Off-Policy Reinforcement Learning for Adaptive Optimal Output Tracking of Unknown Linear Discrete-Time Systems

Ci Chen

2020. 09. 25

Content

- ▲ **Background & Motivation:** **Optimal output tracking**
- ▲ **Problem Formulation:** **Output regulation theory**
- ▲ **Method:** **Off-policy reinforcement learning for output-feedback-based optimal control**
- ▲ **Simulation**
- ▲ **Conclusion**

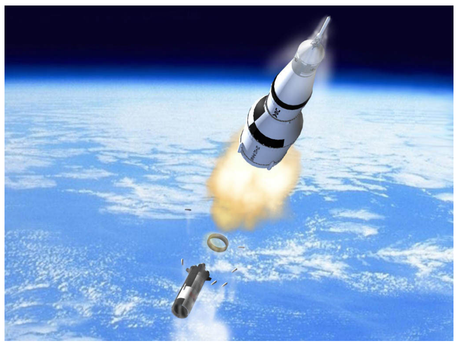
1. Background & Motivation

Challenge

How to design tracking control systems with satisfactory performance without exact model knowledge?

- **Rapid response**
- **Stability/robustness/safety guarantee**
- **Optimality for reduced fuel consumption**

Aeronautics

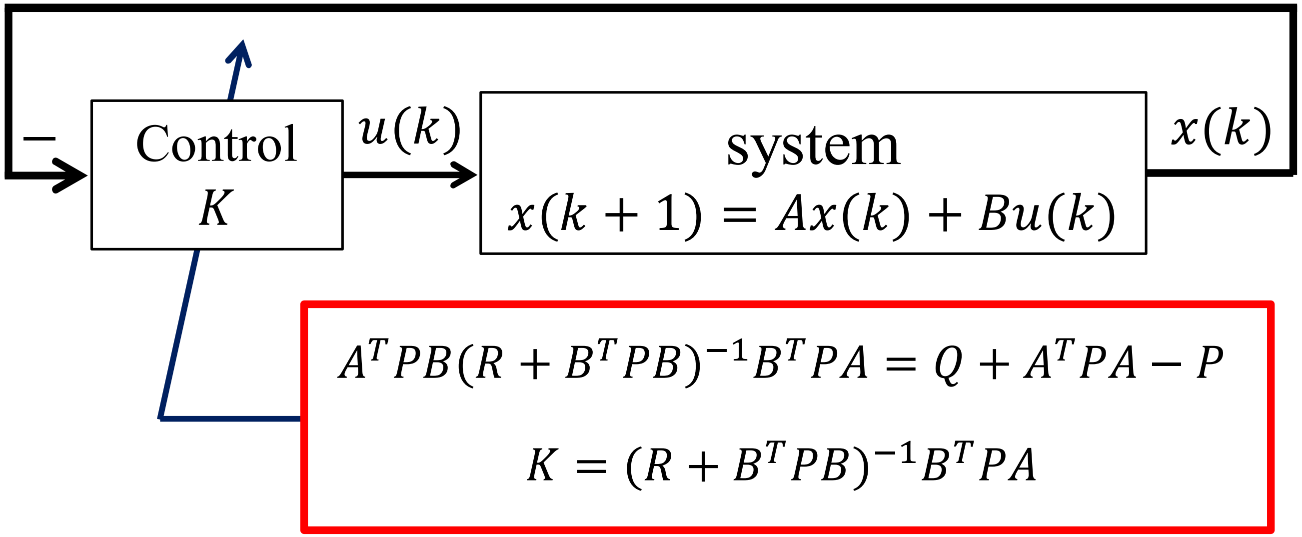


Power systems

Transportation

1. Background & Motivation

Optimal Control--The Linear Quadratic Regulator (LQR)
if full system dynamics are available.



On-line real-time control loop

Off-line design Loop

General goal

We want to find optimal control solutions

a) **Online in real-time**

b) **Using adaptive control techniques**

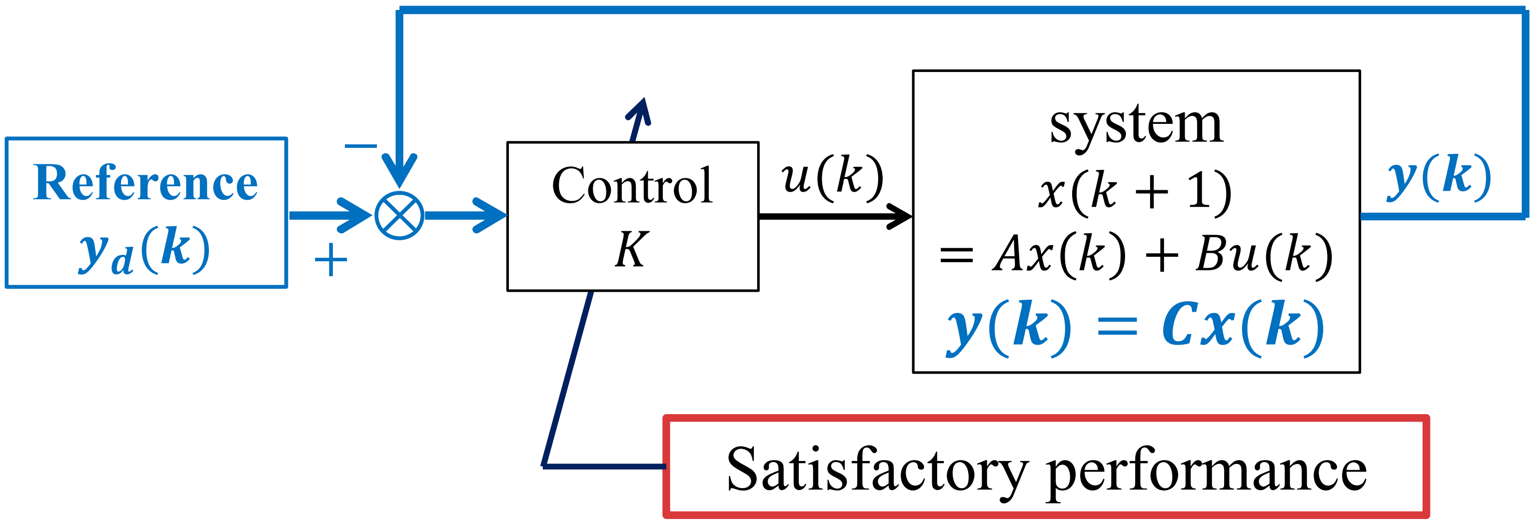
c) **Without knowing the full dynamics**

Reinforcement learning (RL)
turns out to be
the key to this goal!

1. Background & Motivation

Problem to be studied

How to achieve Optimal Output Tracking for DT systems via Output-Feedback-based Reinforcement learning?



On-line real-time control loop

On-line RL-based design Loop

Content

2. Problem Formulation (From Tracking to Regulation)

System to be controlled	Reference	Tracking error
$\boldsymbol{x(k+1) = Ax(k) + Bu(k)}$	$\boldsymbol{x_d(k+1) = Sx_d(k)}$	$\boldsymbol{y_e(k) = y(k) - y_d(k)}$
$\boldsymbol{y(k) = Cx(k)}$	$\boldsymbol{y_d(k) = Rx_d(k)}$	

$r_n \times r_m$

$r_p \times r_n$

Control design by the standard output regulation (full system dynamics)

Standard controller

$$\boldsymbol{z(k+1) = Fz(k) - Gy_e(k)}$$

$$\boldsymbol{u(k) = -Kx(k) - Hz(k)}$$

Feedback gains to be designed

Augmented system

$$\begin{cases} \boldsymbol{e(k+1) = \underline{A}e(k) + \bar{B}u_e(k)} \\ \boldsymbol{y_{e(k)} = \bar{C}e(k)} \end{cases}$$

$\boldsymbol{u_e(k) = -\bar{K}e(k)}$ with $\boldsymbol{\bar{K} = [K, H]}$

$$\underline{A} = \begin{bmatrix} A & O \\ -GC & F \end{bmatrix}$$

$$\bar{C} = \begin{bmatrix} C, & O \end{bmatrix} \quad \bar{B} = \begin{bmatrix} B \\ O \end{bmatrix}$$

Problem:

$\left(\underline{A}, \bar{C}\right)$ may not be detectable (observable).

A **fundamental** problem of detectability (observability) is resulted for output-feedback-based design.

2. Problem Formulation (From Tracking to Regulation)

System to be controlled

$$\boldsymbol{x}(k+1) = \boldsymbol{A}\boldsymbol{x}(k) + \boldsymbol{B}\boldsymbol{u}(k)$$

$$\boldsymbol{y}(k) = \boldsymbol{C}\boldsymbol{x}(k)$$

$r_n \times r_m$

$r_p \times r_n$

Reference

$$\boldsymbol{x}_d(k+1) = \boldsymbol{S}\boldsymbol{x}_d(k)$$

$$\boldsymbol{y}_d(k) = \boldsymbol{R}\boldsymbol{x}_d(k)$$

Tracking error

$$\boldsymbol{y}_e(k) = \boldsymbol{y}(k) - \boldsymbol{y}_d(k)$$

To ensure the detectability (observability), we design

Our controller

$$\boldsymbol{z}(k+1) = \boldsymbol{F}\boldsymbol{z}(k) - \boldsymbol{G}\boldsymbol{y}_e(k)$$

$$\boldsymbol{u}(k) = -\boldsymbol{K}\boldsymbol{x}(k) - \boldsymbol{H}\boldsymbol{z}(k) - \boldsymbol{T}\boldsymbol{z}(k)$$

Augmented system

$$\begin{cases} \boldsymbol{e}(k+1) = \underline{\boldsymbol{A}}\boldsymbol{e}(k) + \bar{\boldsymbol{B}}\boldsymbol{u}_e(k) \\ \boldsymbol{y}_{e(k)} = \bar{\boldsymbol{C}}\boldsymbol{e}(k) \end{cases}$$

$$\boldsymbol{u}_e(k) = -\bar{\boldsymbol{K}}\boldsymbol{e}(k) \text{ with } \bar{\boldsymbol{K}} = [\boldsymbol{K}, \boldsymbol{H}]$$

New feedforward gain Tz

$$\underline{A} = \begin{bmatrix} A & -BT \\ -GC & F \end{bmatrix}$$

$$\bar{C} = [\, C, \quad O \,] \quad \bar{B} = \begin{bmatrix} B \\ O \end{bmatrix}$$

Our Solution:

1. Given (F, T) detectable (observable) and $r_p \geq r_m$, then $(\underline{A}, \bar{C})$ is detectable (observable).

2. Given any matrix T, then $(\underline{A}, \bar{B})$ is stabilizable (controllable).

2. Problem Formulation

A tracking control design problem is now transformed into a regulation-based optimization problem.

Augmented system

$$(*)\ \begin{cases} e(k+1) = \underline{A}e(k) + \bar{B}u_e(k) \\ y_e(k) = \bar{C}e(k) \end{cases}$$

$$\boxed{\begin{array}{l} \textit{Problem} \\ \begin{cases} \min \sum_{i=k}^{\infty} \left(y_e^T(i) Q y_e(i) + u_e^T(i) \bar{R} u_e(i) \right) \\ \quad \textit{subject to } (*) \end{cases} \end{array}}$$

<u>Question:</u> How to use i/o data to learn the optimal controller that solves the optimization problem without exact model knowledge?

Content

3. Method (State-Feedback Case)

Design the behavior policy
$$\begin{cases} \bar{u}(k) = -\bar{K}^0 r(k) + \xi(k) - T\bar{z}(k) \\ \bar{z}(k+1) = F\bar{z}(k) - Gy(k) + G\vartheta(k) \end{cases}$$

Policy-iteration-based Bellman equation solver in state-feedback form

$$\begin{aligned} & r^T(k+1)P^{j+1}r(k+1) - r^T(k)P^{j+1}r(k) \\ &= -r^T(k)(\bar{Q} + (\bar{K}^j)^T \bar{R}\bar{K}^j)r(k) + \vartheta^T(k)\bar{G}^T P^{j+1}\bar{G}\vartheta(k) \\ &\quad + (-\bar{K}^j r(k) + \bar{u}(k))^T \bar{B}^T P^{j+1} \bar{B}(\bar{K}^j r(k) + \bar{u}(k)) \\ &\quad + 2\vartheta^T(k)\bar{G}^T P^{j+1}\bar{B}u(k) + 2r^T(k)\underline{A}^T P^{j+1}\bar{G}\vartheta(k) \\ &\quad + 2r^T(k)\underline{A}^T P^{j+1}\bar{B}(\bar{K}^j r(k) + \bar{u}(k)). \end{aligned}$$

$r = [x^T, \bar{z}^T]^T$

unknown

We seek to reconstruct the state using input and output data.

3. Method (System State Reconstruction)

Reconstruct the state using input and output data.

$$\zeta_{\bar{u}}(k+1) = (I_m \otimes A_\zeta)\zeta_{\bar{u}}(k) + \bar{u}(k) \otimes b,$$
$$\zeta_{y}(k+1) = (I_p \otimes A_\zeta)\zeta_{y}(k) + y(k) \otimes b,$$
$$\zeta_{\vartheta}(k+1) = (I_p \otimes A_\zeta)\zeta_{\vartheta}(k) + \vartheta(k) \otimes b,$$

$$\bar{\zeta}^T = [\zeta_u^T, \zeta_y^T, \zeta_\vartheta^T]^T$$

where $b = [0, 0, \ldots, 0, 1]^T$ and A_ζ is a companion matrix[1]

1. G. Tao, Adaptive control design and analysis. John Wiley & Sons, 2003

***Theorem* 1:** If the matrix pair $(\underline{A}, \bar{B})$ is controllable and $(\underline{A}, \bar{C})$ is observable, then the system state satisfies

$$r = \bar{M}\bar{\zeta} + (\underline{A} - \bar{L}\bar{C})^k r(0)$$

where $\boxed{\bar{M} \text{ is a full row rank matrix}}$ and $\bar{\zeta}$ is a known vector. The re-expression error $r - \bar{M}\bar{\zeta}$ converges to zero asymptotically.

3. Method (System State Reconstruction)

***Theorem* 1:** If the matrix pair $(\underline{A}, \bar{B})$ is controllable and $(\underline{A}, \bar{C})$ is observable, then the system state satisfies

$$\boxed{r = \bar{M}\bar{\zeta} + (\underline{A} - \bar{L}\bar{C})^k r(0)}$$

where $\bar{M}$ is a full row rank matrix and $\bar{\zeta}$ is a known vector. The re-expression error $r - \bar{M}\bar{\zeta}$ converges to zero asymptotically.

Policy iteration-based Bellman equation solver in state-feedback form

$$\begin{aligned}
& r^T(k+1)P^{j+1}r(k+1) - r^T(k)P^{j+1}r(k) \\
&= -r^T(k)(\bar{Q} + (\bar{K}^j)^T\bar{R}\bar{K}^j)r(k) + \vartheta^T(k)\bar{G}^T P^{j+1}\bar{G}\vartheta(k) \\
&\quad + (-\bar{K}^j r(k) + \bar{u}(k))^T \bar{B}^T P^{j+1}\bar{B}(\bar{K}^j r(k) + \bar{u}(k)) \\
&\quad + 2\vartheta^T(k)\bar{G}^T P^{j+1}\bar{B}u(k) + 2r^T(k)\underline{A}^T P^{j+1}\bar{G}\vartheta(k) \\
&\quad + 2r^T(k)\underline{A}^T P^{j+1}\bar{B}(\bar{K}^j r(k) + \bar{u}(k)).
\end{aligned}$$

$r = [x^T, \bar{z}^T]^T$

unknown

3. Method (Output-Feedback Case)

Solve the optimal control gain through output feedback

Solve

$$\varrho_o^j \begin{bmatrix} \text{vecs}(\bar{M}^T P^{j+1} \bar{M}) \\ \text{vec}(\bar{M}^T \underline{A}^T P^{j+1} \bar{B}) \\ \text{vecs}(\bar{B}^T P^{j+1} \bar{B}) \\ \text{vec}(\bar{M}^T \underline{A}^T P^{j+1} \bar{G}) \\ \text{vec}(\bar{G}^T P^{j+1} \bar{B}) \\ \text{vecs}(\bar{G}^T P^{j+1} \bar{G}) \end{bmatrix} = \nu_o^j + \boxed{\mathcal{D}_{\bar{\chi}^{j+1}}},$$

Collect the input-output data over $[k_i, k_{i+1}]$, $i=0,\ldots,f$

$$\mathcal{C}_r = [\text{vecv}(r(k_1)) - \text{vecv}(r(k_0)), \cdots, \text{vecv}(r(k_f)) - \text{vecv}(r(k_{f-1}))]^T$$

$$\mathcal{D}_{\bar{K}^j r} = [\text{vecv}(\bar{K}^j r(k_0)), \text{vecv}(\bar{K}^j r(k_1)), \cdots, \text{vecv}(\bar{K}^j r(k_{f-1}))]^T$$

$$\mathcal{D}_{\bar{u}} = [\text{vecv}(\bar{u}(k_0)), \text{vecv}(\bar{u}(k_1)), \cdots, \text{vecv}(\bar{u}(k_{f-1}))]^T$$

$$\mathcal{D}_r = [\text{vecv}(r(k_0)), \text{vecv}(r(k_1)), \cdots, \text{vecv}(r(k_{f-1}))]^T$$

$$\mathcal{D}_\vartheta = [\text{vecv}(\vartheta(k_0)), \text{vecv}(\vartheta(k_1)), \cdots, \text{vecv}(\vartheta(k_{f-1}))]^T$$

$$\mathcal{D}_{\vartheta r} = [\vartheta(k_0) \otimes r(k_0), \vartheta(k_1) \otimes r(k_1), \cdots, \vartheta(k_{f-1}) \otimes r(k_{f-1})]^T$$

$$\mathcal{D}_{rr} = [r(k_0) \otimes r(k_0), r(k_1) \otimes r(k_1), \cdots, r(k_{f-1}) \otimes r(k_{f-1})]^T$$

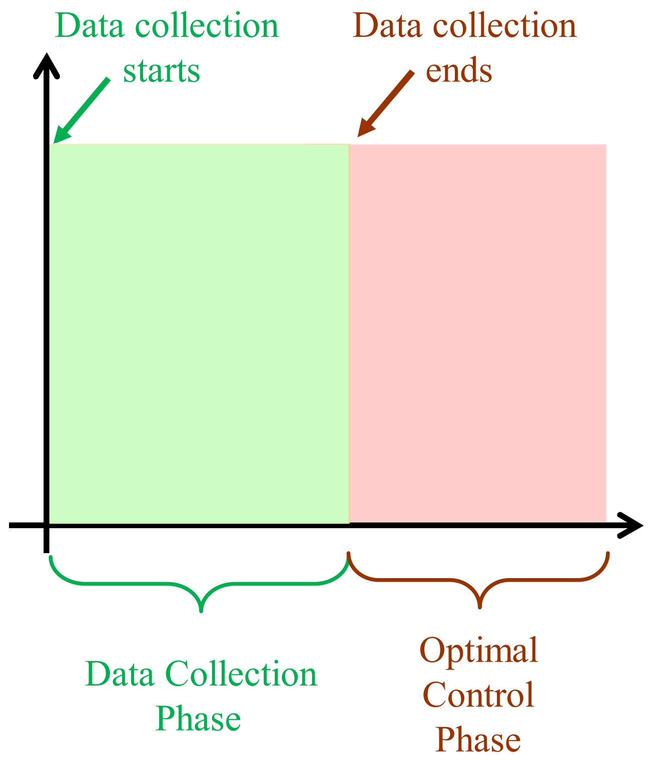


Fig. 1 data processing for off-policy learning

3. Method (Learning From Input-Output Data)

Solve the optimal control gain through output feedback

Solve

$$\varrho_o^j \begin{bmatrix} \mathrm{vecs}(\bar{M}^T P^{j+1} \bar{M}) \\ \mathrm{vec}(\bar{M}^T \underline{A}^T P^{j+1} \bar{B}) \\ \mathrm{vecs}(\bar{B}^T P^{j+1} \bar{B}) \\ \mathrm{vec}(\bar{M}^T \underline{A}^T P^{j+1} \bar{G}) \\ \mathrm{vec}(\bar{G}^T P^{j+1} \bar{B}) \\ \mathrm{vecs}(\bar{G}^T P^{j+1} \bar{G}) \end{bmatrix} = \nu_o^j + \boxed{\mathcal{D}_{\bar{\chi}^{j+1}}},$$

Verifiable criterion for checking how much data.

$$\mathrm{rank}([\mathcal{D}_{\bar{\zeta}\bar{\zeta}}, \mathcal{D}_{\bar{u}\bar{\zeta}}, \mathcal{D}_{\bar{u}}, \mathcal{D}_{\vartheta\bar{\zeta}}, \mathcal{D}_{\bar{u}\vartheta}, \mathcal{D}_{\vartheta}]) = \frac{1}{2}(n_{\bar{\zeta}}(n_{\bar{\zeta}}+1) + r_m(r_m+1) + r_p(r_p+1)) + n_{\bar{\zeta}} r_m + r_p n_{\bar{\zeta}} + r_p r_m$$

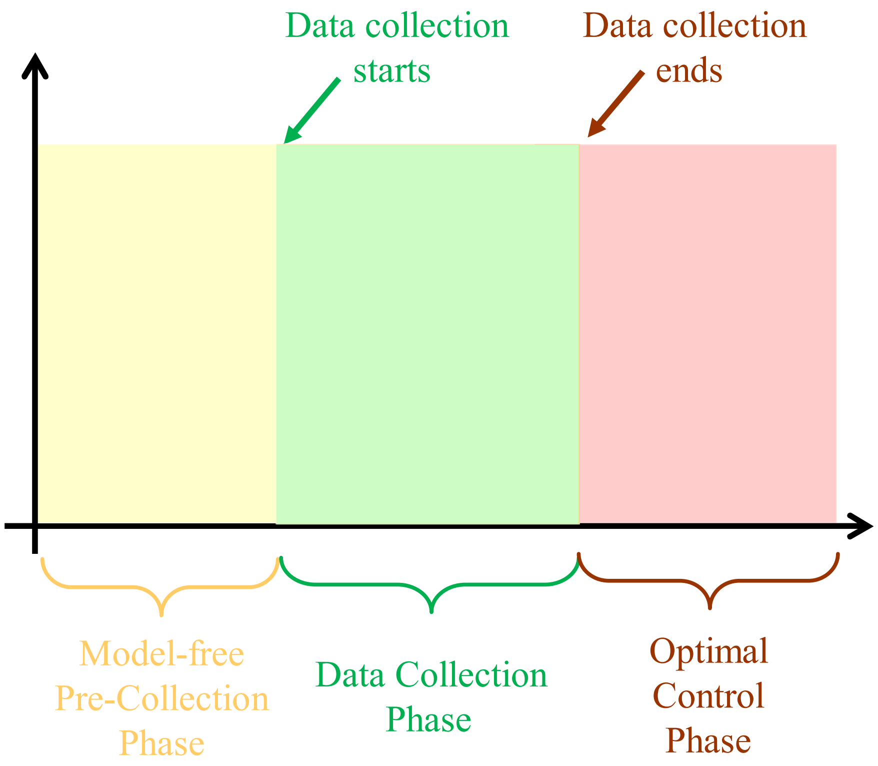


Fig. 2 The **proposed** data processing for off-policy learning

The optimal control gain $\left(\bar{R} + \bar{B}^T P^{j+1} \bar{B}\right)^{-1} \left(\bar{M}^T \underline{A}^T P^{j+1} \bar{B}\right)^T$ is uniquely learned from input-output data.

Content

4. Simulation

F-16 aircraft dynamics after the discretization

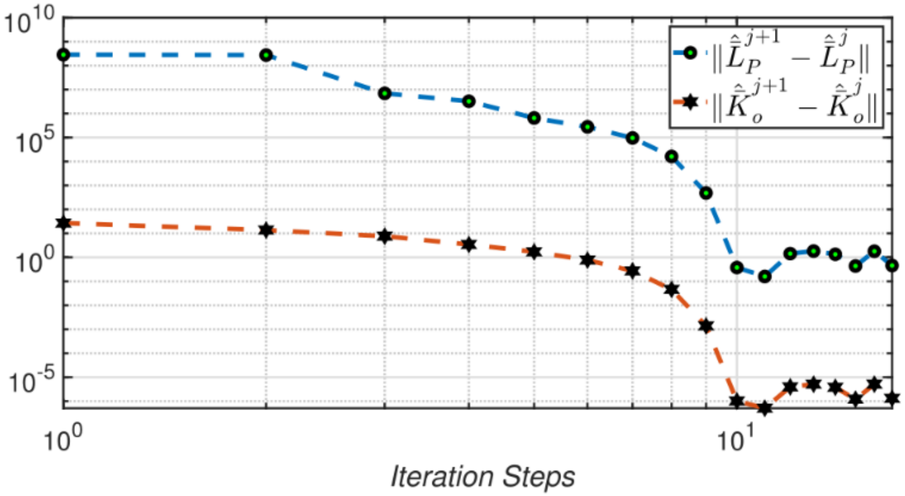
$$A = \begin{bmatrix} 0.887086 & -0.00423047 & -0.00281863 \\ 0.282727 & 0.999573 & -0.00043168 \\ 0.0284386 & 2.93148 & 0.994194 \end{bmatrix},$$

$$B = \begin{bmatrix} 1.77695 \\ 0.272081 \\ 2.82003 \end{bmatrix}, \quad C = \begin{bmatrix} 0 & 57.2958 & 0 \end{bmatrix}.$$

$\boldsymbol{x(1)}$ ***angle attack***; $\boldsymbol{x(3)}$ ***elevator actuator***;
$\boldsymbol{x(2)}$ ***pitch rate***; $\boldsymbol{y = x(2)}$ ***pitch rate***.

Reference dynamics

$$S = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \quad R = \begin{bmatrix} 1 & 0 \end{bmatrix}.$$



(b) $\|\hat{\bar{L}}_P^{j+1} - \bar{L}_P^{j}\|$ and $\|\hat{\bar{K}}_o^{j+1} - \bar{K}_o^{j}\|$

Fig. 3 Convergence of the learned control gain

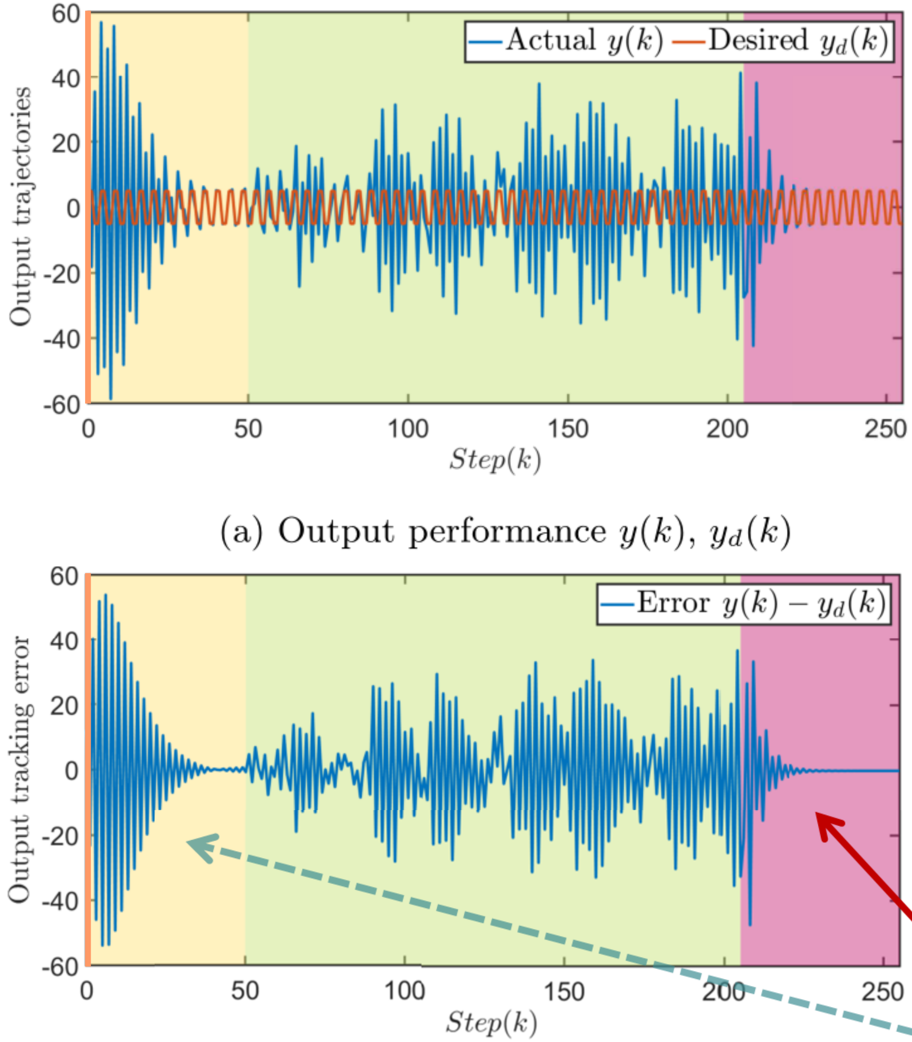
The initial stabilizing control gain as

$$\bar{K}_o^0 = \begin{bmatrix} 16.0669 & 20.11 & 20.3783 \\ 20.8936 & 3.93085 & 7.08518 \\ 19.9149 & 10.5774 & 4.29378 \\ -0.31044 & -8.03354 & -18.0887 \\ -12.2209 & -2.57785 & -0.414603 \end{bmatrix}$$

The learned optimal control gain as

$$\bar{K}_o^{10} = \begin{bmatrix} -3.14906 & -1.06988 & 0.068666 \\ -0.161422 & 2.73117 & -1.38856 \\ -2.63657 & 1.70597 & 1.54444 \\ 0.842932 & 1.57443 & 2.10917 \\ -1.08647 & -2.09949 & -0.47113 \end{bmatrix}$$

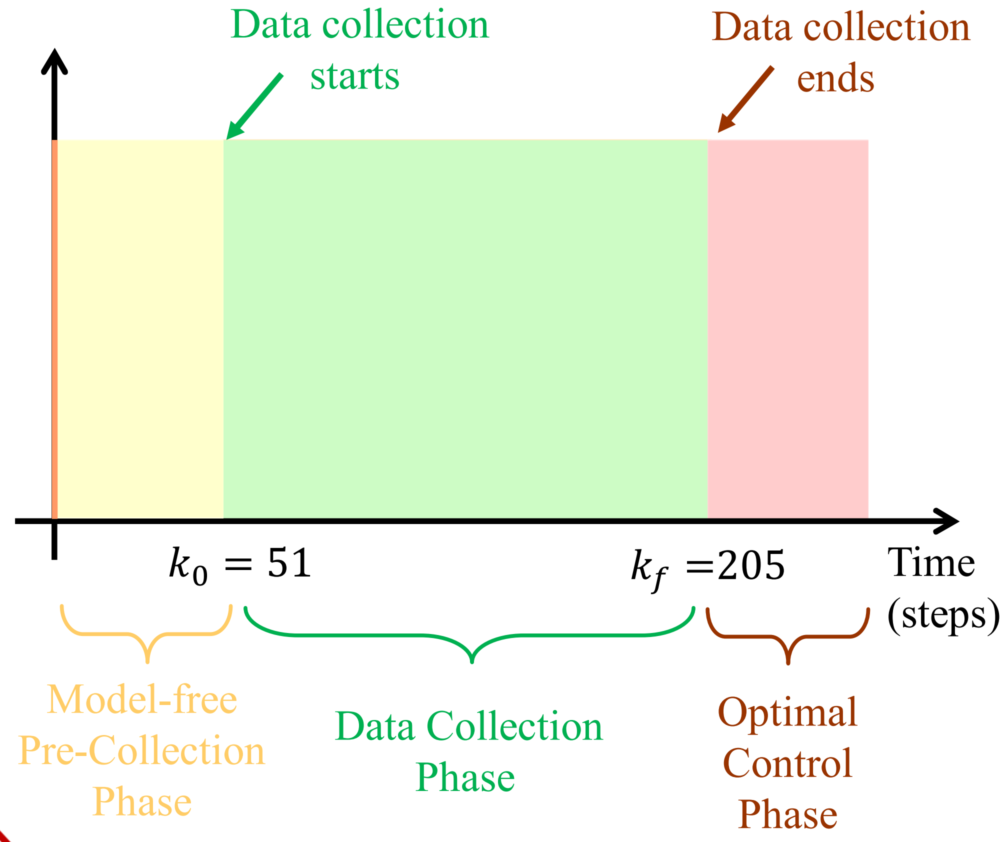
4. Simulation



(a) Output performance $y(k)$, $y_d(k)$

(b) Tracking error $y(k) - y_d(k)$

Fig. 4 Output trajectories over the time



Performance Improved

Content

- Background & Motivation
- Problem Formulation
- Method
- Simulation
- **Conclusion**

5. Conclusion

Adaptive Optimal Output Tracking for DT systems via Output-Feedback-Based Reinforcement Learning

1. **We proposed an output regulation and off-policy RL-based controller to formulate adaptive optimal output tracking problem for DT systems.**
2. **We derived a verifiable rank condition to ensure the uniqueness of the optimal control gain learned from input-output data.**
3. **We proposed a model-free pre-collection phase to supplement the off-policy learning for DT systems.**

Thank you

www.control.lth.se

LUND UNIVERSITY