IEEE Control Systems Letters paper presented at
2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

Error Bounds for Kernel-Based Linear System Identification with Unknown Hyperparameters

Mingzhou Yin, Roy S. Smith

Abstract— **Applying regularization in reproducing kernel Hilbert spaces has been successful in linear system identification using stable kernel designs. From a Gaussian process perspective, it automatically provides probabilistic error bounds for the identified models from the posterior covariance, which are useful in robust and stochastic control. However, the error bounds require knowledge of the true hyperparameters in the kernel design. They can be inaccurate with estimated hyperparameters for lightly damped systems or in the presence of high noise. In this work, we provide reliable quantification of the estimation error when the hyperparameters are unknown. The bounds are obtained by first constructing a high-probability set for the true hyperparameters from the marginal likelihood function. Then the worst-case posterior covariance is found within the set. The proposed bound is proven to contain the true model with a high probability and its validity is demonstrated in numerical simulation.**

I. Introduction

System identification estimates models of dynamical systems from input-output data. Under the assumption that a low-dimensional model structure is known *a priori*, the model parameters can be estimated by maximum likelihood estimation, of which one example is the prediction error method (PEM) [1]. This framework, despite some numerical difficulties, has been very successful in various applications [2].

However, as more complex and large-scale systems are emerging, low-dimensional model structures become less accessible. Following the seminal work [3], system identification can be reformulated as a non-parametric function learning problem, and solved using, among other approaches, the kernel-based method [4], [5], [6]. The kernel-based method can be interpreted as function learning in a reproducing kernel Hilbert space (RKHS), Gaussian process (GP) regression, or ridge regression with basis expansions. In linear system identification, a truncated impulse response model is identified with a weighted l_2 regularization term, with a class of stable kernels designed to identify stable linear systems effectively [3], [7], [8]. The GP interpretation provides the kernel-based method with one of its main advantages: it obtains Gaussian stochastic models and thus high-probability error bounds simultaneously with the nominal model [9]. This enables its application in robust and stochastic control.

However, one often-neglected aspect of kernel-based identification is that the results, including the error bounds, are conditioned on correct hyperparameter selection, in the same way as PEM is conditioned on the correct model structure. The hyperparameters in the kernel-based method are usually selected separately using the maximum marginal likelihood method or cross-validation, and used in identification empirically with certainty equivalence. This makes the GP-based error bounds unreliable when the estimated hyperparameters are inaccurate, and thus detrimental to use in safety-critical applications. This phenomenon has been observed in machine learning literature [10]. In linear system identification, it is demonstrated in this paper that the error bound derived from estimated hyperparameters can be inaccurate, especially for lightly damped systems and in low signal-to-noise ratio scenarios.

Therefore, error bounds for the kernel-based methods are needed in the case of unknown hyperparameters. In kernel-based linear system identification, [11] establishes non-asymptotic bounds for all stable systems with bounded pole magnitudes. However, the bounds are too conservative and thus only useful for sample complexity analysis. Error bounds are also widely studied in GP literature, e.g., [12], [13], which are usually obtained by scaling posterior standard deviations. Such bounds are derived under the assumptions that the prior covariance function is exact and/or an upper bound of the RKHS-induced norm is known. However, both assumptions are impractical in general. Several works provide modified bounds considering the discrepancy between the applied and the true kernel functions [14], [15], [16], [17]. These results depend on knowledge of the magnitude of the discrepancy, which is usually not known *a priori*. Such information is estimated from data in [14] by investigating the maximum marginal likelihood problem in hyperparameter estimation. Unfortunately, these works all consider an identity regressor which is not common in system identification and often consider kernel classes that do not contain the typical stable kernels used in linear system identification. Sampling-based approaches have also been proposed. The sign-perturbed sums approach is used in [18] by perturbing the sign of model residuals randomly. The Markov chain Monte Carlo approach is used in [19] to approximate the full posterior distribution. However, such bounds are based on sampling and thus do not admit an easy-to-use analytic form.

In this work, we provide probabilistic error bounds for kernel-based linear system identification with no prior knowledge of the hyperparameters by extending [14] to general regression problems and stable kernels. The proof in [14] is also simplified with an improved constant. Our approach

This work was supported by the Swiss National Science Foundation under Grant 200021_178890.

The authors are with the Automatic Control Laboratory, Swiss Federal Institute of Technology (ETH Zurich), Physikstrasse 3, 8092 Zurich, Switzerland, {myin,rsmith} @control.ee.ethz.ch.

assumes the correct kernel structure and a known hyperprior that describes the distribution of the hyperparameters. A high-probability set is first estimated for the hyperparameters from the marginal likelihood function. Then, the upper bound of the posterior covariance is found within the range of hyperparameters. A uniform bound is obtained for diagonal and tuned/correlated kernels. For general kernels, element-wise bounds can be found by optimization. Finally, probabilistic error bounds are established by scaling the worst-case posterior standard deviations. Optimization problems to obtain the tightest error bounds are discussed as well. Numerical simulations demonstrate that the proposed error bounds are able to provide high-probability bounds of the estimation error in practice.

Notation. A Gaussian distribution with mean μ and covariance Σ is indicated by $\mathscr{N}(\mu,\Sigma)$. The expectation and the covariance of a random vector x are denoted by $\mathbb{E}(x)$ and $\mathrm{cov}(x)$, respectively. The notation $A \preccurlyeq B$ means $(B-A)$ is positive semidefinite. The symbol $\overset{p}{\leq}$ indicates less than or equal to with probability p. For a vector x, the weighted l_2-norm $(x^\top P x)^{\frac{1}{2}}$ is denoted by $\|x\|_P$. The (i,j)-th element and the trace of a matrix A are denoted by $A_{i,j}$ and $\mathrm{tr}(A)$, respectively.

II. The Kernel-Based Method in Linear System Identification

A. Problem Statement

Consider a causal and stable linear time-invariant single-input single-output discrete-time system $y_t = G(q)u_t + v_t$, where u_t, y_t, v_t are the inputs, outputs, and additive noise respectively, and q is the shift operator. The additive noise is assumed to be zero-mean i.i.d. Gaussian with a variance of σ^2. We are interested in identifying the transfer function $G(q)$. In this work, we consider the finite impulse response model of $G(q)$: $G(q) = \sum_{l=0}^{n_g-1} g_l q^{-l}$, i.e., $y_t = \sum_{l=0}^{n_g-1} g_l u_{t-l} + v_t$.

For this purpose, an input-output sequence of the system

$$\mathbf{u} = \begin{bmatrix} u_{2-n_g} & u_{3-n_g} & \dots & u_N \end{bmatrix}^\top,\ \mathbf{y} = \begin{bmatrix} y_1 & y_2 & \dots & y_N \end{bmatrix}^\top \tag{1}$$

has been collected. This leads to the data equation:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \underbrace{\begin{bmatrix} u_1 & u_0 & \cdots & u_{2-n_g} \\ u_2 & u_1 & \cdots & u_{3-n_g} \\ \vdots & \vdots & \ddots & \vdots \\ u_N & u_{N-1} & \cdots & u_{N-n_g+1} \end{bmatrix}}_{\Phi} \underbrace{\begin{bmatrix} g_0 \\ g_1 \\ \vdots \\ g_{n_g-1} \end{bmatrix}}_{\mathbf{g}} + \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_N \end{bmatrix}. \tag{2}$$

B. The Kernel-Based Method

If no prior knowledge is assumed for $G(q)$, the maximum likelihood estimator of $\mathbf{g}$ is given by the least-squares solution

$$\hat{\mathbf{g}}^{\mathrm{LS}} = \underset{\mathbf{g}}{\mathrm{argmin}}\ \|\mathbf{y}-\Phi\mathbf{g}\|_2^2 = \left(\Phi^\top\Phi\right)^{-1}\Phi^\top\mathbf{y}, \tag{3}$$

where Φ and $\mathbf{g}$ are defined in (2). It is well known that the estimation error is also Gaussian with covariance $\mathrm{cov}(\mathbf{g}) = \sigma^2\left(\Phi^\top\Phi\right)^{-1} =: \Sigma^{\mathrm{LS}}$. Element-wise stochastic error bounds can be obtained for $\hat{\mathbf{g}}^{\mathrm{LS}}$ as

$$\mathbb{P}\left(\left|\hat{g}_l^{\mathrm{LS}} - g_l\right| \leq \mu_\delta\sqrt{\Sigma_{l,l}^{\mathrm{LS}}}\right) \geq 1-\delta, \tag{4}$$

where μ_δ is the two-tailed quantile function of the Gaussian distribution, given by

$$F_{\mathscr{N}}(\mu_\delta) \geq 1-\delta/2, \tag{5}$$

$F_{\mathscr{N}}(\cdot)$ is the cumulative distribution function of the Gaussian distribution.

Note that the impulse response of the stable system $G(q)$ is typically smooth and exponentially converges to zero. This prior knowledge can be encoded as either a prior distribution in GP regression or an RKHS in kernel regression. In both cases, the nominal estimate of $\mathbf{g}$ is given by the ridge-regularized least-squares solution

$$\hat{\mathbf{g}} = \underset{\mathbf{g}}{\mathrm{argmin}}\ \|\mathbf{y}-\Phi\mathbf{g}\|_2^2 + \sigma^2\mathbf{g}^\top K^{-1}\mathbf{g} \tag{6}$$

$$= \left(\Phi^\top\Phi + \sigma^2K^{-1}\right)^{-1}\Phi^\top\mathbf{y}, \tag{7}$$

with K having different interpretations.

In the GP regression interpretation, K is the covariance of the prior distribution of $\mathbf{g}$: $\mathbf{g} \sim \mathscr{N}(\mathbf{0},K)$. Then $\mathbf{y}$ and $\mathbf{g}$ are jointly Gaussian:

$$\begin{bmatrix} \mathbf{g} \\ \mathbf{y} \end{bmatrix} \sim \mathscr{N}\left(\mathbf{0}, \begin{bmatrix} K & K\Phi^\top \\ \Phi K & \Phi K\Phi^\top + \sigma^2\mathbb{I} \end{bmatrix}\right). \tag{8}$$

The distribution of $\mathbf{g}$ given $\mathbf{y}$ is also Gaussian: $\mathbf{g}|\mathbf{y} \sim \mathscr{N}(\hat{\mathbf{g}},\Sigma)$, where the posterior mean is the estimate $\hat{\mathbf{g}}$ and $\Sigma = \sigma^2\left(\Phi^\top\Phi + \sigma^2K^{-1}\right)^{-1}$ is the posterior covariance [9]. From the posterior covariance, the associated element-wise stochastic error bounds are

$$\mathbb{P}\left(|\hat{g}_l - g_l| \leq \mu_\delta\sqrt{\Sigma_{l,l}}\right) \geq 1-\delta, \tag{9}$$

conditioned on the identification data. The bounds assume that the prior distribution of $\mathbf{g}$ is correct.

In the kernel regression interpretation [20, Chapter 3], the *continuous-time* impulse response function $g(t) : [0,+\infty) \to \mathbb{R}$, $g(l) = g_l$, $l = 0,\dots,n_g-1$ is identified by solving the regularized function learning problem within an RKHS $\mathscr{H}$ associated with a kernel function $k(\cdot,\cdot) : [0,+\infty)\times[0,+\infty) \to \mathbb{R}$:

$$\begin{aligned} g^\star(\cdot) = \arg\min_{g(\cdot)\in\mathscr{H}}\ & \|\mathbf{y}-\Phi\mathbf{g}\|_2^2 + \sigma^2\|g(\cdot)\|_{\mathscr{H}}^2 \\ \text{s.t.}\ & \mathbf{g} = \begin{bmatrix} g(0) & \dots & g(n_g-1) \end{bmatrix}^\top, \end{aligned} \tag{10}$$

where $\|g(\cdot)\|_{\mathscr{H}}$ is the induced norm of $g(\cdot)$ in $\mathscr{H}$. From the representer theorem [21], the optimal continuous-time impulse response function for (10) is given by $g^\star(x) = \mathbf{k}_x\left(\Phi^\top\Phi K + \sigma^2\mathbb{I}\right)^{-1}\Phi^\top\mathbf{y}$, where K evaluates the kernel function associated with the RKHS $\mathscr{H}$ at $l = 0,\dots,n_g-1$, i.e., $K_{l,l} = k(l,l)$, and $\mathbf{k}_x = [k(x,0)\ \dots\ k(x,n_g-1)]$. The corresponding optimal discrete-time impulse response vector is $\mathbf{g}^\star = \hat{\mathbf{g}}$. The induced norm of $g^\star$ is calculated as $\|g^\star(\cdot)\|_{\mathscr{H}}^2 = \hat{\mathbf{g}}^\top K^{-1}\hat{\mathbf{g}}$.

C. Kernel Design and Hyperparameter Selection

The matrix K is critical to the performance of the kernel-based method. Extensive studies have been conducted to obtain appropriate structures of K that promote impulse response estimates that are both smooth and exponentially converge to zero [8]. These structures parameterize the kernel with hyperparameters η: $K = K(\eta)$. The most commonly used kernels in linear system identification include:

1) diagonal (DI): $K^{\mathrm{DI}}_{i,i}(\eta) = c\lambda^i$, $K^{\mathrm{DI}}_{i,j}(\eta) = 0$ for $i \neq j$,
2) tuned/correlated (TC): $K^{\mathrm{TC}}_{i,j}(\eta) = c\lambda^{\max(i,j)}$,
3) stable spline (SS):

$$K^{\mathrm{SS}}_{i,j}(\eta) = c\lambda^{2\max(i,j)}\left(\frac{\lambda^{\min(i,j)}}{2} - \frac{\lambda^{\max(i,j)}}{6}\right),$$

where $\eta = [c \;\; \lambda]^\top \in \left\{[c \;\; \lambda]^\top \middle| c \geq 0, 0 \leq \lambda \leq 1\right\} =: \mathbb{H}$ are the hyperparameters. These kernel designs have been shown effective both theoretically and numerically [6].

The hyperparameters need to be selected before applying the estimator (7). The most widely-used approach to hyperparameter selection is the maximum marginal likelihood method. It uses the GP regression interpretation and maximizes the probability of observing $\mathbf{y}$ given the inputs $\mathbf{u}$ and the hyperparameters η: $\hat{\eta} = \underset{\eta}{\operatorname{argmin}} -\log p(\mathbf{y}|\mathbf{u},\eta)$, where

$$p(\mathbf{y}|\mathbf{u},\eta) = \exp\left(-\tfrac{1}{2}\log\det\Psi(\eta) - \tfrac{1}{2}\mathbf{y}^\top\Psi^{-1}(\eta)\mathbf{y} + \text{const.}\right) \tag{11}$$

and $\Psi = \sigma^2\mathbb{I} + \Phi K(\eta)\Phi^\top$. If the prior distribution of the hyperparameters $p(\eta)$, known as the hyperprior, is available, η can also be estimated using a maximum a posterior approach: $\hat{\eta} = \underset{\eta}{\operatorname{argmin}} -\log p(\mathbf{y}|\mathbf{u},\eta)p(\eta)$ [22]. The estimated hyperparameters $\hat{\eta}$ are used, with certainty equivalence, to construct $K(\hat{\eta})$ and then to obtain the estimate $\hat{\mathbf{g}}$.

III. Error Bounds with Unknown Hyperparameters

A. Pitfalls with Error Bounds from Posterior Covariances

The kernel-based method has shown remarkable performance in linear system identification, in terms of the nominal estimate (7). However, the stochastic error bound (9) is only rigorously valid when we consider a random impulse response model subject to an exact prior distribution with exact hyperparameters. On the other hand, in practical system identification applications, a fixed plant is usually considered, and hyperparameters are estimated as the most probable ones if the impulse response is drawn from the prior distribution with the assumed structure. When the estimated hyperparameters $\hat{\eta}$ are used, directly using (9) to provide a stochastic model for a fixed plant can be problematic.

To demonstrate this issue, consider two second-order systems

$$G_1(q) = \frac{0.4888}{q^2 - 1.8q + 0.9^2},\; G_2(q) = \frac{0.0616}{q^2 - q + 0.9^2},$$

with two different noise levels $\sigma^2 = 0.1$ and 0.5. Both systems have two poles of magnitude 0.9: $G_1(q)$ has two real poles at 0.9; $G_2(q)$ has a pair of complex poles with a real part of 0.5. The systems have been normalized to have an $\mathscr{H}_2$-norm of 1.

Stochastic models given by the error bound (9) with $\hat{\eta}$ are analyzed by 1000 Monte Carlo simulations with TC kernels. Different unit Gaussian inputs are used to generate the identification data in each run. Table I shows the empirical probabilities of violating the elementwise bounds with $\delta = 0.1$ and identification parameters $N = 200$ and $n_g = 50$. It can be seen that except for the case of $G_1(q)$ with low noise, the magnitudes of the errors are significantly underestimated in the other three cases, with bound violation probabilities much larger than the target value of $\delta = 0.1$. This indicates that the error bounds based on estimated hyperparameters are not reliable in cases where the impulse response is lightly damped and/or the signal-to-noise ratio is poor.

TABLE I
Empirical probability of bound violations and standard deviations of hyperparameter estimation

$\delta = 0.1$	% bound violations	STD($\hat{c}$)	STD($\hat{\lambda}$)
(a) $G_1, \sigma^2 = 0.1$	13.2%	0.0052	0.0069
(b) $G_2, \sigma^2 = 0.1$	29.8%	0.2191	0.0204
(c) $G_1, \sigma^2 = 0.5$	24.6%	0.0010	0.0313
(d) $G_2, \sigma^2 = 0.5$	60.1%	0.0242	0.0373

To investigate the reason why the error bounds are inaccurate under these cases, Table I also shows the standard deviations of the estimated hyperparameters, and Figure 1 plots the marginal probability density (11) with respect to the hyperparameters in one representative simulation. It can be seen that in cases (b), (c), and (d), where the error bounds based on estimated hyperparameters are inaccurate, the variances of the estimated hyperparameters are larger than those in case (a), and the marginal probability density is not strongly localized. This suggests that the estimated hyperparameters can be inaccurate, which leads to the misspecification of the error bounds.

B. Worst-Case Posterior Variances

To solve the problem of quantifying error bounds with unknown hyperparameters, we first bound the true hyperparameters using the measured data. In this work, we consider the following two assumptions.

Assumption 1: The kernel structure $K(\eta)$ is assumed to be correct with unknown true hyperparameters η_0.

Assumption 2: The hyperprior $p(\eta)$ is known.

The hyperprior $p(\eta)$ can be selected as a uniform distribution if no additional knowledge about the hyperparameters is available. The distribution of hyperparameters conditioned on the measured data is given by

$$p(\eta|\mathbf{u},\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{u},\eta)p(\eta)}{\int_{\eta\in\mathbb{H}} p(\mathbf{y}|\mathbf{u},\eta)p(\eta)\,\mathrm{d}\eta}, \tag{12}$$

where $p(\mathbf{y}|\mathbf{u},\eta)$ is given in (11). This leads to

$$\mathbb{P}(\eta_0 \in [\eta_1,\eta_2]) = \frac{\int_{\eta\in[\eta_1,\eta_2]} p(\mathbf{y}|\mathbf{u},\eta)p(\eta)\,\mathrm{d}\eta}{\int_{\eta\in\mathbb{H}} p(\mathbf{y}|\mathbf{u},\eta)p(\eta)\,\mathrm{d}\eta} =: 1-\delta', \tag{13}$$

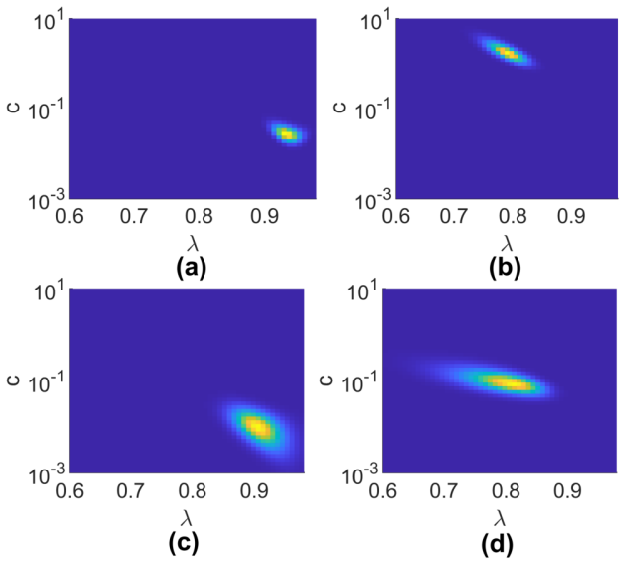

Fig. 1. Marginal probability density with respect to hyperparameters. (a) $G_1, \sigma^2 = 0.1$, (b) $G_2, \sigma^2 = 0.1$, (c) $G_1, \sigma^2 = 0.5$, (d) $G_2, \sigma^2 = 0.5$. Yellow: higher value, blue: lower value.

where $\eta_i = [c_i \ \lambda_i]^\top$, $i = 1, 2$ and

$$[\eta_1, \eta_2] = \left\{ \eta = [c \ \lambda]^\top \,|\, c_1 \le c \le c_2, \lambda_1 \le \lambda \le \lambda_2 \right\}$$

is a rectangular set. By choosing a small δ', (13) establishes a high-probability set for the true hyperparameters.

Then, we investigate the effect of hyperparameters on the stochastic model, to find the worst-case posterior variances $\Sigma_{l,l}$ within the set $[\eta_1, \eta_2]$. For DI and TC kernels, a uniform bound is derived analytically using the following lemma.

Lemma 1: The matrix inequality $\Sigma(\eta_1) \preccurlyeq \Sigma(\eta_2)$ is satisfied when $\left(\frac{\lambda_2}{\lambda_1}\right)^\gamma c_1 \le c_2$, $\lambda_1 \le \lambda_2$, with $\gamma = 0$ for DI kernels and $\gamma = -1/\ln \lambda_2 - 1$ for TC kernels.

Proof: See Appendix I. ■

From Lemma 1, we have

$$\Sigma(\eta_0) \overset{1-\delta'}{\preccurlyeq} \sigma^2 \left(\Phi^\top \Phi + \sigma^2 \left(\frac{\lambda_1}{\lambda_2} \right)^\gamma K^{-1}(\eta_2) \right)^{-1} =: \bar{\Sigma}. \quad (14)$$

So the posterior variances with true hyperparameters η_0 can be uniformly bounded by

$$\Sigma_{l,l}(\eta_0) \overset{1-\delta'}{\le} \bar{\Sigma}_{l,l} =: \sigma_l^2. \quad (15)$$

For a general kernel structure, the bound can be computed element-wise by directly solving the optimization problem:

$$\sigma_l^2 = \max_{\eta \in [\eta_1, \eta_2]} \Sigma_{l,l}(\eta). \quad (16)$$

C. Stochastic Error Bounds

We are now ready to present the main result of the paper.

Theorem 1: The impulse response estimate (7) with estimated hyperparameters $\hat{\eta}$ admits the following stochastic element-wise error bound:

$$\mathbb{P}(|\hat{g}_l(\hat{\eta}) - g_l| \le \bar{\mu} \sigma_l) \ge (1-\delta)(1-\delta'), \quad (17)$$

where $\bar{\mu} = \mu_\delta + \frac{2}{\sigma} \|\mathbf{y}\|_S$ and $S = \Phi \left(\Phi^\top \Phi \right)^{-1} \Phi^\top$, if $\hat{\eta} \in [\eta_1, \eta_2]$.

Proof: The estimation error is decomposed as

$$|\hat{g}_l(\hat{\eta}) - g_l| \ \le \ |\hat{g}_l(\hat{\eta}) - \hat{g}_l(\eta_0)| + |\hat{g}_l(\eta_0) - g_l| \quad (18)$$

$$\overset{1-\delta}{\le} |\hat{g}_l(\hat{\eta}) - \hat{g}_l(\eta_0)| + \mu_\delta \sqrt{\Sigma_{l,l}(\eta_0)}, \quad (19)$$

where the two terms are due to misspecified hyperparameters and measurement noise respectively.

Define the posterior kernel

$$k_\eta^p(x, x') = k_\eta(x, x') - \mathbf{k}_x(\eta) \left(K(\eta) + \sigma^2 \left(\Phi^\top \Phi \right)^{-1} \right)^{-1} \mathbf{k}_x(\eta)^\top.$$

Note that $k_\eta^p(i, j) = \Sigma_{i,j}(\eta)$. The associated RKHS is denoted as $\mathscr{H}_\eta^p$. It is easy to see that $g_\eta^\star(\cdot) \in \mathscr{H}_\eta^p$ and $\left\| g_\eta^\star(\cdot) \right\|_{\mathscr{H}_\eta^p}^2 = \hat{\mathbf{g}}^\top(\eta) \Sigma^{-1}(\eta) \hat{\mathbf{g}}(\eta)$. Note the reproducing property of the RKHS $g_\eta^\star(x) = \langle g_\eta^\star(\cdot), k_\eta^p(\cdot, x) \rangle_{\mathscr{H}_\eta^p}$, where $\langle \cdot, \cdot \rangle_{\mathscr{H}_\eta^p}$ denotes the inner product in $\mathscr{H}_\eta^p$. From the Cauchy–Schwarz inequality, we have $\left| g_\eta^\star(x) \right| \le k_\eta^p(x, x)^{\frac{1}{2}} \left\| g_\eta^\star(\cdot) \right\|_{\mathscr{H}_\eta^p}$. This leads to

$$\begin{aligned} |\hat{g}_l(\eta)|^2 &\le \Sigma_{l,l}(\eta) \hat{\mathbf{g}}^\top(\eta) \Sigma^{-1}(\eta) \hat{\mathbf{g}}(\eta) \\ &= \frac{1}{\sigma^2} \Sigma_{l,l}(\eta) \mathbf{y}^\top \Phi \left(\Phi^\top \Phi + \sigma^2 K^{-1}(\eta) \right)^{-1} \Phi^\top \mathbf{y} \quad (20) \\ &\le \Sigma_{l,l}(\eta) \|\mathbf{y}\|_S^2 / \sigma^2. \end{aligned}$$

Since $\hat{\eta} \in [\eta_1, \eta_2]$, we have $\Sigma_{l,l}(\hat{\eta}) \le \sigma_l^2$. This leads to $|\hat{g}_l(\hat{\eta})|^2 \le \frac{\sigma_l^2}{\sigma^2} \|\mathbf{y}\|_S^2$ and $|\hat{g}_l(\eta_0)|^2 \overset{1-\delta'}{\le} \frac{\sigma_l^2}{\sigma^2} \|\mathbf{y}\|_S^2$. Then,

$$|\hat{g}_l(\hat{\eta}) - \hat{g}_l(\eta_0)| \le |\hat{g}_l(\hat{\eta})| + |\hat{g}_l(\eta_0)| \overset{1-\delta'}{\le} \frac{2\sigma_l}{\sigma} \|\mathbf{y}\|_S. \quad (21)$$

From (13), (15), (16), we have $\mu_\delta \sqrt{\Sigma_{l,l}(\eta_0)} \overset{1-\delta'}{\le} \mu_\delta \sigma_l$. This, together with (19) and (21), proves Theorem 1. ■

Remark 1: For DI and TC kernels, by modifying the last inequality in (20), the bound in Theorem 1 can be tightened by choosing $S = \Phi \left(\Phi^\top \Phi + \sigma^2 \left(\frac{\lambda_1}{\lambda_2} \right)^\gamma K^{-1}(\eta_2) \right)^{-1} \Phi^\top$.

Remark 2: Theorem 1 still holds when more hyperparameters are involved with minor modifications to Lemma 1 if needed. So the proposed approach can be extended to consider unknown noise levels and ARX models with an additional kernel on the autoregressive output terms.

Remark 3: Although Theorem 1 improves the bounds in [14], the constant $\bar{\mu}$ is still quite conservative, mainly due to the triangle equality in (21). Such conservativeness is often observed in GP error bounds, so a much smaller scaling factor is often selected in practical applications [14], [23], [24], despite that this invalidates the theoretical guarantees. As will be seen in Section IV, $\bar{\mu} = \mu_\delta$ is used in numerical simulation.

D. Selecting the Set of Hyperparameters

Theorem 1 holds for any choices of η_1, η_2 that satisfy (13) and $\hat{\eta} \in [\eta_1, \eta_2]$. To obtain the tightest bound, η_1, η_2 can be selected by optimization. For DI and TC kernels, the total

magnitude of the bounds $\sum_{l=0}^{n_g-1} \bar{\mu}\sigma_l$ can be minimized. From (14) and (15), this is equivalent to solving

$$\min_{\eta_1,\eta_2} \quad \left(\frac{\lambda_2}{\lambda_1}\right)^{\gamma} \mathrm{tr}\,(K(\eta_2)) \tag{22a}$$

$$\text{s.t.} \quad \frac{\int_{\eta\in[\eta_1,\eta_2]} p(\mathbf{y}|\mathbf{u},\eta)p(\eta)\,\mathrm{d}\eta}{\int_{\eta\in\mathbb{H}} p(\mathbf{y}|\mathbf{u},\eta)p(\eta)\,\mathrm{d}\eta} \geq 1-\delta',\ \hat{\eta}\in[\eta_1,\eta_2] \tag{22b}$$

For a general kernel structure with element-wise bound (16), η_1,η_2 can be selected individually for each l by solving the minimax problem:

$$\sigma_l^2 = \min_{\eta_1,\eta_2} \max_{\eta\in[\eta_1,\eta_2]} \Sigma_{l,l}(\eta) \quad \text{s.t. (22b).} \tag{23}$$

The algorithm to obtain the error bounds with unknown hyperparameters is summarized as follows.

1: Estimate $\hat{\eta}$ and obtain $\hat{\mathbf{g}}(\hat{\eta})$ from (7).
2: Calculate η_1,η_2 by solving (22) or (23).
3: Calculate σ_l, $l=0,\ldots,n_g-1$ from (15) or (16).
4: Obtain the elementwise error bounds from (17).

IV. Numerical Results

The proposed bound is applied numerically by considering the same examples as in Section III-A. Again, the practical scenario with fixed impulse responses is considered. The error bound (9) with estimated hyperparameters analyzed in Section III-A is termed the *vanilla kernel bound*, whereas the proposed bound in Section III-C is called the *robust kernel bound*. The *least-squares bound* (4) is also compared.

For computational efficiency, the optimization problems to find η_1,η_2 are solved by discretizing η. The nominal estimate and the estimated hyperparameters are obtained by `impulseest` in MATLAB. The inner problem in (23) is solved by `fmincon` in MATLAB. For the robust kernel bound, we select $\delta'=0.1$ and $\bar{\mu}=\mu_\delta$.

Figure 2 presents a comparison of the performance of different error bounds with a TC kernel design. For each example, the left figure shows representative identification results in one simulation, whereas the right figure shows the empirical probability of error bounds containing the true parameters from 1000 Monte Carlo simulations. The results show that the proposed robust kernel bounds are more conservative compared to the vanilla kernel bounds, especially under high noise, but they are much more reliable with much higher empirical probabilities of containing the true parameters. On the other hand, the robust kernel bounds are still much tighter than the least-squares bounds.

Figure 3 shows the empirical probability with a SS kernel design. The robust kernel bounds are derived by selecting σ_l from (23). Similar results to the TC kernel case are obtained, where the robust kernel bounds are much more reliable than the vanilla kernel bounds.

V. Conclusions

In this work, we investigate the problem of quantifying the estimation error in kernel-based linear system identification with unknown hyperparameters. First, it is illustrated that the certainty equivalence principle does not work here: error bounds constructed using the estimated hyperparameters are too optimistic in multiple examples. Instead, a rectangular set of hyperparameters is constructed to contain the true ones with high probability. The error bounds can then be obtained by scaling the worst-case posterior variances within the set. It is shown both theoretically and numerically that the proposed bound is accurate in specifying the estimation error.

This work provides a practical approach to obtaining a reliable stochastic model centered around the nominal estimate of kernel-based system identification. Further research directions include deriving uniform posterior covariance bounds for other kernel structures and improving the constant $\bar{\mu}$ in Theorem 1.

Appendix I
Proof of Lemma 1

The result is trivial for DI kernels. For TC kernels, define $M(\mathbf{m}_n)\in\mathbb{R}^{n\times n}$, $\mathbf{m}_n=[m_1\ m_2\ \ldots\ m_n]^\top$ with $M_{i,j}(\mathbf{m}_n)=m_{\max(i,j)}$. We first prove that

$$\det M(\mathbf{m}_n) = m_n\prod_{i=1}^{n-1}(m_i-m_{i+1}) \tag{24}$$

by induction, the detail of which is omitted due to space constraints.

Using Sylvester's criterion, $M(\mathbf{m}_n)$ is positive semidefinite iff $\det M(\mathbf{m}_l)\geq 0, \forall\, l=1,\ldots,n$. This requires

$$m_i-m_{i+1}\geq 0, \forall\, i=1,\ldots,n-1. \tag{25}$$

Define $\eta_2'=\left[\left(\frac{\lambda_2}{\lambda_1}\right)^{\gamma}c_1\ \lambda_2\right]^\top$. Since $\left(\frac{\lambda_2}{\lambda_1}\right)^{\gamma}c_1\leq c_2$, we have $K(\eta_2)\succcurlyeq K(\eta_2')$. Define $M(\mathbf{m}_{n_g})=K(\eta_2')-K(\eta_1)$ by choosing $m_i=\left(\frac{\lambda_2}{\lambda_1}\right)^{\gamma}c_1\lambda_2^i-c_1\lambda_1^i$. So $K(\eta_2')-K(\eta_1)\succcurlyeq 0$ is equivalent to

$$\lambda_2^{1+\gamma}-\lambda_1^{1+\gamma}\geq\lambda_2^{2+\gamma}-\lambda_1^{2+\gamma}\geq\cdots\geq\lambda_2^{n_g+\gamma}-\lambda_1^{n_g+\gamma}. \tag{26}$$

Note that $f(x)=\lambda_2^x-\lambda_1^x$ is monotonically non-increasing for $x\geq -1/\ln\lambda_2$, $\forall\,\lambda_2\geq\lambda_1$. This indicates that (26) is satisfied for $\gamma\geq -1/\ln\lambda_2-1$. Therefore, $K(\eta_2)\succcurlyeq K(\eta_2')\succcurlyeq K(\eta_1)$ for $\gamma=-1/\ln\lambda_2-1$, which leads to

$$\left(\Phi^\top\Phi+\sigma^2K^{-1}(\eta_2)\right)^{-1}\succcurlyeq\left(\Phi^\top\Phi+\sigma^2K^{-1}(\eta_1)\right)^{-1}.$$

This directly proves Lemma 1.

References

[1] L. Ljung, "Prediction error estimation methods," *Circuits, Systems, and Signal Processing*, vol. 21, no. 1, pp. 11–21, 2002.
[2] ——, *System Identification: Theory for the User*. Upper Saddle River, NJ, USA: Prentice-Hall, 1999.
[3] G. Pillonetto and G. De Nicolao, "A new kernel-based approach for linear system identification," *Automatica*, vol. 46, no. 1, pp. 81–93, 2010.
[4] A. Chiuso and G. Pillonetto, "System identification: a machine learning perspective," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 2, no. 1, pp. 281–304, 2019.
[5] L. Ljung, T. Chen, and B. Mu, "A shift in paradigm for system identification," *International Journal of Control*, vol. 93, no. 2, pp. 173–180, 2019.

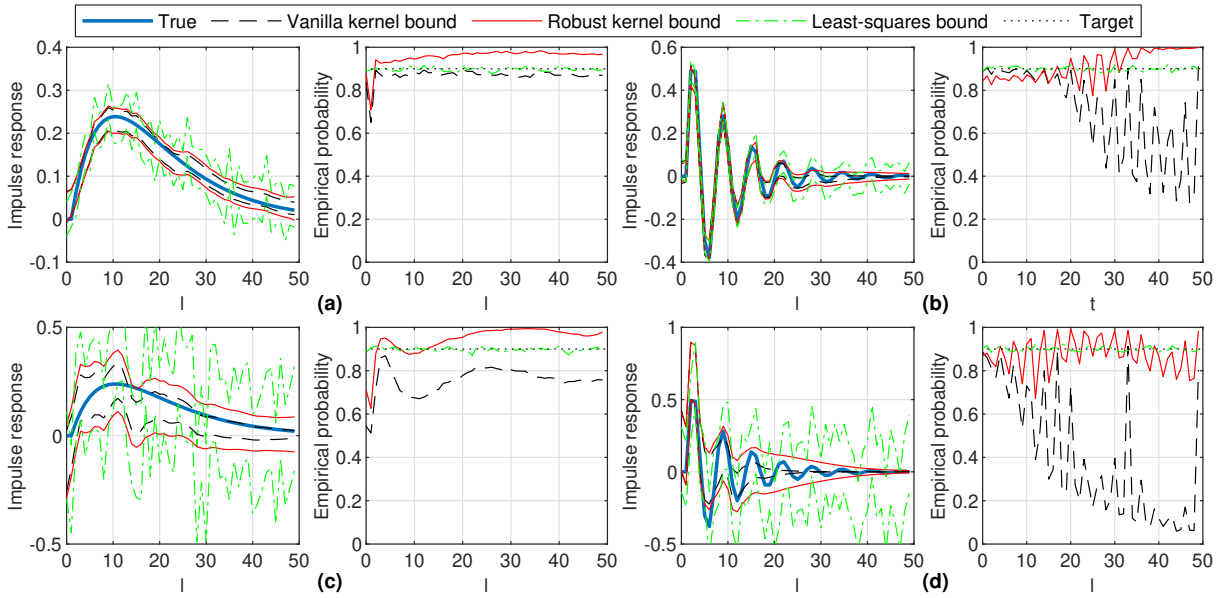

Fig. 2. Comparison of different error bounds with TC kernels. (a) $G_1, \sigma^2 = 0.1$, (b) $G_2, \sigma^2 = 0.1$, (c) $G_1, \sigma^2 = 0.5$, (d) $G_2, \sigma^2 = 0.5$. Left: representative elementwise error bounds, right: the empirical probability of error bounds containing the true parameters. l: index of the impulse response vector.

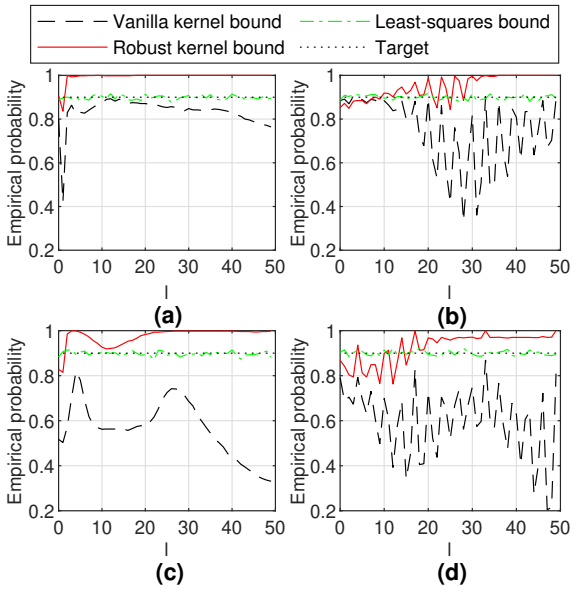

Fig. 3. Empirical probability of error bounds containing the true parameters with SS kernels. (a) $G_1, \sigma^2 = 0.1$, (b) $G_2, \sigma^2 = 0.1$, (c) $G_1, \sigma^2 = 0.5$, (d) $G_2, \sigma^2 = 0.5$. l: index of the impulse response vector.

[6] G. Pillonetto, T. Chen, A. Chiuso, G. De Nicolao, and L. Ljung, *Regularized system identification: learning dynamic models from data*. Springer, 2022.

[7] G. Pillonetto, T. Chen, A. Chiuso, G. D. Nicolao, and L. Ljung, "Regularized linear system identification using atomic, nuclear and kernel-based norms: the role of the stability constraint," *Automatica*, vol. 69, pp. 137–149, 2016.

[8] T. Chen, "On kernel design for regularized LTI system identification," *Automatica*, vol. 90, pp. 109–122, 2018.

[9] T. Chen, H. Ohlsson, and L. Ljung, "On the estimation of transfer functions, regularizations and Gaussian processes—revisited," *Automatica*, vol. 48, no. 8, pp. 1525–1535, 2012.

[10] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. The MIT Press, 2005.

[11] G. Pillonetto and A. Scampicchio, "Sample complexity and minimax properties of exponentially stable regularized estimators," *IEEE Transactions on Automatic Control*, vol. 67, no. 5, pp. 2330–2342, 2022.

[12] E. T. Maddalena, P. Scharnhorst, and C. N. Jones, "Deterministic error bounds for kernel-based learning techniques under bounded noise," *Automatica*, vol. 134, p. 109896, 2021.

[13] N. Srinivas, A. Krause, S. M. Kakade, and M. W. Seeger, "Information-theoretic regret bounds for Gaussian process optimization in the bandit setting," *IEEE Transactions on Information Theory*, vol. 58, no. 5, pp. 3250–3265, 2012.

[14] A. Capone, A. Lederer, and S. Hirche, "Gaussian process uniform error bounds with unknown hyperparameters for safety-critical applications," in *Proceedings of the 39th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 162, 2022, pp. 2609–2624.

[15] T. Beckers, J. Umlauft, and S. Hirche, "Mean square prediction error of misspecified Gaussian process models," in *IEEE Conference on Decision and Control (CDC)*, 2018, pp. 1162–1167.

[16] C. Fiedler, C. W. Scherer, and S. Trimpe, "Practical and rigorous uncertainty bounds for Gaussian process regression," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 8, pp. 7439–7447, 2021.

[17] R. Tuo and W. Wang, "Kriging prediction with isotropic Matérn correlations: Robustness and experimental designs," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 7604–7641, 2022.

[18] G. Baggio, A. Carè, A. Scampicchio, and G. Pillonetto, "Bayesian frequentist bounds for machine learning and system identification," *Automatica*, vol. 146, p. 110599, 2022.

[19] G. Pillonetto and L. Ljung, "Full bayesian identification of linear dynamic systems using stable kernels," *Proceedings of the National Academy of Sciences*, vol. 120, no. 18, 2023.

[20] S. Saitoh and Y. Sawano, *Theory of reproducing kernels and applications*. Singapore: Springer, 2016.

[21] B. Schölkopf, R. Herbrich, and A. J. Smola, "A generalized representer theorem," in *Computational Learning Theory*, 2001, pp. 416–426.

[22] M. Khosravi, M. Yin, A. Iannelli, A. Parsi, and R. S. Smith, "Low-complexity identification by sparse hyperparameter estimation," *IFAC-PapersOnLine*, vol. 53, no. 2, pp. 412–417, 2020.

[23] F. Berkenkamp, M. Turchetta, A. Schoellig, and A. Krause, "Safe model-based reinforcement learning with stability guarantees," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[24] J. Umlauft, T. Beckers, M. Kimmel, and S. Hirche, "Feedback linearization using Gaussian processes," in *IEEE 56th Annual Conference on Decision and Control (CDC)*, 2017.