ETH zürich

Automatic Control Laboratory

Data-driven prediction and control with stochastic data:

A system identification perspective

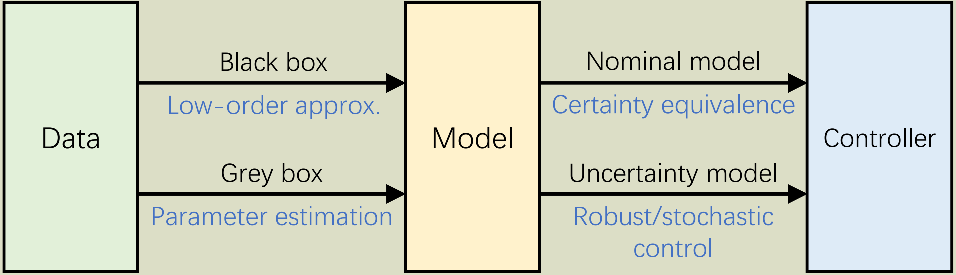
Mingzhou Yin

September 21, 2023

System identification: ‘classical’ data-driven control

- Most control applications are data-driven
- ... but were restricted by control design tools $\rightarrow$ model

Paradigm of system identification



From system identification to learning

- In practice, modeling & identification take up the majority of the budget
- **Challenge:** much more complex systems
- . . . *but*, we also have much more data

- Is it stressed enough? $\sim$ 5 sessions on identification in CDC
- . . . 20+ sessions on 'learning'

Main difference: Do we have/require a compact structure for the model?

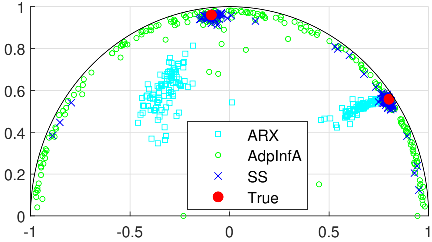
Two paths: 1. Borrow tools from learning theories

2. Accept over-parameterized models

Path 1: Preserve systems theory properties in learning

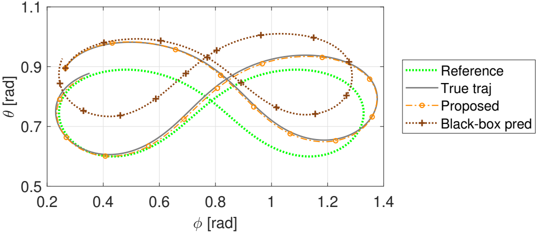
Example 1: Learn pole locations

- First-order model decomposition + sparse learning
- ... but with infinitely many features

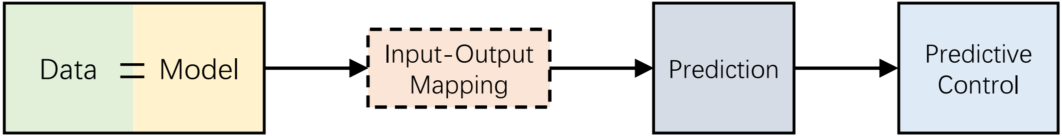


Example 2: Learn limit cycle dynamics

- Local approximation around limit cycle + kernel learning
- ... but with local convergence (stability) and known periodicity



Path 2: Model is merely input-output mapping



Idea: 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

$\Longrightarrow$ **Willems’ Fundamental Lemma**

Willems' fundamental lemma

Data:

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

$$= \underbrace{\left[\begin{array}{cccc} u_{t_1}^d & u_{t_2}^d & \cdots & u_{t_M}^d \\ u_{t_1+1}^d & u_{t_2+1}^d & \cdots & u_{t_M+1}^d \\ \vdots & \vdots & \ddots & \vdots \\ u_{t_1+L-1}^d & u_{t_2+L-1}^d & \cdots & u_{t_M+L-1}^d \\ \hline y_{t_1}^d & y_{t_2}^d & \cdots & y_{t_M}^d \\ y_{t_1+1}^d & y_{t_2+1}^d & \cdots & y_{t_M+1}^d \\ \vdots & \vdots & \ddots & \vdots \\ y_{t_1+L-1}^d & y_{t_2+L-1}^d & \cdots & y_{t_M+L-1}^d \end{array}\right]}_{\text{columns of length-}L\text{ trajectories}}$$

- *Any linear combination of trajectories is still a trajectory*

  $\forall g \in \mathbb{R}^M,\ Zg$ is a valid trajectory

- *If we have sufficiently 'good' data...*

  There are $(n_u L + n_x)$ DoF for a length-L trajectory

  If $\text{rank}(Z) = n_u L + n_x$ covers all DoF

- *... linear combinations of such data cover all possibilities*

  $\forall$ valid trajectory $\mathbf{z}, \exists\, g \in \mathbb{R}^M, \mathbf{z} = Zg$

In a world without noise...

- If we fix all DoF with inputs $\mathbf{u} \in \mathbb{R}^{n_u L'}$ & initial condition $\mathbf{u}_{\text{ini}} \in \mathbb{R}^{n_u L_0}$, $\mathbf{y}_{\text{ini}} \in \mathbb{R}^{n_y L_0}$, we can predict the other outputs
- Input-output mapping based on WFL

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

- ... is a well-defined function since $\operatorname{rank}(Z) = n_u L + n_x$, but implicit & overparametrized

Directly into predictive control

Receding horizon control at time t:

$$\min_{\mathbf{u}^t} \quad J_{\text{ctr}}\left(\mathbf{u}^t, \mathbf{y}^t\right)$$
$$\text{s.t.} \quad \begin{bmatrix} \mathbf{u}^t_{\text{ini}} \\ \mathbf{y}^t_{\text{ini}} \\ \mathbf{u}^t \end{bmatrix} = \begin{bmatrix} U_p \\ Y_p \\ U_f \end{bmatrix} g^t, \quad \mathbf{y}^t = Y_f g^t, \quad \mathbf{u}^t \in \mathcal{U}^t, \quad \mathbf{y}^t \in \mathcal{Y}^t.$$

Today's agenda

What if we have uncertainties?

- What are the paths going from noise-free data to stochastic data?
- Is there an optimal predictor we can use?
- Can we quantify the prediction error and use it to robustify the controller?
- Where is the observer in data-driven predictive control?
- Does the algorithm hold in practice with nonlinearity?

Today's agenda

- What are the paths going from noise-free data to stochastic data?
- Is there an optimal predictor we can use?
- Can we quantify the prediction error and use it to robustify the controller?
- Where is the observer in data-driven predictive control?
- Does the algorithm hold in practice with nonlinearity?

. . . until noise ruins everything

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

Three paths out:

1. **Subspace identification**: recover rank condition $\text{rank}\,(Z) = n_u L + n_x$
2. **Direct data-driven predictive control**: accept ill-defined predictor & regularize prediction in control
3. **Indirect data-driven predictive control**: accept full-rank Z and fix one unique g

The three paths

- **Subspace identification**:
  structured low-rank denoising problem

$$Z = Z_0 + \sigma E, \quad \text{rank}\,(Z_0) = n_u L + n_x,$$

$$\min_{\hat{Z}} \; \mathbb{E}\left(\left\|\hat{Z} - Z_0\right\|_F^2\right) \quad \text{s.t. } \hat{Z} \in \text{struct}(Z_0)$$

- **Direct DDPC**:

$$\min_{\mathbf{u}^t} \; J_{\text{ctr}}\left(\mathbf{u}^t, \mathbf{y}^t\right) + \underbrace{\lambda_g \left\|\Pi\, g^t\right\|_p^p}_{\text{pred. error}} + \underbrace{\lambda_y \left\|Y_p g^t - \bar{\mathbf{y}}_{\text{ini}}^t\right\|_2^2}_{\text{initial cond. mismatch}}$$

Problems

- Computationally hard
- Equivalent to SysID paradigm

- Hyperparameter tuning
- No explicit mapping (interpretability)

Indirect data-driven predictive control

- Predictor as an optimization problem with some useful g criterion

$$g^t = \underset{g}{\text{argmin}} \; \underbrace{\left\| Y_p g - \bar{\mathbf{y}}^t_{\text{ini}} \right\|_S^2}_{\text{initial cond. mismatch}} + \underbrace{\lambda \left\| g \right\|_2^2}_{\text{pred. error}} \quad \textbf{s.t.} \begin{bmatrix} \mathbf{u}^t_{\text{ini}} \\ \mathbf{u}^t \end{bmatrix} = \begin{bmatrix} U_p \\ U_f \end{bmatrix} g \qquad (\star)$$

- Predictive controller as a bi-level optimization problem

$$\min_{\mathbf{u}^t} \quad J_{\text{ctr}}\left(\mathbf{u}^t, \mathbf{y}^t\right) \quad \textbf{s.t. } (\star),\ \mathbf{y}^t = Y_f g^t,\ \mathbf{u}^t \in \mathcal{U}^t,\ \mathbf{y}^t \in \mathcal{Y}^t$$

- Explicit closed-form mapping $\sim$ signal matrix model

$$g^t = \begin{bmatrix} R_1 & R_2 & R_3 \end{bmatrix} \begin{bmatrix} \mathbf{u}^t_{\text{ini}} \\ \mathbf{u}^t \\ \bar{\mathbf{y}}^t_{\text{ini}} \end{bmatrix}, \quad \mathbf{y}^t = Y_f g^t$$

Today's agenda

- What are the paths going from noise-free data to stochastic data?
- Is there an optimal predictor we can use?
- Can we quantify the prediction error and use it to robustify the controller?
- Where is the observer in data-driven predictive control?
- Does the algorithm hold in practice with nonlinearity?

‘Optimal’ g . . . But in what sense?

- Even for very simple uncertainty: i.i.d Gaussian output noise of variance σ^2
- . . . a very special parameter estimation problem
  - Noise on both sides: $\begin{bmatrix} \mathbf{u}_{\text{ini}} \\ \mathbf{y}_{\text{ini}} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} U_p \\ Y_p \\ U_f \end{bmatrix} g$
  - Non-unique true parameter g_0 (constitute a subspace)
  - Error evaluated on an unknown projection $Y_f g$

- Many statistical tools won’t work
- **Our approach**: maximum likelihood estimation

Maximum likelihood estimation

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

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

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

- $\Sigma_y(g) = \left(g^{\mathsf{T}} \otimes \mathbb{I}\right) \text{cov}\left[\text{vec}\left(\begin{bmatrix} Y_p \\ Y_f \end{bmatrix}\right)\right] (g \otimes \mathbb{I}) + \begin{bmatrix} \sigma^2 \mathbb{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$
- Non-convex even for this simple uncertainty

A practical approximation

- Assume independent entries in Y_p, Y_f: cov $\left[\text{vec}\left(\begin{bmatrix} Y_p \\ Y_f \end{bmatrix}\right)\right] = \sigma^2 \mathbb{I}$
- One-step SQP for the MLE program is

$$g^t = \operatorname*{argmin}_g \left\| Y_p g - \bar{\mathbf{y}}^t_{\text{ini}} \right\|_2^2 + \lambda \|g\|_2^2 \quad \text{s.t.} \begin{bmatrix} \mathbf{u}^t_{\text{ini}} \\ \mathbf{u}^t \end{bmatrix} = \begin{bmatrix} U_p \\ U_f \end{bmatrix} g$$

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

- g_{ini}: initialization point, can be selected as g^{t-1} or $\begin{bmatrix} U_p \\ Y_p \\ U_f \end{bmatrix}^\dagger \begin{bmatrix} \mathbf{u}^t_{\text{ini}} \\ \mathbf{y}^t_{\text{ini}} \\ \mathbf{u}^t \end{bmatrix}$

Signal matrix model predictive control

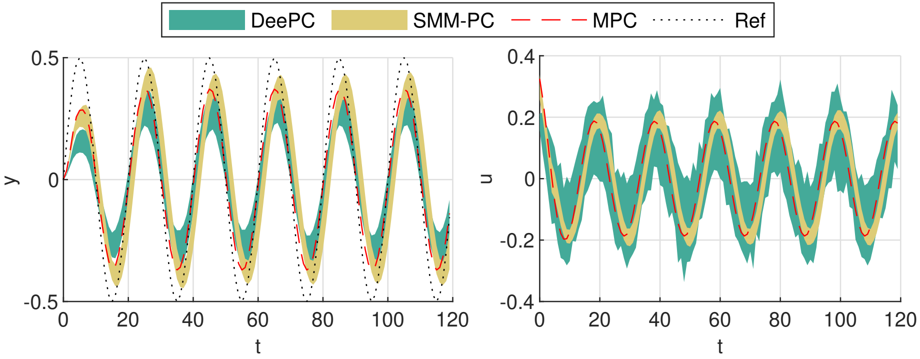


Figure: Closed-loop trajectory comparison. **DeePC**: direct DDPC with optimal tuning, **SMM-PC**: proposed, **MPC**: ideal MPC with no noise

Today's agenda

- What are the paths going from noise-free data to stochastic data?
- Is there an optimal predictor we can use?
- Can we quantify the prediction error and use it to robustify the controller?
- Where is the observer in data-driven predictive control?
- Does the algorithm hold in practice with nonlinearity?

Quantify prediction errors

- Consider the set of all reasonable predictors:

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

$$Y_p = Y_p^0 + E_p, \; Y_f = Y_f^0 + E_f, \; \mathbf{y}_{\text{ini}} = \mathbf{y}_{\text{ini}}^0 + \epsilon_{\text{ini}}$$

- 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

$$\mathbb{E}\left[\mathbf{y}\right] = \bar{\mathbf{y}},\ \text{cov}\left(\mathbf{y}\right) = \Sigma$$

where

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

$$\Sigma = \sigma^2 \left\|g\right\|_2^2 \left(\Gamma\Gamma^\top + \mathbb{I}\right) + \Gamma\, \Sigma_{\text{yini}} \Gamma^\mathsf{T}$$

- Exact distribution requires unknown model parameter Γ
- ... but can be estimated by a data-driven approach (and assume certainty equivalence)
- Linear map $\Gamma\mathbb{I}\mathrm{d} = f(\mathbf{u} = \mathbf{0}; \mathbf{u}_{\text{ini}} = \mathbf{0}, \cdot)$

Chance constraint satisfaction

- Unlike usual uncertainty assumptions, error depends on inputs via g^t
- Chance constraints $\mathbb{P}\left(h_i^t \mathbf{y}^t \leq q_i^t\right) \geq p,\ \forall\, i = 1, \dots, n_c$ $(\triangle)$ is non-convex

Lemma: Convex surrogate of chance constraints

$(\triangle)$ is guaranteed by second-order cone constraints

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

where

$$c_1 = \sqrt{h_i^t \, \Gamma \, \Sigma_{\mathsf{yini}} \Gamma^\top \left(h_i^t\right)^\top}, \quad c_2 = \sigma \sqrt{h_i^t \, \left(\Gamma\Gamma^\top + \mathbb{I}\right) \left(h_i^t\right)^\top}, \quad \mu = \sqrt{\frac{1}{1-p} - 1}$$

Stochastic version of SMM predictive control

$$\begin{aligned}
\min_{\mathbf{u}^t} \quad & \left\|\mathbf{u}^t\right\|_R^2 + \overbrace{\left\|\bar{\mathbf{y}}^t - \mathbf{r}^t\right\|_Q^2 + \lambda_g \left\|g^t\right\|_2^2}^{\text{expected output cost } \mathbb{E}\left[\left\|\mathbf{y}^t - \mathbf{r}^t\right\|_Q^2\right]} \\
\text{s.t.} \quad & g^t = \begin{bmatrix} R_1 & R_2 & R_3 \end{bmatrix} \begin{bmatrix} \mathbf{u}^t_{\text{ini}} \\ \mathbf{u}^t \\ \bar{\mathbf{y}}^t_{\text{ini}} \end{bmatrix} \\
& \bar{\mathbf{y}}^t = Y_f g^t - \Gamma(Y_p g^t - \mathbf{y}^t_{\text{ini}}) \\
& h_i^t \bar{\mathbf{y}}^t \leq q_i^t - \mu \left(c_1 + c_2 \left\|g^t\right\|_2\right), \forall\, i = 1, \dots, n_c, \\
& \mathbf{u}^t \in \mathcal{U}^t.
\end{aligned}$$

- $\lambda_g = \sigma^2 \mathrm{tr}\left(Q\left(\Gamma\Gamma^\top + \mathbb{I}\right)\right)$ resembles the regularization in direct DDPC

Beyond confidence region

- Mean-squared error can also be computed

$$\mathsf{MSE}(g,\delta) = \delta^{\mathsf{T}}\Gamma^{\mathsf{T}}\Gamma\delta + \mathsf{tr}\left(\sigma^2 \|g\|_2^2 \left(\Gamma\Gamma^\top + \mathbb{I}\right) + \Gamma\,\Sigma_{\mathsf{yini}}\Gamma^{\mathsf{T}}\right)$$

- Minimum MSE predictor

$$f(\cdot) = Y_f \operatorname*{argmin}_g \quad \mathsf{MSE}(g,\delta)$$
$$\text{s.t.} \quad \begin{bmatrix} U_p \\ U_f \\ Y_p \end{bmatrix} g = \begin{bmatrix} \mathbf{u}_{\mathsf{ini}} \\ \mathbf{u} \\ \mathbf{y}_{\mathsf{ini}} + \delta \end{bmatrix}$$

Implications:

- Characterize the optimal data-driven predictor in terms of MSE
- Propose a new data-driven predictor by replacing Γ with $\hat{\Gamma}_Z$

Today's agenda

- What are the paths going from noise-free data to stochastic data?
- Is there an optimal predictor we can use?
- Can we quantify the prediction error and use it to robustify the controller?
- Where is the observer in data-driven predictive control?
- Does the algorithm hold in practice with nonlinearity?

Towards better initial condition...

- In standard DDPC, the initial condition $\mathbf{y}_{\text{ini}}^{t}$ is directly measured
  $\Rightarrow$ constant covariance = measurement error
- In MPC, the initial condition x_t is estimated from both measurement y_t and previous prediction $x_{t|t-1}$
  $\Rightarrow$ diminishing error covariance
- **Idea:** Update $\mathbf{y}_{\text{ini}}^{t}$ with Kalman-filtered measurement from previous prediction

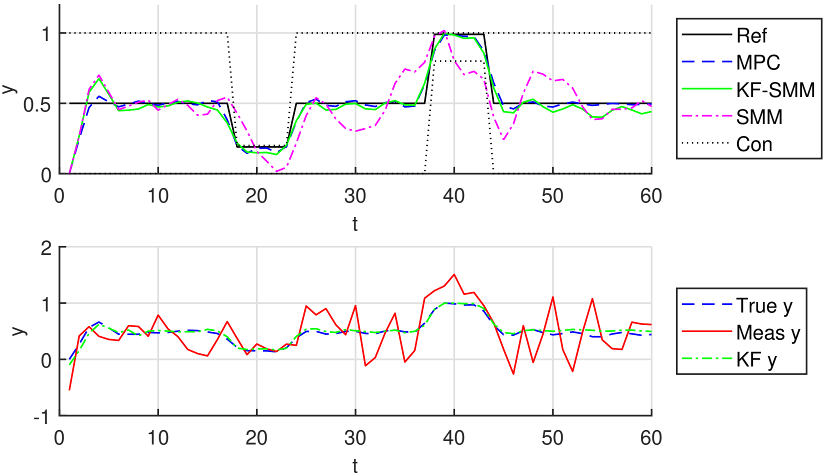
Kalman filter for data-driven input-output mapping

- Data-driven input-output mapping as a non-minimal state-space model

$$\begin{cases} \bar{x}_{t+1} & = \begin{bmatrix} \Lambda^{n_u} & \mathbf{0} \\ \mathbf{0} & \Lambda^{n_y} \end{bmatrix} \bar{x}_t + \begin{bmatrix} \mathbf{0} \\ \hat{u}_0^t \\ \mathbf{0} \\ \bar{y}_0^t \end{bmatrix} + \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ e_0^t \end{bmatrix}, \\ \zeta_{t+1} & = \begin{bmatrix} \mathbf{0} & \mathbb{I}_{n_y} \end{bmatrix} \bar{x}_{t+1} + w_t = y_t^0 + w_t = y_t \end{cases} \qquad \bar{x}_t = \begin{bmatrix} u_{t-L} \\ \vdots \\ u_{t-1} \\ y_{t-L}^0 \\ \vdots \\ y_{t-1}^0 \end{bmatrix}$$

- Λ: upper shift operator
- e_0^t: one-step-ahead prediction error with covariance $\Sigma(1,1)$
- w_t: measurement error with variance σ^2
- Standard Kalman filter design can be done

Signal matrix model predictive control (v2)



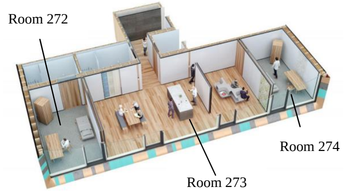
Today's agenda

- What are the paths going from noise-free data to stochastic data?
- Is there an optimal predictor we can use?
- Can we quantify the prediction error and use it to robustify the controller?
- Where is the observer in data-driven predictive control?
- Does the algorithm hold in practice with nonlinearity?

Applications: building control

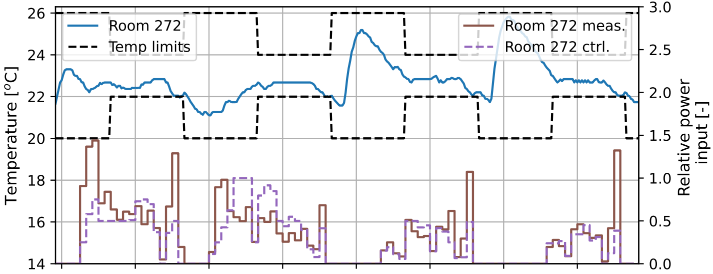
- Space heating
- Domestic hot water heating
- Stationary electric battery

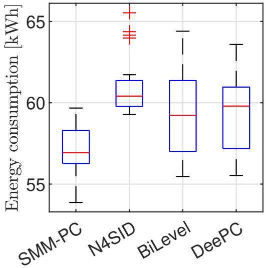
- Stochastic disturbance and measurement noise
- Nonlinearity as disturbance
- The same piece of code used with little tuning (transferability)

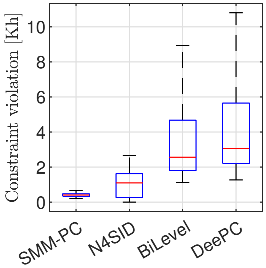


The NEST building in Dübendorf, Switzerland

Space heating





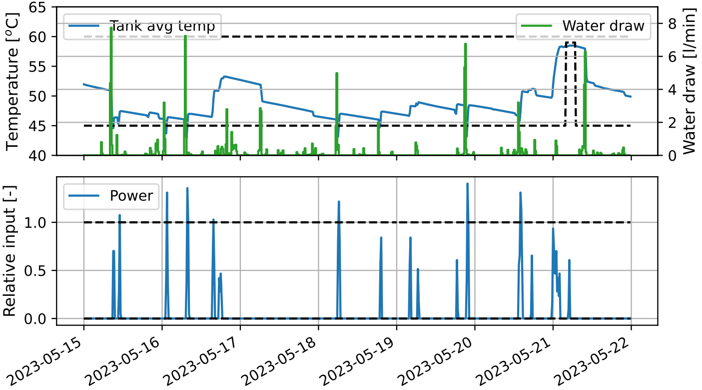


- Experiment: 0.025°C·h constraint violation in 4 days
- High-fidelity simulation: 59% – 90% reduction in constraint violation, 4% – 8% energy saving

SMM-PC: proposed, **N4SID**: subspace ID,

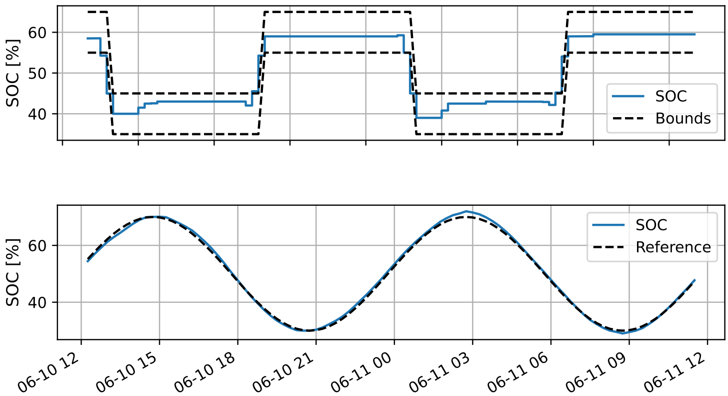
BiLevel: benchmark indirect DDPC, **DeePC**: direct DDPC

Domestic hot water heating



- Very high uncertainty due to the lack of a water draw prediction model
- Infeasible at the decontamination point, but working the most of time

Stationary electric battery



- Model-based control is also fine, but the data-driven method avoids parameter estimation for the whole life cycle

Future research directions

Bayesian perspective of behavioral systems theory

- WFL is based on binary characterization of system behaviors
- With stochastic data, you cannot falsify a trajectory completely
- Bayesian description: posterior probability of system behaviors given the data
- Unify prediction, denoising, and control

Exploration in data-driven predictive control

- Input for minimizing future prediction errors
- Bayesian optimization, upper confidence bound policy?

Future research directions

Nonlinear data-driven predictive control via Koopman operator

- WFL still valid on (inf-dim) eigenfunction space of nonlinear systems
- Learn dominant eigenfunction subspace and apply DDPC
- **Difficulties:** persistency of excitation, prediction error quantification

ETH zürich

- Optimal stochastic predictors in terms of MLE and minimum MSE
- Prediction error quantified & chance constraint satisfaction by SOCP
- Kalman filter to improve initial condition estimation
- Works in multiple building control examples

AUTOMATIC CONTROL LABORATORY

Mingzhou Yin
myin@ethz.ch