Regularization when modeling with biased simulation data as a prior

Håvard B. Bjørkøy * **Hans A. Engmark** * **Adil Rasheed** *
Damiano Varagnolo *,**

* *Norwegian University of Science and Technology, Department of Engineering Cybernetics, Trondheim, Norway (e-mail: havard.bjorkoy@ntnu.no)*
** *University of Padova, Department of Information Engineering, Italy*

Abstract: Embedding physical knowledge in system identification increases the generalization capabilities of the identified models. For complex engineering systems, such as a process plant, the most complete and detailed quantitative description of the existing physical and structural knowledge is often provided by a simulator. We describe the procedure of fusing simulated data with measurement data via L2 regularization for models that are linear in the parameters. We characterize how the MSE minimization problem in this framework is nontrivial, and show that for certain realizations of the data there is no unique minimum of the MSE w.r.t. the regularization parameter. In these cases the MSE can even increase to larger values than both the variance and the bias, which is counter-intuitive. We show how this issue appears less frequently with more data, even though multiple minima can occur for any realization of the data. However, we show also that the Stein effect is present regardless, so that it is always possible to decrease the MSE with careful selection of the regularization parameter, i.e., information fusion may always be beneficial.

Keywords:
Regularized system identification, Data fusion, Gray-box modeling

1. INTRODUCTION

We assume to have noisy measurement data from a real-world process that we want to model as accurately as possible, as well as having domain knowledge about the process that can be used to assist in model estimation. This estimated model will be referred to as "the model", having parameters $\boldsymbol{\beta}$. The domain knowledge will here be simulation data that contains bias due to modeling errors, which poses some challenges when used for regularization. The simulated data implies a prior distribution on $\boldsymbol{\beta}$. Even though the simulator is technically also a model, it will be referred to as "the simulator" to minimize confusion.

In many data-driven modeling problems there exists structural information about the process that is being modeled. A common problem is how to embed physics-based priors in system identification tasks, such as knowing that the exponential decay of the impulse response should be within certain bounds. The question is thus how to utilize this structural information to obtain models that have the highest generalization capabilities possible. Some existing methodologies include gray-box modeling principles such as hybrid modeling, constrained black-box modeling, or semi-physical modeling (Glassey and Von Stosch, 2018; Sohlberg and Jacobsen, 2008). For example, in regularized system identification, kernels can be designed to reflect prior knowledge on the physical system (Pillonetto et al., 2022). In some of these approaches, the physical information is hard coded (e.g., the structure of the semi-physical model), and this constraint may reduce the flexibility of the subsequent identification steps, possibly worsening underfitting phenomena. In other words, prior knowledge about the physics of the system is in most cases uncertain to some degree and should intuitively be treated that way. Many gray-box modeling techniques do not give this flexibility which may inhibit discovering certain features of the process.

We also note how such physical knowledge may be represented by first principles embedded in a simulator (e.g., finite element (FE) or computational fluid dynamics (CFD) methods). The data obtained through this simulator can be seen as an instance of a physical knowledge prior, and is in a sense a prior itself. The idea that *"data from a simulator may be seen as a prior on* $\boldsymbol{\beta}$*"* is not new (Kedem et al., 2017; Saadallah et al., 2022). This interpretation introduces the following information theoretic subtlety: *embedding simulation data may be seen as a data fusion problem.* When fusing data, the different information sources should be weighted according to their trustworthiness, with the weight inversely proportional to the uncertainty. FE and CFD simulators are however typically deterministic (at least for fixed parameters), and their uncertainty lies in their bias. Both the simulator and the model will have modeling errors. The difference between these modeling errors is the bias of the simulator with respect to the parameters $\boldsymbol{\beta}$. In other words, this bias is what constitutes the uncertainty of the simulator when it is used as a prior for $\boldsymbol{\beta}$. Thus, given that the simulation bias is not zero, there is the need to establish a statistically rigorous link between this type of uncertainty and the weight given to the simulation data.

In this article, we address this problem, specifically analyzing the L2 regularized least-squares (ReLS) estimator of the parameters of data-driven models when prior information is biased simulation data. We derive analytical results for models that are linear in the parameters. Our contributions show that even within this relatively restricted model structure set, situations may arise where the mean squared error (MSE) of the ReLS estimator has several local minima w.r.t. the regularization parameter, which contradicts general intuitions about the bias-variance trade-off. For a scalar parameter model one may characterize the minimum of the MSE in a closed form, while for the case of vectorial parameter models the problem is significantly more complex and may not have analytical solutions. Although motivated by regularized system identification with simulation data, the results apply to any data-fusion problem using this ReLS estimator where bias is present.

Regarding how this paper relates to existing literature, we note that regularization is commonly used to reduce overfitting to the training set of a model, and can be understood by the bias-variance trade-off (Pillonetto et al., 2022). The idea is to reduce the variance of the model by introducing some bias, which in total reduces MSE. The relationship between bias and variance of an estimator is typically depicted as in Fig. 1. The standard way of selecting the regularization parameter λ is by cross-validation, a trial-and-error grid search technique, or through Stein Unbiased Risk Estimators. In this paper we focus on understanding how to optimally choose λ for fusing data from a field plant with data from a simulator. However, selecting λ may be non-trivial, since situations as in Fig. 2 may occur, as we describe. The presence of multiple local minima makes selecting the optimal regularization level λ logically more complex than in the standard situation depicted in Fig. 1.

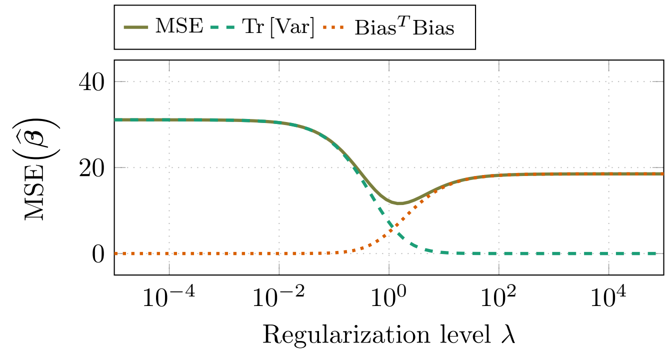

Fig. 1. A typical depiction of the bias-variance trade-off, and the effect it has on MSE. The bias grows as the regularization parameter is increased, while for $\lambda = 0$ the MSE equals the variance of the estimator.

The field of data fusion approaches this problem quite generally. For example, Saadallah et al. (2022) describe different paradigms of fusion, and show conditions for when model-level fusion reduces variance, though they note that bias may eliminate this benefit. It is clear that simulation data is often given a role identical to measured data; for instance Saadallah et al. (2022) give examples where simulation data is even used as the ground truth. However, for complex industrial processes (as an example, submerged arc furnaces (Sparta et al., 2021)), this is not a viable assumption, since FE and CFD models may be too idealistic (which suggests that such simulators provide biased information).

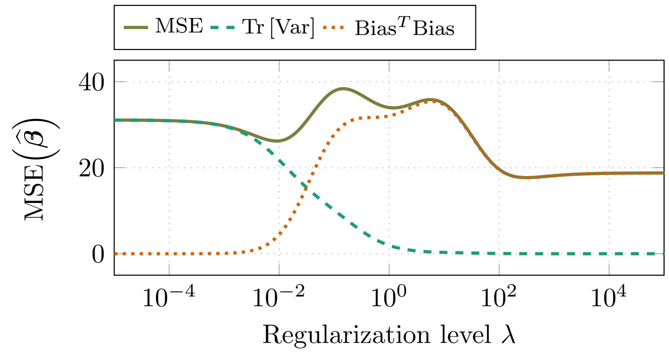

Fig. 2. A depiction of how the MSE might change with increasing regularization for certain data realizations. The bias term may not grow monotonically with the regularization parameter, causing multiple local minima (and, potentially, multiple global minima).

The book by Kedem et al. (2017) also discusses the problem of fusing simulation data with field data, via the concept of *out of sample fusion*. Although this relates to the data fusion field, it has a greater emphasis on statistical theory. Out of sample fusion indeed relies heavily on the validity of a density ratio model, and does not explicitly treat the simulator as a prior (although arguably implicitly), nor addresses the potential bias in simulators.

Regularization applied to system identification problems is well-covered in a recent book (Pillonetto et al., 2022). This book treats regularization more with the purpose of estimating sparse models from flexible model structures that avoid overfitting, particularly kernel-based, and embedding physics through kernel design.

Regularization using prior knowledge on the steady states of a dynamic system using kernel methods, FIR models and subspace methods are addressed in Fujimoto and Sugie (2018); Khosravi and Smith (2021) and Yoshimura et al. (2019), respectively. The former assumes a Gaussian prior on the steady states, and is quite comparable to this work, though our focus is more on how to embed the nature of the uncertainty *within* the problem of choosing the regularization levels. The latter two assume the prior information to be exact, and discuss the problem more as a model-level fusion problem.

Regularizing estimators through embedding prior physical knowledge is the concept behind physics-informed neural networks (Raissi et al., 2019). These methods can be shown to converge to the PDEs that constitute the prior knowledge (Shin et al., 2020). The regularization level is adjusted by varying the number of collocation points where the PDEs are enforced. However, if this prior is inaccurate, this means converging to something biased. In the data-from-different-sources fusion framework we consider here, convergence to the prior is not the goal.

Finally, we note that the developments in this work are useful towards including prior information in models that are linear in the parameters, such as in the SINDy framework (Brunton et al., 2016), or in ARMAX models (Ljung, 1998). This work is a first step towards an information theoretic framework that considers the prior to be simulation data, with the assumption that it is biased (even if the bias is unknown).

2. A NAÏVE APPROACH FOR TREATING SIMULATION BIAS

As mentioned above, the uncertainty of deterministic simulators may be assumed to reside in their bias, and the relative importance of the simulation data when fusing it with field data should be connected to the uncertainty of the simulator (thus to its bias). Generally, both the simulator and the model to be identified will contain modeling errors that are unequal. Since the simulated data is used to estimate parameters $\boldsymbol{\beta}$, the bias we consider is the bias of the simulator with respect to the model to be estimated, thus it is depending on the model structure.

Let $\mathcal{D}_m$ denote data measured from the plant, and $\mathcal{D}_s$ data from the simulator. When $\mathcal{D}_s$ is biased, one may think to: 1) estimate its bias (say $\boldsymbol{\varepsilon}$) by somehow comparing $\mathcal{D}_s$ to $\mathcal{D}_m$, 2) subtract this inferred bias $\boldsymbol{\varepsilon}$ from the data $\mathcal{D}_s$, and 3) treat the detrended data $\mathcal{D}_s - \boldsymbol{\varepsilon}$ as if it was unbiased. There are several reasons for why this may be statistically suboptimal. In fact, we have yet to find rigorous statistical reasons for why this should *add* any information about the process, meaning that it should only be performed if it is convenient for subsequent estimation steps. Furthermore, treating this approach in a statistically rigorous way seems non-trivial: subtracting an estimated bias that correlates with *both* data sources requires careful handling of the relative uncertainties.

Another idea may be to estimate the *variance* of the simulator - e.g., varying the inputs and parameters that generate the simulations, to analyze sensitivity, and thus find a relation between simulation parameter uncertainty and the simulation output uncertainty. Such tests will never reveal uncertainties originating from modeling errors, and may thus give a wrong view of the uncertainty. In other words, making two step approaches, such as those mentioned, statistically rigorous and efficient seems to require systematical execution of a large number of simulations, which may be unfeasible.

It therefore seems to be the need for frameworks where the simulations are taken as they are and treated as an opportune form of prior information.

3. PROBLEM FORMULATION

Our goal is to formulate a framework that enables addressing the simulation bias discussed above in some statistically optimal way. Towards this we assume a scenario where there exists two data sets: $\mathcal{D}_m$ (measurement data) of size n_m, and $\mathcal{D}_s$ (simulated data) of size n_s. Both data sources shall be used to estimate $\boldsymbol{\beta}$. For the sake of deriving analytical results, we consider only models that are linear in the parameters, and without loss of generality (but for the sake of notational clarity) with scalar outputs. The model is assumed to coincide with the measurements, only disturbed by additive noise, meaning that the modeling error is zero. To assist explaining the general formulation below, the following example is given:

Example: Model and metamodel of a separable FIR Let the measurement data $\mathcal{D}_m$ correspond to a number of time-series measurements from a scalar, nonlinear finite impulse response (FIR) model, linear in the parameters, with zero-mean i.i.d. measurement noise w, thus of the kind

$$y_t = \sum_{i=1}^{p} \beta_i f_i(x_{t-i}) + w = \boldsymbol{f}_t^T \boldsymbol{\beta} + w \tag{1}$$

where $x_t, y_t \in \mathbb{R}$ are the input and output signals, the f_i's are generic nonlinear functions, the β_i's are the parameters to be estimated, and $\boldsymbol{f}_t$ is the vector obtainable by stacking the various $f_i(x_{t-i})$'s. Let then X_m be the matrix obtained through stacking the row vectors $\boldsymbol{f}_t^T$, and $\boldsymbol{y}_m$ be the column vector obtained through stacking the outputs y_t. This means letting the measurement data be

$$\boldsymbol{y}_m = X_m \boldsymbol{\beta} + \boldsymbol{w}, \tag{2}$$

where $\boldsymbol{w}$ is a vector whose components are equal to w. In the following the measurement data is indicated by $\mathcal{D}_m = \{\boldsymbol{y}_m, X_m\}$.

As for the simulated data $\mathcal{D}_s$, suppose that there exists a simulator modeling steady states of (1), i.e., assume to be able to simulate, for any generic input x_s that is constant in time, the corresponding steady state outputs y_s. The FIR model implies that the simulation data obeys

$$y_s = \sum_{i=1}^{p} \beta_i f_i(x_s) + \varepsilon, \tag{3}$$

which is thus a metamodel (model of a model) of the steady states of model (1) whose parameters β_i are correct, but whose additive disturbance $\boldsymbol{\varepsilon}$ is the unknown simulation bias. Given a set of steady states $\boldsymbol{x_s} = \{x_{s,1}, \ldots, x_{s,n_s}\}$, applying (3) to each $x_{s,i}$ leads to

$$\boldsymbol{y}_s = X_s \boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{4}$$

where $\boldsymbol{y_s} \in \mathbb{R}^{n_s}$, $X_s \in \mathbb{R}^{n_s \times p}$, the parameters β_i are as in (1), and the term $\boldsymbol{\varepsilon} \in \mathbb{R}^{n_s}$ is the simulation bias. The simulated data $\mathcal{D}_s = \{\boldsymbol{y}_s, X_s\}$ may be seen as uncertain information about the equilibria of (1). In other words, the simulations provide implicit information on $\boldsymbol{\beta}$, despite the bias.

3.1 The model and the simulator metamodel

Based on the assumptions above, the model to be estimated is generally expressed as

$$y(t) = \boldsymbol{f(x(t))^T \beta}\ , \tag{5}$$

where $y \in \mathbb{R}, \boldsymbol{x} \in \mathbb{R}^m, \boldsymbol{\beta} \in \mathbb{R}^p$ and $f : \mathbb{R}^m \to \mathbb{R}^p$. As in the example above, we form the regressor of the model $\boldsymbol{f(x_i)}$ for all n_m measurements in $\mathcal{D}_m$, denoted X_m, and the dependent variables are denoted $\boldsymbol{y_m}$, thus

$$\boldsymbol{y_m} = X_m \boldsymbol{\beta} + \boldsymbol{w}\ , \tag{6}$$

where $\boldsymbol{w}$ is zero-mean noise, and $\boldsymbol{y_m} \in \mathbb{R}^{n_m}, X_m \in \mathbb{R}^{n_m \times p}$. For the analysis later on to hold, we assume that X_m is full rank, making $X_m^T X_m$ positive definite (PD).

The metamodel of the simulator is implied by the model structure, and it is needed in order to fuse the information of the two data sets $\mathcal{D}_m, \mathcal{D}_s$, since it defines the connection between $\boldsymbol{\beta}$ and the simulations. Deriving the metamodel may generally be non-trivial for certain model structures and simulators, but is not the focus of this work. The metamodel is here denoted with

$$y_s = \boldsymbol{h}(\boldsymbol{x_s})^T \boldsymbol{\beta} \ , \tag{7}$$

where $y_s \in \mathbb{R}, \boldsymbol{x_s} \in \mathbb{R}^k, \boldsymbol{\beta} \in \mathbb{R}^p$ and $\boldsymbol{h} : \mathbb{R}^k \to \mathbb{R}^p$. From $\mathcal{D}_s$ we form a regressor, X_s, of the metamodel $\boldsymbol{h}(\boldsymbol{x_i})$ for all n_s simulations, and the dependent variables $\boldsymbol{y_s}$, thus

$$\boldsymbol{y_s} = X_s \boldsymbol{\beta} + \boldsymbol{\varepsilon} \ , \tag{8}$$

where $\boldsymbol{\varepsilon}$ is the bias of the simulator and the dimensionalities of $\boldsymbol{\beta}, \boldsymbol{y_s}, X_s$ are as in the example above. Note that the simulator may simulate the exact same process as what is measured. In that case we have $f = h$ from (5) and (7). We do not assume that X_s is full rank.

3.2 The prior problem

We want to work towards answering the following general questions: *1)* Consider the information about $\boldsymbol{\beta}$ that is encoded in the simulations (8). How shall this information be used as a prior when estimating the model (5) with the data $\mathcal{D}_m$? *2)* How does $\mathcal{D}_m, \mathcal{D}_s$, the bias $\boldsymbol{\varepsilon}$ and noise $\boldsymbol{w}$ influence the posterior distribution of $\boldsymbol{\beta}$?

4. REGULARIZATION BASED ESTIMATION

This section presents an accessible way of solving the problem above. Regularization is well known to both be including prior knowledge and reducing MSE. In fact, one may say that regularization is *the result of* incorporating some (uncertain, possibly biased) prior knowledge into a model. However, regularization is often used to sparsify models and reduce overfitting without an explicit justification of the underlying prior.

Consider the standard least squares (LS) regression problem of estimating the model parameters $\boldsymbol{\beta}$ using data from a physical system via the model (5), i.e.,

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \|\boldsymbol{y_m} - X_m \boldsymbol{\beta}\|_2^2 \ . \tag{9}$$

The most straightforward way to use simulation data for regularizing (9) (assuming that (7) holds) is to include a standard L2 norm penalty, regularizing the solution towards the simulated data, i.e.,

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \|\boldsymbol{y_m} - X_m \boldsymbol{\beta}\|_2^2 + \underbrace{\lambda \|\boldsymbol{y_s} - X_s \boldsymbol{\beta}\|_2^2}_{\text{regularization}} \tag{10}$$

which is the ReLS estimator. The degree of regularization is adjusted via the regularization parameter $\lambda \in \mathbb{R}^+$. The analytical solution to (10) is given by

$$\widehat{\boldsymbol{\beta}} = \left(X_m^T X_m + \lambda X_s^T X_s\right)^{-1} \left(X_m^T \boldsymbol{y_m} + \lambda X_s^T \boldsymbol{y_s}\right) , \tag{11}$$

a solution that shows how $\boldsymbol{\beta}$ fuses data from both the physical system and the simulator, with the fusion level depending on the hyperparameter λ.

In this paper we are specifically interested in how to choose λ. Commonly, λ is chosen to minimize the MSE of the estimator, as minimizing the expected posterior loss with a quadratic loss function is equivalent to minimizing the MSE (Berger, 2013). As highlighted by the Gauss-Markov theorem (Carroll and Ruppert, 2017), this choice becomes obvious for problems where the noise on *all* of the data is zero-mean Gaussian with known covariance. However, and as illustrated in Section 1, whenever one of the data sources contains (unknown) bias, optimally choosing λ becomes a non-trivial minimization task.

The main questions we seek to explore in the remainder of the paper are thus:

Q1 How does the presence of bias from the simulator affect the MSE of (11) as a function of λ?
Q2 Are there analytical solutions for obtaining the λ that minimizes the MSE of the estimator?
Q3 Under what conditions does the MSE as a function of λ have the expected shape as seen in Fig. 1?

5. ANALYZING THE REGULARIZED ESTIMATOR

This section analyzes the MSE of the regularized estimator (11). We analyze its derivative w.r.t. λ and show why it is hard to find general analytical solutions for the optimal λ. We show that the scalar case has a closed form solution, and that it behaves as expected. We show that for multiple parameters, problems such as the one seen in Fig. 2 may happen for any realization of the data, for specific biases $\boldsymbol{\varepsilon}$. We also show how increasing the number of data points reduce the likelihood of several minima. We assume that $\boldsymbol{w}$ is i.i.d. Gaussian noise with variance σ^2.

5.1 The MSE and its derivative

The MSE of the estimator $\widehat{\boldsymbol{\beta}}$ is

$$\text{MSE}(\widehat{\boldsymbol{\beta}}) = \text{Tr}\left[\text{Var}(\widehat{\boldsymbol{\beta}})\right] + \text{Bias}(\widehat{\boldsymbol{\beta}})^T \text{Bias}(\widehat{\boldsymbol{\beta}}) \tag{12a}$$

where

$$\text{Bias}(\widehat{\boldsymbol{\beta}}) = \lambda M^{-1} X_s^T \boldsymbol{\varepsilon} \tag{12b}$$

$$\text{Var}(\widehat{\boldsymbol{\beta}}) = M^{-1} X_m^T \sigma^2 X_m M^{-1} \tag{12c}$$

and

$$M := X_m^T X_m + \lambda X_s^T X_s \ .$$

Note that M is a symmetric PD matrix, i.e., $M = M^T$ and $M^{-T} = M^{-1}$, since X_m is assumed to be full rank. As the model is linear in the parameters, the MSE of $\widehat{\boldsymbol{\beta}}$ does not depend on the true parameter value, but rather on the bias $\boldsymbol{\varepsilon}$.

The problem is to choose the optimal relative weight, namely

$$\lambda^* = \underset{\lambda \in \mathbb{R}^+}{\arg\min} \, \text{MSE}(\widehat{\boldsymbol{\beta}}, \lambda) \ ,$$

an operation that is usually done by solving $\frac{\partial \text{MSE}}{\partial \lambda}\big|_{\lambda=\lambda^*} = 0$, and confirming that the extremum is a minimum. To compute the stationary points, we note that differentiating (12a) leads to

$$\frac{\partial}{\partial \lambda}\text{MSE} = \text{Tr}\left[\frac{\partial}{\partial \lambda}\text{Var}\right] + 2\text{Bias}^T \frac{\partial}{\partial \lambda}\text{Bias} \ . \tag{13a}$$

By applying

$$\frac{\partial}{\partial \lambda} M^{-1} = -M^{-1} X_s^T X_s M^{-1} \tag{13b}$$

we obtain

$$\frac{\partial}{\partial \lambda}\text{Bias}(\widehat{\boldsymbol{\beta}}) = \left(I - \lambda M^{-1} X_s^T X_s\right) M^{-1} X_s^T \boldsymbol{\varepsilon} \tag{13c}$$

$$\frac{\partial}{\partial \lambda}\mathrm{Var}(\widehat{\boldsymbol{\beta}}) = -\sigma^2 M^{-1}\bigg(X_s^T X_s M^{-1} X_m^T X_m + X_m^T X_m M^{-1} X_s^T X_s\bigg) M^{-1} \,. \tag{13d}$$

5.2 Conditions for Uniqueness of Minimum

As discussed in Section 1, certain realizations of the data may lead to multiple minima when $\dim(\boldsymbol{\beta}) > 1$, as in Fig. 2. Equation (13a) is analyzed here, where we will be referring to the terms as the variance and the bias term, respectively. Well-known properties of the matrix inverse and positive semi-definite (PSD) matrices are applied throughout (Horn and Johnson, 2012).

We show that the variance is monotonically decreasing, as expected. This implies that the bias term is causing the issue described. Since the variance term in (13a) is always negative and the bias term is "mostly" positive, it holds that if the bias term in (13a) changes sign, then there *will* be multiple minima for certain σ^2. Thus, a necessary condition for having a unique minimum for arbitrary σ^2 is that the differentiated bias term is strictly positive, making the bias monotonically increasing. Such a necessary condition is described below.

Firstly, the variance differentiated is always negative, which can be seen from applying $A := X_m^T X_m > 0, B_\delta := X_s^T X_s + \delta I > 0$ for $\delta > 0$. This gives that

$$\begin{aligned} & X_s^T X_s M^{-1} X_m^T X_m + X_m^T X_m M^{-1} X_s^T X_s \\ &= \lim_{\delta \to 0^+} B_\delta M^{-1} A + A M^{-1} B_\delta \\ &= \lim_{\delta \to 0^+} 2(B_\delta^{-1} + \lambda A^{-1})^{-1} \geq 0 \,, \forall \lambda \geq 0 \end{aligned} \tag{14}$$

by continuity of the matrix inverse. This makes (13d) negative semi-definite, considering the negative sign, so its trace is negative. Furthermore, considering a larger regularization at $\lambda + \Delta\lambda$, $\Delta\lambda \geq 0$, then

$$0 \leq M \leq \left(X_m^T X_m + (\lambda + \Delta\lambda)\, X_s^T X_s\right) := M^+ \tag{15}$$

giving that $(M^+)^{-1} \leq M^{-1}$. Then, corresponding to (14), at $\lambda + \Delta\lambda$ we have that

$$0 \leq \left(B_\delta^{-1} + (\lambda + \Delta\lambda) A^{-1}\right)^{-1} \leq \left(B_\delta^{-1} + \lambda A^{-1}\right)^{-1} \,. \tag{16}$$

This makes (13d) of the form $-CDC$, where C, D are both PSD matrices, showing that the derivative of the variance increases monotonically to zero as $\lambda \to +\infty$. Thus, the variance is always positive but its derivative is always negative, and the variance term is monotonically decreasing to zero as $\lambda \to +\infty$. This holds for any rank of X_s. Notice that σ^2 only scales the variance term.

The bias term differentiated is less regular, being

$$\mathrm{Bias}^T \frac{\partial \mathrm{Bias}}{\partial \lambda} = \lambda \boldsymbol{\varepsilon}^T X_s M^{-1}\left(I - \lambda M^{-1} X_s^T X_s\right) M^{-1} X_s^T \boldsymbol{\varepsilon} \,. \tag{17}$$

Due to the particular form of (17) ($\sim z^T F z$) it suffices to consider its symmetric part only (as the anti-symmetric part cancels), namely

$$\lambda \boldsymbol{\varepsilon}^T X_s M^{-1}\left(I - \underbrace{\frac{\lambda}{2}\left(M^{-1} X_s^T X_s + X_s^T X_s M^{-1}\right)}_{:=C}\right) M^{-1} X_s^T \boldsymbol{\varepsilon} \tag{18}$$

It is meaningful to consider only $I - C$ in (18) due to the Rayleigh-Ritz theorem (Horn and Johnson, 2012), which says that for a symmetric matrix A,

$$\min \mathrm{eig}(A) z^T z \leq z^T A z \leq \max \mathrm{eig}(A) z^T z \,, \tag{19}$$

where the inequalities are tight. The matrix $I - C$ may have negative eigenvalues, depending on the properties of C, since $\mathrm{eig}\,(I - C) = 1 - \mathrm{eig}(C)$. Therefore, we go on analyzing if $\mathrm{eig}\,(C) \leq 1$. This set is approximated by the (real) matrix' *field of values* (Givens, 1952), via

$$F(A) := \{z^T A z \,|\, z^T z = 1\} \implies \mathrm{eig}(A) \in F(A) \,.$$

The following applies to the fields of values here:

$$\begin{aligned} & F(A) = [\min \mathrm{eig}(A), \max \mathrm{eig}(A)] \\ & F(A) + F(B) := \{a + b \,|\, a \in F(A), b \in F(B)\} \\ & F(A)/F(B) := \{a/b \,|\, a \in F(A), b \in F(B), 0 \notin F(B)\} \,. \end{aligned}$$

By applying theorems for eigenvalues and fields of values (Wielandt, 1972), this yields

$$\mathrm{eig}\,(C) \subseteq F(\lambda X_s^T X_s)/F(M)$$

where

$$F(M) \subseteq F(X_m^T X_m) + F(\lambda X_s^T X_s) \,.$$

Therefore

$$\mathrm{eig}\,(C) \subseteq \frac{F(\lambda X_s^T X_s)}{F(X_m^T X_m) + F(\lambda X_s^T X_s)} \tag{22}$$

which is an interval of positive values. As $\lambda \to +\infty$ the upper bound from (22) on $\mathrm{eig}(C)$ will be larger than 1, however, the size of the eigenspace associated with the eigenvalues above 1 should decrease as more data is added. In other words, there may always be certain biases $\boldsymbol{\varepsilon}$ that make (17) negative, though they are less and less likely to occur. Further investigations towards this are found below.

5.3 Recognizing the Stein effect

When $\lambda = 0$ the bias term and its derivative vanishes. By the above, it is clear that the derivative of the variance term is negative, meaning that we are *guaranteed* that there exists *some* (potentially very small) level of regularization that lowers the MSE of the LS estimator - independently of how biased the simulated data is. This is also known as the Stein effect (Pillonetto et al., 2022).

As $\lambda \to +\infty$ the variance term vanishes, and we are left with $\frac{\partial}{\partial \lambda}\mathrm{MSE} = 2\mathrm{Bias}^T \frac{\partial}{\partial \lambda}\mathrm{Bias}$. As seen in Section 5.2, generally we can not guarantee that the derivative of the bias term is positive (this depends on X_s and $\boldsymbol{\varepsilon}$), meaning that a similar Stein effect (lowering the MSE by data fusion) is not necessarily present for $\lambda \to +\infty$. However, when $\dim(\beta) = 1$, $X_s^T X_s, X_m^T X_m$ are scalars, so by equation (22), $\mathrm{eig}(C)$ is a scalar always smaller than 1. Thus, then the bias term is monotonically increasing. In fact, we may then rewrite (13a) as

$$\frac{\partial}{\partial \lambda}\mathrm{MSE} = -2 X_m^T X_m \frac{\sigma^2 X_s^T X_s - \lambda (X_s^T \boldsymbol{\varepsilon})^2}{M^3} \,.$$

This has a unique stationary point given by

$$\lambda^* = \sigma^2 \frac{X_s^T X_s}{(X_s^T \boldsymbol{\varepsilon})^2} \,, \tag{23}$$

which indeed is a minimum by the Stein effect.

5.4 Increasing the sample size

Any MSE curve depends on the specific $\boldsymbol{\varepsilon}$, but the MSE is *more likely* to have a unique minimum if more data is

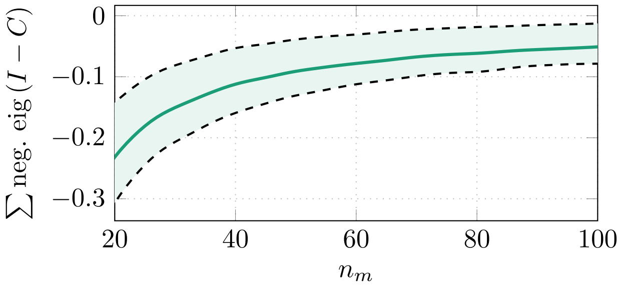

Fig. 3. Sum of negative eigenvalues (worst case over λ) of $I - C$ when increasing the sample size $n_m \in [20, 100]$ for 10000 random realizations of $X_{m,ij} \sim \mathcal{U}_{[0,1]}$ (uniform distribution), where $\dim(\boldsymbol{\beta}) = p = 10$. The mean over all realizations is shown in solid, while the shaded region mark the $[25\%, 75\%]$ quartile.

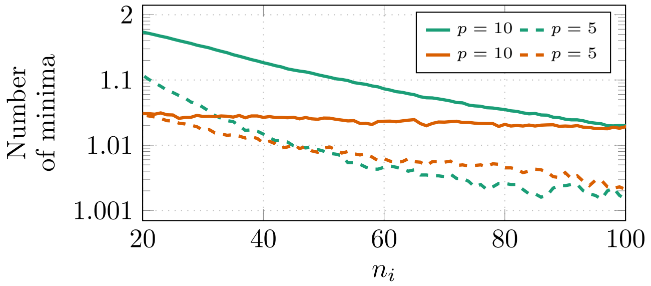

Fig. 4. Number of minima of the MSE when increasing the sample size $n_m, n_s \in [20, 100]$ for 10000 random realizations of $X_{m,ij}, X_{s,ij} \sim \mathcal{U}_{[0,1]}$ (uniform distribution), where $p = \{5, 10\}$, $\sigma = 1$ and $\boldsymbol{\varepsilon}_i \sim \mathcal{U}_{[0,5]}$. $p = \dim(\boldsymbol{\beta})$. The lines are the mean of all realizations. **Green**: n_m increases, $n_s = 50$. **Orange**: n_s increases, $n_m = 50$.

included, as seen from (22). Adding data to the matrices X_m, X_s corresponds to adding rows x^T. Adding x^T to X_i gives that $X_i^T X_i \rightarrow X_i^T X_i + xx^T$. Since xx^T is PSD, adding it to $X_i^T X_i$ increases the eigenvalues of $X_i^T X_i$, but not necessarily all of them or specific ones. This depends on x together with the eigenspace of X_i ($i = \{s, m\}$). Hence, adding (arbitrary) data to $\mathcal{D}_m$ should generally be more beneficial than adding to $\mathcal{D}_s$, since we may raise the minimal eigenvalue of $X_m^T X_m$, thus the minimum of $F(X_m^T X_m)$, while adding data to $\mathcal{D}_s$ gives less precise answers regarding the bounds of (22) since both the enumerator and denominator are affected. This is also observed in numerical tests, indicating that the negative eigenvalues of $I - C$ from (18) both decrease in quantity and magnitude as more data is used, depicted in Fig. 3. This seems to correspond to a decreasing number of minima of the MSE, as depicted in Fig. 4.

6. CONCLUSION

Regularizing towards prior knowledge is, as an approach, more flexible than many gray-box modeling principles. However, as shown here, regularizing optimally towards biased data is a non-trivial problem. The MSE optimization problem may exhibit multiple minima, and for certain regularization levels λ the overall MSE can become larger than both the MSE at $\lambda = 0$ *and* $\lambda \rightarrow +\infty$. This unwanted behaviour could be present in other regularized models too, even if the analysis presented here suggests that this issue diminishes as the dataset size increases. Our main result is that, generally, using simulations as a prior needs caution. A further development may then be adopting Bayesian perspectives by posing the estimation problem as a kernel regression problem and analyzing how the assumptions on the bias ε affect the posterior distribution, to possibly construct more robust approaches for utilizing simulators in machine learning.

REFERENCES

Berger, J.O. (2013). *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media.

Brunton, S.L., Proctor, J.L., and Kutz, J.N. (2016). Sparse identification of nonlinear dynamics with control (sindyc). *IFAC-PapersOnLine*, 49(18), 710–715.

Carroll, R.J. and Ruppert, D. (2017). *Transformation and Weighting in Regression*. Chapman and Hall/CRC.

Fujimoto, Y. and Sugie, T. (2018). Kernel-based impulse response estimation with a priori knowledge on the DC gain. *IEEE Control Systems Letters*, 2(4), 713–718.

Givens, W. (1952). Fields of values of a matrix. *Proceedings of the American Mathematical Society*.

Glassey, J. and Von Stosch, M. (2018). *Hybrid modeling in process industries*. CRC Press.

Horn, R.A. and Johnson, C.R. (2012). *Matrix analysis*. Cambridge university press.

Kedem, B., De Oliveira, V., and Sverchkov, M. (2017). *Statistical data fusion*. World Scientific.

Khosravi, M. and Smith, R. (2021). *Kernel-based impulse response identification with side-information on steady-state gain*. arXiv.2111.00409v1.

Ljung, L. (1998). *System identification*. Springer.

Pillonetto, G., Chen, T., Chiuso, A., De Nicolao, G., and Ljung, L. (2022). *Regularized System Identification: Learning Dynamic Models from Data*. Springer Nature.

Raissi, M., Perdikaris, P., and Karniadakis, G.E. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics*, 378, 686–707.

Saadallah, A., Finkeldey, F., Buß, J., Morik, K., Wiederkehr, P., and Rhode, W. (2022). Simulation and sensor data fusion for machine learning application. *Advanced Engineering Informatics*, 52, 101600.

Shin, Y., Darbon, J., and Karniadakis, G.E. (2020). On the convergence of physics informed neural networks for linear second-order elliptic and parabolic type pdes. *arXiv preprint arXiv:2004.01806*.

Sohlberg, B. and Jacobsen, E.W. (2008). Grey box modelling–branches and experiences. *IFAC Proceedings Volumes*, 41(2), 11415–11420.

Sparta, M., Varagnolo, D., Stråbø, K., Halvorsen, S.A., Herland, E.V., and Martens, H. (2021). Metamodeling of the electrical conditions in submerged arc furnaces. *Metallurgical and Materials Transactions B*.

Wielandt, H. (1972). On the eigenvalues of a+ band ab. *Journal or Research of the National Bureau of Standards, Sec. B*, 1-2.

Yoshimura, S., Matsubayashi, A., and Inoue, M. (2019). System identification method inheriting steady-state characteristics of existing model. *International Journal of Control*, 92(11), 2701–2711.