Combining off-white and sparse black models in multi-step physics-based systems identification

Automatica 179 (2025) 112409

Contents lists available at ScienceDirect

Automatica

journal homepage: www.elsevier.com/locate/automatica

Brief paper

# Combining off-white and sparse black models in multi-step physics-based systems identification☆

Cesare Donati [a,b,*], Martina Mammarella [b], Fabrizio Dabbene [b], Carlo Novara [a], Constantino M. Lagoa [c]

[a] DET, Politecnico di Torino, Corso Duca degli Abruzzi 24, Torino, Italy
[b] CNR-IEIIT, c/o Politecnico di Torino, Corso Duca degli Abruzzi 24, Torino, Italy
[c] EECS, The Pennsylvania State University, University Park, PA, USA

ARTICLE INFO

*Article history:*
Received 12 August 2024
Received in revised form 28 February 2025
Accepted 14 April 2025

*Keywords:*
Nonlinear system identification
Gray-box modeling
Parametric optimization
Time-invariant systems

ABSTRACT

In this paper, we propose a unified framework for identifying interpretable nonlinear dynamical models that preserve physical properties. The proposed approach integrates a model, based on physical principles, with black-box basis functions to compensate for unmodeled dynamics, thus ensuring accuracy over multi-step horizons. Additionally, we introduce penalty terms to enforce physical consistency and stability during training. We provide a comprehensive analysis of theoretical properties related to multi-step nonlinear system identification, establishing bounds on parameter estimation errors and conditions for sparsity recovery. The proposed framework demonstrates significant potential for improving model accuracy and reliability in various engineering applications, making a substantial step towards the effective use of combined off-white and sparse black models in system identification. The effectiveness of the proposed approach is proven on a nonlinear system identification benchmark.



## 1. Introduction

The majority of the systems encountered in modern engineering applications (e.g., aerospace, automotive, energy, or systems biology) exhibit dynamical behaviors that may be too complex to be captured by linear relationships. For this reason, the field of nonlinear system identification has experienced significant growth, and many approaches have been developed aiming at identifying nonlinear dynamical models from collected data. However, while several important developments have been proposed (Ljung, 2010), it still largely remains an open issue (Schoukens & Ljung, 2019). Existing approaches to nonlinear systems identification can be classified into two main groups. On the one side, we have methods arising from *basic principles*: when a model is derived from physical laws and its parameters are known or are obtained from dedicated measurements, it is referred to as *white-box model*. However, this is a rather extreme situation; more generally, some parameters require identification from data, and one enters the wide family of the so-called *gray-box* models. Adopting the classification proposed by L. Ljung in Ljung (2010), we refer to this particular class of models as *off-white*. On the other side of the spectrum, we have the *black-box* models, which aim at describing the system's dynamics using generic linear parameterizations (Mattsson, Zachariah, & Stoica, 2018; Svensson & Schön, 2017) or families of universal approximators (Chiuso & Pillonetto, 2019). This second class of models has gained increasing popularity for their adaptability to a wide range of systems with minimal (or no) prior information about the underlying physical processes. Notably, the majority of published works focus on some variation of the black-box model approach (see, e.g., Chiuso and Pillonetto (2019) and references therein), while only few methodological works specifically discuss off-white methods. However, despite the desirable properties of black-box models, these show several limitations as, for instance, lack of interpretability, poor consistency with physical properties, and absence of shared guidelines for selecting basis functions and model order.

To deal with these issues, the community has adopted and adapted solutions from machine learning and artificial intelligence literature. On the one side, non-parametric approaches (Zorzi & Chiuso, 2017), mainly kernel-based methods (Dalla Libera, Carli, & Pillonetto, 2021; Schoukens & Scheiwe, 2016), have gained popularity for their ability to capture a large diversity of nonlinear behaviors without requiring complicated choices of basis functions. On the other side, techniques based on neural

☆ The material in this paper was partially presented at The 63rd IEEE Conference on Decision and Control (CDC), December 16–19, 2024, Milan, Italy. This paper was recommended for publication in revised form by Associate Editor Dario Piga under the direction of Editor Alessandro Chiuso.

* Corresponding author at: DET, Politecnico di Torino, Corso Duca degli Abruzzi 24, Torino, Italy.

*E-mail addresses:* cesare.donati@polito.it (C. Donati), martina.mammarella@cnr.it (M. Mammarella), fabrizio.dabbene@cnr.it (F. Dabbene), carlo.novara@polito.it (C. Novara), cml18@psu.edu (C.M. Lagoa).

https://doi.org/10.1016/j.automatica.2025.112409

networks (NN), e.g., Mavkov, Forgione, and Piga (2020), have been proposed, showing a remarkable implementation ease and capability of recovering long-term behaviors, becoming the solution-to-go in several application fields. However, lately these techniques have also revealed their main limitations, i.e., the need of large amount of training data and the difficulty of capturing some inherent physical phenomena at the core of such data. The proposed solution to these two drawbacks has been the same: to devise ways to "bring the physics back" into the model. This led to the exponential growth of the family of physics-informed NNs (PINNs) (Karniadakis et al., 2021). The main feature of PINNs lies in exactly incorporating physical information through either physics-based loss functions or structural modifications, ensuring physical consistency between inputs and outputs (Karniadakis et al., 2021).

The present paper is motivated by the recognition that, despite the significant advances in system identification, there is still a need to develop identification algorithms that can: (i) leverage all the information available on a system, such as partial parametric description, measurements, and available information on intrinsic properties of the system, (ii) exploit recent advances in first-order optimization, at the core of the success of NN-based methods, and (iii) deliver reliable estimates over extended time horizons, providing multi-step error minimization guarantees. This latter point is particularly crucial for applications such as model predictive control. With these objectives in mind, in this paper, we take a substantial step towards using off-white models and propose a simple yet effective technique with a twofold goal: (i) *physical interpretability* — we aim at deriving models whose parameters have a physical meaning and whose values are as close as possible to the "real" ones; and (ii) *multi-step horizon* — the derived model should ensure more accurate long-term predictions, minimizing a multi-step prediction cost, to improve the reliability of the identified model.

### 1.1. Setting and contributions

To guarantee physical interpretability, we consider a system described by an off-white model, given by known nonlinear equations derived from physical principles, alongside discrepancies arising from, e.g., modeling errors, uncertainties, and perturbations that can affect parameter accuracy (Mammarella et al., 2024; Quaghebeur, Nopens, & De Baets, 2021). Additionally, we integrate specific penalty terms into a multi-step cost function, to extrapolate the physical characteristics of the system, such as passivity, monotonicity, or stability (Zakwan et al., 2023).

To promote adherence to the real physical parameters, we introduce a mixed approach that combines the multi-step identification method discussed in Donati, Mammarella, Dabbene, Novara, and Lagoa (2024) with a sparse black-box component to compensate for unmodeled dynamics in the physical model. By integrating partial knowledge with black-box basis functions, we aim to improve both accuracy and reliability, addressing output estimation errors and providing better parameter estimates for the physics-based model. To this end, we focus on finding the sparsest coefficient vector selecting the basis functions to obtain a compensation term that only deals with the unmodeled dynamics when necessary and it is easily "interpretable". Indeed, non-sparse coefficient vectors might also capture part of the physical model's dynamics, resulting in predictions aligned with the measurements but with inaccurate physical parameter estimates. Similar model augmentation strategies appear in Liu, Tóth, and Schoukens (2024), which expands the physics-based models via weighted regularization, and in Kaheman et al. (2019), where the SINDy algorithm aims to capture discrepancies between a simplified model and the measurement data enforcing sparsity. However, unlike the approach we propose in this paper, these methods assume *known* physical parameters. Contrary, we focus on identifying physical parameters while leveraging black-box augmentation. The work (Quaghebeur et al., 2021) integrates machine learning into first-principles modeling using NNs. However, unlike what proposed here, no regularization is used, thus leading to reduced interpretability of physical parameters and possible data over-fitting. Furthermore, no theoretical analysis is present.

In this paper, we provide for the first time a general framework with a comprehensive analysis of crucial theoretical properties regarding multi-step, nonlinear system identification. Unlike existing works, we prove explicit parameter estimation bounds and sparsity conditions in a unified setting, bridging these two aspects in a novel way. Specifically, we derive bounds on parameter estimation error and we establish conditions for the exact sparsity recovery in the black-box term, generalizing sparsity results from single-step, linear settings (Gribonval, Figueras i Ventura, & Vandergheynst, 2006; Novara, 2012) to the multi-step, nonlinear case. We show that accurately compensating for unmodeled dynamics enhances the estimation accuracy of physical parameters. These results provide formal guarantees on the interplay between black-box augmentation and physical consistency, offering new insights into the reliability of mixed modeling approaches.

The paper is structured as follows. In Section 2, we formulate the problem and define the identification framework. A theoretical analysis of the parameter error bounds is presented in Section 3, while Section 4 shows an analysis of the sparsity of the black-box term. Numerical results are discussed in Section 5. Main conclusions are drawn in Section 6.

**Notation.** Given a vector $v$, $\|v\|_p$ is its $\ell_p$-norm, $\mathbf{v}_{1:T} \doteq \{v_k\}_{k=1}^{T}$ denotes $\{v_1, \dots, v_T\}$, and $\|\mathbf{v}_{1:T}\|_p \doteq \|[v_1^\top, \dots, v_T^\top]^\top\|_p$. Moreover, $\text{supp}(v) \doteq \{i : v_i \neq 0\}$, $\overline{\text{supp}}(v) \doteq \{i : v_i = 0\}$. Given $\mathbf{v}_{1:T}$ and a function $f(v_k)$, $f(\mathbf{v}_{1:T}) \doteq [f(v_1), \dots, f(v_T)]^\top$. Given $A \in \mathbb{R}^{n,m}$, $\|A\|$ is its spectral norm and $A^\dagger$ its pseudo-inverse. Given $a, b \in \mathbb{N}$, $a \leq b$, $[a, b]$ denotes $\{a, a+1, \dots, b\}$.

## 2. System identification framework

Let us consider a dynamical system $\mathcal{S}$, combining a physical model and an unknown component of the form

$$\mathcal{S}: \; x_{k+1} = f(x_k, u_k; \theta) + \Delta(x_k, u_k); \;\; z_k = h(x_k), \tag{1}$$

where $x_k \in \mathbb{R}^{n_x}$ is the state, $u_k \in \mathbb{R}^{n_u}$ the input, $z_k \in \mathbb{R}^{n_z}$ the observation, and $\theta \in \mathbb{R}^{n_\theta}$ the unknown parameters to be estimated, along with the initial state $x_0$. The known functions $f : \mathbb{R}^{n_x} \times \mathbb{R}^{n_u} \times \mathbb{R}^{n_\theta} \to \mathbb{R}^{n_x}$ and $h : \mathbb{R}^{n_x} \to \mathbb{R}^{n_z}$ are nonlinear, time-invariant, and continuously differentiable. The term $\Delta : \mathbb{R}^{n_x} \times \mathbb{R}^{n_u} \to \mathbb{R}^{n_x}$ represents the unknown dynamics, and will be estimated using a black-box approximation. Noisy measurements of the input and output $\tilde{\mathbf{u}}_{0:T}, \tilde{\mathbf{z}}_{0:T}$ are available. In particular, $\tilde{u}_k = u_k + \eta_k^u$, $\tilde{z}_k = z_k + \eta_k^z$, with $\eta_k^u \in \mathbb{R}^{n_u}$ and $\eta_k^z \in \mathbb{R}^{n_z}$ being the process and measurement noises, respectively. Note that, while $\tilde{u}_k$ is the known input, the actual input received by the system is $u_k$.

The goal is to identify $\theta$ and $x_0$, thus defining a multi-step[1] estimation model $\mathcal{M}$ approximating the system $\mathcal{S}$ while compensating for the unmodeled dynamics. This is achieved by minimizing a cost function $\mathcal{C}_T$ measuring a multi-step prediction error and, possibly, additional penalty terms, e.g., physically-driven or sparsity constraints, as discussed in the following section.

[1] In single-step identification, the model predicts one step ahead, in multi-step it propagates $x_k$ recursively.

### 2.1. Multi-step off-white and black-box model

We define the estimation model $\mathcal{M}$ as the combination of the known physical model (1) with a black-box component $\delta$ that approximates the unmodeled dynamics $\Delta$,

$$\mathcal{M}: \widehat{x}_{k+1} = f(\widehat{x}_k, \tilde{u}_k; \widehat{\theta}) + \delta(\widehat{x}_k, \tilde{u}_k; \omega);\ \widehat{z}_k = h(\widehat{x}_k). \tag{2}$$

Here, $\widehat{x}_k$, $\widehat{z}_k$ are the estimated states and observations, obtained using the identified parameters $\widehat{\theta}$ and initial state $\widehat{x}_0$. The parameters $\omega$, which are learned simultaneously with $\theta$, describe the black model. For instance, they may represent the weights of a NN or the coefficients of an ARX model. Indeed, while in our setup we consider a basis-function representation for $\delta$, the approach is general and extends to more involved families of approximators, including NNs (Mavkov et al., 2020) and kernel methods (Dalla Libera et al., 2021). Formally, let $\varphi \in \mathbb{R}^m$ be the vector of basis functions $\varphi(\widehat{x}_k, \tilde{u}_k) = \left[\varphi_1(\widehat{x}_k, \tilde{u}_k), \dots, \varphi_m(\widehat{x}_k, \tilde{u}_k)\right]^\top$, with $\varphi_j : \mathbb{R}^{n_x} \times \mathbb{R}^{n_u} \to \mathbb{R}$. The $\iota$-th element of $\delta$ is defined as

$$\delta_\iota = \sum_{j=1}^{m} \Omega_{\iota j} \varphi_j \left( \sum_{i=1}^{n_x} W_{ij}^{(\iota)} \widehat{x}_{i,k} + \sum_{i=1}^{n_u} W_{(n_x+i)j}^{(\iota)} \tilde{u}_{i,k} + B_{\iota j} \right),$$

with $\iota \in [1, n_x]$. Here, $\omega = \{\Omega, W, B\}$ consists of: (i) the basis functions weights $\Omega = [\Omega_{\iota j}] \in \mathbb{R}^{n_x, m}$; (ii) the input/state coefficients $W = [W^{(1)}, \dots, W^{(n_x)}] \in \mathbb{R}^{(n_x+n_u), n_x m}$ with $W^{(\iota)} = [W_{ij}^{(\iota)}] \in \mathbb{R}^{n_x+n_u, m}$; and (iii) the bias terms $B = [B_{\iota j}] \in \mathbb{R}^{n_x, m}$. Notice that, when available, domain knowledge and prior system understanding can guide the choice of the initial set of basis functions $\varphi$. In other cases, their choice can be carried out considering the many options in the literature (see, e.g., Sjöberg et al. 1995 for a discussion on basis functions and indications for their choice). Moreover, we note that elements of $\delta$, $\Omega$, $W$, and $B$ can be selectively fixed or set to zero, allowing for arbitrary flexibility in the design of $\delta$.

As previously discussed, since $\delta$ has only a complementary role, $\Omega$ is required to be sparse when the physical model is accurate. Moreover, to extrapolate physical properties in the estimated model $\mathcal{M}$ from the observed data, as done, e.g., in many PINN-based approaches (Karniadakis et al., 2021), we embed explicit physical penalty terms into the cost function. Thus, defining the prediction error sequence $\mathbf{e}_{0:T} = \{e_k\}_{k=0}^T$ with $e_k \doteq \widehat{z}_k - \tilde{z}_k \in \mathbb{R}^{n_z}$, $k \in [0, T]$, the multi-step cost function $\mathcal{C}_T$ is given by

$$\begin{aligned} \mathcal{C}_T(\theta, x_0, \omega; \mathbf{e}_{0:T}) \doteq \sum_{k=0}^{T} \Big( & \mathcal{L}_k(\theta, x_0; e_k) + \gamma \|\Omega\|_{1\beta} \\ & + \lambda p(\widehat{x}_k, \theta) + \nu q^2(\widehat{x}_k, \theta) \Big). \end{aligned} \tag{3}$$

We assume that the local loss $\mathcal{L}_k$ is twice continuously differentiable. The $\ell_1$-norm[2] in the second term $\|\Omega\|_{1\beta} \doteq \sum_{\iota=1}^{n_x} \|\Omega_\iota^\top\|_1$, being $\Omega_\iota^\top$ the $\iota$-th row of $\Omega$, is chosen to promote sparsity and ensure a simpler and interpretable representation of the unmodeled dynamics. The weight $\gamma \in \mathbb{R}$ controls the regularization. Finally, the remaining terms $p, q$ define the physics-based cost. In particular, $\lambda, \nu \in \mathbb{R}$ are the penalty weights, and $p, q : \mathbb{R}^{n_x} \times \mathbb{R}^{n_\theta} \to \mathbb{R}$ are twice continuously differentiable functions enforcing $p(\widehat{x}_k, \theta) \le 0$, $q(\widehat{x}_k, \theta) = 0$, $\forall k \in [0, T]$. Here, properly customizing $p$ and $q$ allows incorporating various physics-based constraints, such as physical limits, energy conservation, symmetry, positive semi-definiteness, or triangle inequality (see, e.g., Donati et al. 2024, Mammarella et al. 2024 for more details). Hence, we formally state the following problem.

**Problem 1** (*Multi-Step Identification Problem*)**.** Given sequences $\tilde{\mathbf{u}}_{0:T}$ and $\tilde{\mathbf{z}}_{0:T}$, functions $f$ and $h$ in (1), and basis functions $\varphi$, estimate the optimal values for the physical parameters $\widehat{\theta}$, initial state $\widehat{x}_0$, and black-box weights $\omega$ over the horizon $T$, such that the estimated model $\mathcal{M}$ is the best physically-consistent approximation of $\mathcal{S}$. Thus, given the multi-step cost $\mathcal{C}_T$ in (3), solve

$$(\theta^\star, x_0^\star, \omega^\star) \doteq \arg\min_{\theta, x_0, \omega}\ \mathcal{C}_T. \tag{4}$$

Within this framework, we aim at solving the optimization problem (4) relying on first-order methods. Such methods, which under suitable conditions are guaranteed to converge to a solution (in general sub-optimal), have recently gained popularity for their ability to tackle complex problems, as discussed in Donati et al. (2024, 2025).

## 3. Parametric error bound

In this section, we propose key theoretical properties regarding the upper bound on the parametric identification error, i.e., the maximum distance between the solution $\theta^\star$ of the optimization problem (4) and the true parameters vector $\bar{\theta}$. In the following, we will highlight the relevant dependencies of $\mathcal{C}_T$, observing that $\mathcal{C}_T(\theta, x_0, \omega; \mathbf{e}_{0:T}) \equiv \mathcal{C}_T(\theta, x_0, \omega; \tilde{\mathbf{u}}_{0:T}, \tilde{\mathbf{z}}_{0:T}) \equiv \mathcal{C}_T(\theta, x_0, \omega; \boldsymbol{\eta})$ with $\boldsymbol{\eta} = \{\boldsymbol{\eta}_{0:T}^u, \boldsymbol{\eta}_{0:T}^z\}$. Moreover, we assume for simplicity that $x_0$ is known and we focus on $\theta$, noting that the general case extends straightforwardly.[3]

First, given that the solution of (4) may not be unique, we recall the concept of system *identifiability* (Bellman & Åström, 1970), which determines the convergence behavior.

**Definition 1** (*System Identifiability*)**.** A system $\mathcal{S}$ with parameters $\bar{\theta}$ and initial state $\bar{x}_0$ is locally identifiable if $\mathcal{C}_T$ has a strict local minimum at $\widehat{\theta} = \bar{\theta}, \widehat{x}_0 = \bar{x}_0$. If the minimum is global, the system is globally identifiable.

We recall that a sufficient condition for local identifiability is that the Hessian matrix with respect to $\widehat{\theta}$ and $\widehat{x}_0$, is positive definite for all $\theta \in \Theta$, $x \in \mathcal{X}$, where $\Theta$ and $\mathcal{X}$ are suitable neighborhoods of $\bar{\theta}, \bar{x}_0$ (Bellman & Åström, 1970).

Then, to derive the parametric error bound, we rely on the following assumptions.

**Assumption 1** (*Basic Assumptions*)**.**

(i) (local identifiability) The system is locally identifiable according to Definition 1. Hence, the Hessian at $\theta^\star$ is positive definite for all noise realizations, i.e.,

$$H \doteq \frac{\partial^2 \mathcal{C}_T(\theta; \boldsymbol{\eta})}{\partial^2 \theta}\Big|_{\theta=\theta^\star} \succ 0, \quad \forall \boldsymbol{\eta}.$$

(ii) (convergence to $\bar{\theta}$) When the noise is null and $\Delta = 0$, minimizing $\mathcal{C}_T$ recovers true parameters $\theta^\star = \bar{\theta}$.

Assumption 1. (i) is rather standard in system identification (Bellman & Åström, 1970) while Assumption 1. (ii) holds when the optimization is initialized within the region of attraction of $\bar{\theta}$. Indeed, if $\Delta = 0$, the noise is null, and the identifiability assumption holds, then we have $\mathcal{C}_T(\bar{\theta}; \mathbf{0}) = 0$, making $\theta^\star = \bar{\theta}$ a local minimum. Next, we introduce the concept of bounded functions with the following lemma.

[2] Being the $\ell_1$-norm not differentiable, gradient computation is allowed via softplus smoothing (Schmidt, Fung, & Rosales, 2007).

[3] Due to nonlinearity and nonconvexity of the considered problem, most results are *local* holding when the optimization is initialized sufficiently close to the optimal solution.

**Lemma 1** (*Bounded Function*). *Let Assumption 1 hold. Assume $\boldsymbol{\eta} \in \mathcal{N}$ with $\mathcal{N}$ closed and bounded. Define $G(\boldsymbol{\eta}) \doteq \frac{\partial^2 \mathcal{C}_T(\theta;\boldsymbol{\eta})}{\partial \boldsymbol{\eta} \partial \theta}\big|_{\theta=\theta^\star}$. Then $H^{-1}G(\boldsymbol{\eta})$ is bounded for all $\boldsymbol{\eta} \in \mathcal{N}$. Specifically, there exist constants $M_u, M_z < \infty$, such that $\max_{\boldsymbol{\eta}\in\mathcal{N}} \|[H^{-1}G(\boldsymbol{\eta})]_u\|_p < M_u$, $\max_{\boldsymbol{\eta}\in\mathcal{N}} \|[H^{-1}G(\boldsymbol{\eta})]_z\|_p < M_z$, with $[H^{-1}G]_u$ and $[H^{-1}G]_z$ corresponding to the columns related to $\boldsymbol{\eta}^u_{0:T}$ and $\boldsymbol{\eta}^z_{0:T}$, respectively.*

**Proof.** From Assumption 1. (i), the Hessian $H$ is positive definite $\forall \boldsymbol{\eta} \in \mathcal{N}$. Thus, it is invertible and its inverse is bounded. Moreover, $\mathcal{C}_T$ is twice continuously differentiable since it is a sum of twice differentiable functions with continuous derivatives $\mathcal{L}_k$. Hence, from Lipschitz continuity it follows that $G(\boldsymbol{\eta})$ is bounded. The lemma follows considering that $\mathcal{N}$ is closed and bounded. ■

The following theorem, whose proof is reported in Appendix A, formalizes the boundness of the parametric error.

**Theorem 1** (*Parametric Error Bound*). *Let Assumption 1 hold. The parametric identification error is bounded as*

$$\left\|\theta^\star - \bar{\theta}\right\|_p \leq M_u \left\|\boldsymbol{\eta}^u_{0:T}\right\|_p + M_z \left\|\boldsymbol{\eta}^z_{0:T}\right\|_p + M_\Delta \|\widetilde{\boldsymbol{\Delta}}\|_p, \tag{5}$$

*where $M_\Delta$ is a finite constant and $\widetilde{\boldsymbol{\Delta}} \doteq \widetilde{\boldsymbol{\Delta}}_{0:T} = \{\widetilde{\Delta}_0, \dots, \widetilde{\Delta}_T\}$, $\widetilde{\Delta}_k \doteq \Delta(x_k, u_k) - \delta(\widehat{x}_k, u_k; \omega)$.*

Some interesting properties follow directly from Theorem 1. First, since the parametric error bound is directly proportional to $H^{-1}$ (see Lemma 1), a well-defined identification problem with an invertible $H$ ensures that the larger is the norm of $H$, the smaller is the parametric error. Second, as the black model $\delta(\cdot)$ better approximates the effect of $\Delta(\cdot)$, i.e., as $\widetilde{\Delta}_k \to 0$, $\forall k$, the parametric error bound decreases, thus enhancing the physical parameters estimation accuracy. This aspect will be further examined in the subsequent discussion in Section 4.4.

## 4. Maximum sparsity recovery

In this section, we provide a theoretical analysis of the recovery of the maximum sparsity of the black-box contribution. The goal is to look for the sparsest representation of the unmodeled dynamics to have a description that it is easier to analyze and interpret. First, we address the single-step, linear-in-parameters model, then we extend the results to the multi-step, nonlinear cases.

### 4.1. Single-step, linear-in-parameters setting

We consider a dynamical system described by the state-equation representation (1), with parameters $\bar{\theta} \in \mathbb{R}^{n_\theta}$, a linear-in-parameters function $f$, full state observations, and, without loss of generality, $n_x = n_z = 1$. The system $\mathcal{S}$ takes the form

$$\mathcal{S}: \; x_{k+1} = \xi^\top(x_k, u_k)\bar{\theta} + \Delta(x_k, u_k); \; \tilde{z}_k = x_k + \eta^z_k, \tag{6}$$

with $\xi(x_k, u_k) : \mathbb{R}^{n_x} \times \mathbb{R}^{n_u} \to \mathbb{R}^{n_\theta}$ a vector of known (possibly nonlinear) physical terms defined as $\xi(x_k, u_k) = [\xi_1(x_k, u_k), \dots, \xi_{n_\theta}(x_k, u_k)]^\top$. A dictionary of basis $\varphi(x_k, u_k) = [\varphi_1(x_k, u_k), \dots, \varphi_m(x_k, u_k)]^\top$ and a set of noise-corrupted data $\mathcal{D} = \{\tilde{\mathbf{u}}_{0:T}, \tilde{\mathbf{z}}_{0:T}\}$, collected from (6), are available. The relationship between $\tilde{u}_k$ and $\tilde{z}_k$ is described by following measurement model

$$\tilde{z}_{k+1} = \xi^\top(x_k, \tilde{u}_k)\bar{\theta} + \Delta(x_k, \tilde{u}_k) + \eta^v_k,$$

where $\eta^v_k$ accounts for both process $\eta^u_k$ and measurement noise $\eta^z_k$ (see Appendix B) and satisfies the following assumption.

**Assumption 2** (*Unknown but Bounded Noise*). The noise sequence $\eta^v = [\eta^v_1, \dots, \eta^v_T]^\top$ is unknown but bounded, i.e., $\|\eta^v\|_2 \leq \mu$.

Now, we move to address the problem of finding a sparse linear combination of the basis functions in the dictionary that, combined with the prior physical model, is consistent with the measured data. First, we define the matrices $\Xi \in \mathbb{R}^{T,n_\theta}$ and $\Phi \in \mathbb{R}^{T,m}$

$$\begin{aligned} \Xi &\doteq \left[\xi(\tilde{x}_0, \tilde{u}_0), \dots, \xi(\tilde{x}_{T-1}, \tilde{u}_{T-1})\right]^\top, \\ \Phi &\doteq \left[\varphi(\tilde{x}_0, \tilde{u}_0), \dots, \varphi(\tilde{x}_{T-1}, \tilde{u}_{T-1})\right]^\top, \end{aligned} \tag{7}$$

where $\tilde{x}_k \doteq \tilde{z}_k$, having initially considered full state measurements. Then, we recall the definitions of feasible parameter set and maximally sparse coefficients.

**Definition 2** (*Feasible Parameter Set*). Consider a dataset $\mathcal{D} = \{\tilde{\mathbf{u}}_{0:T}, \tilde{\mathbf{z}}_{0:T}\}$ satisfying Assumption 2, a sequence of states (predicted or measured) $\mathbf{x}_{0:T-1}$, and the noise bound $\mu$. Let $\tilde{z} = [\tilde{z}_1, \dots, \tilde{z}_T]^\top$. We define the Feasible Parameter Set (FPS) as

$$\text{FPS} \doteq \left\{\theta, \omega : \|\tilde{z} - f(\mathbf{x}_{0:T-1}, \tilde{\mathbf{u}}_{0,T-1}, \theta) - \Phi\omega\|_2 \leq \mu\right\}. \tag{8}$$

Moreover, for a given $\theta_0 \in \mathbb{R}^{n_\theta}$, the vector $\omega_0 \in \mathbb{R}^m$ is said to be feasible for $\theta_0$ if $(\theta_0, \omega_0) \in$ FPS.

**Definition 3** (*Maximally Sparse Coefficients*). Given $\mathcal{D} = \{\tilde{\mathbf{u}}_{0:T}, \tilde{\mathbf{z}}_{0:T}\}$ satisfying Assumption 2, $\mathbf{x}_{0:T-1}$, and the noise bound $\mu$, a feasible coefficient vector is said maximally sparse if it is a solution of

$$\bar{\omega} = \arg\min_{\theta,\omega} \|\omega\|_0 \text{ s.t. } \|\tilde{z} - f(\mathbf{x}_{0:T-1}, \tilde{\mathbf{u}}_{0,T-1}, \theta) - \Phi\omega\|_2 \leq \mu \tag{9}$$

where $\|\cdot\|_0$ is the $\ell_0$ quasi-norm.

Finally, we introduce the simplified sparsity problem.

**Problem 2** (*Simplified Sparsity Problem*). Consider a single-step, state-equation estimation model

$$\mathcal{M}: \quad \widehat{x}_{k+1} = \sum_{i=1}^{n_\theta} \theta_i \xi_i(\tilde{x}_k, \tilde{u}_k) + \sum_{i=1}^{m} \omega_i \varphi_i(\tilde{x}_k, \tilde{u}_k),$$

expressed compactly as $\widehat{x}_{k+1} = \widehat{F}(\tilde{x}_k, \tilde{u}_k, \theta, \omega)$. The goal is to estimate $(\widehat{\theta}, \widehat{\omega})$ from the data set $\mathcal{D}$, such that: (i) $\widehat{\omega}$ is sparse; and (ii) $(\widehat{\theta}, \widehat{\omega})$ is consistent with the data, i.e., the single-step prediction error satisfies $\|\tilde{z} - \widehat{F}(\tilde{\mathbf{x}}_{0:T-1}, \tilde{\mathbf{u}}_{0:T-1}, \widehat{\theta}, \widehat{\omega})\|_2 \leq \mu$, where $\mu$ is the noise bound from Assumption 2.

Note that no assumptions are made on the structural form of the black-box term $\Delta$ in (6). We only suppose that the chosen dictionary of basis function $\varphi$ is sufficiently rich to allow an approximation of $\Delta$ compatible with the given noise level. Moreover, note that we do not explicitly require a statistical characterization of the noise. Instead, we only assume that an upper bound on its norm is given.

Here, solving the sparse identification problem via (9) is intractable due to the nonconvexity of the $\ell_0$ quasi-norm, which makes it NP-hard. Instead, we consider the following convex relaxation

$$\begin{aligned} (\widehat{\theta}, \widehat{\omega}) = \arg\min_{\theta,\omega} &\|\omega\|_1 \\ \text{s.t. } &\|\tilde{z} - \widehat{F}(\tilde{\mathbf{x}}_{0:T-1}, \tilde{\mathbf{u}}_{0:T-1}, \theta, \omega)\|_2 \leq \mu. \end{aligned} \tag{10}$$

This $\ell_1$-regularized problem encourages sparsity in $\widehat{\omega}$, addressing condition (i) in Problem 2. However, it does not necessarily yield a maximally sparse solution in the sense of Definition 3, i.e., the sparsest set of basis functions that still ensures consistency with the data. In the following, we analyze the conditions under which a sparse solution of (10) is also maximally sparse.

### 4.2. Theoretical analysis on sparsity

In Gribonval et al. (2006), Novara (2012), the authors established conditions under which a coefficient vector, representing the solution to a general sparsity problem, is maximally sparse. While their focus is on sparsifying all decision variables, in this section, we consider a setting where only the black-box coefficients $\omega$ (a subset of the decision variables $(\theta, \omega)$), require sparsification. Additionally, we generalize these results from single-step, linear systems to multi-step, nonlinear cases (see Section 4.3). First, in the following, we introduce some key concepts.

**Definition 4** (*Preliminary Notations*)**.** For each integer $n \in \mathbb{R}$ and matrix $Q \in \mathbb{R}^{n_1,n_2}$ we denote

$$\sigma_n^2(Q) \doteq \inf_{\|x\|_0 \le n} \frac{\|Qx\|_2^2}{\|x\|_2^2}, \ \|x\|_{(Q,n)} \doteq \sqrt{\sum_{i \in \mathcal{I}_n(x)} \left(x^\top q_i\right)^2},$$

where $\mathcal{I}_n(x)$ indexes the $n$ largest inner products $|x^\top q_i|$, with $q_i$ the $i$th column of $Q$.

**Definition 5** (*Prediction Errors*)**.** Given observation $z \in \mathbb{R}^T$ and matrices $P \in \mathbb{R}^{T,n_\theta}$ and $Q \in \mathbb{R}^{T,m}$ such that $z = P\theta + Q\omega$, the prediction error is defined as $e_{P,Q}(z, \theta, \omega) \doteq z - P\theta - Q\omega$. Moreover, given $\theta^\star_{P,Q}(z, \omega) \doteq \arg\min_{\theta \in \mathbb{R}^{n_\theta}} \|e_{P,Q}(z, \theta, \omega)\|_2$, the optimal compensation error is defined as $e^\star_{P,Q}(z, \omega) \doteq e_{P,Q}(z, \theta^\star_{P,Q}(z, \omega), \omega)$.

For given $z$, $P$, $Q$, the prediction error is the discrepancy between measurements and predictions when employing black-box augmentation. The corresponding optimal compensation error, $e^\star_{P,Q}(z, \omega)$, represents the remaining prediction error after substituting the optimal estimate $\theta^\star_{P,Q}(z, \omega)$, explicitly depending on $\omega$, into the model. Given Definition 5, the following Lemma holds.

**Lemma 2** (*Optimal Compensation Error*)**.** *Let $\Upsilon(P) = (I_T - PP^\dagger)$. Considering Definition 5, it holds that $\theta^\star_{P,Q}(z, \omega) = P^\dagger(z - Q\omega)$. Moreover, $e^\star_{P,Q}(z, \omega) = \Upsilon(P)(z - Q\omega)$.*

**Proof.** Solving $\min_{\theta \in \mathbb{R}^{n_\theta}} \|e_{P,Q}(z, \theta, \omega)\|_2 = \|(z - Q\omega) - P\theta\|_2$, yields the least squares solution, i.e., $\theta^\star_{P,Q}(z, \omega) = (P^\top P)^{-1}P^\top(z - Q\omega) = P^\dagger(z - Q\omega)$. Thus, $e^\star_{P,Q}(z, \omega)$ simplifies to $e_{P,Q}(z, \theta^\star_{P,Q}(z, \omega), \omega) = z - PP^\dagger(z - Q\omega) - Q\omega = (I_T - PP^\dagger)z - (I_T - PP^\dagger)Q\omega$. ■

**Assumption 3** (*Problem Feasibility*)**.** (i) FPS $\neq \emptyset$, and (ii) for given $z$, $P$, $Q$, a feasible $\omega$ exists for $\theta^\star_{P,Q}(z, \omega)$ according to Definition 2.

Thus, the following theorems provide conditions under which the solution of (10) shares the same support as $\bar{\omega}$, ensuring maximal sparsity. First, using results from Gribonval et al. (2006, Theorem 1, Corollary 1), we formally establish when two vectors have equivalent supports. The proof is given in Appendix C.

**Theorem 2** (*Equivalent Support Conditions*)**.** *Given $\mathcal{D} = \{\tilde{\mathbf{u}}_{0:T}, \tilde{\mathbf{z}}_{0:T}\}$, and $\Xi$, $\Phi$ (7), let Assumptions 1, 2, 3 hold. Let $\widehat{\omega} \in \mathbb{R}^m$ be the solution of (10), with $M \doteq \|\widehat{\omega}\|_0$ the cardinality of its support, and $\Upsilon_0 = \Upsilon(\Xi)$. Consider any other representation $\omega'$ and the noise-corrupted observation vector $\tilde{z} = [\tilde{z}_1, \dots, \tilde{z}_T]^\top$ from (6). If $\|e^\star_{\Xi,\Phi}(\tilde{z}, \omega')\|_2 \le \|e^\star_{\Xi,\Phi}(\tilde{z}, \widehat{\omega})\|_2$, and $\|\omega'\|_0 \le M$, then,*

$$\|\omega' - \widehat{\omega}\|_\infty \le \frac{\|e^\star_{\Xi,\Phi}(\tilde{z}, \widehat{\omega})\|_{(\Upsilon_0,1)} + \|e^\star_{\Xi,\Phi}(\tilde{z}, \widehat{\omega})\|_{(\Upsilon_0,2M)}}{\sigma^2_{2M}(\Upsilon_0)}.$$

*Moreover, if*

$$\|e^\star_{\Xi,\Phi}(\tilde{z}, \widehat{\omega})\|_{(\Upsilon_0,1)} + \|e^\star_{\Xi,\Phi}(\tilde{z}, \widehat{\omega})\|_{(\Upsilon_0,2M)} < \sigma^2_{2M}(\Upsilon_0)\eta(\widehat{\omega}),$$

*with $\eta(\omega) \doteq \min_{i \in \text{supp}(\omega)} |\omega_i|$, then for all $i \in [1, m]$, $\widehat{\omega}$ and $\omega'$ have the same support and sign, i.e., $\text{supp}(\widehat{\omega}) = \text{supp}(\omega')$ and $\text{sign}(\widehat{\omega}_i) = \text{sign}(\omega'_i)$.*

**Remark 1** (*Generality of the Sparsity Conditions*)**.** While Theorem 2 specifically applies results from Gribonval et al. (2006) to (10), its conditions can be verified on any estimate $\widehat{\omega} \in \mathbb{R}^m$ satisfying $\Upsilon_0\tilde{z} = \Upsilon_0\Phi\widehat{\omega} + e^\star_{\Xi,\Phi}(\tilde{z}, \widehat{\omega})$.

Now, let us define the vector $\omega^v$ as the solution of

$$\omega^v \doteq \arg\min_{\omega} \|\omega\|_1 \tag{11a}$$

$$\text{s.t.} \quad \omega_i \ge \text{sign}(\widehat{\omega}_i)\eta(\widehat{\omega}), \quad \forall i \in \text{supp}(\widehat{\omega}) \tag{11b}$$

$$|\omega_i| < \eta(\widehat{\omega}), \quad \forall i \in \overline{\text{supp}}(\widehat{\omega}) \tag{11c}$$

$$\|\Upsilon_0\tilde{z} - \Upsilon_0\Phi\omega\|_2 \le \mu, \tag{11d}$$

where $\eta(\omega) \doteq \min_{i \in \text{supp}(\omega)} |\omega_i|$. Building on the results from Novara (2012, Theorem 1), with the following theorem we define conditions ensuring that a vector is maximally sparse. The proof is reported in Appendix D.

**Theorem 3** (*Maximum Sparsity Recovery*)**.** *Given $\mathcal{D} = \{\tilde{\mathbf{u}}_{0:T}, \tilde{\mathbf{z}}_{0:T}\}$, and $\Xi$, $\Phi$ (7), let Assumptions 1, 2, 3 hold. Let $\bar{\omega}$ and $\omega^v$ be the solutions of (9) and (11), respectively. Let $\widehat{\omega}$ be the parameter vector obtained from (10), $M \doteq \|\widehat{\omega}\|_0$ the cardinality of its support, and $\Upsilon_0 = \Upsilon(\Xi)$. Define the noise-corrupted observation vector from (6) as $\tilde{z} = [\tilde{z}_1, \dots, \tilde{z}_T]^\top$, and let $\kappa_e \doteq \|\widehat{\omega}\|_0 - \|\bar{\omega}\|_0$. Define the set*

$$\lambda \doteq \left\{ i : |\omega^v_i| > \frac{\|e^\star_{\Xi,\Phi}(\tilde{z}, \omega^v)\|_{(\Upsilon_0,1)} + \|e^\star_{\Xi,\Phi}(\tilde{z}, \omega^v)\|_{(\Upsilon_0,2M)}}{\sigma^2_{2M}(\Upsilon_0)} \right\}.$$

*Assume that the constraint (11d) is active, i.e., $\|\Upsilon_0\tilde{z} - \Upsilon_0\Phi\omega^v\|_2 = \mu$. Then,*

$$\kappa_e \le \bar{\kappa}_e \doteq \|\widehat{\omega}\|_0 - \text{card}(\lambda). \tag{12}$$

*Moreover, if*

$$\frac{\|e^\star_{\Xi,\Phi}(\tilde{z}, \omega^v)\|_{(\Upsilon_0,1)} + \|e^\star_{\Xi,\Phi}(\tilde{z}, \omega^v)\|_{(\Upsilon_0,2M)}}{\sigma^2_{2M}(\Upsilon_0)} < \eta(\widehat{\omega}), \tag{13}$$

*then $\widehat{\omega}$ is maximally sparse and $\text{supp}(\widehat{\omega}) = \text{supp}(\bar{\omega})$.*

### 4.3. Multi-step, nonlinear-in-parameters extension

We now consider the more general case where the system follows (1) with parameters $\bar{\theta}$, $n_x = n_z = 1$, and no assumptions on the forms of $f$ and $h$. A dictionary of basis function $\varphi \in \mathbb{R}^m$ and a noise-corrupted dataset $\mathcal{D} = \{\tilde{\mathbf{u}}_{0:T}, \tilde{\mathbf{z}}_{0:T}\}$ are available. The measurement model, detailed in Appendix B, is given by

$$\tilde{z}_{k+1} = h(f(x_k, \tilde{u}_k, \bar{\theta}) + \Delta(x_k, \tilde{u}_k)) + \eta^v_k.$$

**Problem 3** (*Multi-Step Sparsity Problem*)**.** Consider a multi-step estimation model defined as

$$\mathcal{M}: \quad \widehat{x}_{k+1} = f(\widehat{x}_k, \tilde{u}_k, \theta) + \sum_{i=1}^{m} \omega_i\varphi_i(\widehat{x}_k, \tilde{u}_k), \tag{14}$$

expressed compactly as $\widehat{x}_{k+1} = \widehat{F}(\widehat{x}_k, \tilde{u}_k, \theta, \omega)$ with output $\widehat{z}_k = h(\widehat{x}_k)$. The goal is to estimate $(\widehat{\theta}, \widehat{\omega})$ such that: (i) $\widehat{\omega}$ is sparse; and (ii) $(\widehat{\theta}, \widehat{\omega})$ is consistent with the dataset, i.e., $\|\tilde{z} - \widehat{z}\|_2 \le \mu$, with $\widehat{z} = [\widehat{z}_1, \dots, \widehat{z}_T]^\top$.

**Remark 2** (*Multi-Step Nonlinearity*)**.** In the multi-step setting $(T > 1)$, iterating the system over time introduces strong nonlinearity in the parameters, even if $\widehat{F}$ is linear in $\theta$ and $\omega$. Specifically, the $k$th output prediction follows $\widehat{z}_k = h(\widehat{F}^k(\widehat{x}_0, \tilde{u}_1, \dots, \tilde{u}_{k-1}, \theta, \omega))$, where $\widehat{F}^k$ represents the $k$th iteration of (14).

In this case, to solve Problem 3, we consider the multi-step optimization problem (9), using the predicted sequence $\widehat{\mathbf{x}}_{0:T-1}$. As in the single-step case, we solve the following convex relaxation

$$(\widehat{\theta}, \widehat{\omega}) = \quad \arg\min_{\theta,\omega} \|\omega\|_1 \quad \text{s.t. } \|\tilde{z} - \widehat{z}\|_2 \leq \mu. \tag{15}$$

In the following, we extend Theorems 2 and 3 to the multi-step case, under Assumptions 2 and 3. From a conceptual point of view, this is nontrivial, as the parameters appear inside highly nonlinear functions.

**Theorem 4** (*Local Equivalent Support Conditions*)**.** *Given $\mathcal{D} = \{\tilde{\mathbf{u}}_{0:T}, \tilde{\mathbf{z}}_{0:T}\}$, let Assumptions 1, 2, 3(i) hold. Let $\widehat{\omega} \in \mathbb{R}^m$ be a local solution of (15) with $M \doteq \|\widehat{\omega}\|_0$. Consider any other representation $\omega'$ in a local neighborhood of $\widehat{\omega}$. Define the Jacobian matrices*

$$\mathcal{J}_\theta \doteq \frac{\partial h(\widehat{F}(\widehat{\mathbf{x}}_{0:T-1}, \tilde{\mathbf{u}}_{0:T-1}, \widehat{\theta}, \widehat{\omega}))}{\partial \theta} \in \mathbb{R}^{T,n_\theta},$$
$$\mathcal{J}_\omega \doteq \frac{\partial h(\widehat{F}(\widehat{\mathbf{x}}_{0:T-1}, \tilde{\mathbf{u}}_{0:T-1}, \widehat{\theta}, \widehat{\omega}))}{\partial \omega} \in \mathbb{R}^{T,m}.$$

*Let the residual $\tilde{z}_\ell = \tilde{z} - h(\widehat{F}(\widehat{\mathbf{x}}_{0:T-1}, \tilde{\mathbf{u}}_{0:T-1}, \widehat{\theta}, \widehat{\omega})) - \mathcal{J}_\theta\widehat{\theta} - \mathcal{J}_\omega\widehat{\omega}$, $\Upsilon_\ell = \Upsilon(\mathcal{J}_\theta)$, and Assumption 3(ii) hold for $\tilde{z}_\ell, \mathcal{J}_\theta, \mathcal{J}_\omega$. If the following conditions hold, i.e.,*

$$\|e^\star_{\mathcal{J}_\theta,\mathcal{J}_\omega}(\tilde{z}_\ell, \omega')\|_2 \leq \|e^\star_{\mathcal{J}_\theta,\mathcal{J}_\omega}(\tilde{z}_\ell, \widehat{\omega})\|_2, \ \|\omega'\|_0 \leq M,$$
$$\|e^\star_{\mathcal{J}_\theta,\mathcal{J}_\omega}(\tilde{z}_\ell,\widehat{\omega})\|_{(\Upsilon_\ell,1)} + \|e^\star_{\mathcal{J}_\theta,\mathcal{J}_\omega}(\tilde{z}_\ell, \widehat{\omega})\|_{(\Upsilon_\ell,2M)} < \sigma^2_{2M}(\Upsilon_\ell)\eta(\widehat{\omega}),$$

*where $\eta(\omega) \doteq \min_{i\in supp(\omega)} |\omega_i|$, then $\widehat{\omega}$ and $\omega'$ have the same support $\text{supp}(\widehat{\omega}) = \text{supp}(\omega')$ and the same sign $\text{sign}(\widehat{\omega}_i) = \text{sign}(\omega'_i)$ for all $i$.*

**Proof.** The proof follows by linearizing $h(\widehat{F}(\cdot))$ around $(\widehat{\theta}, \widehat{\omega})$ (15), i.e., $h(\widehat{F}(\cdot)) \simeq h(\widehat{F}(\widehat{\mathbf{x}}_{0:T-1}, \tilde{\mathbf{u}}_{0:T-1}, \widehat{\theta}, \widehat{\omega})) + \mathcal{J}_\theta(\theta - \widehat{\theta}) + \mathcal{J}_\omega(\omega - \widehat{\omega})$, applying the proof to Theorem 3 in Appendix C, and Gribonval et al. (2006, Theorem 1, Corollary 1), thus considering $\Upsilon_\ell\tilde{z}_\ell = \Upsilon_\ell\mathcal{J}_\omega\widehat{\omega} + e^\star_{\mathcal{J}_\theta,\mathcal{J}_\omega}(\tilde{z}_\ell, \widehat{\omega})$. ■

Next, we define the conditions under which $(\widehat{\theta}, \widehat{\omega})$ (15) is locally maximally sparse within the linearized neighborhood of the solution. Let $\bar{\omega}_\ell$ be the maximally sparse vector in the local approximation around $(\widehat{\theta}, \widehat{\omega})$, i.e.,

$$\bar{\omega}_\ell = \quad \arg\min_{\theta,\omega} \|\omega\|_0 \quad \text{s.t. } \|\tilde{z} - \mathcal{J}_\theta\theta - \mathcal{J}_\omega\omega\|_2 \leq \mu. \tag{16}$$

Given $\eta(\omega) \doteq \min_{i\in\text{supp}(\omega)} |\omega_i|$, then we compute $\omega^v_\ell$ as

$$\omega^v_\ell \doteq \arg\min_{\omega\in\mathbb{R}^m} \|\omega\|_1 \tag{17a}$$
$$\text{s.t. } \text{sign}(\widehat{\omega}_i)\omega_i \geq \eta(\widehat{\omega}), \quad \forall i \in \text{supp}(\widehat{\omega}) \tag{17b}$$
$$|\omega_i| < \eta(\widehat{\omega}), \quad \forall i \in \overline{\text{supp}}(\widehat{\omega}) \tag{17c}$$
$$\|\Upsilon_\ell\tilde{z}_\ell - \Upsilon_\ell\mathcal{J}_\omega\widehat{\omega}\|_2 \leq \mu. \tag{17d}$$

**Theorem 5** (*Maximum Sparsity Local Recovery*)**.** *Given $\mathcal{D} = \{\tilde{\mathbf{u}}_{0:T}, \tilde{\mathbf{z}}_{0:T}\}$, let Assumptions 1, 2, 3(i) hold. Let $\bar{\omega}_\ell$ and $\omega^v_\ell$ be the solutions of (16) and (17), respectively. Let $\widehat{\omega}$ be the solution of (15), with $M \doteq \|\widehat{\omega}\|_0$. Define the Jacobian matrices $\mathcal{J}_\theta \in \mathbb{R}^{T,n_\theta}$, $\mathcal{J}_\omega \in \mathbb{R}^{T,m}$, and the residual vector $\tilde{z}_\ell$ as in Theorem 4. Let $\Upsilon_\ell = \Upsilon(\mathcal{J}_\theta)$, and Assumption 3(ii) hold for $\tilde{z}_\ell, \mathcal{J}_\theta, \mathcal{J}_\omega$. Define $\kappa_e \doteq \|\widehat{\omega}\|_0 - \|\bar{\omega}_\ell\|_0$. Consider the set*

$$\lambda \doteq \left\{ i : |\omega^v_{\ell,i}| > \frac{\|e^\star_{\mathcal{J}_\theta,\mathcal{J}_\omega}(\tilde{z}_\ell, \omega^v_\ell)\|_{(\Upsilon_\ell,1)} + \|e^\star_{\mathcal{J}_\theta,\mathcal{J}_\omega}(\tilde{z}_\ell, \omega^v_\ell)\|_{(\Upsilon_\ell,2M)}}{\sigma^2_{2M}(\Upsilon_\ell)} \right\}.$$

*Assume that the constraint (17d) is active, i.e., $\|\Upsilon_\ell\tilde{z} - \Upsilon_\ell\Phi\omega^v\|_2 = \mu$. Then, $\kappa_e \leq \bar{\kappa}_e \doteq \|\widehat{\omega}\|_0 - card(\lambda)$.*

*Furthermore, if*

$$\frac{\|e^\star_{\mathcal{J}_\theta,\mathcal{J}_\omega}(\tilde{z}_\ell, \omega^v_\ell)\|_{(\Upsilon_\ell,1)} + \|e^\star_{\mathcal{J}_\theta,\mathcal{J}_\omega}(\tilde{z}_\ell, \omega^v_\ell)\|_{(\Upsilon_\ell,2M)}}{\sigma^2_{2M}(\Upsilon_\ell)} < \eta(\widehat{\omega}),$$

*then $\widehat{\omega}$ is locally maximally sparse, i.e., $\bar{\kappa}_e = 0$, and $\text{supp}(\widehat{\omega}) = \text{supp}(\bar{\omega}_\ell)$.*

The proof follows the same reasoning of the ones of Theorem 3 and Theorem 4.

**Remark 3** (*Relation with the Original Problem*)**.** Although the optimization problems in this section are presented in a constrained form, as in (10) and (15), they can always be reformulated in the unconstrained form (4) within the proposed framework. As discussed in Gribonval et al. (2006), this can be obtained by solving the corresponding Lagrangian problem with appropriate multipliers.

### 4.4. Optimality of the physical parameters

In this section, we show that, under black-box augmentation and maximum sparsity conditions, the identified physical parameters $\theta$ are *provably more accurate*, in the worst-case sense, than standard estimates lacking black-box models with sparsity guarantees. Specifically, given a sufficiently rich basis function dictionary $\varphi$, the unknown term $\Delta(x_k, u_k)$ in (1) with true physical parameters $\bar{\theta}$ can be parameterized as

$$\Delta(x_k, u_k) = \sum_{i=1}^{m} \bar{\omega}_i\varphi_i(x_k, u_k), \tag{18}$$

where $\bar{\omega}$ is the maximally sparse coefficient vector solving (9). This defines the true system parameters as $(\bar{\theta}, \bar{\omega})$. Thus, we define the parameter sets on $\theta$ and $\omega$.

**Definition 6** (*Parameters Sets*)**.** Given dataset $\mathcal{D} = \{\tilde{\mathbf{u}}_{0:T}, \tilde{\mathbf{z}}_{0:T}\}$ satisfying Assumption 2, a state sequence $\mathbf{x}_{0:T-1}$, and the noise bound $\mu$, recalling (8), we have FPS $\doteq \{\theta, \omega : \|\tilde{z} - \widehat{F}(\mathbf{x}_{0:T-1}, \tilde{\mathbf{u}}_{0,T-1}, \theta, \omega)\|_2 \leq \mu\}$. A subset of FPS, the Supported Feasible Parameter Set (SFPS), is SFPS $\doteq$ FPS $\cap \{\theta, \omega : \text{supp}(\omega) = \text{supp}(\bar{\omega})\}$, with $\bar{\omega}$ the maximally sparse coefficients vector.

The bound $\mu$ plays a key role in defining the sets. Thus, we remark that it can be refined through iterative approaches to balance complexity, accuracy, and fit to data.

Given a generic estimate $\widehat{\theta}$ of the true parameter vector $\bar{\theta}$, we now consider the parametric error defined as $e_\theta(\widehat{\theta}) \doteq \|\bar{\theta} - \widehat{\theta}\|_p$. Since $\bar{\theta}$ is unknown in practice, the so-called worst-case parametric error – a tight upper bound on $e_\theta(\widehat{\theta})$ over the set of feasible parameters – provides a measure of the maximum possible deviation of $\widehat{\theta}$ from $\bar{\theta}$. In a "standard" scenario, i.e., without sparsity guarantees or black-box compensation, the FPS defined in (8) is the smallest set that contains $(\bar{\theta}, \bar{\omega})$. Hence, the worst-case parametric error is given by

$$e^{FPS}_\theta(\widehat{\theta}) = \sup_{\theta\in\text{FPS}} \|\theta - \widehat{\theta}\|_p. \tag{19}$$

However, when black-box compensation is applied and $\widehat{\omega}$ satisfies the maximum sparsity conditions, SFPS replaces FPS as the minimal guaranteed set, yielding

$$e^{SFPS}_\theta(\widehat{\theta}) = \sup_{\theta\in\text{SFPS}} \|\theta - \widehat{\theta}\|_p. \tag{20}$$

Note that applying (20) to an estimate that does not satisfy maximum sparsity conditions is meaningless, as its inclusion in SFPS is not guaranteed. In such cases, FPS remains the smallest valid set. Thus, based on these definitions, the following theorem establishes that finding the "correct sparsity" of the black model leads to a more accurate physical estimate, i.e., an estimate with a lower worst-case parametric error, compared to "standard" estimates without sparsity guarantees or obtained without a black model augmentation.

**Theorem 6** (*Maximum Sparsity Optimality*). *Let Assumptions 1, 2, 3(i) hold. Consider $\Delta$ as in (18), and let $(\widehat{\theta}, \widehat{\omega})$ be a solution of (15) in the region of attraction of $(\bar{\theta}, \bar{\omega})$, for which conditions of Theorem 5 hold. Consider any other solution $(\widetilde{\theta}, \widetilde{\omega})$, for which conditions of Theorem 5 do not hold. Then, $e_\theta^{\text{SFPS}}(\widehat{\theta}) \leq e_\theta^{\text{FPS}}(\widetilde{\theta})$.*

**Proof.** By Theorem 5, $\text{supp}(\widehat{\omega}) = \text{supp}(\bar{\omega})$, making $\widehat{\omega}$ maximally sparse and ensuring $(\widehat{\theta}, \widehat{\omega}) \in$ SFPS. On the other hand, being $\widetilde{\omega}$ without sparsity guarantees, it is not ensured that $(\widetilde{\theta}, \widetilde{\omega}) \in$ SFPS. Thus, $(\widetilde{\theta}, \widetilde{\omega}) \in$ FPS. From Definition 6 we have SFPS $\subseteq$ FPS. It follows that $e_\theta^{\text{SFPS}}(\widehat{\theta}) = \sup_{\theta \in \text{SFPS}} \|\theta - \widehat{\theta}\|_p \leq \sup_{\theta \in \text{FPS}} \|\theta - \widetilde{\theta}\|_p = e_\theta^{\text{FPS}}(\widetilde{\theta})$, showing that "correct sparsity" yields a better worst-case estimate. ■

**Remark 4** (*On the Benefits of $\delta$*). The result of Theorem 6 is closely linked to Theorem 1. As evident from (5), the better the black-box model compensates for unknown dynamics, the smaller $\widetilde{\Delta}$, thereby tightening the parametric error bound. In particular, note also that, if $\widehat{\omega} = \bar{\omega}$, then $\widetilde{\Delta} = 0$.

## 5. Numerical examples

### 5.1. Academic example

First, we validate our theoretical results identifying the nonlinear vehicle lateral dynamics model in Novara (2011), i.e.,

$$\tilde{y}_{k+1} = \bar{\theta}^\top [\tilde{y}_k,\ \tilde{y}_{k-1},\ \tilde{y}_k\tilde{p}_k,\ \tilde{s}_{k-1}\tilde{p}_{k-1}]^\top + \Delta_k + \eta_k^v,$$
$$\Delta_k = -9.625\tilde{y}_k\tilde{p}_{k-1} + 10.69\tilde{y}_{k-1}\tilde{p}_{k-1} - 11.52\tilde{y}_{k-1}\tilde{p}_{k-1}^2,$$

where $\tilde{s}_k$ (steering angle) and $\tilde{p}_k$ (inverse longitudinal velocity) are the measured inputs generated as in Novara (2011), and $\tilde{y}_k$ is the yaw rate output. The noise term $\eta_k^v$ accounts for both process and measurement disturbances. The physical parameters are $\bar{\theta} = [2, -1.087, -1.070, 3.715]^\top$, and $\Delta_k$ is the unmodeled dynamic. We recast the known part of the system as a state-space model, i.e.,

$$\begin{aligned} x_{k+1} &= [\theta_1 x_{1,k} + \theta_2 x_{2,k} + \theta_3 x_{1,k}x_{5,k} + \theta_4 x_{4,k}x_{6,k},\\ &\qquad x_{1,k},\ u_{1,k},\ x_{3,k},\ u_{2,k},\ x_{5,k}]^\top = f(x_k, u_k, \theta),\\ z_k &= x_{1,k} = h(x_k), \end{aligned}$$

with $x_k = [\tilde{y}_k, \tilde{y}_{k-1}, \tilde{s}_k, \tilde{s}_{k-1}, \tilde{p}_k, \tilde{p}_{k-1}]^\top$, $u_k = [\tilde{s}_k, \tilde{p}_k]$.

To capture the unmodeled dynamics, we augment $f_1$ with $\delta(x_k, u_k, \omega) = \omega^\top \varphi(x_k, u_k)$, where $\varphi(x_k, u_k) = [1, x_{3,k}, x_{4,k}, x_{5,k}, x_{6,k}, x_{3,k}^2, x_{4,k}^2, x_{5,k}^2, x_{4,k}x_{6,k}^2, x_{1,k}x_{6,k}, x_{2,k}x_{6,k}, x_{2,k}x_{6,k}^2]^\top$. This choice is driven by, e.g., the polynomial composition of the known model $f$.

Thus, the "true" black-box parameter is given by $\bar{\omega} = [0, \ldots, 0, -9.625, 10.69, 11.52]^\top$. Using $T = 2000$ input–output samples, we identify $\widehat{\theta}$ and a sparse $\widehat{\omega}$ by minimizing the regularized cost function $\mathcal{C}_T$ (3) with $\mathcal{L}_k = \frac{1}{T}\|e_k\|_2^2$ and $\gamma = 0.001$.

First, we performed 7500 simulations – 5000 with the black-box model and 2500 without – under varying noise (with fixed upper bound) and input conditions. Fig. 1 shows the parametric errors versus $\|\widetilde{\Delta}\|_\infty$, along with the estimated bound $M_\Delta$ and confidence ellipses. The black-box model consistently reduces the error, confirming its effectiveness, while all data points respect the computed bound, validating Theorem 1. Then, Fig. 2 shows the approximated FPS and SFPS as convex hulls in the space of the two principal components of $\theta$ ($\text{pca}_i(\theta)$). Notably, most parameters identified with $\delta$ lie close to the ones in the SFPS, which contains only those for which conditions of Theorem 5 hold, yielding $\text{supp}(\widehat{\omega}) = \text{supp}(\bar{\omega})$ (see Definition 6). In contrast, estimates without $\delta$ are clearly biased. A statistic on the worst-case error analysis over all the simulations (mean $\pm$ 1$\sigma$) further confirms this. In particular, $e_\theta^{\text{SFPS}}(\widehat{\theta})$ is $1.19 \pm 0.19$ for $\theta \in$ SFPS

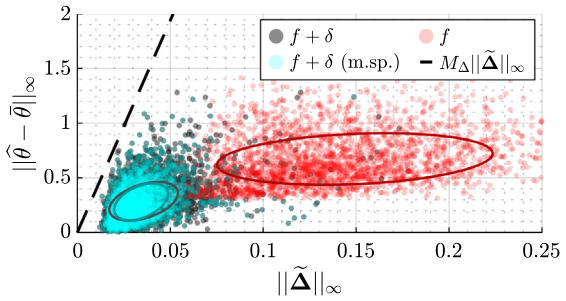


**Fig. 1.** Parametric error with respect to $\|\widetilde{\Delta}\|_\infty$.

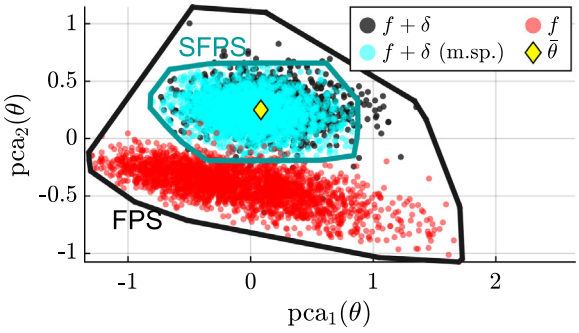


**Fig. 2.** Approximated FPS and SFPS.

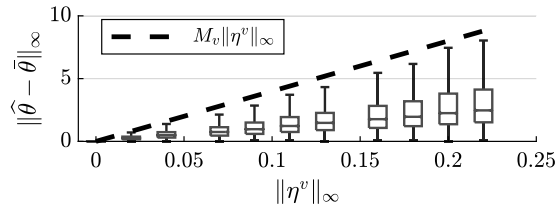


**Fig. 3.** Parametric error with respect to $\|\eta^v\|_\infty$.

(20), compared to $e_\theta^{\text{FPS}} = 1.79 \pm 0.28$ for $\theta \in$ FPS (19), validating Theorem 6. Then, we conducted 5000 simulations for each noise upper bounds $\|\eta^v\|_\infty$, under varying noise and input conditions. Fig. 3 shows the parametric error distribution, with box plots illustrating median, interquartile ranges (IQRs), and whiskers extending to 1.5 $\times$ IQR. The results confirm the linear dependence between error and noise bound, as stated in Theorem 1.

### 5.2. Cascade tank system identification benchmark

In this example, we evaluate the framework on the cascade tank system (CTS) benchmark, described in Schoukens, Mattson, Wigren, and Noël (2016). The CTS consists of two connected tanks, a pump, and free outlets. Water is transferred from a reservoir to the upper tank, drains into the lower tank, and returns. Overflow occurs when a tank is full, with excess water either flowing to the lower tank or leaving the system. The system follows a discrete-time, nonlinear model

$$\begin{aligned} x_{1,k+1} &= x_{1,k} + T_s(-k_1\sqrt{x_{1,k}} + k_4 u_k + v_{1,k}),\\ x_{2,k+1} &= x_{2,k} + T_s(k_2\sqrt{x_{1,k}} - k_3\sqrt{x_{2,k}} + v_{2,k}), \end{aligned} \tag{21}$$

with output $\tilde{z}_k = x_{2,k} + \eta_k^z$, input $u_k \in \mathbb{R}$, states $x_{1,k}, x_{2,k} \in \mathbb{R}$, and noises $v_{1,k}, v_{2,k}, \eta_k^z \in \mathbb{R}$. The unknown physical parameters $k = [k_1, k_2, k_3, k_4]$ and initial state values (equals for both training and validation) must be estimated from training data ($T = 1024$, $T_s = 4$s). Notice that some dynamics remain unmodeled since

**Table 1**
Comparison of identification methods (RMSE on validation).

| Method | RMSE | Method | RMSE |
|---|---|---|---|
| Svensson and Schön (2017) | 0.45 | PWARX (Mattsson et al., 2018) | 0.35 |
| Volt.FB (Schoukens & Scheiwe, 2016) | 0.39 | SED-MPK (Dalla Libera et al., 2021) | 0.48 |
| INN (Mavkov et al., 2020) | 0.41 | PNLSS-I (Relan, Tiels, Marconato, & Schoukens, 2017) | 0.45 |
| NLSS2 (Relan et al., 2017) | 0.34 | NOMAD (Brunot, Janot, & Carrillo, 2017) | 0.37 |
| **Proposed** | 0.26 | | |

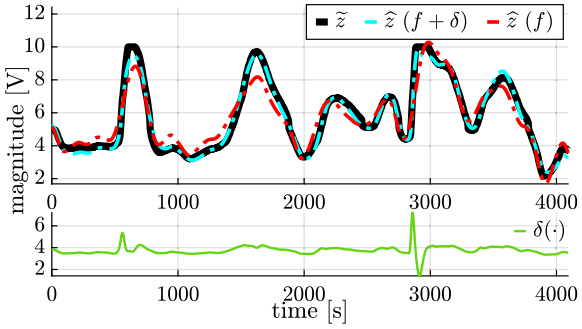


**Fig. 4.** Validation data (simulation).

the model in (21) excludes overflow effects. Thus, the physical model is augmented with a black-box term comprising standard and most-used basis functions (Sjöberg et al., 1995), such as sigmoid, softplus, hyperbolic, trigonometric, and polynomial functions. Then, a regularized cost function $\mathcal{C}_T$ ensures a minimal black box if the physical model (21) adequately describes the system. In this example, $\mathcal{L}_k = \frac{1}{T}\|e_k\|_2^2$ and $\gamma = 0.1$. Parameters are initialized as $\widehat{k}_0 = [0.05, 0.05, 0.05, 0.05]$, $\widehat{x}_0 = [z_0, z_0]^\top$. After the optimization, black-box parameters are set to 0 if $|\Omega_{ij}| \leq 10^{-4}$.

Fig. 4 compares measured and simulated outputs, showing how the black-box augmented model (cyan) effectively compensates for unmodeled dynamics, outperforming the physical-only model (red). This is confirmed by the Root Mean Square Error (RMSE), which improves from 0.603V (training) and 0.670V (validation) to 0.135V and 0.260V, respectively. Thus, to allow a fair comparison between different methods, as indicated in Schoukens et al. (2016), accuracy is compared using the RMSE on the validation dataset, with predictions obtained through simulation.

Table 1 compares our method with state-of-the-art identification approaches, achieving the lowest RMSE[4]. To further assess the identification accuracy, we report the optimal fitness percentage ($\text{fit}_\% = 100(1 - \|\widehat{\boldsymbol{z}}_{0:T-1} - \tilde{\boldsymbol{z}}_{0:T-1}\|_2 / \|\widehat{\boldsymbol{z}}_{0:T-1} - \bar{\boldsymbol{z}}\|_2)$, with $\bar{\boldsymbol{z}} \doteq \frac{1}{T}\sum_{k=0}^{T-1} \tilde{z}_k$): for the physics-only model, $\text{fit}_\%$ is 72.23% (training) and 68.16% (validation), while incorporating $\delta$ improves the accuracy to $\text{fit}_\% = 93.78\%$ (training) and 87.64% (validation). Identified parameters, whose true values are not reported in Schoukens et al. (2016), are $\widehat{k} = [0.076, 0.027, 0.042, 0.039]$, and $\widehat{x}_0 = [3.52, 5.19]^\top$ V.

## 6. Conclusions and future works

This paper presented a unified framework for identifying interpretable nonlinear dynamical models while preserving physical properties. By combining off-white and sparse black-box models, the approach compensates for unmodeled dynamics and provably improves interpretable parameter identification. Theoretical analysis established bounds on parameter estimation errors and identified conditions for sparsity recovery, while a benchmark test demonstrated significant accuracy improvements. Future work will explore adapting the cost function for non-uniform measurements and extending black-box modeling with neural networks or kernel methods.

[4] To ensure metrics consistency, comparison methods provided at https://github.com/MaartenSchoukens/nonlinear_benchmarks are used.

## Acknowledgments

This work was partially supported by the Italian Ministry of Research under the complementary actions to the NRRP "Fit4MedRob - Fit for Medical Robotics" Grant (PNC0000007). C. Donati acknowledges support from PRIN project TECHIE "A control and network-based approach for fostering the adoption of new technologies in the ecological transition", funded by Unione Europea - Next Generation EU, Missione 4 Componente 1 CUP: D53D23001320006.

## Appendix A. Proof of Theorem 1

Consider system (1) with $\Delta = \delta = 0$ and true parameters $\bar{\theta} \in \mathbb{R}^{n_\theta}$. Since $\mathcal{C}_T$ is a function of the noise sequences $\boldsymbol{\eta}$, its minimizer $\theta^\star$ is also a function of $\boldsymbol{\eta}$, i.e., $\theta^\star \equiv \theta^\star(\boldsymbol{\eta})$. Therefore,

$$\theta^\star - \bar{\theta} = \theta^\star(\boldsymbol{\eta}) - \theta^\star(\mathbf{0}). \tag{22}$$

From the Mean Value Theorem, a $\check{\boldsymbol{\eta}}$ exists such that

$$\theta^\star(\boldsymbol{\eta}) - \theta^\star(\mathbf{0}) = \frac{\partial \theta^\star(\check{\boldsymbol{\eta}})}{\partial \boldsymbol{\eta}}\boldsymbol{\eta} = \frac{\partial \theta^\star(\check{\boldsymbol{\eta}})}{\partial \boldsymbol{\eta}^u}\boldsymbol{\eta}^u_{0:T} + \frac{\partial \theta^\star(\check{\boldsymbol{\eta}})}{\partial \boldsymbol{\eta}^z}\boldsymbol{\eta}^z_{0:T} \tag{23}$$

where $\frac{\partial \theta^\star(\check{\boldsymbol{\eta}})}{\partial \boldsymbol{\eta}^u}$ and $\frac{\partial \theta^\star(\check{\boldsymbol{\eta}})}{\partial \boldsymbol{\eta}^z}$ are the matrices splitting the contributions from process and measurement noise. Since $\theta^\star$ minimizes $\mathcal{C}_T$, it satisfies $\frac{\partial \mathcal{C}_T(\boldsymbol{\eta},\theta)}{\partial \theta} = 0$. Hence, it follows from implicit differentiation that $\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\eta}}\frac{\partial \mathcal{C}_T(\boldsymbol{\eta},\theta)}{\partial \theta}\big|_{\theta=\theta^\star} = \frac{\partial^2 \mathcal{C}_T(\boldsymbol{\eta},\theta^\star)}{\partial \boldsymbol{\eta}\partial \theta} + \frac{\partial^2 \mathcal{C}_T(\boldsymbol{\eta},\theta^\star)}{\partial^2 \theta}\frac{\partial \theta^\star}{\partial \boldsymbol{\eta}} = G(\boldsymbol{\eta}) + H(\boldsymbol{\eta})\frac{\partial \theta^\star}{\partial \boldsymbol{\eta}} = 0$. Being the Hessian $H$ invertible by assumption, then $\frac{\partial \theta^\star}{\partial \boldsymbol{\eta}} = -H^{-1}G(\boldsymbol{\eta})$. Thus, recalling the definitions of $M_u$ and $M_z$ in Lemma 1, it follows from (22) and (23) that

$$\left\|\theta^\star - \bar{\theta}\right\|_p \leq M_u \left\|\boldsymbol{\eta}^u_{0:T}\right\|_p + M_z \left\|\boldsymbol{\eta}^z_{0:T}\right\|_p. \tag{24}$$

Now, consider $\Delta \neq 0$. We observe that $\mathcal{C}_T$ is also a function of $\widetilde{\boldsymbol{\Delta}}$ so that $\theta^\star \equiv \theta^\star(\boldsymbol{\eta}, \widetilde{\boldsymbol{\Delta}})$. Thus, we have

$$\theta^\star - \bar{\theta} = \theta^\star(\boldsymbol{\eta}, \widetilde{\boldsymbol{\Delta}}) - \theta^\star(\mathbf{0}, \widetilde{\boldsymbol{\Delta}}) + \theta^\star(\mathbf{0}, \widetilde{\boldsymbol{\Delta}}) - \theta^\star(\mathbf{0}, \mathbf{0}). \tag{25}$$

Applying the implicit function theorem to $P = \frac{\partial \mathcal{C}_T(\mathbf{0}, \widetilde{\Delta}, \theta)}{\partial \theta}\big|_{\theta=\theta^\star}$ it follows that $\theta^\star(\mathbf{0}, \widetilde{\boldsymbol{\Delta}})$ is continuously differentiable with respect to $\widetilde{\boldsymbol{\Delta}}$. Thus, by Lipschitz continuity, there exists a constant $M_\Delta < \infty$ such that

$$\|\theta^\star(\mathbf{0}, \widetilde{\boldsymbol{\Delta}}) - \theta^\star(\mathbf{0}, \mathbf{0})\|_p \leq M_\Delta \|\widetilde{\boldsymbol{\Delta}}\|_p. \tag{26}$$

Now, combining (24) for $\Delta \neq 0$ with (26), we have

$$\begin{aligned} &\left\|\theta^\star(\boldsymbol{\eta}, \widetilde{\boldsymbol{\Delta}}) - \theta^\star(\mathbf{0}, \widetilde{\boldsymbol{\Delta}})\right\|_p + \|\theta^\star(\mathbf{0}, \widetilde{\boldsymbol{\Delta}}) - \theta^\star(\mathbf{0}, \mathbf{0})\|_p \\ &\quad \leq M_u \left\|\boldsymbol{\eta}^u_{0:T}\right\|_p + M_z \left\|\boldsymbol{\eta}^z_{0:T}\right\|_p + M_\Delta \|\widetilde{\boldsymbol{\Delta}}\|_p. \end{aligned} \tag{27}$$

Finally, recalling the triangle inequality, (27) yields (5).

## Appendix B. Measurement model

Given (1), $\tilde{\mathbf{u}}_{0:T}, \tilde{\mathbf{z}}_{0:T}$, $\tilde{u}_k = u_k + \eta_k^u$, $\tilde{z}_k = z_k + \eta_k^z$, we have $\tilde{z}_{k+1} = h(f(x_k, u_k, \theta) + \Delta(x_k, u_k)) + \eta_{k+1}^z$. Let $\bar{h}(x_k, u_k, \theta) \doteq h(f(x_k, u_k, \theta) + \Delta(x_k, u_k))$. Substituting $u_k = \tilde{u}_k - \eta_k^u$, we obtain $\tilde{z}_{k+1} = \bar{h}\left(x_k, \tilde{u}_k - \eta_k^u, \theta\right) + \eta_{k+1}^z$. Thus, applying the mean value theorem, there exists a $\check{u} \in [\tilde{u}_k - \eta_k^u, \tilde{u}_k]$ such that $\bar{h}\left(x_k, \tilde{u}_k - \eta_k^u, \theta\right) - \bar{h}\left(x_k, \tilde{u}_k, \theta\right) = \frac{\partial \bar{h}(\check{u})}{\partial u_k}\eta_k^u$, so that, defining $\eta_k^v = \frac{\partial \bar{h}(\check{u})}{\partial u_k}\eta_k^u + \eta_{k+1}^z$, that accounts for measurement and process noise, we obtain $\tilde{z}_{k+1} = h(f\left(x_k, \tilde{u}_k, \theta\right) + \Delta(x_k, \tilde{u}_k)) + \eta_k^v$.

## Appendix C. Proof of Theorem 2

The optimization problem (10) is a *feasibility problem* in $\theta$. Therefore, we equally reformulate it by selecting $\omega$ for which a $\theta$ exists that satisfies the constraint, i.e.,

$$\widehat{\omega} = \arg\min_{\omega} \|\omega\|_1, \text{ s.t. } \omega \in \left\{\omega : \exists \theta \text{ s.t. } \|\tilde{z} - \Xi\theta - \Phi\omega\|_2 \leq \mu\right\}.$$

Moreover, since at least one $\theta \in \mathbb{R}^{n_\theta}$ satisfies the constraint, we further refine the problem by seeking the $\omega$ for which a suitable $\theta \in \mathbb{R}^{n_\theta}$ minimizes the error, i.e.,

$$\widehat{\omega} = \arg\min_{\omega} \|\omega\|_1 \tag{28a}$$

$$\text{s.t. } \omega \in \{\omega : \min_{\theta \in \mathbb{R}^{n_\theta}} \|\tilde{z} - \Xi\theta - \Phi\omega\|_2 \leq \mu\}. \tag{28b}$$

The optimal $\theta$ minimizing $\|\tilde{z} - \Xi\theta - \Phi\omega\|_2$ is given the least squares optimal solution, i.e. $\theta^\star(\omega) = \Xi^\dagger(\tilde{z} - \Phi\omega)$, that once substituted in (28b) gives $g(\omega) = \|\tilde{z} - \Xi\Xi^\dagger(\tilde{z} - \Phi\omega) - \Phi\omega\|_2 = \|\Upsilon_0(\tilde{z} - \Phi\omega)\|_2 = \|e^\star_{\Xi,\Phi}(\tilde{z}, \omega)\|_2$, recalling $\Upsilon_0 = (I_T - \Xi\Xi^\dagger)$ (see Lemma 2). Thus, (28) simplifies to $\widehat{\omega} = \arg\min_\omega \|\omega\|_1$ s.t. $\|\Upsilon_0(\tilde{z} - \Phi\omega)\|_2 \leq \mu$. By applying Gribonval et al. (2006, Theorem 1, Corollary 1), with $\Upsilon_0\widehat{z} = \Upsilon_0\Phi\widehat{\omega} + e^\star_{\Xi,\Phi}(\tilde{z}, \widehat{\omega})$, the result follows.

## Appendix D. Proof of Theorem 3

As in Appendix C, the optimization problem (9) with respect to $\omega$ can be written as

$$\bar{\omega} = \arg\min_{\omega \in \mathbb{R}^m} \|\omega\|_0 \quad \text{s.t. } \|\Upsilon_0\tilde{z} - \Upsilon_0\Phi\omega\|_2 \leq \mu.$$

By definition, $\bar{\omega}$ is the sparsest vector satisfying the constraint $\|\Upsilon_0\tilde{z} - \Upsilon_0\Phi\omega\|_2 \leq \mu$. Thus, consider $\omega^v$, solution of (11) (Novara, 2012). Since constraint (11d) is active, we have $\|e^\star_{\Xi,\Phi}(\tilde{z}, \bar{\omega})\|_2 \leq \|e^\star_{\Xi,\Phi}(\tilde{z}, \omega^v)\|_2 = \mu$. Moreover, $\|\bar{\omega}\|_0 \leq \|\omega^v\|_0$, since $\bar{\omega}$ is maximally sparse. Applying Theorem 2, $\|\bar{\omega} - \omega^v\|_\infty \leq \frac{\|e^\star(\tilde{z},\omega^v)\|_{(\Upsilon_0,1)} + |e^\star(\tilde{z},\omega^v)|_{(\Upsilon_0,2M)}}{\sigma^2_{2M}(\Upsilon_0)}$ holds, implying that, for all $i \in [1, m]$, $|\bar{\omega}_i - \omega_i^v| \leq \frac{\|e^\star(\tilde{z},\omega^v)\|_{(\Upsilon_0,1)} + \|e^\star(\tilde{z},\omega^v)\|_{(\Upsilon_0,2M)}}{\sigma^2_{2M}(\Upsilon_0)}$. Hence, if $|\omega_i^v| > \frac{\|e^\star(\tilde{z},\omega^v)\|_{(\Upsilon_0,1)} + \|e^\star(\tilde{z},\omega^v)\|_{(\Upsilon_0,2M)}}{\sigma^2_{2M}(\Phi^\Xi)}$, then $\bar{\omega} \neq 0$ and, consequently $\lambda \doteq \{i : |\omega_i^v| > \frac{\|e^\star(\tilde{z},\omega^v)\|_{(\Upsilon_0,1)} + \|e^\star(\tilde{z},\omega^v)\|_{(\Upsilon_0,2M)}}{\sigma^2_{2M}(\Upsilon_0)}\} \subseteq \text{supp}(\bar{\omega})$. If follows that $\|\bar{\omega}\|_0 \geq \text{card}(\lambda)$, which yields (12). Now, the constraints (11b) and (11c) imply that $\text{supp}(\widehat{\omega}) = \{i : |\omega_i^v| \geq \eta(\widehat{\omega})\}$. Moreover, if condition (13) holds, then $\{i : |\omega_i^v| \geq \eta(\widehat{\omega})\} \subseteq \lambda \subseteq \text{supp}(\bar{\omega})$. It follows that $\text{supp}(\widehat{\omega}) \subseteq \text{supp}(\bar{\omega})$. Since $\bar{\omega}$ is the sparsest vector satisfying $\|\Upsilon_0\tilde{z} - \Upsilon_0\Phi\omega\|_2 \leq \mu$ relation $\text{supp}(\widehat{\omega}) = \text{supp}(\bar{\omega})$ follows, concluding the proof.

## References


Bellman, R., & Åström, K. J. (1970). On structural identifiability. *Mathematical Biosciences*, *7*(3–4), 329–339.

Brunot, M., Janot, A., & Carrillo, F. (2017). Continuous-time nonlinear systems identification with output error method based on derivative-free optimisation. *IFAC-PapersOnLine*, *50*(1), 464–469.

Chiuso, A., & Pillonetto, G. (2019). System identification: A machine learning perspective. *Annual Review of Control, Robotics, and Autonomous Systems*, *2*, 281–304.

Dalla Libera, A., Carli, R., & Pillonetto, G. (2021). Kernel-based methods for Volterra series identification. *Automatica*, *129*, Article 109686.

Donati, C., Mammarella, M., Dabbene, F., Novara, C., & Lagoa, C. (2024). One-shot backpropagation for multi-step prediction in physics-based system identification. *IFAC-PapersOnLine*, *58*(15), 271–276.

Donati, C., Mammarella, M., Dabbene, F., Novara, C., & Lagoa, C. (2025). A scalable, gradient-stable approach to multi-step, nonlinear system identification using first-order methods. In *6th IFAC Workshop on Linear Parameter Varying Systems*. (in press, see arXiv:2410.03544).

Gribonval, R., Figueras i Ventura, R., & Vandergheynst, P. (2006). A simple test to check the optimality of a sparse signal approximation. *Signal Processing*, *86*(3), 496–510.

Kaheman, K., Kaiser, E., Strom, B., Kutz, J. N., & Brunton, S. L. (2019). Learning discrepancy models from experimental data. In *58th IEEE conf. decision and control*.

Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., & Yang, L. (2021). Physics-informed machine learning. *Nature Reviews Physics*, *3*(6), 422–440.

Liu, Y., Tóth, R., & Schoukens, M. (2024). Physics-guided state-space model augmentation using weighted regularized neural networks. *IFAC-PapersOnLine*, *58*(15), 295–300.

Ljung, L. (2010). Perspectives on system identification. *Annual Reviews in Control*, *34*, 1–12.

Mammarella, M., Donati, C., Dabbene, F., Novara, C., & Lagoa, C. (2024). A blended physics-based and black-box identification approach for spacecraft inertia estimation. In *2024 IEEE 63rd conference on decision and control* (pp. 8282–8287).

Mattsson, P., Zachariah, D., & Stoica, P. (2018). Identification of cascade water tanks using a PWARX model. *Mechanical Systems and Signal Processing*, *106*, 40–48.

Mavkov, B., Forgione, M., & Piga, D. (2020). Integrated neural networks for nonlinear continuous-time system identification. *IEEE Control Systems Letters*, *4*(4), 851–856.

Novara, C. (2011). Sparse identification of nonlinear functions and parametric set membership optimality analysis. In *Proceedings of the 2011 American control conference* (pp. 663–668).

Novara, C. (2012). Sparse identification of nonlinear functions and parametric set membership optimality analysis. *IEEE Transactions on Automatic Control*, *57*, 3236–3241.

Quaghebeur, W., Nopens, I., & De Baets, B. (2021). Incorporating unmodeled dynamics into first-principles models through machine learning. *IEEE Access*, *9*, 22014–22022.

Relan, R., Tiels, K., Marconato, A., & Schoukens, J. (2017). An unstructured flexible nonlinear model for the cascaded water-tanks benchmark. *IFAC-PapersOnLine*, *50*(1), 452–457.

Schmidt, M., Fung, G., & Rosales, R. (2007). Fast optimization methods for l1 regularization: A comparative study and two new approaches. In *18th European conference on machine learning* (pp. 286–297). Springer.

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

Schoukens, M., Mattson, P., Wigren, T., & Noël, J.-P. (2016). Cascaded tanks benchmark combining soft and hard nonlinearities. In *Workshop on nonlinear system identification benchmarks* (pp. 20–23).

Schoukens, M., & Scheiwe, F. G. (2016). Modeling nonlinear systems using a Volterra feedback model. In *Workshop on nonlinear system identification benchmarks*.

Sjöberg, J., Zhang, Q., Ljung, L., Benveniste, A., Delyon, B., Glorennec, P.-Y., et al. (1995). Nonlinear black-box modeling in system identification: a unified overview. *Automatica*, *31*(12), 1691–1724.

Svensson, A., & Schön, T. B. (2017). A flexible state–space model for learning nonlinear dynamical systems. *Automatica*, *80*, 189–199.

Zakwan, M., Di Natale, L., Svetozarevic, B., Heer, P., Jones, C. N., & Trecate, G. F. (2023). Physically consistent neural ODEs for learning multi-physics systems. *IFAC-PapersOnLine*, *56*(2), 5855–5860.

Zorzi, M., & Chiuso, A. (2017). Sparse plus low rank network identification: A nonparametric approach. *Automatica*, *76*, 355–366.


**Cesare Donati** received the B.Sc. degree in Computer Engineering and the M.Sc. degree in Computer Engineering from Politecnico di Torino, Italy, in 2019 and 2021, respectively. He is currently pursuing the Ph.D. degree in Electrical, Electronics, and Communication Engineering at Politecnico di Torino. In 2024, he was a Visiting Scholar at The Pennsylvania State University. He is also a Research Associate at the Institute of Electronics, Computer and Telecommunication Engineering of the Italian National research Council (CNR-IEIIT).

He is a member of the IEEE committee on System Identification and Adaptive Control. His research interests include system identification, physics-based modeling, machine learning, filtering/estimation, and optimization.

**Martina Mammarella** received her M.Sc. (2015) and Ph.D. (2019) in Aerospace Engineering both from Politecnico di Torino. She is a researcher at CNR-IEIIT and adjunct professor at Politecnico di Torino. Her domain of expertise are spaceflight mechanics and robust and stochastic predictive control techniques for constrained systems, mainly for satellites and aerial drones. Dr. Mammarella has also been involved in various Italian and European projects and collaborates with various international universities and research centers. Currently, she is Associate Editor for the IEEE CSS Technology Conferences Editorial Board and Associate Editor for IEEE Transactions on Aerospace and Electronic Systems and IEEE Transactions on Control Systems Technology journal. She is an active member of IFAC TC on Aerospace, IFAC TC on Control in Agriculture, IFAC TC on Robust Control, IEEE CSS TC on Aerospace Control, IEEE RAS TC on Agricultural Robotics and Automation, and AIAA TC on Guidance, Navigation and Control.

**Fabrizio Dabbene** is a Director of Research at the institute IEIIT of the National Research Council of Italy (CNR), where he coordinates the Information and Systems Engineering Group. He has held visiting and research positions with The University of Iowa, Penn State University, and the Russian Academy of Sciences, Institute of Control Science, Moscow, Russia. He has authored or coauthored more than 150 research papers and two books. Dr. Dabbene was an Elected Member of the Board of Governors, from 2014 to 2016. He has served as the vice president for publications, from 2015 to 2016. He has also served as an Associate Editor for Automatica (2008–2014), and IEEE Transactions on Automatic Control (2008–2012) and as Senior Editor of the IEEE Control Systems Society Letters (2018–2023), and he is currently Senior Editor for the IEEE Transactions on Control Systems Technology. He chaired the IEEE-CSS Italy Chapter (2019–2024) and since 2023 he serves as NMO representative for Italy at the International Federation of Automatic Control (IFAC). He is recipient of the 2024 IEEE CSS Distinguished Member Award.

**Carlo Novara** received the Laurea degree in Physics from Universitá di Torino (Italy) in 1996 and the Ph.D. degree in Computer and System Engineering from Politecnico di Torino (Italy) in 2002. He held a visiting researcher position at the University of California at Berkeley in 2001 and 2004. He is currently a Full Professor at Politecnico di Torino in the field of Automatic Control. He is the author or co-author of about 200 scientific publications in international journals and conference proceedings. He has been involved in several national and international projects, in collaboration with Italian and European institutions/companies. He is the co-author of several patents in the automotive field. He is a member of the IEEE committee on System Identification and Adaptive Control, of the IFAC committee on Modelling, Identification and Signal Processing, of the IFAC committee on Modeling and control of environmental systems, and a founding member of the IEEE-CSS committee on Medical and Healthcare Systems. His research interests include system identification, machine learning, filtering/estimation, nonlinear control, predictive control, data-driven methods, set membership methods, nonlinear optimization, quantum computing, and aerospace, automotive, and energy applications.

**Constantino M. Lagoa** received the Ph.D. degree from the University of Wisconsin at Madison, Madison, WI, USA, in 1998. In 1998, he joined the Electrical Engineering Department, Pennsylvania State University, State College, PA, USA, where he is currently a Professor. His research interests include robust optimization and control, chance constrained optimization, controller design under risk specifications, system identification, and control of computer networks. He was an Associate Editor for IEEE Transactions on Automatic Control during 2012–2017 and IEEE Transactions on Control systems Technology during 2009–2013. He is an Associate Editor for Automatica.