\boldsymbol{k}}^1\}_{\boldsymbol{k} \in \mathcal{I}(V_1)} \cup \bigcup_{\ell=1}^m \{\varphi_{\boldsymbol{s}}^{\ell+1} u_\ell\}_{\boldsymbol{s} \in \mathcal{I}(V_{\ell+1})}$ *in* $L_2(\Omega \times \mathbb{U}, \omega \cdot \widetilde{\omega})$. *Under Assumption 3.11, i. e., if none of the* u_ℓ *is* $\omega_{d+\ell}$*-almost everywhere constant, the functions* $1, u_1, \dots, u_m$ *are linearly independent in* $L_2(\mathbb{U}, \widetilde{\omega})$, *as the* u_ℓ *are independent and non-degenerate. Therefore, the dictionary is linearly independent, its Gram matrix is positive definite, and the limit system is uniquely solvable. Finally, if all control densities have zero mean, i. e.,* $\eta_\ell = 0$ *for* $\ell = 1, \dots, m$, *all off-diagonal blocks, which carry the factors* η_ℓ *or* $\eta_\ell \eta_n$, *vanish. The limit system then becomes block diagonal, and the coupled reconstruction coincides in the limit with the separate ANOVA approximations of* F_j *and* $G_{j\ell}$ *for arbitrary basis systems, in complete analogy to Remark 3.6.*

The error analysis of Section 3.2 carries over to multivariate controls. To formulate it, we denote by

$$\boldsymbol{N} \coloneqq \int_{\mathbb{U}} \begin{pmatrix} 1 \\ \boldsymbol{u} \end{pmatrix} \begin{pmatrix} 1 \\ \boldsymbol{u} \end{pmatrix}^\top \widetilde{\omega}(\boldsymbol{u}) \, \mathrm{d}\boldsymbol{u} = \begin{pmatrix} 1 & \boldsymbol{\eta}^\top \\ \boldsymbol{\eta} & \boldsymbol{\eta}\boldsymbol{\eta}^\top + \operatorname{diag}(\nu_1 - \eta_1^2, \dots, \nu_m - \eta_m^2) \end{pmatrix} \in \mathbb{R}^{(m+1)\times(m+1)} \tag{62}$$

with $\boldsymbol{\eta} \coloneqq (\eta_\ell)_{\ell=1}^m$ the Gram matrix of the functions $1, u_1, \dots, u_m$ in $L_2(\mathbb{U}, \widetilde{\omega})$, where the second identity follows from the independence of the control components. Its Schur complement with respect to the top left entry is the diagonal matrix of the variances $\nu_\ell - \eta_\ell^2$, such that $\boldsymbol{N}$ is positive definite if and only if Assumption 3.11 holds. For $m = 1$, the eigenvalues of $\boldsymbol{N} = \left(\begin{smallmatrix} 1 & \eta \\ \eta & \nu \end{smallmatrix}\right)$ are $\lambda_\pm$ from (48), and for zero-mean controls we have $\boldsymbol{N} = \operatorname{diag}(1, \nu_1, \dots, \nu_m)$.

Corollary 3.13 (Error estimate for multivariate control)**.** *Let $m \geq 1$, $j \in [d]$, let Assumption 3.11 hold, and let $\lambda_- > 0$ denote the smallest eigenvalue of $\boldsymbol{N}$ in (62). Let $F_j, G_{j1}, \dots, G_{jm} \in H^s_{\mathrm{mix}}([-1,1]^d)$ with $s > \frac{1}{2}$, let the sample points be drawn i.i.d. according to $\omega^{\mathrm{cheb}} \cdot \widetilde{\omega}$, and use the Chebyshev basis (6) in all $m+1$ systems of the ansatz (58) with prescribed index sets $\mathcal{I}(V_1), \dots, \mathcal{I}(V_{m+1})$ of cardinalities $n_1, \dots, n_{m+1}$. With $u_{\max,\ell} \coloneqq \max_{\boldsymbol{u} \in \mathbb{U}} |u_\ell|$, $\alpha \coloneqq \max\{|\operatorname{supp} \boldsymbol{k}| \colon \boldsymbol{k} \in \mathcal{I}(V_1) \cup \dots \cup \mathcal{I}(V_{m+1})\}$ and c_δ from (52), let $\delta \in (0,1)$, $r > 0$ and $M > 3$ satisfy*

$$n_1 + \sum_{\ell=1}^{m} u_{\max,\ell}^2 \, n_{\ell+1} \leq \frac{c_\delta \, \lambda_- \, M}{2^\alpha \, (r+1) \log M}.$$

Let $\tilde{H}_j = \tilde{F}_j + \sum_{\ell=1}^m \tilde{G}_{j\ell} u_\ell$ be the approximation (58) with the coefficients obtained from the corresponding least squares problem, where $\tilde{F}_j$ and $\tilde{G}_{j\ell}$ collect the terms with the coefficients $\boldsymbol{c}^1$ and $\boldsymbol{c}^{\ell+1}$, respectively. Then, with probability at least $1 - 3M^{-r}$, we have

$$\begin{aligned}\|H_j - \tilde{H}_j\|^2_{L_2([-1,1]^d \times \mathbb{U}, \omega^{\mathrm{cheb}} \cdot \widetilde{\omega})} \leq{}& \frac{2(2-\delta)}{1-\delta} \Big(\|F_j\|_{H^s_{\mathrm{mix}}} \tau^{1/2}_{\mathcal{I}(V_1)} + \sum_{\ell=1}^{m} \sqrt{\nu_\ell} \, \|G_{j\ell}\|_{H^s_{\mathrm{mix}}} \tau^{1/2}_{\mathcal{I}(V_{\ell+1})} \Big)^2 \\ &+ \frac{2^{d+1}}{1-\delta} \sqrt{\frac{r \log M}{2M}} \Big(\|F_j\|_{H^s_{\mathrm{mix}}} \rho^{1/2}_{\mathcal{I}(V_1)} + \sum_{\ell=1}^{m} u_{\max,\ell} \, \|G_{j\ell}\|_{H^s_{\mathrm{mix}}} \rho^{1/2}_{\mathcal{I}(V_{\ell+1})} \Big)^2\end{aligned}$$

with the abbreviations of Theorem 3.8, as well as

$$\|F_j - \tilde{F}_j\|^2_{L_2([-1,1]^d, \omega^{\mathrm{cheb}})} + \sum_{\ell=1}^{m} \|G_{j\ell} - \tilde{G}_{j\ell}\|^2_{L_2([-1,1]^d, \omega^{\mathrm{cheb}})} \leq \frac{1}{\lambda_-} \, \|H_j - \tilde{H}_j\|^2_{L_2([-1,1]^d \times \mathbb{U}, \omega^{\mathrm{cheb}} \cdot \widetilde{\omega})}.$$

Proof. We indicate the changes in the proofs of Proposition 3.7, Theorem 3.8 and Remark 3.10. Let a, $b_1, \dots, b_m \in L_2([-1,1]^d, \omega^{\mathrm{cheb}})$ and $\boldsymbol{w} \coloneqq (a, b_1, \dots, b_m)^\top$. Then, Fubini's theorem yields

$$\|a + \sum_{\ell=1}^{m} b_\ell u_\ell\|^2_{L_2([-1,1]^d \times \mathbb{U}, \omega^{\mathrm{cheb}} \cdot \widetilde{\omega})} = \int_{[-1,1]^d} \boldsymbol{w}(\boldsymbol{x})^\top \boldsymbol{N} \, \boldsymbol{w}(\boldsymbol{x}) \, \omega^{\mathrm{cheb}}(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \geq \lambda_- (\|a\|^2 + \sum_{\ell=1}^{m} \|b_\ell\|^2) \tag{63}$$

with the norms of $L_2([-1,1]^d, \omega^{\mathrm{cheb}})$ on the right-hand side. Here, we used that

$$\int_{\mathbb{U}} |\boldsymbol{w}(\boldsymbol{x})^\top (1, \boldsymbol{u}^\top)^\top|^2 \, \widetilde{\omega}(\boldsymbol{u}) \, \mathrm{d}\boldsymbol{u} = \boldsymbol{w}(\boldsymbol{x})^\top \boldsymbol{N} \boldsymbol{w}(\boldsymbol{x}) \geq \lambda_- \|\boldsymbol{w}(\boldsymbol{x})\|_2^2 \qquad \text{for every } \boldsymbol{x} \in [-1,1]^d.$$

For a and b_ℓ in the spans of the respective systems, the right-hand side equals λ_- times the squared Euclidean norm of the coefficient vector by orthonormality. Hence, the Gram matrix $\mathcal{G}$ of the dictionary of (58) satisfies $\mathcal{G} \succeq \lambda_- \boldsymbol{I}$, which replaces (48) in the proof of Proposition 3.7. Moreover, the vector $\boldsymbol{v}(\boldsymbol{z})$ now consists of the $\varphi_{\boldsymbol{k}}(\boldsymbol{x})$, $\boldsymbol{k} \in \mathcal{I}(V_1)$, and the $\varphi_{\boldsymbol{s}}(\boldsymbol{x}) \, u_\ell$, $\boldsymbol{s} \in \mathcal{I}(V_{\ell+1})$, $\ell = 1, \dots, m$, such that $\|\boldsymbol{v}(\boldsymbol{z})\|_2^2 \leq 2^\alpha (n_1 + \sum_{\ell=1}^m u^2_{\max,\ell} \, n_{\ell+1})$, and the remaining steps of the proof of Proposition 3.7 carry over. In the proof of Theorem 3.8, we set $g \coloneqq P_{\mathcal{I}(V_1)} F_j + \sum_{\ell=1}^m u_\ell \, P_{\mathcal{I}(V_{\ell+1})} G_{j\ell}$. The triangle inequality together with $\|u_\ell b\|_{L_2([-1,1]^d \times \mathbb{U}, \omega^{\mathrm{cheb}} \cdot \widetilde{\omega})} = \sqrt{\nu_\ell} \, \|b\|_{L_2([-1,1]^d, \omega^{\mathrm{cheb}})}$ and $|u_\ell| \leq u_{\max,\ell}$ yields the analogues of (55), and Steps 2–4 follow analogously. Finally, the second estimate follows from (63) with $a = F_j - \tilde{F}_j$ and $b_\ell = G_{j\ell} - \tilde{G}_{j\ell}$.

□

4. Numerical Examples

In this section, we illustrate the coupled approach of Section 3 by means of four examples. The first one is an academic test function with prescribed low-dimensional structure. The other three are discretized PDEs,

namely a nonlinear advection-diffusion equation, the shallow water equations, and the viscous Burgers equation of Example 3.2. For each example, we illustrate that the relevant ANOVA terms of F and G are identified from the approximated sensitivity indices (21) and analyze the accuracy of the approximation.

Throughout, we proceed rowwise on a single component H_j of (31). We use the Chebyshev basis (6) with i.i.d. samples from the Chebyshev density, and apply the two-stage procedure of Algorithm 2.4 with threshold $\varepsilon = 1/|V_{\mathrm{cand}}|$, where V_{cand} is the family of non-constant candidate terms. For $d = 8$ and $q = 2$, this yields $\varepsilon = 1/73$ for $m = 1$ and $\varepsilon = 1/110$ for $m = 2$. The coupled approach is realized via the preprocessed index sets of Algorithm A.1, such that the ansatz (58) consists of $m+1$ groups of terms with the same number of basis functions each. Correspondingly, we choose the number M of sample points $(\boldsymbol{x}^i, \boldsymbol{u}^i)$ proportional to $m+1$, such that the ratio of sample points to unknowns does not depend on m. Since the coupled approach yields combined indices for $h(\boldsymbol{z}) = H_j(\boldsymbol{x}, \boldsymbol{u})$, we split them for the visualization into a group of terms that depend only on $\boldsymbol{x}$ and groups of terms containing one u_ℓ, where the term depending on u_ℓ only, i. e., the constant part of G_ℓ, is not displayed. We then renormalize the indices within each group, whereas the truncation itself is carried out with the combined indices and the threshold ε. In all figures depicting sensitivity indices, we show the approximated indices with the terms for F on the left and those for the input maps G_ℓ on the right. The computations were carried out with the package [27, 29].

Reconstruction errors are estimated by Monte-Carlo quadrature,

$$\|H_j - \tilde{H}_j\|_{L_2(\Omega\times\mathbb{U},\omega\cdot\widetilde{\omega})} \approx \left(\frac{1}{M_{\mathrm{test}}} \sum_{i=1}^{M_{\mathrm{test}}} |H_j(\boldsymbol{x}^i, \boldsymbol{u}^i) - \tilde{H}_j(\boldsymbol{x}^i, \boldsymbol{u}^i)|^2 \right)^{1/2}, \tag{64}$$

using $M_{\mathrm{test}} = 200$ fresh points that are independent of the training data. In the following, e denotes this error. Unless stated otherwise, we use $\Omega = [-1, 1]^d$, $\mathbb{U} = [-1, 1]^m$, $d = 8$, $q = 2$, $\boldsymbol{N} = (6, 4)$, $M = 10\,000\,(m + 1)$, and no regularization. States and controls are drawn from the Chebyshev density, which gives $\eta_\ell = 0$, $\nu_\ell = \frac{1}{2}$, $\max_{\boldsymbol{u}\in\mathbb{U}} |u_\ell|^2 = 1$, and hence $\lambda_- = \frac{1}{2}$ in Proposition 3.7 and Corollary 3.13.

The Julia code that reproduces all numerical experiments and figures of this section is available at

`https://gitlab.hrz.tu-chemnitz.de/manuels-at-tu-chemnitz.de/anova-for-control`.

4.1. **Academic test function**

As a first test case we consider the example function

$$H_j(\boldsymbol{x}, u) = (x_4 - x_1)\, x_2 - x_3 u$$

with $d = 8$ and $m = 1$. This corresponds to one row of (31), with $F_j(\boldsymbol{x}) = (x_4 - x_1)\, x_2$ and $G_j(\boldsymbol{x}) = -x_3$ in the notation of (32). Hence, the index sets to be identified are $V_{j,0} = \{\{1, 2\}, \{2, 4\}\}$ and $V_{j,1} = \{\{3\}\}$, while the remaining four variables $x_5, \ldots, x_8$ are uninformative. Since this example function contains two-dimensional terms at maximum, it is sufficient to choose $q = 2$.

In Figure 2, we depict the approximated global sensitivity indices. We observe that the relevant index sets are correctly identified, as expected for a function with prescribed structure. The resulting reconstruction is accurate as well. The L_2-error (64) is $e = 1.50 \cdot 10^{-9}$, i. e., the function is reproduced up to the accuracy of the least squares solver.

Moreover, we study the influence of the superposition dimension q (truncation error) and of the number of samples M (estimation error), cf. Remark 2.7. The corresponding results are depicted in Figure 3, where the bandwidths are kept at the default $\boldsymbol{N} = (6, 4)$. Since the example consists of ANOVA terms of order at most two, the truncation error vanishes for $q \geq 2$, as confirmed by the numerical results. When varying $M \in (m+1) \cdot \{10, 20, 50, \ldots, 50\,000\}$, the error remains above $6 \cdot 10^{-1}$ up to $M = 100$, drops below 10^{-6} at $M = 200$, and reaches the linear solver accuracy of 10^{-9}–10^{-8} for larger M.

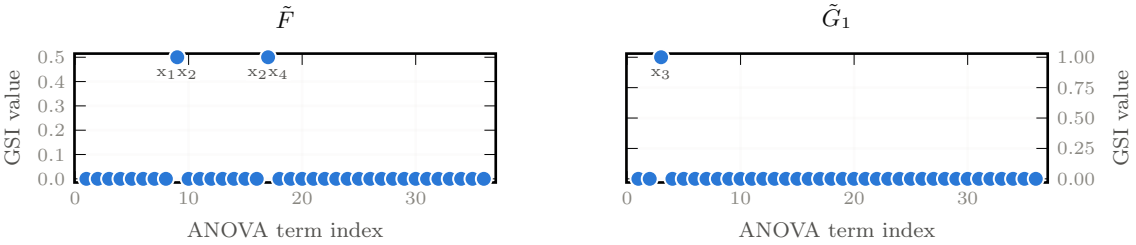


FIGURE 2. Approximated global sensitivity indices (21) for the *academic test function* with $m = 1$, where the relevant terms are $\{1, 2\}$, $\{2, 4\}$ of F and $\{3\}$ of G_1.

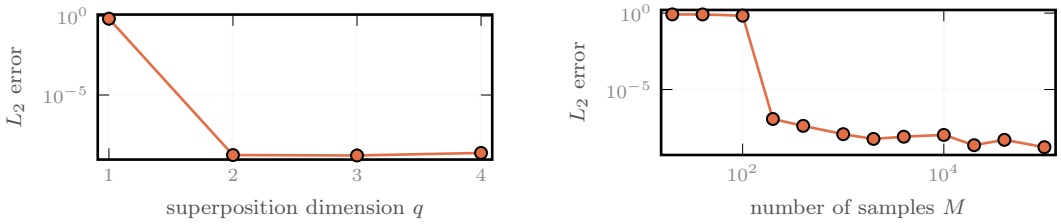


FIGURE 3. Error e for the *academic test function* depending on the superposition dimension q (left) and the number of samples M (right).

4.2. **Diffusion equation with nonlinear advection**

As a second example, we consider the nonlinear advection-diffusion equation

$$\frac{\partial}{\partial t}x \;=\; D\,\frac{\partial^2}{\partial \omega^2}x \;+\; [x(1-x)]\,\frac{\partial}{\partial \omega}x, \qquad \omega \in (0, L),\ t > 0. \tag{65}$$

Here, $D > 0$ denotes the diffusion coefficient and $x(\omega, t)$ represents, e. g., the density of a biological population. The first term on the right-hand side accounts for diffusion. The second term combines the logistic nonlinearity $R(x) = x(1-x)$ with the spatial gradient to form a state-dependent drift, which is strongest at intermediate concentrations and vanishes for $x \in \{0, 1\}$.

Using a finite-difference approximation of the spatial derivatives with step $\Delta\omega > 0$ and an explicit Euler method with time step $\Delta t > 0$, we obtain the dynamics

$$\begin{aligned} X_i^+ &= X_i + \Delta t \left(D\,\frac{X_{i+1} - 2X_i + X_{i-1}}{\Delta\omega^2} + X_i(1 - X_i)\,\frac{X_{i+1} - X_{i-1}}{2\Delta\omega} \right) \\ &= f_1(X_i) + f_2(X_{i+1}) + f_3(X_{i-1}) + f_4(X_i, X_{i+1}) + f_5(X_i, X_{i-1}) \end{aligned}$$

at spatial point i, for suitably chosen f_n, $n = 1, \ldots, 5$.

We now consider two different ways of adding a control to this system, which differ in how the control enters the dynamics and, consequently, in the structure of the resulting input map G.

Gradient-coupled control. First, we add a term $(\frac{\partial x}{\partial \omega})^2 u$, which acts most strongly where the spatial gradient is steep and results in

$$X_i^+ = X_i + \Delta t \left(D \frac{X_{i+1} - 2X_i + X_{i-1}}{\Delta \omega^2} + X_i(1 - X_i) \frac{X_{i+1} - X_{i-1}}{2\Delta \omega} + \left(\frac{X_{i+1} - X_{i-1}}{2\Delta \omega} \right)^2 U_i \right). \tag{66}$$

For this example, the threshold ε for the global sensitivity indices would have to be chosen considerably smaller to capture all relevant terms, see Figure 4, where the indices are renormalized within each group. The weak couplings $\{1\}$ and $\{3\}$ of G_1 have combined indices of $2.2 \cdot 10^{-3}$ each, which is below $\varepsilon = 1/73 \approx 1.4 \cdot 10^{-2}$, such that they are truncated in the second least squares step. As a result, the error $e = 8.87 \cdot 10^{-2}$ is considerably larger than for the academic test function.

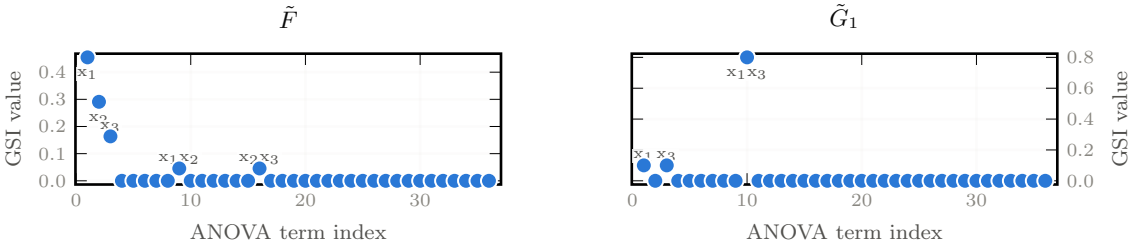


FIGURE 4. Approximated global sensitivity indices for the *diffusion equation with gradient-coupled control* (66), for $i = 2$ and $D = \Delta t = \Delta \omega = 1$, where the relevant terms are $\{1\}, \{2\}, \{3\}, \{1,2\}, \{2,3\}$ of F and $\{1\}, \{3\}, \{1,3\}$ of G_1.

Controlled nonlinear advection. Second, we multiply the advection term by u, i. e., we consider

$$X_i^+ = X_i + \Delta t \left(D \frac{X_{i+1} - 2X_i + X_{i-1}}{\Delta \omega^2} + X_i(1 - X_i) \frac{X_{i+1} - X_{i-1}}{2\Delta \omega} U_i \right). \tag{67}$$

Here, the advection term is multiplied by the control, such that the second-order interactions move from F to G_1. Consequently, the relevant terms of F reduce to the three one-dimensional diffusion terms, while G_1 contains the pair couplings $\{1,2\}$ and $\{2,3\}$ together with the weaker terms $\{1\}$ and $\{3\}$, see Figure 5. As for (66), these weak one-dimensional terms have combined indices of $9.7 \cdot 10^{-3}$ each, fall below the threshold $\varepsilon = 1/73$, and are therefore truncated. As a result, we obtain an error $e = 1.76 \cdot 10^{-1}$.

Finally, Figure 6 shows the influence of superposition dimension q and number of samples M for the two controlled systems (66) and (67). Both require $q \geq 2$, since both dynamics contain pair interactions of neighbouring grid points. Hence, for $q = 1$, the relevant pair interactions cannot be represented and the error stagnates at order 10^{-1}. At $q = 2$, however, the error remains large, since the weak couplings of G discussed above fall below the sparsity threshold and are truncated. For larger q, the threshold $\varepsilon = 1/|V_{\text{cand}}|$ decreases as the number of candidate terms increases, such that the weak couplings are eventually retained and the error drops to the solver accuracy, for (67) at $q = 3$ and for (66) at $q = 6$. The same effect renders e independent of M. It settles at $\approx 9.1 \cdot 10^{-2}$ and $\approx 1.9 \cdot 10^{-1}$ for (66) and (67), respectively, and does not change up to $M = 10^5$, since increasing the number of samples cannot compensate for a term that is not contained in the ansatz space.

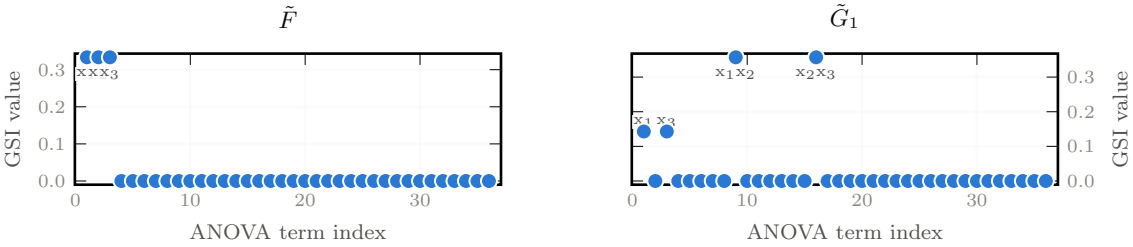


FIGURE 5. Approximated global sensitivity indices for the *diffusion equation with controlled nonlinear advection* (67), for $i = 2$ and $D = \Delta t = \Delta\omega = 1$, where the relevant terms are $\{1\}, \{2\}, \{3\}$ of F and $\{1\}, \{3\}, \{1,2\}, \{2,3\}$ of G_1.

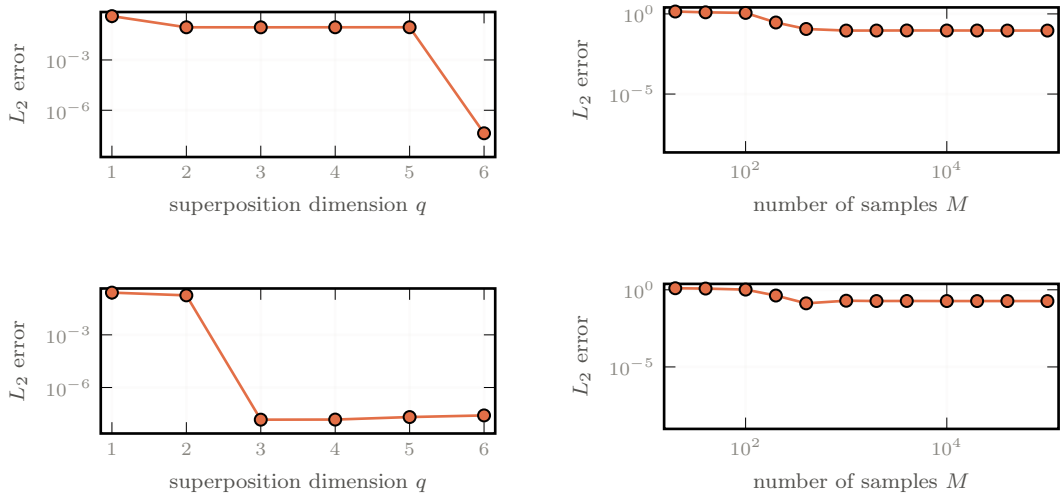


FIGURE 6. Error e for the *diffusion equation with the two control variants* (66) *and* (67) (top to bottom), for $i = 2$, depending on the superposition dimension q (left) and the number of samples M (right).

4.3. **Shallow water equations**

The 1D shallow water equations, see [17, Chapter 13], form a nonlinear system of conservation laws that describes the conservation of mass and momentum. They are given by

$$\begin{aligned} \frac{\partial h}{\partial t} + \frac{\partial (hv)}{\partial \omega} &= 0 \qquad \text{(mass conservation)} \\ \frac{\partial (hv)}{\partial t} + \frac{\partial}{\partial \omega}\left(hv^2 + \frac{1}{2}gh^2\right) &= 0 \qquad \text{(momentum conservation).} \end{aligned} \tag{68}$$

Here, g is the gravitational acceleration, while $h(\omega, t)$ and $v(\omega, t)$ denote the fluid depth and the horizontal velocity, respectively, for $(\omega, t) \in \Omega \times [0, \infty)$. Discretization of the spatial domain Ω, for instance using a Lax–Friedrichs scheme[3] with an equidistant spatial grid consisting of the nodes $\omega_1, \ldots, \omega_n$ with grid size $\Delta\omega$, yields

[3]More advanced finite-volume solvers, e. g., Godunov's method or Riemann solvers [17], lead to a similar structure.

the explicit time stepping scheme

$$\mathbf{Z}_i^+ = \frac{1}{2}\left(\mathbf{Z}_{i+1} + \mathbf{Z}_{i-1}\right) - \tfrac{\Delta t}{2\Delta\omega}\left(\mathbf{F}_{i+1} - \mathbf{F}_{i-1}\right), \tag{69}$$

with temporal step size Δt. The quantities $\mathbf{Z}_i$ and $\mathbf{F}_i$ approximate the conserved quantity and the nonlinearity at space point ω_i, $i \in \{1, \dots, n\}$, respectively, i. e.,

$$\mathbf{Z}_i = \begin{pmatrix} h(\omega_i) \\ h(\omega_i)v(\omega_i) \end{pmatrix} \qquad \text{and} \qquad \mathbf{F}_i = \begin{pmatrix} h(\omega_i)v(\omega_i) \\ h(\omega_i)v(\omega_i)^2 + \frac{1}{2}gh(\omega_i)^2 \end{pmatrix}.$$

Rewriting this scheme in terms of the variables $x = (h, q)$ with the conserved quantity $q = hv$, and denoting its numerical approximation by X, we obtain the discrete-time system

$$X_i^+ = \frac{1}{2}(X_{i+1} + X_{i-1}) - \tfrac{\Delta t}{2\Delta\omega}\left(\mathbf{F}(X_{i+1}) - \mathbf{F}(X_{i-1})\right) = f_2(X_{i+1}) + f_1(X_{i-1})$$

for the spatial grid point i, where

$$\mathbf{F}(x) = \mathbf{F}(h, q) = \begin{pmatrix} q \\ q^2/h + \frac{1}{2}gh^2 \end{pmatrix}, \quad f_1(x) = \frac{1}{2}x + \frac{\Delta t}{2\Delta\omega}\mathbf{F}(x), \quad f_2(x) = \frac{1}{2}x - \frac{\Delta t}{2\Delta\omega}\mathbf{F}(x).$$

At first sight, the two summands look like fully decoupled terms. However, X is composed of the height h and the discharge $q = hv$, and is thus two-dimensional. Applying our scalar-valued approximation approach, we obtain nonlinearly coupled terms, since, e. g., the second line reads

$$Q_i^+ = \frac{1}{2}(Q_{i+1} + Q_{i-1}) - \frac{\Delta t}{2\Delta\omega}\left(Q_{i+1}^2/H_{i+1} + \frac{1}{2}gH_{i+1}^2 - Q_{i-1}^2/H_{i-1} - \frac{1}{2}gH_{i-1}^2\right). \tag{70}$$

Clearly, its right-hand side may be written in terms of $z = (Q_{i+1}, Q_{i-1}, H_{i+1}, H_{i-1})$ and thus involves (multiplicative) couplings of (Q_{i+1}, H_{i+1}) and (Q_{i-1}, H_{i-1}).

Adding control. An acceleration u (e. g., from a pump or turbine) enters the second equation through a term hu. The factor h makes the momentum source stronger where the water is deeper. It also ensures dimensional consistency, since $[h] = \mathrm{m}$ and $[u] = \mathrm{m/s^2}$ give $[hu] = \mathrm{m^2/s^2}$, which matches $[\partial_t q]$ for the discharge per unit width $q = hv$. In the discretization, this amounts to adding the term

$$\Delta t\, H_i U_i$$

to the right-hand side of (70).

Concerning the sampling, note that the flux term Q^2/H is singular at $H = 0$. We therefore draw all eight components on $[-1, 1]$ as before. For the height components, we then use the shifted values

$$H_i \coloneqq X_i + 2 \in [1, 3], \qquad \text{where } i \text{ is a height index}, \tag{71}$$

so that the water height stays bounded away from zero. The discharge components Q_i are used without shifting on $[-1, 1]$.

In contrast to the previous examples, the flux term Q^2/H is not a polynomial, so the ANOVA approximation cannot reproduce the dynamics exactly. The error (64) is $e = 2.83 \cdot 10^{-2}$. This is also reflected in Figure 8, where the two panels show plateaus at approximately $2.8 \cdot 10^{-2}$ and $3.0 \cdot 10^{-2}$, respectively.

The influence of q and M is displayed in Figure 8. In contrast to the previous examples, a non-vanishing error of about $2.8 \cdot 10^{-2}$ remains for *all* parameter choices. In the following, we identify its source among the three error contributions of Remark 2.7. It is not a superposition truncation error, since the flux term depends

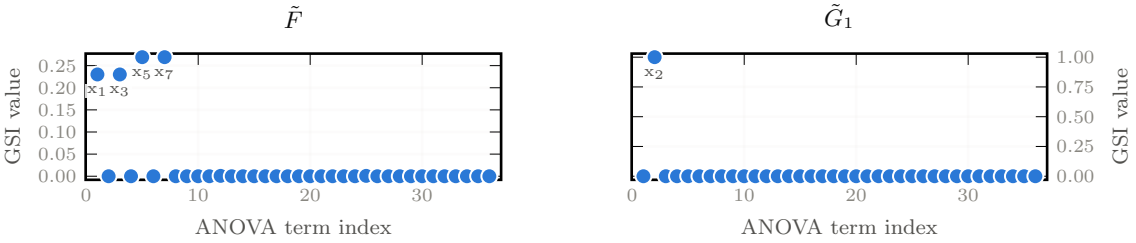


FIGURE 7. Approximated global sensitivity indices for the *shallow water equations* (70), for $i = 2$, $X = (H_1, H_2, H_3, H_4, Q_1, Q_2, Q_3, Q_4)$, $\Delta t = g = 1$, and $\Delta\omega = 2$, with the heights sampled from $[1, 3]$ (affinely transformed to $[-1, 1]$), where the dominant terms are $\{1\}, \{3\}, \{5\}, \{7\}$ of F and $\{2\}$ of G_1.

on two variables only, such that the truncation by q is exact for $q \geq 2$. It is not a projection error either. Expanding the row in the Chebyshev basis, the projection error at $q = 2$ decreases from $2.8 \cdot 10^{-2}$ for $N_2 = 2$ to $2.8 \cdot 10^{-6}$ for $\boldsymbol{N} = (20, 8)$, since $1/H$ is analytic on $[1, 3]$. However, the measured error stays at $2.8 \cdot 10^{-2}$ for all admissible bandwidths. Instead, the plateau stems from the truncation by the sensitivity threshold. The flux-induced pair couplings $\{H_{i\pm1}, Q_{i\pm1}\}$ carry sensitivity indices of $8.5 \cdot 10^{-4}$ each, fall below $\varepsilon = 1/73$ and are hence discarded, cf. Figure 7. Removing these two terms from the exact expansion leaves an error of $2.84 \cdot 10^{-2}$, which is independent of $\boldsymbol{N}$ by construction and matches the measured value. Hence, the error floor of this example is governed by the threshold ε and not by the bandwidths. This is a second instance of the effect already observed for (66) and (67).

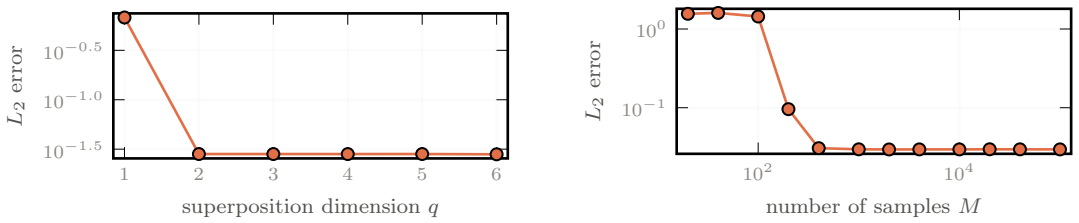


FIGURE 8. Error e for the *shallow water equations* (70), for $i = 2$, depending on the superposition dimension q (left) and the number of samples M (right).

4.4. **Viscous Burgers equation**

As a final example, we return to the controlled viscous Burgers equation of Example 3.2. Unlike the previous examples, it features a multivariate control with $m = 2$, namely the state-affine actuation $\gamma(\omega)\, x\, U_1$ and the purely additive actuation $b(\omega)\, U_2$. Recall from Example 3.2 that the discretization yields the dynamics

$$X_i^+ = X_i + \Delta t \left(D\, \frac{X_{i+1} - 2X_i + X_{i-1}}{\Delta\omega^2} - X_i\, \frac{X_{i+1} - X_{i-1}}{2\Delta\omega} + \gamma_i\, X_i\, U_1 + b_i\, U_2 \right) \tag{72}$$

for spatial point i, i. e., the index sets to be identified are $V_{i,0} = \{\{i-1\}, \{i\}, \{i+1\}, \{i-1, i\}, \{i, i+1\}\}$, $V_{i,1} = \{\{i\}\}$, and $V_{i,2} = \{\varnothing\}$. As before, we consider the row $i = 2$ with $d = 8$ and choose $D = \Delta t = \Delta\omega = 1$ as well as $\gamma_i = b_i = 1$.

The results are shown in Figure 9. The relevant ANOVA terms of F and G_1 are correctly identified. For G_2, the situation is degenerate. Since G_2 consists of the constant term only, all terms coupling u_2 to the state have

vanishing global sensitivity indices, which are of order 10^{-18} in the computation, such that they are correctly discarded. As these indices sum to zero up to rounding errors, they cannot be renormalized within their group, and Figure 9 shows them without renormalization. The constant actuation $b_i\, U_2$ is captured by the ANOVA term in u_2 alone, which does not belong to the displayed group. As expected, the reconstruction is accurate with $e = 8.15 \cdot 10^{-9}$.

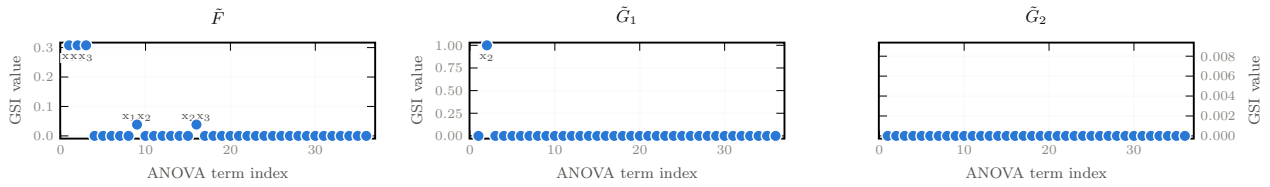


FIGURE 9. Approximated global sensitivity indices for the *viscous Burgers equation* (72), for $i = 2$ and $D = \Delta t = \Delta\omega = \gamma_i = b_i = 1$, where the relevant terms are $\{1\}, \{2\}, \{3\}, \{1,2\}, \{2,3\}$ of F (left), $\{2\}$ of G_1 (middle) and $\varnothing$ of G_2 (right).

Finally, Figure 10 shows the same parameter study for this example with two controls. The behaviour matches the one of the academic test function. The truncation error vanishes for $q \geq 2$ and the estimation error drops below 10^{-6} once $M \gtrsim 300$, since the right-hand side of (72) is polynomial and hence contained in the ansatz space. The additional control component does not change the qualitative behaviour but only increases the number of coefficients to be determined.

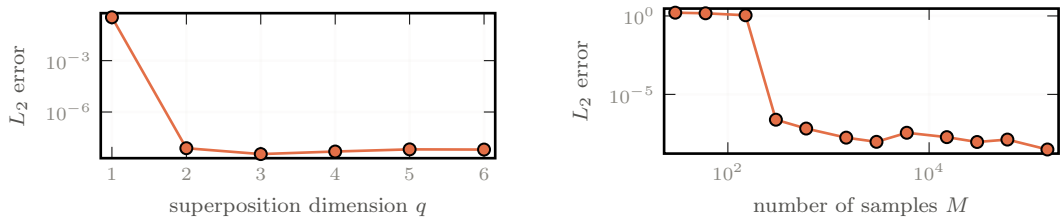


FIGURE 10. Error e for the *viscous Burgers equation* (72) with $i = 2$ and $m = 2$ depending on the superposition dimension q (left) and the number of samples M (right).

5. CONCLUSIONS

We have extended the ANOVA approximation to control-affine systems $\boldsymbol{x}^+ = F(\boldsymbol{x}) + G(\boldsymbol{x})\boldsymbol{u}$ whose components depend on only a few state variables, such that no effort is spent on interactions between control components, which are excluded a priori by the control-affine structure. To this end, we have encoded the control-affine structure into the approximation ansatz itself, which allows to work with given state-control samples. In Proposition 3.4, we have characterized the limit for $M \to \infty$ as an orthogonal projection onto the span of the dictionary consisting of the state basis functions and their products with the control components. For finitely many samples, we have established a Marcinkiewicz–Zygmund inequality for the resulting ansatz space in Proposition 3.7. Based on it, we have derived a probabilistic L_2-error bound with explicit constants in Theorem 3.8 and Corollary 3.13, which also controls the errors of the drift and of the input map separately. Finally, we have illustrated the approach by means of an academic test function and discretizations of a nonlinear advection-diffusion equation, the shallow water equations, and the viscous Burgers equation. In all examples, the dominant couplings are identified, and for polynomial dynamics the error reaches the accuracy of the least squares solver, provided that the relevant terms are not discarded by the sparsity threshold.

Acknowledgments

The authors would like to express their sincere gratitude to their colleague Laura Weidensager for several fruitful discussions. Melanie Kircheis gratefully acknowledges the support of the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 519323897.

Declaration on the use of AI tools

During the preparation of this work, the authors used Claude (Anthropic, model Claude Fable 5.1 and Opus 5.5) for the language editing, the discussion of proof ideas, proof reading and the documentation of code. The authors have reviewed, verified and revised all AI-assisted content and take full responsibility for the content of this publication.

Data Availability Statement

The code associated with this article is available at

`https://gitlab.hrz.tu-chemnitz.de/manuels-at-tu-chemnitz.de/anova-for-control`,

see [14].

References

[1] F. Bartel, K. Lüttgen, N. Nagel, and T. Ullrich. Efficient recovery of non-periodic multivariate functions from few scattered samples. In *14th International Conference on Sampling Theory and Applications (SampTA)*, pages 1–6, 2023. doi:10.1109/SampTA59647.2023.10301396

[2] L. Bold, F. M. Philipp, M. Schaller, and K. Worthmann. Kernel-based Koopman approximants for control: flexible sampling, error analysis, and stability. *SIAM J. Control Optim.*, 63(6):4044–4071, 2025. doi:10.1137/24M1715945

[3] S. L. Brunton, J. L. Proctor, and J. N. Kutz. Sparse identification of nonlinear dynamics with control (SINDYc). *IFAC-PapersOnLine*, 49(18):710–715, 2016. doi:10.1016/j.ifacol.2016.10.249

[4] S. L. Brunton, J. L. Proctor, and J. N. Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proc. Natl. Acad. Sci. USA*, 113(15):3932–3937, 2016. doi:10.1073/pnas.1517384113

[5] F. Bullo and A. D. Lewis. *Geometric control of mechanical systems: modeling, analysis, and design for simple mechanical control systems*, volume 49 of *Texts in Applied Mathematics*. Springer, New York, 2005. doi:10.1007/978-1-4899-7276-7

[6] R. E. Caflisch, W. Morokoff, and A. B. Owen. Valuation of mortgage-backed securities using Brownian bridges to reduce effective dimension. *J. Comput. Finance*, 1(1):27–46, 1997. doi:10.21314/JCF.1997.005

[7] A. Cohen, M. A. Davenport, and D. Leviatan. On the stability and accuracy of least squares approximations. *Found. Comput. Math.*, 13(5):819–834, 2013. doi:10.1007/s10208-013-9142-3

[8] A. Cohen and G. Migliorati. Optimal weighted least-squares methods. *SMAI J. Comput. Math.*, 3:181–203, 2017. doi:10.5802/smai-jcm.24

[9] R. DeVore, G. Petrova, and P. Wojtaszczyk. Approximation of functions of few variables in high dimensions. *Constr. Approx.*, 33:125–143, 2011. doi:10.1007/s00365-010-9105-8

[10] J. Eggl, E. Mindlberger, and M. Ullrich. Sparse grids vs. random points for high-dimensional polynomial approximation. *Adv. Comput. Math.*, 52(3), 2026. Paper No. 44. doi:10.1007/s10444-026-10316-5

[11] M. Hinze, R. Pinnau, M. Ulbrich, and S. Ulbrich. *Optimization with PDE constraints*, volume 23 of *Mathematical Modelling: Theory and Applications*. Springer, Dordrecht, 2009. doi:10.1007/978-1-4020-8839-1

[12] W. Hoeffding. Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, 58(301):13–30, 1963. doi:10.1080/01621459.1963.10500830

[13] L. Kämmerer, T. Ullrich, and T. Volkmer. Worst case recovery guarantees for least squares approximation using random samples. *Constr. Approx.*, 54:295–352, 2021. doi:10.1007/s00365-021-09555-0

[14] M. Kircheis, D. Potts, M. Schaller, and K. Worthmann. Julia code for "ANOVA approximation for high-dimensional control systems". `https://gitlab.hrz.tu-chemnitz.de/manuels-at-tu-chemnitz.de/anova-for-control`, 2026.

[15] F. Köhne, F. M. Philipp, M. Schaller, A. Schiela, and K. Worthmann. L^∞-error bounds for approximations of the Koopman operator by kernel extended dynamic mode decomposition. *SIAM J. Appl. Dyn. Syst.*, 24(1):501–529, 2025. doi:10.1137/24M1650120

[16] F. Y. Kuo, I. H. Sloan, G. W. Wasilkowski, and H. Woźniakowski. On decompositions of multivariate functions. *Math. Comp.*, 79(270):953–966, 2010. doi:10.1090/S0025-5718-09-02319-9

[17] R. J. LeVeque. *Finite volume methods for hyperbolic problems*, volume 31 of *Cambridge Texts in Applied Mathematics*. Cambridge University Press, Cambridge, 2002. doi:10.1017/CBO9780511791253
[18] R. Liu and A. B. Owen. Estimating mean dimensionality of analysis of variance decompositions. *J. Amer. Statist. Assoc.*, 101(474):712–721, 2006. doi:10.1198/016214505000001410
[19] D. C. Montgomery. *Design and analysis of experiments*. John Wiley & Sons, Hoboken, NJ, 9th edition, 2017.
[20] C. C. Paige and M. A. Saunders. LSQR: an algorithm for sparse linear equations and sparse least squares. *ACM Trans. Math. Software*, 8:43–71, 1982. doi:10.1145/355984.355989
[21] S. Peitz, S. E. Otto, and C. W. Rowley. Data-driven model predictive control using interpolated Koopman generators. *SIAM J. Appl. Dyn. Syst.*, 19(3):2162–2193, 2020. doi:10.1137/20M1325678
[22] D. Potts and M. Schmischke. Approximation of high-dimensional periodic functions with Fourier-based methods. *SIAM J. Numer. Anal.*, 59(5):2393–2429, 2021. doi:10.1137/20M1354921
[23] D. Potts and M. Schmischke. Interpretable approximation of high-dimensional data. *SIAM J. Math. Data Sci.*, 3(4):1301–1323, 2021. doi:10.1137/21M1407707
[24] D. Potts and L. Weidensager. ANOVA-boosting for random Fourier features. *Appl. Comput. Harmon. Anal.*, 79:101789, 2025. doi:10.1016/j.acha.2025.101789
[25] H. Rabitz and Ö. F. Aliş. General foundations of high-dimensional model representations. *J. Math. Chem.*, 25:197–233, 1999. doi:10.1023/A:1019188517934
[26] I. Schimperna, K. Worthmann, M. Schaller, L. Bold, and L. Magni. Data-driven model predictive control: asymptotic stability despite approximation errors exemplified in the Koopman framework. *Automatica*, 193:113203, 2026. doi:10.1016/j.automatica.2026.113203
[27] M. Schmischke. ANOVAapprox.jl – Julia package for ANOVA approximation. https://github.com/NFFT/ANOVAapprox.jl. Contributors: F. Bartel, F. Nestler, P. Schröter, L. Weidensager. doi:10.5281/zenodo.5657976
[28] M. Schmischke. *Interpretable approximation of high-dimensional data based on the ANOVA decomposition*. PhD thesis, Chemnitz University of Technology, 2022.
[29] P. Schröter. pyANOVAapprox – Python package for ANOVA approximation. https://github.com/NFFT/pyANOVAapprox. Contributors: F. Bartel, F. Nestler, L. Weidensager.
[30] I. M. Sobol. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Math. Comput. Simulation*, 55(1–3):271–280, 2001. doi:10.1016/S0378-4754(00)00270-6
[31] J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.*, 12(4):389–434, 2012. doi:10.1007/s10208-011-9099-z
[32] M. O. Williams, I. G. Kevrekidis, and C. W. Rowley. A data-driven approximation of the Koopman operator: extending dynamic mode decomposition. *J. Nonlinear Sci.*, 25(6):1307–1346, 2015. doi:10.1007/s00332-015-9258-5

Appendix A. Implementation via preprocessed ANOVA index sets

In this part, we show that the coupled approach may be realized entirely within the existing ANOVA approximation software [27, 29], that is, without implementing the coupled least squares problem (40) from scratch. To this end, we choose the ANOVA index sets (15) and (22) for the joint function $h(\boldsymbol{z}) = H_j(\boldsymbol{x}, \boldsymbol{u})$ a priori such that they enforce the structure (34) and (58), respectively. At the same time, this avoids the computation of ANOVA terms that vanish due to the control-affine structure.

Accordingly, set

$$h(\boldsymbol{z}) \coloneqq H_j(\boldsymbol{x}, \boldsymbol{u}), \quad \boldsymbol{z} \coloneqq (\boldsymbol{x}, \boldsymbol{u}) \in \mathbb{R}^{d+m}.$$

The naive way of applying the machinery of Section 2 to the function h is to consider its full ANOVA decomposition (1), i. e.,

$$h(\boldsymbol{z}) = \sum_{\boldsymbol{v} \subseteq [d+m]} h_{\boldsymbol{v}}(\boldsymbol{z}_{\boldsymbol{v}}).$$

However, this decomposition consists of 2^{d+m} ANOVA terms instead of the $2^d + 2^d m$ terms that would be necessary for the theoretical scenario of approximating F and (the columns of) G separately. This is inefficient because all terms coupling two or more components of $\boldsymbol{u}$ vanish a priori by the control-affine structure of (31). Hence, computing these couplings between the u_ℓ is unnecessary. Moreover, noisy sensitivity indices could incorrectly identify such terms as relevant. Further, despite the terms containing the control variable have to be *linear* in u_ℓ, a generic ANOVA term may depend on u_ℓ in an arbitrary nonlinear fashion.

In the following, we address these two issues by means of a suitable a priori choice of the index sets (15) and (22).

We first address the number of ANOVA terms. To this end, recall that in the standard ANOVA approximation, the first step to reduce the number of ANOVA terms is to introduce a superposition dimension $q \in \mathbb{N}$ and to consider only the terms of lower order, i. e., all terms with $|\boldsymbol{v}| \leq q$, cf. (14). This default is chosen since usually no a priori information about the structure of the given function is available. Here, however, we already know that the system (31) depends linearly on the individual components of $\boldsymbol{u}$, i. e., that there are no couplings between the u_ℓ, $\ell = 1, \dots, m$. Therefore, we can encode this structural knowledge at the very beginning by providing a suitable index set V of relevant ANOVA terms, cf. (22), and computing an approximation of the form

$$h(\boldsymbol{z}) = \sum_{\boldsymbol{k} \in \mathcal{I}(V)} c_{\boldsymbol{k}}(h) \varphi_{\boldsymbol{k}}(\boldsymbol{z})$$

with $\mathcal{I}(V)$ in (23). Thereby, we avoid computing terms that vanish a priori. More precisely, with this choice of the index set, we obtain the 2^d terms purely in $\boldsymbol{x}$ and, for each u_ℓ, its 2^d couplings with $\boldsymbol{x}$. In other words, we work with $2^d + 2^d m$ ANOVA terms, i. e., the same number as in the theoretical scenario of approximating F and G separately.

In practice, we again introduce a superposition dimension q to further reduce the number of terms. If the terms with and without a control component are to admit the same couplings purely in $\boldsymbol{x}$, we have to start with $q+1$, since couplings with u_ℓ are one order higher. We then remove couplings purely in $\boldsymbol{x}$ with $|\boldsymbol{v}| = q+1$ to obtain an ANOVA decomposition of the form

$$h(\boldsymbol{z}) \approx \sum_{\substack{\boldsymbol{v} \subseteq [d] \\ |\boldsymbol{v}| \leq q}} h_{\boldsymbol{v}}(\boldsymbol{x}_{\boldsymbol{v}}) + \sum_{\ell=1}^{m} \sum_{\substack{\boldsymbol{v} \subseteq [d] \\ |\boldsymbol{v}| \leq q}} h_{\boldsymbol{v} \cup \{d+\ell\}}(\boldsymbol{x}_{\boldsymbol{v}}, u_\ell),$$

i. e., terms purely in $\boldsymbol{x}$ up to order q as well as terms of order at most $q+1$ containing exactly one control component. The terms involving two or more control components vanish by the control-affine structure. Consequently, the above is an equality whenever all index sets in (32) satisfy $|\boldsymbol{v}| \leq q$. The converse does not hold in general, since the $f_{\boldsymbol{v}}^j$ in (32) form a structural rather than an ANOVA decomposition. Therefore, an index set with $|\boldsymbol{v}| > q$ may still yield vanishing ANOVA terms of order larger than q.

Note, however, that we have not yet achieved the desired structure (34), since linearity in the u_ℓ is still missing. For general basis functions and index sets (15), the terms $h_{\boldsymbol{v} \cup \{d+\ell\}}(\boldsymbol{x}_{\boldsymbol{v}}, u_\ell)$ may depend on u_ℓ in an arbitrary nonlinear fashion. For instance, already the univariate term $h_{\{d+\ell\}}(u_\ell)$ is in general not proportional to u_ℓ.

A closer look at the definition of the index sets (15) reveals that the ANOVA approximation allows to specify the number of coefficients and basis functions for each individual ANOVA term. More precisely, by choosing $N_j^{\boldsymbol{v}} = 2$ for a certain set of indices $\boldsymbol{v}$, the index set $\mathcal{I}_{\boldsymbol{N}^{\boldsymbol{v}}}$ contains only the value 1 for the corresponding component. Hence, we use only one coefficient c_1 from $\{c_{\boldsymbol{k}}\}$ and one basis function φ_1 from $\{\varphi_{\boldsymbol{k}}\}$.

In addition, note that for the Chebyshev basis (6) we have $\{\varphi_k^{\text{cheb}}\}_{k \in \mathbb{N}_0} = \{1, \sqrt{2}x, \dots\}$. Thus, combining the above restriction of the index sets with the structure of the Chebyshev basis, we reconstruct the component u_ℓ exactly by setting $N_j^{\boldsymbol{v}} = 2$ for all index sets $\boldsymbol{v}$ containing one of the u_ℓ. The normalization constant of φ_1 is then absorbed into the corresponding coefficients. Note that this restriction only concerns the control variables. In the $\boldsymbol{x}$-directions, we are free to choose any orthonormal basis as introduced in Section 2, and we may even employ several different bases depending on the structure of F and G.

In summary, the strategy described above realizes the same approximation (34) as in Section 3.1, provided that the same basis systems $\varphi^1 = \varphi^2$ with possibly different index sets $\mathcal{I}(V_1)$ and $\mathcal{I}(V_2)$ are used. The approach of Section 3.1 is thus slightly more general, since it admits different basis systems for F and G, whereas the present strategy has the advantage that it may be carried out entirely within the existing ANOVA software [27, 29].

Algorithm A.1 (ANOVA approximation for control via preprocessed index sets)**.**

Input:	$d, m \in \mathbb{N}$	*dimension and control dimension*
	$q \in \mathbb{N}$, $q \ll d$	*superposition dimension*
	$\{(\boldsymbol{x}^i, \boldsymbol{u}^i)\}_{i=1}^M \subset \Omega \times \mathbb{U}$, $M \in \mathbb{N}$	*i.i.d. sample points drawn according to the density* $\omega \cdot \widetilde{\omega}$
	$\big(H(\boldsymbol{x}^i, \boldsymbol{u}^i)\big)_{i=1}^M$	*function evaluations at the sample points*
	$\varepsilon \in (0,1)$	*threshold parameter*
	$r > 0$	*confidence parameter, cf. Theorem 2.6*

(1) Set $V = \{\boldsymbol{v} \subseteq [d] \colon |\boldsymbol{v}| \le q\} \cup \{\boldsymbol{v} \cup \{d+\ell\} \colon \boldsymbol{v} \subseteq [d],\, |\boldsymbol{v}| \le q,\, \ell = 1, \dots, m\}$.

(2) Choose the bandwidths $\boldsymbol{N}^{\boldsymbol{v}}$ *in* (15) *with* $N_j^{\boldsymbol{v}} = 2$ *for the component corresponding to* u_ℓ *in every* $\boldsymbol{v} \ni d+\ell$*, using the Chebyshev basis in the control variables such that* $\varphi_1(u_\ell) \propto u_\ell$*, cf. Remark A.2.*

(3) For $j = 1, \dots, d$*: perform steps 2–5 of Algorithm 2.4 for* $h(\boldsymbol{z}) = H_j(\boldsymbol{x}, \boldsymbol{u})$ *with the index set* $\mathcal{I}(V)$ *from steps 1–2, where the truncation preserves the structure of* V *(at most one control component per term, bandwidth* 2 *in the control directions).*

Output: *coefficients, where the terms with* $\boldsymbol{v} \subseteq [d]$ *form* $\tilde{F}_j$ *and the terms with* $d + \ell \in \boldsymbol{v}$ *form* $\tilde{G}_{j\ell}$

We briefly comment on the second step, which is more restrictive than it may seem at first sight.

Remark A.2 (The control measure fixed by the preprocessing)**.** *Using the Chebyshev basis in the control variables fixes the control measure, since the ANOVA approximation of Section 2 requires the basis to be orthonormal with respect to the sampling density. Hence,* $\mathbb{U}_\ell = [-1, 1]$ *is equipped with the Chebyshev density* $\omega_{d+\ell} = \omega^{\mathrm{cheb}}$*, for which* $\varphi_1(u_\ell) = \sqrt{2}\, u_\ell$*. Step 2 of Algorithm A.1 is therefore a* specialization *of the setting of Section 3.1, where* $\widetilde{\omega}$ *was arbitrary. It yields* $\eta_\ell = 0$ *and* $\nu_\ell = \frac{1}{2}$*, that is, the zero-mean situation in which the limit systems of Remarks 3.5 and 3.6 become block diagonal. Note that a different control measure remains admissible. However, the control basis then has to be orthonormalized accordingly, since the ANOVA partition in the* $\boldsymbol{u}$*-direction requires an orthonormal system in* $L_2(\mathbb{U}_\ell, \omega_{d+\ell})$ *whose first non-constant element is proportional to* u_ℓ*. In particular, the control measure has to have zero mean.*

We close this appendix by illustrating the preprocessing strategy of Algorithm A.1 with two examples of increasing complexity.

Example A.3. *As a first example, we consider again the Euler approximation of a controlled Duffing oscillator* (33) *from Example 3.1. Here, we have* $d = 2$ *and* $m = 1$*, i. e., the coupled variable is* $\boldsymbol{z} = (x_1, x_2, u)^\top \in \mathbb{R}^3$ *with* $z_1 = x_1$*,* $z_2 = x_2$*, and* $z_3 = u$*. Since the control is univariate, no unnecessary couplings between control components can occur. We use this simple example to illustrate the proposed choice of the index sets and of the basis. Additionally, as the dimension* $d + m = 3$ *is already small, we do not introduce a superposition dimension* $q < 3$*, but keep all possible ANOVA terms, i. e.,* $V = \{\varnothing, \{1\}, \{2\}, \{3\}, \{1,2\}, \{1,3\}, \{2,3\}, \{1,2,3\}\}$*. Following the preprocessing strategy described above, the bandwidths* (15) *are then chosen as* $N_j^{\boldsymbol{v}} = 2$ *for the component corresponding to the control variable* $z_3 = u$ *and arbitrary even numbers otherwise, i. e., with arbitrary* $a_n \in 2\mathbb{N}$*,* $n = 1, \dots, 8$*, we set*

$$\boldsymbol{N}^{\{1\}} = a_1, \boldsymbol{N}^{\{2\}} = a_2, \boldsymbol{N}^{\{3\}} = 2, \boldsymbol{N}^{\{1,2\}} = \begin{pmatrix} a_3 \\ a_4 \end{pmatrix}, \boldsymbol{N}^{\{1,3\}} = \begin{pmatrix} a_5 \\ 2 \end{pmatrix}, \boldsymbol{N}^{\{2,3\}} = \begin{pmatrix} a_6 \\ 2 \end{pmatrix}, \boldsymbol{N}^{\{1,2,3\}} = \begin{pmatrix} a_7 \\ a_8 \\ 2 \end{pmatrix}.$$

The corresponding index sets (15) *then read*

$$\mathcal{I}_{\boldsymbol{N}^{\{1\}}} = \big\{\boldsymbol{k} = (k_1, 0, 0)^\top \colon k_1 \in \{1, \dots, a_1 - 1\}\big\},$$
$$\mathcal{I}_{\boldsymbol{N}^{\{2\}}} = \big\{\boldsymbol{k} = (0, k_2, 0)^\top \colon k_2 \in \{1, \dots, a_2 - 1\}\big\},$$

$$
\begin{aligned}
\mathcal{I}_{\boldsymbol{N}^{\{3\}}} &= \left\{\boldsymbol{k} = (0,0,k_3)^\top : k_3 \in \{1\}\right\} = \left\{(0,0,1)^\top\right\},\\
\mathcal{I}_{\boldsymbol{N}^{\{1,2\}}} &= \left\{\boldsymbol{k} = (k_1,k_2,0)^\top : k_1 \in \{1,\dots,a_3-1\}, k_2 \in \{1,\dots,a_4-1\}\right\},\\
\mathcal{I}_{\boldsymbol{N}^{\{1,3\}}} &= \left\{\boldsymbol{k} = (k_1,0,1)^\top : k_1 \in \{1,\dots,a_5-1\}\right\},\\
\mathcal{I}_{\boldsymbol{N}^{\{2,3\}}} &= \left\{\boldsymbol{k} = (0,k_2,1)^\top : k_2 \in \{1,\dots,a_6-1\}\right\},\\
\mathcal{I}_{\boldsymbol{N}^{\{1,2,3\}}} &= \left\{\boldsymbol{k} = (k_1,k_2,1)^\top : k_1 \in \{1,\dots,a_7-1\}, k_2 \in \{1,\dots,a_8-1\}\right\}.
\end{aligned}
$$

Denoting the resulting approximation of the component $h(\boldsymbol{z}) = H_j(\boldsymbol{x},u)$ by K and using the Chebyshev basis in the control variable, such that $\varphi_1(u) \propto u$ (where the normalization constant is absorbed into the coefficients), the ansatz takes the form

$$
\begin{aligned}
K(\boldsymbol{z}) = c_{\boldsymbol{0}}(K) &+ \sum_{\boldsymbol{k}\in\mathcal{I}_{\boldsymbol{N}^{\{1\}}}} c_{\boldsymbol{k}}(K)\,\varphi_{k_1}(x_1) + \sum_{\boldsymbol{k}\in\mathcal{I}_{\boldsymbol{N}^{\{2\}}}} c_{\boldsymbol{k}}(K)\,\varphi_{k_2}(x_2) + \sum_{\boldsymbol{k}\in\mathcal{I}_{\boldsymbol{N}^{\{1,2\}}}} c_{\boldsymbol{k}}(K)\,\varphi_{k_1}(x_1)\,\varphi_{k_2}(x_2)\\
&+ c_{(0,0,1)^\top}(K)\cdot u + \sum_{\boldsymbol{k}\in\mathcal{I}_{\boldsymbol{N}^{\{1,3\}}}} c_{\boldsymbol{k}}(K)\,\varphi_{k_1}(x_1)\cdot u + \sum_{\boldsymbol{k}\in\mathcal{I}_{\boldsymbol{N}^{\{2,3\}}}} c_{\boldsymbol{k}}(K)\,\varphi_{k_2}(x_2)\cdot u\\
&+ \sum_{\boldsymbol{k}\in\mathcal{I}_{\boldsymbol{N}^{\{1,2,3\}}}} c_{\boldsymbol{k}}(K)\,\varphi_{k_1}(x_1)\,\varphi_{k_2}(x_2)\cdot u.
\end{aligned}
$$

Defining the sets $V_1 \coloneqq \{\varnothing,\{1\},\{2\},\{1,2\}\}$ and $V_2 \coloneqq \{\{3\},\{1,3\},\{2,3\},\{1,2,3\}\}$, this can be written as

$$
K(\boldsymbol{z}) = \sum_{\boldsymbol{k}\in\mathcal{I}(V_1)} c_{\boldsymbol{k}}(K)\,\varphi_{\boldsymbol{k}}(\boldsymbol{x}) + \sum_{\boldsymbol{k}\in\mathcal{I}(V_2)} c_{\boldsymbol{k}}(K)\,\varphi_{\boldsymbol{k}}(\boldsymbol{x})\cdot u,
$$

which is the desired structure (34).

The following example is more elaborate and shows that the preprocessing indeed reduces the number of ANOVA terms.

Example A.4. *As a second example, we consider a setting in which the naive application of the ANOVA decomposition would indeed introduce unnecessary couplings between the components of $\boldsymbol{u}$. While the previous example primarily illustrated how the choice of the index sets $\boldsymbol{N}^{\boldsymbol{v}}$ in* (15) *affects the number of basis functions, this example illustrates how to choose the set V to avoid redundant ANOVA terms. To this end, we choose $d = 3$ and $m = 2$, i. e., we have $\boldsymbol{z} = (x_1,x_2,x_3,u_1,u_2)^\top$.*

Since the power set of $[d+m]$ contains $2^5 = 32$ elements, we admit only ANOVA terms of order at most 2 in $\boldsymbol{x}$. Thus, in the terminology of Algorithm A.1 we choose $q = 2$, which means that we include terms of total order at most $q+1 = 3$ in $\boldsymbol{z}$. We then manually remove all couplings with more than two components of $\boldsymbol{x}$ as well as couplings with more than one component of $\boldsymbol{u}$, i. e., we utilize

$$
\begin{aligned}
V = \{&\varnothing,\{1\},\{2\},\{3\},\{4\},\{5\},\{1,2\},\{1,3\},\{2,3\},\{1,4\},\{2,4\},\{3,4\},\\
&\{1,5\},\{2,5\},\{3,5\},\{1,2,4\},\{1,3,4\},\{2,3,4\},\{1,2,5\},\{1,3,5\},\{2,3,5\}\}.
\end{aligned}
$$

Thereby, we reduce the number of ANOVA terms from 32 to 21. Moreover, we choose the index sets $\boldsymbol{N}^{\boldsymbol{v}}$ in (15) *accordingly, that is, we restrict the number of basis functions for all sets containing a component of $\boldsymbol{u}$ by setting*

$$
\begin{aligned}
2 &= \boldsymbol{N}^{\{4\}} = \boldsymbol{N}^{\{5\}}\\
&= N_2^{\{1,4\}} = N_2^{\{2,4\}} = N_2^{\{3,4\}} = N_2^{\{1,5\}} = N_2^{\{2,5\}} = N_2^{\{3,5\}}\\
&= N_3^{\{1,2,4\}} = N_3^{\{1,3,4\}} = N_3^{\{2,3,4\}} = N_3^{\{1,2,5\}} = N_3^{\{1,3,5\}} = N_3^{\{2,3,5\}}.
\end{aligned}
$$

In analogy to the previous example, the resulting ansatz then reads

$$
\begin{aligned}
K(\boldsymbol{z}) = \Bigg(& c_{\mathbf{0}}(K) + \sum_{\boldsymbol{k}\in\mathcal{I}_{\boldsymbol{N}}\{1\}} c_{\boldsymbol{k}}(K)\,\varphi_{k_1}(x_1) + \sum_{\boldsymbol{k}\in\mathcal{I}_{\boldsymbol{N}}\{2\}} c_{\boldsymbol{k}}(K)\,\varphi_{k_2}(x_2) + \sum_{\boldsymbol{k}\in\mathcal{I}_{\boldsymbol{N}}\{3\}} c_{\boldsymbol{k}}(K)\,\varphi_{k_3}(x_3) \\
& + \sum_{\boldsymbol{k}\in\mathcal{I}_{\boldsymbol{N}}\{1,2\}} c_{\boldsymbol{k}}(K)\,\varphi_{k_1}(x_1)\,\varphi_{k_2}(x_2) + \sum_{\boldsymbol{k}\in\mathcal{I}_{\boldsymbol{N}}\{1,3\}} c_{\boldsymbol{k}}(K)\,\varphi_{k_1}(x_1)\,\varphi_{k_3}(x_3) \\
& + \sum_{\boldsymbol{k}\in\mathcal{I}_{\boldsymbol{N}}\{2,3\}} c_{\boldsymbol{k}}(K)\,\varphi_{k_2}(x_2)\,\varphi_{k_3}(x_3) \Bigg) \\
+ \Bigg(& c_{(0,0,0,1,0)^\top}(K)\cdot u_1 + \sum_{\boldsymbol{k}\in\mathcal{I}_{\boldsymbol{N}}\{1,4\}} c_{\boldsymbol{k}}(K)\,\varphi_{k_1}(x_1)\cdot u_1 + \sum_{\boldsymbol{k}\in\mathcal{I}_{\boldsymbol{N}}\{2,4\}} c_{\boldsymbol{k}}(K)\,\varphi_{k_2}(x_2)\cdot u_1 \\
& + \sum_{\boldsymbol{k}\in\mathcal{I}_{\boldsymbol{N}}\{3,4\}} c_{\boldsymbol{k}}(K)\,\varphi_{k_3}(x_3)\cdot u_1 + \sum_{\boldsymbol{k}\in\mathcal{I}_{\boldsymbol{N}}\{1,2,4\}} c_{\boldsymbol{k}}(K)\,\varphi_{k_1}(x_1)\,\varphi_{k_2}(x_2)\cdot u_1 \\
& + \sum_{\boldsymbol{k}\in\mathcal{I}_{\boldsymbol{N}}\{1,3,4\}} c_{\boldsymbol{k}}(K)\,\varphi_{k_1}(x_1)\,\varphi_{k_3}(x_3)\cdot u_1 + \sum_{\boldsymbol{k}\in\mathcal{I}_{\boldsymbol{N}}\{2,3,4\}} c_{\boldsymbol{k}}(K)\,\varphi_{k_2}(x_2)\,\varphi_{k_3}(x_3)\cdot u_1 \Bigg) \\
+ \Bigg(& c_{(0,0,0,0,1)^\top}(K)\cdot u_2 + \sum_{\boldsymbol{k}\in\mathcal{I}_{\boldsymbol{N}}\{1,5\}} c_{\boldsymbol{k}}(K)\,\varphi_{k_1}(x_1)\cdot u_2 + \sum_{\boldsymbol{k}\in\mathcal{I}_{\boldsymbol{N}}\{2,5\}} c_{\boldsymbol{k}}(K)\,\varphi_{k_2}(x_2)\cdot u_2 \\
& + \sum_{\boldsymbol{k}\in\mathcal{I}_{\boldsymbol{N}}\{3,5\}} c_{\boldsymbol{k}}(K)\,\varphi_{k_3}(x_3)\cdot u_2 + \sum_{\boldsymbol{k}\in\mathcal{I}_{\boldsymbol{N}}\{1,2,5\}} c_{\boldsymbol{k}}(K)\,\varphi_{k_1}(x_1)\,\varphi_{k_2}(x_2)\cdot u_2 \\
& + \sum_{\boldsymbol{k}\in\mathcal{I}_{\boldsymbol{N}}\{1,3,5\}} c_{\boldsymbol{k}}(K)\,\varphi_{k_1}(x_1)\,\varphi_{k_3}(x_3)\cdot u_2 + \sum_{\boldsymbol{k}\in\mathcal{I}_{\boldsymbol{N}}\{2,3,5\}} c_{\boldsymbol{k}}(K)\,\varphi_{k_2}(x_2)\,\varphi_{k_3}(x_3)\cdot u_2 \Bigg).
\end{aligned}
$$

Again, by defining the sets

$$
\begin{aligned}
V_1 &\coloneqq \{\varnothing, \{1\}, \{2\}, \{3\}, \{1,2\}, \{1,3\}, \{2,3\}\}, \\
V_2 &\coloneqq \{\{4\}, \{1,4\}, \{2,4\}, \{3,4\}, \{1,2,4\}, \{1,3,4\}, \{2,3,4\}\}, \\
V_3 &\coloneqq \{\{5\}, \{1,5\}, \{2,5\}, \{3,5\}, \{1,2,5\}, \{1,3,5\}, \{2,3,5\}\},
\end{aligned}
$$

this can be written as

$$
K(\boldsymbol{z}) = \sum_{\boldsymbol{k}\in\mathcal{I}(V_1)} c_{\boldsymbol{k}}(K)\,\varphi_{\boldsymbol{k}}(\boldsymbol{x}) + \sum_{\boldsymbol{k}\in\mathcal{I}(V_2)} c_{\boldsymbol{k}}(K)\,\varphi_{\boldsymbol{k}}(\boldsymbol{x})\cdot u_1 + \sum_{\boldsymbol{k}\in\mathcal{I}(V_3)} c_{\boldsymbol{k}}(K)\,\varphi_{\boldsymbol{k}}(\boldsymbol{x})\cdot u_2,
$$

i. e., we obtain the multivariate structure (58) *without ever computing couplings between* u_1 *and* u_2.