Do We Need Models for Control? A System Theoretic Study of Data-driven Policy Iteration

Bowen Song and Andrea Iannelli

Institute for Systems Theory and Automatic Control, University of Stuttgart

Abstract

We investigate the role played by the **identification of a model** of unknown systems in **data-driven control**. Specifically, we apply **policy iteration** to the **linear quadratic regulator (LQR)** problem. We consider two iterative procedures to compute the optimal controller. In **indirect policy iteration (IPI)**, data collected from the system are leveraged to update model estimates via **recursive least squares (RLS)**. The estimates are subsequently employed for the model-based policy iteration. In **direct policy iteration (DPI)**, on-policy data are employed to directly approximate the value function and the associated controller. The goal is to analytically study the implications of an indirect and a direct scheme on the **sample complexity** and **convergence rate** of the two algorithms. Introducing the **identification of a model** offers advantages in **sample complexity and robustness** over a purely direct (model-free) approach. Finally, we show further insights into these two methods through numerical simulations.

Problem Setting and Policy Iteration

We consider linear time-invariant (LTI) systems:

$$x_{t+1} = Ax_t + Bu_t, \tag{1}$$

where $A \in \mathbb{R}^{n_x \times n_x}$, $B \in \mathbb{R}^{n_x \times n_u}$ are **unknown** and (A, B) is stabilizable. The objective is to design a state-feedback controller $u_t = Kx_t$ that minimizes the infinite horizon cost:

$$J(x_0, K) = \sum_{k=0}^{+\infty} r(x_k, u_k) = \sum_{k=0}^{+\infty} x_k^T Q x_k + u_k^T R u_k, \quad Q \succeq 0, R \succ 0 \tag{2}$$

If (A, B) is known and a stabilizing gain K_1 is given, a method based on dynamic programming to compute the optimal gain $K^\star$ is **policy iteration**:

- **Policy evaluation (PE)**: $K_i \to P_i$

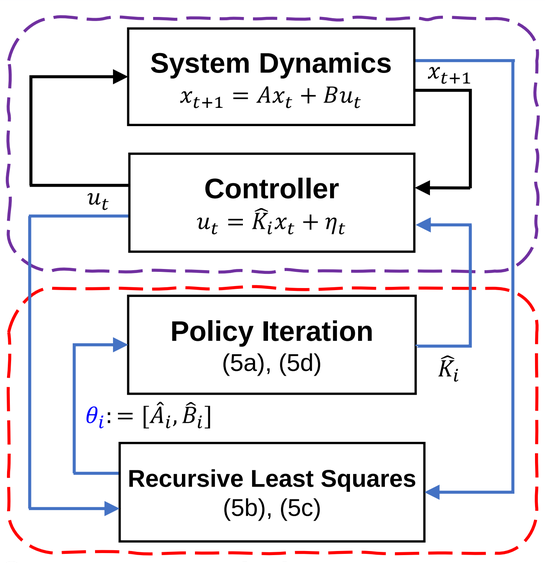
$$P_i = Q + K_i^T R K_i + (A + BK_i)^T P_i (A + BK_i). \tag{3}$$

- **Policy improvement (PI)**: $P_i \to K_{i+1}$

$$K_{i+1} = -(R + B^T P_i B)^{-1} B^T P_i A. \tag{4}$$

In this setting, policy iteration is guaranteed to converge: $\lim_{i\to\infty} P_i = P^*$ and $\lim_{i\to\infty} K_i = K^*$.

Indirect Policy Iteration (IPI): Data → Models → Controllers



Online data collection: Each episode, denoted by index i, comprises τ_{IPI} data points $\{d_t := [x_t^T, u_t^T]^T, x_{t+1}\}_{t=1}^{\tau_{\text{IPI}}}$ obtained with $u_t = \hat{K}_i x_t + \eta_t$, where $\eta_t \sim \mathcal{N}(0, \Sigma)$. $d^{(i)} := \{d_t\}_{t\geq 1}$ is the sequence of d_t within the i-th episode.

Concurrent learning and design:

$$\hat{P}_i \leftarrow \textbf{PE} \text{ based on } \theta_{i-1} = [\hat{A}_{i-1}, \hat{B}_{i-1}], \tag{5a}$$

$$H_i = H_{i-1} + \left(\sum_{t=1}^{\tau_{\text{IPI}}} d_t d_t^T\right), \tag{5b}$$

$$\theta_i = \left(\theta_{i-1} H_{i-1} + \sum_{t=1}^{\tau_{\text{IPI}}} x_{t+1} d_t^T\right) H_i^{-1}, \tag{5c}$$

$$\hat{K}_i \leftarrow \textbf{PI} \text{ based on } [\hat{A}_i, \hat{B}_i]. \tag{5d}$$

Convergence analysis:

Define $\theta := [A, B]$, estimation error $\Delta\theta_i := \theta_i - \theta$, estimation error upper bound $\|\Delta\theta_i\|_F \leq \|\Delta\theta_0 H_0\|_F \|H_i^{-1}\|_F =: \Delta\theta_i^{\text{Upper}}$ and the sequence $\Delta\theta^{\text{Upper}} := \{\Delta\theta_i^{\text{Upper}}\}_{i\geq 0}$.

Assumption 1. The estimated pairs $(\hat{A}_i, \hat{B}_i), i \in \mathbb{Z}_+$ obtained from recursive least squares (5b) and (5c), are all stabilizable.

Theorem 1. If Assumption 1 holds, for any $\tau_{\text{IPI}} \in \mathbb{Z}_+$ system (5) is input-to-state stable (**ISS**) with respect to $\Delta\theta^{\text{Upper}}$, i.e.,

$$\|\hat{P}_i - P^*\|_F \leq \beta(\|\hat{P}_0 - P^*\|, i) + \gamma(\|\Delta\theta^{\text{Upper}}\|_\infty), \tag{6}$$

where $\beta(\|\hat{P}_0 - P^*\|_F, i) = c_1^i \|\hat{P}_0 - P^*\|_F$ is a $\mathcal{KL}$ function and $\gamma(\|\Delta\theta^{\text{Upper}}\|_\infty) = \frac{c_2}{1-c_1}\|\Delta\theta^{\text{Upper}}\|_\infty$ is a $\mathcal{K}$ function with $c_1 \in (0, 1)$ and $c_2 > 0$.

Corollary 1. If the sequence $\{d^{(i)}\}_{i\geq 1}$ is persistent, then $\lim_{i\to\infty} \hat{P}_i = P^*$, $\lim_{i\to\infty} \hat{K}_i = K^*$.

Direct Policy Iteration (DPI): Data → Controllers

Online data collection: Each episode, denoted by index i, comprises τ_{DPI} data points $\{x_t, u_t, x_{t+1}\}_{t=1}^{\tau_{\text{DPI}}}$ obtained with $u_t = \hat{K}_i x_t + \eta_t$, where $\eta_t = \begin{cases} \epsilon_j, & t = 2j-1 \\ -\epsilon_j, & t = 2j \end{cases}, j \in \mathbb{Z}_+$, $\epsilon_j \sim \mathcal{N}(0, \Sigma)$.

PE: Use *data pairs* $\{\bar{\mathbf{x}}_\mathbf{j} := x_{2j-1} + x_{2j}, \hat{K}_i \bar{x}_j, \bar{\mathbf{x}}_\mathbf{j}^+ := x_{2j} + x_{2j+1}\}_{j=1}^{\frac{\tau_{\text{DPI}}}{2}}$ to estimate $\hat{\mathbf{P}}_\mathbf{i}$:

A model-free version of PE (3) can be expressed in terms of data pairs as:

$$(3) \Leftrightarrow \bar{\mathbf{x}}_\mathbf{j}^T \hat{\mathbf{P}}_\mathbf{i} \bar{\mathbf{x}}_\mathbf{j} = r(\bar{\mathbf{x}}_\mathbf{j}, \hat{K}_i \bar{\mathbf{x}}_\mathbf{j}) + (\bar{\mathbf{x}}_\mathbf{j}^+)^T \hat{\mathbf{P}}_\mathbf{i} \bar{\mathbf{x}}_\mathbf{j}^+. \tag{7}$$

PI: Use *data points* $\{x_t, u_t, x_{t+1}\}_{t=1}^{\tau_{\text{DPI}}}$ and $\hat{\mathbf{P}}_\mathbf{i}$ from PE to estimate $\widehat{\mathbf{B}^\mathbf{T}\mathbf{P}_\mathbf{i}\mathbf{B}}$ and $\widehat{\mathbf{B}^\mathbf{T}\mathbf{P}_\mathbf{i}\mathbf{A}}$:

$$(3) \Leftrightarrow x_t^T \hat{\mathbf{P}}_\mathbf{i} x_t = r(x_t, \hat{K}_i x_t) + (x_{t+1} - B\eta_t)^T \hat{\mathbf{P}}_\mathbf{i} (x_{t+1} - B\eta_t),$$

$$\begin{bmatrix} 2x_t \otimes \eta_t \\ vecv(u_t) - vecv(\hat{K}_i x_t) \end{bmatrix}^T \begin{bmatrix} vec(\widehat{\mathbf{B}^\mathbf{T}\mathbf{P}_\mathbf{i}\mathbf{A}}) \\ vecs(\widehat{\mathbf{B}^\mathbf{T}\mathbf{P}_\mathbf{i}\mathbf{B}}) \end{bmatrix} = r(x_t, \hat{K}_i x_t) - x_t^T \hat{\mathbf{P}}_\mathbf{i} x_t + x_{t+1}^T \hat{\mathbf{P}}_\mathbf{i} x_{t+1}. \tag{8}$$

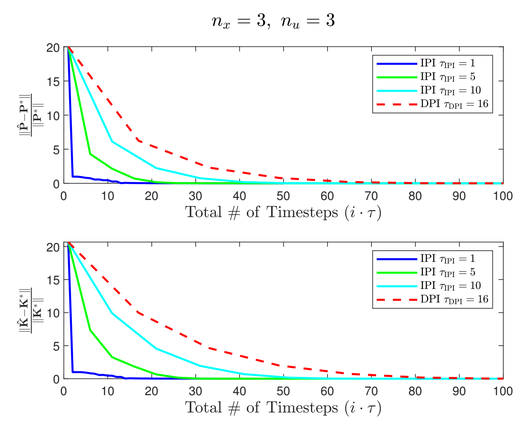
Then update $\hat{K}_{i+1}$ with (4).

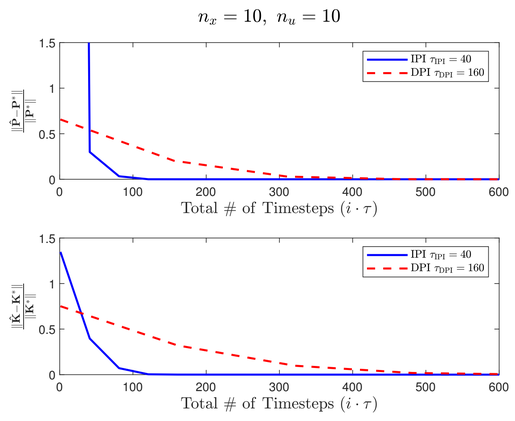
Convergence analysis:

When the minimum sample complexity in each episode is met (see below) and the data is persistently exciting, then both linear equations (7) and (8) can be solved and the estimations of $\hat{\mathbf{P}}_\mathbf{i}$, $\widehat{\mathbf{B}^\mathbf{T}\mathbf{P}_\mathbf{i}\mathbf{B}}$ and $\widehat{\mathbf{B}^\mathbf{T}\mathbf{P}_\mathbf{i}\mathbf{A}}$ are exact. Convergence follows from the case where system is known.

Comparison IPI vs DPI

- **Sample complexity of each episode:**
  - **IPI**: The theoretical minimum sample complexity is $\tau_{\text{IPI}} = 1$. PE and PI updates can be performed multiple times within a single episode.
  - **DPI**: All data points are leveraged for PE and also PI simultaneously. Solving (7) and (8) exactly requires at least $\tau_{\text{DPI}} = \max[n_x(n_x+1), \frac{n_u(n_u+1)}{2} + n_u n_x]$ data points.





- **Convergence rate:**
  - **IPI**:

$$\|\hat{P}_{i_{\text{IPI}}} - P^*\|_F \leq c_1^{i_{\text{IPI}}} \|\hat{P}_0 - P^*\|_F + \gamma(\|\Delta\theta^{\text{Upper}}\|_\infty). \tag{9}$$

  - **DPI**:

$$\|\hat{P}_{i_{\text{DPI}}} - P^*\|_F \leq c_1^{i_{\text{DPI}}} \|\hat{P}_0 - P^*\|_F. \tag{10}$$

Note: Because $\tau_{\text{DPI}} \gg \tau_{\text{IPI}}$, the total number of episodes i_{IPI} is larger than i_{DPI} for a fixed time budget.

- **Selection of excitation u_t:**
  - **IPI:** More flexible, only need to ensure the persistency requirement on $\{d^{(i)}\}_{i\geq 1}$.
  - **DPI:** Must be given in the form of $u_t = \hat{K}_i x_t + \eta_t$ (on-policy excitation).

Takeaways

- New quantitative **convergence guarantees** for indirect policy iteration
- **Advantages of identifying the model** in data-driven policy iteration
- A **system theoretic approach** to analyse **concurrent learning and control** methods

Future Works

- Study of the effect of **process and measurement noise**
- Dual control-inspired **selection of** u_t in indirect policy iteration
- Use of **discount factors** in the cost to relax Assumption 1

References

F. A. Yaghmaie, F. Gustafsson and L. Ljung, "Linear Quadratic Control Using Model-Free Reinforcement Learning," in *IEEE Transactions on Automatic Control*, 2023.

B. Pang, T. Bian and Z. -P. Jiang, "Robust Policy Iteration for Continuous-Time Linear Quadratic Regulation," in *IEEE Transactions on Automatic Control*, 2022.

B. Song, A. Iannelli, "Do We Need Models for Control? A System Theoretic Study of Data-driven Policy Iteration," in *arXiv*, coming soon.

Acknowledgements

Bowen Song acknowledges the support of the International Max Planck Research School for Intelligent Systems (IMPRS-IS).

Institute for Systems Theory and Automatic Control
University of Stuttgart
www.ist.uni-stuttgart.de

Contact:
Bowen Song
bowen.song@ist.uni-stuttgart.de

imprs-is

Universität Stuttgart