ETH zürich
Automatic Control Laboratory

Error bounds for kernel-based linear system identification with unknown hyperparameters

Mingzhou Yin, Roy S. Smith

Sept 26, 2023, ERNSI 2023

System identification as a function learning problem

- Traditionally, SysID is studied as parameter estimation problems with known model structures

$$\min_{\theta \in \Theta} \quad \sum_{k=1}^{N} \|y_k - \hat{y}(k|\theta)\|_2^2 \qquad \text{(PEM)}$$

- Less accessible model structure $\Rightarrow$ non-parametric models

$$y_k = f\left(u_k, u_{k-1}, \ldots, u_{-\infty}\right) + v_k$$

- Restrict to causal linear systems

$$y_k = (g \otimes u)_k + v_k, \quad g_k = 0, \forall\, k < 0, \quad \otimes : \text{discrete convolution}$$

- $\Rightarrow$ a function learning problem of g_k

More concretely...

- Consider finite impulse response model

$$G(q) = \sum_{l=0}^{n_g-1} g_l q^{-l}, \quad y_k = \sum_{l=0}^{n_g-1} g_l u_{k-l} + v_k$$

- Formulate data equation with collected input-output data

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

Kernel-based system identification

- One can try least-squares

$$\hat{\mathbf{g}}^{\mathsf{LS}} = \operatorname*{argmin}_{\mathbf{g}} \|\mathbf{y} - \Phi\,\mathbf{g}\|_2^2 = \left(\Phi^\top\Phi\right)^{-1}\Phi^\top\mathbf{y},\ \text{with covariance } \Sigma^{\mathsf{LS}} = \sigma^2\left(\Phi^\top\Phi\right)^{-1}$$

- ...but usually leads to overfitting — too many unknowns

- The regularized version can be more effective

$$\hat{\mathbf{g}} = \operatorname*{argmin}_{\mathbf{g}} \|\mathbf{y} - \Phi\,\mathbf{g}\|_2^2 + \sigma^2\,\mathbf{g}^\top K^{-1}\mathbf{g} = \left(\Phi^\top\Phi + \sigma^2 K^{-1}\right)^{-1}\Phi^\top\mathbf{y}$$

- ...by inducing prior assumptions on $\mathbf{g}$

Threefold interpretation

- **Ridge regression with basis expansion**: $\mathbf{g} = \sum_{i=1}^{n_b} \alpha_i \mathbf{g}_i = G\alpha$

$$\min_{\alpha} \; \|\mathbf{y} - \Phi G\alpha\|_2^2 + \sigma^2 \|\alpha\|_2^2 \, , \quad G^\top K^{-1} G = \mathbb{I}$$

- **Gaussian process**: Gaussian random design of $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, K)$

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

Posterior distribution: $\mathbf{g}|\mathbf{y} \sim \mathcal{N}\left(\hat{\mathbf{g}}, \Sigma\right)$, $\Sigma = \sigma^2 \left(\Phi^\top \Phi + \sigma^2 K^{-1}\right)^{-1}$

Threefold interpretation

- **Reproducing kernel Hilbert space**: $\mathbf{g}$ is sampled from CT function $g(t) \in \mathcal{H}\left(k(\cdot,\cdot)\right)$

$$g^\star(\cdot) = \underset{g(\cdot)\in\mathcal{H}}{\text{arg min}} \;\; \|\mathbf{y} - \Phi\,\mathbf{g}\|_2^2 + \sigma^2\,\|g(\cdot)\|_{\mathcal{H}}^2$$

$$\text{s.t.}\;\; \mathbf{g} = \begin{bmatrix} g(0) & \dots & g(n_g - 1) \end{bmatrix}^\top,$$

- Representer theorem: $g^\star(x) = \mathbf{k}_x \left(\Phi^\top \Phi K + \sigma^2 \mathbb{I}\right)^{-1} \Phi^\top \mathbf{y} \implies \mathbf{g}^\star = \hat{\mathbf{g}}$

$$K_{l,l} = k(l,l), \quad \mathbf{k}_x = [k(x,0) \;\dots\; k(x,n_g-1)]$$

- Induced norm: $\|g^\star(\cdot)\|_{\mathcal{H}}^2 = \hat{\mathbf{g}}^\top K^{-1} \hat{\mathbf{g}}$

How to choose K?

Extensively studied, the common approach:

- Stable kernel structure:

$$K^{\mathsf{DI}}_{i,i}(\eta) = c\lambda^i, \qquad K^{\mathsf{DI}}_{i,j}(\eta) = 0,\ i \neq j \qquad \text{(diagonal)}$$
$$K^{\mathsf{TC}}_{i,j}(\eta) = c\lambda^{\max(i,j)} \qquad \text{(tuned/correlated)}$$
$$K^{\mathsf{SS}}_{i,j}(\eta) = c\lambda^{2\max(i,j)}\left(\frac{\lambda^{\min(i,j)}}{2} - \frac{\lambda^{\max(i,j)}}{6}\right) \qquad \text{(stable spline)}$$

- Maximum marginal likelihood to estimate hyperparameters η:

$$\hat{\eta} = \underset{\eta}{\text{argmin}} \ -\log p(\mathbf{y}|\mathbf{u},\eta)$$

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

- Certainty equivalence: $\hat{\eta} \to \eta$

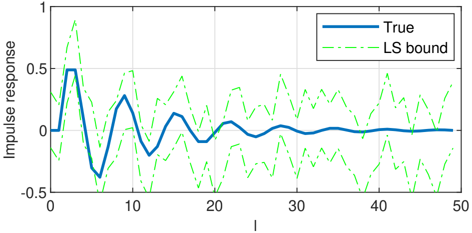
Error bound quantification

- For fixed design of $\mathbf{g}$, LS gives unbiased estimator with minimum variance for i.i.d. Gaussian output noise
- Stochastic high-probability error bounds

$$\mathbb{P}\left(\left|\hat{g}_l^{\mathsf{LS}} - g_l\right| \leq \mu_\delta \sqrt{\Sigma_{l,l}^{\mathsf{LS}}}\right) \geq 1-\delta, \quad F_{\mathcal{N}}(\mu_\delta) \geq 1-\delta/2$$

- Still conservative due to overfitting

$$G_2(q) = \frac{0.0616}{q^2 - q + 0.9^2}, \quad \sigma^2 = 0.5$$



Towards better error bounds

- Hope with random design of $\mathbf{g}$: one of the main advantages of GP interpretation
- If $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, K(\hat{\eta}))$, stochastic bounds associated with posterior covariance

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

- Improvement is guaranteed

$$\Sigma = \sigma^2 \left(\Phi^\top \Phi + \sigma^2 K^{-1}\right)^{-1} \preccurlyeq \sigma^2 \left(\Phi^\top \Phi\right)^{-1} = \Sigma^{\mathsf{LS}}$$

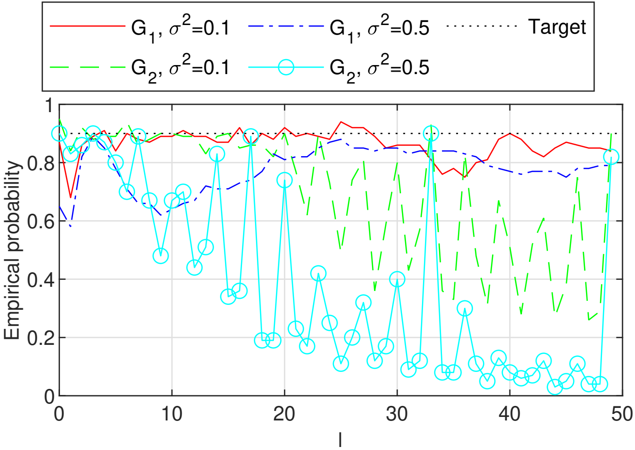
Are the bounds reliable?

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

$$G_2(q) = \frac{0.0616}{q^2 - q + 0.9^2}$$

Target prob.: $1 - \delta = 0.9$

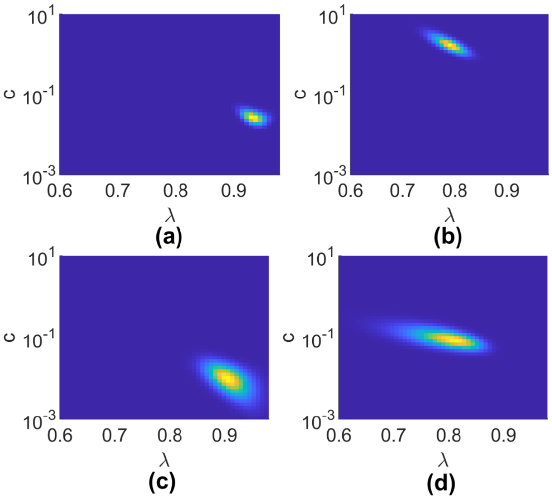
Too optimistic for lightly damped systems and low signal-to-noise ratio



What's the reason behind

- Certainty equivalence: $\hat{\eta} \to \eta$
- …but is it valid?

- Indirect evidence: how localized is the marginal likelihood function?
- $\hat{\eta}$ can be rather inaccurate in (b), (c), (d)



(a) $G_1, \sigma^2 = 0.1$, (b) $G_2, \sigma^2 = 0.1$, (c) $G_1, \sigma^2 = 0.5$, (d) $G_2, \sigma^2 = 0.5$

Toward more reliable error bounds

- Be more conservative in estimating η
- Instead of using the maximum likelihood point $\hat{\eta}$, establishing a high-probability set for η_0

- Assume a hyperprior of η: $p(\eta)$ (uniform distribution if no prior knowledge)

$$\text{Posterior dist. of } \eta\text{: } p(\eta|\mathbf{u},\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{u},\eta)p(\eta)}{\int_{\eta\in\mathbb{H}} p(\mathbf{y}|\mathbf{u},\eta)p(\eta)\,\mathsf{d}\eta}$$

$$\text{High-probability set: } \mathbb{P}\left(\eta_0 \in [\eta_1,\eta_2]\right) = \frac{\int_{\eta\in[\eta_1,\eta_2]} p(\mathbf{y}|\mathbf{u},\eta)p(\eta)\,\mathsf{d}\eta}{\int_{\eta\in\mathbb{H}} p(\mathbf{y}|\mathbf{u},\eta)p(\eta)\,\mathsf{d}\eta} \geq 1-\delta'$$

- $\Longrightarrow$ Bounds robust to the whole set

Worst-case posterior covariance

- For general kernels, direct (non-convex) optimization for the worst case

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

- For DI & TC kernels, analytical results available

Lemma: Uniform worst-case covariance

The posterior covariance with true hyperparameters η_0 can be bounded by

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

where $\gamma = 0$ for DI kernels and $\gamma = -1/\ln \lambda_2 - 1$ for TC kernels.

Select the 'best' high-probability set

- DoF in choosing η_1, η_2 — only a feasibility problem

$$\mathbb{P}\left(\eta_0 \in [\eta_1, \eta_2]\right) = \frac{\int_{\eta \in [\eta_1, \eta_2]} p(\mathbf{y}|\mathbf{u}, \eta) p(\eta) \, \mathsf{d}\eta}{\int_{\eta \in \mathbb{H}} p(\mathbf{y}|\mathbf{u}, \eta) p(\eta) \, \mathsf{d}\eta} \geq 1 - \delta' \qquad (\star)$$

- Select η_1, η_2 that minimizes worst-case covariance $\Rightarrow$ minimax problem

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

- For DI & TC kernels, minimize the sum of uniform worst-case variances

$$\min_{\eta_1, \eta_2} \sum_{l=0}^{n_g - 1} \sigma_l = \mathsf{tr}(\bar{\Sigma}) \iff \min_{\eta_1, \eta_2} \left(\frac{\lambda_2}{\lambda_1}\right)^{\gamma} \mathsf{tr}\left(K(\eta_2)\right) \quad \text{s.t. } (\star)$$

From worst-case covariance to stochastic bounds

Theorem: Stochastic error bounds

The regularized estimate $\hat{\mathbf{g}}$ admits stochastic error bounds:

$$\mathbb{P}\left(|\hat{g}_l(\hat{\eta}) - g_l| \leq \bar{\mu}\sigma_l\right) \geq (1-\delta)(1-\delta'), \quad (1)$$

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

Proof sketch: decompose the error

$$|\hat{g}_l(\hat{\eta}) - g_l| \leq \underbrace{|\hat{g}_l(\hat{\eta}) - \hat{g}_l(\eta_0)|}_{\text{error in nominal estimate}} + \underbrace{|\hat{g}_l(\eta_0) - g_l|}_{\text{error with true hyperparam.}}$$

For $|\hat{g}_l(\eta_0) - g_l|$, we have bounded the worse-case covariance for η_0

$$|\hat{g}_l(\hat{\eta}) - \hat{g}_l(\eta_0)| \overset{1-\delta}{\leq} \mu_\delta\sqrt{\Sigma_{l,l}(\eta_0)} \overset{(1-\delta)(1-\delta')}{\leq} \mu_\delta\sigma_l$$

Still conservative...

- For $|\hat{g}_l(\hat{\eta}) - \hat{g}_l(\eta_0)|$, no good bound yet...
- ...a conservative bound: $|\hat{g}_l(\hat{\eta}) - \hat{g}_l(\eta_0)| \leq |\hat{g}_l(\hat{\eta})| + |\hat{g}_l(\eta_0)|$
- From RKHS theory,

$$|g^\star(l)| \leq k^p(l,l)^{\frac{1}{2}} \left\| g^\star(\cdot) \right\|_{\mathcal{H}^p} \leq \cdots \leq \Sigma_{l,l} \left\| \mathbf{y} \right\|_S^2 / \sigma^2$$

$k^p(x,x)$: posterior kernel with $k^p(i,j) = \Sigma_{i,j}$

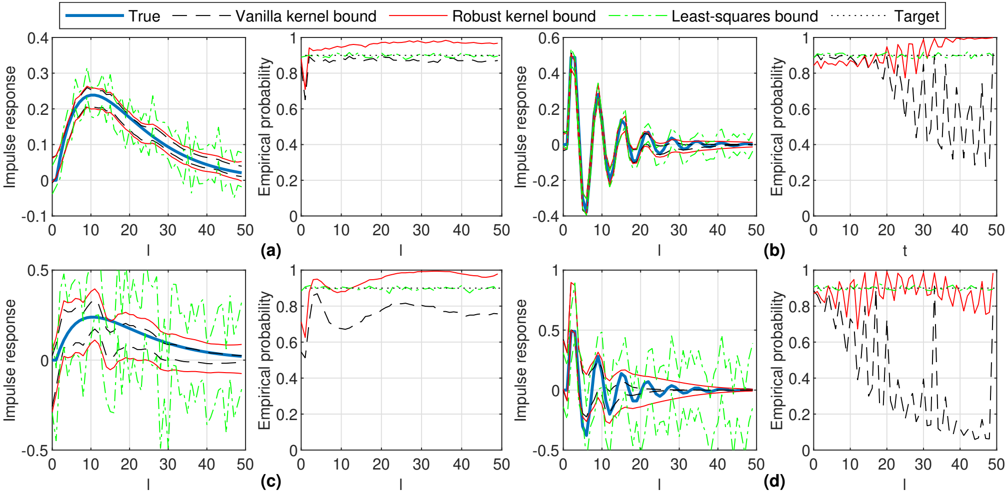
- True for all η

$$|\hat{g}_l(\hat{\eta})| + |\hat{g}_l(\eta_0)| \leq 2\Sigma_{l,l} \left\| \mathbf{y} \right\|_S^2 / \sigma^2 \overset{1-\delta'}{\leq} \frac{2\sigma_l}{\sigma} \left\| \mathbf{y} \right\|_S$$

- Better than existing work in ML[1], but still not directly usable in practice

[1] Capone, A., Lederer, A., & Hirche, S. (2022). Gaussian process uniform error bounds with unknown hyperparameters for safety-critical applications. In International Conference on Machine Learning (pp. 2609-2624).

Numerical verification



(a) $G_1, \sigma^2 = 0.1$, (b) $G_2, \sigma^2 = 0.1$, (c) $G_1, \sigma^2 = 0.5$, (d) $G_2, \sigma^2 = 0.5$

ETH zürich

Error bounds for kernel-based linear system identification with unknown hyperparameters

- Posterior covariance error bounds are not reliable by default
- . . . when hyperparameters are not easy to identify
- Construct high-probability sets for true hyperparameters
- Robust error bounds from worst-case covariance in the set

AUTOMATIC CONTROL LABORATORY

Mingzhou Yin
myin@ethz.ch