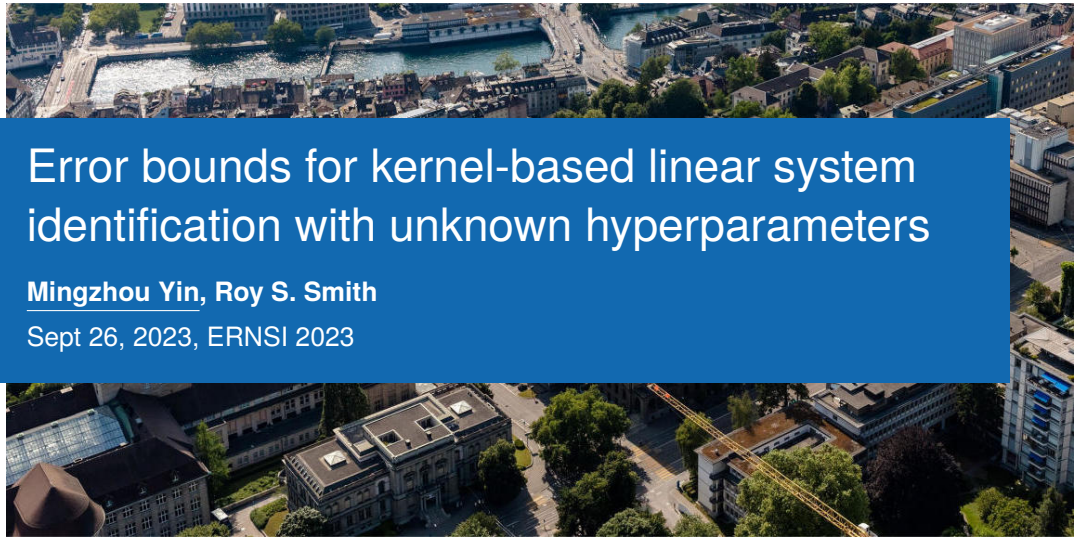
ETH zürich

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

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

- Less accessible model structure non-parametric models

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

- Restrict to causal linear systems

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

- 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

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

$\mathbf{y}$

Kernel-based system identification

- One can try least-squares

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

- . . . but usually leads to overfitting — too many unknowns
- The regularized version can be more effective

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

- . . . by inducing prior assumptions on g

Threefold interpretation

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

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

- **Gaussian process**: Gaussian random design of g $N(0, K)$

$$\begin{bmatrix} g \\ y \end{bmatrix} \quad N \quad 0, \quad \begin{bmatrix} K & K\Phi \\ \Phi K & \Phi K \Phi + \sigma^2 I \end{bmatrix}$$

Posterior distribution: $g|y \quad N(\hat{g}, \Sigma)$, $\Sigma = \sigma^2 \left[\Phi \Phi + \sigma^2 K^{-1} \right]^{-1}$

Threefold interpretation

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

$$g(\cdot) = \arg\min_{g(\cdot)\ H} \quad \mathbf{y} - \Phi\mathbf{g}\ _2^2 + \sigma^2\ g(\cdot)\ _H^2$$

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

- Representer theorem: $g(x) = \mathbf{k}_x \begin{bmatrix} \Phi \\ \Phi K + \sigma^2 \mathbf{I} \end{bmatrix}^{-1} \begin{bmatrix} \Phi \\ \mathbf{y} \end{bmatrix} = \mathbf{g} = \hat{\mathbf{g}}$

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

- Induced norm: $g(\cdot)\ _H^2 = \hat{\mathbf{g}}\ K^{-1}\hat{\mathbf{g}}$

How to choose K?

Extensively studied, the common approach:

- Stable kernel structure:

$$K^{\mathrm{DI}}_{i,i}(\eta) = c\lambda^{i}, \qquad K^{\mathrm{DI}}_{i,j}(\eta) = 0,\ i = j \qquad \text{(diagonal)}$$

$$K^{\mathrm{TC}}_{i,j}(\eta) = c\lambda^{\max(i,j)} \qquad \text{(tuned/correlated)}$$

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

- Maximum marginal likelihood to estimate hyperparameters η:

$$\hat{\eta} = \operatorname*{argmin}_{\eta} \; -\log p(\mathrm{y}|\mathrm{u},\eta)$$

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

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

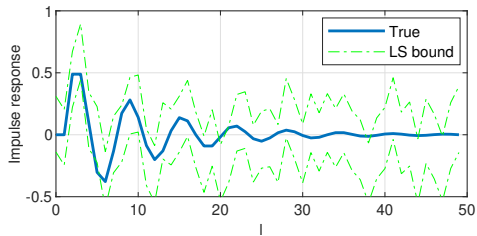
Error bound quantification

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

$$\mathrm{P}\left[\,\left|\hat{g}_l^{\mathrm{LS}} - g_l\right| \le \mu_\delta \sqrt{\Sigma_{l,l}^{\mathrm{LS}}}\,\right] \ge 1-\delta, \quad F_N(\mu_\delta) \ge 1-\delta/2$$

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

$$\mathrm{P}\left[\,|\hat{g}_l - g_l| \le \mu_\delta \sqrt{\Sigma_{l,l}}\,\right] \ge 1 - \delta,$$

- Improvement is guaranteed

$$\Sigma = \sigma^2 \left[\Phi \;\; \Phi + \sigma^2 K^{-1}\right]^{-1} \; 4 \; \sigma^2 \left[\Phi \;\; \Phi\right]^{-1} = \Sigma^{\mathrm{LS}}$$

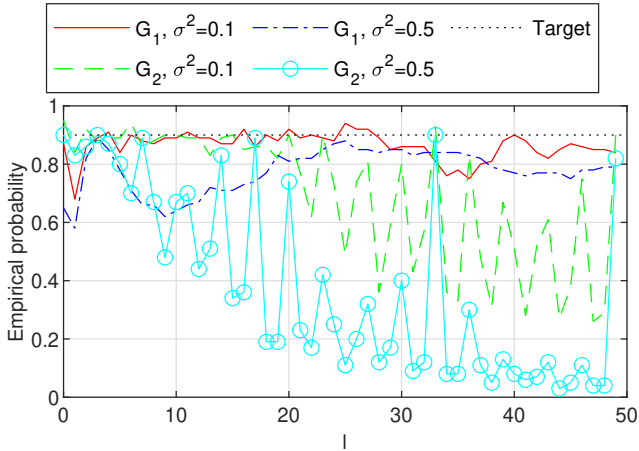
Are the bounds reliable?

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

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

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

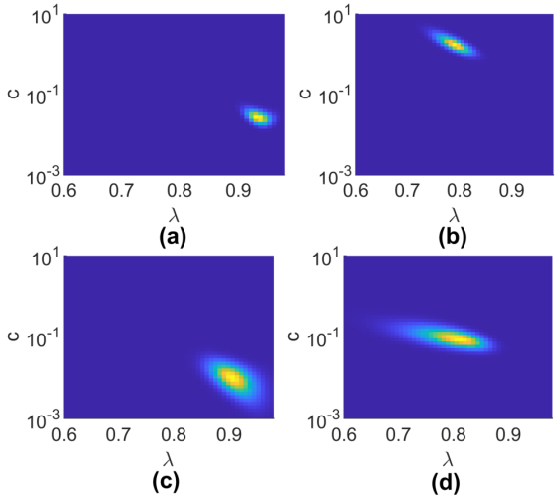
Too optimistic for lightly damped systems and low signal-to-noise ratio



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)

$$\text{Posterior dist. of } \eta\text{: } p(\eta|\mathrm{u}, \mathrm{y}) = \frac{p(\mathrm{y}|\mathrm{u}, \eta)p(\eta)}{\int_{\eta} p(\mathrm{y}|\mathrm{u}, \eta)p(\eta)\, \mathrm{d}\eta}$$

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

- = Bounds robust to the whole set

Worst-case posterior covariance

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

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

- For DI & TC kernels, analytical results available

Lemma: Uniform worst-case covariance

The posterior covariance with true hyperparameters η_0 can be bounded by

$$\Sigma(\eta_0) \leq \left(4^{1-\delta_{\gamma}} \sigma^2 \Phi \Phi + \sigma^2 \left(\frac{\lambda_1}{\lambda_2}\right)^{\gamma} K^{-1}(\eta_2)\right)^{-1} =: \bar{\Sigma}, \quad \sigma_I^2 = \bar{\Sigma}_{I,I}$$

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

$$\mathrm{P}(\eta_0 \quad [\eta_1, \eta_2]) = \frac{\int_{\eta \in [\eta_1, \eta_2]} p(\mathrm{y}|\mathrm{u}, \eta) p(\eta)\, \mathrm{d}\eta}{\int_{\eta \in \mathrm{H}} p(\mathrm{y}|\mathrm{u}, \eta) p(\eta)\, \mathrm{d}\eta} \geq 1 - \delta \qquad (\ )$$

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

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

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

$$\min_{\eta_1, \eta_2} \sum_{I=0}^{n_q - 1} \sigma_I = \mathrm{tr}(\bar{\Sigma}) \qquad \min_{\eta_1, \eta_2} \left(\frac{\lambda_2}{\lambda_1}\right)^{\gamma} \mathrm{tr}\,(K(\eta_2)) \quad \text{s.t. } (\ )$$

From worst-case covariance to stochastic bounds

Theorem: Stochastic error bounds

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

$$\mathrm{P}\left(|\hat{g}_l(\hat{\eta}) - g_l| \le \bar{\mu}\sigma_l\right) \ge (1-\delta)(1-\delta'), \tag{1}$$

where $\bar{\mu} = \mu_\delta + \frac{2}{\sigma}$ y $_S$, $S = \Phi$ Φ Φ^{-1} Φ , if $\hat{\eta}$ $[\eta_1, \eta_2]$.

Proof sketch: decompose the error

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

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

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

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,

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

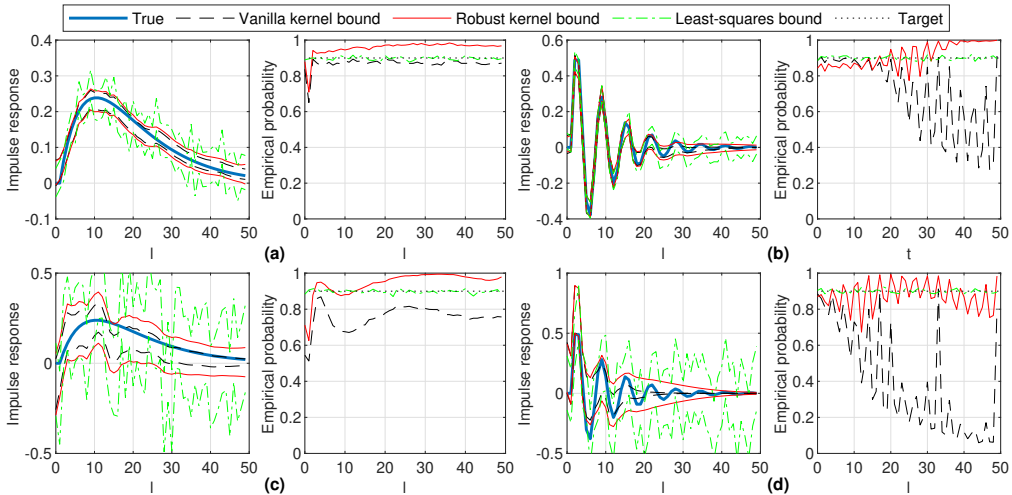
  $k^p(x,x)$: posterior kernel with $k^p(i,j) = \Sigma_{i,j}$
- True for all η

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

- 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