Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnline CONFERENCE PAPER ARCHIVE

IFAC PapersOnLine 58-15 (2024) 187–192

Differentiable multi-ridge regression for system identification

Gabriele Maroni, Loris Cannelli, Dario Piga

IDSIA Dalle Molle Institute for Artificial Intelligence USI-SUPSI, Via la Santa 1, CH- 6962 Lugano-Viganello, Switzerland.

Abstract: Regularization aims to shrink model parameters, reducing complexity and overfitting risk. Traditional methods like LASSO and Ridge regression, limited by a single regularization hyperparameter, can restrict bias-variance trade-off adaptability. This paper addresses system identification in a multi-ridge regression framework, where an ℓ_2-penalty on the model coefficients is introduced, and a different regularization hyperparameter is assigned to each model parameter. To compute the optimal hyperparameters, a cross-validation-based criterion is optimized through gradient descent. Autoregressive and Output Error models are considered. The former requires formulating a regularized least-squares problem. The identification of the latter class is more challenging and is addressed by adopting regularized instrumental variable methods to ensure a consistent parameter estimation.

Keywords: System identification, Multi-ridge regression, Instrumental Variables.

1. INTRODUCTION

Regularization is a key strategy in system identification for managing the trade-off between accuracy and simplicity of the model. By shrinking the model parameters towards zero, regularization effectively simplifies the model, mitigating the risk of overfitting. Additionally, regularization often leads to better numerical properties, particularly in scenarios characterized by multicollinearity, where regressors are highly correlated with each other.

LASSO (Tibshirani (1996)) and Ridge regression (Hoerl and Kennard (1970)) are the most common regularization methods, also commonly adopted for identification of dynamical models with a a sparse structure (see., e.g., Pillonetto et al. (2022); Piga and Tóth (2013b,a); Ohlsson et al. (2010)). These approaches impose penalties on the ℓ_1 and ℓ_2 norms of the model coefficients, respectively. The significance of the penalty term is determined by a single hyperparameter. The usage of only one hyperparameter simplifies its choice, often achieved through random or grid search within a cross-validation setting. However, regularization by its nature induces a bias in the model parameters, and the usage of only one hyperparameter can restrict the adaptability in balancing the bias-variance tradeoff. For instance, in the case of sparse models, a more appropriate approach might be to selectively encourage the parameters linked to less relevant inputs to shrink towards zero, instead of uniformly penalizing all model coefficients. Although theoretically attractive, assigning individual regularization hyperparameters to model coefficient increases the dimensionality of the hyperparameter search space, making grid or random search impractical.

To overcome the problem of the curse of dimensionality, Bengio (2000) proposed a gradient-based approach to optimize the regularization hyperparameters in Ridge regression with multiple hyperparameters (referred to as "multi-ridge" hereafter). A similar problem is addressed by the authors in Maroni et al. (2023a), where a custom approach is proposed to efficiently compute an analytical expression for the gradient of a cross-validation-based criterion with respect to the hyperparameters, leveraging matrix differential calculus.

This paper builds upon the work Maroni et al. (2023a), which is adapted to the setting of dynamical system learning. Two structures are considered: *Autoregressive with Exogenous Inputs* (ARX) and *Output Error* (OE) models. In the ARX case, a regularized least-squares problem is formulated. As for OE models, it is known from classic system identification that the least-squares solution is not a consistent estimate of the true model parameters even with no regularization. To ensure consistency in the estimation of the model parameters, we formulate the multi-ridge problem using *instrumental variable* (IV) methods (Söderström and Stoica (2002)). This necessitates adapting the cross-validation criterion for optimizing the regularization hyperparameters, including the corresponding gradient with respect to these hyperparameters.

2. PROBLEM SETTING

2.1 Data-generating system

Let us consider a data-generating dynamical system S_o with the input-output representation:

$$y(k) = x_o'(k)\theta_o + e(k), \tag{1}$$

⋆ This work was supported in part by the European Union, Grant Agreement n. 101093126 (project ACES "Autopoietic Cognitive Edge-cloud Services") and by the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract number 22.00490. The activities of G. Maroni were supported by the Hasler Foundation under the project "Accurate and interpretable deep learning models for peak load forecasting".

10.1016/j.ifacol.2024.08.526

where k represents the time index; $y(k) \in \mathbb{R}$ is the system output at time k; $e(k) \in \mathbb{R}$ is a random noise; $\theta_{\rm o} \in \mathbb{R}^{n_\theta}$ is a vector of unknown parameters that have to be estimated; and $x_{\rm o}(k) \in \mathbb{R}^{n_\theta}$ is a regressor vector, defined as follows:

- **ARX structure**: $x_{\rm o}(k)$ is composed by past values of the system inputs and *measured* outputs, *i.e.,*
$$x_{\rm o}(k) = [y(k-1), \dots, y(k-n_{\rm a}), u(k-1), \dots, u(k-n_{\rm b})]', \tag{2}$$
where $u(k)$ is the system input at time k, and $n_{\rm a}, n_{\rm b} \in \mathbb{N}$ characterize the dynamical order of the system.
- **Output-error structure**: $x_{\rm o}(k)$ is composed by past values of the system inputs and *noise-free* outputs, *i.e.,*
$$x_{\rm o}(k) = [y_{\rm o}(k-1), \dots, y_{\rm o}(k-n_{\rm a}), u(k-1), \dots, u(k-n_{\rm b})]', \tag{3}$$
where $y_{\rm o}(k)$ is the noise-free output: $y_{\rm o}(k) = y(k) - e(k)$.

As for the noise $e(k)$, we assume that the noise samples are *i.i.d.*, zero mean, with finite variance σ^2, and statistically independent of $u(k)$ and $y_{\rm o}(k)$. Thus, for all indexes k, t:

$$\mathbb{E}[e(k)] = \mathbb{E}[e(k)u(t)] = \mathbb{E}[e(k)y_{\rm o}(t)] = 0, \tag{4a}$$
$$\mathbb{E}[e(k)e(t)] = \delta_{k,t}\sigma^2. \tag{4b}$$

where $\delta_{k,t} = 1$ if $k = t$, 0 otherwise.

2.2 Identification problem

We consider a parametric identification problem, where we aim to estimate the unknown parameters $\theta_{\rm o}$ from a dataset of N input-output pairs $\mathcal{D} = \{(u(k), y(k))\}_{k=1}^{N}$.

To this aim, we minimize a regularized empirical risk:

$$\hat{\theta}(\lambda) = \arg\min_{\theta} L(\theta, \mathcal{D}) + \frac{1}{2}\theta'\Lambda'\Lambda\theta, \tag{5}$$

where $L(\theta, \mathcal{D})$ is a scalar loss function constructed based on the available dataset $\mathcal{D}$, and $\Lambda \in \mathbb{R}^{n_\theta, n_\theta}$ is a diagonal matrix with diagonal elements given by a vector of regularization hyperparameters $\lambda \in \mathbb{R}^{n_\theta}$.

The following quadratic loss will be considered:

- **ARX structure.** In the case of ARX structures, the following quadratic loss $L(\theta, \mathcal{D})$ is considered:
$$L(\theta, \mathcal{D}) = L_{\rm LS}(\theta, \mathcal{D}) = \frac{1}{2N}\sum_{k=1}^{N} \|y(k) - x'(k)\theta\|^2, \tag{6}$$
where $x(k)$ represents the model's regressor which coincides with the system's regressor (2), i.e.,
$$x(k) = [y(k-1)\ \dots y(k-n_{\rm a})\ u(k-1)\ \dots u(k-n_{\rm b})]'. \tag{7}$$
The solution of the optimization problem (5), for $L(\theta, \mathcal{D}) = L_{\rm LS}(\theta, \mathcal{D})$ is given by:
$$\hat{\theta}_{\rm LS}(\lambda) = \left(\frac{1}{N}\sum_{k=1}^{N} x(k)x'(k) + \Lambda'\Lambda\right)^{-1} \frac{1}{N}\sum_{k=1}^{N} x(k)y(k). \tag{8}$$
- **OE model.** In the case of OE structures, the following quadratic loss $L(\theta, \mathcal{D})$ is considered:
$$L(\theta, \mathcal{D}) = L_{\rm IV}(\theta, \mathcal{D}) = \frac{1}{2N^2}\left\|\sum_{k=1}^{N} z(k)(y(k) - x'(k)\theta)\right\|^2, \tag{9}$$
with the same regressor vector $x(k)$ in (7), and with $z(k) \in \mathbb{R}^{n_\theta}$ being a vector of the so-called *instrumental variables*, chosen to satisfy the following properties:
  (1) the matrix $\Xi_\star = \lim_{N\to\infty} \frac{1}{N}\sum_{k=1}^{N} x(k)z'(k)$ exists;
  (2) the instrumental variable vector $z(k)$ is statistically independent of the noise $e(t)$. Thus, for all k, t:
$$\mathbb{E}[z(k)e(t)] = \mathbb{E}[z(k)]\,\mathbb{E}[e(t)] = 0, \tag{10}$$
where the last equality follows from the zero-mean assumption of the noise (see (4)).

The solution of problem (5), for $L(\theta, \mathcal{D}) = L_{\rm IV}(\theta, \mathcal{D})$ can be computed analytically and is given by:

$$\hat{\theta}_{IV}(\lambda) = \left(\frac{1}{N^2}\sum_{k=1}^{N} x(k)z'(k)\sum_{k=1}^{N} z(k)x'(k) + \Lambda'\Lambda\right)^{-1} \times \frac{1}{N^2}\sum_{k=1}^{N} x(k)z'(k)\sum_{k=1}^{N} z(k)y(k). \tag{11}$$

The loss $L_{\rm IV}(\theta, \mathcal{D})$ is used for OE structures to guarantee consistency of the estimate (Section 3).

2.3 Multi-fold cross-validation

In multi-fold cross-validation, the dataset $\mathcal{D}$ is divided into Q partitions $\mathcal{D}_1, \dots, \mathcal{D}_Q$, and, the following evaluation criterion is minimized to select the hyperparameters λ:

$$E(\lambda) = \sum_{q=1}^{Q} L_q\left(\hat{\theta}^{(\backslash q)}(\lambda), \mathcal{D}_q\right), \tag{12}$$

where $\hat{\theta}^{(\backslash q)}(\lambda)$ represents the estimate of the model parameters θ obtained by solving the training problem (5) using the subsets $\mathcal{D}_j$ with $j = 1, \dots, Q$ and $j \neq q$, i.e.,

$$\hat{\theta}^{(\backslash q)}(\lambda) = \arg\min_{\theta} L\left(\theta, \bigcup_{j\neq k} \mathcal{D}_j\right) + \frac{1}{2}\theta'\Lambda'\Lambda\theta. \tag{13}$$

L_q in (12) represents the average loss defined as in (6) and (9) for ARX and OE structures, respectively, but constructed using the subset $\mathcal{D}_q$ (namely, the *validation fold*) instead of the entire set $\mathcal{D}$, i.e,

- **ARX structure:**
$$L_q\left(\hat{\theta}^{(\backslash q)}(\lambda), \mathcal{D}_q\right) = L_{q,\rm LS}\left(\hat{\theta}^{(\backslash q)}(\lambda), \mathcal{D}_q\right) = \frac{1}{2|\mathcal{D}_q|}\sum_{\substack{k=1 \\ k:\{u(k),y(k)\}\in\mathcal{D}_q}}^{N} \left\|y(k) - x'(k)\hat{\theta}^{(\backslash q)}(\lambda)\right\|^2. \tag{14a}$$
where $|\mathcal{D}_q|$ denotes the cardinality of the set $\mathcal{D}_q$.
- **OE structure:**
$$L_q\left(\hat{\theta}^{(\backslash q)}(\lambda), \mathcal{D}_q\right) = L_{q,\rm IV}\left(\hat{\theta}^{(\backslash q)}(\lambda), \mathcal{D}_q\right) = \frac{1}{2|\mathcal{D}_q|^2}\left\|\sum_{\substack{k=1 \\ k:\{u(k),y(k)\}\in\mathcal{D}_q}}^{N} z(k)(y(k) - x'(k)\hat{\theta}^{(\backslash q)}(\lambda))\right\|^2. \tag{14b}$$

To simplify the notation, without loss of generality, we assume that all the subsets have the same size, i.e., $|\mathcal{D}_q| = N_V = \frac{N}{Q}$, for all $q = 1, \dots, Q$. Thus, the length of each training set is given by $\left|\bigcup_{j\neq k} \mathcal{D}_j\right| = N_T = N - N_V$.

The selected hyperparameter vector λ is then given by:

$$\lambda^\star = \arg\min_{\lambda} E(\lambda). \tag{15}$$

The choice of λ turns out to be a bilevel optimization problem, where the loss of the outer problem (15) depends on the solution $\hat{\theta}^{(\setminus q)}(\lambda)$ of the inner problem (13).

3. ASYMPTOTIC BEHAVIOUR

The motivation behind optimizing multiple hyperparameters comes from the limitations of LASSO and Ridge regression. These methods reduce the variance in the estimate, but also introduce bias by shrinking model parameters towards zero. In contrast, multiple hyperparameters allow to obtain an unbiased estimate. To elucidate this concept, let us fix a scalar $\bar{\lambda} \in \mathbb{R}$, $\bar{\lambda} \neq 0$, and let $\theta_{\mathrm{o},j}$ be the j-th element of the true vector θ_o, which is supposed to be sparse. Then, a matrix Λ with diagonal elements:

$$\lambda_j = \begin{cases} \bar{\lambda} & \text{if } \theta_{\mathrm{o},j} = 0 \\ 0 & \text{if } \theta_{\mathrm{o},j} \neq 0 \end{cases}, \quad j = 1, \dots, n_\theta, \tag{16}$$

shrinks towards zero only the parameters θ_j such that the corresponding true parameters are $\theta_{\mathrm{o},j} = 0$.

3.1 Consistency in training

The following propositions provide consistency results for the ARX and OE model structures.

Proposition 1. [Consistency for ARX models: training]

Let us assume that:

- the diagonal elements of the hyperparameter matrix $\Lambda \in \mathbb{R}^{n_\theta \times n_\theta}$ are given and have the structure in (16);
- the following matrices exist and are invertible:

$$\Gamma = \frac{1}{N}\sum_{k=1}^{N} x(k)x'(k) + \Lambda'\Lambda, \quad \Gamma_\star = \lim_{N\to\infty} \Gamma.$$

Then, $\hat{\theta}_{LS}(\lambda)$ in (8) is a consistent estimate of θ_o, i.e.,

$$\lim_{N\to\infty} \hat{\theta}_{LS}(\lambda) = \theta_\mathrm{o}, \quad \text{w.p. 1.} \tag{17}$$

Proof:

Let us substitute the output $y(k)$ (see (1)) into (8):

$$\begin{aligned}\hat{\theta}_{LS}(\lambda) =& \Gamma^{-1}\frac{1}{N}\sum_{k=1}^{N} x(k)(x'(k)\theta_\mathrm{o} + e(k)) \\ =& \underbrace{\Gamma^{-1}\frac{1}{N}\sum_{k=1}^{N} x(k)x'(k)\theta_\mathrm{o}}_{(a)} + \underbrace{\Gamma^{-1}\frac{1}{N}\sum_{k=1}^{N} x(k)e(k)}_{(b)},\end{aligned} \tag{18}$$

where we also used the property that for the ARX model structure, $x_\mathrm{o}(k) = x(k)$.

The term (a) in (18) can be rewritten as:

$$\begin{aligned}&\frac{\Gamma^{-1}}{N}\sum_{k=1}^{N} x(k)x'(k)\theta_\mathrm{o} = \frac{\Gamma^{-1}}{N}\sum_{k=1}^{N}(x(k)x'(k) + \Lambda'\Lambda - \Lambda'\Lambda)\theta_\mathrm{o} \\ &= \Gamma^{-1}\underbrace{\frac{1}{N}\sum_{k=1}^{N}(x(k)x'(k) + \Lambda'\Lambda)}_{\Gamma}\theta_\mathrm{o} - \frac{\Gamma^{-1}}{N}\sum_{k=1}^{N}\Lambda'\Lambda\theta_\mathrm{o} \\ &= \theta_\mathrm{o} - \Gamma^{-1}\sum_{j=1}^{n_\theta}\lambda_j^2\theta_{\mathrm{o},j} = \theta_\mathrm{o},\end{aligned} \tag{19}$$

where the last equation follows from the choice of the hyperparameters in (16).

As for (b) in (18), we have the following asymptotic behaviour:

$$\begin{aligned}&\lim_{N\to\infty}\Gamma^{-1}\frac{1}{N}\sum_{k=1}^{N} x(k)e(k) = \Gamma_\star^{-1}\lim_{N\to\infty}\frac{1}{N}\sum_{k=1}^{N} x(k)e(k) = \\ &= \Gamma_\star^{-1}\mathbb{E}[x(k)e(k)], \quad \text{w.p. 1.}\end{aligned} \tag{20}$$

Since $x(k)$ and $e(k)$ are statistically independent, $\mathbb{E}[e(k)] = 0$ (from Assumptions (4)), and $\Gamma_\star^{-1}$ exists by assumption in Proposition 1, then $\Gamma_\star^{-1}\mathbb{E}[x(k)e(k)] = 0$. Summarizing, taking the limit for $N \to \infty$ in (18), and considering (19) and (20), (17) follows. ■

Proposition 2. [Consistency for OE models: training]

Let us assume that:

- the diagonal elements of the hyperparameter matrix $\Lambda \in \mathbb{R}^{n_\theta \times n_\theta}$ are given and have the structure in (16);
- the following matrices exist and are invertible:

$$\Gamma_{IV} = \frac{1}{N}\sum_{k=1}^{N} x(k)z'(k)\frac{1}{N}\sum_{k=1}^{N} z(k)x'(k) + \Lambda'\Lambda$$

$$\Gamma_{\star,IV} = \lim_{N\to\infty}\frac{1}{N}\sum_{k=1}^{N} x(k)z'(k)\frac{1}{N}\sum_{k=1}^{N} z(k)x'(k) + \Lambda'\Lambda.$$

Then, the estimate $\hat{\theta}_{IV}(\lambda)$ in (11) is a consistent estimate of the true parameter vector θ_o, i.e.,

$$\lim_{N\to\infty}\hat{\theta}_{IV}(\lambda) = \theta_\mathrm{o} \quad \text{w.p. 1.} \tag{21}$$

Proof:

Let us rewrite the observed output $y(k)$ in (1) as follows:

$$y(k) = x_\mathrm{o}'(k)\theta_\mathrm{o} + e(k) = x'(k)\theta_\mathrm{o} + v(k), \tag{22}$$

where $v(k)$ is defined as:

$$v(k) = -[e(k-1),\ \dots,\ e(k-n_\mathrm{a}),\ 0,\ \dots,\ 0]\theta_\mathrm{o} + e(k). \tag{23}$$

Note that $v(k)$ is a zero-mean noise, correlated with $x(k)$.

By substituting the expression of the output $y(k)$ in (22) into the IV solution (11), we obtain:

$$\hat{\theta}_{IV}(\lambda) = \Gamma_{IV}^{-1}\frac{1}{N}\sum_{k=1}^{N} x(k)z'(k)\frac{1}{N}\sum_{k=1}^{N} z(k)(x'(k)\theta_\mathrm{o} + v(k)). \tag{24}$$

Following the same algebraic manipulations already adopted in the proof of Proposition 1, we have:

$$\hat{\theta}_{\mathrm{IV}}(\lambda) = \theta_{\mathrm{o}} + \Gamma_{\mathrm{IV}}^{-1} \frac{1}{N} \sum_{k=1}^{N} x(k) z'(k) \frac{1}{N} \sum_{k=1}^{N} z(k) v(k). \quad (25)$$

Taking the limit as $N \to \infty$, the last term in (25) is:

$$\lim_{N\to\infty} \Gamma_{\mathrm{IV}}^{-1} \frac{1}{N} \sum_{k=1}^{N} x(k) z'(k) \frac{1}{N} \sum_{k=1}^{N} z(k) v(k)$$
$$= \Gamma_{\star,IV}^{-1} \Xi_{\star} \mathbb{E}\left[z(k) v(k)\right] = 0 \ \text{ w.p. 1}, \quad (26)$$

where the last equation follows since the instrumental variable $z(k)$ is not correlated with the measurement noise $e(k)$ (and thus not correlated with the $v(k)$), and $v(k)$ is zero mean. Summarizing, by combing (26) and (25), the consistency property in (21) follows. ■

3.2 Consistency in validation

In this section we discuss the effect of selecting regularization hyperparameters λ by minimizing the evaluation criterion $E(\lambda)$ in (12). We demonstrate that, for both the ARX and OE model structures, the optimal hyperparameters $\lambda^\star$ (see (15)) that minimize $E(\lambda)$ ensure that $\hat{\theta}_{\mathrm{LS}}(\lambda^\star)$ and $\hat{\theta}_{IV}(\lambda^\star)$ are consistent estimates of the true parameter vector θ_{o}, as discussed in the following Propositions.

Proposition 3. [Consistency for ARX models: validation]

Let $\lambda^\star$ be the solution of (15), with $L_q\left(\hat{\theta}^{(\backslash q)}(\lambda), \mathcal{D}_q\right) = L_{q,\mathrm{LS}}\left(\hat{\theta}^{(\backslash q)}(\lambda), \mathcal{D}_q\right)$ according to (14a). Let us assume that the cost $L_{q,\mathrm{LS}}(\theta, \mathcal{D}_q)$ admits a unique minimizer. Then, $\hat{\theta}_{\mathrm{LS}}(\lambda^\star)$ is a consistent estimate of θ_{o}, i.e.,

$$\lim_{N\to\infty} \hat{\theta}_{\mathrm{LS}}(\lambda^\star) = \theta_{\mathrm{o}} \ \text{ w.p. 1}. \quad (27)$$

Proof: The minimizer of the quadratic loss $L_{q,\mathrm{LS}}(\theta, \mathcal{D}_q)$ in (14a) is unique (by assumption) and it is given by the least-squares solution, which converges (w.p. 1) to the true parameter vector θ_{o}, as N (or equivalently N_{V}) goes towards infinity, i.e.,

$$\lim_{N\to\infty} \arg\min_{\theta} L_{q,\mathrm{LS}}(\theta, \mathcal{D}_q) = \theta_{\mathrm{o}} \text{ w.p. 1}, \quad \forall\, q = 1, \ldots, Q.$$

Then, by continuity of $L_{q,\mathrm{LS}}(\theta, \mathcal{D}_q)$, for all $q = 1, \ldots, Q$:

$$\lim_{N\to\infty} \min_{\theta} L_{q,\mathrm{LS}}(\theta, \mathcal{D}_q) = L_{q,\mathrm{LS}}(\theta_{\mathrm{o}}, \mathcal{D}_q) \ \text{ w.p. 1}. \quad (28)$$

Thus, from the definition of the evaluation criterion $E(\lambda)$ in (12), and (28), we have:

$$\sum_{q=1}^{Q} L_{q,\mathrm{LS}}(\theta_{\mathrm{o}}, \mathcal{D}_q) \leq \lim_{N\to\infty} E(\lambda) \text{ w.p. 1 } \forall\ \lambda. \quad (29)$$

From Proposition 1, there exists a value $\tilde{\lambda}$ such that:

$$\lim_{N\to\infty} \hat{\theta}_{\mathrm{LS}}(\tilde{\lambda}) = \theta_{\mathrm{o}} \ \text{ w.p. 1}. \quad (30)$$

This value is also a minimizer of $E(\lambda)$. In fact, by continuity of $L_{q,\mathrm{LS}}(\theta, \mathcal{D}_q)$ and from (30), we have:

$$\sum_{q=1}^{Q} L_{q,\mathrm{LS}}(\theta_{\mathrm{o}}, \mathcal{D}_q) = \lim_{N\to\infty} \underbrace{\sum_{q=1}^{Q} L_{q,\mathrm{LS}}\left(\hat{\theta}_{\mathrm{LS}}(\tilde{\lambda}), \mathcal{D}_q\right)}_{E(\tilde{\lambda})} \text{ w.p. 1},$$

which is, according to (29), the minimum of $E(\lambda)$.

It remains to be proved that any other minimizer $\lambda^\star$ of $E(\lambda)$ is such that $\hat{\theta}_{LS}(\lambda^\star) \to \theta_{\mathrm{o}}$, as $N \to \infty$. This follows since, by assumption, θ_{o} is the unique minimizer of $L_{q,\mathrm{LS}}(\theta, \mathcal{D}_q)$ for all q; thus (31) is satisfied if and only if:

$$\lim_{N\to\infty} \hat{\theta}_{\mathrm{LS}}(\lambda^\star) = \theta_{\mathrm{o}} \ \text{ w.p.1}. \quad (31)$$

■

Proposition 4. [Consistency for OE models: validation]

Let $\lambda^\star$ be the solution of (15), with $L_q\left(\hat{\theta}^{(\backslash q)}(\lambda), \mathcal{D}_q\right) = L_{q,IV}\left(\hat{\theta}^{(\backslash q)}(\lambda), \mathcal{D}_q\right)$ according to (14b). Let us assume that the cost $L_{q,IV}(\theta, \mathcal{D}_q)$ admits a unique minimizer. Then, $\hat{\theta}_{IV}(\lambda^\star)$ is a consistent estimate of θ_{o}, i.e.,

$$\lim_{N\to\infty} \hat{\theta}_{IV}(\lambda^\star) = \theta_{\mathrm{o}} \ \text{ w.p. 1}. \quad (32)$$

The proof of this proposition is based on arguments similar to those used in proving Proposition 3. It relies on the property the minimizer of the loss $L_{q,IV}(\theta, \mathcal{D}_q)$ in (14b) is unique and it is given by the IV solution $\hat{\theta}_{IV}(\lambda)$, and:

$$\lim_{N\to\infty} \arg\min_{\theta} L_{q,IV}(\theta, \mathcal{D}_q) = \theta_{\mathrm{o}} \text{ w.p. 1}, \quad \forall\ q = 1, \ldots, Q.$$

4. GRADIENT-BASED MULTI-HYPERPARAMETER OPTIMIZATION

In this section we provide the gradient, which is essential to compute the optimal solution $\lambda^\star$ of problem (15) through any gradient-based numerical optimization algorithm.

4.1 ARX model structure

Since in the ARX case the model parameters $\hat{\theta}^{(\backslash q)}(\lambda)$ are obtained through the least-squares solution, the formula for the gradient $\nabla_\lambda E$ coincides with the one derived in Maroni et al. (2023a) for (static) linear-in-the-parameter models, and computed leveraging matrix differential calculus. This formula is reported below for self-consistency of the paper.

Let $X_{\backslash q} \in \mathbb{R}^{N_T \times n_\theta}$ and $Y_{\backslash q} \in \mathbb{R}^{N_T}$ be the matrices stacking in their rows, respectively, the regressor $x'(k)$ and the output $y(k)$, with $\{(x(k), y(k))\} \in \bigcup_{j\neq q} \mathcal{D}_j$. Similarly, let $X_q \in \mathbb{R}^{N_V \times n_\theta}$ and $Y_q \in \mathbb{R}^{N_V}$ be matrices stacking in their rows, respectively, the regressor $x'(k)$ and the output sample $y(k)$, with $(x(k), y(k)) \in \mathcal{D}_q$. Then, the estimate $\hat{\theta}^{(\backslash q)}(\lambda)$ (see (13)) can be computed analytically and written in the compact matrix form:

$$\hat{\theta}_{LS}^{(\backslash q)}(\lambda) = \left(X'_{\backslash q} X_{\backslash q} + N_T \Lambda' \Lambda\right)^{-1} X'_{\backslash q} Y_{\backslash k}. \quad (33)$$

The gradient of the evaluation criterion E with respect to the regularization hyperparameters λ can be computed analytically and is given by:

$$\nabla_\lambda E = -\frac{N_T}{Q N_V} \sum_{q=1}^{Q} \mathrm{diag}\left(\Lambda B_q + B_q \Lambda\right), \quad (34)$$

where $B_q = A'_q X'_q R_q \hat{\theta}^{(\backslash q)'}$, $A_q = \left(X'_{\backslash q} X_{\backslash q} + N_T \Lambda' \Lambda\right)^{-1}$, $R_q = X_q \hat{\theta}^{(\backslash q)}(\lambda) - Y_q$, and $\mathrm{diag}(M)$ is the column vector composed by the diagonal of a generic square matrix M.

4.2 OE model structure

Let $Z_{\setminus q} \in \mathbb{R}^{N_T \times n_\theta}$ be the matrix of instrumental variables defined similarly to the matrix of regressors $X_{\setminus q}$. Namely, $Z_{\setminus q}$ stacks in its rows, respectively, the instrumental variable vector $z'(k)$ such that k satisfies $\{(x(k), y(k))\} \in \bigcup_{j \neq q} \mathcal{D}_j$. Similarly, let $Z_q \in \mathbb{R}^{N_V \times n_\theta}$ stacks in its rows, the vector $z'(k)$ with k satisfying $(x(k), y(k)) \in \mathcal{D}_q$.

As for the OE model structure, the expression of the model parameters $\hat{\theta}^{(\setminus q)}(\lambda)$ (see (13)) can be computed analytically and written in the compact matrix form:

$$\hat{\theta}^{(\setminus q)}_{IV}(\lambda) = \left(X'_{\setminus q} Z_{\setminus q} Z'_{\setminus q} X_{\setminus q} + N_T^2 \Lambda' \Lambda \right)^{-1} X'_{\setminus q} Z_{\setminus q} Z'_{\setminus q} Y_{\setminus k}. \tag{35}$$

The gradient $\nabla_\lambda E$ can be computed analytically and is given by: $\nabla_\lambda E = -(N_T^2/QN_V^2) \sum_{q=1}^{Q} \operatorname{diag}(\Lambda B_q + B_q \Lambda)$, with $B_q = A'_q X'_q Z_q R_q \hat{\theta}^{(\setminus q)'}$, $A_q = \left(X'_{\setminus q} Z_{\setminus q} Z'_{\setminus q} X_{\setminus q} + N_T^2 \Lambda' \Lambda \right)^{-1}$ and $R_q = Z'_q \left(X_q \hat{\theta}^{(\setminus q)}(\lambda) - Y_q \right)$. Details on the derivation of $\nabla_\lambda E$ are omitted in this paper due to space constraints, but it can be obtained by the same arguments presented in Maroni et al. (2023a). Intuitively, by considering the structures of the least-square and IV solutions in (33) and (35) respectively, it is evident that the gradients $\nabla_\lambda E$ for both the ARX and OE cases have a similar structure. This similarity is achieved by replacing $X_{\setminus q}$ with $Z'_{\setminus q} X_{\setminus q}$, and $\frac{N_T}{N_V}$ with $\frac{N_T^2}{N_V^2}$ in (34).

5. NUMERICAL EXAMPLES

In this section we present two numerical examples to demonstrate the effectiveness of the approach. The first example, proposed for demonstrative purposes, focuses on the identification of a chaotic Lorenz system. The goal of this example is to show how the proposed algorithm is able to promote sparsity in the estimated model parameter vector, without introducing bias in the non-zero model parameters. Although the methodology was discussed for SISO and ARX systems, the considered system is MIMO, with regressors containing nonlinear terms. The second example aims to demonstrate the effectiveness of the proposed IV-based regularized approach, along with a comparison with standard regularization strategies.
To allow reproducibility of the results, the code is available in the GitHub repository Maroni et al. (2023b).

5.1 Example 1: Lorenz system

As a data-generating system, let us consider the continuous-time Lorenz system taken from Brunton et al. (2016):

$$\dot{x} = \sigma(y - x), \quad \dot{y} = x(\rho - z) - y, \quad \dot{z} = xy - \beta z, \tag{36}$$

with $\sigma = 10$, $\rho = 28$, and $\beta = \frac{8}{3}$. Data samples are generated by integrating (36) from $t = 0$ to $t = 10$, with a time step of $\Delta t = 0.002$, and initial condition $[x(0), \ y(0), \ z(0)]' = [-8, \ 7, \ 27]'$. To demonstrate the sparsifying effect of the proposed multi-ridge approach, we consider an overparameterized model with regressor containing all the polynomial terms of the (measured) system states x, y, z up to the fifth order, for a total of $n_\theta = 55$ elements. The Λ regularization matrix is initialized as an identity matrix and optimized over 500 epochs through Gradient Descent, with $Q = 3$ cross-validation folds.

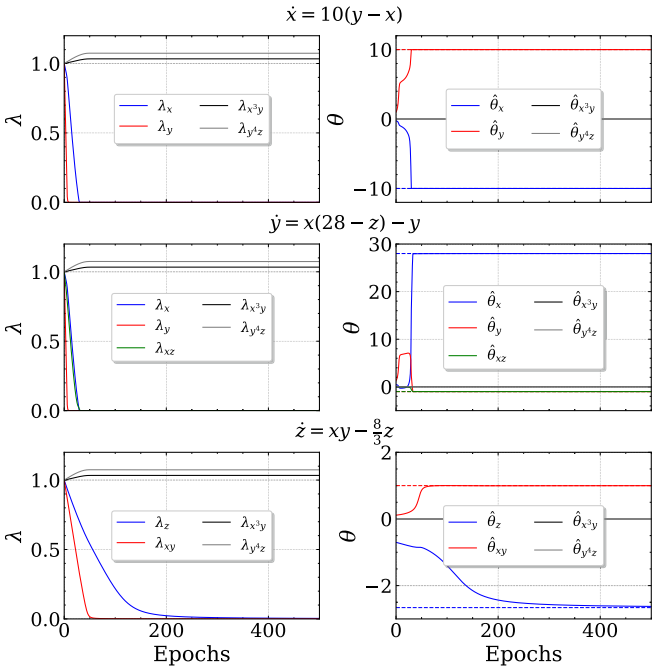

Fig. 1. Example 1: The three rows display the results related to the three state equations (36). Right panels: estimated coefficients *vs.* number of epochs. True model parameters are plotted with dashed lines. Left panels: optimized regularization hyperparameters associated with the model coefficients in the right plots. For better visualization, only non-zero model coefficients and a subset of zero coefficients are shown.

Results are shown in Figure 1. In the top row, corresponding to the first state equation in (36), the estimated parameters are visualized with increasing epochs (right panel). The following model coefficients are plotted: $\hat{\theta}_x$; $\hat{\theta}_y$; $\hat{\theta}_{x^3y}$; and $\hat{\theta}_{y^4z}$. These coefficients are associated with the regressor elements x; y; x^3y; and y^4z. Note that according to (36), only the variables x and y characterize the state equation. The left panel shows the optimized values of the regularization hyperparameters $(\lambda_x, \lambda_y, \lambda_{x^3y}, \lambda_{y^4z})$, associated with the aforementioned model coefficients. It can be observed that the regularization hyperparameters λ_x, λ_y are driven towards zero, and consequently, the estimated coefficients are free to converge to their corresponding true values. On the contrary, the values of the hyperparameters λ_{x^3y} and λ_{y^4z} increase, thus shrinking the associated coefficients θ_{x^3y} and θ_{y^4z} towards zero. Similar considerations apply to the remaining two rows of Figure 1.

5.2 Example 2: OE model structures

In this numerical example, we compare different regularization strategies (multi-ridge *vs* benchmark algorithms: Ridge, LASSO and Elastic Net) in different SNR conditions. For a fair comparison, the same IV-based fitting loss $L(\theta, \mathcal{D}) = L_{IV}(\theta, \mathcal{D})$ in (9) is used by all the algorithms, and the same instrumental variables $z(k)$ are used. Like in Laurain et al. (2015)), the variables $z(k)$ are chosen as:

$z(k) = [\hat{y}_{LS}(k-1), \ldots, \hat{y}_{LS}(k-n_a), u(k-1), \ldots, u(k-n_b)]'$, where $\hat{y}_{LS}$ represents the simulated output of a model with parameters estimated through a least-squares algorithm.

As a data-generating system, we consider:

$$y_o(k) = -a_1 y_o(k-1) - a_2 y_o(k-2) + b_1 u(k-1) + b_2 u(k-2),$$

with $a_1 = -0.2$, $a_2 = -0.3$, $b_1 = 0.1$, $b_2 = -1.2$. The input signal u is a white noise following a zero-mean unitary-variance Gaussian distribution. The measured output $y(k) = y_o(k) + e(k)$ is corrupted by a white noise $e(k)$ generated by a zero-mean Gaussian distribution. In this study, 10 different values of noise variance σ^2 are chosen, thus leading to 10 different SNRs conditions, where the SNR is defined as: $\text{SNR} = 10 \log \left(\frac{\|y_o\|^2}{\sigma^2} \right)$. To increase the statistical significance of the results, 1000 Monte Carlo scenarios are generated. At each run, $N = 100$ training input/samples and 1000 test input/samples are collected. The output noise is only added to the training data and not to the test data. The dynamical order of the model is assumed to be unknown and is bounded by 10 for both the autoregressive and exogenous parts, i.e., $n_a = n_b = 10$.

The baseline algorithms are subjected to a training and evaluation pipeline characterized by the following steps: i) hyperparameter optimization through an exhaustive grid-search with a 5-fold cross-validation scheme, ii) final re-training on the complete training dataset, iii) performance evaluation on the test dataset. For multi-ridge regression, the regularization hyperparameter matrix Λ is initialized as an identity matrix and then optimized with 1000 Gradient Descent epochs. Implementation details are provided in the provided scripts Maroni et al. (2023b).

The performance of the estimated models is evaluated through open-loop simulation on a test sequence (used neither for training nor for validation), and measured through the R^2 index. Figure 2 shows the obtained results. For each regularization algorithm and each SNR condition, the median value obtained on the 1000 Monte Carlo runs is shown. Multi-ridge regression outperforms the benchmark regularization strategies under all SNR conditions. Ridge regression shows the worst performance: this result is expected and due to the mismatch between the real order of the system (i.e., 2) and the order assumed in the experiments (i.e., 10), and Ridge's inability to select the

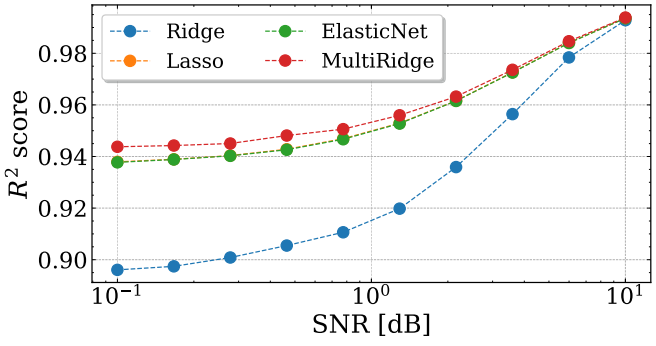

Fig. 2. Example 2: Comparison of regularization algorithms for different SNR conditions. Median values (dots) over 1000 runs are reported. Lines associated with Lasso and Elastic Net are overlapped.

dynamical order through ℓ_2 penalization. On the contrary, algorithms that promote a sparse solution such as LASSO make a better selection of the order of the system through ℓ_1 penalization. Even if using ℓ_2 regularization, multi-ridge regression still surpasses LASSO in performance. This is mainly due to the multi-ridge capabilities to associate a unique regularization penalty to each model parameter, and thus not introducing bias in the non-zero coefficients.

6. CONCLUSIONS

The proposed multi-ridge regression method enables using distinct regularization hyperparameters for each model coefficient. For accurate estimation of true system parameters in output-error model structures, it is essential to use an instrumental-variable-based fitting loss. While consistency is also maintained for nonlinear ARX structures, output error structures require output linearity for the instrumental-variable method's consistency. Future activities focus on extending the approach to nonlinear systems with an output-error noise model. This involves adopting Prediction Error Methods or bias-correction schemes to ensure a consistent parameter estimate.

REFERENCES

Bengio, Y. (2000). Gradient-based optimization of hyperparameters. *Neural computation*, 12(8), 1889–1900.

Brunton, S.L., Proctor, J.L., and Kutz, J.N. (2016). Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the national academy of sciences*, 113(15), 3932–3937.

Hoerl, A.E. and Kennard, R.W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55–67.

Laurain, V., Tóth, R., Piga, D., and Zheng, W.X. (2015). An instrumental least squares support vector machine for nonlinear system identification. *Automatica*, 54, 340–347.

Maroni, G., Cannelli, L., and Piga, D. (2023a). Gradient-based bilevel optimization for multi-penalty Ridge regression through matrix differential calculus. *arXiv preprint arXiv:2311.14182*. Submitted to Automatica.

Maroni, G., Cannelli, L., and Piga, D. (2023b). Multiridge-SYSID: Codebase for multi-ridge identification. `https://github.com/gabribg88/Multiridge-SYSID`.

Ohlsson, H., Ljung, L., and Boyd, S. (2010). Segmentation of ARX-models using sum-of-norms regularization. *Automatica*, 46(6), 1107–1111.

Piga, D. and Tóth, R. (2013a). LPV model order selection in an LS-SVM setting. In *52nd IEEE Conference on Decision and Control*, 4128–4133.

Piga, D. and Tóth, R. (2013b). An SDP approach for ℓ_0-minimization: Application to ARX model segmentation. *Automatica*, 49(12), 3646–3653.

Pillonetto, G., Chen, T., Chiuso, A., De Nicolao, G., and Ljung, L. (2022). *Regularized system identification: Learning dynamic models from data*. Springer Nature.

Söderström, T. and Stoica, P. (2002). Instrumental variable methods for system identification. *Circuits, Systems and Signal Processing*, 21(1), 1–9.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267–288.