ETH zürich

AUTOMATIC CONTROL LABORATORY

Regularized and Nonparametric Approaches in System Identification and Data-Driven Control

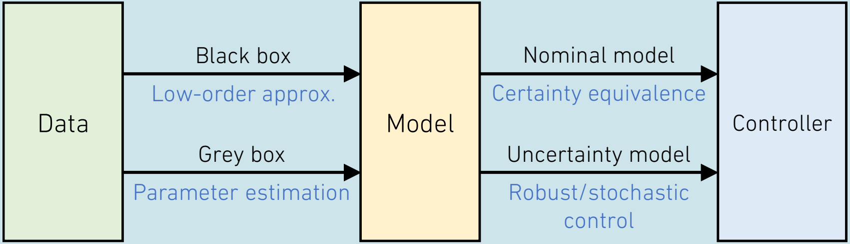
Mingzhou Yin

February 19, 2024

System identification

"classical" data-driven control

Paradigm of system identification: a two-step approach



From system identification to learning

1. Data collection
2. **Selection of model structure**
   - Compact structure
   - Parameter estimation
3. **Determining the "best" model**
   - Prediction error method
4. Model validation

- Similar framework to supervised learning
- **Motivation**: More complex systems & more data
- **Main difference**: Do we have/require a compact structure for the model?
- **Two paths**:
  1. Borrow tools from learning theories $\Rightarrow$ **regularized approaches**
  2. Accept over-parameterized models $\Rightarrow$ **nonparametric approaches**

An overview

Regularized approaches

Preserve system-theoretic properties in learning

- **Learn pole locations in sparse learning** (CDC'22), kernel learning (IFAC'20)
- Reliable uncertainty models in kernel learning (L-CSS'23)

Nonparametric approaches

Stochastic nonparametric prediction without a model

- **Stochastic prediction by MLE** (TAC'21, ECC'22), matrix denoising (SYSID'21)
- **Data-driven predictive control with stochastic predictions** (L4DC'21, SYSID'24)
- **Application to building control** (AE'24)

Periodic systems

Beyond LTI systems: making use of periodicity

- Learn linear periodic systems (IFAC'20, L-CSS'21)
- Kernel learning of nonlinear systems with periodic models (CDC'22)

Regularized approaches

with more and more parameters

Paradigm shift in system identification

Method	**Parameter estimation**	**Kernel learning**	**Sparse learning**
Theory	Classical statistics	Nonparametric statistics	High-dimensional statistics
Regime	$n \ll N$	$n \approx N$	$n \gg N$
Prior info.	Low dimension	**Smoothness**	**Sparsity**
Tool	MLE	RKHS, GP	Lasso, compressive sensing
Algorithm	Prediction error method	Kernel-based identification	Atomic norm regularization
Problem	Model structure selection	Complexity measure	Bias, false positives

Pole location estimation by sparse learning

- Sparse model decomposition: $G_0(q) = \sum_{k \in K} c_k A_k(q)$
  - $A_k(q) = \dfrac{1-|k|^2}{q-k}$: set of first-order model features
  - $K = \left\{k = \alpha \cdot \mathrm{e}^{j\beta} \,|\, \alpha \in [0,1), \beta \in [0, 2\pi)\right\}$: set of stable poles
  - $c_k \in \mathbb{C}$: **sparse** coefficients to be identified
- Simultaneous estimation of model & pole locations: $S = \{k \mid |c_k| > 0\}$
- A sparse learning problem: l_1-norm regularization

$$\min_{\{c_k\}_{k\in K}} \quad \left\| \mathbf{y} - \sum_{k\in K} c_k \, \phi_k \right\|_2^2 + \lambda \sum_{k \in K} |c_k| \,, \qquad \phi_k\text{: response of } A_k(q) \text{ under input } \mathbf{u}$$

- *but*. . . **an infinite-dimensional problem**

Observation from the optimality conditions

- **Equivalent real-valued problem**: a group lasso problem

$$\min_{\{\gamma_k\}_{k\in\hat{K}}} \left\| \mathbf{y} - \sum_{k\in\hat{K}} \zeta_k \gamma_k \right\|_2^2 + 2\lambda \sum_{k\in\hat{K}} \|\gamma_k\|_2$$

$$\gamma_k = \begin{bmatrix} \Re(c_k) & \Im(c_k) \end{bmatrix}^\top, \quad \zeta_k = \begin{bmatrix} 2\Re(\phi_k) & -2\Im(\phi_k) \end{bmatrix}, \quad \hat{K} : \text{upper unit disk}$$

- The **optimality conditions** are

$$\begin{cases} \left\|\zeta_k^\top R\right\|_2 \le \lambda, & \text{if } \|\gamma_k^\star\|_2 = 0 \\ \zeta_k^\top R + \lambda \gamma_k^\star / \|\gamma_k^\star\|_2 = 0, & \text{if } \|\gamma_k^\star\|_2 > 0 \end{cases}, \quad \overbrace{R = \mathbf{y} - \sum_{k\in\hat{K}} \zeta_k \gamma_k^\star}^{\text{output residuals}}$$

The infinite-dimensional algorithm

- For a finite-dimensional solution w.r.t. $\hat{K}_d = \{k_1, k_2, \dots, k_p\}$, if a new pole is added $\hat{K}_d^+ := \hat{K}_d \cup \{k_{p+1}\}$, the trivial solution

$$\gamma_i^\star(\hat{K}_d^+) = \begin{cases} \gamma_i^\star(\hat{K}_d), & i = 1, \dots, p, \\ \mathbf{0}, & i = p+1, \end{cases}$$

  holds iff $\left\| \zeta_{k_{p+1}}^\top R(\hat{K}_d) \right\|_2 \leq \lambda$.
- k_{p+1} is only a meaningful pole when $\left\| \zeta_{k_{p+1}}^\top R(\hat{K}_d) \right\|_2 > \lambda$
- **Greedy algorithm**: Add new pole k_{p+1} that maximizes $\left\| \zeta_{k_{p+1}}^\top R(\hat{K}_d) \right\|_2 \quad (\triangle)$

Properties:

- Optimality conditions satisfied with ϵ-tolerance
- Objective decreases every iteration even if $(\triangle)$ is not solved exactly

Numerical example

4^{th}-order system, data length $N = 100$, 100 simulations, λ selected by cross-validation

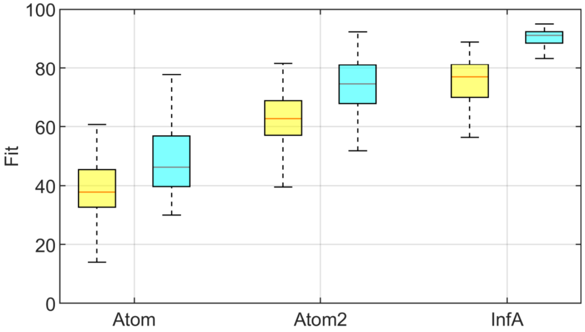
Atom: discretized solution with 50 poles

Atom2: discretized solution with 500 poles

InfA: inf-dim solution starting from 50 poles ($\sim$100 poles at convergence)

Yellow : 20 dB SNR

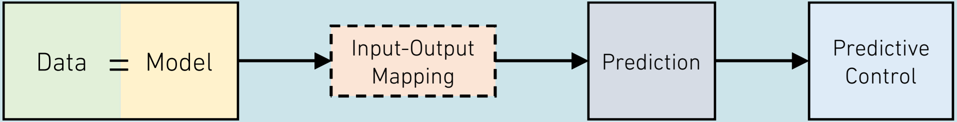
Cyan : 40 dB SNR



Nonparametric approaches

model is merely input-output mapping

Paradigm of data-driven control: prediction via input-output mapping



Idea: (**Willems' Fundamental Lemma**) For linear systems,

- Any linear combination of trajectories is still a trajectory
- If we have sufficiently 'good' data. . .
- . . . linear combinations of such data cover all possibilities

Prediction via nonparametric input-output mapping

Data collection:

$$Z = \begin{bmatrix} z_1^d & \cdots & z_M^d \end{bmatrix} \sim \text{signal matrix}$$

$$= \underbrace{\left[\begin{array}{ccc} u_{t_1}^d & \cdots & u_{t_M}^d \\ u_{t_1+1}^d & \cdots & u_{t_M+1}^d \\ \vdots & \ddots & \vdots \\ u_{t_1+L-1}^d & \cdots & u_{t_M+L-1}^d \\ \hline y_{t_1}^d & \cdots & y_{t_M}^d \\ y_{t_1+1}^d & \cdots & y_{t_M+1}^d \\ \vdots & \ddots & \vdots \\ y_{t_1+L-1}^d & \cdots & y_{t_M+L-1}^d \end{array}\right]}_{\text{columns of length-}L\text{ trajectories}} = \begin{bmatrix} U_p \\ U_f \\ Y_p \\ Y_f \end{bmatrix}$$

- If $\operatorname{rank}(Z) = n_u L + n_x$, all valid trajectory $\mathbf{z}$ can be parametrized by $g \in \mathbb{R}^M$: $\mathbf{z} = Zg$
- By fixing inputs $\mathbf{u} \in \mathbb{R}^{n_u L'}$ & initial condition $\mathbf{u}_{\text{ini}} \in \mathbb{R}^{n_u L_0}$, $\mathbf{y}_{\text{ini}} \in \mathbb{R}^{n_y L_0}$,
- ...the other outputs can be predicted:

$$\mathbf{y} = f(\mathbf{u}; \mathbf{u}_{\text{ini}}, \mathbf{y}_{\text{ini}}) : \begin{bmatrix} \mathbf{u}_{\text{ini}} \\ \mathbf{u} \\ \mathbf{y}_{\text{ini}} \\ \mathbf{y} \end{bmatrix} = \begin{bmatrix} U_p \\ U_f \\ Y_p \\ Y_f \end{bmatrix} g$$

From noise-free data to stochastic data

What if we have uncertainties?

- Z : full row rank almost surely
- $\mathbf{y}$ can be anything

$$\forall \mathbf{y} \in \mathbb{R}^{n_y L'}, \exists g : \begin{bmatrix} \mathbf{u}_{\text{ini}} \\ \mathbf{y}_{\text{ini}} \\ \mathbf{u} \\ \mathbf{y} \end{bmatrix} = \begin{bmatrix} U_p \\ Y_p \\ U_f \\ Y_f \end{bmatrix} g$$

- Ill-defined input-output mapping

Multiple paths out:

1. **Subspace identification**
2. **Direct data-driven predictive control**
3. **Indirect data-driven predictive control**: accept full-rank Z and fix one unique g

A hard "parameter estimation" problem of g

- Noise on both sides: $\mathbf{y}_{\text{ini}} = Y_p g$
- A subspace of true parameters g_0
- Error evaluated on an unknown projection $Y_f g$

The signal matrix model

a maximum likelihood approach

- Find the g that maximizes the likelihood of observing the **predicted output trajectory** $\mathbf{y}$

$$g_{\mathsf{SMM}} = \underset{g}{\mathsf{argmin}} \quad \underbrace{\mathsf{logdet}(\Sigma_y(g))}_{\text{Uncertainty of prediction}} + \underbrace{\begin{bmatrix} Y_p g - \mathbf{y}_{\mathsf{ini}} \\ \mathbf{0} \end{bmatrix}^{\mathsf{T}} \Sigma_y^{-1}(g) \begin{bmatrix} Y_p g - \mathbf{y}_{\mathsf{ini}} \\ \mathbf{0} \end{bmatrix}}_{\text{Deviation from past output measurements}}$$

$$\text{s.t.} \quad \begin{bmatrix} \mathbf{u}_{\mathsf{ini}} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} U_p \\ U_f \end{bmatrix} g$$

- $\Sigma_y(g) = (g^{\mathsf{T}} \otimes \mathbb{I})\, \mathsf{cov} \left[\mathsf{vec} \left(\begin{bmatrix} Y_p \\ Y_f \end{bmatrix} \right) \right] (g \otimes \mathbb{I}) + \begin{bmatrix} \sigma^2 \mathbb{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$
- Predictor as an online lower-level program

The signal matrix model

a stochastic predictor

- Two sources of error:

$$\mathbf{y} - \mathbf{y}_0 = \Gamma \underbrace{(\delta + \epsilon_{\text{ini}} - E_p g)}_{\text{initial condition mismatch}} + \underbrace{E_f g}_{\text{noise in } Y_f}$$

$$\Gamma = \begin{bmatrix} CA^{L_0} \\ \vdots \\ CA^{L-1} \end{bmatrix} \begin{bmatrix} C \\ \vdots \\ CA^{L_0-1} \end{bmatrix}^{\dagger}$$

$\sim$ autonomous transformation matrix from $\mathbf{y}_{\text{ini}}$ to $\mathbf{y}$

Theorem: Statistics of stochastic data-driven predictors

The stochastic predictor is given by $\mathbb{E}[\mathbf{y}] = \bar{\mathbf{y}}$, cov $(\mathbf{y}) = \Sigma$, where

$$\bar{\mathbf{y}} = Y_f g - \Gamma \left(Y_p g - \mathbf{y}_{\text{ini}} \right), \quad \Sigma = \sigma^2 \|g\|_2^2 \left(\Gamma\Gamma^\top + \mathbb{I} \right) + \Gamma \Sigma_{\text{yini}} \Gamma^\top$$

- Exact distribution requires unknown model parameter Γ
- ... but can be estimated by a data-driven approach (and assume certainty equivalence)

Data-driven predictor in stochastic predictive control

$$\min_{\mathbf{u}^t} \quad \left\|\mathbf{u}^t\right\|_R^2 + \mathbb{E}\left[\left\|\mathbf{y}^t - \mathbf{r}^t\right\|_Q^2\right]$$

$$\text{s.t.} \quad g^t = g_{\text{SMM}}(\mathbf{u}_{\text{ini}}^t, \mathbf{y}_{\text{ini}}^t, \mathbf{u}^t)$$

$$\bar{\mathbf{y}}^t = Y_f g^t - \Gamma(Y_p g^t - \mathbf{y}_{\text{ini}}^t)$$

$$\mathbb{P}\left(h_i^t \mathbf{y}^t \leq q_i^t\right) \geq p, \ \forall i$$

$$\mathbf{u}^t \in \mathcal{U}^t$$

- Lower-level program : non-convex even for i.i.d Gaussian output noise

  One-step SQP approximation: linear closed-form solution

$$g^t = \underset{g}{\text{argmin}} \ \left\|Y_p g - \bar{\mathbf{y}}_{\text{ini}}^t\right\|_2^2 + \lambda \left\|g\right\|_2^2$$

$$\text{s.t.} \ \begin{bmatrix} \mathbf{u}_{\text{ini}}^t \\ \mathbf{u}^t \end{bmatrix} = \begin{bmatrix} U_p \\ U_f \end{bmatrix} g$$

  where $\lambda = \left(L' / \left\|g_{\text{ini}}\right\|_2^2 + L\right) \sigma^2$

- Expected output cost $= \left\|\bar{\mathbf{y}}^t - \mathbf{r}^t\right\|_Q^2 + \lambda_g \left\|g^t\right\|_2^2$, $\lambda_g = \sigma^2 \text{tr}\left(Q\left(\Gamma\Gamma^\top + \mathbb{I}\right)\right) \sim$ regularization term

Data-driven predictor in stochastic predictive control

$$\min_{\mathbf{u}^t} \quad \|\mathbf{u}^t\|_R^2 + \mathbb{E}\left[\|\mathbf{y}^t - \mathbf{r}^t\|_Q^2\right]$$

$$\text{s.t.} \quad g^t = g_{\mathsf{SMM}}(\mathbf{u}_{\mathsf{ini}}^t, \mathbf{y}_{\mathsf{ini}}^t, \mathbf{u}^t)$$

$$\bar{\mathbf{y}}^t = Y_f g^t - \Gamma(Y_p g^t - \mathbf{y}_{\text{ini}}^t)$$

$$\mathbb{P}\left(h_i^t \mathbf{y}^t \le q_i^t\right) \ge p, \; \forall\, i$$

$$\mathbf{u}^t \in \mathcal{U}^t$$

Chance constraint : non-convex, error depends on inputs via g^t

Lemma: Convex surrogate of chance constraints

Chance constraints are guaranteed by SOC constraints

$$h_i^t \bar{\mathbf{y}}^t \le q_i^t - \mu\left(c_1 + c_2 \left\|g^t\right\|_2\right), \quad \forall\, i = 1$$

where

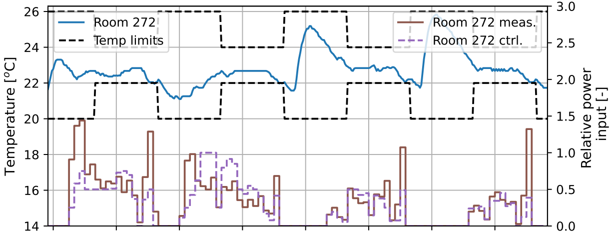
$$c_1 = \sqrt{h_i^t\, \Gamma\, \Sigma_{\mathsf{yini}} \Gamma^\top \left(h_i^t\right)^\top}$$

$$c_2 = \sigma \sqrt{h_i^t \left(\Gamma\Gamma^\top + \mathbb{I}\right)\left(h_i^t\right)^\top}, \quad \mu = \sqrt{\frac{1}{1-p} - 1}$$

Application

Space heating control

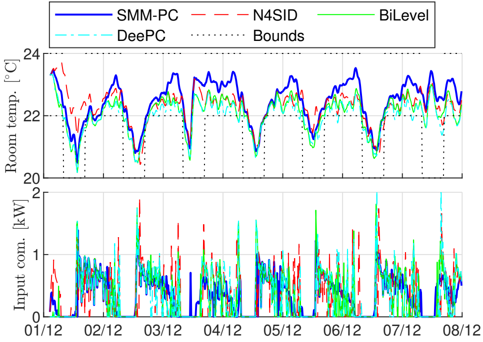
- Stochastic disturbance & measurement noise
- Nonlinearity as disturbance



- **Experiment**: 0.025°C·h constraint violation in 4 days

Benchmarking against competing algorithms

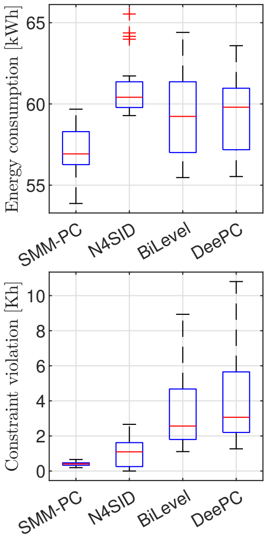
- **High-fidelity simulation**: 59% – 90% reduction in constraint violation, 4% – 8% energy saving, smoother control action



SMM-PC: proposed

N4SID: classical SysID approach

BiLevel, DeePC: existing DDPC approaches



ETH zürich

AUTOMATIC CONTROL LABORATORY

The End.

Mingzhou Yin

~~myin@control.ee.ethz.ch~~

mingzhouyin@gmail.com